Universität zu Köln
Institut für theoretische Physik

Convex reconstruction from structured measurements

Inaugural-Disseration
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln
vorgelegt von

Richard Küng

aus Linz an der Donau

Köln, 2017

Berichterstatter (Erstgutachter):	David Gross
Berichterstatter (Zweitgutachter):	Johannes Berg
Drittgutachterin:	Gitta Kutyniok
Tag der Disputation:	13.12.2016

Contents

1 Introduction and summary of results

1.1 Introduction

1.1.1 The challenge of under-determined inverse problems

Inverse problems have a long-standing history in science. In its simplest form this data analysis problem requires inferring an n-dimensional vector $x \in \mathbb{C}^n$ from linear measurements of the form

$$y_k = \langle a_k, x \rangle. \tag{1.1}$$

Here, $a_1, \ldots, a_m \in \mathbb{C}^n$ denote the measurements and $y_1, \ldots, y_m$ the corresponding data acquired. If the measurements $a_1, \ldots, a_m$ span $\mathbb{C}^n$, this task is trivial: perform linear inversion.

This situation changes drastically if we consider an under-determined set of $m < n$ measurements. Problems of this type arise in many different areas of science, where the problem dimensions n are extremely high and/or massive data acquisition is challenging. Concrete examples include high resolution biomedical imaging, seismology, radio frequency analysis, quantum state/process estimation and many more. See for instance [Gro+10; HFY12; LDP07].

In general, such inverse problems do not have a unique solution. Additional assumptions are required to ensure uniqueness. *Compressibility* is one such assumption. In concrete applications it is often justified, or at least a justifiable approximation.

Sparsity is one of the simplest notions of compressibility. A vector $x \in \mathbb{C}^n$ is s-sparse, if it has s non-vanishing components with respect to a certain basis. In order to exploit such a model assumption, it makes sense to search for the sparsest vector $z \in \mathbb{C}^n$ that agrees with our measurements:

$$\begin{aligned} \text{minimize} \quad & \text{sparsity}(z) \\ \text{subject to} \quad & \langle a_k, z \rangle = y_k \quad 1 \leq k \leq m. \end{aligned} \tag{1.2}$$

However, such a constrained sparsity minimization is known to be NP-hard in general. In fact, there is not even an obvious heuristic algorithm for solving (1.2).

Low rank is another notion of compressibility, applicable to matrices. It may be viewed as a “non-commutative analogue” of sparsity: a matrix has rank r, if and only if its vector of singular values is r-sparse. Low rank matrix reconstruction problems have many applications. Examples include kernel-based learning methods [CST00], principal component analysis [Jol02], and quantum state/process estimation [Gro+10]. Moreover, it gained fame through the *Netflix-prize* of \$1 000 000. This was an open competition for the best algorithm to predict user ratings, based a small set of available ratings. The algorithm which won the prize in 2009 used a low rank model assumption [KB11]. However, once more we face the problem that constrained rank minimization is NP hard in general.

Another important instance of a challenging inverse problem is *phase retrieval*, see for instance [Wal63]. It occurs naturally in X-ray crystallography, astronomy, diffraction imaging—see for example [Mil90]. This problem will feature prominently in this work. Its discrete version asks for inferring a complex signal vector $x \in \mathbb{C}^n$ from m measurements of the form:

$$\tilde{y}_k = |\langle a_k, x \rangle| \quad 1 \leq k \leq m. \tag{1.3}$$

This problem is ill-posed, because all phase information is lost in the measurement process. If one had access to the complex phases ϕ_k of $\langle a_k, x \rangle$ this problem reduces to solving a linear system of equations:

$$\Phi \tilde{y} = Ax, \tag{1.4}$$

where $\Phi = \sum_{k=1}^m \phi_k e_k e_k^*$ and $A = \sum_{k=1}^m e_k a_k^*$ subsumes the measurement process. Here, $e_1, \ldots, e_m$ denotes the standard basis of $\mathbb{R}^m$. Crucially for phase retrieval, we do not know Φ in (1.4). One approach to recovering x is performing a least-squares minimization over both unknowns:

$$\underset{\Phi, x}{\text{minimize}} \quad \| \Phi \tilde{y} - Ax \|_2, \tag{1.5}$$

where $\Phi \in U(m)$ is unitary and diagonal in the standard basis and $x \in \mathbb{C}^n$. Problems of this type are non-convex, and, in fact, NP hard in general.

However, in contrast to, for instance, constrained sparsity minimization (1.2), there are heuristics for solving (1.5). One such heuristics is *alternating minimization*, see e.g [Fie82]. This is an iterative algorithm, where one alternates between keeping x fixed and minimizing Φ and, vice-versa: fixing Φ and optimizing over x. Very few theoretical guarantees regarding its performance are known. Nonetheless, alternating minimization algorithms are used in many applications, see for instance [MCKS99].

Given the importance of the problem and the lack of mathematical understanding, obtaining theoretical guarantees for phase retrieval is highly desirable. In order to do so, we will follow a different direction: Interpret phase retrieval as a particular instance of low rank matrix reconstruction.

1.1.2 Convex signal reconstruction

Convex signal reconstruction is a novel scientific discipline that allows for analyzing underdetermined inverse problems in a mathematically rigorous way. To this end, techniques from various branches of math are combined. Pioneering works include Refs. [Can+06; CR06; CRT06] by Candès, Romberg and Tao, as well as and Ref. [Don06] by Donoho. They show that a sparse vector $x \in \mathbb{C}^n$ may be reconstructed exactly from considerably fewer than n linear measurements of the form (1.1). For instance, a measurement process containing

$$m \geq Cs \log\left(\frac{n}{s}\right) \tag{1.6}$$

standard Gaussian measurement vectors suffices to reconstruct any s-sparse vector $x \in \mathbb{C}^n$ with high probability (w.h.p.). This *sampling rate m* turns out to be essentially tight. Moreover, the actual reconstruction can be achieved by performing

$$\begin{aligned} \underset{z \in \mathbb{C}^n}{\text{minimize}} \quad & \|z\|_{\ell_1} \\ \text{subject to} \quad & \langle a_k, z\rangle = y_k \quad 1 \leq k \leq m. \end{aligned} \tag{1.7}$$

This can be viewed as a convex relaxation of problem (1.2). It may be reformulated as a linear program, see e.g. [Bar02; BV04]. Hence, it is computationally tractable. Today, the idea of using a constrained ℓ_1-minimization to promote sparsity is known as *compressed sensing*. It has received considerable scientific attention over the past decade. We refer to [EK12; FR13] for an overview.

Subsequently, similar ideas have been used to address other important estimation problems. The reconstruction of low rank matrices $X \in M_{n\times n}$ is one of them. It has been shown that

$$\begin{aligned} \underset{Z \in M_{n\times n}}{\text{minimize}} \quad & \|Z\|_1 \\ \text{subject to} \quad & \operatorname{tr}(A_k Z) = y_k \quad 1 \leq k \leq m \end{aligned} \tag{1.8}$$

is a convex optimization problem that does promote low rank, see e.g. [FHB01]. Here, $\|Z\|_1$ denotes the *nuclear norm* of Z, that is the sum singular values: $\|Z\|_1 = \sum_{k=1}^n \sigma_k(Z)$. The nuclear norm may be viewed as a non-commutative analogue of the ℓ_1-norm: It is the ℓ_1-norm of the vector of singular values. Moreover, (1.8) may be re-phrased as a semidefinite program [Bar02; BV04] which assures computational tractability. Similarly to compressed sensing, one can prove that a number of

$$m \geq Crn \tag{1.9}$$

Gaussian measurements $y_k = \mathrm{tr}\left(A_k X\right)$ suffice to reconstruct any rank-r matrix w.h.p. [CP11b; FCRP08]. Note that it requires roughly rn parameters to describe a $n \times n$-matrix with rank r. From this perspective, the sampling rate (1.9) may be viewed as optimal.

Phase retrieval also admits a convex relaxation. To see this, we square the measurements in (1.3):

$$y_k := \tilde{y}_k^2 = |\langle a_k, x\rangle|^2 = \mathrm{tr}\left(a_k a_k^* x x^*\right) \tag{1.10}$$

These quadratic measurements are linear in the outer product $X = xx^*$ of $x \in \mathbb{C}^n$. In turn, an order of $\mathcal{O}\left(n^2\right)$ linearly independent measurements $A_k = a_k a_k^*$ allow for reconstructing $X = xx^*$ via linear inversion [BBCE09]. Knowledge of X then specifies $x \in \mathbb{C}^n$ up to a global phase. On first glance, a quadratic number of measurements seems necessary. Formula (1.10) reinterpretes the n-dimensional non-linear phase retrieval problem as a linear inverse problem on H_n—the n^2-dimensional real-valued vector space of hermitian $n \times n$-matrices. However, lifted phase retrieval does exhibit additional structure: the object of interest $X = xx^* \in H_n$ is guaranteed to have unit rank. In analogy to low rank matrix reconstruction, one may promote this key feature via minimizing the nuclear norm [CESV15]:

$$\begin{aligned} \underset{Z \in H_n}{\text{minimize}} \quad & \|Z\|_1 \\ \text{subject to} \quad & \mathrm{tr}\left(a_k a_k^* Z\right) = y_k \quad 1 \leq k \leq m. \end{aligned} \tag{1.11}$$

Following its inventors [CESV15; CSV13], we call this approach to phase retrieval *PhaseLift*. Subsequently, it was proven that a number of

$$m \geq Cn \tag{1.12}$$

measurements allows for reconstructing any $X = xx^*$ w.h.p., provided that each measurement $a_k \in \mathbb{C}^n$ is chosen uniformly from the complex unit sphere S^{n-1} [CL14]. Random Gaussian measurements also allow for drawing the same conclusion. We emphasize that this sampling rate (1.12) scales linearly in n, the actual problem dimension of phase retrieval.

Note that alternating minimization (1.5) and PhaseLift (1.11) are two very different approaches to the same problem. Alternating minimization is a heuristic for the “vector level”, where phase retrieval is a challenging non-convex problem. PhaseLift, on the other hand, solves an under-determined, but linear, inverse matrix problem by exploiting techniques from low rank matrix reconstruction.

In practical applications, the dimensionality n of phase retrieval is typically very large. For large problem dimensions, alternating minimization heuristics have a considerable lower runtime than PhaseLift, see for instance Table 1 in [NJS13]. Arguably, the merit of PhaseLift is more conceptual than practical: its convex structure allows for a rigorous mathematical analysis. In turn, theoretical reconstruction guarantees obtained via PhaseLift lend credence to commonly employed heuristics, in the sense that they highlight the problem's tractability.

1.1.3 Stability towards noise corruption

Let us now turn our attention to an important issue: noise corruption. In practical applications, linear measurements of the form (1.1) are affected by noise:

$$y_k = \langle a_k, x \rangle + \epsilon_k \quad 1 \leq k \leq m.$$

We measure the *strength* of such corruptions by the ℓ_2-norm of the noise vector $\epsilon = (\epsilon_1, \ldots, \epsilon_m)^T \in \mathbb{C}^m$ and assume that it is bounded by a known constant $\eta \geq \|\epsilon\|_{\ell_2}$. Beyond that, we shall make no further assumptions on ϵ. In particular, we do not require ϵ to be stochastic.

The convex optimization algorithms treated so far are ill-equipped to handle noisy measurements. Their constraints demand exact reproduction of the noise-corrupted measurements y_k. Having access to η allows for overcoming this issue by further relaxing the equality constraints. For instance,

$$\begin{aligned} \underset{z \in \mathbb{C}^n}{\text{minimize}} \quad & \|z\|_{\ell_1} \\ \text{subject to} \quad & \|Az - y\|_{\ell_2} \leq \eta \end{aligned} \tag{1.13}$$

is a noise-robust reformulation of (1.7). Here, $A = \sum_{k=1}^m e_k a_k^*$ subsumes the measurement process and $y = (y_1, \ldots, y_m)^T$ encompasses the acquired data. For $m = Cs \log\left(\frac{n}{s}\right)$ random Gaussian measurements $a_k \in \mathbb{C}^n$ one can prove w.h.p. that such a reconstruction is stable towards noise corruptions. For instance, Ref. [BDDW08] in conjunction with Ref. [C0̀8] assures that the minimizer $z^\sharp$ of (1.13) obeys

$$\left\| z^\sharp - x \right\|_{\ell_2} \leq C\eta.$$

Low rank matrix reconstruction from noisy measurements $y_k = \operatorname{tr}(A_k X) + \epsilon_k$ admits a similar relaxation:

$$\begin{aligned} \underset{Z \in M_{n \times n}}{\text{minimize}} \quad & \|Z\|_{\ell_1} \\ \text{subject to} \quad & \|\mathcal{A}(Z) - y\|_{\ell_2} \leq \eta. \end{aligned} \tag{1.14}$$

Here, $\mathcal{A} : M_{n \times n} \to \mathbb{R}^m$ denotes the measurement operator $\mathcal{A}(Z) = \sum_{k=1}^m e_k \operatorname{tr}(A_k Z)$. For

Gaussian measurement matrices $A_k \in M_{n\times n}$ this reconstruction is again stable w.h.p.:

$$\left\| Z^\sharp - X \right\|_2 \leq C\eta. \tag{1.15}$$

See, for instance, [CP11b; OMFH11] in conjunction with [RFP10]. Here $\|\cdot\|_2$ denotes the Frobenius norm on $M_{n\times n}$: $\|Z\|_2 = \sqrt{\operatorname{tr}(ZZ^*)}$.

1.2 The challenge of structured measurements

While these initial breakthroughs of convex signal reconstruction are truly remarkable, they do have drawbacks. All results mentioned above rely on measurements chosen randomly from "generic" distributions: Gaussian measurements for compressed sensing and matrix reconstruction, and measurements choosen uniformly from S^{n-1} for PhaseLift. The associated reconstruction guarantees only hold with high probability over the particular realizations of these measurements. This is undesirable for several reasons:

(i) While true with high probability, checking whether a concrete measurement instance does indeed allow for convex signal reconstruction is a hard task.

(ii) Relying on generic measurements obscures the specific properties of measurement ensembles that enable convex signal reconstruction.

(iii) Perhaps most importantly, the lack of any structure in generic measurements renders the task of practical implementations hard and, more often than not, even infeasible.

Identifying deterministic sets of highly structured measurement ensembles that allow for proving reconstruction guarantees deterministically would solve all these issues. Unfortunately, this seems to be an extremely hard task and is a major open problem. To this date, essentially all deterministic constructions of measurements are unsatisfactory, because they suffer from at least one of the following drawbacks: (i) they require a considerably larger number of measurements than random measurements: In compressed sensing, this deficit is known as the "quadratic bottleneck", because deterministic constructions require $m \geq Cs^2$ instead of (1.6), (ii) contrived and complicated structure of the measurements, and (iii) weak stability towards noise corruption, see e.g. [Kec15].

Acknowledging the hardness of such a task, we focus on a less ambitious and more realistic goal:

Central goal

Prove convex reconstruction guarantees for measurements that are chosen randomly from *small* and *structured* ensembles.

Ideally, such a compromise has two advantages: (i) the residual amount of randomness allows for employing strong probabilistic proof techniques and (ii) the ensemble's structure facilitates practical implementation.

For compressed sensing, the discrete Fourier basis

$$f_k = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathrm{e}^{-\frac{2\pi i k (j-1)}{n}} e_k \in \mathbb{C}^n, \quad 1 \leq k \leq n \tag{1.16}$$

was early on identified to fulfill this purpose [CRT06]. Fourier basis measurements occur naturally in many applications, where raw data acquisition happens in the Fourier domain. A prominent example for this feature is medical MRI imaging, see for instance [LDP07]. Also, different problems in wideband radio frequency signal analysis are of this form, see e.g. the motivation provided in [Can+06]. with respect to the standard basis, Fourier vectors (1.16) have full support and coefficients with constant modulus $\frac{1}{\sqrt{n}}$. This in turn implies that the *coherence parameter* [CP11a]

$$\mu := \max_{1 \leq j \leq n} \left| \langle e_j, f_k \rangle \right|^2 = \frac{1}{n} \quad \forall 1 \leq k \leq n. \tag{1.17}$$

of the Fourier basis is minimal. Intuitively, this incoherence assures that Fourier measurements are sufficiently "spread out", or "global". In compressed sensing, incoherence rules out the undesirable property that a measurement reveals too little information about the sparse vector of interest[1].

Pauli matrices allow for drawing similar conclusions [Gro11; Liu11] in low rank matrix reconstruction. There are four elementary Pauli matrices:

$$\sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{1.18}$$

In dimension $n = 2^d$, n^2 different Pauli matrices arise by taking all possible d-fold tensor products of (1.18). Such a construction can be generalized to dimensions that are not a power

[1] Let us consider the task of reconstructing a standard basis vector $x = e_j \in \mathbb{C}^n$ from discrete Fourier basis measurements as a concrete example. Then, its basis expansion $x = \sum_{k=1}^{N} \tilde{x}_k f_k$ with respect to the Fourier basis is guaranteed to have full support. In turn, *any* Fourier measurement $\langle f_k, x \rangle$ reveals "some" information about x, because $\tilde{x}_k \neq 0$. For some special cases, this intuition can be made precise via entropic uncertainty relations which we briefly introduce in the outlook-section.

of two, see e.g. Ref. [Gro06] and references therein. These n^2 Pauli matrices are hermitian and unitary. If we re-scale them by $1/\sqrt{n}$, the resulting matrices $W_1, \ldots, W_{n^2}$ form an orthonormal basis of H_n with respect to the Hilbert-Schmidt inner product $(X, Y) = \mathrm{tr}\,(X^* Y)$. Hoelder's inequality then implies

$$\nu := \max_{x \in S^{n-1}} |(xx^*, W_k)|^2 \leq \max_{x \in S^{n-1}} \|xx^*\|_1 \, \|W_k\|_\infty = \frac{1}{n} \quad \forall 1 \leq k \leq n. \tag{1.19}$$

Consequently, the re-scaled Pauli basis is incoherent with respect to any rank-one projector $xx^* \in H_n$. This may be viewed as a non-commutative analogue of the incoherence relation (1.17).

Such incoherence properties facilitate mathematical proofs. Nonetheless, these "derandomized" reconstruction guarantees [CRT06; Gro11; Liu11] typically require much higher technical efforts and deliver slightly weaker results.

For phase retrieval via PhaseLift, the task of identifying the "right" structural properties is more involved. Obviously, Pauli matrices are not applicable, while measuring the discrete Fourier basis does not provide sufficient information to recover phases. Moreover, PhaseLift has the interesting feature that the measurement matrices $A_k = a_k a_k^*$ are constrained to unit rank. If we normalize them to unit Frobenius norm ($\|A_k\|_2 = \|a_k\|_{\ell_2}^2 = 1$), this in turn implies

$$\nu = \max_{x \in S^{n-1}} |(xx^*, a_k a_k^*)|^2 = \max_{x \in S^{n-1}} |\langle x, a_k \rangle|^4 = 1 \quad \forall a_k \in S^{n-1}.$$

On the contrary to Fourier vectors (1.17) and Pauli matrices (1.19), these measurements can never be incoherent. This turns out to be a considerable technical obstacle. For measurements a_k chosen uniformly from S^{n-1} it may be overcome by proving $\left|\left(xx^*, a_k a_k^*\right)\right|^2 = \mathcal{O}\left(\frac{1}{n}\right)$ is true for any fixed $x \in S^{n-1}$ with extremely high probability [CL14; CSV13]. In turn, such a strong notion of "probabilistic incoherence" allows for employing proof techniques from low rank matrix reconstruction [Gro11]. However, this notion of probabilistic incoherence can only worsen, if we move on to smaller and less generic measurement ensembles. And it is not even clear what good candidates for such ensembles would be.

Addressing these open problems regarding PhaseLift is particularly important to this thesis. To this end, we formulate the following objectives:

(I) Identify specific properties of measurement ensembles that enable PhaseLift.

(II) Prove reconstruction guarantees for measurements chosen from such ensembles,

(III) Find concrete examples.

We shall treat these questions separately in the next three sections.

1.3 Spherical designs as a general purpose tool for de-randomization

In this section, we focus on the first objective: finding structural properties on measurement vectors that enable PhaseLift to succeed. As the number of measurements required for detecting lost phases is strictly larger than the signal space dimension n, one cannot expect that measuring a single orthonormal basis suffices. On the other hand, Candès *et al.* [CL14; CSV13] prove that $m = Cn$ measurements chosen uniformly from the complex unit sphere S^{n-1} do enable phase retrieval w.h.p.

The concept of *spherical t-designs* provides an interpolation between these extreme cases. Roughly speaking, a spherical t-design is a finite subset $\{w_1, \ldots, w_N\}$ of the complex unit sphere S^{n-1} in $\mathbb{C}^n$ with the following defining property: Sampling uniformly from this set reproduces the first $2t$ moments of the uniform distribution over S^{n-1}. Many equivalent definitions capture this property, the most explicit one being

$$\frac{1}{N} \sum_{k=1}^{N} \left(w_k w_k^* \right)^{\otimes t} = \int_{v \in S^{n-1}} \left(v v^* \right)^{\otimes t} \mathrm{d}v. \tag{1.20}$$

Here, $\otimes$ denotes the canonical tensor (Kronecker) product of matrices. Introduced in a seminal paper by Delsarte et al. [DGS77], spherical t-designs have since been studied in algebraic combinatorics [Sid99], coding theory [NRS01] and quantum information theory [AE07; RBKSC04; Sco06].

For $t = 1$, Def. (1.20) is equivalent to that of a tight frame. For larger t, they correspond to equally weighted *cubature formulas of the Grassmannian* $\mathcal{G}\left(1, \mathbb{C}^n\right)$ [DLHP05]. In this sense, they may be viewed as rank-one instances of *tight t-fusion frames* [BE13]. We refer to [EGK15] for a concise comparison between spherical designs and tight t-fusion frames. As t scales up, t-designs give better and better approximations to vectors distributed uniformly over S^{n-1}.

For phase retrieval, we already know that the structure of a tight frame ($t = 1$) alone is not sufficient, while S^{n-1}-uniform vectors ($t = \infty$) provably perform optimally [CL14]. Choosing the parameter t appropriately, allows us to interpolate between two extremes in a controlled way. We illustrate this intuition pictorially in Figure 1.1.

After having realized that demanding the structure of a 1-design alone is insufficient, a natural next step is to consider phase retrieval from spherical 2-designs. Interestingly, the defining property of such a set is almost equivalent to a prominent structural requirement in convex optimization: *isotropy*, see e.g. [CP11a]. A matrix-valued ensemble $A \in M_{n \times n}$ is

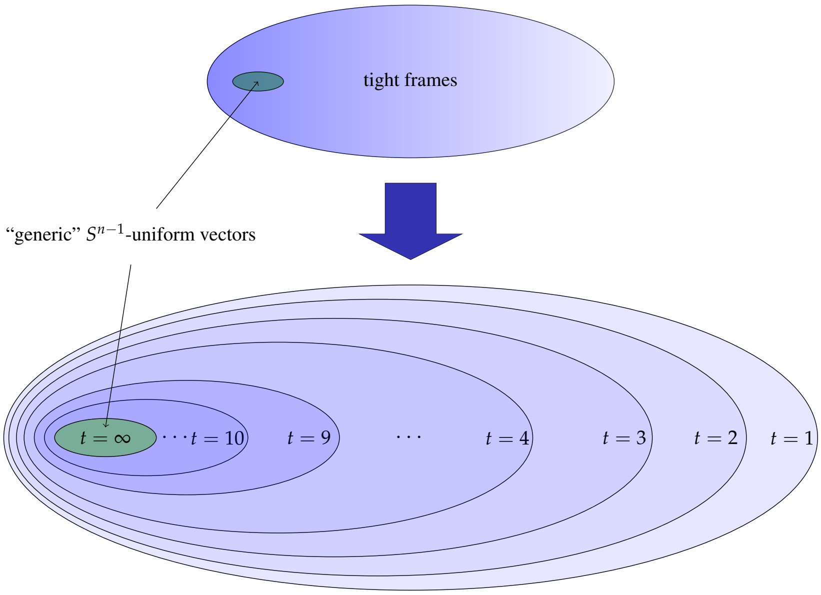


Figure 1.1: Caricature of the intuition behind spherical t-designs: the parameter t endows the set of all tight frames with a finer structure.

isotropic, if

$$\mathbb{E}\left[A \operatorname{tr}(A^* Z)\right] = Z \quad \forall Z \in M_{n \times n}. \tag{1.21}$$

This requirement, which is equivalent to the notion of a (matrix valued) tight frame, assures that choosing measurement matrices $A_1, \ldots, A_m$ uniformly at random from A results in a measurement process that is well-conditioned in expectation. While strict isotropy is in general not necessary, it does usually simplify mathematical proofs[2].

For PhaseLift, strict isotropy in the sense of (1.21) is impossible to attain. To see this, consider measurements chosen uniformly from S^{n-1}. Such an ensemble obeys

$$\mathbb{E}\left[A \operatorname{tr}(AZ)\right] = \mathbb{E}\left[aa^* \operatorname{tr}(aa^* Z)\right] \propto Z + \operatorname{tr}(Z)\mathbb{I} \quad \forall Z \in H_n, \tag{1.22}$$

see for instance Lemma 8 in [GKK15a]. And a similar relation is true for Gaussian measurement vectors $a_k \in \mathbb{C}^n$, see Eq. (4.1) in [CSV13]. The identity-term in (1.22) is unavoidable, because very phaseless measurement obeys $(A_k, \mathbb{I}) = \operatorname{tr}\left(a_k a_k^* \mathbb{I}\right) = \|a_k\|_{\ell_2}^2 > 0$.

Importantly, Formula (1.22) is equivalent to the demand that a is chosen uniformly from a spherical 2-design [AFZ15]. This highlights how well suited the notion of spherical 2-designs seems to be for analyzing phase retrieval [KGK15].

In fact, it was conjectured that Condition (1.22) suffices to assure non-trivial reconstruction

[2] Consider compressed sensing as an illustrative example: Candès and Plan [CP11a] have identified incoherence (1.17) and isotropy as sufficient assumptions to assure sparse reconstruction w.h.p. Subsequently, Gross and myself could show that isotropy is not a necessary assumption and may be further generalized [KG14].

results for phase retrieval [EK13]. However, this turns out to be not the case, see [GKK15a]. By means of a concrete counterexample, we show that choosing phaseless measurements uniformly from a spherical 2-design may require a total number of $m = \mathcal{O}(n^2)$ measurements in order to correctly distinguish two vectors $x, y \in \mathbb{C}^n$.

It is worthwhile to point out that the above no-go result does not exclude the possibility that certain realizations of 2-designs can perform better, if additional structural properties can be exploited. It states that solely demanding a 2-design structure is insufficient. We provide a concrete example for such a measurement process in subsection 1.5.2.

The applicability of spherical t-designs is by no means limited to the problem of phase retrieval. In [GKK15a], it has been one of the intentions of my co-authors and me to advertise spherical designs as a general-purpose tool for partially "de-randomizing" constructive results that initially relied on generic randomness. Already, this has partly come to fruition in [Kue15], where we apply this idea to a particular scenario of matrix reconstruction that takes into account typical features of quantum mechanical experiments.

1.4 Main results on phase retrieval

We are now in a position to describe the first main results obtained as part of this thesis. In this section, I summarize three papers that I have co-authored during my PhD. They provide increasingly tight and stable reconstruction guarantees for PhaseLift from spherical t-designs.

We know from the previous section that a minimal requirement for achieving this goal—without having to make further assumptions on the ensemble—is $t \geq 3$. An important first step was achieved by the following result:

Theorem 1 (Simplified version of Theorem 1 in [GKK15a])**.** *Fix $t \geq 3$ and $x \in \mathbb{C}^n$. Then, performing PhaseLift with*

$$m \geq Ctn^{1+\frac{2}{t}} \log^2(n)$$

measurements chosen uniformly from a spherical t-design allows for reconstructing x with high probability.

Ignoring logarithmic factors, this sampling rate m is proportional to $\mathcal{O}\left(n^{1+\frac{2}{t}}\right)$. Already for $t = 3$, this implies a sub-quadratic scaling which is non-trivial. If we allow the design order t to grow logarithmically with the problem dimension (as $t = 2\log(n)$), a sampling rate $m \geq Cn\log^3(n)$ suffices. Up to logarithmic factors, this scaling is optimal.

However, comparing this statement to the original result of Candès *et al.* reveals that the transition from S^{n-1}-uniform measurements to t-designs comes at a prize:

(i) *Non-optimal sampling rates:* The sampling rate m only becomes optimal (up to logarithmic factors), if we allow the design order t to grow with the problem dimension.

(ii) *Non-uniform reconstruction guarantee:* The result in [CL14] assures that a concrete realization of the measurements w.h.p. allows for reconstructing *any* unknown vector $x \in \mathbb{C}^n$ (*uniform reconstruction*). In contrast, Theorem 1 only promises that a concrete realization of the measurement process is w.h.p. capable of reconstructing a single vector $x \in \mathbb{C}^n$ (*non-uniform reconstruction*).

(iii) *No stability towards noise corruption:* The S^{n-1}-uniform result is stable towards noise corruption. Although highly plausible, in its current form Theorem 1 has no stable reformulation.

While certainly undesirable, drawbacks of this kind are typical for "de-randomizations" of reconstruction statements that initially relied on generic randomness. Moreover, the apparent trade-off between sampling rate m and design-order t seems to reflect our intuition about spherical t-designs: the degree t interpolates between "maximal structure" and "maximal randomness". In consideration of these facts, the next result should come as a surprise:

Theorem 2 (Simplified version of Theorem 2 in [KRT15])**.** *Fix* $1 \leq r \leq n$. *Then, with high probability*

$$m \geq Crn \log(n)$$

4-design measurements $A_k = a_k a_k^*$ *allow for reconstructing any hermitian rank-*r *matrix* $X \in H_n$ *via constrained nuclear norm minimization. This reconstruction is stable under additive noise corruption.*

This is actually a statement about matrix reconstruction. It is uniform in the sense of point (ii) discussed above: one randomly chosen measurement process w.h.p. suffices to reconstruct *any* hermitian rank-r matrix. It reduces to PhaseLift, if we set r to one. This special case overcomes all the drawbacks from Theorem 1. In particular, up to a single log-factor, $m = Cn\log(n)$ scales linearly—and thus optimally—in the problem dimension n. However, unlike Theorem 1, this statement does require a 4-design. Comparing this to the no-go result for 2-designs leaves open the behavior for $t = 3$. We will come back to this in Sec. 1.5.3.

Let us now turn our attention to PhaseLift, and more generally: matrix reconstruction from rank-one projective measurements, in the presence of noise:

$$y_k = \mathrm{tr}\left(a_k a_k^* X\right) + \epsilon_k, \quad \text{or} \quad y = \mathcal{A}(X) + \epsilon, \tag{1.23}$$

where $y, \epsilon \in \mathbb{R}^m$and $\mathcal{A}(Z) = \sum_{k=1}^m e_k \mathrm{tr}\left(a_k a_k^* Z\right)$. Theorem 2 implies stable reconstruction of any rank-r X from noisy measurements (1.23) via

$$\begin{aligned} \underset{Z \in H_n}{\text{minimize}} \quad & \|Z\|_1 \\ \text{subject to} \quad & \|\mathcal{A}(Z) - y\|_{\ell_2} \leq \eta. \end{aligned}$$

In analogy to the results introduced in Section 1.1.3, the minimizer $Z^\sharp$ of this optimization is guaranteed to obey

$$\left\| Z^\sharp - X \right\|_2 \leq C\sqrt{n(n+1)}\eta, \tag{1.24}$$

provided that $\eta \geq \|\epsilon\|_{\ell_2}$. If this parameter is chosen too small, the reconstruction proofs don't apply. Choosing it too large worsens the reconstruction quality (1.24) unnecessarily.

We overcome this issue in [KKRT16] by exploiting an additional structural constraint: Positive-semidefiniteness. PhaseLift re-interprets the task of inferring $x \in \mathbb{C}^n$ from phaseless measurements as a linear inverse problem on H_n: Reconstruct $X = xx^*$ from a particular family of linear measurements $A_k = a_k a_k^*$. Thus, both the matrix of interest and the measurements are positive semidefinite ($X, A_k \geq 0$) by definition.

Theorem 3 (Simplified version of Corollary 6 in[KKRT16])**.** *Fix $r \leq n$ and $1 \leq p \leq \infty$. A number of $m \geq Crn\log(n)$ noisy 4-design measurements* (1.23) *w.h.p. allows for approximating any positive semidefinite matrix $X \in H_n$ with rank at most r via solving*

$$\underset{Z \geq 0}{\textit{minimize}} \quad \|\mathcal{A}(Z) - y\|_{\ell_p}. \tag{1.25}$$

The resulting minimizer $Z^\sharp$ obeys

$$\left\| Z^\sharp - X \right\|_2 \leq C'\sqrt{n(n+1)}\frac{\|\epsilon\|_{\ell_p}}{m^{1-\frac{1}{p}}}. \tag{1.26}$$

Note that this reconstruction guarantee depends on the true noise strength $\|\epsilon\|_{\ell_p}$, rather than on an upper bound η that needs to be guessed in advance. Also, the additional freedom of choosing $1 \leq p \leq \infty$ allows for adjusting reconstructions to the expected noise type. For instance, it may be advantageous to choose $p = 1$ for Poisson noise and $p = \infty$ for quantization errors.

We also point out that the dimensional pre-factors $\sqrt{(n+1)n} \simeq n$ in (1.24) and (1.26) are due to normalization. By definition, 4-design vectors have unit norm and hence $\|A_k\|_2 = \|a_k\|_{\ell_2}^4 = 1$. This is not the case for other "typical" measurements. For instance, the Frobenius norm a random Gaussian $n \times n$ matrix amounts to roughly n. If we re-scale the 4-design measurements by $\sqrt{(n+1)n} \simeq n$, the dimensional factors in (1.26) and (1.24) vanish.

Finally, we want to point out that Theorem 2 and Theorem 3 remain valid, if we replace

4-design measurements with complex standard Gaussian measurements. In fact, a sampling rate of

$$m \geq Crn$$

suffices for rank-one Gaussian measurements [KKRT16; KRT15]. This may be viewed as a generalization of uniform PhaseLift [CL14] to matrix reconstruction. Measurements of this form admit an interpretation as quantum mechanical measurements. Interestingly, for a brief period of time, these results provided actually the best possible known bounds on "sample complexity" for quantum state estimation [Haa+15; OW15]. Refs. [Haa+15; OW15] gave improved constructions—however at the expense of having to employ so-called "coherent measurements across samples", which are seen as more demanding to implement physically. We defer the interested reader to these references for a precise definition of the terms used here.

1.5 Concrete realizations of structured spherical designs

Our results from the previous section highlight that measurement vectors chosen uniformly from S^{n-1} are not required for performing phase retrieval via PhaseLift. However, the practical relevance of these statements hinges on the availability of explicit constructions.

Explicit constructions of spherical designs are known for any degree t and any dimension n. However, these constructions are typically inefficient, in the sense that they require an exponentially large number of vectors, see e.g. [HHH05; SZ84]. Moreover, these constructions typically lack the type of structure that would be important in practical applications.

1.5.1 Approximate spherical designs and randomized constructions

While tight and "structured" 2-designs are widely known [KR05; Kön99; Sch60; Zau99], tighter analytic designs for $t \geq 3$ are notoriously difficult to find. This lack of candidates may be overcome by relaxing the definition of a t-design:

(i) Allow for non-uniform weights $p_k \neq \frac{1}{N}$ in (1.20). Doing so, results in *weighted spherical t-designs* $\{p_k, w_k\}_{k=1}^N$. These are also known as *cubatures of strength* t, see e.g. [EGK15]. Constructions for such sets containing only $\mathcal{O}\left(n^{2t}\right)$ vectors are available [Kup06].

(ii) *Approximate spherical designs* $\{p_k, w_k\}_{k=1}^N$ arise, if one relaxes strict equality in (1.20) to closeness in some norm. Typically, Schatten p-norms $\|\cdot\|_p$ are used to measure the inaccuracy θ_p of a given relaxation:

$$\left\| \sum_{k=1}^{N} p_k \left(w_k w_k^* \right)^{\otimes t} - \int_{v \in S^{n-1}} \left(v v^* \right)^{\otimes t} \mathrm{d}v \right\|_p \leq \theta_p.$$

Such relaxations are well-established in quantum information science and randomized constructions do exist [AE07; BHH12]. We refer to [KRT15] for further information.

Theorem 4 (Simplified version of Theorem 28 in [KKRT16])**.** *The assertions of Theorem 2 and Theorem 3 remain true, if one chooses measurements from a weighted, approximate 4-design with accuracy $\theta_\infty \leq \frac{1}{16r^2}$, or $\theta_1 \leq \frac{1}{4}$. For weighted designs $\{p_k, w_k\}_{k=1}^N$, uniform sampling must be replaced by choosing measurement vectors independently according to the weights p_k.*

While we have not explicitly done the calculations, it is plausible that Theorem 1 also remains true for approximate, weighted t-designs.

1.5.2 Coded diffraction patterns

Coded diffraction patterns are a simplified model of techniques used in diffraction imaging. There, the phase retrieval problem arises naturally, because detectors can only capture light intensities, not phases. A typical diffraction imaging experiment aims at identifying the structure of a microscopic probe, for instance a protein. To this end, the probe is illuminated by coherent X-ray light. The resulting diffraction pattern is then recorded at detectors, or a photographic plate. Fresnel and Fraunhofer approximations to the diffraction equation often allow for relating microscopic features of the probe to its diffraction pattern via a 2D-Fourier transform. However, observing the absolute values of a single Fourier transform is insufficient to recover phases.

To overcome this, one typically repeats this process under different physical conditions. Conceptually, one of the simplest examples for such a procedure is *masked illumination*: one inserts different masks, or phase plates, between the sample and the recording screen, see e.g. [Liu+08] and Figure 1.2 for an illustration. Alternative techniques for achieving similar goals are well-established and we refer to [CESV15] for a concise overview. To illustrate how important these problems are in practice, we note that Watson and Crick used vital information from such diffraction patterns to identify the double-helix structure of DNA.

Motivated by such procedures, Candès, Li and Soltanolkotabi [CLS15] introduced the following measurement model for discrete phase retrieval: They describe modulations (via masks, or otherwise) by random matrices $D_l \in M_{n \times n}$ that are diagonal with respect to the

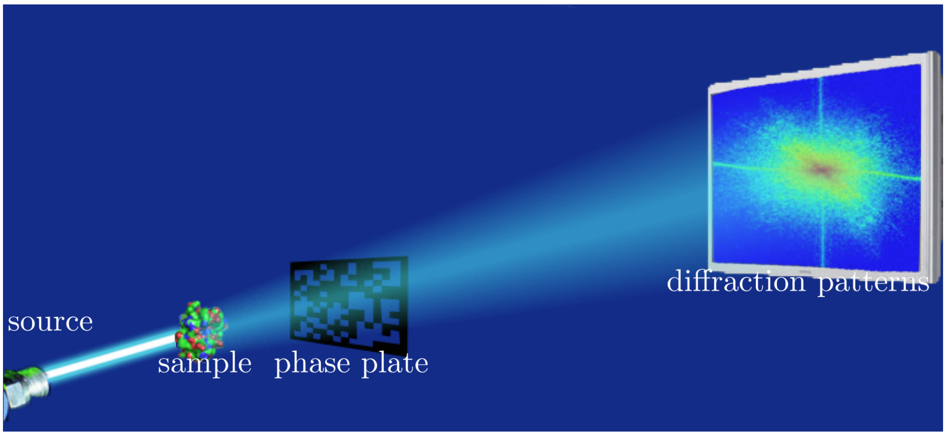


Figure 1.2: Caricature of a typical masked illumination experiment (courtesy of M.Šoltanolkotabi (CLS15)).

standard basis: $D_l = \sum_{k=1}^{n} d_k^{(l)} e_k e_k^*$. In turn, they approximate diffraction patterns by measuring all inner products with discrete Fourier vectors. Let $x \in \mathbb{C}^n$ be a vector which may carry important information about the microscopic structure of a probe. Then, this model associates n measurements

$$y_{k,l} = |\langle f_k, D_l x \rangle|^2 = \operatorname{tr}\left(D_l f_k f_k^* D_l^* x x^* \right) \quad 1 \leq k \leq n \tag{1.27}$$

with the l-th modulated diffraction pattern. Following [CLS15], we call one such measurements a *coded diffraction pattern*. Note that—even if D_l is random—the n different measurements (1.27) exhibit a high degree of structure in that each measurement vector is similar to a Fourier vector:

$$D_l f_k = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} b_k^{(l)} \omega^{jk} e_k \quad \text{with} \quad \omega = \mathrm{e}^{\frac{2\pi i}{n}}.$$

Also, they are correlated in the sense that the same random numbers $b_1^{(l)}, \ldots, b_n^{(l)}$ feature in every $D_l f_k$, $1 \leq k \leq n$. Soltanolkotabi *et al.* could show that for certain random models of D_l,

$$L = C \log^4(n)$$

independent coded diffraction patterns allow for recovering a fixed $x \in \mathbb{C}^n$ with high probability. This amounts to a total sampling rate of $m = Ln = Cn \log^4(n)$.

Their random model assumes that each $d_k^{(l)}$ is an independent instances of a bounded ($|d| \leq c$ almost surely), symmetric random variable $d \in \mathbb{C}$ obeying $\mathbb{E}\left[d^2\right] = 0$ and

$$\mathbb{E}\left[|d|^4\right] = 2\mathbb{E}\left[|d|^2\right]^2. \tag{1.28}$$

A concrete example [CLS15] for a complex-valued random variable that fulfills these properties is $d = b_1 b_2$, where b_1 and b_2 are independent and distributed as

$$b_1 = \begin{cases} 1 & \text{with prob. } \frac{1}{4} \\ -1 & \text{with prob. } \frac{1}{4} \\ -i & \text{with prob. } \frac{1}{4} \\ i & \text{with prob. } \frac{1}{4} \end{cases} \quad \text{and} \quad b_2 = \begin{cases} 1 & \text{with prob. } \frac{4}{5} \\ \sqrt{6} & \text{with prob. } \frac{1}{5} \end{cases}.$$

One key ingredient of proving such a statement is the fact that the b_k's are centered. This allows for applying Hoeffding's inequality [Hoe63] in order to conclude

$$|\langle f_k, D_l x \rangle|^2 = \frac{1}{n} \left| \sum_{j=1}^n b_j x_j \omega^{jk} \right|^2 \leq \frac{C \log(n)}{n}$$

for any fixed $x = \sum_{j=1}^n x_j e_j \in \mathbb{C}^n$ with very high probability. This is a rather strong notion of probabilistic incoherence.

A few months after this original paper appeared on the pre-print server, we succeeded in improving this statement [GKK15b]. In particular, we managed to further reduce the required sampling rate to

$$L = C \log^2(n).$$

This is close to optimal. Indeed, we also established a converse lower bound: $C' \log(n)$ such coded diffraction patterns are necessary to guarantee injectivity [GKK15b].

Also, as a minor improvement, we drop their "simplifying assumption" that d must obey $\mathbb{E}\left[d^2\right] = 0$. Instead, we require d to be a bounded, real-valued random variable obeying (1.28), as well as $\mathbb{E}\left[d\right] = \mathbb{E}\left[d^3\right] = 0$. A particular example for a random variable fulfilling all these requirements is

$$d \sim \begin{cases} \sqrt{2} & \text{with prob. } 1/4, \\ 0 & \text{with prob. } 1/2, \\ -\sqrt{2} & \text{with prob. } 1/4. \end{cases}$$

Finally, we point out that the moment condition (1.28) together with $\mathbb{E}\left[d\right] = 0$ is equivalent to near-isotropy (1.22) of the measurement model. This in turn implies that these coded diffraction patterns form a spherical 2-design, albeit a very particular one. The additional structural properties of this design allow for establishing close-to-optimal reconstruction proofs for PhaseLift.

1.5.3 Stabilizer states

In this section we consider *stabilizer states*—a ubiqutous tool in quantum information theory [Got97; NC10]. They also feature prominently in discrete Weyl-Heisenberg theory, where they correspond to the smallest orbit of the *oscillator group*. Real-valued versions of stabilizer states arise as generators of Barnes-Wall lattices [NRS02] and have been studied extensively in coding theory, see e.g. [NRS06].

In a sense made precise below, stabilizer states can be viewed as a generalization of discrete Fourier vectors. Here, we will introduce them from a quantum information perspective. Stabilizer states are joint eigenvectors of Pauli matrices (1.18). For the particular case of a single qubit ($n = 2$), they form a set $\mathrm{Stab}(2)$ of six normalized vectors:

$$\begin{aligned} s_1 =& e_1,\ s_2 = e_2,\ s_3 = \frac{1}{\sqrt{2}}\left(e_1 + e_2\right),\ s_4 = \frac{1}{\sqrt{2}}\left(e_1 - e_2\right), \\ s_5 =& \frac{1}{\sqrt{2}}\left(e_1 + ie_2\right),\ s_6 = \frac{1}{\sqrt{2}}\left(e_1 - ie_2\right). \end{aligned} \tag{1.29}$$

This is a union of three orthonormal bases which contains both the standard basis (s_1, s_2) and the Fourier basis (s_3, s_4) of $\mathbb{C}^2$. Note that s_1 is the unique joint eigenvector of σ_0 and σ_3 with eigenvalue +1. Also, $\sigma_0 = \mathbb{I}$ and σ_3 commute. Likewise, s_2 is the unique joint +1-eigenvector of the commuting matrices σ_0 and $-\sigma_3$. The remaining stabilizer states $s_3, \ldots, s_6 \in \mathbb{C}^2$ admit a similar unique description.

Such a definition of stabilizer states can be generalized to arbitrary dimensions. However, for the sake of brevity, we shall restrict ourselves to power-of-two dimensions $n = 2^d$. It is useful to introduce the following notation. Let us re-label the elementary Pauli matrices (1.18) in the following way:

$$\sigma_{(0,0)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \sigma_{(0,1)} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ \sigma_{(1,1)} = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix},\ \sigma_{(1,0)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This notation allows us to identify every 2×2 Pauli matrix with a 2-dimensional vector $(p_1, q_1) \in \mathbb{F}_2^2$. Likewise, in dimension $n = 2^d$, every Pauli matrix is uniquely specified by a $2d$-dimensional vector $(p, q) := (p_1, \ldots, p_d; q_1, \ldots, q_d) \in \mathbb{F}_2^{2d}$ of length $2d$:

$$W(p,q) = W(p_1, \ldots, p_n; q_1, \ldots q_n) = \sigma_{(p_1,q_1)} \otimes \cdots \otimes \sigma_{(p_d,q_d)}.$$

Such a description turns out to be extremely useful. For instance, two Pauli matrices $W(p,q), W(p',q') \in H_n$ commute, if and only if the *symplectic inner product* of their description vanishes:

$$[(p;q),(p';q')] := \langle p,q'\rangle - \langle q,p'\rangle = \sum_{k=1}^{d} p_k q_k' - \sum_{k=1}^{d} q_k p_k' = 0. \tag{1.30}$$

The vector space $\mathbb{F}_2^{2d}$ together with the non-degenerate symplectic product (1.30) is called *phase space* due to its resemblance to the phase space appearing in classical mechanics.

In turn, a set of $n = 2^d$ commuting $n \times n$-Pauli matrices corresponds to a n-dimensional subspace $M \subset \mathbb{F}_2^{2d}$ that is *isotropic*: $[(p;q),(p';q')] = 0 \; \forall (p;q),(p';q') \in M$. This observation allows us to generalize the definition of stabilizer states from $\mathbb{C}^2$ to $\mathbb{C}^n$ [Got97]. In the language adopted here and in [KG15] we obtain the following description:

Theorem 5. *Let $n = 2^d$ be a power of two. Then, up to a global phase, every stabilizer state $s \in \mathrm{Stab}(n) \subset \mathbb{C}^n$ is specified by a vector $(v;w) \in \mathbb{F}_2^{2d}$ and a d-dimensional isotropic subspace $M \subset \mathbb{F}_2^{2d}$:*

$$ss^* = \frac{1}{n} \sum_{(p;q)\in M} (-1)^{[(v;w),(p,q)]} W(p;q).$$

We emphasize, that Theorem 5 allows for a succinct description of every n-dimensional stabilizer state in terms of at most $2\left(\log_2^2(n) + \log_2(n)\right)$ bits. In turn, the "low complexity" of stabilizer states allows for generating them algorithmically with relative ease. To this end, let us label the standard basis in $\mathbb{C}^n$ by $y \in \mathbb{F}_2^d$ (each e_k is specified by the binary representation of $1 \leq k \leq n$). In turn, every stabilizer state $s \in \mathbb{C}^n$ is uniquely specified by an affine subspace $S + t \subset \mathbb{F}_2^d$, a vector $l \in \mathbb{F}_2^d$ and a quadratic form $q : \mathbb{F}_2^d \to \mathbb{F}_2$:

$$s = \frac{1}{\sqrt{|S|}} \sum_{y \in S} i^{\langle l,y\rangle} (-1)^{q(y)} \mathrm{e}_y, \tag{1.31}$$

see [DDM03] and also Theorem 5 in [GN07]. Moreover, there is a one-to-one correspondence between this triple $(S+t,l,q)$ and $(M,(v;w))$ from Theorem 5.

If we set $l = 0 \in \mathbb{F}_2^d$, $S = \{0\}$ and $q(y) = 0 \; \forall y \in \mathbb{F}_2^d$, we recover the standard basis: $s_t = e_t \; \forall t \in \mathbb{F}_2^d$. Conversely, if we choose $S = \mathbb{F}_2^n$, $t \in \mathbb{F}_2^d$ becomes irrelevant and setting $l = 0$ and $q(y) = \langle k,y\rangle$ with $k \in \mathbb{F}_2^d$ results in the discrete Fourier basis over $\mathbb{Z}_2^{\times d}$:

$$s_k = \frac{1}{\sqrt{n}} \sum_{y \in \mathbb{F}_2^n}^{n} (-1)^{\langle k,y\rangle} e_y = f_{k_1} \otimes \cdots \otimes f_{k_d},$$

where $f_0 = \frac{1}{\sqrt{2}}(e_1 + e_2)^T$ and $f_1 = \frac{1}{\sqrt{2}}(e_1 - e_2)^T$. In this sense, stabilizer states are a generalization of both standard basis and Fourier basis.

The standard basis description (1.31) of stabilizer states in particular allows for generating random stabilizer states efficiently. We have used such algorithms in different numerical experiments, see for instance Figure 1.3.

Beyond that, multi-qubit ($n = 2^d$) stabilizer states exhibit structural properties similar to their single-qubit counterpart (1.29). The set $\mathrm{Stab}(n) \subset \mathbb{C}^n$ of all stabilizer states is a union of

$$\frac{|\mathrm{Stab}(n)|}{n} = \prod_{j=1}^{d} \left(2^j + 1\right) = \mathcal{O}\left(2^{\frac{1}{2}d^2}\right)$$

different orthonormal bases. Each basis is uniquely determined by a d-dimensional isotropic subspace $M \subseteq \mathbb{F}_2^{2d}$, while different vectors $(v;w) \in \mathbb{F}_2^{2d}$ single out the individual basis vectors. As pointed out above, the standard basis and the discrete Fourier basis over $\mathbb{Z}_2^{\times d}$ are two particular instances of these bases.

Using the rich geometric structure of stabilizer states, we were able to prove the following statement:

Theorem 6 (Simplified version of Corollary 1 in [KG15])**.** *Let $n = 2^d$ be a power of two. Then, the set $\mathrm{Stab}(n) \subseteq \mathbb{C}^n$ of all stabilizer states forms a spherical 3-design. They do, however, not constitute a spherical 4-design.*

This statement is wrong for dimensions n that are not a power of two. Sidelnikov could prove an analogous statement for real-valued stabilizer states [Sid99] which also requires power-of-two dimensions. However, the structure of real-valued spherical designs is surprisingly different from their complex-valued counterparts. In turn, it is not obvious how to generalize the techniques from Sidelnikov to the complex case and our proof technique [KG15] is completely different from [Sid99].

Theorem 6 assures that the set of all stabilizer states obeys the requirements of Theorem 1—our first main result for phase retrieval via PhaseLift. Said result assures that a fixed $x \in \mathbb{C}^n$ may be reconstructed from

$$m \geq C n^{\frac{5}{3}} \log^2(n) \tag{1.32}$$

random stabilizer state measurements w.h.p While non-trivial, the sampling rate (1.34) is far from being optimal. Very recently, in a fruitful collaboration with Zhu, Gross and Grassl we were able to considerably improve this statement. These results are not yet published and we present parts of them in Chapter 3 below. The key idea is to approach stabilizer states via their symmetry group. Stabilizer states are the smallest orbit of the *Clifford group* $C(n) \subset U(n)$:

$$\mathrm{Stab}(n) = \{Ce_1 : \ C \in C(n)\}.$$

The Clifford group is defined as the group of operations that—up to phase factors—map Pauli matrices onto themselves under conjugation. It arises naturally in quantum information. For instance, the important field of *quantum error correction* relies practically exclusively on constructions that arise from Pauli matrices and Clifford actions. We refer to [LB13] and references therein for further information.

This symmetry group features prominently in different fields: For instance it is known as the *oscillator group* in finite Weyl-Heisenberg analysis and the metaplectic representation of $\mathrm{Sp}\left(\mathbb{F}_2, d\right)$ in mathematical physics, see e.g. [Fol16].

In order to improve (1.32), we fully characterized the irreducible representations of the diagonal representation $C \mapsto C^{\otimes 4}$ of the Clifford group. I want to emphasize that this characterization is mainly due to my collaborators, in particular Zhu and Gross. Since stabilizer states are an orbit of the Clifford group, this result allowed us to conclude the following formula for stabilizer states [ZKGG16]:

$$\frac{1}{|\mathrm{Stab}(n)|} \sum_{s \in \mathrm{Stab}(n)} (ss^*)^{\otimes 4} = \binom{n+2}{3} \left(P_1 + \frac{4}{n+4} P_2 \right). \tag{1.33}$$

Here, $P_1, P_2 \in H_n^{\otimes 4}$ denote orthogonal projections that obey $P_1 + P_2 = P_{\mathrm{Sym}^4}$, where P_{Sym^4} denotes the projector onto the totally symmetric subspace of $(\mathbb{C}^n)^{\otimes 4}$. We refer to Section 3.2 in [GKK15a] for a precise definition. This precise knowledge of the fourth moments of stabilizer states allowed us to apply proof techniques similar to [KRT15] and [KKRT16] and establish the following statement:

Theorem 7 (Simplified version of Theorems 2 and 3 in [KZG16b])**.** *Let $n = 2^d$ be a power of two and fix $1 \leq r \leq n$. Then, w.h.p.*

$$m = C r^3 n \log(n) \tag{1.34}$$

random stabilizer measurements $A_k = a_k a_k^*$ allow for reconstructing any rank-r matrix $X \in H_n$ via constrained nuclear norm minimization. This reconstruction is stable towards noise corruptions. If X is in addition positive semidefinite, noise robust reconstruction may be done by solving

$$\underset{Z \geq 0}{\textit{minimize}} \quad \|\mathcal{A}(Z) - y\|_{\ell_p} \quad \forall 1 \leq p \leq \infty.$$

We point out that the sampling rate (1.34) is cubic in the rank parameter r. For phase retrieval via PhaseLift this non-linearity is irrelevant, because every matrix of interest $X = xx^* \geq 0$ is proportional to a rank one projector ($r = 1$). In turn, Theorem 7 reproduces the strongest PhaseLift reconstruction statement available to date [CL14] up to a single log-factor. We emphasize that, unlike measurement vectors chosen uniformly from S^{n0-1} [CL14], stabilizer states have an exceedingly rich structure in the following sense:

(i) They admit a concise description in terms of finite symplectic geometry.

(ii) They form the smallest orbit of a big and well-studied symmetry group—the Clifford

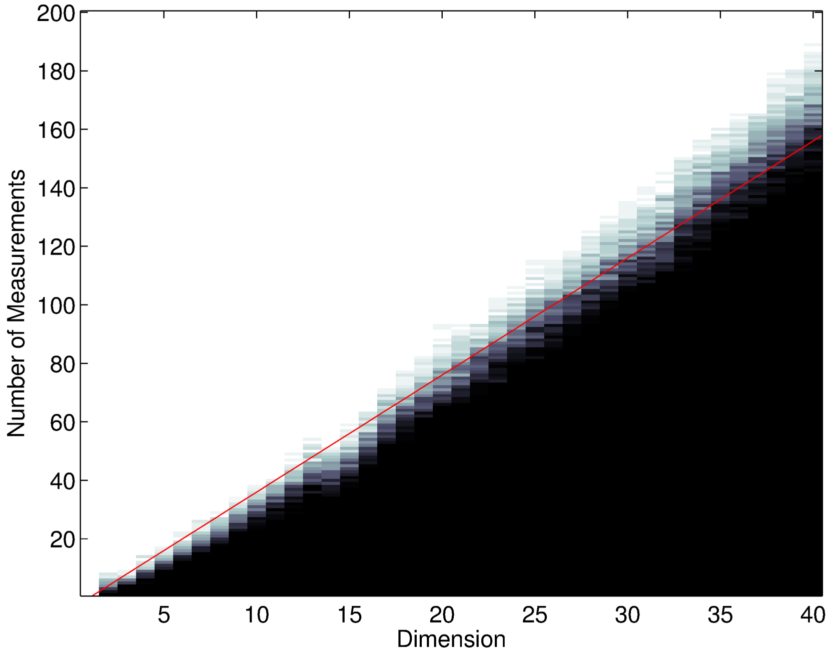


Figure 1.3: Phase Diagram for PhaseLift from (projected) stabilizer states. The red line indicates $m = 4d - 4$—a sufficient criterion for injectivity of generic measurements (CEHV15).

group.

Numerical experiments conducted in [GKK15a] highlight the almost optimal behavior of stabilizer states for phase retrieval, see Figure 1.3.

We conclude this section by pointing out that Theorem 7 remains valid, if we replace stabilizer states by any other Clifford orbit. In fact, several other Clifford orbits admit a better rank scaling in the sampling rate m.

1.5.4 Orthonormal basis measurements

Here, we shall focus on reconstructing a matrix X from a collection of *orthonormal basis measurements*:

$$y_k = \operatorname{tr}\left(b_k b_k^* X\right) \quad 1 \leq k \leq n, \tag{1.35}$$

where $b_1, \ldots, b_n \in \mathbb{C}^n$ denotes an orthonormal basis. Since two orthonormal bases are related via a unitary transformation $U \in U(n)$, we may equivalently write

$$y_k = \operatorname{tr}\left(U f_k f_k^* U^* X\right) \quad 1 \leq k \leq n, \tag{1.36}$$

where $f_1, \ldots, f_n$ denotes the orthonormal basis of discrete Fourier vectors. Viewed from this perspective, orthonormal basis measurements are very similar to coded diffraction patterns (1.27). However, here the modulation is due to a unitary rotation U, instead of a diagonal mask D_l. For maximal randomness, in the sense that each U is chosen Haar-uniformly from $U(n)$, Voroninski could establish reconstruction results [Vor13]. He proved that a constant number of such generic orthonormal basis measurements suffices to reconstruct a rank-one matrix $X = xx^*$ with high probability. We point out that each orthonormal basis measurement encompasses n different measurements. Thus the total number of measurements amounts to $m = Cn$, which is optimal up to multiplicative factors. Although not stated explicitly, it is plausible that this result may be extended to hermitian matrices with higher rank.

By combining the proof techniques from [GKK15b] (coded diffraction patterns) and [GKK15a] (spherical designs) we were able to de-randomize this statement also generalize it to arbitrary rank:

Theorem 8 (Simplified version of Theorem 2 in [Kue15])**.** *Let $X \in H_n$ be a hermitian matrix of rank r an suppose that each U in* (1.36) *is chosen independently from a unitary t-design ($t \geq 3$). Then, with high probability*

$$L \geq Ctn^{\frac{2}{t}} r \log^2(n)$$

orthonormal basis measurements allow for reconstructing X via nuclear norm minimization.

Unitary t-designs are a generalization of the spherical design concept to the unitary group $U(n)$ [DCEL09; GAE07]. On first sight, this result bears strong similarities with Theorem 1 above. However, it is a statement about matrix reconstruction and not not only valid for PhaseLift ($r = 1$).

Also, unlike coded diffraction patterns, the orthonormal basis measurements considered here do not admit a strong notion of probabilistic incoherence. So far, this lack of incoherence together with the fact that that the individual measurements are not independent has prevented us from further improving this result.

A concrete example for such a measurement procedure are stabilizer states in power of two dimension. In turn, Theorem 8 implies that measuring

$$L \geq C' r n^{\frac{2}{3}} \log^2(n)$$

random stabilizer bases allows for rank-r matrix reconstruction. Measurements of this type are not only feasible, but also typical, for several quantum mechanical experiments. Meeting the structural requirements of these types of experiments has been my main motivation to study matrix reconstruction from orthonormal basis measurements.

1.6 Miscellaneous results convex reconstruction problems

The previous two sections were devoted to the main results of this thesis. In this section, I present further results on convex signal reconstruction that were obtained throughout the course of my PhD. Several of these projects address features that are typical for PhaseLift—such as anisotropic measurements and positivity constraints—in more generality.

1.6.1 Compressed sensing from anisotropic measurements

We have already introduced *minimal coherence* (1.17) and *isotropy* (1.21) as desirable properties for performing convex signal reconstruction. For sparse vector reconstruction from measurements that are independent realizations of a random vector $a \in \mathbb{C}^n$ these amount to

$$\mu = \max_{1 \leq k \leq n} |\langle e_k, a \rangle|^2 \text{ (coherence parameter)} \quad \text{and} \quad \mathbb{E}\left[aa^*\right] = \mathbb{I} \text{ (isotropy).}$$

Candès and Plan could show that these two requirements suffice for establishing compressed sensing reconstruction guarantees [CP11a]. They prove that w.h.p. a fixed s-sparse vector can be reconstructed from

$$m \geq C\mu s \log(n)$$

random isotropic measurements with coherence parameter μ.

A concrete example for such a measurement ensemble are re-scaled Fourier basis vectors $\sqrt{n} f_1, \ldots, \sqrt{n} f_n \in \mathbb{C}^n$. They are isotropic and admit a minimal coherence parameter $\mu = 1$. Consequently, $m = Cs\log(n)$ random Fourier basis measurements w.h.p. suffice to reconstruct a s-sparse vector via ℓ_1-norm minimization (1.7).

We could further generalize this result by considerably relaxing the isotropy condition. To this end, we introduce the following *condition number*:

$$\kappa := \kappa\left(\mathbb{E}\left[aa^*\right]^{\frac{1}{2}}\right) = \frac{\lambda_{\max}\left(\mathbb{E}\left[aa^*\right]^{\frac{1}{2}}\right)}{\lambda_{\min}\left(\mathbb{E}\left[aa^*\right]^{\frac{1}{2}}\right)}$$

Note that isotropy is equivalent to demanding $\kappa = 1$. In turn, we need to adjust the coherence parameter

$$\tilde{\mu} := \max\left\{ \max_{1 \leq k \leq n} |\langle e_k, a \rangle|^2, \max_{1 \leq k \leq n} \left|\langle e_k, \mathbb{E}\left[aa^*\right]^{-1} a \rangle\right|^2 \right\}$$

and arrive at the following statement.

Theorem 9 (Simplified version of Theorem 2 in [KG14]). *Let $x \in \mathbb{C}^n$ be an s-sparse vector and suppose that measurement vectors are chosen randomly from an ensemble $a \in \mathbb{C}^n$ with condition number κ and coherence parameter $\tilde{\mu}$. Then, w.h.p.*

$$m \geq C s \kappa \tilde{\mu} \log(n)$$

independent measurements suffice to reconstruct x via constrained ℓ_1-minimization (1.7).

1.6.2 The role of positivity assumptions

In the context of phase retrieval, we have already seen that exploiting its positive semi-definite structure is advantageous for noise-robustness. It allowed us to replace the "usual" constrained nuclear norm minimization [KRT15] by Algorithm (1.25) [KKRT16]. This latter algorithm is considerably simpler. And, perhaps more importantly, posing it does not require an a-priori bound $\eta \geq \|\epsilon\|_{\ell_2}$ on the noise strength.

Here, we show that similar conclusions may be drawn for sparse reconstruction of entry-wise non-negative vectors $x \geq 0$. The study of sparse reconstruction under such a positivity constraint has a long and rich history that actually pre-dates compressed sensing, see e.g. [DT05]. Subsequently, different aspects of non-negativity in compressed sensing have been analyzed, see e.g. [BEZ08] and [SH+13]. To the best of our knowledge, these works focus on the idealized scenario of reconstructing positive, sparse vectors from noiseless measurements.

In [KJ16] we put an emphasis on non-negative compressed sensing from noisy measurements. We combine the geometric insights from [BEZ08] with the notion of a robust null space property[3] [FR13] to arrive at the following conclusion:

Theorem 10 (Simplified version of Theorem 1 in [KJ16]). *Suppose that a real-valued measurement process $A : \mathbb{R}^n \to \mathbb{R}^m$ obeys the robust NSP for s-sparse vectors and its row-span intersects the positive orthant: $\sum_{k=1}^m t_k a_k > 0$ for some $t \in \mathbb{R}^m$. Then, solving*

$$z^\sharp = \underset{z \geq 0}{\arg\min} \|Az - y\|_{\ell_2} \tag{1.37}$$

allows for stably reconstructing any non-negative s-sparse vector $x \in \mathbb{R}^n$ from noisy measurements $Ax = y + \epsilon$:

$$\left\| z^\sharp - x \right\|_{\ell_2} \leq C' \frac{\|\epsilon\|_{\ell_2}}{\sqrt{m}}.$$

Note that Algorithm (1.37) is actually a simple non-negative least squares regression (NNLS). When using standard tools, such as CVX [GB14; GBY08], its runtime is consid-

[3] The null space property is somewhat "folklore". We refer to loc. cit. for a discussion about its origin.

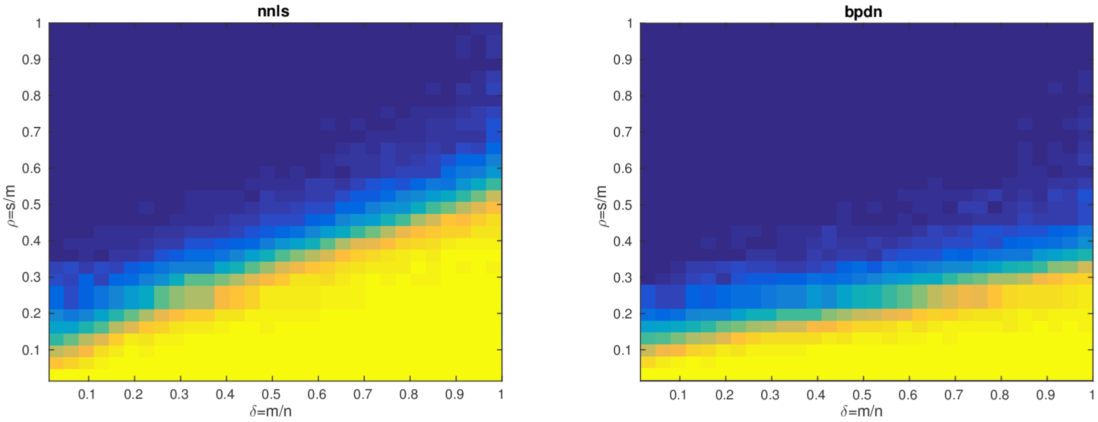


Figure 1.4: Comparison of NNLS and BPDN for $0/1$-Bernoulli matrices in the noisy setting.

erably lower than constrained ℓ_1-minimization (1.7). Perhaps more importantly, NNLS does not require any assumptions on the noise ϵ to assure stable reconstruction. This is not the case for constrained ℓ_1-minimization, where an appropriate choice of η is essential.

As a concrete example, we consider Bernoulli-random measurements:

Theorem 11 (Simplified version of Theorem 2 in [KJ16]). *A measurement process containing*

$$m \geq Cs\log(n)$$

independent $0/1$*-Bernoulli vectors* $a_k \in \mathbb{R}^n$ *meets the requirements of Theorem 10 with high probability.*

It is plausible that the number of measurements required may further be improved to $m \geq Cs\log(n/s)$. For such measurement processes, we have run numerical simulations to compare NNLS to "traditional" ℓ_1-minimization, see Figure 1.4. They highlight the advantage of exploiting positivity. This $0/1$-measurement model also has potential applications in current engineering problems. We discuss one such application—activity detection in large wireless networks—in [KJ16].

Finally, we point out that we proved a robust NSP for $0/1$-Bernoulli matrices in order to arrive at Theorem 11. This result alone allows for concluding strong (i.e. uniform and stable) compressed sensing results—regardless of positivity. To the best of our knowledge, we were the first to derive such strong results for $0/1$-Bernoulli measurements. This lack of results is likely due to the fact that such measurements are not isotropic:

$$\mathbb{E}\left[aa^*\right] = \sum_{k,l=1}^{n} \mathbb{E}\left[b_k b_l\right] e_k e_l^* = \frac{1}{4}\sum_{k\neq l} e_k e_l + \frac{1}{2}\sum_{k=1}^{n} e_k e_k^* = \frac{1}{4}\mathbb{I} + \frac{1}{4}\vec{1}\,\vec{1}^*.$$

Here, $\vec{1} = \sum_{k=1}^{n} e_k$ denotes the "all-ones" vector with respect to the standard basis. This anisotropy renders traditional strong proof techniques, such as establishing the famous restricted isometry property (RIP) [C0̀8], infeasible. However, in accordance with the previous subsection, Theorem 11 highlights that isotropy is not required for deriving strong compressed sensing results.

1.6.3 Matrix reconstruction via minimizing the diamond norm

Low rank matrix reconstruction is typically carried out via a constrained nuclear norm minimization (1.8). In some sense [FHB01], the nuclear norm is the tightest convex relaxation of rank. However, this may not necessarily be the case if we restrict our attention to strict subsets of low rank matrices.

Motivated by applications in quantum information science, we focus on matrices with a *bipartite* structure: $X \in H_{n_1} \otimes H_{n_2}$. For such matrices, we identify a novel convex surrogate for rank. It is based on the *diamond norm*—an important distance measure in quantum information theory:

$$\|X\|_{\square} = \max \left\{ \|(\mathbb{I} \otimes A) X (\mathbb{I} \otimes B)\|_1 : \ \|A\| = \|B\| = \sqrt{n_2} \right\} \tag{1.38}$$

It is easy to see that $\|\cdot\|_{\square}$ is a norm and, although not obvious, it can be computed via a semidefinite program that satisfies strong duality [Wat13]. Also, note that the pair $A = B = \mathbb{I}$ is admissible in the maximization (1.38) and consequently

$$\|X\|_1 \leq \|X\|_{\square} \quad \forall X \in H_{n_1} \otimes H_{n_2}. \tag{1.39}$$

In [KKEG16] we provide analytical evidence in favor of the diamond norm as a convex surrogate for rank. We use the fact that the geometry of convex reconstruction schemes is well understood, see e.g. [Tro15]. Starting with a convex regularizer f (e.g. the nuclear norm), geometric proof techniques like Tropp's Bowling scheme [Tro15] (see also [KRT15]) bound the reconstruction error in terms of the *descent cone* of f at the matrix X that is to be recovered. These arguments suggest that the reconstruction error would decrease, if another convex regularizer with smaller descent cone would be used. In this sense, the following result implies that the diamond norm may be an improved regularizer for certain classes of matrices:

Theorem 12 (Simplified version of Corollary 8 in [KKEG16]). *Let $X \in H_{n_1} \otimes H_{n_2}$ be a matrix that saturates* (1.39) *, i.e. $\|X\|_1 = \|X\|_{\square}$. Then the descent cone of the diamond norm* (1.38) *in X is contained in the descent cone of the nuclear norm at the same point. Moreover, we completely describe the set of all matrices $X \in H_{n_1} \otimes H_{n_2}$ obeying $\|X\|_{\square} = \|X\|_1$.*

We demonstrate this supremacy numerically: The diamond norm indeed outperforms the nuclear norm in a number of relevant applications, including *quantum process reconstruction*. Quantum processes may be described by bipartite matrices $X \in H_{n_1} \otimes H_{n_2}$ which obey $\|X\|_\square = \|X\|_1$ by definition. Moreover, many idealized quantum processes that are relevant in quantum computation are described by matrices X with unit rank. Our results suggest to employ a constrained diamond norm minimization in order to reconstruct such processes from few random measurements. Numerical simulations conducted in [KKEG16] suggest that this is indeed favorable: For correct reconstruction, constrained diamond norm minimization requires fewer measurements than nuclear norm minimization.

1.7 Further results in quantum information theory

The main focus of this project is convex signal reconstruction, with an emphasis on phase retrieval and matrix reconstruction. As a field, convex reconstruction combines techniques from various branches of mathematics, including convex optimization, linear algebra and probability theory. These mathematical techniques lend themselves to tackling various different types of problems. This section is devoted to presenting insights into different problems in quantum information science. These were all obtained by applying such techniques.

1.7.1 A causal interpretation of Bell inequality violations

Bell inequalities are an elegant method to single out certain properties of quantum mechanical systems that cannot be explained classically [Bel64]. At their heart are experiments that involve two experimenters at different locations who simultaneously perform measurements on a shared physical system. Under natural assumptions, such as *locality* (the results obtained by one observer cannot be influenced by any action of the other), *measurement independence* (experimenters are free to choose which properties to measure) and *realism* (one can consistently assign a value to any physical property—independently of whether or not it is measured), the causal structure of this setup alone implies strong constraints on the statistical data that can arise. The resulting constraints are called Bell inequalities. Famously, quantum mechanical experiments can violate these constraints, see e.g. [CS78].

However, when trying to reproduce these results in a quantum experiment, practical limitations make it very challenging to assure that the underlying assumptions – space-like separation, locality and measurement independence – are met exactly. These practical limitations motivate studying how stable Bell inequalities are towards violating one, or more, of these assumptions.

To address this issue, we have re-visited typical "Bell experiments" and tried to explain the observations classically via *Bayesian networks*. Doing so allows us to construct and subsequently analyze alternative causal structures, e.g. one that does allow for non-local interactions. It turns out that the resulting problems can often be re-cast as linear programs. Subsequently, the versatility of linear programming allowed us to draw several conclusions [CKBG15], for instance: Novel "causal" interpretations of Bell inequality violations.

1.7.2 Bayesian quantum state estimation with fidelity

A finite, n-dimensional quantum system is fully described by its *quantum state*: a positive semidefinite matrix that has unit trace. The task of estimating such a description of a physical system from empirical data is called *quantum tomography*.

In recent years it has become increasingly popular to approach this task via Bayesian estimation theory. The key idea is to choose a prior distribution over quantum states and subsequently update it based on the measurement data. Doing so results in a posterior distribution $\mathrm{d}\rho$. Subsequently, point estimators $\hat{\rho}$ are obtained by specifying a loss function $L : H_n \times H_n \to \mathbb{R}$ and minimizing the expected posterior loss:

$$\begin{aligned} \underset{\sigma \geq 0}{\text{minimize}} \quad & \mathbb{E}_\rho \left[L(\rho, \sigma) \right] \\ \text{subject to} \quad & \mathrm{tr}(\sigma) = 1. \end{aligned} \tag{1.40}$$

A prominent loss function is the following generalization of mean square error: $L\left(\rho, \sigma\right) = \|\rho - \sigma\|_2^2$. For such a loss function, (1.40) results in the *Bayesian mean estimator* $\hat{\rho} = \mathbb{E}_\rho\left[\rho\right]$ [BK10].

However, Frobenius distance $\|\rho - \sigma\|_2$ is not a prominent distance measure in quantum information science. Instead, the *fidelity* [Uhl76]

$$F(\rho, \sigma) = \|\sqrt{\rho}\sqrt{\sigma}\|_1^2 \in [0, 1]$$

is by far the most commonly used figure of merit for comparing quantum states. Interestingly, the point estimator maximizing the expected fidelity is only known for a single qubit ($n = 2$) [Bag+06]. In [KF15] we address this lack of knowledge by providing upper bounds on the maximal expected fidelity achievable by any estimator:

Theorem 13 (Simplified version of Theorem 2 in [KF15]). *The maximal average fidelity achieved by any estimator $\hat{\rho}$ obeys*

$$\mathbb{E}_\rho\left[F(\rho, \hat{\rho})\right] \leq 1 - \frac{1}{4}\mathrm{tr}\left(\mathbb{E}_\rho\left[\rho^2\right] - \mathbb{E}_\rho\left[\rho\right]^2\right).$$

Such a result is useful for benchmarking the fidelity performance of different estimation techniques.

We complement our theoretical findings with numerical experiments. These demonstrate the relative tightness of our bounds. Moreover, they reveal that the Bayesian mean estimator [BK10] is an excellent point estimator. Despite not being designed for maximizing expected fidelity, it achieves values that—according to our bounds—are close to optimal.

1.7.3 Comparing experiments to fault-tolerance thresholds

The possibility of eventually constructing a working quantum computer hinges on the availability of robust control mechanisms that allow for compensating faulty computations. A strong theoretical guarantee for being able to do so, is the *Threshold Theorem* [Kit97]. In a nutshell, it states that if the error rate remains below a certain threshold, potentially noisy and even faulty computations can be fully compensated using quantum error correction. The figure of merit that quantifies this threshold is the *diamond norm difference* between the identity map $\mathcal{I} : H_n \to H_n$ and the error channel $\mathcal{E} : H_n \to H_n$ occurring. This makes it imperative to estimate diamond distances $\Delta(\mathcal{E}) = \|\mathcal{I} - \mathcal{E}\|_\diamond$ from experimentally available data. However, doing so is a non-trivial task.

Instead, the *average error rate* of a noise channel

$$r(\mathcal{E}) := 1 - \int_{S^{n-1}} \langle w, \mathcal{E}(ww^*)w \rangle \mathrm{d}w$$

is a prominent figure of merit. It can be estimated efficiently by performing techniques like direct fidelity estimation [FL11] and randomized benchmarking, see e.g. [MGE11].

To date, the best known general bound that relates these two error measures is

$$r(\mathcal{E}) \leq \Delta(\mathcal{E}) \leq \sqrt{n(n+1)r(\mathcal{E})} \quad \forall \mathcal{E} : H_n \to H_n, \tag{1.41}$$

which is rather discouraging. Current results on fault-tolerant quantum computation require a threshold of order 10^{-4} [AC07]. In principle, experimenters need to be able to achieve average error rates of order 10^{-8} in their working devices.

In [KLDF16] we address this problem. Concretely, we consider several realistic models of incoherent quantum noise and show that these admit a linear relation:

$$r(\mathcal{E}) \leq \Delta(\mathcal{E}) \leq 3r(\mathcal{E}).$$

Conversely, we show that coherent noise processes—such as unitary errors $\mathcal{E}(X) = UXU^*$—essentially saturate the upper bound in (1.41). We derived these relations by exploiting the semidefinite program formulation of the diamond norm [Wat09].

From a practical perspective, these results are encouraging. On the contrary to incoherent errors, such as leakage, coherent noise effects can typically be corrected. We point out that our findings from subsection 1.6.3 may be useful for such a task. They allow for reconstructing particular coherent processes from a reduced number of random measurements.

1.7.4 Distinguishing quantum states

Quantum state discrimination is the task of correctly distinguishing between two quantum states, say ρ and σ, by performing a single quantum measurement. Helstrom's theorem [Hel76] states that the maximal probability of success for such a task is bounded by

$$p_{\mathrm{succ}} \leq \frac{1}{2} + \frac{1}{4} \|\rho - \sigma\|_1 ,$$

if both states occur with equal probability. This amounts to a an optimal bias of $\beta_{\mathrm{Helstrom}} = \frac{1}{4} \|\rho - \sigma\|_1$. Moreover, Helstrom showed that this bound is achievable by performing a particular quantum measurement that depends on ρ and σ. However, such a particular measurement is optimized to distinguish ρ from σ and may fail completely at distinguishing other state pairs.

Addressing this lack of universality in Helstrom's theorem, Matthews, Wehner and Winter [MWW09] turned this problem around: instead of fixing the states and optimizing the measurement procedure, they consider the performance of a fixed measurement at distinguishing arbitrary pairs of states $\rho, \sigma \in H_n$. In particular, they could show that a 4-design measurement performs surprisingly well at this task:

$$\beta_{4\mathrm{D}} \geq \frac{1}{6\sqrt{\mathrm{rank}(\rho - \sigma)}} \|\rho - \sigma\|_1 \quad \forall \rho, \sigma.$$

This is close to optimal, in the sense that it reproduces the performance of the *uniform measurement* encompassing all xx^* with $x \in S^{n-1}$. Conversely, a 2-design measurement may perform considerably worse:

$$\beta_{2\mathrm{D}} \geq \frac{1}{2(n+1)} \|\rho - \sigma\|_1 .$$

This bound cannot be further improved in general.

Our novel results about the fourth moments of stabilizer states (1.33) [ZKGG16] have enabled us to infer similar results for measurements that consist of all stabilizer states.

Theorem 14 (Simplified version of Theorem 4 in [KZG16a]). *If the dimension $n = 2^d$ is a* *power of two, stabilizer measurements obey*

$$\beta_{\text{stab}} \geq \frac{1}{4\text{rank}(\rho - \sigma)} \|\rho - \sigma\|_1 \quad \forall \rho, \sigma.$$

Note that this result critically depends on the rank of the states ρ and σ considered. If both states are approximately *pure*, i.e. $\text{rank}(\rho) \simeq \text{rank}(\sigma) \simeq 1$, we can conclude $\beta_{\text{stab}} \geq \frac{1}{8} \|\rho - \sigma\|_1$, which almost reproduces the 4-design behavior. Conversely, if $\rho - \sigma$ has full rank then stabilizer states may perform as bad as 2-designs: $\frac{1}{4n} \|\rho - \sigma\|_1 \leq \beta_{\text{stab}} \leq \frac{1}{n+1} \|\rho - \sigma\|_1$.

2 Publications

2.1 List of publications

This cumulative dissertation is based on thirteen publications. Out of these, eight have been published in peer-refereed journals, three in peer-refereed conference proceedings and one is currently under review at physical review letters. I am the main contributor to eight of these articles and provided important contributions to the remaining five.

[GKK15a] D. Gross, F. Krahmer, R. Kueng, *A partial derandomization of PhaseLift using spherical designs*, Journal of Fourier Analysis and Applications **21**, 229–266 (2015),

[KGK15] R. Kueng, D. Gross, F. Krahmer, *Spherical designs as a tool for derandomization: the case of PhaseLift*, IEEE-endorsed (peer-reviewed and published) proceedings for for the international conference on Sampling Theory and Applications (SampTA) (2015)

[KRT15] R. Kueng, H. Rauhut, U. Terstiege, *Low rank matrix recovery from rank one measurements*, Applied and Computational Harmonic Analysis **42**, 88–116 (2017)

[KKRT16] M. Kabanava, R. Kueng, H. Rauhut, U. Terstiege, *Stable low-rank matrix recovery via null space properties*, Information and Inference **5**, 405–441 (2016)

[GKK15b] D. Gross, F. Krahmer, R. Kueng, *Improved recovery guarantees for phase retrieval from coded diffraction patterns*, Applied and Computational Harmonic Analysis, 37–64 (2017)

[KG15] R. Kueng, D. Gross, *Qubit stabilizer states are complex projective 3-designs* preprint arXiv:1510.02767 (2015)

[Kue15] R. Kueng, *Low rank matrix recovery from few orthonormal basis measurements*, IEEE-endorsed (peer-reviewed and published) proceedings for for the international conference on Sampling Theory and Applications (SampTA) (2015)

[KG14] R. Kueng, D. Gross, *RIPless compressed sensing from anisotropic measurements*, Linear Algebra and its Applications **441**, 119–123 (2014)

[KJ16] R. Kueng, P. Jung, *Robust nonnegative sparse recovery and the null space property of* $0/1$ *measurements*, Proceedings of the IEEE Information Theory Workshop (2016)

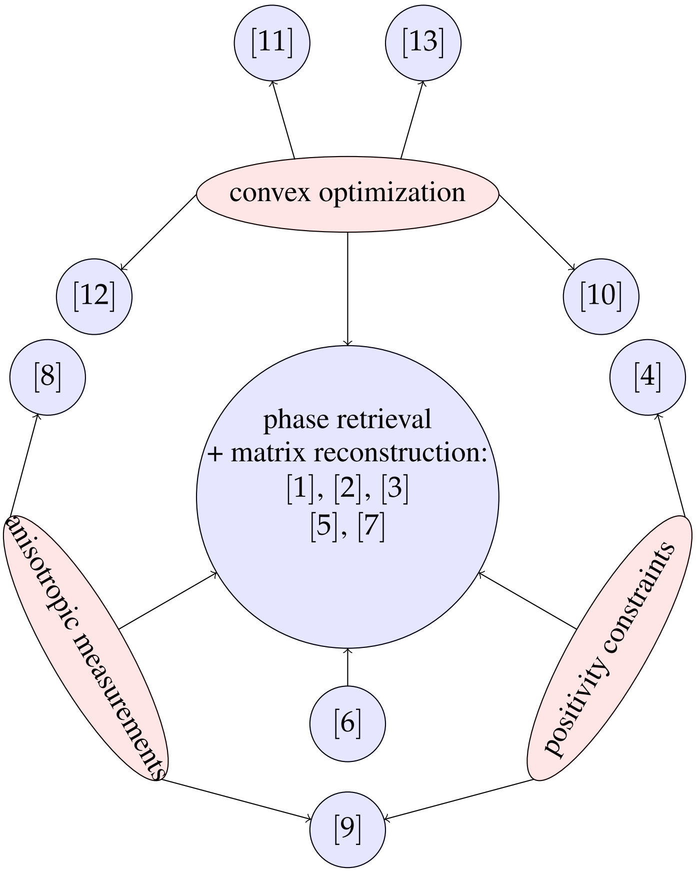
[KKEG16] M. Kliesch, R. Kueng, J. Eisert, D. Gross, *Improving compressed sensing with the diamond norm*, IEEE Transactions on Information Theory **62**, 7445–7463 (2016)

[CKBG15] R. Chaves, R. Kueng, J.B. Brask, D. Gross, *A unifying framework for relaxations of the causal assumptions in Bell's Theorem*, Physical Review Letters **114**, 140403 (2015)

[KF15] R. Kueng, C. Ferrie, *Near-optimal quantum tomography: estimators and bounds*, New Journal of Physics **17**, 123013 (2015)

[KLDF16] R. Kueng, D.M. Long, A.C. Doherty, S.T. Flammia, *Comparing Experiments to the Fault-Tolerance Threshold* , Phys. Rev. Lett. **117**, 170502 (2016)

While the topics of these articles are seemingly diverse, I want to emphasize that the methods and proof techniques are strongly related. This is illustrated in Figure 2.1.



[1] A partial derandomization of PhaseLift using spherical designs, [GKK15a]
[2] Spherical designs as a tool for derandomization: the case of PhaseLift [KGK15]
[3] Low rank matrix recovery from rank one measurements, [KRT15]
[4] Stable low-rank matrix recovery via null space properties, [KKRT16]
[5] Improved recovery guarantees for phase retrieval from coded diffraction patterns, [GKK15b]
[6] Qubit stabilizer states are complex projective 3-designs [KG15]
[7] Low rank matrix recovery from few orthonormal basis measurements [Kue15]
[8] RIPless compressed sensing from anisotropic measurements, [KG15]
[9] Robust nonnegative sparse recovery and the null space property of $0/1$ measurements [KJ16]
[10] Improving compressed sensing with the diamond norm, [KKEG16]
[11] A unifying framework for relaxations of the causal assumptions in Bell's theorem, [CKBG15]
[12] Near optimal quantum tomography: estimators and bounds, [KF15]
[13] Comparing experiments to the fault-tolerance threshold [KLDF16]

Figure 2.1: This figure illustrates the methodical connections between the different publications presented in this cumulative dissertation.

A Partial Derandomization of PhaseLift using Spherical Designs

D. Gross[1,2], F. Krahmer[3], R. Kueng*[1]

[1]Institute for Physics, University of Freiburg, Rheinstraße 10, 79104 Freiburg, Germany
[2]Freiburg Center for Data Analysis Modeling, Eckerstr. 1, 79104 Freiburg, Germany
[3]Institute for Numerical and Applied Mathematics, University of Göttingen, Lotzestraße 16-18, 37083 Göttingen, Germany

November 20, 2014

Abstract. The problem of retrieving phase information from amplitude measurements alone has appeared in many scientific disciplines over the last century. *PhaseLift* is a recently introduced algorithm for phase recovery that is computationally tractable, numerically stable, and comes with rigorous performance guarantees. PhaseLift is optimal in the sense that the number of amplitude measurements required for phase reconstruction scales linearly with the dimension of the signal. However, it specifically demands Gaussian random measurement vectors — a limitation that restricts practical utility and obscures the specific properties of measurement ensembles that enable phase retrieval. Here we present a partial derandomization of PhaseLift that only requires sampling from certain polynomial size vector configurations, called *t-designs*. Such configurations have been studied in algebraic combinatorics, coding theory, and quantum information. We prove reconstruction guarantees for a number of measurements that depends on the degree t of the design. If the degree is allowed to grow logarithmically with the dimension, the bounds become tight up to polylog-factors. Beyond the specific case of PhaseLift, this work highlights the utility of spherical designs for the derandomization of data recovery schemes.

Keywords: Phase retrieval, PhaseLift, Semidefinite relaxations of nonconvex quadratic programs, non-commutative large deviation estimates, spherical designs, quantum information

Mathematics Subject Classification: 90C25 – 49N30 – 62H12 – 60F10

1. Introduction

In this work we are interested in the problem of recovering a complex signal (vector) $x \in \mathbb{C}^d$ from an *intensity* measurement $y_0 = \|x\|_{\ell_2}^2$ and *amplitude* measurements

$$y_i = |\langle a_i, x\rangle|^2 \quad i = 1, \ldots, m,$$

where $a_1, \ldots, a_m \in \mathbb{C}^d$ are sampling vectors. Problems of this type are abundant in many different areas of science, where capturing phase information is hard or even infeasible, but obtaining amplitudes is comparatively easy. Prominent examples for this case occur in X-ray cristallography, astronomy and diffraction imaging – see for example [1]. This inverse problem is called *phase retrieval* and has attracted considerable interest over the last decades.

It is by no means clear how many such amplitude measurements are necessary to allow for recovery. Thus from the very beginning, there have been a number of works regarding injectivity conditions for this problem in the context of the specific applications [2].

More recently this question has been studied in more abstract terms, asking for the minimal number of amplitude measurements of the form (1) – without imposing structural assumptions on the a_i's – that are required to make the above map injective. In [3], the authors showed that in the real case ($x \in \mathbb{R}^d$), at least $2d - 1$ such measurements are

*Corresponding author: richard.kueng@physik.uni-freiburg.de necessary and generically sufficient to guarantee injectivity, while in the complex case a generic sample size of $m \geq 4d - 2$ suffices. Here generic is to be understood in the sense that the sets of measurements of such size which do not allow for recovery form an algebraic variety in the space of all frames. Also, the latter bound is close to optimal: as shown in [4], it follows from the results derived in [5] that a sample size of $m \geq (4 + o(1))\, d$ is necessary (cf. [6]). However, finding the precise bound is still an open problem.

Balan et al. [7] consider the scenario of $\mathcal{O}(d^2)$ measurements, which form a complex projective 2-design (cf. Def. 3 below). They derive an explicit reconstruction formula for this setup based on the following observation well known in conic programming. Namely, the quadratic constraints on x are linear in the outer product xx^*:

$$y_i = |\langle a_i, x\rangle|^2 = \operatorname{tr}\left((a_i a_i^*)(xx^*)\right). \tag{1}$$

This "lifts" the problem to matrix space of dimension d^2, where it becomes linear and can be explicitly solved to find the unique solution.

As we will show in Theorem 2, it is, without making additional assumptions on the 2-design, not possible to use as measurements a random subset of this 2-design which is of size $o(d^2)$. In other words, for the measurement scenario described in [7], the quadratic scaling in d is basically unavoidable.

To contrast these two extreme approaches, ref. [3] works with a number of measurements close to the absolute minimum, but there are no tractable reconstruction schemes provided, the question of numerical stability is not considered, and it is unclear whether non-generic measurements – i.e., vectors with additional structural properties – can be employed. On the other hand, the number of measurements in [7] is much larger, while the measurements are highly structured and there is an explicit reconstruction method. A number of recent works including this paper aim to balance between these two approaches, working with a number of measurements only slightly larger while having at least some of the desired properties mentioned above.

Ref. [8] introduces a reconstruction method called *polarization* that works for $\mathcal{O}(d \log d)$ measurements and can handle structured measurement vectors, including the *masked illumination* setup that appears in diffraction imaging [9], where the measurements are generated by the discrete Fourier transform preceded by a random diagonal matrix. For Gaussian measurements, the polarization approach has also shown to be stable with respect to measurement noise [8]. While simulations seem to suggest stability also for the derandomized masked illumination setup, a proof of stability is – to our knowledge – not available yet.

An alternative approach, which we will also follow in this paper, is the *PhaseLift* algorithm, which is based on the lifted formulation (1). The algorithm was introduced in [10] and reconstruction guarantees have been provided in [11, 12]. The central observation is that the matrix xx^*, while unknown, is certainly of rank one. This connects the phase retrievel problem with the young but already extensive field of *low-rank matrix recovery* [13, 14, 15, 16]. Over the past years, this research program has rigorously identified many instances in which low-rank matrices can be efficiently reconstructed from few linear measurements. The existing results on low-rank matrix recovery were not directly applicable to phase retrieval, because the measurement matrices $a_i a_i^*$ failed to be sufficiently *incoherent* in the sense of [14, 15] (the incoherence parameter captures the well-posedness of a low-rank recovery problem). For the case of Gaussian measurement vectors a_i, Candès, Strohmer, Voroninski and Li were able to circumvent this problem, providing problem-specific stable recovery guarantees [11, 12] for a number of measurements of optimal order

$\mathcal{O}(d)$. For recovery, they use a convex relaxation of the rank minimization problem, which makes the reconstruction algorithm tractable.

It should be noted, however, that because of the significantly increased problem dimensions, PhaseLift is not as efficient as many phase retrieval algorithms developed over the last decades in the physics literature (such as [17]) and the optimization literature (for example [18]). Recently there have been attempts to provide recovery guarantees for alternating minimization algorithms [19], which are somewhat closer to the algorithms used in practice, but this direction of research is only at its beginnings.

While the above mentioned recovery guarantees for PhaseLift address the issues of tractable reconstruction and stability with respect to noise, these results leave open the question of whether measurement systems with additional structure and less randomness still allow for guaranteed recovery. There are both practical and theoretical motivations for pursuing such generalizations: A practitioner may be constrained in the choice of measurements by the application at hand or reduce the amount of randomness required for implementation purposes. The most prominent example are again masked Fourier measurements, which appear as a natural model in diffraction imaging, but a lot of different scenarios imposing different structure are conceivable. From a theoretical point of view, the use of Gaussian vectors obscures the specific properties that make phase retrieval possible. As discussed in the following subsection, it is a common thread in randomized signal processing that results are first established for Gaussian measurements and later generalized to structured ensembles.

A different direction of research, which will not be pursued in this paper, is to ask how additional structural assumptions on the signal to be recovered, such as sparsity, can be incorporated into the theory. A general analysis based on the Gaussian width of how many measurements are needed to allow for stable recovery of a signal known to lie in a set $T \subset \mathbb{R}^d$ is provided in [20]. Notably the results allow for measurements with arbitrary subgaussian rather than just Gaussian entries. Efficient algorithms for recovery, however, are not provided. For the case of s-sparse signals, also tractable recovery algorithms are available: It has been shown in [21] that PhaseLift can recover x with high probability from Gaussian measurements for a number of measurements m proportional to s^2 (up to logarithmic factors), which, for small s, can be considerably less than the dimension. In [22], it is shown that only a number of subgaussian measurements scaling linearly in the sparsity (up to logarithmic factors) is needed if recovery proceeds using certain greedy algorithms.

1.1. **Designs as a general-purpose tool for de-randomization.** In this paper, we focus on the theoretical aspect: which properties of a measurements are sufficient for PhaseLift to succeed? We prove recovery guarantees for ensembles of measurement vectors drawn at random from a finite set whose first $2t$ moments agree with those of Haar-random vectors (or, essentially, Gaussian vectors). A configuration of finite vectors which gives rise to such an ensemble is known as a *complex projective t-design*[2]. Designs were introduced by Delsarte, Goethals and Seidel in a seminal paper [23] and have been studied in algebraic combinatorics [24], coding theory [23, 25], and recently in quantum information theory [26, 27, 28, 29, 30]. Furthermore, complex projective 2-designs were the key ingredient for the reconstruction formula for phase retrieval proposed in [7].

[2] The definition of a t-design varies between authors. In particular, what is called a t-design here (and in most of the physics literature), would sometimes be referred to as a $2t$ or even a $(2t + 1)$-design. See Section 3.3 for our precise definition.

One may see a more general philosophy behind this approach. In the field of sparse and low-rank reconstruction, a number of recovery results had first been established for Gaussian measurements. In subsequent works, it has then been proven that measurements drawn at random from certain fixed orthonormal bases are actually sufficient. Examples include uniform recovery guarantees for compressed sensing ([31, 32] vs. [33, 34]) and low-rank matrix recovery ([13] vs. [16]), respectively. Typically, the de-randomized proofs require much higher technical efforts and deliver slightly weaker results. For a recent survey on structured random measurements in signal processing see [35].

As the number of measurements needed for phase retrieval is larger than the signal space dimension, one cannot expect these results to exactly carry over to the phase retrieval setting. Nevertheless, the question remains whether there is a larger, but preferably not too large, set such that measurements drawn from it uniformly at random allow for phase retrieval reconstruction guarantees. In some sense, the sampling scenario we seek can be interpreted as an interpolation between the maximally random setup of Gaussian measurement with an optimal order of measurements and the construction in [7], which is completely deterministic, but suboptimal in terms of the embedding dimension. While in this paper, we will focus on the phase retrieval problem, we remark that such an interpolating approach between measurements drawn from a basis and maximally random measurements may also be of interest in other situations where constructions from bases are known, but lead to somewhat suboptimal embedding dimensions.

The concept of t-designs, as defined in Section 3.3, provides such an interpolation. The intuition behind that definition is that with growing t, more and more moments of the random vector corresponding to a random selection from the t-design agree with the Haar measure on the unit sphere. In that sense, as t scales up further, t-designs give better and better approximations to Haar-random vectors.

The utility of this concept as a general-purpose de-randomization tool for Hilbert-space valued random construtions has been appreciated for example in quantum information theory [27, 36]. It has been compared [27] to the notion of *t-wise independence*, which plays a role for example in the analysis of discrete randomized algorithms [37], seems to have been long appreciated in coding theory. The smallest t-design in $\mathbb{C}^d$ consists of $\mathcal{O}(d^{2t})$ elements. Thus, whenever that lower bound is met, drawing a single element from a design requires $2t \log d$ bits, as opposed to $2d$ bits for a complex Bernoulli vector – an exponential gap.

From a practical point of view, the usefulness of these concepts hinges on the availability of constructions for designs. Explicit constructions for any order t and any dimension d are known [28, 38, 39, 40] – however, they are typically “inefficient” in the sense that they require a vector set of exponential size. For example, the construction in [28] uses $\mathcal{O}(t)^d$ vectors which is exponential in the dimension d.

Analytic expressions for *exact* designs are notoriously difficult to find. Designs of degree 2 are widely known [41, 42, 43, 44]. A concrete example is used for the converse bound in Section 7 (as well as for the converse bounds for low-rank matrix recovery from Fourier-type bases in [15]). For degree 3, both real[3] [24] and complex [45] designs are known. For higher t, there are numerical methods based on the notion of the *frame potential* [46, 44, 45] , non-constructive existence proofs [40], and constructions in sporadic dimensions (c.f. [47] and references thererin).

[3] While stated only for dimensions that are a power of 2, the results can be used for construtions in arbitrary dimensions [45].

Importantly, almost-tight randomized constructions for *approximate designs* for arbitrary degrees and dimensions are known [27, 28, 30]. The simplest results [28] show that collections of Haar-random vectors form approximate t-designs. This indeed can reduce randomness: One only needs to expend a considerable amount of randomness *once* to generated a design – for subsequent applications it is sufficient to sample small subsets from it[4]. Going further, there have been recent deep results on designs obtained from certain structured ensembles [30]. We do not describe the details here, as they are geared toward quantum problems and may have to be substantially modified to be applicable to the phase retrivial. The only connection to phase retrieval to date is the estimation of pure quantum states [4, 49].

Finally we point out that the notion of the *frame potential* above is no coincidence. In [50] a frame-theoretic approach to designs is provided, underlining their close connection.

1.2. **Main results.** In this paper, we show that spherical designs can indeed be used to partially derandomize recovery guarantees for underdetermined estimation problems; we generalize the recovery guarantee in [11] to measurements drawn uniformly at random from complex projective designs, at the cost of a slightly higher number of measurements.

Theorem 1 (Main Theorem). *Let $x \in \mathbb{C}^d$ be the unknown signal. Suppose that $\|x\|_{\ell_2}^2$ is known and that m measurement vectors $a_1, \ldots, a_m$ have been sampled independently and uniformly at random from a t-design $D_t \subset \mathbb{C}^d$ ($t \geq 3$). Then, with probability at least $1 - \mathrm{e}^{-\omega}$, PhaseLift (the convex optimization problem (24) below) recovers x up to a global phase, provided that the sampling rate exceeds*

$$m \geq \omega\, C t\, d^{1+2/t} \log^2 d. \tag{2}$$

Here $\omega \geq 1$ is an arbitrary parameter and C is a universal constant.

As the discussion of the previous subsection suggests, the bounds on the sampling rate decrease as the order of the design increases. For fixed t, and up to poly-log factors, it is proportional to $\mathcal{O}(d^{1+2/t})$. This is sub-quadratic for the regime $t \geq 3$ where our arguments apply. If the degree is allowed to grow logarithmialy with the dimension (as $t = 2\log d$), we recover an optimal, linear scaling up to a polylog overhead, $m = \mathcal{O}(d \log^3 d)$.

In light of the highly structured, analytical and exact designs known for degree 2 and 3, it is of great interest to ask whether a linear scaling can already be achieved for some small, fixed t. As shown by the following theorem, however, for $t = 2$ not even a subquadratic scaling is possible if no additional assumptions are made, irrespective of the reconstruction algorithm used.

Theorem 2 (Converse bound). *Let d be a prime power. Then there exists a 2-design $D_2 \subset \mathbb{C}^d$ and orthogonal, normalized vectors $x, z \in \mathbb{C}^d$ which have the following property.*

Suppose that m measurement vectors $y_1, \ldots, y_m$ are sampled independently and uniformly at random from D_2. Then, for any $\omega \geq 0$, the number of measurements must obey

$$m \geq \frac{\omega}{4} d(d+1),$$

or the event

$$|\langle a_i, x\rangle|^2 = |\langle a_i, z\rangle|^2 \quad \forall\, i \in \{1, \ldots, m\}$$

will occur with probability at least $\mathrm{e}^{-\omega}$.

[4] The situation is comparable to the use of random graphs as randomness expanders [48].

It is worthwhile to put this statement in perspective with other advances in the field. Throughout our work, we have only demanded that the set of all possible measurement vectors forms a t-design and have not made any further assumptions. Theorem 2 has to be interpreted in this regard: The 2-design property *alone* does not allow for a sub-quadratic scaling when a "reasonably small" probability of failure is required in the recovery process.

Note that this does not exclude the possibility that certain realizations of 2-designs can perform better, if additional structural properties can be exploited. A good example for such a measurement process is the multi-illumination setup provided in [51]. In [52] the authors of this paper verified that the set of all measurement vectors used in the framework of [51] does constitute a 2-design (Lemma 6). Additional structural properties – most notably a certain correlated Fourier basis structure in the individual measurements – allowed for establishing recovery guarantees already for $m = \mathcal{O}\left(d \log^4 d\right)$ measurements [51] and $m = \mathcal{O}(d \log^2 d)$ [52], respectively – which both clearly are sub-quadratic sampling rates.

1.3. **Outlook.** There are a number of problems left open by our analysis. First, recall that our results achieve linear scaling up to logarithmic factors only when samples are drawn from a set of superpolynomial size. Thus it would be very interesting to find out whether there are polynomial size sets such that sampling from them achieves such a scaling, in particular, if t-designs for some fixed t can be used. The case of $t = 3$ seems particularly important in that regard, since the converse bound (Theorem 2) shows that a design order of at least 3 is necessary. Also, highly structued 3-designs are known to exist (see above).

Another important follow-up problem concerns approximate t-designs. While our main result is phrased for exact t-designs, certain scenarios will only exhibit approximate design properties. We expect that our proofs can be generalized to such a setup, but also leave this problem for future work. Lastly, the reconstruction quality for noisy measurements is also an important issue yet to be investigated.

2. Numerical Experiments

In this section we complement our theoretical results with numerical experiments, which we have implemented in Matlab using CVX [53, 54]. As may have been expected, these experiments suggest that PhaseLift from designs actually works much better than our main theorem suggests. To be concrete, we use *stabilizer states* – a highly structured vector set which is very prominent in quantum information theory [55, 56]. Stabilizer states exist in any dimension, though their properties are somewhat better-behaved in prime power dimensions. In this case, there exists $\mathcal{O}(d^{\log d})$ stabilizer state vectors. Due to their rich combinatorial structure, these vectors can be constructed efficiently. For dimensions $d = 2^n$ that are a power of two, it is known [45] that the set of stabilizer states forms a 3-design. This statement is false for other prime power dimensions ($d \neq 2^n$ for some n), where they only form an exact 2-design. However, weighted 3-designs can be constructed for arbitrary dimensions d by projecting down stabilizer states from the next largest power-of-2-dimension 2^n obeying $2^{n-1} < d < 2^n$ [45]. For further clarification of the concept of exact and weighted t-designs we defer the reader to [26] and references therin.

We have used these vectors in our numerical simulations, the results of which are depicted in Figure 1. For each dimension d between 1 and 32 (x-axis) and for each number of measurements m ranging from 1 to 160 (y-axis), we ran a total of 30 independent experiments. Each such experiment consisted in choosing a Haar-random (normalized Gaussian) complex vector x as test signal. Then, we drew m projected stabilizer states uniformly at random and calculated their squared overlap with the test signal. We then ran PhaseLift on this data and declared the recovery a "success" if the Forbenius distance between the reconstructed matrix $\tilde{X}$ and the true projection $X = xx^*$ was smaller than 10^{-3}. Figure 1 depics the empirical success probability: Black corresponds to only failures, white to only successes.

We obtain the picture of a relatively sharp phase transition along a line that scales linearly in the problem dimension. In fact, the transtion seems to occur in the vicinity of the line $m = 4d - 4$ – drawn in red in Figure 1. This seems to agree with the conjecture that $4d - 4$ measurements are required for injectivity (see e.g. [4]). However, there are a few differences in the problem setup: Firstly, the conjecture only asks whether there is a unique solution, while the numerical simulations study whether the PhaseLift algorithm can find it. Secondly, the conjecture concerns unique solutions for all possible inputs, while numerically, we estimate the success probability. And thirdly, the conjecture states that generic measurements work, while our simulations use a specific random procedure (drawn uniformly from a 3-design) to generate them.

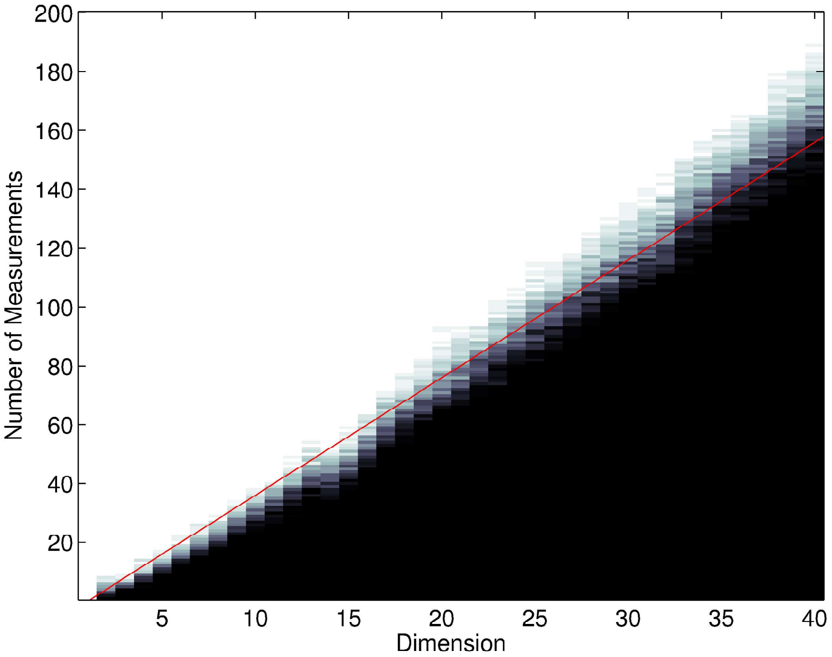


FIGURE 1. Phase Diagram for PhaseLift from (projected) stabilizer states, which form an exact 3-design in dimensions 2^n and a weighted one else. The x-axis indicates the problem's dimension, while the y-axis denotes the number of independent design measurements performed. The frequency of a successful recovery over 30 independent runs of the experiment appears color-coded from black (zero) to white (one). To guide the eye, we have furthermore included a red line indicating $m = 4d - 4$.

3. Technical Background and Notation

3.1. **Vectors, Matrices and matrix valued Operators.** In this work we require three different objects of linear algebra: vectors, matrices and operators acting on matrices.

We will work with vectors in a d-dimensional complex Hilbert space V^d equipped with an inner product $\langle \cdot, \cdot \rangle$. We refer to the associated induced norm by

$$\|z\|_{\ell_2} = \sqrt{\langle z, z \rangle} \quad \forall z \in V^d.$$

We will denote such vectors by latin characters. For $z \in V^d$, we define the dual vector $z^* \in (V^d)^*$ via

$$z^* y = \langle z, y \rangle \quad \forall y \in V^d.$$

On the level of matrices we will exclusively consider $d \times d$ dimensional hermitian matrices, which we denote by capital latin characters. Endowed with the Hilbert-Schmitt (or Frobenius) scalar product

$$(Z, Y) = \operatorname{tr}(ZY), \tag{3}$$

the space H^d becomes a Hilbert space. In addition to that, we will require the 3 different Schatten-norms

$$\begin{aligned} \|Z\|_1 &= \operatorname{tr}(|Z|) \quad \text{(trace norm)}, \\ \|Z\|_2 &= \sqrt{\operatorname{tr}(Z^2)} \quad \text{(Frobenius norm)}, \\ \|Z\|_\infty &= \sup_{y \in V^d} \frac{\|Zy\|_{\ell_2}}{\|y\|_{\ell_2}} \quad \text{(operator norm)}, \end{aligned}$$

where the second one is induced by the scalar product (3). These three norms are related via the inequalities

$$\|Z\|_2 \leq \|Z\|_1 \leq \sqrt{d}\|Z\|_2 \quad \text{and} \quad \|Z\|_\infty \leq \|Z\|_2 \leq \sqrt{d}\|Z\|_\infty \quad \forall Z \in H^d.$$

We call a hermitian matrix Z positive-semidefinite ($Z \geq 0$), if $\langle y, Zy \rangle \geq 0$ for all $y \in V^d$. Positive semidefinite matrices form a cone [57] (Chapter II,12), which induces a partial ordering of matrices. Concretely, for $Z, Y \in H^d$ we write $Y \geq Z$ if $Y - Z$ is positive-semidefinite ($Y - Z \geq 0$).

In this work, the identity matrix $\mathbb{1}$ and rank-1 projectors are of particular importance. They are positive semidefinite and any matrix of the latter kind can be decomposed as $Z = zz^*$ for some $z \in V^d$. Up to a global phase, they correspond to vectors $z \in V^d$. The most important cases are the projection onto the unknown signal x and onto the ith measurement vector a_i respectively. They will be denoted by

$$X = xx^* \quad \text{and} \quad A_i = a_i a_i^*.$$

Finally, we will frequently encouter *matrix-valued operators* acting on the space H^d. We label such objects with capital caligraphic letters and introduce the operator norm

$$\|\mathcal{M}\|_{\mathrm{op}} = \sup_{Z \in H^d} \frac{\|\mathcal{M}Z\|_2}{\|Z\|_2}$$

induced by the Frobenius norm on H^d. It turns out that only very few matrix-valued operators will appear below. These are: the identity map

$$\begin{aligned} \mathcal{I} : H^d &\to H^d \\ Z &\mapsto Z \quad \forall Z \in H^d \end{aligned}$$

and (scalar multiples of) projectors onto some matrix $Y \in H^d$. The latter corresponds to

$$\begin{aligned} \Pi_Y : H^d &\to H^d \\ Z &\mapsto Y(Y,Z) = Y\operatorname{tr}(YZ) \quad \forall Z \in H^d. \end{aligned}$$

The operator

$$\Pi_{\mathbb{1}} : Z \mapsto \mathbb{1}\operatorname{tr}(\mathbb{1}Z) = \mathbb{1}\operatorname{tr}(Z) \quad \forall Z \in H^d,$$

is a very important example for this subclass of operators. Note that it is not a normalized projection, but $\frac{1}{d}\Pi_{\mathbb{1}}$ is. Indeed, for $Z \in H^d$ arbitrary

$$\left(d^{-1}\Pi_{\mathbb{1}}\right)^2 Z = d^{-2}\mathbb{1}\operatorname{tr}(\mathbb{1}\Pi_{\mathbb{1}}Z) = d^{-2}\operatorname{tr}(\mathbb{1})\mathbb{1}\operatorname{tr}(Z) = d^{-1}\Pi_{\mathbb{1}}Z. \tag{4}$$

The notion of positive-semidefiniteness directly translates to matrix valued operators. Concretely, we call $\mathcal{M}$ positive-semidefinite ($\mathcal{M} \geq 0$) if $(Z, \mathcal{M}Z) \geq 0$ for all $Z \in H^d$. Again, this induces a partial ordering. Like in the matrix case, we write $\mathcal{N} \geq \mathcal{M}$, if $\mathcal{N} - \mathcal{M} \geq 0$. It is easy to check that all the operators introduced so far are positive semidefinite and in particular we obtain the ordering

$$0 \leq \Pi_{\mathbb{1}} \leq d\mathcal{I}, \tag{5}$$

by using (4).

3.2. **Multilinear Algebra.** The properties of t-designs are most naturally stated in the framework of (t-fold) tensor product spaces. This motivates recapitulating some basic concepts of multilinear algebra that are going to greatly simplify our analysis later on. The concepts presented here are standard and can be found in any textbook on multilinear algebra. Our presentation has been influenced in particular by [58, 59].

Let $V_1, \ldots, V_k$ be (finite dimensional, complex) vector spaces, and let $V_1^*, \ldots, V_k^*$ be their dual spaces. A function

$$f : V_1 \times \cdots \times V_k \to \mathbb{C}$$

is *multilinear*, if it is linear in each V_i, $i = 1, \ldots, k$. We denote the space of such functions by $V_1^* \otimes \cdots \otimes V_k^*$ and call it the *tensor product* of $V_1^*, \ldots, V_k^*$. Consequently, the tensor product $\left(V^d\right)^{\otimes k} = \bigotimes_{i=1}^k V^d$ is the space of all multilinear functions

$$f : \underbrace{\left(V^d\right)^* \times \cdots \times \left(V^d\right)^*}_{k \text{ times}} \mapsto \mathbb{C}, \tag{6}$$

and we call the elementary elements $z_1 \otimes \cdots \otimes z_k$ the *tensor product* of the vectors $z_1, \ldots, z_k \in V^d$. Such an element can alternatively be defined more concretely via the *Kronecker product* of the individual vectors. However, such a construction requires an explicit choice of basis in V^d which is not the case in (6).

With this notation, the space of linear maps $V^d \to V^d$ ($d \times d$-matrices) corresponds to the tensor product $M^d := V^d \otimes \left(V^d\right)^*$ which is spanned by $\left\{y \otimes z^* : \; y, z \in V^d\right\}$ – the set of all rank-1 matrices. For this generating set of M^d, we define the *trace* to be the natural bilinear map

$$\begin{aligned} \operatorname{tr} : V^d \otimes \left(V^d\right)^* &\to \mathbb{C} \\ (y \otimes z^*) &\mapsto z^* y = \langle z, y \rangle \end{aligned}$$

for all $y, z \in V^d$. The familiar notion of trace is obtained by extending this definition linearly to M^d.

Using $M^d = V^d \otimes \left(V^d\right)^*$ allows us to define the (matrix) tensor product $\left(M^d\right)^{\otimes k}$ to be the space of all multilinear functions

$$f : \underbrace{\left(\left(V^d\right)^* \times V^d\right) \times \cdots \times \left(\left(V^d\right)^* \times V^d\right)}_{k \text{ times}} \to \mathbb{C}$$

in complete analogy to the above. We call the elements $Z_1 \otimes \cdots \otimes Z_k$ the tensor product of the matrices $Z_1, \cdots, Z_k \in M^d$.

On this tensor space, we define the *partial trace* (over the i-th system) to be

$$\begin{aligned} \mathrm{tr}_i : \left(M^d\right)^{\otimes k} &\to \left(M^d\right)^{\otimes(k-1)} \\ Z_1 \otimes \cdots \otimes Z_k &\mapsto \mathrm{tr}(Z_i)\left(Z_1 \otimes \cdots \otimes Z_{i-1} \otimes Z_{i+1} \otimes \cdots \otimes Z_k\right). \end{aligned}$$

Note that with the identification $M^d = V^d \otimes \left(V^d\right)^*$, tr_i corresponds to the natural contraction at position i. The partial trace over more than one system can be obtained by concatenating individual traces of this form, e.g. for $1 \leq i < j \leq k$

$$\mathrm{tr}_{i,j} := \mathrm{tr}_i \circ \mathrm{tr}_j : \left(M^d\right)^{\otimes k} \to \left(M^d\right)^{\otimes(k-2)}.$$

In particular, the *full trace* then corresponds to

$$\begin{aligned} \mathrm{tr} := \mathrm{tr}_{1,\ldots,k} : \left(M^d\right)^{\otimes k} &\to \mathbb{C} \\ (Z_1 \otimes \cdots \otimes Z_k) &\mapsto \mathrm{tr}(Z_1)\ldots\mathrm{tr}(Z_k). \end{aligned}$$

Let us now return to the tensor space $\left(V^d\right)^{\otimes k}$ of vectors. We define the (symmetrizer) map $P_{\mathrm{Sym}^k} : \left(V^d\right)^{\otimes k} \to \left(V^d\right)^{\otimes k}$ via their action on elementary elements:

$$P_{\mathrm{Sym}^k}\left(z_1 \otimes \cdots \otimes z_k\right) := \frac{1}{k!}\sum_{\pi \in S_k} z_{\pi(1)} \otimes \cdots \otimes z_{\pi(k)}, \tag{7}$$

where S_k denotes the group of permutations of k elements. This map projects $\left(V^d\right)^{\otimes k}$ onto the totally symmetric subspace Sym^k of $\left(V^d\right)^{\otimes k}$ whose dimension [58] is

$$\dim \mathrm{Sym}^k = \binom{d+k-1}{k}. \tag{8}$$

3.3. **Complex projective designs.** The idea of (real) spherical designs originates in coding theory [23] and has been extended to more general spaces in [60, 61, 62]. We refer the interested reader to Levenshtein [62] for a unified treatment of designs in general metric spaces and from now on focus on designs in the complex vector space V^d.

Roughly speaking, a complex projective t-design is a finite subset of the complex unit sphere in V^d with the property that the discrete average of any polynomial of degree t or less equals its uniform average. Many equivalent definitions – see e.g. [60, 61, 43] – capture this essence. However, there is a more explicit definition of a t-design that is much more suitable for our purpose:

Definition 3 (Definition 2 in [26])**.** *A finite set $\{w_1, \ldots, w_N\} \subset V^d$ of normalized vectors is called a t-*design *of dimension d if and only if*

$$\frac{1}{N}\sum_{i=1}^{N}(w_i w_i^*)^{\otimes t} = \dim(\mathrm{Sym}^t)^{-1} P_{\mathrm{Sym}^t}, \tag{9}$$

where P_{Sym^t} denotes the projector onto the totally symmetric subspace (7) of $(V^d)^{\otimes t}$ and consequently $\dim(\mathrm{Sym}^t) = \binom{d+t-1}{t}$.

Note that the defining property (9) is invariant under global phase changes $w_i \mapsto \mathrm{e}^{i\phi} w_i$, thus it matches the symmetry of the phase retrieval problem. The definition above is equivalent to demanding

$$\frac{1}{N}\sum_{i=1}^{N}(w_i w_i^*)^{\otimes t} = \int_w \mathrm{d}w\,(ww^*)^{\otimes t},$$

where the right hand side is integrated with respect to the Haar measure. This form makes the statement that t-designs mimic the first $2t$ moments of Haar measure more explicit.

P. Seymor and T. Zaslavsky proved in [40] that t-designs on V^d exist for every $t, d \geq 1$, provided that N is large enough ($N \geq N(d,t)$), but they do not give an explicit construction. A necessary criterion – cf. [61, 43] – for the t-design property is that the number of vectors N obeys

$$N \geq \binom{d + \lceil t/2 \rceil - 1}{\lceil t/2 \rceil}\binom{d + \lfloor t/2 \lfloor - 1}{\lfloor t/2 \rfloor} = \mathcal{O}(d^{2t}). \tag{10}$$

However, the proof in [40] is non-constructive and known constructions are “innefficient” in the sense that the number of vectors required greatly exceeds (10). Hayashi et al. [28] proposed a construction requiring $\mathcal{O}(t)^d$ vectors. For real spherical designs other “inefficient” constructions have been proposed [38, 39] ($N = t^{\mathcal{O}(d^2)}$) which can be used to obtain complex projective designs.

Adressing this apparant lack of efficient constructions, Ambainis and Emerson [27] proposed the notion of *approximate desings*. These vector sets only fulfill property (9) only up to an ϵ-precision, but their great advantage is that they can be constructed efficiently. Concretely, they show that for every $d \geq 2t$, there exists an $\epsilon = \mathcal{O}(d^{-1/3})$ approximate t-design consisting of $\mathcal{O}(d^{3t})$ vectors only.

The great value of t-designs is due to the following fact: If we sample m vectors $a_i, \ldots, a_m$ iid from a t-design $D_t = \{w_1, \ldots, w_N\}$, the design property guarantees (with $A_i = a_i a_i^*$ and $W_i = w_i w_i^*$)

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} A_i^{\otimes k}\right] = \mathbb{E}\left[A_1^{\otimes k}\right] = \frac{1}{N}\sum_{i=1}^{N} W_i^{\otimes k} = \binom{d+k-1}{k}^{-1} P_{\mathrm{Sym}^k}$$

for all $1 \leq k \leq t$. This knowledge about the first t moments of the sampling procedure is the key ingredient for our partial derandomization of Gaussian PhaseLift [11].

3.4. **Large Deviation Bounds.** This approach makes heavy use of operator-valued large deviation bounds. They have been established first in the field of quantum information by Ahlswede and Winter [63]. Later the first author of this paper and his coworkers successfully applied these methods to the problem of low rank matrix recovery [15, 64]. By now these methods are widely used and we borrow them in their most recent (and convenient) form from Tropp [65, 66].

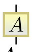
Theorem 4 (Uniform Operator Bernstein inequality, [65, 15])**.** *Consider a finite sequence $\{M_k\}$ of independent, random self-adjoint operators. Assume that each random variable satisfies $\mathbb{E}\left[M_k\right] = 0$ and $\|M_k\|_\infty \leq \overline{R}$ (for some finite constant $\overline{R}$) almost surely and define the norm of the total variance $\sigma^2 := \|\sum_k \mathbb{E}\left[M_k^2\right]\|_\infty$. Then the following chain of inequalities holds for all $t \geq 0$.*

$$\Pr\left[\|\sum_k M_k\|_\infty \geq t\right] \leq d\,\exp\left(-\frac{t^2/2}{\sigma^2 + \overline{R}t/3}\right) \leq \begin{cases} d\,\exp(-3t^2/8\sigma^2) & t \leq \sigma^2/\overline{R} \\ d\,\exp(-3t/8\overline{R}) & t \geq \sigma^2/\overline{R}. \end{cases}$$

Theorem 5 (Smallest Eigenvalue Bernstein Inequality, [66])**.** *Let* $S = \sum_k M_k$ *be a sum of iid random matrices* M_k *which obey* $\mathbb{E}[M_K] = 0$ *and* $\lambda_{min}(M_k) \geq -\underline{R}$ *almost surely for some fixed* $\underline{R}$*. With the variance parameter* $\sigma^2(S) = \| \sum_k \mathbb{E}\left[M_k^2\right] \|_\infty$ *the following chain of inequalities holds for all* $t \geq 0$*.*

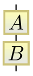
$$\Pr\left[\lambda_{\min}(S) \leq -t\right] \leq d \exp\left(-\frac{t^2/2}{\sigma^2 + \underline{R}t/3}\right) \leq \begin{cases} d\ \exp(-3t^2/8\sigma^2) & t \leq \sigma^2/\underline{R} \\ d\ \exp(-3t/8\underline{R}) & t \geq \sigma^2/\underline{R}. \end{cases}$$

3.5. **Wiring Diagrams.** The defining property (9) of t-designs is phrased in terms of tensor spaces. To work with these notions practically, we need tools for efficiently computing contractions between high-order tensors. The concept of *wiring diagrams* provides such a method – see [58] for an introduction and also [67, 68] (however, they use a slightly different notation). Here, we give a brief description that should suffice for our calculations.

Roughly, the calculus of wiring diagrams associates with every tensor a box, and with every index of that tensor a line emanating from the box. Two connected lines represent contracted indices. (More precisely, we place contravariant indices of a tensor on top of the associated box and covariant ones at the bottom. However, one should be able to digest our calculations without reference to this detail). A matrix $A : V^d \to V^d$ can be seen as a two-indexed tensor $A^i{}_j$. It will thus be represented by a node  with the upper line corresponding to the index i and the lower one to j. Two matrices A, B are multiplied by contracting B's "contravariant" index with A's "covariant" one:

$$(AB)^i{}_j = \sum_k A^i{}_k B^k{}_j$$

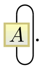
Pictographically, we write

$AB =$ 

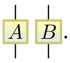
The trace operation

$$A \mapsto \text{tr}\, A = \sum_k A^k{}_k$$

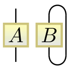
corresponds to a contraction of the two indices of a matrix:

$\text{tr}(A) =$  .

Tensor products are arranged in parallel:

$A \otimes B =$  .

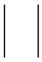
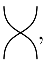
Hence, a partial trace takes the following form:

$\text{tr}_2\,(A \otimes B) =$  .

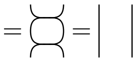
The last ingredient we need are the *transpositions* $\sigma_{(i,j)}$ on $(V^d)^{\otimes t}$ which act by interchanging the ith and the jth tensor factor. For example

$$\sigma_{(1,2)}\,(x \otimes y \otimes \cdots) = y \otimes x \otimes \cdots,$$

with $x, y \in V^d$ arbitrary. Transpositions suffice, because they generate the full group of permutations. For $\left(V^d\right)^{\otimes 2}$ we only have

$\underline{1} =$  (trivial permutation) and $\sigma_{(1,2)} =$  , but for higher tensor systems more permutations can occur. Consequently, permutations act by interchanging different input and output lines and the wiring diagram representation allows one to keep track of this pictorially. In fact, only the input and output position of a line matters. We can use diagrams to simplify expressions by disentangling the corresponding lines. Take $\sigma_{(1,2)}$ on $\left(V^d\right)^{\otimes 2}$ as an example. Using wiring diagrams we can derive the standard result

$$\sigma^2_{(1,2)} =$$



$$= \underline{1}$$

pictorially. We are now ready to prove some important auxiliary results.

Lemma 6. *Let $A, B \in H^d$ be arbitrary. Then it holds that*

$$\mathrm{tr}_2\left(P_{\mathrm{Sym}^2} A \otimes B\right) = \frac{1}{2}\left(\mathrm{tr}(B)A + BA\right). \tag{11}$$

We remark that in general,

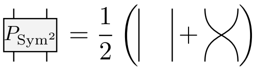
$$P_{\mathrm{Sym}^2}\left(X \otimes Y\right) \neq \frac{1}{2}\left(X \otimes Y + Y \otimes X\right),$$

which is, in our experience, a common misconception.

Proof of Lemma 6. The basic formula (7) for P_{Sym^2} is given by

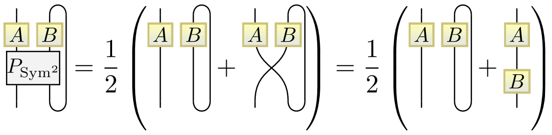
$$P_{\mathrm{Sym}^2} = \frac{1}{2}\sum_{\pi \in S_2} \sigma_{\pi(1),\pi(2)} = \frac{1}{2}\left(\underline{1} + \sigma_{(1,2)}\right),$$

and the concepts from above allow us to translate this into the following wiring diagram:



(Note that this operator acts on the full tensor space $\left(V^d\right)^{\otimes 2}$, hence in the wiring diagram it is represented by a two-indexed box.) Applying the graphical calculus yields

$$\mathrm{tr}_2\left(P_{\mathrm{Sym}^2} A \otimes B\right) =$$



$$= \frac{1}{2}\left(\mathrm{tr}(B)A + BA\right),$$

which is the desired result. □

Obviously, it is also possible to obtain (11) by direct calculation. We have included such a calculation in the appendix (Section 9.1) to demonstrate the complexity of direct calculations as compared to graphical ones.

We conclude this section with the following slightly more involved result.

Lemma 7. *Let $A, B, C \in H^d$ be arbitrary. Then it holds that*

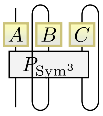
$$\begin{aligned} &\mathrm{tr}_{2,3}\left(P_{\mathrm{Sym}^3} A \otimes B \otimes C\right) \\ = \; &\frac{1}{6}\left(A\,\mathrm{tr}(B)\mathrm{tr}(C) + BA\,\mathrm{tr}(C) + CA\mathrm{tr}(B) + A\,\mathrm{tr}(BC) + CBA + BCA\right). \end{aligned} \tag{12}$$

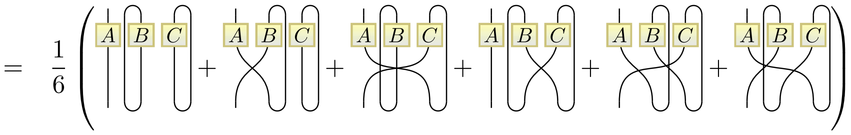
The proof can in principle be obtained by evaluating all permutations of 3 tensor systems algebraically and taking the partial trace afterwards. However, a pictorial calculation using wiring diagrams is much faster and more elegant.

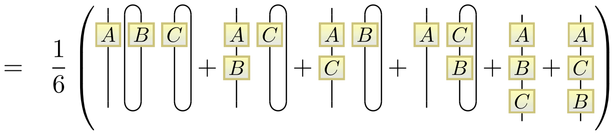
Proof. For permutations of three elements, formula (7) implies

$$P_{\mathrm{Sym}^3} = \frac{1}{6}\sum_{\pi \in S_3} \sigma_{\pi(1),\pi(2),\pi(3)} = \frac{1}{6}\left(\sigma_{1,2,3} + \sigma_{2,1,3} + \sigma_{3,2,1} + \sigma_{1,3,2} + \sigma_{2,3,1} + \sigma_{3,1,2}\right),$$

where. $\sigma_{2,1,3}(u \otimes v \otimes w) = (v \otimes u \otimes w)$, etc. This in turn allows us to write







$$= \frac{1}{6}\left(A\,\mathrm{tr}(B)\mathrm{tr}(C) + BA\,\mathrm{tr}(C) + CA\mathrm{tr}(B) + A\,\mathrm{tr}(BC) + CBA + BCA\right)$$

and we are done. □

4. Problem Setup

4.1. **Modelling the sampling process.** In the sampling process, we start by measuring the intensity of the signal:

$$y_0 = \|x\|_{\ell_2}^2 = \mathrm{tr}(\mathbb{1}X). \tag{13}$$

This allows us to assume w.l.o.g. $\|x\|_{\ell_2} = 1$. Next, we choose m vectors $a_1, \ldots, a_m$ iid at random from a t-design $D_t \subset V^d$ and evaluate

$$y_i = \mathrm{tr}(A_i X) = |\langle x, a_i\rangle|^2 \quad \text{for } i = 1, \ldots m, \tag{14}$$

and consequently the vector $y = (y_1, \ldots, y_m)^T \in \mathbb{R}_+^m$ captures all the information we obtain from the sampling process. This process can be represented by a measurement operator

$$\begin{aligned} \mathcal{A} : H^d &\to \mathbb{R}^m, \\ Z &\mapsto \sum_{i=1}^m \mathrm{tr}(A_i Z) e_i, \end{aligned} \tag{15}$$

where $e_1, \ldots, e_m$ denotes the standard basis of $\mathbb{R}^m$. Therefore $\mathcal{A}(X) = y$ completely encodes the measurement process. For technical reasons we also consider the measurement operator

$$\begin{aligned} \mathcal{R} : H^d &\to H^d, \\ Z &\mapsto m^{-1}\sum_{i=1}^m (d+1)d\,\Pi_{A_i} Z = m^{-1}\sum_{i=1}^m (d+1)d\,A_i\,\mathrm{tr}(A_i Z), \end{aligned} \tag{16}$$

which is a renormalized version of $\mathcal{A}^*\mathcal{A} : H^d \to H^d$. Concretely

$$\mathcal{R} = \frac{(d+1)d}{m}\mathcal{A}^*\mathcal{A}.$$

The scaling is going to greatly simplify our analysis, because it guarantees that $\mathcal{R}$ is "near-isotropic", as the following result shows.

Lemma 8 ($\mathcal{R}$ is near-isotropic)**.** *The operator* $\mathcal{R}$ *defined in (16) is* near-isotropic *in the sense that*

$$\mathbb{E}[\mathcal{R}] = \mathcal{I} + \Pi_{\mathbb{1}} \quad \textit{or} \quad \mathbb{E}\left[\mathcal{R}\right] Z = Z + \operatorname{tr}(Z)\mathbb{1} \quad \forall Z \in H^d \tag{17}$$

Proof. Let us start with deriving (17). For $Z \in H^d$ arbitrary we have

$$\begin{aligned} \mathbb{E}[\mathcal{R}]Z &= \frac{(d+1)d}{m}\sum_{i=1}^{m}\mathbb{E}[A_i \operatorname{tr}(A_i Z)] \\ &= (d+1)d \operatorname{tr}_2\left(\mathbb{E}[A_1^{\otimes 2}]\mathbb{1}\otimes Z\right) && (18) \\ &= 2\operatorname{tr}_2\left(P_{\mathrm{Sym}^2}\mathbb{1}\otimes Z\right) && (19) \\ &= Z + \mathbb{1}(\operatorname{tr} Z) = \left(\mathcal{I} + \Pi_{\mathbb{1}}\right)Z. \end{aligned}$$

Here, (18) follows from the fact that the a_i's are chosen iid from a t-design, (19) uses the fact that $\dim(\mathrm{Sym}^2) = \binom{d+1}{2}^{-1}$ together with Definition 3, and the final line is an application of Lemma 6. □

Let now $x \in V^d$ be the signal we want to recover. As in [11] we consider the space

$$T := \left\{xz^* + zx^* : \ z \in V^d\right\} \subset H^d \tag{20}$$

(which is the tangent space of the manifold of all hermitian matrices at the point $X = xx^*$). This space is of crucial importance for our analysis. The orthogonal projection onto this space can be given explicitly:

$$\begin{aligned} \mathcal{P}_T : H^d &\to T, \\ Z &\mapsto XZ + ZX - XZX && (21) \\ &= XZ + ZX - (X,Z)X. && (22) \end{aligned}$$

We denote the projection onto its orthogonal complement with respect to the Frobenius inner product by $\mathcal{P}_T^\perp$. Then for any matrix $Z \in H^d$ the decomposition

$$Z = \mathcal{P}_T Z + \mathcal{P}_T^\perp Z =: Z_T + Z_T^\perp$$

is valid. We point out that in particular

$$\mathcal{P}_T \Pi_{\mathbb{1}} \mathcal{P}_T = \Pi_X \tag{23}$$

holds. We will frequently use this fact. For a proof, consider $Z \in H^d$ arbitrary and insert the relevant definitions:

$$\begin{aligned} \mathcal{P}_T \Pi_{\mathbb{1}} \mathcal{P}_T Z &= \mathcal{P}_T \mathbb{1} \operatorname{tr}(\mathbb{1}\mathcal{P}_T Z) = (X\mathbb{1} + \mathbb{1}X - X\mathbb{1}X)\operatorname{tr}\left(XZ + ZX - XZX\right) \\ &= X\operatorname{tr}(XZ) = \Pi_X Z. \end{aligned}$$

4.2. **Convex Relaxation.** Following [3, 11, 12] the measurements (13) and (14) can be translated into matrix form by applying the following "lifts":

$$X := xx^*, \quad \text{and} \quad A_i := a_i a_i^*.$$

By doing so the measurements assume the a linear form:

$$\begin{aligned} y_0 &= \|x\|_2^2 = (\mathbb{1}, X) = \operatorname{tr}(X), \\ y_i &= (A_i, X) = \operatorname{Tr}\left(A_i X\right) \quad i = 1, \ldots, m. \end{aligned}$$

Hence, the phase retrivial problem becomes a matrix recovery problem. The solution to this is guaranteed to have rank 1 and encodes (up to a global phase) the unknown vector x via $X = xx^*$. Relaxing the rank minimization problem (which would output the correct solution) to a trace norm minimization yields the now-familiar convex optimization problem

$$\begin{array}{lll} \text{minarg}_{X'} & \|X'\|_1 & \\ \text{subject to} & (A_i, X') = y_i \quad i = 1, \ldots m, & \\ & X' = (X')^\dagger, & \\ & \operatorname{tr}(X') = 1, & \\ & X' \geq 0. & \end{array} \tag{24}$$

While this convex program is formally equivalent to the previously studied general-purpose matrix recovery algorithms [13, 14, 15], there are two important differences:

- The measurement matrices A_i are rank-1 projectors: $A_i = a_i a_i^*$.
- The unknown signal is known to be proportional to a rank-1 projector ($X = xx^*$) as well.

While the second fact is clearly of advantage for us, the first one makes the problem considerably harder: In the language of [15], it means that the "incoherence parameter" $\mu = d \max_{i=1,\ldots,m} \|A_i\|_\infty = d\|a_i\|_{\ell_2}^2 = d$ is as large as it can get! Higher values of μ correspond to more ill-posed problems and as a result, a direct application of previous low-rank matrix recovery results fails. It is this problem that Refs. [11, 12] first showed how to circumvent for the case of Gaussian measurements. Below, we will adapt these ideas to the case of measurements drawn from designs, which necessitates following more closely the approach of [15].

4.3. **Well-posedness / Injectivity.** In this section, we follow [11, 15] to establish a certain injectivity property of the measurement operator $\mathcal{A}$. Compared to [11], our injectivity properties are somewhat weaker. Their proof used the independence of the components of the Gaussian measurement operator, which is not available in this setting, where individual vector components might be strongly correlated. We will pay the price for these weaker bounds in Section 6. There, we construct an "approximate dual certificate" that proves that the sought-for signal indeed minimizes the nuclear norm. Owing to the weaker bounds found here, the construction is more complicated than in [11]. In the language of [15], we will have to carry out the full "golfing scheme", as opposed to the "single leg" that proved sufficient in [11].

Proposition 9. *With probability of failure smaller than $d^2 \exp(-\frac{3m}{384d})$ the inequality*

$$0.25 d^{-2} \|Z\|_2^2 < m^{-1} \|\mathcal{A}(Z)\|_2^2 \tag{25}$$

is valid for all matrices $Z \in T$ simultaneously.

Proof. We aim to show the more general statement

$$\Pr\left[m^{-1}\|\mathcal{A}(Z)\|_2^2 < 0.5(1-\delta)\|Z\|_2^2 \;\; \forall Z \in T\right] \leq d^2 \exp\left(-\frac{3m\delta^2}{96d}\right)$$

for any $\delta \in (0, 1)$.

For $Z \in T$ abritrary use near-isotropicity of $\mathcal{R}$ ($\mathbb{E}[\mathcal{R}] = \mathcal{I} + \Pi_{\mathbb{1}}$) and observe

$$\begin{aligned}
& m^{-1}\|\mathcal{A}(Z)\|_2^2 \\
= \;& m^{-1}\sum_{i=1}^{m}(\operatorname{tr}(ZA_i))^2 = \operatorname{tr}(Zm^{-1}\sum_i A_i \operatorname{tr}(A_i Z)) = \frac{1}{(d+1)d}\operatorname{tr}(Z\mathcal{R}Z) \\
= \;& \frac{1}{(d+1)d}\operatorname{tr}(Z(\mathcal{R}-\mathbb{E}[\mathcal{R}])Z) + \frac{1}{(d+1)d}\operatorname{tr}(Z(\mathcal{I}+\Pi_{\mathbb{1}})Z) \\
= \;& \frac{1}{(d+1)d}\operatorname{tr}(Z\mathcal{P}_T(\mathcal{R}-\mathbb{E}[\mathcal{R}])\mathcal{P}_T Z) + \frac{1}{(d+1)d}(\operatorname{tr}(Z^2) + (\operatorname{tr} Z)^2) \\
\geq \;& 0.5d^{-2}\left(\operatorname{tr}(Z\mathcal{P}_T(\mathcal{R}-\mathbb{E}[\mathcal{R}])\mathcal{P}_T Z) + \operatorname{tr}(Z^2)\right) \\
\text{(26)} \quad \geq \;& 0.5d^{-2}(1 + \lambda_{\min}\left(\mathcal{P}_T(\mathcal{R}-\mathbb{E}[\mathcal{R}])\mathcal{P}_T\right)\|Z\|_2^2,
\end{aligned}$$

where we have used $\mathcal{P}_T Z = Z$ as well as $\mathcal{M} \geq \lambda_{\min}(\mathcal{M})\mathcal{I}$ for any operator $\mathcal{M}$. Therefore everything boils down to bounding the smallest eigenvalue of $\mathcal{P}_T(\mathcal{R}-\mathbb{E}[\mathcal{R}])\mathcal{P}_T$. To this end we aim to apply Theorem 5 and decompose

$$\mathcal{P}_T(\mathcal{R}-\mathbb{E}[\mathcal{R}])\mathcal{P}_T = \sum_{i=1}^{m}(\mathcal{M}_i - \mathbb{E}[\mathcal{M}_i]) \quad \text{with} \quad \mathcal{M}_i = \frac{(d+1)d}{m}\mathcal{P}_T\Pi_{A_i}\mathcal{P}_T.$$

Note that these summands have mean zero by construction. Furthermore observe that the auxiliary result (23) implies

$$\begin{aligned}
-\frac{2}{m}\mathcal{I} &\leq -\frac{1}{m}\mathcal{I} - \frac{1}{m}\Pi_X \leq -\frac{1}{m}\mathcal{P}_T\mathcal{I}\mathcal{P}_T - \frac{1}{m}\mathcal{P}_T\Pi_{\mathbb{1}}\mathcal{P}_T \\
&= -\mathcal{P}_T\mathbb{E}[\mathcal{M}_i]\mathcal{P}_T \leq \mathcal{P}_T(\mathcal{M}_i - \mathbb{E}[\mathcal{M}_i])\mathcal{P}_T
\end{aligned}$$

and the a priori bound

$$\lambda_{\min}(\mathcal{M}_i - \mathbb{E}[\mathcal{M}_i]) \geq -2/m =: -\underline{R}$$

follows. For the variance we use the standard identity

$$0 \leq \mathbb{E}[(\mathcal{M}_i - \mathbb{E}[\mathcal{M}_i])^2] = \mathbb{E}[\mathcal{M}_i^2] - \mathbb{E}[\mathcal{M}_i]^2 \leq \mathbb{E}[\mathcal{M}_i^2]$$

and focus on the last expression. Writing it out explicitly yields

$$\begin{aligned}
0 \leq \mathbb{E}[\mathcal{M}_i^2] &= \frac{(d+1)^2 d^2}{m^2}\mathcal{P}_T\mathbb{E}\left[\Pi_{A_i}\mathcal{P}_T\Pi_{A_i}\right]\mathcal{P}_T \\
&= \frac{(d+1)^2 d^2}{m^2}\mathcal{P}_T\mathbb{E}\left[\operatorname{tr}(A_i\mathcal{P}_T A_i)\Pi_{A_i}\right]\mathcal{P}_T.
\end{aligned}$$

The trace can be bounded from above by

$$\begin{aligned}
\operatorname{tr}(A_i\mathcal{P}_T A_i) &= \operatorname{tr}\left(A_i(XA_i + A_i X - \operatorname{tr}(A_i X)X)\right) \\
&= 2\operatorname{tr}(A_i X) - \operatorname{tr}(A_i X)^2 \leq 2\operatorname{tr}(A_i X),
\end{aligned}$$

where we have used the basic definition of $\mathcal{P}_T$ and $0 \leq \operatorname{tr}(A_i X) = |\langle a_i, x\rangle|^2 \leq 1$. Consequently, for $Z \in T$ arbitrary

$$\begin{aligned}
& \mathcal{P}_T \mathbb{E}[\mathcal{M}_i^2] \mathcal{P}_T Z \\
\leq \; & \frac{2(d+1)^2 d^2}{m^2} \mathcal{P}_T \mathbb{E}\left[A_i \operatorname{tr}(A_i X) \operatorname{tr}(A_i Z)\right] \\
= \; & \frac{2(d+1)^2 d^2}{m^2} \mathcal{P}_T \operatorname{tr}_{2,3}\left(\mathbb{E}[A_i^{\otimes 3}] \mathbb{1} \otimes X \otimes Z\right) \\
= \; & \frac{12(d+1)^2 d^2}{m^2 (d+2)(d+1)d} \mathcal{P}_T \operatorname{tr}_{2,3}\left(P_{\mathrm{Sym}^3} \mathbb{1} \otimes X \otimes Z\right) \\
\leq \; & \frac{2d}{m^2} \mathcal{P}_T \left(\mathbb{1} \operatorname{tr}(Z) + X \operatorname{tr}(Z) + Z + \mathbb{1} \operatorname{tr}(XZ) + ZX + XZ\right) \\
= \; & \frac{2d}{m^2} \left(X \operatorname{tr}(XZ) + X \operatorname{tr}(XZ) + Z + X \operatorname{tr}(XZ) + \mathcal{P}_T Z + X \operatorname{tr}(XZ)\right) \\
= \; & \frac{2d}{m^2} \left(4\Pi_X + 2\mathcal{I}\right) Z \leq \frac{12d}{m^2} \mathcal{I} Z.
\end{aligned}$$

Here we have applied $\dim \mathrm{Sym}^3 = \binom{d+2}{3}^{-1}$ and Lemma 7 in lines 3 and 4, respectively. Furthermore we used $Z \in T$ – hence $\mathcal{P}_T Z = Z$ and $\operatorname{tr}(Z) = \operatorname{tr}(XZ)$ – as well as the basic definition (22) of $\mathcal{P}_T$ to simplify the terms occuring in the fourth line. Putting everything together yields

$$\mathbb{E}[(\mathcal{M}_i - \mathbb{E}[\mathcal{M}_i])^2] \leq \mathbb{E}[\mathcal{M}_i^2] \leq \frac{12d}{m^2} \mathcal{I}$$

and we can safely set $\sigma^2 := \frac{12d}{m}$. Now Theorem 5 tells us

$$\Pr\left[\lambda_{\min}\left(\mathcal{P}_T(\mathcal{R} - \mathbb{E}[\mathcal{R}])\mathcal{P}_T\right) \leq -\delta\right] \leq d^2 \exp\left(-\frac{3m\delta^2}{8 \times 12d}\right)$$

for all $0 \leq \delta \leq 1 \leq 6d = \sigma^2 / \underline{R}$. This gives the desired bound on the event

$$\{\lambda_{\min}(\mathcal{P}_T(\mathcal{R} - \mathbb{E}[\mathcal{R}])\mathcal{P}_T) \leq -\delta\}$$

occuring. If this is not the case, (26) implies

$$m^{-1} \|\mathcal{A}(Z)\|_{\ell_2}^2 > 0.5 d^{-2} (1-\delta) \|Z\|_2^2$$

for all matrices $Z \in T$ simultaneously. This is the general statement at the beginning of the proof and setting $\delta = 1/2$ yields Proposition 9. □

Proposition 10. *Let $\mathcal{A}$ be as above with vectors sampled from a t-design ($t \geq 1$). Then the statement*

$$m^{-1} \|\mathcal{A}(Z)\|_{\ell_2}^2 \leq \|Z\|_2^2 \tag{27}$$

holds with probability one for all matrices $Z \in H^d$ simultaneously.

Proof. Pick $Z \in H^d$ arbitrary and observe

$$\|\mathcal{A}(Z)\|_{\ell_2}^2 = \frac{1}{m} \sum_{i=1}^m (\operatorname{tr}(A_i Z))^2 = \operatorname{tr}\left(Z \left(\frac{1}{m} \sum_{i=1}^m \Pi_{A_i}\right) Z\right) \leq \operatorname{tr}(Z \mathcal{I} Z) = \|Z\|_2^2,$$

where we have used $0 \leq \Pi_{A_i} \leq \mathcal{I}$. □

Note that equation (27) can be improved. Indeed, a standard application of the Operator Bernstein inequality (Theorem 4) gives

$$m^{-1}\|\mathcal{A}(Z)\|_{\ell_2}^2 \leq 2d^{-1}\|Z\|_2^2$$

for all matrices $Z \in T$ with probability of failure smaller than $d^2 \exp(-Cm/d)$ for some $0 < C \leq 1$. However, we actually do not require this tighter bound.

5. Proof of the Main Theorem / Convex Geometry

In this section, we will follow [15, 14] to prove that the convex program (24) indeed recovers the sought for signal x, provided that a certain geometric object – an *approximate dual certificate* – exists.

Definition 11 (Approximate dual certificate)**.** *Assume that the sampling process corresponds to (13) and (14). Then we call $Y \in H^d$ an* approximate dual certificate*, provided that $Y \in \operatorname{span}(\mathbb{1}, A_1, \ldots, A_m)$ and*

$$\|Y_T - X\|_2 \leq \frac{1}{4d} \quad \textit{as well as} \quad \|Y_T^\perp\|_\infty \leq \frac{1}{2}. \tag{28}$$

Proposition 12. *Suppose that the measurement gives us access to $\|x\|_{\ell_2}^2$ and $y_i = |\langle a_i, x\rangle|^2$ for $i = 1, \ldots, m$. Then the convex optimization (24) recovers the unknown x (up to a global phase) provided that (25) holds and an approximate dual certificate Y exists.*

Proof. Let $\tilde{X} \in H^d$ be an arbitrary feasible point of (24) and decompose it as $\tilde{X} = X + \Delta$. Feasibility then implies $\mathcal{A}(\tilde{X}) = \mathcal{A}(X)$ and $\mathcal{A}(\Delta) = 0$ must in turn hold for any feasible displacement Δ. Now the pinching inequality [69] (Problem II.5.4) implies

$$\|\tilde{X}\|_1 = \|X + \Delta\|_1 \geq \|X\|_1 + \operatorname{tr}(\Delta_T) + \|\Delta_T^\perp\|_1.$$

Consequently X is guaranteed to be the unique minimum of (24), if

$$\operatorname{tr}(\Delta_T) + \|\Delta_T^\perp\|_1 > 0 \tag{29}$$

is true for every feasible Δ. In order to show this we combine feasibility of Δ with inequalities (25) and (27) to obtain

$$\|\Delta_T\|_2 < 2dm^{-1/2}\|\mathcal{A}(\Delta_T)\|_{\ell_2} = 2dm^{-1/2}\|\mathcal{A}(\Delta_T^\perp)\|_{\ell_2} \leq 2d\|\Delta_T^\perp\|_2. \tag{30}$$

Feasibility of Δ also implies $(Y, \Delta) = 0$, because by defnition Y is in the range of $\mathcal{A}^*$. Combining this insight with the defining property (28) of Y and (30) yields

$$\begin{aligned} 0 &= (Y,\Delta) = (Y_T - X, \Delta_T) + (X, \Delta_T) + (Y_T^\perp, \Delta_T^\perp) \\ &\leq \|Y_T - X\|_2\|\Delta_T\|_2 + \operatorname{tr}(\Delta_T) + \|Y_T^\perp\|_\infty\|\Delta_T^\perp\|_1 \\ &< \operatorname{tr}(\Delta_T) + \|Y_T - X\|_2 2d\|\Delta_T^\perp\|_2 + \|Y_T^\perp\|_\infty\|\Delta_T^\perp\|_1 r \\ &\leq \operatorname{tr}(\Delta_T) + 1/2\|\Delta_T^\perp\|_2 + 1/2\|\Delta_T^\perp\|_1 \\ &\leq \operatorname{tr}(\Delta_T) + \|\Delta_T^\perp\|_1, \end{aligned}$$

which is just the desired optimality criterion (29). □

6. Constructing the Dual Certificate

A straightforward approach to construct an approximate dual certificate would be to set (31)

$$Y = \mathcal{R}X - \operatorname{tr}(X)\mathbb{1} = \frac{(d+1)d}{m}\sum_{i=1}^{m} A_i \operatorname{tr}(A_i X) - \operatorname{tr}(X)\mathbb{1} \in \operatorname{span}(\mathbb{1}, A_1, \ldots, A_m).$$

In expectation, $\mathbb{E}[Y] = X$, which is the "perfect dual certificate" in the sense that the norm bounds in (28) vanish. The hope would be to use the Operator Bernstein inequality to show that with high probablity, Y will be sufficiently close to its expectation. It has been shown that a slight refinement of the ansatz (31) indeed achieves this goal Ref. [15, 70]. However, the Bernstein bounds depend on the worst-case operator norm of the summands. In our case, they can be as large as $d^2|\langle a_i, x\rangle|^2$, which can reach d^2. This is far larger than in previous low-rank matrix recovery problems. Ref. [11] relied on the fact that large overlaps $|\langle a_i, x\rangle|^2 \gg \mathcal{O}(d^{-1})$ are "rare" for Gaussian a_i.

The key observation here is that the t-design property provides one with useful information about the first t moments of the random variable $|\langle x, a_i\rangle|^2$. This knowledge allows us to explicitly bound the probability of "dangerously large overlaps" or "coherent measurement vectors" occurring.

Lemma 13 (Undesired events)**.** *Let $x \in V^d$ be an arbitrary vector of unit length. If a is chosen uniformly at random from a t-design ($t \geq 1$) $D_t \subset V^d$, then the following is true for every $\gamma \leq 1$:*

$$\Pr\left[|\langle a, x\rangle|^2 \geq 5td^{-\gamma}\right] \leq 4^{-t} d^{-t(1-\gamma)}. \tag{32}$$

Proof. We aim to prove the slightly more general statement

$$\Pr\left[|\langle a, x\rangle|^2 \geq (\delta+1)td^{-\gamma}\right] \leq \delta^{-t} d^{-t(1-\gamma)},$$

which is valid for any $\delta \geq 1$. Setting $\delta = 4$ then yields (32). The t-design property provides us with useful information about the first t moments of the non-negative random variable $\xi = |\langle a, x\rangle|^2$. Indeed, with $A = aa^*$ it holds for every $k \leq t$ that

$$\begin{aligned}
\mathbb{E}\left[\xi^k\right] &= \mathbb{E}\left[\operatorname{tr}(AX)^k\right] \\
&= \operatorname{tr}\left(\mathbb{E}\left[A^{\otimes k}\right] X^{\otimes k}\right) \\
&= \binom{d+k-1}{k}^{-1} \operatorname{tr}\left(P_{\mathrm{Sym}^k} X^{\otimes k}\right) \\
&= \binom{d+k-1}{k}^{-1} \operatorname{tr}\left(X^{\otimes k}\right) \\
&\leq d^{-k} k!,
\end{aligned}$$

because $X^{\otimes k}$ is invariant under P_{Sym^k}. One way of seing this[5] is to note that $\operatorname{range}(X^{\otimes k}) = \operatorname{span}(x^{\otimes k})$ and the latter is already contained in Sym^k. Therefore the k-th moment τ_k of ξ is bounded by

$$\tau_k = \left(\mathbb{E}[\xi^k]\right)^{1/k} \leq (d^{-k}k!)^{1/k} \leq k/d.$$

These inequalities are tight for the mean $\mu = \tau_1$ of ξ and hence

$$\mu = \mathbb{E}[\xi] = d^{-1}.$$

[5] Alternatively one could also rearange tensor systems: $X^{\otimes k} = (xx^*)^{\otimes k} \simeq x^{\otimes k}(x^*)^{\otimes k}$ and use $P_{\mathrm{Sym}^k} x^{\otimes k} = x^{\otimes k}$.

Now we aim to use the well-known t-th moment bound

$$\Pr\left[|\xi - \mu| \geq s\tau_t\right] \leq s^{-t},$$

which is a straightforward generalization of Chebyshev's inequality. Applying it, yields the desired result. Indeed,

$$\begin{aligned}\Pr\left[|\langle a, x\rangle|^2 \geq (\delta+1)td^{-\gamma}\right] &= \Pr\left[\xi - \mu \geq (\delta+1)td^{-\gamma} - d^{-1}\right] \\ &\leq \Pr\left[\xi - \mu \geq \delta t d^{-\gamma}\right] \\ &\leq \Pr\left[|\xi - \mu| \geq \delta d^{1-\gamma}\tau_t\right] \\ &\leq \delta^{-t} d^{-t(1-\gamma)},\end{aligned}$$

and we are done. □

The previous lemma bounds the probability of the undesired events

$$E_i^c = \left\{|\langle a_i, x\rangle|^2 \geq 5td^{-\gamma}\right\}, \tag{33}$$

where $0 \leq \gamma \leq 1$ is a fixed parameter which we refer to as the *truncation rate*. It turns out that a single truncation of this kind does not quite suffice yet for our purpose. We need to introduce a second truncation step.

Definition 14. *Fix $Z \in T$ arbitrary and decompose it as*

$$Z = \zeta\left(xz^* + zx^*\right),$$

for some unique $\zeta > 0$ and $z \in V^d$ with $\|z\|_{\ell_2} = 1$. For this z we introduce the event

$$G_i^c := \left\{|\langle z, a_i\rangle|^2 \geq 5td^{-\gamma}\right\}$$

and define the two-fold truncated operator

$$\mathcal{R}_Z := \mathcal{R}_z = \frac{(d+1)d}{m}\sum_{i=1}^m 1_{E_i} 1_{G_i} \Pi_{A_i}, \tag{34}$$

where 1_{E_i} and 1_{G_i} denote the indicator functions associated with the events E_i and G_i, respectively.

The following result shows that due to Lemma 13 this truncated operator is in expectation close to the original $\mathcal{R}$.

Proposition 15. *Fix $Z \in T$ arbitrary and let $\mathcal{R}_Z$ be as in (34). Then*

$$\|\mathbb{E}[\mathcal{R}_Z - \mathcal{R}]\|_{\mathrm{op}} \leq 4^{1-t} d^{2-t(1-\gamma)} \tag{35}$$

Proof. We start by introducing the auxiliar (singly truncated) operator

$$\mathcal{R}_{\mathrm{aux}} := \frac{(d+1)d}{m}\sum_{i=1}^m 1_{E_i}\Pi_{A_i}$$

and observe

$$\|\mathbb{E}\left[\mathcal{R}_Z - \mathcal{R}\right]\|_{\mathrm{op}} \leq \|\mathbb{E}\left[\mathcal{R} - \mathcal{R}_{\mathrm{aux}}\right]\|_{\mathrm{op}} + \|\mathbb{E}\left[\mathcal{R}_Z - \mathcal{R}_{\mathrm{aux}}\right]\|_{\mathrm{op}}. \tag{36}$$

Now use Lemma 13 to bound the first term:

$$\begin{aligned} \|\mathbb{E}[\mathcal{R}-\mathcal{R}_{\text{aux}}]\|_{\text{op}} &= \left\| \frac{(d+1)d}{m}\sum_{i=1}^m \mathbb{E}\left[(1-1_{E_i})\Pi_{A_i}\right]\right\|_{\text{op}} \\ &\leq \frac{(d+1)d}{m}\sum_{i=1}^m \mathbb{E}\left[1_{E_i^c}\|\Pi_{A_i}\|_{\text{op}}\right] \\ &\leq \frac{2d^2}{m}\sum_{i=1}^m \mathbb{E}\left[1_{E_i^c}\right] = \frac{2d^2}{m}\sum_{i=1}^m \Pr[E_i^c] \\ &\leq 2d^2\times 4^{-t}d^{-t(1-\gamma)} = 2^{1-2t}d^{2-t(1-\gamma)}. \end{aligned}$$

Similarily,

$$\begin{aligned} \|\mathbb{E}\left[\mathcal{R}_{\text{aux}}-\mathcal{R}_Z\right]\|_{\text{op}} &= \frac{(d+1)d}{m}\left\|\sum_{i=1}^m \mathbb{E}\left[1_{G_i^c}\Pi_{A_i}\right]\right\|_{\text{op}} \leq \frac{2d^2}{m}\sum_{i=1}^m \mathbb{E}[1_{G_i^c}] \\ &\leq \frac{2d^2}{m}\sum_{i=1}^m \Pr[G_i^c] \leq 2^{1-2t}d^{2-t(1-\gamma)} \end{aligned}$$

and inserting these bounds into (36) yields the desired statement. □

We now establish a technical result which will allow us to find a suitable approximate dual certificate using the "golfing scheme" construction [15, 70].

Proposition 16. *Fix $Z \in T$ arbitrary, let $\mathcal{R}_Z$ be as in (34). Assume that that the design order t is at least 3 and the truncation rate γ satisfies*

$$\gamma \leq 1 - 2/t.$$

Then for $1/4 \leq b \leq 1$ and $c \geq \sqrt{2}b$ with probability at least $1 - d\exp(-\frac{9mb}{640td^{2-\gamma}})$ one has

$$\begin{aligned} &(37) & \|\mathcal{P}_T^\perp\left(\mathcal{R}_Z Z - \operatorname{tr}(Z)\mathbb{1}\right)\|_\infty &\leq b\|Z\|_2 \quad \textit{and} \\ &(38) & \|\mathcal{P}_T\left(\mathcal{R}_Z - Z - \operatorname{tr}(Z)\mathbb{1}\right)\|_2 &\leq c\|Z\|_2. \end{aligned}$$

Proof. The statement is invariant under rescaling of Z. Therefore it suffices to treat the case $\|Z\|_2 = 1$. In this case we can decompose

$$Z = \zeta(zx^* + xz^*)$$

with some fixed $z \in V^d$ obeying $\|z\|_{\ell_2} = 1$ and $0 < \zeta \leq 1$. Near-Isotropicity (Lemma 8) of $\mathcal{R}$ guarantees $\mathcal{P}_T^\perp \mathbb{E}[\mathcal{R}]Z = \operatorname{tr}(Z)\mathcal{P}_T^\perp Z$ as well as $\mathcal{P}_T\mathbb{E}[\mathcal{R}]Z = Z + \operatorname{tr}(Z)\mathcal{P}_T\mathbb{1}$. Let us now focus on (37) and use Proposition 15 in order to write

$$\begin{aligned} & \|\mathcal{P}_T^\perp\left(\mathcal{R}_Z Z - \operatorname{tr}(Z)\mathbb{1}\right)\|_\infty \\ = \;& \|\mathcal{P}_T^\perp\left(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}]\right)Z\|_\infty \\ \leq \;& \|\mathcal{P}_T^\perp\left(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}_Z]\right)Z\|_\infty + \|\mathcal{P}_T^\perp\mathbb{E}[\mathcal{R}_Z - \mathcal{R}]Z\|_\infty \\ \leq \;& \|\mathcal{P}_T^\perp\|_{\text{op}}\|(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}_Z])Z\|_\infty + 4^{1-t}d^{2-t(1-\gamma)}\|\mathcal{P}_T^\perp\|_{\text{op}}\|Z\|_2 \\ \leq \;& \|(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}_Z])Z\|_\infty + b/4. \end{aligned}$$

Here we have used $\|\mathcal{P}_T^\perp\|_{\text{op}} \leq 1$ as well as

$$(39) \qquad \|\mathbb{E}\left[\mathcal{R}_Z - \mathcal{R}\right]\|_{\text{op}} \leq 4^{1-t}d^{2-t(1-\gamma)} \leq 4^{1-t} \leq 1/16 \leq b/4,$$

which follows from $\gamma \leq 1 - 2/t$, $t \geq 3$ and $b \geq 1/4$. To obtain (38) we use a similar reasoning:

$$\begin{aligned} & \|\mathcal{P}_T\left(\mathcal{R}_Z Z - Z - \operatorname{tr}(Z)\mathbb{1}\right)\|_2 \\ = \; & \|\mathcal{P}_T\left(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}]\right) Z\|_2 \\ \leq \; & \sqrt{2}\|\mathcal{P}_T\left(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}_Z]\right) Z\|_\infty + \|\mathcal{P}_T \mathbb{E}[\mathcal{R}_Z - \mathcal{R}]Z\|_2 \\ \leq \; & \sqrt{2}\|\mathcal{P}_T\|_{\text{op}}\|(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}_Z])Z\|_\infty + b/4\|\mathcal{P}_T\|_{\text{op}}\|Z\|_2 \\ \leq \; & \sqrt{2}\|\left(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}_Z]\right) Z\|_\infty + b/4, \end{aligned}$$

where we have used the fact that $\mathcal{P}_T$ projects onto a subspace of at most rank-2 matrices in the third line and (43) in the fourth. This motivates to define the event

$$E := \{\|\left(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}_Z]\right) Z\|_\infty \leq 3b/4\}$$

which guarantees both (37) and (38) due to the assumption on c and $\|Z\|_2 = 1$. So everything boils down to bounding the probability of E^c. We decompose

$$\left(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}_Z]\right) Z = \sum_{i=1}^m \left(M_i - \mathbb{E}[M_i]\right) \quad \text{with} \quad M_i = \frac{(d+1)d}{m} 1_{E_i} 1_{G_i} A_i \operatorname{tr}(A_i Z).$$

We will estimate this sum using the Operator Bernstein inequality (Theorem 4). Thus we need an a priori bound for the summands

$$\begin{aligned} \|M_i\|_\infty &= \frac{(d+1)d}{m} 1_{E_i} 1_{G_i} \|A_i\|_\infty |\operatorname{tr}(A_i Z)| \leq \frac{2d^2}{m} 1_{E_i} 1_{G_i} 2|\langle x, a_i\rangle||\langle z, a_i\rangle| \\ &\leq \frac{4d^2}{m} 5td^{-\gamma} = \frac{20}{m} td^{2-\gamma} =: \overline{R}, \end{aligned}$$

as well as a bound for the variance. First observe that

$$\mathbb{E}[(M_i - \mathbb{E}[M_i])^2] = \mathbb{E}\left[M_i^2\right] - \mathbb{E}[M_i]^2 \leq \mathbb{E}\left[M_i^2\right].$$

and therefore

$$\begin{aligned} & \mathbb{E}\left[M_i^2\right] \\ = \; & \frac{(d+1)^2 d^2}{m^2} \mathbb{E}\left[1_{E_i} 1_{G_i} \operatorname{tr}(A_i Z)^2 A_i^2\right] \leq \frac{(d+1)^2 d^2}{m^2} \mathbb{E}\left[\operatorname{tr}(A_i Z)^2 A_i^2\right] \\ = \; & \frac{(d+1)^2 d^2}{m^2} \operatorname{tr}_{2,3}\left(\mathbb{E}[A_i^{\otimes 3}] \mathbb{1} \otimes Z \otimes Z\right) = \frac{6(d+1)d}{m^2(d+2)} \operatorname{tr}_{2,3}\left(P_{\text{Sym}^3} \mathbb{1} \otimes Z \otimes Z\right) \\ \leq \; & \frac{d}{m^2}\left(\mathbb{1}\operatorname{tr}(Z)^2 + Z\operatorname{tr}(Z) + Z + \mathbb{1}\operatorname{tr}(Z^2) + 2Z^2\right) \\ \leq \; & \frac{8d}{m^2}\|Z\|_2^2 \mathbb{1} = \frac{8d}{m^2}\mathbb{1}. \end{aligned}$$

Here we have used $\operatorname{tr}(Z) \leq \sqrt{2}\|Z\|_2$, $Z^2 \leq \|Z\|_2^2 \mathbb{1}$ and $\|Z\|_2 = 1$. From this we can conclude

$$\Big\|\sum_i \mathbb{E}[(M_i - \mathbb{E}[M_i])^2\Big\|_\infty \leq m \max_{i=1,\ldots,m} \|\mathbb{E}[M_i^2]\|_\infty \leq \frac{8d}{m} =: \sigma^2.$$

Observing that

$$\frac{\sigma^2}{\overline{R}} \leq \frac{8}{20t} d^{\gamma-1} \leq \frac{2}{15} \leq \frac{3}{4} b,$$

Theorem 4 yields

$$\Pr\left[E^c\right] = \Pr\left[\| \left(\mathcal{R}_Z - \mathbb{E}[\mathcal{R}_Z]\right) Z\|_\infty > 3b/4\right] \leq d \exp\left(- \frac{3 \times 3mb}{8 \times 4 \times 20td^{2-\gamma}} \right),$$

as desired. □

With this ingredient we can now construct a suitable approximate dual certificate Y, closely following [70].

Proposition 17. *Let $x \in V^d$ be an arbitrary normalized vector ($\|x\|_{\ell_2} = 1$), $X = xx^*$ and let $\omega \geq 1$ be arbitrary. If the design order t ($t \geq 3$) and the truncation rate γ is chosen such that*

$$\gamma \leq 1 - 2/t$$

holds and the total number of measurements fulfills

$$m \geq C\omega t d^{2-\gamma} \log^2(d), \tag{40}$$

then with probability larger than $1 - 0.5\mathrm{e}^{-\omega}$, there exists an approximate dual certificate Y as in Def. 11. Here, C is a universal constant (which can in principle be recoverd explicitly from the proof).

Proof. The randomzied construction of Y is summarized in Algorithm 1. If this algorithm succeeds, it outputs three lists

$$\mathbf{Y} = [Y_1, \ldots, Y_r], \qquad \mathbf{Q} = [X, Q_1, \ldots, Q_r], \qquad \text{and} \quad \xi = \{\xi_1, \ldots, \xi_l\}.$$

The recursive construction yields the following expressions (c.f. [70, Lemma 14]):

$$\begin{aligned} Y &:= Y_r = \mathcal{R}_{Q_{r-1}} Q_{r-1} - \mathrm{tr}(Q_{r-1})\mathbb{1} + Y_{r-1} \\ &= \sum_{i=1}^r \left(\mathcal{R}_{Q_{i-1}} Q_{i-1} - \mathrm{tr}(Q_{i-1})\mathbb{1} \right) \quad \text{and} \\ Q_i &= X - \mathcal{P}_T Y_i = \mathcal{P}_T \left(Q_{i-1} + \mathrm{tr}(Q_{i-1})\mathbb{1} - \mathcal{R}_{Q_{i-1}} Q_{i-1} \right) \\ &= \mathcal{P}_T \left(\mathcal{I} + \Pi_{\mathbb{1}} - \mathcal{R}_{Q_{i-1}} \right) Q_{i-1} = \cdots = \prod_{j=1}^i \mathcal{P}_T \left(\mathcal{I} + \Pi_{\mathbb{1}} - \mathcal{R}_{Q_{j-1}} \right) X. \end{aligned}$$

We now set

$$r = \lceil \log_2 d \rceil + 2. \tag{41}$$

[6] The use of pseudo-code allows for a compact presentation of this randomized procedure. However, the reader should keep in mind that the construction is purely part of a proof and should not be confused with the recovery algorithm (which is given in Eq. (24)).

Algorithm 1: Summary of the randomized “golfing scheme” [15] used in the proof of Prop. 17 to show the existence of an approximate dual certificate[6].

```
Input:
    $X \in H^d$                    # signal to be recovered
    $l \in \mathbb{N}$                    # maximum number of iterations
    $\{m_i\}_{i=1}^{l} \subset \mathbb{N}$          # number of measurement vectors used in $i$th iteration
    $r$                          # require $r$ “successful” iterations
                                 # (i.e. iterations where we enter the inner if-block)
Initialize:
    $\mathbf{Y} = []$                   # a list of matrices in $H^d$, initially empty
    $\mathbf{Q} = [X]$                  # a list of matrices in $T$, initialized to hold $X$ as its only element
    $i = 1$                      # number of current iteration
    $\xi = [0, \ldots, 0]$             # array of $l$ zeros; $\xi_i$ will be set to 1 if $i$th iteration succeeds
Body:
  while $i \leq l$ and $\sum_{j=1}^{i} \xi_j \leq r$ do
    set $Q$ to be the last element of $\mathbf{Q}$ and $Y$ to be the last element of $\mathbf{Y}$,
     sample $m_i$ vectors uniformly from the $t$-design; construct $\mathcal{R}_Q$ according to
     Def. 14.
    if (37), (38) hold for $\mathcal{R}_Q$ and $Q \in T$ with parameters $b = 1/8$, $c = 1/2$ then
      $\xi_i = 1$
      $Y \leftarrow \mathcal{R}_Q Q - \mathrm{tr}(Q)\mathbb{1} + Y$,    append $Y$ to $\mathbf{Y}$
      $Q \leftarrow X - \mathcal{P}_T Y$,    append $Q$ to $\mathbf{Q}$
    end
    $i \leftarrow i + 1$
  end
  if $\sum_{i=1}^{l} \xi_i = r$ then
    report success and output $\mathbf{Y}, \mathbf{Q}, \xi$
  else
    report failure
  end
```

Then, in case of success, the validity of properties (37) and (38) for $c = 1/2$ and $b = 1/8$ in each step ($Q_i \to Q_{i+1}$ and $Y_i \to Y_{i+1}$, respectively) guarantee

$$\begin{aligned}
\|Y_T - X\|_2 &= \|Q_r\|_2 \leq \|X\|_2 \prod_{j=1}^{r} \frac{1}{2} = 2^{-\lceil \log_2 d \rceil - 2} \|X\|_2 \leq \frac{1}{4d}, \\
\|Y_T^{\perp}\|_\infty &\leq \sum_{i=1}^{r} \left\| \mathcal{P}_T^{\perp} \left(\mathcal{R}_{Q_{i-1}} Q_{i-1} - \mathrm{tr}(Q_{i-1})\mathbb{1} \right) \right\|_\infty \\
&\leq \sum_{i=1}^{r} \frac{1}{8} \|Q_{i-1}\|_2 \leq \frac{1}{8} \sum_{i=1}^{r} 2^{1-i} \|Q_0\|_\infty \\
&\leq \|X\|_\infty \frac{1}{8} \sum_{i=0}^{\infty} 2^{-i} = \frac{1}{4} \leq \frac{1}{2}.
\end{aligned}$$

Thus, Y_r constitutes an approximate dual certificate in the sense of Def. 11.

What remains to be done is to choose the parameters l and $\{m_i\}_{i=1}^{l}$ such that the probability of the algorithm failing is smaller than $0.5\mathrm{e}^{-\omega}$. Algorithm 1 fails precisely if

$$\sum_{i=1}^{l} \xi_i < r. \tag{42}$$

Recall that the ξ_i's are Bernoulli random variables which indicate whether the i-th iteration of the algorithm has been succesful ($\xi_i = 1$), or failed ($\xi_i = 0$). Our aim is to bound the probability of the event in (42) by a similar expression involving *independent*[7] Bernoulli variables ξ_i'. To this end, write

$$\Pr\left[\sum_{i=1}^{l} \xi_i < r\right] = \mathbb{E}\left[\Pr\left[\xi_l < r - \sum_{i=1}^{l-1} \xi_i \,\Big|\, \xi_{l-1}, \ldots, \xi_1\right]\right]. \tag{43}$$

Conditioned on an arbitrary instance of $\xi_{l-1}, \ldots, \xi_1$, the variable ξ_l follows a Bernoulli distribution with some parameter $p(\xi_{l-1}, \ldots, \xi_1)$. Note that if $\xi \sim \mathrm{B}(p)$ is a Bernoulli variable with parameter p, then for every fixed $t \in \mathbb{R}$, the probability $\Pr_{\xi\sim\mathrm{B}(p)}[\xi < t]$ is non-increasing as a function of p. Consequently, the estimate

$$\Pr\left[\sum_{i=1}^{l} \xi_i < r\right] \leq \Pr\left[\xi_l' + \sum_{i=1}^{l-1} \xi_i < r\right] \tag{44}$$

is valid if ξ_l' is an independent p'-Bernoulli distributed with

$$p' \leq \min_{\xi_{l-1}, \ldots, \xi_1} p(\xi_{l-1}, \ldots, \xi_1).$$

Proposition 16 provides a uniform lower bound on the success probability $p(\xi_{l-1}, \ldots, \xi_1)$. Indeed, there is a universal constant C_1 such that invoking Prop. 16 with

$$m := C_1 t d^{2-\gamma} \log d$$

and $Z = Q$ gives a probability of success of at least $9/10$ for any Q (in particular, independently of the $\xi_{l-1}, \ldots, \xi_1$). Thus, choosing $p' = 9/10$ and $m_i = m$ for all i, we can then iterate the estimate (44) to arrive at

$$\Pr\left[\sum_{i=1}^{l} \xi_i < r\right] \leq \Pr\left[\xi_l' + \sum_{i=1}^{l-1} \xi_i < r\right] \leq \cdots \leq \Pr\left[\sum_{i=1}^{l} \xi_i' < r\right], \tag{45}$$

where the ξ_i' are independent Bernoulli variables with parameter $9/10$. A standard Chernoff bound (e.g. [72, Section Concentration: Theorem 2.1]) gives

$$\Pr\left[\sum_{i=1}^{l} \xi_i' \leq l(9/10 - t)\right] \leq \mathrm{e}^{-2lt^2}.$$

[7] It was pointed out to us by A. Hansen that in some previous papers [15, 70] which involve a similar construction to the one presented here, it was tacitly assumed that the ξ_i are independent. This will of course not be true in general. Fortunately, a more careful argument shows that all conclusions remain valid [71]. Our treatment here is similar to the one presented in [71].

Choosing $t = 9/10 - r/l$ we obtain

$$\begin{aligned} \Pr\left[\sum_{i=1}^{l} \xi_i' < r\right] &\leq \Pr\left[\sum_{i=1}^{l} \xi_i' \leq r\right] = \Pr\left[\sum_{i=1}^{l} \xi_i' \leq l\,(9/10 - t)\right] \\ &\leq \exp\left(-2l\left(\frac{9}{10} - \frac{r}{l}\right)^2\right). \qquad (46) \end{aligned}$$

Setting the number of iterations generously to

$$l = 10\omega r = 10\omega\left(\lceil \log_2 d \rceil + 2\right)$$

implies

$$2l\left(\frac{9}{10} - \frac{r}{l}\right)^2 \geq 20\omega r\left(\frac{8}{10}\right)^2 \geq 12\omega r \geq \omega + \log 2,$$

where we have used $\omega \geq 1 \geq \log 2$ in the last inequality. Together with (42), (45) and (46) this gives the desired bound

$$\Pr\left[\text{algorithm fails}\right] = \Pr\left[\sum_{i=1}^{l} \xi_i < r\right] \leq \Pr\left[\sum_{i=1}^{l} \xi_i' < r\right] \leq \mathrm{e}^{-\omega - \log(2)} = \frac{1}{2}\mathrm{e}^{-\omega},$$

on our construction of Y failing. The total number of measurement vectors sampled is

$$\sum_{i=1}^{l} m_i = l m_l \leq C\omega t d^{2-\gamma} \log^2 d,$$

for some constant C. □

Finally we are ready to put all pieces together and show or main result – Theorem 1.

Proof of the Main Theorem. In section 5 (Proposition 12) we have shown that the algorithm (24) recovers the sought for signal x, provided that (25) holds and a suitable approximate dual certificate Y exists. Proposition 17 – with a maximal truncation rate of $\gamma = (1 - 2/t)$ – implies that the probability that no such Y can be constructed is smaller than $0.5\mathrm{e}^{-\omega}$, provided that the sampling rate m obeys

$$m \geq C\omega t d^{1+2/t} \log^2 d, \qquad (47)$$

for a sufficiently large absolute constant C. Provided that this constant is large enough, Proposition 9 implies that the probability of (25) failing is also bounded by $0.5\mathrm{e}^{-\omega}$. Theorem 1 now follows from the union bound over these two probabilities of failure. □

7. Converse Bound

In this paper, we require designs of order at least three. Here we prove that this criterion is fundamental in the sense that sampling from 2-designs in general cannot guarantee a subquadratic sampling rate. In order to do so, we will use a particular sort of 2-design, called a *maximal set of mutually unbiased bases* (MUBs) [41, 42, 43, 44]. Two orthonormal bases $\{u_i\}_{i=1}^d$ and $\{v_i\}_{i=1}^d$ are called *mutually unbiased* if their overlap is uniformly minimal. Concretely, this means that

$$|\langle u_i, v_j \rangle|^2 = \frac{1}{d} \quad \forall i, j = 1, \ldots, d$$

must hold for all $i, j = 1, \ldots, d$. Note that this is just a generalization of the incoherence property between standard and Fourier basis. In prime power dimensions, a maximal set of $(d + 1)$ such MUBs is known to exist (and can be constructed) [73]. Such a set is maximal in the sense that it is not possible to find more than $(d+1)$ MUBs in any Hilbert space. Among other interesting properties – cf. [74] for a detailed survey – maximal sets of MUBs are known to form 2-designs [42, 44].

The defining properties of a maximal set of MUBs allow us to derive the converse bound – Theorem 2.

Theorem 18 (Converse bound)**.** *Let $d \geq 2$ be a prime power and let $D_2 \subset \mathbb{C}^d$ be a maximal set of MUBs. Then there exist orthogonal, normalized vectors $x, z \in \mathbb{C}^d$ which have the following property.*

Suppose that m measurement vectors $a_1, \ldots, a_m$ are sampled independently and uniformly at random from D_2. Then, for any $\omega \geq 0$, the number of measurements must obey

$$m \geq \frac{\omega}{4} d(d+1), \tag{48}$$

or the event

$$|\langle a_i, x\rangle|^2 = |\langle a_i, z\rangle|^2 \quad \forall\, i \in \{1, \ldots, m\}$$

will occur with probability at least $\mathrm{e}^{-\omega}$.

Consequently a scaling of $\mathcal{O}(d^2)$ in general cannot be avoided when demanding only the property of being a 2-design and simultaneously requiring a "reasonably small" probability of failure in the recovery process.

Proof of Theorem 18. Suppose that $\{u_i\}_{i=1}^d$ is one orthonormal basis contained in the maximal set of MUBs D_2 and set $x := u_1$ as well as $z := u_2$. Note that by definition these vectors are orthogonal and normalized. Due to the particular structure of MUBs, x and z can only be distinguished if either u_1 or u_2 is contained in $\{a_1, \ldots, a_m\}$. Since each a_i is chosen iid at random from D_2 containing $(d+1)d$ elements, the probability of obtaining either u_1 or u_2 is $p = \frac{2}{(d+1)d}$. As a result, the problem reduces to the following standard stopping time problem (cf. for example Example (2) in Chapter 6.2 in [75]):

Suppose that the probability of success in a Bernoulli experiment is p. How many trials m are required in order for the probability of at least one success to be $1 - \mathrm{e}^{\omega}$ or larger?

To answer this question, we have to find the smallest integer m such that

$$1 - (1-p)^m \geq 1 - \mathrm{e}^{-\omega}, \quad \text{or equivalently} \quad -m\log(1-p) \geq \omega. \tag{49}$$

The standard inequality

$$p \leq -\log(1-p) \leq \frac{p}{1-p} \leq 2p$$

for any $p \in [0, 1/2]$ implies that (48) is a necessary criterion for (49) and we are done. □

8. Conclusion

In this paper we have derived a partly derandomized version of Gaussian PhaseLift [11, 12]. Instead of Gaussian random measurements, our method guarantees recovery for sampling iid from certain finite vector configurations, dubbed t-designs. The required sampling rate depends on the design order t:

$$m = \mathcal{O}\left(t d^{1+2/t} \log^2 d\right). \tag{50}$$

For small t this rate is worse than the Gaussian analogue – but still non-trivial. However, as soon as t exceeds $2\log d$, we obtain linear scaling up to a polylogarithmic overhead.

In any case, we feel that the main purpose of this paper is not to present yet another efficient solution heuristics, but to show that the phase retrieval problem can be derandomized using t-designs. These finite vector sets lie in the vast intermediate region between random Fourier vectors and Gaussian random vectors (the Fourier basis is a 1-design, whereas normalized Gaussian random vectors correspond to an ∞-design). Therefore the design order t allows us to gradually transcend between these two extremal cases.

Acknowledgements DG and RK are grateful to the organizers and participants of the Workshop on Phaseless Reconstruction, held as part of the 2013 February Fourier Talks at the University of Maryland, where they were introduced to the details of the problem. This extends, in particular, to Thomas Strohmer.

The work of DG and RK is supported by the Excellence Initiative of the German Federal and State Governments (Grant ZUK 43), by scholarship funds from the State Graduate Funding Program of Baden-Württemberg, and by the US Army Research Office under contracts W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), and the DFG. FK acknowledges support from the German Federal Ministry of Education and Reseach (BMBF) through the cooperative research project ZeMat.

References

[1] R. Millane, "Phase retrieval in crystallography and optics," *JOSA A*, vol. 7, pp. 394–411, 1990.

[2] Y. M. Bruck and L. Sodin, "On the ambiguity of the image reconstruction problem," *Optics Communications*, vol. 30, pp. 304–308, 1979.

[3] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase." *Appl. Comput. Harmon. Anal.*, vol. 20, pp. 345–356, 2006.

[4] T. Heinosaari, L. Mazzarella, and M. M. Wolf, "Quantum tomography under prior information." *Commun. Math. Phys.*, vol. 318, pp. 355–374, 2013.

[5] B. Sanderson, "Immersions and embeddings of projective spaces." *Proc. Lond. Math. Soc. (3)*, vol. 14, pp. 137–153, 1964.

[6] D. Mixon, "Short, fat matrices," blog, 2013. [Online]. Available: http://dustingmixon.wordpress.com/

[7] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, "Painless reconstruction from magnitudes of frame coefficients." *J. Fourier Anal. Appl.*, vol. 15, pp. 488–501, 2009.

[8] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, "Phase retrieval with polarization," *SIAM J. Imaging Sci.*, vol. 7, pp. 35–66, 2014.

[9] A. S. Bandeira, Y. Chen, and D. G. Mixon, "Phase retrieval from power spectra of masked signals," *Information and Inference*, p. iau002, 2014.

[10] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imaging Sci.*, vol. 6, pp. 199–225, 2013.

[11] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: exact and stable signal recovery from magnitude measurements via convex programming." *Commun. Pure Appl. Math.*, vol. 66, pp. 1241–1274, 2013.

[12] E. Candès and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Found. Comput. Math.*, pp. 1–10, 2013.

[13] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization." *SIAM Rev.*, vol. 52, pp. 471–501, 2010.

[14] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inform. Theory*, vol. 56, pp. 2053–2080, 2010.

[15] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inform. Theory*, vol. 57, pp. 1548–1566, 2011.

[16] Y.-K. Liu, "Universal low-rank matrix recovery from pauli measurements," *Adv. Neural Inf. Process. Syst.*, pp. 1638–1646, 2011.

[17] J. R. Fienup, "Phase retrieval algorithms: A comparison," *Applied Optics*, vol. 21, pp. 2758–2769, 1982.

[18] H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Hybrid projection–reflection method for phase retrieval," *JOSA A*, vol. 20, pp. 1025–1034, 2003.

[19] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[20] Y. C. Eldar and S. Mendelson, "Phase retrieval: Stability and recovery guarantees," *Appl. Comput. Harmon. Anal.*, vol. 36, pp. 473–494, 2014.

[21] X. Li and V. Voroninski, "Sparse signal recovery from quadratic measurements via convex programming," *SIAM J. Math Anal.*, vol. 45, pp. 3019–3033, 2013.

[22] M. Ehler, M. Fornasier, and J. Sigl, "Quasi-linear compressed sensing," *Multiscale Model. Simul.*, vol. 12, pp. 725–754, 2014.

[23] P. Delsarte, J. Goethals, and J. Seidel, "Spherical codes and designs." *Geom. Dedicata*, vol. 6, pp. 363–388, 1977.

[24] V. Sidelnikov, "Spherical 7-designs in 2^n-dimensional Euclidean space." *J. Algebr. Comb.*, vol. 10, pp. 279–288, 1999.

[25] G. Nebe, E. Rains, and N. Sloane, "The invariants of the Clifford groups." *Des. Codes Cryptography*, vol. 24, pp. 99–121, 2001.

[26] A. Scott, "Tight informationally complete quantum measurements." *J. Phys. A-Math. Gen.*, vol. 39, pp. 13 507–13 530, 2006.

[27] A. Ambainis and J. Emerson, "Quantum t-designs: t-wise independence in the quantum world," in *22nd Annual IEEE Conference on Computational Complexity, Proceedings*, 2007, pp. 129–140.

[28] A. Hayashi, T. Hashimoto, and M. Horibe, "Reexamination of optimal quantum state estimation of pure states," *Phys. Rev. A*, vol. 72, SEP 2005.

[29] D. Gross, K. Audenaert, and J. Eisert, "Evenly distributed unitaries: on the structure of unitary designs." *J. Math. Phys.*, vol. 48, pp. 052 104, 22, 2007.

[30] F. G. Brandao, A. W. Harrow, and M. Horodecki, "Local random quantum circuits are approximate polynomial-designs," *preprint arXiv:1208.0692*, 2012.

[31] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, pp. 4203–4215, 2005.

[32] R. G. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin, "A simple proof of the Restricted Isometry Property for random matrices," *Constr. Approx.*, vol. 28, pp. 253–263, 2008.

[33] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, pp. 1207–1223, 2006.

[34] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Comm. Pure Appl. Math.*, vol. 61, pp. 1025–1045, 2008.

[35] F. Krahmer and H. Rauhut, "Structured random measurements in signal processing," *preprint arXiv:1401.1106*, 2014.

[36] R. A. Low, "Large deviation bounds for k-designs." *Proc. R. Soc. Lond., Ser. A, Math. Phys. Eng. Sci.*, vol. 465, pp. 3289–3308, 2009.

[37] M. Luby and A. Wigderson, *Pairwise independence and derandomization. Print version of Foundations and Trends in Theoretical Computer Science Vol. 1, No. 4 (2005).*, print version of foundations and trends in theoretical computer science vol. 1, no. 4 (2005) ed. Boston, MA: Now, 2006.

[38] B. Bajnok, "Construction of spherical t-designs." *Geom. Dedicata*, vol. 43, pp. 167–179, 1992.

[39] J. Korevaar and J. Meyers, "Chebyshev-type quadrature on multidimensional domains." *J. Approx. Theory*, vol. 79, pp. 144–164, 1994.

[40] P. Seymour and T. Zaslavsky, "Averaging sets: A generalization of mean values and spherical designs." *Adv. Math.*, vol. 52, pp. 213–240, 1984.

[41] J. Schwinger, "Unitary operator bases." *Proc. Natl. Acad. Sci. USA*, vol. 46, pp. 570–579, 1960.

[42] G. Zauner, "Quantendesigns: Grundzüge einer nichtkommutativen Designtheorie," Ph.D. dissertation, University of Vienna, 1999.

[43] H. König, "Cubature formulas on spheres." in *Advances in multivariate approximation. Proceedings of the 3rd international conference on multivariate approximation theory*. Berlin: Wiley-VCH, 1999, pp. 201–211.

[44] A. Klappenecker and M. Rotteler, "Mutually unbiased bases are complex projective 2-designs," in *2005 IEEE International Symposium on Information Theory (ISIT), Vols 1 and 2*, 2005, pp. 1740–1744.

[45] R. Kueng and D. Gross, "Stabilizer states are complex projective 3-designs in qubit dimensions," in preparation, 201.

[46] J. M. Renes, R. Blume-Kohout, A. Scott, and C. M. Caves, "Symmetric informationally complete quantum measurements." *J. Math. Phys.*, vol. 45, pp. 2171–2180, 2004.

[47] C. Bachoc and B. Venkov, "Modular forms, lattices and spherical designs." in *Euclidean lattices, spherical designs and modular forms. On the works of Boris Venkov*. Genève: L'Enseignement Mathématique, 2001, pp. 87–111.

[48] S. Hoory, N. Linial, and A. Wigderson, "Expander graphs and their applications," *B. Am. Math. Soc.*, vol. 43, pp. 439–561, 2006.

[49] D. Mondragon and V. Voroninski, "Determination of all pure quantum states from a minimal number of observables," *preprint arXiv:1306.1214*, 2013.

[50] C. Bachoc and M. Ehler, "Tight p-fusion frames," *Appl. Comput. Harmon. Anal.*, vol. 35, pp. 1–15, 2013.

[51] E. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval from masked fourier transforms," *Appl. Comput. Harmon. Anal.*, to appear, preprint arXiv:1310.3240.

[52] D. Gross, F. Krahmer, and R. Kueng, "Improved recovery guarantees for phase retrieval from coded diffraction patterns," *preprint arXiv:1402.6286*, 2014.

[53] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[54] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

[55] D. Gottesman, "Stabilizer Codes and Quantum Error Correction," Ph.D. dissertation, California Institute of Technology, 1997.

[56] ——, "The Heisenberg representation of quantum computers," in *Proceedings of the XXII International Colloquium on Group Theoretical Methods in Physics, pp. 32-43. International Press*, 1999.

[57] A. Barvinok, *A course in convexity.* Providence, RI: American Mathematical Society (AMS), 2002.

[58] J. M. Landsberg, *Tensors: geometry and applications.* Providence, RI: American Mathematical Society (AMS), 2012.

[59] J. Watrous, "Theory of quantum information," lecture notes, 2011. [Online]. Available: https://cs.uwaterloo.ca/~watrous/LectureNotes.html

[60] A. Neumaier, "Combinatorial configurations in terms of distances," *Dept. of Mathematics Memorandum*, pp. 81–09, 1981.

[61] S. Hoggar, "t-designs in projective spaces." *European J. Combinatorics*, vol. 3, pp. 233–254, 1982.

[62] V. I. Levenshtein, "Universal bounds for codes and designs." in *Handbook of coding theory. Vol. 1. Part 1: Algebraic coding*. Amsterdam: Elsevier, 1998, pp. 499–648.

[63] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels," *IEEE Trans. Inform. Theory*, vol. 48, pp. 569–579, 2002.

[64] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "state tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, p. 150401, 2010.

[65] J. A. Tropp, "User-friendly tail bounds for sums of random matrices." *Found. Comput. Math.*, vol. 12, pp. 389–434, 2012.

[66] J. A. Tropp, "User-friendly tools for random matrices: An introduction," Notes, 2012. [Online]. Available: http://users.cms.caltech.edu/~jtropp/notes/Tro12-User-Friendly-Tools-NIPS.pdf

[67] V. Turaev, *Quantum invariants of knots and 3-manifolds.* Berlin: Walter de Gruyter, 1994.

[68] P. Cvitanović, *Group theory. Birdtracks, Lie's, and exceptional groups.* Princeton, NJ: Princeton University Press, 2008.

[69] R. Bhatia, *Matrix analysis.* New York, NY: Springer, 1996.

[70] R. Kueng and D. Gross, "RIPless compressed sensing from anisotropic measurements," *Lin. Alg. Appl.*, vol. 441, pp. 110–123, 2014.

[71] B. Adcock and A. C. Hansen, "Generalized sampling and infinite-dimensional compressed sensing," Technical report NA2011/02, DAMTP, University of Cambridge, Tech. Rep., 2011.

[72] M. Habib, C. McDiarmid, J. Ramírez Alfonsín, and B. Reed, Eds., *Probabilistic methods for algorithmic discrete mathematics.* Berlin: Springer, 1998.

[73] A. Klappenecker and M. Rötteler, "Constructions of mutually unbiased bases." in *Finite fields and applications. 7th international conference,* $\mathbb{F}_{q}7$. Berlin: Springer, 2004, pp. 137–144.

[74] T. Durt, B.-G. Englert, I. Bengtsson, and K. Życzkowski, "On mutually unbiased bases." *Int. J. Quantum Inf.*, vol. 8, pp. 535–640, 2010.

[75] W. Feller, "An introduction to probability theory and its applications. I." New York-London-Sydney: John Wiley and Sons, 1968.

9. Appendix

Here we briefly state an elementary proof of Lemma 6. In the main text we proved this result using wiring diagrams. The purpose of this is to underline the relative simplicity of wiring diagram calculations. Indeed, the elementary proof below is considerably more cumbersome than its pictorial counterpart.

9.1. **Elementary proof of Lemma 6.** Let us choose an arbitrary orthonormal basis $b_1, \ldots, b_d$ of V^d. In the induced basis $\{b_i \otimes b_j\}_{i,j=1}^d$ of $V^d \otimes V^d$ the transpositions then correspond to

$$\underline{1} = \mathbb{1} \otimes \mathbb{1} = \sum_{i=1}^d b_i b_i^* \otimes \sum_{j=1}^d b_j b_j^* \quad \text{and} \quad \sigma_{(1,2)} = \sum_{i,j=1}^d b_i b_j^* \otimes b_j b_i^* .$$

This choice of basis furthermore allows us to write down $\operatorname{tr}_2(A)$ for $A \in M^d \otimes M^d$ explicity:

$$\operatorname{tr}_2(A) = \sum_{i=1}^d \left(\mathbb{1} \otimes b_i^*\right) A \left(\mathbb{1} \otimes b_i\right).$$

Consequently we get for $A, B \in H^d$ arbitrary

$$\operatorname{tr}_2\left(P_{\mathrm{Sym}^2} A \otimes B\right) = \frac{1}{2} \operatorname{tr}_2\left(A \otimes B\right) + \frac{1}{2} \operatorname{tr}_2\left(\sigma_{(1,2)} A \otimes B\right).$$

The latter term can be evaluated explicitly:

$$\begin{aligned} \operatorname{tr}_2\left(\sigma_{(1,2)} A \otimes B\right) &= \sum_{k=1}^d \left(\mathbb{1} \otimes b_k^*\right) \sum_{i,j=1}^d b_i b_j^* \otimes b_j b_i^* A \otimes B \left(\mathbb{1} \otimes b_k\right) \\ &= \sum_{i,j,k=1}^d b_i b_j^* A b_k^* b_j b_i^* B b_k = \sum_{i,j=1}^d \langle b_i, B b_j \rangle b_i b_j^* A \\ &= \left(\sum_{i=1}^d b_i b_i^*\right) B \left(\sum_{j=1}^d b_j b_j^*\right) A = \mathbb{1} B \mathbb{1} A = BA, \end{aligned}$$

and the desired result follows. Here we have used the basis representation of the identity, namely $\mathbb{1} = \sum_{i=1}^d b_i b_i^*$.

Spherical designs as a tool for derandomization: The case of PhaseLift

(Invited Paper)

Richard Kueng
Institute for Physics & FDM
University of Freiburg, and
School of Physics
The University of Sydney
richard.kueng@physik.uni-freiburg.de

David Gross
Institute for Theoretical Physics
University of Cologne, and
Institute for Physics & FDM
University of Freiburg
david.gross@physik.uni-freiburg.de

Felix Krahmer
Research Unit M15
Department of Mathematics
Technische Universität München
felix.krahmer@tum.de

***Abstract*—The problem of retrieving phase information from amplitude measurements alone has appeared in many scientific disciplines over the last century. *PhaseLift* is a recently introduced algorithm for phase recovery that is computationally tractable and numerically stable. However, initial rigorous performance guarantees relied specifically on Gaussian random measurement vectors. To date, it remains unclear which properties of the measurements render the problem well-posed. With this question in mind, we employ the concept of *spherical t-designs* to achieve a partial derandomziation of PhaseLift. Spherical designs are ensembles of vectors which reproduce the first $2t$ moments of the uniform distribution on the complex unit sphere. As such, they provide notions of "evenly distributed" sets of vectors, ranging from tight frames ($t=1$) to the full sphere, as t approaches infinity. Beyond the specific case of PhaseLift, this result highlights the utility of spherical designs for the derandomization of data recovery schemes.**

***Index Terms*—phase retrieval, spherical designs, low rank matrix recovery**

I. Introduction

A. The phase retrieval problem and PhaseLift

The problem of retrieving a complex signal $x \in \mathbb{C}^n$ from measurements of the form

$$y_i = |\langle a_i, x\rangle|^2 \quad i = 1,\ldots,m, \tag{1}$$

where $a_1,\ldots,a_m \in \mathbb{C}^n$ are measurement vectors, has long been abundant in many areas of science. Quite recently, several new recovery algorithms have been proposed and first rigorous performance guarantees have been established. Examples include methods based on polarization identities [1], alternating projections [2], or Wirtinger flow [3]. In addition, there are reconstruction methods that are tailored to specific measurement ensembles, such as the approach in [4], which is based on polynomial representations.

The approach we will focus on has been called *PhaseLift* [5]–[7] and relies on formulating the problem as a low-rank matrix recovery task [8]–[10]. To this end, one notes [11] that the y_i's in (1) can equivalently be expressed as

$$y_i = |\langle a_i, x\rangle|^2 = \operatorname{tr}\left((a_i a_i^*)\,(xx^*)\right) =: \operatorname{tr}\left(A_i\,(xx^*)\right). \tag{2}$$

In other words, the measurement results y_i are linear in the outer product $X := xx^*$ of the signal x with itself. This slight reformulation "lifts" phase retrieval to a linear problem on the (non-linear) set of $n\times n$ hermitian rank-one matrices $\{Z : Z = zz^*,\, z \in \mathbb{C}^n\} \subset H^n$:

$$\begin{aligned} &\text{find} && Z \in H^n \\ &\text{subject to} && \operatorname{tr}(ZA_i) = y_i \quad i=1,\ldots,m, \\ & && \operatorname{rank} Z = 1. \end{aligned} \tag{3}$$

Throughout this article, we shall denote the n^2-dimensional real vector space of hermitian $n \times n$ matrices by H^n.

In general, solving linear equations over the set of rank-1 matrices is computationally intractable. However, there are now many situations for which it has been proved that the *nuclear norm* can be employed as an efficiently computable proxy for rank [8]–[10]. (Recall that the nuclear norm $\|X\|_* = \operatorname{tr}(|X|)$ is the sum of the singular values of X. In a sense, it is the natural "non-commutative", basis-independent matrix analogue of the vector ℓ_1 norm). These results are closely related to the use of the ℓ_1-norm as a convex relaxation of sparsity in compressed sensing [12]. In particular, these findings suggest that (3) can be substituted by the semi-definite program

$$\begin{aligned} &\underset{Z\in H^n}{\text{minimize}} && \|Z\|_* \\ &\text{subject to} && \operatorname{tr}(ZA_i) = y_i \quad i=1,\ldots,m. \end{aligned} \tag{4}$$

This ansatz was dubbed the *PhaseLift* algorithm for phase retrieval by its inventors [5]–[7].

The task is now to establish sufficient conditions under which the above convex problem will indeed have the outer product $X = xx^*$ of the sought-for signal as its unique solution. First results proved that this is the case (with high probability) if the number of measurements roughly scales linearly in the problem dimension – i.e. $m = O(n)$ – and the measurement vectors a_i are complex standard Gaussians [6], [7]. Later works concentrated on the practically more important case of "masked Fourier measurements" [13], [14]. However, the use of Gaussian vectors obscures the specific properties of measurement vectors that enable phase retrieval, while the masked Fourier case is highly application-specified. Thus, the question we are interested in is: Can one identify particular properties of measurement ensembles that allow for phase retrieval via PhaseLift, that are sufficiently general to encompass structured measurements (unlike Gaussians), but at the same time are fairly general (unlike masked Fourier)? We will argue below that the defining properties of spherical designs fall into this category.

978-1-4673-7353-1/15/$31.00 ©2015 IEEE

II. Phase retrieval from spherical design measurements

To motivate the notion of spherical designs, we recall that low-rank recovery results [8]–[10] are usually phrased for measurement ensembles that are *isotropic*, or drawn from a *tight frame* (analogous statements apply to compressed sensing [12], but can be generalized – see e.g. [15], [16]). One definition of such a structural property is as follows:

Definition 1 (*Isotropy*)**.** A weighted set $\{\mu(\alpha), B_\alpha\}_{\alpha\in I} \subseteq H^n$ is *isotropic*, if, for all $Z \in H^n$,

$$\int_I B_\alpha \mathrm{tr}\left(B_\alpha Z\right) \mathrm{d}\mu(\alpha) = Z. \tag{5}$$

In the case of PhaseLift, full isotropy for the measurements A_i is impossible to attain. The reason is that the $A_i = a_i a_i^*$ are all outer products and thus have positive Hilbert-Schmitt inner product with the identity matrix:

$$\mathrm{tr} A_i \mathbb{I} = \mathrm{tr} A_i = \|a_i\|_2^2 > 0. \tag{6}$$

Fortunately, the vectors can be chosen in such a way that A_i's are isotropic on the trace-free subspace of H^n, i.e. the "overweighting of the identity component" just noted is the only way in which the A_i's deviate from being a tight frame. Indeed, for complex standard Gaussian vectors the following identity follows from a simple direct calculation:

Proposition 2 (Near-isotropy: Equation (5.1) in [6])**.** *Let $Z \in H^n$ be arbitrary and assume that b is a complex standard Gaussian vector in $\mathbb{C}^n$. Then*

$$\mathbb{E}\left[bb^* \mathrm{tr}\left(bb^* Z\right)\right] = Z + \mathrm{tr}(Z)\mathbb{I}. \tag{7}$$

Note that the "identity component" $\mathrm{tr}(\mathbb{I}xx^*) = \|x\|_2^2$ of the signal is nothing but its squared length, or intensity. From now on, we assume that the intensity $\|x\|_2^2$ is in fact known. Also, while not essential, we have opted to carry out our analysis for measurement vectors a_i with unit length $\|a_i\|_2 = 1$. With these conventions, the entire problem only ever concerns vectors on the complex unit-sphere. The rotation-invariant measure on the unit sphere $S^{n-1} \subset \mathbb{C}^n$ is called the *Haar measure*. One can sample from it e.g. by drawing complex standard Gaussian vectors and normalizing them. The resulting analogue of (7) reads

$$\int_{w\in S^{n-1}} ww^* \mathrm{tr}\left(ww^* Z\right) \mathrm{d}w = \frac{1}{n(n+1)} \left(Z + \mathrm{tr}(Z)\mathbb{I}\right) \tag{8}$$

for all $Z \in H^n$. Near-isotropy for an ensemble $\{\mu(\alpha), b_\alpha\}_{\alpha\in I}$ can easily be seen [17, Lemma 1] to be equivalent to demanding that the ensemble reproduces the 4th moments of the Haar measure:

Proposition 3 (*Necessary and sufficient criterion for near isotropy*)**.** *Let $\{\mu(\alpha), b_\alpha\}_{\alpha\in I} \subseteq S^{n-1}$ be a weighted set of unit vectors. Then*

$$\int_I b_\alpha b_\alpha^* \otimes b_\alpha b_\alpha^* \mathrm{d}\mu(\alpha) = \int_{w\in S^{n-1}} ww^* \otimes ww^* \mathrm{d}w \tag{9}$$

holds if and only if the re-scaled set $\{\mu(\alpha), \sqrt{n(n+1)} b_\alpha b_\alpha^*\}_{\alpha\in I}$ is near-isotropic in the sense of (7).

If the A_i's range over *all* measurements in a near-isotropic set, they essentially[1] form a tight frame in matrix space and thus X can be recovered by simple linear inversion [11]. However, such a tight frame necessarily contains at least $\dim H^n = O(n^2)$ elements – much more than the $O(n)$ degrees of freedom in x. Still, one could hope that PhaseLift could be proved to succeed for linearly many a_i's sampled from such a set. We will prove below that, unfortunately, this is too optimistic. In general, near-isotropy alone is insufficient for reaching an optimal linear scaling in the number of measurements m. However, (9) suggests a generalization which will turn out to be sufficiently strong for achieving such a goal.

To motivate it, it is wortwhile to point out that vectors drawn uniformly from the sphere are proportional to a tight frame in $\mathbb{C}^n$ (as opposed to H^n):

$$\int_{w\in S^{n-1}} ww^* \mathrm{d}w = \frac{1}{n}\mathbb{I}. \tag{10}$$

Combining this with (9) yields the following two structural criteria for a weighted set $\{\mu(\alpha), b_\alpha\}_{\alpha\in I}$ of unit vectors:

$$\int_I b_\alpha b_\alpha^* \mathrm{d}\mu(\alpha) = \int_{S^{n-1}} ww^* \mathrm{d}w \Rightarrow \text{ tight frame}, \tag{11}$$

$$\int_I \left(b_\alpha b_\alpha^*\right)^{\otimes 2} \mathrm{d}\mu(\alpha) = \int_{S^{n-1}} \left(ww^*\right)^{\otimes 2} \mathrm{d}w \Rightarrow \text{ near-isotropy}. \tag{12}$$

Generalizing these equalities to arbitrary t-th tensor powers yields the following definition which is a the heart of our work:

Definition 4 (Spherical t-design)**.** Let $t \in \mathbb{N}$. We call a weighted set $\{\mu(\alpha), b_\alpha\}_{\alpha\in I} \subseteq S^{n-1}$ of unit vectors a *spherical t-design*, if

$$\int_I \left(b_\alpha b_\alpha^*\right)^{\otimes k} \mathrm{d}\mu(\alpha) = \int_{S^{n-1}} \left(ww^*\right)^{\otimes k} \mathrm{d}w \tag{13}$$

is valid for all $1 \leq k \leq t$. The parameter $t \in \mathbb{N}$ is called the design's order.

While this definition underlines the resemblance of a t-design to vectors drawn uniformly from the complex unit sphere, the expression on the right hand side of (13) is not very practical for actual calculations (in particular, if t is large). Fortunately, a straightforward application of Schur's Lemma [18, Lemma 1] yields

$$\int_{S^{n-1}} \left(ww^*\right)^{\otimes k} \mathrm{d}w = \binom{n+k-1}{k}^{-1} P_{\mathrm{Sym}^k}$$

for $t, n \in \mathbb{N}$ arbitrary. Here, P_{Sym^t} denotes the projector onto the totally symmetric subspace of $(\mathbb{C}^n)^{\otimes k}$. In turn, techniques from multilinear algebra – in particular *wiring calculus* [19], [20] – allow for carrying out calculations involving P_{Sym^t} explicitly.

Analytic expressions for *exact* designs are notoriously difficult to find. Designs of degree 2 are widely known [21]–[24] (see also next section). For degree 3, both real [25] and complex [26] designs are known. For higher t, there are numerical methods based on the notion of the *frame potential* [24], [26], [27], non-constructive existence proofs [28], and constructions in sporadic dimensions (c.f. [29] and references thererin).

In Section II-A below, we will show that drawing the measurements from a spherical 2-design does not allow for non-trivial performance guarantees for PhaseLift. Conversly, in Section II-B, we provide such guarantees for designs of order $t \geq 3$.

[1] As already mentioned, (6) implies that the ray $\{Z \in H^n : Z = c\mathbb{I},\ c \in \mathbb{R}\} \in H^n$ is "overweighted". If the signal's intensity is known, however, this distortion can be readily compensated.

A. Phase retrieval from spherical 2-designs

Proposition 3 together with Definition 4 establishes a one-to-one correspondence between spherical 2-designs and weighted sets of unit vectors which fulfill near-isotropy in the sense of (7). This in turn assures that for any $X \in H^n$, measuring all projectors onto elements of a 2-design as well as measuring $\mathrm{tr}\,(\mathbb{I}X) = \mathrm{tr}(X)$ allows one to linearly invert the measurement process and determine X exactly.

In the context of the "lifted" phase retrieval problem, Balan et. al [11] were the first to be aware of this correspondence. In turn, they used a particular instance of a spherical 2-design to recover $X = xx^*$ via a deterministic choice of order $\mathcal{O}\left(n^2\right)$ projective measurements onto elements of this design.

Concretely, their approach uses a maximal set of *mutually unbiased bases* (MUBs). Two orthonormal bases $\{u_1, \ldots, u_n\}$ and $\{v_1, \ldots, v_n\}$ of $\mathbb{C}^n$ are called *mutually unbiased*, if their overlap is uniformly minimal. Concretely, this means that

$$|\langle u_i, v_j\rangle|^2 = \frac{1}{n} \quad \forall i, j = 1, \ldots, n$$

must hold for all $i, j = 1, \ldots, n$. Note that this is just a generalization of the incoherence property between standard and Fourier basis. In prime power dimensions, a maximal set of $(n+1)$ such MUBs is known to exist (and can be constructed) [30]. Such a set is maximal in the sense that it is not possible to find more than $(n+1)$ MUBs in any Hilbert space. It is well-known that equally weighted, maximal sets of MUBs form spherical 2-designs [22], [24].

Ehler and Kunis [31] also identified near isotropy – equation (7) – and its connection to spherical 2-designs (12) (which they call curbatures of strength 2) as a crucial ingredient for performing phase retrieval. Combining this insight with the PhaseLift approach [5], [6] they conjecture that measuring $\mathcal{O}(n)$ projectors onto randomly chosen elements of a spherical 2-design should be sufficient for establishing a recovery guarantee for PhaseLift. However, in [20] a counter-example to this conjecture is provided: without assuming and exploiting additional properties of the measurement ensemble, random subsampling is not sufficient for avoiding a scaling of $m = \mathcal{O}(n^2)$.

At the heart heart of this counter-example is the following observation regarding the injectivity of a random phaseless measurement chosen from a particular spherical 2-design.

Proposition 5. *Suppose that a is chosen uniformly at random from a maximal set of MUBs (which forms a spherical 2-design). Then there exist orthogonal and normalized vectors $x, y \in \mathbb{C}^n$ such that*

$$\Pr\left[\left||\langle a, x\rangle|^2 - |\langle a, y\rangle|^2\right| > 0\right] \leq \frac{2}{n(n+1)}.$$

Proof: Suppose that $\{u_1, \ldots, u_n\} \subset \mathbb{C}^n$ is one orthonormal basis contained in the maximal set of MUBs and set $x := u_1$, as well as $y := u_2$. Note that by definition these vectors are orthogonal and normalized. Due to the particular structure of MUBs, the expression of interest obeys

$$\left||\langle a, u_1\rangle|^2 - |\langle a, u_2\rangle|^2\right| = \begin{cases} 1 & \text{if } a = u_1, \text{ or } a = u_2, \\ 0 & \text{otherwise.} \end{cases}$$

The claim now follows from noticing that a is chosen uniformly at random from the $n(n+1)$ vectors contained in a maximal set of MUBs. ■

Note that this statement implies that the probability of distinguishing the orthogonal vectors x and y by means of a random phaseless measurement (chosen uniformly from the spherical 2-design formed by a maximal set of MUBs) is proportional to $1/n^2$. In [20] the authors use a slightly refined version of this insight together with a stopping time argument to establish the following rigorous counter-example to subsampling from a particular 2-design.

Theorem 6 (Theorem 2 in [20])**.** *Let $n \geq 2$ be a prime power and let $D_2 \subset \mathbb{C}^n$ be a maximal set of MUBs. Then there exist orthogonal, normalized vectors $x, y \in \mathbb{C}^d$ which have the following property:*

Suppose that m measurement vectors $a_1, \ldots, a_m$ are sampled independently and uniformly at random from D_2. Then, for any $\omega \geq 0$, the number of measurements must obey

$$m \geq \frac{\omega}{4} n(n+1), \tag{14}$$

or the event

$$|\langle a_i, x\rangle|^2 = |\langle a_i, y\rangle|^2 \quad \forall\, i \in \{1, \ldots, m\}$$

will occur with probability at least $\mathrm{e}^{-\omega}$.

It is worthwhile to emphasize that this no-go result only applies to specific 2-designs. Particular instances of a 2-design that exhibit additional structural properties may well allow for subsampling. In fact, such a measurement ensemble – "coded diffraction patterns", or "masked Fourier measurements" – was introduced by Candès et al. in [13] and it was proven in the same paper that a total number of $m = \mathcal{O}\left(n \log^4 n\right)$ such measurements actually allows for establishing a Phaselift recovery guarantee. Since these coded-diffraction patterns fulfill near-isotropy in the sense of (7), Proposition 3 assures that they also form a spherical 2-design. This equivalence was pointed out in [14], where the required sampling rate for such a recovery was furthermore reduced to a total of $\mathcal{O}\left(n \log^2 n\right)$ measurements.

B. Phase retrieval from higher-order designs

Since a subsampled collection of random projectors onto a spherical 2-design can in general not be sufficient for recovering a sought for signal $X = xx^*$ via PhaseLift, it is natural to ask if designs of higher order allow for establishing such a recovery guarantee.

Recall that the main problem with subsampling from a spherical 2-design was the injectivity-issue pointed out in Proposition 5. However, this situation changes dramatically for designs of higher order.

Proposition 7. *Suppose that a is chosen uniformly at random from a spherical 4-design. Then for any two distinct vectors $x, y \in \mathbb{C}^n$*

$$\Pr\left[\left||\langle a, x\rangle|^2 - |\langle a, y\rangle|^2\right| > 0\right] \geq \frac{1}{96}$$

is true.

Proof: The claim is an immediate corollary from the auxiliary statement [32, equation (24)] used to establish Proposition 12 there. For a chosen uniformly at random from a spherical 4-design, this statement assures for any $Z \in H^n$

$$\Pr\left[|\mathrm{tr}\,(aa^* Z)| \geq \frac{\xi \|Z\|_2}{\sqrt{n(n+1)}}\right] \geq \frac{\left(1-\xi^2\right)^2}{24} \quad \forall \xi \in [0,1] \tag{15}$$

by means of the Payley-Zygmund inequality. Here, $\|Z\|_2$ denotes the Frobenius norm of Z. Setting $Z := xx^* - yy^*$ (note that since x and

y are distinct, the matrix Z cannot vanish and therefore $\|Z\|_2 > 0$ must hold) and setting $\xi = 2^{-1/2}$ allows us to conclude

$$\Pr\left[\left||\langle a,x\rangle|^2 - |\langle a,y\rangle|^2\right| > 0\right] = \Pr\left[|\mathrm{tr}\,(aa^*Z)| > 0\right]$$
$$\geq \Pr\left[|\mathrm{tr}\,(aa^*Z)| \geq \frac{2^{-1/2}\|Z\|_2}{\sqrt{n(n+1)}}\right] \geq \frac{(1-1/2)^2}{24} = \frac{1}{96}$$

by means of (15). ■

Proposition 7 assures that choosing measurement vectors independently from any spherical 4-design behaves strikingly different from the 2-design case. In particular, this statement guarantees that injectivity issues in the sense of Proposition 5 are much less severe for designs of higher order. In accordance with such a disintegration of the injecivity problem, non-trivial recovery guarantees for PhaseLift can be established for designs of higher order, as the main result in [20] shows.

Theorem 8 (Theorem 1 in [20])**.** *Let $x \in \mathbb{C}^n$ be the unknown signal of interest. Suppose that $\|x\|_{\ell_2}^2$ is known and that m measurement vectors $a_1, \ldots, a_m$ have been sampled independently and uniformly at random from an equally weighted t-design obeying $t \geq 3$. Then, with probability at least $1 - \mathrm{e}^{-\omega}$, PhaseLift (the convex optimization problem (4) above) recovers x up to a global phase, provided that the sampling rate exceeds*

$$m \geq \omega\, C t\, n^{1+2/t} \log^2 n. \tag{16}$$

Here $\omega \geq 1$ is an arbitrary parameter and C is a universal constant.

Already for 3-designs, this result establishes a recovery guarantee from subquadratically many – namely $\mathcal{O}(n^{5/3}\log^2(n))$ – projectors onto randomly selected 3-design elements. The statement furthermore becomes tight – i.e. the required sampling rate scales linearly in the dimension n of the signal x – up to polylog-factors, if the design order t is allowed to grow logarithmically with the dimension ($t = 2\log n$). Note that a similar recovery guarantee for sampling from particular t-designs can be established, even if n sampling vectors are correlated at a time [33]. Recently, the above Theorem was substantially strengthened and generalized in [32].

Theorem 9 (Theorem 3 in [32])**.** *Consider the measurement process described in (2) where the measurement vectors $a_1, \ldots, a_m$ have been sampled independently from a spherical 4-design (according to the design's weights). Furthermore assume that the number of measurements m obeys*

$$m \geq C_1 n r \log n,$$

for $1 \leq r \leq n$ arbitrary. Then with probability at least $1 - \mathrm{e}^{-C_2 m}$ it holds that for any $X \in \mathcal{H}_n$ with rank at most r, any solution $X^\#$ of the convex optimization problem (4) with noisy measurements $y_i = \mathrm{tr}\,(A_i X) + \epsilon_i$, where $\sum_{i=1}^n \epsilon_i^2 \leq \eta^2$, obeys

$$\|X - X^\#\|_2 \leq \frac{C_3 \eta}{\sqrt{m}}. \tag{17}$$

Here, $C_1, C_2, C_3 > 0$ again denote universal positive constants.

This is a uniform recovery guarantee for recovering arbitrary rank-r matrices that is furthermore robust towards noise. Clearly it covers phase retrieval via PhaseLift as a special case – namely the one, where all matrices X of interest are guaranteed to be rank one. Consequently, $\mathcal{O}(n \log n)$ measurements randomly chosen from a 4-design are sufficient to guarantee phaseless recovery of arbitrary signals $x \in \mathbb{C}^n$ via the convex optimization (4). Moreover, such a sampling rate is close to optimal. As shown in [34], it follows from the results derived in [35] that a sample size of $m \geq (4 + o(1))\, n$ is in fact necessary (cf. [36]).

Finally, we want to point out that Theorem 9 is also close to optimal in terms of the design order t required. Indeed, Theorem 6 establishes that a design order of at least $t = 3$ is required without making additional assumptions on the measurement ensemble. Theorem 9 gets by with a design order of $t = 4$ and no further assumptions. Fully closing the gap by establishing an analogue of Theorem 9 which is valid already for 3-designs, or tightening the required sampling rate in Theorem 8 does constitute an intriguing open problem. Numerical studies presented in [20] suggest that this might indeed be feasible. For the sake of completeness we have included the results of this study in Figure 1.

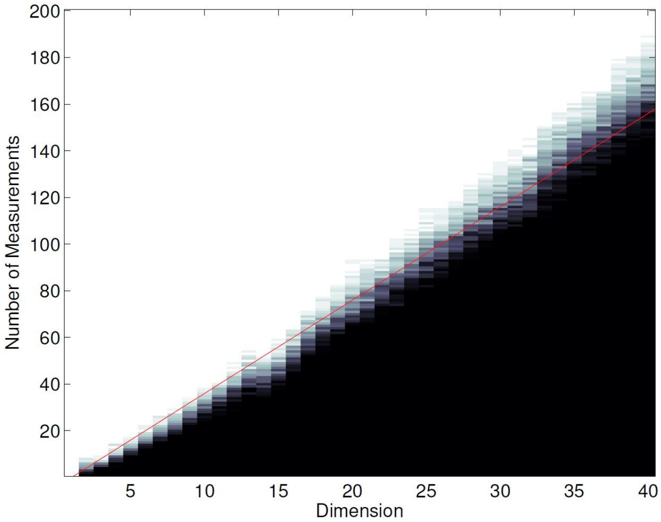


Fig. 1. Phase Diagram for PhaseLift from (projected) stabilizer states, which form an equally weighted 3-design in power-of-two dimensions [26]. The x-axis indicates the problem's dimension, while the y-axis denotes the number of independent design measurements performed. The frequency of a successful recovery over 30 independent runs of the experiment appears color-coded from black (zero) to white (one). To guide the eye, we have furthermore included a red line indicating $m = 4n - 4$.

III. Spherical designs as a general-purpose tool for partial derandomization

In section II we have introduced spherical t-designs as a generalization of natural structural properties (11), (12) which assure that the weighted vector set forms a tight frame on $\mathbb{C}^n$ and the corresponding rank-one projectors obey near-isotropy (essentially meaning that they form a slightly distorted tight-frame on H^n).

Equivalently, one can define spherical t-designs as (usually finite) weighted distributions of vectors that approximate Haar-random vectors (or equivalently: the distribution of complex standard Gaussian vectors renormalized to unit length) up to t-th moments.

Viewing spherical t-designs from this angle reveals that they do constitute a general purpose tool for derandomizing results that initially required generic – i.e. Haar-random or standard Gaussian –

vectors. This utility of the design concept has long been appreciated for example in quantum information theory [37], [38]. It has been compared [37] to the notion of *t-wise independence*, which plays a role for example in the analysis of discrete randomized algorithms.

Acknowledgements

The work of DG and RK is supported by the Excellence Initiative of the German Federal and State Governments (Grants ZUK 43 & 81), the ARO under contracts W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), the Freiburg Research Innovation Fund, the DFG (GRO 4334 & SPP1798), and the State Graduate Funding Program of Baden-Württemberg. FK acknowledges support from the German Federal Ministry of Education and Reseach (BMBF) through the cooperative research project ZeMat.

References

[1] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, “Phase retrieval with polarization,” *SIAM J. Imaging Sci.*, vol. 7, pp. 35–66, 2014.

[2] P. Netrapalli, P. Jain, and S. Sanghavi, “Phase retrieval using alternating minimization,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[3] E. Candes, X. Li, and M. Soltanolkotabi, “Phase retrieval via Wirtinger flow: Theory and algorithms,” *preprint arXiv:1407.1065*, 2014.

[4] B. Bodmann and N. Hammen, “Stable phase retrieval with low-redundancy frames,” *Advances in Computational Mathematics*, to appear.

[5] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, “Phase retrieval via matrix completion,” *SIAM J. Imaging Sci.*, vol. 6, pp. 199–225, 2013.

[6] E. J. Candès, T. Strohmer, and V. Voroninski, “Phaselift: exact and stable signal recovery from magnitude measurements via convex programming.” *Commun. Pure Appl. Math.*, vol. 66, pp. 1241–1274, 2013.

[7] E. Candès and X. Li, “Solving quadratic equations via PhaseLift when there are about as many equations as unknowns,” *Found. Comput. Math.*, pp. 1–10, 2013.

[8] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization.” *SIAM Rev.*, vol. 52, pp. 471–501, 2010.

[9] E. J. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” *IEEE Trans. Inform. Theory*, vol. 56, pp. 2053–2080, 2010.

[10] D. Gross, “Recovering low-rank matrices from few coefficients in any basis,” *IEEE Trans. Inform. Theory*, vol. 57, pp. 1548–1566, 2011.

[11] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, “Painless reconstruction from magnitudes of frame coefficients.” *J. Fourier Anal. Appl.*, vol. 15, pp. 488–501, 2009.

[12] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, ser. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[13] E. Candes, X. Li, and M. Soltanolkotabi, “Phase retrieval from masked fourier transforms,” *Appl. Comput. Harmon. Anal.*, to appear, preprint arXiv:1310.3240.

[14] D. Gross, F. Krahmer, and R. Kueng, “Improved recovery guarantees for phase retrieval from coded diffraction patterns,” *Appl. Comput. Harmon. Anal.*, to appear, preprint arXiv:1402.6286.

[15] M. Rudelson and S. Zhou, “Reconstruction from anisotropic random measurements,” *IEEE Trans. Inform. Theory*, vol. 59, pp. 3434–3447, 2013.

[16] R. Kueng and D. Gross, “RIPless compressed sensing from anisotropic measurements,” *Linear Algebra Appl.*, vol. 441, pp. 110–123, 2014.

[17] D. Appleby, C. A. Fuchs, and H. Zhu, “Group theoretic, Lie algebraic and Jordan algebraic formulations of the SIC existence problem,” *Quantum Inf. Comput.*, vol. 15, no. 1-2, pp. 61–94, 2015.

[18] A. Scott, “Tight informationally complete quantum measurements.” *J. Phys. A-Math. Gen.*, vol. 39, pp. 13 507–13 530, 2006.

[19] J. M. Landsberg, *Tensors: geometry and applications.* Providence, RI: American Mathematical Society (AMS), 2012.

[20] D. Gross, F. Krahmer, and R. Kueng, “A partial derandomization of PhaseLift using spherical designs,” *J. Fourier Anal. Appl.*, vol. 21, pp. 229–266, 2015.

[21] J. Schwinger, “Unitary operator bases.” *Proc. Natl. Acad. Sci. USA*, vol. 46, pp. 570–579, 1960.

[22] G. Zauner, “Quantendesigns: Grundzüge einer nichtkommutativen Designtheorie,” Ph.D. dissertation, University of Vienna, 1999.

[23] H. König, “Cubature formulas on spheres.” in *Advances in multivariate approximation. Proceedings of the 3rd international conference on multivariate approximation theory*. Berlin: Wiley-VCH, 1999, pp. 201–211.

[24] A. Klappenecker and M. Rotteler, “Mutually unbiased bases are complex projective 2-designs,” in *2005 IEEE International Symposium on Information Theory (ISIT), Vols 1 and 2*, 2005, pp. 1740–1744.

[25] V. Sidelnikov, “Spherical 7-designs in 2^n-dimensional Euclidean space.” *J. Algebr. Comb.*, vol. 10, pp. 279–288, 1999.

[26] R. Kueng and D. Gross, “Stabilizer states are complex projective 3-designs in qubit dimensions,” in preparation, 2015.

[27] J. M. Renes, R. Blume-Kohout, A. Scott, and C. M. Caves, “Symmetric informationally complete quantum measurements.” *J. Math. Phys.*, vol. 45, pp. 2171–2180, 2004.

[28] P. Seymour and T. Zaslavsky, “Averaging sets: A generalization of mean values and spherical designs.” *Adv. Math.*, vol. 52, pp. 213–240, 1984.

[29] C. Bachoc and B. Venkov, “Modular forms, lattices and spherical designs.” in *Euclidean lattices, spherical designs and modular forms. On the works of Boris Venkov*. Genève: L’Enseignement Mathématique, 2001, pp. 87–111.

[30] A. Klappenecker and M. Rötteler, “Constructions of mutually unbiased bases.” in *Finite fields and applications. 7th international conference, $\mathbb{F}_{q^7}$*. Berlin: Springer, 2004, pp. 137–144.

[31] “Phase retrieval using time and fourier magnitude measurements,” in *SampTA*.

[32] R. Kueng, H. Rauhut, and U. Terstiege, “Low rank matrix recovery from rank one measurements,” *preprint arXiv:1410.6913*, 2014.

[33] R. Kueng, “Low rank matrix recovery from few orthonormal basis measurements,” in *11th international conference on Sampling Theory and Applications (SampTA 2015)*, Washington, USA, May 2015.

[34] T. Heinosaari, L. Mazzarella, and M. M. Wolf, “Quantum tomography under prior information.” *Commun. Math. Phys.*, vol. 318, pp. 355–374, 2013.

[35] B. Sanderson, “Immersions and embeddings of projective spaces.” *Proc. Lond. Math. Soc. (3)*, vol. 14, pp. 137–153, 1964.

[36] D. Mixon, “Short, fat matrices,” blog, 2013. [Online]. Available: http://dustingmixon.wordpress.com/

[37] A. Ambainis and J. Emerson, “Quantum t-designs: t-wise independence in the quantum world,” in *22nd Annual IEEE Conference on Computational Complexity, Proceedings*, 2007, pp. 129–140.

[38] R. A. Low, “Large deviation bounds for k-designs.” *Proc. R. Soc. Lond., Ser. A, Math. Phys. Eng. Sci.*, vol. 465, pp. 3289–3308, 2009.

LOW RANK MATRIX RECOVERY FROM RANK ONE MEASUREMENTS

RICHARD KUENG, HOLGER RAUHUT, AND ULRICH TERSTIEGE

Abstract. We study the recovery of Hermitian low rank matrices $X \in \mathbb{C}^{n\times n}$ from undersampled measurements via nuclear norm minimization. We consider the particular scenario where the measurements are Frobenius inner products with random rank-one matrices of the form $a_j a_j^*$ for some measurement vectors $a_1, \ldots, a_m$, i.e., the measurements are given by $y_j = \operatorname{tr}(X a_j a_j^*)$. The case where the matrix $X = xx^*$ to be recovered is of rank one reduces to the problem of phaseless estimation (from measurements, $y_j = |\langle x, a_j\rangle|^2$ via the PhaseLift approach, which has been introduced recently. We derive bounds for the number m of measurements that guarantee successful uniform recovery of Hermitian rank r matrices, either for the vectors a_j, $j = 1, \ldots, m$, being chosen independently at random according to a standard Gaussian distribution, or a_j being sampled independently from an (approximate) complex projective t-design with $t = 4$. In the Gaussian case, we require $m \geq Crn$ measurements, while in the case of 4-designs we need $m \geq Crn\log(n)$. Our results are uniform in the sense that one random choice of the measurement vectors a_j guarantees recovery of all rank r-matrices simultaneously with high probability. Moreover, we prove robustness of recovery under perturbation of the measurements by noise. The result for approximate 4-designs generalizes and improves a recent bound on phase retrieval due to Gross, Kueng and Krahmer. In addition, it has applications in quantum state tomography. Our proofs employ the so-called bowling scheme which is based on recent ideas by Mendelson and Koltchinskii.

1. Introduction

1.1. **The phase retrieval problem.** The problem of retrieving a complex signal from measurements that are ignorant towards phases is abundant in many different areas of science, such as X-ray cristallography [40, 57], astronomy [29] diffraction imaging [67, 57] and more [8, 12, 76]. Mathematically formulated, the problem consists of recovering a complex signal (vector) $x \in \mathbb{C}^n$ from measurements of the form

$$|\langle a_j, x\rangle|^2 = b_j \quad \text{for} \quad j = 1, \ldots, m, \tag{1}$$

where $a_1, \ldots, a_m \in \mathbb{C}^n$ are sampling vectors. This ill-posed inverse problem is called *phase retrieval* and has attracted considerable interest over the last few decades. An important feature of this problem is that the signal x enters the measurement process (1) quadratically. This leads to a non-linear inverse problem. Classical approaches to numerically solving it include alternating projection methods [30, 34]. However, these methods usually require extra constraints and careful selection of parameters, and in particular, no rigorous convergence or recovery guarantees seem to be available.

As Balan et al. pointed out in [7], this apparent obstacle of having nonlinear measurements can be overcome by noting that the measurement process – while quadratic in x – is linear in the outer product xx^*:

$$|\langle a_j, x\rangle|^2 = \operatorname{tr}\left(a_j a_j^* x x^*\right).$$

This "lifts" the problem to a matrix space of dimension n^2, where it becomes linear and can be solved explicitly, provided that the number of measurements m is at least n^2 [7]. However, there is additional structure present, namely the matrix $X = xx^*$ is guaranteed to have rank one. This connects the phase retrieval problem to the young but already extensive field of *low-rank matrix recovery*. Indeed, it is just a special case of low-rank matrix recovery, where both the signal $X = xx^*$ and the measurement matrices $A_j = a_j a_j^*$ are constrained to be proportional to rank-one projectors.

Date: October 25, 2014.

It should be noted, however, that such a reduction to a low rank matrix recovery problem is just one possibility to retrieve phases. Other approaches use polarization identities [2] or alternate projections [60]. Yet another recent method is phase retrieval via Wirtinger flow [14].

1.2. **Low rank matrix recovery.** Building on ideas of compressive sensing [18, 27, 33], low rank matrix recovery aims to reconstruct a matrix of low rank from incomplete linear measurements via efficient algorithms [63]. For our purposes we concentrate on Hermitian matrices $X \in \mathbb{C}^{n\times n}$ and consider measurements of the form

$$\operatorname{tr}(XA_j) = b_j \quad j = 1,\ldots,m \tag{2}$$

where the $A_j \in \mathbb{C}^{n\times n}$ are some Hermitian matrices. For notational simplicity, we define the measurement operator

$$\mathcal{A} : \mathcal{H}_n \to \mathbb{R}^m \quad Z \mapsto \sum_{j=1}^{m} \operatorname{tr}(ZA_j)\, e_j,$$

where $e_1,\ldots,e_m$ denotes the standard basis in $\mathbb{R}^m$. This summarizes an entire (possibly noisy) measurement process via

$$b = \mathcal{A}(X) + \epsilon. \tag{3}$$

Here $b = (b_1,\ldots,b_m)^T$ contains all measurement outcomes and $\epsilon \in \mathbb{R}^m$ denotes additive noise. Low rank matrix recovery can be regarded as a non-commutative version of compressive sensing. Indeed, the structural assumption of low rank assures that the matrix is sparse in its eigenbasis. In parallel to the prominent role of ℓ_1-norm minimization in compressive sensing [33], it is by now well-appreciated [1, 17, 16, 63, 35] that in many relevant measurement scenarios, the sought for matrix X can be efficiently recovered via convex programming, although the corresponding rank minimization problem is NP hard in general [28].

In order to formulate this convex program, we introduce the standard ℓ_p-norm on $\mathbb{R}^n$ or $\mathbb{C}^n$ by $\|x\|_{\ell_p} = (\sum_{\ell=1}^n |x_\ell|^p)^{1/p}$ for $1 \le p < \infty$ and the Schatten-p-norm on the space $\mathcal{H}_n$ of Hermitian $n \times n$ matrices as

$$\|Z\|_p = \left(\sum_{\ell=1}^{n} \sigma_\ell(Z)^p\right)^{1/p} = \operatorname{tr}\left(|Z|^p\right)^{1/p}, \quad p \ge 1,$$

where $\sigma_\ell(Z)$, $\ell = 1,\ldots,n$, denote the singular values of Z, tr is the trace and $|Z| = (Z^*Z)^{1/2}$. Important special cases are the nuclear norm $\|Z\|_* = \|Z\|_1$, the Frobenius norm $\|Z\|_F = \|Z\|_2$ and the spectral norm $\|Z\|_\infty = \|Z\|_{2\to 2} = \sigma_{\max}(Z)$ being the largest singular value. More information, concerning Schatten-p norms can be found in Appendix 5.1.

Assuming the upper bound $\|\epsilon\|_{\ell_2} \le \eta$ on the noise for some $\eta \ge 0$, recovery via nuclear norm minimization corresponds to

$$\underset{Z\in\mathcal{H}_n}{\text{minimize}}\ \|Z\|_1 \text{ subject to } \|\mathcal{A}(Z) - b\|_{\ell_2} \le \eta. \tag{4}$$

This is a convex optimization problem which can be solved computationally efficiently with various strategies [33, Chapter 15], [10, 23, 62, 71]. We note that several alternatives to nuclear norm minimization may also be applied including iteratively reweighted least squares [32], iterative hard thresholding [47, 70], greedy approaches [51] and algorithms specialized to certain measurement maps $\mathcal{A}$ [43], but our analysis is geared towards nuclear norm minimization and does not provide guarantees for these other algorithms.

Up to date, a number of measurement instances have been identified for which nuclear norm minimization (4) – and potentially other algorithms – provably recovers the sought for low-rank matrix from considerably fewer than n^2 measurements [17, 16, 20, 35, 32, 52, 63, 74]. All these constructions are based on randomness, the simplest being a random Gaussian measurement map where all entries $\mathcal{A}_{j,k,\ell}$ in the representation $\mathcal{A}(X)_j = \sum_{k,\ell=1}^n \mathcal{A}_{j,k,\ell} X_{k,\ell}$ are independent mean zero variance one Gaussian random variables. It is shown in [16, 63] that

$$m \ge Crn$$

measurements suffice in order to (stably) reconstruct a matrix $X \in \mathbb{C}^{n\times n}$ of rank at most r with probability at least $1 - \exp(-cm)$, where the constants $C, c > 0$ are universal. This result is based on a version of the by-now classical restricted isometry property so that this result is uniform in the sense that a random draw of $\mathcal{A}$ enables reconstruction of *all* rank r matrices simultaneously with high probability. A corresponding nonuniform result, holding only for a fixed rank r matrix X is stated in [20], see also [4, 74], which shows that essentially $m > 6rn$ measurements are sufficient, thus providing also good constants.

While unstructured Gaussian measurements provide optimal guarantees, which are comparably easy to derive, many applications demand for more structure in the measurement process. A particular instance is the matrix completion problem [22, 17, 19, 35, 21], which aims at recovering missing entries of a matrix which is known to be of low rank. Here, the source of randomness is in the selection of the known entries. In contrast to the unstructured measurements, additional incoherence properties of the matrix to be recovered are required and the bounds on the number of measurements are slightly worse [22, 35], namely $m \geq Crn\log^2(n)$. The matrix completion setup generalizes to measurements with respect to an arbitrary operator basis. The incoherence assumption on the matrix to be recovered can be dropped if in turn the operator basis is incoherent, which is the case for the particular example of Pauli measurements arising in quantum tomography [35, 52]. Here, a sufficient and necessary number of measurements scales like $m \geq Crn\log(n)$.

Rank-one measurements, however, in general fail to be sufficiently incoherent for directly applying proof techniques of the same type. For the particular case of phase retrieval (where the matrix of interest is by construction a rank-one projector) this obstacle could be overcome by providing problem specific recovery guarantees that either manifestly rely on (rank one) Gaussian measurements [13, 74] or result in a non-optimal sampling rate [38, 15, 37].

1.3. **Weighted complex projective designs.** The concept of real spherical designs was introduced by Delsarte Goethals and Seidel in a seminal paper [26] and has been studied in algebraic combinatorics [68] and coding theory [26, 59]. Recently, complex projective designs – the natural extension of real spherical designs to the complex unit sphere – have been of considerable interest in quantum information theory [79, 65, 41, 36, 53, 11, 48].

Roughly speaking, a complex projective t-design is a finite subset of the complex unit sphere in $\mathbb{C}^n$ with the particular property that the discrete average of any polynomial of degree (t,t) (i.e., a polynomial $p(z,\bar{z})$ of total degree t both in $z = (z_1,\ldots,z_n)$ and in $\bar{z} = (\bar{z}_1,\ldots,\bar{z}_n)$) or less equals its uniform average. Many equivalent definitions capture this essence, but the following one best serves our purpose.

Definition 1 (*exact, weighted t-design*, Definition 3 in [65])**.** For $t \in \mathbb{N}$, a finite set $\{w_1,\ldots,w_N\} \subset \mathbb{C}^n$ of normalized vectors with corresponding weights $\{p_1,\ldots,p_N\}$ such that $p_i \geq 0$ and $\sum_{i=1}^N p_i = 1$ is called a *weighted complex projective t-design* of dimension n and cardinality N if

$$\sum_{i=1}^N p_i \left(w_i w_i^*\right)^{\otimes t} = \int_{\mathbb{C}P^{n-1}} \left(ww^*\right)^{\otimes t} \mathrm{d}w, \tag{5}$$

where the integral on the right hand side is taken with respect to the unique unitarily-invariant probability measure on the complex projective space $\mathbb{C}P^{n-1}$ and the integrand is computed using arbitrary preimages of the $w \in \mathbb{C}P^{n-1}$ in the unit sphere in $\mathbb{C}^n$. (Note that if w_1 and w_2 are elements of the unit sphere that have the same image w in $\mathbb{C}P^{n-1}$ then $w_1w_1^* = w_2w_2^*$.) This definition in particular shows that uniform sampling from a t-design mimics the first $2t$ moments of sampling uniformly according to the Haar measure, which is equivalent to sampling standard Gaussian vectors followed by renormalization.

A simple application of Schur's Lemma – see e.g. [65, Lemma 1] – reveals that the integral on the right hand side of (5) amounts to

$$\int_{\mathbb{C}P^{n-1}} \left(ww^*\right)^{\otimes t} \mathrm{d}w = \binom{n+t-1}{t}^{-1} P_{\mathrm{Sym}^t}, \tag{6}$$

where P_{Sym^t} denotes the projector onto the totally symmetric subspace Sym^t of $(\mathbb{C}^n)^{\otimes t}$ defined in the appendix – see equation (40).

In accordance with [55], we call a t-design *proper*, if all the weights are equal, i.e., $p_i = 1/N$ for all $i = 1, \ldots, N$.

Although exact, proper t-designs exist and can be constructed in any dimension n for any $t \in \mathbb{N}$ [66, 6, 45, 41], these constructions are typically inefficient in the sense that they require vector sets of exponential size. For example, the construction in [41] requires on the order of $\mathcal{O}(t)^n$ vectors which scales exponentially in the dimension n. Constructions of *exact, proper* designs with significantly smaller number of vectors (scaling only polynomially in n) are notoriously difficult to find.

By introducing weights, it becomes simpler to obtain designs with a number of elements that scales polynomially in the dimension n. Some existence results can be found in [25], where weighted t-designs appear under the notion of cubatures of strength t. It seems that one can construct weighted t-designs by drawing sufficiently many vectors at random and afterwards solving a linear system for the weights. Further note, that generalizations of cubatures to higher dimensional projections were used in [5] in the context of a generalized phase retrieval problem, where the measurements are given as norms of projections onto higher dimensional subspaces.

2. Main results

2.1. Low rank matrix recovery from rank one Gaussian projections.

Our first main result gives a uniform and stable guarantee for recovering rank-r matrices with $\mathcal{O}(rn)$ rank one measurements that are proportional to projectors onto standard Gaussian random vectors.

Theorem 2. *Consider the measurement process described in (3) with measurement matrices $A_j = a_j a_j^*$, where $a_1, \ldots, a_m \in \mathbb{C}^n$ are independent standard Gaussian distributed random vectors. Furthermore assume that the number of measurements m obeys*

$$m \geq C_1 nr,$$

for $1 \leq r \leq n$ arbitrary. Then with probability at least $1 - \mathrm{e}^{-C_2 m}$ it holds that for any positive semidefinite matrix $X \in \mathcal{H}_n$ with rank at most r, any solution $X^\#$ to the convex optimization problem (4) with noisy measurements $b = \mathcal{A}(X) + \epsilon$, where $\|\epsilon\|_{\ell_2} \leq \eta$, obeys

$$\|X - X^\#\|_2 \leq \frac{C_3 \eta}{\sqrt{m}}. \tag{7}$$

Here, C_1, C_2 and C_3 denote universal positive constants. (In particular, for $\eta = 0$ one has exact reconstruction.)

For the rank one case $r = 1$, Theorem 2 essentialy reproduces the main result in [13] which uses completely different proof techniques. (More precisely, for X of rank 1 the estimate in loc. cit. is $\|X - X^\#\|_2 \leq \frac{C\|\epsilon\|_1}{m}$ with high probability.) A variant of the above statement was shown in [74] to hold (in the real case) for a fixed matrix X of rank one. (More precisely, in loc. cit. it is assumed that X is positive semidefinite and the optimization is performed wrt. the function f given by (9) below.) In fact, our proof reorganizes and extends the arguments of [74, Section 8] in such a way, that Theorem 8.1 of loc. cit. is shown to hold even uniformly (that is simultaneously for all X) and for arbitrary rank. On the contrary to [13], we will not need ε-nets to show uniformity.

2.2. Recovery with 4-designs.

As we will see, the proof method for Theorem 2 can also be applied to measurements drawn independently from a weighted complex projective 4-design in the sense of Definition 1. In [38] exact complex projective t-designs have been applied to the problem of phase retrieval. The main result (Theorem 1) in [38] is a non-uniform exact recovery guarantee for phase retrieval via the convex optimization problem (4) that requires $m = \mathcal{O}\left(t n^{1+2/t} \log^2 n\right)$ measurement vectors that are drawn uniformly from a proper t-design ($t \geq 3$). The proof technique which we are going to employ here, allows for considerably generalizing and improving this statement. We will draw the measurement vectors $a_1, \ldots, a_m \in \mathbb{C}^n$ independently at random from a weighted 4-design $\{p_i, w_i\}_{i=1}^N$, which means that for each draw of a_j, the design element w_i is selected with probability p_i. In the sequel we assume that $n \geq 2$.

Theorem 3. *Let $\{p_i, w_i\}_{i=1}^N$ be a weighted 4-design and consider the measurement process described in (3) with measurement matrices $A_j = \sqrt{n(n+1)} a_j a_j^*$, where $a_1, \ldots, a_m \in \mathbb{C}^n$ are drawn independently from $\{p_i, w_i\}_{i=1}^N$. Furthermore assume that the number of measurements m obeys*

$$m \geq C_4 nr \log n,$$

for $1 \leq r \leq n$ arbitrary. Then with probability at least $1 - \mathrm{e}^{-C_5 m}$ it holds that for any $X \in \mathcal{H}_n$ with rank at most r, any solution $X^\#$ to the convex optimization problem (4) with noisy measurements $b = \mathcal{A}(X) + \epsilon$, where $\|\epsilon\|_{\ell_2} \leq \eta$, obeys

$$\|X - X^\#\|_2 \leq \frac{C_6 \eta}{\sqrt{m}}. \tag{8}$$

Here, $C_4, C_5, C_6 > 0$ again denote universal positive constants.

The normalization factor $\sqrt{n(n+1)}$ leads to approximately the same normalization of the A_j (wrt. the Frobenius norm) as in expectation in the Gauss case. The theorem is a stable, uniform guarantee for recovering arbitrary Hermitian matrices of rank at most r with high probability using the convex optimization problem (4) and $m = \mathcal{O}(nr\log(n))$ measurements drawn independently (according to the design's weights) from a weighted 4-design. It obviously covers sampling from *proper* 4-designs as a special case.

Also, Theorem 3 is close to optimal in terms of the design order t required. In the context of the phase retrieval problem[1] it was shown in [38, Theorem 2], that choosing measurements uniformly from a proper 2-design does not allow for a sub-quadratic sampling rate m without additional structural assumptions on the measurement ensemble. It is presently open whether Theorem 3 also holds for 3-designs.

Finally, note that the results for Gaussian measurement vectors and 4-designs are remarkably similar. They only differ by a logarithmic factor. This underlines the usefulness of complex projective designs as a general-purpose tool for de-randomization – see e.g. [38, Section 1.1.] for further reading on this topic. Also, Theorem 3 resembles insights in the context of distinguishing quantum states [55, 3], where it was pointed out that (approximate) 4-designs "perform almost as good" as uniform measurements (projectors onto random Gaussian vectors). Note that we will generalize Theorem 3 to approximate 4-designs in Theorem 5 below.

2.3. **Extensions.**

In this section we state variants of the main theorems which can be proved in a similar way.

2.3.1. *Real-valued case.* Theorem 2 is also valid in the real case, i.e., assuming that the a_j are real standard Gaussian distributed and $\mathcal{H}_n$ is replaced by the space $\mathcal{S}_n$ of real symmetric $n \times n$-matrices. The proof of the corresponding statement is very similar to the one of Theorem 2 and we sketch the necessary adaptations in Subsection 4.3.

2.3.2. *Recovery of positive semidefinite matrices.* The matrix X to be recovered may be known to be positive semidefinite ($X \succcurlyeq 0$) in advance. In this case, one can enforce the reconstructed matrix to be positive semidefinite by considering the optimization program

$$\underset{Z \succcurlyeq 0}{\text{minimize}} \ \operatorname{tr}(Z) \quad \text{subject to} \quad \|\mathcal{A}(Z) - b\|_{\ell_2} \leq \eta$$

instead of the nuclear norm minimization program (4). Then analog versions of Theorems 2, 3 and 5 hold. In particular, the error bounds (7), (8) remain valid. In the noisy case $\eta > 0$, this does not follow directly from these theorems, since the minimizer of the nuclear norm minimization (4) is not guaranteed to be positive semidefinite in the noisy case. The proof proceeds similarly as the ones for the case $X \in \mathcal{H}_n$. Instead of the nuclear norm one has to consider (as in [74]) the function

$$f : \mathcal{H}_n \to \mathbb{R} \cup \{\infty\}, \quad f(X) = \begin{cases} \operatorname{tr}(X), & \text{if } X \succcurlyeq 0 \\ \infty, & \text{otherwise.} \end{cases} \tag{9}$$

[1] i.e., recovering unknown Hermitian matrices of rank one

3. Applications to quantum state tomography

A particular instance of matrix recovery is the task of reconstructing a finite n-dimensional quantum mechanical system which is fully characterized by its *density operator* ρ – an $n \times n$-dimensional positive semidefinite matrix with trace one. Estimating the density operator of an actual (finite dimensional) quantum system is an important task in quantum physics known as *quantum state tomography*.

One is often interested in performing tomography for quantum systems that have certain structural properties. An important structural property – on which we shall focus here – is *purity*. A quantum system is called *pure*, if its density operator has rank one and *almost pure* if it is well approximated by a matrix of low rank $\operatorname{rank}(\rho) = r \ll n$. Assuming this structural property, quantum state tomography is a low-rank matrix recovery problem [39, 35, 31, 52]. An additional requirement for tomography is the fact that the measurement process has to be "experimentally realizable" and – preferably – "efficiently" so.

Any "experimentally realizable" quantum mechanical measurement corresponds to a *positive operator-valued measure* (POVM). In the special case of (finite) n-dimensional quantum systems, a POVM is a set of positive semidefinite matrices $\{M_j\}_{j=1}^N \subset \mathcal{H}_n$ that sum up to the identity, i.e., $\sum_{j=1}^N M_j = \mathrm{id}$ – see e.g. [61, Chapter 2.2.6] for further information.

For practical reasons, it is highly desirable that a quantum measurement (represented by a POVM) can be implemented with reasonable effort. In accordance with [61], we call a POVM-measurement *efficient* (or *practical*), if it can be carried out by performing a number of $\mathcal{O}(\operatorname{polylog}(n))$ elementary steps[2]. Making this notion precise would go beyond the scope of this work and we refer to [3, 61] for further reading.

Below we will concentrate on random constructions of the vectors a_j. We note, however, that implementing the POVM element $a_j a_j^*$ corresponding to the projection onto a Gaussian random vector is *not* efficient as it requires $\mathcal{O}(\operatorname{poly}(n))$ steps. This renders all low rank matrix recovery guarantees which rely on Gaussian measurements – like in Theorem 2 above – inefficient (and therefore impractical) for low rank quantum state tomography. Utilizing a weakened concept of t-designs discussed next, we partly overcome this obstackle with Theorem 5 below and its possible implementations outlined in Sections 3.2.1, 3.2.2.

3.1. An analogue of Theorem 3 for approximate designs.

While Theorem 3 is a substantial derandomization of Theorem 2 and therefore interesting from a theoretical point of view, its usefulness hinges on the availability of constructions of exact weighted 4-designs. Unfortunately, such constructions are notoriously difficult to find unless one relies on randomness, for which, however, the resulting designs are *not* efficient in the sense described in the previous section. One way to circumvent these difficulties is to relax the defining property (5) of a t-design. This approach was – up to our knowledge – introduced by A. Ambainis and J. Emerson [3] and resulted in the notion of approximate designs which is by now well established in quantum information science.

Definition 4 (*Approximate t-design*)**.** We call a weighted set $\{p_i, w_i\}_{i=1}^N$ of normalized vectors an approximate t-design of p-norm accuracy θ_p, if

$$\left\| \sum_{i=1}^N p_i \left(w_i w_i^*\right)^{\otimes t} - \int_{\mathbb{C}P^{n-1}} \left(w w^*\right)^{\otimes t} \mathrm{d}w \right\|_p \leq \binom{n+t-1}{t}^{-1} \theta_p. \tag{10}$$

While accuracy measured in arbitrary Schatten-p-norms is conceivable, the ones measured in operator norm ($p = \infty$) [42, 3, 54, 11] and nuclear norm ($p = 1$) [58] are the ones most commonly used – at least in quantum information theory. For these two accuracies, the definition in particular assures that every approximate t-design is in particular also a k-design for any $1 \leq k \leq t$ with the same p-norm accuracy θ_p [3, 54]. For the sake of being self-contained we provide a proof of this statement in the appendix – see Lemma 16.

[2] This notion is comparable to the *circuit depth* in classical computer science.

A slightly refined analysis reveals that Theorem 3 also holds for sufficiently accurate approximate 4-designs.

Theorem 5. *Fix $1 \leq r \leq n$ arbitrary and let $\{p_i, w_i\}_{i=1}^N$ be an approximate 4-design satisfying*

$$\left\| \sum_{i=1}^N p_i w_i w_i^* - \frac{1}{n}\mathrm{id} \right\|_\infty \leq \frac{1}{n}, \tag{11}$$

that admits either operator norm accuracy $\theta_\infty \leq 1/(16r^2)$, or trace-norm accuracy $\theta_1 \leq 1/4$, respectively. Then, the recovery guarantee from Theorem 3 is still valid (possibly with slightly worse absolute constants $\tilde{C}_4, \tilde{C}_5$ and $\tilde{C}_6$).

3.2. **Protocols for efficient low rank matrix recovery.** Up to now, efficient recovery of low rank density operators by means of the convex optimization problem (4) has been established for random measurements of (generalized) Pauli observables [39, 35]. For this type of measurements, the statistical issues are well understood [31] and Y.K. Liu managed to prove a uniform recovery guarantee [52] which is comparable to the results presented here. Also, this procedure has been tested in experiments [64].

Theorem 5 is similar in spirit and we show here that it permits efficient low rank quantum state tomography for different types of measurements. Indeed, in the field of quantum information theory, various ways of constructing approximate t-designs are known. Most of these methods are inspired by “realistic” quantum mechanical setups (e.g. the circuit model [61, Chapter 4]) and can therefore be – in principle – implemented efficiently in an actual experiment.

Introducing these constructions in full detail would go beyond the scope of this work and we content ourselves with sketching two possible ways of generating approximate 4-design measurements which meet the requirements of Theorem 5. For further clarification on the concepts used here, we refer directly to the stated references.

From now on we shall assume that the dimension $n = 2^d$ is a power of two (d-qubit density operators).

3.2.1. *The Ambainis-Emerson POVM.* In [3], the authors provide a way of constructing a normalized approximate 4-design of operator-norm accuracy $\theta_\infty = \mathcal{O}\left(1/n^{1/3}\right)$, which in addition is a tight frame. They furthermore present a way to generate the corresponding POVM-measurements efficiently – i.e., involving only $\mathcal{O}\left(\mathrm{polylog}(n)\right)$ elementary steps. It therefore meets the requirements of Theorem 5, provided that the maximal rank r of the unknown density operator obeys

$$r \leq C_7 n^{1/6}, \tag{12}$$

where C_7 is a sufficiently small absolute constant. The additional rank requirement stems from the fact that the resulting design only has limited accuracy.

This accuracy can be improved if we construct an approximate design in a much larger space – say $\mathbb{C}^{n^6}$ – and project it down onto an arbitrary n-dimensional subspace. The reason for such an approach is that the projected design's accuracy corresponds to $\theta_\infty = \mathcal{O}\left(\left(n^6\right)^{-1/3}\right) = \mathcal{O}(1/n^2)$. This allows for replacing (12) by the much weaker rank constraint

$$r \leq C_8 n, \tag{13}$$

(where C_8 is again a sufficiently small absolute constant) in order to assure that the design's operator-norm accuracy obeys $\theta_\infty \leq 1/(16r^2)$.

Also, the projected design vectors still form a tight frame, but are sub-normalized, i.e. $\|\tilde{w}_i\|_{\ell_2}^2 = \|P w_i\|_{\ell_2}^2 \leq \|w_i\|_2^2 = 1$. Here, $P : \mathbb{C}^{n^6} \to \mathbb{C}^n$ denotes the projection. However, since they are an approximate design's projection onto a smaller space, they maintain all properties of an approximate 4-design – most notably Lemma 16 – except normalization. In the proof of Theorem 5, normalization is only used once, namely in (28) and sub-normalization is sufficient to guarantee this estimate. Consequently, Theorem 5 is applicable and guarantees universal quantum state tomography via the convex optimization problem (4), provided that (13) holds and $m = C_4 r n \log n$ randomly chosen measurements $\mathrm{tr}\left(\tilde{w}_i \tilde{w}_i^* \rho\right)$ are known.

3.2.2. *Approximate unitary designs.* Another way to generate approximate t-designs is to consider arbitrary orbits of unitary t-designs. Unitary t-designs $\{p_i, U_i\}_{i=1}^N$ are a natural generalization of the spherical design concept to unitary matrices [24, 36]. They have the particular property that every weighted orbit $\{p_i, U_i x\}$ with $\|x\|_{\ell_2} = 1$ of an approximate unitary design forms an approximate complex projective t-design of the same accuracy.

It was shown in [11] that unitary t-designs of arbitrary operator-norm accuracy θ_∞ can be constructed efficiently by using local random circuits. This approach allows for generating an approximate unitary 4-design of operator-norm accuracy $\theta_\infty \leq 1/(16n^2)$ by means of local random circuits of length $C_9 \log(n)^2$, where C_9 is a sufficiently large absolute constant. Consequently, every orbit of the union of all such local random circuits of length $C_9 \log(n)^2$ forms a normalized approximate 4-design which meets the requirements of Theorem 5. One way of implementing such a measurement consists in choosing a local quantum circuit U_i at random, applying its adjoint circuit U_i^* to the density operator ρ and then measuring the two-outcome POVM $\{xx^*, \mathrm{id} - xx^*\}$, where $x \in \mathbb{C}^n$ is arbitrary (but fixed and normalized) to obtain

$$y_i = \mathrm{tr}\left(xx^* U_i^* \rho U_i\right) = \mathrm{tr}\left(U_i xx^* U_i^* \rho\right) = \mathrm{tr}\left(w_i w_i^* \rho\right).$$

According to Theorem 5, $m = \tilde{C}_4 nr \log n$ random measurements of this kind are sufficient to reconstruct any density operator ρ of rank at most r with very high probability via the convex optimization problem (4).

Remark 6. One should note that the approximate unitary designs of [11] are not of a finite nature, because the set of all local random unitaries is continuous. Nevertheless, assuming that such local random unitaries are available as "basic building blocks", local random circuits are efficiently implementable in terms of circuit length. Replacing the atomic expectation values $\sum_{i=1}^N p_i \left(w_i w_i\right)^{\otimes t}$ by their continuous counterparts does not change the argument and Theorem 5 remains valid.

It is worthwhile to point out that the two possible applications of Theorem 5 to the problem of low rank quantum state tomography, as presented here, are not yet optimal. The implementation using the Ambainis-Emerson POVM – presented in 3.2.1 – suffers from the drawback that it demands either a very strong criterion on the density operator's rank – condition (12) – or generating the design in a much larger space and projecting it down. The latter construction is highly unlikely to be optimal and it is furthermore a priori not clear where the corresponding POVM-measurements can be implemented efficiently.

The second approach, on the other hand, suffers from the drawback that carrying out each of the $Crn \log n$ random measurements requires terminating with a very coarse two-outcome POVM measurement. It is very likely that a more fine grained-output statistics could be obtained with comparable effort. The recovery protocol stated here, however, does not allow for advantageously taking into account such refined information about the unknown state.

However, we still feel that mentioning these protocols is worthwhile, as they substantially narrow down the gap between what can be proved (Theorem 5 and the protocols presented in subsection 3.2) and what can be implemented efficiently in an actual quantum state tomography experiment. Next, we provide ideas for further narrowing this gap and finding more protocols that allow for efficient low rank quantum state tomography.

3.3. **Outlook.** The construction of approximate t-designs in Section 3.2.1 via projections from higher-dimensional designs would be much stronger if an efficient protocol for the corresponding POVM measurements could be provided. We leave this for future work. Alternatively, the authors of [3] mention results by Kuperberg [46] who managed to construct exact t-designs containing only $\mathcal{O}\left(n^{2t}\right)$ vectors. They furthermore conjecture that their method of efficiently implementing the corresponding POVM measurement also works for Kuperberg's exact construction. Trying to find such an implementation and combining it with Theorem 3 also does constitute an intriguing follow up-project.

Diagonal-unitary designs are yet another generalization of the spherical design concept to a more restrictive family of unitaries [58]. The notion of a diagonal-unitary design depends on choosing a reference basis and is therefore weaker than the unitary design notation from above. Nevertheless, in [58, Proposition 1] it was shown that the orbit[3] of a particular vector $f_1 \in \mathbb{C}^n$ under a diagonal-unitary t-designs still forms approximate complex projective t-designs with trace-norm accuracy

$$\theta_1 = \binom{n+t-1}{t} \left(\frac{t(t-1)}{n} + \mathcal{O}\left(\frac{1}{n^2} \right) \right). \tag{14}$$

A quick calculation reveals that this orbit forms a normalized tight frame. Unfortunately, the trace-norm accuracy (14) is too weak for a direct application of Theorem 5. However, in [58, Theorem 1] it is shown that the union of all 3-qubit phase-random circuits forms an exact diagonal-unitary 4-design. Similar to local random circuits, such 3-qubit phase-random circuits can in principle be implemented efficiently [58, Proposition 3] in an actual quantum mechanical setup. Furthermore, comparing (14) with the accuracy relation $\theta_\infty \leq \theta_1 \leq n^t \theta_\infty$ – see Lemma 16 in the appendix – suggests that particular orbits of diagonal-unitary designs might possess a much tighter operator-norm accuracy, if the spectrum of their (t-fold tensored) average were sufficiently flat. Such a result, combined with Theorem 5, would lead to a tomography procedure that is similar to the one of Section 3.2.2, but uses random 3-qubit phase gates instead of local random circuits.

4. Proofs

Our proof technique consists in the application of a uniform version of Tropp's bowling scheme, see [74]. The crucial ingredient is a new method due to Mendelson [56] and Koltchiskii, Mendelson [44] (see also [50]) to obtain lower bounds for quantities of the form $\inf_{u \in E} \sum_{j=1}^m |\langle \phi_j, x \rangle|^2$ where the ϕ_j are independent random vectors in $\mathbb{R}^d$ and E is a subset of $\mathbb{R}^d$. We start by recalling from [74] the notions and results underlying this technique.

Suppose we measure $x_0 \in \mathbb{R}^d$ via measurements $y = \Phi x_0 + \epsilon \in \mathbb{R}^m$, where Φ is an $m \times d$ measurement matrix and $\epsilon \in \mathbb{R}^m$ vector of unknown errors. Let $\eta \geq 0$ and assume $\|\epsilon\|_{\ell_2} \leq \eta$. For $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ proper convex we aim at recovering x_0 by solving the convex program

$$\text{minimize } f(x) \quad \text{subject to} \quad \|\Phi x - y\|_{\ell_2} \leq \eta. \tag{15}$$

Here, *proper convex* means that f is convex and attains at least one finite value.
Let $K \subseteq \mathbb{R}^d$ be a cone. Then we define the minimum singular value of Φ with respect to K as

$$\lambda_{\min}(\Phi; K) = \inf\{\|\Phi u\|_{\ell_2} : u \in K \cap \mathbb{S}^{d-1}\},$$

where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. For $x \in \mathbb{R}^d$, we consider the (convex) *descent cone*

$$\mathcal{D}(f, x) = \bigcup_{\tau > 0} \{y \in \mathbb{R}^d : \; f(x + \tau y) \leq f(x)\}.$$

With these notions, the success of the convex program (15) can be estimated as follows.

Proposition 7. *([74], see also [20]) Let $x_0 \in \mathbb{R}^d$, $\Phi \in \mathbb{R}^{m \times d}$ and $y = \Phi x_0 + \epsilon$ with $\|\epsilon\|_{\ell_2} \leq \eta$. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be proper convex and let $x^\sharp$ be a solution of the corresponding convex program (15). Then*

$$\|x^\sharp - x_0\|_{\ell_2} \leq \frac{2\eta}{\lambda_{\min}(\Phi; \mathcal{D}(f, x_0))}.$$

The crucial point for us is that in the situation that Φ is a random matrix with i.i.d. rows, the following theorem can be applied to estimate $\lambda_{\min}(\Phi; \mathcal{D}(f, x_0))$ (see also [44, 74, 56]).

[3] For a diagonal-unitary design with respect to the standard basis $e_1, \ldots, e_n$, their result requires the first Fourier vector $f_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n e_i$ as a fiducial. This vector is isomorphic to the $|+\rangle^{\otimes d} = \left(\frac{1}{\sqrt{2}} (e_1 + e_2) \right)^{\otimes d}$ state which is well-known in quantum information theory.

Theorem 8. *(Koltchinskii, Mendelson; Tropp's version [74]) Fix $E \subset \mathbb{R}^d$ and let $\phi_1, \ldots, \phi_m$ be independent copies of a random vector ϕ in $\mathbb{R}^d$. For $\xi > 0$ let*

$$Q_\xi(E;\phi) = \inf_{u \in E} \mathbb{P}\{|\langle \phi, u\rangle| \geq \xi\}$$

$$\text{and} \quad W_m(E,\phi) = \mathbb{E}\sup_{u\in E}\langle h, u\rangle, \quad \text{where } h = \frac{1}{\sqrt{m}}\sum_{j=1}^m \varepsilon_j \phi_j$$

with (ε_j) being a Rademacher sequence[4]*. Then for any $\xi > 0$ and any $t \geq 0$ with probability at least $1 - e^{-2t^2}$*

$$\inf_{u\in E}\left(\sum_{i=1}^m |\langle \phi_i, u\rangle|^2\right)^{1/2} \geq \xi\sqrt{m}Q_{2\xi}(E;\phi) - 2W_m(E,\phi) - \xi t.$$

Remark 9. We note that the above theorem is stated in [74] to hold with probability $1-e^{-t^2/2}$. Inspecting the proof, however, reveals that the probability estimate can actually be improved to $1 - e^{-2t^2}$.

We will apply the notions in these results in the context of Theorems 2 and 3 as follows:

- identify $\mathcal{H}_n$ with $\mathbb{R}^d = \mathbb{R}^{n^2}$
- Φ is the matrix of $\mathcal{A}$ in the standard basis, i.e., $\Phi(X)_i = \operatorname{tr}(a_i a_i^* X)$
- $f : \mathcal{H}_n \to \mathbb{R} \cup \{\infty\}$ is the nuclear norm, i.e., $f(X) = \|X\|_1$.

In particular,

$$\mathcal{D}(f,X) = \bigcup_{\tau>0}\{Y \in \mathcal{H}_n : \ f(X+\tau Y) \leq f(X)\}.$$

In Topp's original *bowling scheme*, [74, Sections 7 and 8], a positive semidefinite matrix X of rank 1 is fixed and Theorem 8 is then applied to $E_X = \mathcal{D}(f,X) \cap \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} = \{Z \in \mathcal{H}_n : \ \|Z\|_2 = 1\}$. He then uses the Payley-Zygmund inequality to obtain a lower bound for $Q_{2\xi}$ (after choosing some appropriate ξ) and finally applies arguments like conic duality to bound W_m from above.

Our approach differs from the original bowling scheme in one aspect: instead of fixing one rank r-matrix and focusing on E_X, we are going to consider the union $E_r = \{X \in \mathcal{H}_n : \ \operatorname{rank}(X) \leq r, \ X \neq 0\}$ of all low rank matrices. The rest of the proof essentially parallels the bowling scheme from [74]. However, we are going to require an auxiliary statement – Lemma 10 below – in order to obtain a comparable upper bound on W_m. This slightly refined analysis is going to result in a uniform recovery result whose probability of success equals the one for non-uniform recovery of a single fixed X. Note that with such an approach, we do not need to use ε-nets in order to establish uniformity.

For $r \leq n$ let

$$K_r = \bigcup_X \mathcal{D}(f,X),$$

where the union runs over all $X \in \mathcal{H}_n \setminus \{0\}$ of rank at most r. We further define

$$E_r = K_r \cap \mathbb{S}^{d-1} = \bigcup_X E_X,$$

where $E_X = \mathcal{D}(f,X) \cap \mathbb{S}^{d-1}$. We recall that for a convex cone $K \subseteq \mathbb{R}^d$, its *polar cone* is defined to be the closed convex cone

$$K^\circ = \{v \in \mathbb{R}^d : \langle v, x\rangle \leq 0 \text{ for all } x \in K\}.$$

A crucial ingredient for Theorems 2 and 3 is the following lemma.

[4] A Rademacher vector $\epsilon = (\epsilon_j)_{j=1}^m$ is a vector of independent Rademacher random variables, taking the values ±1 with equal probability.

Lemma 10. *Let $A \in \mathcal{H}_n$ be a Hermitian $n \times n$-matrix. Then*

$$\sup_{Y \in E_r} \operatorname{tr}(A \cdot Y) \leq 2\sqrt{r}\|A\|_\infty.$$

By duality and the matrix Hölder inequality this statement is equivalent to

$$\|Y\|_1 \leq 2\sqrt{r} \quad \text{for all } Y \in E_r. \tag{16}$$

The following proof is inspired by [74, Section 8], where similar arguments are used.

Proof. It is enough to show that, for any $X \in \mathcal{H}_n \setminus \{0\}$ of rank at most r, we have

$$\sup_{Y \in E_X} \operatorname{tr}(A \cdot Y) \leq 2\sqrt{r}\|A\|_\infty.$$

We may assume that X has precisely rank $r \geq 1$. By weak duality for cones, see [74, Proposition 4.2] or [33, eq. (B.40)], we have $\sup_{Y \in E_X} \operatorname{tr}(A \cdot Y) \leq \operatorname{dist}_F(A, \mathcal{D}(f, X)^\circ)$, where as usual $\operatorname{dist}_F(A, \mathcal{D}(f, X)^\circ) = \inf_{B \in \mathcal{D}(f,X)^\circ} \|A - B\|_2$. By [74, Fact 4.3], we know that the polar cone $\mathcal{D}(f, X)^\circ$ is the closure of $\bigcup_{\tau \geq 0} \tau \cdot \partial f(X)$. For $S \in \partial f(X)$ and $\tau \geq 0$, it follows that

$$\sup_{Y \in E_X} \operatorname{tr}(A \cdot Y) \leq \|A - \tau \cdot S\|_2.$$

Write $X = \sum_{i=1}^r \lambda_i x_i x_i^*$, where the x_i are orthonormal and the λ_i are non-zero. Extend $x_1, \ldots, x_r$ to an orthonormal basis $x_1, \ldots, x_n$ of $\mathbb{C}^n$ and write A in the form

$$A = \sum \tilde{a}_{i,j} x_i x_j^*.$$

(Hence the $\tilde{a}_{i,j}$ form the matrix obtained from A by a basis change to $x_1, \ldots, x_n$.) Define the four blocks $A_1 = \sum_{i,j \leq r} \tilde{a}_{i,j} x_i x_j^*$, $A_2 = \sum_{i \leq r, j > r} \tilde{a}_{i,j} x_i x_j^*$, $A_3 = \sum_{i > r, j \leq r} \tilde{a}_{i,j} x_i x_j^* = A_2^*$ and $A_4 = \sum_{i,j > r} \tilde{a}_{i,j} x_i x_j^*$. It is well known that $\partial\|X\|_1$ consists of all matrices of the form

$$S = \sum_{i=1}^r \operatorname{sgn}(\lambda_i) x_i x_i^* + S_2,$$

where $S_2 \in \mathcal{H}_n$ has the property that $S_2 x_i = 0$ for all $i \in \{1, \ldots, r\}$ and $\|S_2\|_\infty \leq 1$. (See for example [78], where the real analogue is shown.) Consider now

$$S = \sum_{i=1}^r \operatorname{sgn}(\lambda_i) x_i x_i^* + \tau^{-1} A_4 \in \partial\|X\|_1, \quad \text{where } \tau = \|A_4\|_\infty.$$

(If $\tau = 0$, let $S = \sum_{i=1}^r \operatorname{sgn}(\lambda_i) x_i x_i^*$.) To simplify the notation, write $S_1 = \sum_{i=1}^r \operatorname{sgn}(\lambda_i) x_i x_i^*$. Then

$$\begin{aligned}
\|A - \tau S\|_2 &= \|A - A_4 - \tau S_1\|_2 = \left(\operatorname{tr}(A_1 - \tau S_1)^2 + 2\operatorname{tr}(A_2^* A_2)\right)^{1/2} \\
&= \left(\|A_1 - \tau S_1)\|_2^2 + 2\|A_2^*\|_2^2\right)^{1/2} \leq \left(2\|A_1\|_2^2 + 2\|\tau \cdot S_1\|_2^2 + 2\|A_2^*\|_2^2\right)^{1/2} \\
&= \left(2\|A \cdot x_1\|_2^2 + \ldots + 2\|A \cdot x_r\|_2^2 + 2\|\tau \cdot S_1\|_2^2\right)^{1/2} \\
&\leq \left(2r\|A\|_\infty^2 + 2r\tau^2\right)^{1/2} \leq 2\sqrt{r}\|A\|_\infty,
\end{aligned}$$

since $\tau = \|A_4\|_\infty \leq \|A\|_\infty = \lambda$. □

4.1. **Proof of Theorem 2.** In order to prove both statements of Theorem 2, it is enough by Proposition 7 to show that for $m \geq cnr$ with probability at least $1 - e^{-\gamma m}$

$$\inf_{Y \in E_r} \left(\sum_{j=1}^m \operatorname{tr}(a_j a_j^* Y)^2\right)^{1/2} \geq c_1\sqrt{m}$$

for suitable positive constants c, c_1, γ. For $\xi > 0$ let

$$Q_\xi = \inf_{Z \in E_r} \mathbb{P}(|\operatorname{tr}(a_j a_j^* Z)| \geq \xi). \tag{17}$$

Further let

$$H = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \varepsilon_j a_j a_j^*, \tag{18}$$

where the ε_j form a Rademacher sequence independent of everything else, and introduce

$$W_m = \mathbb{E} \sup_{Y \in E_r} \operatorname{tr}(H \cdot Y).$$

By Theorem 8, for any $\xi > 0$ and any $t \geq 0$ with probability at least $1 - e^{-2t^2}$,

$$\inf_{Y \in E_r} \left(\sum_{j=1}^{m} (\operatorname{tr}(a_j a_j^* Y))^2 \right)^{1/2} \geq \xi \sqrt{m} Q_{2\xi} - 2W_m - \xi t.$$

Following Tropp's bowling scheme, we first estimate $Q_{2\xi}$ for a suitable ξ. As in [74], we conclude from the Payley-Zygmund inequality (see e.g. [33, Lemma 7.16]) that

$$\mathbb{P}\{|\langle aa^*, U\rangle|^2 \geq \frac{1}{2}(\mathbb{E}|\langle aa^*, U\rangle|^2)\} \geq \frac{1}{4} \cdot \frac{(\mathbb{E}|\langle aa^*, U\rangle|^2)^2}{\mathbb{E}|\langle aa^*, U\rangle|^4}. \tag{19}$$

(Here a follows the standard Gaussian distribution on $\mathbb{C}^n$.) Assume now $\|U\|_2 = 1$ and write $U = \sum_i \lambda_i u_i u_i^*$, where $\sum_i \lambda_i^2 = 1$ and the u_i are orthonormal. Then $\langle aa^*, U\rangle = \operatorname{tr}(aa^* U) = \sum_j \lambda_j \operatorname{tr}(aa^* u_j u_j^*) = \sum_j \lambda_j |u_j^* a|^2$ and hence,

$$|\langle aa^*, U\rangle|^2 = \sum_{i,j} \lambda_i \lambda_j |u_i^* a|^2 |u_j^* a|^2.$$

The $u_j^* a$ form independent standard (complex) Gaussian random variables. To compute the moments of a standard complex Gaussian random variable Z, write $Z = X + iY$ where X, Y are independent and $\mathcal{N}(0, \frac{1}{2})$ distributed. The $2k$-th moment of X resp. Y is $\frac{(2k)!}{2^{2k} k!}$, which allows us to compute higher moment of Z, for example, $\mathbb{E}|Z|^2 = \mathbb{E}X^2 + \mathbb{E}Y^2 = 1$ and $\mathbb{E}|Z|^4 = \mathbb{E}X^4 + 2\mathbb{E}X^2\mathbb{E}Y^2 + \mathbb{E}Y^4 = 2$. Similarly, we obtain $\mathbb{E}|Z|^6 = 6$ and $\mathbb{E}|Z|^8 = 24$ (and more generally $\mathbb{E}|Z|^{2k} = k!$). Thus, we conclude that

$$\mathbb{E}|\langle aa^*, U\rangle|^2 = \sum_{i \neq j} \lambda_i \lambda_j + 2\sum_i \lambda_i^2 = \sum_{i,j} \lambda_i \lambda_j + \sum_i \lambda_i^2 = (\sum_i \lambda_i)^2 + 1 \geq 1 \tag{20}$$

and

$$(\mathbb{E}|\langle aa^*, U\rangle|^2)^2 = (\sum_i \lambda_i)^4 + 2(\sum_i \lambda_i)^2 + 1.$$

Expanding $\mathbb{E}|\langle aa^*, U\rangle|^4$ in a similar way, we obtain

$$\begin{aligned}\mathbb{E}|\langle aa^*, U\rangle|^4 &= \sum_{i,j,k,\ell} \lambda_i \lambda_j \lambda_k \lambda_\ell + \sum_{i,k,\ell} \lambda_i^2 \lambda_k \lambda_\ell + 2\sum_{i,k} \lambda_i^2 \lambda_k^2 + 4\sum_{i,k} \lambda_i^3 \lambda_k + 16\sum_i \lambda_i^4 \\ &= (\sum_i \lambda_i)^4 + (\sum_i \lambda_i)^2 + 2 + 4(\sum_i \lambda_i)(\sum_i \lambda_i^3) + 16\sum_i \lambda_i^4,\end{aligned}$$

where we used that $\sum_i \lambda_i^2 = 1$. Again because of $\sum_i \lambda_i^2 = 1$ we have $|\lambda_i| \leq 1$ for all i and hence $|\sum_i \lambda_i^3| \leq \sum_i \lambda_i^2 = 1$ and similarly $\sum_i \lambda_i^4 \leq \sum_i \lambda_i^2 = 1$. Also observe that $|\sum_i \lambda_i| \leq 1 + (\sum_i \lambda_i)^2$. Combining these inequalities with the above expressions for $\mathbb{E}|\langle aa^*, U\rangle|^4$ and $(\mathbb{E}|\langle aa^*, U\rangle|^2)^2$, we obtain the inequality

$$\mathbb{E}|\langle aa^*, U\rangle|^4 \leq 24(\mathbb{E}|\langle aa^*, U\rangle|^2)^2.$$

Combining this with (19) and (20), we obtain

$$Q_{1/\sqrt{2}} \geq \frac{1}{96}.$$

Thus we choose $\xi = \frac{1}{2\sqrt{2}}$.

In order to estimate W_m, we use Lemma 10 to obtain

$$W_m = \mathbb{E} \sup_{Y \in E_r} \operatorname{tr}(H \cdot Y) \leq 2\sqrt{r} \cdot \mathbb{E}\|H\|_\infty. \tag{21}$$

By the arguments in [75, Section 5.4.1] we have $\mathbb{E}\|H\|_\infty \leq c_2\sqrt{n}$ if $m \geq c_3 n$ for suitable constants c_2, c_3, see also [74, Section 8]. Choosing $t = c_4\sqrt{m}$ and $m \geq cnr$ for suitable constants c, c_4, the proof of Theorem 2 is completed.

Remark 11. In [13], a uniform result for phase retrieval in the Gaussian case is proved using an inexact dual certificate. One can write down a generalization of this dual certificate for the rank r-case, but following the arguments of loc. cit., the resulting number of required measurements then seems to depend significantly worse than linearly on r. It might be possible to rather adapt the arguments in [38, 37] based on a different construction of a dual certificate in order to derive linear scaling of m in r, but the resulting proof would be more complicated than ours (and likely lead to more logarithmic factors).

4.2. **Proof of Theorem 3.** Let us now turn to proving the analogous result for complex projective 4-designs. It is convenient to rescale the (normalized) 4-design vectors as

$$\tilde{w}_i := \sqrt[4]{(n+1)n}\, w_i \quad \forall i = 1, \ldots, N. \tag{22}$$

This mimics the expected length of random Gaussian vectors (which corresponds to $\mathbb{E}\|a_j\|_2^2 = n$) and we will call the system $\{\tilde{w}_i\}$ a *super-normalized* 4-design. We can apply the same technique as in the proof of Theorem 2, provided that we can derive a suitable lower bound for $Q_{2\xi}$ for some $0 < \xi < 1/2$ and an upper bound for $\mathbb{E}\|H\|_\infty$. The following two technical propositions serve this purpose.

Proposition 12. *Assume that a is drawn at random from a super-normalized weighted 4-design. Then*

$$Q_\xi = \inf_{Z \in E_r} \mathbb{P}\left(|\mathrm{tr}\,(aa^* Z)\,| \geq \xi\right) \geq \frac{(1-\xi^2)^2}{24} \tag{23}$$

for all $\xi \in [0,1]$.

The proof of this statement is similar to the proof of Theorem 4 in [3] and – likewise – equation (15) in [55]. However, since we are interested in a bound on the probability of an event happening, rather than bounding an expectation value, we use the Payley-Zygmund inequality instead of Berger's one [9] (which states $\mathbb{E}\left[|S|\right] \geq \mathbb{E}\left[S^2\right]^{3/2} \mathbb{E}\left[S^4\right]^{-1/2}$).

Proof. The desired statement follows, if we can show that

$$\mathbb{P}\left(|\mathrm{tr}\,(aa^* Z)\,| \geq \xi\right) \geq \frac{(1-\xi^2)^2}{24} \tag{24}$$

holds for any matrix $Z \in \mathcal{H}_n$ obeying $\|Z\|_2 = 1$. For such Z we define the random variable $S := |\mathrm{tr}\,(aa^* Z)\,|$. Since a is chosen at random from a (super-normalized) complex projective 4-design, we can use the design's defining property (5) together with (6) to evaluate the second and fourth moment of S. Indeed,

$$\begin{aligned}
\mathbb{E}S^2 &= \mathbb{E}\mathrm{tr}\,(aa^*Z)^2 = \mathrm{tr}\left(\mathbb{E}\,(aa^*)^{\otimes 2} Z^{\otimes 2}\right) = \mathrm{tr}\left(\sum_{i=1}^N p_i\,(\tilde{w}_i\tilde{w}_i^*)^{\otimes 2} Z^{\otimes 2}\right) \\
&= (n+1)n\;\mathrm{tr}\left(\sum_{i=1}^N p_i\,(w_i w_i^*)^{\otimes 2} Z^{\otimes 2}\right) = (n+1)n\binom{n+1}{2}^{-1}\mathrm{tr}\left(P_{\mathrm{Sym}^2} Z^{\otimes 2}\right) \\
&= 2\mathrm{tr}\left(P_{\mathrm{Sym}^2} Z^{\otimes 2}\right)
\end{aligned}$$

and likewise

$$\mathbb{E}S^4 = \mathbb{E}\mathrm{tr}\,(aa^*Z)^4 = \mathrm{tr}\left(\sum_{i=1}^N p_i\,(\tilde{w}_i\tilde{w}_i^*)^{\otimes 4} Z^{\otimes 4}\right) = \frac{4!(n+1)n}{(n+3)(n+2)}\mathrm{tr}\left(P_{\mathrm{Sym}^4} Z^{\otimes 4}\right).$$

The remaining right hand sides are standard expressions in multilinear algebra and can for instance be calculated using wiring calculus. Indeed, Lemma 17 in the appendix implies that

$$\mathbb{E}S^2 = 2\mathrm{tr}\left(P_{\mathrm{Sym}^2} Z^{\otimes 2}\right) = \mathrm{tr}(Z)^2 + \mathrm{tr}(Z^2) = \mathrm{tr}(Z)^2 + 1, \tag{25}$$

because $\mathrm{tr}(Z^2) = \|Z\|_F^2 = 1$ by assumption, hence,

$$(\mathbb{E}S^2)^2 \geq \max\{1, \mathrm{tr}(Z)^4\}.$$

Similarly, Lemma 17 assures

$$\begin{aligned}
\mathbb{E}S^4 &= \frac{4!(n+1)n}{(n+3)(n+2)}\mathrm{tr}\left(P_{\mathrm{Sym}^4} Z^{\otimes 4}\right) \\
&= \frac{(n+1)n}{(n+3)(n+2)}\left(6\mathrm{tr}(Z^4) + 8\mathrm{tr}(Z)\mathrm{tr}(Z^3) + 6\mathrm{tr}(Z)^2\mathrm{tr}(Z^2) + 3\mathrm{tr}(Z^2)^2 + \mathrm{tr}(Z)^4\right) \\
&\leq \left(6\mathrm{tr}(Z^4) + 8\mathrm{tr}(Z)\mathrm{tr}(Z^3) + 6\mathrm{tr}(Z)^2 + \mathrm{tr}(Z)^4 + 3\right),
\end{aligned}$$

where the simplifications in the last line are due to $\mathrm{tr}(Z^2) = \|Z\|_F^2 = 1$ and $\frac{(n+1)n}{(n+3)(n+2)} \leq 1$. Using the hierarchy of Schatten-p-norms – in particular $\mathrm{tr}(Z^4) = \|Z\|_4^4 \leq \|Z\|_2^4 = 1$ and $\mathrm{tr}(Z^3) \leq \|Z\|_3^3 \leq \|Z\|_2^3 = 1$ – yields

$$\begin{aligned}
\mathbb{E}S^4 &\leq 6\mathrm{tr}(Z^4) + 8\mathrm{tr}(Z)\mathrm{tr}(Z^3) + 6\mathrm{tr}(Z)^2 + \mathrm{tr}(Z)^4 + 3 \\
&\leq \left(6\|Z\|_4^4 + 8\|Z\|_3^3 + 10\right)\max\left\{1, \mathrm{tr}(Z)^4\right\} \leq 24\max\left\{1, \mathrm{tr}(Z)^4\right\}.
\end{aligned}$$

Having precise knowledge of the second and fourth moments and the trivial fact that $\mathrm{tr}(Z)^2 \geq 0$ allows us to use the Payley-Zygmund inequality (for the random variable S^2) to bound

$$\begin{aligned}
\mathbb{P}\left(|\mathrm{tr}\left(aa^*Z\right)| \geq \xi\right) &= \mathbb{P}\left(S^2 \geq \xi^2\right) \geq \mathbb{P}\left(S^2 \geq \xi^2\left(1 + \mathrm{tr}(Z)^2\right)\right) \\
&= \mathbb{P}\left(S^2 \geq \xi^2\mathbb{E}S^2\right) \geq \left(1-\xi^2\right)^2\frac{(\mathbb{E}S^2)^2}{\mathbb{E}S^4} \\
&\geq (1-\xi^2)^2\frac{\max\{1, \mathrm{tr}(Z)^4\}}{24\max\{1, \mathrm{tr}(Z)^4\}} = \frac{(1-\xi^2)^2}{24}.
\end{aligned}$$

This completes the proof. □

Proposition 13. *Let H be the matrix defined in (18), where the a_j's are chosen independently at random from a super-normalized weighted 1-design. Then it holds that*

$$\mathbb{E}\|H\|_\infty \leq c_4\sqrt{n\log(2n)} \quad \textit{with } c_4 = 3.1049, \tag{26}$$

provided that $m \geq 2n\log n$.

Proof. Since the ϵ_j's in the definition of H form a Rademacher sequence, the non-commutative Khintchine inequality [75, p. 19], see also [33, Exercise 8.6(d)], is applicable and yields

$$\begin{aligned}
\mathbb{E}\|H\|_\infty &= \mathbb{E}_a\mathbb{E}_\epsilon\frac{1}{\sqrt{m}}\left\|\sum_{j=1}^m \epsilon_j a_j a_j^*\right\|_\infty \leq \sqrt{\frac{2\log(2n)}{m}}\mathbb{E}_a\left\|\left(\sum_{j=1}^m \left(a_j a_j^*\right)^2\right)^{1/2}\right\|_\infty \\
&= \sqrt{\frac{2\log(2n)}{m}}\mathbb{E}_a\left\|\sqrt{(n+1)n}\sum_{j=1}^m a_j a_j^*\right\|_\infty^{1/2} \leq \sqrt{\frac{2\sqrt{2}n\log(2n)}{m}}\left(\mathbb{E}_a\left\|\sum_{j=1}^m a_j a_j^*\right\|_\infty\right)^{1/2}.
\end{aligned} \tag{27}$$

Here we have used super-normalization of our design vectors $(a_j a_j^*)^2 = \|a_j\|_2^2 a_j a_j^* = \sqrt{(n+1)n}a_j a_j^*$ according to (22), the fact that $\|Z^{1/2}\|_\infty = \|Z\|_\infty^{1/2}$ holds for $Z \in \mathcal{H}_d$ arbitrary and Jensen's inequality in the last estimate. It remains to bound $\mathbb{E}\|\sum_j a_j a_j^*\|_\infty$. To this end, we will use the matrix Chernoff inequality of Theorem 15 for $X_j = a_j a_j^*$ and calculate

$$\begin{aligned}
\|X_j\|_\infty &= \|a_j a_j^*\|_\infty = \|a_j\|_2^2 \le \max_{1\le i\le N} \|\tilde{w}_i\|_2^2 = \sqrt{(n+1)n} \le \sqrt{2}n =: R, \qquad (28)\\
\|\sum_{j=1}^m \mathbb{E}X_j\|_\infty &= \|\sum_{j=1}^m \sum_{i=1}^N p_i \tilde{w}_i \tilde{w}_i^* \|_\infty = m\sqrt{n(n+1)} \left\| \sum_{i=1}^N p_i w_i w_i^* \right\|_\infty \\
&= m\sqrt{(n+1)n} \left\| \frac{1}{n} \mathrm{id} \right\|_\infty = \frac{m\sqrt{(n+1)n}}{n} \le \sqrt{2}m, \qquad (29)
\end{aligned}$$

where we once more have taken into account super-normalization and used the 1-design property. Theorem 15 together with the assumption $m \ge 2n \log n$ implies that, for any $\tau > 0$,

$$\begin{aligned}
\mathbb{E}\|\sum_{j=1}^m a_j a_j^*\|_\infty &\le \frac{e^\tau - 1}{\tau}\sqrt{2}m + \tau^{-1}\sqrt{2}n\log(n) \le \frac{e^\tau - 1}{\tau}\sqrt{2}m + \tau^{-1}\sqrt{2}m/2 \\
&= \left(\frac{e^\tau - 1}{\tau}\sqrt{2} + \frac{1}{\sqrt{2}\tau} \right) m.
\end{aligned}$$

The choice $\tau = 1.27$ approximately minimizes the above expression and yields

$$\mathbb{E}\|\sum_{j=1}^m a_j a_j^*\|_\infty \le c_5 m \quad \text{with } c_5 = 3.4084.$$

Combining this estimate with (27) yields the desired statement with $c_4 = 2^{3/4}\sqrt{c_5} = 3.1049$. □

Now we are ready to prove the second main theorem of this work.

Proof of Theorem 3. The proof of Theorem 2 shows that we only need suitable bounds for $Q_{2\xi}$ and for $\mathbb{E}\|H\|_\infty$ (both notions are defined analogously to the Gaussian case). Fix $0 < \xi < 1/2$ arbitrary. For any such ξ, a lower bound for $Q_{2\xi}$ is provided by Proposition 12 and an upper bound for $\mathbb{E}\|H\|_\infty$ in this case can be obtained from Proposition 13. Setting $m = C_4 nr \log n$, choosing the constants C_4, C_5 and C_6 appropriately (depending on the particular choice of ξ) and applying Theorem 8 then yields the desired result in complete analogy to the Gaussian case (proof of Theorem 2). □

Remark 14. The difference in the sampling rate m by a factor proportional to $\log n$ in Theorems 2 and 3 stems from the fact that Proposition 13 is by a factor of $\sqrt{\log(n)}$ weaker than its Gaussian analogue [75, Section 5.4.1], where $\mathbb{E}\|H\|_\infty \le c_2\sqrt{n}$.

4.3. **Proof of Theorem 2 for real Gaussian vectors.** As already mentioned in paragraph 2.3.1 the proof of this statement is almost identical to the proof of Theorem 2. The only difference is the estimate of $Q_{2\xi}$. Using the moments of the real instead of the complex standard Gaussian distribution, the reasoning in the proof of Theorem 2 yields the estimates $\mathbb{E}|\langle aa^*, U\rangle|^2 \ge 2$, (compare also with [74]). Using real moments, one further obtains $\mathbb{E}|\langle aa^*, U\rangle|^4 \le 27(\mathbb{E}|\langle aa^*, U\rangle|^2)^2$ (alternatively one can use Gaussian hypercontractivity as done in [74], which gives the factor 81 instead of 27.) This yields $Q_1 \ge \frac{1}{108}$, and the rest of the proof is the same as before.

4.4. **Proof for recovery of positive semidefinite matrices.** The only part in the proof of the recovery result for positive semidefinite matrices stated in Section 2.3.2 that slightly differs from the one for arbitrary Hermitian matrices, is the proof of a corresponding version of Lemma 10. The subdifferential of the function f introduced in (9) slightly differs from the subdifferential of the nuclear norm. For $X = \sum_{i=1}^r \lambda_i x_i x_i^*$, where all λ_i are nonzero, $\partial f(X)$ consists of all matrices of the form

$$S = \sum_{i=1}^r x_i x_i^* + S_2,$$

where $S_2 \in \mathcal{H}_n$ has the property that $S_2 x_i = 0$ for all $i \in \{1, \dots, r\}$ and all eigenvalues of S_2 do not exceed 1. Hence we choose (in the notation of the proof of Lemma 10)

$$S = \sum_{i=1}^{r} x_i x_i^* + \tau^{-1} A_4 \in \partial f(X).$$

Then the remainder of the proof of Lemma 10 is the same.

4.5. **Proof of Theorem 5.** The proof of this generalized statement proceeds along the same lines as the one of Theorem 3. However, Propositions 12 and 13 – as well as their respective proofs – have to be slightly altered due to the weaker requirements imposed by Theorem 5.

4.5.1. *Generalized version of Proposition 12.* Under the assumptions of Theorem 5, a weaker version of (23), namely

$$Q_\xi = \inf_{Z \in E_r} \mathbb{P}\left(|\operatorname{tr}(aa^* Z)| \geq \xi \right) \geq \frac{(1-2\xi^2)^2}{192} \tag{30}$$

for all $0 \leq \xi \leq 1/\sqrt{2}$ is still valid. This statement can be shown analogously to Proposition 12. However, one has to establish bounds on the second and fourth moments in a slightly more involved way, depending also on the type of design accuracy. Let us start with generalizing the second moment estimate of $S := |\operatorname{tr}(aa^* Z)|$ for an approximate 4-design with operator norm accuracy $\theta_\infty \leq 1/(16r^2)$:

$$\begin{aligned}
\mathbb{E}S^2 &= (n+1)n \left(\sum_{i=1}^{N} p_i \left(w_i w_i^* \right)^{\otimes 2}, Z^{\otimes 2} \right) \\
&= 2\left(P_{\mathrm{Sym}^2}, Z^{\otimes 2} \right) + (n+1)n \left(\sum_{i=1}^{N} p_i \left(w_i w_i^* \right)^{\otimes 2} - \binom{n+1}{2}^{-1} P_{\mathrm{Sym}^2}, Z^{\otimes 2} \right) \\
&\geq 2 \left| \left(P_{\mathrm{Sym}^2}, Z^{\otimes 2} \right) \right| - (n+1)n \left\| \sum_{i=1}^{N} p_i \left(w_i w_i^* \right)^{\otimes 2} - \binom{n+1}{2}^{-1} P_{\mathrm{Sym}^2} \right\|_\infty \left\| Z^{\otimes 2} \right\|_1 \qquad (31) \\
&\geq 2 \left| \left(P_{\mathrm{Sym}^2}, Z^{\otimes 2} \right) \right| - 2\theta_\infty \|Z\|_1^2 \geq 2 \left| \left(P_{\mathrm{Sym}^2}, Z^{\otimes 2} \right) \right| - \frac{8r}{16r^2}, \\
&> 2 \left| \left(P_{\mathrm{Sym}^2}, Z^{\otimes 2} \right) \right| - 1/2, \qquad (32)
\end{aligned}$$

where we have used the fact that $\left(P_{\mathrm{Sym}^2}, Z^{\otimes 2} \right) = \left| \left(P_{\mathrm{Sym}^2}, Z^{\otimes 2} \right) \right|$ (see Lemma 17), the matrix Hölder inequality and the fact that $\|Z\|_1 \leq 2\sqrt{r}$ – see (16). The estimates for designs with nuclear norm accuracy $\theta_1 \leq 1/4$ is very similar. Replacing the matrix Hölder inequality in (31) by

$$\left(\sum_{i=1}^{N} p_i \left(w_i w_i^* \right)^{\otimes 2} - \binom{n+1}{2}^{-1} P_{\mathrm{Sym}^2}, Z^{\otimes 2} \right) \geq - \left\| \sum_{i=1}^{N} p_i \left(w_i w_i^* \right)^{\otimes 2} - \binom{n+1}{2}^{-1} P_{\mathrm{Sym}^2} \right\|_1 \left\| Z^{\otimes 2} \right\|_\infty$$

yields the same lower bound (32) due to $\|Z^{\otimes 2}\|_\infty = \|Z\|_\infty^2 \leq \|Z\|_2^2 = 1$ (where the last equality follows from $Z \in E_r$). Applying Lemma 17 then yields

$$\mathbb{E}S^2 \geq \operatorname{tr}(Z)^2 + 1/2 \quad \text{and} \quad (\mathbb{E}S^2)^2 \geq \frac{1}{4} \max\{1, \operatorname{tr}(Z)^4\}.$$

which is the (slightly weaker) analogue of (25). Likewise we derive a fourth moment bound:

$$
\begin{aligned}
\mathbb{E}S^4 &= \left(\mathbb{E}\left[(aa^*)^{\otimes 4}\right], Z^{\otimes 4}\right) = (n+1)^2 n^2 \left(\sum_{i=1}^{N} p_i \left(w_i w_i^*\right)^{\otimes 4}, Z^{\otimes 4}\right) \\
&\le (n+1)^2 n^2 \binom{n+3}{4}^{-1} |\left(P_{\mathrm{Sym}^4}, Z^{\otimes 4}\right)| \\
&+ (n+1)^2 n^2 \left\| \sum_{i=1}^{N} p_i \left(w_i w_i^*\right)^{\otimes 4} - \binom{n+3}{4}^{-1} P_{\mathrm{Sym}^4} \right\|_\infty \|Z^{\otimes 4}\|_1 \\
&\le \frac{4!(n+1)n}{(n+3)(n+2)} \left(|\left(P_{\mathrm{Sym}^4}, Z^{\otimes 4}\right) + \theta_\infty \|Z\|_1^4\right) \le |4! \left(P_{\mathrm{Sym}^4}, Z^{\otimes 4}\right)| + 4! \frac{16r^2}{16r^2}.
\end{aligned}
$$

As above, using the nuclear norm accuracy $\theta_1 \le 1/4$ instead of the operator norm accuracy yields the bound $\mathbb{E}\left[S^4\right] \le |4!\left(P_{\mathrm{Sym}^4}, Z^{\otimes 4}\right)| + 4!/4 < |4!\left(P_{\mathrm{Sym}^4}, Z^{\otimes 4}\right)| + 4!$. Lemma 17 yields then in both cases

$$
\begin{aligned}
\mathbb{E}\left[S^4\right] &\le |4!\mathrm{tr}\left(P_{\mathrm{Sym}^4} Z^{\otimes 4}\right)| + 24 \le 6\mathrm{tr}(Z^4) + 8|\mathrm{tr}(Z)\mathrm{tr}(Z^3)| + 6\mathrm{tr}(Z)^2 + \mathrm{tr}(Z)^4 + 27 \\
&\le 48 \max\{1, \mathrm{tr}(Z)^4\},
\end{aligned}
$$

compare the proof of Proposition 12. Having these bounds at hand, allows for applying the Payley Zygmund inequality to obtain

$$
\begin{aligned}
\mathbb{P}\left(|\mathrm{tr}\left(aa^* Z\right)| \ge \xi\right) &= \mathbb{P}\left(S^2 \ge \xi^2\right) \ge \mathbb{P}\left(S^2 \ge 2\xi^2 \left(1/2 + \mathrm{tr}(Z)^2\right)\right) \ge \mathbb{P}\left(S^2 \ge 2\xi^2 \mathbb{E}\left[S^2\right]\right) \\
&\ge (1-2\xi^2)^2 \frac{(\mathbb{E}S^2)^2}{\mathbb{E}S^4} \ge (1-2\xi^2)^2 \frac{\max\{1, \mathrm{tr}(Z)^4\}/4}{48 \max\{1, \mathrm{tr}(Z)^4\}} = \frac{(1-2\xi^2)^2}{192}.
\end{aligned}
$$

The proof is completed.

4.5.2. *Generalized version of Proposition 13.* The assumptions in Theorem 5 assure that (26) is still valid, possibly with a larger absolute constant c_4. Again, the proof of this generalized statement is very similar to the proof of Proposition 13. Indeed, only the bound (29) for the matrix Chernoff inequality needs to be slightly altered. The assumption (11) implies that

$$
\| \sum_{j=1}^{m} \mathbb{E}\left[X_j\right] \|_\infty \le m\sqrt{(n+1)n} \left(\| \frac{1}{n} \mathrm{id} \|_\infty + \| \sum_{i=1}^{N} p_i w_i w_i^* - \frac{1}{n} \mathrm{id} \|_\infty \right) \le 2\sqrt{2}m.
$$

Consequently, applying the matrix Chernoff inequality yields (26) with a slightly larger absolute constant c_4.

5. Appendix

5.1. **Schatten p-norms.** Recall from Section 1.2 that for $1 \le p < \infty$, the Schatten-p-norm on $\mathcal{H}_n$ is defined as

$$
\|Z\|_p = \mathrm{tr}\left(|Z|^p\right)^{1/p} = \left(\sum_{i=1}^{n} |\lambda_i|^p\right)^{1/p},
$$

where $\lambda_1, \ldots, \lambda_n$ denote the n eigenvalues of $Z \in \mathcal{H}_n$. For $p = \infty$ one defines similarly

$$
\|Z\|_\infty = \max\{|\lambda_1|, \ldots, |\lambda_n|\},
$$

i.e., $\|Z\|_\infty$ is the spectral norm of Z. The Frobenius norm $\|\cdot\|_F = \|\cdot\|_2$ is induced by the the Hilbert-Schmitt (or Frobenius) scalar product

$$
(X, Y) = \mathrm{tr}\left(XY\right),
$$

which makes $\mathcal{H}_n$ a Hilbert space. The Schatten-p norms are non-increasing in p, i.e. for any $0 < p \le p' \le \infty$

$$
\|Z\|_p \ge \|Z\|_{p'} \tag{33}
$$

holds for all $Z \in \mathcal{H}_n$. The following relations provide converse inequalities for particular instances of Schatten p-norms that are used frequently in our work:

$$\|Z\|_1 \leq \sqrt{\operatorname{rank}(Z)}\|Z\|_2 \quad \text{and} \quad \|Z\|_2 \leq \sqrt{\operatorname{rank}(Z)}\|Z\|_\infty \quad \text{for all } Z \in \mathcal{H}_n. \tag{34}$$

In addition, we often use a particular instance of the matrix Hölder inequality, namely

$$|\,(X,Y)\,| \leq \|X\|_1\|Y\|_\infty \quad \text{for all } X, Y \in \mathcal{H}_n. \tag{35}$$

5.2. **Matrix Chernoff inequality.** The matrix version of the classical Chernoff inequality for the expection of a sum of independent random matrices shown in [73, Theorem 5.1.1] (see also [72]) reads as follows.

Theorem 15. *Let $X_1, \ldots, X_m$ be a sequence of independent random positive definite matrices in $\mathcal{H}_n$ satisfying*

$$\|X_\ell\|_\infty \leq L \quad \textit{almost surely for all } \ell = 1, \ldots, m.$$

Then, for any $\tau > 0$, their sum obeys

$$\mathbb{E}\|\sum_{\ell=1}^m X_\ell\|_\infty \leq \frac{e^\tau - 1}{\tau}\|\sum_{\ell=1}^m \mathbb{E}X_\ell\|_\infty + \tau^{-1} L \log n.$$

5.3. **Multilinear algebra.** We briefly repeat some standard concepts in multilinear algebra which are convenient for our proof of Proposition 12. They can be found in any textbook on multilinear algebra – e.g. [49] – but we nonetheless include them here for the sake of being self-contained.

Let $V_1, \ldots, V_k$ be (finite dimensional, complex) vector spaces and let $V_1^*, \ldots, V_k^*$ denote their duals. A function $f : V_1 \times \cdots \times V_k \to \mathbb{C}$ is *multilinear*, if it is linear in each space V_i. We denote the space of such functions by $V_1^* \otimes \cdots \otimes V_k^*$ and call it the *tensor product* of $V_1^*, \ldots, V_k^*$. Consequently, for one fixed n-dimensional vector space V, the tensor product $(V)^{\otimes k} = \bigotimes_{i=1}^k V$ is the space of all multilinear functions

$$f : \underbrace{(V)^* \times \cdots \times (V)^*}_{k \text{ times}} \mapsto \mathbb{C}, \tag{36}$$

and we call the elementary elements $z_1 \otimes \cdots \otimes z_k$ the *tensor product* of the vectors $z_1, \ldots, z_k \in V$.

With this notation, the space of linear maps $V \to V$ ($n \times n$-matrices) corresponds to the tensor product $\mathcal{M}_n := V \otimes V^*$ which is spanned by $\{x \otimes y^* : \, x, y \in V\}$ – the set of all rank-1 matrices. Using this tensor product description of $\mathcal{M}_n$ allows for defining the (matrix) tensor product $\mathcal{M}_n^{\otimes k}$ in complete analogy to above. We refer to its elements $Z_1 \otimes \cdots \otimes Z_k$ as the tensor product of the matrices $Z_1, \ldots, Z_k \in \mathcal{M}_n$.

On this tensor space, we define the *partial trace* (over the i-th tensor system) to be the natural contraction

$$\begin{aligned} \operatorname{tr}_i : \; \mathcal{M}_n^{\otimes k} \quad &\to \quad \mathcal{M}_n^{\otimes(k-1)} \\ Z_1 \otimes \cdots \otimes Z_k \quad &\mapsto \quad \operatorname{tr}(Z_i) Z_1 \otimes \cdots \otimes Z_{i-1} \otimes Z_{i+1} \otimes \cdots \otimes Z_k. \end{aligned}$$

The partial trace over multiple systems can then be obtained by concatenating individual traces of this form, e.g.

$$\operatorname{tr}_{i,j} = \operatorname{tr}_i \circ \operatorname{tr}_j : \mathcal{M}_n^{\otimes k} \to \mathcal{M}_n^{\otimes(k-2)} \tag{37}$$

for $1 \leq i < j \leq k$ arbitrary and so forth. A particular property of arbitrary partial traces is that they preserve positive semidefiniteness – see e.g. [61, Section 8.3.1] or any lecture notes on quantum information theory. If a matrix $Z \in \mathcal{M}_n^{\otimes k}$ is positive semidefinite, then $\operatorname{tr}_i(Z) \in \mathcal{M}^{\otimes(k-1)}$ is again positive semidefinite for any $1 \leq i \leq k$. This behavior naturally extends to multiple partial traces in the sense of (37). The *full trace* corresponds to

$$\begin{aligned} \operatorname{tr} := \operatorname{tr}_{1,\ldots,k} : \; \mathcal{M}_n^{\otimes k} \quad &\to \quad \mathbb{C} \\ Z_1 \otimes \cdots \otimes Z_k \quad &\mapsto \quad \operatorname{tr}(Z_1) \cdots \operatorname{tr}(Z_k). \end{aligned}$$

This implies that the nuclear norm is multiplicative with respect to the tensor structure, i.e.,

$$\|Z_1 \otimes \cdots Z_k\|_1 = \operatorname{tr}(|Z_1| \otimes \cdots \otimes |Z_k|) = \operatorname{tr}(|Z_1|) \cdots \operatorname{tr}(|Z_k|) = \|Z_1\|_1 \cdots \|Z_k\|_1 \tag{38}$$

for $Z_1, \ldots, Z_k \in \mathcal{M}$ arbitrary. A singular value decomposition – see e.g. [77, Lecture 2] – reveals that the same is true for the operator norm, i.e.

$$\|Z_1 \otimes \cdots \otimes Z_k\|_\infty = \|Z_1\|_\infty \cdots \|Z_k\|_\infty. \tag{39}$$

Let us now return to the k-fold tensor space $V^{\otimes k}$ of n-dimensional complex vectors. We define the (symmetrizer) map $P_{\mathrm{Sym}^k} : (V)^{\otimes k} \to (V)^{\otimes k}$ via their action on elementary elements:

$$P_{\mathrm{Sym}^k}\left(z_1 \otimes \cdots \otimes z_k\right) := \frac{1}{k!} \sum_{\pi \in S_k} z_{\pi(1)} \otimes \cdots \otimes z_{\pi(k)}, \tag{40}$$

where S_k denotes the group of permutations of k elements. This map projects $(V)^{\otimes k}$ onto the totally symmetric subspace Sym^k of $(V)^{\otimes k}$ whose dimension [49, Exercise 2.6.3.5] is

$$\dim \mathrm{Sym}^k = \binom{n+k-1}{k}. \tag{41}$$

Using these basic concepts of multilinear algebra and (6), we can show that every approximate t-design is also an approximate design of lower order.

Lemma 16. *Every approximate t-design of accuracy measured either in operator- or trace-norm is also an approximate k-design of the same accuracy for any $1 \leq k \leq t$. Furthermore the accuracies θ_∞ and θ_1 are related via*

$$\theta_\infty \leq \theta_1 \leq n^t \theta_\infty. \tag{42}$$

This statement is implicitly proved in [3], where the authors use an equivalent definition of approximate t-designs as averaging sets of complex polynomials of degree at most (t,t). With this alternative definition, Lemma 16 follows naturally from the fact that every polynomial of degree at most (k,k) with $1 \leq k \leq t$ is a particular instance of a degree-(t,t)-polynomial. Here we provide an alternative proof that uses concepts from multilinear algebra and accesses Definition 4 directly. Such a proof idea is mentioned in [54, Section 2.2.3] and we include the full argument here for the sake of being self-contained.

Proof of Lemma 16. Let us start with proving the statement for the accuracy measured in operator norm. In this case, Definition 4 is equivalent to demanding

$$(1-\theta_\infty) \int_{\mathbb{C}P^{n-1}} (ww^*)^{\otimes t} \mathrm{d}w \leq \sum_{i=1}^N p_i \left(w_i w_i^*\right)^{\otimes t} \leq (1+\theta_\infty) \int_{\mathbb{C}P^{n-1}} (ww^*)^{\otimes t} \mathrm{d}w. \tag{43}$$

The desired statement follows if we can show that (43) implies a corresponding inequality for smaller tensor powers k. Fix $1 \leq k \leq t$ and note that the inequality chain (43) is preserved under taking arbitrary partial traces, because partial traces respect the positive semidefinite ordering. This in particular implies that

$$\begin{aligned}(1-\theta_\infty) \int_{\mathbb{C}P^{n-1}} \mathrm{tr}_{1,\ldots,(t-k)} \left((ww^*)^{\otimes t}\right) \mathrm{d}w \quad &\leq \quad \sum_{i=1}^N p_i \mathrm{tr}_{1,\ldots,(t-k)} \left((w_i w_i^*)^{\otimes t} \right) \\ &\leq \quad (1+\theta_\infty) \int_{\mathbb{C}P^{n-1}} \mathrm{tr}_{1,\ldots,(t-k)} \left((ww^*)^{\otimes t} \right) \mathrm{d}w\end{aligned}$$

remains valid. Due to normalization $\|w_i\|_{\ell_2} = 1$ and and since we calculate the integrals using preimages of the $w \in \mathbb{C}P^{n-1}$ in the unit sphere, these expressions can be readily calculated. Indeed,

$$\mathrm{tr}_{1,\ldots,(t-k)} \left((w_i w_i^*)^{\otimes t} \right) = (w_i w_i^*)^{\otimes k} |\langle w_i, w_i \rangle|^{2(t-k)} = (w_i w_i^*)^{\otimes k}$$

and

$$\int_{\mathbb{C}P^{n-1}} \mathrm{tr}_{1,\ldots,(t-k)} \left((ww^*)^{\otimes t} \right) \mathrm{d}w = \int_{\mathbb{C}P^{n-1}} (ww^*)^{\otimes k} |\langle w, w \rangle|^{2(t-k)} \mathrm{d}w = \int_{\mathbb{C}P^{n-1}} (ww^*)^{\otimes k} \mathrm{d}w.$$

The desired statement follows.

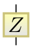
The analogous statement for accuracy measured in trace-norm directly follows from the fact that the nuclear norm is monotonic with respect to partial traces, i.e., $\|\mathrm{tr}_i(Z)\|_1 \leq \|Z\|_1$ for any $Z \in \mathcal{M}_n^{\otimes t}$ and $1 \leq i \leq t$ [77, Lecture 2]. Combining this with the calculations above reveals that

$$\begin{aligned} &\left\| \sum_{i=1}^N p_i \left(w_i w_i^*\right)^{\otimes k} - \int_{\mathbb{C}P^{n-1}} \left(ww^*\right)^{\otimes k} \mathrm{d}w \right\|_1 \\ &= \left\| \mathrm{tr}_{1,\ldots,t-k} \left(\sum_{i=1}^N p_i \left(w_i w_i^*\right)^{\otimes t} - \int_{\mathbb{C}P^{n-1}} \left(ww^*\right)^{\otimes t} \mathrm{d}w \right) \right\|_1 \\ &\leq \left\| \sum_{i=1}^N p_i \left(w_i w_i^*\right)^{\otimes t} - \int_{\mathbb{C}P^{n-1}} \left(ww^*\right)^{\otimes t} \mathrm{d}w \right\|_1 \leq \theta_1 . \end{aligned}$$

Finally, inequality (42) directly follows from comparing trace and operator norm on $\mathcal{M}_n^{\otimes t}$ which is isomorphic to the space of all $n^t \times n^t$-dimensional matrices.

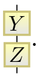
□

5.4. **Wiring calculus in multilinear algebra.** The defining properties (5), (10) of exact and approximate complex projective t-designs are phrased in terms of tensor spaces. For calculations in multilinear algebra – particularly if they involve (partial) traces– *wiring diagrams* [49, Chapter 2.11] are very useful, as they provide a way of computing contractions of tensors pictorially. Here we give a brief introduction that should suffice for our calculations and defer the interested reader to [38] and references therein for further reading.

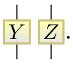
In wiring calculus, every tensor is associated with a box, and every index corresponds to a line emanating from this box. Two connected lines correspond to connected indices. The formalism becomes much clearer when applying it to matrix calculus. A matrix $Z : \mathbb{C}^n \to \mathbb{C}^n$ can be viewed as two-index-tensors $Z^i{}_j$ and is thus represented by a node  with upper line corresponding to the index i and the lower one to j. Two matrices Y, Z are multiplied by contracting Z's upper index with Y's lower one:

$$(YZ)^i{}_j = \sum_{k=1}^n Y^i{}_k Z^k{}_j .$$

In wiring calculus matrix multiplication is therefore represented by

$YZ =$ .

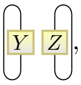
Tensor products of matrices are arranged in parallel, i.e.,

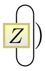
$Y \otimes Z =$ .

Taking traces of tensor products, e.g.,

$$Y \otimes Z \mapsto \mathrm{tr}(Y \otimes Z) = \sum_{i,j=1}^n Y^i{}_i Z^j{}_j$$

just corresponds to contracting parallel matrix indices and therefore

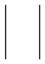
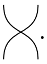
$\mathrm{tr}(Y \otimes Z) =$ ,

which straightforwardly extends to larger (and smaller, namely $\mathrm{tr}(Z) =$ ) tensor systems.

Finally, we are going to require *transpositions* on $(\mathbb{C}^n)^{\otimes t}$ which act by interchanging the i-th and j-th tensor factor. For example

$$\sigma_{(1,2)} \left(x \otimes y \otimes \cdots \right) = y \otimes x \otimes \cdots ,$$

with $x, y \in \mathbb{C}^n$ arbitrary. Note that these transpositions generate the full group of permutations. For $(\mathbb{C}^n)^{\otimes 2}$ there are only two transpositions, namely

$\underline{1} =$  (trivial permutation) and $\sigma_{(1,2)} =$ .

But for higher tensor systems more permutations can occur. In wiring calculus, permutations therefore act by interchanging different input and output lines.

We are now ready to prove the statements required in Proposition 12.

Lemma 17. *For an abritrary Hermitian matrix $Z \in \mathcal{H}_n$ and a positive integer m, it holds*

$$m!\left(P_{\mathrm{Sym}^m} Z^{\otimes m}\right) = \sum_{\substack{(j_1,\ldots,j_m)\in\mathbb{N}_0^m \\ \sum_{k=1}^m k j_k = m}} \frac{m!}{\prod_{k=1}^m j_k!\, k^{j_k}} \prod_{k=1}^m \mathrm{tr}(Z^k)^{j_k}.$$

In particular, for $m = 2$ we obtain

$$2\mathrm{tr}\left(P_{\mathrm{Sym}^2} Z^{\otimes 2}\right) = \mathrm{tr}(Z)^2 + \mathrm{tr}(Z^2),$$

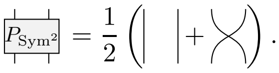
and for $m = 4$ we obtain

$$4!\ \mathrm{tr}\left(P_{\mathrm{Sym}^4} Z^{\otimes 4}\right) = \mathrm{tr}(Z)^4 + 8\mathrm{tr}(Z)\mathrm{tr}(Z^3) + 3\mathrm{tr}(Z^2)^2 + 6\mathrm{tr}(Z)^2\mathrm{tr}(Z^2) + 6\mathrm{tr}(Z^4).$$

Proof. We start with the case $m = 2$ and then extend the argument to the general case. The basic formula for P_{Sym^2} is given by

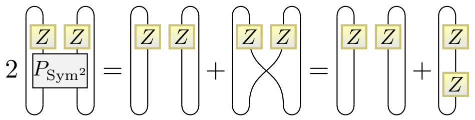
$$P_{\mathrm{Sym}^2} = \frac{1}{2}\sum_{\pi\in S_2}\pi = \frac{1}{2}\left(\underline{1} + \sigma_{(1,2)}\right),$$

and its pictorial counterpart is therefore



Applying the graphical calculus introduced above then yields

$$2\mathrm{tr}\left(P_{\mathrm{Sym}^2} Z^{\otimes 2}\right) =$$



$$= \mathrm{tr}(Z)^2 + \mathrm{tr}(Z^2),$$

which is the desired statement for $m = 2$.

Expanding $m!\,(P_{\mathrm{Sym}^m} Z^{\otimes m})$ analogously in the general case, we obtain for each $\pi \in S_m$ one summand which corresponds to a wiring diagram in which m copies of the node $\boxed{Z}$ are involved. More precisely, the wiring diagram corresponding to π is obtained by connecting for each $i \in \{1, \ldots, m\}$ the output line of the i-th copy of $\boxed{Z}$ to the input line of the $\pi(i)$-th copy of $\boxed{Z}$. If we write π as a product of k cyclic permutations, $\pi = c_1 \cdots c_k$, then the wiring diagram of π consists of k closed loops, one for each of the cyclic permutations $c_1, \ldots, c_k$. Write $c_i = (i_1, \ldots, i_{r_i})$. Then the loop corresponding to c_i connects r_i copies of $\boxed{Z}$. Hence the contribution of π to the whole sum is $\mathrm{tr}(Z^{r_1}) \cdots \mathrm{tr}(Z^{r_k})$. Thus for a given partition $m = r_1 + \ldots + r_k$ of m, any element of S_m which is the product of k cyclic (and disjoint) permutations of lengths $r_1, \ldots, r_k$ respectively gives the same contribution $\mathrm{tr}(Z^{r_1}) \cdots \mathrm{tr}(Z^{r_k})$.

Note that we can rewrite any partition of m in the form $m = j_1 \cdot 1 + \ldots + j_m \cdot m$, where j_i counts how often the summand i appears in that partition. It remains to count for each partition $m = j_1 \cdot 1 + \ldots + j_m \cdot m$ of m how many elements of S_m there are which are a product of precisely j_1 cyclic permutations of length 1, of precisely j_2 cyclic permutations of length 2 and so on (all the cyclic permutations being disjoint). It is easy to see (and well known, see for example [69, Proposition 1.3.2]) that there are precisely $\frac{m!}{\prod_{k=1}^m j_k!\, k^{j_k}}$ such permutations in S_m. Each of them contributes a summand $\mathrm{tr}(Z^1)^{j_1} \ldots \mathrm{tr}(Z^m)^{j_m}$ to $m!\,(P_{\mathrm{Sym}^m} Z^{\otimes m})$. This gives the claimed formula. □

Acknowledgements. RK is glad to acknowledge inspiring discussions with D. Gross and helpful comments from J. Aaberg.

The work of RK is supported by scholarship funds from the State Graduate Funding Program of Baden-Württemberg, the Excellence Initiative of the German Federal and State Governments (Grant ZUK 43), the ARO under contracts, W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), the Freiburg Research Innovation Fund, and the DFG. HR and UT acknowledge funding by the European Research Council through the Starting Grant StG 258926 (SPALORA). RK and HR would like to thank the *Mathematisches Forschungsinstitut Oberwolfach* and the organizers of the Oberwolfach workshop *Mathematical Physics meets Sparse Recovery* (April 2014, Workshop ID: 1416a), where discussions on the topic of this article have started.

References

[1] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *preprint*, 2012.

[2] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon. Phase retrieval with polarization. *SIAM J. Imaging Sci.*, 7:35–66, 2014.

[3] A. Ambainis and J. Emerson. Quantum t-designs: t-wise independence in the quantum world. In *22nd Annual IEEE Conference on Computational Complexity, Proceedings*, pages 129–140, 2007.

[4] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Inform. Inference*, 3(3):224–294, 2014.

[5] C. Bachoc and M. Ehler. Signal reconstruction from the magnitude of subspace components. *arXiv:1209.5986*, 2012.

[6] B. Bajnok. Construction of spherical t-designs. *Geom. Dedicata*, 43:167–179, 1992.

[7] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin. Painless reconstruction from magnitudes of frame coefficients. *J. Fourier Anal. Appl.*, 15:488–501, 2009.

[8] R. Balan, P. Casazza, and D. Edidin. On signal reconstruction without phase. *Appl. Comput. Harmon. Anal.*, 20(3):345–356, 2006.

[9] B. Berger. The fourth moment method. *SIAM J. on Comp.*, 26(4):1188–1207, 1997.

[10] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge Univ. Press, 2004.

[11] F. G. Brandao, A. W. Harrow, and M. Horodecki. Local random quantum circuits are approximate polynomial-designs. *preprint arXiv:1208.0692*, 2012.

[12] O. Bunk, A. Diaz, F. Pfeiffer, C. David, B. Schmitt, D. K. Satapathy, and J. F. van der Veen. Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels. *Acta Crystallographica Section A: Foundations of Crystallography*, 63(4):306–314, 2007.

[13] E. Candès and X. Li. Solving quadratic efquations via PhaseLift when there are about as many equations as unknowns. *Found. Comput. Math.*, pages 1–10, 2013.

[14] E. Candes, X. Li, and M. Soltanolkotabi. Phase Retrieval via Wirtinger Flow: Theory and Algorithms. *ArXiv e-prints*, jul 2014.

[15] E. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Appl. Comput. Harmonic Anal.*, to appear.

[16] E. J. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011.

[17] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[18] E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[19] E. J. Candès and T. Tao. The power of matrix completion: near-optimal convex relaxation. *IEEE Trans. Information Theory*, 56(5):2053–2080, 2010.

[20] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[21] Y. Chen. Incoherence-optimal matrix completion. *preprint arXiv:1310.0154*, 2013.

[22] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Completing any low-rank matrix, provably. *ArXiv:1306.2979*, 2013.

[23] P. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

[24] C. Dankert, R. Cleve, J. Emerson, and E. Livine. Exact and approximate unitary 2-designs: constructions and applications. *arXiv preprint quant-ph/0606161*, 2006.

[25] P. De La Harpe and C. Pache. Cubature formulas, geometrical designs, reproducing kernels, and markov operators. In *Infinite groups: geometric, combinatorial and dynamical aspects*, pages 219–267. Springer, 2005.

[26] P. Delsarte, J. Goethals, and J. Seidel. Spherical codes and designs. *Geom. Dedicata*, 6:363–388, 1977.

[27] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[28] M. Fazel. *Matrix rank minimization with applications.* PhD thesis, 2002.

[29] C. Fienup and J. Dainty. Phase retrieval and image reconstruction for astronomy. In H. Stark, editor, *Image Recovery: Theory and Application*, pages 231–275. Academic Press, San Diego, 1987.

[30] J. Fienup. Phase retrieval algorithms: A comparison. *Appl. Opt.*, 21(15):2758–2769, 1982.

[31] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New J. Phys.*, 14:095022, 2012.

[32] M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM J. Optim.*, 21(4):1614–1640, 2011.

[33] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[34] R. Gerchberg and W. Saxton. Phase retrieval by iterated projection. *Optik*, 35, 1972.

[35] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57:1548–1566, 2011.

[36] D. Gross, K. Audenaert, and J. Eisert. Evenly distributed unitaries: on the structure of unitary designs. *J. Math. Phys.*, 48:052104, 22, 2007.

[37] D. Gross, F. Krahmer, and R. Kueng. Improved recovery guarantees for phase retrieval from coded diffraction patterns. *preprint arXiv:1402.6286*, 2014.

[38] D. Gross, F. Krahmer, and R. Kueng. A partial derandomization of PhaseLift using spherical designs. *J. Fourier Anal. Appl.*, to appear.

[39] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Phys. Rev. Lett.*, 105:150401, Oct 2010.

[40] R. Harrison. Phase problem in crystallography. *JOSA A*, 10(5):1046–1055, 1993.

[41] A. Hayashi, T. Hashimoto, and M. Horibe. Reexamination of optimal quantum state estimation of pure states. *Phys. Rev. A*, 72, SEP 2005.

[42] P. Hayden, D. Leung, P. W. Shor, and A. Winter. Randomizing quantum states: Constructions and applications. *Commun. Math. Phys.*, 250(2):371–391, 2004.

[43] R. KeshavanH., A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Trans. Inform. Theory*, 56(6):2980 – 2998, 2010.

[44] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *ArXiv:1312.3580*, dec 2013.

[45] J. Korevaar and J. Meyers. Chebyshev-type quadrature on multidimensional domains. *J. Approx. Theory*, 79:144–164, 1994.

[46] G. Kuperberg. Numerical cubature using error-correcting codes. *SIAM J. Numer. Anal.*, 44(3):897–907, 2006.

[47] A. Kyrillidis and V. Cevher. Matrix recipes for hard thresholding methods. *J. Math. Imaging Vis.*, 48:235–265, 2014.

[48] C. Lancien and A. Winter. Distinguishing multi-partite states by local measurements. *Commun. in Math.l Phys.*, 323(2):555–573, 2013.

[49] J. M. Landsberg. *Tensors: geometry and applications.* Providence, RI: American Mathematical Society (AMS), 2012.

[50] G. Lecué and S. Mendelson. Sparse recovery under weak moment assumptions. *ArXiv:1401.2188*, 2014.

[51] K. Lee and Y. Bresler. ADMiRA: Atomic decomposition for minimum rank approximation. *IEEE Trans. Image Process.*, 56(9):4402 – 4416, 2010.

[52] Y.-K. Liu. Universal low-rank matrix recovery from Pauli measurements. *Adv. Neural Inf. Process. Syst.*, pages 1638–1646, 2011.

[53] R. A. Low. Large deviation bounds for k-designs. *Proc. R. Soc. Lond., Ser. A, Math. Phys. Eng. Sci.*, 465:3289–3308, 2009.

[54] R. A. Low. *Pseudo-randomness and learning in quantum computation.* PhD thesis, University of Bristol, arXiv:1006.5227, 2010.

[55] W. Matthews, S. Wehner, and A. Winter. Distinguishability of quantum states under restricted families of measurements with an application to quantum data hiding. *Commun. Math. Phys.*, 291(3):813–843, 2009.

[56] S. Mendelson. Learning without Concentration. *ArXiv:1401.0304*, jan 2014.

[57] R. Millane. Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411, 1990.

[58] Y. Nakata, M. Koashi, and M. Murao. Generating a state t-design by diagonal quantum circuits. *New J. Phys.*, 16(5):053043, 2014.

[59] G. Nebe, E. Rains, and N. Sloane. The invariants of the Clifford groups. *Des. Codes Cryptography*, 24:99–121, 2001.

[60] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

[61] M. A. Nielsen and I. L. Chuang. *Quantum computation and quantum information.* Cambridge university press, 2010.

[62] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

[63] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52:471–501, 2010.

[64] C. Schwemmer, G. Tóth, A. Niggebaum, T. Moroder, D. Gross, O. Gühne, and H. Weinfurter. Experimental comparison of efficient tomography schemes for a six-qubit state. *Phys. Rev. Lett.*, 113(4):040503, 2014.

[65] A. Scott. Tight informationally complete quantum measurements. *J. Phys. A-Math. Gen.*, 39:13507–13530, 2006.

[66] P. Seymour and T. Zaslavsky. Averaging sets: A generalization of mean values and spherical designs. *Adv. Math.*, 52:213–240, 1984.

[67] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase Retrieval with Application to Optical Imaging. *Preprint*, feb 2014. arXiv:1402.7350.

[68] V. Sidelnikov. Spherical 7-designs in 2^n-dimensional Euclidean space. *J. Algebr. Comb.*, 10:279–288, 1999.

[69] R. Stanley. *Enumerative Combinatorics, Volume I.* Cambridge University Press, 1997.

[70] J. Tanner and K. Wei. Normalized iterative hard thresholding for matrix completion. *SIAM J. Sci. Comput.*, 59(11):7491–7508, 2013.

[71] K. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pac. J. Optim.*, 6:615–640, 2010.

[72] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[73] J. A. Tropp. User friendly tools for random matrices. An introduction. *Preprint*, 2012.

[74] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. *ArXiv:1405.1102*, 2014.

[75] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge Univ Press, 2012.

[76] A. Walther. The question of phase retrieval in optics. *Journal of Modern Optics*, 10(1):41–49, 1963.

[77] J. Watrous. Theory of quantum information. lecture notes, 2011.

[78] G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.*, 170:33–45, 1992.

[79] G. Zauner. *Quantendesigns: Grundzüge einer nichtkommutativen Designtheorie.* PhD thesis, University of Vienna, 1999.

Institute for Physics, University of Freiburg, Rheinstrasse 10, 79104 Freiburg, Germany
E-mail address: `richard.kueng@physik.uni-freiburg.de`

Lehrstuhl C für Mathematik (Analysis), RWTH Aachen University, Pontdriesch 10, 52062 Aachen, Germany
E-mail address: `rauhut@mathc.rwth-aachen.de`

Lehrstuhl C für Mathematik (Analysis), RWTH Aachen University, Pontdriesch 10, 52062 Aachen, Germany
E-mail address: `terstiege@mathc.rwth-aachen.de`

STABLE LOW-RANK MATRIX RECOVERY VIA NULL SPACE PROPERTIES

MARYIA KABANAVA[1], RICHARD KUENG[2,3,4], HOLGER RAUHUT[1], ULRICH TERSTIEGE[1]

[1] Lehrstuhl C für Mathematik (Analysis), RWTH Aachen University, Germany
[2] School of Physics, The University of Sydney, Australia
[3] Institute for Physics & FDM, University of Freiburg, Germany
[4] Institute for Theoretical Physics, University of Cologne, Germany

Abstract. The problem of recovering a matrix of low rank from an incomplete and possibly noisy set of linear measurements arises in a number of areas such as quantum state tomography, machine learning and the PhaseLift approach to phaseless reconstruction problems. In order to derive rigorous recovery results, the measurement map is usually modeled probabilistically and convex optimization approaches including nuclear norm minimization are often used as recovery method. In this article, we derive sufficient conditions on the minimal amount of measurements that ensure recovery via convex optimization. We establish our results via certain properties of the null space of the measurement map. In the setting where the measurements are realized as Frobenius inner products with independent standard Gaussian random matrices we show that $m > 10r(n_1 + n_2)$ measurements are enough to uniformly and stably recover an $n_1 \times n_2$ matrix of rank at most r. Stability is meant both with respect to passing from exactly rank-r matrices to approximately rank-r matrices and with respect to adding noise on the measurements. We then significantly generalize this result by only requiring independent mean-zero, variance one entries with four finite moments at the cost of replacing 10 by some universal constant. We also study the particular case of recovering Hermitian rank-r matrices from measurement matrices proportional to rank-one projectors. For $r = 1$, such a problem reduces to the PhaseLift approach to phaseless recovery, while the case of higher rank is relevant for quantum state tomography. For $m \geq Crn$ rank-one projective measurements onto independent standard Gaussian vectors, we show that nuclear norm minimization uniformly and stably reconstructs Hermitian rank-r matrices with high probability. Subsequently, we partially de-randomize this result by establishing an analogous statement for projectors onto independent elements of a complex projective 4-designs at the cost of a slightly higher sampling rate $m \geq Crn \log n$. Complex projective t-designs are discrete sets of vectors whose uniform distribution reproduces the first t moments of the uniform distribution on the sphere. Moreover, if the Hermitian matrix to be recovered is known to be positive semidefinite, then we show that the nuclear norm minimization approach may be replaced by the simpler optimization program of minimizing the ℓ_2-norm of the residual subject to the positive semidefinite constraint. This has the additional advantage that no estimate of the noise level is required a priori. We discuss applications of such a result in quantum physics and the phase retrieval problem. Apart from the case of independent Gaussian measurements, the analysis exploits Mendelson's small ball method.

Keywords. low rank matrix recovery, quantum state tomography, phase retrieval, convex optimization, nuclear norm minimization, positive semidefinite least squares problem, complex projective designs, random measurements

MSC 2010. 94A20, 94A12, 60B20, 90C25, 81P50

1. Introduction

In recent years, the recovery of objects (signals, images, matrices, quantum states etc.) from incomplete linear measurements has gained significant interest. While standard compressive sensing considers the reconstruction of (approximately) sparse vectors [26], we study extensions to the recovery of (approximately) low rank matrices from a small number of random measurements. This problem arises in a number of areas such as quantum tomography [30, 24, 6], signal processing [2], recommender systems [16, 11] and phaseless recovery [12, 10, 28, 29]. On the one hand, we consider both random measurement maps generated by independent random matrices with independent entries and on the other hand, measurements with respect to independent rank one measurements. We derive bounds for the number of required measurements in

Date: July 28, 2015.

terms of the matrix dimensions and the rank of the matrix that guarantee successful recovery via nuclear norm minimization. Our results are uniform and stable with respect to noise on the measurements and with respect to passing to approximately rank-r matrices. For rank-one measurements the latter stability result is new.

Let us formally describe our setup. We consider measurements of an (approximately) low-rank matrix $X \in \mathbb{C}^{n_1 \times n_2}$ of the form $b = \mathcal{A}(X)$, where the linear measurement map $\mathcal{A}$ is given as

$$\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m, \quad Z \mapsto \sum_{j=1}^{m} \operatorname{tr}(Z A_j^*) e_j. \tag{1}$$

Here, $e_1, \ldots, e_m$ denote the standard basis vectors in $\mathbb{C}^m$ and $A_1, \ldots, A_m \in \mathbb{C}^{n_1 \times n_2}$ are called measurement matrices. A prominent approach [22, 56] for recovering the matrix X from $b = \mathcal{A}(X)$ consists in computing the minimizer of the convex optimization problem

$$\min_{Z \in \mathbb{C}^{n_1 \times n_2}} \|Z\|_* \quad \text{subject to } \mathcal{A}(Z) = b, \tag{2}$$

where $\|Z\|_* = \|Z\|_1 = \sum_{j=1}^{n} \sigma_j(Z)$ denotes the nuclear norm with $\sigma_j(Z)$ being the singular values of $Z \in \mathbb{C}^{n_1 \times n_2}$ and $n = \min\{n_1, n_2\}$. Efficient optimization methods exist for this problem [55, 8]. In practice the measurements are often perturbed by noise, i.e.,

$$b = \mathcal{A}(X) + w, \tag{3}$$

where $w \in \mathbb{C}^m$ is a vector of perturbations. In this case, we replace (2) by the noise constrained nuclear norm minimization problem

$$\min_{Z \in \mathbb{C}^{n_1 \times n_2}} \|Z\|_* \quad \text{subject to } \|\mathcal{A}(Z) - b\|_{\ell_2} \leq \eta, \tag{4}$$

where η corresponds to a known estimate of the noise level, i.e., $\|w\|_{\ell_2} \leq \eta$ with $\|x\|_{\ell_p} = (\sum_j |x_j|^p)^{1/p}$ being the usual ℓ_p-norm. In some cases it is known a priori that the matrix X of interest is both Hermitian and positive semidefinite ($X \succcurlyeq 0$). Then one may replace (4) by the optimization problem

$$\min_{Z \succcurlyeq 0} \operatorname{tr}(Z) \quad \text{subject to } \|\mathcal{A}(Z) - b\|_{\ell_2} \leq \eta. \tag{5}$$

However, as we will see, the simpler least squares problem

$$\min_{Z \succcurlyeq 0} \|\mathcal{A}(Z) - b\|_{\ell_2} \tag{6}$$

works equally well or even better in terms of recovery under certain natural conditions. Apart from simplicity and computational efficiency it has the additional advantage that no estimate η of the noise level is required. We note that other efficient recovery methods exist as well [46, 25, 64], but we will not go into details here.

A question of central interest concerns the minimal number m of required measurements that guarantees exact (in the noiseless case) or approximate recovery. While it is very hard to study this question for deterministic measurement maps $\mathcal{A}$, several results are available for certain models of random maps. We will study several scenarios which all have in common that the matrices $A_1, \ldots, A_m \in \mathbb{R}^{n_1 \times n_2}$ in (1) are independent draws of a random matrix $\Phi = (X_{ij})_{ij}$. We first consider the real-valued case, where all entries X_{ij} are independent and then move to a complex-valued scenario where $\Phi = aa^* \in \mathbb{C}^{n \times n}$ is a rank one matrix generated by a random vector $a \in \mathbb{C}^n$. For the latter scenario we consider a being a complex Gaussian random vector, or a being randomly drawn from a so-called (approximate) t-design. This last setup has implications for quantum tomography and this part of the article can be seen as a continuation of the investigations in [43]. Next, we describe the present state of the art of of the various setups and present our results.

1.1. **Robust recovery from measurement matrices with independent entries.** We call $\mathcal{A}$ a Gaussian measurement map if the matrices $A_1, \ldots, A_m \in \mathbb{R}^{n_1 \times n_2}$ in (1) are independent realizations of Gaussian random matrices, i.e., all entries of the A_j are independent standard Gaussian random variables. More generally, $\mathcal{A}$ is called subgaussian, if the entries of all the A_j are independent, mean zero, variance one, subgaussian random variables, where we recall that a random variable ξ is called subgaussian if $\mathbb{P}(|\xi| \geq t) \leq 2e^{-ct^2}$ for some constant $c > 0$. If

$$m \geq Cr(n_1 + n_2) \tag{7}$$

for some universal constant $C > 0$, then with probability at least $1 - e^{-cm}$ any rank r matrix $X \in \mathbb{C}^{n_1 \times n_2}$ is reconstructed exactly from subgaussian measurements $b = \mathcal{A}(X)$ via nuclear norm minimization (2) [56, 15]. Moreover, if noisy measurements $b = \mathcal{A}(X) + w$ with $\|w\|_2 \leq \eta$ of an arbitrary matrix $X \in \mathbb{C}^{n_1 \times n_2}$ are taken, then the minimizer $X^\sharp$ of (4) satisfies, again with probability at least $1 - e^{-cm}$,

$$\|X - X^\sharp\|_F \leq \frac{C'}{\sqrt{r}} \inf_{Z:\text{rank}(Z) \leq r} \|X - Z\|_* + \frac{C''\eta}{\sqrt{m}}, \tag{8}$$

where $\|A\|_F = \sqrt{\text{tr}(A^*A)}$ denotes the Frobenius norm, tr being the trace. Note that

$$\inf_{Z:\text{rank}(Z) \leq r} \|X - Z\|_* = \sum_{j=r+1}^{n} \sigma_j(X) = \|X_c\|_*,$$

where the singular values $\sigma_j(X)$ are arranged in decreasing order and for X with singular value decomposition $\sum_{j=1}^n \sigma_j(X) u_j v_j^*$ the matrix $X_c = \sum_{j=r+1}^n \sigma_j(X) u_j v_j^*$. The error estimate (8) means that reconstruction is robust with respect to noise on the measurements and stable with respect to passing to only approximately low rank matrices. These statements are uniform in the sense that they hold for all matrices X simultaneously once the matrix A has been drawn. They have been established in [15, 52, 56] via the rank restricted isometry property (rank-RIP), see e.g. [26] for the standard RIP and its implications.

While the RIP is a standard tool by now, recovery of low rank matrices via nuclear norm minimization is characterized by the so-called null space property [51, 58, 57, 26, 25], see below for details. By using this concept, we are able to significantly relax from subgaussian distributions of the entries to distributions with only four finite moments.

Theorem 1. *Let $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$, $\mathcal{A}(X) = \sum_{j=1}^n \text{tr}(XA_j)e_j$, where the A_j are independent copies of a random matrix $\Phi = (X_{ij})_{i,j}$ with independent mean zero entries obeying $\mathbb{E}X_{ij}^2 = 1$ and*

$$\mathbb{E}X_{ij}^4 \leq C_4 \quad \text{for all } i,j \text{ and some constant } C_4.$$

Fix $1 \leq r \leq \min\{n_1, n_2\}$ and $0 < \rho < 1$ and set

$$m \geq c_1 \rho^{-2} r(n_1 + n_2).$$

Then with probability at least $1 - e^{-c_2 m}$, for any $X \in \mathbb{R}^{n_1 \times n_2}$ the solution $X^\sharp$ of (4) with $b = \mathcal{A}(X) + w$, $\|w\|_{\ell_2} \leq \eta$, approximates X with error

$$\|X - X^\sharp\|_F \leq \frac{2(1+\rho)^2}{(1-\rho)\sqrt{r}} \|X_c\|_* + \frac{(3+\rho)}{(1-\rho)c_3} \cdot \frac{\eta}{\sqrt{m}}. \tag{9}$$

Here c_1, c_2, c_3 are positive constants that only depend on C_4.

In the special case, when Φ has independent standard Gaussian entries, we apply Gordon's escape through a mesh theorem [27] in order to obtain an explicit constant in the estimate for the number of measurements, see Theorem 19. Roughly speaking, with high probability, any $n_1 \times n_2$ matrix of rank r is stably recovered from $m > 10r(n_1 + n_2)$ Gaussian measurements. We remark that the explicit bound $m > 3r(n_1 + n_2)$ has been derived in [18], (see also [49] and [4, Section 4.4] for a phase transition result in this context), but this bound considers nonuniform recovery, i.e. recovery of a fixed low rank matrix with a random draw of a Gaussian measurement matrix with high probability. Moreover, no stability under passing to approximately low rank matrices has been considered there. Our recovery result is therefore stronger than the one in [18], but requires more measurements.

1.2. **Robust recovery of Hermitian matrices from rank-one projective measurements.** Let us now focus on the particular case of recovering complex Hermitian $n \times n$ matrices from noisy measurements of the form (3), where the measurement matrices are proportional to rank-one projectors, i.e.,

$$A_j = a_j a_j^* \in \mathcal{H}_n \tag{10}$$

where $a_j \in \mathbb{C}^n$. Here, $\mathcal{H}_n$ denotes the space of complex Hermitian $n \times n$ matrices, which has real dimension n^2. Measurements of that type occur naturally in convex relaxations of the phase retrieval problem [12, 10, 28, 29]. In fact, suppose phaseless measurements of the form $b_j = |\langle x, a_j \rangle|^2$ of a vector $x \in \mathbb{C}^n$ are given. Then we can rewrite $b_j = \operatorname{tr}(xx^* a_j a_j^*) = \operatorname{tr}(XA_j)$ as linear measurements of the rank one matrix $X = xx^*$. We will expand on this aspect below in Section 2.1. Rank one measurements of low rank matrices feature prominently in quantum state tomography as well, see also below.

The prior information that the desired matrix is Hermitian limits the search space in the convex optimization problem (4) and it simplifies to

$$\min_{Z \in \mathcal{H}_n} \|Z\|_* \quad \text{subject to } \|\mathcal{A}(Z) - b\|_{\ell_2} \leq \eta. \tag{11}$$

Arguably, the most generic measurement matrices of the form (10) result from choosing each a_j to be an independent complex standard Gaussian vector. For the particular case of phase retrieval — i.e., where the matrix of interest $X = xx^*$ is itself proportional to a rank-one projector — uniform recovery guarantees by means of (11) have been established for $m = Cn$ independent measurements in [13]. Recently, this result has been generalized to recovery of any Hermitian rank r-matrix by means of $m = Crn$ such measurements in [43]. Our refined analysis of the null space property enables us to further strengthen this result by additionally guaranteeing stability under passing to approximately low rank matrices:

Theorem 2. *Consider the measurement process described in (1) with m measurement matrices of the form (10),where each a_i is an independent complex standard Gaussian vector. Fix $r \leq n$, $0 < \rho < 1$ and suppose that*

$$m \geq C_1 \rho^{-2} nr.$$

Then with probability at least $1 - \mathrm{e}^{-C_2 m}$ it holds that for any $X \in \mathcal{H}_n$, any solution $X^\sharp$ to the convex optimization problem (11) with noisy measurements $b = \mathcal{A}(X) + \epsilon$, where $\|\epsilon\|_{\ell_2} \leq \eta$, obeys

$$\|X - X^\sharp\|_F \leq \frac{2(1+\rho)^2}{(1-\rho)\sqrt{r}} \|X_c\|_* + \frac{(3+\rho)C_3}{(1-\rho)} \cdot \frac{\eta}{\sqrt{m}}. \tag{12}$$

Here, C_1, C_2 and C_3 denote positive universal constants. (In particular, for $\eta = 0$ and X of rank at most r one has exact reconstruction.)

In addition to the Gaussian measurement setting, we also consider measurement matrices that arise from taking the outer product of elements chosen independently from an approximate complex projective 4-design. Complex projective t-designs are finite sets of unit vectors in $\mathbb{C}^n$ that exhibit a very particular structure. Roughly speaking, sampling independently from a complex projective t-design, reproduces the first t moments of sampling uniformly from the complex unit sphere. Likewise, approximate complex projective t-designs obey such a structural requirement approximately — for a precise introduction, we refer to Definition 27 below. As a consequence, they serve as a general purpose tool for partially de-randomizing results that initially required Gaussian random vectors [42, 28]. This is also the case here and employing complex projective 4-designs allows for partially de-randomizing Theorem 2 at the cost of a slightly larger sampling rate. Here, we content ourselves with presenting and shortened version of this result and refer the reader to Theorem 28 where precise requirements on the approximate design are stated.

Theorem 3. *Let r, ρ be as in Theorem 2 and suppose that each measurement matrix A_j is of the form (10), where a_j, $j = 1, \ldots, m$, are chosen independently from a (sufficiently accurate approximate) complex projective 4-design. If*

$$m \geq C_4 \rho^{-2} nr \log n,$$

then the assertions of Theorem 2 remains valid, possibly with different universal constants.

Note that Theorems 1, 2, 3 resp. Theorem 19 below and their proofs are presented in condensed versions in the conference papers [34] resp. [35].

1.3. **Recovery of positive semidefinite matrices reduces to a feasibility problem.** Imposing additional structure on the matrices to be recovered can further strengthen low rank recovery guarantees. Positive semidefiniteness is one such structural prerequisite that, for instance, occurs naturally in the phase retrieval problem, quantum mechanics and kernel-based learning methods [61]. Motivated by the former, Demanet and Hand [21] pointed out that minimizing the nuclear norm — in the sense of algorithm (4) — can be superfluous for recovering positive semidefinite matrices of rank one. Instead, they propose to reduce the recovery algorithm to a mere feasibility problem and proved that such a reduction works w.h.p. for rank one projective measurements onto Gaussian vectors (the measurement scenario considered in Theorem 2). Subsequently, this recovery guarantee was strengthened by Candès and Li [13]. Here, we go one step further and generalize these results to cover uniform and stable recovery of positive semidefinite matrices of arbitrary rank. Relying on ideas presented in [36], we establish the following statement. (We refer to Section 1.4 for the definition of the Schatten p-norm $\|\cdot\|_p$ used in (13).)

Theorem 4. *Fix $r \leq n$ and consider the measurement processes introduced in Theorem 2 (Gaussian vectors), or Theorem 3 (complex projective 4-designs), respectively. Assume that $m \geq C_1 nr$ (in the Gaussian case) resp. $m \geq C_2 snr \log n$ (in the design case), where $s \geq 1$ is arbitrary. Then, for $1 \leq p \leq 2$ and any two positive semidefinite matrices $X, Z \in \mathcal{H}_n$,*

$$\|Z - X\|_p \leq \frac{C_3}{r^{1-1/p}} \|X_c\|_1 + \frac{C_4 r^{1/p-1/2}}{\sqrt{m}} \|\mathcal{A}(Z) - \mathcal{A}(X)\|_{\ell_2} \tag{13}$$

holds universally with probability exceeding $1 - \mathrm{e}^{-C_5 m}$ for the Gaussian case and $1 - \mathrm{e}^{-sr}$ in the design case. Here, $C_1, \ldots, C_5$ denote suitable positive universal constants.

This statement renders nuclear norm minimization in the sense of (4) redundant and allows for a regularization-free estimation. Moreover, knowledge of a noise bound $\|w\|_{\ell_2} \leq \eta$ for the measurement process (3) is no longer required, since we can estimate any $X \succcurlyeq 0$ by solving a least squares problem of the form (6), i.e.,

$$\min_{Z \in \mathcal{H}_n} \|\mathcal{A}(Z) - b\|_{\ell_2} \quad \text{subject to} \quad Z \succcurlyeq 0. \tag{14}$$

Theorem 4 then in particular assures that the minimizer $Z^\sharp$ of this optimization program obeys

$$\|Z^\sharp - X\|_F \leq \frac{C_3}{\sqrt{r}} \|X_c\|_1 + \frac{C_4}{\sqrt{m}} \|\mathcal{A}(Z^\sharp) - \mathcal{A}(X)\|_{\ell_2} \leq \frac{C_3}{\sqrt{r}} \|X_c\|_1 + \frac{2C_4}{\sqrt{m}} \|w\|_{\ell_2},$$

where $w \in \mathbb{R}^m$ represents additive noise in the measurement process. It is worthwhile to mention that if a matrix X of interest has rank at most r and no noise is present in the sampling process (3), Theorem 4 assures

$$\{Z : \ Z \succcurlyeq 0, \ \mathcal{A}(Z) = \mathcal{A}(X)\} = \{X\} \tag{15}$$

with high probability. Hence, recovering X from noiseless measurements indeed reduces to a feasibility problem.

We emphasize that Theorem 4 is only established for rank one projective measurements. For the other measurement ensembles considered here — matrices with independent entries — one cannot expect such a statement to hold. This pessimistic prediction is due to negative results recently established in [63, Proposition 2]. Focusing on real matrices, the authors show that if the measurement matrices A_j are chosen independently from a Gaussian orthogonal ensemble, then estimating any symmetric, positive semidefinite matrix X via (14) becomes ill-posed, unless the number of measurements obeys

$$m \geq \frac{1}{4} n(n+1) = \mathcal{O}(n^2).$$

Finally, we want to point out that the fruitfulness of plain least squares regression for recovering positive semidefinite matrices was already pointed out and explored by Slawski, Li and Hein [63]. However, there is a crucial difference in the mindset of [63] and the results presented here. The main result [63, Theorem 2] of Slawski et al. assumes a fixed signal $X \succcurlyeq 0$ of interest and provides bounds for the reconstruction error in terms of geometric properties of both X and the measurement ensemble. Conversely, Theorem 4 assumes fixed measurements (e.g. $m = Crn$ projectors onto Gaussian random vectors) and w.h.p. assures robust recovery of all matrices $X \succcurlyeq 0$ having approximately rank-r simultaneously.

1.4. **Notation.** The Schatten p-norm of $Z \in \mathbb{C}^{n_1 \times n_2}$ is given by

$$\|Z\|_p = \left(\sum_{j=1}^{n} \sigma_j(Z)^p \right)^{1/p}, \quad p \geq 1,$$

where $\sigma_j(Z)$, $j = 1, \ldots, n$, denote the singular values of Z. It reduces to the nuclear norm $\| \cdot \|_*$ for $p = 1$ and the Frobenius norm $\| \cdot \|_F$ for $p = 2$. It is a common convention that the singular values of Z are non-increasingly ordered. We write $Z = Z_r + Z_c$, where Z_r is the best rank-r approximation of Z with respect to any Schatten p-norm of Z.

2. Applications

2.1. **Phase retrieval.** The problem of retrieving a complex signal $x \in \mathbb{C}^n$ from measurements that are ignorant towards phase information has long been abundant in many areas of science. Measurements of that type correspond to

$$b_i = |\langle a_i, x \rangle|^2 + w_i \quad i = 1, \ldots, m, \tag{16}$$

where $a_1, \ldots, a_m \in \mathbb{C}^n$ are measurement vectors and w_i denotes additive noise. Recently, the problem's mathematical structure has received considerable attention in its own right. It is clearly ill-posed, since all phase information is lost in the measurement process and, moreover, the measurements (16) are of a non-linear nature. This second obstacle can be overcome by a trick [5] well known in conic programming: the quadratic expressions (16) are linear in the outer products xx^* and $a_i a_i^*$:

$$b_i = |\langle a_i, x \rangle|^2 + w_i = \operatorname{tr}\left((a_i a_i)^* (xx^*) \right) + w_i. \tag{17}$$

Note that such a "lift" allows for reinterpreting the phase-less sampling process as $\mathcal{A}(xx^*) = b + w$. Also, the new object of interest $X := xx^*$ is an Hermitian, positive semidefinite matrix of rank one. In turn, the measurement matrices $A_i = a_i a_i^*$ are constrained to be proportional to rank-one projectors. Consequently, such a "lift" turns the phase retrieval problem into a very particular instance of low rank matrix recovery — a fact that was first observed by Candès, Eldar, Strohmer and Voroninski [12, 10]. Subsequently, uniform recovery guarantees for $m = Cn$ complex standard Gaussian measurement vectors a_i have been established which are stable towards additive noise. The main result in [13] establishes with high probability that for any $X = xx^*$, solving the convex optimization problem (PhaseLift)

$$\min_{Z \in \mathcal{H}_n} \|\mathcal{A}(Z) - b\|_{\ell_1} \quad \text{subject to} \quad Z \succcurlyeq 0 \tag{18}$$

yields an estimator $Z^\sharp$ obeying $\|Z^\sharp - xx^*\|_2 \leq C\|w\|_1/m$. If a bound $\|w\|_{\ell_2} \leq \eta$ on the noise in the sampling process (16) is available, an extension of [43, Theorem 2] (see section 2.3.2 in loc. cit) establishes a comparable recovery guarantee via solving

$$\min_{Z \in \mathcal{H}_n} \operatorname{tr}(Z) \quad \text{subject to} \quad \|\mathcal{A}(Z) - b\|_{\ell_2} \leq \eta, \ Z \succcurlyeq 0 \tag{19}$$

instead of PhaseLift. Our findings allow for establishing novel recovery guarantees for retrieving phases. Indeed, since (17) assures that any signal of interest is positive semidefinite and has precisely rank one, Theorem 4 is applicable and yields the following corollary.

Corollary 5. *Consider $m \geq Cn$ phaseless measurements of the form (16), where each a_i is a complex standard Gaussian vector. Then with probability at least $1 - \mathrm{e}^{-C'm}$ these measurements allow for estimating any signal $x \in \mathbb{C}^n$ via solving*

$$\min_{Z \in \mathcal{H}_n} \|\mathcal{A}(Z) - b\|_{\ell_2} \quad \textit{subject to} \quad Z \succcurlyeq 0. \tag{20}$$

The resulting minimizer $Z^\sharp$ of (20) obeys

$$\|Z^\sharp - xx^*\|_{\ell_2} \leq \frac{C\|w\|_{\ell_2}}{\sqrt{m}},$$

where C denotes a positive constant and $w \in \mathbb{R}^m$ represents additive noise in the sampling process (16).

An analogous statement is true — with a weaker probability of success $1 - \mathrm{e}^{-s}$ for $s \geq 1$ — for $m \geq C' sn \log(n)$ rank one projective measurements onto independent elements of an approximate 4-design.

This recovery procedure is in spirit very similar to (18), but it utilizes an ℓ_2-regression instead of an ℓ_1-norm minimization. Numerical studies indicate that algorithm (20) outperforms (19) as well as (18). These studies were motivated and accompany actual quantum mechanical experiments and will be published elsewhere [41].

Finally, we want to relate Corollary 5 to a non-convex phaseless recovery procedure devised by Candès, Li and Soltanolkotabi [14]. There, the authors refrain from applying the aforementioned “lifting” trick to render the phase retrieval problem linear. Instead, they use a careful initialization step, followed by a gradient descent scheme (based on Wirtinger derivatives) to minimize the problem’s least squares loss function directly over complex vectors $z \in \mathbb{C}^n$. Mathematically, such an optimization is equivalent to solving

$$\min_{Z \in \mathcal{H}_n} \|\mathcal{A}(Z) - b\|_{\ell_2} \quad \text{subject to} \quad Z \succcurlyeq 0, \ \operatorname{rank}(Z) = 1 \tag{21}$$

and the rank-constraint manifests the problem’s non-convex nature. Hence, the convex optimization problem (20) can be viewed as a convex relaxation of (21), obtained by omitting the non-convex rank constraint.

2.2. **Quantum information.** In this section we describe implications and possible applications of our findings to problems in quantum information science. For the sake of being self-contained, we have included a brief introduction to crucial notions of quantum mechanics in the appendix. Quantum mechanics postulates that a finite n-dimensional quantum system is described by an Hermitian, positive semidefinite matrix X with unit trace, called a *density operator*. This “quantum shape constraint” assures that all density operators meet the requirements of Theorem 4. Furthermore, the rank-one projective measurements assumed in that theorem can be recast as valid quantum mechanical measurements — see [43, Section 3] for possible implementations and further discussion on this topic. Note, however, that such a reinterpretation is in general not possible for the measurement matrices with independent entries considered in Theorem 1, because these matrices fail to be Hermitian. With Theorem 4 at hand, we underline its implications for two prominent issues in (finite dimensional) quantum mechanics.

2.2.1. *Quantum state tomography.* Inferring a quantum mechanical description of a physical system is equivalent to assigning it a *density operator* (or quantum state) — a process referred to as *quantum state tomography* [6, 23]. Tomography is now a routine task for designing, testing and tuning qubits in the quest of building quantum information processing devices. Since the size of controllable quantum mechanical systems is ever increasing[1] it is very desirable to exploit additional structure — if present — when performing such a task. One such structural property — often encountered in actual experiments — is *approximate purity*, i.e., the density operator X is well approximated by a low rank matrix. Performing quantum state tomography under such a prior assumption therefore constitutes a particular instance of low rank matrix recovery [30, 24].

The results presented in this paper provide recovery guarantees for tomography protocols that stably tolerate noisy measurements and moreover are *robust* towards the prior assumption of approximate purity. In the context of tomography, results of this type so far have already been established for $m = Cnr \log^6 n$ random (generalized) Pauli measurements [47, Proposition 2.3] via proving a rank-RIP for such measurement matrices and then resorting to [15, Lemma 3.2]. However, this auxiliary result manifestly requires additive Gaussian noise and using a type of Dantzig, or Lasso selector to recover the best rank-r approximation of a given density operator. This is not the case for the result established here, where performing a plain least squares regression of the form (14) is sufficient.

Corollary 6. *Fix $r \leq n$ and suppose that the measurement operator $\mathcal{A} : \mathcal{H}_n \to \mathbb{R}^m$ is of the form*

$$\mathcal{A}(X) = \sum_{i=1}^m \sqrt{\frac{(n+1)n}{m}} \langle a_i, X a_i \rangle e_i + w \in \mathbb{R}^m \quad \text{with} \quad m \geq C_1 rn \log n,$$

[1] Nowadays, experimentalists are able to create and control multi-partite systems of overall dimension $n = 2^8$ in their laboratories [60]. This results in a density operator of size 256 × 256 (a priori 65 536 parameters).

where each $a_i \in \mathbb{C}^n$ is chosen independently from an approximate 4-design and $w \in \mathbb{R}^m$ denotes additive noise. Then, the best rank-r approximation of any density operator X can be obtained from such measurements via solving

$$\min_{Z \in \mathcal{H}_n} \|\mathcal{A}(Z) - \mathcal{A}(X)\|_{\ell_2} \quad \textit{subject to} \quad Z \succcurlyeq 0,\ \mathrm{tr}\,(Z) = 1. \tag{22}$$

With probability at least $1 - \mathrm{e}^{-C_2 m}$, the minimizer $Z^\sharp$ of this optimization obeys

$$\|X - Z^\sharp\|_1 \leq C_3 \|X_c\|_1 + C_4 \sqrt{r} \|w\|_{\ell_2}, \tag{23}$$

where C_1, C_2, C_3 and C_4 denote positive constants.

This statement is a direct consequence of Theorem 4. For the sake of clarity, we have re-scaled each projective measurement with $\sqrt{\frac{(n+1)n}{m}}$. This simplifies the resulting expression (23) and moreover facilitates[2] direct comparison with the main result in [47], as it closely mimics the scaling employed there.

Corollary 6 is valid for any type of additive noise and no a priori knowledge of its magnitude is required. This includes the particularly relevant case of a Bernoulli error model — see e.g. [17, Section 2.2.2] and also [24] — which is particularly relevant for tomography experiments. Also, note that the recovery error is bounded in nuclear norm, instead of Frobenius norm. Such a bound is very meaningful for tomography, since quantum mechanics is a probabilistic theory and the nuclear norm encapsulates total variational distance. Moreover, Helstrom's theorem [32] provides an operational interpretation of the nuclear norm distance bounded in (23): it is proportional to the maximal bias achievable in the task of distinguishing the two quantum states X and $Z^\sharp$, provided that any physical measurement can be implemented.

Finally, note that the bound on the probability of failure in Corollary 6 is much stronger than the one provided in Theorem 4. Such a strengthening is possible, because the trace of any density operator equals one. We comment on this in Remark 34 below.

2.2.2. *Distinguishing quantum states.* One crucial prerequisite in the task of inferring density operators from measurement data, is the ability to faithfully distinguish any two density operators via quantum mechanical measurements. The most general notion of a quantum measurement is a *positive operator valued measure* (POVM) $\mathcal{M} = \{E_m : E_m \succcurlyeq 0, \sum_m E_m = \mathrm{id}\}$ [53, Chapter 2.2]. A POVM $\mathcal{M}$ is called *informationally complete* (IC) [62] if for any two density operators $X \neq Z \in \mathcal{H}_n$ there exists $E_m \in \mathcal{M} \subseteq \mathcal{H}_n$ such that

$$\mathrm{tr}\,(E_m X) \neq \mathrm{tr}\,(E_m Z)\,. \tag{24}$$

This assures the possibility of discriminating any two quantum states via such a measurement in the absence of noise. Without additional restrictions, such an IC POVM must contain at least n^2 elements. However, such a lower bound can be too pessimistic, if the density operators of interest have additional structure. Approximate purity introduced in the previous subsection can serve as such an additional structural restriction:

Definition 7 (Rank-r IC, Definition 1 in [31])**.** For $r \leq n$, we call a POVM $\mathcal{M} = \{E_m\}_{m \in I}$ *rank-r restricted informationally complete* (rank-r IC), if (24) holds for any two density operators of rank at most r.

Bounds for the number m of POVM elements required to assure rank-r-IC have been established in [31, 37, 38]. These approaches exploit topological obstructions of embeddings for establishing lower bounds and explicit POVM constructions for upper bounds. For instance, in [31] a particular rank-r-IC POVM containing $m = 4r(n-r) - 1$ elements is constructed.

Focusing less on establishing tight bounds and more on identifying entire families of rank-r IC measurements, Kalev et al. [36] observed that each measurement ensemble fulfilling the rank-RIP for some $r \leq n$ is also rank-r IC. This in particular applies with high probability to $m = C \log^6 n\, nr$ random (generalized) Pauli measurements [47]. Theorem 4, and likewise Corollary 6, allow us to draw similar conclusions without having to rely on any rank-RIP. Indeed, in the absence of noise, these results guarantee for any rank-r density operator X

$$\{Z : \ Z \succcurlyeq 0,\ \mathcal{A}(Z) = \mathcal{A}(X)\} = \{X\} \tag{25}$$

[2] In fact by resorting to the Frobenius norm bound in Theorem 4 (instead of the nuclear norm bound employed to arrive at Corollary 6), one obtains a performance guarantee that strongly resembles [47, Equation (8)] — the main recovery guarantee in that paper.

with high probability. If this is the case, the measurement operator $\mathcal{A}$ allows for uniquely identifying any rank-r density operator X. This in turn implies that $\mathcal{A}$ is rank-r IC and the following corollary is immediate:

Corollary 8. *Fix $r \leq n$ arbitrary and let C, C' be absolute constants of sufficient size. Then*

(1) Any POVM containing $m = Cnr$ projectors onto Haar[3] random vectors is rank-r IC with probability at least $1 - \mathrm{e}^{C_2 m}$.

(2) Any POVM containing $m = C'nr \log n$ projectors onto random elements of a (sufficiently accurate approximate) 4-design is rank-r IC with probability at least $1 - \mathrm{e}^{-\tilde{C}_2 m}$.

This statement is reminiscent of a conclusion drawn in [3, 48]: In the task of distinguishing quantum states, a POVM containing a 4-design essentially performs as good as as the uniform POVM (the union of all rank-one projectors).

Remark 9. In the process of finishing this article we became aware of recent work by Kech and Wolf [39], who showed that the elements of a generic Parseval frame generate a rank-r IC map $\mathcal{A}$ if $m \geq 4r(n-r)$. In fact, Xu showed in [68] that $m \geq 4r(n-r)$ is both a sufficient and necessary condition for identifiability of complex rank r matrices in $\mathbb{C}^{n \times n}$. We emphasize, however, that these results are only concerned with pure identifiability and do not come with a practical and stable recovery algorithm.

3. The null space property for low-rank matrix recovery

Let $X \in \mathbb{C}^{n_1 \times n_2}$. If X is only approximately of low-rank, then we would like to find a condition on the measurement map $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ that provides the control of the recovery error by the error of its best approximation by low rank matrices. Moreover, it should also take into account that the measurements might be noisy.

Definition 10. We say that $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ satisfies the Frobenius robust rank null space property of order r with constants $0 < \rho < 1$ and $\tau > 0$ if for all $M \in \mathbb{C}^{n_1 \times n_2}$, the singular values of M satisfy

$$\|M_r\|_2 \leq \frac{\rho}{\sqrt{r}} \|M_c\|_1 + \tau \|\mathcal{A}(M)\|_{\ell_2}.$$

The stability and robustness of (4) are established by the following theorem.

Theorem 11. *Let $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ satisfy the Frobenius robust rank null space property of order r with constants $0 < \rho < 1$ and $\tau > 0$. Let $n = \min\{n_1, n_2\}$. Then for any $X \in \mathbb{C}^{n_1 \times n_2}$ any solution $X^\sharp$ of (4) with $b = \mathcal{A}(X) + w$, $\|w\|_{\ell_2} \leq \eta$, approximates X with error*

$$\|X - X^\sharp\|_2 \leq \frac{2(1+\rho)^2}{(1-\rho)\sqrt{r}} \|X_c\|_1 + \frac{2\tau(3+\rho)}{1-\rho} \eta.$$

Theorem 11 can be deduced from the following stronger result.

Theorem 12. *Let $1 \leq p \leq 2$ and $n = \min\{n_1, n_2\}$. Suppose that $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ satisfies the Frobenius robust rank null space property of order r with constants $0 < \rho < 1$ and $\tau > 0$. Then for any $X, Z \in \mathbb{C}^{n_1 \times n_2}$,*

$$\|Z - X\|_p \leq \frac{(1+\rho)^2}{(1-\rho) r^{1-1/p}} \left(\|Z\|_1 - \|X\|_1 + 2\|X_c\|_1\right) + \frac{\tau(3+\rho)}{1-\rho} r^{1/p-1/2} \|\mathcal{A}(Z - X)\|_{\ell_2}. \tag{26}$$

The proof requires some auxiliary lemmas. We start with a matrix version of Stechkin's bound.

Lemma 13. *Let $M \in \mathbb{C}^{n_1 \times n_2}$ and $r \leq \min\{n_1, n_2\}$. Then, for $p > 0$,*

$$\|M_c\|_p \leq \frac{\|M\|_1}{r^{1-1/p}}.$$

[3] Haar random vectors are vectors drawn uniformly from the complex unit sphere in $\mathbb{C}^n$. They can be obtained from complex standard Gaussian vectors by rescaling them to unit length. Property (25) is invariant under such a re-scaling and Theorem 2 therefore assures rank-r IC for both Gaussian and Haar random vectors.

Proof. This follows immediately from [26, Proposition 2.3], but for convenience we give the proof. Since the singular values of M are non-increasingly ordered, it holds

$$\begin{aligned}\|M_c\|_p^p &= \sum_{j=r+1}^{n} (\sigma_j(M))^p \leq (\sigma_r(M))^{p-1} \sum_{j=r+1}^{n} \sigma_j(M) \leq \left[\frac{1}{r}\sum_{j=1}^{r}\sigma_j(M)\right]^{p-1} \sum_{j=r+1}^{n} \sigma_j(M) \\ &\leq \frac{1}{r^{p-1}} \|M\|_1^{p-1} \|M\|_1 = \frac{\|M\|_1^p}{r^{p-1}}.\end{aligned}$$

□

The next result shows that under the Frobenius robust rank null space property the distance between two matrices is controlled by the difference between their norms and the ℓ_2-norm of the difference between their measurements.

Lemma 14. *Suppose that $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ satisfies the Frobenius robust rank null space property of order r with constants $0 < \rho < 1$ and $\tau > 0$. Let $X, Z \in \mathbb{C}^{n_1 \times n_2}$ and $n = \min\{n_1, n_2\}$. Then*

$$\|X - Z\|_1 \leq \frac{1+\rho}{1-\rho}\left(\|Z\|_1 - \|X\|_1 + 2\|X_c\|_1\right) + \frac{2\tau\sqrt{r}}{1-\rho}\|\mathcal{A}(X - Z)\|_{\ell_2}.$$

Proof. Theorem 7.4.9.1 in [33] states that for matrices A, B of the same size over $\mathbb{C}$

$$\|A - B\| \geq \|\Sigma(A) - \Sigma(B)\|,$$

where $\|\cdot\|$ is any unitarily invariant norm and $\Sigma(\cdot)$ denotes the diagonal matrix of singular values of its argument. Hence,

$$\begin{aligned}\|Z\|_1 = \|X - (X - Z)\|_1 &\geq \sum_{j=1}^{n} |\sigma_j(X) - \sigma_j(X - Z)| \\ &= \sum_{j=1}^{r} |\sigma_j(X) - \sigma_j(X - Z)| + \sum_{j=r+1}^{n} |\sigma_j(X) - \sigma_j(X - Z)| \\ &\geq \sum_{j=1}^{r} (\sigma_j(X) - \sigma_j(X - Z)) + \sum_{j=r+1}^{n} (\sigma_j(X - Z) - \sigma_j(X)).\end{aligned}$$

Hence,

$$\begin{aligned}\|(X - Z)_c\|_1 = \sum_{j=r+1}^{n} \sigma_j(X - Z) &\leq \|Z\|_1 - \sum_{j=1}^{r}\sigma_j(X) + \sum_{j=1}^{r}\sigma_j(X - Z) + \|X_c\|_1 \\ &\leq \|Z\|_1 - \|X\|_1 + \sqrt{r}\|(X - Z)_r\|_2 + 2\|X_c\|_1.\end{aligned}$$

Applying the Frobenius robust null space property of $\mathcal{A}$ we obtain

$$\|(X - Z)_c\|_1 \leq \|Z\|_1 - \|X\|_1 + \rho\|(X - Z)_c\|_1 + \tau\sqrt{r}\|\mathcal{A}(X - Z)\|_{\ell_2} + 2\|X_c\|_1.$$

By rearranging the terms in the above inequality we obtain

$$\|(X - Z)_c\|_1 \leq \frac{1}{1-\rho}\left(\|Z\|_1 - \|X\|_1 + \tau\sqrt{r}\|\mathcal{A}(X - Z)\|_{\ell_2} + 2\|X_c\|_1\right).$$

In order to bound $\|X - Z\|_1$ we use Hölder's inequality, the Frobenius robust rank null space property of $\mathcal{A}$ and the inequality above,

$$\begin{aligned}\|X - Z\|_1 &= \|(X - Z)_r\|_1 + \|(X - Z)_c\|_1 \leq \sqrt{r}\|(X - Z)_r\|_2 + \|(X - Z)_c\|_1 \\ &\leq (1 + \rho)\|(X - Z)_c\|_1 + \tau\sqrt{r}\|\mathcal{A}(Z - X)\|_{\ell_2} \\ &\leq \frac{1+\rho}{1-\rho}\left(\|Z\|_1 - \|X\|_1 + \tau\sqrt{r}\|\mathcal{A}(X - Z)\|_{\ell_2} + 2\|X_c\|_1\right) + \tau\sqrt{r}\|\mathcal{A}(X - Z)\|_{\ell_2} \\ &= \frac{1+\rho}{1-\rho}\left(\|Z\|_1 - \|X\|_1 + 2\|X_c\|_1\right) + \frac{2\tau\sqrt{r}}{1-\rho}\|\mathcal{A}(X - Z)\|_{\ell_2}.\end{aligned}$$

This concludes the proof. □

Now we return to the proof of the theorem.

Proof of Theorem 12. By Hölder's inequality, Lemma 13 and the Frobenius robust rank null space property of $\mathcal{A}$

$$\begin{aligned}\|Z-X\|_p &\leq \|(X-Z)_r\|_p + \|(X-Z)_c\|_p \leq r^{1/p-1/2}\|(X-Z)_r\|_2 + \|(X-Z)_c\|_p \\ &\leq \frac{\rho}{r^{1-1/p}}\|(X-Z)_c\|_1 + \tau r^{1/p-1/2}\|\mathcal{A}(X-Z)\|_{\ell_2} + \frac{1}{r^{1-1/p}}\|X-Z\|_1 \\ &\leq \frac{1+\rho}{r^{1-1/p}}\|X-Z\|_1 + \tau r^{1/p-1/2}\|\mathcal{A}(X-Z)\|_{\ell_2}. \end{aligned} \tag{27}$$

Substituting the result of Lemma 14 into (27) yields the desired inequality. □

As a corollary of Theorem 12 we obtain that if $X \in \mathbb{C}^{n_1\times n_2}$ is a matrix of rank at most r and the measurements are noiseless ($\eta = 0$), then the Frobenius robust rank null space property implies that X is the unique solution of

$$\min_{Z\in\mathbb{C}^{n_1\times n_2}} \|Z\|_1 \quad \text{subject to } \mathcal{A}(Z) = b. \tag{28}$$

It was first stated in [57] that a slightly weaker property is actually equivalent to the successful recovery of X via (28).

Theorem 15 (Null space property)**.** *Given $\mathcal{A} : \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^m$, every $X \in \mathbb{C}^{n_1\times n_2}$ of rank at most r is the unique solution of (28) with $b = \mathcal{A}(X)$ if and only if, for all $M \in \ker\mathcal{A}\setminus\{0\}$, it holds*

$$\|M_r\|_1 < \|M_c\|_1. \tag{29}$$

For the proof we refer to [57] and [26, Chapter 4.6]. According to Lemma 14, another implication of the Frobenius robust rank null space property consists in the following error estimate in $\|\cdot\|_1$ for the case of noiseless measurements,

$$\|X - X^\sharp\|_1 \leq \frac{2(1+\rho)}{1-\rho}\|X_c\|_1.$$

The above estimate remains true, if we require that for all $M \in \ker\mathcal{A}$, the singular values of M satisfy

$$\|M_r\|_1 \leq \rho\|M_c\|_1, \quad 0 < \rho < 1.$$

This property is known as the stable rank null space property of order r with constant ρ. It is clear that if $\mathcal{A} : \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^m$ satisfies the Frobenius robust rank null space property, then it satisfies the stable rank null space property. The approach used in [54] to verify that the stable null space property accounts for stable recovery of matrices which are not exactly of low rank, exploits the similarity between the sparse vector recovery and the low-rank matrix recovery. It shows that if some condition is sufficient for stable and robust recovery of any sparse vector with at most r non-zero entries, then the extension of this condition to the matrix case is sufficient for the stable and robust recovery of any matrix up to rank r.

In order to check whether the measurement map $\mathcal{A} : \mathbb{C}^{n_1\times n_2} \to \mathbb{C}^m$ satisfies the Frobenius robust rank null space property, we introduce the set

$$T_{\rho,r} := \left\{M \in \mathbb{C}^{n_1\times n_2} : \|M\|_2 = 1, \|M_r\|_2 > \frac{\rho}{\sqrt{r}}\|M_c\|_1\right\}.$$

Lemma 16. *If*

$$\inf\{\|\mathcal{A}(M)\|_{\ell_2} : M \in T_{\rho,r}\} > \frac{1}{\tau},$$

then $\mathcal{A}$ satisfies the Frobenius robust rank null space property of order r with constants ρ and τ.

Proof. Suppose that

$$\inf\{\|\mathcal{A}(M)\|_{\ell_2} : M \in T_{\rho,r}\} > \frac{1}{\tau}. \tag{30}$$

It follows that for any $M \in \mathbb{C}^{n_1 \times n_2}$ such that $\|\mathcal{A}(M)\|_{\ell_2} \leq \frac{\|M\|_2}{\tau}$ it holds

$$\|M_r\|_2 \leq \frac{\rho}{\sqrt{r}}\|M_c\|_1. \tag{31}$$

For the remaining $M \in \mathbb{C}^{n_1 \times n_2}$ with $\|\mathcal{A}(M)\|_{\ell_2} > \frac{\|M\|_2}{\tau}$ we have

$$\|M_r\|_2 \leq \|M\|_2 < \tau\|\mathcal{A}(M)\|_{\ell_2}.$$

Together with (31) this leads to

$$\|M_r\|_2 \leq \frac{\rho}{\sqrt{r}}\|M_c\|_1 + \tau\|\mathcal{A}(M)\|_{\ell_2}.$$

for any $M \in \mathbb{C}^{n_1 \times n_2}$. □

It is natural to expect that the recovery error gets smaller as the number of measurements increases. This can be taken into account by establishing the null space property for $\tau = \frac{\kappa}{\sqrt{m}}$. Then the error bound reads as follows

$$\|X - X^\sharp\|_2 \leq \frac{2(1+\rho)^2}{(1-\rho)\sqrt{r}}\|X_c\|_1 + \frac{2\kappa(3+\rho)}{\sqrt{m}(1-\rho)}\eta.$$

An important property of the set $T_{\rho,r}$ is that it is imbedded in a set with a simple structure. The next lemma relies on the ideas presented in [59] for the compressed sensing setting.

Lemma 17. *Let D be the set defined by*

$$D := \operatorname{conv}\left\{M \in \mathbb{C}^{n_1 \times n_2} : \|M\|_2 = 1, \operatorname{rank} M \leq r\right\}, \tag{32}$$

where conv *stands for the convex hull.*

(a) Then D is the unit ball with respect to the norm

$$\|M\|_D := \sum_{j=1}^{L}\left[\sum_{i \in I_j}(\sigma_i(M))^2\right]^{1/2},$$

where $L = \lceil \frac{n}{r} \rceil$,

$$I_j = \begin{cases} \{r(j-1)+1, \dots, rj\}, & j = 1, \dots, L-1, \\ \{r(L-1)+1, \dots, n\}, & j = L. \end{cases}$$

(b) It holds

$$T_{\rho,r} \subset \sqrt{1 + (1+\rho^{-1})^2}D. \tag{33}$$

Let us argue briefly why $\|\cdot\|_D$ is a norm. Define $g : \mathbb{C}^n \to [0, \infty)$ by

$$g(x) := \sum_{j=1}^{L}\left(\sum_{i \in I_j}(x_i^*)^2\right)^{1/2},$$

where L and I_j are defined in the same way as in item (a) of Lemma 17. Then g is a symmetric gauge function and $\|M\|_D = g(\sigma(M))$ for any $M \in \mathbb{C}^{n_1 \times n_2}$. The norm property follows from [33, Theorem 7.4.7.2].

Proof of Lemma 17. (a) Any $M \in D$ can be written as

$$M = \sum_i \alpha_i X_i$$

with

$$\operatorname{rank} X_i \leq r,\ \|X_i\|_2 = 1,\ \alpha_i \geq 0,\ \sum_i \alpha_i = 1.$$

Thus

$$\|M\|_D \leq \sum_i \alpha_i \|X_i\|_D = \sum_i \alpha_i \|X_i\|_2 = \sum_i \alpha_i = 1.$$

Conversely, suppose that $\|M\|_D \leq 1$, and let M have a singular value decomposition $M = U\Sigma V^* = \sum_{j=1}^{L}\sum_{i\in I_j} \sigma_i(M)u_i v_i^*$, where $u_i \in \mathbb{C}^{n_1}$ and $v_i \in \mathbb{C}^{n_2}$ are column vectors of U and V respectively. Set $M_j := \sum_{i\in I_j} \sigma_i(M)u_i v_i^*$ and $\alpha_j := \|M_j\|_2$, $j = 1, \ldots, L$. Then each M_j is a sum of r rank-one matrices, so that $\operatorname{rank} M_j \leq r$, and we can write M as

$$M = \sum_{j:\alpha_j\neq 0} \alpha_j \left(\frac{1}{\alpha_j} M_j\right)$$

with

$$\sum_{j:\alpha_j\neq 0} \alpha_j = \sum_j \|M_j\|_2 = \|M\|_D \leq 1 \quad \text{and} \quad \|\frac{1}{\alpha_j} M_j\|_2 = \frac{1}{\alpha_j}\|M_j\|_2 = 1.$$

Hence $M \in D$.

(b) To prove the embedding of $T_{\rho,r}$ into a scaled version of D, we estimate the norm of an arbitrary element M of $T_{\rho,r}$. According to the definition of the $\|\cdot\|_D$-norm

$$\|M\|_D = \sum_{\ell=1}^{L}\left[\sum_{i\in I_\ell}(\sigma_i(M))^2\right]^{\frac{1}{2}} = \|M_r\|_2 + \left[\sum_{i=r+1}^{2r}(\sigma_i(M))^2\right]^{\frac{1}{2}} + \sum_{\ell\geq 3}^{L}\left[\sum_{i\in I_\ell}(\sigma_i(M))^2\right]^{\frac{1}{2}}. \tag{34}$$

To bound the last term in the inequality above, we first note that for each $i \in I_\ell$, $\ell \geq 3$,

$$\sigma_i(M) \leq \frac{1}{r}\sum_{j\in I_{\ell-1}} \sigma_j(M)$$

and hence

$$\left[\sum_{i\in I_\ell}(\sigma_i(M))^2\right]^{1/2} \leq \frac{1}{\sqrt{r}}\sum_{j\in I_{\ell-1}} \sigma_j(M).$$

Summing up over $\ell \geq 3$ yields

$$\sum_{\ell\geq 3}^{L}\left[\sum_{i\in I_\ell}(\sigma_i(M))^2\right]^{\frac{1}{2}} \leq \frac{1}{\sqrt{r}}\sum_{l\geq 2}\sum_{j\in I_\ell}\sigma_j(M) = \frac{1}{\sqrt{r}}\sum_{j=r+1}^{n}\sigma_j(M) = \frac{1}{\sqrt{r}}\|M_c\|_1.$$

and taking into account the inequality for the singular values of $M \in T_{\rho,r}$

$$\sum_{\ell\geq 3}^{L}\left[\sum_{i\in I_\ell}(\sigma_i(M))^2\right]^{\frac{1}{2}} \leq \rho^{-1}\|M_r\|_2.$$

Applying the last estimate to (34) we derive that

$$\|M\|_D \leq (1+\rho^{-1})\|M_r\|_2 + \left[\sum_{i=r+1}^{2r}(\sigma_i(M))^2\right]^{\frac{1}{2}} \leq (1+\rho^{-1})\|M_r\|_2 + \left(1-\|M_r\|_2^2\right)^{\frac{1}{2}}.$$

Set $a = \|M_r\|_2$. The maximum of the function

$$f(a) := (1+\rho^{-1})a + \sqrt{1-a^2}, \quad 0 \leq a \leq 1,$$

is attained at the point

$$a = \frac{1+\rho^{-1}}{\sqrt{1+(1+\rho^{-1})^2}}$$

and is equal to $\sqrt{1+(1+\rho^{-1})^2}$. Thus for any $M \in T_{\rho,r}$ it holds

$$\|M\|_D \leq \sqrt{1+(1+\rho^{-1})^2},$$

which proves (33). □

Remark 18. The previous results hold true in the real-valued case and in the case of Hermitian matrices, when the nuclear norm minimization problem is solved over the set of matrices of that special type. As a set D we then take the convex hull of corresponding matrices of rank r and unit Frobenius norm. The only difference in the proof of Lemma 17 occurs at the point, where we have to show that any M with $\|M\|_D \leq 1$ belongs to D. Say, $M \in \mathbb{C}^{n\times n}$ is Hermitian and $\|M\|_D \leq 1$. Then $M = U\Lambda U^* = \sum_{j=1}^{L} \sum_{i\in I_j} \sigma_i(M) u_i u_i^*$, where $u_i \in \mathbb{C}^n$, and $M_j := \sum_{i\in I_j} \sigma_i(M) u_i u_i^*$ is Hermitian. The rest of the proof remains unchained.

Employing the matrix representation of the measurement map $\mathcal{A}$, the problem of estimating the probability of the event (30) is reduced to the problem of giving a lower bound for the quantities of the form $\inf_{x\in T} \|Ax\|_2$. This is not an easy task for deterministic matrices, but the situation significantly changes for matrices chosen at random.

4. Gaussian measurements

Our main result for Gaussian measurements reads as follows.

Theorem 19. *Let $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ be the linear map (1) generated by a sequence $A_1, \dots, A_m$ of independent standard Gaussian matrices, let $0 < \rho < 1$, $\kappa > 1$ and $0 < \varepsilon < 1$. If*

$$\frac{m^2}{m+1} \geq \frac{r(1+(1+\rho^{-1})^2)\kappa^2}{(\kappa-1)^2} \left[\sqrt{n_1} + \sqrt{n_2} + \sqrt{\frac{2\ln(\varepsilon^{-1})}{r(1+(1+\rho^{-1})^2)}} \right]^2, \tag{35}$$

then with probability at least $1-\varepsilon$, for every $X \in \mathbb{R}^{n_1\times n_2}$, a solution $X^\sharp$ of (4) with $b = \mathcal{A}(X) + w$, $\|w\|_{\ell_2} \leq \eta$, approximates X with error

$$\|X - X^\sharp\|_2 \leq \frac{2(1+\rho)^2}{(1-\rho)\sqrt{r}} \|X_c\|_1 + \frac{2\kappa\sqrt{2}(3+\rho)}{\sqrt{m}(1-\rho)} \eta.$$

In order to prove Theorem 19 we employ Gordon's escape through a mesh theorem that provides an estimate of the probability of the event (30). First we recall some definitions. Let $g \in \mathbb{R}^m$ be a standard Gaussian random vector, that is, a vector of independent mean zero, variance one normal distributed random variables. Then for

$$E_m := \mathbb{E}\,\|g\|_2 = \sqrt{2}\,\frac{\Gamma\left((m+1)/2\right)}{\Gamma\left(m/2\right)}$$

we have

$$\frac{m}{\sqrt{m+1}} \leq E_m \leq \sqrt{m},$$

see [27, 26]. For a set $T \subset \mathbb{R}^n$ we define its Gaussian width by

$$\ell(T) := \mathbb{E} \sup_{x\in T} \langle x, g\rangle,$$

where $g \in \mathbb{R}^n$ is a standard Gaussian random vector.

Theorem 20 (Gordon's escape through a mesh [27])**.** *Let $A \in \mathbb{R}^{m\times n}$ be a Gaussian random matrix and T be a subset of the unit sphere $\mathbb{S}^{n-1}$. Then, for $t > 0$,*

$$\mathbb{P}\left(\inf_{x\in T} \|Ax\|_2 > E_m - \ell(T) - t \right) \geq 1 - e^{-\frac{t^2}{2}}. \tag{36}$$

In order to apply this result to our measurement process (1) we unravel the columns of A_j, $j = 1, \dots, m$, into a single row and collect all of these in a $m \times n_1 n_2$-matrix A, so that $n = n_1 n_2$ when applying (36). In order to give a bound on the number of Gaussian measurements, Theorem 20 requires to estimate the Gaussian width of the set $T_{\rho,r}$ from above. As it was pointed out in the previous section, $T_{\rho,r}$ is a subset of a scaled version of D, which has a relatively simple structure. So instead of evaluating $\ell(T_{\rho,r})$, we consider $\ell(D)$.

Lemma 21. *For the set D defined by (32) it holds*

$$\ell(D) \leq \sqrt{r}(\sqrt{n_1} + \sqrt{n_2}). \tag{37}$$

Proof. Let $\Gamma \in \mathbb{R}^{n_1 \times n_2}$ have independent standard normal distributed entries. Then $\ell(D) = \mathbb{E} \sup_{M \in D} \langle \Gamma, M \rangle$. Since a convex continuous real-valued function attains its maximum value at one of the extreme points, it holds $\ell(D) = \mathbb{E} \sup_{\substack{\|M\|_2 = 1 \\ \operatorname{rank} M \leq r}} \langle \Gamma, M \rangle$. By Hölder's inequality,

$$\ell(D) \leq \mathbb{E} \sup_{\substack{\|M\|_2 = 1 \\ \operatorname{rank} M \leq r}} \|\Gamma\|_\infty \|M\|_1 \leq \sqrt{r} \sup_{\substack{\|M\|_2 = 1 \\ \operatorname{rank} M \leq r}} \|M\|_2 \, \mathbb{E} \, \sigma_1(\Gamma) \leq \sqrt{r}(\sqrt{n_1} + \sqrt{n_2}),$$

where the last inequality follows from an estimate for the expectation of the largest singular value of a Gaussian matrix, see [26, Chapter 9.3]. □

Proof of Theorem 19. Set $t := \sqrt{2 \ln(\varepsilon^{-1})}$. If m satisfies (35), then

$$E_m \left(1 - \frac{1}{\kappa}\right) \geq \sqrt{r(1 + (1 + \rho^{-1})^2)}(\sqrt{n_1} + \sqrt{n_2}) + t.$$

Together with (33) and (37) this yields

$$E_m - \ell(T_{\rho,r}) - t \geq \frac{E_m}{\kappa} \geq \frac{1}{\kappa}\sqrt{\frac{m}{2}}.$$

According to Theorem 20

$$\mathbb{P}\left(\inf_{M \in T_{\rho,r}} \|\mathcal{A}(M)\|_2 > \frac{\sqrt{m}}{\kappa\sqrt{2}}\right) \geq 1 - \varepsilon,$$

which means that with probability at least $1 - \varepsilon$ map $\mathcal{A}$ satisfies the Frobenius robust rank null space property with constants ρ and $\frac{\kappa\sqrt{2}}{\sqrt{m}}$. The error estimate follows from Theorem 11. □

5. Measurement matrices with independent entries and four finite moments

In this section we prove Theorem 1, which is the generalization of Theorem 19 to the case when the map $\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is obtained from m independent samples of a random matrix $\Phi = (X_{ij})_{i,j}$ with the following properties:

- The X_{ij} are independent random variables of mean zero,
- $\mathbb{E}X_{ij}^2 = 1$ and $\mathbb{E}X_{ij}^4 \leq C_4$ for all i, j and some constant C_4.

Note that (by Hölder's inequality) $C_4 \geq 1$.

As before the idea of the proof is to show that the event (30) holds with high probability. In order to do so we apply Mendelson's small ball method [40, 50, 66] in the manner of [66].

Theorem 22 ([40, 50, 66])**.** *Fix $E \subset \mathbb{R}^d$ and let $\phi_1, \dots, \phi_m$ be independent copies of a random vector ϕ in $\mathbb{R}^d$. For $\xi > 0$ let*

$$Q_\xi(E; \phi) = \inf_{u \in E} \mathbb{P}\{|\langle \phi, u \rangle| \geq \xi\}$$

and

$$W_m(E; \phi) = \mathbb{E} \sup_{u \in E} \langle h, u \rangle,$$

where $h = \frac{1}{\sqrt{m}} \sum_{j=1}^m \varepsilon_j \phi_j$ with (ε_j) being a Rademacher sequence [4]*. Then for any $\xi > 0$ and any $t \geq 0$ with probability at least $1 - e^{-2t^2}$*

$$\inf_{u \in E} \left(\sum_{i=1}^m |\langle \phi_i, u \rangle|^2\right)^{1/2} \geq \xi\sqrt{m} Q_{2\xi}(E; \phi) - 2W_m(E; \phi) - \xi t.$$

We start with two lemmas.

[4] i.e., the ε_j are independent and assume the values 1 and -1 with probability 1/2, respectively.

Lemma 23.

$$\inf_{\{Y,\|Y\|_2=1\}} \mathbb{P}(|\langle \Phi, Y\rangle| \geq \frac{1}{\sqrt{2}}) \geq \frac{1}{4C_5},$$

where $C_5 = \max\{3, C_4\}$.

Proof. Assume that Y has Frobenius norm one. The Payley-Zygmund inequality (see e.g. [26, Lemma 7.16], and also [66]), implies

$$\mathbb{P}\{|\langle \Phi, Y\rangle|^2 \geq \frac{1}{2}(\mathbb{E}|\langle \Phi, Y\rangle|^2)\} \geq \frac{1}{4} \cdot \frac{(\mathbb{E}|\langle \Phi, Y\rangle|^2)^2}{\mathbb{E}|\langle \Phi, Y\rangle|^4}. \tag{38}$$

We compute numerator and denominator.

$$\mathbb{E}|\langle \Phi, Y\rangle|^2 = \sum_{i,j,k,l} \mathbb{E}(X_{ij}X_{kl}) \cdot Y_{ij}Y_{kl} = \sum_{i,j} \mathbb{E}X_{ij}^2 \cdot Y_{ij}^2 = \sum_{i,j} Y_{ij}^2 = 1.$$

Likewise,

$$\begin{aligned}
\mathbb{E}|\langle \Phi, Y\rangle|^4 &= \sum_{i_1,\ldots,i_4,j_1,\ldots,j_4} \mathbb{E}(X_{i_1j_1}\cdots X_{i_4j_4}) \cdot Y_{i_1j_1}\cdots Y_{i_4j_4} \\
&= \sum_{i,j} \mathbb{E}X_{ij}^4 \cdot Y_{ij}^4 + 3 \sum_{\substack{i_1,i_2,j_1,j_2 \\ (i_1,j_1)\neq(i_2,j_2)}} \mathbb{E}(X_{i_1j_1}^2 X_{i_2j_2}^2) \cdot Y_{i_1j_1}^2 Y_{i_2j_2}^2 \\
&= \sum_{i,j} \mathbb{E}X_{ij}^4 \cdot Y_{ij}^4 + 3 \sum_{\substack{i_1,i_2,j_1,j_2 \\ (i_1,j_1)\neq(i_2,j_2)}} Y_{i_1j_1}^2 Y_{i_2j_2}^2 \leq \sum_{i,j} C_4 \cdot Y_{ij}^4 + 3 \sum_{\substack{i_1,i_2,j_1,j_2 \\ (i_1,j_1)\neq(i_2,j_2)}} Y_{i_1j_1}^2 Y_{i_2j_2}^2 \\
&\leq C_5 \sum_{i_1,i_2,j_1,j_2} Y_{i_1j_1}^2 Y_{i_2j_2}^2 = C_5(\sum_{i,j} Y_{ij}^2)^2 = C_5.
\end{aligned}$$

Combining this with $(\mathbb{E}|\langle \Phi, Y\rangle|^2)^2 = 1$ and the estimate (38), the claim follows. □

Lemma 24. *Let* $\Phi_1, \ldots, \Phi_m$ *be independent copies of a random matrix* Φ *as above. Let* $\varepsilon_1, \ldots, \varepsilon_m$ *be independent Rademacher variables independent of everything else and let* $H = \frac{1}{\sqrt{m}} \sum_{k=1}^m \varepsilon_k \Phi_k$. *Then*

$$\mathbb{E}\|H\|_\infty \leq C_1\sqrt{n}.$$

Here C_1 *is a constant that only depends on* C_4.

Proof. Let $S = \sum_{k=1}^m \Phi_k$. We first desymmetrize the sum H (see [45, Lemma 6.3]) and obtain

$$\mathbb{E}\|H\|_\infty \leq \frac{2}{\sqrt{m}}\mathbb{E}\|S\|_\infty.$$

Therefore, it is enough to show that $\mathbb{E}\|S\|_\infty \leq c_3\sqrt{mn}$ for a suitable constant c_3. The matrix S has independent mean zero entries, hence by a result Latała (see [44]) the following estimate holds for some universal constant C_2,

$$\mathbb{E}\|S\|_\infty \leq C_2 \left(\max_i \sqrt{\sum_j \mathbb{E}S_{ij}^2} + \max_j \sqrt{\sum_i \mathbb{E}S_{ij}^2} + \sqrt[4]{\sum_{i,j} \mathbb{E}S_{ij}^4} \right).$$

Denoting the entries of Φ_k by $X_{k;ij}$, we have $S_{ij} = \sum_k X_{k;ij}$. Hence, using the independence of the $X_{k;ij}$, we obtain $\mathbb{E}S_{ij}^2 = \mathbb{E}(\sum_k X_{k;ij})^2 = \sum_k \mathbb{E}X_{k;ij}^2 = m$. Thus, $\sqrt{\sum_j \mathbb{E}S_{ij}^2} \leq \sqrt{nm}$ for any i and $\sqrt{\sum_i \mathbb{E}S_{ij}^2} \leq \sqrt{nm}$ for any j. Finally to estimate $\sqrt[4]{\sum_{i,j} \mathbb{E}S_{ij}^4}$ we calculate $\mathbb{E}S_{ij}^4 = \mathbb{E}(\sum_k X_{k;ij})^4$. Using again that the $X_{k;ij}$ are independent and have mean zero we obtain

$$ES_{ij}^4 = \sum_k \mathbb{E}X_{k;ij}^4 + 3 \sum_{k_1\neq k_2} \mathbb{E}X_{k_1;ij}^2 \mathbb{E}X_{k_2;ij}^2.$$

Using that $\mathbb{E}X_{k;ij}^2 = 1$ for all i, j, k, we obtain $\mathbb{E}S_{ij}^4 \leq C_5 m^2$, where $C_5 = \max\{3, C_4\}$ and hence

$$\sqrt[4]{\sum_{i,j} \mathbb{E}S_{ij}^4} \leq \sqrt[4]{C_5 m^2 n^2} = \sqrt[4]{C_4}\sqrt{mn}.$$

Hence, indeed $\mathbb{E}\|S\|_\infty \leq c_3\sqrt{mn}$ for a suitable constant c_3 that depends only on C_4. □

Proof Theorem 1. Let now $T_{\rho,r}$ and D be the sets defined in Section 3, but restricted to the real-valued matrices. By Hölder's inequality, for any $n_1 \times n_2$ matrix Y of Frobenius norm 1 and rank at most r and any $n_1 \times n_2$ matrix H,

$$\langle H, Y\rangle \leq \|Y\|_1 \|H\|_\infty \leq \sqrt{r}\|H\|_\infty.$$

Hence

$$\sup_{Y\in D} \langle H, Y\rangle \leq \sqrt{r}\|H\|_\infty. \tag{39}$$

Let $H = \frac{1}{\sqrt{m}}\sum_{j=1}^m \varepsilon_j \Phi_j$ and let $\xi = \frac{1}{2\sqrt{2}}$ and $E = T_{\rho,r}$. Then it follows from Theorem 22 that for any $t \geq 0$ with probability at least $1 - e^{-2t^2}$

$$\inf_{Y\in T_{\rho,r}} \left(\sum_{i=1}^m |\langle \Phi_i, Y\rangle|^2\right)^{1/2} \geq \frac{\sqrt{m}}{2\sqrt{2}} Q_{\frac{1}{\sqrt{2}}}(T_{\rho,r}; \Phi) - 2W_m(T_{\rho,r}, \Phi) - \frac{1}{2\sqrt{2}}t. \tag{40}$$

Using Lemma 23 and the fact that all elements of $T_{\rho,r}$ have Frobenius norm 1, we obtain

$$Q_{\frac{1}{\sqrt{2}}}(T_{\rho,r}; \Phi) \geq \frac{1}{4C_5}. \tag{41}$$

Combining now the fact that $T_{\rho,r} \subseteq \sqrt{1 + (1+\rho^{-1})^2} D$ (see Lemma 17) with estimate (39) and Lemma 24 leads to

$$W_m(T_{\rho,r}, \Phi) \leq \sqrt{1 + (1+\rho^{-1})^2}\sqrt{r}\ \mathbb{E}\|H\|_\infty \leq C_1\sqrt{1 + (1+\rho^{-1})^2}\sqrt{r}\sqrt{n}. \tag{42}$$

Using (40), (41) and (42) we see that choosing $m \geq c_1\rho^{-2}nr$ and $t = c_4 m$ for suitable constants c_1, c_4, we obtain with probability at least $1 - e^{-c_2 m}$

$$\inf_{Y\in T_{\rho,r}} \left(\sum_{i=1}^m |\langle \Phi_i, Y\rangle|^2\right)^{1/2} \geq c_3\sqrt{m}$$

for suitable constants c_2, c_3. Now the claim follows from Lemma 16 and Theorem 11 (both of which also hold in the real valued version by the same proofs respectively). □

6. Rank one Gaussian measurements

In this section we prove Theorem 2. The proof technique is an application of Mendelson's small ball method analogous to the proof of Theorem 1. Let

$$T_{\rho,r}^{\mathcal{H}} := \left\{ M \in \mathcal{H}_n : \|M\|_2 = 1,\ \|M_r\|_2 > \frac{\rho}{\sqrt{r}}\|M_c\|_1 \right\}.$$

Let $T_{\rho,r}$ be defined as $T_{\rho,r}^{\mathcal{H}}$ but with $\mathcal{H}_n$ replaced by the set of all complex $n \times n$-matrices (i.e. it is defined as before with $n_1 = n_2 = n$). Then $T_{\rho,r}^{\mathcal{H}} \subseteq T_{\rho,r}$. It is enough to show that with high probabiliy

$$\inf_{Y\in T_{\rho,r}^{\mathcal{H}}} \left(\sum_{j=1}^m |\langle a_j a_j^*, Y\rangle|^2\right)^{1/2} \geq \sqrt{m}/C_3 \tag{43}$$

We apply Theorem 22 with $E = T_{\rho,r}^{\mathcal{H}}$. The next lemma estimates the small ball probability $Q_{\frac{1}{\sqrt{2}}}(E; \phi)$ used in Mendelson's method.

Lemma 25 (see [43]). $Q_{\frac{1}{\sqrt{2}}}(E; \phi) := \inf_{u\in E} \mathbb{P}\{|\langle aa^*, u\rangle| \geq \frac{1}{\sqrt{2}}\} \geq \frac{1}{96}$.

Let now (as in [66, 43])

$$H = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \varepsilon_j a_j a_j^*, \tag{44}$$

where the ε_j form a Rademacher sequence. For any $M \in \mathcal{H}_n$ and any $n \times n$ matrix Y of Frobenius norm 1 and rank at most r

$$\langle M, Y \rangle \leq \|Y\|_1 \|M\|_\infty \leq \sqrt{r} \|M\|_\infty.$$

Since $E = T_{\rho,r}^{\mathcal{H}} \subseteq T_{\rho,r} \subseteq \sqrt{1 + (1+\rho^{-1})^2} D$, this implies

$$W_m(E, \phi) = \mathbb{E} \sup_{Y \in E} \langle H, Y \rangle \leq \sqrt{1 + (1+\rho^{-1})^2} \sqrt{r} \mathbb{E} \|H\|_\infty.$$

As in [43] we use now that by the arguments in [67, Section 5.4.1] we have $\mathbb{E}\|H\|_\infty \leq c_2 \sqrt{n}$ if $m \geq c_3 n$ for suitable constants c_2, c_3, see also [66, Section 8]. Now the claim of Theorem 2 follows from Theorem 22, comp. the proof of Theorem 1. □

Remark 26. Inspecting the above proof, resp. the proofs of the cited statements in [43], we see that the real valued analogue of Theorem 2 is also true. We even may assume for this that the a_j are i.i.d. subgaussian with k-th moments, where $k \leq 8$, equal to the corresponding k-th moments of the Gaussian standard distribution. The constants then depend only on the distribution of the a_j. We also note that a similar statement in the real case for the recovery of positive semidefinite matrices using subgaussian measurements has been shown by Chen, Chi and Goldsmith in [19] using the rank restricted isometry property.

7. Rank one measurements generated by 4-designs

Recall the definition of an approximate, weighted t-design.

Definition 27 (*Approximate t-design*, Definition 2 in [3])**.** We call a weighted set $\{p_i, w_i\}_{i=1}^N$ of normalized vectors an approximate t-design of p-norm accuracy θ_p, if

$$\left\| \sum_{i=1}^{N} p_i \left(w_i w_i^* \right)^{\otimes t} - \int_{\|w\|_{\ell_2}=1} \left(w w^* \right)^{\otimes t} \mathrm{d}w \right\|_p \leq \binom{n+t-1}{t}^{-1} \theta_p. \tag{45}$$

A set of unit vectors obeying $\theta_p = 0$ for $1 \leq p \leq \infty$ is called an *exact t-design*, see [62] and also [43, 28].

Theorem 28. *Let $\{p_i, w_i\}_{i=1}^N$ be a an approximate 4-design with either $\theta_\infty \leq 1/(16r^2)$, or $\theta_1 \leq 1/4$ that furthermore obeys $\left\| \sum_{i=1}^N p_i w_i w_i^* - \frac{1}{n} \mathrm{id} \right\|_\infty \leq \frac{1}{n}$. Suppose that the measurement operator $\mathcal{A}$ is generated by*

$$m \geq C_4 \rho^{-2} n r \log n$$

measurement matrices $A_j = \sqrt{n(n+1)} a_j a_j^*$, where each a_j is drawn independently from $\{p_i, w_i\}_{i=1}^N$. Then, with probability at least $1 - \mathrm{e}^{-C_5 m}$, $\mathcal{A}$ obeys the Frobenius robust rank null space property of order r with constants $0 < \rho < 1$ and $\tau = C_6/\sqrt{m}$. Here, C_4, C_5 and C_6 denote positive constants depending only on the design.

Theorem 3 readily follows from combining this statement with Theorem 12.

Proof of Theorem 28. We start by presenting a proof for measurements drawn from an exact 4-design. Paralleling the proof of Theorem 2, the statement can be deduced from Theorem 22 by utilizing results from [43]. Provided that a is randomly chosen from a re-scaled, weighted 4-design (such that each element has Euclidean length $\|w_i\|_{\ell_2} = \sqrt[4]{(n+1)n)}$, [43, Proposition 12] implies that

$$\inf_{Z \in T_{\rho,r}} \mathbb{P}\left(|\mathrm{tr}\left(a a^* Z \right)| \geq \xi \right) \geq \inf_{\|Z\|_2 = 1} \mathbb{P}\left(|\mathrm{tr}\left(a a^* Z \right)| \geq \xi \right) \geq \frac{(1-\xi^2)^2}{24} \tag{46}$$

is valid for all $\xi \in [0,1]$. Now let $H = \sum_{i=1}^m \epsilon_i a_i a_i^*$ be as in Theorem 22. Lemma 17 together with the fact that D is the convex hull of all matrices of rank at most r and Frobenius norm 1 allows us to conclude for $m \geq 2n \log n$, that,

$$\begin{aligned} W_m\left(T_{\rho,r}, aa^*\right) &= \mathbb{E} \sup_{M \in T_{\rho,r}} \operatorname{tr}(HM) \leq \sqrt{1+(1+\rho^{-1})^2}\, \mathbb{E} \sup_{M \in D} \operatorname{tr}(HM) \\ &\leq \sqrt{1+(1+\rho^{-1})^2} \sup_{M \in D} \|M\|_1 \mathbb{E}\|H\|_\infty \leq \sqrt{1+(1+\rho^{-1})^2}\sqrt{r}\, \mathbb{E}\|H\|_\infty \\ &\leq 3.1049\sqrt{1+(1+\rho^{-1})^2 rn\log(2n)}, \end{aligned}$$

where the last bound is due to [43, Proposition 13]. Fixing $0 < \xi < 1/2$ arbitrarily and inserting these two bounds into Theorem 22 completes the proof.

An analogous statement for approximate 4-designs — with slightly worse absolute constants — can be obtained by resorting to the generalized versions of [43, Propositions 12 and 13] presented in Section 4.5.1 in loc. cit. which are valid for approximate 4-designs that satisfy the conditions stated in Theorem 28. □

8. The positive semidefinite case

Finally, we focus on the case, where the matrices of interest are Hermitian and positive semidefinite and establish Theorem 4. In order to arrive at such a statement, we closely follow the ideas presented in [36] which in turn were inspired by [9] containing an analogous statement for a non-negative compressed sensing scenario.

We require two further concepts from matrix analysis. For every positive semidefinite matrix $W \succcurlyeq 0$ with eigenvalue decomposition $W = \sum_{i=1}^n \lambda_i w_i w_i^*$ we define its square root to be $W^{1/2} := \sum_{i=1}^n \sqrt{\lambda_i} w_i w_i^*$. In other words, $W^{1/2}$ is the unique positive semidefinite matrix which acts on the eigenspace corresponding to the eigenvalue λ_i of W by multiplication by $\sqrt{\lambda_i}$. Note that this matrix obeys $W^{1/2} \cdot W^{1/2} = W$. Also, recall that the condition number $\kappa(W)$ of a matrix W is the ratio between its largest and smallest nonzero singular value. For an invertible Hermitian matrix with inverse W^{-1} this number equals

$$\kappa(W) = \|W\|_\infty \|W^{-1}\|_\infty .$$

Suppose that the measurement process (3) is such that there exists $t \in \mathbb{R}^m$ which assures that $W := \sum_{j=1}^m t_j A_j$ is positive definite. We define the artificial measurement map

$$\mathcal{A}_{W^{1/2}} : \mathcal{H}_n \to \mathbb{R}^m, \quad Z \mapsto \mathcal{A}(W^{-1/2} Z W^{-1/2}) \tag{47}$$

and the endomorphism

$$Z \mapsto \tilde{Z} := W^{1/2} Z W^{1/2} \tag{48}$$

of $\mathcal{H}_n$. Note that these definitions assure

$$\mathcal{A}(Z) = \mathcal{A}_{W^{1/2}}(\tilde{Z}) \quad \text{for all } Z \in \mathcal{H}_n \tag{49}$$

and the singular values of Z and $\tilde{Z}$ satisfy

$$\sigma_j(\tilde{Z}) \leq \|W^{1/2}\|_\infty^2 \sigma_j(Z) = \|W\|_\infty \sigma_j(Z), \quad \sigma_j(Z) \leq \|W^{-1/2}\|_\infty^2 \sigma_j(\tilde{Z}) = \|W^{-1}\|_\infty \sigma_j(\tilde{Z}), \tag{50}$$

see [7, p. 75]. Consequently, the mapping (48) preserves the rank of any matrix. The following result assures that the artificial measurement operator $\mathcal{A}_{W^{1/2}}$ obeys the Frobenius robust rank null space property, if the original $\mathcal{A}$ does.

Lemma 29. *Suppose that $\mathcal{A}$ satifies the Frobenius robust rank null space property of order r with constants ρ and τ and suppose that $W = \sum_{j=1}^m t_j A_j$ is positive definite. Then $\mathcal{A}_{W^{1/2}}$ also obeys the Frobenius robust rank null space property of order r, but with constants $\tilde{\rho} = \kappa(W)\rho$ and $\tilde{\tau} = \|W\|_\infty \tau$.*

Proof. Let $\tilde{Z} \in \mathcal{H}_n$. Relations (49), (50) together with the Frobenius robust rank null space property of $\mathcal{A}$ imply that

$$\begin{aligned} \|\tilde{Z}_r\|_2 &\leq \|W^{1/2}\|_\infty^2 \|Z_r\|_2 \leq \|W\|_\infty \left(\frac{\rho}{\sqrt{r}} \|Z_c\|_1 + \tau \|\mathcal{A}(Z)\|_{\ell_2} \right) \\ &\leq \frac{\|W\|_\infty \|W^{-1}\|_\infty \rho}{\sqrt{r}} \|\tilde{Z}_c\|_1 + \|W\|_\infty \tau \|\mathcal{A}_{W^{1/2}}(\tilde{Z})\|_{\ell_2}. \end{aligned}$$

□

Lemma 30. *Suppose there is $t \in \mathbb{R}^m$ such that $W := \sum_{j=1}^m t_j A_j$ is positive definite. Let $\tilde{X}, \tilde{Z}$ be positive semidefinite. Then,*

$$\|\tilde{Z}\|_1 - \|\tilde{X}\|_1 \leq \|t\|_{\ell_2} \|\mathcal{A}_{W^{1/2}}(\tilde{Z} - \tilde{X})\|_{\ell_2}.$$

Proof. The claim follows from positive semidefiniteness of both $\tilde{Z}$ and $\tilde{X}$ and our choice of the endomorphism (48). Indeed,

$$\begin{aligned} \|\tilde{Z}\|_1 &= \operatorname{tr}(\tilde{Z} - \tilde{X}) + \|\tilde{X}\|_1 = \operatorname{tr}(W^{1/2}(Z - X)W^{1/2}) + \|\tilde{X}\|_1 = \operatorname{tr}(W(Z - X)) + \|\tilde{X}\|_1 \\ &= \sum_{j=1}^m t_j \operatorname{tr}(A_j(Z - X)) + \|\tilde{X}\|_1 = \langle t, \mathcal{A}(Z - X)\rangle + \|\tilde{X}\|_1 \\ &= \langle t, \mathcal{A}_{W^{1/2}}(\tilde{Z} - \tilde{X})\rangle + \|\tilde{X}\|_1 \leq \|t\|_{\ell_2} \|\mathcal{A}_{W^{1/2}}(\tilde{Z} - \tilde{X})\|_{\ell_2} + \|\tilde{X}\|_1. \end{aligned}$$

Here X resp. Z denote the preimage of $\tilde{X}$ resp $\tilde{Z}$ under the map (48). □

This simple technical statement allows us to establish the main result of this section.

Theorem 31. *Suppose there exists $t \in \mathbb{R}^m$ such that $W := \sum_{j=1}^m t_j A_j$ is positive definite and $\mathcal{A}$ satisfies the Frobenius robust rank null space property with constants $0 < \rho < \frac{1}{\kappa(W)}$ and $\tau > 0$. Let $1 \leq p \leq 2$. Then, for any $X, Z \succcurlyeq 0$,*

$$\|Z - X\|_p \leq \frac{2C\kappa(W)}{r^{1-1/p}} \|X_c\|_1 + r^{1/p-1/2} \|\mathcal{A}(Z) - \mathcal{A}(X)\|_{\ell_2} \|W^{-1}\|_\infty \left(\frac{C\|t\|_2}{\sqrt{r}} + D\|W\|_\infty \tau \right) \tag{51}$$

with constants $C = \frac{(1+\kappa(W)\rho)^2}{1-\kappa(W)\rho}$ and $D = \frac{3+\kappa(W)\rho}{1-\kappa(W)\rho}$.

Proof. Let $X, Z \succcurlyeq 0$ be arbitrary. Then

$$\|Z - X\|_p = \left\| W^{-1/2} \left(\tilde{Z} - \tilde{X} \right) W^{-1/2} \right\|_p \leq \|W^{-1}\|_\infty \|\tilde{Z} - \tilde{X}\|_p$$

holds and the resulting matrices $\tilde{Z}, \tilde{X}$ are again positive-semidefinite. Also, since $\mathcal{A}$ satisfies the Frobenius robust rank null space property with constants $0 < \rho < \frac{1}{\kappa(W)}$ and $\tau > 0$, Lemma 29 assures that $\mathcal{A}_{W^{1/2}}$ does the same with constants $0 < \tilde{\rho} < 1$ and $\tilde{\tau} = \|W\|_\infty \tau > 0$. Combining this with Theorem 12 and Lemma 30 implies

$$\begin{aligned} \|\tilde{Z} - \tilde{X}\|_p &\leq \frac{C}{r^{1-1/p}} \left(\|\tilde{Z}\|_1 - \|\tilde{X}\|_1 + 2\|\tilde{X}_c\|_1 \right) + D\|W\|_\infty \tau r^{1/p-1/2} \|\mathcal{A}_{W^{1/2}}(\tilde{Z} - \tilde{X})\|_{\ell_2} \\ &\leq \frac{C}{r^{1-1/p}} \left(\|t\|_{\ell_2} \|\mathcal{A}_{W^{1/2}}(\tilde{Z} - \tilde{X})\|_{\ell_2} + 2\|\tilde{X}_c\|_1 \right) + D\|W\|_\infty \tau r^{1/p-1/2} \|\mathcal{A}_{W^{1/2}}(\tilde{Z} - \tilde{X})\|_{\ell_2} \\ &\leq \frac{2C}{r^{1-1/p}} \|\tilde{X}_c\|_1 + r^{1/p-1/2} \|\mathcal{A}_{W^{1/2}}(\tilde{Z} - \tilde{X})\|_{\ell_2} \left(\frac{C\|t\|_{\ell_2}}{\sqrt{r}} + D\|W\|_\infty \tau \right). \end{aligned}$$

The desired statement follows from this estimate by taking into account (49) and (50). □

Note that in contrast to other recovery guarantees established here, Theorem 31 does not require any convex optimization procedure. However, it does require the measurement process to obey an additional criterion: the intersection of the span of measurement matrices with the cone of positive definite matrices must be non-empty. We show that this is the case for the rank-one projective measurements introduced in the previous section with high probability. Since it has already been established that sufficiently many measurements of this kind obey the Frobenius robust rank null space property with high probability (see Theorems 2 and 28 and their respective proofs), Theorem 4 can then be established by taking the union bound over the individual probabilities of failure.

Proposition 32. *Suppose $m \geq 4n$ and let $A_1, \ldots, A_m$ be matrices of the form $a_j a_j^*$, where each $a_i \in \mathbb{C}^n$ is a random complex standard Gaussian vector. Then with probability at least $1 - 2\mathrm{e}^{-C_{10}m}$, $W := \frac{1}{m}\sum_{j=1}^m A_j$ is positive definite and obeys*

$$\max\left\{ \|W\|_\infty, \|W^{-1}\|_\infty, \kappa(W) \right\} \leq C_{11}. \tag{52}$$

Here, $C_9, C_{10}, C_{11} > 0$ denote universal positive constants.

Note that such a construction corresponds to setting $t = \frac{1}{m}(1, \ldots, 1)^T \in \mathbb{R}^m$ which obeys $\|t\|_{\ell_2} = 1/\sqrt{m}$.

Proof. For the sake of simplicity, we are going to establish the statement for real standard Gaussian vectors. Establishing the complex case can be done analogously and leads to slightly different constants. Let $e_1, \ldots, e_m$ denote the standard basis in $\mathbb{R}^m$. We define the auxiliary $m \times n$ matrix $A := \sum_{i=1}^m e_i a_i^*$ which obeys

$$\frac{1}{m} A^T A = \frac{1}{m} \sum_{i=1}^m a_i e_i^* \sum_{j=1}^m e_j a_j^* = \frac{1}{m} \sum_{i=1}^m a_i a_i^* = \frac{1}{m} \sum_{i=1}^m A_i = W.$$

Also, by construction, A is a random matrix with standard Gaussian entries. Essentially, this relation implies that mW is Wishart-distributed. From (8) and the defining properties of eigen- and singular values we infer that

$$\sqrt{\lambda_{\min}(W)} = \frac{1}{\sqrt{m}} \sqrt{\lambda_{\min}\left(A^T A\right)} = \frac{1}{\sqrt{m}} \lambda_{\min}\left(\sqrt{A^T A}\right) = \frac{1}{\sqrt{m}} \sigma_{\min}(A) \tag{53}$$

and an analogous statement is true for the largest eigenvalue $\lambda_{\max}(W)$. Since A is a Gaussian $m \times n$ matrix, concentration of measure implies that for any $\tilde{\tau} > 0$

$$\sqrt{m} - \sqrt{n} - \tilde{\tau} \leq \sigma_{\min}(A) \leq \sigma_{\max}(A) \leq \sqrt{m} + \sqrt{n} + \tilde{\tau} \tag{54}$$

with probability at least $1 - 2\mathrm{e}^{-\tilde{\tau}^2/2}$ — see e.g. [67, Corollary 5.35] or [26, Theorem 9.26]. Combining this with (53), recalling the assumption $m \geq 4n$ and defining $\tau = \tilde{\tau}/\sqrt{m}$ allows for establishing

$$\frac{1}{2} - \tau \leq 1 - \sqrt{\frac{n}{m}} - \tau \leq \sqrt{\lambda_{\min}(W)} \leq \sqrt{\lambda_{\max}(W)} \leq 1 + \sqrt{\frac{n}{m}} + \tau \leq \frac{3}{2} + \tau$$

with probability at least $1 - 2\mathrm{e}^{-m\tau^2/2}$. This inequality chain remains valid, if we square the individual terms. Setting $\tau = 1/4$ thus allows us to conclude

$$\max\left\{\lambda_{\max}(W), \lambda_{\min}^{-1}(W), \frac{\lambda_{\max}(W)}{\lambda_{\min}(W)}\right\} \leq \left(\frac{3/2 + \tau}{1/2 - \tau}\right)^2 = 49 = C_{11}, \tag{55}$$

with probability at least $1 - 2\mathrm{e}^{-m/32}$. □

Alternatively, we could have relied on bounds on the condition number of Gaussian random matrices presented in [20]. While these bounds would be slightly tighter, we feel that our derivation is more illustrative and it suffices for our purpose.

Proposition 33. *Suppose $m \geq \tilde{C}_4 nr \log n$ and let $A_1, \ldots, A_m$ be matrices of the form $a_j a_j^*$, where each $a_j \in \mathbb{C}^n$ is chosen independently from a weighted set $\{p_i, w_i\}_{i=1}^N$ of vectors obeying $\|w_i\|_{\ell_2}^2 = \sqrt{n(n+1)}$ for all $1 \leq i \leq N$ and*

$$\left\| \sum_{i=1}^N p_i w_i w_i^* - \sqrt{\frac{n+1}{n}} \mathrm{id} \right\|_\infty \leq \frac{1}{2}. \tag{56}$$

Then with probability at least $1 - \mathrm{e}^{-\gamma \tilde{C}_4 r}$, the matrix $W := \frac{1}{m} \sum_{j=1}^m A_j$ is positive definite and obeys

$$\max\left\{\|W\|_\infty,\ \|W^{-1}\|_\infty,\ \kappa(W)\right\} \leq 8. \tag{57}$$

Here, $\tilde{C}_4 > 1$ and $0 < \gamma \leq 1$ denote absolute constants of adequate size.

Note that condition (56) is slightly stronger than the corresponding condition in Theorem 28. Also, the construction of W again uses $t = \frac{1}{m}(1 \ldots, 1)^T \in \mathbb{R}^m$.

Proof. In order to show this statement, we are going to employ the matrix Bernstein inequality[5] [65, Theorem 6.1], see also [1], in order to establish

$$\left\| W - \sqrt{\frac{n+1}{n}}\,\mathrm{id} \right\|_\infty \leq \frac{3}{4} \tag{58}$$

with high probability. Let $\lambda_1(W), \ldots, \lambda_n(W)$ denote the eigenvalues of W. Then such a bound together with the definition of the operator norm assures

$$\begin{aligned}
1 - \lambda_{\min}(W) \leq \sqrt{\frac{n+1}{n}} - \lambda_{\min}(W) &\leq \left| \sqrt{\frac{n+1}{n}} - \lambda_{\min}(W) \right| \leq \max_{1 \leq i \leq n} \left| \sqrt{\frac{n+1}{n}} - \lambda_i(W) \right| \\
&= \left\| \sqrt{\frac{n+1}{n}}\,\mathrm{id} - W \right\|_\infty \leq 3/4, \\
\lambda_{\max}(W) - \sqrt{\frac{n+1}{n}} \leq \left| \lambda_{\max}(W) - \sqrt{\frac{n+1}{n}} \right| &\leq \max_{1 \leq i \leq n} \left| \sqrt{\frac{n+1}{n}} - \lambda_i(W) \right| \\
&= \left\| W - \sqrt{\frac{n+1}{n}}\,\mathrm{id} \right\|_\infty \leq 3/4.
\end{aligned}$$

This in turn implies $\lambda_{\min}(W) \geq 1/4$ as well as $\lambda_{\max}(W) \leq 3/4 + \sqrt{\frac{n+1}{n}} \leq 2$ for $n \geq 2$ and the desired bound (57) readily follows.

It remains to assure the validity of (58) with high probability. To this end, for $1 \leq k \leq m$, we define the random matrices $M_k := \frac{1}{m}\left(a_k a_k^* - \mathbb{E}\left[a_k a_k^*\right]\right)$, where each a_k is chosen independently at random from the weighted set $\{p_i, w_i\}_{i=1}^N$. This definition assures

$$\left\| W - \sqrt{\frac{n+1}{n}}\,\mathrm{id} \right\|_\infty = \left\| \sum_{k=1}^m \left(M_k + \mathbb{E}\left[a_k a_k^*\right] \right) - \sqrt{\frac{n+1}{n}}\,\mathrm{id} \right\|_\infty \leq \left\| \sum_{k=1}^m M_k \right\|_\infty + \frac{1}{2} \tag{59}$$

via the triangle inequality and assumption (56) and along similar lines

$$\left\| \mathbb{E}\left[a_k a_k^*\right] \right\|_\infty \leq \frac{1}{2} + \sqrt{\frac{n+1}{n}} \leq 2 \tag{60}$$

readily follows for any $1 \leq k \leq m$. The random matrices M_k have mean-zero by construction and each of them obeys

$$\|M_k\|_\infty = \frac{1}{m} \left\| a_k a_k^* - \mathbb{E}\left[a_k a_k^*\right] \right\|_\infty \leq \frac{1}{m} \max\left\{ \|a_k a_k^*\|_\infty, \left\|\mathbb{E}\left[a_k a_k^*\right]\right\|_\infty \right\} = \frac{1}{m} \|a_k\|_{\ell_2}^2 = \frac{\sqrt{(n+1)n}}{m},$$

as well as

$$\begin{aligned}
\left\|\mathbb{E}\left[M_k^2\right]\right\|_\infty &= \frac{1}{m^2} \left\| \mathbb{E}\left[\left(a_k a_k^*\right)^2 \right] - \mathbb{E}\left[a_k a_k^*\right]^2 \right\|_\infty = \frac{1}{m^2} \left\| \sqrt{(n+1)n}\,\mathbb{E}\left[a_k a_k^*\right] - \mathbb{E}\left[a_k a_k^*\right]^2 \right\|_\infty \\
&= \frac{2}{m^2} \max\left\{ \sqrt{(n+1)n}\, \left\|\mathbb{E}\left[a_k a_k^*\right]\right\|_\infty, \left\|\mathbb{E}\left[a_k a_k^*\right]\right\|_\infty^2 \right\} \leq \frac{2\sqrt{(n+1)n}}{m^2}.
\end{aligned}$$

Hence

$$\left\| \sum_{k=1}^m \mathbb{E}\left[M_k^2\right] \right\|_\infty \leq \frac{2\sqrt{(n+1)n}}{m}.$$

[5] Resorting to the matrix Chernoff inequality would allow for establishing a similar result. However, in the case of an exact tight frame, the numerical constants obtained by doing so are slightly worse.

These bounds allow us to set $R := \frac{\sqrt{(n+1)n}}{m}$, $\sigma^2 := \frac{2\sqrt{(n+1)n}}{m}$ and apply the matrix Bernstein inequality ([65, Theorem 6.1], [1]) in order to establish

$$\Pr\left[\left\|\sum_{k=1}^{m} M_k\right\|_\infty \geq \tau\right] \leq n \exp\left(-\frac{\tau^2/2}{\sigma^2+R\tau}\right) \leq n \exp\left(-\frac{3\tau^2 m}{16\sqrt{(n+1)n}}\right)$$

for $0<\tau \leq \sigma^2/R = 2$. Setting $\tau = 1/4$ and inserting $m \geq \tilde{C}_4 nr\log(n)$ (where $\tilde{C}_4$ is large enough) assures that (58) holds with probability of failure smaller than $\mathrm{e}^{-\gamma \tilde{C}_4 r}$ via (59) for a suitable $\gamma > 0$. □

Finally, we are ready to prove Theorem 4.

Proof of Theorem 4. We content ourselves with establishing the design case and point out that the Gaussian case can be proved analogously (albeit with different constants). Fix $0<\rho<1/8$ and suppose that

$$m \geq C_3\left(1+\left(1+\rho^{-1}\right)^2\right) nr\log n$$

measurement vectors have been chosen independently from an approximate 4-design. Theorem 28 then assures that the resulting measurement operator $\mathcal{A}$ obeys the robust Frobenius rank null space property with constants $\rho < 1/8$ and $\tau \leq \tilde{C}_6/\sqrt{m}$ with probability at least $1-\mathrm{e}^{-\tilde{C}_5 m}$. Likewise, Proposition 33 assures that with probability at least $1-\mathrm{e}^{-\gamma\tilde{C}_4 r}$, setting $t = \frac{1}{\sqrt{m}}(1,\ldots,1)^T \in \mathbb{R}^m$ leads to a positive definite $W = \sum_{j=1}^m t_j A_j$ obeying $\kappa(W) \leq 8$. Note that such a t obeys $\|t\|_{\ell_2} = 1/\sqrt{m}$ and also $0<\rho<1/8 \leq 1/\kappa(W)$ holds by construction. The union bound over these two assertions failing implies that the requirements of Theorem 31 are met with probability at least

$$1-\mathrm{e}^{-\tilde{C}_5 m} - \mathrm{e}^{-\gamma\tilde{C}_4 r} \geq 1-\mathrm{e}^{-\tilde{\gamma}\tilde{C}_4 r},$$

where $\tilde{\gamma}$ denotes a sufficiently small absolute constant and $\tilde{C}_4 = m/nr\log n$. The constants C_4 and s presented in Theorem 4 then amount to $s = \tilde{\gamma}\tilde{C}_4$ and $C_2 \geq \tilde{C}_4$. Inserting $\|t\|_{\ell_2} = 1/\sqrt{m}$ and the bounds on $\|W\|_\infty, \|W^{-1}\|_\infty, \kappa(W)$ from Proposition 33 into (51) yields

$$\begin{aligned}
\|Z-X\|_p &\leq \frac{2C\kappa(W)}{r^{1-1/p}}\|X_c\|_1 + r^{1/p-1/2}\|\mathcal{A}(Z)-\mathcal{A}(X)\|_{\ell_2}\|W^{-1}\|_\infty\left(\frac{C\|t\|_2}{\sqrt{r}} + D\|W\|_\infty \tau\right)\\
&\leq \frac{16C}{r^{1-1/p}}\|X_c\|_1 + 8r^{1/p-1/2}\|\mathcal{A}(Z)-\mathcal{A}(X)\|_{\ell_2}\left(\frac{C}{\sqrt{rm}} + \frac{9D\tilde{C}_6}{\sqrt{m}}\right)\\
&\leq \frac{C_3}{r^{1-1/p}}\|X_c\|_1 + \frac{C_4 r^{1/p-1/2}}{\sqrt{m}}\|\mathcal{A}(Z)-\mathcal{A}(X)\|_{\ell_2}
\end{aligned}$$

with constants $C_3 = 16C$ and $C_4 = 8C + 8D\tilde{C}_6$ (where C, D were introduced in Theorem 31 and $\tilde{C}_6$ is). □

Remark 34. In Corollary 6 we focus on recovering density operators, i.e., positive semidefinite matrices X with trace one. This trace constraint can be re-interpreted as an additional perfectly noiseless measurement

$$b_0 = \operatorname{tr}(\mathrm{id}\, X) = \operatorname{tr}(X) = 1$$

corresponding to the measurement matrix $A_0 = \mathrm{id}$. Setting $t = (1,0,\ldots,0)^T \in \mathbb{R}^{m+1}$ in Theorem 31 then leads to $W = \mathrm{id}$ which obeys $\|W\|_\infty = \|W^{-1}\|_\infty = \kappa(W) = 1$ and furthermore assures that the endomorphism (48) is trivial, i.e. $\tilde{Z} = Z$ for all $Z \in \mathcal{H}_n$. Moreover, these properties render the estimate provided in Lemma 30 redundant, because any two density operators X, Z obey

$$\|\tilde{Z}\|_1 - \|\tilde{X}\|_1 = \|Z\|_1 - \|Z\|_1 = \operatorname{tr}(Z) - \operatorname{tr}(X) = 0.$$

Such a refinement then allows for dropping the term containing $\|t\|_{\ell_2}$ in (51) and by inserting $W = \mathrm{id}$ we arrive at the following conclusion: Any measurement operator $\mathcal{A}$ that obeys the Frobenius robust rank null space property with constants $0<\rho<1$ and $\tau>0$ assures for $1 \leq p \leq 2$ and any two density operators X, Z:

$$\|Z-X\|_p \leq \frac{2(1+\rho)^2}{1-\rho}\|X_c\|_1 + \tau\frac{r^{1/p-1/2}(3+\rho)}{1-\rho}\|\mathcal{A}(Z)-\mathcal{A}(X)\|_{\ell_2}.$$

Corollary 6 then follows from combining this assertion with Theorem 28 and setting $p=1$.

Acknowledgements. MK, HR and UT acknowledge funding by the European Research Council through the Starting Grant StG 258926. The work of RK is supported by the Excellence Initiative of the German Federal and State Governments (Grants ZUK 43 & 81), the ARO under contracts W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), the Freiburg Research Innovation Fund, the DFG (GRO 4334 & SPP1798), and the State Graduate Funding Program of Baden-Württemberg.

Appendix

A brief review of finite-dimensional quantum mechanics. For the sake of being self-contained we briefly recapitulate crucial concepts of (finite dimensional) quantum mechanics without going too much into detail. For further reading on the topics introduced here, we defer the interested reader to [53, Chapter 2.2].

An isolated quantum mechanical system is fully described by its *density operator*. For a finite n-dimensional quantum system, such a density operator corresponds to an Hermitian, positive semidefinite matrix ρ with unit trace.

The most general notion of a measurement is that of a *positive operator-valued measure* (POVM). For an n-dimensional quantum system, a POVM corresponds to a collection $\mathcal{M} = \{E_m\}_{m \in I}$ of positive semidefinite $n \times n$ matrices that sum up to identity, i.e.,

$$\sum_{m \in I} E_m = \mathrm{id}\,.$$

The indices $m \in I$ indicate the possible measurement outcomes of performing such a POVM measurement. Upon performing $\mathcal{M}$ on a system described by ρ, quantum mechanics then postulates that the probability of obtaining the outcome (labeled by) m corresponds to

$$p(m,\rho) = \mathrm{tr}\left(E_m \rho\right).$$

Repeating the same measurement (i.e., preparing ρ and measuring $\mathcal{M}$) many times allows one to estimate the n probabilities $p(\lambda_i,\rho)$ ever more accurately.

Note that the definitions of ρ and $\mathcal{M}$ assure that $p(m,\rho)_{m \in I}$ is in fact a valid probability distribution. Indeed, $p(m,\rho) \geq 0$ follows from positive-semidefiniteness of both ρ and E_m. Unit trace of ρ assures proper normalization via

$$\sum_{m \in I} p(m,\rho) = \sum_{m \in I} \mathrm{tr}\left(E_m \rho\right) = \mathrm{tr}\left(\mathrm{id}\,\rho\right) = \mathrm{tr}(\rho) = 1.$$

References

[1] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569 –579, 2002.

[2] A. Ahmed and J. Romberg. Compressive multiplexing of correlated signals. *IEEE Trans. Inform. Theory*, 61(1):479–498, 2015.

[3] A. Ambainis and J. Emerson. Quantum t-designs: t-wise independence in the quantum world. In *22nd Annual IEEE Conference on Computational Complexity, Proceedings*, pages 129–140, 2007.

[4] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Inform. Inference*, 3(3):224–294, 2014.

[5] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin. Painless reconstruction from magnitudes of frame coefficients. *J. Fourier Anal. Appl.*, 15:488–501, 2009.

[6] K. Banaszek, M. Cramer, and D. Gross. Focus on quantum tomography. *New J. Phys.*, 15:125020, 2013.

[7] R. Bhatia. *Matrix analysis.* Springer, 1997.

[8] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge Univ. Press, 2004.

[9] A. M. Bruckstein, M. Elad, and M. Zibulevsky. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *IEEE Trans. Inform. Theory*, 54(11):4813–4820, 2008.

[10] E. Candès, T. Strohmer, and V. Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.*, 66:1241–1274, 2013.

[11] E. Candes and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010.

[12] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM J. Imaging Sci.*, 6:199–225, 2013.

[13] E. J. Candès and X. Li. Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. *Found. Comput. Math.*, pages 1–10, 2013.

[14] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015.

[15] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011.

[16] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9:717–772, 2009.

[17] A. Carpentier, J. Eisert, D. Gross, and R. Nickl. Uncertainty quantification for matrix compressed sensing and quantum tomography problems. *preprint arXiv:1504.03234*, 2015.

[18] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

[19] Y. Chen, Y. Chi, and A. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Trans. Inform. Theory*, 61(7):4034–4059, July 2015.

[20] Z. Chen and J. J. Dongarra. Condition numbers of Gaussian random matrices. *SIAM J. Matrix Anal. A.*, 27(3):603–620, 2005.

[21] L. Demanet and P. Hand. Stable optimizationless recovery from phaseless linear measurements. *J. Fourier Anal. Appl.*, 20(1):199–221, 2014.

[22] M. Fazel. Matrix rank minimization. *PhD thesis, Stanford University*, 2002.

[23] C. Ferrie and R. Kueng. Have you been using the wrong estimator? These guys bound average fidelity using this one weird trick von Neumann didn't want you to know. *preprint arXiv:1503.00677*, 2015.

[24] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert. Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New J. Phys.*, 14:095022, 2012.

[25] M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM J. Optim.*, 21(4):1614–1640, 2011.

[26] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing.* Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[27] Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric aspects of functional analysis (1986/87)*, volume 1317 of *Lecture Notes in Math.*, pages 84–106. Springer, Berlin, 1988.

[28] D. Gross, F. Krahmer, and R. Kueng. A partial derandomization of Phaselift using spherical designs. *Journal of Fourier Analysis and Applications*, pages 1–38, 2014.

[29] D. Gross, F. Krahmer, and R. Kueng. Improved recovery guarantees for phase retrieval from coded diffraction patterns. *to appear in Appl. Comput. Harmon. Anal., preprint arXiv:1402.6286*, 2015.

[30] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert. State tomography via compressed sensing. *Phys. Rev. Lett.*, 105:150401, 2010.

[31] T. Heinosaari, L. Mazzarella, and M. M. Wolf. Quantum tomography under prior information. *Commun. Math. Phys.*, 318(2):355–374, 2013.

[32] C. W. Helstrom. Quantum detection and estimation theory. *J. Statist. Phys.*, 1(2):231–252, 1969.

[33] R. Horn and C. Johnson. *Topics in matrix analysis.* Cambridge University Press, Cambridge, 1991.

[34] M. Kabanava, H. Rauhut, and U. Terstiege. Analysis of low rank matrix recovery via mendelson's small ball method. In *11th international conference on Sampling Theory and Applications (SampTA 2015)*, Washington, USA, May 2015.

[35] M. Kabanava, H. Rauhut, and U. Terstiege. On the minimal number of measurements in low-rank matrix recovery. In *11th international conference on Sampling Theory and Applications (SampTA 2015)*, Washington, USA, May 2015.

[36] A. Kalev, C. Riofrio, R. Kosut, and I. Deutsch. Informationally complete measurements from compressed sensing methodology. *Bulletin of the American Physical Society*, 60, 2015.

[37] M. Kech, P. Vrana, and M. Wolf. The role of topology in quantum tomography. *preprint arXiv:1503.00506*, 2015.

[38] M. Kech and M. Wolf. From quantum tomography to phase retrieval and back. In *11th international conference on Sampling Theory and Applications (SampTA 2015)*, Washington, USA, May 2015.

[39] M. Kech and M. M. Wolf. Quantum tomography of semi-algebraic sets with constrained measurements. *preprint ArXiv:1507.00903*, 2015.

[40] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *Internat. Math. Res. Notices*, page rnv096, 2015.

[41] R. Kueng, S. Daniel, and D. Gross. Direct characterization of linear-optical networks via PhaseLift. in preparation, 2015.

[42] R. Kueng, D. Gross, and F. Krahmer. Spherical designs as a tool for derandomization: The case of PhaseLift. In *11th international conference on Sampling Theory and Applications (SampTA 2015)*, Washington, USA, May 2015.

[43] R. Kueng, H. Rauhut, and U. Terstiege. Low rank matrix recovery from rank one measurements. *Preprint arXiv:1410.6913*, 2014.

[44] R. Latała. Some estimates of norms of random matrices. *Proc. Amer. Math. Soc.*, 133(5):1273–1282, 2005.

[45] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer-Verlag, 1991.

[46] K. Lee and Y. Bresler. ADMiRA: Atomic decomposition for minimum rank approximation. *IEEE Trans. Image Process.*, 56(9):4402 – 4416, 2010.

[47] Y.-K. Liu. Universal low-rank matrix recovery from Pauli measurements. *Adv. Neural Inf. Process. Syst.*, pages 1638–1646, 2011.

[48] W. Matthews, S. Wehner, and A. Winter. Distinguishability of quantum states under restricted families of measurements with an application to quantum data hiding. *Commun. Math. Phys.*, 291(3):813–843, 2009.

[49] M. B. McCoy and J. A. Tropp. Sharp recovery bounds for convex demixing, with applications. *Found. Comput. Math.*, 14(3):503–567, 2014.

[50] S. Mendelson. Learning without Concentration. *J. ACM*, 62(3):1–25, 2015.

[51] K. Mohan and M. Fazel. Iterative reweighted least squares for matrix rank minimization. In *Proceedings of the Allerton Conference*, pages 653–661, 2010.

[52] K. Mohan and M. Fazel. New restricted isometry results for noisy low-rank recovery. In *Proc. International Symposium Information Theory*, 2010.

[53] M. A. Nielsen and I. L. Chuang. *Quantum computation and quantum information.* Cambridge University Press, 2010.

[54] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi. A simplified approach to recovery conditions for low rank matrices. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2318–2322, July 2011.

[55] N. Parikh and S. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):123–231, 2014.

[56] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010.

[57] B. Recht, W. Xu, and B. Hassibi. Necessary and sufficient conditions for success of the nuclear norm heuristic for rank minimization. In *Proc. 47th IEEE Conference on Decision and Control*, pages 3065–3070, 2008.

[58] B. Recht, W. Xu, and B. Hassibi. Null space conditions and thresholds for rank minimization. *Mathematical Programming*, Ser B, 127:175–211, 2011.

[59] M. Rudelson and R. Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, 2008.

[60] P. Schindler, D. Nigg, T. Monz, J. T. Barreiro, E. Martinez, S. X. Wang, S. Quint, M. F. Brandl, V. Nebendahl, C. F. Roos, et al. A quantum information processor with trapped ions. *New J. Phys.*, 15(12):123012, 2013.

[61] B. Schölkopf and A. J. Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond.* MIT press, 2002.

[62] A. Scott. Tight informationally complete quantum measurements. *J. Phys. A-Math. Gen.*, 39:13507–13530, 2006.

[63] M. Slawski, P. Li, and M. Hein. Regularization-free estimation in trace regression with symmetric positive semidefinite matrices. *preprint arXiv:1504.06305*, 2015.

[64] J. Tanner and K. Wei. Normalized iterative hard thresholding for matrix completion. *SIAM J. Sci. Comput.*, 59(11):7491–7508, 2013.

[65] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[66] J. A. Tropp. Convex recovery of a structured signal from independent random linear measurements. *to appear in Sampling Theory, a Renaissance, preprint arXiv:1405.1102*, 2014.

[67] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge Univ Press, 2012.

[68] Z. Xu. The minimal measurement number for low-rank matrices recovery. *preprint ArXiv:1505.07204*, 2015.

Improved Recovery Guarantees for Phase Retrieval from Coded Diffraction Patterns

D. Gross[1,2], F. Krahmer[3], R. Kueng[*2,4]

[1]Institute for Theoretical Physics, University of Cologne
[2]Institute for Physics & FDM, University of Freiburg
[3]Institute for Numerical and Applied Mathematics, University of Göttingen
[4]ARC Centre for Engineered Quantum Systems, School of Physics, The University of Sydney

January 27, 2016

ABSTRACT. In this work we analyze the problem of phase retrieval from Fourier measurements with random diffraction patterns. To this end, we consider the recently introduced PhaseLift algorithm, which expresses the problem in the language of convex optimization. We provide recovery guarantees which require $\mathcal{O}(\log^2 d)$ different diffraction patterns, thus improving on recent results by Candès et al. [1], which require $\mathcal{O}(\log^4 d)$ different patterns.

1. INTRODUCTION

1.1. **The problem of phase retrieval.** In this work we are interested in the problem of *phase retrieval* which is of considerable importance in many different areas of science, where capturing phase information is hard or even infeasible. Problems of this kind occur, for example, in X-ray crystallography, diffraction imaging, and astronomy.

More formally, *phase retrieval* is the problem of recovering an unknown complex vector $x \in \mathbb{C}^d$ from *amplitude* measurements

$$y_i = |\langle a_i, x\rangle|^2 \quad i = 1, \ldots, m, \tag{1}$$

for a given set of measurement vectors $a_1, \ldots, a_m \in \mathbb{C}^d$. The observations y are insensitive to a global phase change $x \mapsto e^{i\phi}x$ – hence in the following, notions like “recovery” or “injectivity” are always implied to mean “up to a global phase”. Clearly, the most fundamental question is: Which families of measurement vectors $\{a_i\}$ allow for a recovery of x in principle? I.e., for which measurements is the map $x \mapsto y$ defined by (1) injective?

Approaches based on algebraic geometry (for example [2, 3]) have established that for determining x, $4d + o(1)$ *generic* measurements are sufficient and $4d - \mathcal{O}(\log d)$ such observations are necessary. Here, “generic” means that the measurement ensembles for which the property fails to hold lie on a low-dimensional subvariety of the algebraic variety of all tight measurement frames.

This notion of generic success, however, is mainly of theoretical interest. Namely, injectivity alone neither gives an indication on how to recover the unique solution, nor is there any chance to directly generalize the results to the case of noisy measurements. It should be noted, however, that recently the notion of injectivity has been refined to capture aspects of stability with respect to noise [4].

Paralleling these advances, there have been various attempts to find tractable recovery algorithms that yield recovery guarantees. Many of these approaches are based on a linear reformulation in matrix space, which is well-known in convex programming. The crucial

[*]Corresponding author: richard.kueng@physik.uni-freiburg.de

underlying observation is that the quadratic constraints (1) on x are linear in the outer product $X = xx^*$:

$$y_i = |\langle a_i, x\rangle|^2 = \operatorname{tr}\left((a_i a_i^*)X\right).$$

Balan et al. [5] observed that for the right choice of d^2 measurement vectors a_i, this linear system in the entries of X admits for a unique solution, so the problem can be explicitly solved using linear algebra techniques. This approach, however, does not make use of the low-rank structure of X, which is why the required number of measurements is so much larger than what is required for injectivity.

The *PhaseLift* algorithm proposed by Candès et al. [6, 7, 8] uses in addition the property that X is of rank one, so even when the number of measurements is smaller than d^2 and there is an entire affine space of matrices satisfying (1.1), X is the solution of smallest rank. While finding the smallest rank solution of a linear system is, in general, NP hard, there are a number of algorithms known to recover the smallest rank solution provided the system satisfies some regularity conditions. The first such results were based on convex relaxation (see, for example, [9, 10, 11]). PhaseLift is also based on this strategy. For measurement vectors drawn independently at random from a Gaussian distribution, the number of measurements required to guarantee recovery with high probability was shown to be of optimal order, scaling linearly in the dimension [7, 8] – see also [12] for a comparable statement valid for recovering matrices of arbitrary rank. A generalized version of this result—valid for projective measurements onto random subspaces rather than random vectors—was established in [13]. Moreover, Ref. [14] even identifies a deterministic, explicitly engineered set of $4d - 4$ measurement vectors and proves that PhaseLift will successfully recover generic signals from the associated measurements. Conversely, any complex vector is uniquely determined by $4d - 4$ generic phaseless measurements [15].

Since these first recovery guarantees for the phase retrieval problem, recovery guarantees have been proved for a number of more efficient algorithms closer to the heuristic approaches typically used in practice. For example, in [16], an approach based on polarization is analyzed and in [17], the authors study an alternating minimization algorithm. In both works, recovery guarantees are again proved for Gaussian measurements. Further numerical approaches have been proposed and studied in [18].

To relate all these results to practice, the structure of applications needs to be incorporated into the setup, which corresponds to reducing randomness and considering structured measurements. For PhaseLift, the first partial derandomization has been provided by the authors of this paper, considering measurements sampled from spherical designs, that is, polynomial-size sets which generalize the notion of a tight frame to higher-order tensors [19]. Recently, this result has been considerably improved in [12]. Arguably, these derandomized measurement setups are still mainly of theoretical interest.

A structured measurement setup closer to applications is that of coded diffraction patterns. These correspond to the composition of diagonal matrices and the Fourier transform and model the modified application setup where diffraction masks are placed between the object and the screen as originally proposed in [20]. The first recovery guarantees from masked Fourier measurements were provided for polarization based recovery [21], where the design of the masks is very specific and intimately connected to the recovery algorithm. The required number of masks is $\mathcal{O}(\log d)$, which corresponds to $\mathcal{O}(d \log d)$ measurements.

For the PhaseLift algorithm, recovery guarantees from masked Fourier measurements were first provided in [1]. The results require $\mathcal{O}(d \log^4 d)$ measurements and hold with high probability when the masks are chosen at random, which is in line with the observation from [20] that random diffraction patterns are particularly suitable.

In this paper, we consider the same measurement setup as [1], but improve the bound on the required number of measurements to $\mathcal{O}(d \log^2 d)$.

2. Problem Setup and Main Results

2.1. **Coded diffraction patterns.** As in [1], we will work with the following setup:

In every step, we collect the magnitudes of the discrete Fourier transform (DFT) of a random modulation of the unknown signal x. Each such modulation pattern is modeled by a random diagonal matrix. More formally, for $\omega := \exp\left(\frac{2\pi i}{d}\right)$ a d-th root of unity and $\{e_1, \ldots, e_d\}$ the standard basis of $\mathbb{C}^d$, denote by

$$f_k = \sum_{j=1}^{d} \omega^{jk} e_j \tag{2}$$

the k-th discrete Fourier vector, normalized so that each entry has unit modulus. Furthermore, consider the diagonal matrix

$$D_l = \sum_{i=1}^{d} \epsilon_{l,i} e_i e_i^* \tag{3}$$

where the $\epsilon_{l,i}$'s are independent copies of a real-valued[2] random variable ϵ which obeys

$$\begin{aligned} \mathbb{E}[\epsilon] &= \mathbb{E}[\epsilon^3] = 0, \\ |\epsilon| &\leq b \quad \text{almost surely for some } b > 0, \qquad (4)\\ \mathbb{E}[\epsilon^4] &= 2\,\mathbb{E}[\epsilon^2]^2 \quad \text{and we define} \quad \nu := \mathbb{E}\left[\epsilon^2\right]. \qquad (5) \end{aligned}$$

Then the measurements are given by

$$y_{k,l} = |\langle f_k, D_l x\rangle|^2 \quad 1 \leq k \leq d, \quad 1 \leq l \leq L. \tag{6}$$

It turns out (Lemma 7 below) that condition (5) on ϵ ensures that the measurement ensemble forms a spherical 2-design, which draws a connection to [5] and [19].

As an example, the criteria above include the model

$$\epsilon \sim \begin{cases} \sqrt{2} & \text{with prob. } 1/4, \\ 0 & \text{with prob. } 1/2, \\ -\sqrt{2} & \text{with prob. } 1/4. \end{cases} \tag{7}$$

which has been discussed in [1]. In this case, each modulation is given by a Rademacher vector with random erasures.

2.2. **Convex Relaxation.** Following [5], we rewrite the measurement constraints as the inner product of two rank 1 matrices, one representing the signal, the other one the measurement coefficients. In the coded diffraction setup, we obtain, as in [1], that the inner product of (6) can be translated into matrix form by applying the following "lifts":

$$X := xx^* \quad \text{and} \quad F_{k,l} := D_l f_k f_k^* D_l.$$

[2] Ref. [1] also included a strongly related model where ϵ is a complex random variable. We have opted to keep ϵ real, which implies that the D_l are hermitian. This, in turn, has allowed us to slightly simplify notation throughout.

Occasionally, we will make use of the representation with respect to the standard basis, which reads

$$F_{k,l} = \sum_{i,j=1}^{d} \epsilon_{l,i}\epsilon_{l,j}\omega^{k(i-j)} e_i e_j^*. \tag{8}$$

With these definitions, the dL individual linear measurements assume the following form

$$y_{k,l} \quad = \quad \mathrm{Tr}\,(F_{k,l}X) \quad k = 1,\ldots,d,\ 1 \leq l \leq L.$$

and the phase retrieval problem thus becomes the problem of finding rank 1 solutions $X = xx^*$ compatible with these affine constraints. Rank-minimization over affine spaces is NP-hard in general. However, it is now well-appreciated [9, 10, 11, 7] that nuclear-norm based convex relaxations solve this problems efficiently in many relevant instances. Applied to phase retrieval, the relaxation becomes

$$\begin{aligned} \text{argmin}_{X'} \quad & \|X'\|_1 \\ \text{subject to} \quad & \mathrm{tr}\,(F_{k,l}X') = y_{k,l} \quad k = 1,\ldots n,\ 1 \leq l \leq L, \\ & X' = (X')^* \\ & X' \geq 0, \end{aligned} \tag{9}$$

which has been dubbed *Phaselift* by its inventors [6, 7, 8]. For this convex relaxation, recovery guarantees are known for measurement vectors drawn i.i.d. at random from a Gaussian distribution [7, 8], t-designs [19, 12], or in the masked Fourier setting [1].

We want to point out that access to additional information can considerably simplify Phaselift. In particular, knowledge of the signal's *intensity* $y_0 = \|x\|_{\ell_2}^2$ results in an additional trace constraint which together with $X' \geq 0$ implies $\|X'\|_1 = y_0$ for any feasible X'. Consequently, minimizing the nuclear norm becomes redundant and (9) can be replaced by the feasibility problem

$$\begin{aligned} \text{find} \quad & X' \\ \text{subject to} \quad & \mathrm{tr}\,(F_{k,l}X') = y_{k,l} \quad k = 1,\ldots n,\ 1 \leq l \leq L, \\ & X' = (X')^* \\ & \mathrm{tr}(X') = y_0, \\ & X' \geq 0. \end{aligned} \tag{10}$$

2.3. **Our contribution.** In this paper, we adopt the setup from [1]. Our main message is that recovery of x can be guaranteed already for

$$L \geq C \log^2 d$$

random diffraction patterns, provided that the signal's intensity $y_0 = \|x\|_{\ell_2}^2$ is known[3]. This improves the bound given in [1] by a factor of $\mathcal{O}(\log^2 d)$. It is significant, as it indicates that the provably achievable rates are approaching the ultimate limit. Indeed, for the Rademacher masks with random erasures introduced above, a lower bound for the number of diffraction patterns required to allow for recovery with any algorithm is given by $\mathcal{O}(\log d)$. This follows from a standard coupon collector's argument similar to the ones provided in [10, 11]. For completeness, the lower bound is precisely formulated and proved in Lemma 19 in the appendix.

[3] This can, for instance, be achieved by starting the measurement process with a trivial modulation pattern—i.e. D_0 corresponds to the identity matrix—and summing up the d corresponding measurements (6).

Thus there cannot be a recovery algorithm requiring fewer than $O(\log d)$ masks and there is only a single $\log$-factor separating our results from an asymptotically tight solution.

More precisely, our version of [1, Theorem 1.1] reads:

Theorem 1 (Main Theorem). *Let $x \in \mathbb{C}^d$ be an unknown signal with $\|x\|_{\ell_2} = 1$ and let $d \geq 3$ be an odd number. Suppose that L complete Fourier measurements using independent random diffraction patterns (as defined in Section 2.1) are performed.*

Then with probability at least $(1 - \mathrm{e}^{-\omega})$ Phaselift (the convex optimization problem (9) *endowed with the additional constraint $\mathrm{tr}(X') = 1$, or the feasibility problem* (10)*) recovers x up to a global phase, provided that*

$$L \geq C\omega \log^2 d.$$

Here, $\omega \geq 1$ is an arbitrary parameter and C a dimension-independent constant that can be explicitly bounded.

The number C is of the form $C = \tilde{C}\frac{b^8}{\nu^4} \log_2^2 \left(b^2/\nu\right)$, where b and ν were defined in (4) and (5), respectively. Also, $\tilde{C}$ an absolute constant for which an explicit estimate can be extracted from our proof.

For the benefit of the technically-minded reader, we briefly sketch the relation between the proof techniques used here, as compared to References [1] and [19].

- The general structure of this document closely mimics [19] (which bears remarkable similarity to [1], even though the papers were written completely independently and with different aims in mind).
- From [1] we borrow the use of Hoeffding's inequality to bound the probability of "the inner product between the measurement vectors and the signal becoming too large". This is Lemma 13 below. Our previous work also bounded the probability of such events [19, Lemma 13]—however in a weaker way (relying only on certain tth moments as opposed to a Hoeffding bound).
- Both [19, 1] as well as the present paper estimate the condition number of the measurement operator restricted to the tangent space at xx^* ("robust injectivity"). Our Proposition 8 improves over [1, Section 3.3] by using an operator Bernstein inequality instead of a weaker operator Hoeffding bound.
- Finally, we use a slightly refined version of the golfing scheme to construct an approximate dual certificate (following [11, Section III.B]).

2.4. **More general bases and outlook.** The result allows for a fairly general distribution of the masks D_l, but refers specifically to the Fourier basis. An obvious question is how sensitively the statements depend on the properties of this basis.

We begin by pointing out that Theorem 1 immediately implies a corollary for higher-dimensional Fourier transforms. In diffraction imaging applications, for example, one would naturally employ a 2-D Fourier basis

$$f_{k,l} = \sum_{i=1}^{d_x} \sum_{j=1}^{d_y} \omega_{d_x}^{ik} \omega_{d_y}^{jl} e_{i,j}, \tag{11}$$

with d_x and d_y the horizontal and vertical resolution respectively, $\omega_d := \exp\left(\frac{2\pi i}{d}\right)$, and $e_{i,j}$ the position space basis vector representing a signal located at coordinates (i, j). Superficially, (11) looks quite different from the one-dimensional case (2). However, a basic application of the Chinese Remainder Theorem shows that if d_x and d_y are co-prime, then the 2-D transform reduces to the 1-D one for dimension $d_x d_y$ (in the sense that the respective bases agree up to relabeling) [22]. An analogous result holds for higher-dimensional transforms [22], proving the following corollary.

Corollary 2. *Assume $d = \prod_{i=1}^{k} d_i$ is the product of mutually co-prime odd numbers greater than 3. Then Theorem 1 remains valid for the k-dimensional Fourier transform over $d_1, \dots, d_k$.*

More generally speaking, our argument employs the particular properties of Fourier bases in two places: Lemma 7 and Lemma 9.

The former lemma shows that the measurements are drawn from an *isotropic ensemble* (or *tight frame*) in the relevant space of hermitian matrices. A similar condition is frequently used in works on phase retrieval, low-rank matrix completion, and compressed sensing (e.g. [19, 1, 23, 24, 11]). Properties of the Fourier basis are used in the proof of Lemma 7 only for concreteness. Using relatively straight-forward representation theory, one can give a far more abstract version of the result which is valid for any basis satisfying two explicit polynomial relations (cf. the remark below the lemma). The combinatorial structure of Fourier transforms is immaterial at this point.

This contrasts with Lemma 9 which currently prevents us from generalizing the main result to a broader class of bases. Its proof uses explicit coordinate expressions of the Fourier basis to facilitate a series of simplifications. Identifying the abstract gist of the manipulations is the main open problem which we hope to address in future work.

We make use of the condition that d be odd only for Lemma 7. While that particular Lemma fails to hold for even dimensions, we find it plausible that the result as a whole remains essentially true for even dimensions.

It would also be interesting to use the techniques of the present paper to re-visit the problem of quantum state tomography [25, 26, 27, 28] (which was the initial motivation for one of the authors to become interested in low-rank recovery methods). Indeed, the original work on quantum state tomography and low-rank recovery [25] was based on a model where the expectation value of a Pauli matrix is the elementary unity of information exctractable from a quantum experiment. While this correctly describes some experiments, it is arguably more common that the statistics of the eigenbasis of an observable are the objects that can be physically directly accessed. For this practically more relevant case, no recovery guarantees seem to be currently known and the methods used here could be used to amend that situation.

3. Technical Background and Notation

3.1. **Vectors, Matrices, and matrix valued Operators.** The signals x are assumed to live in $\mathbb{C}^d$ equipped with the usual inner product $\langle \cdot, \cdot \rangle$. We denote the induced norm by

$$\|z\|_{\ell_2} = \sqrt{\langle z, z \rangle} \quad \forall z \in \mathbb{C}^d.$$

Vectors in $\mathbb{C}^d$ will be denoted by lower case Latin characters. For $z \in \mathbb{C}^d$ we define the absolute value $|z| \in \mathbb{R}_+^d$ component-wise $|z|_i = |z_i|$.

On the level of matrices we will exclusively encounter $d \times d$ hermitian matrices and denote them by capital Latin characters. Endowed with the Hilbert-Schmidt (or Frobenius) scalar product

$$(Z, Y) = \mathrm{tr}(ZY) \tag{12}$$

the space H^d of all $d \times d$ hermitian matrices becomes a Hilbert space itself. In addition to that, we will require three different operator norms

$$\begin{aligned}
\|Z\|_1 &= \operatorname{tr}(|Z|) \quad \text{(trace or nuclear norm)}, \\
\|Z\|_2 &= \sqrt{\operatorname{tr}(Z^2)} \quad \text{(Frobenius norm)}, \\
\|Z\|_\infty &= = \sup_{y \in \mathbb{C}^d} \frac{|\langle y, Zy \rangle|}{\|y\|_{\ell_2}^2} \quad \text{(operator norm)}.
\end{aligned} \tag{13}$$

In the definition of the trace norm, $|Z|$ denotes the unique positive semidefinite matrix obeying $|Z|^2 = Z^2$ (or equivalently $|Z| = \sqrt{Z^2}$ which is unique). For arbitrary matrices Z of rank at most r, the norms above are related via the inequalities

$$\|Z\|_2 \leq \|Z\|_1 \leq \sqrt{r}\|Z\|_2 \quad \text{and} \quad \|Z\|_\infty \leq \|Z\|_2 \leq \sqrt{r}\|Z\|_\infty.$$

Recall that a hermitian matrix Z is positive semidefinite if one has $\langle y, Zy \rangle \geq 0$ for all $y \in \mathbb{C}^d$. We write $Y \geq Z$ iff $Y - Z$ is positive semidefinite.

In this work, hermitian rank-1 projectors are of particular importance. They are of the form $Z = zz^*$ with $z \in \mathbb{C}^d$. The vector z can then be recovered from Z up to a global phase factor via the singular value decomposition. In this work, the most prominent rank-1 projectors are $X = xx^*$ and $F_{k,l} = D_l f_k (D_l f_k)^*$.

Finally, we will also encounter *matrix-valued operators* acting on the matrix space H^d. Here, we will restrict ourselves to operators that are hermitian with respect to the Hilbert-Schmitt inner product. We label such objects with calligraphic letters. The operator norm becomes

$$\|\mathcal{M}\|_{\mathrm{op}} = \sup_{Z \in H^d} \frac{|\operatorname{tr}(Z\mathcal{M}Z)|}{\|Z\|_2^2}. \tag{14}$$

It turns out that only two classes of such operators will appear in our work, namely the identity map

$$\begin{aligned}
\mathcal{I} : H^d &\to H^d \\
Z &\mapsto Z \quad \forall Z \in H^d
\end{aligned}$$

and (scalar multiples of) projectors onto some matrix $Y \in H^d$ as given by

$$\begin{aligned}
\Pi_Y : H^d &\to H^d \\
Z &\mapsto Y(Y, Z) = Y \operatorname{tr}(YZ) \quad \forall Z \in H^d.
\end{aligned}$$

An important example of the latter class is

$$\Pi_{\mathbb{1}} : \ Z \mapsto \mathbb{1} \operatorname{tr}(\mathbb{1} Z) = \operatorname{tr}(Z)\mathbb{1} \quad \forall Z \in H^d.$$

Note that the normalization is such that $\frac{1}{d}\Pi_{\mathbb{1}}$ is idempotent, i.e. a properly normalized projection. Indeed, for $Z \in H^d$ arbitrary it holds that

$$(d^{-1}\Pi_{\mathbb{1}})^2 Z = d^{-2}\mathbb{1} \operatorname{tr}(\mathbb{1}\Pi_{\mathbb{1}} Z) = d^{-2} \operatorname{tr}(\mathbb{1}) \operatorname{tr}(Z)\mathbb{1} = d^{-1}\Pi_{\mathbb{1}} Z. \tag{15}$$

The notion of positive-semidefiniteness directly translates to matrix valued operators. It is easy to check that all the operators introduced so far are positive semidefinite. From (15) we obtain the ordering

$$0 \leq \Pi_{\mathbb{1}} \leq d\mathcal{I}. \tag{16}$$

3.2. **Tools from Probability Theory.** In this section, we recall some concentration inequalities which will prove useful for our argument. Our first tool is a slight extension of Hoeffding's inequality [29].

Theorem 3. *Let $z = (z_1, \dots, z_d) \in \mathbb{C}^d$ be an arbitrary vector and let ϵ_i, $i = 1, \dots d$, be independent copies of a real-valued, centered random variable ϵ which is almost surely bounded in modulus by $b > 0$. Then*

$$\mathrm{Pr}\left[\left|\sum_{i=1}^{d} \epsilon_i z_i\right| \geq t\|z\|_{\ell_2}\right] \leq 4\exp\left(-t^2/(4b^2)\right). \tag{17}$$

One way to prove this statement, is to split up z into $x + iy$ with $x, y \in \mathbb{R}^d$ and noting that $\|z\|_{\ell_2} \geq \left(\|x\|_{\ell_2} + \|y\|_{\ell_2}\right)/\sqrt{2}$ holds. Splitting up the sum into real and imaginary parts, applying the triangle inequality and bounding $\mathrm{Pr}\left[\left|\sum_{i=1}^{d} \epsilon_i x_i\right| \geq t\|x\|_{\ell_2}/\sqrt{2}\right]$ and $\mathrm{Pr}\left[\left|\sum_{i=1}^{d} \epsilon_i y_i\right| \geq t\|y\|_{\ell_2}/\sqrt{2}\right]$ individually by means of Hoeffding's inequality (or a slightly generalized version of [30, Corllary 7.21]) then establishes (17) via the union bound.

Secondly, we will require two matrix versions of Bernstein's inequality. Such matrix valued large deviation bounds have been established first in the field of quantum information by Ahlswede and Winter [31] and introduced to sparse and low-rank recovery in [25, 11]. We make use of refined versions from [32, 33], see also [30, Chapter 8.5] for the former. Note that as H^d is a finite dimensional vector space, the results also apply to matrix valued operators as introduced in section 3.1.

Theorem 4 (Uniform Operator Bernstein inequality, [32, 11])**.** *Consider a finite sequence $\{M_k\}$ of independent random self-adjoint matrices. Assume that each M_k satisfies $\mathbb{E}\left[M_k\right] = 0$ and $\|M_k\|_\infty \leq \overline{R}$ (for some finite constant $\overline{R}$) almost surely. Then with the variance parameter $\sigma^2 := \|\sum_k \mathbb{E}\left[M_k^2\right]\|_\infty$, the following chain of inequalities holds for all $t \geq 0$.*

(18)

$$\mathrm{Pr}\left[\Big\|\sum_k M_k\Big\|_\infty \geq t\right] \leq d\ \exp\left(-\frac{t^2/2}{\sigma^2 + \overline{R}t/3}\right) \leq \begin{cases} d\ \exp(-3t^2/8\sigma^2) & t \leq \sigma^2/\overline{R} \\ d\ \exp(-3t/8\overline{R}) & t \geq \sigma^2/\overline{R}. \end{cases}$$

Theorem 5 (Smallest Eigenvalue Bernstein Inequality, [33])**.** *Let $S = \sum_k M_k$ be a sum of i.i.d. random matrices M_k which obey $\mathbb{E}\left[M_K\right] = 0$ and $\lambda_{min}(M_k) \geq -\underline{R}$ almost surely for some fixed $\underline{R}$. With the variance parameter $\sigma^2(S) = \|\sum_k \mathbb{E}\left[M_k^2\right]\|_\infty$ the following chain of inequalities holds for all $t \geq 0$.*

$$\mathrm{Pr}\left[\lambda_{\min}(S) \leq -t\right] \leq d\exp\left(-\frac{t^2/2}{\sigma^2 + \underline{R}t/3}\right) \leq \begin{cases} d\ \exp(-3t^2/8\sigma^2) & t \leq \sigma^2/\underline{R} \\ d\ \exp(-3t/8\underline{R}) & t \geq \sigma^2/\underline{R}. \end{cases}$$

Finally, we are also going to require a type of vector Bernstein inequality. Note that, since H^d is a d^2-dimensional real vector space, the statement remains valid for a sum of random hermitian matrices.

Theorem 6 (Vector Bernstein inequality)**.** *Consider a finite sequence $\{M_k\}$ of independent random vectors. Assume that each M_k satisfies $\mathbb{E}\left[M_k\right] = 0$ and $\|M_k\|_2 \leq B$ (for some finite constant B) almost surely. Then with the variance parameter $\sigma^2 := \sum_k \mathbb{E}\left[\|M_k\|_2^2\right]$,*

$$\mathrm{Pr}\left[\Big\|\sum_k M_k\Big\|_2 \geq t\right] \leq \exp\left(-\frac{t^2}{4\sigma^2} + \frac{1}{4}\right)$$

holds for any $t \leq \sigma^2/B$.

This particular vector-valued Bernstein inequality is based on the exposition in [34, Chapter 6.3, equation (6.12)] and a direct proof can be found in [11].

4. Proof Ingredients

4.1. **Near-isotropicity.** In this section we study the *measurement operator*[4]

$$\mathcal{R} : H^d \rightarrow H^d, \quad \mathcal{R} := \sum_{l=1}^{L} \mathcal{M}_l \quad \text{with} \tag{19}$$

$$\mathcal{M}_l Z := \frac{1}{\nu^2 dL} \sum_{k=1}^{d} \Pi_{F_{k,l}} Z = \frac{1}{\nu^2 dL} \sum_{k=1}^{d} F_{k,l} \operatorname{tr}(F_{k,l} Z), \tag{20}$$

which just corresponds to $\mathcal{R} = \frac{1}{\nu^2 dL} \mathcal{A}^* \mathcal{A}$, where ν was defined in (5).

The following result shows that this operator is *near-isotropic* in the sense of [19, 6].

Lemma 7 ($\mathcal{R}$ is *near-isotropic*)**.** *The operator $\mathcal{R}$ defined in (19) is* near-isotropic *in the sense that*

$$\mathbb{E}[\mathcal{R}] = L\mathbb{E}\left[\mathcal{M}_l\right] = \mathcal{I} + \Pi_{\mathbb{1}} \quad \textit{or} \quad \mathbb{E}\left[\mathcal{R}(Z)\right] = Z + \operatorname{tr}(Z)\mathbb{1} \quad \forall Z \in H^d. \tag{21}$$

A proof of Lemma 7 can be found in [1]. However, we still present a proof – which is of a slightly different spirit – in the appendix for the sake of being self-contained.

Two remarks are in order with regard to the previous lemma.

First, it is worthwhile to point out that *near-isotropicity* of $\mathcal{R}$ is equivalent to stating that the set of all possible realizations of $D_l f_k$ form a 2-design. This has been made explicit recently in [35, Lemma 1]. The notion of higher-order spherical designs is the basic mathematical object of our previous work [19] on phase retrieval.

Second, our proof of Lemma 7 uses the explicit representation of the measurement vectors with respect to the standard basis. As alluded to in Section 2.4, a more abstract proof can be given. We sketch the basic idea here and refer the reader to an upcoming work for details [36]. Consider the case where ϵ is a symmetric random variable (i.e., where ϵ has the same distribution as $-\epsilon$). In that case, the distribution of the D_l is plainly invariant under permutations of the main diagonal elements and under element-wise sign changes. These are the symmetries of the d-cube. They constitute the group $\mathbb{Z}_2^d \rtimes S_d$, sometimes refered to as the *hyperoctahedral group*. Using a standard technique [37, 38], conditions for near-isotropicity can be phrased in terms of the representation theory of the hyperoctahedral group acting on $\operatorname{Sym}^2(\mathbb{C}^d)$. This action decomposes into three explicitely identifiable irreducible components, from which one can deduce that near-isotropicity holds for any basis that fulfillls two 4th order polynomial equations [36].

Let now $x \in \mathbb{C}^d$ be the signal we aim to recover. Since the intensity of x (i.e., its ℓ_2-norm) is known by assumption, we can w.l.o.g. assume that $\|x\|_{\ell_2} = 1$. As in [7, 19, 1] we consider the space

$$T := \left\{ xz^* + zx^* : \ z \in \mathbb{C}^d \right\} \subset H^d \tag{22}$$

which is the tangent space of the manifold of all rank-1 hermitian matrices at the point $X = xx^*$. The orthogonal projection onto this space can be given explicitly:

$$\begin{aligned} \mathcal{P}_T : H^d &\rightarrow H^d \\ Z &\mapsto XZ + ZX - XZX && (23) \\ &= XZ + ZX - \operatorname{tr}(XZ)X. && (24) \end{aligned}$$

[4] We are going to use the notations $\mathcal{M}(Z)$ and $\mathcal{M}Z$ equivalently.

The Frobenius inner product allows us to define an ortho-complement $T^\perp$ of T in H^d. We denote the projection onto $T^\perp$ by $\mathcal{P}_T^\perp$ and decompose any matrix $Z \in H^d$ as

$$Z = \mathcal{P}_T Z + \mathcal{P}_T^\perp Z =: Z_T + Z_T^\perp.$$

We point out that, in particular,

$$\mathcal{P}_T \Pi_{\mathbb{1}} \mathcal{P}_T = \Pi_X \quad \text{and} \quad \|\mathcal{P}_T Z\|_\infty \leq 2\|Z\|_\infty \tag{25}$$

holds for any $Z \in H^d$. The first fact follows by direct calculation, while the second one comes from

$$\|Z_T\|_\infty = \|Z - Z_T^\perp\|_\infty \leq \|Z\|_\infty + \|Z_T^\perp\|_\infty \leq 2\|Z\|_\infty$$

where the last estimate used the pinching inequality [39] (Problem II.5.4).

4.2. **Well-posedness/Injectivity.**

In this section, we follow [7, 11, 1] in order to establish a certain injectivity property of the measurement operator $\mathcal{A}$.

Our Proposition 8 is the analogue of Lemma 3.7 in [1]. The latter contained a factor of $\mathcal{O}(\log^2 d)$ in the exponent of the failure probability, which does not appear here. The reason is that we employ a single-sided Bernstein inequality, instead of a symmetric Hoeffding inequality.

Proposition 8 (Robust injectivity, lower bound)**.** *With probability of failure smaller than $d^2 \exp\left(-\frac{\nu^4 L}{C_1 b^8}\right)$ the inequality*

$$\frac{1}{\nu^2 dL}\|\mathcal{A}(Z)\|_{\ell_2}^2 > \frac{1}{4}\|Z\|_2^2 \tag{26}$$

is valid for all matrices $Z \in T$ simultaneously. Here b and ν are as in (4, 5) and C_1 is an absolute constant.

We require bounds on certain variances for the proof of this statement. The technical Lemma 9 serves this purpose.

Lemma 9. *Let $Z \in T$ be an arbitrary matrix and let $\mathcal{M}_l$ be as in (20). Then it holds that*

$$\left\|\mathbb{E}\left[\mathcal{M}_l(Z)^2\right]\right\|_\infty \leq \frac{30 b^8}{\nu^4 L^2}\|Z\|_2^2, \tag{27}$$

and

$$\left\|\mathbb{E}\left[(\mathcal{P}_T \mathcal{M}_l(Z))^2\right]\right\|_\infty \leq \operatorname{tr}\left(\mathbb{E}\left[(\mathcal{P}_T \mathcal{M}_l(Z))^2\right]\right) \leq \frac{60 b^8}{\nu^4 L^2}\|Z\|_2^2. \tag{28}$$

In the following proof we will use that for $a, b \in \mathbb{Z}_d = \{0, \ldots, d-1\}$ one has

$$\frac{1}{d}\sum_{k=1}^{d} \omega^{k(a \ominus b)} = \delta_{a,b} = \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{else.} \end{cases} \tag{29}$$

The symbols $\oplus$ and $\ominus$ denote addition and subtraction modulo d.

Proof of Lemma 9. Let $y, z, v \in \mathbb{C}^d$ be vectors of unit length. Compute:

$$
\begin{aligned}
& \nu^4 L^2 \mathbb{E}\left[\mathcal{M}_l(yy^*)\mathcal{M}_l(zz^*)\right] v \\
(30) = \quad & \frac{1}{d^2} \sum_{k,j=1}^{d} \mathbb{E}\left[\left(\sum_{i_3,i_4=1}^{d} \epsilon_{i_3}\epsilon_{i_4}\omega^{k(i_3-i_4)} \bar{y}_{i_3} y_{i_4} \right) \left(\sum_{i_5,i_6=1}^{d} \epsilon_{i_5}\epsilon_{i_6}\omega^{j(i_5-i_6)} \bar{z}_{i_5} z_{i_6} \right) \right. \\
\times \quad & \left. \sum_{i_1,i_2,i_7,i_8=1}^{d} \epsilon_{i_1}\epsilon_{i_2}\omega^{k(i_2-i_1)}\epsilon_{i_7}\epsilon_{i_8}\omega^{j(i_8-i_7)} e_{i_2} \delta_{i_1,i_8} v_{i_7} \right] \\
= \quad & \sum_{i_1,\dots,i_7} \mathbb{E}\left[\epsilon_{i_1}^2 \epsilon_{i_2} \cdots \epsilon_{i_7}\right] \left(\frac{1}{d} \sum_k \omega^{k(i_2+i_3-i_1-i_4)} \right) \left(\frac{1}{d} \sum_j \omega^{j(i_5+i_1-i_6-i_7)} \right) \\
\times \quad & \bar{y}_{i_3} y_{i_4} \bar{z}_{i_5} z_{i_6} v_{i_7}\, e_{i_2} \\
(31) = \quad & \sum_{i_1,\dots,i_7} \mathbb{E}\left[\epsilon_{i_1}^2 \epsilon_{i_2} \cdots \epsilon_{i_7}\right] \delta_{i_1,(i_2 \oplus i_3 \ominus i_4)} \delta_{i_1,(i_6 \oplus i_7 \ominus i_5)} \bar{y}_{i_3} y_{i_4} \bar{z}_{i_5} z_{i_6} v_{i_7}\, e_{i_2} \\
(32) = \quad & \sum_{i_2,\dots,i_7} \mathbb{E}\left[\epsilon_{i_2 \oplus i_3 \ominus i_4}^2 \epsilon_{i_2} \cdots \epsilon_{i_7}\right] \delta_{i_2,(i_4 \oplus i_6 \oplus i_7 \ominus i_3 \ominus i_5)} \bar{y}_{i_3} y_{i_4} \bar{z}_{i_5} z_{i_6} v_{i_7}\, e_{i_2},
\end{aligned}
$$

where in (30) we have inserted the definition of $\mathcal{M}_l$, in (31) have made use of (29), and in (32) we have eliminated i_1. We now make the crucial observation that the expectation

$$
\mathbb{E}\left[\epsilon_{i_2 \oplus i_3 \ominus i_4}^2 \epsilon_{i_2} \cdots \epsilon_{i_7}\right] \tag{33}
$$

vanishes unless every number in $i_2, \dots, i_7$ appears at least twice. More formally, the expectation is zero unless the set $\{2, \dots, 7\}$ can be partitioned into a disjoint union of pairs $\{2, \dots, 7\} = \bigcup_{\{k,l\} \in E} \{k, l\}$ such that $i_k = i_l$ for every $\{k, l\} \in E$ (in graph theory, E would be a set of edges constituting a *matching*). Indeed, assume to the contrary that there is some j such that i_j is unmatched (i.e., $i_j \neq i_k$ for all $k \neq j$). We distinguish two cases: If $i_j \neq i_2 \oplus i_3 \ominus i_4$, then ϵ_j appears only once in the product in (33) and the expectation vanishes because $\mathbb{E}[\epsilon_j] = 0$ by assumption. If $i_j = i_2 \oplus i_3 \ominus i_4$, then the same conclusion holds because we have also assumed that $\mathbb{E}[\epsilon_j^3] = 0$ (this is the only point in the argument where we need third moments of ϵ to vanish).

With this insight, we can proceed to put a tight bound on the ℓ_2-norm of the initial expression.

$$
\begin{aligned}
& \|\nu^4 L^2 \mathbb{E}\left[\mathcal{M}(yy^*)\mathcal{M}(zz^*)\right] v\|_{\ell_2} \\
= \quad & \Big\| \sum_{i_2,\dots,i_7=1}^{d} \mathbb{E}\left[\epsilon_{i_2 \oplus i_3 \ominus i_4}^2 \epsilon_{i_2} \cdots \epsilon_{i_7}\right] \delta_{i_2,(i_4 \oplus i_6 \oplus i_7 \ominus i_3 \ominus i_5)} \bar{y}_{i_3} y_{i_4} \bar{z}_{i_5} z_{i_6} v_{i_7}\, e_{i_2} \Big\|_{\ell_2} \\
\leq \quad & \Big\| \sum_{i_2,\dots,i_7=1}^{d} \mathbb{E}\left[\epsilon_{i_2 \oplus i_3 \ominus i_4}^2 \epsilon_{i_2} \cdots \epsilon_{i_7}\right] \bar{y}_{i_3} y_{i_4} \bar{z}_{i_5} z_{i_6} v_{i_7}\, e_{i_2} \Big\|_{\ell_2} \\
\leq \quad & \sum_{\text{matchings } E} \Big\| \sum_{\substack{i_2,\dots,i_7 \\ i_k = i_l \text{ for } \{k,l\} \in E}} \left| \mathbb{E}\left[\epsilon_{i_2 \oplus i_3 \ominus i_4}^2 \epsilon_{i_2} \cdots \epsilon_{i_7}\right] \bar{y}_{i_3} y_{i_4} \bar{z}_{i_5} z_{i_6} v_{i_7} \right| e_{i_2} \Big\|_{\ell_2} \\
(34) \quad \leq \quad & b^8 \sum_{\text{matchings } E} \Big\| \sum_{\substack{i_2,\dots,i_7 \\ i_k = i_l \text{ for } \{k,l\} \in E}} \left| \bar{y}_{i_3} y_{i_4} \bar{z}_{i_5} z_{i_6} v_{i_7} \right| e_{i_2} \Big\|_{\ell_2},
\end{aligned}
$$

where the three inequalities follow, in that order, by realizing that making individual coefficients of e_{i_2} larger will increase the norm; restricting to non-zero expectation values as per the discussion above; and using the assumed bound $|\epsilon| \leq b$.

Now fix a matching E. Let $x^{(1)}$ be the vector in $\{v, \bar{y}, y, \bar{z}, z\}$ whose index in (34) is paired with i_2. Label the remaining four vectors in that set by $x^{(2)}, \dots, x^{(5)}$, in such a way that $x^{(2)}$ and $x^{(3)}$ are paired and the same is true for $x^{(4)}$ and $x^{(5)}$. Then the summand corresponding to that matching becomes

$$\begin{aligned} & \| \sum_{a,b,c=1}^{d} \left| x_a^{(1)} x_b^{(2)} x_b^{(3)} x_c^{(4)} x_c^{(5)} \right| e_a \|_{\ell_2} \\ = & \left(\sum_{b=1}^{d} |x_b^{(2)} x_b^{(3)}| \right) \left(\sum_{c=1}^{d} |x_c^{(4)} x_c^{(4)}| \right) \left\| \sum_{a=1}^{d} |x_a^{(1)}| e_a \right\|_{\ell_2} \leq 1, \end{aligned}$$

by the Cauchy-Schwarz inequality and the fact that all the $x^{(i)}$ are of length one. As there are 15 possible matchings of 6 indices, we arrive at

$$\|\mathbb{E}\left[\mathcal{M}(yy^*)\mathcal{M}(zz^*)\right] v\|_{\ell_2} \leq \frac{15 b^8}{\nu^4 L^2}.$$

Finally, let $Z \in T$. As Z has rank at most two, we can choose normalized vectors $y, z \in \mathbb{C}^d$ such that $Z = \lambda_1 yy^* + \lambda_2 zz^*$. Then

$$\left\|\mathbb{E}[\mathcal{M}(Z)^2]\right\|_\infty \leq \sum_{i,j=1}^{2} |\lambda_i|\, |\lambda_j|\, \frac{15 b^8}{\nu^4 L^2} = \|Z\|_1^2 \frac{15 b^8}{\nu^4 L^2} \leq \|Z\|_2^2 \frac{30 b^8}{\nu^4 L^2}.$$

For (28) we start by noting positive-semidefiniteness of $\mathbb{E}\left[\left(\mathcal{P}_T \mathcal{M}_l(Z)\right)^2\right]$ implies the first inequality. In order to bound the trace-term, we insert (23) for $\mathcal{P}_T$, expand the product, cancel terms using $X^2 = X = xx^*$ and use cyclicity of the trace to arrive at

$$\begin{aligned} \operatorname{tr}\left(\mathbb{E}\left[\left(\mathcal{P}_T \mathcal{M}_l(Z)\right)^2\right]\right) &= 2 \operatorname{tr}\left(\mathbb{E}\left[X \mathcal{M}_l(Z)^2\right]\right) - \operatorname{tr}\left(\mathbb{E}\left[(X \mathcal{M}_l(Z))(\mathcal{M}_l(Z) X)\right]\right) \\ &\leq 2 \operatorname{tr}\left(X\, \mathbb{E}\left[\mathcal{M}_l(Z)^2\right]\right) = 2\langle x, \mathbb{E}\left[\mathcal{M}_l(Z)^2\right] x\rangle \\ &\leq 2 \|\mathbb{E}\left[\mathcal{M}_l(Z)^2\right]\|_\infty. \end{aligned}$$

The upper bound in (28) is thus implied by (27). □

With Lemma 9 at hand, we can proceed to the lower bound on robust injectivity.

Proof of Proposition 8. We strongly follow the ideas presented in [19, Proposition 9] and aim to show the more general statement

$$\text{(35)} \qquad \Pr\left[(\nu^2 d L)^{-1} \|\mathcal{A}(Z)\|_{\ell_2}^2 \leq (1-\delta) \|Z\|_2^2 \quad \forall Z \in T\right] \leq d^2 \exp\left(-\frac{\nu^4 \delta^2 L}{\tilde{C}_1 b^8}\right)$$

for any $\delta \in (0,1)$, where $\tilde{C}_1$ is a numerical constant.

Pick $Z \in T$ arbitrary and use *near isotropicity* (21) of $\mathcal{R}$ in order to write

$$\begin{aligned}
& (\nu^2 dL)^{-1} \|\mathcal{A}(Z)\|_{\ell_2}^2 \\
= \ & (\nu^2 dL)^{-1} \sum_{l=1}^{L} \sum_{k=1}^{d} \left(\operatorname{tr}(F_{k,l} Z)\right)^2 = \operatorname{tr}\Big(Z \frac{1}{\nu^2 dL} \sum_{l=1}^{L} \sum_{k=1}^{d} F_{k,l} \operatorname{tr}(F_{k,l} Z)\Big) \\
= \ & \operatorname{tr}(Z\mathcal{R}Z) = \operatorname{tr}\left(Z(\mathcal{R} - \mathbb{E}[\mathcal{R}])Z\right) + \operatorname{tr}\left(Z(\mathcal{I} + \Pi_{\mathbb{1}})Z\right) \\
= \ & \operatorname{tr}\left(Z\mathcal{P}_T(\mathcal{R} - \mathbb{E}[\mathcal{R}])\mathcal{P}_T Z\right) + \operatorname{tr}(Z^2) + \operatorname{tr}(Z)^2 \\
\geq \ & \operatorname{tr}\left(Z\mathcal{P}_T(\mathcal{R} - \mathbb{E}[\mathcal{R}])\mathcal{P}_T Z\right) + \operatorname{tr}(Z^2) \\
\geq \ & (1 + \lambda_{\min}\left(\mathcal{P}_T(\mathcal{R} - \mathbb{E}[\mathcal{R}])\mathcal{P}_T\right))\|Z\|_2^2, \qquad (36)
\end{aligned}$$

where we have used the fact that $\mathcal{M} \geq \lambda_{\min}(\mathcal{M})\mathcal{I}$ for any matrix valued operator $\mathcal{M}$ as well as $\mathcal{P}_T Z = Z$. Therefore it suffices to to bound the smallest eigenvalue of $\mathcal{P}_T(\mathcal{R} - \mathbb{E}[\mathcal{R}])\mathcal{P}_T$ from below. To this end we aim to use the Operator Bernstein inequality – Theorem 5 – and decompose

$$\mathcal{P}_T(\mathcal{R} - \mathbb{E}[\mathcal{R}])\mathcal{P}_T = \sum_{l=1}^{L} \Big(\widetilde{\mathcal{M}}_l - \mathbb{E}[\widetilde{\mathcal{M}}_l]\Big) \quad \text{with} \quad \widetilde{\mathcal{M}}_l = \mathcal{P}_T \mathcal{M}_l \mathcal{P}_T,$$

where $\mathcal{M}_l$ was defined in (20). Note that these summands have mean zero by construction. Furthermore (25) implies

$$\begin{aligned}
-\frac{1}{\nu^2 L}\mathcal{I} - \frac{1}{\nu^2 L}\Pi_X \quad & \leq \quad -\frac{1}{\nu^2 L}\mathcal{P}_T \mathcal{I} \mathcal{P}_T - \frac{1}{\nu^2 L}\mathcal{P}_T \Pi_{\mathbb{1}} \mathcal{P}_T = -\frac{1}{L}\mathcal{P}_T \mathbb{E}[\mathcal{R}]\mathcal{P}_T \\
& = \quad -\mathcal{P}_T \mathbb{E}[\mathcal{M}_l]\mathcal{P}_T \leq \widetilde{\mathcal{M}}_l - \mathbb{E}[\widetilde{\mathcal{M}}_l],
\end{aligned}$$

where the last inequality follows from $\widetilde{\mathcal{M}}_l \geq 0$. This yields an a priori bound

$$\lambda_{\min}(\widetilde{\mathcal{M}}_l - \mathbb{E}[\widetilde{\mathcal{M}}_l]) \geq -2/(\nu^2 L) =: -\underline{R}.$$

For the variance we use the standard identity

$$0 \leq \mathbb{E}\left[(\widetilde{\mathcal{M}}_l - \mathbb{E}[\widetilde{\mathcal{M}}_l])^2\right] = \mathbb{E}\left[\widetilde{\mathcal{M}}_l^2\right] - \mathbb{E}\left[\widetilde{\mathcal{M}}_l\right]^2 \leq \mathbb{E}\left[\widetilde{\mathcal{M}}_l^2\right]$$

and focus on the last expression. For obtaining a bound on the total variance we are going to apply (14) to $\|\mathbb{E}[\widetilde{\mathcal{M}}_l^2]\|_{\mathrm{op}}$. To this end, fix $Z \in T$ arbitrary – this restriction is valid, due to the particular structure of $\widetilde{\mathcal{M}}_l$ – and observe

$$\begin{aligned}
|\operatorname{tr}\Big(Z\,\mathbb{E}\left[\widetilde{\mathcal{M}}_l^2\right] Z\Big)| \quad & = \quad |\mathbb{E}\left[\operatorname{tr}\left(\mathcal{M}_l(Z)\mathcal{P}_T\mathcal{M}_l(Z)\right)\right]| = |\operatorname{tr}\left(\mathbb{E}\left[(\mathcal{P}_T\mathcal{M}_l(Z))^2\right]\right)| \\
& \leq \quad 2\|\mathbb{E}\left[(\mathcal{P}_T\mathcal{M}_l(Z))^2\right]\|_\infty \leq \frac{120 b^8}{\nu^4 L^2}\|Z\|_2^2.
\end{aligned}$$

The first equality follows from inserting the definition (20) of $\mathcal{M}_l$ and rewriting the expression of interest. For the second equality, we have used the fact that $\operatorname{tr}(AB_T) = \operatorname{tr}(A_T B_T)$ for any matrix pair $A, B \in H^d$ ($\mathcal{P}_T$ is an orthogonal projection with respect to the Frobenius inner product) and the last estimate is due to (28) in Lemma 9. Since $Z \in T$ was arbitrary, we have obtained a bound on $\|\mathbb{E}[\widetilde{\mathcal{M}}_l^2]\|_{\mathrm{op}}$ which in turn allows us to set $\sigma^2 := \frac{120 b^8}{\nu^4 L}$ for the variance. Now we are ready to apply Theorem 5 which implies

$$\Pr\left[\lambda_{\min}\left(\mathcal{P}_T\left(\mathcal{R} - \mathbb{E}[\mathcal{R}]\right)\mathcal{P}_T\right) \leq -\delta\right] \leq d^2 \exp\left(-\frac{\nu^4 \delta^2 L}{\tilde{C}_1 b^8}\right)$$

for any $0 \leq \delta \leq 1 < 60b^8/\nu^2 = \sigma^2/\underline{R}$ and $\widetilde{C}_1$ is an absolute constant. This gives a suitable bound on the probability of the undesired event

$$\{\lambda_{\min}\left(\mathcal{P}_T(\mathcal{R} - \mathbb{E}[\mathcal{R}])\mathcal{P}_T\right) \leq -\delta\}.$$

If this is not the case, (36) implies

$$(dL)^{-1}\|\mathcal{A}(Z)\|_{\ell_2}^2 > (1-\delta)\|Z\|_2^2$$

for all matrices $Z \in T$ simultaneously. This proves (35) and setting $\delta = 3/4$ yields Proposition 8 (with $C_1 = \frac{16}{9}\widetilde{C}_1$). □

For our proof we will also require a uniform bound on $\|\mathcal{A}(Z)\|_{\ell_2}$.

Lemma 10 (Robust injectivity, upper bound)**.** *Let $\mathcal{A}$ be as above. Then the statement*

$$\frac{1}{dL}\|\mathcal{A}(Z)\|_{\ell_2}^2 \leq b^4 d\|Z\|_2^2 \tag{37}$$

holds with probability 1 for all matrices $Z \in H^d$ simultaneously.

Proof. Estimate

$$\begin{aligned} \frac{1}{dL}\|\mathcal{A}(Z)\|_{\ell_2}^2 &= \frac{1}{dL}\sum_{k,l}\left(\mathrm{tr}(f_k f_k^* D_l Z D_l)\right)^2 \leq \max_{1\leq k\leq d}\|f_k f_k^*\|_2^2 \frac{1}{dL}\sum_l \|D_l Z D_l\|_2^2 \\ &\leq d\|D_l\|_\infty^4\|Z\|_2^2 \leq db^4\|Z\|_2^2, \end{aligned}$$

where the first inequality holds because the $f_k f_k^*$'s are mutually orthogonal. The second inequality follows from the fact that the Frobenius norm (and more generally: any unitarily invariant norm) is symmetric [39, Proposition IV.2.4] – i.e., $\|ABC\|_2 \leq \|A\|_\infty\|B\|_2\|C\|_\infty$ for any $A, B, C \in H^d$ – and the last one is due to the a-priori bound $\|D_l\|_\infty \leq b$. □

5. Proof of the Main Theorem / Convex Geometry

In this section, we will prove that the convex program (9) indeed recovers the signal x with high probability. A common approach to prove recovery is to show the existence of an *approximate dual certificate*, which in our problem setup can be formalized by the following definition.

Definition 11 (Approximate dual certificate)**.** *Assume that the sampling process corresponds to (6). Then we call $Y \in H^d$ an* approximate dual certificate *if $Y \in \mathrm{range}\mathcal{A}^*$ and*

$$\|Y_T - X\|_2 \leq \frac{\nu}{4b^2\sqrt{d}} \quad \textit{as well as} \quad \|Y_T^\perp\|_\infty \leq \frac{1}{2}. \tag{38}$$

The following proposition, showing that the existence of such a dual certificate indeed guarantees recovery, is just a slight variation of Proposition 12 in [19]. For completeness, we have nevertheless included a proof in the appendix.

Proposition 12. *Suppose that the measurement gives us access to $\|x\|_{\ell_2}^2$ and $y_{k,l} = |\langle f_k, D_l x\rangle|^2$ for $1 \leq k \leq n$ and $1 \leq l \leq L$. Then the convex optimization (9) recovers the unknown x (up to a global phase), provided that (26) holds and an approximate dual certificate Y exists.*

Proposition 12 proves the Main Theorem of this paper, provided that an approximate dual certificate exists. A first approach to construct an approximate dual certificate is to set

$$Y = \mathcal{R}(X) - \operatorname{tr}(X)\mathbb{1}. \tag{39}$$

Note that any such Y is indeed in the range of our measurement process and, in expectation, yields an exact dual certificate, $\mathbb{E}[Y] = X$. One can then show using an operator Bernstein or Hoeffding inequality that Y is close to its expectation, but the number of measurements required is too large to make the result meaningful. This obstacle can be overcome using the golfing scheme, a refined construction procedure originally introduced in [11].

A main difference between our approach and the approach in [1] is that the authors of that paper use Hoeffding's inequality in the golfing scheme, while we employ Bernstein's inequality. The resulting bounds are sharper, but require to estimate an additional variance parameter.

An issue that remains is that such bounds heavily depend on the worst-case operator norm of the individual summands. In this framework these are proportional to $|\langle f_k, D_l x\rangle|^2$, which a priori can reach $b^2 d$ (recall that $\|f_k\|_2^2 = d$). To deal with this issue, we follow the approach from [19, 1] to condition on the event that their maximal value is not too large.

Lemma 13. *For $Z \in T$ abitrary and a parameter $\gamma \geq 1$ we introduce the event*

$$U_{k,l} := \left\{ |\operatorname{tr}(F_{k,l}Z)| \leq 2^{3/2} b^2 \gamma \log d \|Z\|_2 \right\}, \tag{40}$$

If D_l is chosen according to (3) it holds that

$$\max_{1\leq k \leq d} \Pr\left[U_{k,l}^c\right] \leq 4d^{-\gamma}.$$

In the following, we refer to γ as the *truncation rate* (cf. [19]). Here, we fix

$$\gamma = 8 + \log_2\left(b^2/\nu\right), \tag{41}$$

for reasons that shall become clear in the proofs of Propositions 16 and 17. Here b and ν are as in (4) and (5).

Proof of Lemma 13. Fix $Z \in T$ arbitrary and apply an eigenvalue decomposition

$$Z = \lambda_1 y y^* + \lambda_2 z z^*$$

with normalized eigenvectors $u, v \in \mathbb{C}^d$. Then one has for $1 \leq k \leq d$:

$$\begin{aligned} \Pr\left[U_{k,l}^c\right] &\leq \Pr\left[|\operatorname{tr}(F_{k,l}Z)| \geq 2b^2\gamma \log d \|Z\|_1\right] \\ &\leq \Pr\left[|\lambda_1||\langle f_k, D_l, y\rangle|^2 + |\lambda_2||\langle f_k, D_l, z\rangle|^2 \geq (|\lambda_1| + |\lambda_2|) 2b^2\gamma\log d\right] \\ &\leq \Pr\left[|\langle f_k, D_l y\rangle| \geq \sqrt{2b^2\gamma\log d}\right] + \Pr\left[|\langle f_k, D_l z\rangle| \geq \sqrt{2b^2\gamma \log d}\right], \end{aligned}$$

where the last inequality uses a union bound. The desired statement thus follows from

$$\Pr\left[|\langle f_k, D_l u\rangle| \geq b\sqrt{2\gamma\log d}\|u\|_{\ell_2}\right] \leq 2d^{-\gamma} \quad \forall u \in \mathbb{C}^d \ \forall 1 \leq k \leq d,$$

which we now aim to show. Fix $1 \leq k \leq d$ and $z = (z_1, \ldots, z_d) \in \mathbb{C}^d$ arbitrary and insert the definitions of f_k and D_l to obtain

$$|\langle f_k, D_l u\rangle| = |\sum_{i=1}^d \epsilon_i \left(\omega^{ki} u_i\right)| = |\sum_{i=1}^d \epsilon_i \tilde{u}_i|.$$

Here we have defined $\tilde{u} = \left(\omega^k u_1, \ldots, \omega^{k(d-1)} u_{d-1}, u_d\right)$. Note that $\|\tilde{u}\|_{\ell_2} = \|u\|_{\ell_2} = 1$ holds and applying Theorem 3 therefore yields

$$\begin{aligned} \Pr\left[\left|\sum_{i=1}^d \epsilon_i \tilde{u}_i\right| \geq b\sqrt{2\gamma \log d}\right] &= \Pr\left[\left|\sum_{i=1}^d \epsilon_i \tilde{u}_i\right| \geq b\sqrt{2\gamma \log d}\|\tilde{u}\|_2\right] \\ &\leq 2\exp\left(-\gamma \log d\right) = 2d^{-\gamma}. \end{aligned}$$

□

This result will be an important tool to bound the probability of extreme operator norms.

Definition 14. *For $Z \in T$ arbitrary and the corresponding $U_{k,l}$ introduced in (40) we define the truncated measurement operator*

$$\mathcal{R}_Z := \sum_{l=1}^L \mathcal{M}_l^Z \quad \textit{with} \quad \mathcal{M}_l^Z := \frac{1}{\nu^2 dL} \sum_{k=1}^d 1_{U_{k,l}} \Pi_{F_{k,l}}, \tag{42}$$

where $1_{U_{k,l}}$ denotes the indicator function associated with the event $U_{k,l}$.

We now show that in expectation, this truncated operator is close to the original one.

Lemma 15. *Fix $Z \in T$ arbitrary and let $\mathcal{R}_Z$ and $\mathcal{M}_l^Z$ be as in (42). Then*

$$\begin{aligned} \|\mathbb{E}[\mathcal{R} - \mathcal{R}_Z]\|_{\text{op}} &\leq \frac{4b^4}{\nu^2} d^{2-\gamma}, \\ \|\mathbb{E}[(\mathcal{M}_l(W))^2 - (\mathcal{M}_l^Z(W))^2]\|_\infty &\leq \frac{8b^8}{\nu^4 L^2} d^{4-\gamma} \|W\|_\infty^2, \\ \mathbb{E}\left[\|\mathcal{M}_l - \mathcal{M}_l^Z\|_{\text{op}}^2\right] &\leq \frac{4b^8}{\nu^4 L^2} d^{4-\gamma}. \end{aligned}$$

for any $W \in H^d$.

Proof. Note that $\mathbb{E}\left[\mathcal{R}\right] = L\mathbb{E}[\mathcal{M}_l]$ as well as $\mathbb{E}[\mathcal{R}_Z] = L\mathbb{E}\left[\mathcal{M}_l^Z\right]$. For the first statement, we can therefore fix $1 \leq l \leq L$ arbitrary and consider $L\|\mathbb{E}[\mathcal{M}_l - \mathcal{M}_l^Z]\|_{\text{op}}$. Due to Jensen's inequality this expression is majorized by $L\mathbb{E}\left[\|\mathcal{M}_l - \mathcal{M}_l^Z\|_{\text{op}}\right]$. Inserting the definitions and applying Lemma 13 then yields the first estimate via

$$\begin{aligned} L\mathbb{E}\left[\|\mathcal{M}_l - \mathcal{M}_l^Z\|_{\text{op}}\right] &\leq \frac{1}{\nu^2 d}\mathbb{E}\left[\sum_{k=1}^d (1 - 1_{U_{k,l}})\left\|\Pi_{F_{k,l}}\right\|_{\text{op}}\right] \leq \frac{b^4 d^2}{\nu^2 d}\sum_{k=1}^d \mathbb{E}\left[1_{U_{k,l}^c}\right] \\ &= \frac{b^4 d^2}{\nu^2 d}\sum_{k=1}^d \Pr\left[U_{k,l}^c\right] \leq \frac{b^4 d^2}{\nu^2} \max_{1\leq k\leq d} \Pr[U_{k,l}^c] \leq \frac{4b^4}{\nu^2} d^{2-\gamma}, \end{aligned}$$

where the second inequality is due to $\|\Pi_{F_{k,l}}\|_{\mathrm{op}} \leq b^4 d^2$ (which follows by direct calculation). Similarly

$$\begin{aligned} & \left\|\mathbb{E}\left[(\mathcal{M}_l(W))^2 - \left(\mathcal{M}_l^Z(W)\right)^2\right]\right\|_\infty \\ = & \left\| \frac{1}{(\nu^2 dL)^2} \sum_{k,j=1}^d \mathbb{E}\left[(1-1_{U_{k,l}}1_{U_{j,l}}) \operatorname{tr}(F_{k,l}W)\operatorname{tr}(F_{j,l}W)F_{k,l}F_{j,l}\right]\right\|_\infty \\ \leq & \frac{1}{\nu^4 L^2 d^2}\sum_{k,j=1}^d \mathbb{E}\left[1_{U_{k,l}^c \cup U_{j,l}^c} |\operatorname{tr}(F_{k,l}W)\operatorname{tr}(F_{j,l}W)| \|F_{k,l}\|_\infty \|F_{j,l}\|_\infty\right] \\ \leq & \frac{b^8 d^4}{\nu^4 L^2}\|W\|_\infty^2 \max_{1\leq k,j\leq d}\left(\Pr[U_{k,l}^c]+\Pr[U_{j,l}^c]\right) \leq \frac{8b^8}{\nu^4 L^2} d^{4-\gamma}\|W\|_\infty^2 \end{aligned}$$

Here we have used $|\operatorname{tr}(F_{k,l}W)| \leq b^2 d \|W\|_\infty$ for any $W \in H^d$ and $\|F_{k,l}\|_\infty \leq b^2 d$ (both estimates are direct consequences of the definition of $F_{k,l}$). Finally

$$\begin{aligned} \mathbb{E}\left[\|\mathcal{M}_l - \mathcal{M}_l^Z\|_{\mathrm{op}}^2\right] & \leq \frac{1}{(\nu^2 dL)^2}\mathbb{E}\left[\left(\sum_{k=1}^d (1-1_{U_{k,l}})\|\Pi_{F_{k,l}}\|_{\mathrm{op}}\right)^2\right] \\ & \leq \frac{b^8 d^4}{\nu^4 d^2 L^2}\sum_{k,j=1}^d \mathbb{E}\left[1_{U_{k,l}^c} 1_{U_{j,l}^c}\right] \leq \frac{b^8 d^4}{\nu^4 L^2} \max_{1\leq k\leq d} \Pr\left[U_{k,l}^c\right] \\ & \leq \frac{4b^8}{\nu^4 L^2} d^{4-\gamma} \end{aligned}$$

follows in a similar fashion. □

We will now establish two technical ingredients for the golfing scheme.

Proposition 16. *Assume $d \geq 3$, fix $Z \in T$ arbitrary and let $\mathcal{R}_Z$ be as in (42). Then*

$$\Pr\left[\|\mathcal{P}_T^\perp(\mathcal{R}_Z(Z) - \operatorname{tr}(Z)\mathbb{1})\|_\infty \geq t\|Z\|_2\right] \leq d \exp\left(-\frac{t\nu^4 L}{C_2 b^8 \gamma \log d}\right) \tag{43}$$

for any $t \geq 1/4$ and γ defined in (41). Here C_2 denotes an absolute constant.

Proof. Assume w.l.o.g. that $\|Z\|_2 = 1$. By Lemma 7,

$$\mathcal{P}_T^\perp \mathbb{E}[\mathcal{R}(Z)] = \mathcal{P}_T^\perp(Z + \operatorname{tr}(Z)\mathbb{1}) = 0 + \operatorname{tr}(Z)\mathcal{P}_T^\perp \mathbb{1},$$

because $Z \in T$ by assumption. We can thus rewrite the desired expression as

$$\begin{aligned} & \|\mathcal{P}_T^\perp\left(\mathcal{R}_Z(Z) - \mathbb{E}[\mathcal{R}(Z)]\right)\|_\infty \\ \leq & \|\mathcal{P}_T^\perp\left(\mathcal{R}_Z(Z) - \mathbb{E}[\mathcal{R}_Z(Z)]\right)\|_\infty + \|\mathcal{P}_T^\perp \mathbb{E}\left[\mathcal{R}_Z(Z) - \mathcal{R}(Z)\right]\|_2 \\ \leq & \|\mathcal{R}_Z(Z) - \mathbb{E}[\mathcal{R}_Z(Z)]\|_\infty + \|\mathbb{E}[\mathcal{R}_Z - \mathcal{R}]\|_{\mathrm{op}}\|Z\|_2 \\ \leq & \|\mathcal{R}_Z(Z) - \mathbb{E}[\mathcal{R}_Z(Z)]\|_\infty + t/4. \end{aligned} \tag{44}$$

In the third line, we have used that $\|\mathcal{P}_T^\perp W\| \leq \|W\|$ for any $W \in H^d$ and any unitarily invariant norm $\|\cdot\|$ (pinching, cf. [39] (Problem II.5.4)). The last inequality follows from

$$\|\mathbb{E}[\mathcal{R}_Z - \mathcal{R}]\|_{\mathrm{op}} \leq \frac{4b^4}{\nu^2} d^{2-\gamma} \leq \frac{b^4}{\nu^2} 2^{4-\gamma} \leq \frac{1}{16} \leq \frac{t}{4}, \tag{45}$$

which in turn follows from Lemma 15 and the assumptions on d, t and γ. By (44), it remains to bound the probability of the complement of the event

$$E := \{\|\mathcal{R}_Z(Z) - \mathbb{E}[\mathcal{R}_Z(Z)]\|_\infty \le 3t/4\}$$

To this end, we use the Operator Bernstein inequality (Theorem 4). We decompose

$$\mathcal{R}_Z(Z) - \mathbb{E}[\mathcal{R}_Z(Z)] = \sum_{l=1}^{L} (M_l - \mathbb{E}[M_l]) \quad \text{with} \quad M_l := \mathcal{M}_l^Z(Z),$$

where $\mathcal{M}_l^Z$ was defined in (42). To find an a priori bound for the individual summands, we write, using that $F_{k,l} \ge 0$ holds for all $1 \le k \le d$,

$$\begin{aligned} \|M_l - \mathbb{E}[M_l]\|_\infty &\le \|M_l\|_\infty + \|\mathbb{E}\left[\mathcal{M}_l(Z) - \mathcal{M}_l^Z(Z)\right]\|_\infty + \|\mathbb{E}\left[\mathcal{M}_l(Z)\right]\|_\infty \\ &\le \|M_l\|_\infty + \frac{1}{L}\|\mathbb{E}\left[\mathcal{R}_l - \mathcal{R}_l^Z\right]\|_{\mathrm{op}}\|Z\|_2 + \frac{1}{L}\|Z + \mathrm{tr}(Z)\|_\infty \\ &\le \left\| \frac{1}{\nu^2 dL} \sum_{k=1}^{d} 1_{U_{k,l}} |\mathrm{tr}(F_{k,l}Z)| F_{k,l} \right\|_\infty + \frac{1}{L}\left(\frac{b^4}{\nu^2} d^{2-\gamma} + 1 + \sqrt{2}\right)\|Z\|_2 \\ &\le \frac{b^4}{\nu^2 L}\left(2^{3/2}\gamma \log d + d^{2-\gamma} + 3\right)\|Z\|_2 \le \frac{608 b^8 \gamma \log d}{3\nu^4 L} =: \overline{R}. \end{aligned}$$

Here we have employed near-isotropy of $\mathcal{R}$, the first estimate in Lemma 15 and the fact that $Z \in T$ has rank at most two. The last but one inequality follows from $\frac{1}{d}\sum_{k=1}^{d} f_k f_k^* = \mathbb{1}$, $\|D_l^2\|_\infty \le b^2$, and $\nu \le b^2$. The last estimate is far from tight, but will slightly simplify the resulting operator Bernstein bound. For the variance we start with the standard estimate

$$\mathbb{E}\left[(M_l - \mathbb{E}[M_l])^2\right] = \mathbb{E}\left[M_l^2\right] - \mathbb{E}[M_l]^2 \le \mathbb{E}\left[M_l^2\right]$$

and bound this expression via

$$\begin{aligned} &\|\mathbb{E}\left[M_l^2\right]\|_\infty = \left\|\mathbb{E}\left[\left(\mathcal{M}_l^Z(Z)\right)^2\right]\right\|_\infty \\ \le\ & \left\|\mathbb{E}\left[\left(\mathcal{M}_l^Z(Z)\right)^2 - \left(\mathcal{M}_l(Z)\right)^2\right]\right\|_\infty + \left\|\mathbb{E}\left[\left(\mathcal{M}_l(Z)\right)^2\right]\right\|_\infty \\ \le\ & \frac{8b^8}{\nu^4 L^2} d^{4-\gamma}\|Z\|_\infty^2 + \frac{30 b^8}{\nu^4 L^2}\|Z\|_2^2, \end{aligned}$$

where we have used Lemmas 15 and 9. Using $\|Z\|_\infty \le \|Z\|_2 = 1$ and noting that $\nu \le b^2$ entails $\gamma = 8 + 2\log_2(b^2/\nu) \ge 8$ we conclude

$$\|\sum_{l=1}^{L} \mathbb{E}[M_l^2]\|_\infty \le \sum_{l=1}^{L} \|\mathbb{E}[M_l^2]\|_\infty \le \frac{8b^8}{\nu^4 L} d^{-4} + \frac{30 b^8}{\nu^4 L} \le \frac{38 b^8}{\nu^4 L} =: \sigma^2.$$

Our choice for $\overline{R}$ now guarantees $\sigma^2/\overline{R} = 3/(16\gamma\log d) \le 3t/4$ for any $t \ge 1/4$ (here we have used $\gamma \ge 1$ and our assumption $d \ge 3$ which entails $\log d \ge 1$). Consequently

$$\Pr\left[E^c\right] = \Pr\left[\left\|\sum_{l=1}^{L}(M_l - \mathbb{E}[M_l])\right\|_\infty > 3t/4\right] \le d \exp\left(-\frac{t\nu^4 L}{C_2 b^8 \gamma \log d}\right)$$

with C_2 an absolute constant. This completes the proof. □

Proposition 17. *Assume $d \ge 2$ and fix $Z \in T$ arbitrary and let $\mathcal{R}_Z$ be as in* (42) *with γ defined in* (41). *Then*

$$\text{(46)} \qquad \Pr\left[\|\mathcal{P}_T\left(\mathcal{R}_Z(Z) - Z - \mathrm{tr}(Z)\mathbb{1}\right)\|_2 \ge c\|Z\|_2\right] \le \exp\left(-\frac{c^2\nu^4 L}{C_3 b^8 \gamma \log d} + \frac{1}{4}\right)$$

holds for any $1/(2\log d) \leq c \leq 1$. *Here,* C_3 *is again an absolute constant.*

Proof. Similar to the previous proof, we start by assuming $\|Z\|_2 = 1$ and using *near-isotropy* of $\mathcal{R}$ to bound the desired expression by

$$\begin{aligned} & \|\mathcal{P}_T\left(\mathcal{R}_Z(Z) - \mathbb{E}\left[\mathcal{R}(Z)\right]\right)\|_2 \\ \leq\ & \|\mathcal{P}_T\left(\mathcal{R}_Z(Z) - \mathbb{E}\left[\mathcal{R}_Z(Z)\right]\right)\|_2 + \|\mathcal{P}_T \mathbb{E}\left[\mathcal{R}(Z) - \mathcal{R}_Z(Z)\right]\|_2 \\ \leq\ & \|\mathcal{P}_T\left(\mathcal{R}_Z(Z) - \mathbb{E}\left[\mathcal{R}_Z(Z)\right]\right)\|_2 + \|\mathcal{P}_T \mathbb{E}\left[\mathcal{R} - \mathcal{R}_Z\right]\|_{\mathrm{op}} \|Z\|_2 \\ \leq\ & \|\mathcal{P}_T\left(\mathcal{R}_Z(Z) - \mathbb{E}\left[\mathcal{R}_Z(Z)\right]\right)\|_2 + c/4. \end{aligned}$$

Here, we have used $\|\mathcal{P}_T W\|_2 \leq \|W\|_2$ for any matrix W (this follows e.g. from the entry-wise definition of the Frobenius norm) and a calculation similar to (45):

$$\|\mathbb{E}\left[\mathcal{R}_Z - \mathcal{R}\right]\|_{\mathrm{op}} \leq \frac{4b^4}{\nu^2 d} d^{3-\gamma} \leq \frac{b^4}{\nu^2 \log d} 2^{5-\gamma} \leq \frac{1}{8\log d} \leq \frac{c}{4},$$

where we have used $d \geq 2$, $\gamma \geq 8$ and the assumption $c \geq 1/(2\log d)$. Paralleling our idea from the previous proof, we define the event

$$E' := \{\|\mathcal{P}_T(\mathcal{R}_Z(Z) - \mathbb{E}[\mathcal{R}_Z(Z)])\|_\infty \leq 3c/4\}$$

which guarantees that the desired inequality is valid. However, in order to bound the probability of $(E')^c$, this time we are going to employ the vector Bernstein inequality—Theorem 6. Decompose

$$\mathcal{P}_T\left(\mathcal{R}_Z(Z) - \mathbb{E}\left[\mathcal{R}_Z(Z)\right]\right) = \sum_{l=1}^{L} \left(\tilde{M}_l - \mathbb{E}\left[\tilde{M}_l\right]\right).$$

Note that the $\tilde{M}_l$'s are related to M_l in the previous proof via $\tilde{M}_l = \mathcal{P}_T M_l = \mathcal{P}_T \mathcal{M}_l^Z(Z)$. and in particular, $\tilde{M}_l$ has at most rank two. Consequently

$$\begin{aligned} \|\tilde{M}_l - \mathbb{E}\left[\tilde{M}_l\right]\|_2 \leq\ & \sqrt{2}\|\mathcal{P}_T M_l\|_\infty + \|\mathcal{P}_T \mathbb{E}\left[\mathcal{M}_l^Z(Z) - \mathcal{M}_l(Z)\right]\|_2 + \|\mathcal{P}_T \mathbb{E}\left[\mathcal{M}_l(Z)\right]\|_2 \\ \leq\ & 2^{3/2}\|M_l\|_\infty + \|\mathbb{E}\left[\mathcal{M}_l - \mathcal{M}_l^Z\right]\|_{\mathrm{op}}\|Z\|_2 + \frac{1}{L}\|\mathcal{P}_T\left(Z + \mathrm{tr}(Z)\mathbb{1}\right)\|_2 \\ \leq\ & \frac{8b^2\gamma\log d}{\nu^2 L}\|Z\|_2\|D_l^2\|_\infty + \frac{4b^4}{\nu^2 L} d^{2-\gamma}\|Z\|_2 + \frac{\|Z\|_2 + |\,\mathrm{tr}(Z)|}{L} \\ \leq\ & \frac{15b^4\gamma\log d}{\nu^2 L}\|Z\|_2 =: B, \end{aligned}$$

where we have used near-isotropy of $\mathcal{M}_l$, the estimate of $\|M_l\|_\infty$ presented in (46), $\|\mathcal{P}_T\mathbb{1}\|_2 = \|X\|_2 = 1$ and $|\,\mathrm{tr}(Z)| \leq \|Z\|_1 \leq \sqrt{2}\|Z\|_2 = \sqrt{2}$, because $Z \in T$ has rank at most two. For the variance, we estimate

$$\begin{aligned} \mathbb{E}\left[\left\|\tilde{M}_l - \mathbb{E}\left[\tilde{M}_l\right]\right\|_2^2\right] =\ & \mathbb{E}\left[\left\|\mathcal{P}_T\left(\mathcal{M}_l^Z(Z) - \mathbb{E}\left[\mathcal{M}_l^Z(Z)\right]\right)\right\|_2^2\right] \\ \leq\ & \mathbb{E}\left[\|\mathcal{P}_T\mathcal{M}_l(Z)\|_2^2\right] + \mathbb{E}\left[\left\|\mathcal{P}_T\left(\mathcal{M}_l^Z(Z) - \mathcal{M}_l(Z)\right)\right\|_2^2\right] \\ +\ & \left\|\mathcal{P}_T\mathbb{E}\left[\mathcal{M}_l^Z(Z) - \mathcal{M}_l(Z)\right]\right\|_2^2 + \|\mathcal{P}_T\mathbb{E}\left[\mathcal{M}_l(Z)\right]\|_2^2 \\ \leq\ & \mathbb{E}\left[\mathrm{tr}\left(\left(\mathcal{P}_T\mathcal{M}_l(Z)\right)^2\right)\right] + \frac{1}{L^2}\|\mathcal{P}_T\left(Z + \mathrm{tr}(Z)\mathbb{1}\right)\|_2^2 \\ +\ & 2\mathbb{E}\left[\left\|\mathcal{M}_l(Z) - \mathcal{M}_l^Z(Z)\right\|_{\mathrm{op}}^2\right]\|Z\|_2^2 \\ \leq\ & \frac{60b^8}{\nu^4 L^2}\|Z\|_2^2 + \frac{\|Z\|_2^2 + \mathrm{tr}(Z)^2}{L^2} + \frac{8b^8}{\nu^4 L^2} d^{4-\gamma}\|Z\|_2^2. \qquad (47) \end{aligned}$$

Applying $b^2 \geq \nu$, $\mathrm{tr}(Z)^2 \leq 2\|Z\|_2^2 = 2$ and $d^{4-\gamma} \leq 1$ (because we choose $\gamma \geq 8$) allows us to upper-bound (47) by $71b^8/(\nu^4 L^2)$ and set

$$\sum_{l=1}^{L} \mathbb{E}\left[\left\|\tilde{M}_l - \mathbb{E}\left[\tilde{M}_l\right]\right\|_2^2\right] \leq \frac{71b^8}{\nu^4 L} \leq \frac{15b^8\gamma\log d}{\nu^4 L} =: \sigma.$$

Again, the last estimate is far from tight, but assures $\sigma^2/B = b^4/\nu^2 \geq 1$. Applying the vector Bernstein inequality—Theorem 6—for $t = 3c/4$ yields the desired bound on the probability of $(E')^c$ occurring.

□

We are now ready to construct a suitable approximate dual certificate in the sense of Definition 11. The key idea here is an iterative procedure – dubbed the *golfing scheme* – that was first established in [11] (see also [40, 24, 1, 19]).

Proposition 18. *Assume $d \geq 3$ and let $\omega \geq 1$ be arbitrary. If the total number of L of diffraction patterns fulfills*

$$L \geq C\omega \log^2 d, \tag{48}$$

then with probability larger than $1 - 5/6\mathrm{e}^{-\omega}$, an approximate dual certificate Y as in Definition 11 can be constructed using the golfing scheme. Here, C is a constant that only depends on the probability distribution used to generate the random masks D_l.

To be concrete, the constant C depends on the truncation rate γ – which we have fixed in (41) – and the a-priori bound b and ν of the random variable ϵ used to generate the diffraction patterns D_l:

$$C = \tilde{C}\gamma\frac{b^8}{\nu^4}\log_2\left(b^2/\nu\right) = \bar{C}\frac{b^8}{\nu^4}\log_2^2\left(b^2/\nu\right), \tag{49}$$

where $\tilde{C}$ and $\bar{C}$ are absolute constants.

Proof of Proposition 18. This construction is inspired by [24, 40] and [41]. As in [11], our construction of Y follows a recursive procedure of w iterations which can be summarized in the pseudo-code described in Algorithm 1. It depends on a number of parameters – w, L_i, r, c.f. Input section of the algorithm – the values of which will be chosen below. If this algorithm succeeds, it outputs three lists

$$\mathbf{Y} = [Y_1, \ldots, Y_{r+2}], \quad \mathbf{Q} = [Q_0, \ldots, Q_{r+2}], \quad \text{and} \quad \xi = [\xi_1, \ldots, \xi_{w+2}].$$

They obey iterative relations of the following form (c.f. [24, Lemma 14]):

$$\begin{aligned} Y &:= Y_{r+2} = \mathcal{R}_{Q_{r+1}}(Q_{r+1}) - \mathrm{tr}(Q_{r+1})\mathbb{1} + Y_{r+1} \\ &= \cdots = \sum_{i=1}^{r+2}\left(\mathcal{R}_{Q_{i-1}}(Q_{i-1}) - \mathrm{tr}\left(Q_{i-1}\right)\mathbb{1}\right) \quad \text{and} \\ Q_i &= X - \mathcal{P}_T Y_i = \mathcal{P}_T\left(Q_{i-1} + \mathrm{tr}(Q_{i-1})\mathbb{1} - \mathcal{R}_{Q_{i-1}}(Q_{i-1})\right) \\ &= \ldots = \prod_{j=1}^{i}\mathcal{P}_T\left(\mathcal{I} + \Pi_{\mathbb{1}} - \mathcal{R}_{Q_{j-1}}\right)Q_0. \end{aligned}$$

[5] Similar to [19] we use use of pseudo-code for a compact presentation of this randomized procedure. However, the reader should keep in mind that the construction is purely part of a proof and should not be confused with the recovery algorithm (which is given in Eq. (9)).

Algorithm 1: Pseudo-code[5] that summarizes the randomized “golfing scheme” for constructing an approximate dual certificate in the sense of Definition 11.

```
Input:
    $X \in H^d$                   # signal to be recovered
    $w \in \mathbb{N}$                   # maximum number of iterations (after the first two steps)
    $\{L_i\}_{i=1}^{w+2} \subset \mathbb{N}$       # number of masks used in $i$th iteration
    $r$                        # require $r$ “successful” iterations after the first two
                             # (i.e. iterations where we enter the inner if-block)
Initialize:
    $\mathbf{Y} = [\,]$                 # a list of matrices in $H^d$, initially empty
    $\mathbf{Q} = [X]$                # a list of matrices in $T$, initialized to hold $X$ as its only element
    i = 1                    # number of current iteration
    $\xi = [0, \ldots, 0]$         # array of $w+1$ zeros; $\xi_i$ will be set to 1 if $i$th iteration succeeds
Body:
for $1 \le i \le 2$ do
    set $Q$ to be the last element of $\mathbf{Q}$ and $Y$ to be the last element of $\mathbf{Y}$
    Sample $L_i$ masks independently according to (3) and construct $\mathcal{R}_Q$ according to
    Def. 14 if (43),(46) hold for $\mathcal{R}_Q$ and $Q \in T$ with parameters $t = 1/8$,
    $c = 1/\sqrt{2 \log d}$ then
        $\xi_i = 1$
        $Y \leftarrow \mathcal{R}_Q Q - \mathrm{tr}(Q)\mathbb{1} + Y$,    append $Y$ to $\mathbf{Y}$
        $Q \leftarrow X - \mathcal{P}_T Y$,    append $Q$ to $\mathbf{Q}$ $i \leftarrow i + 1$
    else
        abort and report failure
    end
end
while $3 \le i \le w + 2$ and $\sum_{j=3}^{i} \xi_j \le r$ do
    set $Q$ to be the last element of $\mathbf{Q}$ and $Y$ to be the last element of $\mathbf{Y}$,
    sample $L_{i+2}$ masks independently according to (3); construct $\mathcal{R}_Q$ according to
    Def. 14.
    if (43), (46) hold for $\mathcal{R}_Q$ and $Q \in T$ with parameters $t = \log d/4$, $c = 1/2$ then
        $\xi_i = 1$
        $Y \leftarrow \mathcal{R}_Q Q - \mathrm{tr}(Q)\mathbb{1} + Y$,    append $Y$ to $\mathbf{Y}$
        $Q \leftarrow X - \mathcal{P}_T Y$,    append $Q$ to $\mathbf{Q}$
    end
    $i \leftarrow i + 1$
end
if $\sum_{i=3}^{w+2} \xi_i = r$ then
    report success and output $\mathbf{Y}, \mathbf{Q}, \xi$
else
    report failure
end
```

We now set

$$r = \lceil \frac{1}{2} \log_2 d \rceil + \lceil \log_2 (b^2/\nu) \rceil + 1$$

This choice, together with the validity of properties (43) and (46) for $t = 1/8$, $c = 1/\sqrt{2\log d}$ in the first two steps and for $t = \log d/4$, $c = 1/2$ in each remaining update ($Y_i \to Y_{i+1}$ and $Q_i \to Q_{i+1}$, respectively) together with $Q_0 = X$ then guarantee

$$\begin{aligned} \|Y_T - X\|_2 &= \|Q_{r+2}\|_2 \le \|Q\|_0 \frac{1}{2\log d}\prod_{i=3}^{r+2}\frac{1}{2} = \frac{1}{\log d}2^{-(r+1)} \le \frac{\nu}{4b^2\sqrt{d}}, \\ \|Y_T^\perp\|_\infty &\le \sum_{i=1}^{r+2} \left\|\mathcal{P}_T\left(\mathcal{R}_{Q_{i-1}}(Q_{i-1}) - \operatorname{tr}(Q_{i-1})\mathbb{1}\right)\right\|_\infty \\ &\le \frac{1}{8}\|Q_0\|_2 + \frac{1}{8}\|Q_1\| + \sum_{i=3}^{r+2}\frac{\log d}{4}\|Q_{i-1}\|_2 \\ &\le \left(\frac{1}{8} + \frac{1}{8\sqrt{2\log d}} + \sum_{i=3}^{r+2}\frac{\log d}{4}\left(\frac{1}{\sqrt{2\log d}}\right)^2 \prod_{j=1}^{i-2}\frac{1}{2}\right)\|Q_0\|_2 \\ &\le \frac{1}{4}\sum_{i=0}^{\infty} 2^{-i} = \frac{1}{2} \end{aligned}$$

which are precisely the requirements (38) on Y.

What remains to be done now is to choose parameters w and $\{L_i\}_{i=1}^{w+2}$ such that the probability of the algorithm failing is smaller than $\frac{5}{6}\mathrm{e}^{-\omega}$. Recall that the ξ_i's are Bernoulli random variables that indicate whether the i-th iteration of the algorithm failed ($\xi_i = 0$) or has been successful ($\xi_i = 1$). The complete Algorithm 1 fails exactly if one of the first two iterations fails

(50) $$\xi_1 = 0 \qquad \text{or} \qquad \xi_2 = 0$$

or fewer than r of the remaining ones succeed

(51) $$\sum_{i=3}^{w+2}\xi_i < r.$$

We start by estimating the probability of (50) occuring. Setting

$$L_1 = L_2 = C_5\frac{b^8}{\nu^4}\omega\gamma\log^2 d$$

for a sufficiently large absolute constant C_5, and using the union bound over Propositions 16 and 17 (for $Z = X$), one obtains

$$\begin{aligned} & \Pr\left[\xi_1 = 0\right] \\ \le\ & \Pr\left[\text{(43) fails to hold in the first step}\right] + \Pr\left[\text{ (46) fails to hold in the first step}\right] \\ \text{(52)}\ \le\ & \exp\left(-\frac{(1/\sqrt{2\log d})^2\nu^4 L_1}{C_3 b^8\gamma\log d} + \frac{1}{4}\right) + d\exp\left(-\frac{4^{-1}\nu^4 L_1}{C_2 b^8\gamma\log d}\right) \le \frac{1}{6}\mathrm{e}^{-\omega}. \end{aligned}$$

An analogous bound holds for the probability of $\xi_2 = 0$.

We turn to (51). Our aim is to bound $\Pr\left[\sum_{i=3}^{w+2}\xi_i < r\right]$ by a similar expression involving *independent* Bernoulli variables ξ_i'. To achieve this, we observe

$$\Pr\left[\sum_{i=3}^{w+2}\xi_i < r\right] = \mathbb{E}\left[\Pr\left[\xi_{w+2} < r - \sum_{i=3}^{w+1}\xi_i | \xi_{w+1},\dots,\xi_3\right]\right].$$

Conditioned on an arbitrary instance of $\xi_{w+1}, \ldots, \xi_3$, the variable ξ_{w+2} follows a Bernoulli distribution with some parameter $p\left(\xi_w, \ldots, \xi_2\right)$. Now note that if $\xi \sim \mathrm{B}(p)$ is a Bernoulli variable with parameter p, then for every fixed $t \in \mathbb{R}$, the probability $\Pr_{\xi\sim\mathrm{B}(p)}\left[\xi < t\right]$ is non-increasing as a function of p. This observation implies that the estimate

$$\Pr\left[\sum_{i=3}^{w+2} \xi_i < r\right] \leq \Pr\left[\xi'_{w+2} + \sum_{i=3}^{w+1} \xi_i < r\right] \tag{53}$$

is valid, provided that ξ'_{w+1} is an independent p'-Bernoulli distributed random variable with

$$p' \leq \min_{\xi_{w+1},\ldots,\xi_3} p\left(\xi_{w+1}, \ldots, \xi_3\right).$$

A combination of Propositions 16 and 17 provides a uniform lower bound on $p\left(\xi_{w+1}, \ldots, \xi_3\right)$. Indeed, setting $Z = Q_w$ and invoking them with

$$L := C_4 \frac{b^8}{\nu^4} \gamma \log d$$

– where C_4 is a sufficiently large constant – assures a probability of success of at least $9/10$ for any Q. This estimate is in particular independent of $\xi_{w+1}, \ldots, \xi_3$. Consequently, by choosing $p' = 9/10$ and $L_i = L$ for all $3 \leq i \leq w+2$, we can iterate the estimate (53) and arrive at

$$\Pr\left[\sum_{i=3}^{w+2} \xi_i < r\right] \leq \Pr\left[\xi'_{w+2} + \sum_{i=3}^{w+1} \xi_i < r\right] \leq \cdots \leq \Pr\left[\sum_{i=3}^{w+2} \xi'_i < r\right], \tag{54}$$

where the ξ_i''s on the right hand side are independent Bernoulli variables with parameter $9/10$. A standard one-sided Chernoff bound (e.g. e.g [42, Section Concentration: Theorem 2.1]) gives

$$\Pr\left[\sum_{i=3}^{w+2} \xi'_i \leq w(9/10 - t)\right] \leq \mathrm{e}^{-2wt^2}.$$

Choosing $t = 9/10 - r/w$, we then obtain

$$\begin{aligned} \Pr\left[\sum_{i=3}^{w+2} \xi'_i < r\right] &\leq \Pr\left[\sum_{i=3}^{w+2} \xi'_i \leq r\right] = \Pr\left[\sum_{i=3}^{w+2} \xi'_i \leq w\left(9/10 - t\right)\right] \\ &\leq \exp\left(-2w\left(\frac{9}{10} - \frac{r}{w}\right)^2\right). \end{aligned} \tag{55}$$

Setting the number of iterations generously to

$$w = 10\omega r = 10\omega\left(\lceil \frac{1}{2}\log_2 d\rceil + \lceil \log_2(b^2/\nu)\rceil + 1\right)$$

guarantees

$$2w\left(\frac{9}{10} - \frac{r}{w}\right)^2 \geq 20\omega r\left(8/10\right)^2 \geq 12\omega r \geq \omega + \log 2,$$

where we have used $\omega \geq 1$ in the first and last step. From this estimate we can conclude

$$\Pr\left[\sum_{i=3}^{w+2} \xi_i < r\right] \leq \mathrm{e}^{-\omega - \log 2} = \frac{1}{2}\mathrm{e}^{-\omega} \tag{56}$$

which suffices for our purpose.

The desired bound of $\frac{5}{6}\mathrm{e}^{-\omega}$ on the probability of the algorithm failing now follows from taking the union bound over (52) and two times (56).

Finally we note that with our construction the total amount of masks obeys

$$\begin{aligned} L &= \sum_{i=1}^{w+2} L_i = 2C_5 \frac{b^8}{\nu^4}\omega\gamma \log^2 d + 10\omega \left(\lceil 0.5 \log_2 d \rceil + \lceil \log_2 (b^2/\nu \rceil \right) C_4 \frac{b^8}{\nu^4} \gamma \log d \\ &\leq \tilde{C}\gamma \frac{b^8}{\nu^4} \log_2 \left(b^2/\nu \right) \omega \log^2 d = C\omega \log^2 d \end{aligned}$$

for a sufficiently large absolute constant $\tilde{C}$ (recall that we have chosen $\gamma = 8 + \log_2 \left(b^2/\nu \right)$ in (41)) and C as in (49). □

We now have all the ingredients for the proof of our main result, Theorem 1.

Proof of the Main Theorem. With probability at least $1 - 5/6\mathrm{e}^{-\omega}$, the construction of Proposition 18 yields an approximate dual certificate provided that the total number of masks L obeys

$$L \geq \bar{C} \frac{b^8}{\nu^4} \log_2^2 \left(n^2/\nu \right) \omega \log^2 d,$$

where $\bar{C}$ is a sufficiently large constant. In addition, by Proposition 8, one has (26) with probability at least $1 - 1/6\mathrm{e}^{-\omega}$, potentially with an increased value of $\bar{C}$. Thus the result follows from Proposition 12 and a union bound over the two probabilities of failure. □

Acknowledgements:

DG and RK are grateful to the organizers and participants of the Workshop on Phaseless Reconstruction, held as part of the 2013 February Fourier Talks at the University of Maryland, where they were introduced to the details of the problem. This extends, in particular, to Thomas Strohmer. RK is pleased to acknolwedge extremely helpful advice he received from Johan Aberg troughout the course of the project. FK thanks the organizers and participants of the AIM workshop "Frame theory intersects geometry", in particular Thomas Strohmer, Götz Pfander, and Nate Strawn, for stimulating conversations on the topic of this paper. The authors also acknowledge inspiring discussions with Emmanuel Candès, Yonina Eldar, and David James. We would also like to thank the anonymous referees for extremely helpful comments and suggestions which allowed us to further improve the presentation of our results.

The work of DG and RK is supported by the Excellence Initiative of the German Federal and State Governments (Grants ZUK 43 & 81), by scholarship funds from the State Graduate Funding Program of Baden-Württemberg, by the US Army Research Office under contracts W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), and the DFG (GRO 4334 & SPP 1798). FK acknowledges support from the German Federal Ministry of Education and Reseach (BMBF) through the cooperative research project ZeMat.

References

[1] E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval from coded diffraction patterns," *Appl. Comput. Harmon. Anal.*, 2014.

[2] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase." *Appl. Comput. Harmon. Anal.*, vol. 20, no. 3, pp. 345–356, 2006.

[3] T. Heinosaari, L. Mazzarella, and M. M. Wolf, "Quantum tomography under prior information." *Commun. Math. Phys.*, vol. 318, no. 2, pp. 355–374, 2013.

[4] Y. C. Eldar and S. Mendelson, "Phase retrieval: Stability and recovery guarantees," *Appl. Comput. Harmon. Anal.*, vol. 36, no. 3, pp. 473–494, 2014.

[5] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, "Painless reconstruction from magnitudes of frame coefficients." *J. Fourier Anal. Appl.*, vol. 15, no. 4, pp. 488–501, 2009.

[6] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imaging Sci.*, vol. 6, no. 1, pp. 199–225, 2013.

[7] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: exact and stable signal recovery from magnitude measurements via convex programming." *Commun. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013.

[8] E. Candès and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Found. Comput. Math.*, pp. 1–10, 2013.

[9] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization." *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[10] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inform. Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[11] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inform. Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[12] R. Kueng, H. Rauhut, and U. Terstiege, "Low rank matrix recovery from rank one measurements," *preprint arXiv:1410.6913*, 2014.

[13] C. Bachoc and M. Ehler, "Signal reconstruction from the magnitude of subspace components," *preprint arXiv:1209.5986*, 2012.

[14] V. Pohl, F. Yang, and H. Boche, "Phase retrieval from low rate samples," *preprint arXiv:1311.7045*, 2013.

[15] A. Conca, D. Edidin, M. Hering, and C. Vinzant, "An algebraic characterization of injectivity in phase retrieval," *Appl. Comput. Harmon. Anal.*, 2014.

[16] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, "Phase retrieval with polarization," *SIAM J. Imaging Sci.*, vol. 7, no. 1, pp. 35–66, 2014.

[17] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[18] M. Ehler, M. Fornasier, and J. Sigl, "Quasi-linear compressed sensing," *Multiscale Model. Simul.*, vol. 12, no. 2, pp. 725–754, 2014.

[19] D. Gross, F. Krahmer, and R. Kueng, "A partial derandomization of PhaseLift using spherical designs," *J. Fourier Anal. Appl.*, pp. 1–38, 2014.

[20] A. Fannjiang and W. Liao, "Phase retrieval with random phase illumination," *JOSA A*, vol. 29, no. 9, pp. 1847–1859, 2012.

[21] A. S. Bandeira, Y. Chen, and D. G. Mixon, "Phase retrieval from power spectra of masked signals," *Inf. Inference*, p. iau002, 2014.

[22] I. J. Good, "The interaction algorithm and practical fourier analysis," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 20, no. 2, pp. pp. 361–372.

[23] M. Rudelson and S. Zhou, "Reconstruction from anisotropic random measurements," *IEEE Trans. Inform. Theory*, vol. 59, no. 6, pp. 3434–3447, 2013.

[24] R. Kueng and D. Gross, "RIPless compressed sensing from anisotropic measurements," *Linear Algebra Appl.*, vol. 441, pp. 110–123, 2014.

[25] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "state tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, no. 15, p. 150401, 2010.

[26] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, "Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators," *New J. Phys.*, vol. 14, no. 9, p. 095022, 2012.

[27] C. Schwemmer, G. Tóth, A. Niggebaum, T. Moroder, D. Gross, O. Gühne, and H. Weinfurter, "Efficient tomographic analysis of a six photon state," *preprint arXiv:1401.7526*, 2014.

[28] K. Banaszek, M. Cramer, and D. Gross, "Focus on quantum tomography," *New J. Phys.*, vol. 15, no. 12, p. 125020, 2013.

[29] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, 1963.

[30] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, ser. Appl. Numer. Harmon. Anal. Birkhäuser, 2013.

[31] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels." *IEEE Trans. Inform. Theory*, vol. 48, no. 3, pp. 569–579, 2002.

[32] J. A. Tropp, "User-friendly tail bounds for sums of random matrices." *Found. Comput. Math.*, vol. 12, no. 4, pp. 389–434, 2012.

[33] J. A. Tropp, "User-friendly tools for random matrices: An introduction," Notes, 2012. [Online]. Available: http://users.cms.caltech.edu/~jtropp/notes/Tro12-User-Friendly-Tools-NIPS.pdf

[34] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

[35] D. Appleby, C. A. Fuchs, and H. Zhu, "Group theoretic, lie algebraic and jordan algebraic formulations of the sic existence problem," *Quantum Inf. Comput.*, vol. 15, no. 1-2, pp. 61–94, 2015.

[36] C. Majenz, R. Kueng, and D. Gross, in preparation, 2015.

[37] D. Gross, K. Audenaert, and J. Eisert, "Evenly distributed unitaries: on the structure of unitary designs." *J. Math. Phys.*, vol. 48, no. 5, pp. 052 104, 22, 2007.

[38] R. Kueng and D. Gross, "Stabilizer states are complex projective 3-designs," in preparation, 2015.

[39] R. Bhatia, *Matrix analysis.* New York, NY: Springer, 1996.

[40] E. J. Candes and Y. Plan, "A probabilistic and ripless theory of compressed sensing," *IEEE Trans. Inform. Theory*, vol. 57, no. 11, pp. 7235–7254, 2011.

[41] B. Adcock and A. C. Hansen, "Generalized sampling and infinite-dimensional compressed sensing," Technical report NA2011/02, DAMTP, University of Cambridge, Tech. Rep., 2011.

[42] M. Habib, C. McDiarmid, J. Ramírez Alfonsín, and B. Reed, Eds., *Probabilistic methods for algorithmic discrete mathematics.* Berlin: Springer, 1998.

6. Appendix

Lemma 19. *Consider as signal the first standard basis vector $e_1 \in \mathbb{C}^d$. Let a_ℓ, $\ell = 1, \dots, m = Ld$. Then for every $\delta > 0$ there exists $c > 0$ such that the following holds for the measurement vectors corresponding to $L < c \log_2 d$ masked Fourier measurements of e_1 as introduced in Section 2.1 with random masks ϵ_ℓ drawn independently at random according to the distribution given in* (7)*. With probability at least $1 - \delta$, there exists another signal that produces the exact same measurements. Thus no algorithm will be able to distinguish these signals based on their measurements.*

Proof. As e_1 as well as any other standard basis vector e_ℓ is 1-sparse, their phaseless measurements corresponding to one mask will just consist of the entry-wise absolute values first (or ℓ-th, respectively) column of the corresponding masked Fourier transform matrix. As all entries of the Fourier transform matrix are of unit modulus, the measurements of e_ℓ are hence completely determined by the vector v_ℓ consisting of the ℓ-th entry of every mask. As a consequence, e_1 and e_ℓ produce the same measurements if the entries of v_1 and v_ℓ have the same absolute value. There are L masks, and each entry's absolute value can be either 0 or $\sqrt{2}$. So there are 2^L possible choices for $|v_\ell|$. For each $\ell > 1$, one of them is drawn uniformly at random. Hence by the coupon collector's problem, a v_ℓ with the same absolute values as v_1 appears again with high probabilty within the first $\Theta(L2^L)$ draws, where by increasing the constant, one can make the probability arbitrarily small. For $L < c\log_2(d)$, we obtain $L2^L < cd^c \log_2(d)$, which for c small enough is less than $d-1$. Thus there will exist another v_ℓ with $|v_\ell| = |v_1|$, which proves the lemma. □

Proof of Lemma 7. We prove formula (21) in a way that is slightly different from the proof provided in [1]. We show that the set of all possible $D_l f_k$'s is in fact proportional to a 2-design and deduce *near-isotropicity* of $\mathcal{R}$ from this. We refer to [19] for further clarification of the concepts used here. Concretely, for $1 \le l \le L$ we aim to show

$$\frac{1}{\nu^2 d} \sum_{k=1}^{d} \mathbb{E}\left[F_{k,l}^{\otimes 2}\right] = 2P_{\mathrm{Sym}^2}, \tag{57}$$

where P_{Sym^2} denotes the projector onto the totally symmetric subspace of $\mathbb{C}^d \otimes \mathbb{C}^d$. Near isotropicity of $\mathcal{R}$ directly follows from (57) by applying [35, Lemma 1] (with $\alpha = \beta = 1$):

$$\mathbb{E}\left[\mathcal{R}\right] Z = \frac{1}{\nu^2 dL} \sum_{k=1}^{d} \sum_{l=1}^{L} \mathbb{E}\left[F_{k,l} \operatorname{tr}(F_{k,l} Z)\right] = \frac{1}{\nu^2 d} \sum_{k=1}^{d} \mathbb{E}\left[F_{k,1} \operatorname{tr}(F_{k,1} Z)\right] = (\mathcal{I} + \Pi_{\mathbb{1}}) Z.$$

So let us proceed to deriving equation (57). We do this by exploring the action of the equation's left hand side on a tensor product $e_i \otimes e_j$ ($1 \leq i, j \leq d$) of two standard basis vectors in $\mathbb{C}^d$. Here it is important to distinguish two special cases, namely $i = j$ and $i \neq j$. For the former we get by inserting standard basis representations

$$\begin{aligned}\frac{1}{\nu^2 d}\sum_{k=1}^{d} \mathbb{E}\left[F_k^{\otimes 2}\right](e_i \otimes e_i) &= \frac{1}{\nu^2 d}\sum_{k=1}^{d} \mathbb{E}\left[\epsilon_i^2 \langle f_k, e_i\rangle^2 D^{\otimes 2}(f_k \otimes f_k)\right] \\ &= \frac{1}{\nu^2}\sum_{a,b=1}^{d} \mathbb{E}\left[\epsilon_i^2 \epsilon_a \epsilon_b\right]\left(\frac{1}{d}\sum_{k=1}^{d}\omega^{k(a+b-2i)}\right)(e_a \otimes e_b) \\ &= \frac{1}{\nu^2}\sum_{a,b=1}^{d} \delta_{(a\oplus b),(2i)}\mathbb{E}\left[\epsilon_i^2 \epsilon_a \epsilon_b\right](e_a \otimes e_b),\end{aligned}$$

where we have used (29) and the fact that for odd d, there is a multiplicative inverse of 2 modulo d. Now $\mathbb{E}[\epsilon_a] = \mathbb{E}[\epsilon_b] = 0$ implies that one obtains a non-vanishing summand only if $a = b$. Therefore one in fact gets

$$\frac{1}{\nu^2 d}\sum_{k=1}^{d} \mathbb{E}\left[F_k^{\otimes 2}\right](e_i \otimes e_i) = \frac{1}{\nu^2}\sum_{a=1}^{d} \delta_{(2a),(2i)}\mathbb{E}\left[\epsilon_i^2 \epsilon_a^2\right](e_a \otimes e_b) = \frac{1}{\nu^2}\mathbb{E}\left[\epsilon_i^4\right](e_i \otimes e_i) = 2(e_i \otimes e_i),$$

where we have used the moment condition (5) in the last step. This however is equivalent to the action of $2P_{\mathrm{Sym}^2}$ on symmetric basis states.

Let us now focus on the second case, namely $i \neq j$. A similar calculation then yields

$$\frac{1}{\nu^2 d}\sum_{k=1}^{d} \mathbb{E}\left[F_k^{\otimes 2}\right](e_i \otimes e_j) \quad = \quad \frac{1}{\nu^2}\sum_{a,b=1}^{d} \mathbb{E}\left[\epsilon_i \epsilon_j \epsilon_a \epsilon_b\right]\delta_{(a+b),(i+j)}(e_a \otimes e_b).$$

Again, $\mathbb{E}[\epsilon] = 0$ demands that the ϵ's have to "pair up". Since $i \neq j$ by assumption, there are only two such possibilities, namely $(i = a, j = b)$ and $(i = b, j = a)$. Both pairings obey the additional delta-constraint and we therefore get

$$\frac{1}{\nu^2 d}\sum_{k=1}^{d} \mathbb{E}\left[F_k^{\otimes 2}\right](e_i \otimes e_j) = \frac{1}{\nu^2}\mathbb{E}\left[\epsilon_i^2 \epsilon_j^2\right](e_i \otimes e_j + e_j \otimes e_i) = (e_i \otimes e_j) + (e_j \otimes e_i),$$

where we have once more used (5) in the final step. This, however is again just the action of $2P_{\mathrm{Sym}^2}$ on vectors $e_i \otimes e_j$ with $i \neq j$. Since the extended standard basis $\{(e_i \otimes e_j)\}_{1 \leq i,j \leq d}$ forms a complete basis of $\mathbb{C}^d \otimes \mathbb{C}^d$, we can deduce equation (57) from this.

□

Proof of Proposition 12. Let X' be an arbitrary feasible point of (9) and we decompose it as $X' = X + \Delta$, where Δ is a feasible displacement. Feasibility then implies $\mathcal{A}(X') = \mathcal{A}(X)$ and consequently $\mathcal{A}(\Delta) = 0$ must hold. The pinching inequality [39] (Problem II.5.4) now implies

$$\|X'\|_1 = \|X + \Delta\|_1 \geq \|X\|_1 + \mathrm{tr}(\Delta_T) + \|\Delta_T^\perp\|_1$$

and X is guaranteed to be the minimum of (9) if

$$\mathrm{tr}(\Delta_T) + \|\Delta_T^\perp\|_1 > 0 \tag{58}$$

is true for any feasible displacement Δ. Therefore it suffices to show that (58) is guaranteed to hold under the assumptions of the proposition. In order to do so, we combine feasibility of Δ with Proposition 8 and Lemma 10 to obtain

$$\|\Delta_T\|_2 < \frac{2}{\sqrt{\nu^2 dL}}\|\mathcal{A}(\Delta_T)\|_{\ell_2} = \frac{2}{\nu\sqrt{dL}}\|\mathcal{A}(\Delta_T^\perp)\|_{\ell_2} \leq \frac{2b^2\sqrt{d}}{\nu}\|\Delta_T^\perp\|_2. \tag{59}$$

Feasibility of Δ also implies $(Y, \Delta) = 0$, because $Y \in \mathrm{range}(\mathcal{A}^*)$ by definition. Combining this insight with (59) and the defining property (38) of Y now yields

$$\begin{aligned} 0 &= (Y, \Delta) = (Y_T - X, \Delta_T) + (X, \Delta_T) + (Y_T^\perp, \Delta_T^\perp) \\ &\leq \|Y_T - X\|_2\|\Delta_T\|_2 + \mathrm{tr}(\Delta_T) + \|Y_T^\perp\|_\infty\|\Delta_T^\perp\|_1 \\ &< \mathrm{tr}(\Delta_T) + \|Y_T - X\|_2 2b^2\sqrt{d}/\nu\|\Delta_T^\perp\|_2 + \|Y_T^\perp\|_\infty\|\Delta_T^\perp\|_1 \\ &\leq \mathrm{tr}(\Delta_T) + 1/2\|\Delta_T^\perp\|_2 + 1/2\|\Delta_T^\perp\|_1 \\ &\leq \mathrm{tr}(\Delta_T) + \|\Delta_T^\perp\|_1, \end{aligned}$$

which is just the optimality criterion (58). □

Qubit stabilizer states are complex projective 3-designs

Richard Kueng[1,2] and David Gross[1,2]

[1]Institute for Theoretical Physics, University of Cologne
[2]Institute for Physics & FDM, University of Freiburg

(Dated: October 12, 2015)

A *complex projective t-design* is a configuration of vectors which is "evenly distributed" on a sphere in the sense that sampling uniformly from it reproduces the moments of Haar measure up to order $2t$. We show that the set of all n-qubit stabilizer states forms a complex projective 3-design in dimension 2^n. Stabilizer states had previously only been known to constitute 2-designs. The main technical ingredient is a general recursion formula for the so-called *frame potential* of stabilizer states. To establish it, we need to compute the number of stabilizer states with pre-described inner product with respect to a reference state. This, in turn, reduces to a counting problem in discrete symplectic vector spaces for which we find a simple formula. We sketch applications in quantum information and signal analysis.

I. INTRODUCTION AND MAIN RESULTS

A. Introduction

In its simplest incarnation, a D-dimensional *complex projctive t-design* is a set of unit-length vectors in $\mathbb{C}^D$ that is evenly distributed on the sphere in the sense that sampling uniformly from this set reproduces the moments of Haar measure up to order $2t$ [1–5] (see Definition 1 below for a precise definition). In a variety of contexts such a design structure is important:

In *numerical integration*, designs are known as *cubatures*. It follows from the definition that the average of a homogeneous polynomial p of order $2t$ over the complex unit sphere equals p's average over the design. If the design has small order, this realization can be made the basis for fast numerical procedures that compute integrals of smooth functions over high-dimensional spheres.

In *quantum information theory*, designs are a widely-employed tool for *derandomizing probabilistic constructions*. Recall that the probabilistic method [6] is a powerful proof technique originally designed to tackle problems in combinatorics. At its core is the observation that the existence of certain extremal combinatorial structures often can be be proved by showing that a suitably chosen random construction would produce an example with high probability. In quantum information, randomized construction often rely on randomly chosen Hilbert space vectors [7]. While this method has brought about spectacular successes (such as the the celebrated proof of strict sub-additivity of entanglement of formation [8]), it suffers e.g. from the problem that generic Haar-random states of large quantum systems are *unphysical*: they cannot be prepared from separable inputs using a polynomial number of operations [9]. Designs, in contrast, *can* be chosen to consist solely of highly-structured and efficiently preparable vectors, while retaining "generic" properties in a precise sense. Thus considerable efforts have been expended at designing complex projective designs (and their unitary cousins) [3, 10–13].

Lastly, randomized constructions in Hilbert spaces have completely classical applications, e.g. in *signal analysis*. Take for instance the highly active field of compressed sensing and related topics [14]: There, one is interested in reconstructing objects that possess some non-trivial structure (e.g. sparsity, or low rank) from a small number of linear measurements. Strong recovery guarantees can be proven for randomly constructed measurement vectors. Once more, this raises the problem of finding sets of structured and well-understood measurements that sufficiently resemble the properties of generic random vectors. The use of designs for this purpose has been proposed in [15–17].

Despite this wealth of applications and non-constructive existence proofs [18], explicit constructions for complex designs remain rare. There are varios infinite families of complex projective 2-designs (e.g. maximal sets of mutually unbiased bases [19, 20], stabilizer states, or symmetric informationally complete POVMs [2]); sporadic solutions for higher orders [11, 21, 22]; and approximate constructions involving random circuits [13]. To the best of our knowledge, an infinite set of explicit complex projective 3-designs has not been identified before.

Here, we show that the set of all *stabilizer states* in dimension 2^n forms a complex projective 3-design for all $n \in \mathbb{N}$.

Recall that the stabilizer formalism is a ubiqutous tool in quantum information theory [9, 23]. Stabilizer states (and, slightly more general, stabilizer codes) are joint eigenvectors of generalized Pauli matrices. Constituting the main realization of quantum error correcting codes [23], they can be efficiently prepared [24] and described in terms of polynomially many parameters [9]. Yet they exhibit non-trivial properties like multi-partite entanglement [25]. Stabilizer states were instrumental in the development of measurement-based quantum computation [26, 27]. In several precise ways, they can be seen as the discrete analogue of Gaussian states [28]. Beyond quantum information, stabilizer states have proved to be versatile enough to provide powerful models for one of the most influential recent development in theoretical condensed mater physics: the study of topological order [29, 30].

Our main result thus identifies yet another aspect according to which stabilizer states capture properties of generic state vectors.

B. Designs and frame potential

In order to state our results more precisely, we need to give a formal definition of complex projective designs and introduce the related notion of *frame potential*. Following [4, 31, 32], we define

Definition 1. *Fix a dimension D and let μ be a probability measure on the unit sphere in $\mathbb{C}^D$. The measure μ is a* complex projective t-design *if, for any order-t polynomial p, we have*

$$\mathbb{E}_{x,y\sim\mu}\left[p\left(|\langle x,y\rangle|^2\right)\right] = \int_{x,y} p\left(|\langle x,y\rangle|^2\right) \mathrm{d}x\mathrm{d}y, \quad (1)$$

where the right-hand-side integration is with respect to the uniform (Haar) measure on the sphere.

In other words, sampling according to μ should give the same expectation values as sampling according to the uniform measure for any random variable that is a polynomial in $|\langle x,y\rangle|^2$ of order at most t. From now on, we will only be concerned with the case where μ is the uniform measure on a finite set of unit vectors.

It is not hard to see that μ fulfills (1) for all polynomials of order t or less, if equality holds for the specific case of $p(z) = z^t$. The resulting value is the t-th order *frame potential* [33]

$$\mathcal{F}_t(\mu) := \mathbb{E}_{x,y\sim\mu}\left[|\langle x,y\rangle|^{2t}\right]. \quad (2)$$

It is known that the Haar integral on the r.h.s. of (1) minimizes the frame potential over the set of all measures μ and that, in fact, its value is given by

$$\mathcal{F}_t(\mu) \geq \mathcal{W}_t(D) := \binom{D+t-1}{t}^{-1}. \quad (3)$$

This relation is known as *Welch bound* [34] or *Sidelnikov inequality* [35]. In summary, we have:

Theorem 1 ([4, 31–33]). *Fix a dimension D and let μ be a probability measure on the unit sphere in $\mathbb{C}^D$. The measure μ is a complex projective t-design if and only if its frame potential meets the Welch bound*

$$\mathcal{F}_t(\mu) = \mathcal{W}_t(D).$$

C. Main results

At the heart of this work is an explicit characterization of the frame potential assumed by the uniform distribution over stabilizer states in prime power dimensions $D = d^n$. We denote the set of stabilizer states on $\left(\mathbb{C}^d\right)^{\otimes n} \simeq \mathbb{C}^D$ by $\mathrm{Stabs}(d,n)$. The unitary symmetry group of the set of stabilizer states is the *Clifford group* (for a precise definition, see Section II C). All results are then implied by the following recursion formula over the dimension's exponent $n = \log_d(D)$.

Theorem 2 (Main Theorem). *Let d be a prime number and let $t \in \mathbb{N}_+$. Then for all dimensions $D = d^n$, the frame potential $\mathcal{F}_t(\mathrm{Stabs}(d,n))$ of stabilizer states in $\mathbb{C}^D$ is determined by the following recursion formula over n:*

$$\mathcal{F}_t(\mathrm{Stabs}(d,1)) = \frac{d^{2-t}+1}{(d+1)d}, \quad (4)$$

$$\frac{\mathcal{F}_t\left(\mathrm{Stabs}(d,n+1)\right)}{\mathcal{F}_t\left(\mathrm{Stabs}(d,n)\right)} = \frac{d^{n-(t-2)}+1}{d\left(d^{n+1}+1\right)}. \quad (5)$$

Comparing this explicit characterization of the frame potential to the Sidelnikov inequality (3) allows us to draw the following conclusions:

Corollary 1. *Let d^n be a prime-power dimension. Then the following statements are true*

1. $\mathrm{Stabs}(d,n)$ *forms a complex projective 2-design.*
2. $\mathrm{Stabs}(d,n)$ *constitutes a complex projective 3-design if and only if $d = 2$.*
3. *The set* $\mathrm{Stabs}(d,n)$ *does not constitute a complex projective 4-design.*
4. *The Clifford group does not act irreducibly on* $\mathrm{Sym}^4(\mathbb{C}^D) \subset \left(\mathbb{C}^D\right)^{\otimes 4}$. *In particular, it is not a unitary 4-design.*

As indicated before, the first fact was already widely known [11, 19, 20]. The other results, however, are new to the best of our knowledge. We reemphasize that these assertions follow immediately form the Main Theorem, which may be of independent interest.

D. Applications and Outlook

Here, we sketch relations of the result to problems from signal analysis and quantum physics. Elaborating on these connections will be the focus of future work.

In *low-rank recovery* [14, 36–38], a low-rank matrix X is to be reconstructed from few linear measurements of the form $y_i = \mathrm{tr}\left(XA_i\right)$. In the *phase retrieval problem* [15, 39, 40] one aims to recover a complex vector $x \in \mathbb{C}^D$ from the absolute value of a small number of measurements $y_i = |\langle x, a_i\rangle|$ that are ignorant towards phase information. This task can be reduced to a particular instance of rank-one matrix recovery by rewriting the measurements as [41, 42]

$$y_i^2 = \operatorname{tr}\left(|x\rangle\langle x|\, |a_i\rangle\langle a_i|\right),$$

i.e. by setting $X = |x\rangle\langle x|$ and $A_i = |a_i\rangle\langle a_i|$. For both problems, strong recovery guarantees for randomly constructed measurements are known. Oftentimes these rely on generic (e.g. Gaussian) measurement ensembles and employing complex projective designs to partially derandomize these result has been proposed in both contexts [15, 16, 43].

Regarding both low rank matrix recovery and phase retrieval, it is known that sampling measurement vectors independently from a 2-design does not do the job [15], while 4-designs already have an essentially optimal performance [43, 44]. However, the remaining intermediate case for $t = 3$ is not yet fully understood. Numerical studies conducted by Drave and Rauhut [45] indicate that random stabilizer-state measurements perform surprisingly well at that task. The combinatorial properties of prime power stabilizer states – e.g. Theorem 2 – may help to clarify this situation. We believe this to be a potentially very insightful open problem.

Finally, we want to point out that one nice structural property of stabilizer states is that they come in bases, i.e. the set of all stabilizer states is a union of different orthonormal bases (see e.g. Theorem 3 below). This allows for a considerably more structured random measurement protocol: Select one such basis at random and iteratively measure the trace inner product of an unknown low rank matrix with all projectors onto the individual basis vectors. After having acquired D data points that way, choose a new stabilizer basis at random and repeat. We refer to [46] for a detailed description of such a protocol. It should be clear that it has immediate applications to *quantum state tomography*. In the above paper, non-trivial recovery statements have been announced for t-designs that admit such a basis structure and have strength $t \geq 3$. Again, stabilizer states obey these criteria and have been used for the numerical experiments conducted there. However the announced recovery statement suffers from a non-optimal sampling rate for 3-designs and the rich combinatorial structure of stabilizer bases might help to amend that situation.

E. Relation to previous work and history

After completion of this work (first announced at the QIP 2013 conference [47]), we became aware of the fact that a close analogue of our main result follows from a statement proved in the field of algebraic combinatorics [48] in 1999. The object of study there is a *real* version of stabilizer states in $\mathbb{R}^{2^n}$, as well as their symmetries, which are given by a real version of the Clifford group. The key result is that under the action of the real Clifford group, the space $\mathrm{Sym}^3(\mathbb{R}^{2^n})$ decomposes into irreps in exactly the same way as it does under the action of the full orthogonal group $O(2^n)$ [48, 49]. This implies [50, 51] that *any* orbit of the real Clifford group gives rise to a set that reproduces moments of Haar measure up to order 6 (the established – if confusing – terminology is to refer to such sets as *spherical* 6*-designs* [1], while the complex-valued analogue would be called a *complex projective* 3*-design* [2]).

The findings of [48] are formulated in the language of algebraic invariant theory. While the present authors were trying to relate them to the results we had established in the context of quantum information, we became aware of yet another development. Huangjun Zhu [52] independently derived a very simple and elegant proof showing that the complex Clifford group in dimensions $d = 2^n$ actually forms a *unitary 3-design* [10, 11]. This means that the the irreducible representation spaces of the action of the Clifford group on $\left(\mathbb{C}^{2^n}\right)^{\otimes 3}$ coincide with those of the full unitary group $U(d)$. In particular, the Clifford group acts irreducibly on $\mathrm{Sym}^3(\mathbb{C}^d)$ which, in turn, implies that that any orbit of the group constitutes a complex projective 3-design. The work of Zhu thus fully implies our main result. What is more, the proof is simpler.

The appeal of the question treated here was underscored even more, when we learned a few days prior to submission of this paper to the arxiv e-print server, that yet another researcher – Zak Webb – had independently obtained results related to the ones of Zhu [53].

In comparision to these works, our proof methods are completely different: We rely on counting structures in discrete symplectic vector spaces in order to compute the angle set between stabilizer states, whereas [48] is based on algebraic invariant theory and [52] on character theory. As a corollary, we derive an expression for the number of stabilizer states with prescribed inner product to a reference state. This finding might be of independent interest. Also, we show that the set of stabilizer states fails to be a 4-design in dimensions 2^n and that stabilizer states in dimensions other than powers of two do not even constitute a 3-design. The simultaneously submitted papers seem to have left this possibility open.

II. PROOF OF THE MAIN STATEMENT

A. Outline

We already mentioned in the introduction that there is a geometric approach to stabilizer states building on the theory of discrete symplectic vector spaces[1]. This *phase space formalism* will be introduced in Section II B. We formally define stabilizer states and explain how to compute inner products in this language in Section II C. We then move on to briefly introducing Grassmannians and some core concepts of discrete symplectic geometry. These tools will be used to establish Theorem 2 in Section III.

B. Phase Space Formalism

We start by considering a d-dimensional Hilbert space $\mathcal{H}$, equipped with a basis $\{|q\rangle \,|\, q \in Q\}$, where the *configuration space* Q is given by $Q := \{0, \ldots, d-1\} \subset \mathbb{Z}$ with arithmetics modulo d. Following [54, 55], we define two phase factors $\tau := e^{\pi i(d^2+1)/d} = (-1)^d e^{\pi i/d}$ and $\omega := \tau^2 = e^{2\pi i/d}$. For $q, p \in Q$, we introduce the *shift* and *boost* operators defined by the relations

$$\text{shift: } \hat{x}(q)|x\rangle = |x+q\rangle, \quad \text{boost: } \hat{z}(p)|x\rangle = \omega^{px}|x\rangle \quad (6)$$

for all $x \in Q$.

For $p, q \in Q$, the corresponding *Weyl operator* (or *generalized Pauli operator*) is defined as

$$w(p,q) = \tau^{-pq}\hat{z}(p)\hat{x}(q). \quad (7)$$

Again following [54, 55], we adopt the convention that any artihmetic expression *in the exponent of* τ is *not* understood to be modulo d, but rather as taking place in the integers. This makes a difference for even dimensions (see below). One could argue that it would be slightly cleaner to syntactically distinguish the modular operations appearing in (6) from the non-modular arithmetic in (7). However, the implicit convention does declutter notation and we feel it is ultimately benefitial.

This definition is consistent with established conventions. For example, one recovers the usual Pauli matrices for the qubit case $d = 2$. We use the notation $V := Q \times Q$ and consequently write $w(v) := w(v_p, v_q)$ for elements $v = (v_p, v_q) \in V$. Furthermore we define the *standard symplectic form*

$$[u, v] := u_p v_q - u_q v_p = u^T J v \quad (8)$$

where

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

and $u = (u_p, u_q), v = (v_p, v_q) \in V$. If d is prime, the space V together with the non-degenerate symplectic product (8) forms a symplectic vector space which is called *phase space* due to its resemblance to the phase space appearing in classical mechanics.

The Weyl operators obey the *composition* and *commutation relations*

$$\begin{aligned} w(u)w(v) &= \tau^{[u,v]}w(u+v), && (9)\\ w(u)w(v) &= \omega^{[u,v]}w(v)w(u) \quad \forall u, v \in V. && (10) \end{aligned}$$

which can be verified by direct computation.

It is worthwhile to point out that for odd d, the ring $\mathbb{Z}_d$ contains a multiplicative inverse of 2, namely $2^{-1} = \frac{1}{2}(d+1) \in \mathbb{Z}_d$. This in particular assures that τ is a dth root of unity and hence the phase factors in (7, 9) depend only on, respectively, pq and $[u, v]$ *modulo* d. In even dimensions, however, τ has order $2d$. This somewhat complicates the theory of stabilizer states in the even-d case – c.f. Section II C.

The preceeding definitions have been made with a single d-dimensional system in mind. We now extend our formalism to n such systems. The corresponding configuration space is $Q = \mathbb{Z}_d^n$ with elements $q = (q_1, \ldots, q_n)$ and $q_i \in \mathbb{Z}_d$. The associated phase space will be denoted by $V := Q \times Q \simeq \mathbb{Z}_d^{2n}$ ($\dim V = 2n$). It carries a symplectic form given by the natural multi-dimensional analogue of (8):

$$[u, v] := u^T J v, \qquad J = \begin{pmatrix} 0_{n\times n} & \mathbb{I}_{n\times n} \\ -\mathbb{I}_{n\times n} & 0_{n\times n} \end{pmatrix}.$$

With elements $(p, q) \in V$, we associate Weyl operators

$$\begin{aligned} w(p,q) =& w(p_1, \ldots, p_n, q_1, \ldots q_n) \\ =& w(p_1, q_1) \otimes \ldots \otimes w_(p_n, q_n) \end{aligned}$$

acting on the tensor product space $(\mathbb{C}^d)^{\otimes n}$. With these definitions, the composition and commutation relations (9, 10) remain valid for $n > 1$.

We conclude this section with two formulas that will be important in what follows and can both be verified immediately. First, the Weyl operators are trace-less, with the exception of the trivial one:

$$\operatorname{tr}(w(v)) = d^n \delta_{v,0}. \quad (11)$$

Second, for any vector $v \in V$ and any subspace $W \subseteq V$ one has

$$\sum_{w \in W} \omega^{[v,w]} = \begin{cases} |W| & \text{if } [v, w] = 0 \ \forall w \in W, \\ 0 & \text{else.} \end{cases} \quad (12)$$

C. Stabilizer States

Here, we will cast the established theory [9, 23] of stabilizer states into the language of symplectic geometry required for our proof. For previous similar expositions, see [28, 56].

[1] This is connected to the fact that stabilizer states are the natural discrete analogue of *Gaussian states* of bosonic systems, where the symplectic structure is well-appreciated. For a concise introduction of this point of view, see [28].

Note that Equation (10) implies that two Weyl operators $w(u)$ and $w(v)$ commute if and only if $[u,v] = 0$. Now consider the image of an entire subspace $M \subseteq V$ under the Weyl representation. We define

$$w(M) = \{w(m) : \ m \in M\}$$

and observe that $w(M)$ consists of mutually commuting operators if and only if $[m,m'] = 0$ holds for all $m,m' \in M$. Spaces having this property are called *isotropic*. Assume now that M is isotropic.

If d is odd, then the $w(M)$ not only commute, but actually form a group $w(u)w(v) = w(u+v)$. That's because in (9), the phase factor depends on $[u,v]$ modulo d, which is zero by assumption for $u,v \in M$. For even dimensions, however, $[u,v]$ might equal d and in this case, the product $w(u)w(v) = -w(u+v)$ does not lie in $w(M)$ (in other words, $v \mapsto w(v)$ is only a *projective* representation of the additive group of M). This would create problems in our analysis below. Fortunately, it turns out that one can choose phases $c(v) \in \{\pm 1\}$ such that $v \mapsto c(v)w(v)$ does become a true representation of M. We will now describe this construction.

To this end, choose a basis $\mathcal{B} = \{u_1, \dots, u_{\dim M}\}$ of M. For a given element $m \in M$, let $m = \sum_i m_i u_i$ be the expansion of m with respect to this basis. Define the (basis-dependent) Weyl operators to be:

$$w_{\mathcal{B}}(m) := \prod_{i=1} w(u_i)^{m_i}. \tag{13}$$

Using the fact that the $w(u_i)$ commute, one then obtains for $m, m' \in M$

$$\begin{aligned} w_{\mathcal{B}}(m)w_{\mathcal{B}}(m') &= \prod_{i=1}^{n} w(u_i)^{m_i} \prod_{i=1}^{n} w(u_i)^{m'_i} \\ &= \prod_{i=1}^{n} w(u_i)^{m_i + m'_i} = w_{\mathcal{B}}(m+m'). \end{aligned}$$

This is the desired representation of M.

Stabilizer states turn out to be related to *maximal* isotropic spaces M. We call a subspace $M \subseteq V$ *Lagrangian* (LAG) – or *maximally isotropic* – if every vector $v \in V$ that commutes with all elements of M is already contained in M. This is precisely the case if

$$M = \{v \in V : \ [v,m] = 0 \ \forall m \in M\} =: M^{\perp},$$

where $M^{\perp}$ denotes the symplectic complement of M. A basic result of symplectic geometry (e.g. Satz 9.11 in [57]) states that this condition is fulfilled if and only if $\dim M = \frac{1}{2} \dim V = n$, or equivalently $|M| = d^n$.

We are now ready to state the relation between Lagrangian subspaces and state vectors in Hilbert space:

Theorem 3 (Stabilizer States)**.** *Let $M \subset V$ be a Lagrangian subspace, let $\mathcal{B}$ be a basis of M. Then the following assertions are valid:*

1. *Up to a global phase, every $v \in M$ singles out one unit vector $|M,v\rangle \in \mathcal{H}$ – called a* stabilizer state *that fulfills the eigenvalue equations*

$$\omega^{[v,m]} w_{\mathcal{B}}(m) |M,v\rangle = |M,v\rangle \quad \forall m \in M. \tag{14}$$

2. *Two elements $u,v \in M$ define the same stabilizer state if and only if they belong to the same affine space $[v]_M := \{v+m, \ m \in M\}$ modulo M. If this is not the case, the resulting stabilizer states are orthogonal, i.e. $\langle M,u|M,v\rangle = 0$.*

3. *V can be decomposed into a union of $d^n = \dim(\mathcal{H})$ different affine spaces modulo M. Via (14), this union defines an orthonormal basis of stabilizer states associated with M.*

This statement implies that each stabilizer state is uniquely characterized by a Lagrangian subspace $M \subset V$ and one particular affine space $[v]_M$ modulo M. In the remainder of this article it will be convenient to represent each such affine space by a representative $\zeta \in [v]_M \in V$ contained in it. We have opted to denote such representatives of cosets $\zeta, \iota \in V$ by greek letters to notationally underline their origin.

Proof of Theorem 3. Define

$$\rho_{M,v} := d^{-n} \sum_{m \in M} \omega^{[v,m]} w_{\mathcal{B}}(m)$$

and compute

$$\begin{aligned} \rho^2_{M,v} =& d^{-2n} \sum_{m,m' \in M} \omega^{[v,m]} \omega^{[v,m']} w_{\mathcal{B}}(m) w_{\mathcal{B}}(m') \\ =& d^{-2n} \sum_{m,m' \in M} \omega^{[v,m+m']} w_{\mathcal{B}}(m+m') \\ =& d^{-n} \sum_{m \in M} \omega^{[v,m]} w_{\mathcal{B}}(m) = \rho_{M,v}, \end{aligned}$$

as well as

$$\begin{aligned} \operatorname{tr} \rho_{M,v} =& d^{-n} \sum_{m, \in M} \omega^{[v,m]} \operatorname{tr} w_{\mathcal{B}}(m) \\ =& d^{-n} \operatorname{tr} w_{\mathcal{B}}(0) = 1 \end{aligned}$$

where we have employed (11). The first relation implies that $\rho_{M,v}$ is a projection and the second one that is has rank one. One can check by direct calculation that

$$\omega^{[v,m]} w_{\mathcal{B}}(m) \rho_{M,v} = \rho_{M,v},$$

holds for every $m \in M$. Consequently, the so that the any vector from the range of $\rho_{M,v}$ fulfills all eigenvalue equations. However, since $\rho_{M,v}$ has rank one, its range corresponds to a single vector that we can associate with $|M,v\rangle \in \mathcal{H}$ up to a global phase. This proves the first claim up to uniqueness which we are going to establish later on.

For the second claim, fix $u, v \in V$ and observe

$$\begin{aligned} \operatorname{tr}\left(\rho_{M,u}\rho_{M,v}\right) =& d^{-2n} \sum_{m,m'\in M} \omega^{[u,m]}\omega^{[v,m']}\operatorname{tr}\left(w_{\mathcal{B}}(m+m')\right) \\ =& d^{-2n}\sum_{m,m'\in M}\omega^{[u,m]}\omega^{[v,m']} d^n \delta_{m+m',0} \\ =& d^{-n}\sum_{m\in M}\omega^{[u-v,m]} \\ =& d^{-n}\begin{cases} |M| & \text{if } [u-v,m]=0\ \forall m\in M,\\ 0 & \text{else,}\end{cases}\end{aligned}$$

where we have used (12). But because M is *maximally* isotropic, $[u-v, m] = 0\ \forall m \in M$ implies $u - v \in M$. Thus, there is one $\rho_{M,u}$ for each affine space $u + M \subset V$, and two distinct affine spaces give rise to othogonal states which is just the second claim.

Finally, note that there are $|V/M| = d^n = \dim \mathcal{H}$ such affine spaces, which proves that one obtains an ortho-normal basis in this way. Moreover, this establishes the uniqueness part of the first statement and implies, justifying that $|M, v\rangle$ is well-defined up to a global phase. □

In the remainder of this section, we will show how to choose consistent bases for two, possibly intersecting, Lagrangian spaces M, N and use these results to come up with formulas for the inner product between two arbitrary stabilizer states.

Lemma 1 (Compatible bases). *Let $M, N \subset V$ be two Lagrangian subspaces. Then there exists bases $\mathcal{B}_M$ of M and $\mathcal{B}_N$ of N such that $w_{\mathcal{B}_K}(m) = w_{\mathcal{B}_M}(m) = w_{\mathcal{B}_N}(m)$ for any $m \in M \cap N$. What is more, for $m \in M$ and $n \in N$, it holds that*

$$\operatorname{tr}\left(w_{\mathcal{B}_M}(m)w_{\mathcal{B}_N}(-n)\right) = d^n\delta_{m,n}. \tag{15}$$

Proof. Choose a basis $\{u_1, \ldots, u_{\dim M\cap N}\}$ of $M \cap N$. By elementary linear algebra, it can be extended both to a basis $\mathcal{B}_M$ of M and to a basis $\mathcal{B}_N$ of N. The first claim follows immediately from (13). For the second claim, note that for from (9), we have that $w_{\mathcal{B}_M}(m)w_{\mathcal{B}_N}(-n) = \pm w(m-n)$. Thus, by (11), the trace in (15) vanishes unless $m = -n$. In that case, however, $m, n \in K$ and thus, by construction of the bases, $w_{\mathcal{B}_M}(m) = w_{\mathcal{B}_K}(m)$ and $w_{\mathcal{B}_N}(-n) = w_{\mathcal{B}_K}(-n)$. Thus

$$w_{\mathcal{B}_M}(m)w_{\mathcal{B}_N}(-n) = w_{\mathcal{B}_K}(m-n) = w_{\mathcal{B}_K}(0) = w(0).$$

The claim then follows from (11). □

We conclude this subsection with an important observation: The overlap of different stabilizer states is fully characterized by the geometric intersection of their underlying Lagrangian subspaces.

Lemma 2 (Overlap of stabilizer states). *Let $|M, \zeta\rangle, |N, \iota\rangle \in \mathcal{H}$ be two stabilizer states characterized by Lagrangian subspaces $M, N \subset V$ (as well as corresponding bases $\mathcal{B}_M$ and $\mathcal{B}_N$ if d is even) and representatives $\zeta, \iota \in V$ of cosets $[\zeta]_M \in V/M$ and $[\iota]_N \in V/N$, respectively. Then, setting $K = M \cap N$, their inner product is given by*

$$|\langle M,\zeta|N,\iota\rangle|^2 = \begin{cases} d^{-n}|K| & \text{if } [\zeta,m] = [\iota,m]\ \forall m\in K,\\ 0 & \text{else.}\end{cases} \tag{16}$$

Proof. The claim follows from direct computation. According to Lemma 1 we can pick bases $\mathcal{B}_K$ of $K := M \cap N$, $\mathcal{B}_M$ of M and $\mathcal{B}_N$ of N that are compatible with each other. With respect to these bases we can write

$$\begin{aligned} |M,\zeta\rangle\langle M,\zeta| &= d^{-n}\sum_{m\in M}\omega^{[\zeta,m]}w_{\mathcal{B}_M}(m), \\ |N,\iota\rangle\langle N,\iota| &= d^{-n}\sum_{m'\in N}\omega^{-[\iota,m']}w_{\mathcal{B}_N}(-m'). \end{aligned}$$

Formula (15) now implies

$$\begin{aligned} & |\langle M,\zeta|N,\iota\rangle|^2 \\ =& \operatorname{tr}\left(|M,\zeta\rangle\langle M,\zeta||N,\iota\rangle\langle N,\iota|\right) \\ =& d^{-2n}\sum_{m\in M}\sum_{m'\in N}\omega^{[\zeta,m]-[\iota,m']}\operatorname{tr}\left(w_{\mathcal{B}_M}(m)w_{\mathcal{B}_N}(-m')\right) \\ =& d^{-n}\sum_{m\in M\cap N}\omega^{[\zeta-\iota,m]} \\ =& d^{-n}\begin{cases}|M\cap N| & \text{if } [\zeta-\iota,m]=0\ \forall m\in M\cap N\\ 0 & \text{else,}\end{cases}\end{aligned}$$

where the last equation follows from formula (12). □

D. Grassmannian subspaces and discrete symplectic geometry

Let Q be a n-dimensional vector space over the finite field $\mathbb{Z}_d$. The *Grassmannian* $\mathcal{G}(d, n, k)$ is the set of k-dimensional subspaces of V. A standard result – e.g formula (9.2.2) in [58] – says that the size of $\mathcal{G}$ is given by the *Gaussian binomial coefficient*:

$$|\mathcal{G}(d,n,k)| = \binom{n}{k}_d := \begin{cases}\prod_{i=0}^{k-1}\frac{d^{n-i}-1}{d^{k-i}-1} & \text{if } k\le n,\\ 0 & \text{else.}\end{cases} \tag{17}$$

This is the analogue of the familiar binomial coefficient for the finite field $\mathbb{Z}_d$. As such it exhibits similar properties, such as $\binom{n}{k}_d = \binom{n}{n-k}_d$ (symmetry), $\binom{n}{n}_d = \binom{n}{0}_d = 1$ (trivial coefficients) and Pascal's identity

$$\binom{n}{k}_d = d^k\binom{n-1}{k}_d + \binom{n-1}{k-1}_d. \tag{18}$$

For further reading and proofs of these identities we refer to Chapter 9 in [58] and move on to introducing some core concepts of symplectic geometry:

Let V be a $2n$-dimensional symplectic vector space over the finite field $\mathbb{Z}_d$. A *polarization* (M, N) of V is the choice of two Lagrangian subspaces M, N which are transverse in the sense that their direct sum spans the entire space, i.e $M \oplus N = V$. For a fixed Lagrangian M we define the set

$$\mathcal{T}(M) = \{N \mid N \text{ Lagrangian; } (M,N) \text{ is a polarization of } V\}$$

of all Lagrangian subspaces transverse to M. The set $\mathcal{T}(M)$ appears in various contexts. For instance it labels all *graph states* (in a sense explain below) in quantum information theory [25]

For the purpose of our counting argument, we need to compute the size of $\mathcal{T}(M) \in V$.

Proposition 1. *Let V be a $2n$-dimensional symplectic space over $\mathbb{Z}_d$ and let M be an arbitrary Lagrangian subspace. Then, the cardinality of $\mathcal{T}(M)$ amounts to*

$$\mathcal{T}(d,n) := |\mathcal{T}(M)| = d^{\frac{1}{2}n(n+1)}.$$

Proof. Fix M and note that a subset $N \subset V$ has to be both Lagrangian and transverse to M in order to lie in $\mathcal{T}(M)$. These conditions can be made more explicit if we choose a basis $b_1, \ldots, b_{2n}$ of V which obeys

$$M = \operatorname{span}\{b_1, \ldots, b_n\} \quad \text{and} \quad [b_i, b_j] = \delta_{n \oplus i, j},$$

where $\oplus$ denotes addition modulo $2n$. Such a basis allows us to fully characterize any subspace N by a $n \times 2n$-generator matrix G_N with column vectors $a_1, \ldots, a_n$ obeying $\operatorname{span}\{a_1, \ldots, a_n\} = N$. Moreover, it will be instructive to partition each generator matrix into two $n \times n$ blocks A and B, i.e. $G_N = \begin{pmatrix} A \\ B \end{pmatrix}$. Due to our choice of basis the generator matrix G_M of M is particularly simple, namely $G_M = \begin{pmatrix} \mathbb{I}_{n\times n} & 0_{n\times n} \end{pmatrix}^T$. Transversality can be restated in terms of these generator matrizes: $M \oplus N = V$ if and only if the $2n \times 2n$-matrix $\begin{pmatrix} G_M & G_N \end{pmatrix}$ has full rank. Due to the particular form of G_M this is however equivalent to demanding $\operatorname{rank}(B) = n$. Thus we can convert G_N into the equivalent generator matrix $\tilde{G}_N = \begin{pmatrix} \tilde{A}^T & \mathbb{I}_{n\times n} \end{pmatrix}^T$ (and generators $\tilde{a}_1, \ldots \tilde{a}_n$ as above) by applying a Gauss-Jordan elimination in the columns of G_N.

The generator matrix $\tilde{G}_N$ characterizes a Lagrangian subspace if and only if $[\tilde{a}_i, \tilde{a}_j] = 0$ holds for all $i, j = 1, \ldots, n$. These requirements can be summarized in a single matrix equality, namely that $\tilde{G}_N^T J \tilde{G}_N$ must identically vanish. Inserting the particular form of $\tilde{G}_N$ and carrying out the math reveals that this is equivalent to demanding that $\tilde{A}^T - \tilde{A}$ must be the zero matrix. Hence, a subspace N is a polarization of M if and only if its generator matrix (with respect to the basis chosen above) is Gauss-Jordan equivalent to $G_N = \begin{pmatrix} A & \mathbb{I}_{n\times n} \end{pmatrix}^T$, where A is a symmetric $n \times n$-matrix over $\mathbb{Z}_d$. Therefore there is a one-to-one correspondence between polarizations N of M and symmetric $n \times n$-matrizes over $\mathbb{Z}_d$. The dimensionality of the latter is $\frac{1}{2}n(n+1)$ which completes the proof. □

The one-to-one correspondence between polarizations of M and symmetric matrices in this proof gives additional meaning to the set $\mathcal{T}(M)$. Recall that a stabilizer state $|N, \zeta\rangle$ is a *graph state* if N possesses a generator matrix of the form $\begin{pmatrix} A & \mathbb{I}_{n\times n} \end{pmatrix}^T$, where A is a symmetric $n \times n$-matrix. Hence, $\mathcal{T}(M)$ is the set of all Lagrangian subspaces N which lead to graph states.

The name graph state pays tribute to the fact that A can be interpreted as the adjacency matrix of a (possibly weighted) graph. Graph states possess a rich structure and many properties of $|N, \zeta\rangle$ can be deduced from the corresponding graph alone. However, here we content ourselves with pointing out the analogy between graph states and $\mathcal{T}(M)$. For further reading we defer the reader to [25].

Let us now turn to subspaces of the symplectic vector space V. It is clear that a proper subspace $W \subset V$ is itself a vector space, however in general it fails to be symplectic. This is due to the fact that the standard symplectic inner product (8) of V becomes degenerate if we restrict it to W. Therefore important tools – such as Proposition 1 – cannot be directly applied to the proper subspace W. However, this problem can be (partly) circumvent by applying a *linear symplectic reduction*. For $W \subsetneq V$ we define the quotient

$$\hat{W} = W/(W^{\perp} \cap W). \tag{19}$$

This space carries the non-degenerate symplectic form

$$\big[[v], [w]\big]_{\hat{W}} := [v, w]_V \tag{20}$$

which is easily seen not to depend on the representatives for $[v]$ and $[w]$. Consequently, the space $\hat{W}$ endowed with $[\cdot,\cdot]_{\hat{W}}$ is a symplectic vector space. We will need such a reduction in the proof of Theorem 4.

III. PROOF OF THE MAIN THEOREM

In this section we show our main result – Theorem 2 – which provides an explicit recursion fully characterizing the frame potential $\mathcal{F}_t(\operatorname{Stabs}(d,n))$ of stabilizer states in prime power dimensions $D = d^n$. We denote the set of all stabilizer states by $\operatorname{Stabs}(d,n) = \left\{x_1, \ldots, x_{S(d,n)}\right\} \subset \mathbb{C}^D$, where $S(d,n) := |\operatorname{Stabs}(d,n)|$ is just the cardinality of that set. Recall that in our framework each stabilizer state $x_i \in \mathbb{C}^D$ is specified by a Lagrangian subspace M in $V = \mathbb{Z}_d^{2n}$ and a representative $\zeta \in V$ of the coset $[\zeta]_M \in V/M$. The Clifford invariance [28] of stabilizer states allows us to calculate any frame potential $\mathcal{F}_t(\operatorname{Stabs}(d,n))$ by counting intersections of Lagrangian subspaces. This is the content of the following result that considerably simplifies the expression for frame potentials.

Lemma 3. *Let $D = d^n$ be a prime power. The t-th frame potential of the set of all stabilizer states in dimension D is* *given by*

$$\mathcal{F}_t(\mathrm{Stabs}(d,n)) = \frac{1}{S(d,n)} \sum_{k=1}^{n} \kappa_M(d,n,k)\, d^{(1-t)(n-k)}, \tag{21}$$

where $\kappa_M(d,n,k)$ is the number of Lagrangian subspaces N whose intersection with an arbitrary fixed Lagrangian subspace M is k-dimensional.

Proof. Stabilizer states constitute an orbit of a particular finite unitary group – the Clifford group. Due to this symmetry, the second summation in $\mathcal{F}_t(\mathrm{Stabs}(d,n))$ is superfluous and we can write

$$\begin{aligned} \mathcal{F}_t(\mathrm{Stabs}(d,n)) =& \frac{1}{S(d,n)^2} \sum_{i,j=1}^{N} |\langle x_i, x_j \rangle|^{2t} \\ =& \frac{1}{S(d,n)} \sum_{i=1}^{N} |\langle x_k, x_i \rangle|^{2t}, \end{aligned} \tag{22}$$

where $x_k \in \mathrm{Stabs}(d,n)$ is an arbitrary fixed stabilizer state. Theorem 3 assures that any such x_k is unambiguously specified by a Lagrangian subspace M of V and coset $[\zeta]_M \in M/V$. Since the choice of x_k in (22) was arbitrary, we can choose $x_k = |M,0\rangle$ – i.e. it is specified by M and the particularly simple representative $0 \in V$ of the coset $[0]_M$. Such a choice of x_k together with Theorem 3 allows us to rewrite (22) as

$$\mathcal{F}_t(\mathrm{Stabs}(d,n)) = \frac{1}{S(d,n)} \sum_{N\ \mathrm{LAG}} \sum_{[\zeta]_N \in V/N} |\langle N,\zeta | M,0\rangle|^{2t}, \tag{23}$$

because instead of summing over stabilizer states, we may as well sum over their characterizing Lagrangian subspaces and cosets instead. Such a reformulation allows us to employ Lemma 2 which implies

$$|\langle N,\zeta|M,0\rangle|^{2t} = \begin{cases} d^{-nt}|K|^t & \text{if } [\zeta,m]=0\ \forall m \in K, \\ 0 & \text{else,} \end{cases}$$

where $K = M \cap N$ denotes the intersection. If this intersection is k-dimensional, $|K| = d^k$ and consequently $|\langle N,\zeta|M,0\rangle|^{2t} = d^{-t(n-k)}$, provided that $[\zeta,m] = 0$ for all elements $m \in K$. This requirement for a non-vanishing overlap is met if and only if $\zeta \in K^{\perp}$. The number of representatives ζ which obey this property (and single out different stabilizer states) is given by the order of the quotient space $|K^{\perp}/N|$. Since $N \subseteq K^{\perp}$ (which follows from $K \subseteq N$ and $N^{\perp} = N$), such a quotient space is well defined and its order amounts to

$$|K^{\perp}/N| = d^{\dim(K^{\perp}/N)} = d^{2n-k-n} = d^{n-k}.$$

Consequently, for each pair of Lagrangians M,N with k-dimensional intersection, d^{n-k} out of a total of d^n stabilizer states specified by N give rise to a non-vanishing overlap $|\langle N,\zeta|M,0\rangle|^{2t} = d^{-t(n-k)}$ with the fixed stabilizer state $x_k = |M,0\rangle$. Inserting this insight into (23) reveals

$$\begin{aligned} \mathcal{F}_t(\mathrm{Stabs}(d,n)) =& \frac{1}{S(d,n)} \sum_{N\ \mathrm{LAG}} d^{(1-t)(n-\dim(N\cap M))} \\ =& \frac{1}{S(d,n)} \sum_{k=1}^{N} \kappa_M(d,n,k)\, d^{(1-t)(n-k)}, \end{aligned}$$

where we have replaced the summation over the different Lagrangian subspaces with an equivalent summation over the dimension k of the intersections $M \cap N$. □

Lemma 3 shows that we can compute the stabilizer frame potential $\mathcal{F}_t(\mathrm{Stabs}(d,n))$ provided that the number $\kappa_M(d,n,k)$ is known for any Lagrangian subspace M and any intersection space dimesion $k \in \{0,\ldots,n\}$. The following two statements characterize that number.

Theorem 4. *Let V be a $2n$-dimensional symplectic space over $\mathbb{Z}_d$. Fix an arbitrary Lagrangian subspace M and a k-dimensional subspace K of M. The number of Lagrangian subspaces N that obey $M \cap N = K$ equals*

$$\mathcal{T}(d,n-k) = d^{\frac{1}{2}(n-k)(n-k+1)}.$$

The fact that each Lagrangian M admits $|\mathcal{G}(d,n,k)| = \binom{n}{k}_d$ different k-dimensional subspaces K (formula (17)) immediately yields the following corollary.

Corollary 2 (Expression for $\kappa_M(d,n,k)$)**.** *Let V be a $2n$-dimensional symplectic space over $\mathbb{Z}_d$. For an arbitrary Lagrangian subspace $M \subset V$ and $k \in \{0,\ldots,n\}$, the number of Lagrangian subspaces N whose intersection with M is k-dimensional amounts to*

$$\kappa_M(d,n,k) = \binom{n}{k}_d d^{\frac{1}{2}(n-k)(n-k+1)}. \tag{24}$$

Proof of Theorem 4. We need to count in how many ways one can choose a Lagrangian space $N \subset V$ that intersects M exactly in K. Our strategy will be to relate the set of such extensions N of K to a set $\mathcal{T}$ as in Proposition 1. To that end, set $\hat{W} := K^{\perp}/K$. Note that $K \subseteq K^{\perp}$ (because $K \subseteq M$ and M is Lagrangian) implies

$$\hat{W} = K^{\perp}/K = K^{\perp}/(K \cap K^{\perp}) = K^{\perp}/((K^{\perp})^{\perp} \cap K^{\perp}).$$

Therefore $\hat{W}$ is the linear symplectic reduction of $K^{\perp}$ as defined in (19). The space $\hat{W}$ endowed with the induced symplectic product $[\cdot,\cdot]_{\hat{W}}$ defined in (20) forms a symplectic vector space with dimension

$$\dim \hat{W} = \dim K^{\perp}/K = 2n-k-k = 2(n-k).$$

Note that any isotropic space N containing K is in particular contained in $K^{\perp}$. The canonical projection $N \mapsto N/K$ sets up a one-to-one correspondence between n-dimensional subspaces of $K^{\perp}$ containing K and $(n-k)$-dimensional subspaces of $\hat{W}$. We need two properties of this correspondence:

(i) $N/K \subset \hat{W}$ is isotropic if and only if $N \subset V$ is. Proof: This follows immediately from (20).

(ii) $N/K \subset \hat{W}$ is transverse to M/K if and only if $M \cap N = K$. Proof: Basic linear algebra shows

$$(M+N)/K \simeq M/K + N/K.$$

For the left hand side:

$$\begin{aligned}\dim(M+N) &= \dim(M) + \dim(N) - \dim(M \cap N)\\ &\leq 2n-k\end{aligned}$$

with equality if and only if $M \cap N = K$. Hence $\dim(M+N)/K \leq 2(n-k)$ with the same condition for equality. For the right hand side:

$$\begin{aligned}\dim(M/K) + \dim(N/K) &\leq \dim M + \dim N - 2\dim K\\ &= 2(n-k)\end{aligned}$$

with equality if and only if the two spaces are transverse.

It follows that M/K is a Lagrangian subspace of $\hat{W}$ and there is a one-to-one correspondence between Lagrangian spaces N intersecting M in K and Lagrangian subspaces of $\hat{W}$ transverse to M/K. Employing Proposition 1 then yields the desired result. □

Finally, we are going to require an explicit characterization of the number $S(d,n)$ of stabilizer states. We borrow it from [28, Corollary 21]:

Proposition 2 (Number of stabilizer states). *For $\mathcal{H} = \left(\mathbb{C}^d\right)^{\otimes n}$, the cardinality $S(d,n)$ of* Stabs$(d,n) \subset \mathcal{H}$ *amounts to*

$$S(d,n) = |\mathrm{Stabs}(d,n)| = d^n \prod_{j=1}^{n} \left(d^j + 1\right) \qquad (25)$$

and thus obeys the recursion

$$\frac{S(d,n)}{S(d,n+1)} = \frac{1}{(d^{n+1}+1)d}. \qquad (26)$$

Formula (25) combined with Corollary 2 allows us to write down the frame potential (Lemma 3) explicitly:

$$\mathcal{F}_t(\mathrm{Stabs}(d,n)) = \frac{1}{S(d,n)} \sum_{k=0}^{n} \binom{n}{k}_d d^{\frac{1}{2}(n-k)(n-k+3-2t)} \qquad (27)$$

with $S(d,n)$ defined in (25). Note that this is a purely combinatorical expression that depends solely on d and n. Analyzing its recursive dependence on n allows us to establish the main result of this work – Theorem 2.

Proof of Theorem 2. Let us start with the base case (4) which is readily established. Indeed, setting $n = 1$ and evaluating formula (27) reveals that for any d and t $\mathcal{F}_t(\mathrm{Stabs}(d,n))$ amounts to

$$\frac{1}{(d+1)d}\left(\binom{1}{0}_d d^{\frac{1}{2}(4-2t)} + \binom{1}{1}_d\right) = \frac{d^{2-t}+1}{(d+1)d},$$

where we have used $\binom{n}{0}_d = \binom{n}{n}_d = 1$. Let us now move on to establishing the recursive behavior. Replacing n by $(n+1)$ in formula (27) and employing Pascal's identity (18) as well as trivial coefficients for Gaussian binomials yields

$$\begin{aligned}\mathcal{F}_t\left(\mathrm{Stabs}(d,n+1)\right) =& \frac{1}{S(d,n+1)} \sum_{k=0}^{n+1} \binom{n+1}{k}_d d^{\frac{1}{2}(n+1-k)(n+1-k+3-2t)}\\ =& \frac{1}{S(d,n+1)} \left(\binom{n+1}{0}_d d^{\frac{1}{2}(n+1)(n+4-2t)} + \sum_{k=1}^{n} \binom{n+1}{k}_d d^{\frac{1}{2}(n+1-k)(n-k+4-2t)} + \binom{n+1}{n+1}_d \right)\\ =& \frac{1}{S(d,n+1)} \left(d^0 \binom{n}{0}_d d^{\frac{1}{2}(n+1)(n+4-2t)} + \sum_{k=1}^{n} \left(d^k \binom{n}{k}_d + \binom{n}{k-1}_d \right) d^{\frac{1}{2}(n+1-k)(n-k+4-2t)} + \binom{n}{n}_d \right)\\ =& \frac{1}{S(d,n+1)} \left(\sum_{k=0}^{n} d^k \binom{n}{k}_d d^{\frac{1}{2}(n+1-k)(n-k+4-2t)} + \sum_{k=1}^{n+1} \binom{n}{k-1} d^{\frac{1}{2}(n-(k-1))(n-(k-1)+3-2t)} \right),\end{aligned} \qquad (28)$$

where we have encorporated the first and last terms in the first and second summation, respectively. Note that the second summation just corresponds to $\sum_{k=0}^{n} \binom{n}{k}_d d^{\frac{1}{2}(n-k)(n-k+3-2t)}$ – which in that very form also appears in (27). Importantly, a similar equivalence is true for the first sum appearing in (28). Taking a closer look at the overall exponent of d in that summation reveals

$$\begin{aligned} & k+\frac{1}{2}(n+1-k)(n-k+4-2t) \\ = & n-(t-2)+\frac{1}{2}(n-k)(n-k+3-2t) \end{aligned}$$

and the first term is independent of the summation index. Consequently the first sum in (28) actually corresponds to $d^{n-(t-2)}\sum_{k=0}^{n}\binom{n}{k}_d d^{\frac{1}{2}(n-k)(n-k+3-2t)}$ and we can conclude

$$\begin{aligned} & \mathcal{F}_t(\mathrm{Stabs}(d,n+1)) \\ = & \frac{S(d,n)}{S(d,n+1)}\left(d^{n-(t-2)}+1\right) \\ & \times\frac{1}{S(d,n)}\sum_{k=0}^{n}\binom{n}{k}_d d^{\frac{1}{2}(n-k)(n-k+3-2t)} \\ = & \frac{d^{n-(t-2)}+1}{d(d^{n+1}+1)}\frac{1}{S(d,n)}\sum_{k=0}^{n}\binom{n}{k}_d d^{\frac{1}{2}(n-k)(n-k+3-2t)} \\ = & \frac{d^{n-(t-2)}+1}{d(d^{n+1}+1)}\mathcal{F}_t\left(\mathrm{Stab}(d,n)\right) \end{aligned}$$

where we have employed (26). □

We conclude this article with presenting a proof of Corollary 1 which establishes some substantial insights into the structure of stabilizer states.

Proof of Corollary 1. Start with the case $t=2$. Then the result of Theorem 2 reads

$$\begin{aligned} \mathcal{F}_2(\mathrm{Stabs}(d,1)) = & \frac{2}{d(d+1)} \\ \frac{\mathcal{F}_2(\mathrm{Stabs}(d,n+1))}{\mathcal{F}_2(\mathrm{Stabs}(d,n))} = & \frac{d^n+1}{d(d^{n+1}+1)}. \end{aligned}$$

But the Welch Bound (3) satisfies identical relations:

$$\mathcal{W}_2(d) = \frac{2}{d(d+1)} \tag{29}$$

$$\frac{\mathcal{W}_2(d,n+1)}{\mathcal{W}_2(d,n)} = \frac{\binom{d^n+1}{2}}{\binom{d^{n+1}+1}{2}} = \frac{(d^n+1)d^n}{(d^{n+1}+1)d^{n+1}} \tag{30}$$

The 3-design case can be proved along similar lines. We have

$$\mathcal{F}_3(\mathrm{Stabs}(d,1)) = \frac{1+d^{-1}}{(d+1)d} \tag{31}$$

$$\frac{\mathcal{F}_3(\mathrm{Stabs}(d,n+1))}{\mathcal{F}_3(\mathrm{Stabs}(d,n))} = \frac{d^{n-1}+1}{d(d^{n+1}+1)} \tag{32}$$

and the Welch bound satisfies

$$\mathcal{W}_3(d) = \binom{d+2}{3}^{-1} = \frac{6}{(d+2)(d+1)d} \tag{33}$$

$$\frac{\mathcal{W}_3(d^{n+1})}{\mathcal{W}_3(d^n)} = \frac{(d^n+2)(d^n+1)}{(d^{n+1}+2)(d^{n+1}+1)d}. \tag{34}$$

The two base values (31) and (33) coincide for $d\leq 2$. Otherwise, the former is strictly larger than the latter. Comparing the recursion factors yields

$$\frac{\text{Eq. (34)}}{\text{Eq. (32)}} = \frac{(d^n+2)(d^n+1)}{(d^{n+1}+2)((d^{n-1}+1)} \tag{35}$$

$$= \frac{d^{2n}+3d^n+2}{d^{2n}+(2/d+d)d^n+2}\leq 1 \tag{36}$$

with equality if and aonly if $d=1,2$. Consequently we have $\mathcal{F}_3(\mathrm{Stabs}(d,n)) = \mathcal{W}_3(d^n)$ for any $n\in\mathbb{N}_+$ if and only if $d\leq 2$.

Finally, let us move on the the 4-design case, where we have

$$\mathcal{F}_4\left(\mathrm{Stabs}(d,1)\right) = \frac{1+d^{-2}}{(d+1)d}, \tag{37}$$

$$\frac{\mathcal{F}_4\left(\mathrm{Stabs}(d,n+1)\right)}{\mathcal{F}_4\left(\mathrm{Stabs}(d,n)\right)} = \frac{d^{n-2}+1}{(d^{n+1}+1)d}, \tag{38}$$

$$\tag{39}$$

and

$$\mathcal{W}_4(d) = \frac{24}{(d+3)(d+2)(d+1)d} \tag{40}$$

$$\frac{\mathcal{W}_4(d^{n+1})}{\mathcal{W}_4(d^n)} = \frac{(d^n+3)(d^n+2)(d^n+1)}{(d^{n+1}+3)(d^{n+1}+2)(d^{n+1}+1)d}. \tag{41}$$

Comparing (37) to (40) reveals $\mathcal{F}_4\left(\mathrm{Stabs}(d,1)\right)\geq\mathcal{W}_4\left(d\right)$ with equality if and only if $d=1$. An analogous relation holds for (38) and (41) which assures that $\mathcal{F}_4(\mathrm{Stabs}(d,n))$ and $\mathcal{W}_4(d^n)$ only ever coincide in the trivial case $d=1$.

For the final claim of Corollary 1, note that the set of stabilizer states in prime-power dimensions form one orbit under the action of the Clifford group [28]. Also, any orbit of a unitary t-design is a complex projective t-design [10, 11]. Thus Claim 3 implies that the Clifford group is not a 4-design. Peter Turner has made us aware of the fact that the frame potential of group orbits only depends on the action of that group on the totally symmetric space $\mathrm{Sym}^t(\mathbb{C}^D)$. Following the reasoning of [11], a group acting irreducibly on that space has the property that any orbit constitutes a complex projective t-design. Thus, the stronger statement in Claim 4 is also implied by Claim 3. □

Acknowledgements: The authors want to thank P. Turner for insightful discussions and H. Zhu, as well as Z. Webb for informing us of their impeding work [52, 53].

The work of DG and RK is supported by the Excellence Initiative of the German Federal and State Governments (Grants ZUK 43 & 81), the ARO under contract W911NF-14-1-0098 (Quantum Characterization, Verification, and Validation), the DFG projects GRO 4334/1,2 (SPP1798 CoSIP), and the State Graduate Funding Program of Baden-Württemberg.

[1] P. Delsarte, J. Goethals, and J. Seidel, "Spherical codes and designs," *Geom. Dedicata*, vol. 6, no. 3, pp. 363–388, 1977.

[2] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, "Symmetric informationally complete quantum measurements," *J. Math. Phys.*, vol. 45, no. 6, pp. 2171–2180, 2004.

[3] A. Ambainis and J. Emerson, "Quantum t-designs: t-wise independence in the quantum world," in *Computational Complexity, 2007. CCC '07. Twenty-Second Annual IEEE Conference on*, pp. 129–140, June 2007.

[4] A. J. Scott, "Tight informationally complete quantum measurements," *J. Phys A*, vol. 39, no. 43, p. 13507, 2006.

[5] W. Matthews, S. Wehner, and A. Winter, "Distinguishability of quantum states under restricted families of measurements with an application to quantum data hiding," *Commun. Math. Phys.*, vol. 291, no. 3, pp. 813–843, 2009.

[6] N. Alon and J. Spencer, *The Probabilistic Method*. Wiley Series in Discrete Mathematics and Optimization, Wiley, 2004.

[7] P. Hayden, D. Leung, P. W. Shor, and A. Winter, "Randomizing quantum states: Constructions and applications," *Commun. Math. Phys.*, vol. 250, no. 2, pp. 371–391, 2004.

[8] M. B. Hastings, "Superadditivity of communication capacity using entangled inputs," *Nat. Phys.*, vol. 5, no. 4, pp. 255–257, 2009.

[9] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information 10th Anniversary Edition*. Cambridge, UK: Cambridge University Press, 2010.

[10] C. Dankert, R. Cleve, J. Emerson, and E. Livine, "Exact and approximate unitary 2-designs and their application to fidelity estimation," *Phys. Rev. A*, vol. 80, p. 012304, 2009.

[11] D. Gross, K. Audenaert, and J. Eisert, "Evenly distributed unitaries: On the structure of unitary designs," *J. Math. Phys*, vol. 48, no. 5, p. 052104, 2007.

[12] R. A. Low, "Large deviation bounds for k-designs," *P. Roy. Soc. Lond. A Mat.*, vol. 465, no. 2111, pp. 3289–3308, 2009.

[13] F. G. Brandao, A. W. Harrow, and M. Horodecki, "Local random quantum circuits are approximate polynomial-designs," *preprint arXiv:1208.0692*, 2012.

[14] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis, Basel, Switzerland: Birkhäuser, 2013.

[15] D. Gross, F. Krahmer, and R. Kueng, "A partial derandomization of PhaseLift Using spherical designs," *J. Fourier Anal. Appl.*, vol. 21, no. 2, pp. 229–266, 2015.

[16] M. Ehler, M. Graef, and F. J. Kiraly, "Phase retrieval using random cubatures and fusion frames of positive semidefinite matrices," *preprint arXiv:1505.05003*, 2015.

[17] R. Kueng, D. Gross, and F. Krahmer, "Spherical designs as a tool for derandomization: the case of PhaseLift," in *11th International Conference on Sampling Theory and Applications (SampTA 2015)*, (Washington, USA), May 2015.

[18] A. V. Bondarenko and M. S. Viazovska, "Spherical designs via brouwer fixed point theorem," *SIAM J. Discrete Math.*, vol. 24, no. 1, pp. 207–217, 2010.

[19] A. Klappenecker and M. Rotteler, "Mutually unbiased bases are complex projective 2-designs," in *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, pp. 1740–1744, Sept 2005.

[20] I. Bengtsson and K. Zyczkowski, *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge, UK: Cambridge University Press, 2006.

[21] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*. New York, USA: Springer Science & Business Media, 1999.

[22] C. Bachoc and B. Venkov, "Modular forms, lattices and spherical designs.," in *Euclidean lattices, spherical designs and modular forms. On the works of Boris Venkov*, pp. 87–111, Genève: L'Enseignement Mathématique, 2001.

[23] D. Gottesman, *Stabilizer codes and quantum error correction*. PhD thesis, California Institute of Technology, Pasadena, CA, 1997.

[24] E. Hostens, J. Dehaene, and B. De Moor, "Stabilizer states and Clifford operations for systems of arbitrary dimensions and modular arithmetic," *Phys. Rev. A*, vol. 71, p. 042315, Apr 2005.

[25] M. Hein, W. Dür, J. Eisert, R. Raussendorf, M. Nest, and H.-J. Briegel, "Entanglement in graph states and its applications," in *Proceedings of the International School of Physics Enrico Fermi (Varenna, Italy), Quantum computers, algorithms and chaos 162*, (IOS Press), 2006.

[26] R. Raussendorf and H. J. Briegel, "A one-way quantum computer," *Phys. Rev. Lett.*, vol. 86, pp. 5188–5191, May 2001.

[27] D. Gross and J. Eisert, "Novel schemes for measurement-based quantum computation," *Phys. Rev. Lett.*, vol. 98, p. 220503, May 2007.

[28] D. Gross, "Hudsons theorem for finite-dimensional quantum systems," *J. Math. Phys.*, vol. 47, no. 12, p. 122107, 2006.

[29] A. Kitaev, "Fault-tolerant quantum computation by anyons," *Ann. Phys.*, vol. 303, no. 1, pp. 2 – 30, 2003.

[30] B. Zeng, X. Chen, D.-L. Zhou, and X.-G. Wen, "Quantum information meets quantum matter–from quantum entanglement to topological phase in many-body systems," *preprint arXiv:1508.02595*, 2015.

[31] V. Levenshtein, "On designs in compact metric spaces and a universal bound on their size," *Discrete Math.*, vol. 192, no. 13, pp. 251 – 271, 1998.

[32] H. König, "Cubature formulas on spheres," in *Advances in Multivariate Approximation: Proceedings of the Third Internation Conference on Multivariate Approximation Theory*, vol. 107 of *Math. Res.*, (Berlin, Germany), pp. 201 – 211, Wiley-VCH, 1999.

[33] J. Benedetto and M. Fickus, "Finite normalized tight frames," *Adv. Comput. Math.*, vol. 18, no. 2-4, pp. 357–385, 2003.

[34] L. Welch, "Lower bounds on the maximum cross correlation of signals (corresp.)," *IEEE Trans. on Inf. Theory*, pp. 397–399, 1974.

[35] V. Sidelnikov, "Upper bounds on the cardinality of a binary code with a given minimum distance," *Inform. Control*, vol. 28, no. 4, pp. 292 – 303, 1975.

[36] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[37] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, p. 150401, Oct 2010.

[38] D. Gross, "Recovering low-rank matrices from few co-

efficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, pp. 1548–1566, March 2011.

[39] A. Walther, "The question of phase retrieval in optics," *Journal of Modern Optics*, vol. 10, no. 1, pp. 41–49, 1963.

[40] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: exact and stable signal recovery from magnitude measurements via convex programming.," *Commun. Pure Appl. Math.*, vol. 66, pp. 1241–1274, 2013.

[41] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, "Painless reconstruction from magnitudes of frame coefficients.," *J. Fourier Anal. Appl.*, vol. 15, pp. 488–501, 2009.

[42] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Review*, vol. 57, no. 2, pp. 225–251, 2015.

[43] R. Kueng, H. Rauhut, and U. Terstiege, "Low rank matrix recovery from rank one measurements," *Appl. Comput. Harmon. Anal.,* doi:10.1016/j.acha.2015.07.007, 2015.

[44] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, "Stable low-rank matrix recovery via null space properties," *preprint arXiv:1507.07184*, 2015.

[45] I. H. Drave, "Numerische Experimente zur Niedrigrang-Rekonstruktion." Bachelor thesis, 2015. RWTH Aachen, Germany.

[46] R. Kueng, "Low rank matrix recovery from few orthonormal basis measurements," in *11th International Conference on Sampling Theory and Applications (SampTA 2015)*, (Washington, USA), May 2015.

[47] R. Kueng and D. Gross, "Qubit stabilizer states are spherical 3-designs," in *16th Workshop on Quantum Information Processing (Poster)*, (Beijing, China), 2013.

[48] V. Sidelnikov, "Spherical 7-designs in 2^n -dimensional Euclidean space," *J. Algebraic Combin.*, vol. 10, no. 3, pp. 279–288, 1999.

[49] G. Nebe, E. M. Rains, and N. J. A. Sloane, *Self-dual codes and invariant theory*, vol. 17. Berlin, Germany: Springer, 2006.

[50] J. Goethals and J. Seidel, "Spherical designs," in *P. Symp. Pure Math*, vol. 34, pp. 255–272, 1979.

[51] E. Bannai, "On some spherical t-designs," *J. Combin. Theory Ser. A*, vol. 26, no. 2, pp. 157 – 161, 1979.

[52] H. Zhu, "Multiqubit Clifford groups are unitary 3-designs," *to appear on arXiv*, 2015.

[53] Z. Webb, "The Clifford group forms a unitary 3-design," *to appear on arXiv*, 2015.

[54] N. De Beaudrap, "A linearized stabilizer formalism for systems of finite dimension," *Quant. Inf. Comp.*, vol. 13, pp. 73 – 115, 2013.

[55] D. M. Appleby, "Symmetric informationally completepositive operator valued measures and the extended clifford group," *J. Math. Phys.*, vol. 46, no. 5, p. 052107, 2005.

[56] D. Gross and M. Walter, "Stabilizer information inequalities from phase space distributions," *Journal of Mathematical Physics*, vol. 54, no. 8, p. 082201, 2013.

[57] B. Huppert, *Endliche Gruppen: Vol.: 1*. Berlin, Germany: Springer-Verlag, 1967.

[58] P. J. Cameron, *Combinatorics: topics, techniques, algorithms*. Cambridge, UK: Cambridge University Press, 1994.

Low rank matrix recovery from few orthonormal basis measurements

(Invited Paper)

Richard Kueng
Institute for Physics & FDM
University of Freiburg, and
School of Physics
The University of Sydney
richard.kueng@physik.uni-freiburg.de

***Abstract*—Recent insights concerning the *PhaseLift* algorithm for retrieving phases have furthered our understanding of low rank matrix recovery from rank-one projective measurements. Motivated by the structure of certain quantum mechanical experiments, we introduce a particular class of such rank-one measurements: *orthonormal basis measurements*. One such measurement corresponds to choosing an orthonormal basis and treating all the rank-one projectors onto different basis elements as a series of consecutive measurement matrices. We elaborate on performing low-rank matrix recovery from few, sufficiently random orthonormal basis measurements and sketch applications of such a procedure in quantum physics. We conclude this article by presenting numerical experiments testing such an approach.**

***Index Terms*—Low rank matrix recovery, quantum information theory, phase retrieval**

I. Introduction

A. Low rank matrix recovery

The young but already extensive field of low rank matrix recovery uses ideas from compressed sensing to reconstruct a given matrix of low rank from highly incomplete data in a computationally efficient way. Here we shall restrict our attention to hermitian $n \times n$ matrices which form a real n^2-dimensional vector space H^n. Let $X \in H^n$ be a rank-r matrix of interest ($r \ll n$) and suppose that we have access to m linear measurements of the form

$$y_i = \operatorname{tr}(A_i X) \quad i = 1, \ldots, m, \tag{1}$$

where $A_1, \ldots, A_m \in H^n$ denote measurement matrices. Having data of this form at hand, the analogy to compressed sensing [1] suggests to exploit the low-rank structure of X by means of a constrained nuclear-norm[1] minimization

$$\begin{aligned} \underset{Z \in H^n}{\text{minimize}} \quad & \|Z\|_* \\ \text{subject to} \quad & \operatorname{tr}(Z A_i) = y_i \quad i = 1, \ldots, m, \end{aligned} \tag{2}$$

which can be solved computationally efficiently. One aim of low-rank matrix recovery is to identify instances for which $m = Crn\text{polylog}(n)$ measurements of the form (1) suffice to prove that the convex program (2) recovers the sought for X with high probability. Up to date many such instances have been identified [2]–[8].

B. The phase retrieval problem

The problem of retrieving a complex signal $x \in \mathbb{C}^n$ from measurements of the form

$$y_i = |\langle a_i, x \rangle|^2 \quad i = 1, \ldots, m, \tag{3}$$

where $a_1, \ldots, a_m \in \mathbb{C}^n$ are measurement vectors, has long been abundant in many areas of science. Recently, its mathematical structure has received considerable attention in its own right. The problem is clearly ill-posed, since all phase information is lost in the measurement process and the measurements (3) are furthermore of a non-linear nature. This second obstacle can be overcome by a trick [9] well known in conic programming: the quadratic expressions (3) are linear in the outer products xx^* and $a_i a_i^*$:

$$y_i = |\langle a_i, x \rangle|^2 = \operatorname{tr}\left((a_i a_i)^* (xx^*) \right).$$

Since the object of interest – $X := xx^* \in H^n$ – is proportional to a rank-one projector, such a "lift" turns the phase retrieval problem into a particular instance of low rank matrix recovery – a fact that was first observed by Candès, Eldar, Strohmer and Voroninski [10]. In turn, the measurement matrices $A_i = a_i a_i^*$ are constrained to be proportional to rank-one projectors as well. These structural constraints prevent a direct application of results from low-rank matrix recovery, because signal and measurements fail to be sufficiently *incoherent*[2] in the sense of [5], [11]. Nonetheless, phase retrieval recovery guarantees by means of the optimization (2) – dubbed *PhaseLift* for this particular setting – have been established for different types of measurements.

The chronologically first result [12] of this kind proves a non-uniform recovery guarantee for $m = Cn\log(n)$ measurement vectors a_i sampled independently and uniformly from the complex unit sphere. This recovery guarantee was partially derandomized (at the cost of a larger sampling rate) in [13] using the concept of spherical t-designs. Both results were improved by means of uniform counterparts [14], [15] getting by with lower sampling rates[3].

Motivated by actual experimental setups, Candès, Li and Soltanolkotabi [16] furthermore established a non-uniform recovery guarantee for $L = C\log^4(n)$ complete Fourier basis measurements

[1] In a sense, the nuclear norm $\|X\|_* = \operatorname{tr}(|X|)$ is the natural non-commutative analogue of the ℓ_1-norm which features prominently in compressed sensing [1]. Furthermore, low rank assures that the matrix of interest is sparse in its eigenbasis.

[2] Roughly, the incoherence paramter caputures the well-posedness of the inverse problem.

[3] In fact, both references establish an optimal sampling rate (up to a multiplicative constant) for measurement vectors drawn independently and uniformly from the complex unit sphere.

978-1-4673-7353-1/15/$31.00 ©2015 IEEE

that are randomly distorted. To be more concrete, one such measurement encompasses n distorted Fourier vectors of the form $a_{k,l} = D_l f_k$ ($1 \leq i \leq k$), where each D_l ($1 \leq l \leq L$) is an instance of a random matrix diagonal in the standard basis – e.g. a diagonally embedded Rademacher vector with random erasures. Subsequently, a recovery guarantee requiring fewer – namely $L = C \log^2(n)$ – such coded diffraction patterns was established in [17].

We conclude this section by pointing out that PhaseLift is just one possibility for solving the phase retrieval problem. Other approaches rely on polarization identities [18], alternate projections [19], or Wirtinger flow methods [20].

II. Low rank matrix recovery from orthonormal basis measurements

Given these recent advances regarding the phase retrieval problem, it seems natural to ask, whether these insights can be translated to general low rank matrix recovery from certain types of rank-one projective measurements. A first step in this direction was done in [15], where uniform recovery guarantees for inferring hermitian rank r-matrices from $m = Crn$ projectors onto i.i.d Gauss-random vectors, or from $m = Crn \log(n)$ projectors onto randomly chosen elements of a spherical 4-design, were established.

Here – inspired by coded diffraction patterns [16], [17] – we shall focus on randomly distorted basis measurements instead. More formally: let $X \in H^n$ be a rank-r matrix of interest and consider n consecutive measurements of the form

$$\begin{aligned} y_{1,l} &= \operatorname{tr}\left(\left(D_l b_1 b_1^* D_l^* \right) X \right) = \langle D_l b_1, X \; D_l b_1 \rangle, \\ &\vdots \\ y_{n,l} &= \operatorname{tr}\left(\left(D_l b_n b_n^* D_l^* \right) X \right) = \langle D_l b_n, X \; D_l b_n \rangle, \end{aligned} \tag{4}$$

where $b_1, \ldots, b_n$ denotes an arbitrary orthonormal basis of $\mathbb{C}^n$ and D_l is an instance of a random $n \times n$ matrix. Motivated by typical quantum mechanical experiments – see Section III – we consider the special case, where each D_l is unitary. Consequently each distorted orthonormal basis measurement corresponds to measuring a different orthonormal basis $b_1^{(l)}, \ldots, b_n^{(l)}$:

$$\begin{aligned} y_{1,l} &= \operatorname{tr}\left(b_1^{(l)} \left(b_1^{(l)} \right)^* X \right) = \langle b_i^{(l)}, X \; b_1^{(l)} \rangle, \\ &\vdots \\ y_{n,l} &= \operatorname{tr}\left(b_1^{(l)} \left(b_1^{(l)} \right)^* X \right) = \langle b_n^{(l)}, X \; b_n^{(l)} \rangle. \end{aligned} \tag{5}$$

Regarding such types of measurements, the following question is imminent:

> *Are there unitary transformations D_l – or equivalently: orthonormal bases $b_1^{(l)}, \ldots, b_n^{(l)}$ – such that the convex optimization* (2) *allows for recovering an unknown rank-r matrix X from $L = Cr\mathrm{polylog}(n)$ orthonormal basis measurements of the form* (4)*, or* (5)*, respectively?*

It is highly conceivable, that this is the case for unitaries D_l chosen uniformly from the Haar measure – or equivalently: orthonormal bases $b_1^{(l)}, \ldots, b_n^{(l)}$ obtained by choosing n standard complex Gaussian vectors independently at random and orthonormalizing them (e.g. by means of Gram-Schmitt). Clearly, such a procedure requires one to be able to choose from a continuous, very generic union of bases. However, the results in [13], [15] suggest that such a requirement might not be necessary and that more structured, finite unions of bases may suffice to establish low rank matrix recovery guarantees by means of nuclear norm minimization. For going further into that direction – and, by doing so, partially derandomizing the recovery scheme proposed above – we rely on the concept of spherical designs. These finite sets of unit vectors were first introduced in [21] and serve as a general purpose tool for partial derandomization – see [13], [22] for further reading on this aspect of spherical designs. To mimic the problem's structure, we need to equip spherical designs with an additional structural property. This results in the following definition.

Definition 1 (spherical t-design with basis structure)**.** We call a finite union $\Lambda_t = \{b_1^{(i)}, \ldots, b_n^{(i)}\}_{i=1}^N \subset \mathbb{C}^n$ of orthonormal bases a *spherical t-design with basis structure*, if the uniform distribution over the Nn elements of Λ_t reproduces the first $2t$ moments of standard complex Gaussian vectors renormalized to unit length.

Although demanding such an orthonormal basis structure in addition to the t-design property might seem alien at first sight, there are numerous examples for designs that admit it. Examples include arbitrary orthonormal bases (1-designs), maximal sets of mutually unbiased bases (2-designs) [23] which exist in prime power dimensions, stabilizer states (3-designs in power-of-two-dimensions) [24] and orbits $\{U_i b_1, \ldots, U_i b_n\}_{i=1}^N$ of an arbitrary orthonormal basis under the action of a unitary t-design[4] $\{U_i\}_{i=1}^N$ [25], [26] (which constitute a spherical t-designs of the same order).

Similar to [13], t-designs with basis structure suffice to establish a non-uniform recovery guarantee for measurements of the form (5):

Theorem 2 (Low rank matrix recovery from orthonormal basis measurements)**.** *Let $X \in H^n$ be an arbitrary matrix of rank r and let Λ_t be a t-design ($t \geq 3$) with basis structure in the sense of Definition 1. Then choosing*

$$L = Ctn^{2/t} r \log^2(n) \tag{6}$$

different bases independently and uniformly at random from Λ_t and performing the corresponding orthonormal basis measurements of the form (5) *suffices to recover X by means of the convex optimization* (2) *with high probability.*

Note that the requirement $t \geq 3$ on the design order is in fact necessary. Similar to [13, Theorem 2], a counter-example can be constructed for $t = 2$ using mutually unbiased bases. While Theorem 2 is non-trivial – as it allows for recovering X from a total of $m = Ln = Ctn^{1+2/t} r \log^2(n) \ll n^2$ measurements (provided that $r \ll n$ and n is large enough) – the required number L of orthonormal basis measurements contains the term $n^{2/t}$. As a consequence, the sampling rate only becomes optimal up to polylog-factors, if we allow the design order t to grow logarithmically with the dimension ($t = 2\log(n)$). However, we believe that the factor $n^{2/t}$ in (6) is an artifact of the proof technique employed. It uses ideas presented in [13] which resulted in a similar non-optimal factor appearing in the sampling rate. There, employing different techniques allowed for eradicating such a factor and substantially strengthening the statement [15]. In turn, we believe that a more careful analysis will allow for

[4] Unitary t-design are a natural generalization of the spherical design concept to unitaries. These finite sets of unitary matrices reproduce the Haar-uniform distribution over the unitary group up to $2t$-th moments.

establishing a recovery guarantee getting by with a sampling rate that already for $t = 3$, or $t = 4$, is optimal up to $\mathrm{polylog}$-factors.

Furthermore, we want to point out that there is crucial difference between Theorem 2 and the main result established in [13] (and its generalization presented in [15]). There it is assumed that each measurement is sampled independently and uniformly from Λ_t – which strongly resembles the design's defining property (see Definition 1). In Theorem 2, on the other hand, one entire basis is selected at random and n corresponding orthonormal basis measurements of the form (5) are carried out. Evidently, these n measurements are correlated. Such a situation bears more similarity to coded diffraction patterns [16], [17] than it does to independent sampling of individual design elements. In order to establish Theorem 2 we pay tribute to this fact and combine proof techniques from [17] (which can handle such correlated measurements) with others from [13] (that exploit the underlying design-structure). A detailed presentation of such a proof would go beyond the scope of this article and will be presented elsewhere.

Finally we want to point out that Theorem 2 is stated for noiseless measurements only. We leave establishing stability towards noise for future work.

III. Motivation: Quantum state tomography

In this section, we briefly want to motivate the measurement setups introduced in (4) and (5) without going into too much detail. For further reading on the topics introduced here, we defer the interested reader to [27, Chapter 2.2]. In quantum mechanics, the state of an isolated finite n-dimensional quantum system is fully described by a positive-semidefinite hermitian matrix $\rho \in H^n$ with unit trace. Such a matrix is called a *density operator*. Estimating the density operator of a given quantum system is an important task in quantum physics known as *quantum state tomography* [28], [29]. When performing this task, it is highly desirable to exploit additional structure – if present – especially when n is large[5]. One such structural property – often encountered in actual experiments – is *approximate purity*, i.e. the density operator ρ is well approximated by a low rank matrix. Performing quantum state tomography under the prior assumption of approximate purity therefore constitutes a particular instance of low rank matrix recovery [31], [32].

The dynamics of an isolated quantum system – i.e. some physical evolution – corresponds to a unitary transformation $\rho \mapsto \rho' = U \rho U^*$ of the system's density operator ρ.

Finally, after preparing a quantum system ρ and letting it undergo some physical evolution U, a typical experiment is terminated by performing a measurement on the resulting system ρ'. While substantially more general types of measurements are possible, *non-degenerate projective measurements* constitute a particular important subclass. Each such measurement is described by a non-degenerate hermitian matrix $M = \sum_{i=1}^n \lambda_i b_i b_i^*$ with eigenvalues $\lambda_i \in \mathbb{R}$ and a corresponding orthonormal eigenbasis $\{b_1, \ldots, b_n\} \subset \mathbb{C}^n$. Upon performing such a measurement on a system described by ρ, quantum mechanics postulates that the probability of obtaining the outcome λ_i is given by

$$p(\lambda_i, \rho) = \mathrm{tr}\left(b_i b_i^* \rho\right) = \langle b_i, \rho\, b_i \rangle .$$

Repeating such an experiment (i.e. preparing ρ and measuring M) many times allows one to estimate the n probabilities $p(\lambda_i, \rho)$ ever more accurately.

Consequently, combining a certain unitary evolution U of a density operator ρ with performing a non-degenerate projective measurement $M = \sum_i \lambda_i b_i b_i^*$, results in estimating n numbers of the form

$$p\left(\lambda_i, \rho'\right) = \mathrm{tr}\left(b_i b_i^* U X U^*\right) = \langle U^* b_i, X U^* b_i \rangle \quad i = 1, \ldots, n. \quad (7)$$

This exactly corresponds to one orthonormal basis measurement introduced in (4) (with $D_l = U^*$) and the corresponding experiment is sketched in Figure 1. With such a setup at hand, Theorem 2 yields the following corollary relevant for quantum state tomography.

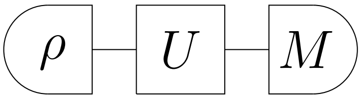


Fig. 1. Pictorial description of a typical quantum mechanical experiment. In a first step, a quantum system described by a density operator ρ is produced. The system then undergoes some physical evolution characterized by a unitary matrix U: $\rho \mapsto \rho' = U \rho U^*$. The experiment is concluded by performing a measurement M. If $M = \sum_{i=1}^n \lambda_i b_i b_i^*$ is a non-degenerate projective measurement, information about ρ' can be gained via (7) by repeating the experiment many times and inferring the probabilities $p(\lambda_i, \rho')$ of the individual measurement outcomes λ_i occurring.

Corollary 3 (Quantum state tomography from sufficiently random evolutions). *Let ρ be a density operator of rank $r \leq n$ and let $M = \sum_{i=1}^n \lambda_i b_i b_i^*$ denote a fixed non-degenerate projective measurement. Then, $L = C r \log^3(n)$ independent instances of the basic experimental protocol described in Figure 1 suffice to recover ρ via* (2) *with high probability, provided that the unitary evolutions are chosen from a sufficiently generic set – e.g. a unitary $2\log(n)$-design.*

Some remarks on the practicality of the protocol presented in Corollary 3 may be appropriate: The postulates of quantum mechanics demand that each instance of the scenario depicted in Figure 1 needs to be repeated many times in order to infer the resulting probability distribution. This obstacle is of a fundamental nature and cannot easily be overcome. However, when it comes to imposing evolutions, some unitaries are considerably more challenging to realize than others. While the effort for implementing a generic Haar-random unitary evolution is considerable, implementing an evolution corresponding to a random element of a weighted, approximate unitary t-design can be done much more easily [33]. Practicality issues of this type were our main motivation for focusing on t-designs with basis structure, as they include orbits $\{U_i b_1, \ldots, U_i b_n\}_{i=1}^N$ of any orthonormal basis under a the action of a unitary t-design as a special case. Consequently, the results in [33] assure that the L different instances of the experiment proposed in Corollary 3 can be implemented in a practical way[6].

Note that Corollary 3 is not the first approach to use low rank matrix recovery techniques for quantum state tomography. Up to now, recovery of approximately pure density operators by means of the convex optimization problem (2) has been established for

[5] Nowadays, experimental physicists are able to create and control multi-partite quantum systems of overall dimension $n = 2^8$ in their labs [30]. This results in a density operator of size 256×256 (a priori 65 536 parameters).

[6] Technically, this conclusion is only valid if Theorem 2 remains true for weighted, approximate t-designs with basis structure. That this is indeed the case, will be established elsewhere.

independently chosen (generalized) Pauli measurements [5], [31] which can be implemented in a practical way for various experimental setups. For this type of measurements, the statistics is well understood [32], uniform recovery guarantees have been established [6] and the procedure has been tested in experiments [34]. However, all the existing results manifestly require performing at least $m = C'rn\log(n)$ independently chosen Pauli-type measurements, each of which can be interpreted as a highly degenerate projective measurement[7]. Here, we propose and establish a novel approach that goes beyond the Pauli setting and exploits a much more fine-grained measurement outcome statistics. Arguably, our protocol requires a more complicated experimental setup and the theoretical assertions are weaker (so far), but it gets by with only $L = Cr\log^3(n)$ different measurement settings.

IV. Numerical Experiments

Finally, we complement our theoretical observations and claims with numerical experiments. These were implemented in Matlab, using CVX [35]. To this end, we used *stabilizers states* [27, Chapter 10.5] – a highly structured union of orthonormal bases that forms a 3-design in power-of-two-dimensions [24] (this is false for other dimensions). Due to their rich combinatorial structure, choosing one stabilizer basis independently at random can be implemented efficiently and we have used this in our numerical simulations. The results for dimensions $n = 16$ and $n = 32$ are depicted in Figure 2. In each case we ran a total of 30 independent experiments for matrix ranks between 1 and $3n/4$ (x-axis) and the number L of measured stabilizer bases ranging from 1 to 70 and 1 to 120, respectively (y-axis). For each experiment we first constructed a rank-r test matrix $X = \sum_{i=1}^{r} v_i v_i^*$, where each $v_i \in \mathbb{C}^n$ was a standard Gaussian random vector and renormalized X to Frobenius norm one. We then chose L stabilizer bases uniformly at random and for each such basis, we evaluated the n measurement outcomes $y_{1,l}, \ldots, y_{n,l}$ according to (5). Using these Ln data points, we ran the convex optimization (2) and declared the recovery a "success" if the Frobenius-norm distance between the reconstructed matrix $X^\sharp$ and the true test signal X was smaller than 10^{-3}. Figure 2 illustrates the resulting empirical success probability for dimensions $n = 16$ and $n = 32$: black corresponds to only failures, white to exclusively successes.

Acknowledgements

My work is supported by scholarship funds from the State Graduate Funding Program of Baden-Württemberg, the Excellence Initiative of the German Federal and State Governments (Grant ZUK 43), the ARO under contracts, W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), the Freiburg Research Innovation Fund, and the DFG (GRO 4334-1-1). I would like to thank Johan Aberg, Martin Kliesch, Ulrich Terstiege and Holger Rauhut for inspiring discussions. This gratitude in particular extends to Yi-Kai Liu and David Gross who provided valuable inputs that ultimately led to the ideas presented here.

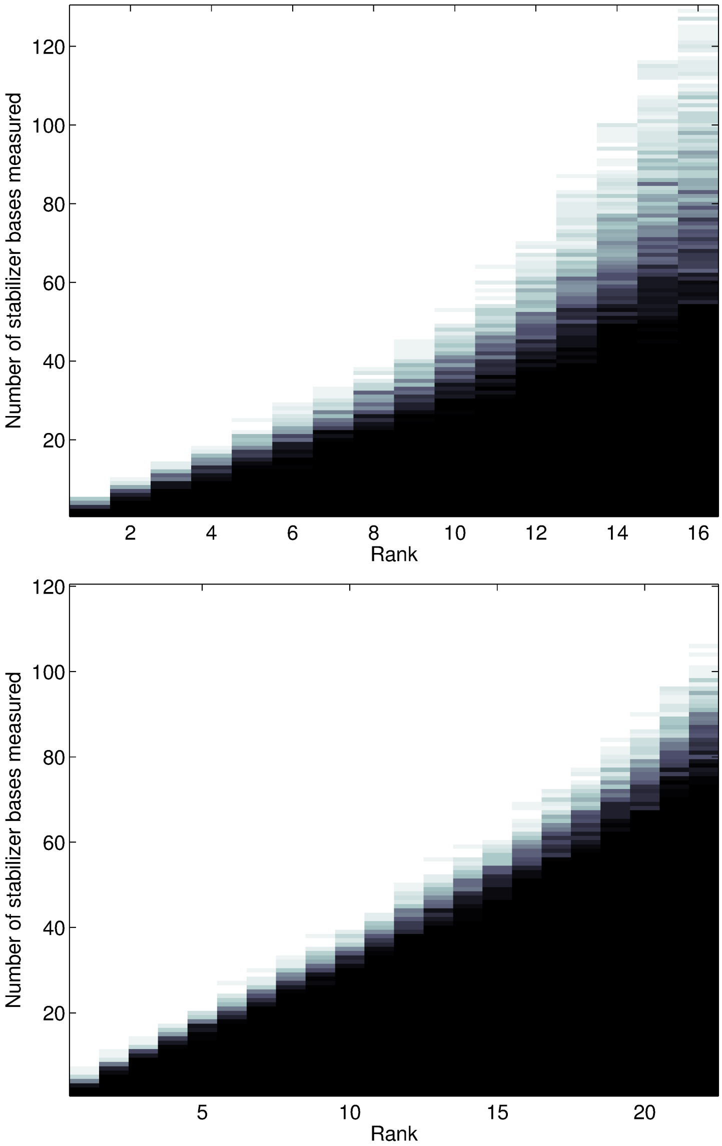


Fig. 2. Phase diagrams for rank-r matrix recovery via (2) from measuring random (orthonormal) stabilizer bases in dimensions $n = 16$ and $n = 32$. The x axis specifies the rank of the test matrices – ranging from 1 to $3n/4$ – while the y-axis denotes the number L of independently chosen stabilizer bases measured. The frequency of successful recovery via (2) over 30 independent runs is color-coded from black (zero) to white (one).

References

[1] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, ser. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[2] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[3] E. J. Candès and Y. Plan, "Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements," *IEEE Trans. Inform. Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[4] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization." *SIAM Rev.*, vol. 52, pp. 471–501, 2010.

[5] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inform. Theory*, vol. 57, pp. 1548–1566, 2011.

[6] Y.-K. Liu, "Universal low-rank matrix recovery from pauli measurements," *Adv. Neural Inf. Process. Syst.*, pp. 1638–1646, 2011.

[7] In power of two dimensions, for instance, each non-trivial Pauli measurement has two eigenvalues ±1 with associated eigenspaces of dimension $n/2$ each.

[7] A. Ahmed, B. Recht, and J. Romberg, “Blind deconvolution using convex programming,” *preprint*, 2012.

[8] Y. Chen, “Incoherence-optimal matrix completion,” *preprint arXiv:1310.0154*, 2013.

[9] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, “Painless reconstruction from magnitudes of frame coefficients.” *J. Fourier Anal. Appl.*, vol. 15, pp. 488–501, 2009.

[10] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, “Phase retrieval via matrix completion,” *SIAM J. Imaging Sci.*, vol. 6, pp. 199–225, 2013.

[11] E. J. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” *IEEE Trans. Inform. Theory*, vol. 56, pp. 2053–2080, 2010.

[12] E. J. Candès, T. Strohmer, and V. Voroninski, “Phaselift: exact and stable signal recovery from magnitude measurements via convex programming.” *Commun. Pure Appl. Math.*, vol. 66, pp. 1241–1274, 2013.

[13] D. Gross, F. Krahmer, and R. Kueng, “A partial derandomization of PhaseLift using spherical designs,” *J. Fourier Anal. Appl.*, vol. 21, pp. 229–266, 2015.

[14] E. Candès and X. Li, “Solving quadratic equations via PhaseLift when there are about as many equations as unknowns,” *Found. Comput. Math.*, pp. 1–10, 2013.

[15] R. Kueng, H. Rauhut, and U. Terstiege, “Low rank matrix recovery from rank one measurements,” *preprint arXiv:1410.6913*, 2014.

[16] E. J. Candes, X. Li, and M. Soltanolkotabi, “Phase retrieval from coded diffraction patterns,” *Appl. Comput. Harmon. Anal.*, 2014.

[17] D. Gross, F. Krahmer, and R. Kueng, “Improved recovery guarantees for phase retrieval from coded diffraction patterns,” *Appl. Comput. Harmon. Anal.*, to appear, preprint arXiv:1402.6286.

[18] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, “Phase retrieval with polarization,” *SIAM J. Imaging Sci.*, vol. 7, pp. 35–66, 2014.

[19] P. Netrapalli, P. Jain, and S. Sanghavi, “Phase retrieval using alternating minimization,” in *Advances in Neural Information Processing Systems*, 2013, pp. 2796–2804.

[20] E. Candes, X. Li, and M. Soltanolkotabi, “Phase retrieval via wirtinger flow: Theory and algorithms,” *preprint arXiv:1407.1065*, 2014.

[21] P. Delsarte, J. Goethals, and J. Seidel, “Spherical codes and designs.” *Geom. Dedicata*, vol. 6, pp. 363–388, 1977.

[22] R. Kueng, D. Gross, and F. Krahmer, “Spherical designs as a tool for derandomization: The case of PhaseLift,” in *11th international conference on Sampling Theory and Applications (SampTA 2015)*, Washington, USA, May 2015.

[23] A. Klappenecker and M. Rotteler, “Mutually unbiased bases are complex projective 2-designs,” in *2005 IEEE International Symposium on Information Theory (ISIT), Vols 1 and 2*, 2005, pp. 1740–1744.

[24] R. Kueng and D. Gross, “Stabilizer states are complex projective 3-designs in qubit dimensions,” in preparation, 2015.

[25] C. Dankert, R. Cleve, J. Emerson, and E. Livine, “Exact and approximate unitary 2-designs and their application to fidelity estimation,” *Phys. Rev. A*, vol. 80, no. 1, p. 012304, 2009.

[26] D. Gross, K. Audenaert, and J. Eisert, “Evenly distributed unitaries: on the structure of unitary designs.” *J. Math. Phys.*, vol. 48, pp. 052 104, 22, 2007.

[27] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge university press, 2010.

[28] K. Banaszek, M. Cramer, and D. Gross, “Focus on quantum tomography,” *New J. Phys.*, vol. 15, p. 125020, 2013.

[29] C. Ferrie and R. Kueng, “Have you been using the wrong estimator? These guys bound average fidelity using this one weird trick von Neumann didn’t want you to know,” *preprint arXiv:1503.00677*, 2015.

[30] P. Schindler, D. Nigg, T. Monz, J. T. Barreiro, E. Martinez, S. X. Wang, S. Quint, M. F. Brandl, V. Nebendahl, C. F. Roos *et al.*, “A quantum information processor with trapped ions,” *New J. Phys.*, vol. 15, no. 12, p. 123012, 2013.

[31] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, “state tomography via compressed sensing,” *Phys. Rev. Lett.*, vol. 105, p. 150401, 2010.

[32] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, “Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators,” *New J. Phys.*, vol. 14, p. 095022, 2012.

[33] F. G. Brandao, A. W. Harrow, and M. Horodecki, “Local random quantum circuits are approximate polynomial-designs,” *preprint arXiv:1208.0692*, 2012.

[34] C. Schwemmer, G. Tóth, A. Niggebaum, T. Moroder, D. Gross, O. Gühne, and H. Weinfurter, “Experimental comparison of efficient tomography schemes for a six-qubit state,” *Phys. Rev. Lett.*, vol. 113, no. 4, p. 040503, 2014.

[35] M. Grant and S. Boyd, “CVX: Matlab software for disciplined convex programming, version 2.1,” http://cvxr.com/cvx, Mar. 2014.

RIPless compressed sensing from anisotropic measurements

R. Kueng[a,b,1], D. Gross[b,1]

[a]Institute for Theoretical Physics, ETH Zürich, Wolfgang-Pauli-Strasse 27, 8093 Zürich, Switzerland
[b]Institute for Physics, University of Freiburg, Rheinstrasse 10, 79104 Freiburg, Germany

Abstract

Compressed sensing is the art of reconstructing a sparse vector from its inner products with respect to a small set of randomly chosen measurement vectors. It is usually assumed that the ensemble of measurement vectors is in *isotropic position* in the sense that the associated covariance matrix is proportional to the identity matrix. In this paper, we establish bounds on the number of required measurements in the *anisotropic* case, where the ensemble of measurement vectors possesses a non-trivial covariance matrix. Essentially, we find that the required sampling rate grows proportionally to the condition number of the covariance matrix. In contrast to other recent contributions to this problem, our arguments do not rely on any *restricted isometry properties* (RIP's), but rather on ideas from convex geometry which have been systematically studied in the theory of low-rank matrix recovery. This allows for a simple argument and slightly improved bounds, but may lead to a worse dependency on noise (which we do not consider in the present paper).

Keywords: Compressed sensing, ℓ_1 minimization, the LASSO, the Dantzig selector, restricted isometries, anisotropic ensembles, sparse regression, operator Bernstein inequalities, non-commutative large deviation estimates, the golfing scheme. Subject Classification: (94A12, 60D05, 90C25).

1. Introduction and Results

Compressed sensing is a highly active research field in statistics and signal analysis [1, 2, 3, 4]. It can be thought of as being concerned with establishing *Nyquist*-type sampling theorems for signals which are sparse, rather than band-limited.

More precisely, let $x \in \mathbb{C}^n$ be a vector with no more than s non-zero entries (i.e. x is s-*sparse*). Suppose we have no information about x apart from its sparsity and the inner products $\langle a_i, x\rangle, i = 1, \ldots, m$ between x and $m \ll n$ vectors a_i. The central question is: under what conditions on m and the a_i's is it possible to uniquely and computationally efficiently recover x? Early celebrated results [1, 2, 3] established e.g. that if the measurement vectors $\{a_i\}_{i=1}^m$ are randomly chosen discrete Fourier vectors

[1] Contact: www.qc.uni-freiburg.de

and $m = O(s \log n)$, then, with high probability, the unknown vector x is the unique minimizer of the ℓ_1-norm in the affine space defined by the known inner products.

The precise statement of our results in this introductory section will follow very closely the exhibition in [5]. The reason for this approach, and the relation of the present paper with other work (in particular [6]), is stated in Section 2.

We make the following definitions: Let F be a distribution of random vectors on $\mathbb{C}^n$. Let $a_1, \ldots, a_m$ be a sequence of i.i.d. random vectors drawn from F. Define the *sampling matrix*

$$A := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} e_i a_i^*,$$

where $e_1, \ldots, e_m$ denote the canonical basis vectors of $\mathbb{C}^m$. Once more, let x be an s-sparse vector. We aim to prove that with high probability the solution $x^\star$ to the convex optimization problem

$$\min_{\bar{x} \in \mathbb{C}^n} \|\bar{x}\|_1 \quad \text{subject to} \quad A\bar{x} = Ax, \tag{1}$$

is unique and equal to x given that the number of measurements m is large enough.

It turns out that the required size of m depends only on two simple properties of the ensemble F. These are identified below:

Completeness We require that the ensemble F is *complete* in the sense that the *covariance matrix* $\Sigma = \mathbb{E}[aa^*]^{1/2}$ is invertible. The *condition number* [2] of Σ will be denoted by κ.

Most of the previous work has focused on the case where the covariance matrix is proportional to the identity matrix $\Sigma \propto \mathbb{1}$ (however, see Section 2). We refer to this case as the *isotropic* one.

In order to describe the second relevant property of the ensemble, we have to fix a scale. Indeed, note that the minimizer of the convex problem (1) is invariant under re-scaling of the ensemble (i.e. substituting a_i by νa_i for a number $\nu \neq 0$). The same is true for the condition number κ. Thus, we are free to pick an advantageous scale, without affecting the notions introduced so far. In the isotropic case, a natural normalization convention [5] consists in requiring that $\mathbb{E}[aa^*] = \mathbb{1}$. This option is not available in the more general, anisotropic case, we are interested in here. Instead, we will implicitly demand from now that

$$\lambda_{\max}(\mathbb{E}[aa^*]) = \lambda_{\min}(\mathbb{E}[aa^*])^{-1}, \tag{2}$$

where $\lambda_{\max}, \lambda_{\min}$ denote the maximal and the minimal eigenvalue respectively. In the isotropic case, this reduces to the normalization $\mathbb{E}[aa^*] = \mathbb{1}$ used in [5].

The fact that (2) can always be achieved (and further properties that follow from it) will be established in Lemma 8 below. With this convention, we define:

[2] Recall that the condition number of a matrix is the ratio between its largest and its smallest singular value.

Incoherence The *incoherence parameter* is the smallest number μ such that

$$\max_{1\leq i\leq n} \left|\langle a, e_i\rangle\right|^2 \leq \mu, \qquad \max_{1\leq i\leq n} \left|\left\langle a, \mathbb{E}[aa^*]^{-1}\, e_i\right\rangle\right|^2 \leq \mu \tag{3}$$

holds almost surely.

The previously known isotropic result we aim to generalize is:

Theorem 1 ([5]). *Let x be an s-sparse vector in $\mathbb{R}^n$. If we demand isotropy ($\mathbb{E}[aa^*] = \mathbb{1}$) and if the number of measurements fulfills*

$$m \geq C_\omega \mu s \log n,$$

then the solution $x^\star$ of the convex program (1) is unique and equal to x with probability at least $1 - \frac{5}{n} - \mathrm{e}^{-\omega}$.

In the statement above, C_ω may be chosen as $C_0\,(1+\omega)$ for some positive numerical constant C_0

Our main theorem reads:

Theorem 2 (Main Theorem). *Let $x \in \mathbb{C}^n$ be an s-sparse vector, let $\omega \geq 1$. If the number of measurements fulfills*

$$m \geq C\kappa\mu\omega^2 s \log n,$$

then the solution $x^\star$ of the convex program (1) is unique and equal to x with probability at least $1 - \mathrm{e}^{-\omega}$.

In the statement above, C is a constant less than 18044. For n, s sufficiently large, the value may be improved to $C \leq 228$. We have made no attempts to optimize these constants.

Comparing these two theorems, we see that the effect of dropping the isotropy constraint on the ensemble can essentially be captured in a single, simple quantity: the condition number κ of the covariance matrix. All other minor differences between Theorem 1 and Theorem 2 result from slightly different proof techniques.

1.1. Improvements

A first way of improving the result is based on a definition borrowed from [6, Def. 1.2] [3]:

Definition 3. *The* largest and smallest s-sparse eigenvalue *of a matrix X are given by*

$$\lambda_{\max}(s, X) := \max_{v, \|v\|_0 \leq s} \frac{\|Xv\|_2}{\|v\|_2}, \qquad \lambda_{\min}(s, X) := \min_{v, \|v\|_0 \leq s} \frac{\|Xv\|_2}{\|v\|_2},$$

[3] In fact, our definition differs very slightly from [6]: their $\rho_{\max}(s, X)$ is the square of our $\lambda_{\max}(s, X)$. We opted for this change because the notions defined here reduce to the ordinary eigenvalues in the case of $s = n$.

*where $\|v\|_0 = |\mathrm{supp}(v)|$ denotes the cardinality of the support (i.e. the sparsity) of v. The s-*sparse condition number [4] *of X is*

$$\mathrm{cond}(s, X) := \frac{\lambda_{\max}(s, X)}{\lambda_{\min}(s, X)}.$$

Based on this notion, one can state a strictly stronger version of the Main Theorem (which is the form we will prove in Section 3):

Theorem 4. *With*

$$\kappa_s := \max\left\{\mathrm{cond}(s, \Sigma), \mathrm{cond}(s, \Sigma^{-1})\right\},$$

the conclusion of the main Theorem 2 continues to hold if the lower bound on m is weakened to

$$m \geq C\mu\, \kappa_s\, \omega^2 s \log n,$$

for the same constant C.

We further suspect that the second incoherence condition in (3) can be relaxed. Two alternative bounds not relying on that condition are stated in Proposition 5 below. (The modifications of our proof necessary to arrive at these improved estimates will be sketched after Lemma 9).

Proposition 5. *Let K be a constant such that*

$$2\left\|[aa^*, \mathbb{E}[aa^*]^{-1}]\right\|_\infty \leq K$$

holds almost surely, where $[\cdot, \cdot]$ denotes the commutator ($[A, B] = AB - BA$) and $\|\cdot\|_\infty$ is the operator norm.

If the requirement (3) is not necessarily fulfilled, the conclusions of Theorem 2 remain valid if the sampling rate is bounded below by either

$$m \geq C\kappa\mu\omega^2 s^2 \log n \tag{4}$$

or

$$m \geq C(\kappa\mu s + K)\omega^2 \log n. \tag{5}$$

The commutator bound (5) is particularly relevant for ensembles corresponding to non-uniform samples from an orthogonal basis. In that case, $\mathbb{E}[aa^*]$ and aa^* commute with probability one, so that K may be chosen to be zero.

There is another degree of freedom which we have not yet systematically explored: Note that the minimizer of the convex optimization (1) does not change if we re-scale

[4] Estimating $\mathrm{cond}(s, X)$ is equivalent to computing the RIP constants of X (c.f. e.g. [7]). There are currently no tractable methods known for computing these numbers for any concrete set of matrices. We want to emphasize that while the mathematical concept of "RIP constants" appears in our sharpened result, its use here is completely different from the way it would be employed in RIP-based approaches to compressed sensing. To wit, we apply the concept to the *expected sensing matrix* (and its inverse), but not to any actual instances.

individual vectors $a_i \mapsto \nu_i a_i$ for some set of non-zero numbers ν_i. While we have chosen a *global* scale for the covariance matrix (c.f. Lemma 8), the individual weights remain free parameters that may be used to optimize the sampling rate. Pursuing this problem further seems likely to be fruitful.

We remark that the incoherence conditions can be relaxed to hold only with high probability. This opens up our results to, for example, the case of Gaussian measurement vectors. The details can be developed in complete analogy to Ref. [5, Appendix B].

Lastly, all statements remain true if the measurement vectors are drawn "without replacement" instead of independently – c.f. [8] for details.

2. Relation with previous work and history

Most results on sparse vector recovery have relied on certain conditions that quantify how much a given sampling matrix A distorts the geometry of the set of all sparse vectors. By far the most prominent example in that regard is the *restricted isometry property* (RIP) [3, 6] which measures the extent to which A deviates from preserving Euclidean distances between sparse vectors. Conceputally close variations of the RIP include the *restricted eigenvalue condition* introduced in [9], or the *restricted correlation assumption* [10]. Another example is the *width property* advanced in [11]—a Banach space-theoretic condition that seems to be weaker than the RIP.

From roughly 2008 on, the conceptually strongly related problem of recovering a low-rank matrix from few expansion coefficients with respect to a fixed matrix basis has come more and more into focus [12, 13]. There seems to be no easy way to directly translate the geometric approaches mentioned above to the general low-rank matrix recovery problem. Instead, the pioneering publications on the matrix problem used fairly elaborate methods from convex duality theory [12, 13]. (However, c.f. [14, 15, 16] for interesting special cases where RIP-based techniques *are* applicable to low-rank matrix recovery problems; and [17] for a related "restricted strong convexity" property with consequences for matrix recovery).

In [18, 19] the second author and his collaborators introduced a simplified approach to the low-rank matrix recovery problem. While these works still build on the convex framework of [12, 13], they incorporate several new ideas. These include the use of non-commutative large deviation theorems originating from quantum information theory [20, 21], randomized constructions based on i.i.d. samples of the measurement vectors, and a certain iterative "golfing scheme" for the construction of inexact dual certificates. These techniques were later modified and adapted to the original sparse-vector setting in [5]. This showed that the conceptual closeness of the matrix and the vector theory may be used to devise very similar proofs.

This "RIPless" approach to compressed sensing leads arguably to simpler proofs and gives tighter bounds at least for the noise-free recovery problem. As far as we know, RIP-based arguments still perform superior in the important noisy regime.

The work [5] did not include a systematic study of non-isotropic ensembles (however, "small" deviations from isotropy were discussed in Appendix B). In fact, E. Candès [5] suggested to us the problem of finding a generalization of the golfing scheme that could cope with anisotropic ensembles. This has been achieved by the first author of this paper during a research project under the supervision of the second author [22]. This explains the close relation between [5] and the present work.

An analysis of anisotropic compressed sensing within the original RIP framework has been carried out by other authors, most notably in [6]. Since their paper does not directly address the noise-free case, a direct comparison of statements is difficult. The closest result to ours seems to appear in Section 1.3, where a bound of

$$m \geq \mathcal{O}\big(sM^2 \log n \, \log^3(s \log n)\big)$$

for the sampling rate is given. The quantity M is an upper bound on the largest coefficient for the measurement vectors a_i, related to our parameter μ. The *big-Oh* notation hides a constant proportional to κ (ρ^{-1} in the language of [6]). Thus, the basic structure of the solutions is very similar. However, some important differences are these:

- We do not incur the $\log^3$-term, which is a major advantage of our method. Up to a constant factor, our required sampling rate corresponds to the theoretical lower limit.
- The result in [6] holds *uniformly* in the sense that with their probability of success, one obtains a sampling matrix which works simultaneously for all sparse vectors. This is not the case for us.
- We have proved no results on noise-resilience. While, following [5], it should be straight-forward to do so, the results may be worse than the RIP-based ones in [6].
- The proof methods are completely different.

3. Proof

The proof is conceptually close to [5], which in turn closely resembles [19]. Here we give a largely self-contained presentation.

3.1. Notation

Throughout this paper, we will use the following conventions:
If a statements holds almost surely, we will abbreviate this by a.s. In the case of vectors, $\|\cdot\|_p$ denotes the ℓ_p-norm, whereas in the operator case $\|\cdot\|_p$ refers to the Schatten-p norm (i.e. the ℓ_p-norm of the singular values). The letter z will always denote a vector in $\mathbb{C}^n$ supported on a set T of cardinality at most s (i.e. z is s-sparse). T^c shall denote the complement of T, and P_T (P_{T^c}) refers to the orthogonal projector onto the set of all vectors supported on T (T^c). Finally we will use the following technical definitions:

$$X = (\mathbb{E}[aa^*])^{-1} = \Sigma^{-2}, \qquad X_T = P_T X P_T.$$

3.2. Large deviation bounds

A central role in the argument is played by certain large deviation bounds for sums of matrix-valued random variables. These have been introduced in [20] in the context of quantum information theory. The first application to matrix completion and compressed sensing problems, as well as the first "Bernstein version" taking variance information into account, appeared in [18, 19]. The version we will be making use of derives from Theorem 1.6 in [21].

Proposition 6 (Matrix Bernstein inequality [21])**.** *Consider a finite sequence* $\{M_k\} \in \mathbb{C}^{d\times d}$ *of independent, random matrices. Assume that each random matrix satisfies* $\mathbb{E}\left[M_k\right] = 0$ *and* $\|M_k\|_\infty \leq B$ *a.s. and define*

$$\sigma^2 := \max\left\{\|\sum_k \mathbb{E}\left(M_k M_k^*\right)\|_\infty, \|\sum_k \mathbb{E}\left(M_k^* M_k\right)\|_\infty\right\}.$$

Then for all $t \geq 0$,

$$\Pr\left(\|\sum_k M_k\|_\infty \geq t\right) \leq 2d \exp\left(-\frac{t^2/2}{\sigma^2 + Bt/3}\right). \tag{6}$$

We will also require a vector-valued deviation estimate. While one could in principle obtain such a statement by applying Proposition 6 to diagonal matrices, a direct argument does away with the dimension factor d on the r.h.s. of (6). This will save a logarithmic factor in the sampling rate of the Main Theorem. The particular vector-valued Bernstein inequality below is based on the exposition in [23] (Chapter 6.3, equation (6.12)), with a direct proof appearing in [19].

Proposition 7 (Vector Bernstein inequality)**.** *Let* $\{g_k\} \in \mathbb{C}^d$ *be a finite sequence of independent random vectors. Suppose that* $\mathbb{E}\left[g_k\right] = 0$ *and* $\|g_k\|_2 \leq B$ *a.s. and put* $\sigma^2 \geq \sum_k \mathbb{E}\left[\|g_k\|_2^2\right]$*. Then for all* $0 \leq t \leq \sigma^2/B$*:*

$$\Pr\left(\left\|\sum_k g_k\right\|_2 \geq t\right) \quad \leq \quad \exp\left(-\frac{t^2}{8\sigma^2} + \frac{1}{4}\right).$$

3.3. Fundamental estimates

We adopt the structure and nomenclature of this section from [5]. The following elementary bounds will be used repeatedly:

$$|\langle a_k, z\rangle|^2 \leq s\mu\|z\|_2^2, \qquad |\langle a_k, Xz\rangle|^2 \leq s\mu\|z\|_2^2, \tag{7}$$

$$\|P_T a_k\|_2^2 \leq \mu s, \qquad \|P_T X a_k\|_2^2 \leq \mu s. \tag{8}$$

Also, we will always assume that $m \geq s$.

Lemma 8 (Scaling). *Let $\tilde{a}$ be a random vector such that $\mathbb{E}[\tilde{a}\tilde{a}^*]$ is invertible.*

There is a number ν such that, with $a := \nu\tilde{a}$, it holds that

$$\kappa_s = \lambda_{\max}(s, \mathbb{E}[aa^*]) = \lambda_{\min}(s, \mathbb{E}[aa^*])^{-1}$$

for all $1 \leq s \leq n$. This resealed ensemble fulfills:

$$\kappa_s \mu \geq 1. \tag{9}$$

Proof. The first assertion follows immediately for

$$\nu = \left(\lambda_{\max}(s, \mathbb{E}[\tilde{a}\tilde{a}^*])\lambda_{\min}(s, \mathbb{E}[\tilde{a}\tilde{a}^*])\right)^{-\frac{1}{4}}.$$

For the second claim: By definition $\mu \geq \max_i |\langle a, e_i\rangle|^2$ holds almost surely, so that in particular

$$\mu \quad \geq \quad \mathbb{E}\left[\max_i |\langle a, e_i\rangle|^2\right].$$

For every i, the function

$$a \mapsto |\langle a, e_i\rangle|^2$$

is convex, which implies that

$$a \mapsto \max_i |\langle a, e_i\rangle|^2 = \max_i e_i^*(aa^*)e_i$$

is convex (as the pointwise maximum of convex functions). Hence, by Jensen's inequality,

$$\begin{aligned} \mathbb{E}\left[\max_i |\langle a, e_i\rangle|^2\right] &\geq \max_i e_i^* \mathbb{E}[aa^*]e_i = \max_i \langle e_i, \mathbb{E}[aa^*]e_i\rangle \\ &\geq \lambda_{\min}\left(1, \mathbb{E}[aa^*]\right) \geq \lambda_{\min}\left(s, \mathbb{E}[aa^*]\right). \end{aligned}$$

Therefore $\mu \geq \lambda_{\min}\left(s, \mathbb{E}[aa^*]\right)$. Together with $\kappa_s = \lambda_{\min}^{-1}\left(s, \mathbb{E}[aa^*]\right)$, this implies $\mu\kappa_s \geq 1$. □

The estimates in this proof are tight in the sense that there are ensembles for which each inequality above turns into an equality. A straightforward example for such an ensemble is given by picking super-normalized Fourier basis vectors f_k (with coefficients $(f_k)_l = e^{2\pi \mathrm{i}\frac{kl}{n}}$) according to the uniform probability distribution.

Lemma 9 (Local isometry). *Let T and P_T be as in the notation section. Then for each $0 \leq \tau \leq \frac{1}{2}$:*

$$\Pr\left(\|P_T\left(XA^*A - \mathbb{1}\right)P_T\|_\infty \geq \tau\right) \leq 2s \exp\left(-\frac{m}{s\mu\kappa_s}\frac{\tau^2}{2\left(1 + 2\tau/3\right)}\right)$$

Proof. Let us decompose the relevant expression:

$$P_T\left(XA^*A - \mathbb{1}\right)P_T = \frac{1}{m}\sum_{i=1}^m M_k,$$

where $M_k := P_T \left(X a_k a_k^* - \mathbb{1} \right) P_T$. Note that $\mathbb{E}[M_k] = 0$.
We aim to apply the Matrix Bernstein inequality. To this end, we estimate

$$\begin{aligned} \|M_k\|_\infty &\leq \|P_T X a_k a_k^* P_T\|_2 + 1 \\ &= \|P_T X a_k\|_2 \, \|a_k^* P_T\|_2 + 1 \\ &\leq \mu s + 1 \leq 2\mu s \kappa_s =: B. \end{aligned}$$

Furthermore:

$$\begin{aligned} & \left\| \mathbb{E} \left[M_k M_k^* \right] \right\|_\infty \\ = \;& \left\| \mathbb{E} \left[\left(P_T \left(X a_k a_k^* - \mathbb{1} \right) P_T \right) \left(P_T \left(a_k a_k^* X - \mathbb{1} \right) P_T \right) \right] \right\|_\infty \\ = \;& \left\| \mathbb{E} \left[P_T X a_k a_k^* P_T a_k a_k^* X P_T \right] - \mathbb{E} \left[P_T X a_k a_k^* P_T \right] - \mathbb{E} \left[P_T a_k a_k^* X P_T \right] + P_T \right\|_\infty \\ = \;& \left\| \mathbb{E} \left[P_T \left(X a_k \left\langle a_k, P_T a_k \right\rangle a_k^* X - \mathbb{1} \right) P_T \right] \right\|_\infty \\ \leq \;& \max \left(\left\| \mu s \, \mathbb{E} \left[P_T X a_k a_k^* X P_T \right] \right\|_\infty , 1 \right) \\ \leq \;& \max \left(\mu s \left\| X_T \right\|_\infty , 1 \right) \leq \max \left(\mu s \kappa_s, 1 \right) = \mu s \kappa_s. \end{aligned}$$

Similarly,

$$\begin{aligned} \left\| \mathbb{E} \left[M_k^* M_k \right] \right\|_\infty &= \left\| \mathbb{E} \left[P_T \left(a_k \left\langle a_k, X P_T X a_k \right\rangle a_k^* - I \right) P_T \right] \right\|_\infty \\ &\leq \max \left(\left\| s\mu \mathbb{E} \left[P_T a_k a_k^* P_T \right] \right\|_\infty , 1 \right) \qquad (10) \\ &\leq \max \left(s\mu \left\| P_T X^{-1} P_T \right\|_\infty , 1 \right) \leq \mu s \kappa_s. \end{aligned}$$

Thus:

$$\max \left\{ \| \sum_{k=1}^{m} \mathbb{E} \left(M_k M_k^* \right) \|_\infty , \| \sum_{k=1}^{m} \mathbb{E} \left(M_k^* M_k \right) \|_\infty \right\} \leq m s \mu \kappa_s =: \sigma^2.$$

Applying the Matrix Bernstein inequality for s-dimensional matrices ($P_T \left(X A^* A - \mathbb{1} \right) P_T$ has rank at most s) with $t = m\tau$ yields the desired result. □

The estimate (10) is the only place in the proof where the second incoherence property in (3) is essentially used. A careful analysis shows that in all other cases, one can do without it, possibly at the price of replacing κ_s by κ (which is the reason why we have not spelled it out). In order to obtain the results of Proposition 5, the bound (10) has to be modified. To arrive at (4), use

$$\begin{aligned} \left\| \mathbb{E} \left[M_k^* M_k \right] \right\|_\infty &\leq \mathbb{E} \left[\| [M_k^* M_k] \|_\infty \right] \\ &\leq \mathbb{E} \left[\| P_T a_k a_k^* P_T \left\langle a_k, X P_T X a_k \right\rangle \|_\infty \right] \\ &\leq s\mu \, \mathbb{E} \left[\left\langle a_k, X P_T X a_k \right\rangle \right] = s\mu \, \mathbb{E} \left[\operatorname{tr} \left(a_k a_k^* X P_T X \right) \right] \\ &= s\mu \operatorname{tr} \left(X^{-1} X P_T X \right) = s\mu \operatorname{tr} \left(P_T X \right) \leq s^2 \mu \kappa_s. \end{aligned}$$

And for (5):

$$\begin{aligned}
& \left\|\mathbb{E}\left[P_T a_k a_k^* X P_T X a_k a_k^* P_T\right]\right\|_\infty \\
= \; & \big\|\mathbb{E}\left[P_T X a_k a_k^* P_T X a_k a_k^* P_T\right] + \mathbb{E}\left[P_T [a_k a_k^*, X] P_T X a_k a_k^* P_T\right] \big\|_\infty \\
\leq \; & \left\|\mathbb{E}\left[P_T X a_k a_k^* P_T a_k a_k^* X P_T\right]\right\|_\infty + 2 \left\|\mathbb{E}\left[P_T [a_k a_k^*, X] P_T X a_k a_k^* P_T\right]\right\|_\infty \\
\leq \; & \mu s \kappa_s + K \left\|\mathbb{E}\left[P_T X a_k a_k^* P_T\right] \right\|_\infty \\
= \; & \mu s \kappa_s + K \left\|P_T X X^{-1} P_T\right\|_\infty \\
= \; & \mu s \kappa_s + K.
\end{aligned}$$

Lemma 10 (Low-distortion)**.** *Let z, T, P_T be as in the notation section. For each $0 \leq \tau \leq 1$ it holds that*

$$\Pr\big(\left\|P_T \left(\mathbb{1} - A^* A X\right) z\right\|_2 \geq \tau \left\|z\right\|_2 \big) \leq \exp\left(-\frac{m\tau^2}{16 s \mu \kappa_s} + \frac{1}{4} \right).$$

Proof. The structure of the proof closely follows the one of Lemma 9. Set

$$g_k := P_T \left(\mathbb{1} - a_k a_k^* X\right) z.$$

We bound

$$\begin{aligned}
\|g_k\|_2 &= \|P_T(\mathbb{1} - a_k a_k^* X) z\|_2 \\
&\leq \|z\|_2 + \|P_T a_k \langle a_k, Xz\rangle\|_2 \\
&\leq \|z\|_2 + s\mu \|z\|_2 \leq 2 s\mu\kappa_s \|z\|_2 =: B
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\|g_k\|_2^2] &\leq \mathbb{E}[\|P_T a_k \langle a_k, Xz\rangle\|_2^2] + \|z\|_2^2 \\
&= \mathbb{E}\big[\|P_T a_k\|_2^2 |\langle a_k, Xz\rangle|^2\big] + \|z\|_2^2 \\
&\leq s\mu \mathbb{E}\big[\langle Xz, a_k\rangle \langle a_k, Xz\rangle\big] + \|z\|_2^2 \\
&= s\mu \langle Xz, \mathbb{E}[a_k a_k^*] Xz\rangle\big] + \|z\|_2^2 \\
&= s\mu \langle Xz, z\rangle + \|z\|_2^2 \leq 2 s\mu\kappa_s \|z\|_2^2
\end{aligned}$$

so that

$$\sum_{k=1}^{m} \mathbb{E}[\|g_k\|_2^2] \leq 2 m s\mu\kappa_s \|z\|_2^2 =: \sigma^2$$

and thus $\frac{\sigma^2}{B} = m\|z\|_2$. The advertised statement follows by applying the vector Bernstein inequality for $t = m\tau$. □

Lemma 11 (Off-support incoherence)**.** *Let z, P_{T^c} again be as in the notation section. Then for each $\tau \geq 0$:*

$$\Pr\big(\left\|P_{T^c} A^* A X z\right\|_\infty \geq \tau \left\|z\right\|_2 \big) \leq 2n \exp\left(-\frac{3m\tau^2}{2\mu\kappa_s(3 + \sqrt{s}\tau)} \right)$$

Proof. Fix $i \in T^c$ and use the following decomposition:

$$\langle e_i, A^*AXz\rangle = \frac{1}{m}\sum_{i=1}^{m} M_k,$$

where $M_k := \langle e_i, a_k a_k^* Xz\rangle = \langle e_i, a_k\rangle\langle a_k, Xz\rangle$. Note that we have:

$$\mathbb{E}[M_k] = \langle e_i, \mathbb{E}[a_k a_k^*]Xz\rangle = \langle e_i, z\rangle = 0,$$

because $e_i \in T^c$. Bound

$$|M_k| = |\langle e_i, a_k\rangle\langle a_k, Xz\rangle| \leq \sqrt{s}\mu\kappa_s\|z\|_2 =: B,$$

and

$$\begin{aligned}\mathbb{E}[M_k M_k^*] &= \mathbb{E}[M_k^* M_k] = \mathbb{E}[|\langle a_k, e_i\rangle|^2|\langle a_k, Xz\rangle|^2] \\ &\leq \mu\mathbb{E}[\langle Xz, a_k a_k^* Xz\rangle] = \mu\langle Xz, z\rangle \\ &\leq \mu\|X_T\|_\infty\|z\|_2^2 \leq \mu\kappa_s\|z\|_2^2.\end{aligned}$$

Therefore we can set $\sigma^2 := m\mu\kappa_s\|z\|_2^2$. Applying the Matrix Bernstein inequality for $d = 1$ and the union bound over all $i \in T^c$ yields the claim. □

Lemma 12 (Uniform off-support incoherence). *Let T^c, P_T be as in the notation section. For $0 \leq \tau \leq 1$ we have*

$$\Pr\left(\max_{i\in T^c}\|P_T XA^*Ae_i\|_2 \geq \tau\right) \leq n\exp\left(-\frac{m\tau^2}{8s\mu\kappa_s} + \frac{1}{4}\right)$$

Proof. Fix $i \in T^c$ and decompose:

$$P_T XA^*Ae_i = \frac{1}{m}\sum_{k=1}^{m} g_k,$$

where $g_k := \langle a_k, e_i\rangle P_T X a_k$. It holds that $\mathbb{E}[g_k] = 0$. Next, bound

$$\|g_k\|_2 = |\langle a_k, e_i\rangle|\|P_T X a_k\|_2 \leq s\mu =: B.$$

Furthermore:

$$\mathbb{E}[\|g_k\|_2^2] \leq \sum_{i\in T}\mu\mathbb{E}[\langle e_i, Xa_k a_k^* Xe_i\rangle] \leq \sum_{i\in T}\mu\|X_T\|_\infty \leq s\mu\kappa_s.$$

We can therefore set $\sigma^2 := ms\mu\kappa_s$ and apply the Vector Bernstein inequality for $t = m\tau$. Noting that $\sigma^2/B = m\kappa_s \geq m$ finishes the proof. □

3.4. Convex geometry

Our aim is to prove that the solution $x^\star$ to the optimization problem (1) equals the unknown vector x. One way of assuring this is by exhibiting a *dual certificate* [24]. This method was first introduced in [2] and is now standard. We will use a relaxed version of this first introduced in [19] and later adapted from matrices to vectors in [5]. Our version further adapts the statement to the anisotropic setting.

Lemma 13 (Inexact duality). *Let $x \in \mathbb{C}^n$ be a s-sparse vector, let $T = \operatorname{supp}(x)$.*

Assume that

$$\begin{aligned} \| \left(P_T X A^* A P_T\right)^{-1} \|_\infty &\leq 2, & (11) \\ \max_{i \in T^c} \| P_T X A^* A e_i \|_2 &\leq 1 & (12) \end{aligned}$$

and that there is a vector v in the row space of A obeying

$$\begin{aligned} \| v_T - \operatorname{sgn}(x) \|_2 &\leq \frac{1}{4} & (13) \\ \| v_{T^c} \|_\infty &\leq \frac{1}{4}. & (14) \end{aligned}$$

Then the solution $x^\star$ of the convex program (1) is unique and equal to x.

Proof. Let $\hat{x} = x + h$ be a solution of the minimization procedure. We note that feasibility requires $Ah = 0$. To prove the claim it suffices to show $h = 0$. Observe:

$$\begin{aligned} \|\hat{x}\|_1 &= \|x + h_T\|_1 + \|h_{T^c}\|_1 \\ &= \langle \operatorname{sgn}(x + h_T), x + h_T \rangle + \|h_{T^c}\|_1 \\ &\geq \langle \operatorname{sgn}(x), x \rangle + \langle \operatorname{sgn}(x), h_T \rangle + \|h_{T^c}\|_1 \\ &\geq \|x\|_1 - |\langle \operatorname{sgn}(x), h_T \rangle| + \|h_{T^c}\|_1. \end{aligned}$$

Feasibility requires $\langle v, h \rangle = 0$ (since v is in the row space of A) and therefore:

$$\begin{aligned} |\langle \operatorname{sgn}(x), h_T \rangle| &= |\langle \operatorname{sgn}(x) - v_T, h_T \rangle + \langle v_T, h_T \rangle| \\ &= |\langle \operatorname{sgn}(x) - v_T, h_T \rangle - \langle v_{T^c}, h_{T^c} \rangle| \\ &\leq |\langle \operatorname{sgn}(x) - v_T, h_T \rangle| + |\langle v_{T^c}, h_{T^c} \rangle| \\ &\leq \|\operatorname{sgn}(x) - v_T\|_2 \|h_T\|_2 + |\langle v_{T^c}, h_{T^c} \rangle| \\ &\leq \frac{1}{4}\|h_T\|_2 + |\langle v_{T^c}, h_{T^c} \rangle|, \end{aligned}$$

where we have used (13). Together with:

$$|\langle v_{T^c}, h_{T^c} \rangle| \leq \|v_{T^c}\|_\infty \|h_{T^c}\|_1 \leq \frac{1}{4}\|h_{T^c}\|_1,$$

this implies:

$$|\langle \operatorname{sgn}(x), h_T \rangle| \leq \frac{1}{4}\left(\|h_T\|_2 + \|h_{T^c}\|_1\right).$$

Furthermore due to (11) and (12) it holds that

$$\begin{aligned}
\|h_T\|_2 &= \| \left(P_T X A^* A P_T\right)^{-1} \left(P_T X A^* A P_T\right) h_T \|_2 \\
&= \| \left(P_T X A^* A P_T\right)^{-1} \left(P_T X A^* A\right) \left(h - h_{T^c}\right) \|_2 \\
&= \| - \left(P_T X A^* A P_T\right)^{-1} \left(P_T X A^* A\right) h_{T^c} \|_2 \\
&\leq 2 \| P_T X A^* A P_{T^c} h \|_2 \\
&\leq 2 \mathrm{max}_{i \in T^c} \| P_T X A^* A e_i \|_2 \| h_{T^c} \|_1 \\
&\leq 2 \| h_{T^c} \|_1,
\end{aligned}$$

All this together implies:

$$\begin{aligned}
\|\hat{x}\|_1 &\geq \|x\|_1 - \frac{1}{4} \|h_T\|_2 + \frac{3}{4} \|h_{T^c}\|_1 \\
&\geq \|x\|_1 + \frac{1}{4} \|h_{T^c}\|_1.
\end{aligned}$$

Consequently $\|\hat{x}\|_1 = \|x\|_1$ demands $\|h_{T^c}\|_1 = 0$, which in turn implies $\|h_T\|_2 = 0$, because $\|h_T\|_2 \leq 2\|h_{T^c}\|_1$. Therefore $h = 0$ which corresponds to a unique minimizer ($\hat{x} = x$). □

3.5. Construction of the certificate

It remains to show that a dual certificate v as described in Lemma 13 can indeed be constructed. We will prove:

Lemma 14. *Let $x \in \mathbb{C}^n$ be an s-sparse vector, let $\omega \geq 1$. If the number of measurements fulfills*

$$m \geq 18044 \kappa_s \mu \omega^2 s \log n,$$

then with probability at least $1 - \mathrm{e}^{-\omega}$, the constraints (11, 12) will hold and a vector v with the properties required for Lemma 13 exists.

This lemma immediately implies the Main Theorem.

The proof employs a recursive procedure (dubbed the “golfing scheme”) to construct a sequence v_i of vectors converging to a dual certificate with high probability. The technique has been developed in [18, 19] in the context of low-rank matrix recovery problems and has later been refined for compressed sensing in [5]. Here, we further modify the construction to handle anisotropic ensembles.

Proof. The recursive scheme consists of l iterations. The i-th iteration depends on three parameters: $m_i \in \mathbb{N}; c_i, t_i \in \mathbb{R}$ which will be chosen in the course of the later analysis. To initialize, set

$$v_0 = 0$$

(the v_i for $1 \leq i \leq l$ will be defined iteratively below). We will use the notation

$$q_i = \mathrm{sgn}\left(x\right) - P_T v_i.$$

The i-th step of the scheme proceeds according to the following protocol: We sample m_i vectors from the ensemble F. Let $\tilde{A}$ be the $m_i \times n$-matrix whose rows consists of these vectors. We check whether the following two conditions are met:

$$\begin{aligned} \|P_T \left(\mathbb{1} - \frac{m}{m_i}\tilde{A}^* \tilde{A} X\right) P_T q_{i-1}\|_2 &\leq c_i \|q_{i-1}\|_2, \qquad (15) \\ \|\frac{m}{m_i} P_{T^c} \tilde{A}^* \tilde{A} X P_T q_{i-1}\|_\infty &\leq t_i \|q_{i-1}\|_2. \qquad (16) \end{aligned}$$

If so, set

$$A_i = \tilde{A}, \qquad v_i = \frac{m}{m_i} A_i^* A_i X P_T \left(\operatorname{sgn}(x) - v_{i-1}\right) + v_{i-1}$$

and proceed to step $i+1$. If either of (15), (16) fails to hold, repeat the i-th step with a fresh batch of m_i vectors drawn from F. Denote the number of repetitions of the i-th step by r_i.

We now analyze the properties of the above recursive construction. The following identities are easily verified by repeating the given transformations inductively:

$$\begin{aligned} v &:= v_l = \frac{m}{m_l} A_l^* A_l X P_T \left(\operatorname{sgn}(x) - v_{l-1}\right) + v_{l-1} \\ &= \frac{m}{m_l} A_l^* A_l X P_T q_{l-1} + v_{l-1} \\ &= \ldots = \sum_{i=1}^{l} \frac{m}{m_i} A_i^* A_i X P_T q_{i-1}, \qquad (17) \\ q_i &= \operatorname{sgn}(x) - P_T v_i \\ &= \operatorname{sgn}(x) - P_T \left(\frac{m}{m_i} A_i^* A_i X P_T \left(\operatorname{sgn}(x) - v_{i-1}\right) + v_{i+1}\right) \\ &= \left(\operatorname{sgn}(x) - P_T v_{i-1}\right) - \frac{m}{m_i} A_i^* A_i X P_T \left(\operatorname{sgn}(x) - v_{i-1}\right) \\ &= P_T \left(\mathbb{1} - \frac{m}{m_i} A_i^* A_i X\right) q_{i-1} \\ &= \ldots = \prod_{j=1}^{i} P_T \left(\mathbb{1} - \frac{m}{m_i} A_j^* A_j X\right) P_T \operatorname{sgn}(x). \qquad (18) \end{aligned}$$

Together with (15) and (16), one obtains

$$\begin{aligned}
\|q_l\|_2 &\leq c_l\|q_{l-1}\|_2 \leq \prod_{i=1}^{l} c_i\|q_0\|_2 = \prod_{i-1}^{l} c_i\|\operatorname{sgn}(x)\|_2 = \sqrt{s}\prod_{i=1}^{l} c_i, \\
\|v_{T^c}\|_\infty &= \left\| P_{T^c}\left(\sum_{i=1}^{l} \frac{m}{m_i} A_i^* A_i X P_T q_{i-1}\right)\right\|_\infty \\
&\leq \sum_{i=1}^{l} \left\| \frac{m}{m_i} P_{T^c} A_i^* A_i X P_T q_{i-1}\right\|_2 \\
&\leq \sum_{i=1}^{l} t_i\|q_{i-1}\|_2 \leq \sqrt{s}\left(t_1 + \sum_{i=2}^{l} t_i \prod_{j=1}^{i-1} c_j\right).
\end{aligned}$$

Following [19], we choose the parameters l, c_i, t_i as

$$l = \left\lceil \frac{1}{2}\log_2 s \right\rceil + 2, \qquad c_1 = c_2 = \frac{1}{2\sqrt{\log n}}, \qquad t_1 = t_2 = \frac{1}{8\sqrt{s}},$$

and for $i \geq 3$

$$t_i = \frac{\log n}{8\sqrt{s}}, \qquad c_i = \frac{1}{2}.$$

A short calculation then yields

$$\|v_{T^c}\|_\infty \leq \frac{1}{4}, \qquad \|v - \operatorname{sgn}(x_T)\|_2 = \|q_l\|_2 \leq \frac{1}{4},$$

which are conditions (13) and (14).

Next, we need to establish that the total number

$$\sum_{i=1}^{l} m_i r_i$$

of sampled vectors remains small with high probability. More precisely, we will bound the probability

$$p_3 := \Pr\left((r_1 > 1) \text{ or } (r_2 > 1) \text{ or } \sum_{i=1}^{l} r_i \geq l'\right)$$

for some l' to be chosen later.

To that end, denote by $p_1(i)$ the probability that (15) fails to hold in any given batch of the i-th step. Analogously, let $p_2(i)$ be the probability of failure for (16). Lemmas 10 and 11 give the estimates

$$p_1(i) \leq \exp\left(-\frac{m_i c_i^2}{16 s\mu\kappa_s} + \frac{1}{4}\right), \qquad p_2(i) \leq 2n\exp\left(-\frac{3 m_i t_i^2}{2\mu\kappa_s\left(3 + \sqrt{s}t_i\right)}\right).$$

We choose

$$l' = 4(\omega + \log 12 + \frac{2}{3}l), \qquad m_1 = m_2 = 694\kappa_s\mu\omega s \log n,$$

and for $i \geq 3$

$$m_i = 694\kappa_s\mu\omega s.$$

Such a choice can be guaranteed by a total sampling rate $m \geq 18044\kappa_s\mu\omega^2 s \log n$ and ensures

$$p_1(i) + p_2(i) \leq \frac{1}{6}e^{-\omega} \leq \frac{1}{12}$$

for all i. (It is easily seen that for for $n \gg 1$, a bound of $m \geq 228\kappa_s\mu\omega^2 s \log n$ is sufficient. The constants appearing here are highly unlikely to be optimal.) Note that

$$\sum_{i=1}^{l} r_i \geq l'$$

only if fewer than l of the first l' batches of vectors satisfied both (15) and (16). This implies that

$$\Pr\left(\sum_{i=1}^{l} r_i \geq l'\right) \leq \Pr(N \leq l-1)_{\mathrm{Bin}(l',\frac{11}{12})},$$

where the r.h.s. is the probability of obtaining fewer than l outcomes in a binomial process with l' repetitions and individual success probability $11/12$. We bound this quantity using a standard concentration bound from [25] (C. McDiarmid's section "Concentration"):

$$\Pr\left(|\mathrm{Bin}\,(n,p) - np| > \tau\right) \leq 2\exp\left(-\frac{\tau^2}{3np}\right).$$

This yields $\Pr\left(\sum_{i=1}^{l} r_i \geq l'\right) \leq \frac{1}{6}\mathrm{e}^{-\omega}$ for our choice of $l^{'}$. Putting things together, we have

$$p_3 \leq 3\,\frac{1}{6}\mathrm{e}^{-\omega} = \frac{1}{2}\mathrm{e}^{-\omega}$$

according to the union bound. In addition, we have to take into account that properties (11) and (12) can fail as well. We denote these probabilities of failure by p_4 and p_5. Lemmas 9 and 12 give:

$$p_4 \leq 2s\exp\left(-\frac{6m}{7s\mu\kappa_s}\right), \qquad p_5 \leq n\exp\left(-\frac{m}{8s\mu\kappa_s} + \frac{1}{4}\right).$$

Our sampling rate m guarantees $p_4 \leq \frac{1}{4}\mathrm{e}^{-\omega}$ as well as $p_5 \leq \frac{1}{4}\mathrm{e}^{-\omega}$. Applying the union bound now yields our desired overall error bound ($p_3 + p_4 + p_5 \leq \mathrm{e}^{-\omega}$). □

4. Conclusion and Outlook

In this paper, we have shown that proof techniques based on duality theory and the "golfing scheme" are versatile enough to handle the situation where the ensemble of measurement vectors is not isotropic.

An obvious future line of research would be to translate these results to the low-rank matrix recovery problem. Given the high degree of similarity between [19] and [5], this should be a conceptually straight-forward task. This would further generalize the scope of this proof method, beyond ortho-normal operator bases [19] and tight frames [26].

Also, Proposition 5 suggests that the second incoherence property (3) can be relaxed or maybe even disposed of. We leave this as an open problem.

5. Acknowledgments

We thank E. Candès for suggesting the problem treated here, and him, Y. Plan, P. Jung, and P. Walk for insightful discussions. Financial support from the Excellence Initiative of the German Federal and State Governments (grant ZUK 43), the German Science Foundation (DFG grants CH 843/1-1 and CH 843/2-1), the Swiss National Science Foundation, and the Swiss National Center of Competence in Research "Quantum Science and Technology" is gratefully acknowledged.

References

[1] E. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Transactions on Information Theory 52 (2006) 489 – 509.

[2] E. Candes, T. Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, IEEE Transactions on Information Theory 52 (2006) 5406 – 5425.

[3] D. Donoho, Compressed sensing, IEEE Transactions on Information Theory 52 (2006) 1289 – 1306.

[4] M. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso), IEEE Transactions on Information Theory 55 (2009) 2183 –2202.

[5] E. J. Candes, Y. Plan, A probabilistic and RIPless theory of compressed sensing, IEEE Transactions on Information Theory 57 (2011) 7235–7254.

[6] M. Rudelson, S. Zhou, Reconstruction from anisotropic random measurements, preprint: arXiv:1106.1151 (2011).

[7] A. Juditsky, A. Nemirovski, On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization, Mathematical Programming 127 (2011) 57–88.

[8] D. Gross, V. Nesme, Note on sampling without replacing from a finite collection of matrices, preprint: arXiv:1001.2738 (2010).

[9] P. J. Bickel, Y. Ritov, A. B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, The Annals of Statistics 37 (2009) 1705–1732.

[10] P. J. Bickel, Discussion: The Dantzig selector: Statistical estimation when p is much larger than n, The Annals of Statistics (2007) 2352–2357.

[11] B. S. Kashin, V. N. Temlyakov, A remark on compressed sensing, Mathematical notes 82 (2007) 748–755.

[12] E. Candes, B. Recht, Exact matrix completion via convex optimization, Foundations of Computational Mathematics 9 (2009) 717–772.

[13] E. Candes, T. Tao, The power of convex relaxation: Near-Optimal matrix completion, IEEE Transactions on Information Theory 56 (2010) 2053 –2080.

[14] Y. Liu, Universal low-rank matrix recovery from Pauli measurements, Adv. in Neural Information Processing Systems 24 (2011) 1638–1646.

[15] E. J. Candes, Y. Plan, Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements, IEEE Transactions on Information Theory 57 (2011) 2342 –2359.

[16] S. T. Flammia, D. Gross, Y.-K. Liu, J. Eisert, Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators, New Journal of Physics 14 (2012) 095022.

[17] S. Negahban, M. J. Wainwright, Restricted strong convexity and weighted matrix completion: Optimal bounds with noise, The Journal of Machine Learning Research 13 (2012) 1665–1697.

[18] D. Gross, Y. Liu, S. T. Flammia, S. Becker, J. Eisert, Quantum state tomography via compressed sensing, Physical Review Letters 105 (2010) 150401.

[19] D. Gross, Recovering low-rank matrices from few coefficients in any basis, IEEE Transactions on Information Theory 57 (2011) 1548–1566.

[20] R. Ahlswede, A. Winter, Strong converse for identification via quantum channels, IEEE Transactions on Information Theory 48 (2002) 569–579.

[21] J. Tropp, User-Friendly tail bounds for sums of random matrices, Foundations of Computational Mathematics (2011) 1–46.

[22] R. Kueng, Efficient recovery of sparse vectors using anisotropic ensembles, research project, ETH Zürich, June 2011.

[23] M. Ledoux, M. Talagrand, Probability in Banach Spaces: Isoperimetry and Processes, Springer, Berlin, 1991.

[24] D. P. Bertsekas, A. Nedi, A. E. Ozdaglar, Convex analysis and optimization, Athena Scientific, Belmont, MA, 2003.

[25] M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, B. Reed, Probabilistic methods for algorithmic discrete mathematics, Vol. 16, Springer, 1998.

[26] M. Ohliger, V. Nesme, D. Gross, Y. Liu, J. Eisert, Continuous-variable quantum compressed sensing, preprint arXiv:1111.0853 (2011).

Robust Nonnegative Sparse Recovery and the Nullspace Property of 0/1 Measurements

Richard Küng*, Peter Jung†

*Institute for Theoretical Physics, University of Cologne

†Communications and Information Theory Group, Technische Universität Berlin

rkueng@thp.uni-koeln.de, peter.jung@tu-berlin.de

Abstract

We investigate recovery of nonnegative vectors from non–adaptive compressive measurements in the presence of noise of unknown power. It is known in literature that under additional assumptions on the measurement design recovery is possible in the noiseless setting with nonnegative least squares without any regularization. We show that such known uniquenes results carry over to the noisy setting. We present guarantees which hold instantaneously by establishing the relation to the robust nullspace property. As an important example, we establish that an $m \times n$ random iid. $0/1$–valued Bernoulli matrix has with overwhelming probability the robust nullspace property for $m = \mathcal{O}(s\log(n))$ and is applicable in the nonnegative case. Our analysis is motivated by applications in wireless network activity detection.

I. Introduction

Recovery of lower complexity objects by observations far below the Nyquist rate has applications in physics, applied math, and many engineering disciplines. Moreover,it is one of the key tools for facing challenges in data processing (like big data and the Internet of Things), wireless communications (the 5th generation of the mobile cellular network) and large scale network control. Compressed Sensing (CS), with its origin in the recovery of sparse or compressible vectors has, in particular, stimulated the research community to investigate further directions of compressibility and low-dimensional structures which allow the recovery from low-rate samples and with efficient algorithms. In many applications, the objects of interest exhibit further structural constraints which should by exploited in reconstution algorithms. Take, for instance, the following setting which appears naturally in communication protocols: the components of sparse information carrying vectors are taken from a finite alphabet or the data vectors are lying in specific subspaces. Similarly, in network traffic estimation and anomaly detection from end-to-end measurements, the parameters are restricted to particular lower-dimensional domains. Finally, the signals occurring in imaging problems are typically constrained to non-negative intensities.

Our work is partially inspired by the task of identifiying sparse network activation patterns in a large-scale asynchronous wireless network: suppose that, in order to indicate its presence, each active device node transmits an individual sequence into a noisy wireless channel. All such sequences are multiplied with individual, but unknown, channel amplitudes[1] and finally superimposed at the receiver. The receiver's task then is to detect all active devices and the corresponding channel amplitudes from this global superposition (note that each device is uniquely characterized by the sequene it transmits). This problem can be re-cast as the task of estimating non-negative sparse vectors from noisy linear observations.

Such non-negative and sparse structures also arise naturally in certain empirical inference problems, like network tomography [1], [2], statistical tracking (see e.g. [3]) and compressed imaging of intensity patterns [4]. The underlying mathematical problem has received considerable attention in its own right [5], [6], [7], [8], [9]. It has been shown that measurement matrices $\mathbf{A} \in \mathbb{R}^{m \times n}$ coming from *outwardly s–neighborly polytopes* [10] and matrices $\mathbf{A} \in \mathcal{M}^{+}$ whose *row span intersects the positive orthant*[2] [11] maintain an intrinsic uniqueness property for non-negative, s-sparse vectors even in the underdetermined setting ($m < n$). Such uniqueness properties in turn allow for entirely avoiding CS algorithms in the reconstruction step. From an algorithmic point of view, this is highly beneficial. However, all the statements mentioned above are manifestly focussed on idealized scenarios, where no noise is present in the sampling procedure.

Motivated by device detection, we shall overcome this idealization and devise recovery protocols that are robust towards any form of additive noise. Our results have the added benefit that no a-priori bound on the noise step is required in the reconstruction algorithm.

A. *Main Results*

Let us introduce some notation and then state our main findings. Throughout our work we endow $\mathbb{R}^n$ with the partial ordering induced by the nonnegative orthant, i.e. $\mathbf{x} \leq \mathbf{z}$ if and only if $x_i \leq z_i$ for all $1 \leq i \leq n$. Here, $x_i = \langle \mathbf{e}_i, \mathbf{x} \rangle$ are the components of $\mathbf{x}$ with respect to the standard basis $\{\mathbf{e}_i\}_{i=1}^{n}$. Similarly, we write $\mathbf{x} < \mathbf{z}$ if strict inequality holds in each component. Consequently, we write $\mathbf{x} \geq \mathbf{0}$ to indicate that $\mathbf{x}$ is (entry-wise) nonnegative. For $1 \leq p \leq \infty$, we denote the ℓ_p–norms of vectors by $\|\cdot\|_{\ell_p}$ and $\|\cdot\|$ is the usual operator/matrix norm. The sparsity of a vector $\mathbf{x}$ is denoted by $\|\mathbf{x}\|_{\ell_0} := |\mathrm{supp}(\mathbf{x})| \leq s$ where $\mathrm{supp}(\mathbf{x}) := \{i : x_i \neq 0\}$ is its support in the standard basis.

Mathematically, we are interested in recovering sparse, nonnegative vectors $\mathbf{x} \in \mathbb{R}^n$ from $m \ll n$ erronous linear measurements of the form $y_i = \mathbf{a}_i^T \mathbf{x} + e_i$. Here, the vectors $\mathbf{a}_i \in \mathbb{R}^n$ model the different measurement operations and e_i is additive noise of arbitrary size and nature. By encompassing all $\mathbf{a}_i$'s as rows of a sampling matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and defining $\mathbf{y} = (y_1, \ldots, y_m)^T$, as well as $\mathbf{e} = (e_1, \ldots, e_m)^T$, such a sampling procedure can succingtly be written as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}. \tag{1}$$

Several conditions on $\mathbf{A}$ are known which are sufficient to ensure that a sparse vector $\mathbf{x}$ can be robustly estimated from measurements $\mathbf{y}$. A famous condition is the *restricted isometry property* (RIP). A matrix $\tilde{\mathbf{A}}$ is said to

[1] This can be justified under certain assumptions like pre-multiplications using channel reciprocity in time–division multiplexing.

[2] See (7) below for a precise definition.

be s–RIP, if it acts almost isometrically on s–sparse vectors, meaning that there exists a $\delta_s \in [0,1)$ such that $|\|\tilde{\mathbf{A}}\mathbf{x}\|_{\ell_2}^2 - \|\mathbf{x}\|_{\ell_2}^2| \leq \delta_s \|\mathbf{x}\|_{\ell_2}^2$ for all s–sparse $\mathbf{x}$. When dealing with random matrices $\mathbf{A}$, one has also to distinguish between uniform and non–uniform guarentees[3]. It is a well-known fact that RIP is only sufficient but not necessary for uniform recovery. Overcoming this asymmetry, the notion of a *nullspace property* assures that no s-sparse vectors lie in the kernel of $\mathbf{A}$. Hence, the NSP is both a sufficient and necessary condition for recovery. Proving that (1) indeed allows for *robustly* recovering any s-sparse $\mathbf{x}$ in the presence of noise therefore is equivalent to establishing that $\mathbf{A}$ obeys a *robust nullspace property of order s* (NSP) [12, Chapter 4]. Our first main technical contribution is a substantial strengthening of the implications of such an NSP for reconstructing nonnegative sparse vectors:

Theorem 1. *Suppose that $\mathbf{A}$ obeys the NSP of order $s \leq n$ from Def. 3 and moreover admits a strictly–positive linear combination of its rows ($\mathbf{A} \in \mathcal{M}^+$, i.e., $\exists \mathbf{t} \in \mathbb{R}^m$ such that $\mathbf{w} = \mathbf{A}^T\mathbf{t} > \mathbf{0}$). Then, the following bound holds for any s-sparse $\mathbf{x} \geq \mathbf{0}$ and any $\mathbf{z} \geq \mathbf{0}$:*

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq \frac{D'}{\sqrt{m}} \|\mathbf{A}(\mathbf{z} - \mathbf{x})\|_{\ell_2} . \tag{2}$$

The constant D' only[4] depends on the quality of NSP and the conditioning of the strictly positive vector $\mathbf{w}$.

We are interested in retrieving $\mathbf{x}$ from the measurements $\mathbf{y}$ in (1). Inserting this equation into the r.h.s of (2) and applying the triangle inequality reveals

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq \frac{D'}{\sqrt{m}} (\|\mathbf{A}\mathbf{z} - \mathbf{y}\|_{\ell_2} + \|\mathbf{e}\|_{\ell_2}) \quad \forall \mathbf{z} \geq \mathbf{0}.$$

This data-dependent bound suggests to minimize its right hand side over the “free parameter” $\mathbf{z} \geq \mathbf{0}$ in order to get an estimator $\mathbf{x}^\sharp$ of $\mathbf{x}$, i.e.

$$\mathbf{x}^\sharp = \arg\min_{\mathbf{0} \leq \mathbf{z} \in \mathbb{R}^n} \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_{\ell_2} . \tag{3}$$

This is a simple *nonnegative least squares regression* (NNLS) that does not require any assumptions on the noise $\mathbf{e}$. Since the target vector $\mathbf{x}$ is itself nonnegative and therefore a feasible point of (3), we can furthermore conclude

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}^\sharp\|_{\ell_2} \leq & \frac{D'}{\sqrt{m}} (\arg\min_{\mathbf{z} \geq \mathbf{0}} \|\mathbf{A}\mathbf{z} - \mathbf{y}\|_{\ell_2} + \|\mathbf{e}\|_{\ell_2}) \\ \leq & \frac{D'}{\sqrt{m}} (\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\ell_2} + \|\mathbf{e}\|_{\ell_2}) = \frac{2D'}{\sqrt{m}} \|\mathbf{e}\|_{\ell_2}, \end{aligned} \tag{4}$$

where we have once more resorted to (1). Consequently, Theorem 1 assures that solving (3) yields an estimator of any s-sparse vector $\mathbf{x} \geq \mathbf{0}$. Moreover, this estimator is robust towards additive noise in the sampling process. Such a recovery guarantee is (up to multiplicative constants) as strong as existing ones for different reconstruction algorithms, including the LASSO and Dantzig selectors, as well as basis pursuit denoising (BPDN) (see [12] and references therein). However, on the contrary to them, algorithms for solving (3) require neither an explicit a-priori bound

[3] Non-uniform guarantees hold w.h.p. for priorly fixed vectors $\mathbf{x}$, while uniform guarantees assure recovery of all s-sparse vectors simultaneously. RIP is an example for the latter.

[4] See Theorem 4 below for explicit dependencies.

$\eta \geq \|\mathbf{e}\|_{\ell_2}$ on the noise, nor an $\|\cdot\|_{\ell_1}$ regression term. This *remarkable simplicity* is caused by the non-negativity constraint $\mathbf{z} \geq \mathbf{0}$ and the geometric restrictions it imposes. Also, these assertions stably remain true, if we consider approximately sparse target vectors instead of perfectly sparse ones (see Theorem 4 below).

In order to underline the applicability of Theorem 1, we consider nonnegative $0/1$–Bernoulli sampling matrices and prove that they meet the requirements of said statemnt with high probability (w.h.p).

Theorem 2. *Let* $\mathbf{A}$ *be a sampling matrix whose entries are independently chosen from a* $0/1$*–Bernoulli distribution with parameter* $p \in [0,1]$*, i.e.* $\Pr[1] = p$ *and* $\Pr[0] = 1 - p$*. Fixing* $s \leq n$ *and setting*

$$m \geq \frac{C}{(p(1-p))^2} s \left(\log(n) + \frac{p}{1-p} \right) \tag{5}$$

assures that $\mathbf{A}$ *obeys the NSP from Definition 3 and the vector* $\mathbf{w} := \mathbf{A}^T \left(\frac{1}{pm} \mathbf{1} \right)$ *obeys* $\max_{1 \leq i \leq n} |w_i - 1| < 1/2$ *(and is thus strictly positive) with probability at least* $1 - (n+1)\mathrm{e}^{-C' p^2 (1-p)^2 m}$.

Combining this statement with (4) implies that w.h.p. such Bernoulli matrices allow for uniformly and stably reconstructing sparse, nonnegative vectors $\mathbf{x}$ via Alg. (3). We demonstrate this numerically in Figure 1. Up to our knowledge, this is the first rigorous proof that $0/1$–matrices tend to obey a strong version of the nullspace property. The challenging difference to existing NSP and RIP results is the fact that the individual random entries of $\mathbf{A}$ are not centered, ($\mathbb{E}\left[\mathbf{A}_{k,j}\right] = p \neq 0$). Thus, the covariance matrix of $\mathbf{A}$ admits a condition number of $\kappa(\mathbb{E}[\mathbf{A}^T\mathbf{A}]) = 1 + \frac{pn}{1-p}$, which underlines the ensemble's anisotropy. Traditional proof techniques, like establishing an RIP, are either not applicable in such a setting, or yield sub-optimal results [13], [14]. This is not true for Mendelson's small ball method [15], [16] (see also [17]), which we employ in our proof. This method is a strong general purpose tool whose applicability only requires row-wise independence, not centeredness. In the conceptually similar problem of reconstructing low rank matrices from rank-one projective measurements (which arises e.g. from the PhaseLift approach for phase retrieval [18]), applying this technique allowed for establishing strong null space properties, despite a similar degree of anisotropy in the sampling model [19]. A detailed survey of the applicability of Mendelson's small ball method for compressed sensing was recently presented in [20].

Organization of the Paper: In Section II we explain our motivating application in more detail and rephrase activity detection as a nonnegative sparse recovery problem. Then, we provide an overview on prior work and known results regarding this topic. In Section III we show that recovery guarentees in the presence of noise are governed here by the *robust nullspace property* (see here [12]) *under nonnegative constraints* which hasn't been fully analyzed so far in literature. It turns out that this property assures that any nonnegative s–sparse vector can be robustly recovered using conventional nonnegative least–squares. We stress out that such an algorithm requires *no apriori–knowledge* on the norm of the noise vector. Finally, in Section IV we analyze binary measurements matrices having iid. random $0/1$–valued entries and we show that with overwhelming probability such matrices admit the robust nullspace property on nonnegative vectors. We obtain this result make use of a recent tool, known as "Mendelson's small ball method" which has already used by one of the authors in a related matrix recovery problem [19].

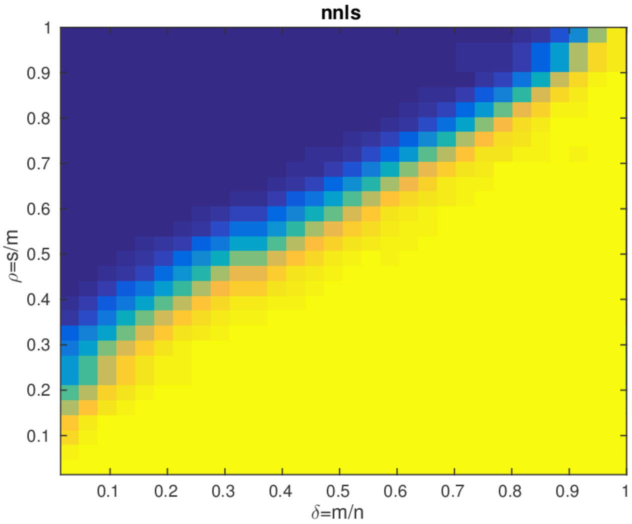


Fig. 1: Phase transition for NNLS in (3) – for iid. $0/1$–Bernoulli measurement matrices in the noiseless case. More details are given in Section V.

II. System Model and Problem Statement

A. Activity Detection in Wireless Networks

Let $\mathbf{A} = (\mathbf{s}_j)_{j=1}^n \in \mathbb{R}^{m \times n}$ be a matrix with n real columns $\mathbf{s}_j \in \mathbb{R}^m$. In our network application [21], the columns $\mathbf{s}_j$ are the individual sequences of length m transmitted by the active devices. These sequences are transmitted simultaneously and each of them is multiplied by an individual amplitude that depends on transmit power and other channel conditions. In practice this can be achieved for example using the channel reciprocity principle in time-division multiplexing so that the devices have knowledge about the complex channel coefficients and perform a corresponding pre-multiplication to correct for the phase. At a single receiver, all these modulated sequences are superimposed, because a single wireless medium is shared by all devices. We model such a situation by an unknown non-negative vector $\mathbf{0} \leq \mathbf{x} \in \mathbb{R}^n$, where $x_i > 0$ indicates that a device with sequence i is active with amplitude x_i ($x_i = 0$ implies that a device is inactive). We point out that, due to path loss in the channel, the individual received amplitudes x_i of each active device are unknown to the receiver as well. Here, we focus on networks that contain a large number n of registered devices, but, at any time, only a small unknown fraction, say $s \ll n$, of these devices are active.

Communicating activity patterns, that is $\text{supp}(\mathbf{x}) = \{i : x_i \neq 0\}$, and the corresponding list of received amplitudes/powers ($\mathbf{x} \geq \mathbf{0}$ itself) in a traditional way would require an $\mathcal{O}(n)$ resources to perform this task. We aim therefore for a reduction of the signaling time m by exploiting the facts that (i) $\mathbf{x} \geq \mathbf{0}$ is non-negative and (ii) the vector $\mathbf{x}$ is s-sparse, i.e. $\|\mathbf{x}\|_{\ell_0} \leq s$. Hence, we assume that $s \leq m \ll n$. Obviously, in such a scenario the resulting system of linear equations cannot be directly inverted. A reasonable approach towards recovery is to consider the program:

$$\arg\min \|\mathbf{x}\|_{\ell_0} \;\text{ s.t. }\; \mathbf{A}\mathbf{x} = \mathbf{y} \;\&\; \mathbf{x} \geq 0$$

Combinatorial problems of this type are infamous for being NP-hard in general. A common approach to circumvent this obstacle is to consider convex relaxations. A prominent relaxation is to replace $\|\cdot\|_{\ell_0}$ with the ℓ_1-norm. The resulting algorithm can then be re-cast as an efficiently solvable linear program. However, such approaches become more challenging when robustness towards additive noise is required, in particular if the type and the strength of the noise is itself unknown. In our application, noisy contributions inevitable arises due to quantization, thermal noise and other interferences. If the noisy measurements are of the form (1) (i.e. $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ where the vector $\mathbf{e}$ is an additive distortion) a well–known modification is then to consider

$$\arg\min \|\mathbf{x}\|_{\ell_1} \;\text{ s.t. }\; \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\ell_2} \leq \eta \;\&\; \mathbf{x} \geq \mathbf{0}. \tag{6}$$

While this is not a linear problem anymore, it is still convex and is computationally tractable. In practice further modifications are necessary to solve such problems also sufficiently fast and efficiently, see [21]. However, having access to an apriori bound η on $\|\mathbf{e}\|_{\ell_2}$ is essential for (i) posing this problem and (ii) solving it using certain algorithms (stopping conditions etc.). Suppose, for instance, that $\mathbf{e}$ is iid normal distributed. Then $\|\mathbf{e}\|_{\ell_2}^2$ admits a χ^2-distribution of order m and feasibility is assured w.h.p., when taking η in terms of second moments. However, much less is known for different noise distributions or for situations, where second moment information about the noise is challenging to acquire.

One option to tackle problems of this kind is to establish a *quotient property* for the measurement matrix $\mathbf{A}$ [12]. However, this property is geared towards Gaussian measurements and it is challenging to establish it, if $\mathbf{A}$ follows a different random model. We shall show below that, interestingly, requiring $\mathbf{A} \in \mathcal{M}^+$ instead allows for drawing similar conclusions.

B. Prior Work on Recovery of Nonnegative Sparse Vectors

One of the first works in the noiseless setting is due to Donoho et al.n [4] on the "nearly black object". It furthers understanding of the "maximum entropy inversion" method to recover sparse (nearly–black) images in radio astronomy. In [10], Donoho and Tanner investigated this subject more directly. The question is, when $\mathbf{A}$ intrinsically ensures that for each s–sparse $\mathbf{x}^{(0)}$ only one solution is feasible:

$$\{\mathbf{y} \,|\, \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x}^{(0)} \,\&\, \mathbf{x} \geq \mathbf{0}\} = \{\mathbf{x}^{(0)}\}$$

At the center of their work is the notion of *outwardly s–neighborly polytopes*. Assume w.l.o.g. that all columns $\mathbf{s}_j$ of $\mathbf{A}$ are non-zero and define their convex hull

$$P_{\mathbf{A}} := \text{conv}(\mathbf{s}_1, \ldots, \mathbf{s}_n).$$

This polytope is called *s-neighborly*, if every set of s vertices spans a face of $P_{\mathbf{A}}$. If this is the case, the polytope $P^0_{\mathbf{A}} := \text{conv}(P_{\mathbf{A}} \cup \{0\})$ is called then *outwardly s-neighborly*. They then move on to prove that the solution to

$$\arg\min \|\mathbf{x}\|_{\ell_0} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{y}$$

is unique if and only if $P_{\mathbf{A}}^0$ is outwardly s–neighborly (see [10]). Another notion is the set of full-rank $m \times n$-matrices having *intersection of its row space with the positive orthant* as introduced in [11]:

$$\mathcal{M}^+ = \{\mathbf{A} : \exists \mathbf{t} \in \mathbb{R}^m \ \mathbf{A}^* \mathbf{t} > 0\}. \tag{7}$$

Note that both structures are related in the sense that $\mathbf{A} \in \mathcal{M}^+$, if and only if $0 \notin P_{\mathbf{A}}$ [22]. In [11] Bruckstein et al. investigated the recovery of nonnegative vectors by (6) and modifications of OMP using a coherence-based approach. They obtained numerical evidence for unique recovery in the regime $s = \mathcal{O}(\sqrt{n})$. Later, Wang and coauthors [22] have analyzed non-negativity priors for vector and matrix recovery using an RIP-based analysis. Concretely, they translated the well–known RIP–result of random iid. ± 1–Bernoulli matrices (see for example [23]) to $0/1$-measurements in the following way. Let

$$\mathbf{1}_n := (1, \dots, 1)^T$$

denote the “all-ones” vector in $\mathbb{R}^n$. Perform measurements using an $(m+1) \times n$ matrix $\mathbf{A}^1 = \left(\mathbf{1}_n^T | \mathbf{A}^T\right)^T$ which consists of an all-ones row $\mathbf{1}_n$ appended by a random iid. $0/1$–valued $m \times n$ matrix $\mathbf{A}$. By construction, the first noiseless measurement on a nonnegative vector $\mathbf{x}$ returns its ℓ_1–norm $\|\mathbf{x}\|_{\ell_1} = \langle \mathbf{1}_n, \mathbf{x} \rangle$. Rescaling and substracting this value from the m remaining measurements then results in ± 1–measurements. This insight allows for an indirect nullspace characterization of $\mathbf{A}$ in terms of the RIP–constant δ_{2s} (see above, paragraph below (1)) of iid ± 1–Bernoulli random matrices $\tilde{\mathbf{A}}$. More precisely [24]: For each $\mathbf{v} \in \mathcal{N}(\tilde{\mathbf{A}})$ in the nullspace $\mathcal{N}(\tilde{\mathbf{A}})$ of $\tilde{\mathbf{A}}$, an (ℓ_1, ℓ_1)–nullspace property is valid. Mathematically this means

$$\|\mathbf{v}_S\|_{\ell_1} \leq \frac{\sqrt{2}\delta_{2s}}{1 - \delta_{2s}} \|\mathbf{v}_{\bar{S}}\|_{\ell_1} \tag{8}$$

for all $\mathbf{v} \in \mathcal{N}(\tilde{\mathbf{A}})$ and $|S| \leq s$. Combining this with $\mathcal{N}(\mathbf{A}^1) \subset \mathcal{N}(\tilde{\mathbf{A}})$ then allows for proving unique recovery in regime $s = \mathcal{O}(n)$ with overwhelming probability.

However, so far, all these results manifestly focus on noiseless measurements. Thus, the robustness of these approaches towards noise corruption needs to be examined. Foucart, for instance, considered the *ℓ_1–squared nonnegative regularization* [9]:

$$\min_{\mathbf{x} \geq 0} \|\mathbf{x}\|_{\ell_1}^2 + \lambda^2 \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_{\ell_2}^2 \tag{9}$$

which can be re-cast as nonnegative least-squares problem. He then showed that for stochastic matrices[5] the solution of (9) converges to the solution of (6) for $\lambda \to \infty$.

Contrary to this, we aim at establishing even stronger recovery guarantees that, among other things, *do not require an a priori noise bound*. We have already mentioned that the quotient property would assure such bounds for Gaussian matrices in the optimal regime. But $m \times n$ Gaussian matrices fail to be in $\mathcal{M}^+$ with probability approaching one as long as $\lim_{n \to} m/n < \frac{1}{2}$ [22]. On the algorithmic side, there exists variations of certain regression methods where the regularization parameter can be choosen independent of the noise power – see the overview article [25]

[5] Recall that a matrix is stochastic, if all entries are non-negative and all columns sum up to one.

for more details. For the LASSO selector, in particular, such modifications are known as the "scaled LASSO" and "square root LASSO" [26], [27].

Non-negativity as a further structural constraint has also been investigated in the statistics community. But these works focus on the averaged case with respect to (sub–)Gaussian additive noise, whereby we consider instantaneous guarantees. Slawski and Hein [8], as well as Meinshausen [7] have recently investigated this averaged setting.

III. Nullspace Property with Nonnegative constraints

We use the following notation. For a given vector $\mathbf{x} \in \mathbb{R}^n$ and a set $S \subset [n] := [1 \dots n]$ we denote by $\mathbf{x}_S$ the vector containing only the coefficients of $\mathbf{x}$ in S. Let $\bar{S}$ the complement of S in $[1 \dots n]$ such that $\mathbf{x} = \mathbf{x}_S + \mathbf{x}_{\bar{S}}$. The ℓ_q–error of the best s–term approximation of a vector $\mathbf{x}$ will be denoted by $\sigma_k(\mathbf{x})_{\ell_q}$. The well–known convex relaxation of the ℓ_0-minimization with respect to an apriori ℓ_2–bound η on the residual $\mathbf{Ax} - \mathbf{y}$ is *basis pursuit denoising* (BPDN):

$$\Delta_\eta(\mathbf{y}) = \arg\min \|\mathbf{x}\|_{\ell_1} \quad \text{s.t.} \quad \|\mathbf{Ax} - \mathbf{y}\|_{\ell_2} \leq \eta \tag{10}$$

A. The robust nullspace property

Let us recall the definition of the ℓ_2*–robust nullspace property* with respect to the ℓ_2–norm [12, Def. 4.21].

Definition 3 (ℓ_2–robust nullspace property). *A $m \times n$ matrix $\mathbf{A}$ satisfies the* ℓ_2-robust null space property *of order s with parameters $\rho \in (0,1)$ and $\tau > 0$, if:*

$$\|\mathbf{v}_S\|_{\ell_2} \leq \frac{\rho}{\sqrt{s}} \|\mathbf{v}_{\bar{S}}\|_{\ell_1} + \tau \|\mathbf{Av}\|_{\ell_2} \quad \textit{for all} \quad \mathbf{v} \in \mathbb{R}^n \tag{11}$$

holds for all $S \subset [n]$ with $|S| \leq s$.

The ℓ_2-robust nullspace property order s (s–NSP) allows for drawing the following conclusion [12, Theorem 4.25]: for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq \frac{C}{\sqrt{s}} \left(\|\mathbf{z}\|_{\ell_1} - \|\mathbf{x}\|_{\ell_1} + 2\sigma_s(\mathbf{x})_{\ell_1}\right) + D \|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2} \tag{12}$$

is true, where $C = \frac{(1+\rho)^2}{1-\rho}$ and $D = \frac{3+\rho}{1-\rho}\tau$. Replacing $\mathbf{z}$ with the BPDN minimizer $\mathbf{x}_\eta = \Delta_\eta(\mathbf{y})$ from (10) for the sampling model $\mathbf{y} = \mathbf{Ax} + \mathbf{e}$ then implies

$$\begin{aligned} \|\mathbf{x} - \mathbf{x}_\eta\|_{\ell_2} &\leq \frac{2C}{\sqrt{s}} \sigma_s(\mathbf{x})_{\ell_1} + D \|\mathbf{y} - \mathbf{e} - \mathbf{Ax}_\eta\|_{\ell_2} \leq \frac{2C}{\sqrt{s}} \sigma_s(\mathbf{x})_{\ell_1} + D \|\mathbf{y} - \mathbf{Ax}_\eta)\|_{\ell_2} + \|\mathbf{e}\|_{\ell_2} \\ &\leq \frac{2C}{\sqrt{s}} \sigma_s(\mathbf{x})_{\ell_1} + (D+1)\eta, \end{aligned} \tag{13}$$

provided that $\|\mathbf{e}\|_{\ell_2} \leq \eta$ is true. This estimate follows from combining $\|\mathbf{x}_\eta\|_{\ell_1} \leq \|\mathbf{x}\|_{\ell_1}$ and with $\|\mathbf{y} - \mathbf{Ax}_\eta)\|_{\ell_2} \leq \eta$. Once more, we point out that this estimate is only valid if an appropriate η is *known*.

B. Nonnegative Constraints

Here we will prove now a variation of (12) (Theorem 4.25 in [12]) which holds for nonnegative vectors and matrices in $\mathcal{M}^+$. For such matrices we define a condition number by

$$\kappa(\mathbf{A}) = \min\{\|\mathbf{W}\|\|\mathbf{W}^{-1}\| \ \ |\exists \mathbf{t} \text{ with } \mathbf{W} = \text{diag}(\mathbf{A}^T\mathbf{t}) > 0\} \tag{14}$$

Note that for diagonal matrices $\mathbf{W}$ with non-negative entries $\kappa(\mathbf{W}) = \|\mathbf{W}\|\|\mathbf{W}^{-1}\|$.

Theorem 4. *Let $\mathbf{A} \in \mathcal{M}^+$ obeying the s-NSP with parameters ρ and τ, and let $\kappa = \kappa(\mathbf{A})$ be its condition number. If $\kappa\rho < 1$, then*

$$\|\mathbf{x}-\mathbf{z}\|_{\ell_2} \leq \frac{2C}{\sqrt{s}}\sigma_s(\mathbf{x}) + D\left(\|\mathbf{t}\|_{\ell_2} + \tau\right)\|\mathbf{A}(\mathbf{x}-\mathbf{z})\|_{\ell_2}$$

is true for all nonnegative vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$. The constants amount to

$$C = \kappa\frac{(1+\kappa\rho)^2}{1-\kappa\rho} \text{ and } D = \kappa\frac{3+\kappa\rho}{1-\kappa\rho}. \tag{15}$$

Comparing this to (12) reveals, that the ℓ_1–term $(\|\mathbf{z}\|_{\ell_1} - \|\mathbf{x}\|_{\ell_1})$ has disappeared. Let us exploit this by reproducing the steps in (13). If we once more use $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, and apply the triangle inequality, we obtain

$$\|\mathbf{x}-\mathbf{z}\|_{\ell_2} \leq \frac{c_1}{\sqrt{s}}\sigma_s(\mathbf{x})_{\ell_1} + c_2\|\mathbf{y}-\mathbf{A}\mathbf{z}\|_{\ell_2} + \|\mathbf{e}\|_{\ell_2} \tag{16}$$

This simple observation already highlights that CS–oriented algorithms, which essentially minimize the ℓ_1–norm, are not required anymore in the non–negative case. Instead, in order to get good estimates it makes sense to minimize the r.h.s. of the bound over the "free" parameter $\mathbf{z} \geq \mathbf{0}$. Doing so, results in s *nonnegative least–squares* estimate for $\mathbf{x}$ by minimizing $\|\mathbf{y}-\mathbf{A}\mathbf{z}\|_{\ell_2}$ subject to $\mathbf{z} \geq \mathbf{0}$. To prove this theorem, we will need two auxiliar statements.

Lemma 5. *Suppose that $\mathbf{A}$ obeys the s–NSP with parameters ρ and τ, and set $\mathbf{W} = diag(\mathbf{w})$, where $\mathbf{w} > \mathbf{0}$ is strictly positive. Then, $\mathbf{A}\mathbf{W}^{-1}$ also obeys the s–NSP with parameters $\tilde{\rho} = \kappa(\mathbf{W})\rho$ and $\tilde{\tau} = \|\mathbf{W}\|\tau$.*

Proof: First, since $\mathbf{W}$ is diagonal we can conclude for any vector $\mathbf{v} \in \mathbb{R}^n$ and any set $S \subset [n]$ that $\mathbf{W}^{-1}\mathbf{v}_S = (\mathbf{W}^{-1}\mathbf{v})_S$ (same for $\bar{S}$). Also, $\mathbf{A}$ obeys the s-NSP which in turn implies for any $|S| \leq s$:

$$\begin{aligned}\|\mathbf{v}_S\|_{\ell_2} &= \|\mathbf{W}\mathbf{W}^{-1}\mathbf{v}_S\|_{\ell_2} \leq \|\mathbf{W}\|\|(\mathbf{W}^{-1}\mathbf{v})_S\|_{\ell_2} \leq \|\mathbf{W}\|\left(\frac{\rho}{\sqrt{s}}\|(\mathbf{W}^{-1}\mathbf{v})_{\bar{S}}\|_{\ell_2} + \tau\|\mathbf{A}\mathbf{W}^{-1}\mathbf{v}\|_{\ell_2}\right)\\ &= \tilde{\rho}\sigma_s(\mathbf{v}) + \tilde{\tau}\|\mathbf{A}\mathbf{W}^{-1}\mathbf{v}\|_{\ell_2}.\end{aligned}$$

■

Lemma 6. *Suppose that $\mathbf{W} := \text{diag}\left(\mathbf{A}^T\mathbf{t}\right) > 0$ for some $\mathbf{t} \in \mathbb{R}^m$. Then any pair $\mathbf{x}, \mathbf{z} \geq 0$ obeys*

$$\|\mathbf{W}\mathbf{z}\|_{\ell_1} - \|\mathbf{W}\mathbf{x}\|_{\ell_1} \leq \|\mathbf{t}\|_{\ell_2}\|\mathbf{A}\left(\mathbf{x}-\mathbf{z}\right)\|_{\ell_2} \tag{17}$$

Proof: Note that, by construction, $\mathbf{W}$ is symmetric and preserves positivity of vectors. These features together with positivity of $\mathbf{z}$ imply

$$\|\mathbf{W}\mathbf{z}\|_{\ell_1} = \langle\mathbf{1}, \mathbf{W}\mathbf{z}\rangle = \langle\mathbf{W}\mathbf{1}, \mathbf{z}\rangle = \langle\text{diag}(\mathbf{A}^T\mathbf{t})\mathbf{1}, \mathbf{z}\rangle = \langle\mathbf{A}^T\mathbf{t}, \mathbf{z}\rangle = \langle\mathbf{t}, \mathbf{A}\mathbf{z}\rangle.$$

An analogous reformulation is true for $\|\mathbf{W}\mathbf{x}\|_{\ell_1}$ and combining these two reveals

$$\|\mathbf{W}\mathbf{z}\|_{\ell_1} - \|\mathbf{W}\mathbf{x}\|_{\ell_1} = \langle \mathbf{t}, \mathbf{A}\left(\mathbf{z} - \mathbf{x}\right)\rangle \leq \|\mathbf{t}\|_{\ell_2}\|\mathbf{A}(\mathbf{z} - \mathbf{x})\|_{\ell_2}$$

due to Cauchy-Schwarz. ■

Proof of Theorem 4: The assumption $\mathbf{A} \in \mathcal{M}^+$ assures that there exists $\mathbf{t} \in \mathbb{R}^m$ such that $\mathbf{w} = \mathbf{A}^T\mathbf{t} > \mathbf{0}$ and we define $\mathbf{W} := \mathrm{diag}(\mathbf{w})$. By construction, $\mathbf{W}$ is invertible and admits a condition number $\kappa = \|\mathbf{W}\|\|\mathbf{W}^{-1}\|$. Thus, we can write

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} = \|\mathbf{W}^{-1}\mathbf{W}\left(\mathbf{x} - \mathbf{z}\right)\|_{\ell_2} \leq \|\mathbf{W}^{-1}\|\|\mathbf{W}(\mathbf{x} - \mathbf{z})\|_{\ell_2}$$

for any pair $\mathbf{x}, \mathbf{z} > \mathbf{0}$. Now, since $\mathbf{A}$ obeys the s–NSP, Lemma 5 assures that $\mathbf{A}\mathbf{W}^{-1}$ has s–NSP as well, with parameters $\tilde{\rho} = \kappa\rho$ and $\tilde{\tau} = \|\mathbf{W}\|\tau$. Thus, from (12) we conclude that for vectors $\mathbf{W}\mathbf{x}$ and $\mathbf{W}\mathbf{z}$ we have

$$\begin{aligned}\|\mathbf{W}(\mathbf{x} - \mathbf{z})\|_{\ell_2} &\leq \frac{C'}{\sqrt{s}}\left(\|\mathbf{W}\mathbf{z}\|_{\ell_1} - \|\mathbf{W}\mathbf{x}\|_{\ell_1} + 2\sigma_s(\mathbf{W}\mathbf{x})_{\ell_1}\right) + D'\|\mathbf{A}(\mathbf{x} - \mathbf{z})\|_{\ell_2} \\ &\overset{(17)}{\leq} \frac{2C'\|\mathbf{W}\|}{\sqrt{s}}\sigma_s(\mathbf{x})_{\ell_1} + (\frac{C'\|\mathbf{t}\|_{\ell_2}}{\sqrt{s}} + D')\|\mathbf{A}(\mathbf{x} - \mathbf{z}\|_{\ell_2}.\end{aligned}$$

Here, we invoked Lemma 6 in the last step, as well as the relation $\sigma_s(\mathbf{W}\mathbf{x})_{\ell_1} \leq \|\mathbf{W}\|\sigma_s(\mathbf{x})_{\ell_1}$. The constants above amount to $C' = \frac{(1+\tilde{\rho})^2}{1-\rho} = \frac{(1+\kappa\rho)^2}{1-\kappa\rho}$ and $D' = \frac{3+\tilde{\rho}}{1-\tilde{\rho}}\tilde{\tau} = \frac{3+\kappa\rho}{1-\kappa\rho}\|\mathbf{W}\|\tau$. So, in summary we obtain

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq \frac{2C'\kappa}{\sqrt{s}}\sigma_s(\mathbf{x})_{\ell_1} + \|\mathbf{W}^{-1}\|(\frac{C'\|\mathbf{t}\|_{\ell_2}}{\sqrt{s}} + D')\|\mathbf{A}(\mathbf{x} - \mathbf{z}\|_{\ell_2}.$$

We shall simplify the second term further by using the fact that $(1 + x)^2 \leq 3 + x$ for any $x \in [0, 1]$, i.e., C' and D'/τ are both upper bounded by $\frac{3+\kappa\rho}{1-\kappa\rho}\|\mathbf{W}\|$. Consequently,

$$\|\mathbf{x} - \mathbf{z}\|_{\ell_2} \leq 2\frac{\kappa C'}{\sqrt{s}}\sigma_s(\mathbf{x}) + \frac{3 + \kappa\rho}{1 - \kappa\rho}\kappa\left(\|\mathbf{t}\|_{\ell_2} + \tau\right)\|A\left(\mathbf{x} - \mathbf{z}\right)\|_{\ell_2},$$

and setting $C := \kappa C'$ and $D = \frac{3+\kappa\rho}{1-\kappa\rho}\kappa$ proves the claim. ■

IV. Robust NSP for 0/1-Bernoulli matrices

In this section, we prove our second main result, namely Theorem 2. Said statements summarizes two results, namely (i) 0/1-Bernoulli matrices $\mathbf{A}$ with $m = Cs\log(n)$ rows obey the robust null space property of order s and (ii) the row space of $\mathbf{A}^T$ allows for constructing a strictly positive vector $\mathbf{w} = \mathbf{A}^T\mathbf{t} > \mathbf{0}$ (that is sufficiently well-conditioned). We will first state the main ideas and prove both statements in subsequent subsections.

A. Sampling model and overview of main proof ideas

Let us start by formally defining the concept of a 0/1-Bernoulli matrix.

Definition 7. *We call $\mathbf{A} \in \mathbb{R}^{m\times n}$ a 0/1-Bernoulli matrix with parameter $p \in [0, 1]$, if every matrix element $[\mathbf{A}]_{i,j}$ is an independent realization of a Bernoulli random variable b with parameter p, i.e.*

$$\Pr\left[b = 1\right] = p \quad \textit{and} \quad \Pr\left[b = 0\right] = 1 - p.$$

Recall that such a Bernoulli variable obeys $\mathbb{E}\left[b\right] = p$ and $\mathrm{Var}(b) = \mathbb{E}\left[\left(b - \mathbb{E}[b]\right)^2\right] = p(1-p)$. By construction, the m rows $\mathbf{a}_1, \ldots, \mathbf{a}_m$ of such a Bernoulli matrix are independent and obey

$$\mathbb{E}\left[\mathbf{a}_k\right] = \sum_{j=1}^{n} \mathbb{E}\left[\mathbf{A}_{k,j}\right] \mathbf{e}_j = p \sum_{j=1}^{n} \mathbf{e}_j = p\mathbf{1}.$$

This expected behavior of the individual rows will be crucial for addressing the second point in Theorem 2: setting

$$\mathbf{w} := \frac{1}{pm} \sum_{k=1}^{m} \mathbf{a}_k = \mathbf{A}^T \left(\frac{1}{pm}\mathbf{1}\right)$$

results in a random vector $\mathbf{w} \in \mathbb{R}^n$ that obeys $\mathbb{E}\left[\mathbf{w}\right] = \mathbf{1} > \mathbf{0}$. Applying a large deviation bound will in turn imply that a realization of $\mathbf{w}$ will w.h.p. not deviate too much from its expectation $\mathbf{1}$ and thus remains strictly positive. We will do this in Subsection IV-C.

However, when turning our focus to establishing null space properties for $\mathbf{A}$, working with 0/1-Bernoulli entries renders such a task more challenging. The simple reason for such a complication is that the individual random entries of $\mathbf{A}$ are not centered, i.e. $\mathbb{E}\left[\mathbf{A}_{k,j}\right] = p \neq 0$. Combining this with independence of the individual entries yields

$$\mathbb{E}\left[\mathbf{a}_k \mathbf{a}_k^T\right] = p^2 \mathbf{1}\mathbf{1}^T + p(1-p)\mathbb{I}.$$

This matrix admits a condition number of $\kappa\left(\mathbb{E}\left[\mathbf{a}_k \mathbf{a}_k^T\right]\right) = 1 + \frac{pn}{1-p}$ which underlines the ensemble's anisotropy. Traditional proof techniques, e.g. establishing an RIP, are either not applicable in such a setting, or yield sub-optimal results [13], [14]. This is not true for Mendelson's small ball method [15], [16] (see also [17]) – a strong general purpose tool whose applicability only requires row-wise independence. In the conceptually similar problem of reconstructing low rank matrix from rank-one projective measurements (which arises e.g. from the PhaseLift approach for phase retrieval [28], [18]) applying this technique allowed for establishing strong null space properties, despite a similar degree of anisotropy in the sampling model [19]. In the next subsection, we adapt the ideas from said paper to our Bernoulli model and succeed in establishing the NSP presented in Theorem 2.

Finally, we point out that a detailed survey of the applicability of Mendelson's small ball method for compressed sensing was recently presented in [20]. However, there centeredness of the individual matrix entries is a key assumption which is not met in our 0/1-Bernoulli model.

B. Null Space Properties for $0/1$-Bernoulli matrices

Recall that Definiton 3 states that a $m \times n$ matrix $\mathbf{A}$ obeys the robust null space property with parameters $\rho \in (0,1)$ and $\tau > 0$, if

$$\|\mathbf{v}_S\|_{\ell_2} \leq \frac{\rho}{\sqrt{s}} \|\mathbf{v}_{\bar{S}}\|_{\ell_1} + \tau \|\mathbf{A}\mathbf{v}\|_{\ell_2} \tag{18}$$

is true for all vectors $\mathbf{v} \in \mathbb{R}^n$ and support sets $S \in [n]$ with support size $|S| \leq s$. Demanding such generality in the choice of the support set is in fact not necessary, see e.g. [12, Remark 4.2]. For a fixed vector $\mathbf{v}$, the above condition holds for any index set S, if it holds for an index set $S_{\max}$ containing the s largest (in modulus) entries of $\mathbf{v}$. Introducing the notation $\mathbf{v}_s := \mathbf{v}_{S_{\max}}$ and $\mathbf{v}_c := \mathbf{v}_{\bar{S}_{\max}}$, the robust null space property (18) holds, provided that every vector $\mathbf{v} \in \mathbb{R}^n$ obeys

$$\|\mathbf{v}_s\|_{\ell_2} \leq \frac{\rho}{\sqrt{s}} \|\mathbf{v}_c\|_{\ell_1} + \tau \|\mathbf{A}\mathbf{v}\|_{\ell_2}. \tag{19}$$

Note that this requirement is invariant under re-scaling and we may w.l.o.g. assume $\|\mathbf{v}\|_{\ell_2} = 1$. Moreover, for fixed parameters s and ρ, any vector $\mathbf{v}$ obeying $\|\mathbf{v}_s\|_{\ell_2} \leq \frac{\rho}{\sqrt{s}} \|\mathbf{v}_c\|_{\ell_1}$ is guaranteed to fulfill (19) by default. Consequently, when aiming to establish null space properties, it suffices to establish condition (19) for the set of unit-norm vectors that do not obey this criterion:

$$T_{\rho,s} := \left\{ \mathbf{v} \in \mathbb{R}^n : \ \|\mathbf{v}\|_{\ell_2} = 1, \ \|\mathbf{v}_s\|_{\ell_2} > \frac{\rho}{\sqrt{s}} \|\mathbf{v}_c\|_{\ell_1} \right\}.$$

As a result, a matrix $\mathbf{A}$ obeys the NSP (18), if

$$\inf \left\{ \|\mathbf{A}\mathbf{v}\|_{\ell_2} : \ \mathbf{v} \in T_{\rho,s} \right\} > \frac{1}{\tau}, \tag{20}$$

holds, where $\tau > 0$ is the second parameter appearing in (18). The task of establishing this is somewhat simplified by the observation that the set $T_{\rho,s}$ exclusively contains vectors that are effectively sparse:

Lemma 8. *For fixed s and ρ, every vector $\mathbf{v} \in T_{\rho,s}$ obeys*

$$\|\mathbf{v}\|_{\ell_1} \leq \sqrt{s} \frac{1+\rho}{\rho} \|\mathbf{v}\|_{\ell_2}.$$

Proof: Note that any $\mathbf{v}_s$ is s-sparse by construction and thus obeys $\|\mathbf{v}_s\|_{\ell_1} \leq \sqrt{s} \|\mathbf{v}_s\|_{\ell_2}$. Combining this with the triangle inequality and the defining feature of the set $T_{\rho,s}$ yields

$$\|\mathbf{v}\|_{\ell_1} = \|\mathbf{v}_s + \mathbf{v}_c\|_{\ell_1} \leq \|\mathbf{v}_s\|_{\ell_1} + \|\mathbf{v}_c\|_{\ell_1} \leq \sqrt{s} \|\mathbf{v}_s\|_{\ell_2} + \frac{\sqrt{s}}{\rho} \|\mathbf{v}_s\|_{\ell_2}$$

and the claim readily follows from $\|\mathbf{v}_s\|_{\ell_2} \leq \|\mathbf{v}\|_{\ell_2}$. ■

Despite such a geometric insight, proving (20) for a given $\mathbf{A}$ is still a daunting task. This situation greatly changes, if we assume that our sampling matrix $\mathbf{A}$ consists of m rows $\mathbf{a}_1, \ldots, \mathbf{a}_m$ that are independent instances of a random vector $\mathbf{a} \in \mathbb{R}^n$. Assuming this, (20) is equivalent to showing

$$\inf_{\mathbf{v} \in T_{\rho,r}} \left(\sum_{k=1}^{m} |\langle \mathbf{a}_k, \mathbf{v} \rangle|^2 \right)^{1/2} > \frac{1}{\tau}. \tag{21}$$

Independence of the $\mathbf{a}_k$'s then allows for establishing this (w.h.p.) by resorting to Mendelson's small ball method [15], [16], [17]:

Theorem 9 (Koltchinskii, Mendelson; Tropp's version [17])**.** *Fix $E \subset \mathbb{R}^n$ and let $\mathbf{a}_1, \ldots, \mathbf{a}_m$ be independent copies of a random vector $\mathbf{a} \in \mathbb{R}^n$. Set $\mathbf{h} = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} \epsilon_k \mathbf{a}_k$, where $\epsilon_1, \ldots, \epsilon_m$ is a Rademacher sequence, and for $\xi > 0$ define*

$$Q_\xi (E, \mathbf{a}) = \inf_{\mathbf{u} \in E} \Pr \left[|\langle \mathbf{a}, \mathbf{u} \rangle| \geq \xi \right], \quad \textit{as well as} \quad W_m (E, \mathbf{a}) = \mathbb{E} \left[\sup_{\mathbf{u} \in E} \langle \mathbf{h}, \mathbf{u} \rangle \right].$$

Then, for any $\xi > 0$ *and* $t \geq 0$ *the following is true with probability at least* $1 - \mathrm{e}^{-2t^2}$*:*

$$\inf_{\mathbf{u} \in E} \left(\sum_{k=1}^{m} |\langle \mathbf{a}_k, \mathbf{u} \rangle|^2 \right)^{1/2} \geq \xi \sqrt{m} Q_{2\xi}(E, \mathbf{a}) - \xi t - 2 W_m(E, \mathbf{a}). \tag{22}$$

In order to establish (21), we can set $E = T_{\rho,s}$, choose ξ and t appropriately and establish suitable bounds for $Q_\xi(T_{\rho,s}, \mathbf{a})$ and $W_m(T_{\rho,r}, \mathbf{a})$. Note that the geometric insight provided in Lemma 8 considerably simplifies this last task. It assures

$$W_m(T_{\rho,s}, \mathbf{a}) = \mathbb{E}\left[\sup_{\mathbf{u} \in T_{\rho,s}} \langle \mathbf{h}, \mathbf{u} \rangle \right] \leq \sup_{\mathbf{u} \in T_{\rho,s}} \|\mathbf{u}\|_{\ell_1} \mathbb{E}\left[\|\mathbf{h}\|_{\ell_\infty} \right] \leq \sqrt{s} \frac{1+\rho}{\rho} \mathbb{E}\left[\|\mathbf{h}\|_{\ell_\infty} \right]$$

and it suffices bound $\mathbb{E}\left[\|\mathbf{h}\|_{\ell_\infty}\right]$ from above. We do this by adapting the techniques from [29, Proposition 13] to the vector case. The calculations are detailed in the appendix and yield

$$\mathbb{E}\left[\|\mathbf{h}\|_{\ell_\infty} \right] \leq \sqrt{4p(1-p)\left(3\log(2n) + \frac{p}{1-p} \right)} \tag{23}$$

under the assumption that the sampling rate m exceeds $\frac{\log(n)}{p^2(1-p)^2}$. Such a bound allows us to deduce

$$W_m\left(T_{\rho,s}, \mathbf{a}\right) \leq \frac{1+\rho}{\rho} \sqrt{4sp(1-p)\left(3\log(2n) + \frac{p}{1-p} \right)} \tag{24}$$

without having to pay too much attention to the complicated geometry of the set $T_{\rho,s}$. Likewise, said set is strictly contained in the unit sphere $S^{n-1} \in \mathbb{R}^n$. For fixed $\xi > 0$, this allows us to bound $Q_{2\xi}(T_{\rho,s}, \mathbf{a})$ from below by establishing a global lower bound on $\Pr\left[|\langle \mathbf{a}, \mathbf{u} \rangle| \geq 2\xi \right]$ that is valid for any $\mathbf{u} \in S^{n-1}$. We do this in the appendix and obtain

$$\Pr\left[|\langle \mathbf{a}, \mathbf{z} \rangle| \geq \theta \sqrt{p(1-p)} \right] \geq \frac{4}{13} p(1-p)(1-\theta^2)^2 \quad \forall \mathbf{z} \in S^{n-1} \text{ and } \theta \in [0,1].$$

The structure of such a global bound suggests choosing $\xi = \frac{1}{4}\sqrt{p(1-p)}$ for which we can conclude

$$Q_{2\xi}(T_{\rho,s}, \mathbf{a}) \geq \frac{4p(1-p)(\frac{3}{4})^2}{13} > \frac{p(1-p)}{6}. \tag{25}$$

Such a choice of ξ, setting $t = \frac{p(1-p)}{12}\sqrt{m}$ and envoking the bounds (24) and (25) into (22) implies

$$\begin{aligned}
\inf_{\mathbf{v} \in T_{\rho,s}} \|\mathbf{A}\mathbf{v}\|_{\ell_2} \geq & \frac{\sqrt{p(1-p)}^3}{24} \sqrt{m} - \frac{\sqrt{p(1-p)}^3}{48} \sqrt{m} - 2 \frac{1+\rho}{\rho} \sqrt{4sp(1-p)\left(3\log(2n) + \frac{p}{1-p} \right)} \\
= & \sqrt{p(1-p)} \left(\frac{p(1-p)}{48} \sqrt{m} - \sqrt{16 \frac{(1+\rho)^2}{\rho^2} s \left(3\log(2n) + \frac{p}{1-p} \right)} \right)
\end{aligned}$$

with probability at least $1 - \mathrm{e}^{-\frac{p^2(1-p)^2}{72} m}$. This prompts us to demand

$$m \geq \frac{C_1 (1+\rho)^2}{p^2(1-p)^2 \rho^2} s \left(\log(n) + \frac{p}{1-p} \right), \tag{26}$$

where C_2 is a sufficiently large constant (note that this justifies the assumption $m \geq \frac{\log(n)}{p^2(1-p)^2}$ made before). Then the above inequality implies that there is another constant $C_2 > 0$ (whose size depends on the choice of C) such that

$$\inf_{\mathbf{v} \in T_{\rho,s}} \|\mathbf{A}\mathbf{v}\|_{\ell_2} \geq \frac{\sqrt{p(1-p)}^3}{C_2} \sqrt{m}. \tag{27}$$

with probability of failure bounded by $\mathrm{e}^{-\frac{p^2(1-p)^2}{72}m}$. Comparing this bound to (21) allows us to set $\tau = \frac{C_2}{\sqrt{p(1-p)}^3\sqrt{m}}$ and we arrive at the main result of this section:

Theorem 10. *Let* $\mathbf{A}: \mathbb{R}^n \to \mathbb{R}^m$ *be a* $0/1$*-Bernoulli matrix with parameter* $p \in [0,1]$ *and fix* $s \leq n$ *and* $\rho \in [0,1]$. *Then, there are constants* C_1 *and* C_2 *such that choosing the number of rows to be*

$$m = C_1 \frac{(1+\rho)^2}{p^2(1-p)^2\rho^2} s \left(\log(n) + \frac{p}{1-p} \right) \tag{28}$$

assures that $\mathbf{A}$ *obeys the robust NSP of order* s *with parameters* ρ *and* $\tau = \frac{C_2}{\sqrt{p(1-p)}^3\sqrt{m}}$. *Hereby, the probability of failure is bounded by* $\mathrm{e}^{-\frac{p^2(1-p)^2}{72}m}$.

This is a more detailed version of the first claim presented in Theorem 2. We see that sampling rate, size of the NSP-parameter τ and the probability bound all depend on the Bernoulli parameter $p \in [0,1]$. Factoring out the p-dependence of m by writing $m = \frac{\tilde{m}}{p^2(1-p)^2}$ we obtain a probability bound of $\mathrm{e}^{-\frac{\tilde{m}}{72}}$ which is independent of p. On the other hand $\tau = \frac{C_2}{\sqrt{p(1-p)\tilde{m}}}$ still exhibits a p-dependence.

Finally, we point out that when opting for a standard Bernoulli process, i.e. $p = \frac{1}{2}$, the assertions of Theorem 10 considerably simplify, because $p(1-p) = \frac{1}{4}$. Inserting this, we obtain:

Corollary 11. *Fix* $s \leq n$, $\rho \in [0,1]$ *and let* $\mathbf{A}$ *be a standard* $(m \times n)$ $0/1$*-Bernoulli matrix (i.e.* $p = \frac{1}{2}$*) with*

$$m \geq 17 C_1 \frac{(1+\rho)^2}{\rho^2} s \log(n).$$

Then with probability at least $1 - \mathrm{e}^{-\frac{m}{1152}}$ *this matrix obeys the NSP of order* s *with parameters* ρ *and* $\tau = \frac{C_2}{2\sqrt{m}}$. *Herce,* C_1 *and* C_2 *are the constants from Theorem 10.*

C. 0/1-Bernoulli matrices lie in $\mathcal{M}_+$

We now move on to showing that $0/1$-Bernoulli matrices are very likely to admit the second requirement of Theorem 4. Namely, that there exists a vector $\mathbf{w} = \mathbf{A}^T\mathbf{t}$ that is strictly positive which is equivalent to demanding $\mathbf{A} \in \mathcal{M}_+$. Concretely, we show that setting $\mathbf{t} = \frac{1}{pm}\mathbf{1} \in \mathbb{R}^m$ results in a strictly positive vector $\mathbf{w} \in \mathbb{R}^n$ whose conditioning obeys

$$\kappa(\mathbf{w}) = \frac{\max_k |\langle \mathbf{e}_k, \mathbf{w} \rangle|}{\min_k |\langle \mathbf{e}_k, \mathbf{w} \rangle|} \leq 3. \tag{29}$$

To do so, we note that $\mathbf{w} = \frac{1}{pm}\sum_{k=1}^m \mathbf{a}_k$ has expectation $\mathbb{E}[\mathbf{w}] = \mathbf{1}$, which is – up to re-scaling – the unique non-negative vector admitting $\kappa(\mathbf{1}) = 1$. After having realized this, it suffices to use a concentration inequality to prove that w.h.p. $\mathbf{w}$ does not deviate too much from its expectation $\mathbf{1}$. We do this by invoking a Bernstein inequality which implies:

Theorem 12. *Suppose that* $\mathbf{A}: \mathbb{R}^n \to \mathbb{R}^m$ *is a* $0/1$*-Bernoulli matrix with parameter* $p \in [0,1]$ *and set*

$$\mathbf{w} = \mathbf{A}^T\mathbf{t} \in \mathbb{R}^n \quad \textit{with} \quad \mathbf{t} = \frac{1}{pm}\mathbf{1} \in \mathbb{R}^m. \tag{30}$$

Then with probability at least $1 - n\mathrm{e}^{-\frac{3}{8}p(1-p)m}$ $\max_i |\langle \mathbf{e}_i, \mathbf{w}\rangle| \leq \frac{3}{2}$ *and* $\min_i |\langle \mathbf{e}_i, \mathbf{w}\rangle| \leq \frac{1}{2}$ *which in turn implies* (29).

Proof: Instead of showing the claim directly, we prove that a stronger statement, namely

$$|\langle \mathbf{e}_i, \mathbf{w}\rangle - 1| \leq \frac{1}{2} \quad 1 \leq i \leq n, \tag{31}$$

is true with probability of failure bounded by $n\mathrm{e}^{-\frac{3}{8}p(1-p)m}$. If such a bound is true for all i, it is also valid for maximal and minimal components and we obtain

$$\max_i |\langle \mathbf{e}_i, \mathbf{w}\rangle| \leq \max_k |\langle \mathbf{e}_i, \mathbf{w}\rangle - 1| + 1 \leq \frac{3}{2} \quad \text{and} \quad \min_k |\langle \mathbf{e}_i, \mathbf{w}\rangle| \geq 1 - \min_i |\langle \mathbf{e}_i, \mathbf{w}\rangle - 1| \geq \frac{1}{2},$$

as claimed. In order to prove (31), we fix $1 \leq i \leq n$ and focus on

$$|\langle \mathbf{e}_i, \mathbf{w}\rangle - 1| = \left| \frac{1}{pm} \sum_{k=1}^{m} \langle \mathbf{e}_i, \mathbf{a}_k\rangle - 1 \right| = \frac{1}{pm} \left| \sum_{k=1}^{m} \left(b_{k,i} - \mathbb{E}\left[b_{k,i}\right]\right) \right|.$$

Here, we have used $\langle \mathbf{e}_i, \mathbf{a}_k\rangle = \langle \mathbf{e}_k, \mathbf{A}\mathbf{e}_i\rangle = b_{k,i}$, which is an indepenent instance of a Bernoulli random variable with parameter p. Thus we are faced with bounding the deviation of a sum of m centered, independent random variables $c_k := b_{k,i} - \mathbb{E}\left[b_{k,i}\right]$ from its mean. Each such variable obeys

$$|c_k| \leq \max\left\{p, 1-p\right\} \leq 1 \quad \text{and} \quad \mathbb{E}\left[c_k^2\right] = \mathrm{Var}(b_{k,i}) = p(1-p).$$

Applying a Bernstein inequality [12, Theorem 7.30] reveals

$$\Pr\left[|\langle \mathbf{e}_i, \mathbf{w}\rangle - 1| \geq \frac{1}{2}\right] \leq \Pr\left[|\langle \mathbf{e}_i, \mathbf{w}\rangle - 1| \geq \frac{1-p}{2}\right] = \Pr\left[\left|\sum_{k=1}^{m} c_k\right| \geq \frac{mp(1-p)}{2}\right] \leq \exp\left(-\frac{3}{8}p(1-p)m\right).$$

Combining this with a union bound assures that $|\langle \mathbf{e}_i, \mathbf{w}\rangle - 1| < \frac{1}{2}$ is simultaneously true for all $1 \leq i \leq n$ with probability at least $1 - n\mathrm{e}^{-\frac{3}{8}p(1-p)m}$. ■

D. Proof of Theorem 2

Finally, these two results can be combined to yield Theorem 2. It readily follows from taking a union bound over the individual probabilities of failure. Theorem 10 requires a sampling rate of

$$m \geq C_1 \frac{(1+\rho)^2}{p^2(1-p)^2\rho^2} s \left(\log(n) + \frac{p}{1-p}\right) \tag{32}$$

to assure that a corresponding 0/1-Bernoulli matrix obeys a strong version of the NSP with probability at least $1 - \mathrm{e}^{-\frac{p^2(1-p)^2}{72}m}$. On the other hand, Theorem 12 asserts that choosing $\mathbf{w} = \mathbf{A}^T \frac{1}{pm}\mathbf{1}$ for 0/1-Bernoulli matrices $\mathbf{A}$ results in a well-conditioned and strictly positive vector $\mathbf{w}$ with probability at least $1 - n\mathrm{e}^{-\frac{3}{8}p(1-p)m}$. The probability that either of these assertions fails to hold can be controlled by the union bound over both probabilities of failure:

$$\begin{aligned}\Pr\left[\text{Thm. 10 fails to hold} \cup \text{Thm. 12 fails to hold}\right] \leq & \Pr\left[\text{Thm. 10 fails to hold}\right] + \Pr\left[\text{Thm. 12 fails to hold}\right] \\ \leq & \mathrm{e}^{-\frac{p^2(1-p)^2}{72}m} + n\mathrm{e}^{-\frac{3p(1-p)}{8}m} \leq (n+1)\mathrm{e}^{-\frac{p^2(1-p)^2}{72}m}.\end{aligned}$$

Finally, we focus on 0/1-Bernoulli matrices $\mathbf{A}$ for which both statements are true and whose sampling rate exceeds (32). Theorem 10 then implies that $\mathbf{A}$ obeys the s-NSP with a pre-selected parameter $\rho \in [0,1]$ and

$\tau = \frac{C_2}{\sqrt{p(1-p)}^3\sqrt{m}}$. Moreover, the vector selection $\mathbf{t} = \frac{1}{pm}\mathbf{1}$ in Theorem 12 obeys $\|\mathbf{t}\|_{\ell_2} = \frac{1}{p\sqrt{m}}$. As a result, the implication of Theorem 4 reads for any $\mathbf{x}, \mathbf{z} \geq \mathbf{0}$:

$$\begin{aligned}\|\mathbf{x}-\mathbf{z}\|_{\ell_2} \leq & \frac{2C}{\sqrt{s}}\sigma_s(\mathbf{x}) + D\left(\|\mathbf{t}\|_{\ell_2} + \tau\right)\|\mathbf{A}(\mathbf{x}-\mathbf{z})\|_{\ell_2} \\ = & \frac{2C}{\sqrt{s}}\sigma_s(\mathbf{x}) + D\left(\frac{1}{p\sqrt{m}} + \frac{C_2}{\sqrt{p(1-p)}^3\sqrt{m}}\right)\|\mathbf{A}(\mathbf{x}-\mathbf{z})\|_{\ell_2} \\ \leq & \frac{2C}{\sqrt{s}}\sigma_s(\mathbf{x}) + \frac{D(1+C_2)}{\sqrt{p(1-p)}^3}\frac{\|\mathbf{A}(\mathbf{x}-\mathbf{z})\|_{\ell_2}}{\sqrt{m}}.\end{aligned}$$

The constant $\frac{D(1+C_2)}{\sqrt{p(1-p)}^3}$ is the explicit value of D' in Theorem 1 for the case of $0/1$-Bernoulli matrices with parameter $p \in [0,1]$.

V. Numerical Experiments

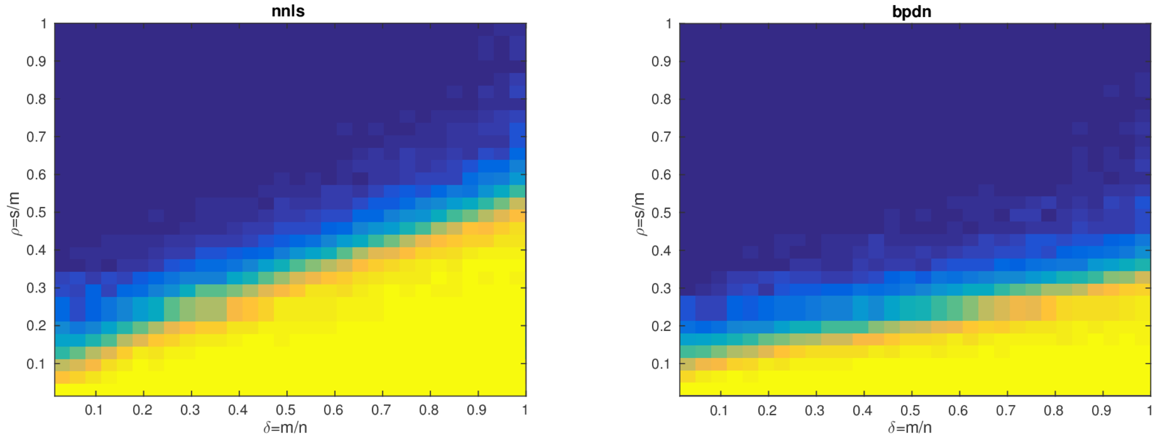
In the following we evaluate the *nonnegative least squares* (NNLS) in (3) and we compare this to the results obtained with *basis pursuit denoising* (BPDN) in (10). The NNLS has been computed using the `lsqnonneg` function in MATLAB which implements the "active-set" Lawson–Hanson algorithm [30]. For the BPDN the SPGL1 toolbox has been used [31].

In a first test we have evaluated numerically the phase transition of NNLS in the $0/1$–Bernoulli setting for the noiseless case. The dimension and sparsity parameters are generated uniformely (in this order) in the ranges $n \in [10\ldots500]$, $m \in [10\ldots n]$ and $s \in [1\ldots m]$. Thus, the sparsity/density variable is $\rho = s/m$ and the subsampling ratio is $\delta = m/n$. The $m \times n$ measurement matrix $\mathbf{A}$ is generated using the iid. $0/1$–Bernoulli model with $p = 1/2$. The nonnegative s–sparse signal $0 \leq \mathbf{x} \in \mathbb{R}^n$ to recover is created as follows: the random support $\text{supp}(\mathbf{x})$ is obtained from taking the first s elements of a random (uniformely–distributed) permutation of the indices $(1\ldots n)$. On this support each value is the absolute value of an iid. standard (zero mean, unit variance) Gaussian, i.e., $x_i = |g_i|$ with $g_i \sim N(0,1)$ for all $i \in \text{supp}(\mathbf{x})$. An event counts as successful once $\|\mathbf{x}-\hat{\mathbf{x}}\|_{\ell_2} \leq 10^{-3}\|\mathbf{x}\|_{\ell_2}$. The resulting phase transition diagram, shown in Figure 1 above, demonstrates that NNLS indeed reliable recovers nonnegative sparse vectors without any ℓ_1–regularization.

In the second experiment we consider the noisy case. Beside its simplicity, the important feature of NNLS is that no a-priori norm assumptions on the noise are necessary as it is required for the BPDN. As illustrated in (4), a result of Theorem 1 is that the NNLS estimate $\mathbf{x}^\sharp$ fullfils:

$$\|\mathbf{x}-\mathbf{x}^\sharp\|_{\ell_2} \leq \frac{2C}{\sqrt{m}}\|\mathbf{e}\|_{\ell_2} \tag{33}$$

A similar bound is valid for the BPDN (see (13)) estimate $\mathbf{x}_\eta$ when $\|\mathbf{e}\|_{\ell_2} \leq \eta$, i.e., once $\|\mathbf{e}\|_{\ell_2}$ is known. Interestingly, even under this prerequisites the performance of NNLS is considerable better then BPDN in our setting. This is visualized in Figure 2 where each component e_j of $\mathbf{e}$ is iid. Gaussian distributed with zero mean and variance $\sigma_e^2 = 1/100$. There recovery has been identified as "successful" if (33) is fulfilled for $2C = \sqrt{10}$.



(a)

(b)

Fig. 2: Comparison of NNLS in (3) with BPDN in (10) for iid. $0/1$–Bernoulli matrices in the noisy setting.

VI. Conclusions

In this work we have shown that nonnegativity is a tremendeous important additional property when recoving sparse vectors. This situation is relevant in many applications and we are motivated here by activity detection in wireless networks using individual sequences. Designing measurement matrices such that convex hull of its columns (the sequences) is sufficiently well-separated from the origin recovery allow for remarkable simple recovery algorithms which are prone to noise and blind in a sense that no regularization and a priori information on the noise is required. We have demonstrated this feature by strengthen the implications of the robust nullspace property for the nonnegative setting. Furthermore, we have shown that iid. binary measurements fullfill w.h.p. this property and are simultaneously well-conditioned and can be used therefore for recovering nonnegative and sparse vectors in the optimal regime.

Acknowledgements

The authors want to thank S. Dirksen, H. Rauhut and D. Gross for inspiring discussions and helpful comments. This work has been been supported by “Mathematics of Signal Processing” trimester program at the Hausdorff Research Institute for Mathematics (HIM). RK acknowledges support from the Excellence Initiative of the German Federal and State Governments (Grants ZUK 43 & 81), the ARO under contract W911NF-14-1-0098 (Quantum Characterization, Verification, and Validation), and the DFG. PJ is supported by DFG grant JU 2795/2.

Appendix: Proofs of Equations (23) and (25)

Here we provide proofs of the two bounds (23) and (25) on which we built our argument that $0/1$-Bernoulli matrices obey the robust NSP. Since both are rather technical and not essential for understanding the main ideas, we decided to present them in this appendix.

Preliminaries

In order to prove the remaining estimates we rely on a couple of probabilistic standard tools which we shall summarize here. Recall that a Rademacher sequence $(\epsilon_1, \ldots, \epsilon_m)$ is a sequence of m independent dichotomic random variables obeying $\Pr[\epsilon_k = 1] = \Pr[\epsilon_k = -1] = \frac{1}{2}$.

Theorem 13 (Khintchine Inequality, Corollary 8.7 in [12])**.** *Let* $\mathbf{c} \in \mathbb{C}^m$ *and* $\epsilon_1, \ldots, \epsilon_m$ *be a Rademacher sequence. Then for all* $q > 0$

$$\left(\mathbb{E}\left[\left| \sum_{k=1}^m \epsilon_k c_k \right|^q \right] \right)^{1/q} \leq 2^{3/(4q)} \mathrm{e}^{-1/2} \|\mathbf{c}\|_{\ell_2}.$$

Theorem 14 (Non-commutative Khintchine inequality: Exercise 8.6 (d) in [29])**.** *Let* $M_1, \ldots, M_m$ *be hermitian* $n \times n$ *matrices and suppose that* $(\epsilon_1, \ldots, \epsilon_m)$ *is a Rademacher sequence. Then*

$$\mathbb{E}\left[\left\| \sum_{k=1}^m \epsilon_k M_k \right\| \right] \leq \sqrt{2\log(2n)} \left\| \sum_{k=1}^m M_k^2 \right\|^{1/2}.$$

Theorem 15 (Matrix Chernoff for expectation values: Theorem 5.1.1 in [32] (see also [33])**.** *Let* $\{M_k\}_{k=1}^m$ *be a sequence of independent, random, non-negative* $n \times n$ *matrices obeying* $\|M_k\| \leq R$ *almost surely. Then, for any* $t > 0$ *their sum obeys*

$$\mathbb{E}\left[\left\| \sum_{k=1}^m M_k \right\| \right] \leq \frac{\mathrm{e}^t - 1}{t} \left\| \sum_{k=1}^m \mathbb{E}[M_k] \right\|_\infty + \frac{R}{t} \log(n).$$

Theorem 16 (Paley-Zygmund Inequality)**.** *Let* X *be a non-negative random variable with bounded second moment. Then*

$$\Pr[X \geq \theta \mathbb{E}[X]] \geq \frac{(1-\theta)^2 \mathbb{E}[X]}{\mathrm{Var}(X) + \mathbb{E}[X]^2},$$

where $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ *is the variance of* X*.*

Bounding $\mathbb{E}[\|\mathbf{h}\|_{\ell_\infty}]$ *for* $0/1$*-Bernoulli matrices*

In this section, we prove that the bound presented in (23) holds in the Bernoulli setting. Let $\mathbf{A}$ be a $0/1$-Bernoulli matrix with parameter p and m rows $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$. The vector $\mathbf{h} := \frac{1}{\sqrt{m}} \sum_{k=1}^m \epsilon_k \mathbf{a}_k$ was introduced in Theorem 9 and in (23) we claimed that this vector obeys

$$\mathbb{E}[\|\mathbf{h}\|_{\ell_\infty}] \leq \sqrt{4p(1-p)\left(3\log(2n) + \frac{p}{1-p}\right)}, \tag{34}$$

provided that $m \geq \frac{\log(n)}{p^2(1-p)^2}$. When aiming to prove this, we first minimize the anisotropic impact of $\mathbf{A}$'s rows. Recalling $\mathbb{E}[\mathbf{a}_k] = p\mathbf{1}$, we introduce $\tilde{\mathbf{a}}_k := \mathbf{a}_k - p\mathbf{1}$, and likewise $\tilde{\mathbf{h}} := \frac{1}{\sqrt{m}} \sum_{k=1}^m \epsilon_k \tilde{\mathbf{a}}_k$, which obey

$$\mathbf{h} = \tilde{\mathbf{h}} + \frac{p}{\sqrt{m}} \left(\sum_{k=1}^m \epsilon_k \right) \mathbf{1} \tag{35}$$

by construction. Applying the triangle inequality reveals

$$\mathbb{E}[\|\mathbf{h}\|_{\ell_\infty}] \leq \mathbb{E}\left[\left\| \tilde{\mathbf{h}} \right\|_{\ell_\infty} \right] + \frac{p}{\sqrt{m}} \mathbb{E}\left[\left| \sum_{k=1}^m \epsilon_k \right| \right] \|\mathbf{1}\|_{\ell_\infty} \tag{36}$$

and we may bound the two terms individually. For the second term, we resort to the classical Khintchine inequality (with $q = 1$ and $\mathbf{c} = \mathbf{1}$) and obtain

$$\frac{p}{\sqrt{m}}\left(\sum_{k=1}^m \epsilon_k\right)\mathbf{1} \leq \frac{p2^{3/4}\mathrm{e}^{-1/2}}{\sqrt{m}}\|\mathbf{1}\|_{\ell_2}\|\mathbf{1}\|_{\ell_\infty} \leq \sqrt{2}p, \tag{37}$$

because $\|\mathbf{1}\|_{\ell_2} = \sqrt{m}\|\mathbf{1}\|_{\ell_\infty} = \sqrt{m}$ and $2^{3/4}\mathrm{e}^{-1/2} \simeq 1.02.006 < \sqrt{2}$. For the remaining estimate of $\mathbb{E}\left[\left\|\tilde{\mathbf{h}}\right\|_{\ell_\infty}\right]$, we find it advantageous to work with an equivalent matrix problem

$$\mathbb{E}\left[\left\|\tilde{\mathbf{h}}\right\|_{\ell_\infty}\right] = \mathbb{E}\left[\left\|\mathrm{diag}\left(\tilde{\mathbf{h}}\right)\right\|_{\infty}\right] = \frac{1}{\sqrt{m}}\mathbb{E}\left[\left\|\sum_{k=1}^m \epsilon_k \mathrm{diag}\left(\tilde{\mathbf{a}}_k\right)\right\|_{\infty}\right]$$

that can be tackled by consecutively applying matrix Khintchine and Chernoff inequalities. Exploiting the randomness in $(\epsilon_1,\ldots,\epsilon_m)$, by applying Theorem 14 assures

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\mathbf{h}}\right\|_{\ell_\infty}\right] =& \frac{1}{\sqrt{m}}\mathbb{E}_{\mathbf{a}}\mathbb{E}_\epsilon\left[\left\|\sum_{k=1}^m \epsilon_k \mathrm{diag}\left(\tilde{\mathbf{a}}\right)_k\right\|_{\infty}\right] \leq \sqrt{\frac{2\log(2n)}{m}}\mathbb{E}_{\mathbf{a}}\left[\left\|\sum_{k=1}^m \mathrm{diag}\left(\tilde{\mathbf{a}}_k\right)^2\right\|_{\infty}^{1/2}\right] \\ \leq& \sqrt{\frac{2\log(2n)}{m}}\left(\mathbb{E}_{\mathbf{a}}\left[\left\|\sum_{k=1}^m \mathrm{diag}\left(\tilde{\mathbf{a}}_k\right)^2\right\|_{\infty}\right]\right)^{1/2},\end{aligned} \tag{38}$$

where we have also employed Jensen's inequality. Now, note thate the matrices $\mathrm{diag}\left(\tilde{\mathbf{a}_k}\right)^2$ are all positive semidefinite and obey

$$\begin{aligned}\left\|\mathrm{diag}\left(\tilde{\mathbf{a}_k}\right)^2\right\| =& \left\|\mathrm{diag}\left(\mathbf{a}_k - p\mathbf{1}\right)^2\right\| \leq \max\left\{p^2,(1-p)^2\right\}, \\ \mathbb{E}\left[\mathrm{diag}\left(\tilde{\mathbf{a}_k}\right)^2\right] =& \sum_{i=1}^n \mathbb{E}\left[\left(\langle \mathbf{e}_i,\mathbf{a}_k\rangle - p\right)^2\right]\mathbf{e}_i\mathbf{e}_i^T = p(1-p)\mathbb{I}.\end{aligned}$$

This is true, because each $\langle \mathbf{e}_i,\mathbf{a}_k\rangle$ is an independent instance of a Bernoulli variable with parameter p. Thus, Theorem 15 is applicable and setting $t = 1$ implies for

$$\begin{aligned}\mathbb{E}_{\mathbf{a}}\left[\left\|\sum_{k=1}^m \mathrm{diag}\left(\tilde{\mathbf{a}}_k\right)^2\right\|_{\infty}\right] \leq& (\mathrm{e}-1)\left\|\sum_{k=1}^m p(1-p)\mathbb{I}\right\|_{\infty} + \max\left\{p^2,(1-p)^2\right\}\log(n) \\ \leq& \mathrm{e}p(1-p)m + \max\left\{p^2,(1-p)^2\right\}\log(n).\end{aligned}$$

Inserting this into (38) yields

$$\mathbb{E}\left[\left\|\tilde{\mathbf{h}}\right\|_{\ell_\infty}\right] \leq \sqrt{2\log(2n)\left(\mathrm{e}p(1-p) + \frac{\log(n)}{m}\right)} \tag{39}$$

and turning back to (36), we see that

$$\mathbb{E}\left[\|\mathbf{h}\|_{\ell_\infty}\right] \leq \sqrt{2\log(2n)\left(\mathrm{e}p(1-p) + \frac{\log(n)}{m}\right)} + \sqrt{2}p$$

holds. In order to simplify this further, we now use the prior assumption $m \geq \frac{\log(n)}{p^2(1-p)^2}$ which assures

$$\frac{\log(n)}{m} \leq p^2(1-p)^2 \leq \frac{1}{4}p(1-p),$$

because $p(1-p) \leq \frac{1}{4}$ for all $p \in [0,1]$. Combining this with $\mathrm{e} + \frac{1}{4} < 3$ allows us to deduce

$$\mathbb{E}\left[\|\mathbf{h}\|_{\ell_\infty}\right] \leq \sqrt{6p(1-p)\log(2n)} + \sqrt{2}p.$$

Finally, we use the elementary inequality $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ $\forall a, b \geq 0$ to obtain

$$\mathbb{E}\left[\|\mathbf{h}\|_{\ell_\infty}\right] \leq \sqrt{2\left(6p(1-p)\log(2n) + 2p^2\right)} = \sqrt{4p(1-p)\left(3\log(2n) + \frac{p}{1-p}\right)},$$

which is the estimate presented in (34).

A. *Bounding* $\Pr\left[|\langle \mathbf{a}, \mathbf{z}\rangle| \geq \theta \|\mathbf{z}\|_{\ell_2}\right]$ *for* $0/1$-*Bernoulli vectors*

In this final section we prove that for any unit vector $\mathbf{z} = (z_1, \ldots, z_n)^T \in S^{n-1}$ and any $\theta \in [0, 1/2]$, the bound

$$\Pr\left[|\langle \mathbf{a}, \mathbf{z}\rangle| \geq \theta\sqrt{p(1-p)}\right] \geq \frac{4}{13} p(1-p)(1-\theta^2)^2 \tag{40}$$

holds in the Bernoulli setting. Here, the probability is taken over instances $\mathbf{a} \in \mathbb{R}^n$ of the i.i.d. row distribution in a $0/1$-Bernoulli matrix. Hence, $\mathbf{a} = \sum_{i=1}^n b_i \mathbf{e}_i$, where each b_i is an independent Bernoulli random variable with parameter p. This estimate is going to rely on the Paley-Zygmund inequality and a few standard, but rather tedious, moment calculations for Bernoulli processes. We start by exploiting

$$\Pr\left[|\langle \mathbf{a}, \mathbf{z}\rangle| \geq \theta\sqrt{p(1-p)}\right] = \Pr\left[\langle \mathbf{a}, \mathbf{z}\rangle^2 \geq \theta^2 p(1-p)\right], \tag{41}$$

because the latter expression is easier to handle. Introducing the nonnegative random variable $S := \langle \mathbf{a}, \mathbf{z}\rangle^2 = \sum_{i,j=1}^n b_i b_j z_i z_j$,, we see

$$\mathbb{E}[S] = \sum_{i \neq j} \mathbb{E}[b_i]\,\mathbb{E}[b_j]\, z_i z_j + \sum_{i=1}^n \mathbb{E}\left[b_i^2\right] z_i^2 = p^2 \langle \mathbf{1}, \mathbf{z}\rangle + p(1-p)\|\mathbf{z}\|_{\ell_2}^2 \geq p(1-p) \tag{42}$$

(recall that each b_i is an independent Bernoulli variable with parameter p). This calculation together with (41) implies

$$\Pr\left[|\langle \mathbf{a}, \mathbf{z}\rangle| \geq \theta\sqrt{p(1-p)}\right] \geq \Pr\left[S \geq \theta^2 \mathbb{E}[S]\right]. \tag{43}$$

Since $S \geq 0$ by definition, the requirements for Paley-Zygmund – Theorem 16 – are met and said Theorem implies

$$\Pr\left[S \geq \theta^2 \mathbb{E}[S]\right]^2 \geq \frac{(1-\theta^2)^2 \mathbb{E}[S]}{\mathrm{Var}(S) + \mathbb{E}[S]^2}. \tag{44}$$

We have already computed $\mathbb{E}[S]$ in (42), but we still have to compute its variance. We defer this calculation to the very end of this section and for now simply state its result:

$$\mathrm{Var}(S) = 2\mathbb{E}[S]^2 - 2p^4\langle \mathbf{1}, \mathbf{z}\rangle + 4p^2(1-p)(1-2p)\langle \mathbf{1}, \mathbf{z}\rangle \sum_{i=1}^n z_i^3 + p(1-p)(1-6p(1-p))\|\mathbf{z}\|_{\ell_4}^4. \tag{45}$$

We now move on to bound these contributions individually by a multiple of $\mathbb{E}[S]^2$. We can omit the second term and obtain

$$\begin{aligned}4p^2(1-p)(1-2p)\langle \mathbf{1}, \mathbf{z}\rangle \sum_{i=1}^n z_i^3 \leq & 4p^2(1-p)^2\langle \mathbf{1}, \mathbf{z}\rangle \|\mathbf{z}\|_{\ell_2}^3 = 4p^2(1-p)^2\langle \mathbf{1}, \mathbf{z}\rangle \leq 4p^2(1-p)^2 \max\left\{\langle \mathbf{1}, \mathbf{z}\rangle^2, 1\right\} \\ \leq & \frac{2}{p}\left(p^2\langle \mathbf{1}, \mathbf{z}\rangle^2 + p(1-p)\right)^2 = \frac{2}{p}\mathbb{E}[S]^2\end{aligned}$$

for the third term. The fourth term can be bounded cia

$$p(1-p)(1-6p(1-p))\|\mathbf{z}\|_{\ell_4}^4 \leq p(1-p)\|\mathbf{z}\|_{\ell_2}^4 \leq \frac{1}{p(1-p)}\mathbb{E}\left[S\right]^2.$$

and combining all these bounds implies

$$\mathrm{Var}(S) \leq \left(2+\frac{2}{p}+\frac{1}{p(1-p)}\right)\mathbb{E}\left[S\right]^2 = \frac{3-2p^2}{p(1-p)}\mathbb{E}\left[S\right]^2 \leq \frac{3}{p(1-p)}\mathbb{E}\left[S\right]^2.$$

Inserting this upper bound into the Paley-Zygmund estimate (44) yields

$$\Pr\left[|\langle \mathbf{a},\mathbf{z}\rangle| \geq \theta\sqrt{p(1-p)}\right] \geq \frac{(1-\theta^2)^2\mathbb{E}\left[S\right]^2}{\mathrm{Var}(S)+\mathbb{E}\left[S\right]^2} \geq \frac{(1-\theta^2)^2\mathbb{E}\left[S\right]^2}{(\frac{3}{p(1-p)}+1)\mathbb{E}\left[S\right]^2} \geq \frac{4}{13}p(1-p)(1-\theta^2)^2,$$

as claimed in (25) and (40), respectively. In the last line, we have used $p(1-p) \leq \frac{1}{4}$ for any $p \in [0,1]$.

Finally, we provide the derivation of Equation (45). We use our knowledge of $\mathbb{E}[S] = p^2\langle \mathbf{1},\mathbf{z}\rangle + p(1-p)\|\mathbf{z}\|_{\ell_2}^2$ together with the elementary formula

$$(b_i-p)(b_j-p) = (b_ib_j-p^2) - pb_i - pb_j + 2p^2$$

to rewrite $S-\mathbb{E}[S]$ as

$$\begin{aligned}
S-\mathbb{E}\left[S\right] =& \sum_{i,j=1}^n b_ib_jz_iz_j - p^2\sum_{i\neq j} z_iz_j - p\sum_{i=1}^n z_i^2 = \sum_{i\neq j}\left(b_ib_j-p^2\right)z_iz_j + \sum_{i=1}^n\left(b_i^2-p\right)z_i^2 \\
=& \sum_{i\neq j}\left((b_i-p)(b_j-p)+pb_i+pb_j-2p^2\right)z_iz_j + \sum_{i=1}^n\left(b_i^2-p\right)z_i^2 \\
=& \sum_{i\neq j}\left(b_i-p\right)\left(b_j-p\right)z_iz_j + \sum_{i=1}^n\left(b_i^2-p\right)z_i^2 + p\sum_{i\neq j}b_iz_iz_j + p\sum_{j\neq i}b_jz_jz_i - 2p^2\sum_{i\neq j}z_iz_j \\
=& \sum_{i\neq j}\left(b_i-p\right)\left(b_j-p\right)z_iz_j + \sum_{i=1}^n\left(b_i^2-p\right)z_i^2 + 2p\sum_{i,j=1}^n b_iz_iz_j - 2p\sum_{i=1}^n b_iz_i^2 - 2p^2\sum_{i,j=1}^n z_iz_j + 2p^2\sum_{i=1}^n z_i^2 \\
=& \sum_{i\neq j}\left(b_i-p\right)\left(b_j-p\right)z_iz_j + \sum_{i=1}^n\left(b_i^2-p\right)z_i^2 + 2p\sum_{i,j=1}^n\left(b_i-p\right)z_iz_j - 2p\sum_{i=1}^n\left(b_i-p\right)z_i^2 \\
=& 2\sum_{i<j}\left(b_i-p\right)\left(b_j-p\right)z_iz_j + 2p\langle \mathbf{1},\mathbf{z}\rangle\sum_{i=1}^n\left(b_i-p\right)z_i + (1-2p)\sum_{i=1}^n\left(b_i-p\right)z_i^2.
\end{aligned}$$

Here we have exploited symmetry in the first term and $b_i^2 = b_i$ to further simplify that expression. For notational simplicity, it makes sense to define the random variable $\tilde{b}_i := b_i - p$ which obeys $\mathbb{E}\left[\tilde{b}_i\right] = 0$ and $\mathbb{E}\left[\tilde{b}_i^2\right] = \mathrm{Var}(b_i) = p(1-p)$. Introducing such a notation simplifies the above expression to

$$S-\mathbb{E}\left[S\right] = 2\sum_{i<j}\tilde{b}_i\tilde{b}_jz_iz_j + 2p\langle \mathbf{1},\mathbf{z}\rangle\sum_{i=1}^n\tilde{b}_iz_i + (1-2p)\sum_{i=1}^n\tilde{b}_iz_i^2.$$

Employing the binomial formula $(a+b+c)^2 = a^2 + 2ab + 2ac + b^2 + 2bc + c^2$, we obtain

$$\begin{aligned}\mathrm{Var}(S) =&\mathbb{E}\left[(S-\mathbb{E}\left[S\right])^2\right] = 4\sum_{i<j}\sum_{k<l}\mathbb{E}\left[\tilde{b}_i\tilde{b}_j\tilde{b}_k\tilde{b}_l\right] z_i z_j z_k z_l + 8p\langle\mathbf{1},\mathbf{z}\rangle\sum_{i<j}\sum_{k=1}^{n}\mathbb{E}\left[\tilde{b}_i\tilde{b}_j\tilde{b}_k\right] z_i z_j z_k \\ &+4(1-2p)\sum_{i<j}\sum_{k=1}^{n}\mathbb{E}\left[\tilde{b}_i\tilde{b}_j\tilde{b}_k\right] z_i z_j z_k^2 + 4p^2\langle\mathbf{1},\mathbf{z}\rangle^2\sum_{i,j=1}^{n}\mathbb{E}\left[\tilde{b}_i\tilde{b}_j\right] z_i z_j \\ &+4p(1-2p)\langle\mathbf{1},\mathbf{z}\rangle\sum_{i,j=1}^{n}\mathbb{E}\left[\tilde{b}_i\tilde{b}_j\right] z_i z_j^2 + (1-2p)^2\sum_{i,j=1}^{n}\mathbb{E}\left[\tilde{b}_i\tilde{b}_j\right] z_i^2 z_j^2.\end{aligned}$$

Centeredness of $\tilde{b}$ together with the summation constraints $(i<j)$ and $(k<l)$ implies that summands in the first term vanish, unless $i=k$ and$j=l$. This in turn implies

$$\begin{aligned}4\sum_{i<j}\sum_{k<l}\mathbb{E}\left[\tilde{b}_i\tilde{b}_j\tilde{b}_k\tilde{b}_l\right] z_i z_j z_k z_l =&4\sum_{i<j}\mathbb{E}\left[\tilde{b}_i^2\right]\mathbb{E}\left[\tilde{b}_j^2\right] z_i^2 z_j^2 = 2p^2(1-p)^2\sum_{i\neq j} z_i^2 z_j^2 \\ =&2p^2(1-p)^2\left(\sum_{i,j=1}^{n} z_i^2 z_j^2 - \sum_{i=1}^{n} z_i^4\right) = 2p^2(1-p)^2\left(\|\mathbf{z}\|_{\ell_2}^4 - \|\mathbf{z}\|_{\ell_4}^4\right).\end{aligned}$$

Using a similar argument allows us to conclude that the second and third term must identically vanish (because the index constraints $i<j$ prevents $i=j=k$ and, consequently, at least one index must always remain unpaired). We can exploit $\mathbb{E}\left[\tilde{b}_i\tilde{b}_j\right] = p(1-p)\delta_{i,j}$ in the remaining terms to conclude

$$\begin{aligned}\mathrm{Var}(S) =&2p^2(1-p)^2\left(\|\mathbf{z}\|_{\ell_2}^4 - \|\mathbf{z}\|_{\ell_4}^4\right) + 4p^3(1-p)\langle\mathbf{1},\mathbf{z}\rangle^2\|\mathbf{z}\|_{\ell_2}^2 \\ &+4p^2(1-p)(1-2p)\langle\mathbf{1},\mathbf{z}\rangle\sum_{i=1}^{n} z_i^3 + p(1-p)(1-2p)^2\|\mathbf{z}\|_{\ell_4}^4.\end{aligned}$$

Slightly rewriting this expression then yields the result presented in (45)

References

[1] Y. Vardi, "Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 365–377, 1996.

[2] R. Castro, M. Coates, G. Liang, R. Nowak, and B. Yu, "Network Tomography: Recent Developments," *Statistical Science*, vol. 19, pp. 499–517, 2004.

[3] J. E. Boyd and J. Meloche, "Evaluation of statistical and multiple-hypothesis tracking for video traffic surveillance," *Machine Vision and Applications*, vol. 13, no. 5-6, pp. 344–351, 2003.

[4] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and S. A. S, "Maximum Entropy and the Nearly Black Object," *Journal of the Royal Statistical Society B*, vol. 54, no. 1, 1992.

[5] G. Zhang, S. Jiao, X. Xu, and L. Wang, "Compressed sensing and reconstruction with Bernoulli matrices," in *2010 IEEE International Conference on Information and Automation, ICIA 2010*, 2010, pp. 455–460.

[6] M. A. Khajehnejad, A. G. Dimakis, W. Xu, and B. Hassibi, "Sparse recovery of nonnegative signals with minimal expansion," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 196–208, 2011.

[7] N. Meinshausen, "Sign-constrained least squares estimation for high-dimensional regression," *Electronic Journal of Statistics*, vol. 7, no. 1, pp. 1607–1631, 2013.

[8] M. Slawski and M. Hein, "Sparse recovery by thresholded non-negative least squares," *Electronic Journal of Statistics*, vol. 7, 2013.

[9] S. Foucart and D. Koslicki, "Sparse Recovery by Means of Nonnegative Least Squares," *IEEE Signal Processing Letters*, vol. 21, no. 4, 2014.

[10] D. L. Donoho and J. Tanner, “Sparse nonnegative solution of underdetermined linear equations by linear programming.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 27, pp. 9446–9451, 2005.

[11] A. M. Bruckstein, M. Elad, and M. Zibulevsky, “On the uniqueness of non-negative sparse & redundant representations,” *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, no. 796, pp. 5145–5148, 2008.

[12] S. Foucart and H. Rauhut, “A mathematical introduction to compressive sensing,” *Appl. Numer. Harmon. Anal. Birkhäuser, Boston, in ...*, 2013.

[13] M. Rudelson and S. Zhou, “Reconstruction from anisotropic random measurements,” *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3434–3447, June 2013.

[14] R. Kueng and D. Gross, “{RIPless} compressed sensing from anisotropic measurements,” *Linear Algebra and its Applications*, vol. 441, pp. 110 – 123, 2014, special Issue on Sparse Approximate Solution of Linear Systems.

[15] S. Mendelson, “Learning without concentration,” *J. ACM*, vol. 62, no. 3, pp. 21:1–21:25, Jun. 2015.

[16] V. Koltchinskii and S. Mendelson, “Bounding the smallest singular value of a random matrix without concentration,” *International Mathematics Research Notices*, vol. 2015, no. 23, pp. 12 991–13 008, 2015.

[17] J. A. Tropp, *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*. Cham: Springer International Publishing, 2015, ch. Convex Recovery of a Structured Signal from Independent Random Linear Measurements, pp. 67–101.

[18] E. J. Candes, T. Strohmer, and V. Voroninski, “Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming,” *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.

[19] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, “Stable low-rank matrix recovery via null space properties,” *arXiv preprint arXiv:1507.07184*, 2015.

[20] S. Dirksen, G. Lecué, and H. Rauhut, “On the gap between rip-properties and sparse recovery conditions,” *arXiv preprint arXiv:1504.05073*, 2015.

[21] Y. Chang, P. Jung, C. Zhou, and S. Stanczak, “Block Compressed Sensing based Distributed Resource Allocation for M2M Communications,” *accepted for International Conference on Acoustics, Speech, and Signal Processing, ICASSP16*, 2016.

[22] M. Wang, W. Xu, and A. Tang, “A unique ”nonnegative” solution to an underdetermined system: From vectors to matrices,” *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 1007–1016, 2011.

[23] R. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin, “A Simple Proof of the Restricted Isometry Property for Random Matrices,” *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, jan 2008.

[24] E. J. Candes, “The restricted isometry property and its implications for compressed sensing,” *Compte Rendus de l’Academie des Sciences*, vol. 346, no. Paris, Serie I, pp. 589–592, may 2008.

[25] C. Giraud, S. Huet, and N. Verzelen, “High-Dimensional Regression with Unknown Variance,” *Statistical Science*, vol. 27, pp. 500–518, 2012. [Online]. Available: http://projecteuclid.org/euclid.ss/1356098553

[26] T. Sun and C.-H. Zhang, “Scaled Sparse Linear Regression,” 2011. [Online]. Available: http://arxiv.org/abs/1104.4595

[27] N. Städler, P. Bühlmann, and S. van de Geer, “Rejoinder: l1-penalization for mixture regression models,” *Test*, vol. 19, pp. 209–256, 2010.

[28] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, “Phase retrieval via matrix completion,” *SIAM Review*, vol. 57, no. 2, pp. 225–251, 2015.

[29] R. Kueng, H. Rauhut, and U. Terstiege, “Low rank matrix recovery from rank one measurements,” *Applied and Computational Harmonic Analysis*, pp. –, 2015.

[30] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Prentice-Hall, 1974.

[31] E. van den Berg and M. P. Friedlander, “Probing the pareto frontier for basis pursuit solutions,” *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

[32] J. A. Tropp, “User friendly tools for random matrices. An introduction.” *Preprint*, 2012.

[33] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2011.

Improving compressed sensing with the diamond norm

M. Kliesch[1], R. Kueng[2], J. Eisert[1], and D. Gross[2]

[1]Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, Germany
[2]Institute for Theoretical Physics, University of Cologne, Germany

Abstract

In low-rank matrix recovery, one aims to reconstruct a low-rank matrix from a minimal number of linear measurements. Within the paradigm of compressed sensing, this is made computationally efficient by minimizing the nuclear norm as a convex surrogate for rank. In this work, we identify an improved regularizer based on the so-called diamond norm, a concept imported from quantum information theory. We show that – for a class of matrices saturating a certain norm inequality – the descent cone of the diamond norm is contained in that of the nuclear norm. This suggests superior reconstruction properties for these matrices and we explicitly characterize this set. We demonstrate numerically that the diamond norm indeed outperforms the nuclear norm in a number of relevant applications: These include signal analysis tasks such as blind matrix deconvolution or the retrieval of certain unitary basis changes, as well as the quantum information problem of process tomography with random measurements. The diamond norm is defined for matrices that can be interpreted as order-4 tensors and it turns out that the above condition depends crucially on that tensorial structure. In this sense, this work touches on an aspect of the notoriously difficult tensor completion problem.

1 Introduction

The task of recovering an unknown low-rank matrix from a small number of measurements appears in a variety of contexts. Examples of this task are provided by collaborative filtering in machine learning [1], quantum state tomography in quantum information [2, 3], the estimation of covariance matrices [4, 5], or face recognition [6]. If the measurements are linear, the technical problem reduces to identifying the lowest-rank element in an affine space of matrices. In general, this problem is NP-hard and it is thus unclear how to approach it algorithmically [7].

In the wider field of compressed sensing [8], the strategy for treating such problems is to replace the complexity measure – here the rank – with a tight convex relaxation. Often, it can be rigorously proved that the resulting convex optimization problem has the same solution as the original problem for many relevant problems, while at the same time allowing for an efficient algorithm. The tightest (in some sense [9]) convex relaxation of rank is the *nuclear norm*, i.e. the sum of singular values. Minimizing the nuclear norm subject to linear constraints is a semi-definite program and great number of rigorous performance guarantees have been provided for low-rank reconstruction using nuclear norm minimization [2, 10–17].

The geometry of convex reconstruction schemes is now well-understood (c.f. Figure 2). Starting with a convex regularizer f (e.g. the nuclear norm), geometric proof techniques like Tropp's Bowling scheme [18] or Mendelson's small ball method [19, 20] bound the reconstruction error in terms of the descent cone of f at the matrix that is to be recovered. Moreover, these arguments suggest that the error would decrease if another convex regularizer with smaller descent cone would be used. This motivates the search for new convex regularizers that (i) are efficiently computable and (ii) have a smaller descent cone at particular points of interest.

In this work, we introduce such an improved regularizer based on the *diamond norm* [21]. This norm plays a fundamental role in the context of quantum information and operator theory [22]. For this work, it is convenient to also use a variant of the diamond norm that we call the *square norm*. While not obvious from its definition, it has been found that the diamond norm can be efficiently computed by means of a semidefinite program (SDP) [23–25]. Starting from one such SDP characterization [25], we identify the set of matrices for which the square norm's descent cone is contained in the corresponding one of the nuclear norm. As a result, low-rank matrix recovery guarantees that have been established via analyzing the nuclear norm's descent cone [16, 18] are also valid for square norm regularization, provided that the matrix of interest belongs to said set. What is more, bearing in mind the reduced size of the square norm's descent cone, we actually expect an improved recovery. Indeed, with numerical studies we show an improved performance.

Going beyond low-rank matrix recovery, we identify several applications. In physics, we present numerical experiments that show that the diamond norm offers improved performance for *quantum process tomography* [26]. The goal of this important task is to reconstruct a quantum process from suitable preparations of inputs and measurements on outputs (generalizing quantum *state* tomography, for which low-rank methods have been studied extensively [2, 3, 27, 28]. We then identify applications to problems from the context of signal processing. These include matrix versions of the *phase retrieval problem* [29–36], as well as a matrix version of the *blind deconvolution problem* [15]. Recently, a number of *bi-linear problems* combined with sparsity or low-rank structures have been investigated in the context of compressed sensing, with first progress on recovery guarantees being reported [15, 37]. The present work can be seen as a contribution to this recent development.

We conclude the introduction on a more speculative note. The diamond norm is defined for linear maps taking operators to operators – i.e., for objects that can also be viewed as order-4 tensors. We derive a characterization of those maps for which the diamond norm offers improved recovery, and find that it depends on the order-4 tensorial structure. In this sense, the present work touches on an aspect of the notoriously difficult *tensor recovery problem* (no canonic approach or reference seems to have emerged yet, but see Ref. [38] for an up-to-date list of partial results). In fact, the "tensorial nature" of the diamond norm was the original motivation for the authors to consider it in more detail as a regularizer – even though the eventual concrete applications we found do not seem to have a connection to tensor recovery. It would be interesting to explore this aspect in more detail.

2 Preliminaries

In this section, we introduce notation and mathematical preliminaries used to state our main results. We start by clarifying some notational conventions. In particular, we introduce certain matrix norms and the partial trace for operators acting on a tensor product space. Moreover, we summarize a general geometric setting for the convex recovery of structured signals.

2.1 Vectors and operators

Throughout this work we focus exclusively on finite dimensional mostly complex vector spaces $\mathcal{V}, \mathcal{W}$ whose elements we mostly denote by lower case latin letters, e.g. $x \in \mathcal{V}$. Furthermore we assume that each vector space $\mathcal{V}$ is equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ – or simply $\langle \cdot, \cdot \rangle$ for short – that is linear in the second argument. Such an inner product induces the Euclidean norm

$$\|x\|_{\mathrm{F}} := \sqrt{\langle x, x \rangle_{\mathcal{V}}} \quad \forall x \in \mathcal{V} \tag{1}$$

and moreover defines a conjugate linear bijection from $\mathcal{V}$ to its dual space $\mathcal{V}^*$: to any $x \in \mathcal{V}$ we associate a dual vector $x^\dagger \in \mathcal{V}^*$ which is uniquely defined via $x^\dagger y = \langle x, y \rangle_{\mathcal{V}}$ $\forall y \in \mathcal{V}$. The vector space of linear maps from $\mathcal{V}$ to $\mathcal{W}$ is denoted by $\mathrm{L}(\mathcal{V} \to \mathcal{W})$. Its elements being *operators* are denoted by capital latin letters (e.g. X, Y, U, V) and often we also refer to them as matrices. When dealing with endomorphisms, we write $\mathrm{L}(\mathcal{V}) = \mathrm{L}(\mathcal{V} \to \mathcal{V})$ for the sake of notational brevity. The adjoint $X^\dagger \in \mathrm{L}(\mathcal{W} \to \mathcal{V})$ of an operator $X \in \mathrm{L}(\mathcal{V} \to \mathcal{W})$ is determined by $\langle X^\dagger x, y \rangle_{\mathcal{V}} = \langle x, Xy \rangle_{\mathcal{W}}$ for all $x \in \mathcal{V}$ and $y \in \mathcal{W}$ and we call an operator $X \in \mathrm{L}(V)$ self-adjoint, or Hermitian, if $X^\dagger = X$. A self-adjoint operator X is positive semidefinite, if it has a non-negative spectrum. A particularly simple example for such an operator is the identity operator $\mathbb{1}_{\mathcal{V}} \in \mathrm{L}(\mathcal{V})$. The set of positive semidefinite operators in $\mathrm{L}(\mathcal{V})$ forms a convex cone which we denote by $\mathrm{Pos}(\mathcal{V})$ [39]. This cone induces a partial ordering on $\mathrm{L}(\mathcal{V})$ and we write $X \succeq Y$ if $X - Y \in \mathrm{Pos}(\mathcal{V})$. On $\mathrm{L}(\mathcal{V})$ we define the Frobenius (or Hilbert-Schmidt) inner product to be

$$\langle X, Y \rangle_{\mathrm{L}(\mathcal{V})} := \mathrm{Tr}(X^\dagger Y) \quad \forall X, Y \in \mathrm{L}(\mathcal{V}), \tag{2}$$

where $\mathrm{Tr}(Z)$ denotes the trace of an operator $Z \in \mathrm{L}(V)$. In addition to that, we are going to require three different matrix norms

$$\begin{aligned} \|X\|_* &:= \mathrm{Tr}\big(\sqrt{X^\dagger X}\,\big) && \text{(nuclear norm/trace norm)}, && (3)\\ \|X\|_{\mathrm{F}} &:= \sqrt{\langle X, X \rangle} && \text{(Frobenius norm)}, && (4)\\ \|X\| &:= \sup_{x \in \mathcal{V}} \frac{\|Xx\|_{\mathrm{F}}}{\|x\|_{\mathrm{F}}} && \text{(spectral norm)}. && (5) \end{aligned}$$

The Frobenius norm is induced by the inner product (2), while the nuclear norm requires the operator square root: for $X \in \mathrm{Pos}(\mathcal{V})$ we let $\sqrt{X} \in \mathrm{Pos}(\mathcal{V})$ be the unique positive semidefinite operator obeying $\sqrt{X}^2 = X$. Note that these norms correspond to the Schatten 1-, Schatten 2- and Schatten ∞-norms, respectively. All Schatten norms are multiplicative under taking tensor products. The Frobenius norm is preserved under any re-grouping of indices, the prime example of such an operation being the vectorization of matrices. This fact justifies our convention to extend the notation $\|\cdot\|_{\mathrm{F}}$ to the 2-norms of vectors and (later on) tensors.

A crucial role is played by the space of *bipartite operators* $\mathrm{L}(\mathcal{W} \otimes \mathcal{V})$, by which we refer operators that act on a tensor product space. For such operators we define the *partial trace* $\mathrm{Tr}_{\mathcal{W}} : \mathrm{L}(\mathcal{W} \otimes \mathcal{V}) \to \mathrm{L}(\mathcal{V})$ as the linear extensions of the map given by

$$\mathrm{Tr}_{\mathcal{W}}(Y \otimes X) := \mathrm{Tr}(Y)\, X\,, \tag{6}$$

where $X \in \mathrm{L}(\mathcal{V})$ and $Y \in \mathrm{L}(\mathcal{W})$, see also Figure 1. Finally, we define our improved regularizer on $\mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ to be

$$\|X\|_{\square} := \max\{\|(\mathbb{1}_{\mathcal{W}} \otimes A) X (\mathbb{1}_{\mathcal{W}} \otimes B)\|_* : A, B \in \mathrm{L}(\mathcal{V}),\, \|A\|_{\mathrm{F}} = \|B\|_{\mathrm{F}} = \sqrt{\dim(\mathcal{V})}\}\,. \tag{7}$$

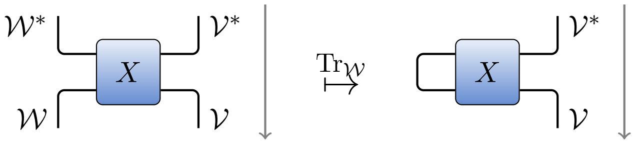


Figure 1: Tensor network diagrams: tensors are denoted by boxes with one line for each index. Contraction of two indices corresponds to connection of the corresponding lines.
Left: A bipartite operator $X \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ viewed as a tensor in $\mathcal{W} \otimes \mathcal{V} \otimes \mathcal{W}^* \otimes \mathcal{V}^*$, i.e., as a tensor with four indices.
Right: Its partial trace $\mathrm{Tr}_{\mathcal{W}}(X)$ as an operator on $\mathcal{V}$.

It is easy to see that $\|\cdot\|_\square$ is a norm and we call it the *square norm*. It will become clear later on that the square norm is closely related to the diamond norm $\|\cdot\|_\diamond$ from quantum information theory [23]. As we will discuss in Section 5.1, $\|X\|_\square = \dim(\mathcal{V}) \left\|J^{-1}(X)\right\|_\diamond$, where J denotes the so-called Choi-Jamiołkowski isomorphism. Both square and diamond norm can be calculated by a semidefinite program (SDP) satisfying strong duality [25]. Also, note that the pair $A = B = \mathbb{1}_{\mathcal{V}}$ is admissible in the maximization (7). Inserting it recovers $\|X\|_*$ and establishes the bound $\|X\|_* \leq \|X\|_\square$. This bound plays a crucial role for our results.

2.2 Convex recovery of structured signals

In this section, we summarize a recent but already widely used geometric proof technique for low-rank matrix recovery. Mainly following Ref. [18], we devote this section to explaining the general reconstruction idea.

In the setting of convex recovery of structured signals, one obtains a *measurement vector* $y \in \mathbb{C}^m$ of a *signal* $x_0 \in \mathcal{V}$ in some vector space $\mathcal{V}$ via a *measurement map* $\mathcal{A} : \mathcal{V} \to \mathbb{C}^m$,

$$y = \mathcal{A}(x_0) + \epsilon\,, \tag{8}$$

where $\epsilon \in \mathbb{C}^m$ represents additive noise in the sampling process. Throughout, we assume linear data acquisition, i.e., $\mathcal{A}$ is linear.

The goal is to efficiently obtain a good approximation to x_0 given $\mathcal{A}$ and y for the case where one only has knowledge about some structure of x_0. Of course, it is desirable that the number m of measurements y_i required for a successful reconstruction is as small as possible. For several different structures of the signal x_0 a general approach of the following form has proven to be very successful [40]. One chooses a convex function $f : \mathcal{V} \to \overline{\mathbb{R}}$ that reflects the structure of x_0 and performs the following convex minimization

$$x^f_\eta = \arg\min\{f(x) : \ \|\mathcal{A}(x) - y\|_{\mathrm{F}} \leq \eta\}\,, \tag{9}$$

where $\eta \geq 0$ is some anticipated error bound.

Next, we give two definitions and a general error bound that has proven to be helpful to find such recovery guarantees. The *descent cone* of a convex function is the set of non-increasing directions u. From the convexity of the function, it follows that the descent cone is a convex cone. The following definitions can also be found, e.g., in Ref. [18].

Definition 1 (Descent cone)**.** *The* descent cone $\mathscr{D}(f,x)$ *of a proper convex function* $f : \mathcal{V} \to \overline{\mathbb{R}}$ *at the point* $x \in \mathcal{V}$ *is*

$$\mathscr{D}(f,x) := \bigcup_{\tau>0} \{u \in \mathcal{V} : \ f(x + \tau u) \leq f(x)\}\,. \tag{10}$$

The *minimum singular value* of a linear map $\mathcal{A}$ is the minimal value of $\|\mathcal{A}(x)\|_\mathrm{F}$ taken over all x with $\|x\|_\mathrm{F} = 1$. Restricting this minimization to a cone yields the *minimum conic singular value.*

Definition 2 (Minimum conic singular value)**.** *Let $\mathcal{A} : \mathcal{V} \to \mathbb{C}^m$ be a linear map and $K \subset \mathcal{V}$ be a cone. The minimum singular value of $\mathcal{A}$ with respect to the cone K is defined as*

$$\lambda_{\min}(\mathcal{A}; K) \coloneqq \inf_{x \in K} \frac{\|\mathcal{A}(x)\|_\mathrm{F}}{\|x\|_\mathrm{F}} . \tag{11}$$

Proposition 3 (Error bound for convex recovery, Tropp's version [18])**.** *Let $x_0 \in \mathcal{V}$ be a signal, $\mathcal{A} \in \mathrm{L}(\mathcal{V} \to \mathbb{C}^m)$ be a measurement map, $y = \mathcal{A}(x_0) + \epsilon$ a vector of m measurements with additive error $\epsilon \in \mathbb{C}^m$, and x^f_η be the solution of the optimization* (9)*. If $\|\epsilon\|_\mathrm{F} \leq \eta$ then*

$$\|x^f_\eta - x_0\|_\mathrm{F} \leq \frac{2\eta}{\lambda_{\min}(\mathcal{A}; \mathscr{D}(f, x_0))} . \tag{12}$$

Note that the statement in Ref. [18] shows this result for real vector spaces only. However, taking a closer look at the proof reveals that it also holds for complex vector spaces as well. We make the following simple but important observation:

Observation 4 (Improved recovery)**.** *The smaller the descent cone the better the recovery guarantee.*

An important example is low-rank matrix recovery. Here, $x_0 = X_0$ is some $n_1 \times n_2$ matrix with $\mathrm{rank}(X_0) = r$ and a low-rank provides structure that allows for reconstruction from a dimension sufficient number of measurements. For this case, choosing $f = \|\cdot\|_*$ to be the nuclear norm has proven very successful, as the nuclear norm is the convex envelope of the matrix rank [9]. In order to give a concrete bound, consider a real matrix X_0 and m measurements $y_j = \mathrm{Tr}\big(A_j^\dagger X_0\big) + \epsilon_j$ with each A_j being a real random matrix with entries drawn independently from a normalized Gaussian distribution. Then one can show that (see, e.g., Ref. [18])

$$\lambda_{\min}(\mathcal{A}; \mathscr{D}(\|\cdot\|_*, X_0)) \geq \sqrt{m-1} - \sqrt{3r(n_1+n_2-r)} - t \tag{13}$$

with probability $1 - \mathrm{e}^{-t^2/2}$ (over the random measurements). As a consequence, a number of $\gtrsim 3\,\mathrm{rank}(X_0)(n_1+n_2-\mathrm{rank}(X_0))$ measurements are enough for a successful reconstruction of the real-valued matrix X_0 with high probability.

3 Results

We show that for certain structured recovery problems, replacing the regularizer f in a convex recovery (9) by an optimized regularizer $f_\square$ can potentially improve performance; see also Figure 2. For the case where f is the nuclear norm and $f_\square$ the square norm, we show such an improvement with numerical simulations in Section 5.

Proposition 5 (Optimizing descent cones)**.** *Let $C \subset \mathcal{V}$ be a convex set and I be a compact index set. Moreover, let $\{f_i\}_{i \in I}$ be a family of upper semi-continuous convex functions $f_i : C \to \mathbb{R}$. Define another convex function $f_\square$ as the point-wise supremum $f_\square(x) \coloneqq \sup_{i \in I} f_i(x)$. Then*

$$\mathscr{D}(f_\square; x) \subset \bigcap_{i \in I(x)} \mathscr{D}(f_i; x) \tag{14}$$

for any $x \in C$, where $I(x) \coloneqq \{i \in I : f_i(x) = f_\square(x)\}$ is the active index set at x with the convention $\bigcap_{i \in \emptyset} \mathscr{D}(f_i; x) \coloneqq \mathcal{V}$.

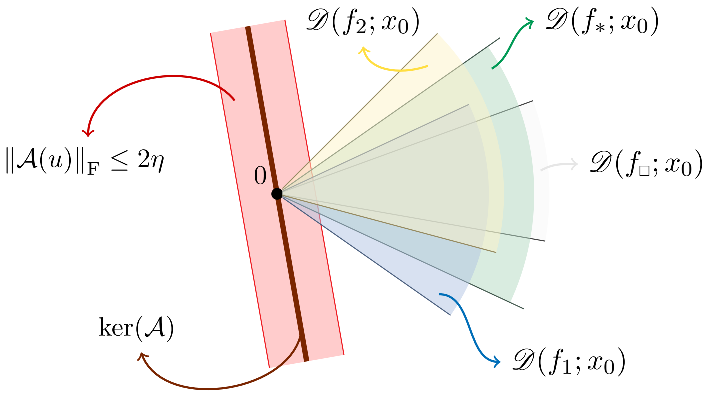


Figure 2: Extension of the geometric arguments [18] used to establish Proposition 3. The descent cone $\mathscr{D}(f_\square;x_0)$ of the optimized regularizer $f_\square$ is contained in an intersection of descent cones.

Proof of Proposition 5. By $\operatorname{cone}(S) := \bigcup_{\tau>0}\{\tau s : s \in S\}$ we will denote the cone generated by a set S. According to Definition 1 of the descent cone, we have

$$\mathscr{D}(f_\square;x) = \bigcup_{\tau>0}\{u \mid \sup_{i\in I} f_i(x+\tau u) \leq f_\square(x)\}\,. \tag{15}$$

Writing the supremum as an intersection yields

$$\mathscr{D}(f_\square;x) = \bigcup_{\tau>0}\bigcap_{i\in I}\{\tau u \mid f_i(x+u) \leq f_\square(x)\} \tag{16}$$

$$\subset \bigcap_{i\in I}\operatorname{cone}\{u \mid f_i(x+u) \leq f_\square(x)\}\,. \tag{17}$$

By $B_\epsilon \subset \mathcal{V}$ we denote the ball around the origin of radius ϵ. Now, consider a non-active index $i \in I \setminus I(x)$. As f_i is upper semi-continuous, there exists $\epsilon > 0$ such that for all $u \in B_\epsilon$ we have $f_i(x+u) < f_\square(x)$. Hence, the set $B_\epsilon \subset \{u \mid f_i(x+u) \leq f_\square(x)\}$ and hence the corresponding cone in Eq. (17) is the entire space. Therefore, every non-active index i can be omitted in the intersection,

$$\mathscr{D}(f_\square;x) \subset \bigcap_{i\in I(x)}\operatorname{cone}\{u \mid f_i(x+u) \leq f_i(x)\}\,. \tag{18}$$

The definition of the descent cone of f_i finishes the proof. □

The square norm (7) is a particular instance of such a supremum over nuclear norms. Thanks to the following nuclear norm bound (20), Proposition 5 can lead to an improved recovery for any bipartite operator $X \in \mathrm{L}(\mathcal{W}\otimes\mathcal{V})$ satisfying

$$\|X\|_* = \|X\|_\square\,. \tag{19}$$

Here, we will only need the lower bound on the square norm but, in order to fully relate it to the usual matrix norms, we also provide two upper bounds.

Proposition 6 (Bounds to the square norm)**.** *For any $X \in \mathrm{L}(\mathcal{W}\otimes\mathcal{V})$*

$$\|X\|_* \leq \|X\|_\square\,, \tag{20}$$

$$\|X\|_\square \leq \dim(\mathcal{V})\,\|X\|_*\,, \tag{21}$$

$$\|X\|_\square \leq \dim(\mathcal{W}\otimes\mathcal{V})\,\|X\|\,. \tag{22}$$

Our second main result fully characterizes the set of operators satisfying Eq. (19).

Theorem 7 (Extremal operators)**.** *Let $X \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ be a bipartite operator. Then Eq. (19) holds if and only if*

$$\mathrm{Tr}_{\mathcal{W}}\big(\sqrt{XX^\dagger}\big) = \mathrm{Tr}_{\mathcal{W}}\big(\sqrt{X^\dagger X}\big) = \frac{\|X\|_*}{\dim(\mathcal{V})}\,\mathbb{1}_{\mathcal{V}}. \tag{23}$$

For now, we content ourselves with sketching the proof idea and present the full proof later.

Proof idea. For the case where Eq. (19) is satisfied, we exploit it to single out a primal feasible optimal point. Exact knowledge of this point together with complementary slackness then allow to severely restrict the range of possible dual optimal points. Relation (23) is an immediate consequence of these restrictions.

To show the converse, we insert a particular feasible point into the dual SDP of the square norm. Eq. (23) enables us to explicitly evaluate the objective function at this point. Doing so yields $\|X\|_*$ which in turn implies $\|X\|_\square \le \|X\|_*$ by weak duality. Combining this implication with the converse bound from Proposition 6 establishes $\|X\| = \|X\|_\square$, as claimed. □

As an implication of Theorem 7 and Proposition 5 we obtain the following.

Corollary 8 (Intersection of descent cones)**.** *Let $X \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ satisfy Eq. (19). Then*

$$\mathscr{D}(\|\cdot\|_\square\,; X) \subset \bigcap_{(A,B)\in I(X)} \mathscr{D}(\|(\mathbb{1}_{\mathcal{W}} \otimes A)(\,\cdot\,)(\mathbb{1}_{\mathcal{W}} \otimes B)\|_*\,; X)\,, \tag{24}$$

where $I(X)$ contains all $A, B \in \mathrm{L}(V)$ with $\|A\|_{\mathrm{F}} = \|B\|_{\mathrm{F}} = 1$ and being active in the sense that $\|X\|_\square = \|(\mathbb{1}_{\mathcal{W}} \otimes A)X(\mathbb{1}_{\mathcal{W}} \otimes B)\|_*$.

For instance, setting $A = B = \mathbb{1}_{\mathcal{V}}$ gives an element of $I(X)$ and yields the inclusion $\mathscr{D}(\|\cdot\|_\square\,; X) \subset \mathscr{D}(\|\cdot\|_*\,; X)$ for any X satisfying Eq. (19). As an immediate application, we will see in the next section that the square norm inherits recovery guarantees from the nuclear norm.

4 Applications to low-rank matrix recovery

In this section we focus on low-rank matrix recovery of Hermitian bipartite operators $X_0 \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ that are either real-valued or complex-valued. As already mentioned in Section 2.2, there the task is to efficiently recover an unknown matrix X_0 of low-rank r from m noisy linear measurements of the form

$$y_i = \mathrm{Tr}\,(A_i X_0) + \epsilon_i, \quad i = 1, \dots, m\,, \tag{25}$$

where $A_1, \dots, A_m \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ are the measurement matrices and $\epsilon_1, \dots, \epsilon_m \in \mathbb{R}^m$ denotes additive noise in the sampling process. By introducing a measurement map $\mathcal{A} : \mathrm{L}(\mathcal{W} \otimes \mathcal{V}) \to \mathbb{R}^m$ of the form $\mathcal{A}(X_0) = \sum_{i=1}^m \mathrm{Tr}\,(A_i X_0)\, e_i$, where $e_1, \dots, e_m$ denotes the standard basis in $\mathbb{R}^m$, the entire measurement process can be summarized as

$$y = \mathcal{A}(X_0) + \epsilon\,. \tag{26}$$

Here, $y = (y_1, \dots, y_m)^T \in \mathbb{R}^m$ contains all measurement outcomes and $\epsilon \in \mathbb{R}^m$ denotes the noise vector. If a bound $\|\epsilon\|_{\mathrm{F}} \leq \eta$ on the noise is available, many measurement scenarios have been identified where estimating X_0 by

$$X_\eta^* := \arg\min\{\|X\|_* : \|\mathcal{A}(X) - y\|_{\mathrm{F}} \leq \eta\} \tag{27}$$

from noisy data of the form (26) stably recovers X_0. Note that by employing the well-known SDP formulation of the nuclear norm [41] this optimization can be recast as

$$\begin{aligned} X_\eta^* = \quad & \underset{X,Y,Z}{\arg\min} && \tfrac{1}{2}\big(\mathrm{Tr}(Y) + \mathrm{Tr}(Z)\big) \\ & \text{subject to} && \begin{pmatrix} Y & -X \\ -X^\dagger & Z \end{pmatrix} \succeq 0\,, \\ &&& Y, Z \in \mathrm{Pos}(\mathcal{W} \otimes \mathcal{V})\,, \\ &&& \|\mathcal{A}(X) - y\|_{\mathrm{F}} \leq \eta\,. \end{aligned} \tag{28}$$

What is more, several of these recovery guarantees can be established using the geometric proof techniques presented in Section 2.2. For results established that way, combining Observation 4 with Corollary 8 allows us to draw the following conclusion.

Implication 9 (Inheriting recovery guarantees)**.** *For bipartite operators $X_0 \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ that satisfy $\|X\|_* = \|X\|_\square$, any recovery guarantee for nuclear norm minimization, which is based on the nuclear norm's descent cone, also holds for square norm minimization.*

This insight indicates that replacing nuclear norm regularization (27) by

$$X_\eta^\square := \arg\min\{\|X\|_\square : \|\mathcal{A}(X) - y\|_{\mathrm{F}} \leq \eta\} \tag{29}$$

results in an estimation procedure that performs at least as well whenever $\|X_0\| = \|X_0\|_*$. In fact, Observation 4 suggests that it may actually outperform traditional recovery procedures. Also, the SDP formulation for the square norm [25] allows one to recast the optimization (29) as

$$\begin{aligned} X_\eta^\square = \quad & \underset{X,Y,Z}{\arg\min} && \tfrac{\dim(\mathcal{V})}{2}\big(\|\mathrm{Tr}_{\mathcal{W}}(Y)\| + \|\mathrm{Tr}_{\mathcal{W}}(Z)\|\big) \\ & \text{subject to} && \begin{pmatrix} Y & -X \\ -X^\dagger & Z \end{pmatrix} \succeq 0\,, \\ &&& Y, Z \in \mathrm{Pos}(\mathcal{W} \otimes \mathcal{V})\,, \\ &&& \|\mathcal{A}(X) - y\|_{\mathrm{F}} \leq \eta\,, \end{aligned} \tag{30}$$

which, just like the optimization (28), is a convex optimization problem that can be solved computationally efficiently. In the remainder of this section, we present three measurement scenarios for which Implication 9 holds. The first one is a version of Ref. [18, Example 4.4] which is valid for reconstructing real-valued matrices. In its original formulation with nuclear norm minimization, it follows from combining Proposition 3 and Eq. (13).

Proposition 10 (Stable recovery of real matrices via Gaussian measurements)**.** *Let $X_0 \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ be a real valued, bipartite matrix of rank r that obeys $\|X_0\|_\square = \|X_0\|_*$. Also, suppose that each measurement matrix A_i is a real-valued standard Gaussian matrix and the overall noise is bounded as $\|\epsilon\|_{\mathrm{F}} \leq \eta$. Then, $m \geq Cr \dim(\mathcal{W} \otimes \mathcal{V})$ noisy measurements of the form (26) suffice to guarantee*

$$\left\|X_\eta^\square - X_0\right\|_{\mathrm{F}} \leq \frac{C'\eta}{\sqrt{m}} \tag{31}$$

with probability at least $1 - \mathrm{e}^{-C''m}$. Here, C, C' and C'' denote absolute constants.

With high probability (w.h.p.), this statement assures *stable* recovery, meaning that the reconstruction error (31) scales linearly in the noise bound η and inversely proportional to $\sqrt{m}$.

For the sake of clarity, we have refrained from providing explicit values for the constants C, C' and C'' in Proposition 10. However, resorting to Tropp's bound (13) on the minimal conical eigenvalue of a Gaussian sampling matrix reveals that stably recovering any rank-r matrix obeying Eq. (19) requires roughly

$$m \gtrsim 6r(\dim(\mathcal{V})\dim(\mathcal{W}) - r) \tag{32}$$

independently selected Gaussian measurements.

Proposition 10 is a prime example for a *non-uniform* recovery guarantee: For any fixed rank-r matrix X_0 obeying Eq. (19), m randomly chosen measurements of the form (8) suffice to stably reconstruct X_0 w.h.p. For some measurement scenarios, stronger recovery guarantees can be established. Called *uniform* recovery guarantees, these results assure that one choice of sufficiently many random measurements w.h.p. suffices to reconstruct all possible matrices of a given rank.

A uniform recovery statement can be established for the following real-valued measurement scenario [17]: suppose that with respect to an arbitrary orthonormal basis of $\mathcal{W} \otimes \mathcal{V}$, each matrix element of A_i is an independent instance of a real-valued random variable a obeying

$$\mathbb{E}[a] = 0, \quad \mathbb{E}\left[a^2\right] = 1 \quad \text{and} \quad \mathbb{E}\left[a^4\right] \leq F, \tag{33}$$

where $F \geq 1$ is an arbitrary constant. Measurement matrices of this form can be considered as a generalization of Gaussian measurement matrices, where each matrix element corresponds to a standard Gaussian random variable. In Ref. [17] – see also Refs. [42, 43] – a uniform recovery guarantee for such measurement matrices has been established by means of the *Frobenius robust rank null space property* [17, Definition 10]. Such a proof technique is different from the geometric one introduced in Section 2.2. However, as laid out in the appendix, some auxiliary statements allow for reassembling technical statements from these works to yield a slightly weaker, but still uniform, statement by means of analyzing descent cones. Implication 9 is applicable for such a result and yields the following.

Proposition 11 (Stable, uniform recovery of real matrices via measurement matrices with finite fourth moments)**.** *Consider the measurement process described in Eq.* (26)*, where each* $A_i \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ *is an independent random matrix of the form* (33)*. Fix* $r \geq 1$ *and suppose that* $m \geq C_F\, r \dim(\mathcal{W} \otimes \mathcal{V})$*. Then, w.h.p., every real-valued matrix* $X_0 \in \mathrm{L}(\mathcal{V} \otimes \mathcal{W})$ *of rank at most* r *and obeying* $\|X\|_* = \|X\|$ *can be stably reconstructed from the measurements* (26) *by means of square norm minimization* (29)*. Here,* C_F *is a constant that only depends on the fourth-moment bound* F*.*

We conclude this section with two uniform recovery guarantees for Hermitian low-rank matrices from measurement matrices A_i that are proportional to rank-one projectors, i.e., $A_i = a_i a_i^*$ for some $a_i \in \mathcal{W} \otimes \mathcal{V}$. Originally established for nuclear norm minimization in Ref. [16], by using an extension of the geometric proof techniques presented in Section 2.2, Implication 9 is directly applicable to such measurements.

Proposition 12 (Stable, uniform recovery of Hermitian matrices from rank-one measurements)**.** *Consider recovery of Hermitian rank-*r *matrices* $X \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ *that obey* $\|X\|_\square = \|X\|_*$ *from rank-one measurements of the form* $A_i = a_i a_i^*$*. Let* $n = \dim(\mathcal{W} \otimes \mathcal{V})$*. Then stable and uniform recovery guarantees for square norm minimization* (29) *analogous to Proposition 11 hold if either*

1. *the measurements a_i are $m \geq C_G rn$ random Gaussian vectors in $\mathcal{W} \otimes \mathcal{V}$ or*

2. *the measurements a_i are $m \geq C_{4D} rn \log(n)$ vectors drawn uniformly from a complex projective 4-design.*

Once more, C_G and C_{4D} denote absolute constants of sufficient size.

In the statement above, a *complex projective t-design* is a configuration of vectors which is "evenly distributed" on a sphere in the sense that sampling uniformly from it reproduces the moments of Haar measure up to order $2t$ [44–46]. More precisely,

$$\frac{1}{N}\sum_{j=1}^{N}(w_j w_j^\dagger)^{\otimes t} = \int_{\|w\|_{\mathrm{F}}=1} (ww^\dagger)^{\otimes t} \mathrm{d}w. \tag{34}$$

The second statement in Proposition 12 can be seen as "partial derandomization" of the first one [35]

5 Application to the recovery of linear maps on operators

Now we come to three concrete applications concerning linear maps that take operators in $\mathrm{L}(\mathcal{V})$ to operators in $\mathrm{L}(\mathcal{W})$. Our reconstruction based on the square norm can be applied to such maps by identifying them with operators in $\mathrm{L}(\mathcal{W}\otimes\mathcal{V})$. We start with introducing some relevant notation and explain such an identification, the Choi-Jamiołkowski isomorphism, in more detail. Then we present numerical results on retrieval of certain unitary basis changes, quantum process tomography, and blind matrix deconvolution.

5.1 Notation concerning linear maps on operators

Our square norm is closely related to the diamond norm, which is defined for linear operators $M : \mathrm{L}(\mathcal{V}) \to \mathrm{L}(\mathcal{W})$ that map operators to operators. We call such objects *maps* and denote their space by $\mathbb{L}(\mathcal{V},\mathcal{W}) := \mathrm{L}\left(\mathrm{L}(\mathcal{V}) \to \mathrm{L}(\mathcal{W})\right)$, or simply by $\mathbb{L}\left(\mathcal{V}\right) := \mathbb{L}\left(\mathcal{V},\mathcal{V}\right)$. We also denote maps by capital latin letters. Concretely, for $M \in \mathbb{L}(\mathcal{V},\mathcal{W})$ and $X \in \mathrm{L}(\mathcal{V})$ we write $M(X) \in \mathrm{L}(\mathcal{W})$. A particularly simple example is the identity map $\mathbb{1}_{\mathrm{L}(\mathcal{V})} \in \mathbb{L}(\mathcal{V})$ which obeys $\mathbb{1}_{\mathrm{L}(\mathcal{V})}(X) = X$ for all $X \in \mathrm{L}(\mathcal{V})$.

We would like to identify maps in $\mathbb{L}(\mathcal{V},\mathcal{W})$ with operators in $\mathrm{L}(\mathcal{W}\otimes\mathcal{V})$, for which we have discussed certain reconstruction schemes. For this purpose, we employ a very useful isomorphism, called the *Choi-Jamiołkowski isomorphism* [47, 48]. In order to explicitly define this isomorphism, we fix an orthogonal basis (e_i) of $\mathcal{V}$. This also gives rise to an operator basis

$$E_{i,j} := e_i e_j^T \in \mathrm{L}(\mathcal{V}) \tag{35}$$

and we define *vectorization* $\mathrm{vec} : \mathrm{L}(\mathcal{V}) \to \mathcal{V}\otimes\mathcal{V}$ by the linear extension of

$$\mathrm{vec}(E_{i,j}) := e_i \otimes e_j\,. \tag{36}$$

Then the Choi-Jamiołkowski isomorphism J is defined by

$$\begin{aligned} J : \mathbb{L}(\mathcal{V},\mathcal{W}) &\to \mathrm{L}(\mathcal{W}\otimes\mathcal{V}) \\ M &\mapsto \sum_{i,j=1}^{\dim(\mathcal{V})} M(E_{i,j}) \otimes E_{i,j}\,. \end{aligned} \tag{37}$$

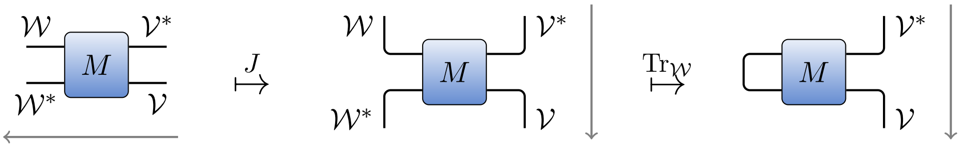


Figure 3: Tensor network diagrams: tensors are denoted by boxes with one line for each index. Contraction of two indices corresponds to connection of the corresponding lines.
Left: Order-4 tensor M as a map from $\mathrm{L}(\mathcal{V}) \cong \mathcal{V} \otimes \mathcal{V}^*$ to $\mathrm{L}(\mathcal{W}) \cong \mathcal{W} \otimes \mathcal{W}^*$.
Middle: Its Choi-matrix $J(M)$ as an operator on $\mathcal{W}^* \otimes \mathcal{V} \cong \mathcal{W} \otimes \mathcal{V}$.
Right: Its partial trace $\mathrm{Tr}_{\mathcal{W}}(J(M))$ as an operator on $\mathcal{V}$.

The resulting operator $J(M)$ is called the *Choi matrix* of M. It can be straightforwardly checked that Eq. (37) is equivalent to setting

$$J(M) = \Big(M \otimes \mathbb{1}_{\mathrm{L}(\mathcal{V})}\Big)\big(\mathrm{vec}(\mathbb{1}_{\mathcal{V}})\,\mathrm{vec}(\mathbb{1}_{\mathcal{V}})^T\big). \tag{38}$$

Although not evident from Eq. (37), this isomorphism is actually basis independent. Indeed, it is just an instance of the natural isomorphism $\mathcal{W}\otimes\mathcal{W}^*\otimes\mathcal{V}^*\otimes\mathcal{V} \cong \mathcal{W}\otimes\mathcal{V}^*\otimes\mathcal{W}^*\otimes\mathcal{V}$. This identification is illustrated in Figure 3, and discussed in more detail in the appendix.

Similarly to the definition of the spectral norm (5), the nuclear norms on $\mathrm{L}(\mathcal{V})$ and $\mathrm{L}(\mathcal{W})$ induce a norm on $\mathbb{L}(\mathcal{V}, \mathcal{W})$,

$$\|M\|_{*\to*} := \sup_{X\in\mathrm{L}(\mathcal{V})} \frac{\|M(X)\|_*}{\|X\|_*}\,. \tag{39}$$

Perhaps surprisingly, the induced nuclear norm of maps of the form $M \otimes \mathbb{1}_{\mathrm{L}(\mathcal{V})}$ can be computed efficiently [23–25], as explained in detail below. This motivates studying the *diamond norm* [21]

$$\|M\|_\diamond := \left\|M \otimes \mathbb{1}_{\mathrm{L}(\mathcal{V})}\right\|_{*\to*}. \tag{40}$$

It plays an important role in quantum mechanics [21] and is also the core concept of this work. Using the Choi-Jamiołkowski isomorphism, the diamond norm (40) can indeed be written [25] as

$$\|M\|_\diamond = \frac{\|J(M)\|_\square}{\dim(\mathcal{V})}\,, \tag{41}$$

where the square norm was defined variationally in Eq. (7). Hence, for the case of a measurement map $\mathcal{A} : \mathbb{L}(\mathcal{V}, \mathcal{W}) \to \mathbb{C}^m$, the reconstruction based on the square norm (30) can also be written as

$$\begin{aligned} M_\eta^\diamond = \quad & \arg\min && \tfrac{1}{2}\,\|\mathrm{Tr}_{\mathcal{W}}(Y)\| + \tfrac{1}{2}\,\|\mathrm{Tr}_{\mathcal{W}}(Z)\| \\ & \text{subject to} && \begin{pmatrix} Y & -J(M) \\ -J(M)^\dagger & Z \end{pmatrix} \succeq 0\,, \\ &&& Y, Z \in \mathrm{Pos}(\mathcal{W}\otimes\mathcal{V})\,, \\ &&& \|\mathcal{A}(M) - y\|_{\mathrm{F}} \le \eta\,. \end{aligned} \tag{42}$$

5.2 Retrieval of certain unitary basis changes

Our problem of retrieval of unitary basis changes is motivated by the phase retrieval problem. Retrieving phases from measurements that are ignorant towards them has a long-standing history in various scientific disciplines [29]. A discretized version of this problem

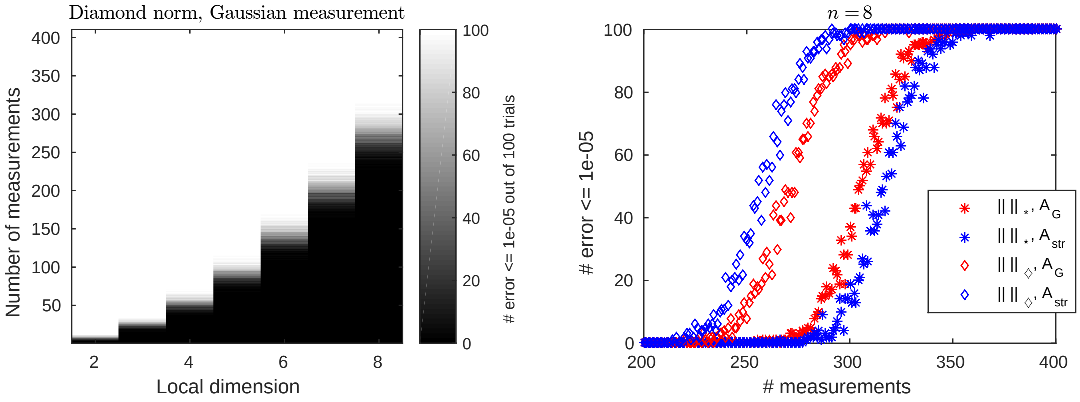


Figure 4: Retrieval of $M(X) = UXV$ for the case of real numbers and $\epsilon = 0$. $U, V \in \mathrm{O}(n)$ are orthogonal matrices. The plots show the number of trails out of 100 with small errors, $\|M^\diamond_{\text{eps}} - M_0\|_{\mathrm{F}} \leq 10^{-5}$ and $\|M^*_{\text{eps}} - M_0\|_{\mathrm{F}} \leq 10^{-5}$, respectively, and with η chosen as machine precision eps.
Left: Diamond norm minimization with Gaussian measurements for different local dimensions n.
Right: Comparison of diamond norm and nuclear norm with Gaussian and structured measurements. Note that the structured measurements improve the reconstruction based on the diamond norm while for the reconstruction based on the nuclear norm Gaussian measurements turn out to work better. The computation time needed for the recovery is approximately the same for both methods.

can be phrased as the task of inferring a complex vector $x \in \mathbb{C}^n$ from measurements of the form

$$y_i = |\langle a_i, x_i \rangle|^2 \,, \tag{43}$$

where $a_1, \dots, a_m \in \mathbb{C}^n$. Recently, the mathematical structure of this problem has received considerable attention [29–36]. One way of approaching this problem is to recast it as a matrix problem which has the benefit that the measurements (43) become linear. Indeed, setting $X := xx^\dagger$ and $A_i = a_i a_i^\dagger$ reveals that

$$y_i = |\langle a_i, x \rangle|^2 = \mathrm{Tr}\big(a_i a_i^\dagger x x^\dagger\big) = \mathrm{Tr}\left(A_i X\right). \tag{44}$$

This “lifting” trick allows for re-casting the phase retrieval problem as the task of recovering a Hermitian rank-one matrix $X = xx^\dagger$ from linear measurements of the form $A_i = a_i a_i^\dagger$.

Recently, Ling and Strohmer [49] used similar techniques to recast the important problem of self-calibration in hardware devices as the task to recover a non-Hermitian rank-one matrix $X = xy^\dagger$ from similar linear measurements.

In this section, we consider the matrix-analogue of such a task and set $\mathcal{V} = \mathbb{C}^n = \mathcal{W}$ but keep $\mathcal{V}$ and $\mathcal{W}$ as labels. Concretely, we consider maps $M \in \mathbb{L}(\mathcal{V}, \mathcal{W})$ of the form

$$M(X) = UXV \,, \tag{45}$$

where U and V are fixed unitaries. Note that any such map has a Choi matrix of the form

$$\begin{aligned} J(M) =& M \otimes \mathbb{1}_{\mathrm{L}(\mathcal{V})} \operatorname{vec}(\mathbb{1}_{\mathrm{L}(\mathcal{V})}) \operatorname{vec}(\mathbb{1}_{\mathrm{L}(\mathcal{V})})^\dagger \\ =& \left(U \otimes \mathbb{1}_{\mathrm{L}(\mathcal{V})} \operatorname{vec}(\mathbb{1}_{\mathrm{L}(\mathcal{V})})\right) \left(V \otimes \mathbb{1}_{\mathrm{L}(\mathcal{V})} \operatorname{vec}(\mathbb{1}_{\mathrm{L}(\mathcal{V})})\right)^\dagger , \end{aligned} \tag{46}$$

which corresponds to an outer product of the form $xy^\dagger$. Moreover, unitarity of both U and V assures that all such maps meet the requirements of Theorem 7.

We aim to numerically recover such maps from two different types of measurements: (i) Gaussian measurements and (ii) structured measurements. The Gaussian measurements are given by a measurement map $\mathcal{A}_G : \mathbb{L}(\mathcal{V}, \mathcal{W}) \to \mathbb{C}^m$ with real and imaginary parts of all of its components drawn from a normal distribution with zero mean and unit variance. In the case of structured measurements, M receives rank-1 inputs and then inner products with regular measurement matrices are measured. More precisely, the measurement map $\mathcal{A}_{\text{str}} : \mathbb{L}(\mathcal{V}, \mathcal{W}) \to \mathbb{C}^m$ is given by

$$\mathcal{A}_{\text{str}}(M)_j := \operatorname{Tr}\bigl(A_j M(x_j y_j^\dagger)\bigr) , \quad j \in [m] , \tag{47}$$

where x_j, y_j are chosen uniformly from the complex unit sphere $\{z \in \mathcal{V} : \|z\|_{\text{F}} = 1\} \subset \mathcal{V}$. The matrices A_j, on the other hand, are independent instances of the random matrix UDV, where $D \in \mathrm{L}(\mathcal{V})$ is a fixed, real-valued diagonal matrix and both U and V are chosen independently from the unique unitarily invariant Haar measure over $U(\dim(\mathcal{V}))$. For our numerical studies, we restrict ourselves to even dimensions $n = \dim(\mathcal{V})$ and set $D = \frac{2}{n}(1, -1, 2, -2, \ldots, n/2, -n/2)$. This in particular assures $\|D\| = 1$. As we will see, similar types of measurements can be used in quantum process tomography and blind matrix deconvolution.

For both measurement setups, we find that diamond norm reconstruction outperforms nuclear norm reconstruction; see Figure 4. Interestingly, the structured measurements are better than the Gaussian measurements for the diamond norm reconstruction, while for the nuclear norm reconstruction we find the converse.

Finally, we would like to to point out that Ling and Strohmer introduced a new algorithm – dubbed "SparseLift" – to efficiently reconstruct the signals they consider and simultaneously promote sparsity [49]. It is an intriguing open problem to compare the performance of SparseLift to the constrained diamond norm minimization advocated here for different types of practically relevant measurement ensembles. We leave this to future work.

5.3 Quantum process tomography

The problem of reconstructing quantum mechanical processes from measurements is referred to as *quantum process tomography*. As explained in the next paragraph, quantum processes are described by maps that saturate the norm inequality (19) and thus are natural candidates for diamond norm-based methods.

Preliminaries. A positive semidefinite operator $\rho \in \operatorname{Pos}(\mathcal{V})$ with unit trace $\operatorname{Tr}(\rho) = \|\rho\|_* = 1$ is called a *density operator* and a matrix representation is a *density matrix*. The convex space of density operators is denoted by $\mathcal{D}(\mathcal{V}) \subset \operatorname{Pos}(\mathcal{V})$ and its elements are referred to as *quantum states*. The extreme elements of $\mathcal{D}(\mathcal{V})$ are called *pure states* and are given by rank-one operators of the form $\psi\psi^\dagger$ with 2-norm normalized *state vectors* $\psi \in \mathcal{V}$. An *observable* is a self-adjoint operator $A \in \operatorname{Herm}(\mathcal{V})$ and the *expectation value* of A in state $\rho \in \mathcal{D}(\mathcal{V})$ is $\operatorname{Tr}(\rho A)$. Note that in the case where ρ and A are diagonal, ρ corresponds to a classical probability vector and A to a random variable also with expectation value $\operatorname{Tr}(\rho A)$. For the following definitions it is helpful to know that quantum systems are composed to larger quantum systems by taking tensor products of operators. A map $M \in \mathbb{L}(\mathcal{V}, \mathcal{W})$ is called *completely positive* if $J(M) \in \operatorname{Pos}(\mathcal{W} \otimes \mathcal{V})$ with J from Eq. (38). This is the case if and only if for every vector space $\mathcal{V}$ the map $M \otimes \mathbb{1}_{\mathrm{L}(\mathcal{V})}$ preserves the cone $\operatorname{Pos}(\mathcal{V} \otimes \mathcal{W})$ of positive semidefinite operators. $M \in \mathbb{L}(\mathcal{V}, \mathcal{W})$ is called *trace preserving* if $\operatorname{Tr}(M(X)) = \operatorname{Tr}(X)$ for all $X \in \mathrm{L}(\mathcal{V})$. The convex space of maps that are both, completely positive and trace preserving is denoted by $\mathrm{CPT}(\mathcal{V},\mathcal{W}) \subset \mathbb{L}(\mathcal{V},\mathcal{W})$ and its elements are quantum operations as they map density operators to density operators and they are also called *quantum channels*. The *Kraus rank* of a quantum channel $M \in \mathrm{CPT}(\mathcal{V},\mathcal{W})$ is the rank of its Choi matrix $J(M)$. A channel $M \in \mathrm{CPT}(\mathcal{V},\mathcal{W})$ of Kraus rank r can be written as

$$M(\rho) = \sum_{j=1}^{r} K_j \rho K_j^\dagger \,, \tag{48}$$

where $K_j \in \mathrm{L}(\mathcal{V} \to \mathcal{W})$ are so-called *Kraus operators* satisfying $\sum_{j=1}^{r} K_j^\dagger K_j = \mathbb{1}_{\mathcal{V}}$, and no other such decomposition has fewer terms. A special role is played by *unitary channels*, which are channels of unit Kraus rank. In this case, the single Kraus operator in the Kraus representation (48) has to be unitary. Unitary quantum channels describe coherent operations in the sense that for isolated quantum systems (i.e., systems that are decoupled from anything else) one can only have unitary quantum channels. Quantum channels describing situations where the system is affected by noise have Kraus ranks larger than one. In many experimental situations, one aims at the implementation of a unitary channel, but actually implements a channel whose Kraus rank is larger than one, but is still approximately low. Therefore, process tomography of quantum channels with low Kraus rank is an important task in quantum experiments. Also, in the context of *quantum error corretion*, low-rank deviations turn out to have a particularly adverse impact [50]. This underscores the need to design efficient estimation protocols for this case.

In the next paragraph, we present numerical results showing that, indeed, replacing the nuclear norm with the diamond norm in a straightforward "compressive process tomography" improves the results. We find it plausible that using the diamond norm as a "drop in replacement" for the nuclear norm will also lead to improvements in other, more advanced process tomography schemes. For example, Kimmel and Liu [51] combine compressed process tomography with ideas from *randomized benchmarking* [52, 53]. This combination allows recovery using only Clifford measurements that are robust to state preparation and measurement (SPAM) errors. Their recovery guarantees are based on the geometric arguments presented in Section 2.2. It thus seems fruitful to conduct numerical experiments using the diamond norm in their setting.

Numerical results for quantum process tomography. The task is to reconstruct $M_0 \in \mathrm{CPT}(\mathcal{V},\mathcal{W})$ from measurements of the form

$$y = \mathcal{A}(M_0) + \epsilon, \tag{49}$$

where $\mathcal{A} : \mathbb{L}(\mathcal{V},\mathcal{W}) \to \mathbb{R}^m$ encodes linear data acquisition, $y \in \mathbb{R}^m$ summarizes the measurement outcomes, and $\epsilon \in \mathbb{R}^m$ represents additive noise. The most general measurements conceivable in this context are so-called *process POVMs* [54]. However, here we consider the case where $\mathcal{A}$ is given by the preparation of pure states given by state vectors $\psi_j \in \mathcal{V}$ and measurements of observables $A_j \in \mathrm{Herm}(\mathcal{W})$, where $j \in [m]$. This yields similar measurements as in Section 5.2,

$$y_j = \mathcal{A}(M_0)_j \coloneqq \mathrm{Tr}\big(A_j M_0(\psi_j \psi_j^\dagger)\big) + \epsilon_j \,, \quad j \in [m] \,, \tag{50}$$

where each $\psi_j \in \mathcal{V}$ is chosen uniformly and independently from the complex unit sphere in $\mathcal{V}$. Each observable $A_j \in \mathrm{Herm}(\mathcal{W})$ is of the form $A_j = U_j D U_j^\dagger$, where each $U_j \in U(\dim(\mathcal{W}))$ is drawn independently from the Haar measure over all unitaries. Once more, $D \in \mathrm{Herm}(\mathcal{W})$ is a fixed Hermitian operator with non-degenerate spectrum. With this

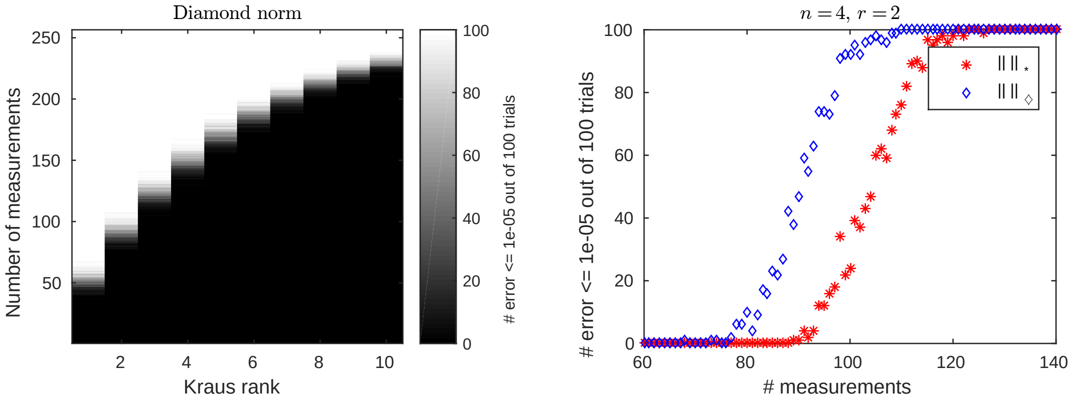


Figure 5: Retrieval of quantum channels acting on two qubits ($n = 4$) with $\epsilon = 0$. The plots show the number of trails out of 100 with small errors, $\left\|M^{\diamond}_{\text{eps}} - M_0\right\|_{\text{F}} \leq 10^{-5}$ and $\left\|M^{*}_{\text{eps}} - M_0\right\|_{\text{F}} \leq 10^{-5}$, respectively, and with η chosen as machine precision eps.
Left: Diamond norm recovery for different Kraus ranks.
Right: Comparison of diamond norm and trace norm of the Choi matrix for Kraus rank $r = 2$. The diamond norm recovery works with fewer measurements than the conventional nuclear norm recovery, while the computation time is approximately the same.

measurement setup, quantum channels can be recovered from few measurements. Once more, diamond reconstruction outperforms the conventional nuclear norm reconstruction, see Figure 5.

5.4 Blind matrix deconvolution

The *blind deconvolution scheme* as considered in Ref. [15] aims to reconstruct unknown vectors $h \in \mathbb{R}^k$ and $m \in \mathbb{R}^n$. From this, length L signals are being generated as

$$w = Bh \qquad \text{and} \qquad x = Cm, \tag{51}$$

for known $B \in \mathrm{L}\left(\mathbb{R}^k \to \mathbb{R}^L\right)$ and $C \in \mathrm{L}\left(\mathbb{R}^n \to \mathbb{R}^L\right)$. The observed quantity is the circular convolution of w and x,

$$y = w * x = \sum_{i=1}^{L} \left(\sum_{j=1}^{L} w_j \, x_{i-j+1 \mod L} \right) e_i \,, \tag{52}$$

where $(e_1, \ldots, e_L)$ denotes the standard basis of $\mathbb{R}^L$. This gives rise to a bi-linear problem, which can still be solved using a lifting technique to a variant of the matrix completion problem.

The type of problem considered in this work allows for the *blind matrix deconvolution*, in which not vectors h, w, but orthogonal or unitary matrices U, V reflecting unknown rotations are reconstructed.

In this new problem, for known $B, C \in \mathrm{L}\left(\mathbb{C}^N \to \mathbb{C}^L\right)$ and real vectors $h^{(q)}, m^{(q)} \in \mathbb{R}^N$ with $q \in [Q]$, that are an input to the problem, we seek to reconstruct $U, V \in U(n)$ from the circular convolutions $y^{(q)} = w^{(q)} * x^{(q)}$ of $w^{(q)}$ and $x^{(q)}$, where now

$$\begin{aligned} w^{(q)} &= BUh^{(q)}, \\ x^{(q)} &= CVm^{(q)}, \end{aligned} \tag{53}$$

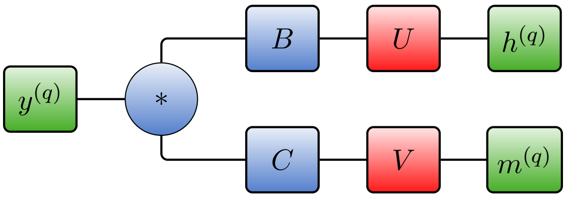


Figure 6: Blind matrix deconvolution: Measurement vectors in green, fixed operations in blue, and unknown signal in red.

see also Figure 6. The observations are given by the Q vectors $y^{(q)} = w^{(q)} * x^{(q)}$ or, equivalently, by

$$\begin{aligned} \hat{y}^{(q)} &= \hat{w}^{(q)} \circ \hat{x}^{(q)} \\ &= (FBUh^{(q)}) \circ (FCVm^{(q)}) , \end{aligned} \tag{54}$$

where $F_{j,k} := \mathrm{e}^{2\pi \mathrm{i}\,(j-1)(k-1)/N}/\sqrt{N}$ defines the Fourier transform F and $(a \circ b)_j := a_j b_j$ the Hadamard product of vector a and b. Let us denote the j-th rows of FB and FC by $\hat{b}_l$ and $\hat{c}_l$, respectively. Then

$$y_l^{(q)} = \hat{b}_l^T U h^{(q)}\, \hat{c}_l V m^{(q)} = \mathrm{Tr}(E_l\, U \rho^{(q)} V^T) \tag{55}$$

with the unit rank matrices $E_l := \hat{c}_l^T \hat{b}_l$ and $\rho^{(q)} := h^{(q)} m^{(q)^T}$.

Indeed, this is precisely a problem of the form discussed here,

$$y_l^{(q)} = \langle E_l^\dagger, M(\rho^{(q)}) \rangle \tag{56}$$

with $\mathcal{V} = \mathcal{W} = \mathbb{C}^n$ and

$$M(X) = UX\overline{V}. \tag{57}$$

Up to a phase, U and V can be trivially reconstructed from M up to phase. That is to say, a matrix version of blind deconvolution can readily be cast into the form of problems considered in this work. Numerically, we find a recovery from few samples and that the diamond norm reconstruction outperforms the nuclear norm based reconstruction from Ref. [15] adapted to our setting; see Figure 7. Many practical application of this problem are conceivable: The reconstruction of an unknown drift of a polarization degree of freedom in a channel problem is only one of the many natural ramifications of this setup.

6 Proofs

In this section, we prove Proposition 6 and an extension of Theorem 7. In order to do so, we first define a generalization of the sign matrix to matrices that are not necessarily Hermitian. This will give rise to the left and right absolute values of arbitrary matrices. Then we introduce SDPs, complementary slackness, and state the SDP for the square norm in standard form. Combining all these concepts, this section cumulates in the proofs of Proposition 6 and Theorem 7.

6.1 Auxiliary statements

The singular value decomposition of a matrix $X \in \mathrm{L}(\mathbb{C}^n)$ is

$$X = U\Sigma V^\dagger, \tag{58}$$

where $U, V \in \mathrm{U}(n)$ are unitaries and $\Sigma \in \mathrm{Pos}(\mathbb{C}^n)$ is positive-semidefinite and diagonal. This decomposition allows one to define a “sign matrix” of X:

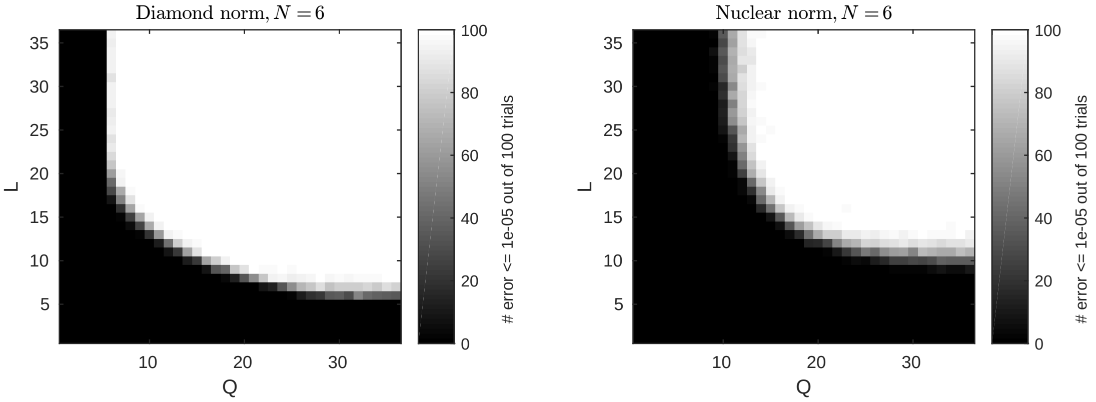


Figure 7: Blind matrix deconvolution with $N = 6$ and $\epsilon = 0$. The plots show the number of trials out of 100 with small errors, $\left\|M^{\diamond}_{\text{eps}} - M_0\right\|_{\text{F}} \leq 10^{-5}$ and $\left\|M^{*}_{\text{eps}} - M_0\right\|_{\text{F}} \leq 10^{-5}$, respectively, and with η chosen as machine precision eps.
Left: Recovery via diamond norm. **Right:** Recovery via nuclear norm.
The diamond norm recovery works with fewer measurements than the nuclear norm recovery, while the computation time is approximately the same.

Definition 13 (Sign matrix)**.** *For any matrix $X \in \mathrm{L}(\mathbb{C}^n)$ with singular value decomposition* (58) *we define its* sign matrix *to be $S_X := VU^\dagger$.*

Note that the sign matrix is in general not unique, but always unitary and it obeys

$$XS_X = U\Sigma U^\dagger = \sqrt{XX^\dagger}, \tag{59}$$

$$X^\dagger S_X^\dagger = V\Sigma V^\dagger = \sqrt{X^\dagger X}. \tag{60}$$

Therefore, S_X indeed generalizes the sign-matrix $\operatorname{sign}(X)$ (which is defined exclusively for Hermitian matrices) upon right multiplication.

The following auxiliary statement will be required later on and follows from a Schur complement rule.

Lemma 14. *For every $A \in \mathrm{L}(\mathcal{V} \to \mathcal{W})$, one has*

$$\begin{pmatrix} \|A\| \, \mathbb{1}_{\mathcal{W}} & \pm A \\ \pm A^\dagger & \|A\| \, \mathbb{1}_{\mathcal{V}} \end{pmatrix} \succeq 0 \,. \tag{61}$$

6.2 Semidefinite programming

Semidefinite programs (SDPs) are a class of optimization problems that can be evaluated efficiently, e.g. by using CVX [55, 56].

Definition 15 (Semidefinite program)**.** *A* semidefinite program *is specified by a triple (Ξ, C, D), where $C \in \mathrm{Herm}(\mathcal{V})$ and $D \in \mathrm{Herm}(\mathcal{W})$ are self-adjoint operators and $\Xi : \mathrm{L}(\mathcal{V}) \to \mathrm{L}(\mathcal{W})$ is a Hermiticity preserving linear map. With such a triple, one associates* *a pair of optimization problems:*

$$\begin{aligned}
\textbf{Primal:}\quad & \text{maximize} && \operatorname{Tr}(CZ) && (62)\\
& \text{subject to} && \Xi(Z) = D\,, && (63)\\
& && Z \succeq 0\,, && (64)\\
\textbf{Dual:}\quad & \text{minimize} && \operatorname{Tr}(DY) && (65)\\
& \text{subject to} && \Xi^\dagger(Y) \succeq C\,, && (66)\\
& && Y \in \operatorname{Herm}(\mathcal{W})\,. && (67)
\end{aligned}$$

$Z^\sharp \in \operatorname{Herm}(\mathcal{V})$ *is called* primal feasible *if it satisfies Eq.* (63) *and Eq.* (64)*. It is called* optimal primal feasible *if, additionally, for* $Z = Z^\sharp$ *in Eq.* (62) *the maximum is attained. Similarly,* $Y^\sharp \in \operatorname{Herm}(\mathcal{W})$ *is called* dual feasible *if it satisfies Eq.* (66) *and* optimal dual feasible, *if for* $Y = Y^\sharp$ *the minimum in Eq.* (65) *is attained.*

SDPs that exactly reproduce the problem structure outlined in this definition are said to be in *standard form*. But for specific SDPs, equivalent formulations might often be more handy.

Weak duality refers to the fact that the value of the primal SDP cannot be larger than the value of the dual SDP, i.e., that $\operatorname{Tr}(CZ) \leq \operatorname{Tr}(DY)$ for any primal feasible point Z and dual feasible point Y. An SDP is said to satisfy *strong duality* if the optimal values coincide, i.e., if for some optimal primal feasible and dual feasible points $Z^\sharp$ and $Y^\sharp$ it hols that $\operatorname{Tr}(CZ^\sharp) = \operatorname{Tr}(DY^\sharp)$. In fact, from a weak condition, called Slater's condition, strong duality follows.

Lemma 16 (Complementary slackness [39])**.** *Suppose that* (Ξ, C, D) *characterizes an SDP that obeys strong duality and let* $Z^\sharp \in \operatorname{Herm}(\mathcal{V})$ *and* $Y^\sharp \in \operatorname{Herm}(\mathcal{W})$ *denote optimal primal and dual feasible points, respectively (i.e.* $\operatorname{Tr}\left(CZ^\sharp\right) = \operatorname{Tr}\left(DY^\sharp\right)$*). Then*

$$\Xi^\dagger(Y^\sharp)Z^\sharp = CZ^\sharp \quad \textit{and} \quad \Xi(Z^\sharp)Y^\sharp = DY^\sharp. \tag{68}$$

The following, somewhat exhaustive, classification of the square norm's SDP will be instrumental later on.

Lemma 17 (SDP for the diamond norm in standard form [25])**.**
Let $X \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V}))$ *be a bipartite operator. Then its square norm* $\|X\|_\square$ *can be evaluated by means of an SDP* (Ξ, C, D) *that satisfies strong duality. In standard form, it is given by the block-wise defined matrices*

$$C = \frac{\dim(\mathcal{V})}{2}\begin{pmatrix} \mathbb{0}_\mathcal{V} & \mathbb{0}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0}_\mathcal{V} & \mathbb{0}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0} & \mathbb{0} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & X \\ \mathbb{0} & \mathbb{0} & X^\dagger & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \end{pmatrix} \in \operatorname{Herm}\left(\mathcal{V} \oplus \mathcal{V} \oplus (\mathcal{W}\otimes\mathcal{V}) \oplus (\mathcal{W}\otimes\mathcal{V})\right), \tag{69}$$

$$D = \begin{pmatrix} 1 & 0 & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} \\ 0 & 1 & 0^T_{\mathcal{W}\otimes\mathcal{V}} & 0^T_{\mathcal{W}\otimes\mathcal{V}} \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \end{pmatrix} \in \operatorname{Herm}\left(\mathbb{C} \oplus \mathbb{C} \oplus (\mathcal{W}\otimes\mathcal{V}) \oplus (\mathcal{W}\otimes\mathcal{V})\right)\,, \tag{70}$$

where $0_{\mathcal{W}\otimes\mathcal{V}} \in \mathcal{W}\otimes\mathcal{V}$ *denotes the zero-vector, and* $\mathbb{0} \in \mathrm{L}(\mathcal{V} \to \mathcal{W}\otimes\mathcal{V})$*, as well as* $\mathbb{0}_\mathcal{V} \in \mathrm{L}(\mathcal{V})$ *represent zero matrices of appropriate dimension. Finally, the map*

$$\Xi : \operatorname{Herm}\left((\mathcal{V} \oplus \mathcal{V} \oplus (\mathcal{W}\otimes\mathcal{V}) \oplus (\mathcal{W}\otimes\mathcal{V})\right) \to \operatorname{Herm}\left(\mathbb{C} \oplus \mathbb{C} \oplus (\mathcal{W}\otimes\mathcal{V}) \oplus (\mathcal{W}\oplus\mathcal{V})\right) \tag{71}$$

acts as

$$\Xi\begin{pmatrix} W_0 & \cdot & \cdot & \cdot \\ \cdot & W_1 & \cdot & \cdot \\ \cdot & \cdot & Z_0 & \cdot \\ \cdot & \cdot & \cdot & Z_1 \end{pmatrix} = \begin{pmatrix} \mathrm{Tr}(W_0) & 0 & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} \\ 0 & \mathrm{Tr}(W_1) & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & Z_0 - \mathbb{1}_{\mathcal{W}}\otimes W_0 & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & Z_1 - \mathbb{1}_{\mathcal{W}}\otimes W_1 \end{pmatrix} \tag{72}$$

and has an adjoint map given by

$$\Xi^\dagger\begin{pmatrix} \lambda_0 & \cdot & \cdot & \cdot \\ \cdot & \lambda_1 & \cdot & \cdot \\ \cdot & \cdot & Y_0 & \cdot \\ \cdot & \cdot & \cdot & Y_1 \end{pmatrix} = \begin{pmatrix} \lambda_0\mathbb{1}_{\mathcal{V}} - \mathrm{Tr}_{\mathcal{W}}(Y_0) & \mathbb{0}_{\mathcal{V}} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0}_{\mathcal{V}} & \lambda_1\mathbb{1}_{\mathcal{V}} - \mathrm{Tr}_{\mathcal{W}}(Y_1) & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0} & \mathbb{0} & Y_0 & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \\ \mathbb{0} & \mathbb{0} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & Y_1 \end{pmatrix}, \tag{73}$$

where $\mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \in \mathrm{L}(\mathcal{W}\otimes\mathcal{V})$, once more, represents a zero-matrix.

Lemma 17 presents an SDP for the square norm in standard form. Although this standard form is going to be important for our proofs, it is somewhat unwieldy. Fortunately, elementary modifications [25] allow to reduce the SDP to the following pair.

$$\begin{aligned} \textbf{Primal:} \qquad \|X\|_{\square} = \quad & \max && \tfrac{1}{2}\mathrm{Tr}(XZ) + \tfrac{1}{2}\mathrm{Tr}(X^\dagger Z^\dagger) \\ & \text{subject to} && \begin{pmatrix} \mathbb{1}_{\mathcal{W}}\otimes\rho & Z \\ Z^\dagger & \mathbb{1}_{\mathcal{W}}\otimes\sigma \end{pmatrix} \succeq 0\,, \\ &&& \mathrm{Tr}(\rho) = \mathrm{Tr}(\sigma) = \dim(\mathcal{V})\,, \\ &&& \rho,\sigma \in \mathrm{Pos}(\mathcal{V})\,, \\ &&& Z \in \mathrm{L}(\mathcal{W}\otimes\mathcal{V}) \end{aligned} \tag{74}$$

$$\begin{aligned} \textbf{Dual:} \qquad \|X\|_{\square} = \quad & \min && \tfrac{\dim(\mathcal{V})}{2}\left(\|\mathrm{Tr}_{\mathcal{W}}(Y)\| + \|\mathrm{Tr}_{\mathcal{W}}(Z)\|\right) \\ & \text{subject to} && \begin{pmatrix} Y & -X \\ -X^\dagger & Z \end{pmatrix} \succeq 0\,, \\ &&& Y, Z \in \mathrm{Pos}(\mathcal{W}\otimes\mathcal{V})\,. \end{aligned} \tag{75}$$

This simplified SDP pair for the square norm comes in handy for establishing the final claim in Proposition 6. For Hermitian matrices, the first two bounds presented there were already established in Ref. [57, Lemma 7]. Here, we show that an analogous strategy remains valid for matrices that need not be Hermitian.

Proof of Proposition 6. Let us start with recalling the variational definition (7) of the square norm:

$$\|X\|_{\square} := \max\{\|(\mathbb{1}_{\mathcal{W}}\otimes A)X(\mathbb{1}_{\mathcal{W}}\otimes B)\|_* : A, B \in \mathrm{L}(\mathcal{V}),\ \|A\|_{\mathrm{F}} = \|B\|_{\mathrm{F}} = \sqrt{\dim(\mathcal{V})}\}.$$

As already mentioned, inserting $A = B = \mathbb{1}$ into Eq. (7) establishes the lower bound (20) ($\|X\|_* \leq \|X\|_{\square}$). Also, a generalized version of Hölder's inequality assures

$$\|(\mathbb{1}_{\mathcal{W}}\otimes A)X(\mathbb{1}_{\mathcal{W}}\otimes B)\|_* \leq \|\mathbb{1}_{\mathcal{W}}\otimes A\|\,\|X\|_*\,\|\mathbb{1}_{\mathcal{W}}\otimes B\| \leq \|A\|_{\mathrm{F}}\,\|B\|_{\mathrm{F}}\,\|X\|_* \tag{76}$$

for any $A, B \in \mathrm{L}(\mathcal{V})$ and $X \in \mathrm{L}(\mathcal{W}\otimes\mathcal{V})$. Inserting this bound into the variational definition of $\|X\|_{\square}$ results in $\|X\|_{\square} \leq \dim(\mathcal{V})\,\|X\|_*$, which is the second bound.

For the final bound ($\|X\|_{\square} \leq \dim(\mathcal{W}\otimes\mathcal{V})\,\|X\|$) we consider the simplified version of the square norm's dual SDP (75). Lemma 14 assures that setting $Y = Z = \|X\|\,\mathbb{1}_{\mathcal{W}\otimes\mathcal{V}}$ results in a feasible point of this program. Inserting this point into the objective function yields a value of $\dim(\mathcal{W})\dim(\mathcal{V})\,\|X\|$, because $\|\mathrm{Tr}_{\mathcal{W}}(\mathbb{1}_{\mathcal{W}\otimes\mathcal{V}})\| = \|\dim(\mathcal{W})\mathbb{1}_{\mathcal{V}}\| = \dim(\mathcal{W})$. The bound follows from this value and the structure of the optimization problem (75). □

6.3 Proof of Theorem 7

In this section, we prove an extension of Theorem 7. In particular, this more general result relates Theorem 7 to optimal feasible points in Watrous' SDP from Lemma 17. These will contain the generalizations of the sign matrix from Definition 13.

Theorem 18 (Extremal operators as optimal feasible points)**.** *Let $X \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$ be a bipartite operator and set $n := \dim(\mathcal{V})$. Then the points (i)–(v) are equivalent:*

(i) X satisfies

$$\|X\|_{\square} = \|X\|_{*}\,, \tag{77}$$

(ii) Some $Z^{\sharp} \in \mathrm{Herm}((\mathcal{V} \oplus \mathcal{V} \oplus (\mathcal{W} \otimes \mathcal{V}) \oplus (\mathcal{W} \otimes \mathcal{V}))$ of the form

$$Z^{\sharp} := \frac{1}{n}\begin{pmatrix} \mathbb{1}_{\mathcal{V}} & \mathbb{0}_{\mathcal{V}} & \mathbb{0}^{\dagger} & \mathbb{0}^{\dagger} \\ \mathbb{0}_{\mathcal{V}} & \mathbb{1}_{\mathcal{V}} & \mathbb{0}^{\dagger} & \mathbb{0}^{\dagger} \\ \mathbb{0} & \mathbb{0} & \mathbb{1}_{\mathcal{W}\otimes\mathcal{V}} & S_X^{\dagger} \\ \mathbb{0} & \mathbb{0} & S_X & \mathbb{1}_{\mathcal{W}\otimes\mathcal{V}} \end{pmatrix} \tag{78}$$

is a primal optimal feasible point for Watrous' SDP (Ξ, C, D) from Lemma 17.

(iii) Some $Y^{\sharp} \in \mathrm{Herm}\left(\mathbb{C} \oplus \mathbb{C} \oplus (\mathcal{W} \otimes \mathcal{V}) \oplus (\mathcal{W} \oplus \mathcal{V})\right)$ of the form

$$Y^{\sharp} = \frac{1}{2}\begin{pmatrix} \|X\|_{*} & \cdot & \cdot & \cdot \\ \cdot & \|X\|_{*} & \cdot & \cdot \\ \cdot & \cdot & n\sqrt{XX^{\dagger}} & \cdot \\ \cdot & \cdot & \cdot & n\sqrt{X^{\dagger}X} \end{pmatrix}. \tag{79}$$

is a dual optimal feasible point for Watrous' SDP (Ξ, C, D) from Lemma 17.

(iv) X satisfies

$$\mathrm{Tr}_{\mathcal{W}}\left(\sqrt{XX^{\dagger}}\right) \propto \mathrm{Tr}_{\mathcal{W}}\left(\sqrt{X^{\dagger}X}\right) \propto \mathbb{1}_{\mathcal{V}}\,. \tag{80}$$

(v) X satisfies

$$\mathrm{Tr}_{\mathcal{W}}\left(\sqrt{XX^{\dagger}}\right) = \mathrm{Tr}_{\mathcal{W}}\left(\sqrt{X^{\dagger}X}\right) = \frac{\|X\|_{*}}{n}\,\mathbb{1}_{\mathcal{V}}\,. \tag{81}$$

Similar to the actual SDP, the optimal feasible points presented in Theorem 18 have simplified counterparts that correspond to optimal feasible points of the simplified SDPs (74) and (75). For the sake of completeness, we present them in the following corollary.

Corollary 19. *For any $X \in \mathrm{L}(\mathcal{W} \otimes \mathcal{V})$, optimal feasible points of the primal SDP (74) and the dual SDP (75) for the square norm are given by the following.*

$$\textit{Primal optimal feasible point:} \quad Z = S_X, \rho = \sigma = \mathbb{1}_{\mathcal{V}} \quad \textit{and} \tag{82}$$

$$\textit{Dual optimal feasible point:} \quad Y = \sqrt{XX^{\dagger}},\ Z = \sqrt{X^{\dagger}X}\,. \tag{83}$$

This statement follows straightforwardly from Theorem 18 by considering the reduced formulations (74) and (75) of the SDP from Lemma 17.

Proof of Theorem 18. For $X = 0$ all statements are evident. From now on, we assume that $X \neq 0$, or, equivalently, $X \neq 0$.

Proof of (i) ⇒ (ii). Note that $Z^\sharp \succeq 0$ by Lemma 14. Straightforward evaluation of $\Xi(Z^\sharp)$ from Lemma 17 reveals that $Z^\sharp$ is indeed a primal feasible point:

$$\begin{aligned}\Xi(Z^\sharp) &= \begin{pmatrix} \frac{1}{n}\operatorname{Tr}(\mathbb{1}_\mathcal{V}) & 0 & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} \\ 0 & \frac{1}{n}\operatorname{Tr}(\mathbb{1}_\mathcal{V}) & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & \frac{1}{n}\mathbb{1}_{\mathcal{W}\otimes\mathcal{V}} - \mathbb{1}_\mathcal{W}\otimes\frac{1}{n}\mathbb{1}_\mathcal{V} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & \frac{1}{n}\mathbb{1}_{\mathcal{W}\otimes\mathcal{V}} - \mathbb{1}_\mathcal{W}\otimes\frac{1}{n}\mathbb{1}_\mathcal{V} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0^\dagger_{\mathcal{W}\otimes\mathcal{W}} & 0^\dagger_{\mathcal{W}\otimes\mathcal{W}} \\ 0 & 1 & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} & 0^\dagger_{\mathcal{W}\otimes\mathcal{V}} \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \end{pmatrix} = D\,. \end{aligned} \tag{84}$$

In order to show optimality, we evaluate the primal SDP's objective function given by C in Eq. (69). Employing formulas (59) and (60) to express the absolute values of X, we obtain

$$\begin{aligned}\operatorname{Tr}\left(CZ^\sharp\right) &= \frac{n}{2}\operatorname{Tr}\left(\begin{pmatrix} \mathbb{0}_\mathcal{V} & \mathbb{0}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0}_\mathcal{V} & \mathbb{0}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0} & \mathbb{0} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & X \\ \mathbb{0} & \mathbb{0} & X^\dagger & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \end{pmatrix} \frac{1}{n} \begin{pmatrix} \mathbb{1}_\mathcal{V} & \mathbb{0}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0}_\mathcal{V} & \mathbb{1}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0} & \mathbb{0} & \mathbb{1}_{\mathcal{W}\otimes\mathcal{V}} & S_X^\dagger \\ \mathbb{0} & \mathbb{0} & S_X & \mathbb{1}_{\mathcal{W}\otimes\mathcal{V}} \end{pmatrix}\right) \\ &= \frac{1}{2}\left(\operatorname{Tr}\left(XS_X\right) + \operatorname{Tr}\left(X^\dagger S_X^\dagger\right)\right) \\ &= \frac{1}{2}\left(\operatorname{Tr}\left(\sqrt{XX^\dagger}\right) + \operatorname{Tr}\left(\sqrt{X^\dagger X}\right)\right) \\ &= \frac{1}{2}\left(\|X\|_* + \|X\|_*\right) = \|X\|_*\,. \end{aligned} \tag{85}$$

By assumption (77), this is indeed optimal.

Proof of (ii) ⇒ (iii) and (iv): Strong duality of Watrous' SDP from Lemma 17 assures that an optimal dual solution $Y^\sharp$ exists and that complementary slackness holds. Since $\Xi^\dagger$ from Eq. (73) does not depend on block off-diagonal terms, optimal feasibility only depends on the block diagonal parts. Hence, we write $Y^\sharp$ as

$$Y^\sharp = \begin{pmatrix} \lambda_0 & \cdot & \cdot & \cdot \\ \cdot & \lambda_1 & \cdot & \cdot \\ \cdot & \cdot & Y_0 & \cdot \\ \cdot & \cdot & \cdot & Y_1 \end{pmatrix}. \tag{86}$$

Complementary slackness (Lemma 16) implies that

$$\begin{aligned}\Xi^\dagger(Y^\sharp)Z^\sharp &= \frac{1}{n}\begin{pmatrix} \lambda_0\mathbb{1}_\mathcal{V} - \operatorname{Tr}_\mathcal{W}(Y_0) & \mathbb{0}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0}_\mathcal{V} & \lambda_1\mathbb{1}_\mathcal{V} - \operatorname{Tr}_\mathcal{W}(Y_1) & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0} & \mathbb{0} & Y_0 & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \\ \mathbb{0} & \mathbb{0} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & Y_1 \end{pmatrix} \begin{pmatrix} \mathbb{1}_\mathcal{V} & \mathbb{0}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0}_\mathcal{V} & \mathbb{1}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0} & \mathbb{0} & \mathbb{1}_{\mathcal{W}\otimes\mathcal{V}} & S_X^\dagger \\ \mathbb{0} & \mathbb{0} & S_X & \mathbb{1}_{\mathcal{W}\otimes\mathcal{V}} \end{pmatrix} \\ &= \frac{1}{n}\begin{pmatrix} \lambda_0\mathbb{1}_\mathcal{V} - \operatorname{Tr}_\mathcal{W}(Y_0) & \mathbb{0}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0}_\mathcal{V} & \lambda_1\mathbb{1}_\mathcal{V} - \operatorname{Tr}_\mathcal{W}(Y_1) & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0} & \mathbb{0} & Y_0 & Y_0\,S_X^\dagger \\ \mathbb{0} & \mathbb{0} & Y_1\,S_X & Y_1 \end{pmatrix} \end{aligned} \tag{87}$$

and

$$CZ^\sharp = \frac{1}{2}\begin{pmatrix} \mathbb{0}_\mathcal{V} & \mathbb{0}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0}_\mathcal{V} & \mathbb{0}_\mathcal{V} & \mathbb{0}^\dagger & \mathbb{0}^\dagger \\ \mathbb{0} & \mathbb{0} & XS_X & X \\ \mathbb{0} & \mathbb{0} & X^\dagger & X^\dagger S_X^\dagger \end{pmatrix} \tag{88}$$

must equal each other. This in turn demands

$$Y_0 = \frac{n}{2} X S_X = \frac{n}{2}\sqrt{XX^\dagger} \quad \text{as well as} \tag{89}$$

$$Y_1 = \frac{n}{2} X^\dagger S_X^\dagger = \frac{n}{2}\sqrt{X^\dagger X}, \tag{90}$$

where we have once more employed identities (59) and (60) for S_X to obtain the absolute values of X. Equality of (87) and (88) in the first two diagonal entries (also guaranteed by complementary slackness) furthermore assures

$$\lambda_0 \mathbb{1}_\mathcal{V} - \mathrm{Tr}_\mathcal{W}\left(Y_0\right) = \lambda_0 \mathbb{1}_\mathcal{V} - \frac{n}{2}\mathrm{Tr}_\mathcal{W}\left(\sqrt{XX^\dagger}\right) = \mathbb{0}_\mathcal{V} \quad \text{and} \tag{91}$$

$$\lambda_1 \mathbb{1}_\mathcal{V} - \mathrm{Tr}_\mathcal{W}\left(Y_1\right) = \lambda_0 \mathbb{1}_\mathcal{V} - \frac{n}{2}\mathrm{Tr}_\mathcal{W}\left(\sqrt{X^\dagger X}\right) = \mathbb{0}_\mathcal{V}. \tag{92}$$

Hence,

$$\lambda_0\, n = \frac{n}{2}\left\|\mathrm{Tr}_\mathcal{W}(Y_0)\right\|_* = \frac{n}{2}\left\|X\right\|_* \quad \text{and} \tag{93}$$

$$\lambda_1\, n = \frac{n}{2}\left\|\mathrm{Tr}_\mathcal{W}(Y_1)\right\|_* = \frac{n}{2}\left\|X\right\|_* \tag{94}$$

and both, (iii) and (iv) follow.

Proof of (iv) ⇒ (v): Let $c_1, c_2 > 0$ be constants such that

$$\mathrm{Tr}_\mathcal{W}\left(\sqrt{XX^\dagger}\right) = c_1 \mathbb{1}_\mathcal{V} \quad \text{and} \tag{95}$$

$$\mathrm{Tr}_\mathcal{W}\left(\sqrt{X^\dagger X}\right) = c_2 \mathbb{1}_\mathcal{V}. \tag{96}$$

Taking the trace of both equations and recognizing the nuclear norm reveals that

$$\left\|X\right\|_* = \mathrm{Tr}\left(\sqrt{XX^\dagger}\right) = \mathrm{Tr}\left(\mathrm{Tr}_\mathcal{W}\left(\sqrt{XX^\dagger}\right)\right) \left\|X\right\|_* = c_1\,\mathrm{Tr}\left(\mathbb{1}_\mathcal{V}\right) = c_1\, n \tag{97}$$

and, similarly,

$$\left\|X\right\|_* = c_2\, n\,, \tag{98}$$

which proves the claimed implication.

Proof of (v) ⇒ (i): The crucial observation for this implication is that Assumption (v) alone assures that $Y^\sharp$ defined in Eq. (79) with all off-diagonal blocks set to zero is a feasible point of Watrous' dual SDP, albeit not necessarily an optimal one. This claim is easily verified by direct computation. Inserting this dual feasible point into the SDP's objective function results in

$$\begin{aligned}\mathrm{Tr}\left(DY^\sharp\right) =& \mathrm{Tr}\left(\begin{pmatrix} 1 & 0 & 0_{\mathcal{W}\otimes\mathcal{V}}^\dagger & 0_{\mathcal{W}\otimes\mathcal{V}}^\dagger \\ 0 & 1 & 0_{\mathcal{W}\otimes\mathcal{V}}^T & 0_{\mathcal{W}\otimes\mathcal{V}}^T \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \end{pmatrix} \frac{1}{2} \begin{pmatrix} \left\|X\right\|_* & 0 & 0_{\mathcal{W}\otimes\mathcal{V}}^\dagger & 0_{\mathcal{W}\otimes\mathcal{V}}^\dagger \\ 0 & \left\|X\right\|_* & 0_{\mathcal{W}\otimes\mathcal{V}}^T & 0_{\mathcal{W}\otimes\mathcal{V}}^T \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & n\sqrt{XX^\dagger} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} \\ 0_{\mathcal{W}\otimes\mathcal{V}} & 0_{\mathcal{W}\otimes\mathcal{V}} & \mathbb{0}_{\mathcal{W}\otimes\mathcal{V}} & n\sqrt{X^\dagger X} \end{pmatrix}\right) \\ =&\frac{1}{2}\left(\left\|X\right\|_* + \left\|X\right\|_*\right) = \left\|X\right\|_*\,. \end{aligned} \tag{99}$$

Since every dual SDP corresponds to a constrained minimization, evaluating the dual objective function at any feasible point results in an upper bound on the optimal value. In our case, obtain the upper bound $\left\|X\right\|_\square \leq \left\|X\right\|_*$, which together with the converse bound from Proposition 10, implies equality between the two.

□

7 Discussion and outlook

We conclude by mentioning several observations and research directions that may merit further attention.

Measurement errors. In our analysis we considered reconstructed matrices $X_\eta^\square$ and X_η^* from Eqs. (29) and (27) that are required to be η-close to the ideal operator X_0. Such a reconstruction stably tolerates additive errors ϵ as in Eq. (8) as long as they obey $\|\epsilon\|_{\mathrm{F}} \leq \eta$. For operators X_0 satisfying the extremality (19) we prove that recovery guarantees for X_η^* are inherited by $X_\eta^\square$. A similar situation is true for the reconstruction of maps M_0 by means of diamond norm minimization. For the idealized setting of noiseless measurements ($\epsilon = 0$), we demonstrate numerically that often $\left\|M_\eta^\diamond - M_0\right\|_{\mathrm{F}}$ vanishes while $\left\|M_\eta^* - M_0\right\|_{\mathrm{F}}$ is large. A numerical analysis for the noisy case $\epsilon > 0$ yields similar results as for $\epsilon = 0$. For the noisy case the phase transition from having no recovery to almost always recovering the signal up to $\eta \gtrsim \|\epsilon\|_{\mathrm{F}}$ broadens equally for both diamond and nuclear norm regularization.

Partial derandomizations. While initial theoretical results often rely on measurements that follow a Gaussian distribution, later on significant effort has been put into derandomizing the measurement process. On the one hand, recovery guarantees for structured measurements were proven [32]. On the other, also the distributions form which the measurements are drawn were partially derandomized [16, 17, 35] (see also Section 4), relying on above mentioned t-designs. The later methods rely on an analysis of the measurement map's descent cone. Hence, such recovery guarantees for partially derandomized measurements are also inherited by our reconstruction via diamond norm minimization. In a similar setting, a partial derandomization of the random unitaries used as part of the measurements for the retrieval of unitary basis changes (Section 5.2) and for quantum process tomography (Section 5.3) seems very promising. Here, structural insights [58–61] on unitary designs could be used in future work.

Improvement from structured measurements. We numerically performed the reconstruction of unitary basis changes in Section 5.2 for two different measurement settings: Gaussian measurements and certain structured measurements. For the nuclear norm, the reconstruction from Gaussian measurements performed slightly better than the one from structured measurements, just as expected. Perhaps surprisingly, we observed the converse for the diamond norm reconstructions. Here, the structure of the measurements seems to be favourable for the reconstruction process. This observation motivates the search for recovery guarantees for diamond norm reconstruction with structured measurements. Such structured measurements are also crucial for the quantum process tomography in Section 5.3 and blind matrix convolution in Section 5.4.

CPT as a constraint in the quantum channel reconstructions. A map $M \in \mathbb{L}(\mathcal{V}, \mathcal{W})$ is a quantum channel if and only if

$$M^\dagger(\mathbb{1}_\mathcal{W}) = \mathbb{1}_\mathcal{V} \quad \text{and} \quad J(M) \succeq 0\,. \tag{100}$$

When aiming at reconstructing quantum channels, these additional constraints can, in principle, be included in the SDPs (29) and (27) for the diamond norm and nuclear norm reconstructions. Doing so leads to a significant overhead in the numerical reconstruction process. Numerically, one can observe that the recovery success of the diamond norm reconstruction (29) is unchanged, while the nuclear norm reconstruction (27) performs significantly better. In fact, it seems to perform roughly as good as the diamond norm reconstruction when these constraints are included in the SDP (27). In this sense, the CPT structure can be used in the nuclear norm reconstruction at the expense of a longer computation time to reduce the number of measurements, while in the diamond norm reconstruction the CPT structure is already inbuilt. The run-time of the diamond norm reconstruction and the nuclear norm reconstruction are practically the same for a given number of measurements and scales polynomially with the number of constraints. Therefore, the diamond norm reconstruction can help to render larger quantum systems accessible to quantum process tomography.

8 Acknowledgements

We thank C. Riofrío, I. Roth, and D. Suess for insightful discussions and S. Kimmel and Y.K. Liu for advanced access to Ref. [51]. The research in Berlin has been supported by the EU (SIQS, AQuS, RAQUEL), the BMBF (Q.com), the DFG project EI 519/9-1 (SPP1798 CoSIP) and the ERC (TAQ). The work of DG and RK is supported by the Excellence Initiative of the German Federal and State Governments (Grants ZUK 43 & 81), the ARO under contract W911NF-14-1-0098 (Quantum Characterization, Verification, and Validation), the DFG project GRO 4334/2-1 (SPP1798 CoSIP), and the State Graduate Funding Program of Baden-Württemberg.

A Appendix

In this appendix we provide known material to make this work more self contained. We provide SDPs for the nuclear norm and the spectral norm, and introduction to tensor products and a basis independent definition of the Choi-Jamiołkowski isomorphism. Also, we devote a subsection to low-rank matrix recovery. There we show how the statements presented in Section 4 can be derived using geometric proof techniques. On the contrary to the other supplementary chapters, this section does include technical novelties.

A.1 Basic concepts of multilinear algebra and the Choi-Jamiołkowski isomorphism

The core objects of this work are tensors of order four and naturally fall into the realm of multilinear algebra. Here we give a brief introduction on core concepts of multilinear algebra that can be found in any textbook on that topic. Our presentation here is influenced by [62]. Let $\mathcal{V}_1, \ldots, \mathcal{V}_k$ be (finite dimensional, complex) vector spaces with associated dual spaces $\mathcal{V}_1^*, \ldots, \mathcal{V}_k^*$. A function

$$f : \mathcal{V}_1 \times \cdots \times \mathcal{V}_k \to \mathbb{C} \tag{101}$$

is *multilinear*, if it is linear in each $\mathcal{V}_i$. The space of such functions constitutes the *tensor product* of $\mathcal{V}_1^*, \ldots, \mathcal{V}_k^*$ and we denote it by $\mathcal{V}_1^* \otimes \cdots \otimes \mathcal{V}_k^*$. By reflexivity $\mathcal{V} \cong \mathcal{V}^{**}$, the tensor product $\mathcal{V}_1 \otimes \cdots \otimes \mathcal{V}_k$ is the space of all multilinear functions

$$f : \mathcal{V}_1^* \times \cdots \times \mathcal{V}_k^* \to \mathbb{C}. \tag{102}$$

Its elementary elements $z_1 \otimes \cdots \otimes z_k$ are the *tensor product* of vectors $x_1 \in \mathcal{V}_1, \ldots, x_k \in \mathcal{V}_k$ which alternatively can be constructed by means of the Kronecker product – however, such an explicit construction requires explicit choices of bases in $\mathcal{V}_1, \ldots, \mathcal{V}_k$.

With such a notation, the space of linear maps $\mathcal{V} \to \mathcal{W}$ (matrices) corresponds to the tensor product $\mathcal{W} \otimes \mathcal{V}^*$ which is spanned by rank-one operators $\{y \otimes x^* : x \in \mathcal{V}, y \in \mathcal{W}\}$. With this identification, it is straightforward to define the tensor product of $\mathrm{L}(\mathcal{W}_1 \to \mathcal{W}_2)$ and $\mathrm{L}(\mathcal{V}_1 \to \mathcal{V}_2)$ to be

$$\mathrm{L}(\mathcal{W}_1 \to \mathcal{W}_2) \otimes \mathrm{L}(\mathcal{V}_1 \to \mathcal{V}_2) \cong (\mathcal{W}_2 \otimes \mathcal{W}_1^*) \otimes (\mathcal{V}_2 \otimes \mathcal{V}_1^*) \cong \mathrm{L}(\mathcal{V}_1 \otimes \mathcal{W}_1 \to \mathcal{V}_2 \otimes \mathcal{W}_2)\,. \tag{103}$$

Analogously to before, the elementary $Y \otimes X$ of this space are the *tensor product* of maps $Y \in \mathrm{L}(\mathcal{W}_1 \to \mathcal{W}_2)$ and $X \in \mathrm{L}(\mathcal{V}_1 \to \mathcal{V}_2)$. Restricting to tensor products of endomorphisms, i.e. $\mathcal{W}_2 \cong \mathcal{W}_1$ and $\mathcal{V}_2 \cong \mathcal{V}_1$, the *partial trace* (over the first tensor factor) for elementary elements to be

$$\begin{aligned} \mathrm{Tr}_{\mathcal{W}} : \mathrm{L}(\mathcal{W}) \otimes \mathrm{L}(\mathcal{V}) &\to \mathrm{L}(\mathcal{W}) \\ Y \otimes X &\mapsto \mathrm{Tr}(X)\, Y \end{aligned} \tag{104}$$

and extend it linearly to $\mathrm{L}(\mathcal{W}) \otimes \mathrm{L}(\mathcal{V})$. Note that with the identification $\mathrm{L}(\mathcal{W}) \cong \mathcal{W} \otimes \mathcal{W}^*$, $\mathrm{Tr}_{\mathcal{W}}$ corresponds to the natural contraction between $\mathcal{W}$ and $\mathcal{W}^*$. This is illustrated in Figure 3.

Similarly to $\mathrm{L}(\mathcal{V}_1 \to \mathcal{V}_2)$, the maps $\mathrm{L}(\mathrm{L}(\mathcal{V}_1 \to \mathcal{V}_2) \to \mathrm{L}(\mathcal{W}_1 \to \mathcal{W}_2))$ introduced in Section 5.1 can be viewed as elements of the tensor product space

$$(\mathcal{W}_2 \otimes \mathcal{W}_1^*) \otimes (\mathcal{V}_2 \otimes \mathcal{V}_1^*)^* \cong \mathcal{W}_2 \otimes \mathcal{W}_1^* \otimes \mathcal{V}_2^* \otimes \mathcal{V}_1\,, \tag{105}$$

which can be seen as a four-linear vector space. There are several equivalent ways to interpret its elements. For the given applications of our work, we have made heavy use of the *Choi-Jamiołkowski isomorphism* which acts on four-linear tensors by permuting tensor factors:

$$\begin{aligned} J : \mathcal{V}_1 \otimes \mathcal{V}_2 \otimes \mathcal{V}_3 \otimes \mathcal{V}_4 &\to \mathcal{V}_1 \otimes \mathcal{V}_3 \otimes \mathcal{V}_2 \otimes \mathcal{V}_4 \\ v_1 \otimes v_2 \otimes v_3 \otimes v_4 &\mapsto v_1 \otimes v_3 \otimes v_2 \otimes v_4\,. \end{aligned} \tag{106}$$

Applied to the four-linear space of maps (105) we obtain

$$\begin{aligned} \mathrm{L}(\mathrm{L}(\mathcal{V}_1 \to \mathcal{V}_2) \to \mathrm{L}(\mathcal{W}_1 \to \mathcal{W}_2)) &\cong \mathrm{L}(\mathcal{V}_2 \otimes \mathcal{V}_1^* \to \mathcal{W}_2 \otimes \mathcal{W}_1^*) \\ &\cong \mathcal{W}_2 \otimes \mathcal{W}_1^* \otimes \mathcal{V}_2^* \otimes \mathcal{V}_1\,, \end{aligned} \tag{107}$$

and

$$\mathrm{L}(\mathcal{W}_1 \otimes \mathcal{V}_1^* \to \mathcal{W}_2 \otimes \mathcal{V}_2^*) \cong \mathcal{W}_2 \otimes \mathcal{V}_2^* \otimes \mathcal{W}_1^* \otimes \mathcal{V}_1 \tag{108}$$

which are basis independent. Consequently the Choi-Jamiołkowski isomorphism is linear bijection from maps to operators

$$J : \mathrm{L}(\mathrm{L}(\mathcal{V}_1 \to \mathcal{V}_2) \to \mathrm{L}(\mathcal{W}_1 \to \mathcal{W}_2)) \to \mathrm{L}(\mathcal{V}_1^* \otimes \mathcal{W}_1 \to \mathcal{V}_2^* \otimes \mathcal{W}_2)\,. \tag{109}$$

Its explicit definitions (37) and (38) in the main text are just basis-dependent realization of this more general identification. We illustrated this fact pictorially in Figure 3 by resorting to *tensor network* [63] or *wiring diagrams* [64].

A.2 Uniform recovery guarantees and partial derandomizations

Our main geometric insight – Corollary 8 – asserts that any square norm descent cone is always contained in the corresponding one of the nuclear norm, provided that the operators in question obey $\|X\|_{\square} = \|X\|_*$. When applying this idea to low-rank matrix recovery, we started with mentioning Proposition 10. This is a non-uniform recovery guarantee that is stable towards additive noise. However, with some additional work, Corollary 8 allows for stronger conclusions. Some of them are summarized in Proposition 11 and Proposition 12, respectively. Here, we outline how these results are obtained. In Section 2.2 we introduced widely used geometric proof techniques for low-rank matrix recovery mainly following Ref. [18]. These aim at recovery of a fixed object X_0 of interest and thus it suffices to focus on precisely one descent cone, namely $\mathscr{D}(X_0, \|\cdot\|_*)$, or $\mathscr{D}(X_0, \|\cdot\|_\diamond)$, respectively. By taking a closer look at the actual proof techniques – most notably Mendelson's small ball method [19], or Tropp's bowling scheme [18] – one can see that such a restriction to a single object of interest is not necessary. Up to our knowledge, this was first pointed out in Ref. [16] and at the heart of this observation is the following technical statement.

Lemma 20. *Fix $1 \leq r \leq n$ and let $K_r = \bigcup_X \mathscr{D}(\|\cdot\|_*, X) \subset \mathrm{L}(\mathcal{V})$ be the union of all descent cones anchored in nonzero matrices $0 \neq X \in \mathrm{L}(\mathcal{V})$ of rank at most r. Then, every element $Y \in K_r$ obeys*

$$\|Y\|_* \leq (1+\sqrt{2})\sqrt{r}\,\|Y\|_{\mathrm{F}} \,. \tag{110}$$

For Hermitian matrices, a slightly stronger statement of this type was presented in [16, Lemma 10]. Here wo provide a different proof that does not require Hermiticity and exploits a variant of pinching.

Lemma 21 (Pinching inequality)**.** *Let $P, Q \in \mathrm{L}(\mathcal{V})$ be orthogonal projectors with complements $P^\perp = \mathbb{1}_{\mathrm{L}(\mathcal{V})} - P$ and $Q^\perp = \mathbb{1}_{\mathrm{L}(\mathcal{V})} - Q$. Also, let $\|\cdot\|_p$ be any Schatten-p norm. Then, every $Z \in \mathrm{L}(\mathcal{V})$ obeys*

$$\|PZQ\|_p^p + \|P^\perp Z Q^\perp\|_p^p \leq \|Z\|_p^p \,. \tag{111}$$

Proof. Note that for any $Z \in \mathrm{L}(V)$ it follows from the definition of the Schatten-p norms that the left hand side of Eq. (111) coincides with $\|PZQ + P^\perp Z Q^\perp\|_p^p$. Using this identity and the decomposition

$$PZQ + P^\perp Z Q^\perp = \frac{1}{2}Z + \frac{1}{2}(P - P^\perp)Z(Q - Q^\perp) \tag{112}$$

allows us to conclude

$$\begin{aligned} \|PZQ + P^\perp Z Q^\perp\|_p^p =& \|\frac{1}{2}Z + \frac{1}{2}(P-P^\perp)Z(Q-Q^\perp)\|_p^p \\ \leq& \frac{1}{2}\|Z\|_p^p + \frac{1}{2}\|(P-P^\perp)Z(Q-Q^\perp)\|_p^p \\ =& \frac{1}{2}\|Z\|_p^p + \frac{1}{2}\|Z\|_p^p = \|Z\|_p^p \,, \end{aligned} \tag{113}$$

where we have exploited unitary invariance of Schatten-p norms and the fact that both $P - P^\perp$ and $Q - Q^\perp$ are unitary matrices. □

Proof of Lemma 20. It suffices to prove this statement for any fixed descent cone K_X, where $X \in \mathrm{L}(\mathcal{V})$ has rank at most r. Let $\mathcal{C} := \mathrm{ran}(X)$ and $\mathcal{R} := \mathrm{ran}(X^\dagger)$ be the column and row ranges of X (these need not coincide, since X need not necessarily be Hermitian) and let $P_\mathcal{C}, P_\mathcal{R} \in \mathrm{L}(\mathcal{V})$ be orthogonal projections onto these subspaces. Note that if X has a singular value decomposition $X = U\Sigma V^\dagger$, then $P_\mathcal{C} = U\Sigma^0 U^\dagger$ and $P_\mathcal{R} = V\Sigma^0 V^\dagger$, where Σ^0 is defined component-wise by $\Sigma^0_{i,j} := 1$ if $\Sigma_{i,j} \neq 0$ and $\Sigma^0_{i,j} := 0$ otherwise. Introducing orthogonal complements $P_\mathcal{C}^\perp = \mathbb{1}_{\mathcal{V}(\mathrm{L})} - P_\mathcal{C}$ and $P_\mathcal{R}^\perp = \mathbb{1}_{\mathrm{L}(\mathcal{V})} - P_\mathcal{R}$ allows us to define

$$\mathcal{P}_T^\perp : \mathrm{L}(\mathcal{V}) \to \mathrm{L}(\mathcal{V}), \quad Z \mapsto P_\mathcal{C}^\perp Z P_\mathcal{R}^\perp \,. \tag{114}$$

This is an orthogonal projection with respect to the Frobenius inner product (2) and obeys $\mathcal{P}_T^\perp X = 0$ by construction. Its complement amounts to

$$\mathcal{P}_T Z = Z - \mathcal{P}_U^\perp Z P_\mathcal{R}^\perp = P_\mathcal{C} Z + Z P_\mathcal{R} - P_\mathcal{C} Z P_\mathcal{R} \tag{115}$$

which obeys $\mathcal{P}_T X = X$. Note that this is a straightforward generalization of the T-space introduced in [13, Equation (2)] to non-Hermitian matrices. Analogously to there, a decomposition $Z = Z_T + Z_T^\perp := \mathcal{P}_T Z + \mathcal{P}_T^\perp Z$ is valid for every $Z \in \mathrm{L}(Z)$ and every $Z_T := \mathcal{P}_T Z$ has rank at most $2r$ by construction.

Now choose $Y \in K_X$ and note that by definition $\|X\|_* \geq \|X + \tau Y\|_*$ must be valid for some $\tau > 0$. Combining this with Lemma 21 (Pinching) assures

$$\begin{aligned} \|X\|_* &\geq \|X + \tau Y\|_* \geq \|P_\mathcal{C}(X+\tau Y)P_\mathcal{R}\|_* + \left\|P_\mathcal{C}^\perp(X+\tau Y)P_\mathcal{R}^\perp\right\|_* \\ &= \|X + \tau P_\mathcal{C} Y P_\mathcal{R}\|_* + \left\|\mathcal{P}_T^\perp(X+\tau Y)\right\|_* = \|X + \tau P_\mathcal{C} Y P_\mathcal{R}\|_* + \tau\left\|Y_T^\perp\right\|_*, \end{aligned} \tag{116}$$

where we have employed $P_\mathcal{C} X P_\mathcal{R} = X$ and $\mathcal{P}_T^\perp X = 0$. Also, note that Hölder's inequality assures $|\operatorname{Tr}(UZ)| \leq \|Z\|_*$ for any $Z \in \mathrm{L}(\mathcal{V})$ and unitary U. Employing this for $U = S_X$, where the sign matrix S_X of X was defined in Def. 13, reveals

$$\begin{aligned} \|X + \tau P_\mathcal{C} Y P_\mathcal{R}\|_* &\geq \operatorname{Tr}(S_X X) + \tau\ |\operatorname{Tr}(S_X P_\mathcal{C} Y P_\mathcal{R})| \geq \|X\|_* - \tau\,\|S_X\|\,\|P_\mathcal{C} Y P_\mathcal{R}\|_* \\ &\geq \|X\|_* - \tau\sqrt{r}\,\|P_\mathcal{C} Y P_\mathcal{R}\|_\mathrm{F} \geq \|X\|_* - \tau\sqrt{r}\,\|Y\|_\mathrm{F}, \end{aligned} \tag{117}$$

where we have in addition used that $P_\mathcal{C} Y P_\mathcal{R}$ has rank at most r and Frobenius norm smaller than or equal to $\|Y\|_\mathrm{F}$. Combining the bounds (116) and (117) implies

$$\|X\|_* \geq \|X\|_* + \tau\left(\left\|Y_T^\perp\right\|_* - \sqrt{r}\,\|Y\|_\mathrm{F}\right). \tag{118}$$

Since $\tau > 0$, this bound implies $\left\|Y_T^\perp\right\|_* \leq \sqrt{r}\,\|Y\|_\mathrm{F}$. Finally, this relation allows us to infer the result,

$$\begin{aligned} \|Y\|_* &= \left\|Y_T + Y_T^\perp\right\|_* \leq \|Y_T\|_* + \left\|Y_T^\perp\right\|_* \\ &\leq \sqrt{2r}\,\|Y_T\|_\mathrm{F} + \sqrt{r}\,\|Y\|_\mathrm{F} = (1+\sqrt{2})\sqrt{r}\,\|Y\|_\mathrm{F}\ , \end{aligned} \tag{119}$$

where we also exploited the fact that Y_T has rank at most $2r$.

□

Lemma 20 asserts that any matrix that lies in the nuclear norm's descent cone of any low-rank matrix, is "effectively" a low-rank matrix as well. This structural property together with Mendelson's small ball method is enough to bound the minimal conic singular value of a measurement map $\mathcal{A}$ with respect to the union of all possible descent cones. Here we provide a particular realization of Mendelson's small ball method that is directly applicable to low-rank matrix recovery (see e.g. Ref. [16, Section 4]).

Theorem 22 (A variant of Mendelson's small ball method)**.** *Let $\mathcal{L} \subset \mathrm{L}(\mathcal{V})$ be real subspace of linear maps and let $\mathcal{A} : \mathcal{L} \to \mathbb{R}^m$ be a measurement map $\mathcal{A}(X) = \sum_{i=1}^m \operatorname{Tr}(A_i X)\, e_i$, where each A_i is an independent copy of a random matrix $A \in \mathrm{L}(\mathcal{V})$ and $e_1, \ldots, e_m$ denotes the standard basis in $\mathbb{R}^m$. Also, let $E_r = \{Y \in K_r :\ \|Y\|_\mathrm{F} = 1\}$, where K_r was defined in Lemma 20. Then for any $\xi, t > 0$, the bound*

$$\lambda_{\min}(\mathcal{A}, K_r) \geq \xi\sqrt{m}Q_{2\xi}(E_r; A) - 2W_m(E_r, \mathcal{A}) - \xi t \tag{120}$$

holds with probability at least $1-\mathrm{e}^{-2t^2}$. *Here*

$$Q_\xi\left(E_r, A\right) = \inf_{Y\in E_r} \Pr\left[\left|\mathrm{Tr}(A^\dagger Y)\right| \geq \xi\right] \tag{121}$$

$$W_m\left(E_r, \mathcal{A}\right) = \mathbb{E}\left[\sup_{Y\in E_r} \mathrm{Tr}(H^\dagger Y)\right], \quad \textit{where} \quad H = \frac{1}{\sqrt{m}}\sum_{j=1}^{m} \epsilon_j A_j \tag{122}$$

and $\epsilon_1,\ldots,\epsilon_m$ *being a Rademacher sequence*[1].

Important examples for the space of considered operators are $\mathcal{L} = \mathrm{Herm}(\mathcal{V})$ and real matrices.

Thanks to Lemma 20 and Hölder's inequality we can bound $W_m\left(E_r, \mathcal{A}\right)$ in Theorem 22 by

$$\begin{aligned} W_m\left(E_r, \mathcal{A}\right) = \mathbb{E}\left[\sup_{Y\in E_r} \mathrm{Tr}\left(H^\dagger Y\right)\right] &\leq \mathbb{E}\left[\sup_{Y\in E_r} \|Y\|_* \left\|H^\dagger\right\|\right] \\ &\leq \mathbb{E}\left[\sup_{Y\in E_r} (1+\sqrt{2})\sqrt{r}\,\|Y\|_{\mathrm{F}}\,\|H\|\right] = (1+\sqrt{2})\sqrt{r}\,\mathbb{E}\left[\|H\|\right], \end{aligned} \tag{123}$$

which is much easier to handle. This simplification together with Mendelson's small ball method – Theorem 22 – and the geometric error bound for convex recovery – Proposition 3 – provide a convenient sufficient means to assure that a given measurement process $\mathcal{A}$ allows for uniform and stable low-rank matrix recovery via nuclear norm minimization:

Proposition 23 (Sufficient criteria for uniform recovery)**.** *Let* $\mathcal{A} : \mathrm{L}(\mathcal{V}) \to \mathbb{C}^m$ *be a measurement map as defined in Theorem 22 and fix* $1 \leq r \leq n$. *Suppose that this measurement map obeys* $Q_{2\xi}(E_r; \mathcal{A}) \geq C_1$ *for some* $\xi > 0$ *and also* $\mathbb{E}\left[\|H\|\right] \leq C_2\sqrt{m/r}$, *where* C_1 *and* C_2 *are positive constants obeying* $\xi C_1 > 2(1+\sqrt{2})C_2$.

Then, with probability at least $1-\mathrm{e}^{-C_4^* m}$, *this measurement map is capable of stably reconstructing any matrix* X_0 *of rank at most* r *from noisy measurements of the form* $y = \mathcal{A}(X_0) + \epsilon$ *obeying* $\|\epsilon\|_{\mathrm{F}} \leq \eta$ *by means of nuclear norm minimization. Concretely, the solution* X_η^* *of the optimization* (27) *obeys*

$$\left\|X_\eta^* - X_0\right\|_{\mathrm{F}} \leq \frac{\eta}{C_3^*\sqrt{m}}. \tag{124}$$

Here $C_3^*, C_4^* > 0$ *denote sufficiently small absolute constants.*

Note that unlike Proposition 10, such a recovery statement is *uniform*, in the sense that with high probability a single measurement map suffices to recover any low-rank matrix. However, it still relies on the geometric proof technique of bounding the widths of nuclear norm descent cones. This is because the set K_r is just the union over all possible nuclear norm descent cones anchored at matrices of rank at most r. As a result, Observation 4 ("the smaller the descent cone, the better the recovery") is also valid in this setting and Corollary 8 allows us to draw the following conclusion.

Corollary 24 (Uniform recovery from square norm regularization)**.** *The assertions of Proposition 23 remain true for recovery via square norm regularization* (29), *for the case of uniform recovery of rank-r maps* $X_0 \in \mathrm{L}\left(\mathcal{V}\otimes\mathcal{W}\right)$ *satisfying* $\|X_0\|_\square = \|X_0\|_*$. *Moreover, the corresponding constants obey* $C_3^\square \geq C_3^*$ *and* $C_4^\square \geq C_3^*$, *meaning that the recovery statement cannot be worse.*

[1] A Rademacher sequence is a sequence of independent random variables that take the values ±1 with equal probability.

Proof of Proposition 23. Theorem 22 together with Eq. (123) and the assumptions on $\mathcal{A}$ assure for any $t>0$

$$\begin{aligned}\lambda_{\min}\left(\mathcal{A}, K_r\right) \geq & \xi\sqrt{m}Q_{2\xi}(E_r;\mathcal{A}) - 2W_m(E_r,\mathcal{A}) - \xi t \\ \geq & \xi\sqrt{m}Q_{2\xi}(E_r;\mathcal{A}) - 2(1+\sqrt{2})\sqrt{r}\mathbb{E}\left[\|H\|\right] - \xi t \\ \geq & \xi C_1\sqrt{m} - 2(1+\sqrt{2})C_2\sqrt{m} - \xi t\end{aligned} \tag{125}$$

with probability at least $1-\mathrm{e}^{-2t^2}$. Introducing $C_3 = (\xi C_1 - 2(1+\sqrt{2})C_2)/2$ – which is strictly positive by assumption – and setting $t = C_3\sqrt{m}/\xi$ then implies

$$\lambda_{\min}\left(\mathcal{A}, K_r\right) \geq C_3\sqrt{m} \tag{126}$$

with probability at least $1-\mathrm{e}^{-C_4 m}$, where $C_4 = C_3^2/\xi^2 > 0$. With such an estimate at hand, the claim follows from applying Proposition 3. □

We conclude this section with presenting a selection of measurement ensembles that meet the criteria of Proposition 23 and as a consequence also the ones of Corollary 24. We start with measurement ensembles that allow for recovering real-valued matrices $X \in \mathrm{L}(\mathcal{V})$.

Corollary 25. *Suppose that $\mathcal{V}$ is a real-valued vector spaces and let $\mathcal{A}:\mathrm{L}(\mathcal{V}) \to \mathbb{R}^m$ be the measurement map $\mathcal{A}(X) = \sum_{i=1}^m \operatorname{Tr}\left(A_i X\right) e_i$, where each A_i is a random matrix with independent entries obeying*

$$\mathbb{E}\left[a_{i,j}\right] = 0, \quad \mathbb{E}\left[a_{i,j}^2\right] = 1, \quad \mathbb{E}\left[a_{i,j}^4\right] \leq F, \tag{127}$$

where F is a constant. Then a sampling rate of $m \geq Crn$ suffices to meet the requirements of Proposition 23.

The result quoted in Corollary 25 was not established as a subroutine of a geometric proof technique for nuclear norm recovery, but consists of auxiliary statements that help to establish the Frobenius stable null space property [17, Definition 10] – a powerful alternative to geometric proof techniques relying on Proposition 3. However, if embedded properly into the framework of geometric recovery proof techniques, the auxiliary statements in Ref. [17] – see also Ref. [42, 43] – can still be used to establish recovery guarantees that rely on bounding the widths of descent cones. For our purposes, such a geometric proof environment is crucial, and this entire section is devoted to develop it. However, we point out that introducing and analyzing the square norm analogue of the Frobenius stable null space property – which is geared towards nuclear norm minimization – does constitute an intriguing follow-up problem. We leave this to future work.

Proof of Corollary 25. For a proof of this statement, we utilize auxiliary statements from Ref. [42]. Lemma 11 in loc. cit. asserts that such random matrices with bounded fourth moments obey $Q_{1/\sqrt{2}} \geq 1/4\max\{3,F\}$, where F is the fourth-moment bound. Also, Ref. [42, Lemma 12] assures $\mathbb{E}\left[\|H\|\right] \leq C_F\sqrt{n}$, where C_F is a constant that only depends on F. This in particular assures

$$\mathbb{E}\left[\|H\|\right] \leq C_F\sqrt{n} \leq \frac{C_F}{\sqrt{C}}\sqrt{\frac{m}{r}} \tag{128}$$

and we can set $\xi = 2^{-3/2}$, $C_2 = C_F/\sqrt{C}$ and $C_1 = 1/4\max\{3,F\}$. Choosing the constant C in the sampling rate large enough assures that these constants obey $\xi C_1 > 2(1+\sqrt{2})C_2$ for $\xi = 2^{-3/2}$ and all the requirements of Proposition 23 are met. The claim then follows from applying this statement. □

We conclude this section with embedding the main results of Ref. [16] into this framework. In fact, the entire apparatus presented in this section is a condensed version of the proofs in loc. cit. However, the reader's convenience, we include the corresponding statement here as well.

Corollary 26. *Consider measurement maps $\mathcal{A} : \mathrm{Herm}(\mathcal{V}) \to \mathbb{R}^m$ of the form $\mathcal{A}(X) = \sum_{i=1}^m \mathrm{Tr}\left(A_i X\right) e_i$. Then the following measurement ensembles meet the requirements of Proposition 23, if restricted to the recovery of Hermitian matrices:*

1. *$m \geq C_G rn$ and each $A_i = a_i a_i^\dagger$ corresponds to the outer product of a complex standard Gaussian vector $a_i \in \mathcal{V}$ with itself,*
2. *$m \geq C_{4D} rn \log(2n)$ and each $A_i = a_i a_i^\dagger$ is the outer product of a randomly selected element a_i of a complex projective 4-design.*

Once more, C_G and C_{4D} denote sufficiently large constants.

Proof. Let us start with the Gaussian case. In Ref. [16, Section 4.1.] the bounds $Q_{1/\sqrt{2}} \geq 1/96$ and $\mathbb{E}\left[\|H\|\right] \leq c_1\sqrt{n}$ are derived under the assumption $m \geq c_2 n$, where c_1 is sufficiently large. Thus, similarly to the proof of Corollary 25, setting $\xi = 2^{-3/2}$ and choosing the constant C_G in m sufficiently large indeed meets the requirements of Proposition 23.

For the 4-design case, [16, Proposition 12] assures that the bound $Q_\xi\left(E_r, \mathcal{A}\right) \geq \left(1-\xi^2\right)^2/24$ is valid for any $\xi \in [0,1]$. Also, Ref. [16, Proposition 13] implies

$$\mathbb{E}\left[\|H\|\right] \leq 3.1049\sqrt{n\log(2n)} \leq \frac{3.1049}{\sqrt{C_{4D}}}\sqrt{\frac{m}{r}}, \tag{129}$$

where we have inserted $m \geq C_{4D} rn \log(2n)$. Thus, choosing ξ appropriately and the constant C_{4D} in the sampling rate m large enough again assures that the requirements of Proposition 23 are met. □

References

[1] Y. Koren, R. Bell, and C. Volinsky, *Matrix factorization techniques for recommender systems,* Computer **42**, 30 (2009).

[2] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, *Quantum state tomography via compressed sensing,* Phys. Rev. Lett. **105**, 150401 (2010), arXiv:0909.3304 [quant-ph].

[3] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, *Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators,* New J. Phys. **14**, 095022 (2012), arXiv:1205.2300 [quant-ph].

[4] P. Bickel and E. Levina, *Regularized estimation of large covariance matrices,* Ann. Statist. **36**, 199 (2008), arXiv:0803.1909.

[5] Y. Chen, Y. Chi, and A. Goldsmith, *Exact and stable covariance estimation from quadratic sampling via convex programming,* IEEE Trans. Inf. Th. **61**, 4034 (2015), arXiv:1310.0807 [cs.IT].

[6] R. Basri and D. W. Jacobs, *Lambertian reflectance and linear sub-spaces,* IEEE Trans. Pattern Anal. Mach. Intell. **25**, 218 (2003).

[7] B. K. Natarajan, *Sparse approximate solutions to linear systems,* SIAM J. Comp. **24**, 227 (1995).

[8] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing* (Springer, 2013).

[9] M. Fazel, H. Hindi, and S. Boyd, *A rank minimization heuristic with application to minimum order system approximation,* in *Proceedings American Control Conference*, Vol. 6 (2001) pp. 4734–4739.

[10] E. Candès and B. Recht, *Exact matrix completion via convex optimization,* Found. Comput. Math. **9**, 717 (2009), arXiv:0805.4471 [cs.IT].

[11] E. J. Candès and T. Tao, *The power of convex relaxation: near-optimal matrix completion,* IEEE Trans. Inf. Th. **56**, 2053 (2010), arXiv:0903.1476 [cs.IT].

[12] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,* SIAM Rev. **52**, 471 (2010), arXiv:0706.4138 [math.OC].

[13] D. Gross, *Recovering low-rank matrices from few coefficients in any basis,* IEEE Trans. Inf. Th. **57**, 1548 (2011), arXiv:0910.1879 [cs.IT].

[14] E. J. Candès and Y. Plan, *Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements,* IEEE Trans. Inform. Theory **57**, 2342 (2011), arXiv:1001.0339 [cs.IT].

[15] A. Ahmed, B. Recht, and J. Romberg, *Blind deconvolution using convex programming,* IEEE Trans. Inf. Th. **60**, 1711 (2014), arXiv:1211.5608 [cs.IT].

[16] R. Kueng, H. Rauhut, and U. Terstiege, *Low rank matrix recovery from rank one measurements,* Appl. Comp. Harm. Anal. (2015), arXiv:1410.6913 [cs.IT].

[17] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, *Stable low-rank matrix recovery via null space properties,* arXiv:1507.07184 [cs.IT].

[18] J. A. Tropp, *Convex recovery of a structured signal from independent random linear measurements,* arXiv:1405.1102 [cs.IT].

[19] S. Mendelson, *Learning without concentration,* J. ACM **62**, 21:1 (2015), arXiv:1401.0304 [cs.LG].

[20] V. Koltchinskii and S. Mendelson, *Bounding the smallest singular value of a random matrix without concentration,* International Mathematics Research Notices , rnv096 (2015), arXiv:1312.3580 [math.PR].

[21] A. Y. Kitaev, A. Shen, and M. N. Vyalyi, *Classical and quantum computation*, Vol. 47 (American Mathematical Society, 2002).

[22] V. Paulsen, *Completely bounded maps and operator algebras*, Vol. 78 (Cambridge University Press, 2002).

[23] J. Watrous, *Semidefinite programs for completely bounded norms,* Theory of Computing **5**, 217 (2009), arXiv:0901.4709 [quant-ph].

[24] A. Ben-Aroya and A. Ta-Shma, *On the complexity of approximating the diamond norm,* Quantum Info. Comput. **10**, 77 (2010), arXiv:0902.3397 [quant-ph].

[25] J. Watrous, *Simpler semidefinite programs for completely bounded norms,* arXiv:1207.5726 [quant-ph].

[26] A. Shabani, R. L. Kosut, M. Mohseni, H. Rabitz, M. A. Broome, M. P. Almeida, A. Fedrizzi, and A. G. White, *Efficient measurement of quantum dynamics via compressive sensing,* Phys. Rev. Lett. **106**, 100401 (2011), arXiv:0910.5498 [quant-ph].

[27] V. Voroninski, *Quantum tomography from few full-rank observables,* arXiv:1309.7669 [math-ph].

[28] R. Kueng, *Low rank matrix recovery from few orthonormal basis measurements,* in *Sampling Theory and Applications (SampTA), 2015 International Conference on* (2015) pp. 402–406.

[29] A. Walther, *The question of phase retrieval in optics,* J. Mod. Opt. **10**, 41 (1963).

[30] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, *Phase retrieval via matrix completion,* SIAM Journal on Imaging Sciences **6**, 199 (2013), arXiv:1109.0573 [cs.IT].

[31] E. Candès and X. Li, *Solving quadratic equations via phaselift when there are about as many equations as unknowns,* Found. Comput. Math. **14**, 1017 (2014), arXiv:1208.6247 [cs.IT].

[32] E. J. Candès, T. Strohmer, and V. Voroninski, *Phaselift: exact and stable signal recovery from magnitude measurements via convex programming.* Commun. Pure Appl. Math. **66**, 1241 (2013), arXiv:1109.4499 [cs.IT].

[33] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon, *Phase retrieval with polarization,* SIAM J. Imaging Sci. **7**, 35 (2014), arXiv:1210.7752 [cs.IT].

[34] E. J. Candès, X. Li, and M. Soltanolkotabi, *Phase retrieval from coded diffraction patterns,* Appl. Comput. Harmon. Anal. **39**, 277 (2015), arXiv:1310.3240 [cs.IT].

[35] D. Gross, F. Krahmer, and R. Kueng, *A partial derandomization of phaselift using spherical designs,* J. Fourier Anal. Appl. **21**, 229 (2015), arXiv:1310.2267 [cs.IT].

[36] D. Gross, F. Krahmer, and R. Kueng, *Improved recovery guarantees for phase retrieval from coded diffraction patterns,* Appl. Comput. Harmon. Anal. , (2015), arXiv:1402.6286 [cs.IT].

[37] P. Walk and P. Jung, *On a reverse l_2-inequality for sparse circular convolutions,* in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (2013) pp. 4638–4642.

[38] H. Rauhut and Ž. Stojanac, *Tensor theta norms and low rank recovery,* arXiv:1505.05175 [cs.IT].

[39] A. Barvinok, *A course in convexity*, Graduate Studies in Mathematics, Vol. 54 (American Mathematical Society, 2002).

[40] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky, *The convex geometry of linear inverse problems,* Found. Comput. Math. **12**, 805 (2012), arXiv:1012.0621 [math.OC].

[41] S. Boyd and L. Vandenberghe, *Convex optimization* (Cambridge University Press, New York, 2004).

[42] M. Kabanava, H. Rauhut, and U. Terstiege, *Analysis of low rank matrix recovery via mendelson's small ball method,* in *Sampling Theory and Applications (SampTA), 2015 International Conference on* (2015) pp. 387–391.

[43] M. Kabanava, H. Rauhut, and U. Terstiege, *On the minimal number of measurements in low-rank matrix recovery,* in *Sampling Theory and Applications (SampTA), 2015 International Conference on* (2015) pp. 382–386.

[44] P. Delsarte, J. Goethals, and J. Seidel, *Spherical codes and designs,* Geom. Dedicata **6**, 363 (1977).

[45] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, *Symmetric informationally complete quantum measurements,* J. Math. Phys. **45**, 2171 (2004).

[46] A. Ambainis and J. Emerson, *Quantum t-designs: t-wise independence in the quantum world,* in *Computational Complexity, 2007. CCC '07. Twenty-Second Annual IEEE Conference on* (2007) pp. 129–140, quant-ph/0701126.

[47] M.-D. Choi, *Completely positive linear maps on complex matrices,* Lin. Alg. App. **10**, 285 (1975).

[48] A. Jamiolkowski, *Linear transformations which preserve trace and positive semidefiniteness of operators,* Rep. Math. Phys. **3**, 275 (1972).

[49] S. Ling and T. Strohmer, *Self-calibration and biconvex compressive sensing,* ArXiv e-prints (2015), arXiv:1501.06864 [cs.IT].

[50] R. Kueng, D. M. Long, A. C. Doherty, and S. T. Flammia, *Comparing experiments to the fault-tolerance threshold,* ArXiv e-prints (2015), arXiv:1510.05653 [quant-ph].

[51] S. Kimmel and Y.-K. Liu, *Quantum compressed sensing using 2-designs,* ArXiv e-prints (2015), arXiv:1510.08887 [quant-ph].

[52] J. Emerson, R. Alicki, and K. Życzkowski, *Scalable noise estimation with random unitary operators,* J. Opt. B **7**, S347 (2005), arXiv:quant-ph/0503243.

[53] C. Granade, C. Ferrie, and D. G. Cory, *Accelerated randomized benchmarking,* New J. Phys. **17**, 013042 (2015), arXiv:1404.5275.

[54] M. Ziman, *Process positive-operator-valued measure: A mathematical framework for the description of process tomography experiments,* Phys. Rev. A **77**, 062112 (2008), arXiv:0802.3862 [quant-ph].

[55] M. Grant and S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1, `http://cvxr.com/cvx` (2014).

[56] M. Grant and S. Boyd, in *Recent advances in learning and control*, Lecture Notes in Control and Information Sciences, edited by V. Blondel, S. Boyd, and H. Kimura (Springer-Verlag Limited, 2008) pp. 95–110, `http://stanford.edu/~boyd/graph_dcp.html`.

[57] J. J. Wallman and S. T. Flammia, *Randomized benchmarking with confidence,* New J. Phys. **16**, 103032 (2014), arXiv:1404.6025 [quant-ph].

[58] C. Dankert, R. Cleve, J. Emerson, and E. Livine, *Exact and approximate unitary 2-designs and their application to fidelity estimation,* Phys. Rev. A **80**, 012304 (2009).

[59] D. Gross, K. M. R. Audenaert, and J. Eisert, *Evenly distributed unitaries: on the structure of unitary designs,* J. Math. Phys. **48**, 052104 (2007).

[60] H. Zhu, *Multiqubit clifford groups are unitary 3-designs,* arXiv:1510.02619 [quant-ph].

[61] Z. Webb, *The clifford group forms a unitary 3-design,* arXiv:1510.02769 [quant-ph].

[62] J. Watrous, CS 766 Theory of Quantum Information, `https://cs.uwaterloo.ca/~watrous/LectureNotes.html` (2011).

[63] A. H. Werner, D. Jaschke, P. Silvi, M. Kliesch, T. Calarco, J. Eisert, and S. Montangero, *A positive tensor network approach for simulating open quantum many-body systems,* arXiv:1412.5746 [quant-ph].

[64] J. M. Landsberg, *Tensors: geometry and applications* (American Mathematical Society, 2012).

A unifying framework for relaxations of the causal assumptions in Bell's theorem

R. Chaves,[1] R. Kueng,[1] J.B. Brask,[2] and D. Gross[1]

[1]*Institute for Physics, University of Freiburg, Rheinstrasse 10, D-79104 Freiburg, Germany*
[2]*Département de Physique Théorique, Université de Genève, 1211 Genève, Switzerland*

(Dated: November 19, 2014)

Bell's Theorem shows that quantum mechanical correlations can violate the constraints that the causal structure of certain experiments impose on any classical explanation. It is thus natural to ask to which degree the causal assumptions – e.g. "locality" or "measurement independence" – have to be relaxed in order to allow for a classical description of such experiments. Here, we develop a conceptual and computational framework for treating this problem. We employ the language of Bayesian networks to systematically construct alternative causal structures and bound the degree of relaxation using quantitative measures that originate from the mathematical theory of causality. The main technical insight is that the resulting problems can often be expressed as computationally tractable linear programs. We demonstrate the versatility of the framework by applying it to a variety of scenarios, ranging from relaxations of the measurement independence, locality and bilocality assumptions, to a novel causal interpretation of CHSH inequality violations.

The paradigmatic Bell experiment [1] involves two distant observers, each with the capability to perform one of two possible experiments on their shares of a joint system. Bell observed that even absent of any detailed information about the physical processes involved, the *causal structure* of the setup alone implies strong constraints on the correlations that can arise from any *classical* description [2]. The physically well-motivated causal assumptions are: (i) *measurement independence*: experimenters can choose which property of a system to measure, independently of how the system has been prepared; (ii) *locality*: the results obtained by one observer cannot be influenced by any action of the other (ideally space-like separated) experimenter. The resulting constraints are Bell's inequalities [1]. Quantum mechanical processes subject to the same causal structure can violate these constraints – a prediction that has been abundantly verified experimentally [3–7]. This effect is commonly referred to as *quantum non-locality*.

It is now natural to ask how stable the effect of quantum non-locality is with respect to relaxations of the causal assumptions. Which "degree of measurement dependence", e.g., is required to reconcile empirically observed correlations with a classical and local model? Such questions are not only, we feel, of great relevance to foundational questions – they are also of interest to practical applications of non-locality, e.g. in cryptographic protocols. Indeed, eavesdroppers can (and do [8]) exploit the failure of a given cryptographic device to be constrained by the presumed causal structure to compromise its security. At the same time, it will often be difficult to ascertain that causal assumptions hold *exactly* – which makes it important to develop a systematic quantitative theory.

Several variants of this question have recently attracted considerable attention [9–20]. For example, measurement dependence has been found to be a very strong resource: If no restrictions are imposed on possible correlations between the measurement choices and the source producing the particles to be measured, any nonlocal distribution can be reproduced [21]. What is more, only about about $1/15$ of a bit of correlation between the source and measurements is sufficient to reproduce all correlations obtained by projective measurements on a singlet state [10, 12, 13]. In turn, considering relaxations of the locality assumption, Toner and Bacon showed that one bit of communication between the distant parties is again sufficient to simulate the correlations of singlet states [9].

In this paper we provide a unifying framework for treating relaxations of the measurement independence and locality assumptions in Bell's theorem. To achieve this, we borrow several concepts from the mathematical theory of *causality*, a relatively young subfield of probability theory and statistics [22, 23]. With the aim of describing the causal relations (rather than mere correlations) between variables that can be extracted from empirical observations, this community has developed a systematic and rigorous theory of causal structures and quantitative measures of causal influence.

Our framework rests on three observations (details are provided below): (i) Alternative causal structures can systematically be represented using the graphical notation of Bayesian networks [22]. There, variables are associated with nodes in a graph, and directed edges represent functional dependencies. (ii) These edges can be weighted by quantitative measures of causal influence [22, 24]. (iii) Determining the minimum degree of influence required for a classical explanation of observable distributions can frequently be cast as a computationally tractable linear program.

The versatility of the framework is demonstrated in a variety of applications. We give an operational meaning to the violation of the CHSH inequality [25] as the minimum amount of direct causal influence between the parties required to reproduce the observed correlations. Considering the Collins-Gisin scenario [26], we show that quantum correlations are incompatible with a classical description, even if we allow one of the parties to communicate its outcomes. We also show that the results in [10, 13] regarding measurement-independence relaxations can be improved by considering different Bell scenarios. Finally, we study the bilocality assumption [27, 28] and show that although it defines a non-convex set, its relaxation can also be cast as a linear program, naturally quantifying the degree of non-bilocality.

Bayesian networks and measures for the relaxation of causal assumptions— The causal relationships between n jointly distributed discrete random variables $(X_1, \ldots, X_n)$ are specified by means of a *directed acyclic graph* (DAG). To this end, each variable is associated with one of the nodes of the graph. One then says that the X_i's form a *Bayesian network* with respect to the graph, if every variable can be expressed as a deterministic function $X_i = f_i(\mathrm{PA}_i, N_i)$ of its graph-theoretic parents PA_i and an unobserved noise term N_i, such that the N_i's are jointly independent [29]. This is the case if and only if the probability $p(\mathbf{x}) = p(x_1, \ldots, x_n)$ is of the form

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i|\mathrm{pa}_i). \qquad (1)$$

This identity encodes the causal relationships implied by the DAG [22].

As a paradigmatic example of a DAG, consider a bipartite Bell scenario (Fig. 1a). In this scenario, two separated observers, Alice and Bob, each perform measurements according to some inputs, here represented by random variables X and Y respectively, and obtain outcomes, represented by A and B. The causal model involves an explicit shared hidden variable Λ which mediates the correlations between A and B. From (1) it follows that $p(x, y, \lambda) = p(x)p(y)p(\lambda)$ — which reflects the measurement independence assumption. It also follows that $a = f_A(x, \lambda, n_A)$, $b = f_B(y, \lambda, n_B)$. We incur no loss of generality by absorbing the local noise terms N_A, N_B into Λ and will thus assume from now on that $a = f_A(x, \lambda), b = f_B(y, \lambda)$ for suitable functions f_A, f_B. This encodes the locality assumption. Together, these relations imply the well-known local hidden variable (LHV) model of Bell's theorem:

$$p(a, b|x, y) = \sum_{\lambda} p(a|x, \lambda)p(b|y, \lambda)p(\lambda). \qquad (2)$$

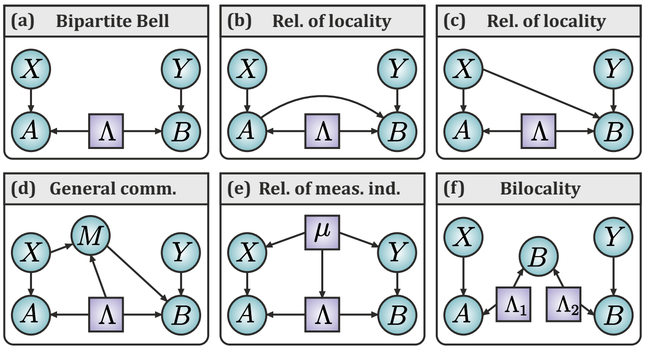


FIG. 1. **(a)** LHV model for the bipartite Bell scenario. **(b)** A relaxation of locality, where A may have direct causal influence on B. **(c)** Another relaxation in which X may have direct causal influence on B. **(d)** The most general communication scenario from Alice to Bob. **(e)** A relaxation of measurement independence, where the two inputs may be correlated, via a common ancestor, with the hidden variable Λ. **(f)** The bilocality scenario for which the two sources Λ_1 and Λ_2 are assumed to be independent. Round edges stand for observable variables while squares represent non-observable (hidden) ones.

Causal mechanisms relaxing locality (Fig. 1b–d) and measurement independence (Fig. 1e) can be easily expressed using Bayesian networks. The networks themselves, however, do not directly quantify the degree of relaxation. Thus, one needs to devise ways of checking and quantifying such causal dependencies. To define a sensible measure of causal influence we introduce a core concept from the causality literature – *interventions* [22].

An intervention is the act of forcing a variable, say X_i, to take on some given value x_i' and is denoted by $do(x_i')$. The effect is to erase the original mechanism $f_i(pa_i, n_i)$ and place X_i under the influence of a new mechanism that sets it to the value x_i' while keeping all other functions f_j for $j \neq i$ unperturbed. The intervention $do(x_i')$ amounts to a change in the decomposition (1), given by [30]

$$p(\mathbf{x}|do(x_i')) = \begin{cases} \prod_{j \neq i}^{n} p(x_j|\mathrm{pa}_j) & \text{if } x_i = x_i', \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

Considering locality relaxations, we can now define a measure $\mathcal{C}_{A \to B}$ for the *direct causal influence* of A into B for the model in Fig. 1b:

$$\mathcal{C}_{A \to B} = \sup_{b,y,a,a'} \sum_{\lambda} p(\lambda)|p(b|do(a), y, \lambda) - p(b|do(a'), y, \lambda)|. \qquad (4)$$

It is the maximum shift (averaged over the unobservable Λ) in the probability of B caused by interventions in A. Similarly, one can define $\mathcal{C}_{X \to B}$ for the DAG in Fig. 1c and in other situations. To highlight the relevance of this measure, we note that a variation of it, known as *average causal effect*, can be used to quantify the effect of a drug in remedying a given symptom [22]. We are also interested in relaxations of measurement independence. Considering the case of a bipartite scenario (illustrated in Fig. 1e and that can be easily extended to multipartite versions), we can define the measure

$$\mathcal{M}_{X,Y:\lambda} = \sum_{x,y,\lambda} |p(x,y,\lambda) - p(x,y)p(\lambda)|. \tag{5}$$

This can be understood as a measure of how much the inputs are correlated with the source, i.e. how much the underlying causal model fails to comply with measurement independence.

The linear programing framework—Given some observed probabilities and a particular measure of relaxation, our aim is to compute the minimum value of the measure compatible with the observations. As sketched below, this leads to a tractable linear program as long as there is only one unobserved variable Λ. (However, even in case of several hidden variables, variants of these ideas can still be used). Details are given in the Appendix.

For simplicity we consider the usual Bell scenario of Fig. 1a. The most general observable quantity is the joint distribution $p(a,b,x,y) = p(a,b|x,y)p(x)p(y)$. Since we control the "inputs" X and Y, their distribution carries no information and we may thus restrict attention to $p(a,b|x,y)$. This conditional probability is, in turn, a linear function of the distribution of Λ. To make this explicit, represent $p(a,b|x,y)$ as a vector $\mathbf{p}$ with components $\mathbf{p}_j$ labeld by the multi-index $j = (a,b,x,y)$. Similarly, identify the distribution of Λ with a vector with components $\mathbf{q}_\lambda = p(\Lambda = \lambda)$. Then from the discussion above, we have that $\mathbf{p} = T\mathbf{q}$ where T is a matrix with elements $T_{j,\lambda} = \delta_{a,f_A(x,\lambda)}\delta_{b,f_B(y,\lambda)}$. Conditional expectations that include the application of a *do*-operation are obtained via a modified T matrix. E.g., $\mathbf{q}'_j = p(a,b|x,y,do(a')) = T'q$ for $T'_{j,\lambda} = \delta_{a,a'}\delta_{b,f_B(y,\lambda)}$. The measures $\mathcal{C}$ and $\mathcal{M}$ are easily seen to be convex functions of the conditional probabilities $p(a,b|x,y)$ and their variants arising from the application of *do*'s – and thus convex functions of $\mathbf{q}$. Hence their minimization subject to the linear constraint $T\mathbf{q} = \mathbf{p}$ for an empirically observed distribution $\mathbf{p}$ is a convex optimization problem. This remains true if only some linear function $V\mathbf{p} = VT\mathbf{q}$ (e.g. a Bell inequality) of the distribution $\mathbf{p}$ is constrained. The problem is not manifestly a (computationally tractable) linear program (LP), since neither objective function is linear in $\mathbf{q}$. However, we establish in the appendix that it can be cast as such:

Theorem 1. *The constrained minimization of the measures $\mathcal{C}$ and $\mathcal{M}$ over hidden variables reproducing any observed probability distribution can be reformulated as a primal linear program (LP). Its solution is equivalent to*

$$\max_{1\leq i\leq K} \langle \mathbf{v}_i, V\mathbf{p}\rangle, \tag{6}$$

where the $\{\mathbf{v}_i\}_{i=1}^K$ are the vertices of the LP's dual feasible region.

This result highlights another nice aspect of our framework. Unlike the results in [11–17], (6) is a closed form-expression valid for any distribution (or observation derived from it by a linear function $V\mathbf{p}$), not just the value of a specific Bell inequality. This allows for a much more detailed description.

In the following sections, we apply our framework to a variety of applications. We focus on the results while the more technical proofs are given in the Appendices.

Novel interpretation of the CHSH inequality— As a first application, we show that a violation of the CHSH inequality can be interpreted as the minimal direct causal influence between the parties required to simulate the observed correlations.

Intuitively, the more nonlocal a given distribution is, the more direct causal influence between Alice and Bob should be required to simulate it. We make this intuition precise by considering the models in Fig. 1b–c and the CHSH scenario (two inputs, two outputs for both Alice and Bob). For any observed distribution $p(a,b|x,y)$, we establish in the Appendix that

$$\min \mathcal{C}_{A\to B} = \min \mathcal{C}_{X\to B} = \max\left[0, \mathrm{CHSH}\right], \tag{7}$$

where the maximum should be taken over all the eight symmetries under relabelling of inputs, outputs, and parties of the CHSH quantity [25]

$$\begin{aligned} CHSH = {} & p(00|00) + p(00|01) + p(00|10) \\ & - p(00|11) - p^A(0|0) - p^B(0|0), \end{aligned} \tag{8}$$

where the last two terms represent the marginals for Alice and Bob respectively. The CHSH inequality stipulates that for any LHV model, $CHSH \leq 0$. Eq. (7) shows that, regardless of the particular distribution, the minimum direct causal influence is exactly quantified by the CHSH inequality violation.

Inspired by the communication scenario of Toner and Bacon [9] (Fig. 1d), we can also quantify the relaxation of the locality assumption as the minimum amount of communication required to simulate a given distribution. We measure the communication by the Shannon entropy $H(m)$ of the message m which is sent. For a binary message, we can use our framework to prove, in complete analogy with (7), that

$$\min H(m) = h(\mathrm{CHSH}) \tag{9}$$

if CHSH > 0 and 0 otherwise, where $h(v) = -v\log_2 v - (1-v)\log_2(1-v)$ is the binary entropy. We note that for maximal quantum violation CHSH $= 1/\sqrt{2} - 1/2$, as produced by a single state, a message with $H(m) \approx 0.736$ bits is required. This is less than the 1 bit of communication required by the protocol of Toner and Bacon [9] for reproducing arbitrary correlations of a singlet.

Quantum nonlocality is incompatible with some locality relaxations— Given that violation of CHSH can be directly related to relaxation of locality, one can ask whether similar interpretations exists for other scenarios. For example, we can consider a setting with three inputs and two outputs for Alice and Bob, and consider the causal model in Fig. 1b. Similar to the usual LHV model (2), the correlations compatible with this model form a polytope. One facet of this polytope is

$$\langle E_{00}\rangle - \langle E_{02}\rangle - \langle E_{11}\rangle + \langle E_{12}\rangle - \langle E_{20}\rangle + \langle E_{21}\rangle \leq 4, \quad (10)$$

where $E_{xy} = \langle A_x B_y\rangle = \sum_{a,b}(-1)^{a+b}p(a,b|x,y)$. This inequality can be violated by any quantum state $|\psi\rangle = \sqrt{\epsilon}|00\rangle + \sqrt{(1-\epsilon)}|11\rangle$ with $\epsilon \neq 0, 1$. Consequently, any pure entangled state – no matter how close to separable – generates correlations that cannot be explained even if we allow for a relaxation of the locality assumption, where one of the parties communicates its measurement outcomes to the other.

How much measurement dependence is required to causally explain nonlocal correlations?— The results in Refs. [10, 12, 13] show that measurement dependence is a very strong resource for simulating nonlocal correlations. In fact, a mutual information as small as $I(X,Y:\lambda) \approx 0.0663$ is already sufficient to simulate all correlations obtained by (any number of) projective measurements on a single state [12, 13]. Given the fundamental implication and practical relevance of increasing these requirements, we aim to find larger values for $I(X,Y:\lambda)$ by means of our framework. The result of [12, 13] leaves us with three options, regarding the quantum states: either non-maximally entangled states of two qubits, two-qudit states, or states with more than two parties.

Regarding non-maximally entangled two-qubit states, we were unable to improve the minimal mutual information. Regarding qudits, we have considered relaxations in the CGLMP scenario [31] – a bipartite scenario, where Alice and Bob each have two inputs and d outcomes. The CGLMP inequality is of the form $I_d \leq 2$. Assuming that a particular I_d-value is observed in the setting of Fig. 1e, we numerically obtain the very simple relation

$$\min \mathcal{M} = \max\left[0, (I_d - 2)/4\right] \quad (11)$$

up to $d = 8$. Via the Pinsker inequality [32, 33], (11) provides a lower bound on the minimum mutual information $I(X,Y:\lambda) \geq \mathcal{M}^2 \log_2 \mathrm{e}$. This bound implies that for any $I_d \geq 3.214$, the mutual information required exceeds the 0.0663 obtained in Ref. [13]. Using the results in Ref. [34] for the scaling of the optimal quantum violation with d, one sees that this requires $d \geq 16$. However, we note that the bounds provided by the Pinsker inequality are usually far from tight, leaving a lot of room for improvement. Moreover – as detailed in the Appendix – a corresponding upper bound (obtained via the solution to the minimization of $\mathcal{M}$) is larger than the values obtained in [12, 13] as soon as $d \geq 5$. Though this upper bound is not necessarily tight, we highlight the fact that for $d = 2$ it gives exactly $I(X,Y:\lambda) = 0.0463$, the value analytically obtained in [12, 13].

Regarding multipartite scenarios, we have considered GHZ correlations [35] in a tri-partite scenario where each party has two inputs and two outputs. We numerically obtain $0.090 \leq I(X,Y,Z:\lambda) \leq 0.207$. This implies that increasing the number of parties can considerably increase the measurement dependence requirements for reproducing quantum correlations.

Bilocality scenario— To illustrate how the formalism can also be used in generalized Bell scenarios [27, 28, 36, 37], we briefly explore the entanglement swapping scenario [38] of Fig. 1f (a more detailed discussion is given in the Appendix). As can be seen from the DAG, the hidden variables in this scenario are independent $p(\lambda_1,\lambda_2) = p(\lambda_1)p(\lambda_2)$, the so-called *bilocality* assumption [27, 28].

As in Ref. [27, 28], we take the inputs x, z and the outputs a, c to be dichotomic while b takes four values which we decompose in two bits as $b = (b_0, b_1)$. The distribution of hidden variables can be organized in a 64-dimensional vector $\mathbf{q}$ with components $q_{\alpha_0,\alpha_1,\beta_0,\beta_1,\gamma_0,\gamma_1}$, where α_x specifies the value of a for a given x (and analogously for γ, c and z) and β_i specifies the value of b_i. Thus together the indices label all the deterministic functions for A, B, C given their parents. As shown in [27, 28], the bilocality assumption is equivalent to demanding $q^{ac}_{\alpha_0,\alpha_1,\gamma_0,\gamma_1} = q^{a}_{\alpha_0,\alpha_1}q^{c}_{\gamma_0,\gamma_1}$, where $q^{ac}_{\alpha_0,\alpha_1,\gamma_0,\gamma_1} = \sum_{\beta_0,\beta_1} q_{\alpha_0,\alpha_1,\beta_0,\beta_1,\gamma_0,\gamma_1}$ is the marginal for AC etc. Similar to (5) a natural measure $\mathcal{M}_{\mathrm{BL}}$ of non-bilocality quantifies by how much the underlying hidden variable distribution fails to comply with this constraint:

$$\mathcal{M}_{\mathrm{BL}} = \sum_{\alpha_0,\alpha_1,\gamma_0,\gamma_1} |q^{ac}_{\alpha_0,\alpha_1,\gamma_0,\gamma_1} - q^{a}_{\alpha_0,\alpha_1}q^{c}_{\gamma_0,\gamma_1}|. \quad (12)$$

Clearly $\mathcal{M}_{\mathrm{BL}} = 0$, if and only if the bilocality constraint is fulfilled. However, demanding bilocality imposes a quadratic constraint on the hidden variables. This results in a non-convex set which is extremely difficult characterize [27, 28, 36, 37]. Nevertheless, our framework is still useful, as using the marginals for a given observed distribution to constrain the problem further, the minimization of $\mathcal{M}_{\mathrm{BL}}$ can be cast in terms of a linear program with a single free parameter, which is then further minimised over (see Appendix).

As an illustration we consider the non-bilocal distribution found in Refs. [27, 28]. It can be obtained by projective measurements on a pair of identical two-qubit entangled states $\varrho = v|\Psi^-\rangle\langle\Psi^-| + (1-v)\mathbb{I}/4$. This distribution violates the bilocality inequality $\mathcal{B} = \sqrt{|I|} + \sqrt{|J|} \leq 1$ giving a value $\mathcal{B} = \sqrt{2}v$. Using our framework we find $\mathcal{M}_{\mathrm{BL}} = \max(2v^2 - 1, 0)$. Thus, for this specific distribution, $\mathcal{M}_{\mathrm{BL}} = \mathcal{B}^2 - 1$, so there is a one-to-one correspondence between the violation of the bilocality inequality and the minimum relaxation of the bilocality constraint required to reproduce the correlations. This assigns an operational meaning to $\mathcal{B}$.

Conclusion— In this work we have revisited nonlocality from a causal inference perspective and provided a linear programming framework for relaxing the measurement independence and locality assumptions in Bell's theorem. Using the framework, we have given a novel causal interpretation of violations of the CHSH inequality, and we have shown that quantum correlations are still incompatible with classical causal models even if one allows for the communication of measurement outcomes. This implies that quantum nonlocality is even stronger than previously thought. Considering a variety of scenarios, we also have shown that the results in Refs. [10, 12, 13] regarding the minimal measurement dependence required to simulated nonlocal correlations can be extended. Finally we explained how the relaxation of the bilocality assumption naturally quantifies the degree of non-bilocality in an entanglement swapping experiment.

In addition to these results, we believe the generality of our framework motivates and – more importantly – provides a basic tool for future research. For instance, it would be interesting to understand how our framework can be generalized in order to derive useful inequalities in the context of randomness expansion, following the ideas in [14]. Another natural possibility, inspired by [39, 40], would be to look for a good measure of genuine multipartite nonlocality, by considering specific underlying signalling models. Finally, it would be interesting to understand how our treatment of the bilocality problem could be generalized and applied to the characterization of the non-convex compatibility regions of more complex quantum networks [36, 41–44].

We thank R. Luce and D. Cavalcanti for useful discussions. Research in Freiburg is supported by the Excellence Initiative of the German Federal and State Governments (Grant ZUK 43), the Research Innovation Fund of the University of Freiburg, the US Army Research Office under contracts W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), and the DFG. JB was supported by the Swiss National Science Foundation (QSIT director's reserve) and SEFRI (COST action MP1006).

[1] J. S. Bell, Physics **1**, 195 (1964).

[2] Here enters the third assumption in Bell's theorem, that of *realism*. It states that one can consistently assign a value to any physical property – independently of whether or not it is measured. In the Bayesian network language, this is expressed by the fact that variables are assumed to be deterministic functions of its parents.

[3] S. J. Freedman and J. F. Clauser, Phys. Rev. Lett. **28**, 938 (1972).

[4] A. Aspect, P. Grangier, and G. Roger, Phys. Rev. Lett. **49**, 91 (1982).

[5] B. G. Christensen, K. T. McCusker, J. B. Altepeter, B. Calkins, T. Gerrits, A. E. Lita, A. Miller, L. K. Shalm, Y. Zhang, S. W. Nam, N. Brunner, C. C. W. Lim, N. Gisin, and P. G. Kwiat, Phys. Rev. Lett. **111**, 130406 (2013).

[6] M. A. Rowe, D. Kielpinski, V. Meyer, C. A. Sackett, W. M. Itano, C. Monroe, and D. J. Wineland, Nature **409**, 791 (2001).

[7] M. Giustina, A. Mech, S. Ramelow, B. Wittmann, J. Kofler, J. Beyer, A. Lita, B. Calkins, T. Gerrits, S. W. Nam, *et al.*, Nature **497**, 227 (2013).

[8] L. Lydersen, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, Nature photonics **4**, 686 (2010).

[9] B. F. Toner and D. Bacon, Phys. Rev. Lett. **91**, 187904 (2003).

[10] J. Barrett and N. Gisin, Phys. Rev. Lett. **106**, 100406 (2011).

[11] M. J. W. Hall, Phys. Rev. A **82**, 062117 (2010).

[12] M. J. W. Hall, Phys. Rev. Lett. **105**, 250404 (2010).

[13] M. J. W. Hall, Phys. Rev. A **84**, 022102 (2011).

[14] D. E. Koh, M. J. W. Hall, Setiawan, J. E. Pope, C. Marletto, A. Kay, V. Scarani, and A. Ekert, Phys. Rev. Lett. **109**, 160404 (2012).

[15] M. Banik, Phys. Rev. A **88**, 032118 (2013).

[16] A. Rai, M. R. Gazi, M. Banik, S. Das, and S. Kunkri, Journal of Physics A: Mathematical and Theoretical **45**, 475302 (2012).

[17] B. Paul, K. Mukherjee, and D. Sarkar, Phys. Rev. A **88**, 014104 (2013).

[18] L. P. Thinh, L. Sheridan, and V. Scarani, Phys. Rev. A **87**, 062121 (2013).

[19] G. Pütz, D. Rosset, T. J. Barnea, Y.-C. Liang, and N. Gisin, arXiv preprint arXiv:1407.5634 (2014).

[20] K. Maxwell and E. Chitambar, Phys. Rev. A **89**, 042108 (2014).

[21] C. H. Brans, International Journal of Theoretical Physics **27**, 219 (1988).

[22] J. Pearl, *Causality* (Cambridge University Press, 2009).

[23] P. Spirtes, N. Glymour, and R. Scheienes, *Causation, Prediction, and Search, 2nd ed.* (The MIT Press, 2001).

[24] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Scholkopf, The Annals of Statistics **41**, 2324 (2013).

[25] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, Phys. Rev. Lett. **23**, 880 (1969).

[26] D. Collins and N. Gisin, Journal of Physics A: Mathematical and General **37**, 1775 (2004).

[27] C. Branciard, N. Gisin, and S. Pironio, Phys. Rev. Lett. **104**, 170401 (2010).

[28] C. Branciard, D. Rosset, N. Gisin, and S. Pironio, Phys. Rev. A **85**, 032119 (2012).

[29] We adopt the convention that uppercase letters label random variables while their values are denoted in lower case. For brevity, we will sometimes suppress explicit mention of the random variables – e.g. write $p(x_i, x_j)$ instead of the more precise $p(X_i = x_i, X_j = x_j)$.

[30] We note that the *do*-operation is defined only relative to a causal model as encoded in the DAG. (The graph structure enters (3) through the reference to parent nodes pa_j). In particular, $p(y|do(x))$ is in general different from the usual conditional probability $p(y|x)$ – these notions only coincide if the set of parents PA_X and PA_Y are disjoint. For example: The variables X and Y can be maximally correlated, i.e. $p(y|x) \propto \delta_{x,y}$, and still $p(y|do(x)) = p(y)$. This would occur e.g. if all the correlations between the variables are mediated via a common parent u, such that $p(x,y|u) = p(x|u)p(x|u)$.

[31] D. Collins, N. Gisin, N. Linden, S. Massar, and S. Popescu, Phys. Rev. Lett. **88**, 040404 (2002).

[32] A. A. Fedotov, P. Harremoës, and F. Topsoe, Information Theory, IEEE Transactions on **49**, 1491 (2003).

[33] M. J. Hall, Entropy **15**, 3698 (2013).

[34] J.-L. Chen, C. Wu, L. C. Kwek, C. H. Oh, and M.-L. Ge, Phys. Rev. A **74**, 032106 (2006).

[35] D. M. Greenberger, M. Horne, and A. Zeilinger, "Bells theorem, quantum theory, and conceptions of the universe," (Kluwer Academic Publishers, Dordrecht, NL).

[36] T. Fritz, New Journal of Physics **14**, 103001 (2012).

[37] R. Chaves and T. Fritz, Phys. Rev. A **85**, 032113 (2012).

[38] M. Zukowski, A. Zeilinger, M. Horne, and A. Ekert, Physical Review Letters **71**, 4287 (1993).

[39] R. Gallego, L. E. Wurflinger, A. Acin, and M. Navascues, Phys. Rev. Lett. **109**, 070401 (2012).

[40] J.-D. Bancal, J. Barrett, N. Gisin, and S. Pironio, Phys. Rev. A **88**, 014102 (2013).

[41] D. Cavalcanti, M. L. Almeida, V. Scarani, and A. Acin, Nature communications **2**, 184 (2011).

[42] R. Chaves, L. Luft, and D. Gross, New J. Phys. **16**, 043001 (2014).

[43] R. Chaves, L. Luft, T. O. Maciel, D. Gross, D. Janzing, and B. Schölkopf, Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence , 112 (2014).

[44] R. Chaves, C. Majenz, and D. Gross, arXiv preprint arXiv:1407.3800 (2014).

[45] S. Boyd and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2009).

[46] A. Barvinok, *A course in convexity* (American Mathematical Society, 2002).

[47] A. A. Balke and J. Pearl, *Probabilistic counterfactuals: semantics, computation, and applications*, Tech. Rep. (DTIC Document, 1997).

[48] T. Christof and A. Löbel, "PORTA – POlyhedron Representation Transformation Algorithm," (2009).

[49] T. Fritz and R. Chaves, IEEE Trans. Inform. Theory **59**, 803 (2013).

[50] R. Chaves, Phys. Rev. A **87**, 022102 (2013).

APPENDIX

For the sake of being as self-contained as possible, we start the appendix with reviewing basic concepts in convex optimization. We then use these concepts to establish Theorem 1 – our main technical result. As detailed below, the measures of direct causal influence (4) and measurement dependence (5), respectively, can be recast as vector norms. Their minimization, subject to the specific constraints of each of the causal models in Fig. 1 is then explored in detail.

REVIEW OF LINEAR PROGRAMMING

Linear Programming (LP) is a very powerful and widely used tool for dealing – both practically and theoretically – with certain families of convex optimization problems. We refer to [45, 46] and references therein for an overview. From now on we assume that vectors $\mathbf{x} \in \mathbb{R}^n$ are represented in the standard basis $\{\mathbf{e}_i\}_{i=1}^n$, i.e. $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i$. In this representation, the two vectors $\mathbf{0}_n := (0,\ldots,0)^T$ (the "zero"-vector) and $\mathbf{1}_n := (1,\ldots,1)^T$ (the "all-ones" vector) will be of particular importance. Furthermore, we are frequently going to concatenate vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$ via $\mathbf{x} \oplus \mathbf{y} := \sum_{i=1}^n x_i \mathbf{e}_i + \sum_{j=1}^m y_i \mathbf{e}_{n+j} \in \mathbb{R}^{n+m}$. Also, $\langle \cdot, \cdot \rangle$ shall denote the standard inner product of finite dimensional real vector spaces.

There are many equivalent ways of defining the standard form of primal/dual LP's. Here we adopt the formalism of [47]. A convex optimization problem fits the framework of linear programming, if it can be reformulated as

$$\begin{aligned} \gamma = \min_{\boldsymbol{\xi} \in \mathbb{R}^n} \quad & \langle \mathbf{c}, \boldsymbol{\xi} \rangle \\ \text{subject to} \quad & \Phi \boldsymbol{\xi} \geq \mathbf{b} \\ & \boldsymbol{\xi} \geq \mathbf{0}_n, \end{aligned} \tag{13}$$

where $\mathbf{c} \in \mathbb{R}^n$ as well as $\mathbf{b} \in \mathbb{R}^m$ are vectors and $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ corresponds to an arbitrary real $m \times n$-matrix. The inequality signs here denote generalized inequalities on $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. To be concrete, two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ obey $\mathbf{y} \geq \mathbf{x}$ if and only if $y_i \geq x_i$ holds for all $i = 1, \ldots, n$.

It is very useful to consider linear programming problems in pairs. An optimization of the form (13) is called a *primal problem in standard form* and is accompanied by its *dual problem (in standard form)*:

$$\begin{aligned} \beta = \max_{\boldsymbol{\zeta} \in \mathbb{R}^m} \quad & \langle \boldsymbol{\zeta}, \mathbf{b} \rangle \\ \text{subject to} \quad & \Phi^T \boldsymbol{\zeta} \leq \mathbf{c} \\ & \boldsymbol{\zeta} \geq \mathbf{0}_m. \end{aligned} \tag{14}$$

Here, $\Phi^T : \mathbb{R}^m \to \mathbb{R}^n$ denotes the transpose of Φ (with respect to the standard basis). For a given pair of linear programs, we call $\boldsymbol{\xi} \in \mathbb{R}^n$ *primal feasible* if it obeys the constraints $\Phi \boldsymbol{\xi} \geq \mathbf{b}$ and $\boldsymbol{\xi} \geq \mathbf{0}_n$. Likewise, we call $\boldsymbol{\zeta} \in \mathbb{R}^m$ *dual feasible* if $\Phi^T \boldsymbol{\zeta} \leq \mathbf{c}$ and $\boldsymbol{\zeta} \geq \mathbf{0}_m$ hold. Furthermore, we call an LP *primal feasible*, if it admits at least one primal feasible variable $\boldsymbol{\xi}$ and *dual feasible*, if there exists at least one dual feasible $\boldsymbol{\zeta}$. One crucial feature of linear programming problems is the following theorem (see e.g. [46, Theorem IV.6.2 and Theorem IV.7.2])

Theorem 2 (Weak+Strong Duality). *Any primal feasible $\boldsymbol{\xi}$ and any dual feasible $\boldsymbol{\zeta}$ obey*

$$\langle \mathbf{c}, \boldsymbol{\xi} \rangle \geq \langle \boldsymbol{\zeta}, \mathbf{b} \rangle \quad \textit{(weak duality)}. \tag{15}$$

Furthermore, if a given LP is either primal or dual feasible, problems (13) and (14) are equivalent, i.e.

$$\gamma = \beta \quad \textit{(strong duality)}. \tag{16}$$

Strong duality is a very powerful tool, as it allows one to switch between solving primal and dual problems at will. Moreover, the general framework of linear programming is surprsingly versatile, because many nonlinear convex optimization problems can be converted into a corresponding LP. Here, we content ourselves with two examples which will turn out to be important for our analysis.

Example 3 (ℓ_1-norm calculation, [45] p. 294). *Let $\mathbf{x} \in \mathbb{R}^n$ be an arbitrary vector. Then*

$$\|\mathbf{x}\|_{\ell_1} = \min_{\mathbf{t} \in \mathbb{R}^n} \quad \langle \mathbf{1}_n, \mathbf{t} \rangle \tag{17}$$

$$\textit{subject to} \quad -\mathbf{t} \leq \mathbf{x} \leq \mathbf{t}. \tag{18}$$

Note that the constraint (18) *implicitly assures* $\mathbf{t} \geq \mathbf{0}_n$.

Example 4 (ℓ_∞-norm calculation, [45] p. 293). *Let $\mathbf{x} \in \mathbb{R}^n$ be an arbitrary vector. Then*

$$\|\mathbf{x}\|_{\ell_\infty} = \min_{v \in \mathbb{R}} v \tag{19}$$

$$\textit{subject to} \quad -v\mathbf{1}_n \leq \mathbf{x} \leq v\mathbf{1}_n. \tag{20}$$

Note that the constraint $-v\mathbf{1}_n \leq \mathbf{x}$ is redundant if the vector of interest obeys $\mathbf{x} \geq \mathbf{0}_n$. Also, (20) implicitly assures $v \geq 0$.

The primal LPs in examples 3 and 4 are not yet in standard form (13). However, they can be converted into it by applying some straightforward reformulations – we will come back to this later.

Another useful feature of LPs is that different minimization procedures of the above kind can be combined in order to yield an LP for a more complicated optimization problem. An instance of such a combination is the following result which will turn out to be crucial for our analysis.

Lemma 5. *Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_L\} \subset \mathbb{R}^n$ be an arbitrary family of L vectors. Then*

$$\max_{1\leq i\leq L} \|\mathbf{x}_i\|_{\ell_1} = \underset{\substack{\mathbf{t}_1,\ldots,\mathbf{t}_L\in\mathbb{R}^n \\ v\in\mathbb{R}}}{\textit{minimize}} \quad v \qquad \textit{subject to} \quad \left.\begin{array}{l} v \geq \langle \mathbf{1}_n, \mathbf{t}_i\rangle \\ -\mathbf{t}_i \leq \mathbf{x}_i \leq \mathbf{t}_i \end{array}\right\} 1 \leq i \leq L$$

which is a primal LP, albeit not yet in standard form. Also, the constraints implicitly assure $\mathbf{t}_1, \ldots, \mathbf{t}_L \geq \mathbf{0}_n$ and $v \geq 0$.

Proof. We introduce the non-negative auxiliary vector

$$\mathbf{u} := \sum_{i=1}^{L} \|\mathbf{x}_i\|_{\ell_1} \mathbf{e}_i \in \mathbb{R}^L.$$

The equivalence

$$\max_{i=1,\ldots,L} \|\mathbf{x}_i\|_{\ell_1} = \|\mathbf{u}\|_{\ell_\infty}$$

then follows from the definition of the ℓ_∞-norm. Replacing this ℓ_∞-norm calculation by the corresponding LP (example 4 for non-negative vectors) and including L unconstrained ℓ_1-norm calculations – one for each component of $\mathbf{u}$ – as "subroutines" (example 3) yields the desired statement. □

Finally it is worthwhile to mention that constrained norm-minimization, e.g.

$$\beta = \min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_{\ell_1} \quad \text{subject to} \quad A\mathbf{x} \geq \mathbf{c},$$

can also be reformulated as a LP, because the constraint is linear. To this end, simply include the additional linear constraint in the LP for calculating $\|\mathbf{x}\|_{\ell_1}$:

$$\begin{aligned} \gamma = \min_{\mathbf{x},\mathbf{t}\in\mathbb{R}^n} \quad & \langle \mathbf{1}_n, \mathbf{t}\rangle \\ \text{subject to} \quad & -\mathbf{t} \leq \mathbf{x} \leq \mathbf{t} \\ & A\mathbf{x} \geq \mathbf{c}. \end{aligned} \tag{21}$$

Clearly, this is a LP. Pushing this further, one can also handle certain types of non-linear constraints, e.g.

$$\tilde{\gamma} = \min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x}\|_{\ell_p} \quad \text{subject to} \quad \|A\mathbf{x}\|_{\ell_q} \leq c$$

for $p, q \in \{1, \infty\}$ within the linear programming formalism.

USEFUL RESULTS REGARDING LP'S

We can now use these concepts and techniques to obtain a linear programming formalism for a particular family of convex optimization problems that is relevant for our analysis. As detailed in the following two sections, the measures of direct causal influence (4) and of measurement dependence (5) can be cast as a ℓ_∞-norm and ℓ_1-norm, respectively. This in turn allows us to state the associated equivalent dual problem for the minimization of each of these two measures, which is the scope of the following theorems.

Theorem 6. *Let A be a real $m \times n$-matrix, $\{M_i\}_{i=1}^L$ a family of L real valued $k \times n$-matrices and let $\mathbf{p} \in \mathbb{R}^m$ be an arbitrary vector. Then, the convex optimization problem*

$$\begin{aligned} \gamma = \min_{\mathbf{q}\in\mathbb{R}^n} \quad & \max_{1\leq i\leq L} \|M_i\mathbf{q}\|_{\ell_1} \\ \textit{subject to} \quad & A\mathbf{q} = \mathbf{p} \\ & \langle \mathbf{1}_n, \mathbf{q}\rangle = 1 \\ & \mathbf{q} \geq 0 \end{aligned}$$

can be reformulated as a primal LP. Its associated dual problem is given by

$$\begin{aligned} \underset{\substack{\mathbf{y}_i\in\mathbb{R}^k,\mathbf{z}\in\mathbb{R}^m \\ w_i,u\in\mathbb{R}}}{\text{maximize}} \quad & \langle \mathbf{p}, \mathbf{z}\rangle + u \\ \textit{subject to} \quad & A^T\mathbf{z} + u\mathbf{1}_n \leq \sum_{i=1}^{L} M_i^T \mathbf{y}_i \\ & -w_i\mathbf{1}_k \leq \mathbf{y}_i \leq w_i\mathbf{1}_k \quad i = 1, \ldots, L \\ & \sum_{i=1}^{L} w_i \leq 1, \\ & w_1, \ldots, w_L \geq 0. \end{aligned}$$

Proof. Combining Lemma 5 – for $\mathbf{x}_i = M_i\mathbf{q} \in \mathbb{R}^k$ for $i = 1, \ldots, L$ – with the constrained minimization argument from (21) shows that the convex optimization problem (22) is equivalent to solving

$$\begin{aligned} \underset{\substack{\mathbf{t}_1,\ldots,\mathbf{t}_L\in\mathbb{R}^k,\mathbf{q}\in\mathbb{R}^n \\ v\in\mathbb{R}}}{\text{minimize}} \quad & v \\ \text{subject to} \quad & A\mathbf{q} = \mathbf{p} \\ & \langle \mathbf{1}_n, \mathbf{q}\rangle = 1 \\ & \left.\begin{array}{l} v \geq \langle \mathbf{1}_k, \mathbf{t}_i\rangle \\ -\mathbf{t}_i \leq M_i\mathbf{q} \leq \mathbf{t}_i \end{array}\right\} i = 1, \ldots, L \\ & \mathbf{q} \geq 0 \end{aligned} \tag{22}$$

which is clearly a LP. Note that the remaining optimization variables $v \in \mathbb{R}$ and $\mathbf{t}_i \in \mathbb{R}^k$ are also implicitly constrained to be non-negative. So, in order to convert (22) into a primal LP in standard form (13), we define

$$\boldsymbol{\xi} := v \oplus \bigoplus_{i=1}^{L} \mathbf{t}_i \oplus \mathbf{q}, \quad \mathbf{c} := 1 \bigoplus_{i=1}^{L} \mathbf{0}_k \oplus \mathbf{0}_n \quad \text{and}$$

$$\mathbf{b} := (0)^{\oplus L} \oplus (\mathbf{0}_k \oplus \mathbf{0}_k)^{\oplus L} \oplus \mathbf{p} \oplus (-\mathbf{p}) \oplus 1 \oplus (-1).$$

Counting the dimensions of the resulting vector spaces reveals $\boldsymbol{\xi}, \mathbf{c} \in \mathbb{R}^{1+Lk+n}$ and $\mathbf{b} \in \mathbb{R}^{L+2Lk+2m+2}$. Also, the (implicit and explicit) non-negativity constraints on $v, \mathbf{t}_1, \ldots, \mathbf{t}_L$ and $\mathbf{q}$ guarantee $\boldsymbol{\xi} \geq \mathbf{0}_{1+Lk+n}$. Due to our choice of $\mathbf{b}$, we can incorporate all relevant constraints of (22) in the compact expression

$$\Phi \boldsymbol{\xi} \geq \mathbf{b},$$

where Φ is the $(L + 2Lk + 2m + 2) \times (1 + Lk + n)$-matrix defined by

$$\Phi = \begin{pmatrix} 1 & -\mathbf{1}_k^T & \mathbf{0}_k^T & \cdots & \mathbf{0}_k^T & \mathbf{0}_n^T \\ \vdots & & & & & \vdots \\ 1 & \mathbf{0}_k^T & \cdots & \mathbf{0}_k^T & -\mathbf{1}_k^T & \mathbf{0}_n^T \\ \mathbf{0}_k & \mathbb{1}_{k\times k} & \mathbb{O}_{k\times k} & \cdots & \mathbb{O}_{k\times k} & M_1 \\ \mathbf{0}_k & \mathbb{1}_{k\times k} & \mathbb{O}_{k\times k} & \cdots & \mathbb{O}_{k\times k} & -M_1 \\ \vdots & & & & & \vdots \\ \mathbf{0}_k & \mathbb{O}_{k\times k} & \cdots & \mathbb{O}_{k\times k} & \mathbb{1}_{k\times k} & M_L \\ \mathbf{0}_k & \mathbb{O}_{k\times k} & \cdots & \mathbb{O}_{k\times k} & \mathbb{1}_{k\times k} & -M_L \\ \mathbf{0}_m & \mathbb{O}_{m\times k} & \cdots & \cdots & \mathbb{O}_{m\times k} & A \\ \mathbf{0}_m & \mathbb{O}_{m\times k} & \cdots & \cdots & \mathbb{O}_{m\times k} & -A \\ 0 & \mathbf{0}_k^T & \cdots & \cdots & \mathbf{0}_k^T & \mathbf{1}_n^T \\ 0 & \mathbf{0}_k^T & \cdots & \cdots & \mathbf{0}_k^T & -\mathbf{1}_n^T \end{pmatrix}$$

in the (extended) standard bases of the spaces $\mathbb{R}^{1+Lk+n}$ and $\mathbb{R}^{L+2Lk+2m+2}$. Our definitions of $\boldsymbol{\xi}, \mathbf{c}, \mathbf{b}$ and Φ now indeed convert (22) into primal standard form (13). Its dual then simply corresponds to (14) which can be further simplified. The structure of $\mathbf{b}$ suggests decomposing the dual variable $\boldsymbol{\zeta} \in \mathbb{R}^{L+2Lk+2m+2}$ into

$$\boldsymbol{\zeta} := \bigoplus_{i=1}^{L} w_i \bigoplus_{i=1}^{L} \left(\mathbf{y}_i' \oplus \mathbf{y}_i''\right) \oplus \mathbf{z}' \oplus \mathbf{z}'' \oplus u' \oplus u'' \qquad (23)$$

with $w_i, u', u'' \in \mathbb{R}$, $\mathbf{y}_i', \mathbf{y}_i'' \in \mathbb{R}^k$ and $\mathbf{z}', \mathbf{z}'' \in \mathbb{R}^m$. Using this decomposition of $\boldsymbol{\zeta}$, we obtain the following constraints from $\Phi^T \boldsymbol{\zeta} \leq \mathbf{c}$:

$$\begin{aligned} A^T(\mathbf{z}' - \mathbf{z}'') + \mathbf{1}_n(u' - u'') &\leq \sum_{i=1}^{L} M_i \left(\mathbf{y}_i'' - \mathbf{y}_i'\right), \\ \mathbf{y}_i' + \mathbf{y}_i'' &\leq w_i \mathbf{1}_k \quad \text{for } i = 1, \ldots, L, \\ \sum_{i=1}^{L} w_i &\leq 1. \end{aligned}$$

Also, due to $\boldsymbol{\zeta} \geq \mathbf{0}_{L+2Lk+2l+2}$, all the optimization variables are non-negative. The objective function corresponds to

$$\langle \boldsymbol{\zeta}, \mathbf{b} \rangle = \langle \mathbf{p}, \mathbf{z}' - \mathbf{z}'' \rangle + u' - u''.$$

The particular form of objective function and constraints suggests to replace the non-negative variables $\mathbf{z}', \mathbf{z}'' \in \mathbb{R}^m$ and $u', u'' \in \mathbb{R}$ by

$$\mathbf{z} := \mathbf{z}' - \mathbf{z}'' \quad \text{and} \quad u := u' - u''$$

which are not constrained to be non-negative anymore. Also, $\mathbf{y}_i' + \mathbf{y}_i'' \leq w_i \mathbf{1}_k$ together with $\mathbf{y}_i', \mathbf{y}_i'' \geq 0$ implies the equivalent constraint

$$-w_i \mathbf{1}_k \leq \mathbf{y}_i'' - \mathbf{y}_i' \leq w_i \mathbf{1}_k$$

for all $1 \leq i \leq L$. This motivates to define $\mathbf{y}_i := \mathbf{y}_i'' - \mathbf{y}_i'$ which is bounded by the above inequality chain, but also not constrained to be non-negative. Putting everything together yields the desired statement □

Theorem 7. *Let A be a real valued $m \times n$ matrix, $\{M_i\}_{i=1}^{L}$ be a family of real valued $k \times n$-matrices, N a real valued $l \times n$-matrix and let $\mathbf{p} \in \mathbb{R}^m$ as well as $c \in \mathbb{R}$ be arbitrary. The convex optimization problem*

$$\begin{aligned} \gamma = \min_{\mathbf{q} \in \mathbb{R}^n} \quad & \|N\mathbf{q}\|_{\ell_\infty} \\ \textit{subject to} \quad & \max_{1 \leq i \leq L} \|M_i \mathbf{q}\|_{\ell_1} \leq c \\ & A\mathbf{q} = \mathbf{p} \\ & \langle \mathbf{1}_n, \mathbf{q} \rangle = 1 \\ & \mathbf{q} \geq 0 \end{aligned} \qquad (24)$$

can be converted into a primal LP. Its associated dual LP corresponds to

$$\begin{aligned} \beta = \max_{\substack{\mathbf{x} \in \mathbb{R}^l, \mathbf{y}_i \in \mathbb{R}^k, \mathbf{z} \in \mathbb{R}^m \\ u, v, w_i \in \mathbb{R}}} \quad & \langle p, z \rangle + u - cv \\ \textit{subject to} \quad & A^T \mathbf{z} + u \mathbf{1}_n \leq \sum_{i=1}^{L} M_i^T \mathbf{y}_i + N^T \mathbf{x} \\ & -w_i \mathbf{1}_k \leq \mathbf{y}_i \leq w_i \mathbf{1}_k \quad i = 1, \ldots, L \\ & \sum_{i=1}^{L} w_i \leq v \\ & \|\mathbf{x}\|_{\ell_1} \leq 1 \\ & w_1, \ldots, w_L, v \geq 0. \end{aligned} \qquad (25)$$

Proof. Proceeding along similar lines as in the previous proof one can show that (24) is equivalent to solving

$$\begin{aligned} \underset{\substack{\mathbf{t}_1, \ldots, \mathbf{t}_L \in \mathbb{R}^k, \mathbf{q} \in \mathbb{R}^n \\ v, \tilde{v} \in \mathbb{R}}}{\text{minimize}} \quad & \tilde{v} \\ \text{subject to} \quad & -\tilde{v} \mathbf{1}_l \leq N\mathbf{q} \leq \tilde{v} \mathbf{1}_l \\ & v \leq c \\ & \left.\begin{aligned} & v \geq \langle \mathbf{1}_k, \mathbf{t}_i \rangle \\ & -\mathbf{t}_i \leq M_i \mathbf{q} \leq \mathbf{t}_i \end{aligned}\right\} i = 1, \ldots, L \\ & A\mathbf{q} = \mathbf{p} \\ & \langle \mathbf{1}_n, \mathbf{q} \rangle = 1 \\ & \mathbf{q} \geq \mathbf{0}_n, \end{aligned} \qquad (26)$$

which is again clearly a primal LP. Moreover, it strongly resembles the linear program (22). Indeed, defining

$$\tilde{\mathbf{c}} := 1 \oplus 0 \bigoplus_{i=1}^{L} \mathbf{0}_k \oplus \mathbf{0}_n,$$

and extending $\boldsymbol{\xi}, \mathbf{b}$, as well as Φ from the proof of Theorem 6 to

$$\tilde{\boldsymbol{\xi}} := \tilde{v} \oplus \boldsymbol{\xi}, \quad \tilde{\mathbf{b}} := \mathbf{0}_l \oplus \mathbf{0}_l \oplus (-c) \oplus \mathbf{b}$$

and

$$\widetilde{\Phi} = \begin{pmatrix} \mathbf{1}_l \oplus \mathbf{1}_l \oplus 0 & B \\ \mathbf{0}_{L+2Lk+2m+2} & \Phi \end{pmatrix},$$

where

$$B := \begin{pmatrix} \mathbf{0}_l & \mathbb{O}_{l\times k} & \cdots & \cdots & \mathbb{O}_{l\times k} & N \\ \mathbf{0}_l & \mathbb{O}_{l\times k} & \cdots & \cdots & \mathbb{O}_{l\times k} & -N \\ -1 & \mathbf{0}_k^T & \cdots & \cdots & \mathbf{0}_k^T & \mathbf{0}_n^T \end{pmatrix}$$

converts (26) into primal standard form. Going to the dual and simplifying it in a similar way as shown in the previous proof – decompose $\tilde{\boldsymbol{\xi}}$ into $\mathbf{x}' \oplus \mathbf{x}'' \oplus v \oplus \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ was defined in (23) – yields the desired statement upon noticing that $\langle \mathbf{1}_l, \mathbf{x}' + \mathbf{x}'' \rangle \leq 1$ together with $\mathbf{x}', \mathbf{x}'' \geq \mathbf{0}_l$ is equivalent to demanding that $\mathbf{x} := \mathbf{x}' - \mathbf{x}''$ obeys $\|\mathbf{x}\|_{\ell_1} \leq 1$, but is not constrained to be non-negative anymore. □

Corollary 8. *Suppose the ℓ_1-norm constraint in the convex optimization (24) is omitted, then the corresponding dual LP simplifies to*

$$\begin{aligned} \beta = \max_{\mathbf{x}\in\mathbb{R}^l, \mathbf{z}\in\mathbb{R}^m, u\in\mathbb{R}} \quad & \langle \mathbf{p}, \mathbf{z} \rangle + u \\ \textit{subject to} \quad & A^T \mathbf{z} + u \mathbf{1}_n \leq N^T \mathbf{x} \\ & \|\mathbf{x}\|_{\ell_1} \leq 1. \end{aligned} \tag{27}$$

If the normalization condition $\langle \mathbf{1}_n, \mathbf{q} \rangle = 1$ is dropped as well, the optimization parameter u assumes 0 and need not be considered in the dual optimization.

Proof. Omitting the ℓ_1-norm constraint is equivalent to letting the constraint c go to infinity. Since $(-cv)$ is part of the dual's objective function (25), this limit enforces $v = 0$. This in turn demands $w_i = 0$ and consequently $\mathbf{y}_i = \mathbf{0}_k$ for all $i = 1, \ldots, L$. As a result, we obtain the first desired statement.

The second simplification requires a closer look at the proof of Theorem 7. Doing so reveals that the constraint $\langle \mathbf{1}_n, \mathbf{q} \rangle = 1$ results in the additional dual optimization parameter u. Omitting this constraint in the primal therefore implies that u has to be dropped accordingly. □

Finally we are going to present the derivation of the second part of Theorem 1, namely that solving an arbitrary feasible primal LP (in standard form), is equivalent to maximizing the dual problem over finitely many points – the vertices of the dual feasible set.

Proposition 9. *Consider a primal feasible LP whose optimal value γ is bounded from below. Then this optimum is attained at one vertex $\mathbf{d}_i$ of the dual feasible region $\mathcal{D} := \{\zeta \in \mathbb{R}^m : \Phi^T \zeta \leq \mathbf{c}, \zeta \geq \mathbf{0}_m\}$:*

$$\gamma = \beta = \max_{1\leq i\leq K} \langle \mathbf{d}_i, \mathbf{b} \rangle,$$

Possible unbounded directions (rays) of $\mathcal{D}$ can be safely ignored.

Note that all the measures we consider – (4), (5) and (12) in the main text – are non-negative by construction. Consequently, any reformulation of calculating (or optimizing over) these measures as a primal LP results in a bounded optimal value $\gamma \geq 0$. Hence, Proposition 9 is applicable, provided there is at least one hidden variable that reproduces the observed distribution, thus establishing that the LP is primal feasible.

Proposition 9 establishes that the relevant part of the dual feasible region is bounded. It can be deduced from duality – Theorem 2 – and is standard. In order to be self-contained, we provide a slightly different proof that exploits the geometry of linear programs more explicitly.

Proof of Proposition 9. The fact that the primal LP is feasible and bounded assures that there is at least one dual feasible point via strong duality – Theorem 2. The dual feasible region $\mathcal{D}$ is defined by $n + m$ linear inequalities and therefore has the structure of a convex polyhedron. We have just established that this polyhedron is non-empty, but it is not necessarily bounded. To see this, suppose for now that $\mathbf{c} \geq \mathbf{0}_n$ holds (this is not necessary, but will simplify our argument). If Φ^T has a non-trivial kernel, then each element $\bar{\boldsymbol{\zeta}} \in \ker(\Phi^T) \cap \mathbb{R}_+^m$ is not affected by the linear inequalities, because

$$\bar{\boldsymbol{\zeta}} \geq \mathbf{0}_m \quad \text{and} \quad \Phi^T \bar{\boldsymbol{\zeta}} = \mathbf{0}_n \leq \mathbf{c}.$$

Consequently, $\mathcal{D}$ contains the convex cone $\mathcal{C} := \ker(\Phi^T) \cap \mathbb{R}_+^m$. Conversely, it is easy to show that the unbounded part of $\mathcal{D}$ is fully contained in $\mathcal{C}$. This allows us to make a Minkowski decomposition

$$\mathcal{D} = \mathcal{C} + \mathcal{P} = \{c + p : c \in C, p \in P\},$$

where $\mathcal{C}$ is the unbounded conic part and $\mathcal{P}$ denotes the polyhedron's remaining part. We now aim to show that elements $\bar{\zeta} \in \mathcal{C}$ do not contribute to the actual optimization procedure and can therefore safely be ignored.

To this end, we combine the primal problem's (13) constraint $\Phi\boldsymbol{\xi} - \mathbf{b} \geq \mathbf{0}_m$ with the dual constraint $\boldsymbol{\zeta} \geq \mathbf{0}_m$ to obtain $\langle \boldsymbol{\zeta}, \mathbf{b} \rangle \leq \langle \boldsymbol{\zeta}, \Phi\boldsymbol{\xi} \rangle$ for any primal feasible $\boldsymbol{\xi} \in \mathbb{R}^n$. Such a $\boldsymbol{\xi}$ is guaranteed to exist due to Theorem 2 and in particular implies for any $\bar{\boldsymbol{\zeta}} \in \mathcal{C}$:

$$\langle \bar{\boldsymbol{\zeta}}, \mathbf{b} \rangle \leq \langle \bar{\boldsymbol{\zeta}}, \Phi\boldsymbol{\xi} \rangle = \langle \Phi^T \bar{\boldsymbol{\zeta}}, \boldsymbol{\xi} \rangle = 0.$$

Here, the last equality is due to $\bar{\boldsymbol{\zeta}} \in \ker\left(\Phi^T\right)$. Therefore elements of $\mathcal{C}$ manifestly do not contribute to the maximization and we can focus on the remaining set $\mathcal{P}$. By construction, $\mathcal{P}$ is a bounded polyhedron and thus a polytope which can be characterized as the convex hull $\mathrm{conv}(\mathbf{d}_1, \ldots, \mathbf{d}_K)$ of its extremal points (Weyl-Minkowski Theorem [46, Corollary 4.3]). However, it is a well known fact that the maximum of a linear (or more generally: any concave) function over a convex polytope is attained at one of its extreme pointes, i.e. vertices. □

RELAXATION OF LOCALITY

In this section we will analyze the relaxation of the locality assumption, as exemplified by the DAGs depicted in Fig. 1b–d. In particular, we will show that evaluating the minimal direct causal influence – see equation (4) in the main text – that is required to simulate a given non-local distribution can be recast as a LP. Consequently, it can be determined efficiently for any observed probability distribution.

We begin analyzing in details the scenario depicted in Fig. 1c. There, the input X of Alice has a direct causal influence over the outcome B of Bob. We consider the general, finite case where Alice has m_x inputs and o_a outputs, that is, $x = 0, \ldots, m_x - 1$ and $a = 0, \ldots, o_a - 1$ (and analogously for Bob). Variations of this signalling model can be easily constructed and will be briefly discussed at the end of this section.

The signalling model in Fig. 1c requires a hidden variable λ assuming $n = o_a^{m_x} o_b^{m_x m_y}$ possible values. The causal structure assures $a = f_A(x, \lambda)$ which resembles the LHV model (Fig. 1a). This is not the case for b, which can depend on x,y and λ – i.e. $b = f_B(x, y, \lambda)$. Consequently there are $o_a^{m_x}$ possible deterministic functions f_A and $o_b^{m_x m_y}$ possible deterministic functions f_B. In turn, we can split up the hidden variable into $\lambda = (\lambda_a, \lambda_b) = (\alpha_0, \ldots, \alpha_{m_x - 1}, \beta_{0,0}, \beta_{0,1}, \ldots, \beta_{m_x-1, m_y-1})$ where $\alpha_x = 0, \ldots, o_a - 1$ determines the value of a given x. Similarly, $\beta_{x,y} = 0, \ldots, o_b - 1$ specifies the value of b given x and y. Following (1) the observed distribution can be decomposed in the following way:

$$p(a, b|x, y) = \sum_{\lambda} p(a|x, \lambda) p(b|x, y, \lambda) p(\lambda). \tag{28}$$

Given such a signalling model and some observed constraints, our task is to find the minimum value of $\mathcal{C}_{X \to B}$. Similarly to (4), this quantity can be defined as

$$\mathcal{C}_{X \to B} = \sup_{b, y, x, x'} \sum_{\lambda} p(\lambda) |p(b|do(x), y, \lambda) - p(b|do(x'), y, \lambda)|, \tag{29}$$

which quantifies the amount of signalling required to explain the observation. Moving on, we note that

$$\begin{aligned} &\sum_{\lambda} p(\lambda) |p(b|do(x), y, \lambda) - p(b|do(x'), y, \lambda)| \\ &= \sum_{\lambda} p(\lambda) |\delta_{b, f_B(x, y, \lambda)} - \delta_{b, f_B(x', y, \lambda)}| \\ &= \sum_{i} q_i v_i = \langle \mathbf{v}, \mathbf{q} \rangle, \end{aligned} \tag{30}$$

where we have identified $p(\lambda)$ with the n-dimensional vector $\mathbf{q}$ via $\langle \mathbf{e}_i, \mathbf{q} \rangle = p(\lambda_i)$. The vector $\mathbf{v} = \mathbf{v}(x, x', y, b)$ only consists of 1's and 0's and fully characterizes the action of the Kronecker-symbols in (30). By doing so, the measure of causal influence (29) can be recast as

$$\mathcal{C}_{X \to B} = \max_{i = 1, \ldots, L} \langle \mathbf{q}, \mathbf{v}_i \rangle = \|C\mathbf{q}\|_{\infty}. \tag{31}$$

Here, the index i parametrizes one of the L possible instances of (x, x', y, b) with $x \neq x'$ and $\mathbf{v}_i = v(x, x', y, b)$ denotes the vector corresponding to that instance. The last equality in (31) then follows from introducing $C := \sum_{i=1}^{L} |\mathbf{e}_i\rangle\langle \mathbf{v}_i|$ and the definition of the ℓ_∞-norm. Consequently, minimizing $\mathcal{C}_{X \to B}$ over all hidden variables that are compatible with our observations is equivalent to solving

$$\begin{aligned} \underset{\mathbf{q} \in \mathbb{R}^n}{\text{minimize}} \quad & \|C\mathbf{q}\|_{\infty} && (32) \\ \text{subject to} \quad & VT\mathbf{q} = V\mathbf{p} \\ & \langle \mathbf{1}_n, \mathbf{q} \rangle = 1 && (33) \\ & \mathbf{q} \geq \mathbf{0}_n. \end{aligned}$$

Corollary 8 assures that this optimization problem can be translated into a LP in standard form. As already mentioned in the main text, $V\mathbf{p}$ denotes the vector representing the correlations under consideration – the probability distribution itself ($V = \mathbb{1}$) or a function of it, e.g., a Bell inequality ($V = |\mathbf{e}_1\rangle\langle \mathbf{b}|$ for some $b \in \mathbb{R}^m$) – and the matrix VT maps the underlying hidden variable states to the actually observed vector $V\mathbf{p}$.

Given any observed distribution $V\mathbf{p}$ of interest, one can easily implement this linear program and solve it efficiently. However, we are also interested in deriving an analytical solution which is valid for any vector $\mathbf{p}$ encoding the full probability distribution $p(a, b|x, y)$. Subjecting to the full probability distribution $\mathbf{p}$ in particular guarantees that the normalization constraint (33) is already assured by $T\mathbf{q} = \mathbf{p}$. This allows for dropping this constraint without loss of generality. Proposition 9 serves precisely the purpose of obtaining such an analytical expression, as it – in combination with Corollary 8 – assures that solving (32) is equivalent to evaluating

$$\max_{1\leq i\leq K} \langle \mathbf{d}_i, V\mathbf{p}\rangle,$$

where $\{\mathbf{d}_i\}_{i=1}^K$ denotes the vertices of the dual feasible region in (27). Standard algorithms like PORTA [48] allow for evaluating these extremal points. We have performed such an analysis for the particular case of the CHSH scenario ($m_x = m_y = o_a = o_b = 2$). We list all the 13 vertices of the LP's dual feasible region in Table I. Nicely, we see that all the extremal points can be divided into three types: i) the trivial vector $\mathbf{0}_m$, ii) the symmetries of the CHSH inequality vector, for example

$$p^{AB}_{00|00} + p^{AB}_{00|01} + p^{AB}_{00|10} - p^{AB}_{00|11} - p^{A}_{0|0} - p^{B}_{0|0} \tag{34}$$

and iii) the non-signalling conditions, for instance

$$-p^{AB}_{01|00} - p^{AB}_{11|00} + p^{AB}_{01|10} + p^{AB}_{11|10}. \tag{35}$$

Here, we have used the short hand notation $p^{AB}_{ab|xy} = p(a,b|x,y)$ and similarly for the marginals.

For any non-signalling distribution, the conditions of the third type vanish and the corresponding vertices need not be considered. Therefore we arrive at the result stated in the main text, namely

$$\min \mathcal{C}_{X\to B} = \max\left[0, \text{CHSH}\right],$$

where the maximum is taken over all the eight symmetries of the CHSH inequality.

Having such a causal interpretation of the CHSH inequality at hand, one can wonder the same holds true for other Bell inequalities, for instance the ($I_{3322} \leq 0$)-inequality [26] (three inputs for Alice and Bob with two outcomes each). Dwelling on the model in Fig. 1c we show that the I_{3322} inequality only provides a lower bound to the actual value of $\mathcal{C}_{X\to B}$ required to simulate a given nonlocal distribution. This is illustrated in Fig. 2. To be more concrete, we consider the particular full probability distribution

$$p(a,b|x,y) = vp_{\text{PR}} + (1-v)p_{\text{W}}, \tag{36}$$

where

$$p_{\text{PR}}(a,b|x,y) = \begin{cases} 1/2 & \text{if } a+b=1 \mod 2,\ x+y=3,\\ 1/2 & \text{if } a+b=0 \mod 2,\ x+y\neq 3,\\ 0 & \text{otherwise,}\end{cases}$$

denotes the generalization of the PR box maximally violating the I_{3322}-inequality (achieving $I_{3322} = 1$) and

$$p_{\text{W}}(a,b|x,y) = 1/4$$

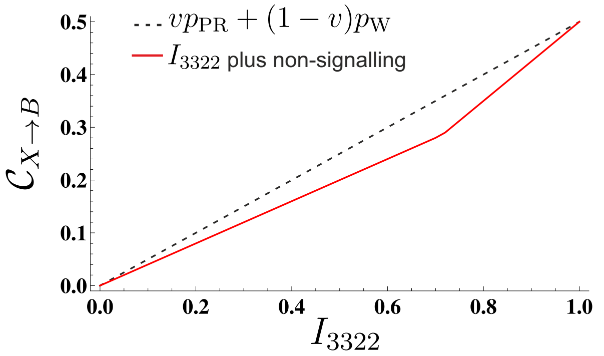


FIG. 2. The value of $\min \mathcal{C}_{X\to B}$ as function of the I_{3322} value. The black curve represents the case where the full probability distribution defined in (36) is taken into account. The red curve is obtained by minimizing $\mathcal{C}_{X\to B}$ for a given value of I_{3322} subject to non-signalling and normalization constraints.

denotes the uniform distribution (achieving $I_{3322} = -1$). Such a full probability distribution results in $I_{3322} = 2v - 1$. We numerically see that

$$\mathcal{C}_{X\to B} = \max\left[0, (2v-1)/2\right] = \max\left[0, I_{3322}/2\right]$$

holds, if we take into account the full probability distribution. However, if we instead only impose a fixed value of the I_{3322}-inequality (plus nonsignalling and normalization constraints) we numerically (see Fig. 2) arrive at

$$\min \mathcal{C}_{X\to B} = \begin{cases} 0 & \text{for } I_{3322} \leq 0,\\ (2/5) * I_{3322} & \text{for } 0 \leq I_{3322} \leq 0.714,\\ (1/4) * (3I_{3322} - 1) & \text{for } 0.714 \leq I_{3322} \leq 1.\end{cases}$$

This shows that different distributions achieving the same value for I_{3322} may have quite different requirements in order to be simulated. Moreover, this result highlights another nice aspect of our framework. Unlike the results in [11–17], it can take into account the full probability distribution, not just the value of a specific Bell inequality. This allows for a much more accurate description.

An almost identical analysis can be done for the model displayed in Fig. 1b. Using (1), the observed distribution can be decomposed as:

$$p(a,b|x,y) = \sum_{\lambda} p(a|x,\lambda)p(b|a,y,\lambda). \tag{37}$$

Using the measure of direct causal influence (4) for $\mathcal{C}_{A\to B}$, revisiting the CHSH scenario, we can once more conclude

$$\min \mathcal{C}_{A\to B} = \max\left[0, \text{CHSH}\right]. \tag{38}$$

In particular, this implies that such a model – where one of the parties communicates its outcomes – is capable

List of extremal points																
#	p^{00}_{00}	p^{01}_{00}	p^{00}_{10}	p^{00}_{11}	p^{01}_{00}	p^{01}_{01}	p^{01}_{10}	p^{01}_{11}	p^{10}_{00}	p^{10}_{01}	p^{10}_{10}	p^{10}_{11}	p^{11}_{00}	p^{11}_{01}	p^{11}_{10}	p^{11}_{11}
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	-1/2	0	1/2	0	-1/2	-1	-1/2	0	-1/2	0	1/2	0	1/2	0	-1/2
3	0	-1/2	0	1/2	0	-1/2	-1	-1/2	0	1/2	0	-1/2	0	-1/2	0	1/2
4	0	-1/2	0	1/2	0	1/2	0	-1/2	0	-1/2	0	1/2	0	-1/2	-1	-1/2
5	0	-1/2	0	1/2	0	1/2	0	-1/2	0	1/2	0	-1/2	-1	-1/2	0	-1/2
6	0	1/2	0	-1/2	0	-1/2	0	1/2	0	-1/2	0	1/2	0	-1/2 -	1	-1/2
7	0	1/2	0	-1/2	0	-1/2	0	1/2	0	1/2	0	-1/2	-1	-1/2	0	-1/2
8	0	1/2	0	-1/2	0	1/2	1	1/2	0	-1/2	0	1/2	-1	-1/2	-1	-3/2
9	0	1/2	0	-1/2	0	1/2	1	1/2	0	1/2	0	-1/2	-1	-3/2	-1	-1/2
10	0	-1	0	-1	0	0	0	0	0	1	0	1	0	0	0	0
11	0	0	0	0	0	-1	0	-1	0	0	0	0	0	1	0	1
12	0	0	0	0	0	1	0	1	0	0	0	0	0	-1	0	-1
13	0	1	0	1	0	0	0	0	0	-1	0	-1	0	0	0	0

TABLE I. Extremal points for the feasible region in the dual problem (27) associated with the CHSH scenario. In the notation above, p^{xy}_{ab} corresponds to $p(a,b|x,y)$. The extremal points 2-9 can be easily seen to correspond to the symmetries of the CHSH inequality. Take for instance point 2 which can be written as the CHSH operator in (34). The extremal points 10-13 correspond to the non-signalling conditions. For instance, point 10 corresponds to (35) and is zero for any non-signalling distribution.

of simulating any nonlocal distributions in the CHSH scenario.

Interestingly, things change drastically if we move on to the I_{3322} scenario. It is worthwhile to point out that model (37) restricts the hidden variables to a region characterized by finitely many inequalities. Therefore, analogously to the usual LHV model (2), the feasible region is a polytope. Using the software PORTA we found different classes of non-trivial inequalities that define the compatibility region of this model. As shown in the main text – equation (10) – one of these inequalities corresponds to

$$I_{A\to B} = \langle E_{00}\rangle - \langle E_{02}\rangle - \langle E_{11}\rangle + \langle E_{12}\rangle - \langle E_{20}\rangle + \langle E_{21}\rangle \leq 4,$$

where $E_{xy} = \langle A_x B_y\rangle = \sum_{a,b}(-1)^{a+b}p(a,b|x,y)$. We now show that this inequality can be violated by any quantum state $|\psi\rangle = \sqrt{\epsilon}|00\rangle + \sqrt{(1-\epsilon)}|11\rangle$ with $\epsilon \neq 0,1$. To arrive at such a statement, it suffices to consider that Alice and Bob perform projective measurements on the X-Z plane of the Bloch sphere. More concretely, Alice measures observables of the form $O^A_x = \cos(\theta^A_x)Z + \sin(\theta^A_x)X$ and so does Bob whose observables we denote by O^B_x. Here, X and Z refer to the Pauli matrices. For such particular measurements, the correlators $E_{xy} = \langle A_x B_y\rangle$ simply correspond to

$$E_{xy} = \cos(\theta^A_x)\cos(\theta^B_y) + 2\sqrt{\epsilon(1-\epsilon)}\sin(\theta^A_x)\sin(\theta^B_y).$$

Choosing the angles such that $\theta^A_0 = 0$, $\theta^A_1 = \pi$, $\theta^A_2 = \pi/2$, $\theta^B_0 = 0$ and $\theta^B_2 = -\pi$ we obtain

$$I_{A\to B} = 3 + \cos(\theta^B_1) + 2\sqrt{\epsilon(1-\epsilon)}\sin(\theta^B_1).$$

This expression exceeds 4 for any $\epsilon \neq 0,1$, provided that we choose θ^B_1 sufficiently small compared to $2\sqrt{\epsilon(1-\epsilon)}$. This result shows that even relaxing some of assumptions in Bell's theorem – in this particular case, the fact that Alice outcomes cannot have a direct causal influence over Bob outcomes – may not be enough to causally explain quantum correlations.

A similar analysis can be performed for the communication model of Fig. 1d. Such a model implies the following decomposition of the distribution observed:

$$p(a,b|x,y) = \sum_{\lambda,m} p(a|x,\lambda)p(m|x,a,\lambda)p(b|m,y,\lambda)p(\lambda).$$

Such an expression suggests to decompose the hidden variable into $\lambda = (\lambda_\alpha, \lambda_\beta, \lambda_m)$. By doing so, one can perform an analysis similar to the one above and define a measure of causal influence similar to (4). However, inspired by the communication model of Toner and Bacon [9], we directly proceed to analyzing the amount of communication between Alice and Bob required to classically reproduce the distribution observed. We quantify the information content of a binary message m sent from Alice to Bob via its Shannon entropy $H(m)$. Due to the highly non-linear character of entropies, the optimizations involving $H(m)$ are quite hard in general. Fortunately in the particular case of binary messages, minimizing $H(m)$ is equivalent to minimizing

$$\begin{aligned} p(m=0) &= \sum_{a,x,\lambda} p(m=0|x,a,\lambda)p(a|x,\lambda)p(x)p(\lambda) \quad (39)\\ &= (1/m_x)\sum_{a,\lambda} p(m=0|x,a,\lambda)p(a|x,\lambda)p(\lambda)\\ &= \langle \mathbf{v}, \mathbf{q}\rangle. \end{aligned}$$

Here, we have once more identified $p(\lambda)$ with the vector $\mathbf{q}$ and the components of $\mathbf{v}$ correspond to $v_i =$

$\sum_a p(m=0|x,a,\lambda_i)p(a|x,\lambda_i)$. Also, we have without loss of generality considered a uniform distribution of Alice's inputs – i.e. $p(x) = 1/m_x$ – in the second line. Consequently, the constrained minimization of $p(m=0)$ (and thus $H(m)$) simply corresponds to

$$\begin{aligned} \underset{\mathbf{q}\in\mathbb{R}^n}{\text{minimize}} \quad & \langle \mathbf{v}, \mathbf{q}\rangle \\ \text{subject to} \quad & T\mathbf{q} = \mathbf{p} \\ & \langle \mathbf{1}_n, \mathbf{q}\rangle = 1 \\ & \mathbf{q} \geq \mathbf{0}_n, \end{aligned}$$

which is clearly a primal LP. Computing the extremal points of the dual problem allows us to infer a novel relation between the degree of nonlocality and the minimum communication required to simulate it. Namely, $\min p(m=0) = \max\left[0, \mathrm{CHSH}_\Pi\right]$ which in turn implies

$$\min H(m) = \begin{cases} h\,(\mathrm{CHSH}) & \text{for CHSH} \geq 0, \\ 0 & \text{else.} \end{cases}$$

Here, h denotes the binary entropy given by $h(v) = -v\log_2 v - (1-v)\log_2(1-v)$.

These results on the relaxation of the locality assumption, in addition to fundamental implications and relevance in nonlocal protocols, can also be used to compute the minimum causal influences/communication required to causally explain the nonlocal correlations observed in experimental realizations of Bell's tests where the space-like separation is not achieved [6, 7].

MEASUREMENT DEPENDENCE MODELS

In this section we focus on the measure $\mathcal{M}_{X,Y:\Lambda}$ – see equation (5) in the main text – which quantifies the degree of measurement dependence in a given causal model. Similar to the previous section, we are going to show that determining the minimal degree of measurement dependence required to reproduce a given nonlocal distribution can be done via solving a LP.

To illustrate this, we consider the simplest scenario of measurement dependence in detail. Such a model is displayed in Fig. 1e and involves a bipartite Bell scenario, where the measurement inputs X of Alice and Y of Bob, respectively, can be correlated with the source Λ producing the particles to be measured.

Without loss of generality, we model such correlations by introducing an additional hidden variable μ which serves as a common ancestor for x, y and λ. This suggests to decompose this common ancestor into $\mu = (\mu_x, \mu_y, \mu_\lambda)$. We can assume $x = \mu_x$, $y = \mu_y$ and $\lambda = \mu_\lambda$ without loss of generality (x, y and λ are deterministic functions of their common ancestor μ). If Alice's apparatus has m_x inputs (i.e. $x = 0,\ldots,m_x-1$) and o_a outputs (i.e. $a = 0,\ldots,o_a-1$), and similarly for Bob, $n = m_x m_y o_a^{m_x} o_a^{m_y}$ different instances of μ suffice to fully characterize the common ancestor's influence. Similar to the previous section, we can use this discrete nature of μ to identify any probability distribution $p(\mu) : \Xi \to [0,1]$ uniquely with a non-negative, real vector $\mathbf{q}$ via

$$q_i = \langle \mathbf{e}_i, \mathbf{q}\rangle = p(\mu_i) \quad i = 1,\ldots,n. \tag{40}$$

Likewise, we can rewrite the observed probability distribution $p(a,b|x,y)$ as

$$\begin{aligned} & p(a,b|x,y) \\ &= \frac{1}{p(x,y)}\sum_{\mu,\lambda} p(a|x,\lambda)p(b|y,\lambda)p(x|\mu)p(y|\mu)p(\lambda|\mu)p(\mu) \\ &= \frac{1}{p(x,y)}\sum_{\mu_\lambda} p(a|x,\mu_\lambda)p(b|y,\mu_\lambda)p(\mu_\lambda) \\ &= \langle \mathbf{v}(x,y,a,b,\lambda), \mathbf{q}\rangle. \end{aligned}$$

The usefulness of such vectorial identifications becomes apparent when taking a closer look at the measure of correlation (5). Indeed,

$$\begin{aligned} \mathcal{M} &= \sum_{x,y,\lambda} |p(x,y,\lambda) - p(x,y)p(\lambda)| \\ &= \sum_{x,y,\lambda} |\sum_{\mu} \delta_{\lambda,\mu_\lambda}(\delta_{x,\mu_x}\delta_{y,\mu_y} - p(x,y))p(\mu)| \\ &= \sum_{x,y,\lambda} |\langle \mathbf{v}(x,y,\lambda), \mathbf{q}\rangle| \\ &= \|M\mathbf{q}\|_{\ell_1}, \end{aligned} \tag{41}$$

where M denotes the real $k \times n$ matrix $M = \sum_{j=1}^{k} |\mathbf{e}_j\rangle\langle \mathbf{v}(x,y,\lambda)|$. Note that this matrix implicitly depends on $p(x,y)$. However, $p(x,y)$ is an observable quantity and thus available. Moreover, one is typically interested in the case, where said distribution for the inputs is uniformly distributed – i.e. $p(x,y) = 1/(m_x m_y)$.

It is worthwhile to point out that different measures of measurement dependence have been considered in the literature. For instance, in Ref. [12] the following measure of correlation has been proposed:

$$\mathcal{M}_{\text{Hall}} = \sup_{x,x',y,y'} \sum_{y} |p(\lambda|x,y) - p(\lambda|x',y')|.$$

Similarly to (41), we can rewrite this measure as a ℓ_1-norm, namely

$$\mathcal{M}_{\text{Hall}} = \max_{i=1,\ldots,L} \|M_i\mathbf{q}\|_{\ell_1}.$$

The constrained minimization of both $\mathcal{M}$ and $\mathcal{M}_{\text{Hall}}$ consequently corresponds to the following optimization:

$$\begin{aligned} \underset{\mathbf{q}\in\mathbb{R}^n}{\text{minimize}} \quad & \max_{i=1,\dots,L} \|M_i \mathbf{q}\|_{\ell_1} \\ \text{subject to} \quad & V\mathbf{q} = \tilde{\mathbf{p}} \\ & \langle \mathbf{1}_n, \mathbf{q}\rangle = 1 \\ & \mathbf{q} \geq \mathbf{0}_n, \end{aligned} \tag{42}$$

Theorem 6 assures that such an optimization can be recast as a primal LP in standard form.

In this work we have opted to focus on the measure defined in (41). The reason for that is two-fold. Firstly, such a choice assures $L = 1$ and numerically solving the corresponding LP is substantially faster. The second reason stems from the fact that (41) is proportional to the variational distance between the distributions $p(x, y, \lambda)$ and $p(x, y)p(\lambda)$. Knowledge of the total variational distance allows to lower-bound the mutual information between (X, Y) and Λ via the Pinsker inequality [32, 33]:

$$I(X, Y : \Lambda) \geq \mathcal{M}^2 \log_2 \mathrm{e}.$$

A converse bound on $I(X, Y : \Lambda)$ is obtained by noting that the (linear program) solution to the minimization of $\mathcal{M}$ returns a specific hidden variable model, for which we can readily compute the mutual information.

Using measure (41), we have considered many different Bell scenarios. This was already mentioned in the main text. In particular we refer to Fig. 3 where we consider the CGLMP scenario [31] – a bipartite model, where Alice and Bob measure one out of two observables each of them having d possible outcomes. The corresponding CGLMP inequality is of the form $I_d \leq 2$, where the local bound of 2 and the maximal violation of 4 are independent of the number of possible outcomes d. Imposing the value of the I_d inequality ad imposing non-signalling and the normalization constraints we numerically obtain a very simple relation up to $d = 8$, namely

$$\min \mathcal{M} = \max\left[0, (I_d - 2)/4\right].$$

Conversely, we have also considered specific quantum realizations. For $d = 2, 5, 7$ we have numerically optimized over quantum states and projective measurements maximizing the corresponding I_d inequality. With the resulting quantum probability distribution at hand, we computed $\mathcal{M}$ and inferred lower and upper bounds for $I(X, Y : \Lambda)$ in turn. These results are depicted in Fig. 3 and we refer to the corresponding section in the main text for further insights concerning measurement dependence.

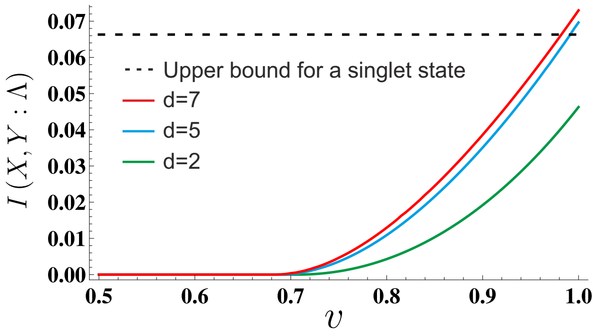


FIG. 3. Upper bound for $I(X, Y : \lambda)$ computed as a function of the visibility V for $d = 2, 5, 7$ (green, blue and red curves, respectively). The black dashed curve correspond to the upper bound $I(X, Y : \Lambda) \approx 0.0663$ obtained in [13] for singlet states. The solid curves correspond to $vp^{Q}_{max} + (1 - v)p_{\mathrm{W}}$ were p^{Q}_{max} was obtained by maximizing the quantum violation of I_d over pure states and projective measurements.

BILOCALITY SCENARIO

In LHV models for multipartite Bell scenarios, it is usually assumed that the same hidden variable is shared among all the parties. That is, a Bell inequality violation rules out any shared LHV. However, in quantum information protocols it is often the case that different parties receive particles produced by independent sources, e.g. in quantum networks [36, 41–44]. It is then natural to focus on LHV models which reproduce the independence structure of the sources. That is, each hidden variable can only be shared between parties receiving particles from the same source. Such models are weaker than general LHV models, i.e. they form a subset of all the models where the hidden variables can be shared arbitrarily among the parties.

A particular case is an entanglement swapping scenario [38] involving three parties A, B and C which receive entangled states from two independent sources. The DAG of Fig. 1f shows an LHV model with independent variables for this scenario. The assumption that the sources are independent, $p(\lambda_1, \lambda_2) = p(\lambda_1)p(\lambda_2)$, is known as bilocality [27, 28]. With this assumption, in analogy with the usual LHV decomposition (2), the correlations for this scenario must fulfil

$$p(a, b, c|x, z) = \sum_{\lambda_1, \lambda_2} p(\lambda_1)p(\lambda_2) \\ p(a|x, \lambda_1)p(b|\lambda_1, \lambda_2)p(c|z, \lambda_2). \tag{43}$$

Note that the set of bilocal correlations is non-convex because of the nonlinearity of the bilocality assumption. This makes the set extremely difficult characterize [27, 28, 36, 37, 49, 50]. In the following, we introduce a measure of relaxation of bilocality, and we show that, despite the non-convex nature of the measure, it can nevertheless be computed by means of a linear program.

For fixed numbers m_x, m_z and o_a, o_b, o_c of the input x, z and output a, b, c values, there is a finite number $n = o_a^{m_x} o_b o_c^{m_z}$ of deterministic strategies. We can label the deterministic strategies for a by symbols $\bar{\alpha} = \alpha_0, \ldots, \alpha_{m_x}$ where α_x is the value of a when the input is x. Similarly, we label the functions for b by β and for c by $\bar{\gamma} = \gamma_0, \ldots, \gamma_{m_z}$. Thus, the distribution over the deterministic strategies can be identified with an n-dimensional vector $\mathbf{q}$, analogous to the case in the main text for usual LHV models. The vector $\mathbf{q}$ then has components $q_{\bar{\alpha},\beta,\bar{\gamma}}$. Defining the marginals

$$\begin{gathered} q^{ac}_{\bar{\alpha},\bar{\gamma}} = \sum_{\beta} q_{\bar{\alpha},\beta,\bar{\gamma}} \\ q^{a}_{\bar{\alpha}} = \sum_{\beta,\bar{\gamma}} q_{\bar{\alpha},\beta,\bar{\gamma}}, \quad q^{c}_{\bar{\gamma}} = \sum_{\beta,\bar{\alpha}} q_{\bar{\alpha},\beta,\bar{\gamma}}, \end{gathered} \tag{44}$$

the bilocality assumption is equivalent to the requirement

$$q^{ac}_{\bar{\alpha},\bar{\gamma}} = q^{a}_{\bar{\alpha}} q^{c}_{\bar{\gamma}}. \tag{45}$$

In analogy with the measure (5) of measurement dependence, the degree of non-bilocality can be measured by how much the distribution over the LHVs fail to comply with this criterion. We define the measure of non-bilocality as

$$\mathcal{M}_{\mathrm{BL}} = \sum_{\bar{\alpha},\bar{\gamma}} |q^{ac}_{\bar{\alpha},\bar{\gamma}} - q^{a}_{\bar{\alpha}} q^{c}_{\bar{\gamma}}|. \tag{46}$$

Clearly $\mathcal{M}_{\mathrm{BL}} = 0$ if and only if the bilocality constraint is fulfilled.

The non-bilocality measure is quadratic in the distribution over the the deterministic strategies. Thus, it is not obvious that linear programming will be helpful in computing $\mathcal{M}_{\mathrm{BL}}$ or that the computation can be made efficient. However, we notice that, for given observed correlations, there are restrictions on the marginals $q^{a}_{\bar{\alpha}}$ and $q^{c}_{\bar{\gamma}}$ imposed by the observed distribution $p(a, b, c|x, z)$ because of the constraint (43) that the LHV must reproduce the observations. This constraint can be written

$$p(a, b, c|x, z) = \sum_{\bar{\alpha},\beta,\bar{\gamma}} \delta_{a,\alpha_x} \delta_{b,\beta} \delta_{c,\gamma_z} q_{\bar{\alpha},\beta,\bar{\gamma}}. \tag{47}$$

Depending on the observed distribution, there may be no or just a few free parameters ν which determine $q^{a}_{\bar{\alpha}} = f_{\bar{\alpha}}(\nu)$. We can then rewrite $\mathcal{M}_{\mathrm{BL}}$ as

$$\mathcal{M}_{\mathrm{BL}}(\nu) = \sum_{\bar{\alpha},\bar{\gamma}} |q^{ac}_{\bar{\alpha},\bar{\gamma}} - f_{\bar{\alpha}}(\nu) q^{c}_{\bar{\gamma}}|. \tag{48}$$

For fixed ν the measure $\mathcal{M}_{\mathrm{BL}}(\nu)$ is linear and its minimum can be found via a linear program, as we now show.

As previously, the first step is to write $\mathcal{M}_{\mathrm{BL}}(\nu)$ as an ℓ_1-norm. For a given value of ν, we can write

$$\begin{aligned} \mathcal{M}_{\mathrm{BL}}(\nu) &= \sum_{\bar{\alpha},\bar{\gamma}} |\sum_{\beta} q_{\bar{\alpha},\beta,\bar{\gamma}} - f_{\bar{\alpha}}(\nu) \sum_{\bar{\alpha}',\beta} q_{\bar{\alpha}',\beta,\bar{\gamma}}| && (49) \\ &= \sum_{\bar{\alpha},\bar{\gamma}} | \sum_{\bar{\alpha}'\beta'\bar{\gamma}'} M^{\nu}_{\bar{\alpha}\bar{\gamma},\bar{\alpha}'\beta'\bar{\gamma}'} q_{\bar{\alpha}'\beta'\bar{\gamma}'} | && (50) \\ &= \|M^{\nu} \mathbf{q}\|_{\ell_1}, && (51) \end{aligned}$$

where M^{ν} is a matrix of dimension $l \times n$, with $l = o_a^{m_x} o_c^{m_z}$ and entries $M^{\nu}_{\bar{\alpha}\bar{\gamma},\bar{\alpha}'\beta'\bar{\gamma}'} = \delta_{\bar{\alpha},\bar{\alpha}'} \delta_{\bar{\gamma},\bar{\gamma}'} - f_{\bar{\alpha}}(\nu) \delta_{\bar{\gamma},\bar{\gamma}'}$ (where $\delta_{\bar{\alpha},\bar{\alpha}'} = \delta_{\alpha_0,\alpha'_0} \cdots \delta_{\alpha_{o_x},\alpha'_{o_x}}$ etc.). Minimisation of $\mathcal{M}_{\mathrm{BL}}(\nu)$ for given, observed correlations $p(a, b, c|x, z)$ is then equivalent to

$$\begin{aligned} \underset{\mathbf{q} \in \mathbb{R}^n}{\text{minimize}} \ & \|M^{\nu} \mathbf{q}\|_1 \\ \text{subject to } & A\mathbf{q} = \mathbf{p} \\ & \langle \mathbf{1}_n, \mathbf{q} \rangle = 1 \\ & \mathbf{q} \geq \mathbf{0}_n, \end{aligned} \tag{52}$$

where $\mathbf{p}$ is the k-dimensionsal vector representing the observed correlations, with $k = o_a o_b o_c m_x m_z$, and A is a $k \times n$ matrix which encodes the constraint (47) that the LHV must reproduce the observations. The entries of A are $A_{abcxz,\bar{\alpha}\beta\bar{\gamma}} = \delta_{a,\alpha_x} \delta_{b,\beta} \delta_{c,\gamma_z}$. From Theorem 6, the minimisation (52) is equivalent to the linear program

$$\begin{aligned} \underset{\mathbf{t} \in \mathbb{R}^l}{\text{minimize}} \ & \langle \mathbf{1}_l, \mathbf{t} \rangle \\ \text{subject to } & -\mathbf{t} \leq M^{\nu} \mathbf{q} \leq \mathbf{t}, \\ & A\mathbf{q} = \mathbf{p}, \\ & \langle \mathbf{1}_n, \mathbf{q} \rangle = 1, \\ & \mathbf{q} \geq \mathbf{0}_n \end{aligned} \tag{53}$$

Thus, minimising $\mathcal{M}_{\mathrm{BL}}(\nu)$ for fixed ν is indeed a linear program. To find the minimum of the measure $\mathcal{M}_{\mathrm{BL}}$ we must minimise also over ν and hence we have an optimisation over a linear program. In order to verify non-bilocality of a given distribution we need to check that the minimum over ν is non-zero, or equivalently that the minimum of $\mathcal{M}_{\mathrm{BL}}(\nu)$ is non-zero for all values of ν in the allowed range. On the other hand, if we find a value of ν such that $\mathcal{M}_{\mathrm{BL}}(\nu) = 0$ this is sufficient to show that the distribution is bilocal (and as a by-product we get an explicit bilocal decomposition).

Bilocality with binary inputs

To illustrate our framework, and to compare with previous results, we now consider the case where the

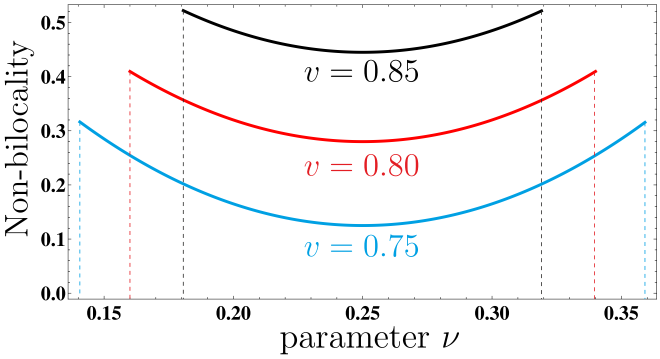


FIG. 4. $\mathcal{M}_{\rm BL}(\nu)$ as a function of ν for three different values of the visibility ($v = 0.75$ (blue curve), $v = 0.80$ (red curve) and $v = 0.85$ (black curve)). The dashed lines correspond to the minimum and maximum values of the parameter ν that are compatible with the probability distribution. We observe that for the specific distribution considered the minimum of $\mathcal{M}_{\rm BL}(\nu)$ is achieved for $\nu = 1/4$.

inputs and ouputs of A and C are all dichotomic ($o_a = o_c = m_x = m_z = 2$), and the output of B takes four values ($o_b = 4$) that we decompose as $b = (b_0, b_1)$ where b_0, b_1 are bits. Furthermore, we consider the distribution [27, 28]

$$p_v\left(a,b,c|x,z\right) = v^2 p\left(a,b,c|x,z\right) + (1-v^2)\frac{1}{16} \tag{54}$$

with

$$p\left(a,b,c|x,z\right) = \frac{1}{16}\left(1 + (-1)^{a+c}\frac{(-1)^{b_0} + (-1)^{x+z+b_1}}{2}\right) \tag{55}$$

This distribution can be obtained by using shared Werner states with visibility v, that is $\varrho = v|\Psi^-\rangle\langle\Psi^-| + (1-v)\mathbb{I}/4$, on which Alice and Charlie perform measurements given by $A_0 = C_0 = \frac{1}{\sqrt{2}}(Z+X)$ and $A_1 = C_1 = \frac{1}{\sqrt{2}}(Z-X)$, while Bob measures in the Bell basis assigning $b_0b_1 = 00, 01, 10, 11$ to $|\Phi^+\rangle$, $|\Phi^-\rangle$, $|\Psi^+\rangle$ and $|\Psi^-\rangle$. As shown in [27, 28] this distribution is non-bilocal. Taking the marginal of (47) gives $p(a|x) = \sum_{\vec{\alpha}} \delta_{a,\alpha_x} q_{\vec{\alpha}}$, explicitly for the distribution (54)

$$\begin{aligned}
p(a=0|x=0) &= q^a_{0,0} + q^a_{0,1} = \frac{1}{2}\\
p(a=0|x=1) &= q^a_{0,0} + q^a_{1,0} = \frac{1}{2}\\
p(a=1|x=0) &= q^a_{1,0} + q^a_{1,1} = \frac{1}{2}\\
p(a=1|x=1) &= q^a_{0,1} + q^a_{1,1} = \frac{1}{2}.
\end{aligned} \tag{56}$$

This implies that $q^a_{0,0} = q^a_{1,1}$ and $q^a_{1,0} = q^a_{0,1} = 1/2 - q^a_{0,0}$ and thus we have a single free parameter $\nu = q^a_{0,0}$. The parameter is further constrained by the full distribution $p(a,b,c|x,z)$. To determine its range we run the following two linear programs

$$\begin{aligned}
&\text{minimize } \langle \mathbf{c}, \mathbf{q}\rangle\\
&\text{subject to } A\mathbf{q} = \mathbf{p},\\
&\qquad\qquad \mathbf{q} \geq \mathbf{0}_n,
\end{aligned} \tag{57}$$

and

$$\begin{aligned}
&\text{maximize } \langle \mathbf{c}, \mathbf{q}\rangle\\
&\text{subject to } A\mathbf{q} = \mathbf{p},\\
&\qquad\qquad \mathbf{q} \geq \mathbf{0}_n,
\end{aligned} \tag{58}$$

where $\langle \mathbf{c}, \mathbf{q}\rangle = q^a_{0,0}$. These two linear programs define a range $\nu_{min} \leq \nu \leq \nu_{max}$. In some particular cases $\nu_{max} = \nu_{min}$, in which case the minimisation over ν is superfluous and the minimum of $\mathcal{M}_{\rm BL}$ is directly given by a linear program and is thus analytical. However, in general these bounds are different. For the distribution (54) with $v = 1$, we have $\nu_{max} = \nu_{min} = 1/4$, while for $v = 0.8$ we have $\nu_{min} = 0.16$ and $\nu = 0.34$. In general what we observe is that for any v, the minimum $\mathcal{M}_{\rm BL}(\nu)$ occurs at $\nu = 1/4$. This is illustrated Fig. 4.

In Fig. 5 we show how the minimum of $\mathcal{M}_{\rm BL}$ depends on the visibility. We also show the value of the bilocality quantity $\mathcal{B} = \sqrt{|I|} + \sqrt{|J|}$ given in [27, 28], where

$$\begin{aligned}
I &= \frac{1}{4}\sum_{x,z=0}^{1} \langle A_x B^0 C_z\rangle,\\
J &= \frac{1}{4}\sum_{x,z=0}^{1} (-1)^{x+z}\langle A_x B^1 C_z\rangle,
\end{aligned} \tag{59}$$

and

$$\langle A_x B^y C_z\rangle = \sum_{a,b_0,b_1,c} (-1)^{a+b_y+c} p(a,b_0,b_1,c|x,z). \tag{60}$$

In [27, 28] it was shown that bilocality implies $\mathcal{B} \leq 1$. For the distribution (54) on the other hand, $I = J = \frac{1}{2}v^2$ and therefore $\mathcal{B} = \sqrt{2}v$. From the numerical results in Fig. 5 one can easily fit the data and find $\min \mathcal{M}_{\rm BL} = \mathcal{B}^2 - 1$. Thus the violation of the bilocality corresponds exactly to how much bilocality must be relaxed to reproduce the observed distribution.

Bilocality with ternary inputs

To sketch how the linear framework could be used in more general bilocality scenarios we consider the case where Alice and Charlie can perform three different measurements. We again consider the case of trivial marginals $p(a|x) = 1/2$. This imposes the following

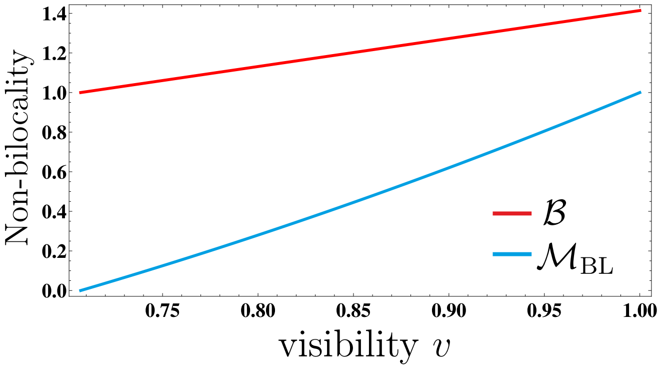


FIG. 5. Minimum of the non-bilocality measure $\mathcal{M}_{\rm BL}$ vs. visibility v (blue). We also show the bilocality quantity $\mathcal{B}$ (red). Our measure can be understood as the amount of correlation between the sources required to simulate the observed correlations.

constraints on $q^a_{\alpha_0,\alpha_1,\alpha_2}$

$$\begin{aligned}
p(a=0|x=0) &= q^a_{0,0,0}+q^a_{0,0,1}+q^a_{0,1,0}+q^a_{0,1,1}=\frac{1}{2}\\
p(a=0|x=1) &= q^a_{0,0,0}+q^a_{0,0,1}+q^a_{1,0,0}+q^a_{1,0,1}=\frac{1}{2}\\
p(a=0|x=2) &= q^a_{0,0,0}+q^a_{0,1,0}+q^a_{1,0,0}+q^a_{1,1,0}=\frac{1}{2}\\
p(a=1|x=0) &= q^a_{1,0,0}+q^a_{1,0,1}+q^a_{1,1,0}+q^a_{1,1,1}=\frac{1}{2}\\
p(a=1|x=1) &= q^a_{0,1,0}+q^a_{0,1,1}+q^a_{1,1,0}+q^a_{1,1,1}=\frac{1}{2}\\
p(a=1|x=2) &= q^a_{0,0,1}+q^a_{0,1,1}+q^a_{1,0,1}+q^a_{1,1,1}=\frac{1}{2},
\end{aligned} \tag{61}$$

which implies that

$$\begin{aligned}
q^a_{011} &= \frac{1}{2}-q^a_{000}-q^a_{001}-q^a_{010}\\
q^a_{101} &= \frac{1}{2}-q^a_{000}-q^a_{001}-q^a_{100}\\
q^a_{110} &= \frac{1}{2}-q^a_{000}-q^a_{010}-q^a_{100}\\
q^a_{111} &= -\frac{1}{2}+2q^a_{000}+q^a_{001}+q^a_{010}+q^a_{100}.
\end{aligned} \tag{62}$$

This means that we now have four free parameters $\nu = (q^a_{000}, q^a_{001}, q^a_{010}, q^a_{100})$. To linearize $\mathcal{M}_{\rm BL}$ in this case, we need to optimize over these four variables.

In practice, given a certain distribution $p(a,b,c|x,z)$, we first fix a certain value for $q^a_{000} = c_0$ in the range $q^{min}_{000} \leq q^a_{000} \leq q^{max}_{000}$. We then solve a linear program to find the bounds for the next free parameter $q^{min}_{001} \leq q^a_{001} \leq q^{max}_{001}$ but now imposing also the constraint that $q^a_{000} = c_0$. Fixed $q^a_{000} = c_0$ and $q^a_{001} = c_1$ we look for the bounds of the next free parameter $q^{min}_{010} \leq q^a_{010} \leq q^{max}_{010}$. We now run the linear program for the remaining free parameter in the range $q^{min}_{100} \leq q^a_{100} \leq q^{max}_{100}$ determined by the probability distribution and the constraints $q^a_{000} = c_0$, $q^a_{001} = c_1$, $q^a_{010} = c_2$.

For a sufficiently good discretization of these continuous free parameters, we can be quite confident about the non-bilocality of the distribution if we find no values for which $\mathcal{M}_{\rm BL} \neq 0$. On the other hand, if we find any values for the free parameters such that $\mathcal{M}_{\rm BL} = 0$, then we can immediately conclude that the distribution is bilocal. To illustrate this we have tested the distribution obtained with two maximally entangled states $|\Psi^-\rangle$ when Alice and Charlie measure the three observables X, Y, Z while Bob measures in the Bell basis. It is possible to show that this distribution is bilocal by setting $q^a_{000} = 0$ and $q^a_{001} = q^a_{010} = q^a_{100} = 1/4$.

Near-optimal quantum tomography: estimators and bounds

Richard Kueng[1,2,3] and Christopher Ferrie[4,1]

[1]Centre for Engineered Quantum Systems, School of Physics, The University of Sydney, Sydney, NSW, Australia
[2]Institute for Theoretical Physics, University of Cologne, Germany
[3]Institute for Physics, University of Freiburg, Germany
[4]Center for Quantum Information and Control, University of New Mexico, Albuquerque, New Mexico, 87131-0001

(Dated: November 24, 2015)

We give bounds on the average fidelity achievable by any quantum state estimator, which is arguably the most prominently used figure of merit in quantum state tomography. Moreover, these bounds can be computed online—that is, while the experiment is running. We show numerically that these bounds are quite tight for relevant distributions of density matrices. We also show that the Bayesian mean estimator is ideal in the sense of performing close to the bound without requiring optimization. Our results hold for all finite dimensional quantum systems.

CONTENTS

I. INTRODUCTION

Inferring a quantum mechanical description of a physical system is equivalent to assigning it a quantum state—a process referred to as *tomography*. Tomography is now a routine task for designing, testing and tuning qubits in the quest of building quantum information processing devices [1]. In determining how "good" one is performing this task, a figure of merit must be reported. By far the most commonly used figure of merit for quantum states is *fidelity* [2, 3]. Nowadays, fidelity is used to compare quantum states and processes in a wide variety of tasks, from quantum chaos to quantum control to the continuous monitoring of quantum systems [4–10]. One might find it surprising, then, that the technique which optimizes performance with respect to fidelity is not known.

For d-dimensional state space,

$$\mathcal{S} := \left\{ \sigma \in L\left(\mathbb{C}^d\right) :\ \sigma \geq 0,\ \mathrm{Tr}(\sigma) = 1 \right\}, \tag{1}$$

the *fidelity* between two states $\rho, \sigma \in \mathcal{S}$ is defined to be [2, 3],

$$F(\rho,\sigma) := \|\sqrt{\rho}\sqrt{\sigma}\|_1^2 = \left[\mathrm{Tr}\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}} \right]^2 . \tag{2}$$

Define the *average fidelity* with respect to some measure $\mathrm{d}\rho$ as $\mathbb{E}_\rho[F(\rho,\sigma)]$[1]. We want the average of this to be as large as possible. Thus, the problem can be succinctly stated as follows:

$$\begin{aligned} \text{maximize} \quad & \mathbb{E}_\rho[F(\rho,\sigma)] \\ \text{subject to} \quad & \mathrm{Tr}(\sigma)=1, \\ & \sigma \geq 0. \end{aligned} \tag{3}$$

In the context of tomography, we think of ρ as the "true state" and σ as the estimated state. An *estimator* is a function from the space of data to quantum states $\sigma : \mathtt{data} \mapsto \sigma(\mathtt{data}) \in \mathcal{S}$, where data are the results of a sequence of quantum measurements. Since both the true state and data are unknown, we take the expected value with respect to the joint distribution of $(\rho, \mathtt{data})$ to obtain the average fidelity:

$$f(\sigma) = \mathbb{E}_{\rho,\mathtt{data}}[F(\rho,\sigma(\mathtt{data}))]. \tag{4}$$

We want this to be as large as possible. The estimator which maximizes this quantity is equivalent to the estimator maximizing the following posterior average fidelity for every data set:

$$f(\sigma|\mathtt{data}) = \mathbb{E}_{\rho|\mathtt{data}}[F(\rho,\sigma(\mathtt{data}))]. \tag{5}$$

An estimator which maximizes this is called a *Bayes estimator*[2]. Bayes estimators are useful both to understand Bayesian optimality and to provide upper bounds for the worst case performance.

Now here is the subtle and important point: the measurements performed, the data themselves and the distribution from which they were generated are not important once the posterior distribution has been calculated. If we know the solution for every measure $\mathrm{d}\rho$, then we know the solution for the posterior measure $\mathrm{d}\rho|\mathtt{data}$. For brevity, then, we will drop this conditional information from now on and the problem reduces again to (3).

II. SUMMARY OF RESULTS

In this work, we provide absolute benchmarks for the average fidelity performance of any tomographic estimation strategy by way of upper and lower bounds. This is important because, in the field of quantum tomography, a common theme is to compare estimators. Up to date many options are available: linear inversion [1], maximum likelihood [12], Bayesian mean [13], hedged maximum likelihood [14], and compressed sensing [15, 16]—to name a few. Often estimators are compared by simulating measurements on ensembles of states drawn according to some measure and averaging the fidelity. This can only provide conclusions about the relative performance of estimators. Thus, our bounds can be used to benchmark the fidelity performance of other candidate estimators.

We complement our theoretical findings with numerical experiments. These demonstrate the relative tightness of our bounds and, in particular, reveal that the Bayesian mean estimator is an excellent choice—owing to its near-optimal performance and ease of implementation. Importantly, both the mean of the distribution and our bounds can be computed *online*—that is, the estimator and its performance can be computed while data is being taken. In the context of Bayesian quantum information theory [13], our findings lend credence to the standard approach of using the mean of the posterior distribution as an estimator is a near-optimal one.

We note that this problem has been solved for the case of a single qubit ($d = 2$). Bagan *et al* [6] have given the optimal estimator (and measurement!) for any isotropic prior measure. Unfortunately, by making heavy use of the Bloch representation of a qubit, the methods do not generalize. Whereas, our bound holds for all distributions of states in any dimension and coincides with the results of [6] for the case of a single qubit.

A. Ensembles of pure states

We first present the analytically soluble case of measures supported only on pure states. Such a case is common in theoretical studies which average the performance of their protocols over the popular choice of the unique Haar invariant measure on pure states. The solution is organized into the following theorem:

[1] Expectation values will always be denote with a subscript which specifies the implicit distribution of variables being averaged over.

[2] The terminology and objective functions used here can be seen as standard generalizations of those familiar in decision theory. See, e.g., [11].

Theorem 1. *Choose an arbitrary dimension d and assume that the integration measure $\mathrm{d}\rho$ is supported only on pure states. Then, the state which solves the optimization problem (3) is the eigenvector of $\mathbb{E}_\rho[\rho]$ with maximal eigenvalue. It achieves a maximal fidelity of $\left\|\mathbb{E}_\rho\left[\rho\right]\right\|_\infty$.*

The proof is a simple exercise in linear programming. When ρ is a pure state, the fidelity simplifies to $F(\rho,\sigma) = \mathrm{Tr}(\rho\sigma)$. Linearity allows us to bring the expectation inside the trace so that the problem becomes

$$\begin{aligned} \text{maximize} \quad & \mathrm{Tr}(\mathbb{E}_\rho[\rho]\sigma) \\ \text{subject to} \quad & \mathrm{Tr}(\sigma) = 1, \\ & \sigma \geq 0. \end{aligned} \tag{6}$$

The solution can be found in many textbooks covering linear programming—e.g. [17]. This solution also coincides with the one noted for a distribution supported on two states in [18].

B. General measures on mixed states

For measures with support on mixed states, the situation is markedly different. Our main technical contribution are new upper bounds for this case. We obtain them by replacing the fidelity function—which is notoriously difficult to grasp—in the main optimization problem (3) by quantities that are easier to handle in full generality. One rather straightforward approach to do so is to relate the fidelity function $f(\rho,\sigma)$ between arbitrary states $\rho,\sigma \in \mathcal{S}$ to corresponding Schatten-p-norm distances

$$\|\rho-\sigma\|_p = \left(\mathrm{Tr}\left(|\rho-\sigma|^p\right)\right)^{1/p},$$

with $1 \leq p \leq \infty$ and $|X| = \sqrt{X^* X}$ for any $X \in L\left(\mathbb{C}^d\right)$. This can be done by employing the well-known and often used Fuchs-van de Graaf inequalities [19]

$$1-\sqrt{F\left(\rho,\sigma\right)} \leq \frac{1}{2}\|\rho-\sigma\|_1 \leq \sqrt{1-F(\rho,\sigma)} \quad \forall \rho,\sigma \in \mathcal{S}.$$

This inequality together with the hierarchy of Schatten-p-norms assures

$$F\left(\rho,\sigma\right) \leq 1-\frac{1}{4}\,|\,\rho-\sigma\|_1^2 \leq 1-\frac{1}{4}\,\|\rho-\sigma\|_2^2, \tag{7}$$

for any two quantum states $\rho,\sigma \in \mathcal{S}$. Replacing the objective function in the central optimization problem (3) by such an upper bound results in a different optimization which admits a general analytic solution. Clearly, such a relaxed optimum bounds the original figure of merit from above and allows us to establish our second main result.

Theorem 2. *For any finite dimension d and any distribution $\mathrm{d}\rho$, the maximal average fidelity achieved by any estimator $\sigma \in \mathcal{S}$ obeys*

$$\max_{\sigma\in\mathcal{S}} \mathbb{E}_\rho\left[F(\rho,\sigma)\right] \leq 1-\frac{1}{4}\mathrm{Tr}\left(\mathbb{E}_\rho\left[\rho^2\right]-\mathbb{E}_\rho\left[\rho\right]^2\right). \tag{8}$$

Note that the expression on the right hand side of (8) can be interpreted as a non-commutative generalization of the variance of a probability distribution. Having already outlined the main ideas necessary to establish such a result, we refer to Section IV B for a complete proof.

Another way of establishing upper bounds on the average fidelity involves the concept of *super-fidelity*, which provides the following upper bound on the fidelity [20]:

$$F(\rho,\sigma) \leq \mathrm{Tr}\left(\rho\sigma\right)+\sqrt{1-\mathrm{Tr}\left(\rho^2\right)}\sqrt{1-\mathrm{Tr}\left(\sigma^2\right)}. \tag{9}$$

Although more involved, we shall see that such an approach yields strictly better bounds than the ones presented in Theorem 2. For brevity, we define $\hat{\rho} := \mathbb{E}_\rho[\rho]$ and $p_\rho := \mathbb{E}_\rho\left[\sqrt{1-\mathrm{Tr}(\rho^2)}\right]$, such that inequality (9) assures

$$\max_{\sigma\in\mathcal{S}} \mathbb{E}_\rho\left[F\left(\rho,\sigma\right)\right] \leq \max_{\sigma\in\mathcal{S}}\left(\mathrm{Tr}\left(\hat{\rho}\sigma\right)+p_\rho\sqrt{1-\mathrm{Tr}\left(\sigma^2\right)}\right), \tag{10}$$

for any distribution $\mathrm{d}\rho$. Although more tractable than the original problem, the optimization on the right hand side still requires solving a non-commutative maximization over all quantum states $\sigma \in \mathcal{S}$. However, applying a corollary of the famous Birkhoff-von Neumann theorem—see e.g. [21, Theorem 8.7.6]—allows for restricting this optimization to density operators σ that commute with the distribution's mean $\hat{\rho}$—see Lemma 1 below. If $\hat{r}_1, \ldots, \hat{r}_d$ denote the eigenvalues of $\hat{\rho}$ such a restriction assures that solving the right hand side of (10) is equivalent to

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{d} \hat{r}_i s_i + p_\rho \sqrt{1 - \sum_{i=1}^{d} s_i^2} \\ \text{subject to} \quad & \sum_{i=1}^{d} s_i = 1, \\ & s_i \geq 0 \quad 1 \leq i \leq d, \end{aligned} \tag{11}$$

which is a commutative convex optimization problem. We refer to Lemma 1 below for a detailed proof of this assertion. Note that, if the measure $\mathrm{d}\rho$ is supported exclusively on pure states, p_ρ vanishes and (11) reduces to Theorem 1 which is tight.

In order to obtain analytical bounds for mixed states, we further relax (11) by replacing the non-negativity constraints ($s_i \geq 0$) by the weaker demand that the optimization vector $(s_1, \ldots, s_d)^T \in \mathbb{R}^d$ is contained in the Euclidean unit ball—i.e. $\sum_{i=1}^d s_i^2 \leq 1$. As we shall show in Section V, such a simplification is the tightest possible ellipsoidal relaxation of (11) and allows us to apply the method of Lagrangian multipliers in a straightforward fashion. Doing so results in the main theoretical statement of this paper.

Theorem 3. *For any finite dimension d and any distribution $\mathrm{d}\rho$ over states, the fidelity achieved by any estimator $\sigma \in \mathcal{S}$ is bounded from above by*

$$\mathbb{E}_\rho[F(\rho,\sigma)] \leq \frac{1}{d}\left(1 + \sqrt{d-1}\sqrt{d\left(\mathbb{E}_\rho\left[\sqrt{1 - \operatorname{Tr}(\rho^2)}\right]^2 + \operatorname{Tr}\left(\mathbb{E}_\rho[\rho]^2\right)\right) - 1}\right). \tag{12}$$

The matrix achieving this optimum corresponds to

$$\sigma^\sharp = \frac{1}{d}\mathbb{1} + \sqrt{\frac{d-1}{d\left(p_\rho^2 + \operatorname{Tr}(\hat{\rho}^2)\right) - 1}}\left(\hat{\rho} - \frac{1}{d}\mathbb{1}\right), \tag{13}$$

where $\mathbb{1} \in L\left(\mathbb{C}^d\right)$ denotes the identity matrix.

Again, we content ourselves here with outlining the proof architecture necessary to establish such a result and refer to Section IV for a detailed analysis.

Note that since we relaxed the maximization constraints, $\sigma^\sharp$ in general fails to be positive-semidefinite and is thus not a valid density operator, though we do not use it as such. In particular, the bound is not tight when $\mathrm{d}\rho$ is supported only on pure states—as might be evident from the possibility of non-positive states arising from the $(\hat{\rho} - \frac{1}{d}\mathbb{1})$ term in (13). On the other hand, the distribution is known and thus in the case of a distribution supported only on pure states, one should consult the exact solution in Theorem 1.

Conversely, if $\sigma^\sharp$ happens to be a state, it also solves the optimization (11) and the analytical bound (12) exactly reproduces an a priori tighter one. In all of our numerical experiments, some of which are presented below, this was indeed the case.

It is also worthwhile to point out that super-fidelity—the bound in (9)—and the actual fidelity coincide for one qubit, i.e. for $d = 2$ [20]. Also replacing positive semidefiniteness by bounded purity yields the same feasible set for that particular case. Consequently the bound (12) reproduces one of the main results in [6]:

Corollary 1. *In the single-qubit case (i.e. $d = 2$) the bound* (12) *exactly reproduces the maximum average fidelity in [6, Equation (2.9)] and $\sigma^\sharp$ is the optimal estimator.*

Finally, we want to emphasize that establishing bounds on the average fidelity by using the super-fidelity instead of the Fuchs-van de Graaf inequalities leads to strictly better results:

Corollary 2. *Let the dimension d and the distribution $\mathrm{d}\rho$ over states be arbitrary. Then, the bound presented in Theorem 2 (Fuchs van-de Graaf inequality) is either trivial—i.e. equal to one—or it strictly majorizes the one presented Theorem 3 (super-fidelity).*

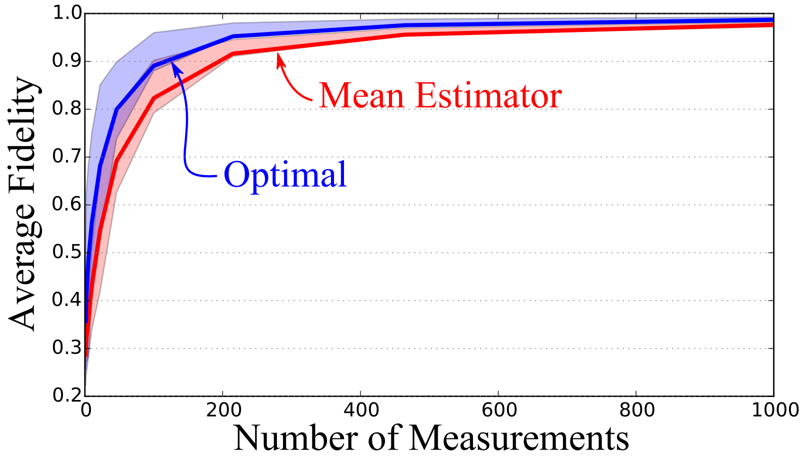


FIG. 1. The average fidelity as a function of the number of single-shot measurements of the Haar uniform measurement. The prior distribution is here is also the Haar uniform measure on two qubits. The lines are the medians and shaded areas the interquartile ranges over 100 trials.

III. NUMERICAL EXPERIMENTS

Note that fidelity achieved by *any* estimator is a lower bound on the one achieved by the optimal estimator. A particularly convenient and generally well motivated [18] estimator is the mean of the distribution $\hat{\rho} = \mathbb{E}_\rho[\rho]$. Our findings underline that for distributions of states relevant to tomography, the mean is very near-optimal. In the context of tomography the mean is furthermore arguably the most convenient estimator, since every other quantity of interest requires its calculation anyway.

Finding an analytical expression for the posterior distribution is a very challenging problem, let alone performing the multidimensional integrals required for the calculation of the expectations above. Thus, we turn to numerics. In particular, we use the Sequential Monte Carlo (SMC) algorithm, which has been successfully applied to quantum statistical problems in the context of dynamical parameter estimation [22–24] and quantum state estimation [25–27]. Also, this algorithm is available as an open-source implementation in Python [28].

Employing SMC allows us to perform the Bayesian updating and averaging. A complete and detailed discussion of the algorithm appears in Ref. [23] and thus we will not repeat the details here, but we will sketch the idea. The algorithm starts with a set of quantum states $\{\rho_j\}_{j=0}^n$, the elements of which are called *particles*. Here, $n = |\{\rho_j\}|$ is the number of particles and controls the accuracy of the approximation. By approximating the prior distribution by a weighted sum of Dirac delta-functions,

$$\Pr(\rho) \approx \sum_{j=1}^{n} w_j \delta(\rho - \rho_j), \tag{14}$$

Bayes' rule then becomes

$$w_j \mapsto \Pr(\texttt{data}|\rho_j) w_j, \tag{15}$$

followed by a normalization step. The SMC algorithm is designed to approximate expectation values, such that

$$\mathbb{E}_\rho[f(\rho)] \approx \sum_{j=1}^{n} w_j f(\rho_j), \tag{16}$$

for any function f. In other words, the SMC algorithm allows us to efficiently compute the multidimensional integrals with respect to the measure defined by the posterior probability distribution. We use this algorithm, as implemented by [28], to numerically compute averages arising in simulated tomography experiments. By doing so, we explore the efficacy of our claims for a variety of distributions relevant to practice and found natural in experimentation.

Recall the sharp distinction between measures supported on pure states and those with full support. We use the fact that Theorem 1 provides us with the optimal estimator in the former case to lend support to the claim that

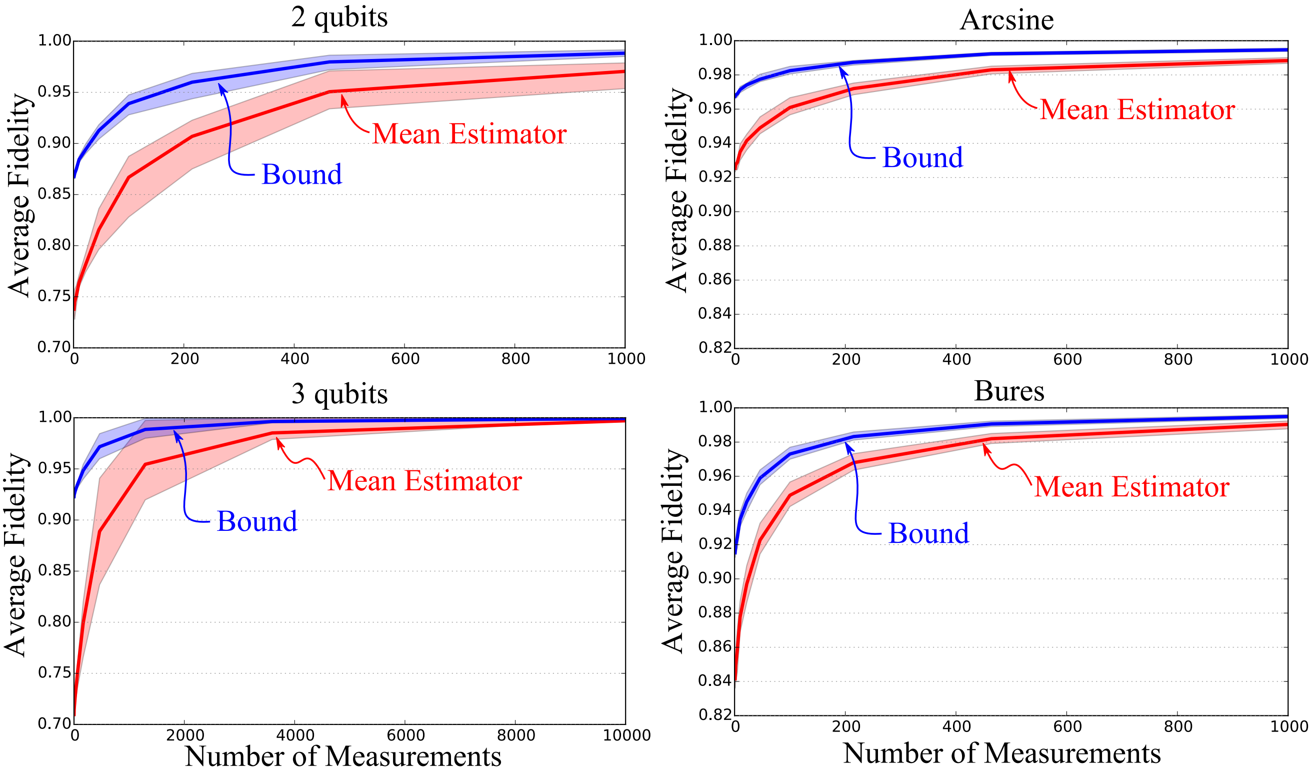


FIG. 2. These plots depict the average fidelity as a function of the number of single-shot measurements of the Haar uniform measurements.
First column: The prior distribution is here is Hilbert-Schmidt measure on two and three qubit mixed quantum states.
Second column: The prior distribution for the upper plot is the *Arcsine* distribution while for the lower plot the *Bures* distribution was used—both are supported on two qubit mixed quantum states (again, see [30] for a review of distributions of density matrices).
In all cases, the solid lines are the medians and shaded areas illustrate the interquartile ranges over 100 trials.

the mean estimator is a good candidate for a computationally simple, yet still near-optimal, alternative to solving the optimization problem in general. In Figure 1, we present the results of numerical simulations on two qubits. Plotted is the average fidelity achieved by the optimal estimator (see Theorem 1) and the mean estimator $\mathbb{E}_\rho[\rho]$. The average is taken with respect to a distribution that begins as the Haar invariant measure on pure states and is updated through simulated measurement data, where the measurement is the "uniform POVM" consisting of all pure states, distributed uniformly according to the Haar measure. For independent measurements—i.e. local, non-adaptive ones—this measurement is optimal [29, Theorem 3.1]. We see that the mean estimator's fidelity tracks the optimal fidelity quite well.

In Figure 2, we plot the average fidelity of the mean estimator against our bound (12) for measures supported also on mixed quantum states. Again, we simulate measurement data to get an accurate sense of how well the average fidelity of the mean estimator performs with respect to our bound for distributions relevant to tomography. In this case, the *prior* distribution is either the Hilbert-Schmidt measure (left column), or the *arcsine* and *Bures* distributions [30] for two qubits (right column). In each case, many other natural distributions appear as we update our prior through Bayes' rule. We see again that the mean estimator is a "good" estimator in that it comes close to the bound on the optimal fidelity and is the easiest non-trivial average quantity to evaluate.

IV. PROOFS

In this section we provide detailed derivations and proofs of the statements presented in Section II.

A. A detailed proof of Theorem 2

Recall that in Theorem 2 we have claimed that the bound

$$\max_{\sigma\in\mathcal{S}} \mathbb{E}_\rho \left[F(\rho,\sigma)\right] \leq 1-\frac{1}{4}\mathrm{Tr}\left(\mathbb{E}_\rho\left[\rho^2\right]-\mathbb{E}\left[\rho\right]^2\right), \tag{17}$$

is valid for any prior distribution $\mathrm{d}\rho$. In order to derive such a statement, we start with inequality (7)

$$F(\rho,\sigma)\leq 1-\frac{1}{4}\|\rho-\sigma\|_2^2,$$

which is a direct combination of the Fuchs-van de Graaf inequalities and the norm inequality $\|\cdot\|_2\leq\|\cdot\|_1$. As such it is valid for any two states $\rho,\sigma\in\mathcal{S}$ which in turn assures that it remains valid upon taking expectations over $\mathrm{d}\rho$ on both sides:

$$\mathbb{E}_\rho\left[F(\rho,\sigma)\right]\leq 1-\frac{1}{4}\mathbb{E}_\rho\left[\|\rho-\sigma\|_2^2\right]. \tag{18}$$

Moreover, we can optimize over σ on both sides to obtain

$$\max_{\sigma\in\mathcal{S}}\mathbb{E}_\rho\left[F(\rho,\sigma)\right]\leq 1-\frac{1}{4}\min_{\sigma\in\mathcal{S}}\mathbb{E}_\rho\left[\|\rho-\sigma\|_2^2\right]. \tag{19}$$

The minimum on the right-hand side can in fact be calculated analytically. To this end, we define the function

$$f(\sigma):=\mathbb{E}_\rho\left[\|\rho-\sigma\|_2^2\right]=\mathrm{Tr}\left(\mathbb{E}_\rho\left[\rho^2\right]\right)-2\mathrm{Tr}\left(\mathbb{E}_\rho[\rho]\sigma\right)+\mathrm{Tr}\left(\sigma^2\right).$$

Note that $f(\sigma)$ is convex, because it corresponds to a weighted average of convex norm-functions $\|\sigma-\rho\|_2^2$ and its matrix-valued derivative corresponds to

$$f'(\sigma)=-2\mathbb{E}_\rho\left[\rho\right]+2\sigma. \tag{20}$$

This derivative vanishes if and only if $\sigma^\sharp=\mathbb{E}_\rho\left[\rho\right]$ holds and convexity of $f(\sigma)$ implies that this critical state corresponds to the unique minimum. The corresponding function value amounts to

$$f\left(\sigma^\sharp\right)=\mathrm{Tr}\left(\mathbb{E}_\rho\left[\rho^2\right]\right)-\mathrm{Tr}\left(\mathbb{E}_\rho\left[\rho\right]^2\right) \tag{21}$$

and reinserting this global minimum into (19) yields the desired bound (17).

B. A detailed derivation of Theorem 3

Our main theoretical statement—Theorem 3—follows from a three step procedure which was already briefly outlined in Section II.

The first step invokes the concept of super-fidelity [20] which assures

$$\max_{\sigma\in\mathcal{S}}\mathbb{E}_\rho\left[F\left(\rho,\sigma\right)\right]\leq\max_{\sigma\in\mathcal{S}}\left(\mathrm{Tr}\left(\hat{\rho}\sigma\right)+p_\rho\sqrt{1-\mathrm{Tr}\left(\sigma^2\right)}\right),$$

with $\hat{\rho}=\mathbb{E}_\rho\left[\rho\right]$ and $p_\rho=\mathbb{E}_\rho\left[\sqrt{1-\mathrm{tr}\left(\rho^2\right)}\right]$ for any distribution $\mathrm{d}\rho$. As it turns out, the optimization on the right hand side of this equation is much more tractable than the original problem on the left hand side. This is manifested by the following technical statement which is a direct consequence of the celebrated Birkhoff-von Neumann theorem.

Lemma 1. *Fix any $p_\rho\geq 0$ and suppose that $\hat{\rho}\in\mathcal{S}$ is an arbitrary density operator with eigenvalue decomposition $\hat{\rho}=\sum_{i=1}^d\hat{r}_i|b_i\rangle\langle b_i|$. Then the optimization*

$$\begin{aligned}\underset{\sigma\in L(\mathbb{C}^d)}{\mathit{maximize}}\quad & \mathrm{Tr}\left(\hat{\rho}\sigma\right)+p_\rho\sqrt{1-\mathrm{Tr}\left(\sigma^2\right)},\\ \mathit{subject\ to}\quad & \sigma\geq 0,\ \mathrm{Tr}(\sigma)=1.\end{aligned} \tag{22}$$

is equivalent to solving

$$\begin{aligned} \underset{s_1,\ldots,s_d\in\mathbb{R}}{\textit{maximize}} \quad & \sum_{i=1}^d \hat{r}_i s_i + p_\rho \sqrt{1-\sum_{i=1}^d s_i^2}, \\ \textit{subject to} \quad & \sum_{i=1}^d s_i = 1, \\ & s_i \geq 0 \quad 1 \leq i \leq d. \end{aligned} \tag{23}$$

Moreover, there is a one-to-one correspondence between any feasible array $(s_1,\ldots,s_d)$ *of this problem and the density operator* $\tilde{\sigma} = \sum_{i=1}^d s_i |b_i\rangle\langle b_i|$.

Proof. At the heart of this statement is an immediate corollary of the Birkhoff-von Neumann Theorem—see e.g. [21, Theorem 8.7.6]. For $d \times d$ Hermitian matrices ρ, σ this corollary assures

$$\mathrm{Tr}\,(\rho\sigma) \leq \sum_{i=1}^d r_i s_i, \tag{24}$$

where r_i and s_i denote the eigenvalues of ρ and σ, respectively, arranged in non-increasing order. If $\hat{\rho}$ has eigenvalue decomposition $\hat{\rho} = \sum_{i=1}^d \hat{r}_i |b_i\rangle\langle b_i|$, the right hand side of (24) corresponds to $\mathrm{Tr}\,(\hat{\rho}\tilde{\sigma})$ where $\tilde{\sigma} = \sum_{i=1}^d s_i |b_i\rangle\langle b_i|$. Clearly, if $\sigma \in \mathcal{S}$ was a quantum state to begin with, so is $\tilde{\sigma}$, because the spectra of σ and $\tilde{\sigma}$ coincide. Moreover, such a definition assures that both states have equal purity, i.e. $\mathrm{Tr}(\sigma^2) = \mathrm{Tr}(\tilde{\sigma}^2)$. Consequently, for any feasible point σ of the optimization (22), there is a $\tilde{\sigma}$ of the above form which admits a larger value in the optimization. Inserting the particular form of $\tilde{\sigma}$ into this program results in (23). □

In order to arrive at the bound presented in Theorem 3, we employ one more relaxation which is going to allow us to solve the resulting problem analytically in full generality. To be concrete, we replace the non-negativity constraints ($s_i \geq 0$) in (23) by the weaker demand that the optimization vector $(s_1,\ldots,s_d)^T \in \mathbb{R}^d$ is contained in the Euclidean unit ball—i.e. $\sum_{i=1}^d s_i^2 \leq 1$. Note that we explore the geometric properties of such a relaxation in Section V. In a nutshell it corresponds to the tightest possible elliptical relaxation of the feasible set in (22). By doing so, we arrive at the problem

$$\begin{aligned} \underset{s_1,\ldots,s_d\in\mathbb{R}}{\text{maximize}} \quad & \sum_{i=1}^d \hat{r}_i s_i + p_\rho \sqrt{1-\sum_{i=1}^d s_i^2}, \\ \text{subject to} \quad & \sum_{i=1}^d s_i = 1,\ \sum_{i=1}^d s_i^2 \leq 1, \end{aligned} \tag{25}$$

which can be solved analytically via the method of Lagrangian multipliers:

Lemma 2. *Let* $\hat{r}_1,\ldots,\hat{r}_d$ *denote the eigenvalues of any density operator and fix* $p_\rho > 0$. *Then the problem* (25) *has a unique solution. The optimal value corresponds to*

$$\frac{1}{d}\left(1+\sqrt{d-1}\sqrt{d\left(p_\rho^2+\mathrm{Tr}\,(\hat{\rho}^2)\right)-1}\right)$$

and the array $(s_1^\sharp,\ldots,s_d^\sharp)$ *achieving this optimum corresponds to the particular matrix*

$$\sigma^\sharp = \frac{1}{d}\mathbb{1} + \sqrt{\frac{d-1}{d\left(p_\rho^2+\mathrm{Tr}\,(\hat{\rho}^2)\right)-1}}\left(\hat{\rho}-\frac{1}{d}\mathbb{1}\right). \tag{26}$$

Note that this result together with the relaxations outlined in this section immediately implies Theorem 3 upon inserting the definitions of p_ρ and $\hat{\rho}$. The assumption $p_\rho > 0$ is furthermore non-critical, because, by definition, $p_\rho = 0$ if and only if $\mathrm{d}\rho$ is supported exclusively on pure states. This particular case, however, is already fully covered by Theorem 1.

Proof of Lemma 2. Throughout this proof we shall represent the eigenvalues of the density operator $\hat{\rho}$ as a vector $\hat{\boldsymbol{r}} = (\hat{r}_1, \ldots, \hat{r}_d)^T \in \mathbb{R}^d$. Likewise we shall encompass the scalar optimization variables s_i in the vector $\boldsymbol{s} \in \mathbb{R}^d$. Furthermore, let $\mathbf{0} = (0, \ldots, 0)^T$ and $\mathbf{1} = (1, \ldots, 1)^T$ denote the "all-zeros" and "all-ones" vectors on $\mathbb{R}^d$, respectively. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we will also make use of the standard inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^d x_i y_i$ and the vectorial inequality $\boldsymbol{x} \geq \boldsymbol{y}$ shall indicate component-wise inequality, i.e. $x_i \geq y_i$ for all $1 \leq i \leq d$.

In such a vectorial form, the optimization problem (23) corresponds to

$$\begin{aligned} \text{maximize} \quad & f(\boldsymbol{s}) = \langle \hat{\boldsymbol{r}}, \boldsymbol{s} \rangle + p_\rho \sqrt{1 - \langle \boldsymbol{s}, \boldsymbol{s} \rangle}, \\ \text{subject to} \quad & g(\boldsymbol{s}) = \langle \mathbf{1}, \boldsymbol{s} \rangle = 0. \\ & \langle \boldsymbol{s}, \boldsymbol{s} \rangle \leq 1. \end{aligned} \tag{27}$$

Note that (27) is a convex optimization problem, as it requires maximizing a concave function over a convex set. As such, it has a unique maximum. One way of finding this maximum is to apply standard techniques such as the Karush-Kuhn-Tucker (KKT) multiplier method [17] which are designed to take into account the inequality constraint (28).

However, here we opt for a less direct but considerably more convenient and less cumbersome approach: we ignore the inequality constraint in (27) for now and employ the standard technique of Lagrangian multipliers (for equality constraints) in order to find the unique critical point $\boldsymbol{s}^\sharp$ of the optimization. In a second step, we are going to verify that this vector strictly obeys the additional inequality constraint, we have ignored so far, i.e. $\langle \boldsymbol{s}^\sharp, \boldsymbol{s}^\sharp \rangle < 1$. This in turn implies that said inequality constraint is not active at the critical point which in retrospect confirms that we were in fact right to ignore it in the first place. Finally, the fact that we face a convex optimization problem assures that this unique critical point indeed yields the sought for global maximum of (27).

In order to find the critical point $\boldsymbol{s}^\sharp$ in question we define the Lagrangian function

$$L(\boldsymbol{s}) = f(\boldsymbol{s}) + \lambda g(\boldsymbol{s}), \tag{28}$$

where we have—as already announced—ignored the inequality constraint $\langle \boldsymbol{s}, \boldsymbol{s} \rangle \leq 1$. As a consequence, $\lambda \in \mathbb{R}$ denotes the single Lagrangian multiplier associated with the remaining normalization constraint. The necessary condition for an optimal solution of (27) then reads

$$\hat{\boldsymbol{r}} - \frac{p_\rho \boldsymbol{s}}{\sqrt{1 - \langle \boldsymbol{s}, \boldsymbol{s} \rangle}} + \lambda \mathbf{1} = \mathbf{0}. \tag{29}$$

Taking the inner product of this vector-identity with the "all-ones" vector $\mathbf{1}$ results in

$$0 = \langle \mathbf{1}, \mathbf{0} \rangle = \langle \mathbf{1}, \hat{\boldsymbol{r}} \rangle - \frac{p_\rho \langle \mathbf{1}, \boldsymbol{s} \rangle}{\sqrt{1 - \langle s, s \rangle}} + \lambda \langle \mathbf{1}, \mathbf{1} \rangle = 1 - \frac{p_\rho}{\sqrt{1 - \langle \boldsymbol{s}, \boldsymbol{s} \rangle}} + d\lambda, \tag{30}$$

where we have used $\langle \mathbf{1}, \hat{\boldsymbol{r}} \rangle = \sum_{i=1}^n \hat{r}_i = \operatorname{Tr}(\hat{\rho}) = 1$ and the normalization constraint, which likewise assures $\langle \mathbf{1}, \boldsymbol{s} \rangle = 1$. This equation allows us to replace $\sqrt{1 - \langle \boldsymbol{s}, \boldsymbol{s} \rangle}$ by $\frac{p_\rho}{1 + d\lambda}$ and reinserting this into (29) results in the equivalent vector equation

$$\hat{\boldsymbol{r}} - (1 + d\lambda) \boldsymbol{s} + \lambda \mathbf{1} = \mathbf{0}. \tag{31}$$

This can be readily inverted to yield

$$\boldsymbol{s} = \frac{1}{1 + d\lambda} (\hat{\boldsymbol{r}} + \lambda \mathbf{1}). \tag{32}$$

In order to determine the value of λ, we revisit (30) which in combination with (32) demands

$$p_\rho^2 = (1 + d\lambda)^2 (1 - \langle \boldsymbol{s}, \boldsymbol{s} \rangle) = (1 + d\lambda)^2 - \langle \hat{\boldsymbol{r}}, \hat{\boldsymbol{r}} \rangle - 2\lambda \langle \mathbf{1}, \hat{\boldsymbol{r}} \rangle - \lambda^2 \langle \mathbf{1}, \mathbf{1} \rangle, \tag{33}$$

$$= d(d-1)\lambda^2 + 2(d-1)\lambda + 1 - \operatorname{Tr}\left(\hat{\rho}^2\right), \tag{34}$$

where we have once more used $\langle \mathbf{1}, \hat{\boldsymbol{r}} \rangle = 1$ as well as $\langle \hat{\boldsymbol{r}}, \hat{\boldsymbol{r}} \rangle = \sum_{i=1}^n \hat{r}_i^2 = \operatorname{Tr}\left(\hat{\rho}^2\right)$. This results in the quadratic equation

$$\lambda^2 + \frac{2}{d}\lambda - \frac{1}{d(d-1)} \left(p_\rho^2 + \operatorname{Tr}\left(\hat{\rho}^2\right) - 1\right), \tag{35}$$

for λ whose two possible solutions correspond to

$$\lambda_{\pm} = -\frac{1}{d}\left(1 \mp \sqrt{\frac{d\left(p_\rho^2 + \mathrm{Tr}\left(\hat{\rho}^2\right)\right) - 1}{d-1}}\right). \tag{36}$$

Note that the argument of the square-root is non-negative, because the purity $\mathrm{Tr}\left(\hat{\rho}^2\right)$ of any quantum state is lower-bounded by $1/d$. Also, the second solution λ_- is vacuous, since it leads to an immediate contradiction. Indeed, it follows by inspection that $\lambda_- < -1/d$ holds. Together with (30) this implies the contradictory relation

$$\sqrt{1 - \langle \boldsymbol{s}, \boldsymbol{s} \rangle} = \frac{p_\rho}{1 + d\lambda_-} < 0, \tag{37}$$

because p_ρ is positive by assumption.

Consequently we are left with one meaningful value λ_+ for the Lagrangian multiplier and inserting it into (32) yields the unique critical solution

$$\boldsymbol{s}^\sharp = \frac{1}{d}\mathbf{1} + \sqrt{\frac{d-1}{d\left(p_\rho^2 + \mathrm{Tr}\left(\hat{\rho}^2\right)\right) - 1}}\left(\hat{\boldsymbol{r}} - \frac{1}{d}\mathbf{1}\right). \tag{38}$$

Recall that throughout this proof we are exploiting a one-to-one correspondence between vectors $\boldsymbol{s} = (s_1, \ldots, s_n)^T \in \mathbb{R}^d$ and hermitian $d \times d$-matrices $\sigma = \sum_{i=1}^n s_i |b_i\rangle\langle b_i|$ that commute with $\hat{\rho}$. Consequently, the critical vector $\boldsymbol{s}^\sharp$ corresponds to the critical matrix presented in (26).

Plugging the critical point $\boldsymbol{s}^\sharp$ into the objective function $f(\boldsymbol{s})$ furthermore yields the corresponding critical function value

$$\begin{aligned} f\left(\boldsymbol{s}^\sharp\right) &= \langle \hat{\boldsymbol{r}}, \boldsymbol{s}^\sharp \rangle + p_\rho \sqrt{1 - \langle \boldsymbol{s}^\sharp, \boldsymbol{s}^\sharp \rangle} = \frac{\langle \hat{\boldsymbol{r}}, \hat{\boldsymbol{r}} \rangle + \lambda_+ \langle \mathbf{1}, \hat{\boldsymbol{r}} \rangle}{1 + d\lambda_+} + \frac{p_\rho^2}{1 + d\lambda_+} = \frac{d\left(p_\rho^2 + \mathrm{Tr}\left(\hat{\rho}^2\right)\right) - 1 + 1 + d\lambda_+}{d(1 + d\lambda_+)}, \\ &= \frac{1}{d}\left(1 + \frac{d\left(p_\rho^2 + \mathrm{Tr}\left(\hat{\rho}^2\right)\right) - 1}{1 + d\lambda_+}\right) = \frac{1}{d}\left(1 + \sqrt{d-1}\sqrt{d\left(p_\rho^2 + \mathrm{Tr}\left(\hat{\rho}^2\right)\right) - 1}\right), \end{aligned}$$

where we have once more replaced $\sqrt{1 - \langle \boldsymbol{s}_+^\sharp, \boldsymbol{s}_+^\sharp \rangle}$ by $\frac{p_\rho}{(1 + d\lambda_+)}$ and combined that with the fact that $(1 + d\lambda_+) = \sqrt{\frac{d\left(p_\rho^2 + \mathrm{Tr}\left(\hat{\rho}^2\right)\right) - 1}{d-1}}$ holds.

With such a unique critical point $\boldsymbol{s}^\sharp$ at hand, we are now ready to show that it strictly obeys the inequality constraint $\langle \boldsymbol{s}^\sharp, \boldsymbol{s}^\sharp \rangle$ we have ignored so far. By employing the same equalities we have used in the previous paragraph, we can readily establish such a claim:

$$\langle \boldsymbol{s}^\sharp, \boldsymbol{s}^\sharp \rangle = 1 - (1 - \langle \boldsymbol{s}^\sharp, \boldsymbol{s}^\sharp \rangle) = 1 - \frac{p_\rho^2}{(1 + d\lambda_+)^2} < 1. \tag{39}$$

The strict inequality on the right follows from the fact that $p_\rho > 0$ holds by assumption. This indeed establishes, that $\boldsymbol{s}^\sharp$ is also a critical point of the optimization problem (27). Since this optimization corresponds to maximizing a concave function over a convex set, the unique critical point $\boldsymbol{s}^\sharp$ must correspond to the unique maximum of (27). □

C. Detailed proofs of Corollary 1 and Corollary 2

We conclude the proof section with providing detailed proofs of the remaining statements, namely that Theorem 3 reproduces the main result in [6] for the particular case of a single qubit, i.e. $d = 2$ (Corollary 1) and that the bounds presented in Theorem 3 are strictly better than the ones outlined in Theorem 2 (Corollary 2).

Proof of Corollary 1. We start this section by pointing out that in the particular case of dimension $d = 2$, the two relaxations we have employed in the previous subsection are not relaxations at all. Indeed, for dimension two, fidelity and super-fidelity coincide, and moreover the sets $\left\{(y_1, y_2)^T \in \mathbb{R}^2 : \ y_1 + y_2 = 1, \ y_1, y_2 \geq 0\right\}$ and $\{(y_1, y_2) \in \mathbb{R}^2 : \ y_1 + y_2 = 1, \ y_1^2 + y_2^2 \leq 1\}$ coincide (this one-to-one correspondence is illustrated in Figure 3 below). These low-dimensional equivalences assure that all the relaxations employed in the derivation of Theorem 3 are actually tight. Consequently, in this particular low-dimensional case, we solve the actual problem of interest.

For deducing the claimed statement from this fact, we consider Equation (2.9) in [6]:

$$F = \frac{1}{2}\left(1 + \sum_{\chi} \|\mathbf{V}_{\chi}\|_2\right). \tag{40}$$

Here χ simply means the the data generated via the measurement. The vector $\mathbf{V}_{\chi}$ is defined as follows:

$$\mathbf{V}_{\chi} = \mathbb{E}_{\rho}[\underline{\mathbf{r}} \, \mathrm{Pr}(\chi|\rho)], \tag{41}$$

where $\underline{\mathbf{r}}$ is related to the usual Bloch vector $\boldsymbol{r} = (x, y, z)$ via

$$\underline{\mathbf{r}} = \left(\sqrt{1 - \|\hat{\boldsymbol{r}}\|_2^2}, \boldsymbol{r}\right). \tag{42}$$

We point out that this F is not the same average fidelity we have considered but the following quantity (which corresponds to our Eq. (4) above):

$$F = \max_{\sigma} \mathbb{E}_{\rho}\left[\mathbb{E}_{\chi|\rho}[F(\rho, \sigma(\chi))]\right]. \tag{43}$$

Note however that, by employing Bayes' rule, this is equal to

$$F = \max_{\sigma} \mathbb{E}_{\chi}\left[\mathbb{E}_{\rho|\chi}[F(\rho, \sigma(\chi))]\right], \tag{44}$$

and thus maximizing the posterior average fidelity is equivalent to maximizing the total average fidelity. Our bound applies directly to the former but trivially extends to the latter.

Thus, to establish Corollary 1, we need to extract the posterior average fidelity from the expressions above. First, using Bayes' rule, we calculate

$$\mathbf{V}_{\chi} = \mathrm{Pr}(\chi)\mathbb{E}_{\rho|\chi}[\underline{\mathbf{r}}]. \tag{45}$$

Using the fact that $\|\boldsymbol{r}\|_2^2 = 2\mathrm{Tr}(\rho^2) - 1$ and

$$\mathrm{Tr}(\mathbb{E}_{\rho|\chi}[\rho]^2) = \frac{1}{2}\left(1 + \left\|\mathbb{E}_{\rho|\chi}\left[\boldsymbol{r}\right]\right\|_2^2\right), \tag{46}$$

we find

$$\|\mathbf{V}_{\chi}\|_2^2 = \mathrm{Pr}(\chi)^2\left(2\mathbb{E}_{\rho|\chi}\left[\sqrt{1 - \mathrm{Tr}(\rho^2)}\right]^2 + 2\mathrm{Tr}\left(\mathbb{E}_{\rho|\chi}[\rho]^2\right) - 1\right). \tag{47}$$

Plugging this back into (40), we have

$$F = \frac{1}{2}\left(1 + \sum_{\chi} \mathrm{Pr}(\chi)\sqrt{2\mathbb{E}_{\rho|\chi}\left[\sqrt{1 - \mathrm{Tr}(\rho^2)}\right]^2 + 2\mathrm{Tr}\left(\mathbb{E}_{\rho|\chi}[\rho]^2\right) - 1}\right), \tag{48}$$

$$= \frac{1}{2}\left(1 + \mathbb{E}_{\chi}\left[\sqrt{2\mathbb{E}_{\rho|\chi}\left[\sqrt{1 - \mathrm{Tr}(\rho^2)}\right]^2 + 2\mathrm{Tr}\left(\mathbb{E}_{\rho|\chi}[\rho]^2\right) - 1}\right]\right), \tag{49}$$

$$= \mathbb{E}_{\chi}\left[\frac{1}{2}\left(1 + \sqrt{2\mathbb{E}_{\rho|\chi}\left[\sqrt{1 - \mathrm{Tr}(\rho^2)}\right]^2 + 2\mathrm{Tr}\left(\mathbb{E}_{\rho|\chi}[\rho]^2\right) - 1}\right)\right]. \tag{50}$$

Thus, implied by the results of [6], the maximum posterior average fidelity (dropping the χ for parallelism) is

$$\max_{\sigma} \mathbb{E}_\rho[F(\rho,\sigma)] = \frac{1}{2}\left(1+\sqrt{2\left(\mathbb{E}_\rho\left[\sqrt{1-\mathrm{Tr}(\rho^2)}\right]^2 + \mathrm{Tr}\left(\mathbb{E}_\rho[\rho]^2\right)\right)-1}\right). \tag{51}$$

This coincides with our main result (12) for dimension $d = 2$. □

Proof of Corollary 2. For notational simplicity, let us introduce the short-hand notation

$$s_\rho := \mathrm{Tr}\left(\mathbb{E}_\rho\left[\rho^2\right]\right) - \mathrm{Tr}\left(\mathbb{E}_\rho\left[\rho\right]^2\right), \tag{52}$$

such that the bound presented in Theorem 2 simply reads $\max_{\sigma\in\mathcal{S}} \mathbb{E}_\rho\left[F(\rho,\sigma)\right] \leq 1-\frac{s_\rho}{4}$. Note furthermore that $0 \leq s_\rho \leq 1$ holds. As already mentioned, the lower bound follows from invoking Jensen's inequality, while the upper bound is a simple consequence of the fact that the purity of any state is at most one. A vanishing s_ρ would correspond to a trivial Fuchs-van de Graaf bound of one which is the first case instance covered by Corollary 2. Therefore we can from now on safely assume that $s_\rho > 0$ holds. Under this assumption we prove the second claim by starting with the bound presented in Theorem 3 and upper-bounding it via a chain of inequalities which will ultimately lead to the bound presented in Theorem 2. Indeed, pick any dimension d and an arbitrary distribution $\mathrm{d}\rho$ over states. Then Jensen's inequality assures

$$\mathbb{E}_\rho\left[\sqrt{1-\mathrm{Tr}\left(\rho^2\right)}\right]^2 \leq 1-\mathbb{E}_\rho\left[\mathrm{Tr}\left(\rho^2\right)\right], \tag{53}$$

and the right hand side of expression (12) in Theorem 3 can be upper-bounded by

$$\frac{1}{d}+\frac{\sqrt{d-1}}{d}\sqrt{d-1-ds_\rho}, \tag{54}$$

because the square root function is monotonically-increasing on the positive reals. Adding and subtracting s_ρ in the last square root and once more invoking monotonicity allows us to continue via

$$\frac{1}{d}+\frac{\sqrt{d-1}}{d}\sqrt{(d-1)(1-s_\rho)-s_\rho} < \frac{1}{d}+\frac{d-1}{d}\sqrt{1-s_\rho}, \tag{55}$$

where we have used $s_\rho > 0$ in the last line to obtain strict inequality. Since the square root is a concave function, the inequality $\sqrt{1-s_\rho} \leq 1-\frac{1}{2}s_\rho$ is valid for any $s_\rho \leq 1$ and consequently

$$\frac{1}{d}+\frac{d-1}{d}\sqrt{1-s_\rho} \leq 1-\frac{d-1}{2d}s_\rho, \tag{56}$$

is true. Finally, we use the simple fact that $\frac{d-1}{d} \geq \frac{1}{2}$ holds for any $d \geq 2$ to arrive at $1-\frac{1}{4}s_\rho$ which is just the Fuchs-van de Graaf bound. Since a strict inequality sign connects the expressions in (55), the claimed strict majorization follows. □

V. GEOMETRIC INTERPRETATION OF THE RELAXATION LEADING TO Equation 25

Recall that in order to arrive at Theorem 3, we have replaced the feasible set

$$\Delta^{d-1} = \left\{ \boldsymbol{s} \in \mathbb{R}^d : \langle \boldsymbol{1}, \boldsymbol{s}\rangle = 1, \ \boldsymbol{s} \geq \boldsymbol{0} \right\}, \tag{57}$$

of the optimization problem (11) by

$$\mathcal{E}_{\Delta^{d-1}} = \left\{ \boldsymbol{s} \in \mathbb{R}^d : \langle \boldsymbol{1}, \boldsymbol{s}\rangle = 1, \ \langle \boldsymbol{s}, \boldsymbol{s}\rangle \leq 1 \right\}, \tag{58}$$

which is a convex outer approximation of Δ^{d-1}. This follows from the basic fact that $x^2 \leq x$ holds for any $x \in [0,1]$. Since the vector components s_i of any $\boldsymbol{s} \in \Delta^{d-1}$ have to obey $s_i \in [0,1]$, we can readily conclude

$$\langle \boldsymbol{s}, \boldsymbol{s}\rangle = \sum_{i=1}^{d} s_i^2 \leq \sum_{i=1}^{d} s_i = 1. \tag{59}$$

Note that the converse is true if and only if $d = 1, 2$—a fact which we have exploited in proving Corollary 1.

Geometrically, the former set corresponds to the standard simplex in $\mathbb{R}^d$. In this section we prove that the latter one is in fact the minimum volume covering ellipsoid of the standard simplex which furthermore corresponds to a $(d-1)$-dimensional Euclidean ball. For dimensions two and three this situation is illustrated in Figure 3.

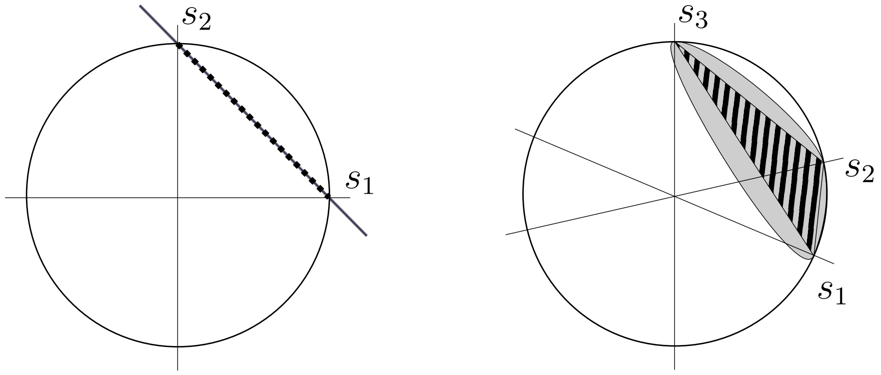


FIG. 3. *Geometric relation between the standard simplex Δ^{d-1} and its outer approximation $\mathcal{E}_{\Delta^{d-1}}$*: Geometrically, the latter set corresponds to the minimum volume outer ellipsoid of the standard simplex. The figure illustrates this relation for dimensions $d = 2$ and $d = 3$. Note that for $d = 2$, the two sets coincide.

Proposition 1 (Geometric nature of $\mathcal{E}_{\Delta^{d-1}}$)**.** *The convex outer-approximation $\mathcal{E}_{\Delta^{d-1}}$ of the d-simplex corresponds to a $(d-1)$-dimensional Euclidean ball with radius $\sqrt{\frac{d-1}{d}}$ and center $\frac{1}{\sqrt{d}}\mathbf{1}$ which is contained in the $(d-1)$-dimensional hyperplane $\mathcal{H}_{\mathbf{1},1} := \left\{ \boldsymbol{s} \in \mathbb{R}^d : \langle \mathbf{1}, \boldsymbol{s} \rangle = 1 \right\}$.*

Proof. By definition, the set $\mathcal{E}_{\Delta^{d-1}}$ corresponds to the intersection of the Euclidean unit ball $\mathcal{B}_1(0) = \left\{ \boldsymbol{s} \in \mathbb{R}^d : \langle \boldsymbol{s}, \boldsymbol{s} \rangle \leq 1 \right\}$ and the hyperplane $\mathcal{H}_{\mathbf{1},1}$. This assures $\mathcal{E}_{\Delta^{d-1}} \subseteq \mathcal{H}_{\mathbf{1},1}$ by construction.

One way to establish that $\mathcal{E}_{\Delta^{d-1}}$ is furthermore itself an Euclidean ball, is using "generalized cylindrical coordinates" for the Euclidean unit ball $\mathcal{B}^d(\mathbf{0}, 1)$: Such coordinates use the fact that $\mathcal{B}^d(\mathbf{0}, 1)$ is equivalent to the union of a family of $(d-1)$-dimensional unit balls. More concretely: let $\boldsymbol{z} \in \mathbb{R}^d$ be an arbitrary unit vector and let $\zeta \in \mathbb{R}$ denote a parameter. For each value of this parameter, we define the hyperplane $\tilde{\mathcal{H}}_{z,\zeta} = \left\{ \boldsymbol{s} \in \mathbb{R}^d : \langle \boldsymbol{z}, \boldsymbol{s} \rangle = \zeta \right\}$ which in particular contains the vector $\zeta \boldsymbol{z}$ by construction. Furthermore, let $\tilde{\mathcal{B}}^{d-1}(\boldsymbol{z}, \zeta) \subset \tilde{\mathcal{H}}_{z,\zeta}$ be the $(d-1)$-dimensional Euclidean ball with radius $\sqrt{1-\zeta^2}$ and center $\zeta \boldsymbol{z}$ that is contained in the hyperplane $\tilde{\mathcal{H}}_{z,\zeta}$. Clearly each element in such a union of sets is contained in the d-ball, and letting ζ range from -1 to 1 covers the entire d-ball. In order to see this, decompose any $\boldsymbol{s} \in \mathcal{B}^d(\mathbf{0}, 1)$ as $\boldsymbol{s} = \langle \boldsymbol{s}, \boldsymbol{z} \rangle \boldsymbol{z} + \boldsymbol{z}^{\perp}$ such that $\langle \boldsymbol{z}^{\perp}, \boldsymbol{z} \rangle = 0$ and set $\zeta = \langle \boldsymbol{s}, \boldsymbol{z} \rangle$. Pythagoras' theorem then assures $\|\boldsymbol{z}^{\perp}\|_2 \leq \sqrt{1-\zeta^2}$ and consequently $\boldsymbol{s} \in \tilde{\mathcal{B}}^{d-1}(\boldsymbol{z}, \zeta)$.

The structure of the particular problem at hand suggests to fix $\boldsymbol{z} = \frac{1}{\sqrt{d}}\mathbf{1}$. Indeed, such a particular choice of $\boldsymbol{z}$ assures equality of the hyperplane $\mathcal{H}_{\mathbf{1},1}$ which contains $\mathcal{E}_{\Delta^{d-1}}$ and the hyperplane $\tilde{\mathcal{H}}_{\frac{1}{\sqrt{d}}\mathbf{1}, \frac{1}{\sqrt{d}}}$, we have just introduced. Consequently, the "cylindrical representation" of the Euclidean unit ball assures that the intersection $\mathcal{E}_{\Delta^{d-1}} = \mathcal{B}_1(0) \cap \mathcal{H}_{\mathbf{1},1}$ corresponds to the $(d-1)$-ball $\tilde{\mathcal{B}}^{d-1}(\frac{1}{\sqrt{d}}\mathbf{1}, \frac{1}{\sqrt{d}})$ associated with the hyperplane $\tilde{\mathcal{H}}_{\frac{1}{\sqrt{d}}\mathbf{1}, \frac{1}{\sqrt{d}}}$ and a parameter value $\zeta = \frac{1}{\sqrt{d}}$. By definition, this ball has center $\frac{1}{d}\mathbf{1}$ and radius $\sqrt{1-\zeta^2} = \sqrt{\frac{d-1}{d}}$ which completes the proof. □

The next statement establishes that our choice of replacing the original feasible set Δ^{d-1} in the proof of Theorem 3 by the larger convex set $\mathcal{E}_{\Delta^{d-1}}$ is in a precise sense the tightest possible elliptic relaxation of the original optimization problem.

Proposition 2. *The set $\mathcal{E}_{\Delta^{d-1}}$ is the unique minimal volume covering ellipsoid of the standard simplex Δ^{d-1}.*

The proof exploits the following standard result about Löwner-John ellipsoids that is originally due to John. However, here we make use of a slightly more general version presented in [31].

Theorem 4 (Theorem 2.1 in [31])**.** *Let $K \subset \mathbb{R}^d$ be a convex body and let K be contained in the Euclidean unit ball $\mathcal{B}^d(0)$. Then the following statements are equivalent:*

1. *$\mathcal{B}^d(0)$ is the unique minimum volume ellipsoid containing K.*

2. *There exist contact points $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ lying both in the boundary of K and $\mathcal{B}^d(0)$, and positive numbers $\lambda_1, \ldots, \lambda_m$, $m \geq d$, such that*

$$\sum_{i=1}^{m} \lambda_i \boldsymbol{u}_i = \boldsymbol{0} \quad \textit{and} \quad \sum_{i=1}^{m} \lambda_i |\boldsymbol{u}_i\rangle\langle\boldsymbol{u}_i| = \mathbb{1}. \tag{60}$$

Proof. In Proposition 1 we have established that the set $\mathcal{E}_{\Delta^{d-1}}$ corresponds to a $(d-1)$-ball with radius $\sqrt{\frac{d-1}{d}}$ and center $\frac{1}{d}\boldsymbol{1}$ that (like the standard simplex) is contained in the hyperplane $\mathcal{H}_{\boldsymbol{1},1}$. A quick calculation reveals that all vertices of the standard simplex Δ^{d-1}—which are just the standard basis vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$—have Euclidean distance $\sqrt{\frac{d-1}{d}}$ to the ball's center. Consequently they are contained in the boundary of the ball $\mathcal{E}_{\Delta^{d-1}}$ and we have found sufficiently many contact points for applying Theorem 4. Since volume is translationally invariant we can furthermore shift the coordinate's origin into the point $\frac{1}{d}\boldsymbol{1}$ (which is the center of the ball $\mathcal{E}_{\Delta^{d-1}}$). This has the advantage that the affine space $\mathcal{H}_{\boldsymbol{1},1}$ containing both Δ^{d-1} and $\mathcal{E}_{\Delta^{d-1}}$ turns into $\mathcal{H}_{\boldsymbol{1},0}$ which is a linear subspace. Note that with respect to the (translated) standard basis, the orthogonal projection onto this subspace is given by

$$P = \mathbb{1} - \frac{1}{d}|\boldsymbol{1}\rangle\langle\boldsymbol{1}|.$$

With respect to this new coordinate system, the d contact points (vertices of the simplex) amount to $\tilde{\boldsymbol{e}}_i = \boldsymbol{e}_i - \frac{1}{d}\boldsymbol{1}$. Choosing unit weights $\lambda_i = 1$ for all $m = d$ contact points $\boldsymbol{u}_i = \tilde{\boldsymbol{e}}_i$ and calculating

$$\sum_{i=1}^{m} \lambda_i \boldsymbol{u}_i = \sum_{i=1}^{n} \tilde{\boldsymbol{e}}_i = \sum_{i=1}^{n} \left(\boldsymbol{e}_i - \frac{1}{d}\boldsymbol{1}\right) = \boldsymbol{0} \tag{61}$$

reveals that the first condition for Theorem 4 is fulfilled. A similar calculation reveals

$$\sum_{i=1}^{m} \lambda_i |\boldsymbol{u}_i\rangle\langle\boldsymbol{u}_i| = \mathbb{1} - \frac{1}{d}|\boldsymbol{1}\rangle\langle\boldsymbol{1}|.$$

This, however equals just the projector P onto the subspace $\mathcal{H}_{\boldsymbol{1},0}$ which contains the entire $(d-1)$-dimensional problem of interest. Restricted to its range, a projector corresponds to the identity which establishes the second condition for Theorem 4. Since this statement is invariant under re-scaling, we can also apply it here, where the radius of the $(d-1)$-dimensional surrounding Euclidean ball is not one but $\sqrt{\frac{d-1}{d}}$. □

VI. CONCLUSION

In this work we have derived upper bounds on the average fidelity of any estimator with no restrictions on the dimension or the distribution being averaged over. Furthermore, we have shown a sharp distinction in the optimization problems of maximizing average fidelity between measures supported only on pure states and those with full support. In the former case, we have provided the exact optimal estimator, while in both cases we argued based on numerical evidence that the mean estimator is a good proxy for the optimal solution.

Interestingly, we found that the analytical bound (12) (which is based on super-fidelity [20]) is strictly tighter than a corresponding one obtained using the well known, and often used, Fuchs-van de Graaf inequalities [19].

These results have obvious applications to practical Bayesian quantum tomography [13], since the bound can be computed *online*—that is, it is only a property of the current distribution under consideration. But we also expect our bound to be of interest in other theoretical work on tomography, where a benchmark is needed to make statements about absolute average performance of some candidate protocol.

ACKNOWLEDGMENTS

We thank Robin Blume-Kohout, Steve Flammia and Patrick Hayden for fruitful discussions. CF was supported by National Science Foundation grant number PHY-1212445, the Canadian Government through the NSERC PDF program, the IARPA MQCO program, the ARC via EQuS project number CE11001013, and by the US Army Research Office grant numbers W911NF-14-1-0098 and W911NF-14-1-0103. RK was supported by scholarship funds from the State Graduate Funding Program of Baden-Württemberg, the Excellence Initiative of the German Federal and State Governments (Grant ZUK 43), the ARO under contracts, W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), the Freiburg Research Innovation Fund, and the DFG (GRO 4334-1-1).

Furthermore, the authors would like to thank the *Mathematisches Forschungsinstitut Oberwolfach* and the organizers of the Oberwolfach workshop *New Horizons in Statistical Decision Theory* (September 2014, Workshop ID: 1437) where work on this project commenced.

[1] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge University Press (2010).
[2] W. K. Wootters, *Statistical distance and Hilbert space*, Physical Review D **23**, 357 (1981).
[3] R. Jozsa, *Fidelity for mixed quantum states*, Journal of Modern Optics **41**, 2315 (1994).
[4] Joseph Emerson, Yaakov S. Weinstein, Seth Lloyd and D. G. Cory, *Fidelity Decay as an Efficient Indicator of Quantum Chaos*, Physical Review Letters **89**, 284102 (2002).
[5] Navin Khaneja, Timo Reiss, Cindie Kehlet, Thomas Schulte-Herbruggen and Steffen J. Glaserb, *Optimal control of coupled spin dynamics: design of NMR pulse sequences by gradient ascent algorithms*, Journal of Magnetic Resonance 172, 296 (2005).
[6] E. Bagan, M. A. Ballester, R. D. Gill, A. Monras and R. Munoz-Tapia, *Optimal full estimation of qubit mixed states*, Physical Review A **73**, 032301 (2006).
[7] J. Emerson, M. P. da Silva, O. Moussa, C. A. Ryan, M. Laforest, J. Baugh, D. G. Cory and R. Laflamme *Symmetrized characterization of noisy quantum processes*, Science **317**, 1893 (2007).
[8] Steven T. Flammia and Yi-Kai Liu, *Direct fidelity estimation from few Pauli measurements*, Physical Review Letters **106**, 230501 (2011).
[9] Marcus P. da Silva, Olivier Landon-Cardinal and David Poulin, *Practical characterization of quantum devices without tomography*, Physical Review Letters 107, 210404 (2011).
[10] Robert L. Cook, Carlos A. Riofrio and Ivan H. Deutsch, *Single-shot quantum state estimation via a continuous measurement in the strong backaction regime*, Physical Review A **90**, 032113 (2014).
[11] J. Berger, *Statistical Decision Theory and Bayesian Analysis*. Springer (1985).
[12] Z. Hradil, *Quantum-state estimation*, Physical Review A **55**, R1561(R) (1997).
[13] Robin Blume-Kohout, *Optimal, reliable estimation of quantum states*, New Journal of Physics, **12**, 043034 (2010).
[14] Robin Blume-Kohout, *Hedged Maximum Likelihood Quantum State Estimation*, Physical Review Letters 105, 200504 (2010).
[15] David Gross, Yi-Kai Liu, Steven T. Flammia, Stephen Becker and Jens Eisert, *Quantum State Tomography via Compressed Sensing*, Physical Review Letters **105**, 150401 (2010).
[16] Richard Kueng, Holger Rauhut and Ulrich Terstiege, *Low rank matrix recovery from rank-one measurements*, arXiv:1410.6913 (2014)
[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press (2004).
[18] Robin Blume-Kohout and Patrick Hayden, *Accurate quantum state estimation via "Keeping the experimentalist honest"*, arXiv:quant-ph/0603116.
[19] Christopher A Fuchs and Jeroen Van De Graaf, *Cryptographic distinguishability measures for quantum-mechanical states*, IEEE Transactions on Information Theory **45**, 1216 (1999).
[20] J. A. Miszczak, Z. Puchaa, P. Horodecki, A. Uhlmann, K. Zyczkowski, *Sub- and super-fidelity as bounds for quantum fidelity*, Quantum Information & Computation **9**, 103 (2009).
[21] Roger A. Horn and Charles R. Johnson, *Matrix analysis (Second edition)*. Cambridge University Press (2013).
[22] Bradley A. Chase and J. M. Geremia, *Single-shot parameter estimation via continuous quantum measurement*, Physical Review A **79**, 022314 (2009).
[23] Christopher E Granade, Christopher Ferrie, Nathan Wiebe and D G Cory, *Robust online Hamiltonian learning*, New Journal of Physics **14**, 103013 (2012).
[24] Nathan Wiebe, Christopher Granade, Christopher Ferrie and D.G. Cory, *Hamiltonian Learning and Certification Using Quantum Resources*, Physical Review Letters **112**, 190501 (2014).
[25] F. Huszar and N. M. T. Houlsby, *Adaptive Bayesian quantum tomography*, Physical Review A **85**, 052120 (2012).
[26] Christopher Ferrie, *High posterior density ellipsoids of quantum states*, New Journal of Physics **16**, 023006 (2014).
[27] Christopher Ferrie, *Quantum model averaging*, New Journal of Physics **16**, 093035 (2014).
[28] Christopher Granade and Christopher Ferrie, *QInfer: Library for Statistical Inference in Quantum Information*, (2012).
[29] A. S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory*. North-Holland Publishing Company (1982).
[30] Karol Zyczkowski, Karol A. Penson, Ion Nechita and Benoit Collins, *Generating random density matrices*, Journal of Mathe-

matical Physics **52**, 062201 (2011).

[31] Martin Henk, *Löwner-John ellipsoids*, Documenta Mathematica, Extra Volume: Optimization Stories, 95 (2012)

Comparing Experiments to the Fault-Tolerance Threshold

Richard Kueng,[1,2,3] David M. Long,[1] Andrew C. Doherty,[1] and Steven T. Flammia[1]

[1] *Centre for Engineered Quantum Systems, School of Physics, University of Sydney, Sydney, NSW, Australia*
[2] *Institute for Theoretical Physics, University of Cologne, Germany*
[3] *Institute for Physics & FDM, University of Freiburg, Germany*

(Dated: July 20, 2016)

Achieving error rates that meet or exceed the fault-tolerance threshold is a central goal for quantum computing experiments, and measuring these error rates using randomized benchmarking is now routine. However, direct comparison between measured error rates and thresholds is complicated by the fact that benchmarking estimates average error rates while thresholds reflect worst-case behavior when a gate is used as part of a large computation. These two measures of error can differ by orders of magnitude in the regime of interest. Here we facilitate comparison between the experimentally accessible average error rates and the worst-case quantities that arise in current threshold theorems by deriving relations between the two for a variety of physical noise sources. Our results indicate that it is coherent errors that lead to an enormous mismatch between average and worst case, and we quantify how well these errors must be controlled to ensure fair comparison between average error probabilities and fault-tolerance thresholds.

The fault-tolerance threshold theorem is a fundamental result that justifies the tremendous interest in building large-scale quantum computers despite the formidable practical difficulties imposed by noise and imperfections. This theorem gives a theoretical guarantee that quantum computers can be built in principle if the noise strength and correlation are below some threshold value [1–3].

To make precise statements of threshold theorems, we must quantify the strength of errors in noisy quantum operations. Ideally we would do this in terms of quantities that can be measured in experiments. A standard measure for quantifying errors in quantum gates is given by the *average error rate*, which is defined as the infidelity between the output of an ideal unitary gate $\mathcal{U}$ and a noisy version $\mathcal{E}\mathcal{U}$ with noise process $\mathcal{E}$, uniformly averaged over all pure states,

$$r(\mathcal{E}) = 1 - \int \mathrm{d}\psi \, \langle\psi|\mathcal{E}\big(|\psi\rangle\langle\psi|\big)|\psi\rangle \,. \qquad (1)$$

This quantity has many virtues: it can be estimated efficiently for any ideal gate $\mathcal{U}$, and in a manner that is independent of state preparation and measurement (SPAM) errors by using the now standard method of randomized benchmarking [4–7]. Recent experimental implementations include [8–17].

The major drawback of using Eq. (1) to quantify gate errors is that it is only a proxy for the actual quantity of interest, the fault-tolerance threshold. This is because r captures average-case behavior for a single use of the gate, while fault tolerance theorems characterize noise in terms of *worst-case* performance when the gate is used repeatedly in a large computation. The importance of this distinction has recently been emphasized by Sanders *et al* [18]. For some noise types (such as pure dephasing and depolarizing noise) the worst- and average-case behavior essentially coincide [19]. However for other classes of errors, notably errors in detuning and calibration that lead to over or under rotation, the worst-case behavior is proportional to $\sqrt{r}$ and can be *orders of magnitude* worse than the average in the relevant regime of $r \ll 1$, as we will discuss in more detail below. Thus it is not possible to directly compare a measured value of r to a threshold result. Despite this, experimentalists are increasingly wishing to relate the results of benchmarking experiments to fault tolerance thresholds. There is thus a pressing need for techniques that allow for direct comparison between experimentally measurable error rates and fault-tolerance thresholds.

In this Letter, we investigate the relationship between worst-case and average-case error for a wide range of error models that are relevant to experiments. Firstly, we show that while closed form expressions do not typically exist, well-established theoretical techniques of convex optimization are often sufficient to determine the relationship between average-case and worst-case errors for models of physical interest. The details of these computations are largely relegated to the Supplementary Material. Secondly, we study a wide range of error models for one-qubit gates. Our main example is of a one-qubit gate with combined dephasing and calibration error. This allows us to demonstrate the crossover between a regime dominated by dephasing, where average-case and worst-case errors are not too different, and the limit of a unitary noise, where the worst-case error scales like $\sqrt{r}$. We then turn to general bounds on worst-case error, showing that it scales as $\sqrt{r}$ for all unitary errors and that for a wide class of errors it can be accurately estimated in terms of r and a recently introduced measure of how close an error process is to being unitary. Finally, conventional benchmarking experiments contain a lot more information than is required just to extract r. We find that this information can often be used to show that the worst-case error has an unfavourable scaling. This is an area that we hope will attract much more study in future.

Fault-tolerance thresholds. A wide range of fault-tolerance thresholds have been reported. The value of the threshold depends greatly on the fault tolerant procedures that are used, on the noise model that is assumed, and whether the threshold is determined from (possibly conservative) analytic bounds on the error, or from (possibly optimistic) numerical simulations. We emphasize that the errors that are given in theoretical fault tolerance papers typically refer to some measure of worst-case error. For example the widely known results of Aliferis and collaborators [20–22] use concatenated error correcting codes and consider a stochastic adversarial noise model that includes all of the noise processes that we will discuss in this paper. These papers find that large-scale quantum computation can be performed for errors below a few times 10^{-4}, when that error is quantified by a measure of worst-case error such as the diamond distance that we discuss below. For more optimistic noise models and for fault-tolerant protocols such as the widely known surface code approaches, the threshold is around 10^{-2} based on numerical simulations of Pauli errors [23]. For Pauli noise however there is no significant difference between worst-case and average-case errors [19]. The performance of these schemes in the presence of coherent errors is not yet understood.

It is possible to state a version of the threshold theorem directly in terms of r, but given current knowledge the thresholds in these theorems would be roughly the square of current thresholds (around 10^{-8} for [20–22]). It is unclear if this can be significantly improved upon since it may be that it is the worst-case error that is physically relevant to the success of the computation. However, our results here motivate research into whether current fault tolerance results could be strengthened to provide significantly improved thresholds when expressed in terms of r for error models sufficiently general to include coherent errors.

Diamond distance. We will now describe the most commonly used metric of worst-case error for quantum processes. Any candidate measure of distance $\Delta(\mathcal{E}, \mathcal{F})$ between noise operations $\mathcal{E}$ and $\mathcal{F}$ should satisfy certain desirable properties [24]. (The operation $\mathcal{F}$ should be thought of as a perfect identity gate for our purposes.) First, like any good distance measure it should have the structure of a metric, which in particular means it should be symmetric, positive, and obey the triangle inequality. Less obviously, but even more importantly, it should obey two additional properties: *chaining* and *stability*. The chaining property,

$$\Delta(\mathcal{E}_2\mathcal{E}_1, \mathcal{F}_2\mathcal{F}_1) \leq \Delta(\mathcal{E}_1, \mathcal{F}_1) + \Delta(\mathcal{E}_2, \mathcal{F}_2)\,, \tag{2}$$

says that composing two noisy operations cannot amplify the error by more than the sum of the two individual errors. Thus, errors can grow at most linearly in the number of operations. The stability property states that the error metric for a single gate should be independent of whether that gate is embedded in a larger computation. So we require

$$\Delta(\mathcal{I} \otimes \mathcal{E}, \mathcal{I} \otimes \mathcal{F}) = \Delta(\mathcal{E}, \mathcal{F})\,, \tag{3}$$

where $\mathcal{I}$ is the identity operation. This ensures that our measure is robust even if the input to the gate is entangled with other qubits in the computation.

The diamond distance, whose formal definition is

$$D(\mathcal{E}, \mathcal{F}) = \tfrac{1}{2} \max_{\rho} \|\mathcal{I} \otimes \mathcal{F}(\rho) - \mathcal{I} \otimes \mathcal{E}(\rho)\|_1\,, \tag{4}$$

satisfies each of these physically motivated desiderata [1]. It also has an appealing operational interpretation as the maximum probability of distinguishing the output of the noisy gate from the ideal output [1, 25]. It is not obvious from the definition how to do practical computations with this quantity, but it can be computed efficiently using the methods of semidefinite programming [26–28]. Because of these properties, the diamond distance is an ideal measure for quantifying noise for the purposes of a fault-tolerance threshold, although in principle other quantities could be employed as well [2].

The only drawback of this quantity is that it is not known how to measure it directly in experiments. It is therefore of interest to have a conversion to, or at least bounds for, diamond distance in terms of the average gate fidelity. To date, the best known bounds for a d-level quantum gate are [29]

$$\tfrac{d+1}{d} r \leq D \leq \sqrt{d(d+1)r}\,,$$

but it is unknown for what conditions these bounds are tight.

Single-qubit calibration and dephasing errors. In order to discuss the relationship between average-case and worst-case errors in quantum computing demonstration experiments we will now analyze in detail a simple but physically relevant noise model for a single-qubit gate. Suppose that the gate is implemented by the noisy control Hamiltonian $H_c = J(t)\sigma_z$. Due to experimental imperfections the control $J(t)$ that is implemented is distinct from the nominal control $J_0(t)$ that would perfectly implement the required gate. Physically, this noise results in two distinct types of errors: *dephasing*, where $\delta J(t) = J - J_0$ varies stochastically between uses of the gate, and *calibration error* where δJ takes the same fixed value each time the gate is used. Where $\delta J(t)$ is stochastically varying we assume that the noise level does not change with time, and that that the noise spectrum for $\delta J(t)$ is mainly confined to frequencies $f > 1/t_g$, where t_g is the time required to perform the gate. When averaged over uses of the gate the resulting noisy operation is $\mathcal{E}\mathcal{U}$ where $\mathcal{U}$ is the desired gate and the noise process amounts to

$$\mathcal{E}(\rho) = p\sigma_z e^{-i\delta\sigma_z}\rho e^{i\delta\sigma_z}\sigma_z + (1-p)e^{-i\delta\sigma_z}\rho e^{i\delta\sigma_z}. \tag{5}$$

In this noise model the dephasing noise rate p arises from the time-varying noise on the gate, while the unitary over rotation δ results from the fixed miscalibration of the control pulse $J(t)$. (Although we speak here in terms of calibration errors, this also approximately captures the effects of highly non-Markovian errors arising from very low-frequency noise in $J(t)$.)

This noise model roughly captures many experimental gates, but more importantly it will demonstrate the range of behaviors that can be expected in terms of the relationship between average-case and worst-case error. Specifically when $\delta = 0$ we have a pure dephasing process. For such errors [19] the worst case error scales like r, so this is the most favorable possible behavior. On the other hand for $p = 0$ we have a purely unitary rotation error that has the worst possible behavior, where the worst-case error scales like $\sqrt{r}$.

Using well-known techniques [30, 31] we find the average error rate for this calibration and dephasing (CD) noise

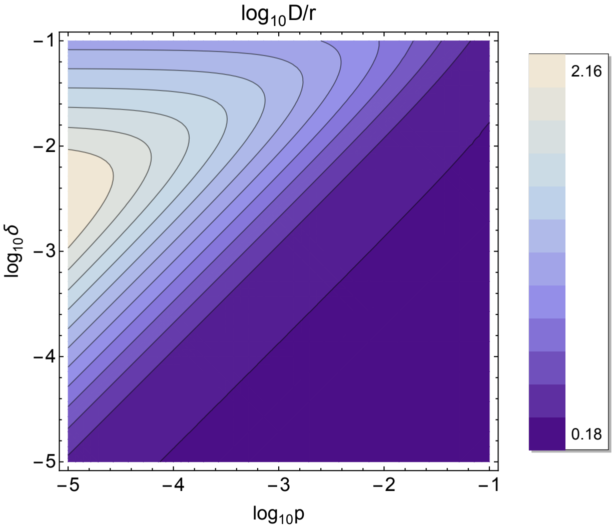


FIG. 1. Average error rate r and worst-case error rate (diamond distance) D for a combination of dephasing and unitary errors. The logarithmic plot is of D/r, which quantifies how much greater the worst-case error is than the average case as a function of a unitary over rotation angle δ and a dephasing probability p, where the exact noise process is given in Eq. (5). When $p \geq \delta$, then D and r are comparable to within a small factor, but as soon as $\delta > p$ then D rapidly becomes much greater than r.

to be $r_{\mathrm{CD}} = \frac{2}{3}\left(p\cos(2\delta) + \sin^2\delta\right)$. Employing the semidefinite programming approach of Refs. [19, 26], we can evaluate the diamond distance for this noise channel and find $D_{\mathrm{CD}} = \sqrt{\frac{3}{2}r_{\mathrm{CD}} - p(1-p)}$. A logarithmic plot of the ratio $D_{\mathrm{CD}}/r_{\mathrm{CD}}$ is shown in Figure 1.

In the interesting regime of low error we find $r_{\mathrm{CD}} \simeq 2(p+\delta^2)/3$, while $D_{\mathrm{CD}} \simeq \sqrt{p^2+\delta^2}$. From this we can see that when $p \gg |\delta|$ we have $D_{\mathrm{CD}} \simeq 3r_{\mathrm{CD}}/2$, as for a pure dephasing process, and there is no great difference between worst-case and average-case errors. But as the calibration error grows, the worst-case error grows significantly. When calibration error dominates, $|\delta| \gg p$, we find $D_{\mathrm{CD}} \simeq \sqrt{3r_{\mathrm{CD}}/2}$. In this regime an average error rate r_{CD} of around 10^{-4} corresponds to a more than one percent worst-case error. Physically then, we see that as dephasing error is reduced in a particular experimental setting, this places more stringent demands on the calibration required if the average error rate r is to be compared directly to a fault-tolerance threshold.

Single-qubit relaxation errors. Another natural single-qubit noise process to consider is qubit relaxation or amplitude damping errors (spontaneous emission or a T_1 process in NMR language), at finite temperature. In this process a qubit with energy splitting E is coupled to a bath at temperature T. Define as in [32] the probability for a decay process during the action of the gate is γp and the probability to go from the ground to the excited state is $\gamma(1-p)$. The ratio of upgoing to downgoing transition rates $p/(1-p) = \exp(-E/k_BT)$ is the Boltzmann factor, which allows us to identify $p = 1/2$ as infinite temperature and $p = 1$ as zero temperature. For this amplitude damping (AD) noise channel we find $r_{\mathrm{AD}} = \left(1 - \sqrt{1-\gamma} + \gamma/2\right)/3$. Although we have no closed form expression for the worst-case error for these channels, we have adapted standard techniques in the analysis of semidefinite programs to find the bound $D_{\mathrm{AD}} \leq 3r_{\mathrm{AD}}\max\{p, 1-p\}$. Therefore we have a guarantee that the average-case and worst-case errors are not too different. Comparing with a direct evaluation of the semidefinite program we find $D_{\mathrm{AD}} \simeq 3r_{\mathrm{AD}}$ for zero temperature ($p = 1$) and low noise $r_{\mathrm{AD}} \ll 1$, so this is the tightest bound possible. In the limit of high temperature $p \to 1/2$ we approach a dephasing channel and recover the formula $D_{\mathrm{AD}} = 3r_{\mathrm{AD}}/2$. This behavior is illustrated in Figure 2.

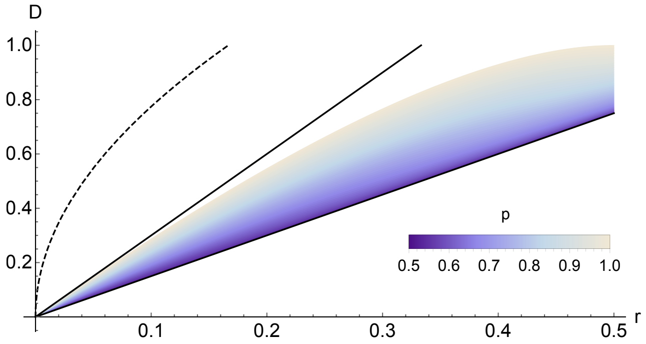


FIG. 2. Tradeoff between average error rate r and the worst-case error rate in terms of the diamond distance D for the thermal amplitude damping channel, where the parameter p controls the temperature with $p = 1$ corresponding to zero temperature and $p = 1/2$ corresponding to infinite temperature. The dashed line is the previous best upper bound [29], while the upper black line is the new bound derived here. The zero-temperature limit ($p = 1$) gives the least favorable scaling of D with r, but in every case the bound $D \leq 3r$ holds. The infinite-temperature limit ($p = 1/2$) recovers the known value of $D = 1.5r$.

Leakage errors Another important class of errors encountered in experiment is leakage errors. Modified randomized benchmarking protocols for leakage errors are proposed in [33, 34]. In Ref. [33] it was shown that a nearly trivial modification of a standard benchmarking protocol in the presence of leakage errors can still be used to determine the average error rate r, so we again use this figure of merit for comparison. For a leakage model we need to consider a larger space of states, so we add a leakage level $|l\rangle$ to the two-qubit states $|0\rangle, |1\rangle$. We follow [34] in distinguishing coherent and incoherent leakage errors and compare the average-case error to the true worst-case error; this will be the diamond distance on the *full* state space including both the leakage and qubit states. Fault-tolerance theorems also exist for leakage error processes [35] and this is the appropriate noise measure to compare with the numerical values found in those papers.

As an example of incoherent leakage (IL) we will consider the case where the qubit state $|1\rangle$ relaxes to $|l\rangle$ with probability p. A benchmarking experiment (following [33]) then obtains the average-case error $r_{\rm IL} = [1 - \sqrt{1-p} + p]/3$ where this is now the infidelity averaged over states initially in the qubit subspace. Since this process is so similar to the amplitude damping channel we can use analogous techniques to find the inequality $D_{\rm IL} \leq 2r_{\rm IL}$. Thus for this error process the average-case and worst-case error again almost coincide.

As an example of coherent leakage (CL), consider the unitary noise process $\mathcal{E}_{\rm CL}(\rho) = U(\delta)\rho U(\delta)^\dagger$ given by $U(\delta) = \exp[-i\delta(|1\rangle\langle l| + |l\rangle\langle 1|)]$. For this noise process one obtains $r_{\rm CL} = [1 - \cos\delta - \cos^2\delta]/3$. However, as for the unitary errors discussed above, the worst-case error can be much larger than this. We find $\sqrt{3r_{\rm CL}/2} \leq D_{\rm CL} = |\sin\delta| \leq \sqrt{2r_{\rm CL}}$ for all $\delta \in [-\pi/2, \pi/2]$ and consequently the worst case error scales like $\sqrt{r_{\rm CL}}$. Where leakage errors are possible, it would be important to use the methods of [34], or some other method to bound coherent leakage errors, before comparing the average-case error r to a fault-tolerance threshold.

Unitary errors. In looking at these examples we have found that unitary or nearly unitary errors appear to result in the largest difference between average-case and worst-case errors. This is true in general. For unitary errors in a d-dimensional space we find the following inequalities

$$\sqrt{\tfrac{d+1}{d}}\sqrt{r_{\rm U}} \leq D_{\rm U} \leq \sqrt{(d+1)d}\sqrt{r_{\rm U}}.$$

Thus any unitary error has a worst-case error scaling like $\sqrt{r_{\rm U}}$.

A general inequality. For a large and important class of noise processes, the worst-case error can be *directly* estimated from benchmarking-type data without side information about the type of error, which generally requires doing full quantum process tomography [36], or one of its SPAM-resistant variants [37, 38]. In Ref. [39] a quantity called the unitarity $u(\mathcal{E})$ of a noise process $\mathcal{E}$ was defined (see the Supplementary Material for a precise definition), and it was shown that this can be estimated efficiently and accurately using benchmarking. We find that for all unital noise (i.e. noise where the maximally mixed state is a fixed point) with no leakage, the unitarity and the average error rate together give a characterization of the worst-case error via the inequality [40]

$$c_d\sqrt{u + \frac{2dr}{d-1} - 1} \leq D \leq d^2 c_d\sqrt{u + \frac{2dr}{d-1} - 1}\,, \qquad (6)$$

where $c_d = \frac{1}{2}(1 - \frac{1}{d^2})^{1/2}$. Since the unitarity generally obeys the inequality $u \geq (1 - dr/(d-1))^2$ (see Ref. [39]) we find (for unital noise without leakage) that the worst-case error scaling matches the average-case *if and only if* $u = 1 - 2dr/(d-1) + O(r^2)$.

To illustrate the power of Inequality (6), we immediately find that for the single-qubit calibration and dephasing noise model, the condition $1 - u_{\rm CD} = 4r_{\rm CD} + O(r_{\rm CD}^2)$ is both necessary and sufficient to recover the favorable linear scaling between the worst- and average-case errors. In fact, the worst-case error for this channel can be expressed directly in terms of the unitarity as $D_{\rm CD} = \sqrt{\frac{3}{2}r_{\rm CD} - \frac{3}{8}(1 - u_{\rm CD})}$. And because the unitarity can be estimated from a benchmarking-type experiment, this gives direct experimental access to worst-case errors for this family of noise models without the need for expensive tomographic methods.

Moreover, Inequality (6) allows us to get insights into generalizing our conclusions for single-qubit models to few-qubit systems such as those required for entangling quantum gates. A natural generalization of our CD model to two-qubit calibration and dephasing errors would be an independent dephasing rate p on each qubit and unitary noise given by $\mathrm{e}^{iH_{\rm CD2}}$ where $H_{\rm CD2} = \delta_1\sigma_z^{(1)} + \delta_2\sigma_z^{(2)} + \epsilon\sigma_z^{(1)}\sigma_z^{(2)}$. The semidefinite programming approach is possible here, but becomes unwieldy because there are so many free parameters. However, both the average error rate and the unitarity are readily computed as in the appendix. Inequality (6) then allows one to easily and generally explore the tradeoffs in the calibration accuracy of the δ and ϵ parameters such that the overall error remains roughly consistent between average and worst case. Furthermore, since $u_{\rm CD2}$ can be measured efficiently in a benchmarking experiment, large values of u can be used to herald that an experiment has left the favorable scaling regime and more characterization and calibration must be done.

Conclusion and Outlook. We have seen that many realistic noise processes admit a linear relation between the average error rate (which is accessible experimentally) and the worst-case error (which is the relevant figure of merit for fault tolerance). The exceptions to this rule are highly coherent errors, where the worst-case error scales proportionally to the square root of the average error rate.

While our methods and results are very general, there are noise sources that we have not tried to fit into our error taxonomy. Errors such as crosstalk [41], time-dependent or non-Markovian noise [42, 43] should be amenable to these methods, however, and extending our results to cover such noise is an important avenue for future work.

Finally, we reiterate that it is an interesting open question if it is possible to prove a fault-tolerance threshold result directly in terms of r without the lossy conversion to D. Fault-tolerant circuits are not perfectly coherent since measuring error syndromes necessarily removes certain coherences, and this may provide an avenue to develop stronger theorems.

Acknowledgments. We thank R. Blume-Kohout, D. Gottesman, J. Gambetta, C. Granade and J. Wallman for discussions. This work was supported by the ARC via EQuS project number CE11001013, by IARPA via the MQCO program, and by the US Army Research Office grant numbers W911NF-14-1-0098 and W911NF-14-1-0103. STF acknowledges support from an ARC Future Fellowship FT130101744.

[1] A. Y. Kitaev, Russian Math. Surv. **52**, 1191 (1997).

[2] D. Aharonov and M. Ben-Or, in *29th ACM Symp. on Theory of Computing (STOC)* (New York, 1997) pp. 176–188,

arXiv:quant-ph/9906129.

[3] E. Knill, R. Laflamme, and W. H. Zurek, Proc. R. Soc. A **454**, 365 (1998), arXiv:quant-ph/9702058.

[4] J. Emerson, R. Alicki, and K. Życzkowski, J. Opt. B **7**, S347 (2005), arXiv:quant-ph/0503243.

[5] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. Blakestad, J. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. Wineland, Phys. Rev. A **77**, 012307 (2008), arXiv:0707.0963.

[6] E. Magesan, J. M. Gambetta, B. R. Johnson, C. A. Ryan, J. M. Chow, S. T. Merkel, M. P. da Silva, G. A. Keefe, M. B. Rothwell, T. A. Ohki, M. B. Ketchen, and M. Steffen, Phys. Rev. Lett. **109**, 080505 (2012), arXiv:1203.4550.

[7] C. Granade, C. Ferrie, and D. G. Cory, New J. Phys. **17**, 013042 (2015), arXiv:1404.5275.

[8] J. M. Chow, J. M. Gambetta, L. Tornberg, J. Koch, L. S. Bishop, A. A. Houck, B. R. Johnson, L. Frunzio, S. M. Girvin, and R. J. Schoelkopf, Phys. Rev. Lett. **102**, 090502 (2009), arXiv:0811.4387.

[9] K. Brown, A. Wilson, Y. Colombe, C. Ospelkaus, A. Meier, E. Knill, D. Leibfried, and D. Wineland, Phys. Rev. A **84**, 030303 (2011), arXiv:1104.2552.

[10] J. Gaebler, A. Meier, T. R. Tan, R. Bowler, Y. Lin, D. Hanneke, J. Jost, J. Home, E. Knill, D. Leibfried, *et al.*, Phy. Rev. Lett. **108**, 260503 (2012), arXiv:1203.3733.

[11] J. M. Chow, J. M. Gambetta, A. D. Córcoles, S. T. Merkel, J. A. Smolin, C. Rigetti, S. Poletto, G. A. Keefe, M. B. Rothwell, J. R. Rozen, M. B. Ketchen, and M. Steffen, Phys. Rev. Lett. **109**, 060501 (2012), arXiv:1202.5344.

[12] A. D. Córcoles, J. M. Gambetta, J. M. Chow, J. A. Smolin, M. Ware, J. Strand, B. L. T. Plourde, and M. Steffen, Phys. Rev. A **87**, 030301 (2013), arXiv:1210.7011.

[13] J. M. Chow, J. M. Gambetta, E. Magesan, D. W. Abraham, A. W. Cross, B. Johnson, N. A. Masluk, C. A. Ryan, J. A. Smolin, S. J. Srinivasan, *et al.*, Nature Commun. **5** (2014), 10.1038/ncomms5015, arXiv:1311.6330.

[14] R. Barends, J. Kelly, A. Megrant, A. Veitia, D. Sank, E. Jeffrey, T. C. White, J. Mutus, A. G. Fowler, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, C. Neill, P. O'Malley, P. Roushan, A. Vainsencher, J. Wenner, A. N. Korotkov, A. N. Cleland, and J. M. Martinis, Nature **508**, 500 (2014), arXiv:1402.4848.

[15] T. Harty, D. Allcock, C. Ballance, L. Guidoni, H. Janacek, N. Linke, D. Stacey, and D. Lucas, Phys. Rev. Lett. **113**, 220501 (2014), arXiv:1403.1524.

[16] J. T. Muhonen, A. Laucht, S. Simmons, J. P. Dehollain, R. Kalra, F. E. Hudson, S. Freer, K. M. Itoh, D. N. Jamieson, J. C. McCallum, A. S. Dzurak, and A. Morello, J. Phys.: Condens. Matter **27**, 154205 (2015), arXiv:1410.2338.

[17] T. Xia, M. Lichtman, K. Maller, A. W. Carr, M. J. Piotrowicz, L. Isenhower, and M. Saffman, Phys. Rev. Lett. **114**, 100503 (2015), arXiv:1501.02041.

[18] Y. R. Sanders, J. J. Wallman, and B. C. Sanders, New J. Phys. **18**, 012002 (2015), arXiv:1501.04932.

[19] E. Magesan, J. M. Gambetta, and J. Emerson, Phys. Rev. A **85**, 042311 (2012), arXiv:1109.6887.

[20] P. Aliferis, D. Gottesman, and J. Preskill, Quant. Inf. Comput. **6**, 97 (2006), arXiv:quant-ph/0504218.

[21] P. Aliferis and A. W. Cross, Phys. Rev. Lett. **98**, 220502 (2007), arXiv:quant-ph/0610063.

[22] P. Aliferis, F. Brito, D. P. DiVincenzo, J. Preskill, M. Steffen, and B. M. Terhal, New J. Phys. **11**, 013061 (2009), arXiv:0806.0383.

[23] D. S. Wang, A. G. Fowler, and L. C. L. Hollenberg, Phys. Rev. A **83**, 020302 (2011), arXiv:1009.3686.

[24] A. Gilchrist, N. K. Langford, and M. A. Nielsen, Phys. Rev. A **71**, 062310 (2005), arXiv:quant-ph/0408063.

[25] C. W. Helstrom, Information and Control **10**, 254 (1967).

[26] J. Watrous, Theory of Computation **5**, 217 (2009), arXiv:0901.4709.

[27] A. Ben-Aroya and A. Ta-Shma, Quant. Inf. Comp. **10**, 77 (2010), arXiv:0902.3397.

[28] J. Watrous, Chicago J. Theo. Comp. Sci. **2013**, 1 (2013), arXiv:1207.5726.

[29] J. J. Wallman and S. T. Flammia, New J. Phys. **16**, 103032 (2014), arXiv:1404.6025.

[30] M. A. Nielsen, Phys. Lett. A **303**, 249 (2002), arXiv:quant-ph/0205035.

[31] M. Horodecki, P. Horodecki, and R. Horodecki, Phys. Rev. A **60**, 1888 (1999), quant-ph/9807091.

[32] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, New York, 2000).

[33] J. M. Epstein, A. W. Cross, E. Magesan, and J. M. Gambetta, Phys. Rev. A **89**, 062321 (2014), arXiv:1308.2928.

[34] J. J. Wallman, M. Barnhill, and J. Emerson, Phys. Rev. Lett. **115**, 060501 (2015), arXiv:1412.4126.

[35] P. Aliferis and B. M. Terhal, Quant. Inf. Comp. **7**, 139 (2007), quant-ph/0511065.

[36] I. L. Chuang and M. A. Nielsen, J. Mod. Opt. **44**, 2455 (1997), arXiv:quant-ph/9610001.

[37] S. Kimmel, M. P. da Silva, C. A. Ryan, B. R. Johnson, and T. Ohki, Phys. Rev. X **4**, 011050 (2014), arXiv:1306.2348.

[38] R. Blume-Kohout, J. K. Gamble, E. Nielsen, J. Mizrahi, J. D. Sterk, and P. Maunz, (2013), arXiv:1310.4492.

[39] J. Wallman, C. Granade, R. Harper, and S. T. Flammia, New J. Phys. **17**, 113020 (2015), arXiv:1503.07865.

[40] A similar bound was recently derived independently by J. Wallman [44].

[41] J. M. Gambetta, A. D. Córcoles, S. T. Merkel, B. R. Johnson, J. A. Smolin, J. M. Chow, C. A. Ryan, C. Rigetti, S. Poletto, T. A. Ohki, M. B. Ketchen, and M. Steffen, Phys. Rev. Lett. **109**, 240504 (2012), arXiv:1204.6308.

[42] M. A. Fogarty, M. Veldhorst, R. Harper, C. H. Yang, S. D. Bartlett, S. T. Flammia, and A. S. Dzurak, Phys. Rev. A **92**, 022326 (2015), arXiv:1502.05119.

[43] H. Ball, T. M. Stace, S. T. Flammia, and M. J. Biercuk, Phys. Rev. A **93**, 022303 (2015), arXiv:1504.05307.

[44] J. J. Wallman, (2015), arXiv:1511.00727.

[45] J. Watrous, "CS 766 Theory of Quantum Information," Available online at `https://cs.uwaterloo.ca/~watrous/LectureNotes.html` (2011).

[46] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, New York, 2004).

[47] L. Vandenberghe and S. Boyd, SIAM Rev. **38**, 49 (1996).

[48] A. Barvinok, *A Course in Convexity*, Vol. 54 (American Mathematical Society, Providence, 2002).

[49] $\mathcal{E}(\rho)$ and $\tilde{\mathcal{E}}(\rho) = U\mathcal{E}(U^\dagger \rho U)U^\dagger$ have equal diamond distance and average error rate.

[50] M. Kliesch, R. Kueng, J. Eisert, and D. Gross, (2015), arXiv:1511.01513.

SUPPLEMENTARY MATERIAL

Quantum states and operations

A d-level quantum system is fully characterized by its is density operator ρ, which is a Hermitian, positive semidefinite $d \times d$ matrix obeying $\mathrm{Tr}(\rho) = 1$. A *quantum operation* or *channel* $\mathcal{E}$ is a completely positive linear map from density operators to density operators [32, 45].

There are a number of representations of a completely positive operator, each of which is useful for different purposes. The most well known is the representation in terms of Kraus operators. These are a set of operators $\{K_i\}$ that encapsulate the channel's action via $\mathcal{E}(\rho) = \sum_i K_i \rho K_i^\dagger$. Moreover, $\sum_i K_i^\dagger K_i \leq I$ holds, where I is the identity matrix, and equality occurs when $\mathcal{E}$ is trace preserving.

Other representations include the Liouville operator $L(\mathcal{E}) = \sum_i \overline{K_i} \otimes K_i$ where $\otimes$ denotes the tensor product. The Liouville operator is also known as the transition matrix, or natural representation. It is a matrix that acts on the vector obtained by stacking the columns of ρ, which we denote $|\rho)$ as in [29], in the same way that $\mathcal{E}$ acts on the density operator ρ. That is $L(\mathcal{E})|\rho) = |\mathcal{E}(\rho))$.

Lastly, we will have cause to use the Choi-Jamiołkowski matrix of a quantum operation $\mathcal{E}$, $J(\mathcal{E}) = d(\mathcal{I}_A \otimes \mathcal{E}_B)(|\psi_{\text{Bell}}\rangle\langle\psi_{\text{Bell}}|)$. Here $\mathcal{I}$ is the identity channel and $|\psi_{\text{Bell}}\rangle = \frac{1}{\sqrt{d}}\sum_{j=1}^{d} |j\rangle \otimes |j\rangle$ is the maximally entangled state between systems A and B (this definition differs by a factor of d to that in [29], instead we use the definition found in [19, 45] so as to be consistent with the semidefinite program in [26], which would otherwise require minor modification). It can be computed from the Kraus operators $\{K_i\}$ with the formula $J(\mathcal{E}) = \sum_i |K_i)(K_i|$ (where $(K_i| = |\bar{K_i})^T$).

This representation is useful because, unlike the other representations mentioned here, $J(\mathcal{E})$ is positive semidefinite for any completely positive quantum operation (the Kraus operators and Liouville operator need not even have a complete set of eigenvectors).

We will be interested in relating the average infidelity $r(\mathcal{E})$ to the diamond distance $D(\mathcal{E})$ as defined in the main text in Eqs. (1) and (4), respectively. (We will always be comparing a noise process to the identity channel, so we write the diamond distance with only one argument for brevity.) A useful formula is provided by the following relation which is a generalization of the main results in [30, 31] to completely positive maps that are not necessarily trace preserving.

Proposition 1. *Let $\mathcal{E}$ be a completely positive (but not necessarily trace preserving) map with Liouville representation $L(\mathcal{E})$. Then*

$$F_{\text{avg}}(\mathcal{E}) = \frac{\mathrm{Tr}[L(\mathcal{E})] + \mathrm{Tr}[\mathcal{E}(I)]}{d(d+1)}, \tag{7}$$

where $F_{\text{avg}}(\mathcal{E}) = 1 - r(\mathcal{E})$ is the average fidelity and d is the system size.

Note that this formula covers the main results in [30, 31] as a special case. Indeed, any trace preserving map obeys $\mathrm{Tr}\left(\mathcal{E}(I)\right) = d$ and Eq. (7) reduces to [31][Proposition 1] and [30][Equation (3)], respectively. For the scope of our work, such a generalization is very useful, since it will allow us to evaluate the fidelity of leakage processes averaged over qubit states.

Proof of Proposition 1. One way of proving the generalized formula (7) is to follow Nielsen's simplified proof steps [30] of the original formula [31] without assuming that $\mathcal{E}$ is trace preserving. At the core of this proof is the fact that the average fidelity is invariant under twirling, i.e. $F_{\text{avg}}\left(\mathcal{E}\right) = F_{\text{avg}}\left(\mathcal{E}_T\right)$ for $\mathcal{E}_T(\rho) := \int \mathrm{d}U U^\dagger \mathcal{E}\left(U\rho U^\dagger\right) U^\dagger$. Here $\mathrm{d}U$ denotes the unique unitarily invariant (Haar) measure over the unitary group $U(d)$ normalized to one ($\int \mathrm{d}U = 1$). The same is true for the r.h.s. of Eq. (7). Indeed, suppose that $\mathcal{E}$ has Kraus representation $\mathcal{E}(\rho) = \sum_i K_i \rho K_i^\dagger$.

Twirling it results in the map $\mathcal{E}_T(\rho) = \int \mathrm{d}U U^\dagger \sum_i (K_i U \rho U^\dagger K_i^\dagger) U$ whose Liouville representation obeys

$$\begin{aligned}\mathrm{Tr}\left(L\left(\mathcal{E}_T\right)\right) =& \mathrm{Tr}\left(\int \mathrm{d}U \sum_i \bar{U}^\dagger \bar{K}_i \bar{U} \otimes U^\dagger K_i U\right) = \int \mathrm{d}U \sum_i \mathrm{Tr}\left(\bar{U}^\dagger \bar{K}_i \bar{U}\right) \mathrm{Tr}\left(U^\dagger K_i U\right) \\ =& \sum_i \mathrm{Tr}\left(\bar{K}_i\right) \mathrm{Tr}\left(K_i\right) \int \mathrm{d}U = \mathrm{Tr}\left(\sum_i \bar{K}_i \otimes K_i\right) = \mathrm{Tr}\left(L\left(\mathcal{E}\right)\right).\end{aligned}$$

Also

$$\mathrm{Tr}\left(\mathcal{E}_T(I)\right) = \int \mathrm{d}U \mathrm{Tr}\left(U^\dagger \mathcal{E}\left(U I U^\dagger\right) U\right) = \mathrm{Tr}\left(\mathcal{E}\left(I\right)\right)$$

which establishes twirl invariance of the r.h.s. of (7). As a result, it suffices to establish the claimed equality for twirled maps only. However, due to Schur's Lemma, every twirl of a completely positive map is proportional to a depolarizing operation

$$\mathcal{E}_T(\rho) = \mathcal{D}_{p,q}(\rho) := p\rho + q\mathrm{Tr}(\rho) I \quad \forall \rho \tag{8}$$

with parameters p, q that may depend upon the original map $\mathcal{E}$. Nielsen [30] established this by using the following elementary argument based on the observation that any twirled channel obeys

$$V \mathcal{E}_T\left(\rho\right) V^\dagger = \mathcal{E}_T\left(V \rho V^\dagger\right) \quad \forall V \in U(d), \forall \rho \tag{9}$$

which is readily established by direct computation. Now let $X = |x\rangle\langle x|$ be a rank one projector, set $X^\perp = I - X$ and let V be an arbitrary unitary operator obeying $V X V^\dagger = X$. Inserting these particular choices into (9) reveals $\mathcal{E}_T(X) = \mathcal{E}_T\left(V X V^\dagger\right) = V \mathcal{E}_T\left(X\right) V^\dagger$ which in turn implies $\mathcal{E}_T(X) = (p+q)X + qX^\perp = pX + qI$ for some $p, q \in \mathbb{R}$. A priori, the parameters p, q may depend on the choice of X, but (9) implies that they are actually the same for any choice of X. From this, Formula (8) is readily deduced, e.g. by inserting eigenvalue decompositions $\rho = \sum_{i=1}^d \lambda_i |x_i\rangle\langle x_i|$ of arbitrary density operators and exploiting linearity.

As a result, it suffices to establish Formula (7) exclusively for depolarizing maps $\mathcal{D}_{p,q}$ of the form (8) with parameters p, q. Noting that such a map has Liouville representation $L\left(\mathcal{D}_{p,q}\right) = pI \otimes I + qd|\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}|$, where $|\psi_{\mathrm{Bell}}\rangle = \frac{1}{\sqrt{d}} \sum_{i=1}^d |i\rangle \otimes |i\rangle$ denotes a maximally entangled state, and calculating

$$\begin{aligned}F_{\mathrm{avg}}\left(\mathcal{D}_{p,q}\right) =& p \int \mathrm{d}\psi \langle\psi|\psi\rangle\langle\psi|\psi\rangle + q \int \mathrm{d}\psi \mathrm{Tr}\left(|\psi\rangle\langle\psi|\right) \langle\psi|I|\psi\rangle = p + q, \\ \mathrm{Tr}\left(L\left(\mathcal{D}_{p,q}\right)\right) =& p\mathrm{Tr}\left(I \otimes I\right) + qd\mathrm{Tr}\left(|\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}|\right) = d^2 p + dq, \\ \mathrm{Tr}\left(\mathcal{D}_{p,q}\left(I\right)\right) =& p\mathrm{Tr}\left(I\right) + q\mathrm{Tr}\left(I\right)^2 = dp + d^2 q\end{aligned}$$

reveals

$$\mathrm{Tr}\left(L\left(\mathcal{D}_{p,q}\right)\right) + \mathrm{Tr}\left(\mathcal{D}_{p,q}\right) = (d+1)d(p+q) = (d+1)dF_{\mathrm{avg}}\left(\mathcal{D}_{p,q}\right),$$

thus establishing the desired statement. □

Semidefinite Programming

It is possible to efficiently calculate the diamond norm of a linear operator through the use of a *semidefinite program* if a full description of the channel is known [26–28].

A semidefinite program (SDP) is a form of mathematical optimization problem (specifically a convex optimization problem; see [46, 47] for a review). A mathematical optimization problem is very generally a specification of some objective function to be maximized (or minimized), subject to some constraints on allowed variables in the form of inequalities involving constraint functions. This can be stated in the form

$$\begin{aligned}\text{Maximize: } & f_0(z) \\ \text{Subject to: } & f_i(z) \leq b_i, \quad i = 1, ..., m.\end{aligned}$$

where f_0 is the objective function, the f_i's and b_i's encode the constraint functions, and z is the variable to be changed so as to maximize $f_0(z)$. Any value of z which meets the constraints of the problem is called feasible. In some contexts these problem specifications are called programs.

A convex optimization problem is a mathematical optimization problem in which the set of all feasible points is a convex set and the objective function to be maximized is concave, i.e. it satisfies $f(\tau\ x\ +\ (1-\tau)y) \geq \tau\ f\ (x) + (1-\tau)\ f(y)$ for any $\tau \in [0,1]$ and feasible x, y. Note that minimising a convex function f_0 over a convex set also fits this framework, because it is equivalent to maximising $(-f_0)$ which is concave. Concave functions have many desirable properties that render convex optimization tasks easier than general optimization problems (e.g. concavity assures that any local maximum is also a global maximum) [48].

Finally, a semidefinite program is a particular instance of a convex optimization problem where one aims to maximize a linear function (which is both concave and convex) over a convex subset of the cone of positive semidefinite matrices [48]. This cone induces a partial ordering on the space of all hermitian $d \times d$ matrices. Concretely, we write $X \geq Y$ if and only if $X - Y$ is positive semidefinite. With this notational convention, every SDP is of the form

$$\begin{aligned}\text{Maximize: } & \mathrm{Tr}\,(CX) \\ \text{Subject to: } & \Xi(X) \leq B, \\ & X \geq 0\,.\end{aligned} \tag{10}$$

and is specified by a triple (Ξ, B, C): B and C are hermitian matrices (not necessarily of the same dimensions) and Ξ is a linear map between these matrices spaces. An SDP of the form (10) is called a *primal program*. In a geometric sense, the problem here is to move as far along the direction of C as possible, while remaining inside the convex region specified by the matrix inequalities [46–48]. A wide variety of problems can be cast in terms of semidefinite programs and efficient methods are known that can solve them. Thus, finding an expression for a problem in terms of a semidefinite program reduces it to one in which the solution is easily found numerically, and sometimes even analytically.

Attached to every primal problem is another semidefinite program (10), known as its *dual program*. In a sense, it corresponds to a reverse problem and is given by

$$\begin{aligned}\text{Minimize: } & \mathrm{Tr}\,(ZB) \\ \text{Subject to: } & \Xi^*(Z) \geq C \\ & Z \geq 0,\end{aligned} \tag{11}$$

which is again completely specified by the triple (Ξ, C, B). Here, Ξ^* denotes the adjoint map of Ξ with respect to the trace-inner product, i.e. the unique map obeying $\mathrm{Tr}\,(\Xi^*(Z)X) = \mathrm{Tr}\,(Z\ \Xi(X))$ for all hermitian matrices X and Z.

Primal and dual SDP's are intimately related. In particular they have the property that any feasible value of the primal objective $\mathrm{Tr}(CX)$ is less than or equal to any feasible value of the dual objective $\mathrm{Tr}(ZB)$. Using the fact that positive semidefinite matrices $A, B, C \geq 0$ obey $\mathrm{Tr}(AB) \leq \mathrm{Tr}(AC)$ if and only if $B \leq C$ allows for an easy proof of this feature [48] via

$$\mathrm{Tr}\,(CX) \leq \mathrm{Tr}\,(\Xi^*(Z)X) = \mathrm{Tr}\,(Z\ \Xi(X)) \leq \mathrm{Tr}\,(ZB)\,,$$

where we also have employed the constraints in (11) and (10), respectively. This result is known as *weak duality*. Typically an even stronger relation – called *strong duality* – is true, namely that the optimum values of both problems coincide.

Weak duality allows us to find an upper bound for the optimum value of (10) in the form of any feasible value of (11). To be more explicit, if Z is feasible, then $\mathrm{Tr}(ZB)$ must be larger than or equal to any feasible $\mathrm{Tr}(CX)$. This in particular includes the maximal value $\mathrm{Tr}(CX^\sharp)$ of (10). However, since $\mathrm{Tr}(CX^\sharp)$ is maximal, it is by definition larger than or equal to any feasible value of $\mathrm{Tr}(CX)$. Consequently, the feasible values $\mathrm{Tr}(CX)$ and $\mathrm{Tr}(ZB)$ certify that the optimum primal value $\mathrm{Tr}(CX^\sharp)$ is in a certain range. These bounds are said to be *certificates*. Throughout this work, we will employ such certificates in order to find bounds for the diamond norm. What is more, if we can find a pair of feasible points X, Z that obey $\mathrm{Tr}(CX) = \mathrm{Tr}(ZB)$, then weak duality dictates that we have analytically found the optimum value for the program. We will also appeal to this fact later.

Semidefinite programs for the diamond distance

Watrous has provided several characterisations of the diamond distance in terms of semidefinite programs [26, 28]. We reproduce here a simplified version that can be used when the operator in question is a difference of quantum channels $\Delta = \mathcal{E} - \mathcal{F}$ [26], as this will always be the case for us. Given this condition, the following pair of primal and dual SDP's has an optimal value of $D = \frac{1}{2}\|\Delta\|_\diamond$:

Primal problem

$$\begin{array}{ll} \text{Maximize:} & \langle J(\Delta), W\rangle \\ \text{Subject to:} & W \le \rho \otimes I_d, \\ & \mathrm{Tr}(\rho) = 1, \\ & W \in \mathrm{Pos}(A \otimes B), \\ & \rho \in \mathrm{Pos}(A). \end{array} \tag{12}$$

Dual problem

$$\begin{array}{ll} \text{Minimize:} & \|\mathrm{Tr}_B(Z)\|_\infty \\ \text{Subject to:} & Z \ge J(\Delta), \\ & Z \in \mathrm{Pos}(A \otimes B). \end{array} \tag{13}$$

Here $\langle X, Y\rangle = \mathrm{Tr}(X^\dagger Y)$ is the Hilbert-Schmidt inner product of the matrices X and Y, $\mathrm{Pos}(A \otimes B)$ denotes the cone of positive semidefinite operators acting on the system $A \otimes B$ and $\mathrm{Tr}_B(X)$ is the partial trace of X over system B, i.e. the subsystem of X obtained when subsystem B is discarded. Also, $\|X\|_\infty$ denotes the operator norm of X, which is the maximum eigenvalue of X (if $X \ge 0$). Further information on these functions and spaces can be found in [32, 45].

Note that, stated as it is, the primal problem is almost, but not quite, of the primal SDP form introduced in (10). However, some straightforward manipulations allow one to convert this problem into such a standard form. Perhaps a bit surprisingly, the same is true for the dual problem which can also be recast as an instance of a dual SDP problem [26].

Finally, note that if Π_+ is the projector onto the positive eigenspace of $J(\Delta)$, then $\rho = \frac{1}{d}I$, $W = \frac{1}{d}\Pi_+$ are valid primal feasible values and $Z = \Pi_+ J(\Delta)\Pi_+$ is dual feasible. These feasible points were identified by Magesan, Gambetta, and Emerson [19], and inspired by their approach we will use similar constructions of primal and dual feasible points to get bounds on the diamond norm for various noise processes.

Dephasing and calibration errors for a single qubit

The channel described in the main text has Kraus operators $K_0 = \sqrt{1-p}U(\delta)$ and $K_1 = \sqrt{p}U(\delta)\sigma_z$, where $U(\delta) = \exp(-i\delta\sigma_z)$. Using the formula above for the average fidelity it is straightforward to show that

$$r_{\mathrm{CD}} = \frac{2}{3}\left[p\cos(2\delta) + \sin^2\delta\right].$$

Likewise evaluating the upper and lower bounds on D_{CD} arising from the primal and dual feasible solutions of Ref. [19] we find them to be equal and so obtain the result

$$D_{\mathrm{CD}} = \frac{1}{2}\left|1 - (1-2p)e^{2i\delta}\right|.$$

A simple algebraic manipulation then shows the result claimed in the main text

$$D_{\mathrm{CD}} = \sqrt{\frac{3}{2}r_{\mathrm{CD}} - p(1-p)}.$$

Thermal relaxation of a single qubit

This one-qubit channel $\mathcal{E}_{\text{AD}}$ is characterized by 2 parameters $p, \gamma \in [0,1]$ and four Kraus operators [32, Chapter 8.3.5]

$$K_0 = \sqrt{p}\begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\gamma} \end{pmatrix}, \quad K_1 = \sqrt{p}\begin{pmatrix} 0 & \sqrt{\gamma} \\ 0 & 0 \end{pmatrix}, \quad K_2 = \sqrt{1-p}\begin{pmatrix} \sqrt{1-\gamma} & 0 \\ 0 & 1 \end{pmatrix}, \quad K_3 = \sqrt{1-p}\begin{pmatrix} 0 & 0 \\ \sqrt{\gamma} & 0 \end{pmatrix}.$$

Repeating the procedure outlined in the previous subsection, we will use a refined dual feasible point to find a bound on the diamond distance in terms of the average fidelity. This feasible point improves over what can be obtained using the Magesan-Gambetta-Emerson feasible solution [19].

Theorem 1. *For the one-qubit amplitude damping channel defined above, the following relation is valid for any choice of parameters* $p, \gamma \in [0,1]$*:*

$$D_{\text{AD}} \leq 3 r_{\text{AD}} \max\{p, 1-p\}.$$

Proof. We first compute the Choi-Jamiołkowski matrix $J(\Delta)$ for $\Delta = \mathcal{I} - \mathcal{E}_{\text{AD}}$. In the basis $|00\rangle, |01\rangle, |10\rangle, |11\rangle$, this matrix is

$$J(\Delta) = \begin{pmatrix} (1-p)\gamma & 0 & 0 & 1-\sqrt{1-\gamma} \\ 0 & -(1-p)\gamma & 0 & 0 \\ 0 & 0 & -p\gamma & 0 \\ 1-\sqrt{1-\gamma} & 0 & 0 & p\gamma \end{pmatrix}. \tag{14}$$

The middle block is already negative semidefinite and so our dual feasible point Z can afford to have zero support on this subspace and still meet the constraints of Eq. (13). Let us therefore make the ansatz that

$$Z = \begin{pmatrix} x+y_0 & 0 & 0 & x \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ x & 0 & 0 & x+y_1 \end{pmatrix} = 2x|\psi_{\text{Bell}}\rangle\langle\psi_{\text{Bell}}| + y_0|00\rangle\langle 00| + y_1|11\rangle\langle 11| \tag{15}$$

where $x = \left(1-\sqrt{1-\gamma}+\gamma/2\right)/2$ and we will determine the parameters $y_0, y_1 \geq 0$. Such a choice of parameters assures that Z is positive semidefinite.

The only other constraint that must be respected is that $Z - J(\Delta)$ must be positive semidefinite. Let us define $x_- = \left(1-\sqrt{1-\gamma}-\gamma/2\right)/2 \geq 0$. Here we have used the elementary relation $1-\sqrt{1-\gamma} \geq \frac{\gamma}{2}$ (which follows from concavity of the square root). Secondly we can define $|\psi_{-\text{Bell}}\rangle = (|00\rangle - |11\rangle)/\sqrt{2}$. In terms of this we may write

$$\begin{aligned} Z - J(\Delta) =& 2x_-|\psi_{-\text{Bell}}\rangle\langle\psi_{-\text{Bell}}| + [y_0 - (1/2-p)\gamma]|00\rangle\langle 00| + [y_1 + (1/2-p)\gamma]|11\rangle\langle 11| \\ &+ (1-p)\gamma|01\rangle\langle 01| + p\gamma|10\rangle\langle 10|. \end{aligned} \tag{16}$$

Accordingly, this difference is positive semidefinite, if both

$$y_0 - (1/2-p)\gamma \geq 0 \quad \text{and} \quad y_1 + (1/2-p)\gamma \geq 0$$

hold. Setting $y_0 = \max\{\gamma/2 - p\gamma, 0\}$ and $y_1 = \max\{0, p\gamma - \gamma/2\}$ satisfies the requirements. The two cases correspond to $p \leq 1/2$ and $p \geq 1/2$ respectively. Such a choice of parameters assures that Z is a valid feasible point of the dual SDP (13) of the channel's diamond distance. Its objective function value amounts to

$$\begin{aligned} \|\mathrm{tr}_B(Z)\|_\infty =& \|2x\,\mathrm{tr}_B(|\psi_{\text{Bell}}\rangle\langle\psi_{\text{Bell}}|) + \mathrm{tr}_B(y_0|00\rangle\langle 00| + y_1|11\rangle\langle 11|)\|_\infty \\ =& \max\{x+y_0, x+y_1\} \\ =& (1-\sqrt{1-\gamma}+\gamma/2)/2 + \gamma|1-2p|/2 \\ \leq& (1-\sqrt{1-\gamma}+\gamma/2)(1+|1-2p|)/2 \\ =& (1-\sqrt{1-\gamma}+\gamma/2)\max\{p, 1-p\}. \end{aligned}$$

The inequality arises because $1-\sqrt{1-\gamma} \geq \gamma/2$ as noted above.

Using the formula of Eq. (7), one easily obtains $r_{\mathrm{AD}} = \frac{1}{3}\big(1-\sqrt{1-\gamma}+\frac{\gamma}{2}\big)$. From this we may conclude

$$D_{\mathrm{AD}} = \frac{1}{2}\|\Delta_{\mathrm{AD}}\|_{\diamond} \leq \|\mathrm{tr}_B(Z)\|_\infty \leq (1-\sqrt{1-\gamma}+\gamma/2)\max\{p,1-p\} = 3r_{\mathrm{AD}}\max\{p,1-p\}.$$

This is the inequality that was to be proven. □

Incoherent leakage errors

Our model of incoherent leakage errors for a single qubit is typical of a physical leakage process that may occur. We assume that the qubit state $|1\rangle$ can relax to a leakage state $|l\rangle$. We specify the noise process in terms of a leakage probability p and Kraus operators

$$K_0 = |0\rangle\langle 0| + \sqrt{1-p}|1\rangle\langle 1| + |l\rangle\langle l|, \quad K_1 = \sqrt{p}|l\rangle\langle 1|.$$

To compute the average fidelity over initial qubit states we note that this average fidelity is unchanged if we replace the noise process with the a noise map where the Kraus operators are $\Pi_q K_i \Pi_q$ and $\Pi_q = |0\rangle\langle 0| + |1\rangle\langle 1|$ is the projector on the qubit subspace. The resulting process maps the qubit subspace to the qubit subspace and is completely positive but not trace preserving. We can thus evaluate the average fidelity using Proposition 1 which is valid for non-trace-preserving maps. Given this we find $r_{\mathrm{IL}} = [4-(1+\sqrt{1-p})^2+p]/6 = [1-\sqrt{1-p}+p]/3$.

Note that if the average fidelity is computed over the full three-level space, the answer is slightly different and corresponds to $[1-\sqrt{1-p}+p/4]/3$. Using this alternate characterization of average error rate gives only minor quantitative and no qualitative changes to our conclusions. We therefore choose the average only over the qubit space as a more physically motivated quantity.

To bound the diamond norm error we modify the dual feasible solution that worked for the thermal relaxation process above. The Choi matrix of the channel difference is

$$J\left(\Delta_{\mathrm{IL}}\right) = -p|11\rangle\langle 11| + p|1l\rangle\langle 1l| + \left(\sqrt{1-p}-1\right)\left(|00\rangle\langle 11| + |11\rangle\langle 00| + |ll\rangle\langle 11| + |11\rangle\langle ll|\right).$$

We choose

$$Z = \left(1-\sqrt{1-p}\right)\left(|00\rangle\langle 00| + |ll\rangle\langle ll| + |00\rangle\langle ll| + |ll\rangle\langle 00|\right) + p|1l\rangle\langle 1l|.$$

as dual feasible point. It is clear that $Z \geq 0$ and the second feasibility condition follows from

$$Z - J\left(\Delta_{\mathrm{IL}}\right) = 3\left(1-\sqrt{1-p}\right)|\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}| + \left[p-\left(1-\sqrt{1-p}\right)\right]|11\rangle\langle 11|,$$

where here $|\psi_{\mathrm{Bell}}\rangle := \sum_{i=1}^{3}(|i\rangle\otimes|i\rangle)/\sqrt{3}$. A routine calculation verifies that the coefficient in front of $|11\rangle\langle 11|$ is nonnegative for any $p \in [0,1]$ and $Z - J\left(\Delta_{\mathrm{IL}}\right)$ is thus positive semidefinite. Inserting Z into the dual problem's objective function (13) yields

$$D_{\mathrm{IL}} \leq \|\mathrm{Tr}_B\left(Z\right)\|_\infty = \left\|\left(1-\sqrt{1-p}\right)\left(|l\rangle\langle l| + |0\rangle\langle 0|\right) + p|1\rangle\langle 1|\right\|_\infty = p \leq 2r_{\mathrm{IL}}. \tag{17}$$

This is the inequality that we wished to show. (The final inequality follows because $p = (p+2p)/3 \leq 2(1-\sqrt{1-p}+p)/3 = 2r_{IL}$ since $p/2 \leq 1-\sqrt{1-p}$ as noted above.)

We may also consider the following alternative model for incoherent leakage in d-dimensional quantum systems:

$$\mathcal{E}_{\mathrm{IL}}(\rho) = pP\rho P + (1-p)\rho,$$

with $p \in [0,1]$ and P is a rank-deficient orthogonal projection (i.e. $P \geq 0$, $P^2 = P$ and $1 \leq \mathrm{tr}(P) \leq d-1$). For single qubits ($d=2$), P necessarily coincides with a pure quantum state and we recover the incoherent leakage model examined in [34, Eq. (25)]. This channel model has the advantage that we can exactly determine its diamond distance:

$$D_{\mathrm{IL}} = p. \tag{18}$$

The related computations greatly simplifies if we exploit unitary invariance of both diamond distance and average error rate [49]. This unitary invariance allows us to w.l.o.g. assume that P is diagonal in the computational basis: $P = \sum_{k=1}^{\mathrm{rank}(P)} |k\rangle\langle k|$. The Choi matrix of $\Delta_{\mathrm{IL}} = \mathcal{I} - \mathcal{E}_{\mathrm{IL}}$ then amounts to

$$J\left(\Delta_{\mathrm{IL}}\right) = d\mathcal{I} \otimes \left(\mathcal{I} - \mathcal{E}_{\mathrm{IL}}\right)\left(|\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}|\right) = pd\left(|\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}| - |\psi_P\rangle\langle\psi_P|\right),$$

where $|\psi_P\rangle = \frac{1}{\sqrt{d}}\sum_{k=1}^{d}|kk\rangle = \frac{1}{\sqrt{d}}\sum_{k=1}^{\mathrm{rank}(P)}|kk\rangle$. In order to obtain an upper bound, we choose the following feasible point of the diamond distance's dual SDP: $Z = pd|\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}|$. Clearly, this matrix is a feasible point, because $Z \geq 0$ and $Z - J(\Delta_{\mathrm{IL}}) = pd|\psi_P\rangle\langle\psi_P| \geq 0$. It's corresponding objective function value amounts to

$$\|\mathrm{tr}_B(Z)\|_\infty = pd\|\mathrm{tr}_B\left(|\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}|\right)\|_\infty = pd\|\frac{1}{d}\mathbb{I}\|_\infty = p,$$

which serves as our upper bound on D_{IL}.

For a lower bound, we turn to the primal SDP of the diamond distance. We set $\rho = |d\rangle\langle d|$ and $W = |dd\rangle\langle dd|$ which is a feasible pair of primal variables ($W \leq \rho \otimes \mathbb{I}$, $\mathrm{tr}(\rho) = 1$ and $W, \rho \geq 0$). Evaluating the primal objective function at this point results in

$$(J(\Delta_{\mathrm{IL}}), W) = dp\left|\langle dd|\psi_{\mathrm{Bell}}\rangle\right|^2 - pd\left|\langle dd|\psi_P\rangle\right|^2 = p.$$

Note that this lower bound on D_{IL} coincides with the upper bound established below. Weak duality allows us to conclude (18).

Finally, the average error rate of $\mathcal{E}_{\mathrm{IL}}$ can be readily computed via Formula (7) and amounts to

$$r_{\mathrm{IL}} = p\left(1 - \frac{\mathrm{tr}(P)(\mathrm{tr}(P)+1)}{(d+1)d}\right) \in \left[\frac{2p}{d+1}, p\left(1 - \frac{2}{d(d+1)}\right)\right].$$

The upper bound is saturated for rank-one projectors P, while the lower bound is achieved for projectors with $\mathrm{rank}(P) = d-1$. Comparing this to $D_{\mathrm{IL}} = p$ reveals

$$D_{\mathrm{IL}} = \left(1 - \frac{\mathrm{tr}(P)(\mathrm{tr}(P)+1)}{(d+1)d}\right) r_{\mathrm{IL}} \leq \frac{d+1}{2} r_{\mathrm{IL}}.$$

The upper bound provided here is tight for $(d-1)$-dimensional projections and becomes increasingle loose for more rank-deficient ones. For single qubits ($d = 2$), however, the upper bound is tight and we obtain $D_{\mathrm{IL}} = \frac{3}{2} r_{\mathrm{IL}}$.. Finally, choosing $d = 3$ and $\mathrm{rank}(P) = 2$ mimics the dimensionalities ocurring in our previous model for incoherent leakage. For such a choice, we obtain

$$D_{\mathrm{IL}} = 2r_{\mathrm{IL}},$$

which agrees with (17), but is slightly stronger.

Coherent leakage errors

The coherent leakage process that we consider is a unitary error process

$$U(\delta) = \exp[-i\delta(|1\rangle\langle l| + |l\rangle\langle 1|)] = |0\rangle\langle 0| + \cos(\delta)(|1\rangle\langle 1| + |l\rangle\langle l|) - i\sin(\delta)(|1\rangle\langle l| + |l\rangle\langle 1|), \tag{19}$$

where $\delta \in [-\pi, \pi]$ mediates the error strength. We can derive the average-case error using the same trick as above of projecting onto the qubit subspace. Note that $\Pi_q U \Pi_q = |0\rangle\langle 0| + \cos(\delta)|1\rangle\langle 1|$. As a result we find $r_{\mathrm{CL}} =$

$[2 - \cos\delta - \cos^2\delta]/3$. Unlike the incoherent case, the average error rate here is by coincidence the same if we compute it in the projected space or in the three-level space.

On the other hand the computation of the diamond norm distance is more straightforward for unitary error models such as this, since the optimization over input states entangled with an ancilla in the definition is not required. More details of the computation of the diamond norm distance for general unitary errors are given in the following subsection. The result of Corollary 1 is that $D_{\mathrm{CL}} = |\sin\delta|$.

To relate worst-case and average-case error, we employ the relation $4\sin^2(\delta/2) \geq \sin^2\delta$ which assures

$$r_{\mathrm{CL}} = (1-\cos\delta)/3 + (1-\cos^2\delta)/3 = 2\sin^2(\delta/2)/3 + \sin^2(\delta)/3 \geq \sin^2(\delta)/2 = D_{\mathrm{CL}}^2/2.$$

On the other hand we can place a lower bound on the diamond norm distance. To tighten it, we will consider the case of moderately small error with $\delta \in [-\pi/2, \pi/2]$. This assures $\cos^2\delta \leq \cos\delta$ and we obtain

$$r_{\mathrm{CL}} = (1-\cos\delta)/3 + (1-\cos^2\delta)/3 \leq 2[1-\cos^2(\delta)]/3 = 2\sin^2(\delta)/3 = 2D_{\mathrm{CL}}^2/3.$$

So for the restricted range of $\delta \in [-\pi/2, \pi/2]$ we have

$$\sqrt{3r_{\mathrm{CL}}/2} \leq D_{\mathrm{CL}} \leq \sqrt{2r_{\mathrm{CL}}}$$

which is the inequality we intended to show and demonstrates that the diamond norm distance scales with $\sqrt{r_{\mathrm{CL}}}$.

Unitary errors

In this section we do not restrict ourselves to qubits anymore and consider d-dimensional unitary channels, i.e.

$$\rho \mapsto U\rho U^\dagger$$

where $U : \mathbb{C}^d \to \mathbb{C}^d$ is a unitary matrix ($UU^\dagger = U^\dagger U = I$). As we will show now, all channels of this form admit the unfavorable “square root” behavior where the worst-case error is roughly equal to the square root of the average case error. We summarize our results as follows.

Theorem 2. *Fix a dimension d and let $\mathcal{E}_{\mathrm{U}}$ be a unitary channel. Then*

$$\sqrt{\frac{d+1}{d}}\sqrt{r_{\mathrm{U}}} \leq D_{\mathrm{U}} \leq \sqrt{(d+1)d}\sqrt{r_{\mathrm{U}}}. \tag{20}$$

Moreover, for single-qubit unitary channels, the lower bound holds with equality, i.e. $D_{\mathrm{U}} = \sqrt{3r_{\mathrm{U}}/2}$.

While the lower bound in (20) is tight, we do not know if the dimensional dependence in the upper bound can be further improved and leave this for future work.

Proof of Theorem 2. Every unitary matrix U is normal and as such has an eigenvalue decomposition

$$U = \sum_{k=1}^{d} \mathrm{e}^{i\delta_k} |k\rangle\langle k|,$$

with eigenvalues $\mathrm{e}^{i\delta_k}$ on the complex unit circle and an orthogonal eigenbasis $\{|k\rangle\}_{k=1}^d$ of $\mathbb{C}^d$. It greatly facilitates our work if we define the maximally entangled state $|\psi_{\mathrm{Bell}}\rangle = \frac{1}{\sqrt{d}}\sum_{k=1}^d |k,k\rangle$ with respect to this eigenbasis. With such a choice, the channel's Choi matrix simply corresponds to

$$J(\mathcal{E}_{\mathrm{U}}) = d(\mathcal{E}_{\mathrm{U}} \otimes \mathcal{I})\left(|\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}|\right) = d\left(U \otimes \mathbb{1}\right)|\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}|\left(U^\dagger \otimes \mathbb{1}\right) = d|\phi_{\mathrm{U}}\rangle\langle\phi_{\mathrm{U}}|,$$

where $|\phi_{\mathrm{U}}\rangle = \frac{1}{\sqrt{d}} \sum_{k=1}^{d} \mathrm{e}^{i\delta_k} |kk\rangle$ is again a maximally entangled state. The channel's average error rate then corresponds to

$$r_{\mathrm{U}} = \frac{d - \langle \psi_{\mathrm{Bell}} | J\left(\mathcal{E}_{\mathrm{U}}\right) | \psi_{\mathrm{Bell}} \rangle}{d+1} = \frac{d - d\left|\langle \psi_{\mathrm{Bell}} | \phi_{\mathrm{U}} \rangle\right|^2}{d+1} = \frac{d^2 - \left|\sum_{k=1}^{d} \mathrm{e}^{i\delta_k}\right|^2}{d(d+1)}. \tag{21}$$

For the upper bound in (20), we use the fact that the Choi matrix of the channel difference $\Delta_{\mathrm{U}} = \mathcal{E}_{\mathrm{U}} - \mathcal{I}$ assumes the form

$$J\left(\Delta_{\mathrm{U}}\right) = d\left(\left(U \otimes \mathbb{1}\right) |\psi_{\mathrm{Bell}}\rangle\langle \psi_{\mathrm{Bell}}| \left(U^\dagger \otimes \mathbb{1}\right) - |\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}| \right)$$

which is proportional to the difference of two rank-one projectors. Such a matrix has two non-zero eigenvalues

$$\lambda_\pm = \pm d \sqrt{1 - \left|\langle \psi_{\mathrm{Bell}} | \phi_{\mathrm{U}} \rangle\right|^2} = \pm \sqrt{(d+1)d} \sqrt{r_{\mathrm{U}}}$$

and corresponding normalized eigenvectors $|v_+\rangle, |v_-\rangle \in \mathbb{C}^{d^2}$ – see e.g. [45, Example 2.3]. Setting $Z = \lambda_+ |v_+\rangle\langle v_+| \geq 0$ yields a valid dual feasible point for the diamond norm's dual SDP (13) and inserting it into the program's objective function reveals

$$D_U \leq \left\| \mathrm{Tr}_B\left(Z\right) \right\|_\infty \leq \left\| \mathrm{Tr}_B\left(Z\right) \right\|_1 = \mathrm{Tr}\left(Z\right) = \lambda_+ \langle v_+ | v_+ \rangle = \lambda_+ = \sqrt{(d+1)d} \sqrt{r_{\mathrm{U}}},$$

as claimed. Here we have made use of the basic norm inequality $\| \cdot \|_\infty \leq \| \cdot \|_1$ and the fact that the partial trace preserves positive semidefiniteness which in turn assures $\|\mathrm{Tr}_Y(Z)\|_1 = \mathrm{Tr}\left(\mathrm{Tr}_Y(Z)\right) = \mathrm{Tr}(Z)$.

For the lower bound, we use the fact that for the difference of two unitary channels, diamond norm and induced trace norm coincide [45, Theorem 20.7]. This in turn assures

$$D_{\mathrm{U}} = \frac{1}{2} \left\| \mathcal{E}_{\mathrm{U}} - \mathcal{I} \right\|_{1 \to 1} = \frac{1}{2} \max_{\|x\|_{\ell_2} = 1} \left\| U|x\rangle\langle x|U^\dagger - |x\rangle\langle x| \right\|_1 = \max_{\|x\|_{\ell_2}=1} \sqrt{1 - |\langle x|U|x\rangle|^2}, \tag{22}$$

where the last simplification once more exploits that the matrix of interest is a difference of two rank-one projectors. Choosing the particular vector $\tilde{x} = \sum_{k=1}^{n} |k\rangle / \sqrt{d}$ allows us to also conclude

$$D_{\mathrm{U}} \geq \sqrt{1 - |\langle \tilde{x}|U|\tilde{x}\rangle|^2} = \frac{1}{d} \sqrt{d^2 - \left| \sum_{k=1}^{d} \mathrm{e}^{i\delta_k} \right|^2} = \sqrt{\frac{d+1}{d}} \sqrt{r_{\mathrm{U}}}, \tag{23}$$

which is the lower bound presented in (20).

For single-qubit unitary channels this argument can be substantially strengthened: in fact the inequality sign in (23) can be replaced with actual equality. To see this, we first note that any unitary channel $\mathcal{E}_{\mathrm{U}}$ is invariant under a global phase change $U \mapsto \mathrm{e}^{i\phi} U$ in the defining unitary matrix. For two-dimensional unitaries, this gauge freedom assures that we can w.l.o.g. assume that U is of the form $\mathrm{e}^{i\delta}|0\rangle\langle 0| + \mathrm{e}^{-i\delta}|1\rangle\langle 1|$ with $\delta \in [-\pi, \pi]$. This in turn assures that any vector $x = x_1|0\rangle + x_2|1\rangle \in \mathbb{C}^2$ obeys

$$|\langle x, Ux\rangle|^2 = \left| \mathrm{e}^{i\delta}|x_1|^2 + \mathrm{e}^{-i\delta}|x_2|^2 \right|^2 = |x_1|^4 + 2\cos\left(2\delta\right)|x_1|^2|x_2|^2 + |x_2|^4.$$

Clearly, this function is ignorant towards individual phases of x_1, x_2 and when attempting to minimize it, we may focus on real coefficients only. Taking into account normalization allows us to restrict x_1 to the interval $[0,1]$ and setting $x_2^2 = 1 - x_1^2$. Doing so reveals

$$\min_{\|x\|_{\ell_2}=1} |\langle x, Ux\rangle|^2 = \min_{x_1 \in [0,1]} \left(x_1^4 + 2\cos(2\delta) x_1^2 (1 - x_1^2) + \left(1 - x_1^2\right)^2 \right) = \min_{x_1 \in [0,1]} \left(4 \sin^2(\delta) \left(x_1^4 - x_1^2\right) + 1 \right) \tag{24}$$

and maximizing the expression on the r.h.s. of (22) is therefore equivalent to finding the minimum of the particularly simple double-well potential (24). The minimal value of the latter is achieved for $x_1 = 1/\sqrt{2}$, which in turn assures that the vector $\tilde{x} = \left(|0\rangle + |1\rangle\right)/\sqrt{2}$ in fact minimizes $|\langle x, Ux\rangle|^2$ and – as claimed – the inequality sign in (23) can be replaced by equality. □

Similar techniques can be employed to exactly characterize the diamond distance of single qubit coherent leakage, as it was introduced in the previous subsection.

Corollary 1 (Diamond distance of coherent leakage)**.** *Consider the three-level coherent leakage channel $U(\delta)$ with $\delta \in [-\pi, \pi]$ introduced in* (19)*. Then, its diamond distance amounts to $D_{\mathrm{CL}} = |\sin(\delta)|$.*

Proof. We start by noting that $U(\delta)$ as introduced in (19) admits an eigenvalue decomposition of the form $U(\delta) = \left(|v_0\rangle\langle v_0| + \mathrm{e}^{i\delta}|v_+\rangle\langle v_+| + \mathrm{e}^{-i\delta}|v_-\rangle\langle v_-|\right)$, where $|v_0\rangle, |v_+\rangle, |v_-\rangle$ form an orthonormal basis of $\mathbb{C}^3$. Since this channel is unitary, we can employ the particularly simple formula (22) to calculate it's diamond distance:

$$D_{\mathrm{CL}} = \max_{\|x\|_{\ell_2}=1} \sqrt{1 - |\langle x|U(\delta)|x\rangle|^2} \tag{25}$$

Now note that for any vector $x = x_1|v_0\rangle + x_2|v_+\rangle + x_3|v_-\rangle$ (represented with respect to the eigenbasis of $U(\delta)$), we have

$$|\langle x|U(\delta)|x\rangle|^2 = \left| |x_1|^2 + |x_2|^2 \mathrm{e}^{i\delta} + |x_3|^2 \mathrm{e}^{-i\delta} \right|^2 .$$

An analysis similar to the one presented at the end of the proof of Theorem 2 reveals that such an expression is minimal for $x_1 = 0$ and $|x_2|^2 = |x_3|^2 = 1/2$. Inserting such an optimal vector into (25) implies

$$D_{\mathrm{CL}} = \max_{\|x\|_{\ell_2}=1} \sqrt{1 - |\langle x|U(\delta)|x\rangle|^2} = \sqrt{1 - \cos^2(\delta)} = |\sin(\delta)| \, ,$$

as claimed. □

The unitarity and average error rate for two-qubit processes

We now consider the noise process on two qubits in the main text, generated by $\mathrm{e}^{iH_{\mathrm{CD2}}}$ where $H_{\mathrm{CD2}} = \delta_1 \sigma_z^{(1)} + \delta_2 \sigma_z^{(2)} + \epsilon \sigma_z^{(1)} \sigma_z^{(2)}$. Because the unitarity and average error rate can be computed directly, without the need of analyzing a semidefinite program, we can simply use the formulas (7) and (31) (below) and do a direct computation. The average error rate is given by

$$r_{\mathrm{CD2}} = \frac{1}{10}\left[4(2p-1)\cos(2\delta)\cos(2\epsilon) - (1-2p)^2\cos(4\delta) + 4p(1-p) + 5\right] ,$$

and the unitarity is given by

$$u_{\mathrm{CD2}} = \tfrac{1}{15}\left([8p(1-p) - 4]^2 - 1\right) .$$

Here for simplicity we have choosen $\delta_1 = \delta_2 = \delta$. This computation is routine, so we omit the details.

The unitarity as a witness for unfavorable scaling

The key message of this work is that the diamond distance $D(\mathcal{E})$ of an error channel $\mathcal{E}$ may be proportional to the square root of its average error rate $r(\mathcal{E})$. This is undesirable, since it underlines that $D(\mathcal{E})$ – which is the crucial number for fault tolerance – may be orders of magnitude larger than $r(\mathcal{E})$ – a quantity that is routinely estimated via randomized benchmarking techniques. However, in our case studies we have found that for many channels this worst case behavior does not occur and there is a linear relationship $D(\mathcal{E}) = \mathcal{O}(r(\mathcal{E}))$. In this section, we provide a necessary and sufficient criterion for such a desirable relationship. It is based on the *unitarity*, a scalar that was introduced in [39] and quantifies the coherence (i.e. the "unitarity") of a given noise channel $\mathcal{E}$. To properly define it, we associate $\mathcal{E}$ with a reduced map $\mathcal{E}'$ that obeys $\mathcal{E}'(I) = 0$ as well as $\mathcal{E}'(X) = \mathcal{E}(X) - \frac{\mathrm{Tr}(\mathcal{E}(X))}{\sqrt{d}} I$ for every traceless X. We define the unitarity of $\mathcal{E}$ to be the following averaged quantity of the reduced map $\mathcal{E}'$:

$$u\left(\mathcal{E}\right) := \frac{d}{d-1} \int \mathrm{d}\psi \, \mathrm{Tr}\Big(\mathcal{E}'(|\psi\rangle\langle\psi|)^\dagger \mathcal{E}'(|\psi\rangle\langle\psi|)\Big). \tag{26}$$

Defined that way, the unitarity obeys $u(\mathcal{I}) = 1$ and its definition in terms of $\mathcal{E}'$ makes it sensitive towards possible non-unital and trace decreasing features of $\mathcal{E}$. In particular, it is also *insensitive* to unitary rotations, in the sense that if $\mathcal{U}$ and $\mathcal{V}$ are unitary quantum channels, then $u\left(\mathcal{U}\mathcal{E}\mathcal{V}\right) = u\left(\mathcal{E}\right)$ holds true for any quantum channel $\mathcal{E}$. As a result, the unitarity is independent of unitary pre- and post-rotations on the noise [39]. The unitarity boasts many other desirable properties and – perhaps most importantly – can be efficiently estimated via a modified randomized benchmarking experiment [39]. Moreover, it is related to the average error rate by means of the following inequality.

Proposition 2. *Let $\mathcal{E}$ be a not necessarily trace preserving quantum operation obeying $\mathrm{Tr}\big(\mathcal{E}(I)\big) \leq \mathrm{Tr}(I)$. Then the unitarity and average error rate of $\mathcal{E}$ obey*

$$u(\mathcal{E}) \geq \left(1 - \frac{dr(\mathcal{E})}{d-1}\right)^2, \tag{27}$$

where d denotes the dimension of the system.

This is a slightly more general version of the inequality in [39][Proposition 8] and we provide a new proof based on fundamental Schatten-norm inequalities below. For now, we content ourselves with stating the main result of this section: for a large family of error channels, nearly saturating the bound (27) is a necessary and sufficient condition for the desirable scaling relation $D(\mathcal{E}) = \mathcal{O}(r(\mathcal{E}))$.

Theorem 3. *Let $\mathcal{E}$ be an arbitrary unital and trace-preserving channel. Then the diamond distance $D(\mathcal{E})$ scales linearly in the average error rate $r = r(\mathcal{E})$, if and only if the bound (27) is saturated up to second order in $r(\mathcal{E})$, i.e.*

$$u(\mathcal{E}) = \left(1 - \frac{dr}{d-1}\right)^2 + \mathcal{O}\left(r^2\right). \tag{28}$$

Since both $r(\mathcal{E})$ and $u(\mathcal{E})$ can be efficiently estimated in actual experiments, Theorem 3 provides an efficient means to check whether or not $D(\mathcal{E})$ and $r(\mathcal{E})$ are of the same magnitude. It immediately follows from the following technical result.

Proposition 3. *Let $\mathcal{E}$ be a unital and trace-preserving quantum operation. Then $D := D(\mathcal{E})$, $r := r(\mathcal{E})$ and $u := u(\mathcal{E})$ are related via*

$$c_d \sqrt{u + \frac{2dr}{d-1} - 1} \leq D \leq d^2 c_d \sqrt{u + \frac{2dr}{d-1} - 1}, \tag{29}$$

where $c_d = \frac{1}{2}\left(1 - \frac{1}{d^2}\right)^{1/2} \in \left[\frac{\sqrt{3}}{4}, \frac{1}{2}\right]$ that only depends on the system dimension d.

To deduce Theorem 3 from this statement, let us start with assuming that (28) holds. Inserting this expression for u into the upper bound provided by Proposition 3 yields

$$D \leq d^2 c_d \sqrt{\left(1 - \frac{dr}{d-1}\right)^2 + \mathcal{O}(r^2) + \frac{2dr}{d-1} - 1} = d^2 c_d \sqrt{\frac{d^2}{(d-1)^2} r^2 + \mathcal{O}\left(r^2\right)} = \mathcal{O}\left(r\right),$$

as claimed. Conversely, suppose by contradiction that $u = \left(1 - \frac{dr}{d-1}\right)^2 + \mathcal{O}(r)$. Employing the lower bound provided by Proposition 3 in a similar fashion assures $D(\mathcal{E}) = \mathcal{O}(\sqrt{r})$ which definitely does not scale linearly in r.

In order to establish the remaining statements – Proposition 3 and Proposition 2 – it is very useful to choose a particular Liouville representation of error channels $\mathcal{E}$. Concretely, we let $\{B_1, \ldots, B_{d^2}\}$ be a unitary operator basis obeying $B_1 = \frac{1}{\sqrt{d}} I$ and $\mathrm{Tr}\left(B_i^\dagger B_j\right) = \delta_{i,j}$ (e.g. the normalized Pauli's with the identity as first element). If defined with respect to such a basis, $L(\mathcal{E})$ admits the following block structure

$$L(\mathcal{E}) = \begin{pmatrix} \frac{1}{d}\mathrm{Tr}\left(\mathcal{E}(I)\right) & e_{\mathrm{sdl}} \\ e_{\mathrm{nu}} & E_{\mathrm{u}} \end{pmatrix}, \tag{30}$$

where $e_{\mathrm{sdl}}, e_{\mathrm{nu}} \in \mathbb{C}^{d^2-1}$ encapsulate state dependent leakage and non-unitarity, respectively. With such a Liouville representation, the unitarity of $\mathcal{E}$ is proportional to the squared Frobenius (or Hilbert-Schmidt) norm of the unital block E_u [39][Proposition 1]:

$$u(\mathcal{E}) = \frac{1}{d^2-1}\left\|E_{\mathrm{u}}\right\|_2^2. \tag{31}$$

Moreover, such a block-matrix structure lets us establish the following relation [39][Proposition 9]

$$\|J(\mathcal{E})\|_2^2 = (d^2+1)u(\mathcal{E}) + \|e_{\mathrm{nu}}\|_{\ell_2}^2 + \|e_{\mathrm{sdl}}\|_{\ell_2}^2 + \frac{1}{d}\mathrm{Tr}\left(\mathcal{E}(I)\right). \tag{32}$$

between the unitarity and the channel's associated Choi matrix. Having laid out these relations, we are ready to prove the main technical result of this section.

Proof of Proposition 3. We start with pointing out that the statement's assumptions assure that both e_{nu} and e_{sdl} vanish. This considerably simplifies the block structure (30) of $L(\mathcal{E})$ as well as relation (32). At the heart of this statement is an inequality that relates the diamond norm of any map $\mathcal{M}$ to different Schatten-norms of its corresponding Choi matrix:

$$\frac{1}{d}\|J(\mathcal{M})\|_1 \leq \|\mathcal{M}\|_\diamond \leq \|J(\mathcal{M})\|_1, \tag{33}$$

see e.g. [29][Lemma 7]. Recalling $D(\mathcal{E}) = \frac{1}{2}\|\Delta\|_\diamond$ and weakening this estimate by employing the Schatten norm inequalities $\|X\|_2 \leq \|X\|_1 \leq \mathrm{rank}(X)\|X\|_2$ allows us to deduce

$$\frac{1}{2d}\|J(\Delta)\|_2 \leq D(\mathcal{E}) \leq \frac{d}{2}\|J(\Delta)\|_2, \tag{34}$$

because $J(\Delta)$ has at most rank d^2. Note that an analogous relation can be derived using the diamond norm bound presented in [50] instead of (33). As a matter of fact, the assumptions on $\mathcal{E}$ allow us to calculate $\|J(\Delta)\|_2$ explicitly. To do so, start with

$$\|J(\Delta)\|_2^2 = \|J(\mathcal{I}-\mathcal{E})\|_2^2 = \|d|\psi_{\mathrm{Bell}}\rangle\langle\psi_{\mathrm{Bell}}| - J(\mathcal{E})\|_2^2 = d^2\langle\psi_{\mathrm{Bell}},\psi_{\mathrm{Bell}}\rangle^2 - 2d\langle\psi_{\mathrm{Bell}}|J(\mathcal{E})|\psi_{\mathrm{Bell}}\rangle + \|J(\mathcal{E})\|_2^2 \tag{35}$$

and note that the second term is related to the average error rate via

$$\langle\psi_{\mathrm{Bell}}|J(\mathcal{E})|\psi_{\mathrm{Bell}}\rangle = (d+1)F_{\mathrm{avg}}(\mathcal{E}) - 1 = (d+1)(1-r(\mathcal{E})) - 1.$$

This can readily be deduced from (7) by inserting the identity $\mathrm{Tr}\left(L(\mathcal{E})\right) = d\langle\psi_{\mathrm{Bell}}|J(\mathcal{E})|\psi_{\mathrm{Bell}}\rangle$ and noting that $\mathrm{Tr}\left(\mathcal{E}(I)\right) = \mathrm{Tr}(I) = d$ holds, because $\mathcal{E}$ is trace-preserving. In turn, equation (32) allows to replace the last term in Eq. (35) by

$$\|J(\mathcal{E})\|_2^2 = (d^2-1)u(\mathcal{E}) + \frac{1}{d^2}\mathrm{Tr}\left(\mathcal{E}(I)\right)^2 + \|e_{\mathrm{sdl}}\|_{\ell_2}^2 + \|e_{\mathrm{n}}\|_{\ell_2}^2 = (d^2-1)u(\mathcal{E}) + 1,$$

where we have used our assumptions that $\mathcal{E}$ is both unital and trace preserving to considerably simplify this expression. Inserting these identities into Eq. (35) reveals

$$\begin{aligned}\|J(\Delta)\|_2^2 =& d^2 - 2d(d+1)(1-r(\mathcal{E})) + 2d + (d^2-1)u(\mathcal{E}) + 1 \\ =&(d^2-1)u(\mathcal{E}) + 2d(d+1)r(\mathcal{E}) + -d^2 + 1 \\ =&(d^2-1)\left(u(\mathcal{E}) + \frac{2dr(\mathcal{E})}{d-1} - 1\right).\end{aligned}$$

Plugging this explicit expression into the inequality chain Equation 34 then establishes the claim. □

Finally, we provide a proof of Propostion 2.

Proof of Proposition 2. The claim can be deduced from the fundamental norm inequality $\|X\|_1^2 \leq \operatorname{rank}(X)\|X\|_2^2$. Now, let $L(\mathcal{E})$ be the particular block matrix representation (30). By construction E_{u} has rank at most (d^2-1) and we infer that

$$\operatorname{Tr}(E_{\mathrm{u}})^2 \leq \|E_{\mathrm{u}}\|_1^2 \leq \operatorname{rank}(E_{\mathrm{u}})\|E_{\mathrm{u}}\|_2^2 = (d^2-1)^2 u(\mathcal{E}) \tag{36}$$

must hold, where we have employed Eq. (31). Also, Formula (7) together with the definition of the error rate implies

$$\operatorname{Tr}(L(\mathcal{E})) + \operatorname{Tr}(\mathcal{E}(I)) = d(d+1)F_{\mathrm{avg}}(\mathcal{E}) = d(d+1)(1-r(\mathcal{E})).$$

This in turn allows us to calculate

$$\begin{aligned}\operatorname{Tr}(E_{\mathrm{u}}) =& \operatorname{Tr}(L(\mathcal{E})) - \frac{1}{d}\operatorname{Tr}(\mathcal{E}(I)) = \operatorname{Tr}(L(\mathcal{E})) + \operatorname{Tr}(\mathcal{E}(I)) - \frac{d+1}{d}\operatorname{Tr}(\mathcal{E}(I))\\ =& d(d+1)(1-r(\mathcal{E})) - \frac{d+1}{d}\operatorname{Tr}(\mathcal{E}(I)) \geq d(d+1)(1-r(\mathcal{E})) - (d+1)\\ =& d(d+1)\left(\frac{d-1}{d} - r(\mathcal{E})\right),\end{aligned}$$

and combining this estimate with (36) readily yields the claimed bound. □

3 Additional results

This section is devoted to three novel results that were obtained in a recent collaboration with Zhu, Grassl and Gross [KZG16a; KZG16b; ZKGG16]. The publication drafts are ready for presentation and I include them in this chapter. I want to emphasize that the first paper [ZKGG16]—where the main representation theoretical result is derived—is in large parts the work of Zhu and Gross. However, Grassl and myself did provide relevant contributions. In contrast to this, I am the main contributor to the follow-up results presented in the other two drafts [KZG16a; KZG16b].

The Clifford group fails gracefully to be a unitary 4-design

Huangjun Zhu,[1] Richard Kueng,[1] Markus Grassl,[2] and David Gross[1]

[1] *Institute for Theoretical Physics, University of Cologne, Germany*
[2] *Max Planck Institute for the Science of Light, Leuchs Division, 91058 Erlangen, Germany*

(Dated: September 5, 2016)

A *unitary t-design* is a set of unitaries that is "evenly distributed" in the sense that the average of any t-th order polynomial over the design equals the average over the entire group. In various fields – e.g. quantum information theory – one frequently encounters constructions that rely on matrices drawn uniformly at random from the unitary group. It can often been shown that it suffices to sample these matrices from a t-design, for sufficiently high t. This results in more explicit, *derandomized* constructions. The most prominent unitary t-design considered in quantum information is the multi-qubit *Clifford group*. It is known that if forms a 3-design, but, unfortunately, not a 4-design. Here, we give a simple, explicit characterization of the way in which the Clifford group fails to constitute a 4-design. Our results show that for various applications in quantum information theory and in the theory of convex signal recovery, Clifford orbits perform almost as well as true 4-designs. Technically, it turns out that in a precise sense, the 4th tensor power of the Clifford group affords only one more invariant subspace than the 4th tensor power of the unitary group. That additional subspace is a *stabilizer code* – a structure extensively studied in the field of quantum error correction codes. This allows for an explicit analysis.

CONTENTS

I. INTRODUCTION

A. Designs and derandomizations

A d-dimensional *complex projective design* is a configuration of vectors that are "evenly distributed" on the unit sphere in $\mathbb{C}^d$. More precisely, a set of unit-length vectors is a *complex projective t-design*, if sampling uniformly from the set gives rise to a random vector whose first $2t$ moments agree with the moments of the uniform measure on the sphere. This property makes designs a useful tool for the derandomization of constructions that rely on random vectors. To motivate our work, we mention one example from signal analysis and one from quantum information theory.

1. *Application: Phase Retrieval*

The signal analysis example is the problem of *phase retrieval*: Let x be an unknown vector in $\mathbb{C}^d$. Assume we have access to a set of "phase insensitive linear measurements"

$$y_i = |(a_i, x)|, \qquad i = 1, \dots, m. \tag{1}$$

Here, the $a_i \in \mathbb{C}^d$ are a given set of *measurement vectors*. The task now is to recover x given $y_1, \dots, y_n$. There are many practical applications – for example in optical microscopy, where information about a sample is encoded in the electro-magnetic light field, but where only phase-insensitive intensity measurments are usually feasible. From a mathematical point of view, the absolute value in Eq. (1) means that we are facing a *non-linear inverse problem* – which are often difficult to solve in theory in practice.

A recent research program has investigated the use of algorithms based on *convex optimization* for the purpose of solving the phase retrieval problem. First theoretical results have shown that certain convex algorithms do indeed recover x with high probability, if the measurements a_i are random Gaussian vectors or drawn uniformly from the unit-sphere in $\mathbb{C}^d$ [1, 2]. However, in many practical applications, such measurements cannot be realized. Therefore, we are facing the task of re-proving those guarantees for measurements that are ideally deterministic, or, if randomized, at least drawn from a "smaller" and "more highly structured" set of vectors than from the entire unit-sphere. Such *derandomized* versions of have indeed been established for a variety of models–see e.g. Refs. [3, 4].

Starting with [5], some of the present authors have been interested in using spherical designs as "general-purpose" tool for derandomzing phase retrieval algorithms. The basic insight is that protocols that ostensibly require Gaussian vectors often only rely on certain measure concentration estimates that can be derived already from information about finite moments. Case in point is Ref. [6], which was proven initially for Gaussian measurements and then generalized – with comparatively few additional efforts – to any set of vectors which forms a 4-design.

2. *Application: POVM norm constants*

We take a related example from *quantum information theory.* In quantum mechanics, the *state* of a d-level system is encoded in a positive semi-definite $d\times d$-matrix, the so-called *density operator.* A *measurement* maps density operators to classical probability distributions over a space of outcomes. The fundamental property of *quantum complementarity* means that classical measurements necessarily entail a loss of information about the quantum system.

One way of precisely measuring this information loss is as follows: The (single-shot) statistical distinguishability of two classical probability distributions p, q is measured by the *total variational distance*, or half their ℓ_1-norm distance $d_c(p,q) := \frac{1}{2}\|p-q\|_{\ell_1}$. Analogously, the optimal probability of distinguishing between two quantum states ρ, σ is given by one half the *Schatten-1 norm* (or *trace norm* or *nuclear norm*) of their distance: $d_q(\rho,\sigma) := \frac{1}{2}\|\rho-\sigma\|_*$. Quantum measurements are represented by (certain) linear maps Λ from the set of density matrices to the set of classical probability distributions. The fact that "information is lost" in such a process can e.g. be made precise by stating that Λ is a strict contraction:

$$d_c(\Lambda(\rho),\Lambda(\sigma)) \leq C_\Lambda d_q(\rho,\sigma),$$

for some *POVM norm constant* $C_\Lambda < 1$. It thus makes sense to ask for an optimal measurement, i.e. one that maximises C_Λ. It has been shown that the *uniform POVM* achieves this goal [7]. This measurement maps quantum states to probability distributions on the complex unit sphere, where the density $p(\psi)$ at the vector ψ is proportional to $\operatorname{tr}\rho\,|\psi\rangle\langle\psi|$.

The situation is now very similar to the one considered in the phase retrieval example above: The uniform POVM is optimal, but impractical to implement in large quantum experiments. However, as has been shown already in Ref. [7], restricing the uniform POVM to a set of vectors that form a 4-design gives rise to a quantum measurement which matches the optimal scaling behavior.

3. *Outline of result: Overcoming the "$t = 3$-barrier"*

One major drawback of the program of using complex projective designs for derandomization is that there has been little progress in constructing explicit families of t-designs for $t > 3$. There are are various constructions using "structured randomness" – most notably the *random circuit model* that yields approximate designs in any dimension and of any degree [8, 9]. While the resulting designs are sufficiently well-structured for some tasks in quantum information theory, they are arguably not as explicit as one could hope for.

This situation seems all the more unsatisfactory, as there are various applications – including the two examples given above – where 2-designs are essentially useless (c.f. [7, 10]), 3-designs give first non-trivial improvements [10], and 4-designs show already optimal behavior.

The only explicit infinite family of complex projective 3-designs known to us are the orbits of the complex Clifford group [11–13]. Unfortunately, it has also been shown that Clifford orbits are not, in general, 4-designs [11–13].

The main result of the present work is that while Clifford orbits fall short of constituting 4-designs, their 4th moments can be explicitly calculated. The results are sufficiently well-behaved that for several applications, Clifford orbits turn out to perform nearly as well as 4-designs or Gaussian random vectors would. In order to establish these statements, we give an explicit description of the irreducible representations of the 4th tensor power of the Clifford group. In a precise sense, it turns out that the 4th tensor power of the Clifford group affords only one more invariant subspace than the 4th tensor power of the unitary group. That additional subspace is a *stabilizer code* – a structure extensively studied in the field of quantum error correction codes [14, 15]. This allows for an explicit analysis.

This paper contains only the representation-theoretic analysis of the 4th tensor power of the Clifford group. In two companion papers, we apply this technical result to the applications mentioned in the introduction: In [16], we establish performance guarantees for phase retrieval from stabilizer measurements; while [17] discuss the norm constants of stabilizer POVMs. The reason for splitting our discussion three-ways is that we target both problems form theoretical physics and from applied mathematics and that the respective communities employ very different language.

II. MATHEMATICAL BACKGROUND

In this section we review the mathematical background on complex projective designs and unitary designs.

A. Projective t-designs

Complex projective t-designs are of interest to a number research areas, such as approximation theory, combinatorics, experimental designs etc. Recently, they have also found increasing applications in many quantum information processing tasks, such as quantum state estimation [18–20], quantum state discrimination [7], and derandomization [21]. Here we review three equivalent definitions of (complex projective) t-designs; cf. [18, 22, 23].

Let $\mathrm{Hom}_{(t,t)}(\mathbb{C}^d)$ be the space of polynomials homogeneous of degree t in the coordinates of $|\psi\rangle \in \mathbb{C}^d$ (with respect to a given basis) and homogeneous of degree t in the coordinates of $\langle\psi|$.

Definition 1. A set of K pure states $\{|\psi_j\rangle\}$ in dimension d is a (complex projective) *t-design* if

$$\frac{1}{K}\sum_j p(\psi_j) = \int p(\psi)\mathrm{d}\psi \quad \forall p \in \mathrm{Hom}_{(t,t)}(\mathbb{C}^d), \tag{2}$$

where the integral is taken with respect to the normalized Haar measure induced by the action of the unitary group.

To derive simpler criteria on t-designs, we need to introduce several additional concepts. Let $\mathrm{Sym}_t(\mathbb{C}^d)$ be the t-partite symmetric subspace of $(\mathbb{C}^d)^{\otimes t}$ with corresponding projector P_t^{Sym}. The dimension of $\mathrm{Sym}_t(\mathbb{C}^d)$ reads

$$D_t^{\mathrm{Sym}} = \binom{d+t-1}{t}. \tag{3}$$

The tth frame potential of $\{|\psi_j\rangle\}$ is defined by

$$\Phi_t(\{|\psi_j\rangle\}) := \frac{1}{K^2}\sum_{j,k} |\langle\psi_j|\psi_k\rangle|^{2t}. \tag{4}$$

Proposition 1. *The following statements are equivalent:*

1. $\{|\psi_j\rangle\}$ *is a t-design.*
2. $\frac{1}{K}\sum_j(|\psi_j\rangle\langle\psi_j|)^{\otimes t} = P_t^{\mathrm{Sym}}/D_t^{\mathrm{Sym}}$.
3. $\Phi_t(\{|\psi_j\rangle\}) = 1/D_t^{\mathrm{Sym}}$.

Remark 1. In general, $\Phi_t(\{|\psi_j\rangle\}) \geq 1/D_t^{\mathrm{Sym}}$, and the lower bound is saturated iff $\{|\psi_j\rangle\}$ is a t-design.

Proof. Let $L(\mathrm{Sym}_t(\mathbb{C}^d))$ be the space of linear operators acting on $\mathrm{Sym}_t(\mathbb{C}^d)$. There is a one-to-one correspondence between polynomials $p \in \mathrm{Hom}_{(t,t)}(\mathbb{C}^d)$ and operators $A \in L(\mathrm{Sym}_t(\mathbb{C}^d))$,

$$A \mapsto p_A, \quad p_A(\psi) := \mathrm{tr}\big[A(|\psi\rangle\langle\psi|)^{\otimes t}\big]. \tag{5}$$

Therefore,

$$\frac{1}{K}\sum_j p_A(\psi_j) = \frac{1}{K}\,\mathrm{tr}\Big[A\sum_j(|\psi_j\rangle\langle\psi_j|)^{\otimes t}\Big], \quad \int p_A(\psi)\mathrm{d}\psi = \mathrm{tr}\Big[A\int(|\psi\rangle\langle\psi|)^{\otimes t}\mathrm{d}\psi\Big]. \tag{6}$$

It follows that $\{|\psi_j\rangle\}$ is a t-design iff

$$\frac{1}{K}\sum_j (|\psi_j\rangle\langle\psi_j|)^{\otimes t} = \int (|\psi\rangle\langle\psi|)^{\otimes t} \mathrm{d}\psi = \frac{1}{D_t^{\mathrm{Sym}}} P_t^{\mathrm{Sym}}. \tag{7}$$

Here the second equality follows from the fact the tth symmetric subspace is irreducible under the action of the unitary group. This observation confirms the equivalence of statements 1 and 2. The equivalence of statements 2 and 3 is a consequence of the following equation,

$$\left\| \frac{1}{K}\sum_j (|\psi_j\rangle\langle\psi_j|)^{\otimes t} - \frac{1}{D_t^{\mathrm{Sym}}} P_t^{\mathrm{Sym}} \right\|_2^2 = \Phi_t(\{|\psi_j\rangle\}) - \frac{1}{D_t^{\mathrm{Sym}}}, \tag{8}$$

where $\|\cdot\|_2$ denotes the Hilbert-Schmidt norm or the Frobenius norm. This equation implies that $\Phi_t(\{|\psi_j\rangle\}) \geq 1/D_t^{\mathrm{Sym}}$, and the lower bound is saturated iff Eq. (7) is satisfied. □

Any t-design in dimension d has at least

$$\binom{d+\lceil t/2\rceil - 1}{\lceil t/2\rceil}\binom{d+\lfloor t/2\rfloor - 1}{\lfloor t/2\rfloor} \tag{9}$$

elements, where $\lceil t/2\rceil$ denotes the smallest integer not smaller than $t/2$, and $\lfloor t/2\rfloor$ the largest integer not larger than $t/2$ [18, 24, 25]. The bound is equal to $d, d^2, d^2(d+1)/2, d^2(d+1)^2/4$ for $t = 1, 2, 3, 4$, respectively. A t-design is tight if the lower bound is saturated. A 1-design is tight iff it defines an orthonormal basis; a 2-design is tight if and only it defines a symmetric informationally complete measurement (SIC) [18, 22, 23, 26, 27]. Other prominent examples of 2-designs include complete sets of mutually unbiased bases (MUB) [28–30].

B. Unitary t-designs

Let $\mathrm{Hom}_{(t,t)}(\mathrm{U}(d))$ be the space of polynomials homogeneous of degree t in the matrix elements of $U \in \mathrm{U}(d)$ and homogeneous of degree t in thhe matrix elements of U^* (the complex conjugate of U; the Hermitian conjugate of U is denoted by $U^\dagger$).

Definition 2. A set of K unitary operators $\{U_j\}$ is a *unitary t-design* if

$$\frac{1}{K}\sum_j p(U_j) = \int \mathrm{d}U p(U) \quad \forall p \in \mathrm{Hom}_{(t,t)}(\mathrm{U}(d)), \tag{10}$$

where the integral is taken over normalized Haar measure. This equation remains intact even if U_j are multiplied by arbitrary phase factors, so what we are concerned are actually projective unitary t-designs.

The tth frame potential of $\{U_j\}$ is defined as

$$\Phi_t(\{U_j\}) := \frac{1}{K^2}\sum_{j,k} |\operatorname{tr}(U_j U_k^\dagger)|^{2t}. \tag{11}$$

As shown in the proof of Proposition 2 below,

$$\Phi_t(\{U_j\}) \geq \gamma(t,d) := \int \mathrm{d}U |\operatorname{tr}(U)|^{2t}, \tag{12}$$

and the lower bound is saturated iff $\{U_j\}$ is a unitary t-design [31–33]. The value of $\gamma(t,d)$ has been computed explicitly: it is equal to the number of permutations of $\{1, 2, \ldots, t\}$ with no increasing subsequence of length larger than d [34, 35]. Here we only need the formula in the following two special cases [32],

$$\gamma(t,d) = \begin{cases} \frac{(2t)!}{t!(t+1)!} & d = 2, \\ t! & d \geq t. \end{cases} \tag{13}$$

Like projective t-designs, there are many equivalent definitions of unitary t-designs.

Proposition 2. *The following statements are equivalent:*

1. *$\{U_j\}$ is a unitary t-design.*
2. *$\frac{1}{K}\sum_j \operatorname{tr}\bigl[BU_j^{\otimes t}A(U_j^{\otimes t})^\dagger\bigr] = \int \mathrm{d}U \operatorname{tr}\bigl[BU^{\otimes t}A(U^{\otimes t})^\dagger\bigr]$ for all $A, B \in L((\mathbb{C}^d)^{\otimes t})$.*
3. *$\frac{1}{K}\sum_j U_j^{\otimes t}A(U_j^{\otimes t})^\dagger = \int \mathrm{d}U U^{\otimes t}A(U_j^{\otimes t})^\dagger$ for all $A \in L((\mathbb{C}^d)^{\otimes t})$.*
4. *$\frac{1}{K}\sum_j U_j^{\otimes t}\otimes(U_j^{\otimes t})^\dagger = \int \mathrm{d}U U^{\otimes t}\otimes(U_j^{\otimes t})^\dagger$.*
5. *$\frac{1}{K}\sum_j U_j^{\otimes t}\otimes(U_j^{\otimes t})^* = \int \mathrm{d}U U^{\otimes t}\otimes(U_j^{\otimes t})^*$.*
6. *$\Phi_t(\{U_j\}) = \gamma(t,d)$.*

Proof. Note that $\operatorname{tr}\bigl[BU^{\otimes t}A(U^{\otimes t})^\dagger\bigr]$ is a homogeneous polynomial in $\mathrm{Hom}_{(t,t)}(\mathrm{U}(d))$ and that all polynomials of this form for $A, B \in L((\mathbb{C}^d)^{\otimes t})$ span $\mathrm{Hom}_{(t,t)}(\mathrm{U}(d))$. Therefore, statements 1 and 2 are equivalent. The equivalence of statements 2 and 3 is obvious.

The equivalence of statements 1 and 4 follows from the following equation,

$$\operatorname{tr}\bigl\{V(B\otimes A)[U^{\otimes t}\otimes(U^{\otimes t})^\dagger]\bigr\} = \operatorname{tr}\bigl\{BU^{\otimes t}A(U^{\otimes t})^\dagger\bigr\}, \tag{14}$$

where V is the swap operator of parties $1, 2, \ldots, t$ with the parties $t+1, t+2, \ldots, 2t$. The equation in statement 5 is a partial transposition of the one in statement 4.

The equivalence of statements 5 and 6 follows from the following equation

$$\left\|\frac{1}{K}\sum_j U_j^{\otimes t}\otimes(U_j^{\otimes t})^* - \int \mathrm{d}U U^{\otimes t}\otimes(U^{\otimes t})^*\right\|_2 = \Phi_t(\{U_j\}) - \gamma(t,d). \tag{15}$$

□

Most known examples of unitary designs are constructed from subgroups of the unitary group, which are referred to as (unitary) group designs henceforth. Given a finite group G of unitary operators, the frame potential of G takes on the form

$$\Phi_t(G) = \frac{1}{|G|}\sum_{U\in G}|\operatorname{tr}(U)|^{2t}. \tag{16}$$

Let $\overline{G}$ be the quotient of G over the phase factors. Then

$$\Phi_t(G) = \Phi_t(\overline{G}) = \frac{1}{|\overline{G}|}\sum_{U\in\overline{G}}|\operatorname{tr}(U)|^{2t}. \tag{17}$$

This formula is applicable whenever $\overline{G}$ is a finite group even if G is not. Note that $\Phi_t(G)$ is equal to the sum of squared multiplicities of irreducible components of $\tau^t(G) := \{U^{\otimes t}|U\in G\}$ [31], which coincides with the dimension of the commutant of $\tau^t(G)$. Recall that the commutant $\mathcal{A}'$ of a set of operatators $\mathcal{A}$ is the algebra of all operators commuting with every element of $\mathcal{A}$:

$$\mathcal{A}' = \{B|[A,B] = 0\ \forall A\in\mathcal{A}\}. \tag{18}$$

Let H be a subgroup in G. It is clear that every irreducible representation of $\tau^t(G)$ on $\left(\mathbb{C}^d\right)^{\otimes t}$ is also invariant under $\tau^t(H)$ and thus forms a representation space of H. However, these spaces need not be irreducible under the action of H. As a consequence, $\Phi_t(H)\le\Phi_t(G)$ for any subgroup H in G, and the equality is saturated iff every irreducible component of $\tau^t(G)$ is also irreducible when restricted to $\tau^t(H)$; that is, $\tau^t(G)$ and $\tau^t(H)$ decompose into the same number of irreducible components.

At this point, it is instructive to review the representation theory of $\mathrm{U}(d)$ on the space of all tensors $(\mathbb{C}^d)^{\otimes t}$ from the point of view of *Schur-Weyl duality*. By definition the unitary group $\mathrm{U}(d)$ acts on $\mathbb{C}^d$. The action extends to the *diagonal action* on $(\mathbb{C}^d)^{\otimes t}$,

$$U\mapsto\tau^t(U) : |\psi_1\rangle\otimes|\psi_2\rangle\otimes\cdots\otimes|\psi_t\rangle\mapsto U|\psi_1\rangle\otimes U|\psi_2\rangle\otimes\cdots\otimes U|\psi_t\rangle\quad\forall|\psi_j\rangle\in\mathbb{C}^d,\ U\in\mathrm{U}(d). \tag{19}$$

Meanwhile, the symmetric group S_t acts on the tensor product space $(\mathbb{C}^d)^{\otimes t}$ by permuting tensor factors:

$$\pi(|\psi_1\rangle \otimes |\psi_2\rangle \otimes \cdots \otimes |\psi_t\rangle) = |\psi_{\pi_1}\rangle \otimes |\psi_{\pi_2}\rangle \otimes \cdots \otimes |\psi_{\pi_t}\rangle \qquad \forall |\psi_j\rangle \in \mathbb{C}^d,\ \pi \in S_t. \tag{20}$$

The diagonal action of $\mathrm{U}(d)$ and the permutation action of S_t on $(\mathbb{C}^d)^{\otimes t}$ commute with each other. Schurl-Weyl duality states that $(\mathbb{C}^d)^{\otimes t}$ decomposes into multiplicity-free irreducible representations of $\mathrm{U}(d) \times S_t$ [36]. More precisely,

$$\left(\mathbb{C}^d\right)^{\otimes t} = \bigoplus_\lambda H_\lambda = \bigoplus_\lambda W_\lambda \otimes S_\lambda. \tag{21}$$

Here the λ's are non-increasing partitions of t into no more than d parts, W_λ is the *Weyl module* carrying the irrep of $\mathrm{U}(d)$ associated with λ, and S_λ the *Schur module* on which S_t acts irreducibly. We denote the dimensions of S_λ and W_λ by d_λ and D_λ, respectively. Note that d_λ equals the multiplicity of the Weyl module W_λ, and, likewise, D_λ is the multiplicity of the Schur module S_λ. As an implication, the commutant of the diagonal action of the unitary group is generated by all permutations of the tensor factors. When $\lambda = [t]$ is the trivial partition, then $W_\lambda = \mathrm{Sym}_t(\mathbb{C}^d)$ and S_t acts trivially on $S_\lambda \simeq \mathbb{C}$. In particular, it follows that the space $\mathrm{Sym}_t(\mathbb{C}^d)$ carries an irreducible representation of $\mathrm{U}(d)$.

The discussion above leads to a number of equivalent characterizations of t-designs constructed from groups.

Proposition 3. *The following statements concerning $G \leq \mathrm{U}(d)$ are equivalent:*

1. *G is a unitary t-design.*
2. *$\Phi_t(G) = \gamma(t, d)$.*
3. *$\tau^t(G)$ decomposes into the same number of irreps as $\tau^t(\mathrm{U}(d))$.*
4. *Every irreducible component in $\tau^t(\mathrm{U}(d))$ is still irreducible when restricted to $\tau^t(G)$.*
5. *$\tau^t(G)$ and $\tau^t(\mathrm{U}(d))$ has the same commutant.*
6. *The commutant of $\tau^t(G)$ is generated by all the permutations of tensor factors.*

For example, G is a 1-design iff it is irreducible, in that case, $\overline{G}$ has at least d^2 elements, and the lower bound is saturated iff it defines a nice error basis, that is, $\mathrm{tr}(U_j U_k) = d\delta_{jk}$ for $U_j, U_k \in \overline{G}$ [37]. The group G is a unitary 2-design iff $\tau^t 2(G)$ has only two irreducible components, which correspond to the symmetric and antisymmetric subspaces of the bipartite Hilbert space. Prominent examples of unitary group 2-designs include Clifford groups and restricted Clifford groups in prime power dimensions [31, 38–41].

Proposition 4. *Any orbit of pure states of a unitary group t-design forms a complex projective t-design.*

Proof. Let G be a unitary group t-design, then $\tau^t(G)$ acts irreducibly on $\mathrm{Sym}_t(\mathbb{C}^d)$. Therefore,

$$\sum_{U \in \overline{G}} \left(U|\psi\rangle\langle\psi|U^\dagger\right)^{\otimes t} = \sum_{U \in \overline{G}} U^{\otimes t}(|\psi\rangle\langle\psi|)^{\otimes t}(U^{\otimes t})^\dagger \propto P_t^{\mathrm{Sym}} \tag{22}$$

for any pure state $|\psi\rangle$. It follows that any orbit of pure states of G forms a complex projective t-design. □

III. DECOMPOSITION OF THE FOURTH TENSOR POWER OF THE CLIFFORD GROUP

A. Pauli group and Clifford group

Let $\mathbb{F}_2 = \mathbb{Z}_2 = \{0, 1\}$ be the finite field of integers with arithmetic modulo 2. We label the *Pauli matrices* on a single qubit by elements of $\mathbb{F}_2^2$ in the following way:

$$\sigma_{(0,0)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \sigma_{(0,1)} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_{(1,0)} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \sigma_{(1,1)} = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}.$$

A *Pauli operator* on n qubits is defined as the tensor product of n Pauli matrices. Concretely, each $a \in \mathbb{F}_2^{2n}$ defines a Pauli operator as follows,

$$W_a := \sigma_{(a_1,a_2)} \otimes \cdots \otimes \sigma_{(a_{2n-1},a_{2n})}.$$

Every pair of Pauli operators either commute or anticommute,

$$W_a W_b = (-1)^{\langle a,b\rangle} W_b W_a, \tag{23}$$

where $\langle a, b\rangle = a^{\mathrm{T}} J b$ is the symplectic product with J being the $2n \times 2n$ block diagonal matrix over $\mathbb{F}_2$ with n blocks of $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ on the diagonal. Let

$$\bar{\mathcal{P}}_n = \{W_a \,|\, a \in \mathbb{F}_2^{2n}\}$$

be the set of all n qubit Pauli operators. The Pauli group on n-qubits is the group generated by all the Pauli operators in $\bar{\mathcal{P}}_n$,

$$\mathcal{P}_n = \langle \bar{\mathcal{P}}_n \rangle = \{\mathrm{i}^j W_a \,|\, a \in \mathbb{F}_2^{2n}, j \in \mathbb{Z}_4\}.$$

In the following discussion $\bar{\mathcal{P}}_n$ is also identified as the projective Pauli group, the quotient group of $\mathcal{P}_n$ with respect to the phase factors. As a group, $\bar{\mathcal{P}}_n$ is isomorphic to $\mathbb{F}_2^{2n}$.

Let $\mathrm{Sp}(2n, \mathbb{F}_2)$ be the symplectic group composed of all $2n \times 2n$ matrices over $\mathbb{F}_2$ that satisfy the following equation

$$FJF^{\mathrm{T}} = J. \tag{24}$$

The n-qubit Clifford group C_n is the normalizer of the n-qubit Pauli group $\mathcal{P}_n$. For every Clifford unitary $U \in \mathrm{C}_n$, there is a unique symplectic matrix $F \in \mathrm{Sp}(2n, \mathbb{F}_2)$ such that

$$UW_aU^\dagger = (-1)^{f(a)} W_{Fa} \qquad \forall a, \tag{25}$$

for a suitable sign function f from $\mathbb{F}_2^{2n}$ to $\mathbb{F}_2$. Conversely, for each $F \in \mathrm{Sp}(2n, \mathbb{F}_2)$ there exists a Clifford unitary and a suitable function f such that the above equation is satisfied. Let $\overline{\mathrm{C}}_n$ be the projective Clifford group. Then $\overline{\mathrm{C}}_n/\bar{\mathcal{P}}_n$ is isomorphic to $\mathrm{Sp}(2n, \mathbb{F}_2)$.

The Clifford group plays an important role in quantum computation [14, 15, 42, 43], quantum error correction [14, 15], and randomized benchmarking [44–46]. Many nice properties of the Clifford group are closely related to the fact that the group is a unitary 2-design [8, 31–33, 38–41, 47]. Recently, it is shown that the Clifford group C_n forms a unitary 3-design but not a 4-design [11–13]. The fourth frame potential of C_n reads [12],

$$\Phi_4(\mathrm{C}_n) = \begin{cases} 15 & n = 1, \\ 29 & n = 2, \\ 30 & n \geq 3. \end{cases} \tag{26}$$

Comparison with Eq. (13) shows that the frame potential of the Clifford group is quite close to that of a 4-design. This observation indicates that the fourth tensor power of the Clifford group has only a few more irreducible components than that of the whole unitary group, as spelled out more precisely in the next section.

B. A special stabilizer code

To state our main result precisely, we need to define a certain stabilizer code. Recall that a stabilizer group is an abelian subgroup of the Pauli group that does not contain -1. The order of any n-qubit stabilizer group is a divisor of $d = 2^n$. Those n-qubit stabilizer groups of order d are called maximal. A stabilizer code is the common eigenspace of operators in a stabilizer group [14, 48]. If the stabilizer group has order 2^m with $m \leq n$, then the stabilizer code has dimension 2^{n-m}. When the stabilizer group is maximal, the stabilizer code has dimension 1 and reduces to a stabilizer state.

Whenever k is even, the following set of Pauli operators

$$S_{n,k} = \{\tau^k(W_a) \,|\, a \in \mathbb{F}_2^{2n}\}$$

commute with each other. The set is also invariant under the diagonal action of the Clifford group. If in addition k is a multiple of 4, then $S_{n,k}$ is also closed under multiplication and thus forms a stabilizer group. Let $V_{n,k}$ be the stabilizer code defined by the joint $+1$ eigenspace of operators in $S_{n,k}$. The dimension of the stabilizer code is d^{k-2}, and the projector onto it is given by

$$P_{n,k} = \frac{1}{|S_{n,k}|} \sum_{a \in \mathbb{F}_2^{2n}} \tau^k(W_a). \tag{27}$$

The stabilizer code $V_{n,k}$ and projector $P_{n,k}$ are invariant under the action of the symmetric group S_k, which acts on $\left(\mathbb{C}^d\right)^{\otimes k}$ by permuting the k tensor factors. Meanwhile, they are also invariant under the diagonal action of the Clifford group. In other words, $V_{n,k}$ affords a representation of the Clifford group C_n and also $\overline{\mathrm{C}}_n$. Given that $V_{n,k}$ is a common $+1$ eigenspace of $\tau^k(W_a)$ for all Pauli operators W_a, it follows that the Pauli group $\overline{\mathcal{P}}_n$ acts trivially on $V_{n,k}$. Therefore, $V_{n,k}$ affords a projective representation of the symplectic group $\mathrm{Sp}(2n, \mathbb{F}_2)$, which is isomorphic to $\overline{\mathrm{C}}_n/\overline{\mathcal{P}}_n$. When $n \neq 2, 3$, $\mathrm{Sp}(2n, \mathbb{F}_2)$ has trivial Schur cover [49], so the projective representation can be turned into ordinary representation with a suitable choice of phase factors.

In the rest of this section, we construct an orthonormal basis for $V_{n,k}$, though this is essential to understand our main result. First consider the special case $n = 1$. Let $u \in \mathbb{F}_2^k$ and define $\tilde{u} := a + (1, 1, \ldots, 1)$ as the bitwise "NOT" of u. If k is a multiple of 4 and a has even number of digits equal to 1, then the state $|\phi_u\rangle := (|u\rangle + |\tilde{u}\rangle)/\sqrt{2}$ is a common $+1$ eigenstate of $\tau^k(W_a)$ for all single qubit Pauli operators W_a; that is, $|\phi_u\rangle \in V_{1,k}$. Now it is straightforward to verify that the follow set

$$\{|\phi_u\rangle | u \in \mathbb{F}_1^k, \sum_{j=1}^{k} u_j = 0, u_1 = 0\} \tag{28}$$

forms an orthonormal basis of $V_{1,k}$.

Simple analysis shows that $V_{n,k}$ and $P_{n,k}$ can be written as tensor products as follows,

$$V_{n,k} = V_{1,k}^{\otimes n}, \quad P_{n,k} = P_{1,k}^{\otimes n}. \tag{29}$$

So an orthonormal basis of $V_{n,k}$ can be constructed by taking tensor product of the basis of $V_{1,k}$.

C. Main results

The most concise way to state our main result is in terms of the *commutant* of $\tau^4(\mathrm{C}_n)$. The classic *Schur-Weyl duality* states that the commutant of $\tau^k(\mathrm{U}(d))$ is generated by the symmetric group S_k with permutation action [36]. If $d = 2^n$ and we restrict from $\tau^4(\mathrm{U}(d))$ to the subgroup $\tau^4(\mathrm{C}_n)$, the commutant becomes larger. Our main result says that there is only one additional generator: the stabilizer projector $P_{n,4}$ introduced above.

Theorem 1 (Main Theorem)**.** *The commutant $\tau^4(\mathrm{C}_n)'$ of the diagonal action of the Clifford group on $\left(\mathbb{C}^d\right)^{\otimes 4}$ is generated as an algebra by S_4 (permuting tensor factors) and the stabilizer projection $P_{n,4}$.*

Next, we will give a more concrete formulation of the main result. To this end, recall that Schur-Weyl duality can be used to find the decompositoin

$$\left(\mathbb{C}^d\right)^{\otimes 4} = \bigoplus_{\lambda} H_\lambda = \bigoplus_{\lambda} W_\lambda \otimes S_\lambda \tag{30}$$

of $\left(\mathbb{C}^d\right)^{\otimes 4}$ into irreps of $\mathrm{U}(d) \times S_4$. Here, the λ's are partitions of 4 into no more than d parts, W_λ is the Weyl module carrying an irrep of $\mathrm{U}(d)$ and S_λ the Schur module on which S_4 acts irreducibly; the group $\mathrm{U}(d) \times S_4$ acts irreducibly on each H_λ. The dimensions of S_λ and W_λ are denoted by d_λ and D_λ, respectively, as listed in Table I. Note that d_λ equals the multiplicity of the Weyl module W_λ, and, likewise, D_λ is the multiplicity of the Schur module S_λ. Let G be a subgroup of $\mathrm{U}(d)$, then the number of irreducible components of $G \times S_4$ on H_λ is equal to the number of irreducible components of G on W_λ. In particular, $G \times S_4$ is irreducible on H_λ iff G is irreducible on W_λ. The multiplicity of each irrep of G appearing in H_λ is always a multiple of d_λ.

Now recall that $V_{n,4}$ is the stabilizer code defined above. We denote its orthgonal complement by $V_{n,4}^{\perp}$ and define the spaces

$$H_\lambda^+ := H_\lambda \cap V_{n,4}, \qquad H_\lambda^- := H_\lambda \cap V_{n,4}^{\perp}.$$

Because $V_{n,4}$ is invariant under the action of S_4, and because the S_λ are irreducible under the same action, it follows that for each λ, there is a subspace $W_\lambda^+ \subset W_\lambda$ such that

$$H_\lambda^+ = W_\lambda^+ \otimes S_\lambda.$$

Likewise,

$$H_\lambda^- = W_\lambda^- \otimes S_\lambda,$$

where W_λ^- is the ortho-complement, within W_λ, of W_λ^+.

The dimensions of these spaces can be computed explicitly.

TABLE I. Dimensions of the Schur modules, Weyl modules and irreducible components of the n-qubit Clifford group appearing on $(\mathbb{C}^d)^{\otimes 4}$, where $d = 2^n$.

λ	d_λ	D_λ	D_λ^+	D_λ^-
$[4]$	1	$\frac{d(d+1)(d+2)(d+3)}{24}$	$\frac{(d+1)(d+2)}{6}$	$\frac{(d-1)(d+1)(d+2)(d+4)}{24}$
$[1,1,1,1]$	1	$\frac{d(d-1)(d-2)(d-3)}{24}$	$\frac{(d+1)(d+2)}{6}$	$\frac{(d+1)(d-1)(d-2)(d-4)}{24}$
$[2,2]$	2	$\frac{d^2(d^2-1)}{12}$	$\frac{(d^2-1)}{3}$	$\frac{(d^2-4)(d^2-1)}{12}$
$[2,1,1]$	3	$\frac{d(d-2)(d^2-1)}{8}$	0	$\frac{d(d-2)(d^2-1)}{8}$
$[3,1]$	3	$\frac{d(d+2)(d^2-1)}{8}$	0	$\frac{d(d+2)(d^2-1)}{8}$

Lemma 1. *Let $D_\lambda^\pm = \dim W_\lambda^\pm$. The values of $D_\lambda^\pm$ for partitions λ of 4 are given in Table I. In addition, $\dim H_\lambda^\pm = d_\lambda D_\lambda^\pm$.*

Then the main theorem can be expressed equivalently in each of the following two ways.

Corollary 1. *Whenever they are non-trivial, the spaces $W_\lambda^\pm$ carry irreducible representations of the n-qubit Clifford group C_n, while $H_\lambda^\pm$ carry irreducible representations of $\mathrm{C}_n \times S_4$.*

Corollary 2. *Under the action of $\mathrm{C}_n \times S_4$, the space $\left(\mathbb{C}^d\right)^{\otimes 4}$ decomposes into these irreps:*

$$\left(\mathbb{C}^d\right)^{\otimes 4} = \bigoplus_{\lambda; s=\pm \,|\, D_\lambda^s \neq 0} W_\lambda^s \otimes S_\lambda.$$

D. Proofs

In this section, we prove Lemma 1 and derive from it our main result Theorem 1.

Proof of Lemma 1. Let $H_\lambda, W_\lambda, S_\lambda$ be the representation spaces appearing in the Schur-Weyl decomposition in Eq. (30). Let P_λ be the projector onto H_λ. We have

$$P_\lambda = \frac{d_\lambda}{24} \sum_{\sigma \in S_4} \chi_\lambda(\sigma) U_\sigma, \tag{31}$$

where U_σ is the unitary operator that realizes the permutation of the tensor factors corresponding to σ, and χ_λ is the character of the irrep of S_4 corresponding to the partition λ; see Table II. For example, the projectors onto the symmetric and antisymmetric subspaces are respectively given by

$$P_{[4]} = \frac{1}{24} \sum_{\sigma \in S_4} U_\sigma, \tag{32}$$

$$P_{[1^4]} = \frac{1}{24} \sum_{\sigma \in S_4} \operatorname{sgn}(\sigma) U_\sigma, \tag{33}$$

where $\operatorname{sgn}(\sigma)$ is equal to 1 for even permutations and -1 for odd permutations.

Note that P_λ commutes with the projector $P_{n,4}$ onto the stabilizer code, so the dimension of $V_{n,4} \cap H_\lambda$ is given by $d_\lambda D_\lambda^+ = \operatorname{tr}(P_{n,4} P_\lambda)$. Therefore,

$$\begin{aligned} D_\lambda^+ &= \frac{1}{d_\lambda} \operatorname{tr}(P_{n,4} P_\lambda) = \frac{1}{d^2 d_\lambda} \sum_a \operatorname{tr}(W_a^{\otimes 4} P_\lambda) \\ &= \frac{1}{d^2} \Big[D_\lambda + \frac{1}{24} \sum_{\sigma \in S_4} \sum_{0 \neq a \in \mathbb{F}_2^{2n}} \chi_\lambda(\sigma) \operatorname{tr}\big(U_\sigma W_a^{\otimes 4}\big) \Big]. \end{aligned} \tag{34}$$

Let $l(\sigma)$ be the number of cycles in σ with even lengths. If $a \neq 0$, then

$$\operatorname{tr}(U_\sigma W_a^{\otimes 4}) = \begin{cases} 0 & \sigma \text{ contains a cycle of odd length,} \\ d^{l(\sigma)} & \text{otherwise.} \end{cases} \tag{35}$$

TABLE II. Characters of the symmetric group S_4.

cycle type	(1^4)	(2^2)	$(2,1^2)$	$(3,1)$	(4)
order	1	2	2	3	4
#	1	3	6	8	6
$\chi_1=[4]$	1	1	1	1	1
$\chi_2=[1,1,1,1]$	1	1	-1	1	-1
$\chi_3=[2,2]$	2	2	0	-1	0
$\chi_4=[2,1,1]$	3	-1	-1	0	1
$\chi_5=[3,1]$	3	-1	1	0	-1

According to Table II, the symmetric group S_4 has 3 permutations of cycle type (2^2) and six permutations of cycle type (4), while all other permutations contain at least one cycle of odd length. In conjunction with the above two equations, this observation enables us to compute D_λ^+ and then D_λ^-, with the result shown in Table I. □

An alternative proof – which is slightly longer, but may give a better feeling for the spaces involved – is presented in Appendix C.

Proof of Theorem 1. Note that Theorem 1 and Corollary 1 are equivalent. To prove Theorem 1, it suffices to prove Corollary 1, which states that the spaces $W_\lambda^\pm$ carry irreducible representations of the n-qubit Clifford group C_n whenever $W_\lambda^\pm$ are non-trivial.

The sum of squared multiplicities of irreducible components in $\tau^4(\mathrm{C}_n)$ is equal to the fourth frame potential of the Clifford group C_n, as shown in Eq. (26). When $n \geq 3$, both $V_{n,4} \cap H_\lambda$ and $V_{n,4}^\perp \cap H_\lambda$ are nontrivial invariant subspaces of $\mathrm{C}_n \times S_4$ for $\lambda = [4], [1,1,1,1], [2,2]$. So the frame potential of C_n is at least

$$\Phi_4(\mathrm{C}_n) \geq d_{[4]}^2 + d_{[1,1,1,1]}^2 + d_{[2,2]}^2 + \sum_\lambda d_\lambda^2 = 30. \tag{36}$$

The lower bound is saturated iff all the representations of C_n afforded by $W_\lambda^\pm$ for $D_\lambda^\pm \neq 0$ are irreducible and inequivalent. If any of $W_\lambda^\pm$ is reducible, then $\Phi_4(\mathrm{C}_n)$ would be strictly larger than 30, in contradiction with Eq. (26). This contradiction confirms Corollary 1 in the case $n \geq 3$, from which Theorem 1 follows. The proofs for the special cases $n = 1, 2$ are similar. □

IV. t-DESIGNS FROM CLIFFORD ORBITS

In this section we determine all Clifford covariant t-designs in the case of a single qubit. We then show that random orbits of the Clifford group are very good approximation of 4-designs. Furthermore, we introduce several simple and efficient methods for constructing exact and approximate fiducial states of 4-design.

A. Clifford covariant t-designs for qubit

Now the t-partite symmetric subspace has dimension $t+1$, so the frame potential of a qubit t-design is equal to $1/(t+1)$. Since the Clifford group is a unitary 3-design, every orbit of the Clifford group forms a complex projective 3-design. The unique shortest orbit is composed of six stabilizer states, which form a complete set of mutually unbiased bases. When represented on the Bloch sphere, the six states form the vertices of the octahedron.

To derive a simple criterion on those orbits that form 4-design, suppose the fiducial state has Bloch vector (x, y, z) with $x^2 + y^2 + z^2 = 1$. Then the fourth frame potential of the Clifford orbit is given by

$$\Phi_4(x,y,z) = \frac{21 - 6(x^4+y^4+z^4) + 5(x^4+y^4+z^4)^2}{96}. \tag{37}$$

The orbit forms a 4-design iff $x^4 + y^4 + z^4 = 3/5$, in which the case $\Phi_4(x, y, z)$ attains the minimum $1/5$. The orbit forms a 5-design under the same condition. One explicit solution is given by

$$x = \sqrt{\frac{5+2\sqrt{10}}{15}}, \quad y = z = \sqrt{\frac{5-\sqrt{10}}{15}}. \tag{38}$$

By contrast, $\Phi_4(x,y,z)$ is maximized when $x^4+y^4+z^4=1$, in which case the Bloch vector corresponds to a stabilizer state.

When the condition $x^4+y^4+z^4=3/5$ is satisfied, the sixth and seventh frame potential satisfy the following equation

$$\begin{aligned}8\Phi_7(x,y,z)-1=4[7\Phi_6(x,y,z)-1]&=\frac{11(1-21x^2+105x^4-105x^6)}{2400}\\&=\frac{11(1-21y^2+105y^4-105y^6)}{2400}=\frac{11(1-21z^2+105z^4-105z^6)}{2400}\\&=\frac{11[3-7(x^6+y^6+z^6)}{480}.\end{aligned}\tag{39}$$

The orbit forms a 6-design iff x^2,y^2,z^2 are distinct roots of the equation $1-21u+105u^2-105u^3=0$, which are given by

$$u_j=\frac{1}{3}\Bigl(1+2\sqrt{\frac{2}{5}}\cos\frac{\theta+2j\pi}{3}\Bigr),\quad \theta=\arctan\frac{3\sqrt{10}}{20},\quad j=1,2,3.\tag{40}$$

Equivalently, the orbit forms a 6-design iff $x^6+y^6+z^6=3/7$ or if $x^2y^2z^2=1/105$ (assuming $x^4+y^4+z^4=3/5$). The same condition also guarantees that the orbit forms a 7-design. There are 48 solutions in total, which compose two Clifford orbits. When represented on the Bloch sphere, the two orbits can be converted to each other by inversion. The two orbits are not unitarily equivalent, but are equivalent under antiunitary transformations. Actually the 48 solutions form one orbit under the action of the extended Clifford group, the group generated by the Clifford group and complex conjugation with respect to the computational basis. Since any qubit 8-design has at least 25 elements according to Eq. (9), no Clifford orbit can form an 8-design.

Calculation shows that a random Clifford orbit is approximately a t-design for t up to 7. The ratio of the average frame potential over the minimum potential is given by

$$(t+1)\mathrm{E}[\Phi_t(x,y,z)]=\begin{cases}1 & t=3,\\ \frac{127}{126} & t=4,\\ \frac{43}{42} & t=5,\\ \frac{1795}{1716} & t=6,\\ \frac{1381}{1287} & t=7.\end{cases}\tag{41}$$

B. Random Clifford orbits are good approximation to 4-designs

In this section we show that random Clifford orbits are very good approximation to projective 4-designs. Recall that $\tau^4(\mathrm{C}_n)$ has two irreducible components $W_{[4]}^{\pm}$ in the totally symmetric space $W_{[4]}=\mathrm{Sym}_4(\mathbb{C}^{2^n})$. According to Table I, the dimensions of $W_{[4]}$ and $W_{[4]}^{\pm}$ are

$$\begin{aligned}D_{[4]}&=\frac{d(d+1)(d+2)(d+3)}{24},\\ D_+:=D_{[4]}^{+}&=\frac{(d+1)(d+2)}{6},\\ D_-:=D_{[4]}^{-}&=\frac{(d-1)(d+1)(d+2)(d+4)}{24}.\end{aligned}\tag{42}$$

The projectors $P_\pm$ onto the two irreps $W_{[4]}^{\pm}$ read

$$P_+=P_{n,4}P_{[4]},\quad P_-=(1-P_{n,4})P_{[4]}.\tag{43}$$

where $P_{n,4}$ is the projector onto the stabilizer code defined in Eq. (27) and $P_{[4]}$ is the projector onto $W_{[4]}$.

As an implication of Corollary 2, we have

Corollary 3. *Let X be the orbit of any normalized vector $\psi\in\mathbb{C}^{2^n}$ under the action of the Clifford group $\overline{\mathrm{C}}_n$. Then*

$$\frac{1}{|X|}\sum_{\phi\in X}\bigl(|\phi\rangle\langle\phi|\bigr)^{\otimes 4}=\alpha_+P_++\alpha_-P_-,$$

where

$$\alpha_+ = \frac{1}{D_+}\operatorname{tr}\bigl[P_+(|\psi\rangle\langle\psi|)^{\otimes 4}\bigr] = \frac{1}{D_+}\operatorname{tr}\bigl[P_{n,4}(|\psi\rangle\langle\psi|)^{\otimes 4}\bigr], \quad D_+\alpha_+ + D_-\alpha_- = 1. \tag{44}$$

The state $|\psi\rangle$ is a fiducial state of a 4-design iff $\alpha_- = \alpha_+ = 1/D_{[4]}$, that is,

$$\beta_+(\psi) := \operatorname{tr}\bigl[P_{n,4}(|\psi\rangle\langle\psi|)^{\otimes 4}\bigr] = \frac{D_+}{D_{[4]}} = \frac{4}{d(d+3)}. \tag{45}$$

The deviation of the Clifford orbit of ψ from 4-design can be characterized by

$$\epsilon(\psi) = \frac{D_{[4]}}{D_+}\beta_+(\psi) - 1, \tag{46}$$

which satisfies $\beta_+(\psi) = D_+[1+\epsilon(\psi)]/D_{[4]}$. Note that $|\epsilon(\psi)|$ is the operator norm of the deviation

$$\frac{D_{[4]}}{|X|}\sum_{\phi\in X}\bigl(|\phi\rangle\langle\phi|\bigr)^{\otimes 4} - P_{[4]}.$$

It also determines the fourth frame potential of the Clifford orbit of ψ as follows,

$$\Phi_4(\mathrm{orb}(\psi)) = \frac{\beta_+(\psi)^2}{D_+} + \frac{\beta_-(\psi)^2}{D_-} = \frac{1 + D_+\epsilon(\psi)^2/D_-}{D_{[4]}}, \tag{47}$$

where $\beta_-(\psi) = 1 - \beta_+(\psi)$.

To determine potential deviation, note that $d^2 P_{n,4} = \sum_a W_a^{\otimes 4}$. So $\beta_+(\psi)$ is proportional to the second frame potential of the orbit of $|\psi\rangle$ under the action of the Pauli group. Define *characteristic function* $\Xi(\psi)$ as the vector composed of the d^2 elements

$$\Xi_a(\psi) = \operatorname{tr}(W_a|\psi\rangle\langle\psi|). \tag{48}$$

Then $\beta_+(\psi) = \|\Xi(\psi)\|_{l_4}^4/d^2$. Since $\{W_a\}$ forms a nice error basis and Hermitian operator basis, we have

$$\|\Xi(\psi)\|_{l_2}^2 = \sum_a \operatorname{tr}\bigl[W_a^{\otimes 2}(|\psi\rangle\langle\psi|)^{\otimes 2}\bigr] = d, \quad \|\Xi(\psi)\|_{l_\infty} = \max_a \operatorname{tr}(W_a|\psi\rangle\langle\psi|) = 1. \tag{49}$$

Consequently,

$$\frac{2d}{d+1} \le \|\Xi(\psi)\|_{l_4}^4 \le d, \tag{50}$$

which implies

$$\frac{2}{d(d+1)} \le \beta_+(\psi) \le \frac{1}{d}, \quad -\frac{d-1}{2(d+1)} \le \epsilon(\psi) \le \frac{d-1}{4}. \tag{51}$$

The upper bound in Eq. (50) follows from the Hölder inequality; it is saturated iff $\Xi(\psi)$ has d entries equal to 1 and all other entries equal to 0; this can happen iff $|\psi\rangle$ is a stabilizer state. The lower bound is saturated iff

$$\Xi_a(\psi) = \frac{1}{\sqrt{d+1}} \quad \forall a \neq 0, \tag{52}$$

in which case the d^2 states $W_a|\psi\rangle$ for $a \in \mathbb{F}_2^{2n}$ form a symmetric informationally complete measurement (SIC), which happens to be a minimal 2-design [18]. According to Godsil and Roy [50], SIC fiducial states of the n-qubit Pauli group can exist only for $n = 1, 3$. As an implication of Eq. (51), the frame potential satisfies

$$\frac{1}{D_{[4]}} \le \Phi_4(\mathrm{orb}(\psi)) \le \frac{1}{D_{[4]}}\left(1 + \frac{d-1}{4(d+4)}\right), \tag{53}$$

where the lower bound is saturated iff the orbit forms a 4-design, and the upper bound is saturated iff the orbit consists of stabilizer states, that is, ψ is a stabilizer state.

In the rest of this section we compute the variance of the deviation parameter $\epsilon(\psi)$ of a random Clifford orbit and thereby show that random Clifford orbits are very good approximation to 4-designs. Suppose ψ is distributed according to the Haar measure. The first and second moments of $\beta_+(\psi)$ are given by

$$\mathrm{E}[\beta_+(\psi)] = \mathrm{tr}(P_{n,4}\mathrm{E}[(|\psi\rangle\langle\psi|)^{\otimes 4}]) = \frac{1}{D_{[4]}}\,\mathrm{tr}(P_{n,4}P_{[4]}) = \frac{4}{d(d+3)}, \tag{54}$$

$$\mathrm{E}[\beta_+(\psi)^2] = \frac{1}{D_{[8]}}\,\mathrm{tr}(P_{n,4}^{\otimes 2}P_{[8]}) = \frac{16(d^2+15d+68)}{d^2(d+3)(d+5)(d+6)(d+7)}, \tag{55}$$

where the last equality was derived in the appendix. The variance reads

$$\mathrm{Var}[\beta_+] = \mathrm{E}[\beta_+^2] - \mathrm{E}[\beta_+]^2 = \frac{96(d-1)}{d^2(d+3)^2(d+5)(d+6)(d+7)}. \tag{56}$$

As a consequence of the above equations,

$$\mathrm{E}[\epsilon(\psi)] = 0, \quad \mathrm{E}[\epsilon(\psi)^2] = \frac{\mathrm{Var}[\beta_+]}{\mathrm{E}[\beta_+]^2} = \frac{6(d-1)}{(d+5)(d+6)(d+7)}. \tag{57}$$

This equation enables us to determine the ratio of the average fourth frame potential over the minimum frame potential (the potential for a 4-design),

$$D_{[4]}\mathrm{E}[\Phi_4] = 1 + \frac{D_+}{D_-}\mathrm{E}[\epsilon(\psi)^2] = 1 + \frac{24}{(d+4)(d+5)(d+6)(d+7)}. \tag{58}$$

Equations (57) and (58) show that random Clifford orbits are very good approximation to 4-designs.

The following lemma is useful for deriving large-deviation bound for the frame potential of random Clifford orbits.

Lemma 2 (Levy)**.** *Let $f: S^{2d-1} \to \mathbb{R}$ be Lipschitz-continuous with Lipschitz constant η, that is,*

$$|f(x) - f(y)| \le \eta\|x-y\|, \tag{59}$$

where $\|x-y\|$ is the Euclidean norm in the surrounding space $\mathbb{R}^{2d}$ of S^{2d-1}. Drawing a point in S^{2d-1} at random with respect to the uniform measure on the sphere yields

$$\mathrm{Prob}\{|f(x) - \mathrm{E}[f]| \ge \epsilon\} \le 2\exp\Big(\frac{-d\epsilon^2}{9\pi^3\eta^2}\Big). \tag{60}$$

Lemma 3. *The function $\psi \to \beta_+(\psi)$ is Lipschitz-continuous with Lipschitz constant $8/d$, that is,*

$$\beta_+(\psi) - \beta_+(\varphi) \le \frac{8}{d}\||\psi\rangle - |\varphi\rangle\|. \tag{61}$$

Question: How much can we improve this lemma.

Lemma 4. *Suppose ψ is drawn randomly according to the Haar measure. Then*

$$\mathrm{Prob}\{|\beta_+(\psi) - \mathrm{E}[\beta_+(\psi)]| \ge \epsilon\} \le 2\exp\Big(-\frac{d^3\epsilon^2}{576\pi^3}\Big). \tag{62}$$

This large deviation bound is not very good. We expect that there is a much better bound, but how can we get better bounds?

C. Fiducial states of exact 4-designs up to five qubits

In this section we propose a method for constructing exact fiducial states of 4-designs of the Clifford group. Exact fiducial states up to five qubits are constructed explicitly.

Recall that an n-qubit state $|\psi\rangle$ is a fiducial state for a 4-design iff $\|\Xi(\psi)\|_{l_4}^4 = 4d/(d+3)$ Suppose $\psi = \psi_1 \otimes \psi_2$ is a tensor product of an n_1-qubit state and an n_2-qubit state with $n_1 + n_2 = n$. Then $\|\Xi(\psi)\|_{l_4}^4 = \|\Xi(\psi_1)\|_{l_4}^4\|\Xi(\psi_2)\|_{l_4}^4$

since $P_{n,4}$ can be written as a tensor product $P_{n,4} = P_{n_1,4} \otimes P_{n_2,4}$. In the case of a single qubit, let $\psi(x, y, z)$ be a fiducial state with Bloch vector (x, y, z), where $x^2 + y^2 + z^2 = 1$; then

$$\|\Xi(\psi)\|_{l_4}^4 = 1 + x^4 + y^4 + z^4. \tag{63}$$

The state generates a 4-design iff $x^4 + y^4 + z^4 = 3/5$ as pointed out in Sec. IV A. Let ψ_{T} be the magic state (also a SIC fiducial) with Bloch vector $(1, 1, 1)/\sqrt{3}$. Then fiducial states of 4-designs for $n = 2, 3, 4$ can be constructed as follows,

$$\begin{cases} \psi_{\mathrm{T}} \otimes \psi(x, y, z), & x^4 + y^4 + z^4 = 5/7, \quad n = 2; \\ \psi_{\mathrm{T}}^{\otimes 2} \otimes \psi(x, y, z), & x^4 + y^4 + z^4 = 7/11, \quad n = 3; \\ \psi_{\mathrm{T}}^{\otimes 3} \otimes \psi(x, y, z), & x^4 + y^4 + z^4 = 8/19, \quad n = 4. \end{cases} \tag{64}$$

There are also many other constructions.

In dimension 8, the set of Hoggar lines forms a SIC that is covariant with respect to the three-qubit Pauli group [26, 51, 52]. One fiducial state of the SIC is given by

$$|\psi_{\mathrm{Hog}}\rangle = \frac{1}{\sqrt{6}}(1 + \mathrm{i}, 0, -1, 1, -\mathrm{i}, -1, 0, 0)^{\mathrm{T}}. \tag{65}$$

Note that $\|\Xi(\psi_{\mathrm{Hog}})\|_{l_4}^4 = 16/9$ attains the minimum over all three-qubit states. This observation enables us to construct fiducial states of 4-designs for $n = 4, 5$,

$$\begin{cases} \psi_{\mathrm{Hog}} \otimes \psi(x, y, z), & x^4 + y^4 + z^4 = 17/19, \quad n = 4; \\ \psi_{\mathrm{Hog}} \otimes \psi_{\mathrm{T}} \otimes \psi(x, y, z), & x^4 + y^4 + z^4 = 8/19, \quad n = 5. \end{cases} \tag{66}$$

The tensor-product construction of fiducial states of 4-designs also has a limitation. Consider tensor products of qubit magic states and ψ_{Hog} for example,

$$\begin{aligned} \|\Xi(\psi_{\mathrm{T}}^{\otimes n})\|_{l_4}^4 &= \left(\frac{4}{3}\right)^n. \\ \|\Xi(\psi_{\mathrm{Hog}}^{\otimes n})\|_{l_4}^4 &= \left(\frac{16}{9}\right)^{n/3} = \left(\frac{4}{3}\right)^{2n/3}, \quad 3|n. \end{aligned} \tag{67}$$

As n increases, $\|\Xi(\psi_{\mathrm{T}}^{\otimes n})\|_{l_4}^4$ and $\|\Xi(\psi_{\mathrm{Hog}}^{\otimes n})\|_{l_4}^4$ increase exponentially with n. By contrast, the value required for a 4-design approaches the constant 4. The following proposition clarifies this limitation; see the appendix for a proof.

Proposition 5. *Suppose a 4-design fiducial state of the n-qubit Clifford group is a tensor product of m states $\psi = \otimes_{j=1}^m \psi_j$, where ψ_j is an n_j-qubit state with $\sum_j n_j = n$ and $n_1 \geq n_2 \geq \cdots \geq n_m$. Then $m \leq 3$ except when $n = 4$. If $m = 3$, then $n_2 = n_3 = 1$, except when $(n_1, n_2, n_3) = (2, 2, 1)$ or $(n_1, n_2, n_3) = (3, 2, 1)$.*

D. Algorithms for constructing fiducial states of 4-designs

Here we present two algorithms for constructing fiducial states of 4-designs. Let $|\psi\rangle$ be an n-qubit state. Recall that $|\psi\rangle$ is a fiducial state of a 4-design iff $\beta_+(\psi) = 4/[d(d+3)]$ or, equivalently, iff $\epsilon(\psi) = 0$; cf. Eqs. (45) and (46). Given two n-qubit states $|\psi_1\rangle, |\psi_2\rangle$, if $\gamma(|\psi_1\rangle) > 0$ and $\gamma(|\psi_2\rangle) < 0$, then any continuous curve of pure states joining $|\psi_1\rangle$ and $|\psi_2\rangle$ contains a 4-design fiducial. The following bisection algorithm is based on this simple observation. Let ϵ_0 be the precision required.

Algorithm 1:

1. Generate two states $|\psi_1\rangle, |\psi_2\rangle$ such that $\epsilon(\psi_1) > 0$, $\epsilon(\psi_2) < 0$, and $\langle\psi_1|\psi_2\rangle \neq 0$. Choose suitable phase factors so that $\langle\psi_1|\psi_2\rangle > 0$.
2. Let $|\psi_3'\rangle = (|\psi_1\rangle + |\psi_2\rangle)/2$ and $|\psi_3\rangle = |\psi_3'\rangle/\sqrt{\langle\psi_3'|\psi_3'\rangle}$. If $|\epsilon(\psi_3)| \leq \epsilon_0$, stop.
3. If $\gamma(\psi_3) \geq 0$, then replace $|\psi_1\rangle$ with $|\psi_3\rangle$; otherwise, replace $|\psi_2\rangle$ with $|\psi_3\rangle$. Repeat Steps 2,3.

Remark 2. $|\psi_1\rangle$ can be chosen to be a stabilizer state, while a potential candidate for $|\psi_2\rangle$ is an eigenstate of a Singer unitary introduced in the next section. In Step 2 we may also use weighted sum of $|\psi_1\rangle, |\psi_2\rangle$, say

$$|\psi_3'\rangle = \frac{|\psi_1\rangle\epsilon(\psi_1) - |\psi_2\rangle\epsilon(\psi_2)}{\epsilon(\psi_1) - \epsilon(\psi_2)}. \tag{68}$$

The second algorithm is based on the tensor product construction discussed in the previous section.

Algorithm 2:

1. Generate an $(n-1)$-qubit state $|\psi_{n-1}\rangle$ such that $\beta_+(\psi_{n-1}) \leq 3/d(d+3)$, where $d = 2^n$.

2. Let $c = 4/[d(d+3)\beta_+(\psi_{n-1})]$. Choose a qubit state $|\psi\rangle$ with Bloch vector (x, y, z) such that $x^4+y^4+z^4 = c-1$. Then $|\psi_{n-1}\rangle \otimes |\psi\rangle$ is a fiducial state of a 4-design.

Remark 3. The qubit state required in Step 2 can always be found. Note that $1/3 \leq c-1 \leq 2(d+2)/(d+3)-1 < 1$ since $\beta_+(|\psi_{n-1}\rangle) \geq 1/2^n(2^{n-1}+1) = 2/d(d+2)$, where the lower bound is saturated iff $|\psi_{n-1}\rangle$ is a SIC fiducial state of the $(n-1)$-qubit Pauli group; cf. Eq. (51).

In general, it is still not clear that there exists an $(n-1)$-qubit state $|\psi_{n-1}\rangle$ such that $\beta_+(\psi_{n-1}) \leq 3/d(d+3)$, but we believe that the answer is positive. Actually, any eigenstate of a Singer unitary might satisfy the requirement; see the next section.

E. Approximate fiducial states of 4-designs from MUB cycler

Let $\{|\psi_j^r\rangle\}_{r,j}$ be a set of mutually unbiased bases, where r labels the basis, and j labels each element in a basis. A *balanced state* $|\psi\rangle$ with respect to $\{|\psi_j^r\rangle\}_{r,j}$ is a state that looks the same from every basis in the set, that is, the set of probabilities $\{|\langle\psi_j^r|\psi\rangle|^2\}_j$ is independent of r. If there exists a unitary operator that cycles through all the bases, then any eigenstate of the unitary operator is a balanced state. For example, the complete set of MUB constructed by Wootters and Fields [29] has a cycler when the dimension is a power of 2, that is $d = 2^n$. Each MUB cycler is a special element in the Clifford group, which is also known as a Singer unitary [53]. The group generated by a Singer unitary is called Singer unitary group. All Singer unitary groups are conjugated to each other in the Clifford group, all of them have the same order of $d+1$ (modular phase factors). In addition, each Singer unitary has a nondegenerate spectrum, so the eigenbasis is well-defined. In the case of a qubit, each Singer unitary has order 3, and each eigenstate of a Singer unitary is a SIC fiducial and a magic state.

When n is a power of 2, a simple construction of Singer unitaries (MUB cyclers) was presented in Ref. [54]. Here we are interested in constructing approximate fiducial states of 4-designs from the eigenstates of a Singer unitary. For $n = 1, 2, 4, 8$, numerical calculation shows that all eigenstates $|\psi_n\rangle$ of a Singer unitary for given n have the same value of $\epsilon(\psi_n)$ [cf. Eq. (46)]. Let $|\psi_\mathrm{T}\rangle$ be a single qubit magic state. Calculation shows that

$$-\epsilon(|\psi_n\rangle \otimes |\psi_\mathrm{T}\rangle) = \begin{cases} \frac{2}{9} & n = 1, \\ 0.12 & n = 2, \\ 0.0312 & n = 4, \\ 0.0020 & n = 8. \end{cases} \tag{69}$$

The magnitude of the deviation $\epsilon(\psi_n \otimes \psi_\mathrm{T})$ is around $1/2^{n+1}$, which has the same order of magnitude as the standard deviation of $\epsilon(\psi)$ for a random $(n+1)$-qubit state $|\psi\rangle$; cf. Eq. (57). The orbit generated from $|\psi_n\rangle \otimes |\psi_\mathrm{T}\rangle$ is a very good approximation to a 4-design. Exact 4-design fiducial state can be constructed using algorithm 2 in the previous section. In addition, $|\psi_n\rangle$ or $|\psi_n\rangle \otimes |\psi_\mathrm{T}\rangle$ can serve as an input to Algorithm 1 presented in the previous section.

Conjecture 1. *Suppose $|\psi_n\rangle$ is any eigenstate of a Singer unitary operator in the n-qubit Clifford group. Then*

$$\lim_{n\to\infty} \epsilon(\psi_n \otimes \psi_\mathrm{T}) = 0. \tag{70}$$

This conjecture implies that the orbit generated by $(n+1)$-qubit Clifford group from $|\psi_n\rangle \otimes |\psi_\mathrm{T}\rangle$ converges to a 4-design with respect to the operator norm. Equation (70) has several equivalent formulations, one of which reads

$$\lim_{n\to\infty} \|\Xi(\psi_n)\|_{l_4}^4 = 3. \tag{71}$$

V. OPEN PROBLEMS

1. Is there any orbit of the Clifford that forms a t-design for $t > 4$? The answer is positive when $n = 1$. It seems that the same should hold for larger n.

2. What is the maximum t such that there is an orbit of the Clifford group that forms a t-design. The answer is 7 when $n = 1$. How about approximate t-designs?

3. Prove Conjecture 1.

TABLE III. Permutations of S_8 without cycle of odd length. N_1 is the number of permutations of a given cycle type; N_2 is the number of balanced permutations (those in $\mathscr{A}$) of a given cycle type; $N_3 = N_{3+} - N_{3-}$, where $N_{3\pm}$ is the number of permutations of a given cycle type that belong to $\mathscr{A}_\pm$. The sets $\mathscr{A}$ and $\mathscr{A}_\pm$ are defined in the text. Note that $N_{3+} + N_{3-} = N_2$.

cycle type	(2^4)	$(2^2,4)$	(4^2)	$(2,6)$	(8)
N_1	105	1260	1260	3360	5040
N_2	9	252	684	1440	5040
N_3	9	108	108	288	432

Appendix A: Derivation of Eq. (55)

In this appendix, we derive the second moment of $\beta_+(\psi)$, as presented in Eq. (55).

$$\mathrm{E}[\beta_+(\psi)^2] = \frac{1}{D_{[8]}}\operatorname{tr}(P_{n,4}^{\otimes 2}P_{[8]}) = \frac{1}{d^4 D_{[8]}}\sum_{a,b}\operatorname{tr}(P_{[8]}W_a^{\otimes 4}\otimes W_b^{\otimes 4}) = \frac{16(d^2+15d+68)}{d^2(d+3)(d+5)(d+6)(d+7)}, \tag{A1}$$

where W_a are n-qubit Pauli operators, $P_{[k]}$ is the projector onto k-partite symmetric subspace of $(\mathbb{C}^d)^{\otimes k}$ with $d = 2^n$, and $D_{[k]}$ is the rank of $P_{[k]}$ or the dimension of the k-partite symmetric subspace. In deriving the las equality in Eq. (A1), we have made use of the following formula

$$\operatorname{tr}(P_{[8]}W_a^{\otimes 4}\otimes W_b^{\otimes 4}) = \begin{cases} D_{[8]} & W_a = W_b = 1, \\ \frac{D_{[8]}}{D_{[4]}}\frac{3d^2+6d}{24} & W_a = 1, W_b \neq 1 \text{ or } W_b = 1, W_a \neq 1, \\ \frac{1}{2688}(7d^4+84d^3+308d^2+336d) & W_a = W_b \neq 1, \\ \frac{1}{4480}(d^4+28d^3+236d^2+560d) & W_a, W_b \neq 1, W_aW_b = W_bW_a, \\ \frac{1}{4480}(d^4+12d^3+44d^2+48d) & W_a, W_b \neq 1, W_aW_b = -W_bW_a. \end{cases} \tag{A2}$$

To derive Eq. (A2), we recall the following facts,

$$P_{[k]} = \frac{1}{k!}\sum_{\sigma\in S_k} U_\sigma, \quad \operatorname{tr}_k P_{[k]} = \frac{D_{[k]}}{D_{[k-1]}}P_{[k-1]}, \tag{A3}$$

where tr_k means the partial trace over party k. If $a \neq 0$, then

$$\operatorname{tr}(U_\sigma W_a^{\otimes k}) = \begin{cases} 0 & \sigma \text{ contains an cycle of odd length}, \\ d^{l(\sigma)} & \text{otherwise}. \end{cases} \tag{A4}$$

where $l(\sigma)$ is the number of cycles in σ with even lengths. The cycle types of elements in S_8 without cycle of odd length are listed in Table III.

The first case in Eq. (A2) is trivial. When $W_b = 1, W_a \neq 1$,

$$\operatorname{tr}\left(P_{[8]}W_a^{\otimes 4}\otimes W_b^{\otimes 4}\right) = \frac{D_{[8]}}{D_{[4]}}\operatorname{tr}\left(P_{[4]}W_a^{\otimes 4}\right), \tag{A5}$$

recall that the symmetric group S_4 has three permutations of cycle type 2^2, six permutations of cycle type 4, and all other permutations contain at least one cycle of odd length (cf. II). The case $W_a = 1, W_b \neq 1$ has the same result. When $W_b = W_a \neq 1$, the result follows from Eqs. (A3), (A4), and Table III.

To settle the last two cases in Eq. (A2), we need to introduce some terminology. A permutation in S_8 is balanced if each cycle involves even number of parties both in the first four parties and in the second four parties. Define $\mathscr{A}$ as the subset of balanced permutations in S_8. Each permutation in S_8 induces a permutation on the vector $v = (a, a, a, a, b, b, b, b)$. Define

$$\mathscr{A}_+ = \{\sigma \in \mathscr{A} |\ \sigma \text{ induces even number of transpositions between } a \text{ and } b\}. \tag{A6}$$

$$\mathscr{A}_- = \{\sigma \in \mathscr{A} |\ \sigma \text{ induces odd number of transpositions between } a \text{ and } b\}. \tag{A7}$$

Note that $\mathscr{A} = \mathscr{A}_+ \cup \mathscr{A}_-$.

If $W_b, W_a \neq 1$, $W_b \neq W_a$, and $W_b W_a = W_a W_b$, then

$$\operatorname{tr}(U_\sigma W_a^{\otimes 4} \otimes W_b^{\otimes 4}) = \begin{cases} d^{l(\sigma)} & \sigma \in \mathscr{A}, \\ 0 & \sigma \notin \mathscr{A}. \end{cases} \tag{A8}$$

If $W_b W_a = -W_a W_b$, then

$$\operatorname{tr}(U_\sigma W_a^{\otimes 4} \otimes W_b^{\otimes 4}) = \begin{cases} d^{l(\sigma)} & \sigma \in \mathscr{A}+, \\ -d^{l(\sigma)} & \sigma \in \mathscr{A}_-, \\ 0 & \text{otherwise}. \end{cases} \tag{A9}$$

Now the last two cases in Eq. (A2) can be determined by virtue of Table III and the above two equations.

Appendix B: Proof of Proposition 5

Proof. Let $d_j = 2^{n_j}$ and suppose $m = 4$. Then

$$\|\Xi(\psi)\|_{l_4}^4 = \prod_{j=1}^{4} \|\Xi(\psi_j)\|_{l_4}^4 \geq \prod_{j=1}^{4} \frac{2d_j}{d_j+1} \geq \left(\frac{4}{3}\right)^3 \frac{2^{n-2}}{2^{n-3}+1}. \tag{B1}$$

If $n \geq 5$, then

$$\left(\frac{4}{3}\right)^3 \frac{2^{n-2}}{2^{n-3}+1} - \frac{2^{n+2}}{2^n+3} = \frac{5 \times 2^{n+2}(2^n - 24)}{27(2^n+3)(2^n+8)} > 0. \tag{B2}$$

So the state ψ cannot generate a 4-design.

Now suppose $m = 3$, so that $n_1 + n_2 + n_3 = n$. If $n_3 = 2$, then

$$\|\Xi(\psi)\|_{l_4}^4 \geq \left(\frac{8}{5}\right)^3 = \frac{512}{125} \geq 4 > \frac{4d}{d+3}, \tag{B3}$$

which leads to a contradiction. If $n_3 = 1, n_1, n_2 \geq 3$, then $n \geq 7$,

$$\|\Xi(\psi)\|_{l_4}^4 - \frac{2^{n+2}}{2^n+3} \geq \prod_{j=1}^{4} \frac{2d_j}{d_j+1} - \frac{2^{n+2}}{2^n+3} \geq \frac{4}{3}\frac{16}{9}\frac{2^{n-3}}{2^{n-4}+1} - \frac{2^{n+2}}{2^n+3} = \frac{2^{n+2}(5 \times 2^n - 336)}{27(2^n+3)(2^n+16)} > 0. \tag{B4}$$

So ψ cannot be a 4-design fiducial. If $n_3 = 1, n_1, n_2 \geq 2$, then $n \geq 5$,

$$\|\Xi(\psi)\|_{l_4}^4 - \frac{2^{n+2}}{2^n+3} \geq \prod_{j=1}^{4} \frac{2d_j}{d_j+1} - \frac{2^{n+2}}{2^n+3} \geq \frac{4}{3}\frac{8}{5}\frac{2^{n-2}}{2^{n-3}+1} - \frac{2^{n+2}}{2^n+3} = \frac{2^{n+2}(2^n - 72)}{15(2^n+3)(2^n+8)}. \tag{B5}$$

If in addition $n \geq 7$, then $\|\Xi(\psi)\|_{l_4}^4 \geq 2^{n+2}/(2^n+3)$, so that ψ cannot be a 4-design fiducial. This observation completes the proof of the proposition. □

Appendix C: Alternative proof of Lemma 1

When $n = 1$, the following four states

$$\begin{aligned} |\phi_0\rangle &:= |0000\rangle + |1111\rangle, \\ |\phi_1\rangle &:= |1001\rangle + |0110\rangle, \\ |\phi_2\rangle &:= |0101\rangle + |1010\rangle, \\ |\phi_3\rangle &:= |0011\rangle + |1100\rangle. \end{aligned} \tag{C1}$$

form an orthonormal basis of $V_{n,4}$. The symmetric group S_4 (permuting the four tensor factors) fixes $|\phi_0\rangle$ and acts like S_3 on the $|\phi_1\rangle, |\phi_1\rangle, |\phi_2\rangle$. For general n, we have that

$$V_{n,4} = V_{1,4}^{\otimes n}.$$

One orthonormal basis of $V_{n,4}$ is composed of the following 4^n states

$$|\phi_{i_1 i_2,\dots,i_n}\rangle = |\phi_{i_1}\rangle \otimes \cdots \otimes |\phi_{i_n}\rangle, \quad i_1, i_2, \dots, i_n \in \{0, 1, 3, 4\}. \tag{C2}$$

Each state is labeled by a length-n word $i_1, \dots, i_n$ with $i_j \in \{0, 1, 3, 4\}$. Each permutation in the symmetric group S_4 induces a permutation on the basis states and a corresponding permutation on the words, which acts on all letters simultaneously. We get these orbits:

1. One orbit containing $0^{\times n}$, referred to as type I orbit below.
2. Any string in $\{0, i\}^{\times n}$ (for given $i \in \{1, 2, 3\}$) excluding $0^{\times n}$ generates an orbit of size 3 . There are $2^n - 1$ such orbits of length three, referred to as type II orbits below..
3. We have accounted for $3 \times (2^n - 1) + 1 = 3 \times 2^n - 2$ strings. The remaining ones have either two or three distinct non-zero letters. These strings are partitioned into orbits of length 6, referred to as type III orbits below.

The three type of strings are referred to as type I, II, III strings below; the corresponding orbits are referred to with similar names. The total number of orbits is

$$2^n + \frac{4^n - 3 \times 2^n + 2}{6} = \frac{4^n + 3 \times 2^n + 2}{6} = \frac{(2^n + 2)(2^n + 1)}{6} = \frac{(d + 2)(d + 1)}{6}. \tag{C3}$$

Now we are ready to construct an orthonormal basis for the totally $W^+_{[4]} = V_{n,4} \cap \mathrm{Sym}_4(\mathbb{C}^d)$. For each string $s \in \{1, 2, 3\}^n$, let $\mathrm{orb}(s)$ be the orbit of the string under the action of S_4. Then

$$P_{[4]}|\phi_s\rangle = \frac{1}{|\mathrm{orb}(s)|} \sum_{r \in \mathrm{orb}(s)} |\phi_s\rangle \tag{C4}$$

Note that $P_{[4]}|\phi_s\rangle \in W^+_{[4]}$ only depends on $\mathrm{orb}(s)$ and that the states corresponding to different orbits are orthogonal. Let $\mathscr{S}$ be a subset of $\{0, 1, 3, 4\}^n$ that contains exactly one string from each orbit. Then

$$\{\sqrt{|\mathrm{orb}(s)|} P_{[4]}|\phi_s\rangle | s \in \mathscr{S}\} \tag{C5}$$

is an orthonormal basis for $W^+_{[4]}$. In particular, the dimension of $W^+_{[4]}$ is equal to the total number of orbits, that is,

$$D^+_{[4]} = \dim(W^+_{[4]}) = \frac{(d + 2)(d + 1)}{6}. \tag{C6}$$

Now consider the subspace $W^+_{[1^4]}$. Note that $P_{[1^4]}|\phi_s\rangle = 0$ when s is an type I or type II string. An orthonormal basis for $W^+_{[1^4]}$ is

$$\{\sqrt{|\mathrm{orb}(s)|} P_{[1^4]}|\phi_s\rangle | s \in \mathscr{S} \text{ is of type III}\}. \tag{C7}$$

The dimension of $W^+_{[1^4]}$ is equal to the number of type III orbits, that is,

$$D^-_{[1^4]} = \dim(W^+_{[1^4]}) = \frac{(d - 2)(d - 1)}{6}. \tag{C8}$$

It is more involved to compute the dimension $W^+_{[2^2]}$, but it is also unnecessary if we can compute the dimension of $W^+_{[2,1,1]}$ and $W^+_{[3,1]}$. With Lemma 5 below one can show that $P_{[2,1,1]}|\phi_s\rangle = 0$ and $P_{[3,1]}|\phi_s\rangle = 0$ for all strings s. So both $W^+_{[2,1,1]}$ and $W^+_{[3,1]}$ have dimension 0. It follows that

$$D^+_{[4]} + D^+_{[1^4]} + 2D^+_{[2^2]} = d^2, \tag{C9}$$

which implies that $D^+_{[2^2]} = (d^2 - 1)/3$.

Lemma 5. *Let H be the unique order-4 normal subgroup of S_4. Then $\sum_{\sigma \in gH} \chi_\lambda(\sigma) = 0$ for $\lambda = [2, 1, 1], [3, 1]$ and all $g \in S_4$.*

[1] E. J. Candes, T. Strohmer, and V. Voroninski, Communications on Pure and Applied Mathematics **66**, 1241 (2013).

[2] E. J. Candès and X. Li, Foundations of Computational Mathematics **14**, 1017 (2014).

[3] E. J. Candes, X. Li, and M. Soltanolkotabi, Applied and Computational Harmonic Analysis **39**, 277 (2015).

[4] D. Gross, F. Krahmer, and R. Kueng, Applied and Computational Harmonic Analysis (2015).

[5] D. Gross, F. Krahmer, and R. Kueng, J. Fourier Anal. Appl. **21**, 229 (2015).

[6] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, "Stable low-rank matrix recovery via null space properties," (2015), arXiv:1507.07184.

[7] W. Matthews, S. Wehner, and A. Winter, Commun. Math. Phys. **291**, 813 (2009).

[8] A. W. Harrow and R. A. Low, Commun. Math. Phys. **291**, 257 (2009).

[9] F. G. S. L. Brandao, A. W. Harrow, and M. Horodecki, "Local random quantum circuits are approximate polynomial-designs," (2015), arXiv:1208.0692.

[10] D. Gross, F. Krahmer, and R. Kueng, Journal of Fourier Analysis and Applications **21**, 229 (2015).

[11] R. Kueng and D. Gross, "Qubit stabilizer states are complex projective 3-designs," (2015), poster at QIP 2013, arXiv:1510.02767.

[12] H. Zhu, "Multiqubit Clifford groups are unitary 3-designs," (2015), arXiv:1510.02619.

[13] Z. Webb, "The Clifford group forms a unitary 3-design," (2015), arXiv:1510.02769.

[14] D. Gottesman, *Stabilizer Codes and Quantum Error Correction*, Ph.D. thesis, California Institute of Technology (1997), available at `http://arxiv.org/abs/quant-ph/9705052`.

[15] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, UK, 2000).

[16] R. Kueng, H. Zhu, and D. Gross, preprint (2016).

[17] R. Kueng, H. Zhu, and D. Gross, preprint (2016).

[18] A. J. Scott, J. Phys. A: Math. Gen. **39**, 13507 (2006).

[19] H. Zhu and B.-G. Englert, Phys. Rev. A **84**, 022327 (2011).

[20] H. Zhu, Phys. Rev. A **90**, 032309 (2014).

[21] A. Ambainis and J. Emerson, in *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)* (2007) pp. 129–140.

[22] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, J. Math. Phys. **45**, 2171 (2004).

[23] D. M. Appleby, C. A. Fuchs, and H. Zhu, Quantum Inf. Comput. **15**, 61 (2015).

[24] S. G. Hoggar, Eur. J. Combinator. **3**, 233 (1982).

[25] V. Levenshtein, Discrete Math. **192**, 251 (1998).

[26] G. Zauner, Int. J. Quant. Inf. **9**, 445 (2011).

[27] A. J. Scott and M. Grassl, J. Math. Phys. **51**, 042203 (2010).

[28] I. D. Ivanović, J. Phys. A: Math. Gen. **14**, 3241 (1981).

[29] W. K. Wootters and B. D. Fields, Ann. Phys. **191**, 363 (1989).

[30] T. Durt, B.-G. Englert, I. Bengtsson, and K. Życzkowski, Int. J. Quant. Inf. **8**, 535 (2010).

[31] D. Gross, K. Audenaert, and J. Eisert, J. Math. Phys. **48**, 052104 (2007).

[32] A. J. Scott, J. Phys. A: Math. Theor. **41**, 055308 (2008).

[33] A. Roy and A. J. Scott, Des. Codes Cryptogr. **53**, 13 (2009).

[34] M. S. Persi Diaconis, Journal of Applied Probability **31**, 49 (1994).

[35] E. M. Rains, Electron. J. Combin. **5**, R12 (1998).

[36] R. Goodman and N. R. Wallach, *Symmetry, Representations, and Invariants*, Graduate Texts in Mathematics, Vol. 255 (Springer, 2009).

[37] A. Klappenecker and M. Rötteler, IEEE Trans. Inf. Theory **48**, 2392 (2002), supplementary information including a catalogue of nice error bases available at `http://www.cs.tamu.edu/faculty/klappi/ueb/ueb.html`.

[38] D. P. DiVincenzo, D. Leung, and B. Terhal, IEEE Trans. Inf. Theory **48**, 580 (2002).

[39] H. F. Chau, IEEE Trans. Inf. Theory **51**, 1451 (2005).

[40] C. Dankert, *Efficient Simulation of Random Quantum States and Operators*, Master thesis, University of Waterloo (2005), available online at `http://arxiv.org/abs/quant-ph/0512217`.

[41] C. Dankert, R. Cleve, J. Emerson, and E. Livine, Phys. Rev. A **80**, 012304 (2009).

[42] D. Gottesman and I. L. Chuang, Nature **402**, 390 (1999).

[43] S. Bravyi and A. Kitaev, Phys. Rev. A **71**, 022316 (2005).

[44] E. Knill, D. Leibfried, R. Reichle, J. Britton, R. B. Blakestad, J. D. Jost, C. Langer, R. Ozeri, S. Seidelin, and D. J. Wineland, Phys. Rev. A **77**, 012307 (2008).

[45] E. Magesan, J. M. Gambetta, and J. Emerson, Phys. Rev. Lett. **106**, 180504 (2011).

[46] J. J. Wallman and S. T. Flammia, New J. Phys. **16**, 103032 (2014).

[47] R. Cleve, D. Leung, L. Liu, and C. Wang, "Near-linear constructions of exact unitary 2-designs," (2015), arXiv:1501.04592.

[48] D. Gross, J. Math. Phys. **47**, 122107 (2006).

[49] R. A. Wilson, *The Finite Simple Groups*, Graduate Texts in Mathematics, Vol. 251 (Springer, London, 2009).

[50] C. Godsil and A. Roy, Eur. J. Combinator. **30**, 246 (2009).

[51] S. G. Hoggar, Geom. Dedicata **69**, 287 (1998).

[52] H. Zhu, Ann. Phys. **362**, 311 (2015).
[53] H. Zhu, "Sharply covariant mutually unbiased bases," (2015), arXiv:1503.00003.
[54] U. Seyfarth and K. S. Ranade, Phys. Rev. A **84**, 042327 (2011).

Low rank matrix recovery from Clifford orbits

Richard Kueng,[1] Huangjun Zhu,[1] and David Gross[1]

[1]*THP Cologne, Germany*

(Dated: September 5, 2016)

We study the reconstruction of hermitian low rank matrices from an undersampled number of measurements via nuclear norm minimization. We consider the particular scenario, where the measurements correspond to rank-one projectors onto orbits of the Clifford group chosen uniformly at random. This includes stabilizer states as a particular special case. Novel results about the higher moments of the multi-qubit Clifford group [1] allow us to establish reconstruction guarantees for $m \geq Cnr\kappa(r)\log(n)$ measurements. The factor $\kappa(r)$ depends on the choice of fiducial. For stabilizer states it amounts to r. This reconstruction is stable towards both additive noise and the model assumption of low rank. If the matrix of interest is in addition positive semidefinite, reconstruction may be performed by a constrained nuclear norm minimization.

Our results in particular imply near-optimal performance guarantees for phase retrieval via PhaseLift.

I. INTRODUCTION AND MAIN RESULTS

A. Phase retrieval and low rank matrix recovery

The problem of retrieving phases has a long history in many different scientific disciplines. Accordingly, the problem's mathematical structure has received considerable attention in its own right. Mathematically, the discrete version of the phase retrieval problem asks to reconstruct an unknown complex vector $\mathbf{x} \in \mathbb{C}^d$ from measurements of the form

$$y_k = |\langle \mathbf{a}_k, \mathbf{x}\rangle|^2 + e_k \quad 1 \leq k \leq d. \tag{1}$$

Here, $\mathbf{a}_, \ldots, \mathbf{a}_m \in \mathbb{C}^d$ model linear measurements, while the absolute values assure that the corresponding observations y_k are ignorant towards complex phases. Finally, the e_k's model additive noise of unknown size and structure.

Importantly, a measurement process of the form (1) is not linear in $\mathbf{x}$. This non-linearity can be overcome by "lifting" the problem to the outer-product $\mathbf{x}\mathbf{x}^*$ of $\mathbf{x}$ [2, 3]:

$$y_k = |\langle \mathbf{a}_k, \mathbf{x}\rangle|^2 + e_k = \mathrm{tr}\left(\mathbf{a}_k\mathbf{a}_k^*\mathbf{x}\mathbf{x}^*\right) + e_k. \tag{2}$$

Known as "*PhaseLift*" [4], such a trick seemingly changes the problem's nature and asks for estimating the positive semidefinite, rank-one matrix $\mathbf{x}\mathbf{x}^*$ from linear measurements. Such a task is a particular instance of *low rank matrix recovery*. Building on ideas from compressed sensing, low rank matrix recovery aims at estimating an unknown $d \times d$ matrix $\mathbf{X} \in M_d$ from $m \ll d^2$ linear measurements of the form

$$y_k = \mathrm{tr}\left(\mathbf{A}_k\mathbf{X}\right) + e_k \quad 1 \leq k \leq m, \tag{3}$$

under the prior assumption that $\mathbf{X}$ is (approximately) low rank. By defining the measurement operator

$$\mathcal{A} : M_d \to \mathbb{R}^m \tag{4}$$
$$\mathbf{Z} \mapsto \sum_{k=1}^m \mathrm{tr}\left(\mathbf{A}_k\mathbf{Z}\right)\mathbf{e}_k,$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_m$ denotes the standard basis in $\mathbb{R}^m$, an entire measurement process can succingtly be written as

$$\mathbf{y} = \mathcal{A}(\mathbf{X}) + \mathbf{e} \tag{5}$$

with $\mathbf{y} = (y_1, \ldots, y_m)^T$ and $\mathbf{e} = (e_1, \ldots, e_m)^T$. Up to date, several measurement ensembles $\mathbf{A}_1, \ldots, \mathbf{A}_m \in M_d$ have been identified [5, 6] for which any rank r matrix can be stably estimated from

$$m = Crd\mathrm{polylog}(d)$$

noisy measurements of the form (3). With a notable exception [7], these results rely on randomly selected measurements. In order to deal with noise corruption, an a priori bound $\eta \geq \|\mathbf{e}\|_{\ell_q}$ on that noise is required. Subsequently, the actual reconstruction is carried out by solving

$$\begin{aligned} \mathbf{Z}^\sharp = \arg\min \quad & \|\mathbf{Z}\|_1 \\ \text{subject to} \quad & \|\mathcal{A}(\mathbf{Z}) - \mathbf{y}\|_{\ell_q} \leq \eta \end{aligned} \tag{6}$$

algorithmically. This is a computationally tractable convex optimization task. With high probability, the quality of such a reconstruction is then bounded by this a-priori chosen noise bound η. A typical uniform recovery guarantee (see also Theorem 2 below) for a measurement operator $\mathcal{A}$ assures

$$\|\mathbf{X} - \mathbf{Z}^\sharp\|_2 \leq C\frac{\eta}{\sqrt{m}}$$

for all target matrices $\mathbf{X} \in M_d$ with $\mathrm{rank}(\mathbf{X}) \leq r$.

B. Clifford Orbits

Throughout this section (and the remainder of this paper) we shall assume that the dimension d is a power of two.

Let $\mathbf{U}_1, \ldots, \mathbf{U}_{d^2}$ denote the Pauli matrices and $\mathbf{W}_k := \frac{1}{\sqrt{d}}\mathbf{W}_k$ their re-normalized counterparts. These matrices form a unitary operator basis of H_d with respect to the Frobenius inner product:

$$(\mathbf{W}_k, \mathbf{W}_l) = \frac{1}{d}(\mathbf{U}_k, \mathbf{U}_l) = \frac{\delta_{k,l}}{d}\|\mathbf{U}_k\|_2^2 = \delta_{k,l}.$$

The *characteristic function*

$$\mathbf{w} : H_d \to \mathbb{R}^{d^2} \tag{7}$$

$$\mathbf{X} \mapsto \sum_{k=1}^{d^2} (\mathbf{W}_k, \mathbf{X})\, \mathbf{e}_k, \tag{8}$$

maps every hermitian matrix to the vector of expansion coefficients with respect to the basis $\mathbf{W}_1, \ldots, \mathbf{W}_{d^2}$. We point out that the characteristic function is an isometry, i.e. $\|\mathbf{w}(\mathbf{Z})\|_{\ell_2} = \|\mathbf{Z}\|_2 \; \forall \mathbf{Z} \in H_d$. In addition, it obeys

$$\|\mathbf{w}(\mathbf{X})\|_{\ell_\infty} = \max_{1 \leq k \leq d^2} \frac{1}{\sqrt{d}} |(\mathbf{U}_k, \mathbf{X})| \leq \frac{\|\mathbf{X}\|_1}{\sqrt{d}} \quad \forall \mathbf{X} \in H_d \tag{9}$$

according to the Hoelder inequality.

The *Clifford group* $C(d)$ is the group of unitary transformations that—up to a global phase—maps Pauli matrices to Pauli matrices under conjugation. It has many remarkable properties. One of them is that in qubit dimensions it forms a *unitary 3-design* [8, 9]. Roughly speaking, unitary t-designs are discrete subsets of the unitary group $U(d)$ that reproduce the Haar measure up to t-th moments. We refer to [1] for further information. This in turn implies that every orbit $O_{\mathbf{z}} = \{\mathbf{C}\mathbf{z} : \mathbf{C} \in C(d)\}$ forms a complex projective 3-design. Similar to unitary 3-designs, complex projective 3-designs reproduce the first $2t$ moments of the uniform distribution over the complex unit sphere S^{d-1}.

Stabilizer states form the smallest Clifford orbit. The fact that they also constitute a complex projective 3-design was independently derived by a subset of the authors [14].

The 3-design property of any Clifford orbit in particular implies that a random vector $\mathbf{a} \in \mathbb{C}^d$ chosen uniformly from such an orbit obeys

$$\mathbb{E}_{\mathbf{a} \in O_{\mathbf{z}}}[\mathbf{a}\mathbf{a}^*] = \frac{1}{d}\mathbb{1}, \tag{10}$$

$$\mathbb{E}_{\mathbf{a} \in O_{\mathbf{z}}}\left[\mathrm{Ten}^3(\mathbf{a}\mathbf{a}^*)\right] = \binom{d+1}{2}^{-1} \mathbf{P}_{\mathrm{Sym}^2}, \tag{11}$$

$$\mathbb{E}_{\mathbf{a} \in O_{\mathbf{z}}}\left[\mathrm{Ten}^3(\mathbf{a}\mathbf{a}^*)\right] = \binom{d+2}{3}^{-1} \mathbf{P}_{\mathrm{Sym}^3}.$$

Here, $\mathrm{Ten}^k(\mathbf{Z}) = \mathbf{Z}^{\otimes k}$ denote the canonical k-fold tensor product of a matrix $\mathbf{Z} \in H_d$.

Recently, we were able to extend this knowledge about moments to order four:

Theorem 1 (Corollary 3 in [1])**.** *For any power of two d, uniformly sampling $\mathbf{a} \in \mathbb{S}^{d-1}$ from a Clifford orbit with fiducial $\mathbf{z} \in \mathbb{S}^{d-1}$ results in a distribution obeying*

$$\mathbb{E}\left[\mathrm{Ten}^4(\mathbf{a}\mathbf{a}^*)\right] = d\binom{d+2}{3}^{-1}(\alpha_1(\mathbf{z})\mathbf{P}_1 + \alpha_2(\mathbf{z})\mathbf{P}_2),$$

where $\mathbf{P}_1, \mathbf{P}_2 \in \mathrm{Ten}^4(H_d)$ are orthogonal projections that commute with $\mathbf{P}_{\mathrm{Sym}^4}$– the projector onto the totally symmetric subspace. Defining $\mathbf{Q} = \sum_{k=1}^{d^2} \mathrm{Ten}^4(\mathbf{W}_k)$ allows to characterize them explicitly by

$$\mathbf{P}_1 = \mathbf{P}_{\mathrm{Sym}^4}\mathbf{Q} \quad \textit{and} \quad \mathbf{P}_2 = \mathbf{P}_{\mathrm{Sym}^4}(\mathbb{I} - \mathbf{Q})$$

and the coefficients amount to

$$\alpha_1(\mathbf{z}) = \|\mathbf{w}(\mathbf{z}\mathbf{z}^*)\|_{\ell_4}^4 \quad \textit{and} \quad \alpha_2(\mathbf{z}) = 4\frac{1 - \|\mathbf{w}(\mathbf{z}\mathbf{z}^*)\|_{\ell_4}^4}{(d+4)(d-1)},$$

where $\mathbf{w}(\cdot) : H_d \to \mathbb{R}^{d^2}$ is the isometry introduced in (7).

C. Main results

The results in [10, 11] highlight that random measurement projectors onto elements of a 4-design admit a required sampling rate of $m = rd\log(d)$. This is optimal up to a single log-factor. In general, Clifford orbits $O_{\mathbf{z}} \subset \mathbb{S}^{d-1}$ fail to constitute a 4-design. In our main results, we pay the prize for this lack of structure by requiring a (potentially trivial) oversampling factor. Depending only on the Clifford orbit's fiducial $\mathbf{z} \in \mathbb{S}^{d-1}$ and a parameter $\rho \in]0,1[$ it amounts to

$$\kappa(\mathbf{z}, \rho) := \frac{1}{\rho}\max\left\{1, rd\|\mathbf{w}(\mathbf{z}\mathbf{z}^*)\|_{\ell_4}^4\right\} \in \left[\frac{1}{\rho}, \frac{r}{\rho}\right], \tag{12}$$

where $\mathbf{w} : H_d \to \mathbb{R}^{d^2}$ is the i The lower bound on $\kappa(\mathbf{z}, \rho)$'s range is trivial, while the upper bound follows from Lemma 3 below. In our reconstruction guarantee, this factor will feature not only in the sampling rate, but also in the reconstruction bound and the failure probability.

Theorem 2. *Let d be a power of two, $\mathbf{z} \in \mathbb{S}^{d-1}$ and fix $1 \leq r \leq d$, $\rho \in \,]0,1[$. Consider a measurement operator $\mathcal{A}$ containing*

$$m \geq \frac{C_1}{\rho^2}\kappa_{\mathbf{z}}^2 rd\log(d). \tag{13}$$

projectors $\mathbf{A}_k = \mathbf{a}_k\mathbf{a}_k^*$ onto uniformly sampled elements of $\mathbf{z}$'s Clifford orbit. Then, with probability $1 - \mathrm{e}^{-C_2\frac{m}{\kappa_{\mathbf{z}}^2}}$, a noisy measurement process of the form $\mathcal{A}(\mathbf{X}) = \mathbf{y} + \mathbf{e}$ suffices to stably reconstruct the best rank-r approximation of any $\mathbf{X} \in H_d$. Concretely, the solution $\mathbf{Z}^\sharp$ of (6) obeys

$$\|\mathbf{X} - \mathbf{Z}^\sharp\|_2 \leq \frac{C(\rho)}{\sqrt{r}}\sigma_r(\mathbf{X}) + D(\rho)\kappa(\mathbf{z}, \rho)$$

Motivated by the structure of PhaseLift (1) we focus our attention on estimating low rank matrices $\mathbf{X}$ that are in addition positive-semidefinite (psd) and rank-one projective measurements $\mathbf{A}_k = \mathbf{a}_k\mathbf{a}_k^*$ which are also psd.

Such a restriction to psd target matrices $\mathbf{X}$ and measurements $\mathbf{A}_k$ has a crucial advantage. As pointed out in [11–13], such positive semidefinite shape constraints often render nuclear norm minimization superfluous in the algorithmic reconstruction step. This in particular allows for replacing (6) by a simple constrained ℓ_q-regression ($q \geq 1$):

$$\mathbf{Z}^\sharp = \underset{\mathbf{Z}\text{ is psd}}{\arg\min} \left\| \mathcal{A}(\mathbf{Z}) - \mathbf{y} \right\|_{\ell_q}. \tag{14}$$

Compared to (6), Algorithm (14) has the considerable advantage of not requiring an a-priori bound η on the noise corrupting the measurement process. From a practical perspective, this feature is highly desirable and our main result has this feature:

Theorem 3. *Let d be a power of two, $\mathbf{z} \in \mathbb{S}^{d-1}$, fix $1 \leq r \leq d$ and $\rho \in \left]0, \frac{1}{2}\right[$. Let $\mathbf{w}(\cdot) : H_d \to \mathbb{R}^{d^2}$ be the isometry introduced in (7) and define*

$$\kappa_{\mathbf{z}} := \frac{1}{\rho} \max\left\{1, rd \|\mathbf{w}(\mathbf{z}\mathbf{z}^*)\|_{\ell_4}^4\right\} \in \left[\frac{1}{\rho}, \frac{r}{\rho}\right]. \tag{15}$$

Consider a noisy measurement operator $\mathcal{A}$ containing

$$m \geq \frac{C_1}{\rho^2} \kappa_{\mathbf{z}}^2 rd \log(d). \tag{16}$$

projectors $\mathbf{A}_k = \mathbf{a}_k \mathbf{a}_k^*$ onto uniformly sampled elements of $\mathbf{z}$'s Clifford orbit. Then, with probability $1 - de^{-\frac{C_2 m}{\max\left\{\kappa_{\mathbf{z}}^2, d\right\}}}$, a noisy measurement process of the form $\mathcal{A}(\mathbf{X}) = \mathbf{y} + \mathbf{e}$ suffices to stably reconstruct the best rank-r approximation of any psd matrix $\mathbf{X} \in H_d$. Concretely, for any $q \geq 1$, the solution $\mathbf{Z}^\sharp$ of Algorithm (14) obeys

$$\|\mathbf{X} - \mathbf{Z}^\sharp\|_2 \leq \frac{C_3(\rho)}{\sqrt{r}} \sigma_r(\mathbf{X})_1 + C_4(\rho) \frac{\kappa_{\mathbf{z}} \sqrt{(d+1)d}}{m^{1/q}} \|\mathbf{e}\|_{\ell_q},$$

where $\sigma_r(\mathbf{X})_1 = \inf\{\|\mathbf{X} - \mathbf{Z}\|_1,\ \mathbf{Z}$ has rank $r\}$ is the nuclear norm error of best rank-r approximation to the matrix $\mathbf{X}$. Here, C_1, C_2 are absolute constants and $C_3(\rho), C_4(\rho)$ depend exclusively on the choice of ρ (see (46) and (47) below for explicit dependencies).

We note that Lemma 3 below assures that

$$\left\| \mathbf{w}\left(\mathbf{z}\mathbf{z}^*\right) \right\|_{\ell_4}^4 \leq \frac{1}{d} \quad \forall \mathbf{z} \in \mathbb{C}^d \tag{17}$$

is true. So, in the worst case the sampling rate required in (16) amounts to

$$m \geq \frac{C_1}{\rho^4} r^3 d \log(d). \tag{18}$$

While sub-optimal in the rank parameter, its dependence on the ambient dimension d is optimal up to a single log-factor. Clearly, different choices of fiducials $\mathbf{z} \in \mathbb{C}^d$ lead to different requirements on the sampling rate:

1. *Stabilizer states:* (e.g. $\mathbf{z} = \mathbf{e}_1$) (17) is tight, which results in the worst case (18).

2. *"Magic states:"* For $d = 2^n$ set $\mathbf{z} = \mathrm{Ten}^n(\mathbf{m})$ and choose $\mathbf{m} \in \mathbb{C}^d$ such that

   $$\mathbf{m}\mathbf{m}^* = \frac{1}{2}\left(\mathbb{I} \pm \frac{1}{\sqrt{3}}\mathbf{U}_1 \pm \frac{1}{\sqrt{3}}\mathbf{U}_2 \pm \frac{1}{\sqrt{3}}\mathbf{U}_3\right),$$

   where $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3 \in H_2$ denote the non-identity Pauli matrices[21]. By construction such a $\mathbf{x}$ obeys $\|\mathbf{w}(\mathbf{z}\mathbf{z}^*)\|_{\ell_4}^4 < d^{-\frac{3}{2}}$, which results in a required sampling rate

   $$m \geq \frac{C_1}{\rho^4} \max\left\{1, \frac{r^2}{d}\right\} rd \log(d)$$

   that is order-optimal for any rank parameter $1 \leq r \leq \sqrt{d}$.

3. *4-design fiducial:* if $\mathbf{z} \in \mathbb{C}^d$ obeys $\|\mathbf{w}(\mathbf{z}\mathbf{z}^*)\|_{\ell_4}^4 = \frac{4}{d(d+3)}$, the corresponding Clifford orbit forms a complex projective 4-design [1]. This work also shows that such a choice is always feasible and inserting such a $\|\mathbf{w}(\mathbf{z}\mathbf{z})\|_{\ell_4}^4$ into (16) results in a sampling rate requirement

   $$m \geq \frac{4C_1}{\rho^4} rd \log(d) \tag{19}$$

   which is always order-optimal. This result should not come as a surprise, since order optimal uniform recovery guarantees have already been established in [10, 11] based on the 4-design property alone.

Finally, let us turn our attention to the particular case of phase retrieval. The PhaseLift approach by construction assures that the target signal is a psd matrix with rank-one, i.e. $\sigma_1(\mathbf{x}\mathbf{x}^*)_1 = 0$. Setting $r = 1$ and employing (17) we thus may conclude the following from Theorem 3:

Corollary 1 (PhaseLift with Clifford orbits)**.** *Let d be a power of two and suppose that $\mathbf{a}_1, \ldots, \mathbf{a}_m$ are*

$$m \geq C_1 d \log(d) \tag{20}$$

uniformly sampled elements of any Clifford orbit. Then, with probability at least $1 - de^{-\tilde{C}_2 m}$, the associated phaseless measurements $y_k = |\langle \mathbf{a}_k, \mathbf{x} \rangle|^2 + e_k$ allow for estimating any $\mathbf{x}\mathbf{x}^*$ (and thus $\mathbf{x}$) by employing Algorithm (14) with any $q \geq 1$. Its minimizer is guaranteed to obey

$$\|\mathbf{Z}^\sharp - \mathbf{x}\mathbf{x}^*\|_{\ell_2} \leq C_3 \frac{\sqrt{(d+1)d}}{m^{1/q}} \|\mathbf{e}\|_{\ell_q}. \tag{21}$$

Any Clifford orbit forms a 3-design [8, 9, 14]. Viewed from this angle, Corollary 1 may be viewed as a substantial strengthening of the main result in [15] for particular 3-designs.

Note that the factor $\sqrt{(d+1)d}$ in (21) is an artifact of our normalization. If we change the normalization of the sampling vectors from one to $\sqrt[4]{(d+1)d}$ – a length that closely resembles the expected length of random Gaussian vectors – we obtain measurements of the form

$$\tilde{y}_k = |\langle \tilde{\mathbf{a}}_k, \mathbf{x}\rangle|^2 + \tilde{e}_k \quad \text{with} \quad \tilde{\mathbf{a}}_k = \sqrt{(d+1)d}\,\mathbf{a}_k \tag{22}$$

and $\tilde{\mathbf{e}} = \sqrt{(d+1)d}\mathbf{e}$, because the noise term is amplified as well. For re-scaled measurements of this form and $q=1$, Corollary 1 assures whp that for any $\mathbf{x} \in \mathbb{C}^d$, solving

$$\mathbf{Z}^\sharp = \underset{\mathbf{Z} \text{ is psd}}{\arg\min} \sum_{k=1}^m |\langle \tilde{\mathbf{a}}_k, \mathbf{Z}\tilde{\mathbf{a}}_k\rangle - \tilde{y}_k| \tag{23}$$

yields a matrix obeying

$$\|\mathbf{Z}^\sharp - \mathbf{x}\mathbf{x}^*\|_2 \le C_3 \frac{\|\tilde{\mathbf{e}}\|_{\ell_2}}{m}. \tag{24}$$

Up to a single log-factor in the sampling rate (20) and a slightly weaker bound on the probability of failure ($d\mathrm{e}^{-\tilde{C}_2 m}$ vs. $\mathcal{O}(\mathrm{e}^{-\gamma m})$) this special case reproduces the main result in [16] – the strongest recovery guarantee for PhaseLift with Gaussian measurements available.

II. Proofs

A. Null space properties under positive semidefinite constraints

Low rank matrix recovery aims to reconstruction rank-r matrices $\mathbf{X} \in H_d$ form an incomplete collection of m linear, and potentially noisy, measurements (3). A necessary and sufficient criterion for this to be uniformly possible (i.e. all matrices of rank $\le r$ may be recovered), is that the measurement operator $\mathcal{A}$ obeys a *null space property* [17]. A strong matrix version thereof was introduced in [11]:

Definition 1 (Definition 3.1 in [11] for hermitian matrices). *For fixed r and $q \ge 1$, a measurement operator $\mathcal{A}: H_d \to \mathbb{R}^m$ obeyes the ℓ_q-*robust null space property *of order r (r/ℓ_q-NSP) with constants $\rho \in]0,1[$ and $\tau > 0$, if*

$$\|\mathbf{Z}_r\|_2 \le \frac{\rho}{\sqrt{r}}\|\mathbf{Z}_c\|_1 + \tau\|\mathcal{A}(\mathbf{Z})\|_{\ell_q} \quad \forall \mathbf{Z} \in H_d. \tag{25}$$

Here $\mathbf{Z}_r$ denotes the best rank-r approximation of $\mathbf{Z}$ and $\mathbf{Z}_c = \mathbf{Z} - \mathbf{Z}_r$ is the error matrix of best rank-r approximation. Consequently, $\|\mathbf{Z}_c\|_1$ equals $\sigma_r(\mathbf{Z})_1$ introduced in Theorem 3.

Validity of a r/ℓ_q-NSP assures that any matrix $\mathbf{Z}$ with rank at most r need obey $\|\mathbf{Z}\|_2 \le \tau\|\mathcal{A}(\mathbf{Z})\|_{\ell_q}$. This assures that no such matrix can lie in $\mathcal{A}$'s null space. While such a criterion is clearly neccessary for uniform rank-r matrix recovery, the following statements shows that it is also sufficient.

Theorem 4 (Theorem 3.3 in [11] for hermitian matrices). *Fix $r, q \ge 1$ and suppose that $\mathcal{A}: H_d \to \mathbb{R}^m$ obeys a r/ℓ_q-NSP with constants $\rho \in]0,1[$ and $\tau > 0$. Then*

$$\begin{aligned}\|\mathbf{Z} - \mathbf{X}\|_2 \le& \frac{C_\rho}{\sqrt{r}}(\|\mathbf{Z}\|_1 - \|\mathbf{X}\|_1 + 2\|\mathbf{X}_c\|_1) \\ &+ D_\rho \tau \|\mathcal{A}(\mathbf{Z} - \mathbf{X})\|_{\ell_q} \quad \forall \mathbf{X}, \mathbf{Z} \in H_d,\end{aligned} \tag{26}$$

with $C_\rho = \frac{(1+\rho)^2}{1-\rho}$ and $D_\rho = \frac{3+\rho}{1-\rho}$.

The nuclear norm difference $\|\mathbf{Z}\|_1 - \|\mathbf{X}\|_1$ appearing in (26) motivates to perform a constrained nuclear norm minimization (6) in order to estimate a matrix $\mathbf{X}$ from noisy measurements $\mathbf{y} = \mathcal{A}(\mathbf{X}) + \mathbf{e}$ with $\|\mathbf{e}\|_{\ell_q} \le \eta$. By construction, the target matrix $\mathbf{X}$ is a feasible point of this algorithm which implies $\|\mathbf{Z}^\sharp\|_1 \le \|\mathbf{X}\|_1$ and

$$\begin{aligned}\|\mathcal{A}(\mathbf{X}) - \mathbf{Z})\|_{\ell_q} \le& \|\mathcal{A}(\mathbf{X}) - \mathbf{y}\|_{\ell_q} + \|\mathcal{A}(\mathbf{Z}) - \mathbf{y}\|_{\ell_q} \\ \le& \|\mathbf{e}\|_{\ell_q} + \eta \le 2\eta.\end{aligned}$$

Inserting these features into (26) assures

$$\|\mathbf{Z}^\sharp - \mathbf{X}\|_2 \le \frac{2C_\rho}{\sqrt{r}}\sigma_r(\mathbf{X})_1 + 2D_\rho\tau\eta, \tag{27}$$

provided that $\mathcal{A}$ obeys a r/ℓ_q-NSP. This stable and uniform recovery guarantee for constrained nuclear norm minimization underlines that a NSP is also sufficient for matrix recovery.

In this work we shall assume more structure: namely that the target matrices are also psd. For a pair of psd matrices $\mathbf{X}, \mathbf{Z}$, the nuclear norm difference in (26) amounts to

$$\|\mathbf{Z}\|_1 - \|\mathbf{X}\|_1 = \mathrm{tr}(\mathbf{Z}) - \mathrm{tr}(\mathbf{X}) = \mathrm{tr}\left(\mathbb{I}(\mathbf{Z} - \mathbf{X})\right). \tag{28}$$

In addition, a measurement operator $\mathcal{A}$ containing m random Clifford orbit measurements $\mathbf{A}_k = \mathbf{a}_k\mathbf{a}_k^*$ obeys

$$\frac{d}{m}\mathbb{E}\left[\mathcal{A}^*(\mathbf{1})\right] = \mathbb{E}\left[\sum_{k=1}^m \frac{d}{m}\mathbf{A}_k\right] = \mathbb{I}, \tag{29}$$

because every Clifford orbit forms a tight frame (10). Here, $\mathcal{A}^*: \mathbb{R}^m \to H_d$ denotes the adjoint of $\mathcal{A}$ and $\mathbf{1} = (1,\ldots,1)^T \in \mathbb{R}^m$ is the "all-ones" vector. Combining these two statements assures

$$\begin{aligned}\|\mathbf{Z}\|_1 - \|\mathbf{X}\|_1 =& \frac{d}{m}\langle \mathbf{1}, \mathbb{E}\left[\mathcal{A}\right](\mathbf{X} - \mathbf{Z})\rangle \\ \le& \frac{d}{m^{\frac{1}{q}}}\|\mathbb{E}\left[\mathcal{A}\right](\mathbf{X} - \mathbf{Z})\|_{\ell_q}\end{aligned}$$

for any pair of psd matrices $\mathbf{X}, \mathbf{Z}$. So, at least in expectation, the nuclear norm difference ($\|\mathbf{X} - \mathbf{Z}\|_1$) between psd matrices is controllable by means of $\|\mathcal{A}(\mathbf{X} - \mathbf{Z})\|_{\ell_q}$ – the second term featuring in (26). This is a strong indication that for psd matrices, the first term in said bound is superfluous. And it certainly is, if (29) were also true for $\mathcal{A}$'s concrete realization – this realization is due to Kalev[22] et. al. [12]. One of the main results in [11] further generalizes this statement: for measurements forming a tight frame, any actual realization of $\mathcal{A}^*(\mathbf{1})$ does not deviate much from its expectation whp [11, Proposition 8.3]:

$$\Pr\left[\left\|\frac{d}{m}\mathcal{A}^*(\mathbf{1})-\mathbb{I}\right\|_\infty \geq \beta\right] \leq de^{-\frac{3\beta^2 m}{8(d-1)}} \;\forall \beta \in [0,1[. \quad (30)$$

This readily follows from applying a matrix Bernstein deviation inequality (see e.g. proof of Proposition 8.3 in [11]). Theorem 8.1 in [11] then assures that validity of $\|\mathcal{A}^*(\mathbf{1})-\mathbb{I}\|_\infty < \beta$ suffices to omit the nuclear norm terms in (26) , provided that both $\mathbf{X}$ and $\mathbf{Z}$ are psd. This omission comes at the price of larger constants C_ρ, D_ρ and tighter conditions on ρ that all depend on $\beta \in [0,1[$. These dependences is particularly simple for $\beta = \frac{\sqrt{2}-1}{\sqrt{2}+1}$, where we obtain:

Corollary 2. *Suppose that $\mathcal{A}: H_d \to \mathbb{R}^m$ obeys a r/ℓ_q-NSP with parameters $\rho \in \left]0,\frac{1}{2}\right[$ and $\tau > 0$ that in addition obeys*

$$\left\|\frac{d}{m}\mathcal{A}^*(\mathbf{1})-\mathbb{I}\right\|_\infty < \frac{\sqrt{2}-1}{\sqrt{2}+1}. \quad (31)$$

Then,

$$\|\mathbf{Z}-\mathbf{X}\|_2 \leq \frac{\tilde{C}_\rho}{\sqrt{r}}\|\mathbf{X}_c\|_1 + \tilde{D}_\rho\left(\frac{d}{m^{\frac{1}{q}}}+\tau\right)\|\mathcal{A}(\mathbf{X}-\mathbf{Z})\|_{\ell_q}$$

is true for any pair of psd matrices $\mathbf{X},\mathbf{Z} \in H_d$. The constants amount to $\tilde{C}_\rho = 4\frac{(1+2\rho)^2}{1-2\rho}$ and $\tilde{D}_\rho = 2\frac{3+2\rho}{1-2\rho}$.

Moreover, if $\mathcal{A}$ consists of m projectors onto uniformly sampled elements of a tight frame, then (31) holds with probability at least $1-de^{-C_4\frac{m}{d-1}}$.

B. A null space property for Clifford orbits

Recall that a measurement operator $\mathcal{A}: H_d \to \mathbb{R}^m$ obeys a r/ℓ_q-NSP, if

$$\|\mathbf{Z}_r\|_2 \leq \frac{\rho}{\sqrt{r}}\|\mathbf{Z}_c\|_1 + \tau\|\mathcal{A}(\mathbf{Z})\|_{\ell_q} \quad \forall \mathbf{Z} \in H_d.$$

For any fixed r and ρ, all matrices $\mathbf{Z} \in H_d$ obeying $\|\mathbf{Z}_r\|_2 \leq \frac{\rho}{\sqrt{r}}\|\mathbf{Z}_c\|_1$ meet this requirement by default. Also, the NSP is invariant under scaling, which allows us to set $\|\mathbf{Z}\|_2 = 1$ without loss of generality. So, when aiming to establish a r/ℓ_q-NSP with constant $\rho \in [0,1[$ for any $\mathcal{A}$, we may restrict our attention to

$$T_{\rho,r} = \left\{\mathbf{Z} \in H_d : \|\mathbf{Z}_r\|_2 > \frac{\rho}{\sqrt{r}}\|\mathbf{Z}_c\|_1, \|\mathbf{Z}\|_2 = 1\right\} \subset H_d. \quad (32)$$

And a measurement operator $\mathcal{A}$ obeys the r/ℓ_q-NSP with constants $\rho \in]0,1[$ and $\tau > 0$, if

$$\inf_{\mathbf{Z}\in T_{\rho,r}} \|\mathcal{A}(\mathbf{Z})\|_{\ell_q} \geq \frac{1}{\tau}. \quad (33)$$

Note that the parameters r,ρ implicitly feature in the definition of $T_{\rho,r}$, while τ is inversly proportional to the best lower bound we manage to establish in (33). The space $T_{\rho,r}$ is contained in H_d – a d^2-dimensional real vector space. Moreover, the "effective rank" of any $\mathbf{Z} \in T_{\rho,r}$ cannot be too large:

Lemma 1. *Let $T_{\rho,r} \subset H_d$ be the set introduced in (32) for some $\rho \in]0,1[$ and $1 \leq r \leq d$. Then:*

$$\|\mathbf{Z}\|_1^2 = \frac{\|\mathbf{Z}\|_1^2}{\|\mathbf{Z}\|_2^2} \leq \left(\frac{\rho+1}{\rho}\right)^2 r \quad \forall \mathbf{Z} \in T_{\rho,r}. \quad (34)$$

Proof. Combining $\|\mathbf{Z}_r\|_1 \leq \sqrt{r}\|\mathbf{Z}_r\|_2$ with the defining property of $\mathbf{Z} \in T_{\rho,r}$ reveals

$$\|\mathbf{Z}\|_1 = \|\mathbf{Z}_r\|_1 + \|\mathbf{Z}_c\|_1 \leq \frac{\rho+1}{\rho}\sqrt{r}\|\mathbf{Z}_r\|_2,$$

and the claim follows from $\|\mathbf{Z}_r\|_2 \leq \|\mathbf{Z}\|_2 = 1$. □

Also, $\mathcal{A}: H_d \to \mathbb{R}^m$ is comprised of m independently selected projectors onto random elements of a Clifford orbit. The real-valued structure of the underlying vector space together with independence of the individual measurement matrices allows us to employ Mendelson's small ball method [18–20]. This strong probabilistic anti-concentration inequality will enable us to establish (33) with high probability.

Theorem 5 (Mendelson's small ball method)**.** *Fix $E \subset \mathbb{R}^d$ arbitrary and let $\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_m \in \mathbb{R}^d$ be independent copies of a random vector $\boldsymbol{\phi}$. For $\xi > 0$ define*

$$Q_\xi(E;\boldsymbol{\phi}) = \inf_{\mathbf{z}\in E}\Pr\left[|\langle\boldsymbol{\phi},\mathbf{z}\rangle| \geq \xi\right], \quad \textit{and} \quad (35)$$

$$W_m(E;\boldsymbol{\phi}) = \mathbb{E}\left[\sup_{\mathbf{z}\in E}\langle\mathbf{h},\mathbf{z}\rangle\right] \quad \textit{with} \quad (36)$$

$$\mathbf{h} = \frac{1}{\sqrt{m}}\sum_{k=1}^m \epsilon_k\boldsymbol{\phi}_k \in \mathbb{R}^d, \quad (37)$$

where $\epsilon_1,\ldots,\epsilon_m$ is a Rademacher sequence. Then for any $\xi > 0$ and $t \geq 0$, the following bound is true with probability at least $1-e^{-2t^2}$:

$$\frac{1}{\sqrt{m}}\inf_{\mathbf{z}\in E}\sum_{k=1}^m|\langle\boldsymbol{\phi}_k,\mathbf{z}\rangle| \geq \xi\sqrt{m}Q_{2\xi}(E;\boldsymbol{\phi}) - 2W_m(E;\boldsymbol{\phi}) - \xi t. \quad (38)$$

We emphasize that this is not the standard result known as "Mendelson's small ball method". The latter establishes a lower bound on $\inf_{\mathbf{z}\in E}\sqrt{\sum_{k=1}^m|\langle\boldsymbol{\phi}_k,\mathbf{z}\rangle|^2}$.

As such, the assertion of Theorem 5 is stronger, but is also implied by Mendelson's original proof. Adapting this statement to the cause at hand yields the following corollary which we are going to employ in order to establish (33).

Corollary 3 (Consequence of Remark 5.1 in [11])**.** *Fix r, ρ and let $T_{\rho,r} \subset H_d$ be the set introduced in (32). Suppose that $\mathcal{A} : H_d \to \mathbb{R}^m$ is a measurement operator containing m independent instances of a single random matrix $\mathbf{A} \in H_d$ as individual measurements. Then for any $q \geq 1$, $\xi > 0$ and $t \geq 0$*

$$\begin{aligned}&\inf_{\mathbf{Z}\in T_{\rho,r}} \|\mathcal{A}(\mathbf{Z})\|_{\ell_q}\\ &\geq m^{\frac{1}{q}-\frac{1}{2}} \left(\xi\sqrt{m}Q_{2\xi}(T_{\rho,r};\mathbf{A}) - 2W_m(T_{\rho,r},\mathbf{A}) - \xi t\right)\end{aligned}$$

is true with probability at least $1 - \mathrm{e}^{-2t^2}$. Here $Q_{2\xi}(E;\mathbf{A})$ and $W_m(E,\mathbf{A})$ are the parameters defined in (35) and (36).

Proof. H_d is a real-valued vector space isomorphic to $\mathbb{R}^{d^2}$. By assumption, we may also identify each $\mathbf{A}_k$ with an instance $\boldsymbol{\phi}_k$ of the random "vector" $\mathbf{A} := \boldsymbol{\phi}_k \in \mathbb{R}^{d^2} \simeq H_d$. We may also identify any $T_{\rho,r} \subset H_d \simeq \mathbb{R}^{d^2}$ with the set E in Theorem 5. Said theorem is applicable and allows us to bound

$$\inf_{\mathbf{Z}\in T_{\rho,r}} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} |(\mathbf{A}_k,\mathbf{z})| = \inf_{\mathbf{Z}\in T_{\rho,r}} \frac{1}{\sqrt{m}} \|\mathcal{A}(\mathbf{Z})\|_{\ell_1}$$

from below. Finally, we employ the basic norm inequality $\|\mathbf{z}\|_1 \leq m^{1-\frac{1}{q}}\|\mathbf{z}\|_{\ell_q}$ $\forall \mathbf{z} \in \mathbb{R}^m$, $\forall q \geq 1$ [17, Equation A.3] to conclude

$$\inf_{\mathbf{Z}\in T_{\rho,r}} \|\mathcal{A}(\mathbf{Z})\|_{\ell_q} \geq m^{\frac{1}{q}-\frac{1}{2}} \inf_{\mathbf{Z}\in T_{\rho,r}} \frac{1}{\sqrt{m}} \|\mathcal{A}(\mathbf{Z})\|_{\ell_1}$$

and the claim follows with Mendelson's bound (38). □

In our sampling model, $\mathcal{A}$ consists of m projectors $\mathbf{A}_k = \mathbf{a}_k\mathbf{a}_k^*$, where each $\mathbf{a}_k$ is sampled uniformly from a Clifford orbit $O_{\mathbf{z}} \subset \mathbb{C}^d$. The choice of a fiducial $\mathbf{z} \in \mathbb{S}^{d-1}$ is arbitrary. According to Equation 10 each Clifford orbit forms a tight frame. Combining this feature with the structural insights from Lemma 1 allows us to establish the following bound:

Proposition 1. *For any $\mathbf{z} \in \mathbb{S}^{d-1}$, let $\mathbf{A} = \mathbf{a}\mathbf{a}^*$ be a projector onto $\mathbf{a} \in O_{\mathbf{z}}$ chosen uniformly at random. Also, fix $1 \leq r \leq d$, $\rho \in]0,1[$ and $m \geq 2d\log(d)$. Then the parameter $W_m(T_{\rho,r},\mathbf{A})$ featuring in Corollary 3 obeys*

$$W_m(T_{\rho,r};\mathbf{A}) \leq \frac{6.2098}{\rho}\sqrt{\frac{r\log(2d)}{d+1}}.$$

Proof. This proof closely resembles a comparable analysis provided in [10]. Matrix Hoelder together with Lemma 1 assures

$$\begin{aligned}W_m(T_{\rho,r};\mathbf{A}) &= \mathbb{E}\left[\sup_{\mathbf{Z}\in T_{\rho,r}} (\mathbf{H},\mathbf{Z})\right] \leq \sup_{\mathbf{Z}\in T_{\rho,r}} \|\mathbf{Z}\|_1 \mathbb{E}\left[\|\mathbf{H}\|_\infty\right]\\ &\leq \frac{\rho+1}{\rho}\sqrt{r}\mathbb{E}\left[\|\mathbf{H}\|_\infty\right] \leq \frac{2}{\rho}\sqrt{r}\mathbb{E}\left[\|\mathbf{H}\|_\infty\right],\end{aligned}$$

where $\mathbf{H} = \frac{1}{\sqrt{m}}\sum_{k=1}^m \epsilon_k \mathbf{a}_k\mathbf{a}_k^*$. Each $\mathbf{a}_k$ is by assumption chosen from a tight frame and normalized to one. This, together with the assumption $m \geq 2d\log(d)$, allows for bounding $\mathbb{E}\left[\|\mathbf{H}\|_\infty\right]$ by means of [10, Proposition 13]. Adapting said statement to unit normalization of the $\mathbf{a}_k$'s yields

$$\mathbb{E}\left[\|\mathbf{H}\|_\infty\right] \leq 3.1049\sqrt{\frac{\log(2d)}{d+1}}$$

and the claim readily follows. □

Establishing a sufficiently tight lower bound on the other parameter – $Q_\xi\left(T_{\rho,r};\mathbf{a}\mathbf{a}^*\right)$ defined in (35) – for Clifford orbits is considerably more challenging. The reason for this complication is that Clifford orbits in general do not constitute a complex projective 4-design. As demonstrated in [10, 11], a 4-design property alone allows for achieving the task at hand by applying a Paley-Zygmund argument. Unfortunately, Clifford orbits in general do not have this structural property.

However, novel insights about the structure of the Clifford group [1] – see also section I B – allow us to still carry out a similar Paley-Zygmund argument.

Lemma 2. *Fix $\mathbf{Z} \in T_{\rho,r}$ for some $1 \leq r \leq d$, $\rho \in]0,1[$ and define the random variable $S := \langle \mathbf{a}, \mathbf{Z}\mathbf{a}\rangle$, where $\mathbf{a}$ is uniformly chosen from a Clifford orbit $O_{\mathbf{z}}$ with fiducial $\mathbf{z} \in \mathbb{S}^{d-1}$. Then*

$$\mathbb{E}\left[S^2\right] = \frac{1+\mathrm{tr}(\mathbf{Z})^2}{(d+1)d} \geq \frac{1}{(d+1)d} \quad \textit{and} \tag{39}$$

$$\mathbb{E}\left[S^4\right] \leq 84\kappa_{\mathbf{z}}\mathbb{E}\left[S^2\right]^2, \tag{40}$$

$$\tag{41}$$

where $\kappa_{\mathbf{z}} = \max\left\{1, rd\|\mathbf{w}(\mathbf{z}\mathbf{z}^*)\|_{\ell_4}^4\right\}$ was defined in (15).

While (39) directly follows from the 2-design property of Clifford orbits, establishing the bound (40) is considerably more challenging. Said bound constitutes this work's main technical contribution. We devote section II D to proving it. We point out that a comparable bound for 4-designs would read $\mathbb{E}\left[S^4\right] \leq 24\mathbb{E}\left[S^2\right]^2$ [10, Proof of Proposition 12]. Inequality (40) is weaker than such a 4-design bound. The nature of this bound will ultimately result in the additional factor $\kappa_{\mathbf{z}}$ featuring in Theorem 3.

Proposition 2. *For any* $\mathbf{z} \in \mathbb{S}^{d-1}m$ *let* $\mathbf{A} = \mathbf{a}\mathbf{a}^*$ *be a projector onto* $\mathbf{a} \in O_{\mathbf{z}}$ *chosen uniformly at random. Then the parameter* $Q_\xi(T_{\rho,r}, \mathbf{A})$*, featuring in Corollary 3, obeys*

$$Q_\xi \left(T_{\rho,r}; \mathbf{A}\right) \geq \frac{1}{\kappa_{\mathbf{z}}} \left(1 - \left(\sqrt{(d+1)d}\xi\right)^2\right)^2$$

for any $0 \leq \xi \leq \frac{1}{\sqrt{(d+1)d}}$, $1 \leq r \leq d$ *and* $\rho \in]0,1[$.

Proof. Fix $\mathbf{Z} \in T_{\rho,r}$, $\xi \geq 0$ and define the real-valued random variable $S = \langle \mathbf{a}, \mathbf{Z}\mathbf{z} \rangle$, where $\mathbf{a}$ is chosen uniformly from $O_{\mathbf{z}}$. Then

$$\begin{aligned}\Pr\left[|\langle \mathbf{a}, \mathbf{Z}\mathbf{a}\rangle| \geq \xi\right] =& \Pr\left[|S| \geq \xi\right] = \Pr\left[\mathbb{S}^2 \geq \xi^2\right] \\ \geq& \Pr\left[\mathbb{S}^2 \geq (d+1)d\xi^2 \mathbb{E}\left[\mathbb{S}^2\right]\right],\end{aligned}$$

where the last inequality is due to (39). Applying the Paley-Zygmund inequality to the non-negative random variable $\mathbb{S}^2$ assures

$$\begin{aligned}\Pr\left[\mathbb{S}^2 \geq (d+1)d\xi^2\mathbb{E}\left[\mathbb{S}^2\right]\right] \geq& \left(1 - (d+1)d\xi^2\right)^2 \frac{\mathbb{E}\left[\mathbb{S}^2\right]^2}{\mathbb{E}\left[\mathbb{S}^4\right]} \\ \geq& \frac{1}{\kappa_{\mathbf{z}}}\left(1 - (d+1)d\xi^2\right)^2,\end{aligned}$$

where the last line is due to (40). Since such a lower bound is valid for any $\mathbf{Z} \in T_{\rho,r}$ we may conclude that it also holds for

$$Q_\xi\left(T_{\rho,r}; \mathbf{A}\right) = \inf_{\mathbf{Z} \in T_{\rho,r}} \Pr\left[|\langle \mathbf{a}, \mathbf{Z}\mathbf{a}\rangle| \geq \xi\right].$$

□

We have now assembled all necessary ingredients to lower bound $\inf_{\mathbf{Z} \in T_{\rho,r}} \|\mathcal{A}(\mathbf{Z})\|_{\ell_q}$ for any choice of q, r, d. Applying Corollary 3 with $\xi = \frac{1}{4\sqrt{(d+1)d}}$ and $t = \frac{\sqrt{\tilde{C}_2 m}}{\kappa_{\mathbf{z}}}$ – where $\tilde{C}_2$ is a sufficiently small constant – implies

$$\begin{aligned}\inf_{\mathbf{Z} \in T_{\rho,r}} \|\mathcal{A}(\mathbf{Z})\|_{\ell_q} \geq& m^{\frac{1}{q}-\frac{1}{2}} \left(\sqrt{m}\frac{Q_{\frac{1}{2\sqrt{(d+1)d}}}(T_{\rho,r}; \mathbf{A})}{4\sqrt{(d+1)d}} - 2W_m(T_{\rho,r}; \mathbf{A}) - \frac{\sqrt{\tilde{C}_2 m}}{4\sqrt{(d+1)d}\kappa_{\mathbf{z}}}\right) \\ \geq& \frac{m^{\frac{1}{q}-\frac{1}{2}}}{\kappa_{\mathbf{z}}\sqrt{(d+1)d}} \left(\frac{9\rho\sqrt{m}}{1344} - \frac{13}{\rho}\sqrt{\kappa_{\mathbf{z}}^2 rd \log(2d)} - \frac{\sqrt{\tilde{C}_2 m}}{4}\right) \qquad (42)\end{aligned}$$

with probability at least $1 - e^{-\frac{2\tilde{C}_2 m}{\kappa_{\mathbf{z}}^2}}$. In the last line, we have inserted the bounds provided by Proposition 1 and Proposition 2, respectively. Let us now fix

$$m \geq \frac{C_1}{\rho^2}\kappa_{\mathbf{z}}^2 rd \log(2d),$$

where C_1 is a sufficiently large constant. Such a choice assures that the bracket in (42) is lower bounded by $\frac{\sqrt{m}}{C_3}$, where C_3 is constant. Inserting this novel bound into (42) allows us to conclude

$$\inf_{\mathbf{Z} \in T_{\rho,r}} \|\mathcal{A}(\mathbf{Z})\|_{\ell_q} \geq \frac{m^{\frac{1}{q}-\frac{1}{2}}}{\sqrt{(d+1)d}} \frac{\sqrt{m}}{C_3\kappa_{\mathbf{z}}} = \frac{m^{\frac{1}{q}}}{C_3\kappa_{\mathbf{z}}\sqrt{(d+1)d}}$$

with high probability. Comparing this bound to (33) reveals that it in turn establishes a NSP for $\mathcal{A}$:

Theorem 6. *Fix* $1 \leq r \leq d$, $\rho \in]0,1[$, $q \geq 1$ *and* $\mathbf{z} \in \mathbb{S}^{d-1}$ *and* $\kappa_{\mathbf{z}}$ *defined in (15). Suppose that* $\mathcal{A} : H_d \to \mathbb{R}^m$ *contains*

$$m \geq \frac{C_1}{\rho^2}\kappa_{\mathbf{z}}^2 rd \log(2d)$$

projectors onto randomly selected elements of the Clifford orbit $O_{\mathbf{z}}$. *Then, with probability at least* $1 - e^{-\frac{\tilde{C}_2 m}{\kappa_{\mathbf{z}}^2}}$, *this* $\mathcal{A}$ *obeys the* r/ℓ_q*-NSP from Definition 1 with paramters*

$$\rho \quad \textit{and} \quad \tau = \frac{C_3\kappa_{\mathbf{z}}\sqrt{(d+1)d}}{m^{\frac{1}{q}}}.$$

Here, $C_1, \tilde{C}_2, C_3$ *denote constants of sufficient size.*

C. Algorithmic Implications and proof of Theorem 3

If Corollary 2 is valid, it has profound algorithmic implications for recovering psd matrices with low rank [11] (see also [13] for similar approach to compressed sensing). If we fix $\rho \in \left]0, \frac{1}{2}\right[$ and $1 \leq r \leq d$, a combination of said statement with Theorem 6 assures that $\mathcal{A}$ consisting of $m = C_1 \kappa_{\mathbf{z}} r d \log(2d)$ random Clifford measurements obeys

$$\begin{aligned} &\|\mathbf{X}-\mathbf{Z}\|_2 \\ \leq& \frac{\tilde{C}_\rho}{\sqrt{r}} \|\mathbf{X}_1\|_1 + \tilde{D}_\rho \left(\frac{d}{m^{\frac{1}{q}}} + \tau \right) \|\mathcal{A}(\mathbf{X}-\mathbf{Z})\|_{\ell_q} \\ \leq& \frac{\tilde{C}_\rho}{\sqrt{r}} \|\mathbf{X}_1\|_1 + \tilde{D}_\rho \left(\frac{d}{m^{\frac{1}{q}}} + \frac{C_3 \kappa_{\mathbf{z}} \sqrt{(d+1)d}}{m^{\frac{1}{q}}} \right) \|\mathcal{A}(\mathbf{X}-\mathbf{Z})\|_{\ell_q} \\ \leq& \frac{\tilde{C}_\rho}{\sqrt{r}} \|\mathbf{X}_1\|_1 + \frac{2 C_3 \tilde{D}_\rho}{m^{\frac{1}{q}}} \kappa_{\mathbf{z}} \sqrt{(d+1)d} \, \|\mathcal{A}(\mathbf{X}-\mathbf{Z})\|_{\ell_q} \quad \forall \mathbf{X}, \mathbf{Z} \text{ psd} \end{aligned} \tag{43}$$

with probability of failure bounded by

$$d \mathrm{e}^{-\frac{C_4 m}{d+1}} + \mathrm{e}^{-\frac{\tilde{C}_2 m}{\kappa_{\mathbf{z}}^2}} \leq (d+1) \mathrm{e}^{-\frac{\min\{C_4, \tilde{C}_2\}}{\max\left\{\kappa_{\mathbf{z}}^2, d+1\right\}}} \leq d \mathrm{e}^{-\frac{C_2 m}{\max\left\{\kappa_{\mathbf{z}}^2, d\right\}}} \tag{44}$$

according to the union bound. Once more, C_2 is a constant.

If we now aim at recover a psd target matrix $\mathbf{X}$ from noisy measurements of the form

$$\mathcal{A}(\mathbf{X}) = \mathbf{y} + \mathbf{e}$$

(see (5)), this sampling process implies

$$\|\mathcal{A}(\mathbf{X}-\mathbf{Z})\|_{\ell_q} = \|\mathcal{A}(\mathbf{Z}) - \mathbf{y} - \mathbf{e}\|_{\ell_q} \leq \|\mathcal{A}(\mathbf{Z}) - \mathbf{y}\|_{\ell_q} + \|\mathbf{e}\|_{\ell_q}$$

for any psd $\mathbf{Z}$. Fixing the target matrix $\mathbf{X}$ and inserting this bound into (43) assures

$$\begin{aligned} &\|\mathbf{X}-\mathbf{Z}\|_{\ell_q} \\ \leq& \frac{\tilde{C}_\rho}{\sqrt{r}} \|\mathbf{X}_c\|_1 + \frac{2 C_3 \tilde{D}_\rho}{m^{\frac{1}{q}}} \kappa_{\mathbf{z}} \left(\|\mathcal{A}(\mathbf{Z}) - \mathbf{y}\|_{\ell_q} + \|\mathbf{e}\|_{\ell_q} \right) \quad \forall \mathbf{Z} \text{ psd.} \end{aligned} \tag{45}$$

In order to obtain a good estimate, it thus makes sense to minimize the r.h.s. of this bound over the free parameter $\mathbf{Z}$:

$$\mathbf{Z}^\sharp = \underset{\mathbf{Z} \text{ psd}}{\arg\min} \|\mathcal{A}(\mathbf{X}) - \mathbf{y}\|_{\ell_q}.$$

This is the psd least squares regression advertised in (14). Crucially, this program does only depend on $\mathcal{A}$ and the data $\mathbf{y}$. It does not require any a priori assumptions on the noise term $\mathbf{e}$. Also, $\mathbf{Z} = \mathbf{X}$ is a feasible point of this optimization, and so

$$\|\mathcal{A}(\mathbf{Z}^\sharp) - \mathbf{y}\|_{\ell_q} \leq \|\mathcal{A}(\mathbf{X}) - \mathbf{y}\|_{\ell_q} = \|\mathbf{e}\|_{\ell_q}.$$

From this and (45) we thus may conclude

$$\|\mathbf{X} - \mathbf{Z}^\sharp\|_{\ell_q} \leq \frac{\tilde{C}_\rho}{\sqrt{r}} \|\mathbf{X}_c\|_1 + \frac{4 C_3 \tilde{D}_\rho}{m^{\frac{1}{q}}} \kappa_{\mathbf{z}} \|\mathbf{e}\|_{\ell_q}.$$

Setting

$$C_3(\rho) = \tilde{C}_\rho = 4 \frac{(1+2\rho)^2}{1-2\rho} \quad \text{and} \tag{46}$$

$$C_4(\rho) = 4 C_3 \tilde{D}_\rho = 8 C_3 \frac{3+2\rho}{1-2\rho} \tag{47}$$

yields the main assertation of Theorem 3. The expression in (44) bounds the probability of this conclusion failing and the proof of Theorem 3 is complete.

D. Proof of Lemma 2

The first statement follows directly from the fact that, endowed with uniform weights, every Clifford orbit forms a 2-design – equation (11). For any $\mathbf{Z} \in H_d$ this implies

$$\begin{aligned} \mathbb{E}\left[\langle \mathbf{a}, \mathbf{Z}\mathbf{a} \rangle^2\right] =& \mathrm{tr}\left(\mathbb{E}\left[\mathrm{Ten}^2(\mathbf{a}\mathbf{a}^*)\right] \mathrm{Ten}^2(\mathbf{Z}) \right) \\ =& \binom{d+1}{2}^{-1} \mathrm{tr}\left(\mathbf{P}_{\mathrm{Sym}^2} \mathrm{Ten}^2(\mathbf{Z}) \right) \\ =& \frac{\mathrm{tr}\left(\mathbf{Z}^2\right) + \mathrm{tr}\left(\mathbf{Z}\right)^2}{(d+1)d}, \end{aligned}$$

where the last line e.g. follows from [10, Lemma 17]. Equation (11) is equivalent to this statement, because every $\mathbf{Z} \in T_{\rho,r}$ obeys $\mathrm{tr}(\mathbf{Z}^2) = \|\mathbf{Z}\|_2^2 = 1$.

For the second bound, we heavily rely on Theorem 1. Said statement assures that choosing $\mathbf{a}$ uniformly from a Clifford orbit $O_{\mathbf{z}}$ with $\mathbf{z} \in \mathbb{S}^{d-1}$ assures

$$\begin{aligned} &\frac{1}{d}\binom{d+2}{3} \mathbb{E}\left[\langle \mathbf{a}, \mathbf{Z}\mathbf{a} \rangle^4\right] \\ =& \frac{1}{d}\binom{d+2}{3} \left(\mathbb{E}\left[\mathrm{Ten}^4\left(\mathbf{a}\mathbf{a}^*\right)\right], \mathrm{Ten}^4(\mathbf{Z}) \right) \\ =& \alpha_1(\mathbf{z}) \left(\mathbf{P}_1, \mathrm{Ten}^4(\mathbf{Z}) \right) + \alpha_2(\mathbf{z}) \left(\mathbf{P}_2, \mathrm{Ten}^4(\mathbf{Z}) \right), \end{aligned} \tag{48}$$

for any $\mathbf{Z} \in H_d$. Recall that $\alpha_1(\mathbf{z}), \alpha_2(\mathbf{z})$, as well as $\mathbf{P}_1$ and $\mathbf{P}_2$ were introduced in said theorem. For the first inner product we may conclude

$$\begin{aligned} \left| \left(\mathbf{P}_1, \mathrm{Ten}^4(\mathbf{Z}) \right) \right| \leq& \left(\mathbf{P}_1, \left| \mathrm{Ten}^4(\mathbf{Z}) \right| \right) \leq \left(\mathbf{Q}, \mathrm{Ten}^4\left(|\mathbf{Z}|\right) \right) \\ =& \sum_{k=1}^{d^2} \left(\mathrm{Ten}^4(\mathbf{W}_k), \mathrm{Ten}^4(|\mathbf{Z}|) \right) \\ =& \sum_{k=1}^{d} \left(\mathbf{W}_k, |\mathbf{Z}| \right)^4 = \|\mathbf{w}(|\mathbf{Z}|)\|_{\ell_4}^4, \end{aligned} \tag{49}$$

using several inequalities valid for positive-semidefinite matrices, as well as $\mathbf{P}_1 \leq \mathbf{Q} = \sum_{k=1}^{d^2} \mathrm{Ten}^4(\mathbf{W}_k)$. The following Lemma allows us to relate $\|\mathbf{w}(|\mathbf{Z}|)\|_{\ell_4}^4$ to the "effective rank" of $\mathbf{Z}$.

Lemma 3. *Any $\mathbf{Z} \in H_d$ with $\|\mathbf{Z}\|_2 = 1$ obeys*

$$\|\mathbf{w}(\mathbf{Z})\|_{\ell_p}^4 \leq \frac{1}{d}\|\mathbf{Z}\|_1^2.$$

Proof. Fix $\mathbf{Z} \in H_d$. Then $\mathbf{w}(\mathbf{Z})$ is a vector in $\mathbb{R}^{d^2}$ which in particular obeys

$$\|\mathbf{w}(\mathbf{Z})\|_{\ell_4}^4 \leq \|\mathbf{w}(\mathbf{Z})\|_{\ell_2}^2 \|\mathbf{w}(\mathbf{Z})\|_{\ell_\infty}^2.$$

Since $\mathbf{w} : H_d \to \mathbb{R}^{d^2}$ is an isometry, we have $\|\mathbf{w}(\mathbf{Z})\|_{\ell_2}^2 = \|\mathbf{Z}\|_2^2$. Also, Matrix Hoelder assures

$$\|\mathbf{w}(\mathbf{Z})\|_{\ell_\infty}^2 = \max_{1 \leq k \leq d^2} (\mathbf{W}_k, \mathbf{Z})^2 \leq \|\mathbf{W}_k\|_\infty^2 \|\mathbf{Z}\|_1^2 \leq \frac{1}{d}\|\mathbf{Z}\|_2^2,$$

and the claim follows. □

We note in passing that this bound also assures

$$\kappa_{\mathbf{z}} = \frac{1}{\rho} \max\left\{1, \|\mathbf{w}(\mathbf{z}\mathbf{z})^*\|_{\ell_4}^4\right\} \leq \frac{r}{\rho} \quad \forall \mathbf{z} \in \mathbb{S}^{d-1},$$

because $\|\mathbf{z}\mathbf{z}^*\|_1 = \|\mathbf{z}\mathbf{z}^*\|_2 = 1$.

At this point the restriction $\mathbf{Z} \in T_{\rho,r}$ becomes important. Under this assumption, a combination of (49), Lemma 3 and Lemma 1 assures

$$\left|\left(\mathbf{P}_1, \mathrm{Ten}^4(\mathbf{Z})\right)\right| \leq \frac{1}{d}\||\mathbf{Z}\|_1^2 = \frac{1}{d}\|\mathbf{Z}\|_1^2 \leq \frac{2r}{\rho d} \tag{50}$$

for any $\mathbf{Z} \in T_{\rho,r}$.

Let us now move on to bound the second inner product in (48). Using $\mathbf{P}_2 = \mathbf{P}_{\mathrm{Sym}^4} - \mathbf{P}_1$ and (50) allows us to conclude

$$\begin{aligned}\left|\left(\mathbf{P}_2, \mathrm{Ten}^4(\mathbf{Z})\right)\right| \leq& \left|\left(\mathbf{P}_{\mathrm{Sym}^4}, \mathrm{Ten}^4(\mathbf{Z})\right)\right| + \left|\left(\mathbf{P}_1, \mathrm{Ten}^4(\mathbf{Z})\right)\right| \\ \leq& \left|\left(\mathbf{P}_{\mathrm{Sym}^4}, \mathrm{Ten}^4(\mathbf{Z})\right)\right| + \frac{2r}{\rho d}.\end{aligned}$$

The remaining inner product is a standard expression in multilinear algebra and can for instance be computed using [10, Lemma 17]. Further bounding the resulting expressions results in

$$\left|\left(\mathbf{P}_{\mathrm{Sym}^4}, \mathrm{Ten}^4(\mathbf{Z})\right)\right| \leq \max\left\{\|\mathbf{Z}\|_2^4, \mathrm{tr}(\mathbf{Z})^4\right\} \quad \forall \mathbf{Z} \in H_d$$

as is shown, for instance, in [10, Proof of Proposition 12]. Employing the trivial bound $r \leq d$, as well as $\|\mathbf{Z}\|_2 = 1$ allows us to conclude

$$\begin{aligned}\left|\left(\mathbf{P}_1, \mathrm{Ten}^4(\mathbf{Z})\right)\right| \leq& \max\left\{1, \mathrm{tr}(\mathbf{Z})^4\right\} + \frac{2r}{\rho d} \\ \leq& \frac{3}{\rho}\left(1 + \mathrm{tr}(\mathbf{Z})^2\right)^2 \quad \forall \mathbf{Z} \in T_{\rho,r}.\end{aligned}$$

Finally, let us turn to the constants featuring in (48). Lemma 3 assures

$$0 \leq \alpha_2(\mathbf{z}) = 4\frac{1 - \|\mathbf{w}(\mathbf{z}\mathbf{z}^*)\|_{\ell_4}^4}{(d+4)(d-1)} \leq \frac{4}{(d+4)(d-1)}.$$

Fixing $\mathbf{Z} \in T_{\rho,r}$ and putting together all these individual bounds yields

$$\begin{aligned}\mathbb{E}\left[\langle \mathbf{a}, \mathbf{Z}\mathbf{a}\rangle^4\right] \leq& \binom{d+2}{3}^{-1} d\alpha_1(\mathbf{z}) \left|\left(\mathbf{P}_1, \mathrm{Ten}^4(\mathbf{Z})\right)\right| + \binom{d+2}{3} d\alpha_2(\mathbf{z}) \left|\left(\mathbf{P}_2, \mathrm{Ten}^4(\mathbf{Z})\right)\right| \\ \leq& \binom{d+2}{3}^{-1} \left(\frac{2rd\|\mathbf{w}(\mathbf{z}\mathbf{z}^*)\|_{\ell_4}^4}{d\rho} + \frac{12d\left(1 + \mathrm{tr}(\mathbf{Z})^2\right)^2}{(d+4)(d-1)\rho}\right) \\ \leq& \frac{84}{\rho} \max\left\{1, rd\|\mathbf{w}(\mathbf{z}\mathbf{z}^*)\|_{\ell_4}^4\right\} \left(\frac{1 + \mathrm{tr}(\mathbf{Z})^2}{(d+1)d}\right)^2 = 84\kappa_{\mathbf{z}} \mathbb{E}\left[\langle \mathbf{a}, \mathbf{Z}\mathbf{a}\rangle^2\right]^2,\end{aligned}$$

where the last equality is due to (39).

[1] H. Zhu, R. Kueng, M. Grassl, and D. Gross, "The clifford group fails gracefully to be a unitary 4-design," *to appear*, 2016.

[2] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin, "Painless reconstruction from magnitudes of frame coefficients.," *J. Fourier Anal. Appl.*, vol. 15, pp. 488–501, 2009.

[3] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imaging Sci.*, vol. 6, pp. 199–225, 2013.

[4] E. Candès, T. Strohmer, and V. Voroninski, "PhaseLift:

Exact and stable signal recovery from magnitude measurements via convex programming," *Comm. Pure Appl. Math.*, vol. 66, pp. 1241–1274, 2013.

[5] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[6] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inform. Theory*, vol. 57, pp. 1548–1566, 2011.

[7] M. Kech, "Explicit frames for deterministic phase retrieval via phaselift," *arXiv preprint arXiv:1508.00522*, 2015.

[8] H. Zhu, "Multiqubit clifford groups are unitary 3-designs," *arXiv preprint arXiv:1510.02619*, 2015.

[9] Z. Webb, "The clifford group forms a unitary 3-design," *arXiv preprint arXiv:1510.02769*, 2015.

[10] R. Kueng, H. Rauhut, and U. Terstiege, "Low rank matrix recovery from rank one measurements," *Appl. Comput. Harmonic Anal.*, to appear. DOI:10.1016/j.acha.2015.07.007.

[11] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege, "Stable low-rank matrix recovery via null space properties," *arXiv preprint arXiv:1507.07184*, 2015.

[12] A. Kalev, C. Riofrio, R. Kosut, and I. Deutsch, "Informationally complete measurements from compressed sensing methodology," *Bulletin of the American Physical Society*, vol. 60, 2015.

[13] R. Kueng and P. Jung, "Robust nonnegative sparse recovery and the nullspace property of 0/1 measurements," *arXiv preprint arXiv:1603.07997*, 2016.

[14] R. Kueng and D. Gross, "Qubit stabilizer states are complex projective 3-designs," *arXiv preprint arXiv:1510.02767*, 2015.

[15] D. Gross, F. Krahmer, and R. Kueng, "A partial derandomization of Phaselift using spherical designs," *J. Fourier Anal. Appl.*, pp. 1–38, 2014.

[16] E. J. Candès and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Found. Comput. Math.*, pp. 1–10, 2013.

[17] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis, Birkhäuser/Springer, New York, 2013.

[18] V. Koltchinskii and S. Mendelson, "Bounding the smallest singular value of a random matrix without concentration," *Internat. Math. Res. Notices*, p. rnv096, 2015.

[19] S. Mendelson, "Learning without Concentration," *J. ACM*, vol. 62, no. 3, pp. 1–25, 2015.

[20] J. A. Tropp, "Convex recovery of a structured signal from independent random linear measurements," in *Sampling Theory, a Renaissance*, pp. 67–101, Birkhäuser/Springer, 2015.

[21] Choosing exclusively "+"-signs in that construction, for instance, results in $\mathbf{m} = \frac{1-i}{2}\sqrt{1+\frac{1}{\sqrt{3}}}\mathbf{e}_1 + \frac{1}{\sqrt{3+\sqrt{3}}}\mathbf{e}_2 \in \mathbb{C}^2$

[22] Actually, they show that the more general demand $\exists \mathbf{t} \in \mathbb{R}^m : \mathcal{A}^*(\mathbf{t}) \propto \mathbb{I}$ suffices to draw such a conclusion

Distinguishability of quantum states under Clifford orbits

Richard Küng,[1] Huangjun Zhu,[1] and David Gross[1]

[1] *Institute for Theoretical Physics, University of Cologne, Germany*

(Dated: September 5, 2016)

Helstrom's Theorem assert that the maximal probability of correctly distinguishing two quantum states is proportional to their trace distance. However, achieving this bound requires one to be able to perform arbitrary measurements that depend on the particular choice of state.

Following Matthews *et al.* [1], we consider the task of distingushing arbitrary quantum states via a fixed measurement. In particular, we focus on multi-qubit dimensions and POVMs that correspond to orbits of the Clifford group. We show that the distinguishing capabilities of such measurements depend on the rank of the states to be distinguished: if both states are approximately pure, Clifford orbits perform essentially optimally. However, if the states are close to maximally mixed, the maximal bias achievable becomes considerably worse.

I. INTRODUCTION

A. Distinguishing quantum states

On the space of d-dimensional quantum states $\mathcal{S}_d$, the *trace distance*

$$d\left(\rho,\sigma\right) = \frac{1}{2}\left\|\rho-\sigma\right\|_1$$

constitutes a very prominent and meaningful distance measure. It features prominently in *Helstrom's Theorem* [2]. Seid theorem asserts that the maximal probability of correctly identifying one out of two known quantum states $\rho,\sigma \in \mathcal{S}_d$ with a single measurement ("single-shot") amounts to

$$\mathrm{Pr}_{\mathrm{Helstrom}} = \frac{1}{2} + d\left(\frac{1}{2}\rho, \frac{1}{2}\sigma\right), \tag{1}$$

provided that both ρ and σ appear with equal probability. A slight generalization of this statement takes into account the possibility that ρ occurs with probality $\tau \in [0,1]$ and σ with probability $1-\tau$. The optimal success probability then becomes

$$\mathrm{Pr}_{\mathrm{Helstrom}} = \frac{1}{2} + d\left(\tau\rho, (1-\tau)\sigma\right),$$

which reduces to (1), if $\tau = \frac{1}{2}$. This corresponds to a maximal bias

$$\beta_{\mathrm{Helstrom}}(\rho,\sigma,\tau) = d\left(\tau\rho,(1-\tau)\sigma\right) \leq 1$$

towards correctly identifying the state. If $\tau \in [0,1]$ and $\rho,\sigma \in \mathcal{S}\left(\mathbb{S}^d\right)$ are known, this maximal bias is achievable by an optimal strategy involving a two-outcome projective measurement[16]. However, such a measurement is optimized to distinguish ρ from σ and may perform considerably worse (or even fail completely) at distinguishing other pairs of states.

Addressing this lack of universality, Matthews, Wehner and Winter [1] turned this problem around: instead of fixing the state pair ρ,σ and varying the measurement, they consider a fixed POVM measurement — i.e. a family of positive semidefinite operators $\{M_k\}_{k=1}^N$ that sum up to unity: $\sum_{k=1}^N M_k = \mathbb{I}$ — and analyze its performance at distinguishing all possible pairs of states $\rho,\sigma \in \mathcal{S}\left(\mathbb{C}^d\right)$. Born's rule asserts that such a POVM maps any state $\rho \in \mathcal{S}_d$ to a discrete probability vector

$$p_\rho = \mathcal{M}(\rho) = \sum_{k=1}^N |e_k\rangle \mathrm{tr}\left(M_k\rho\right) \in \mathbb{R}^N$$

which encompasses all the classical information about ρ that is accessible to us. So, distinguishing ρ from σ necessarily reduces to the task of distinguishing p_ρ from p_σ. If ρ and σ are equiprobable, the optimal decision rule for doing so is the *maximum likelihood rule*[17]. It results in a bias proportional to the total variational distance of p_ρ and p_σ:

$$\beta_{\mathcal{M}}\left(\rho,\sigma,\frac{1}{2}\right) = \frac{1}{4}\left\|p_\rho - p_\sigma\right\|_{\ell_1} = \left\|\mathcal{M}\left(\frac{1}{2}\rho - \frac{1}{2}\sigma\right)\right\|_{\ell_1}.$$

If ρ occurs instead with probability $\tau \neq \frac{1}{2}$ this bias generalizes to

$$\beta_{\mathcal{M}}\left(\rho,\sigma,\tau\right) = \left\|\mathcal{M}\left(\tau\rho - (1-\tau)\sigma\right)\right\|_{\ell_1}.$$

Helstrom's Theorem demands

$$\beta_{\mathcal{M}}\left(\rho,\sigma,\tau\right) \leq \beta_{\mathrm{Helstrom}}(\rho,\sigma,\tau)$$

for all $\rho,\sigma \in \mathcal{S}_d$ and any $\tau \in [0,1]$. On the contrary, if $\mathcal{M}$ is informationally completete,

$$\beta_{\mathcal{M}}\left(\rho,\sigma,\tau\right) > 0 \quad \forall \rho,\sigma \in \mathcal{S}_d,\ \forall \tau \in [0,1]$$

follows by definition. What is more, informational completeness of $\mathcal{M}$ assures that

$$\|\cdot\|_{\mathcal{M}} = \left\|\mathcal{M}\left(\cdot\right)\right\|_{\ell_1}$$

does consitute a norm on $\mathcal{H}_d$ — the vector space of all hermitian $d \times d$-matrices. Since all norms are equivalent on finite dimensional Hilbert spaces, there is a constant $\lambda_{\mathcal{M}}$ such that

$$\lambda_{\mathcal{M}}\|X\|_1 \leq \|X\|_{\mathcal{M}} \quad \forall X \in \mathcal{H}_d.$$

This constant $\lambda_{\mathcal{M}}$ allows us to relate the optimal distinguishability bias achievable by the fixed measurement $\mathcal{M}$ to Helstrom's optimal one:

$$\beta_{\mathcal{M}}(\rho,\sigma,\tau) \geq \lambda_{\mathcal{M}}\, \beta_{\text{Helstrom}}(\rho,\sigma,\tau) \tag{2}$$

for any $\rho,\sigma \in \mathcal{S}_d$ and any $\tau \in [0,1]$. Matthews et al. then moved on to derive sharp bounds on $\lambda_{\mathcal{M}}$ for different families of informationally complete POVMs $\mathcal{M} : \mathcal{H}_d \to \mathbb{R}^N$ [1]. While these bounds have the advantage of being universal, they may be too pessimistic, if the states $\rho,\sigma \in \mathcal{S}_d$ have additional structure. We will come back to this in section II.

B. Complex projective designs

A natural candidate for a single POVM that should perform well at distinguishing quantum states is the *uniform POVM* $\mathcal{M}_{\text{unif}}$ consisting of all projectors onto elements of the complex unit sphere in $\mathbb{C}^d$. The associated distinguishability norm of this POVM obeys [1, Theorem 8]

$$\|\rho-\sigma\|_{\mathcal{M}_{\text{unif}}} \geq \frac{1}{\sqrt{d}}\left(\sqrt{\frac{2}{\pi}} - o(1)\right)\|\rho-\sigma\|_1 \quad \forall \rho,\sigma \in \mathcal{S}_d$$

This in turn implies $\lambda_{\mathcal{M}_{\text{unif}}} = \mathcal{O}\left(d^{-\frac{1}{2}}\right)$ for the constant defined in (2), provided that $\tau = \frac{1}{2}$.

While the uniform POVM is excellent to provide a benchmark for the capability of distinguishing quantum states by means of a fixed quantum measurement, the POVM is far too big for all practical purposes. Natural candidates for "coarse-graining" the uniform POVM are complex projective t-designs [3–5]. A t-design POVM may be viewed as a subset of the uniform POVM that "evenly" approximates the latter up to a certain degree:

Definition 1 (Complex projective design). *A (proper)* complex projective t-design *is a set of unit vectors* $\{|x_k\rangle\}_{k=1}^N \subset \mathbb{C}^d$ *whose outer products obey*

$$\frac{1}{N}\sum_{k=1}^{N} |x_k\rangle\langle x_k|^{\otimes t} = \int_{\|v\|_{\ell_2}=1} \mathrm{d}v |v\rangle\langle v|^{\otimes t}, \tag{3}$$

where integration on the r.h.s. is taken with respect to the uniform measure on the complex unit sphere. Likewise, we call the set $\left\{\frac{d}{N}|x_k\rangle\langle x_k|\right\}_{k=1}^N$ *a* t-design POVM.

Ambainis and Emerson [4] observed that 4-design POVMs already essentially match the distinguishability capacity of the uniform POVM, see also [1, Eq. (15)]. Our first contribution consists of a slight generalization of these results:

Theorem 1 (Performance of 4-designs)**.** *Let $\mathcal{M}_{4D}$ be a 4-design POVM. Then*

$$\|X\|_{\mathcal{M}_{4D}} > \frac{0.32}{\sqrt{\text{rank}(X)}}\|X\|_1 \quad \forall X \in \mathcal{H}_d. \tag{4}$$

This in particular implies that the distinguishability constant (2) *obeys $\lambda_{\mathcal{M}_{4D}} > \frac{0.32}{\sqrt{d}}$. If X has rank 2, then*

$$\|X\|_{\mathcal{M}_{4D}} > \frac{1}{\sqrt{12.2\,\text{rank}(X)}}\|X\|_1. \tag{5}$$

If X has rank 2 and is traceless, then

$$\|X\|_{\mathcal{M}_{4D}} > \frac{1}{\sqrt{12\,\text{rank}(X)}}\|X\|_1. \tag{6}$$

The original statements in [1, 4] require X to be traceless, which is not the case here. This generalization comes at the prize of a sligthly smaller constant in (4) (0.32 vs. $\frac{1}{3}$ for traceless matrices).

In stark contrast to this almost optimal behaviour of 4-designs, 2-design POVMs perform remarkably bad at distinguishing quantum states:

Theorem 2 (Theorem 12 in [1])**.** *Let $\mathcal{M}_{2D}$ be a 2-design POVM. Then*

$$\|X\|_{\mathcal{M}_{2D}} \geq \frac{1}{2(d+1)}\|X\|_1 \tag{7}$$

for any traceless operator $X \in \mathcal{H}_d$ which in turn implies $\lambda_{\mathcal{M}_{2D}} \geq \frac{1}{2(d+1)}$, provided that $\tau = \frac{1}{2}$ in (2).

The factor $\frac{1}{d+1}$ in (7) is unavoidable without further assumptions on the 2-design POVM [1, Section 2.C].

II. MAIN RESULTS

Very little is known about the distinguishability quality of POVMs in the intermediate regime between 2- and 4-designs. The main scope of this work is to fill this gap.

To this end we consider POVMs that correspond to orbits of the multi-qubit Clifford group ($d = 2^n$). The Clifford group plays a fundamental role in many areas of quantum information science, such as quantum computing, quantum error correction, tomography and randomized benchmarking. In qubit dimensions $d = 2^n$, the Clifford group $C(d) \subset U(d)$ corresponds to the normalizer of the Pauli group — i.e. up to global phases every $C \in C(d)$ maps Pauli operators onto Pauli operators under conjugation.

Definition 2 (Clifford POVM). *Set $d = 2^n$ and fix $|z\rangle \in \mathbb{C}^d$ with unit length. Let $\{C_k|z\rangle : C_k \in C(d)\}$ denote the orbit of z under the Clifford group and N its cardinality. We then define the asociated* Clifford POVM (anchored at $|z\rangle$) *to be*

$$\mathcal{M}_{C,z} = \left\{\frac{d}{N}C_k|z\rangle\langle z|C_k^\dagger : C_k \in C(d)\right\}.$$

The multi-qubit Clifford group has a very rich structure. Among other things it forms a *unitary* 3*-design* [6, 7]. Unitary t-design are a generalization of the complex projective t-design concept to unitary matrices [8, 9]. They have the particular property that every orbit of a unitary t-design is proportional to a complex projective t-design. This in turn implies that every Clifford POVM is also a 3-design POVM (provided that $d = 2^n$). Multiqubit *stabilizer states*—arguably the most prominent Clifford orbit which arises e.g. from choosing z to be any vector in the (extended) computational basis)—are a particularly prominent example. For this particular orbit, the 3-design property was established independently [10].

A. Main technical results

In a recent survey we have analyzed the fourth moments of the multiqubit Clifford group [11] from a representation theoretic perspective. It turns out that these moments are very similar to the corresponding moments of the full unitary group, although the group does not constitute a unitary 4-design [6]. In turn, this insight allows us to compute the first four moments of Clifford orbits. They behave very similarly to the corresponding moments of a complex projective 4-design—see Theorem 6 below. This information allows us to adapt the proof technique from Theorem 1 [1, 4] and prove a corresponding statement for Clifford orbits. Interestingly, the capacity for distinguishing different states depends on the choice of the Clifford orbit's fiducial z. Recall that the *characteristic function* of a quantum state $\rho \in \mathcal{S}_d$ amounts to

$$w(\rho) = \frac{1}{\sqrt{d}} \sum_{k=1}^{d^2} \mathrm{tr}\left(W_k \rho\right) |e_k\rangle \in \mathbb{R}^{d^2}, \tag{8}$$

where $W_1, \ldots, W_{d^2} \in \mathcal{H}_d$ denote the d-dimensional Pauli matrices ($d = 2^n$) and $|e_1\rangle, \ldots, |e_{d^2}\rangle$ is the standard basis of $\mathbb{R}^{d^2}$. Our main technical result reads as follows:

Theorem 3. *Fix $d = 2^n$ and let $\mathcal{M}_{C,z}$ be a Clifford POVM resulting from a unit-length fiducial $|z\rangle \in \mathbb{C}^d$. Then*

$$\|X\|_{\mathcal{M}_{C,z}} \geq \frac{\|X\|_1}{\sqrt{\left[6d\|w(|z\rangle\langle z|)\|_{\ell_4}^4 \,\mathrm{rank}(X) + 10\right]\mathrm{rank}(X)}}$$

for any $X \in \mathcal{H}_d$. Here the constant 10 may be replaced by 9 if X is traceless.

According to the theorem,

$$\|X\|_{\mathcal{M}_{C,z}} \geq \frac{\|X\|_1}{4\sqrt{\mathrm{rank}(X)}} \tag{9}$$

for any $X \in \mathcal{H}_d$ obeying $\mathrm{rank}(X) \leq 1/(d\|w(|z\rangle\langle z|)\|_{\ell_4}^4)$ and

$$\|X\|_{\mathcal{M}_{C,z}} \geq \frac{\|X\|_1}{4\sqrt{d}\|w(|z\rangle\langle z|)\|_{\ell_4}^2 \mathrm{rank}(X)} \tag{10}$$

otherwise.

For a typical Clifford orbit, the value of $\|w(|z\rangle\langle z|)\|_{\ell_4}^4$ is usually very close to $\|_{\ell_4}^4 \leq 4/(d(d+3))$ [11]. Such orbits behavior almost exactly as 4-designs according to the following theorem.

Theorem 4. *Fix $d = 2^n$ and let $\mathcal{M}_{C,z}$ be a Clifford POVM resulting from a unit-length fiducial $|z\rangle \in \mathbb{C}^d$ which obeys $\|w(|z\rangle\langle z|)\|_{\ell_4}^4 \leq 6/(d(d+3))$. Then*

$$\|X\|_{\mathcal{M}_{C,z}} \geq \frac{\|X\|_1}{\sqrt{22\,\mathrm{rank}(X)}}$$

for any $X \in \mathcal{H}_d$. Here the constant 22 may be replaced by 21 if X is traceless.

It is worthwhile to point out that, unlike its counterparts Theorem 1 for 4- and 2-design POVMs (Theorem 1 and Theorem 2), the statement in Theorem 3 is very sensible towards the rank of the matrix X considered. If $\mathrm{rank}(X)$ is below a certain threshold (which depends on the choice of fiducial), the favourable bound (9) applies. Such a situation is comparable to the 4-design case. However, above this threshold one needs to resort to the much weaker bound (10) whose scaling is comparable to the 2-design case, provided that $\mathrm{rank}(X)$ approaches d.

The following converse statement shows that such a behavior is essentially unavoidable

Theorem 5. *Fix $d = 2^n$, let $\mathcal{M}_{C,z}$ denote a Clifford POVM with fiducial $|z\rangle \in \mathbb{C}^d$ and fix $W \in \mathcal{H}_d$ to be any Pauli matrix ($W \neq \mathbb{I}$). Then*

$$\|W\|_{\mathcal{M}_{C,z}} = \frac{\sqrt{d}\|w(|z\rangle\langle z|)\|_{\ell_1} - 1}{(d+1)(d-1)}\|W\|_1. \tag{11}$$

The coefficient in the theorem satisfies

$$\frac{1}{d+1} \leq \frac{\sqrt{d}\|w(|z\rangle\langle z|)\|_{\ell_1} - 1}{(d+1)(d-1)} \leq \frac{1}{\sqrt{d+1}}, \tag{12}$$

which follows from the property of the characteristic function for a pure state. Here the lower bound is saturatd if and only z is a stabilizer state, and the upper bound is saturated iff

$$|\langle z|W_k|z\rangle| = \frac{1}{\sqrt{d+1}}, \quad \forall 2 \leq k \leq d^2, \tag{13}$$

in which case the orbit of z under the action of the Pauli group forms a symmetric informationally complete POVMs []. The lower bound in (12) can be improved if $\|w(|z\rangle\langle z|)\|_{\ell_4}$ is known,

$$\frac{\sqrt{d}\|w(|z\rangle\langle z|)\|_{\ell_1} - 1}{(d+1)(d-1)} \geq \frac{\frac{\sqrt{d}}{\|w(|z\rangle\langle z|)\|_{\ell_4}^2} - 1}{(d+1)(d-1)} \tag{14}$$

which follows from the equation below,

$$\|w(|z\rangle\langle z|)\|_{\ell_1} \geq \sqrt{\frac{\|w(|z\rangle\langle z|)\|_{\ell_2}^6}{\|w(|z\rangle\langle z|)\|_{\ell_4}^4}} = \frac{1}{\|w(|z\rangle\langle z|)\|_{\ell_4}^2}. \tag{15}$$

We now move on to discussing the implications of our findings for three different Clifford orbits:

(i) *Stabilizer states:* multi-qubit stabilizer states form a particular Clifford orbit with $N = 2^n \prod_{j=1}^n \left(2^j + 1\right)$ elements. The characteristic function of any stabilizer state has precisely d non-vanishing components with constant modulus $\frac{1}{\sqrt{d}}$—see section III G below. This in turn implies $d\|w(|z\rangle\langle z|)\|_{\ell_4}^4 = 1$ for any stabilizer state fiducial $|z\rangle \in \mathbb{C}^d$. Consequently, (9) is only valid for rank-one matrices X, where $\sqrt{\operatorname{rank}(X)}$ and $\operatorname{rank}(X)$ coincide. In turn we need to conclude

$$\|X\|_{\mathcal{M}_{\text{stab}}} \geq \frac{1}{4\operatorname{rank}(X)} \|X\|_1, \tag{16}$$

for any $X \in \mathcal{H}_d$. This is a worst case behavior for any Clifford orbit. However, Theorem 5 assures that such a scaling is unavoidable: the characteristic function of any stabilizer state obeys $\|w(|z\rangle\langle z|)\|_{\ell_1} = \sqrt{d}$ and inserting this into (11) reveals

$$\|W\|_{\mathcal{M}_{\text{stab}}} = \frac{d}{d+1} \frac{\|W\|_1}{\operatorname{rank}(W)} \tag{17}$$

for any Pauli matrix $W \neq \mathbb{I}$. This equation implies that (16) is actually tight up to multiplicative constants.

(ii) *Magic state fiducial:* Let $|z\rangle\langle z| = \rho_{\text{magic}}^{\otimes n}$ be the n-fold tensor product of a the single qubit "magic state"

$$\rho_{\text{magic}} = \frac{1}{2}\left(\mathbb{I} + \frac{1}{\sqrt{3}}\left(W_1 + W_2 + W_3\right)\right) \in \mathcal{S}_2.$$

Such a fiducial obeys $d\|w\left(|z\rangle\langle z|\right)\|_{\ell_4}^4 = \left(\frac{2}{3}\right)^n < \frac{1}{\sqrt{d}}$ (see Eq. (38) below). This is considerably smaller than the analogous quantity for stabilizer states. In turn, Theorem 3 implies that Clifford orbit POVMs with a magic state fiducial obey

$$\|X\|_{\mathcal{M}_{C,\text{magic}}} \geq \frac{1}{4\sqrt{\operatorname{rank}(X)}} \|X\|_1$$

for any $X \in \mathcal{H}_d$ with $\operatorname{rank}(X) \leq \left(\frac{3}{2}\right)^n$. For matrices X whose rank exceeds $\left(\frac{3}{2}\right)^n$, (10) still assures

$$\|X\|_{\mathcal{M}_{C,\text{magic}}} \geq \frac{\left(\frac{3}{2}\right)^{n/2}\|X\|_1}{4\operatorname{rank}(X)} > \frac{d^{0.29}\|X\|_1}{4\operatorname{rank}(X)} \tag{18}$$

which outperforms the analogous bound for stabilizer states by a factor of $d^{0.29}$. Conversely, Theorem 5 assures

$$\|W\|_{C,\text{magic}} \leq \frac{d^{0.45}\|W\|_1}{\operatorname{rank}(W)},$$

because $\sqrt{d}\|w(|z\rangle\langle z|)\|_{\ell_1} = (1+\sqrt{3})^n \leq d^{1.45}$ (see Eq. (39) below). Unlike before, this bound is to weak to assure tightness of (18) (up to multiplicative constants). However, asymptotically it does rule out the possibility of an (optimal) 4-design scaling for this type of Clifford orbits.

(iii) *4-design fiducial:* As pointed out in [11], particular choices of fiducials $|z\rangle \in \mathbb{C}^d$ result in Clifford orbits that actually form a complex projective 4-design. The necessary and sufficient requirement for such fiducials is $\|w(|z\rangle\langle z|)\|_{\ell_4}^4 = \frac{4}{d(d+3)}$. According to Theorem 1,

$$\|X\|_{\mathcal{M}_{C,4D}} \geq \frac{0.32}{\sqrt{\operatorname{rank}(X)}} \|X\|_1 \quad \forall X \in \mathcal{H}_d.$$

This bound is optimal up to a small multiplicative constant since

$$\|W\|_{\mathcal{M}_{C,4D}} \leq \frac{1}{\sqrt{d+1}} \|W\|_1 < \frac{1}{\sqrt{\operatorname{rank}(W)}} \|W\|_1$$

for any Pauli matrix W that is not proportional to the identity, according to Theorem 5 and (12).

B. Implications for distinguishing quantum states

Let us now turn back our attention to the task of distinguishing different quantum states in the single shot limit. Matthews et al. introduced the proportionality constant $\lambda_{\mathcal{M}}$ (2) to compare the performance of a fixed POVM $\mathcal{M}$ directly to Helstrom's optimal strategy. Without putting further restrictions on the states $\rho, \sigma \in \mathcal{S}_d$ to be distinguished, Theorem 3 only allows us to infer

$$\lambda_{\mathcal{M}_{C,z}} \geq \frac{1}{\sqrt{d\left[6d^2\|w(|z\rangle\langle z|)\|_{\ell_4}^4 + 10\right]}} \tag{19}$$

When $\|w(|z\rangle\langle z|)\|_{\ell_4}^4 \leq 6/(d(d+3))$, Theorem 4 implies that

$$\lambda_{\mathcal{M}_{C,z}} \geq \frac{1}{\sqrt{22d}}. \tag{20}$$

For Clifford POVMs with fiducial $|z\rangle \in \mathbb{C}^d$. For the particular case of multi-qubit stabilizer states, we have

$$\frac{1}{\sqrt{6d}} \leq \lambda_{\mathcal{M}_{\text{stab}}} \leq \frac{1}{d+1}. \tag{21}$$

Here the lower bound is derived in section III; the upper bound follows from (17)[18] and is strictly speaking only valid for $\tau = \frac{1}{2}$. This highlights that the constant $\lambda_{\mathcal{M}_{\mathrm{stab}}}$ scales like $\lambda_{\mathcal{M}_{\mathrm{2D}}}$ from Theorem 2—despite the fact that multi-qubit stabilizer form in fact a 3-design.

For Clifford orbits with a magic state fiducial we obtain

$$\frac{1}{4d^{0.71}} \leq \lambda_{\mathcal{M}_{C,\mathrm{magic}}} \leq \frac{1}{d^{0.55}}.$$

This bound is more reassuring. Qualitatively, it assures that the capacity of such POVMs to distinguish quantum states is at least "half way" between the existing 2-design ($\lambda_{\mathrm{2d}} \geq \frac{1}{2(d+1)}$) and 4-design guarantees ($\lambda_{\mathrm{4d}} \geq \frac{1}{4\sqrt{d}}$). Naively, one may expect precisely such a behavior for 3-designs.

We emphasize that the constant $\lambda_{\mathcal{M}}$ is a worst case promise for correctly distinguishing *any* pair of states $\rho, \sigma \in \mathcal{S}_d$. In particular, it may be too pessimistic for more concrete scenarios where additional structure is present. One model assumption, which is often met in practice, is approximate purity. In the extreme case, where both ρ and σ are assumed to be pure, Theorem 3 assures $\lambda_{\mathcal{M}_{C,z}}|_{\rho,\sigma \text{ pure}} \geq 1/\sqrt{44}$ for any Clifford orbit, including stabilizer states. A slightly better bound is derived in section III,

$$\lambda_{\mathcal{M}_{C,z}}|_{\rho,\sigma \text{ pure}} \geq \frac{1}{6}. \tag{22}$$

Up to a multiplicative constant, this reproduces the 4-design behavior. It is worthwhile to point out that 2-design POVMs do not allow for exploiting purity at all. Matthews et al. [1, Section 2.C] the following bound showed

$$\lambda_{\mathcal{M}_{\mathrm{2D}}}|_{\rho,\sigma \text{ pure}} \leq \frac{1}{d+1}$$

without further assumptions on the 2-design POVM. Similar conclusions may be drawn if we relax the model assumption of purity to low rank. As the rank constraint r increases, the bounds on $\lambda_{\mathcal{M}_{C,z}}|_{\rho,\sigma \text{ rank } r}$ become gradually weaker until they approach (19) for $r = d$.

Finally, we point out that the rank(X)-parameter in Theorem 3 may be replaced by $\frac{\|X\|_1^2}{\|X\|_2^2}$—see (34) and (36) below. This ratio may be viewed as a robust measure for "effective rank". One particular scenario, where such a generalization is useful is the task of deciding whether a pure state $\rho = |\phi\rangle\langle\phi|$, or the maximally mixed state $\frac{1}{d}\mathbb{I}$ was prepared under the assumption of equiprobability ($\tau = \frac{1}{2}$). Lemma 1 below assures that $X = \frac{1}{2}\phi - \frac{1}{2d}\mathbb{I}$ has "effective rank" less than 4 and consequently

$$\beta_{\mathcal{M}_{C,z}}\left(\rho, \frac{1}{d}\mathbb{I}, \frac{1}{2}\right) \geq \frac{1}{12}\beta_{\mathrm{Helstrom}}\left(\rho, \frac{1}{d}\mathbb{I}, \frac{1}{2}\right)$$

for any Clifford orbit. This means that the optimal bias achievable with such a POVM measurement is directly comparable to Helstrom's optimal one.

III. PROOFS

A. Mathematical preliminaries

Throughout this work we will exclusively consider multi-qubit dimensions $d = 2^n$. Let $W_1, \ldots, W_{d^2} \in \mathcal{H}_d$ denote the d^2 Pauli operators and $w(\cdot)$ the associated characteristic function introduced in (8). Also, note that $d = 2^n$ assures that every W_k is actually a tensor product $W_k = \otimes\sigma_{k_1} \otimes \cdots \otimes \sigma_{k_n}$ of single qubit Pauli matrices $\sigma_0, \sigma_1, \sigma_2, \sigma_3 \in \mathcal{H}_2$.

We endow the vector spaces $\mathbb{C}^{d'}$ and $\mathbb{R}^{d'}$ with the usual ℓ_p-norms among which the ℓ_4-norm of the characteristic function (8) will be the most prominent:

$$\|w(\rho)\|_{\ell_4}^4 = \frac{1}{d^2}\sum_{k=1}^{d^2} \operatorname{tr}\left(W_k\rho\right)^4$$

On the level of hermitian matrices $X \in \mathcal{H}_d$, let $|X| = \sqrt{XX^\dagger}$ denote the matrix absolute value. We then define the Schatten-p-norms to be $\|X\|_p = \left(\operatorname{tr}\left(|X|^p\right)\right)^{1/p}$. These are related via $\|X\|_q \leq \|X\|_p$ for all $X \in \mathcal{H}_d$ and $p \leq q$. Moreover, the trace norm ($p = 1$) and the Hilbert-Schmidt norm ($p = 2$) obey the following converse relation: $\|X\|_1 \leq \sqrt{\operatorname{rank}(X)}\|X\|_2 \; \forall X \in \mathcal{H}_d$.

The main technical prerequisite for Theorem 3 is the following statement.

Theorem 6 (from [11]). *Fix $d = 2^n$ and let $\mathcal{M}_{C,z} = \{|x_k\rangle : \; x_k = C_k z, \; C_k \in C(d)\} \subseteq \mathbb{C}^d$ be a Clifford orbit with fiducial $z \in \mathbb{C}^d$ and N elements. Then*

$$\frac{1}{N}\sum_{k=1}^{N}\left(|x_k\rangle\langle x_k|\right)^{\otimes 4} = d\binom{d+2}{3}^{-1}\left(\alpha_1(z)P_1 + \alpha_2(z)P_2\right),$$

where P_1, P_2 are orthogonal projections obeying $P_1 + P_2 = P_{\mathrm{Sym}^4}$—the projector onto the totally symmetric subspace of $\mathcal{H}_d^{\otimes 4}$. Defining $Q = \frac{1}{d^2}\sum_{k=1}^{d^2} W_k^{\otimes 4}$ allows to characterize them explicitly by

$$P_1 = P_{\mathrm{Sym}^4}Q \quad \text{and} \quad P_2 = P_{\mathrm{Sym}^4}\left(\mathbb{I} - Q\right)$$

and the coefficients amount to

$$\alpha_1(z) = \|w(|z\rangle\langle z|)\|_{\ell_4}^4, \quad \alpha_2(z) = 4\frac{1 - \|w(|z\rangle\langle z|)\|_{\ell_4}^4}{(d+4)(d-1)}.$$

According to [11], the coefficient $\alpha_1(z)$ satisfies

$$\frac{2}{d(d+1)} \leq \alpha_1(z) \leq \frac{1}{d}, \tag{23}$$

which implies that

$$\begin{aligned}\frac{4}{d(d+4)} &\leq \alpha_2(z) \leq \frac{4(d+2)}{d(d+1)(d+4)},\\ -\frac{2}{d(d+1)} &\leq \alpha_1(z) - \alpha_2(z) \leq \frac{1}{d+4},\\ -\frac{d}{d+4} &\leq \frac{\alpha_1(z) - \alpha_2(z)}{\alpha_1(z)} \leq \frac{4}{d+4}.\end{aligned} \tag{24}$$

It is insightful to compare this statement to the defining property (3) of a complex projective 4-design:

$$\frac{1}{N}\sum_{k=1}^{n}(|x_k\rangle\langle x_k|)^{\otimes 4} = \int_{\|v\|_{\ell_2}=1} \mathrm{d}w\,(|w\rangle\langle w|)^{\otimes 4} \tag{25}$$
$$= \binom{d+3}{4}^{-1} P_{\mathrm{Sym}^4}.$$

The last equality is a consequence of Schur's Lemma well known in quantum information science—see e.g. [3, Lemma 1].

From such a comparison it becomes apparent that Clifford orbit fiducials $|z\rangle \in \mathbb{C}^d$ result in a complex projective 4-design, precisely if $\|w(|z\rangle\langle z|)\|_{\ell_4}^4 = \frac{4}{d(d+3)}$. Indeed, such a choice assures $\alpha_1(z) = \alpha_2(z) = \frac{4}{d(d+3)}$ for the constants occurring in Theorem 6 which in turn implies the defining property (25) of a 4-design.

However, Theorem 6 also implies that Clifford orbits in general do not have this very particular behavior and consequently fall short of being complex projective 4-designs. Fortunately, the deviation from this ideal behavior is benign: the fourth moment average decomposes into exactly two projectors P_1, P_2 instead of just P_{Sym^4}. As we shall see in the next subsection, this deviation is mild enough to adapt the proof technique from Theorem 1 by Ambainis and Emerson [4] (see also [1, Section 2.B] to Clifford orbits.

B. Proof of Theorem 1

At the heart of the proof of Theorem 1 (see [1, 4]) is the following moment inequality by Berger [12]:

$$\mathbb{E}[|S|] \geq \sqrt{\frac{\mathbb{E}[S^2]^3}{\mathbb{E}[S^4]}} \tag{26}$$

is true for any real valued random variable S.

Now, let $\mathcal{M}_{4D} = \left\{\frac{d}{N}|x_k\rangle\langle x_k|\right\}_{k=1}^{N}$ be a 4-design POVM, fix $X \in \mathcal{H}_d$ arbitrary and define the N-variate random variable

$$S_X = \langle x_k|X|x_k\rangle \quad \text{with probability} \quad \frac{1}{N}. \tag{27}$$

Accordingly,

$$\|X\|_{\mathcal{M}_{4d}} = \frac{d}{N}\sum_{k=1}^{N}|\langle x_k|X|x_k\rangle| = d\mathbb{E}[|S_X|] \geq d\sqrt{\frac{\mathbb{E}\left[S_X^2\right]^3}{\mathbb{E}\left[S_X^4\right]}}. \tag{28}$$

Accordingly, it suffices to bound the moments $\mathbb{E}\left[S_X^2\right]$, as well as $\mathbb{E}\left[S_X^4\right]$ appropriately. Since any complex projective 4-design in particular also constitutes a 2-design, the first quantity amounts to

$$\begin{aligned}
\mathbb{E}\left[S_X^2\right] =& \frac{1}{N}\sum_{k=1}^{N}\mathrm{tr}\left(|x_k\rangle\langle x_k|X\right)^2 \\
=& \mathrm{tr}\left(\frac{1}{N}\sum_{k=1}^{N}(|x_k\rangle\langle x_k|)^{\otimes 2}X^{\otimes 2}\right) \\
=& \binom{d+1}{2}^{-1}\mathrm{tr}\left(P_{\mathrm{Sym}^2}X^{\otimes 2}\right) \\
=& \frac{\mathrm{tr}\left(X^2\right)+\mathrm{tr}(X)^2}{(d+1)d},
\end{aligned} \tag{29}$$

where the last equation follows from $P_{\mathrm{Sym}^2} = \frac{1}{2}(\mathbb{I}+\mathbb{F})$ with $\mathbb{F}$ denoting the Flip-operator on a bi-partite system (see e.g. [13, Lemma 6], or [14, Lemma 17]).

For a corresponding bound of $\mathbb{E}\left[S^4\right]$ the 4-design property of the POVM is of crucial importance. Without requiring further assumptions (25) assures

$$\begin{aligned}
\mathbb{E}\left[S_X^4\right] =& \mathrm{tr}\left(\frac{1}{N}\sum_{k=1}^{N}(|x_k\rangle\langle x_k|)^{\otimes 4}X^{\otimes 4}\right) \\
=& \binom{d+3}{4}^{-1}\mathrm{tr}\left(P_{\mathrm{Sym}^4}X^{\otimes 4}\right) \\
\leq& \frac{10.1}{d(d+1)(d+2)(d+3)}\left[\mathrm{tr}\left(X^2\right)+\mathrm{tr}(X)^2\right] \\
=& \frac{10.1d(d+1)}{(d+2)(d+3)}\mathbb{E}\left[S^2\right]^2,
\end{aligned} \tag{30}$$

where the inequality follows from Lemma 2 in the appendix. Here we content ourselves to state that [14, Lemma 17] allows for evaluating $\mathrm{tr}\left(P_{\mathrm{Sym}^4}X^{\otimes 4}\right)$ explicitly without requiring X to have vanishing tace. Ambainis and Emerson [4], as well as Matthews, Wehner and Winter [1] made this assumption ($\mathrm{tr}(X) = 0$) to considerably simplify evaluating this fourth moment bound and obtain a slightly better constant ($\frac{1}{3}$ vs $\frac{1}{5}$ obtained here). Inserting these bounds into (28) reveals

$$\begin{aligned}
\|X\|_{\mathcal{M}_{4d}} \geq& d\sqrt{\frac{\mathbb{E}\left[S^2\right]^3}{\mathbb{E}\left[S^4\right]}} \\
=& \sqrt{\frac{(d+2)(d+3)}{(d+1)^2}\frac{(\|X\|_2^2+\mathrm{tr}(X)^2)^3}{24\|X\|_2^2\mathrm{tr}(P_{\mathrm{Sym}^4}X^{\otimes 4})}}\|X\|_2 \\
\geq& \sqrt{\frac{(\|X\|_2^2+\mathrm{tr}(X)^2)^3}{24\|X\|_2^2\mathrm{tr}(P_{\mathrm{Sym}^4}X^{\otimes 4})}}\|X\|_2 \\
\geq& \frac{1}{\sqrt{9.673}}\|X\|_2 > \frac{0.32}{\sqrt{\mathrm{rank}(X)}}\|X\|_1,
\end{aligned}$$

where the third inequality follows from Lemma 3 in the appendix. Since the choice of $X \in \mathcal{H}_d$ is arbitrary, (4) in Theorem 1 readily follows.

To derive (5) and (6) in Theorem 1, note that

$$\begin{aligned}\|X\|_{\mathcal{M}_{4d}} &\geq \sqrt{\frac{(\|X\|_2^2+\mathrm{tr}(X)^2)^3}{24\|X\|_1^2\mathrm{tr}(P_{\mathrm{Sym}^4}X^{\otimes 4})}}\|X\|_1 \\ &\geq \frac{1}{\sqrt{12.12}}\|X\|_1 > 0.287\|X\|_1,\end{aligned}$$

where the second inequality follows from Lemma 4 given that X has rank 2. This equation confirms (5). If in addition X is traceless, then

$$\frac{24\|X\|_1^2\mathrm{tr}\left(P_{\mathrm{Sym}^4}X^{\otimes 4}\right)}{[\|X\|_2^2+\mathrm{tr}(X)^2]^3} = 12. \tag{31}$$

according to Lemma 4 in the appendix, from which (6) follows.

C. Proof of Theorem 4 and Theorem 3

Now let us move on to prove Theorem 3—a similar statement for Clifford POVMs. Fix $d = 2^n$ and let $\mathcal{M}_{C,z} = \left\{\frac{d}{N}|x_k\rangle\langle x_k|\right\}_{k=1}^{N}$ be a Clifford orbit POVM with fiducial $|z\rangle \in \mathbb{C}^d$ (i.e. $|x_1\rangle = |z\rangle$, $|x_k\rangle = C_k|z\rangle, \ldots$ with $C_k \in C(d)$). We fix $X \in \mathcal{H}_d$ and define define the random variable S_X in analogy to (27). Doing so assures

$$\|X\|_{\mathcal{M}_{C,z}} = \mathbb{E}\left[|S_X|\right] \geq d\sqrt{\frac{\mathbb{E}\left[S_X^2\right]^3}{\mathbb{E}\left[S_X^4\right]}}$$

via Berger's inequality.

As already pointed out in section II, any Clifford orbit does constitute a complex projective 3-design. This in turn implies that (29) remains valid, because said derivation just requires a 2-design structure:

$$\mathbb{E}\left[S_X^2\right] = \frac{\|X\|_2^2+\mathrm{tr}(X)^2}{(d+1)d}.$$

However, deriving a corresponding bound for $\mathbb{E}\left[S_X^4\right]$ is considerably more challenging, because Clifford orbits in general fall short of being a complex projective 4-design. Instead, we restort to Theorem 6 which implies

$$\begin{aligned}\mathbb{E}\left[S_X^4\right] &= \mathrm{tr}\left(\frac{1}{N}\left(|x_k\rangle\langle x_k|\right)^{\otimes 4}X^{\otimes 4}\right) \\ &= d\binom{d+2}{3}^{-1}\left(\alpha_1(z)\mathrm{tr}\left(P_1X^{\otimes 4}\right)+\alpha_2\mathrm{tr}\left(P_2X^{\otimes 2}\right)\right),\end{aligned} \tag{32}$$

where $P_1, P_2 \in \mathcal{H}_d^{\otimes 4}$ and $\alpha_1(z), \alpha_2(z)$ were introduced in said theorem.

We bound the two occurring terms individually. For the first term, we obtain

$$\begin{aligned}\mathrm{tr}(P_1X^{\otimes 4}) =&\mathrm{tr}\left(P_1\left|X^{\otimes 4}\right|\right) = \mathrm{tr}\left(P_{\mathrm{Sym}}Q|X|^{\otimes 4}\right) \\ \leq&\mathrm{tr}\left(Q|X|^{\otimes 4}\right) = \frac{1}{d^2}\sum_{k=1}^{d^2}\mathrm{tr}\left(W_k^{\otimes 4}|X|^{\otimes 4}\right) \\ =&\frac{1}{d^2}\sum_{k=1}^{d^2}\mathrm{tr}\left(W_k|X|\right)^4\end{aligned}$$

by invoking some standard trace inequalities. Hoelder's inequality together with the fact that the characteristic function (8) is an isometry allows us to simplify further:

$$\begin{aligned}\mathrm{tr}(P_1X^{\otimes 4}) \leq&\frac{1}{d^2}\sum_{k=1}^{d^2}\mathrm{tr}\left(W_k|X|\right)^4 \\ \leq&\frac{1}{d^2}\sum_{k=1}^{d^2}\|X\|_1^2\|W_K\|_\infty^2\mathrm{tr}\left(W_k|X|\right)^2 \\ =&\frac{\|X\|_1^2}{d}\|w(|X|)\|_{\ell_2}^2 = \frac{\|X\|_1^2\|X\|_2^2}{d}.\end{aligned} \tag{33}$$

The last equation is due to the fact that the Schatten-p norms of X and $|X|$ coincide by definition. Together with (24) and (32), this equation implies

$$\begin{aligned}\mathbb{E}\left[S_X^4\right] &= \mathrm{tr}\left(\frac{1}{N}\left(|x_k\rangle\langle x_k|\right)^{\otimes 4}X^{\otimes 4}\right) = d\binom{d+2}{3}^{-1}\left((\alpha_1(z)-\alpha_2(z))\mathrm{tr}\left(P_1X^{\otimes 4}\right)+\alpha_2\mathrm{tr}\left(P_{\mathrm{Sym}^4}X^{\otimes 4}\right)\right), \\ &\leq \binom{d+2}{3}^{-1}|\alpha_1(z)-\alpha_2(z)|\|X\|_1^2\|X\|_2^2 + \frac{24}{d(d+1)^2(d+4)}\mathrm{tr}\left(P_{\mathrm{Sym}^4}X^{\otimes 4}\right).\end{aligned}$$

Consequently,

$$\|X\|_{\mathcal{M}_{C,z}} \geq d\sqrt{\frac{\mathbb{E}\left[S_X^2\right]^3}{\mathbb{E}\left[S_X^4\right]}} \geq \frac{\|X\|_2}{\sqrt{\kappa(X,z)}} \geq \frac{\|X\|_1}{\sqrt{\kappa(X,z)\,\mathrm{rank}(X)}}, \tag{34}$$

where

$$\begin{aligned}\kappa(X,z) =&\frac{\frac{6(d+1)^2}{d+2}|\alpha_1(z)-\alpha_2(z)|\|X\|_1^2\|X\|_2^4+\frac{24(d+1)}{d+4}\|X\|_2^2\mathrm{tr}\left(P_{\mathrm{Sym}^4}X^{\otimes 4}\right)}{[\|X\|_2^2+\mathrm{tr}(X)^2]^3}\\ \leq&\frac{\frac{6d(d+1)^2}{(d+2)(d+4)}\alpha_1(z)\|X\|_1^2\|X\|_2^4+\frac{24(d+1)}{d+4}\|X\|_2^2\mathrm{tr}\left(P_{\mathrm{Sym}^4}X^{\otimes 4}\right)}{[\|X\|_2^2+\mathrm{tr}(X)^2]^3} \leq \frac{d+1}{d+4}\left(6d\alpha_1(z)\frac{\|X\|_1^2}{\|X\|_2^2}+9.673\right)\\ \leq& 6d\alpha_1(z)\frac{\|X\|_1^2}{\|X\|_2^2}+10 \leq 6d\alpha_1(z)\,\mathrm{rank}(X)+10 = 6d\|w(|z\rangle\langle z|)\|_{\ell_4}^4\,\mathrm{rank}(X)+10.\end{aligned} \tag{35}$$

Here the second inequality in (36) follows from Lemma 3 in the appendix. The above two equations confirm Theorem 3. If X is traceless, the bound on $K(X,z)$ can be improved slightly,

$$\kappa(X,z) =\leq \frac{d+1}{d+4}\left(6d\alpha_1(z)\frac{\|X\|_1^2}{\|X\|_2^2}+9\right) \leq 6d\alpha_1(z)\frac{\|X\|_1^2}{\|X\|_2^2}+9 \leq 6d\alpha_1(z)\,\mathrm{rank}(X)+9. \tag{36}$$

If $\|w(|z\rangle\langle z|)\|_{\ell_4}^4 = \alpha_1(z) \leq 6/(d(d+3))$, then

$$-\frac{2}{d(d+1)} \leq \alpha_1(z)-\alpha_2(z) \leq \frac{2}{(d-1)(d+4)}. \tag{37}$$

Therefore,

$$\kappa(X,z) \leq \frac{\frac{12\|X\|_1^2}{d\|X\|_2^2}\|X\|_2^6+24\|X\|_2^2\mathrm{tr}\left(P_{\mathrm{Sym}^4}X^{\otimes 4}\right)}{[\|X\|_2^2+\mathrm{tr}(X)^2]^3} \leq \frac{12\|X\|_1^2}{d\|X\|_2^2}+10 \leq \frac{12\,\mathrm{rank}(X)}{d}+10 \leq 22,$$

from which Theorem 4 follows. If X is traceless, the bound on $K(X,z)$ can be improved slightly,

$$\kappa(X,z) \leq \frac{12\|X\|_1^2}{d\|X\|_2^2}+9 \leq \frac{12\,\mathrm{rank}(X)}{d}+9 \leq 21.$$

D. Proof of (21) and (22)

If z is a stabilizer state, then $d\alpha_1(z) = 1$, so that

$$\begin{aligned}\kappa(X,z) &\leq \frac{d+1}{d+4}(6d+10) \leq 6d,\\ \|X\|_{\mathcal{M}_{C,z}} &\geq \frac{\|X\|_1}{\sqrt{\kappa(X,z)\,\mathrm{rank}(X)}} \geq \frac{\|X\|_1}{\sqrt{6d}}.\end{aligned}$$

This equation confirms the lower bound in (21); the upper bound follows from (17) as mentioned before.

According to the same reasoning that leads to (34),

$$\|X\|_{\mathcal{M}_{C,z}} \geq d\sqrt{\frac{\mathbb{E}\left[S_X^2\right]^3}{\mathbb{E}\left[S_X^4\right]}} \geq \frac{\|X\|_1}{\sqrt{\mu(X,z)}}$$

where

$$\mu(X,z) = \frac{\frac{6(d+1)^2}{d+2}|\alpha_1(z)-\alpha_2(z)|\|X\|_1^4\|X\|_2^2 + \frac{24(d+1)}{d+4}\|X\|_1^2\mathrm{tr}\left(P_{\mathrm{Sym}^4}X^{\otimes 4}\right)}{[\|X\|_2^2+\mathrm{tr}(X)^2]^3} \leq \frac{6\|X\|_1^2\|X\|_2^4 + 24\|X\|_1^2\mathrm{tr}\left(P_{\mathrm{Sym}^4}X^{\otimes 4}\right)}{[\|X\|_2^2+\mathrm{tr}(X)^2]^3}$$
$$\leq 36.$$

Here the second inequality follows from Lemma 5 in the appendix given that X has rank 2. As an immediate consequence, $\|X\|_{\mathcal{M}_{C,z}} \geq \|X\|_1/6$, from which (22) follows.

E. Proof of Theorem 5

At the heart of this proof is the fact that by definition the multi-qubit Clifford group is the normalizer of the Pauli group $P(d) = \{\pm W_k, \pm i W_k\}_{k=1}^{d^2}$ and it acts transitively on Pauli operators up to overall phase factors. This fact in particular implies that

$$\begin{aligned}\|W\|_{\mathcal{M}_{C,z}} =& \frac{d}{|C(d)|}\sum_{j=1}^{|C(d)|}|\langle C_j z|W|C_j z\rangle| \\ =& \frac{d}{|C(d)|}\sum_{j=1}^{|C(d)|}\left|\langle z|C_j^\dagger W C_j|z\rangle\right| \\ =& \frac{d}{d^2-1}\sum_{k=2}^{d^2}|\langle z|W_k|z\rangle|\,.\end{aligned}$$

Using $\langle z|W_1|z\rangle = \langle z|z\rangle = 1$ and the definition (8) of the characteristic function this expression amounts to

$$\begin{aligned}\|W\|_{\mathcal{M}_{C,z}} =& \frac{d(\sum_{k=1}^{d^2}|\mathrm{tr}\left(W|z\rangle\langle z|\right)|-1)}{d^2-1} \\ =& \frac{d\left(\sqrt{d}\|w(|z\rangle\langle z|)\|_{\ell_1}-1\right)}{d^2-1} \\ =& \frac{\sqrt{d}\|w(|z\rangle\langle z|)\|_{\ell_1}-1}{(d+1)(d-1)}\|W\|_1,\end{aligned}$$

because $\|W\|_1 = d$ for any Pauli matrix.

F. Proof of the certainty relation for stabilizer bases

Our derivation closely resembles a similar analysis for 2-designs presented in [1]. Fix $d = 2^n$ and let $\mathcal{B}_{\mathrm{stab}}^{(k)} = \left\{b_1^{(k)}, \ldots, b_d^{(k)}\right\}$ denote the $M = \prod_{j=1}^{n}\left(d^j+1\right)$ different multi-qubit stabilizer bases. Note that this implies that there are $N = dM$ different stabilizer states in total. Now, we fix $\phi = |\phi\rangle\langle\phi|$ and apply Jensen's inequality to conclude

$$\begin{aligned}\frac{1}{M}\sum_{k=1}^{M}S_2\left(\mathcal{B}_{\mathrm{stab}}^{(k)}(\phi)\right) =& \frac{1}{M}\sum_{k=1}^{M}-\log\left(\sum_{j=1}^{d}\left|\langle b_j^{(k)}|\phi\rangle\right|^4\right) \\ \leq& -\log\left(\frac{1}{M}\sum_{k=1}^{M}\sum_{j=1}^{d}\left|\langle b_j^{(k)}|\phi\rangle\right|^4\right) \\ =& -\log\left(\frac{d}{N}\sum_{k=1}^{N}|\langle x_j|\phi\rangle|^4\right)\end{aligned}$$

G. Characteristic function of different fiducial vectors

The characteristic functions of stabilizer states is well-known. Nonetheless, we shall also derive them here for the sake completeness. In qubit dimensions $d = 2^n$, every stabilizer state $|z\rangle \in \mathbb{C}^d$ is defined to be the unique common eigenvector of d commuting elements of the Pauli group $P(d) = \{\pm W_k, \pm i W_k\}_{k=1}^{d}$ that must not contain $-\mathbb{I}$. This in turn implies (see e.g. [15, Exercise 10.34])

$$|z\rangle\langle z| = \frac{1}{d}\sum_{k\in S}\phi_k W_k \quad \phi_k \in \{\pm 1\}\,.$$

Here $S \subset \left\{1,\ldots,d^2\right\}$ denotes a subset of cardinality $|S| = d$. Mutual orthogonality of the Pauli matrices with respect to the Hilbert-Schmidt inner product then implies

$$\begin{aligned}w(|z\rangle\langle z|) =& \frac{1}{\sqrt{d}}\sum_{j=1}^{d^2}\mathrm{tr}\left(W_j\frac{1}{d}\sum_{k\in S}\phi_k W_k\right)|e_j\rangle \\ =& d^{-\frac{3}{2}}\sum_{j=1}^{d^2}\sum_{k\in S}\phi_k\mathrm{tr}\left(W_k W_j\right)|e_j\rangle \\ =& \sum_{k\in S}\frac{\phi_k}{\sqrt{d}}|e_k\rangle.\end{aligned}$$

Accordingly,

$$\|w(|z\rangle\langle z|)\|_{\ell_p}^p = \sum_{k\in S}\left|\frac{\phi_k}{\sqrt{d}}\right|^p = d^{1-\frac{p}{2}}$$

for any $1 \leq p < \infty$.

Let us now turn our attention to the characteristic function of the "magic product state" $|z\rangle\langle z| = \rho^{\otimes n} \in$

$\mathcal{H}_{2^n}$ with $\rho = \frac{1}{2}\left(\sigma + \frac{1}{\sqrt{3}}\left(\sigma_1 + \sigma_2 + \sigma_3\right)\right) \in \mathcal{H}_d$. Here $\sigma_0, \ldots, \sigma_3 \in \mathcal{H}_2$ denote the single qubit Pauli matrices with the convention $\sigma_0 = \mathbb{I}$. We will content ourselves with directly computing ℓ_p norms of the characteristic function. To this end, we use the fact that every $d = 2^n$-dimensional Pauli matrix admits a tensor product decomposition

$$W_k = \sigma_{k_1} \otimes \cdots \otimes \sigma_{k_n} \quad k_j \in \{0,1,2,3\}$$

into single qubit Pauli's. Doing so implies

$$\begin{aligned}
\|w(\rho^{\otimes n})\|_{\ell_p}^p =& d^{-\frac{p}{2}} \sum_{k_1,\ldots,k_n=0}^{3} \left|\operatorname{tr}\left(W_{k_1} \otimes \cdots \otimes W_{k_n} \rho^{\otimes n}\right)\right|^p \\
=& d^{-\frac{p}{2}} \sum_{k_1,\ldots,k_n=0}^{3} \left|\operatorname{tr}\left(W_{k_1}\rho\right) \cdots \operatorname{tr}\left(W_{k_n}\rho\right)\right|^p \\
=& d^{-\frac{p}{2}} \prod_{j=1}^{n} \sum_{k_j=0}^{3} \left|\operatorname{tr}\left(W_{k_j}\rho\right)\right|^p \\
=& d^{-\frac{p}{2}} \prod_{j=1}^{n} \left(1 + 3\left(\frac{1}{\sqrt{3}}\right)^p\right) \\
=& d^{-\frac{p}{2}} \left(1 + 3\left(\frac{1}{\sqrt{3}}\right)^p\right)^n,
\end{aligned}$$

where the last line is due to "magic state's" particular structure. For $p = 4$ we thus obtain

$$\begin{aligned}
\|w(|z\rangle\langle z|)\|_{\ell_4}^4 =& \frac{1}{d^2}\left(1 + \frac{3}{9}\right)^n = \frac{4^n}{2^{2n}3^n} \\
=& \frac{1}{3^n} = \left(\frac{1}{9}\right)^{\frac{n}{2}} < \left(\frac{1}{8}\right)^{\frac{n}{2}} = d^{-\frac{3}{2}}. \qquad (38)
\end{aligned}$$

Similarly:

$$\sqrt{d}\|w(|z\rangle\langle z|)\|_{\ell_1} = \left(1 + \sqrt{3}\right)^n < d^{1.45}. \qquad (39)$$

H. Entropic uncertainty and certainty relations for stabilizer bases

Lemma 1. *Let $\rho \in \mathcal{S}_d$ be quantum state with* $\operatorname{rank}(\rho) = r$. *Then the "effective rank" of $X = \rho - \frac{1}{d}\mathbb{I}$ amounts to*

$$\begin{aligned}
r_{\text{eff}}(X) :=& \frac{\|X\|_1^2}{\|X\|_2^2} \leq \frac{4\operatorname{rank}(\rho)(d - \operatorname{rank}(\rho))}{d} \\
\leq& 4\max\{r, d - r\}.
\end{aligned}$$

The first bound is saturated by quantum states ρ that are maximally mixed on an r-dimensional subspace, while the second bound is saturated, if ρ is pure.

We provide a proof of this elementary statement in the appendix.

[1] W. Matthews, S. Wehner, and A. Winter, "Distinguishability of quantum states under restricted families of measurements with an application to quantum data hiding," *Communications in Mathematical Physics*, vol. 291, no. 3, pp. 813–843, 2009.

[2] C. W. Helstrom, "Quantum detection and estimation theory," *Journal of Statistical Physics*, vol. 1, no. 2, pp. 231–252, 1969.

[3] A. J. Scott, "Tight informationally complete quantum measurements," *Journal of Physics A: Mathematical and General*, vol. 39, no. 43, p. 13507, 2006.

[4] A. Ambainis and J. Emerson, "Quantum t-designs: t-wise independence in the quantum world," in *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*, pp. 129–140, June 2007.

[5] R. Kueng, D. Gross, and F. Krahmer, "Spherical designs as a tool for derandomization: The case of phaselift," in *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pp. 192–196, May 2015.

[6] H. Zhu, "Multiqubit clifford groups are unitary 3-designs," *arXiv preprint arXiv:1510.02619*, 2015.

[7] Z. Webb, "The clifford group forms a unitary 3-design," *arXiv preprint arXiv:1510.02769*, 2015.

[8] C. Dankert, R. Cleve, J. Emerson, and E. Livine, "Exact and approximate unitary 2-designs and their application to fidelity estimation," *Phys. Rev. A*, vol. 80, p. 012304, Jul 2009.

[9] D. Gross, K. Audenaert, and J. Eisert, "Evenly distributed unitaries: On the structure of unitary designs," *Journal of Mathematical Physics*, vol. 48, no. 5, 2007.

[10] R. Kueng and D. Gross, "Qubit stabilizer states are complex projective 3-designs," *arXiv preprint arXiv:1510.02767*, 2015.

[11] M. Grassl, D. Gross, R. Kueng, and H. Zhu, "The clifford group fails gracefully to be a unitary 4-design," *to appear*, 2016.

[12] B. Berger, "The fourth moment method," *SIAM Journal on Computing*, vol. 26, no. 4, pp. 1188–1207, 1997.

[13] D. Gross, F. Krahmer, and R. Kueng, "A partial derandomization of phaselift using spherical designs," *Journal of Fourier Analysis and Applications*, vol. 21, no. 2, pp. 229–266, 2015.

[14] R. Kueng, H. Rauhut, and U. Terstiege, “Low rank matrix recovery from rank one measurements,” *Appl. Comput. Harmonic Anal.*, to appear. DOI:10.1016/j.acha.2015.07.007.

[15] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information 10th Anniversary Edition*. Cambridge University Press, 2010.

[16] Choose P_0 to be the projector onto the non-negative range of $\tau\rho - (1-\tau)\sigma$ and $P_1 = \mathbb{I} - P_0$.

[17] Observing $k \in \{1, \dots, N\}$, we guess ρ if $p_\rho(k) \geq p_\sigma(k)$ and σ otherwise.

[18] Every Pauli matrix W has vanishing trace and is therefore proportional to a particular difference $\tau\rho - (1-\tau)\sigma$ of quantum states $\rho, \sigma \in \mathcal{S}_d$ with $\tau = \frac{1}{2}$.

Appendix

I. Derivation of the 4-design bound presented in (4)

Previous derivations [1, 4] of the fourth moment bound presented in (4) have assumed X to be traceless. This additional assumption considerably simplifies the task at hand. Here, we prove a similar bound valid for arbitrary $X \in \mathcal{H}_d$ at the cost of a slightly larger multiplicative constant. At the heart of this derivation is [14, Lemma 17] which provides a closed-form expression for the object at hand:

Lemma 2. *Suppose X is a nonzero Hermitian operator and $y = |\mathrm{tr}(X)|/\|X\|_2$. Then*

$$\frac{24\mathrm{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right)}{(\mathrm{tr}(X^2)+\mathrm{tr}(X)^2)^2} \leq 3 + \frac{6+8y-2y^4}{(1+y^2)^2} \leq \frac{3}{5}(7 + 4\cdot 2^{1/3} + 3\cdot 2^{2/3}) \approx 10.08113. \tag{40}$$

Here the second inequality is saturated iff $y = 2^{1/3}-1$; the first one cannot be saturated except when $y=1$ and X has rank 1, but it can be approached with arbitrarily small gap.

When X is trace less, Lemma 2 implies that

$$\frac{24\mathrm{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right)}{(\mathrm{tr}(X^2)+\mathrm{tr}(X)^2)^2} < 9, \tag{41}$$

where the upper bound can be approached with arbitrarily small gap.

Proof. According to [14, Lemma 17],

$$\begin{aligned} 24\mathrm{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right) &= \left(\mathrm{tr}(X)^4 + 8\mathrm{tr}(X)\mathrm{tr}(X^3) + 3\mathrm{tr}(X^2)^2 + 6\mathrm{tr}(X)^2\mathrm{tr}(X^2) + 6\mathrm{tr}(X^4)\right) \\ &= 3\left(\mathrm{tr}(X^2)+\mathrm{tr}(X)^2\right)^2 + 8\mathrm{tr}(X)\mathrm{tr}(X^3) + 6\mathrm{tr}(X^4) - 2\mathrm{tr}(X)^4 \\ &\leq 3\left(\|X\|_2^2+\mathrm{tr}(X)^2\right)^2 + 8|\mathrm{tr}(X)|\|X\|_3^3 + 6\|X\|_4^4 - 2\mathrm{tr}(X)^4 \\ &\leq 3\left(\|X\|_2^2+\mathrm{tr}(X)^2\right)^2 + 8|\mathrm{tr}(X)|\|X\|_2^3 + 6\|X\|_2^4 - 2\mathrm{tr}(X)^4, \end{aligned} \tag{42}$$

where the first inequality is saturated iff $X \geq 0$ or $X \leq 0$, and the second one is saturated iff $\|X\|_4 = \|X\|_3 = \|X\|_2$, that is, X has rank 1. Consequently,

$$\begin{aligned} \frac{24\mathrm{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right)}{(\mathrm{tr}(X^2)+\mathrm{tr}(X)^2)^2} &\leq 3 + \frac{8|\mathrm{tr}(X)|\|X\|_2^3 + 6\|X\|_2^4 - 2\mathrm{tr}(X)^4}{\left(\|X\|_2^2+\mathrm{tr}(X)^2\right)^2} = f(y) := 3 + \frac{6+8y-2y^4}{(1+y^2)^2} \\ &\leq \frac{3}{5}(7+4\cdot 2^{1/3}+3\cdot 2^{2/3}) \approx 10.08113. \end{aligned} \tag{43}$$

Here the first inequality is saturated iff X has rank 1 (in which case $y=1$). To derive the second inequality, note that

$$f'(y) = \frac{8(1-3y-3y^2-y^3)}{(1+y^2)^3}, \tag{44}$$

which is positive when $0 \leq y < 2^{1/3}-1$ and negative when $y > 2^{1/3}-1$. So the maximum of $f(y)$ for $y \geq 0$ is attained when $y = 2^{1/3}-1$, in which case

$$f(2^{1/3}-1) = \frac{3}{5}(7+4\cdot 2^{1/3}+3\cdot 2^{2/3}). \tag{45}$$

Although the first inequality in 41 can not be saturated except when $y = 1$, the bound can be approached arbitrarily close if we do not impose any restriction on the rank of X. To show this point, suppose $X = \mathrm{diag}(ak, -1, -1, \ldots, -1)$ has rank $k+1$, where a is a real constant to be determined later. Then

$$\mathrm{tr}(X) = k(a-1), \quad \|X\|_2^2 = a^2k^2 + k, \quad \mathrm{tr}(X^3) = a^3k^3 - k \quad \|X\|_4^4 = a^4k^4 + k. \tag{46}$$

Assuming $y \geq 0, y \neq 1, k \geq y^2$, and let

$$a = \frac{k + \sqrt{ky^2(1+k-y^2)}}{k(1-y^2)}. \tag{47}$$

Then $\mathrm{tr}(X)\mathrm{tr}(X^3) \geq 0$, $|\mathrm{tr}(X)|/\|X\|_2 = y$,

$$\lim_{k\to\infty} a = \frac{1}{1-y}, \quad \lim_{k\to\infty} \frac{|\mathrm{tr}(X^3)|}{\|X\|_2^3} = 1, \quad \lim_{k\to\infty} \frac{\|X\|_4}{\|X\|_2} = 1, \tag{48}$$

which implies that

$$\lim_{k\to\infty} \frac{24\mathrm{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right)}{(\mathrm{tr}(X^2) + \mathrm{tr}(X)^2)^2} = 3 + \frac{6 + 8y - 2y^4}{(1+y^2)^2}. \tag{49}$$

□

Lemma 3. *Suppose X is a nonzero Hermitian operator and $y = |\mathrm{tr}(X)|/\|X\|_2$. Then*

$$\frac{24\mathrm{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right)\mathrm{tr}(X^2)}{(\mathrm{tr}(X^2) + \mathrm{tr}(X)^2)^2} \leq \frac{3(1+y^2)^2 + 6 + 8y - 2y^4}{(1+y^2)^3} < 9.673. \tag{50}$$

Here the first inequality cannot be saturated except when $y = 1$ and X has rank 1, but it can be approached with arbitrarily small gap.

Proof. The lemma follows from Lemma 2 except for the second inequality in Equation 50. To derive this inequality, let

$$f(y) = \frac{3(1+y^2)^2 + 6 + 8y - 2y^4}{(1+y^2)^3}; \tag{51}$$

then

$$f'(y) = -\frac{2(-4 + 21y + 20y^2 + 10y^3 + y^5)}{(1+y^2)^4}. \tag{52}$$

Note that $(1+y^2)^4 f'(y)$ is monotonic decreasing with y when $y \geq 0$ and has a unique real root $y_0 > 0$. Therefore, the maximum of $f(y)$ is attained when $y = y_0$. Now it is straightforward to verify that $f(y_0) < 9.673$. Calculation shows that

$$y_0 \approx 0.163078, \quad f(y_0) \approx 9.67249. \tag{53}$$

□

Lemma 4. *Suppose X is a rank-2 Hermitian operator. Then*

$$\frac{24\|X\|_1^2 \mathrm{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right)}{[\|X\|_2^2 + \mathrm{tr}(X)^2]^3} \leq \frac{5}{81}(95 + 32\sqrt{10}) \approx 12.1107. \tag{54}$$

If X is in addition traceless, then

$$\frac{24\|X\|_1^2 \mathrm{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right)}{[\|X\|_2^2 + \mathrm{tr}(X)^2]^3} = 12. \tag{55}$$

Proof. Note that the left hand side of (60) is invariant when X is multiplied by any nonzero real constant. Without loss of generality, we may assume that the two nonzero eigenvalues of X are equal to $1, x$ with $-1 \leq x \leq 1$. Then

$$\|X\|_1 = 1 + |x|, \quad \|X\|_2 = 1 + x^2, \quad \operatorname{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right) = 1 + x + x^2 + x^3 + x^4, \tag{56}$$

so that

$$\frac{24\|X\|_1^2 \operatorname{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right)}{[\|X\|_2^2 + \operatorname{tr}(X)^2]^3} = f(x) := \frac{3(1+|x|)^2(1+x+x^2+x^3+x^4)}{(1+x+x^2)^3}. \tag{57}$$

If $x \geq 0$, then $f(x) \leq 3$ according to the following equation,

$$(1+|x|)^2(1+x+x^2+x^3+x^4) - (1+x+x^2)^3 = -x^2(2+3x+2x^2) \leq 0. \tag{58}$$

If $-1 \leq x < 0$, then

$$f(x) := \frac{3(1-x)^2(1+x+x^2+x^3+x^4)}{(1+x+x^2)^3}, \quad f'(x) = \frac{3(-1+x)(1+x)(4+4x-x^2+4x^3+4x^4)}{(1+x+x^2)^4}.$$

Let x_0 be the unique real root of $4 + 4x - x^2 + 4x^3 + 4x^4$ which lies between -1 and 0, then $f'(x) \geq 0$ if $-1 \leq x \leq x_0$ and $f'(x) \leq 0$ if $x_0 \leq x \leq 0$. Therefore, the maximum of $f(x)$ is attained when $x = x_0$, in which case

$$f(x_0) = \frac{5}{81}(95 + 32\sqrt{10}). \tag{59}$$

If X is in addition traceless, then $x = -1$, so (55) follows from Equation 57. □

Lemma 5. *Suppose X is a rank-2 Hermitian operator. Then*

$$\frac{6\|X\|_1^4\|X\|_2^2 + 24\|X\|_1^2 \operatorname{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right)}{[\|X\|_2^2 + \operatorname{tr}(X)^2]^3} \leq 36, \tag{60}$$

where the upper bound is saturated iff X is traceless.

Proof. As in the proof of Lemma 4, we may assume that the two nonzero eigenvalues of X are equal to $1, x$ with $-1 \leq x \leq 1$. Then

$$\frac{6\|X\|_1^4\|X\|_2^2 + 24\|X\|_1^2 \operatorname{tr}\left(P_{\mathrm{Sym}^4} X^{\otimes 4}\right)}{[\|X\|_2^2 + \operatorname{tr}(X)^2]^3} = f(x) := \frac{6(1+|x|)^4(1+x^2) + 24(1+|x|)^2(1+x+x^2+x^3+x^4)}{8(1+x+x^2)^3}.$$

When $x \geq 0$, it is straightforward to verify that $f(x) \leq 9$. When $-1 \leq x < 0$,

$$f(x) = \frac{6(1-x)^4(1+x^2) + 24(1+x+x^2+x^3+x^4)}{8(1+x+x^2)^3} = \frac{3(1-x)^2(5+2x+6x^2+2x^3+5x^4)}{4(1+x+x^2)^3},$$

whose derivative is given by

$$f'(x) = \frac{3(-23+9x^2-9x^4+23x^6)}{4(1+x+x^2)^4} \leq 0,$$

Therefore, $f(x) \leq f(-1) = 36$, and the upper boud is saturated iff $x = -1$, in which case X is traceless. □

J. Proof of Lemma 1

We start by computing the Hilbert-Schmidt norm of $X = \frac{\rho}{-}\frac{1}{d}\mathbb{I}$:

$$\|X\|_2^2 = \operatorname{tr}\left(\rho^2\right) + \frac{1}{d^2}\operatorname{tr}(\mathbb{I}) = \operatorname{tr}\left(\rho^2\right) - \frac{1}{d}.$$

Recall that the minimal purity of any rank-r state is $\frac{1}{r}$ which in turn impies

$$\|X\|_2^2 \geq \frac{d-r}{dr}. \tag{61}$$

For computing the trace norm, we employ an eigenvalue decomposition $\rho = \sum_{k=1}^r \lambda_k |k\rangle\langle k|$ of ρ and in turn write $\mathbb{I} = \sum_{k=1}^d |k\rangle\langle k|$. Consequently

$$\begin{aligned} \|X\|_1 &= \sum_{k=1}^r \left|\lambda_k - \frac{1}{d}\right| + \sum_{k=r+1}^d \frac{1}{d} \\ &\leq \sqrt{r \sum_{k=1}^r \left(\lambda_k - \frac{1}{d}\right)^2} + \frac{d+r}{d}, \end{aligned}$$

because $\|x\|_{\ell_1} \leq \sqrt{r}\|x\|_{\ell_2}$ for any $x \in \mathbb{C}^r$. Applying $\sum_{k=1}^r \lambda_k^2 = \mathrm{tr}(\rho^2)$, $\sum_{k=1}^r \lambda_k = \mathrm{tr}(\rho) = 1$ and resorting to (61) we obtain

$$\begin{aligned} \|X\|_1 \leq& \sqrt{r \sum_{k=1}^r \left(\lambda_k - \frac{1}{d}\right)^2} + \frac{d+r}{d} \\ =& \sqrt{r \left(\mathrm{tr}(\rho^2) - \frac{1}{d} - \frac{d-r}{d}\right)} + \frac{d-r}{d^2} \\ =& \sqrt{r \left(\|X\|_2^2 - \frac{r}{d}\frac{d-r}{rd}\right)} + \sqrt{r}\sqrt{\frac{d-r}{d}}\sqrt{\frac{d-r}{dr}} \\ \leq& \sqrt{r\left(1 - \frac{r}{d}\right)}\, \|X\|_2^2 + \sqrt{r\frac{d-r}{d}}\|X\|_2 \\ =& 2\sqrt{r\frac{d-r}{d}}\|X\|_2. \end{aligned}$$

Combining these two relations implies

$$r_{\mathrm{eff}}(X) = \frac{\|X\|_1^2}{\|X\|_2^2} = \frac{4r(d-r)}{d},$$

as claimed. The second bound follows from the fact that $\max\{r, d-r\} \leq \frac{d-1}{d} \leq d-1$ for any $1 \leq r \leq d-1$ (the case $r = d$ is trivial, because it implies $X = 0$). Consequently:

$$\begin{aligned} \frac{4r(d-r)}{d} =& \frac{4}{d}\max\{r, d-r\}\min\{r, d-r\} \\ \leq& 4\frac{d-1}{d}\min\{r, d-r\}. \end{aligned}$$

The fact that both bounds are saturated, follows from a straightforward computation for $\rho = \sum_{k=1}^r |k\rangle\langle k|$ (first bound) and then setting $r = 1$ and $r = d-1$, respectively (second bound).

4 Conclusion and Outlook

4.1 Summary

Convex signal reconstruction combines techniques from linear algebra, convex optimization and probability theory. The aim is to solve ill-posed inverse problems via convex optimization. In many instances, rigorous mathematical performance guarantees can be obtained for such procedures.

The most prominent examples are *compressed sensing* and *low rank matrix reconstruction*. Strong reconstruction guarantees typically require "generic cases", for example situations where the measurements correspond to random Gaussian vectors and matrices, respectively. See for instance [BDDW08; RFP10]. In addition, proofs of convergence can be obtained for more restricted sets of measurements if the obey particular properties, such as *incoherence* and *isotropy*. Examples include discrete Fourier vectors in compressed sensing [CRT06] and Pauli matrices in matrix reconstruction [Gro11; Liu11].

The main objective of this thesis was to devise novel proof techniques that are able to handle further structural restrictions on the measurement process.

An important special case, where this is necessary, is *phase retrieval*. This is the task of reconstructing a complex vector $x \in \mathbb{C}^n$ from quadratic measurements that are ignorant towards phase information. This problem is ubiquitous in many scientific disciplines, including X-ray crystallography, astronomy and quantum mechanics. As pointed out by Candès *et al.* [CESV15], this quadratic inverse problem can be re-cast as a particular instance of low rank matrix reconstruction: Both, the signal and the measurements are proportional to rank-one projectors. Measurements of this type fail to obey *incoherence* and *isotropy*. These issues may be overcome for Gaussian measurement vectors, or vectors chosen uniformly from the complex unit sphere S^{n-1}. It turns out that strict isotropy is not a necessary requirement, and such "generic instances" guarantee a strong notion of probabilistic incoherence. This in turn allowed Candès and Li to prove phaseless reconstruction guarantees that scale linearly in the dimension n [CL14].

In order to partially derandomize this result, Gross, myself and Krahmer have resorted to spherical t-designs [GKK15a; KGK15]. These amount to finite configurations of vectors that are "evenly distributed" in the sense that they reproduce the first $2t$ moments of the uniform distribution over S^{n-1}. In turn, we could relate a relaxed notion of isotropy to the defining property of a spherical 2-design [GKK15a]. The structure of a spherical 2-design alone, however, is insufficient for strong constructive results. Complementing this no-go result, we could prove that spherical 4-designs already enable optimal reconstruction [KRT15]. This leaves the case $t = 3$ as an intriguing open case. In [KG15], we could identify a particular instance of a spherical 3-design: *stabilizer states* in power-of-two dimensions. This family of vectors has several descriptions:

(i) They may be viewed as a generalization of both the standard basis and the discrete Fourier basis.

(ii) In quantum information theory, stabilizer states arise naturally as the joint eigenvectors of n commuting Pauli matrices.

(iii) They form the smallest orbit of a prominent symmetry group. This group is known as *Clifford group* in quantum information, the *oscillator group* in finite Weyl-Heisenberg analysis and the metaplectic representation of $\mathrm{Sp}(\mathbb{F}_2, 2n)$ in mathematical physics.

In an ongoing collaboration with Zhu, Gross and Grassl we could prove close-to-optimal convex reconstruction guarantees for phase retrieval from random stabilizer states [KZG16b; ZKGG16].

Shifting focus more towards practical applicability of PhaseLift, Gross, Krahmer and myself considered *random diffraction patterns*. Introduced Candès, Li and Soltanolkotabi [CLS15], this structured measurement setup mimics diffraction imaging experiments that utilize "masks". These authors then proved that $C \log^4(n)$ randomly chosen diffraction patterns (each of which contains n correlated measurement vectors) suffice w.h.p. to recover a given vector $x \in \mathbb{C}^n$ via PhaseLift. Gross, Krahmer and myself improved on this result by showing that already $C \log^2(n)$ such patterns suffice to derive an analogous statement [GKK15b]. This improvement is only a single log-factor away from the information theoretic lower bound for such types of measurements.

Our occupation with the particular aspects of phase retrieval has led to further insights within the field of convex signal reconstruction. These results include reconstruction proofs applicable to compressed sensing from anisotropic measurements [KG14], an improved noise-robustness for compressed sensing of non-negative vectors [KJ16], and identifying the diamond norm as an improved regularizer for certain low rank matrix reconstruction problems [KKEG16].

The mathematical techniques that are typically employed in convex signal reconstruction lend themselves to tackling a great variety of different problems in many scientific disciplines. Being a physicist by training, I have focused on several open problems in quantum information theory. By applying techniques from convex optimization, I could contribute novel insights to the study of Bell inequalities [CKBG15], the task of comparing experiments to the Threshold Theorem in quantum error correction [KLDF16], quantum state discrimination from stabilizer state measurements [KZG16a] and a novel benchmark for fidelity optimization in Bayesian quantum estimation [KF15].

4.2 Outlook

We conclude this work by mentioning several observations and research directions that may merit further attention.

4.2.1 Sparse reconstruction of positive vectors

Positivity constraints can have profound impacts on compressed sensing, see e.g. [BEZ08; DT05; SH+13]. Bruckstein *et al.* for instance showed that positivity renders ℓ_1-minimization superfluous, if the row span of the measurement matrix A intersects the positive orthant [BEZ08]. It is phrased for noiseless measurements and we generalize it to the noisy case [KJ16]. We prove that solving

$$\underset{z \geq 0}{\text{minimize}} \quad \|Az - y\|_{\ell_2}. \tag{4.1}$$

allows for stably reconstructing any positive s-sparse vector if the measurement matrix A obeys a *nullspace property* and its row span intersects the positive orthant. An analogous statement holds true for positive semidefinite matrix reconstruction [KKRT16].

Unlike constrained ℓ_1-norm minimization, such a constrained least-squares regression does not require an a-priori noise bound $\eta \geq \|\epsilon\|_{\ell_2}$. Moreover, the minimal function value $f(z^\sharp) = \left\|Az^\sharp - t\right\|_{\ell_2}$ always provides a lower bound on the noise strength:

$$f(z^\sharp) \leq \|\epsilon\|_{\ell_2}. \tag{4.2}$$

This already provides some information about the noise present in the sampling process. However, a converse bound

$$f(z^\sharp) \geq c\,\|\epsilon\|_{\ell_2} \tag{4.3}$$

would be considerably more desirable. Such a relation would allow for inferring an upper noise bound $\eta := \frac{f(z^\sharp)}{c}$ directly from the reconstruction algorithm (4.2). In turn, this would allow for inferring a "confidence region": If the measurement matrix A admits a strong notion of the NSP, the original vector x is contained in $B = \left\{z \in \mathbb{C}^n : \left\|z - s^\sharp\right\|_{\ell_2} \leq \frac{C'}{c} f(z^\sharp)\right\}$.

Importantly, this region B would be completely specified by the solution of (4.2).

Clearly, the desired relation (4.3) cannot hold in full generality. It is possible to violate it by choosing ϵ adversely with respect to A and x. However, numerical experiments conducted with i.i.d. Gaussian noise suggest that (4.3) holds w.h.p. for *stochastic* noise. I believe that the prospect of a "self certifying" compressed sensing protocol, in the sense of the previous paragraph, merits further attention.

4.2.2 Tensor reconstruction

Our original motivation for [KKEG16] was the seemingly tensorial nature of the *diamond norm*. The diamond norm is defined for linear maps $\mathcal{M} : M_n \to M_m$ that map square matrices onto square matrices:

$$\|\mathcal{M}\|_\diamond = \sup_{N \geq 0} \sup_{X \in M_n \otimes M_N} \frac{\|\mathcal{M} \otimes \mathcal{I}_N(X)\|_1}{\|X\|_1} \tag{4.4}$$

Here, $\mathcal{I}_N : Z \to Z$ denotes the identity on M_N. This is a stabilized version of the induced nuclear norm. Among other things, this stabilization is responsible for the fact that the diamond norm can be evaluated by means of a semidefinite program (SDP). Our working definition as a particular matrix norm, results from choosing an appropriate matrix representation of $\mathcal{M}$. In turn, we have focused on the implications of such a norm for matrix reconstruction [KKEG16]. Alternatively, the diamond norm (4.4) may be viewed as a norm for maps $\mathcal{M}$ with an order-four tensorial structure. It would be interesting to explore this aspect of the diamond norm in the future.

Another promising objective is to consider tensor reconstruction of highly structured tensor families. Permutation invariant tensors, i.e. elements of $H_n^{\otimes N}$ (where N is much larger than n) that are invariant under any permutation of the individual matrix spaces H_n, are highly promising. In quantum mechanics, such tensors describe bosonic systems. Classically, they stand in one-one correspondence to homogeneous polynomials of degree N in n variables. One strong indication that this special case is considerably simpler to treat than the general problem is given by the fact that the dimension of the totally symmetric space grows only polynomially in N, as opposed to exponential. Results might shed light onto conceptual problems like the relation between rank and *symmetric rank* of symmetric tensors, and might have applications for relevant problems such as learning polynomial functions.

4.2.3 Entropic uncertainty relations for stabilizer states

Heisenberg's uncertainty relation is one of the most famous aspects of quantum mechanics. Roughly speaking, it show that it is impossible to prepare quantum systems with sharply defined position and momentum ("preparation uncertainty"):

$$\Delta P \Delta Q \geq \frac{\hbar}{2}.$$

Here, ΔP and ΔQ denote the standard deviations of momentum and position, respectively. Subsequently, Robertson [Rob29] generalized this relation to arbitrary observables A, B and pure quantum states $\psi\psi^*$:

$$\Delta A \Delta B \geq |\langle \psi, [A,B]\, \psi \rangle| \,. \tag{4.5}$$

Here $[\cdot,\cdot]$ denotes the commutator. Importantly, this relation depends on the quantum state. In fact, for finite dimensional quantum systems ($\psi \in \mathbb{C}^n$, $A, B \in H_n$) it is always possible to choose ψ such that that (4.5) becomes trivial [Deu83].

One way to overcome this drawback is to use entropies as a quantitative measure of uncertainty, rather than standard deviations. This approach has become increasingly popular in quantum information science, see for instance [CBTW15], and plays a key role in quantum cryptography.

A particularly strong entropic uncertainty relation is true for two measurements E and F which correspond to the standard and Fourier basis of $\mathbb{C}^n$, respectively. Born's rule implies that performing such measurements on a quantum state $\rho \in H_n$ results in n-variate probability distributions, e.g. $p(E,\rho) = (\langle e_1, \rho e_1 \rangle, \ldots, \langle e_n, \rho e_n \rangle)^T$. If we quantify the *uncertainty* associated with such outcome probability distributions by means of the Shannon entropy $H(p) = -\sum_{k=1}^n p_k \log_2(p_k)$ these two measurements obey

$$H(E,\rho) + H(F,\rho) \geq \frac{1}{2}\log_2(n) \quad \forall \rho.$$

This is an extreme case of a more general entropic uncertainty relation by Maassen and Uffink [MU88]. It is a consequence of minimal coherence (1.17) between standard and Fourier basis and, in some sense, highlights the validity of incoherence as a structural requirement for compressed sensing.

Entropic uncertainty relations may be phrased for more than two basis measurements. For instance, a complete set of $(n+1)$ mutually unbiased bases $M_1, \ldots, M_{n+1} \subset \mathbb{C}^n$ obey

$$\frac{1}{n+1}\sum_{k=1}^{n+1} H(M_k,\rho) \geq \log_2(n+1) - 1 \quad \forall \rho. \tag{4.6}$$

This is a very strong bound, since the individual entropic terms on the left hand side are upper-bounded by $\log_2(n)$. Interestingly, this relation can be derived by exploiting the fact that a complete set of mutually unbiased forms a spherical 2-design, see e.g. [WW10].

In prime power dimensions $n = a^d$, stabilizer states form another prominent set of $N = \prod_{j=1}^{n} \left(a^j + 1\right)$ orthonormal bases S_k whose union is a spherical 2-design. In analogy to (4.6) one can prove

$$\frac{1}{N} \sum_{k=1}^{N} H\left(S_k, \rho\right) \geq \log_2(n+1) - 1 \quad \forall \rho. \tag{4.7}$$

In [KG15], we could show that, in power of two dimensions ($a = 2$), stabilizer states actually constitute a spherical 3-design. Moreover, the results presented in Chapter 3 exactly characterize the fourth moment of such an ensemble. This additional information may be sufficient for further improving the already strong entropic uncertainty relation (4.7).

4.2.4 Clustering

Clustering is a prominent problem in unsupervised learning theory. Given a finite set of points the task is to partition it into k disjoint subsets such that an a priori chosen dissimilarity function is minimized. A particularly illustrative example for such a task is Euclidean clustering: All points are elements of $\mathbb{R}^n$ and their similarity is mediated by the Euclidean distance. A popular choice for the dissimilarity function is then minimizing the squared pairwise distances between points within a cluster. This problem is NP hard in general, but Lloyd's algorithm provides a popular heuristic for solving it. The undesirable fact that this computationally fast heuristic tends to not always converge to the true solution has prompted further investigation. Recently, LP and SDP relaxations of the k-means problem have been proposed [Awa+15; IMPV15]. For certain distributions of the data points – the stochastic ball model – these methods provably recover the underlying cluster structure w.h.p., provided that the individual balls admit a minimal separation distance. The minimal required distances put forth in [Awa+15] and [IMPV15], respectably, differ from each other and neither seems to be optimal. It is plausible that employing more sophisticated proof techniques—most notably the "golfing scheme" for constructing approximate dual certificates (see e.g. [GKK15a; GKK15b])—would allow for further tightening this separation criterion.

5 Bibliography

[AC07] P. Aliferis and A. W. Cross. "Subsystem fault tolerance with the Bacon-Shor code". In: *Phys. Rev. Lett.* 98.22 (2007), p. 220502.

[AE07] A. Ambainis and J. Emerson. "Quantum t-designs: t-wise independence in the quantum world". In: *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*. 2007, pp. 129–140.

[AFZ15] D. M. Appleby, C. A. Fuchs, and H. Zhu. "Group theoretic, Lie algebraic and Jordan algebraic formulations of the SIC existence problem". In: *Quantum Inf. Comput.* 15.1-2 (2015), pp. 61–94.

[Awa+15] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward. "Relax, no need to round: Integrality of clustering formulations". In: *Proceedings of the Conference on Innovations in Theoretical Computer Science*. ACM. 2015, pp. 191–200.

[Bag+06] E Bagan, M. Ballester, R. D. Gill, A Monras, and R Munoz-Tapia. "Optimal full estimation of qubit mixed states". In: *Phys. Rev. A* 73.3 (2006), p. 032301.

[Bar02] A. Barvinok. *A course in convexity*. Vol. 54. American Mathematical Society Providence, RI, 2002.

[BBCE09] R. Balan, B. G. Bodmann, P. G. Casazza, and D. Edidin. "Painless reconstruction from magnitudes of frame coefficients". In: *J. Fourier Anal. Appl.* 15.4 (2009), pp. 488–501.

[BDDW08] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. "A simple proof of the restricted isometry property for random matrices". In: *Constr. Approx.* 28.3 (2008), pp. 253–263.

[BE13] C. Bachoc and M. Ehler. "Tight p-fusion frames". In: *Appl. Comput. Harmon. Anal.* 35.1 (2013), pp. 1–15.

[Bel64] J. S. Bell. "On the Einstein–Podolsky–Rosen Paradox". In: *Physics* 1 (3 1964), p. 195.

[BEZ08] A. M. Bruckstein, M. Elad, and M. Zibulevsky. "On the uniqueness of non-negative sparse solutions to underdetermined systems of equations". In: *IEEE Trans. Inform. Theory* 54.11 (2008), pp. 4813–4820.

[BHH12] F. G. Brandao, A. W. Harrow, and M. Horodecki. “Local random quantum circuits are approximate polynomial-designs”. In: *preprint arXiv:1208.0692* (2012).

[BK10] R. Blume-Kohout. “Optimal, reliable estimation of quantum states”. In: *New J. Phys.* 12.4 (2010), p. 043034.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[C0̀8] E. J. Càndes. “The restricted isometry property and its implications for compressed sensing”. In: *Compt. Rend. Math.* 346.9 (2008), pp. 589–592.

[Can+06] E. J. Candès et al. “Compressive sampling”. In: *Proceedings of the international congress of mathematicians*. Vol. 3. 2006, pp. 1433–1452.

[CBTW15] P. J. Coles, M. Berta, M. Tomamichel, and S. Wehner. “Entropic uncertainty relations and their applications”. In: *preprint arXiv:1511.04857* (2015).

[CEHV15] A. Conca, D. Edidin, M. Hering, and C. Vinzant. “An algebraic characterization of injectivity in phase retrieval”. In: *Appl. Comput. Harmon. Anal.* 38.2 (2015), pp. 346–356.

[CESV15] E. J. Càndes, Y. C. Eldar, T. Strohmer, and V. Voroninski. “Phase retrieval via matrix completion”. In: *SIAM Rev.* 57.2 (2015), pp. 225–251.

[CKBG15] R. Chaves, R. Kueng, J. B. Brask, and D. Gross. “Unifying framework for relaxations of the causal assumptions in Bells theorem”. In: *Phys. Rev. Lett.* 114.14 (2015), p. 140403.

[CL14] E. J. Candès and X. Li. “Solving quadratic equations via PhaseLift when there are about as many equations as unknowns”. In: *Found. Comput. Math.* 14.5 (2014), pp. 1017–1026.

[CLS15] E. J. Càndes, X. Li, and M. Soltanolkotabi. “Phase retrieval from coded diffraction patterns”. In: *Appl. Comput. Harmon. Anal.* 39.2 (2015), pp. 277–299.

[CP11a] E. J. Càndes and Y. Plan. “A probabilistic and RIPless theory of compressed sensing”. In: *IEEE Trans. Inform. Theory* 57.11 (2011), pp. 7235–7254.

[CP11b] E. J. Càndes and Y. Plan. “Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements”. In: *IEEE Trans. Inform. Theory* 57.4 (2011), pp. 2342–2359.

[CR06] E. J. Càndes and J. Romberg. “Quantitative robust uncertainty principles and optimally sparse decompositions”. In: *Found. Comput. Math.* 6.2 (2006), pp. 227–254.

[CRT06] E. J. Càndes, J. K. Romberg, and T. Tao. “Stable signal recovery from incomplete and inaccurate measurements”. In: *Commun. Pure Appl. Math.* 59.8 (2006), pp. 1207–1223.

[CS78] J. F. Clauser and A. Shimony. “Bell’s theorem. Experimental tests and implications”. In: *Rep. Progr. Phys.* 41.12 (1978), p. 1881.

[CST00] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[CSV13] E. J. Càndes, T. Strohmer, and V. Voroninski. “Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming”. In: *Commun. Pure Appl. Math.* 66.8 (2013), pp. 1241–1274.

[DCEL09] C. Dankert, R. Cleve, J. Emerson, and E. Livine. “Exact and approximate unitary 2-designs and their application to fidelity estimation”. In: *Phys. Rev. A* 80 (1 2009), p. 012304.

[DDM03] J. Dehaene and B. De Moor. “Clifford group, stabilizer states, and linear and quadratic operations over GF (2)”. In: *Phys. Rev. A* 68.4 (2003), p. 042318.

[Deu83] D. Deutsch. “Uncertainty in Quantum Measurements”. In: *Phys. Rev. Lett.* 50 (9 1983), pp. 631–633.

[DGS77] P. Delsarte, J.-M. Goethals, and J. J. Seidel. “Spherical codes and designs”. In: *Geom. Dedicata* 6.3 (1977), pp. 363–388.

[DLHP05] P. De La Harpe and C. Pache. “Cubature formulas, geometrical designs, reproducing kernels, and Markov operators”. In: *Infinite groups: geometric, combinatorial and dynamical aspects*. Springer, 2005, pp. 219–267.

[Don06] D. L. Donoho. “Compressed sensing”. In: *IEEE Trans. Inform. Theory* 52.4 (2006), pp. 1289–1306.

[DT05] D. L. Donoho and J. Tanner. “Sparse nonnegative solution of underdetermined linear equations by linear programming”. In: *Proc. Nat. Acad. Sci.* 102.27 (2005), pp. 9446–9451.

[EGK15] M. Ehler, M. Gräf, and F. J. Király. “Phase retrieval using random cubatures and fusion frames of positive semidefinite matrices”. In: *Waves, Wavelets and Fractals* 1.1 (2015).

[EK12] Y. C. Eldar and G. Kutyniok. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[EK13] M. Ehler and S. Kunis. “Phase retrieval using time and Fourier magnitude measurements”. In: *10th International Conference on Sampling Theory and Applications (SampTA)*. 2013.

[FCRP08] M Fazel, E Càndes, B Recht, and P Parrilo. “Compressed sensing and robust recovery of low rank matrices”. In: *42nd Asilomar Conference on Signals, Systems and Computers*. 2008, pp. 1043–1047.

[FHB01] M. Fazel, H. Hindi, and S. P. Boyd. “A rank minimization heuristic with application to minimum order system approximation”. In: *Proceedings of the American Control Conference*. Vol. 6. 2001, pp. 4734–4739.

[Fie82] J. R. Fienup. “Phase retrieval algorithms: a comparison”. In: *Appl. Opt.* 21.15 (1982), pp. 2758–2769.

[FL11] S. T. Flammia and Y.-K. Liu. “Direct fidelity estimation from few Pauli measurements”. In: *Phys. Rev. Lett.* 106.23 (2011), p. 230501.

[Fol16] G. B. Folland. *Harmonic Analysis in Phase Space.(AM-122)*. Vol. 122. Princeton University Press, 2016.

[FR13] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Vol. 1. 3. Springer, 2013.

[GAE07] D. Gross, K. Audenaert, and J. Eisert. “Evenly distributed unitaries: on the structure of unitary designs”. In: *J. Math. Phys.* 48.5 (2007), p. 052104.

[GB14] M. Grant and S. Boyd. *CVX: Matlab Software for Disciplined Convex Programming, version 2.1*. Mar. 2014.

[GBY08] M. Grant, S. Boyd, and Y. Ye. *CVX: Matlab software for disciplined convex programming*. 2008.

[GKK15a] D. Gross, F. Krahmer, and R. Kueng. “A partial derandomization of PhaseLift using spherical designs”. In: *J. Fourier Anal. Appl.* 21.2 (2015), pp. 229–266.

[GKK15b] D. Gross, F. Krahmer, and R. Kueng. “Improved recovery guarantees for phase retrieval from coded diffraction patterns”. In: *Appl. Comput. Harmon. Anal.* (2015).

[GN07] D. Gross and M Nest. “The LU-LC conjecture, diagonal local operations and quadratic forms over GF (2)”. In: *preprint arXiv:0707.4000* (2007).

[Got97] D. Gottesman. “Stabilizer codes and quantum error correction”. PhD thesis. California Institute of Technology, Pasadena, CA, 1997.

[Gro+10] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert. “Quantum state tomography via compressed sensing”. In: *Phys. Rev. Lett.* 105.15 (2010), p. 150401.

[Gro06] D. Gross. “Hudsons theorem for finite-dimensional quantum systems”. In: *J. Math. Phys.* 47.12 (2006), p. 122107.

[Gro11] D. Gross. “Recovering low-rank matrices from few coefficients in any basis”. In: *IEEE Trans. Inform. Theory* 57.3 (2011), pp. 1548–1566.

[Haa+15] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu. “Sample-optimal tomography of quantum states”. In: *preprint arXiv:1508.01797* (2015).

[Hel76] C. W. Helstrom. *Quantum detection and estimation theory*. Vol. 123. Academic press, 1976.

[HFY12] F. J. Herrmann, M. P. Friedlander, and O. Yilmaz. “Fighting the curse of dimensionality: Compressive sensing in exploration seismology”. In: *IEEE Signal Processing Magazine* 29.3 (2012), pp. 88–100.

[HHH05] A Hayashi, T Hashimoto, and M Horibe. “Reexamination of optimal quantum state estimation of pure states”. In: *Phys. Rev. A* 72.3 (2005), p. 032325.

[Hoe63] W. Hoeffding. “Probability inequalities for sums of bounded random variables”. In: *J. Amer. Statist. Assoc.* 58.301 (1963), pp. 13–30.

[IMPV15] T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar. “On the tightness of an SDP relaxation of k-means”. In: *preprint arXiv:1505.04778* (2015).

[Jol02] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[KB11] Y. Koren and R. Bell. “Advances in collaborative filtering”. In: *Recommender systems handbook*. Springer, 2011, pp. 145–186.

[Kec15] M. Kech. “Explicit Frames for Deterministic Phase Retrieval via PhaseLift”. In: *preprint arXiv:1508.00522* (2015).

[KF15] R. Kueng and C. Ferrie. “Near-optimal quantum tomography: estimators and bounds”. In: *New J. Phys.* 17.12 (2015), p. 123013.

[KG14] R. Kueng and D. Gross. “RIPless compressed sensing from anisotropic measurements”. In: *Lin. Alg. Appl.* 441 (2014), pp. 110–123.

[KG15] R. Kueng and D. Gross. “Qubit stabilizer states are complex projective 3-designs”. In: *preprint arXiv:1510.02767* (2015).

[KGK15] R. Kueng, D. Gross, and F. Krahmer. “Spherical designs as a tool for derandomization: The case of PhaseLift”. In: *International Conference on Sampling Theory and Applications (SampTA)*. 2015, pp. 192–196.

[Kit97] A. Y. Kitaev. “Quantum computations: algorithms and error correction”. In: *Russian Math. Surveys* 52.6 (1997), pp. 1191–1249.

[KJ16] R. Kueng and P. Jung. “Robust Nonnegative Sparse Recovery and the Nullspace Property of 0/1 Measurements”. In: *preprint arXiv:1603.07997* (2016).

[KKEG16] M. Kliesch, R. Kueng, J. Eisert, and D. Gross. “Improving compressed sensing with the diamond norm”. In: *IEEE Trans. Inform. Theory* 62.12 (2016), pp. 7445–7463.

[KKRT16] M. Kabanava, R. Kueng, H. Rauhut, and U. Terstiege. “Stable low-rank matrix recovery via null space properties”. In: *Inf. Inference* 5.4 (2016), pp. 405–441.

[KLDF16] R. Kueng, D. M. Long, A. C. Doherty, and S. T. Flammia. "Comparing experiments to the fault-tolerance threshold". In: *Phys. Rev. Lett.* 117.17 (2016), p. 170502.

[KR05] A Klappenecker and M Rotteler. "Mutually unbiased bases are complex projective 2-designs". In: *IEEE International Symposium on Information Theory (ISIT), Vols 1 and 2*. 2005, 1740–1744.

[KRT15] R. Kueng, H. Rauhut, and U. Terstiege. "Low rank matrix recovery from rank one measurements". In: *Appl. Comput. Harmon. Anal.* (2015).

[Kue15] R. Kueng. "Low rank matrix recovery from few orthonormal basis measurements". In: *International Conference on Sampling Theory and Applications (SampTA)*. 2015, pp. 402–406.

[Kup06] G. Kuperberg. "Numerical cubature using error-correcting codes". In: *SIAM J. Numer. Anal.* 44.3 (2006), pp. 897–907.

[KZG16a] R. Kueng, H. Zhu, and D. Gross. "Distinguishability of quantum states under Clifford orbits". In: *preprint* (2016).

[KZG16b] R. Kueng, H. Zhu, and D. Gross. "Low rank matrix recovery from Clifford orbits". In: *preprint* (2016).

[Kön99] H. König. "Cubature formulas on spheres". In: *Math. Res.* 107 (1999), pp. 201–212.

[LB13] D. A. Lidar and T. A. Brun. *Quantum error correction*. Cambridge University Press, 2013.

[LDP07] M. Lustig, D. Donoho, and J. M. Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging". In: *Magn. Reson. Med. Sci.* 58.6 (2007), pp. 1182–1195.

[Liu+08] Y. Liu, B Chen, E. Li, J. Wang, A Marcelli, S. Wilkins, H Ming, Y. Tian, K. Nugent, P. Zhu, et al. "Phase retrieval in x-ray imaging based on using structured illumination". In: *Phys. Rev. A* 78.2 (2008), p. 023817.

[Liu11] Y.-K. Liu. "Universal low-rank matrix recovery from Pauli measurements". In: *Advances in Neural Information Processing Systems*. 2011, pp. 1638–1646.

[MCKS99] J. Miao, P. Charalambous, J. Kirz, and D. Sayre. "Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens". In: *Nature* 400.6742 (1999), pp. 342–344.

[MGE11] E. Magesan, J. M. Gambetta, and J. Emerson. "Scalable and robust randomized benchmarking of quantum processes". In: *Phys. Rev. Lett.* 106.18 (2011), p. 180504.

[Mil90] R. P. Millane. "Phase retrieval in crystallography and optics". In: *JOSA A* 7.3 (1990), pp. 394–411.

[MU88] H. Maassen and J. B. Uffink. “Generalized entropic uncertainty relations”. In: *Phys. Rev. Lett.* 60.12 (1988), p. 1103.

[MWW09] W. Matthews, S. Wehner, and A. Winter. “Distinguishability of quantum states under restricted families of measurements with an application to quantum data hiding”. In: *Commun. Math. Phys.* 291.3 (2009), pp. 813–843.

[NC10] M. A. Nielsen and I. L. Chuang. *Quantum computation and quantum information*. Cambridge University Press, 2010.

[NJS13] P. Netrapalli, P. Jain, and S. Sanghavi. “Phase retrieval using alternating minimization”. In: *Advances in Neural Information Processing Systems*. 2013, pp. 2796–2804.

[NRS01] G. Nebe, E. M. Rains, and N. J. Sloane. “The invariants of the Clifford groups”. In: *Designs, Codes and Cryptography* 24.1 (2001), pp. 99–122.

[NRS02] G. Nebe, E. Rains, and N. Sloane. “A simple construction for the Barnes-Wall lattices”. In: *Codes, Graphs, and Systems*. 2002, pp. 333–342.

[NRS06] G. Nebe, E. M. Rains, and N. J. A. Sloane. *Self-dual codes and invariant theory*. Vol. 17. Springer, 2006.

[OMFH11] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi. “A simplified approach to recovery conditions for low rank matrices”. In: *IEEE International Symposium on Information Theory Proceedings (ISIT)*. 2011, pp. 2318–2322.

[OW15] R. O’Donnell and J. Wright. “Efficient quantum tomography”. In: *preprint arXiv:1508.01907* (2015).

[RBKSC04] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves. “Symmetric informationally complete quantum measurements”. In: *J. Math. Phys.* 45.6 (2004).

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo. “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization”. In: *SIAM Rev.* 52.3 (2010), pp. 471–501.

[Rob29] H. P. Robertson. “The Uncertainty Principle”. In: *Phys. Rev.* 34 (1 1929), pp. 163–164.

[Sch60] J. Schwinger. “Unitary operator bases”. In: *Proc. Natl. Acad. Sci.* 46.4 (1960), pp. 570–579.

[Sco06] A. J. Scott. “Tight informationally complete quantum measurements”. In: *J. Phys. A.* 39.43 (2006), p. 13507.

[SH+13] M. Slawski, M. Hein, et al. “Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization”. In: *Electron. J. Stat.* 7 (2013), pp. 3004–3056.

[Sid99] V. M. Sidelnikov. “Spherical 7-designs in 2^n-dimensional Euclidean space”. In: *J. Algebraic Combin.* 10.3 (1999), pp. 279–288.

[SZ84] P. Seymour and T. Zaslavsky. “Averaging sets: a generalization of mean values and spherical designs”. In: *Adv. Math.* 52.3 (1984), pp. 213–240.

[Tro15] J. A. Tropp. “Convex recovery of a structured signal from independent random linear measurements”. In: *Sampling Theory, a Renaissance*. Springer, 2015, pp. 67–101.

[Uhl76] A. Uhlmann. “The “transition probability in the state space of a -algebra”. In: *Rep. Math. Phys.* 9.2 (1976), pp. 273–279.

[Vor13] V. Voroninski. “Quantum tomography from few full-rank observables”. In: *preprint arXiv:1309.7669* (2013).

[Wal63] A. Walther. “The question of phase retrieval in optics”. In: *J. Mod. Opt.* 10.1 (1963), pp. 41–49.

[Wat09] J. Watrous. “Semidefinite Programs for Completely Bounded Norms”. In: *Theory of Computing* 5.11 (2009), pp. 217–238.

[Wat13] J. Watrous. “Simpler semidefinite programs for completely bounded norms”. In: *Chicago Journal of Theoretical Computer Science* 2013.8 (2013).

[WW10] S. Wehner and A. Winter. “Entropic uncertainty relationsa survey”. In: *New J. Phys.* 12.2 (2010), p. 025009.

[Zau99] G. Zauner. “Quantendesigns: Grundzüge einer nichtkommutativen Designtheorie”. PhD thesis. University of Vienna, 1999.

[ZKGG16] H. Zhu, R. Kueng, M. Grassl, and D. Gross. “The Clifford group fails gracefully to be a unitary 4-design”. In: *preprint* (2016).

6 Back matter

6.1 Acknowledgements

First, I would like to thank David Gross for his dedicated mentoring, his enthusiasm, his advice and the fact that he gave me the liberty, even encouraged me, to pursue scientific goals of my own. I also thank all other current and former members of the Cologne group who all contributed to the friendly and productive atmosphere. This extends in particular to Johan Aaberg, Rafael Chaves, Daniel Suess and Łukasz Rudnicki who provided me with helpful advice both inside and outside of academia.

Next, I would like to thank Steve Flammia for his hospitality, advice and support during my stay in Sydney. This extends to the entire Sydney quantum group who openly welcomed me and provided a very enjoyable and productive atmosphere.

I would also like to thank Felix Krahmer, Holger Rauhut, Andrew Doherty, Jens Eisert, Chris Ferrie, Rafael Chaves and Peter Jung for productive collaborations with great learning effect on my side. This, of course, extends to all my collaborators: Martin Kliesch, Chris Granade, Ulrich Terstiege, Maria Kabanova, Robin Harper, Jonatan Bohr Brask, Daniel Naoumenko, David Long, Alberto Peruzzo and Robert Chapman.

Furthermore, I am grateful for the advice and encouragement I have received from Philipp Walk, Dustin Mixon, Marco Tomamichel, Fernando Brandao, Sjoerd Dirksen, Holger Boche, Rayan Saab, Matthew Fickus, Michael Sandbichler, Martin Ehler, Mario Ziman, Andreas Winter, Michael Walter, Mario Berta, Sandra Keiper, Afonso Bandeira, and many more throughout the course of my scientific career.

I am also deeply grateful for the constant support and encouragement of my family: Erik, Felix, Josef and Gabriela Küng.

Last but foremost, I want to thank Lolita Ammann for for giving me strength and occasionally reminding me that science is not everything. She is also responsible for designing the cover of this thesis. It depicts her artistic view on my work. The featured picture of “Mattei Athena” (Louvre) is courtesy of Aaron Atsma, creator of the Theoi Project (www.theoi.com).

I acknowledge funding by scholarship funds from the State Graduate Funding Program of Baden-Württemberg for the duration of my stay in Freiburg.

6.2 Abstract

Convex signal reconstruction is the art of solving ill-posed inverse problems via convex optimization. It is applicable to a great number of problems from engineering, signal analysis, quantum mechanics and many more. The most prominent example is *compressed sensing*, where one aims at reconstructing sparse vectors from an under-determined set of linear measurements. In many cases, one can prove rigorous performance guarantees for these convex algorithms. The combination of practical importance and theoretical tractability has directed a significant amount of attention to this young field of applied mathematics.

However, rigorous proofs are usually only available for certain “generic cases”—for instance situations, where all measurements are represented by random Gaussian vectors. The focus of this thesis is to overcome this drawback by devising mathematical proof techniques can be applied to more “structured” measurements. Here, *structure* can have various meanings. E.g. it could refer to the type of measurements that occur in a given concrete application. Or, more abstractly, *structure* in the sense that a measurement ensemble is small and exhibits rich geometric features.

The main focus of this thesis is phase retrieval: The problem of inferring phase information from amplitude measurements. This task is ubiquitous in, for instance, in crystallography, astronomy and diffraction imaging. Throughout this project, a series of increasingly better convex reconstruction guarantees have been established. On the one hand, we improved results for certain measurement models that mimic typical experimental setups in diffraction imaging. On the other hand, we identified *spherical t-designs* as a general purpose tool for the derandomization of data recovery schemes. Loosely speaking, a t-design is a finite configuration of vectors that is ”evenly distributed” in the sense that it reproduces the first 2t moments of the uniform measure. Such configurations have been studied, for instance, in algebraic combinatorics, coding theory, and quantum information. We have shown that already spherical 4-designs allow for proving close-to-optimal convex reconstruction guarantees for phase retrieval.

The success of this program depends on explicit constructions of spherical t-designs. In this regard, we have studied the design properties of *stabilizer states*. These are configurations of vectors that feature prominently in quantum information theory. Mathematically, they can be related to objects in discrete symplectic vector spaces—a structure we use heavily. We have shown that these vectors form a spherical 3-design and are, in some sense, close to a spherical 4-design. Putting these efforts together, we establish tight bounds on phase retrieval from stabilizer measurements.

While working on the derandomization of phase retrieval, I obtained a number of results on other convex signal reconstruction problems. These include *compressed sensing from anisotropic measurements*, *non-negative compressed sensing in the presence of noise* and

identifying *improved convex regularizers for low rank matrix reconstruction*. Going even further, the mathematical methods I used to tackle ill-posed inverse problems can be applied to a plethora of problems from quantum information theory. In particular, *the causal structure behind Bell inequalities*, new ways to *compare experiments to fault-tolerance thresholds* in quantum error correction, a novel benchmark for *quantum state tomography via Bayesian estimation*, and the task of *distinguishing quantum states*.

6.3 Kurzzusammenfassung

Konvexe Signalrekonstruktion ist die Kunst des Lösens schlecht gestellter inverser Probleme mittels konvexer Optimierung. Sie ist auf eine große Anzahl von Problemen im Ingenieurwesen, der Signalanalyse, der Quantenmechanik und vielen weiteren anwendbar. Der bekannteste Anwendungsfall ist *Compressed Sensing*, dessen Ziel es ist, dünnbesetzte Vektoren aus einer unterbestimmten Menge an linearen Messungen zu rekonstruieren. In vielen Fällen ist es möglich, rigorose Leistungsgarantien für diese konvexen Algorithmen zu beweisen. Die Kombination aus praktischer Bedeutung und theoretischer Beweisbarkeit hat zu einem beträchtlichen Interesse an diesem jungen Teilgebiet der angewandten Mathematik geführt.

Nichtsdestotrotz, sind rigorose mathematische Beweise für gewöhnlich nur für gewisse “generische Fälle” vorhanden—zum Beispiel Instanzen, wo alle Messungen zufälligen Gauss-Vektoren entsprechen. Das Thema dieser Arbeit ist es diese Beeinträchtigung durch das Entwickeln neuer mathematische Beweistechniken zu beheben, welche auf “strukturiertere” Messinstanzen anwendbar sind. Wohlgemerkt, kann *Struktur* hier mannigfaltig ausgelegt werden. Zum Beispiel könnte sie auf Messprozesse in konkreten Anwendungen hindeuten. Oder, abstrakter, *Struktur* im Sinne eines kleinen Messensembles, welches besondere geometrische Eigenschaften aufweist.

Ein wichtiger Aspekt dieser Arbeit ist *Phase Retrieval.* Darunter versteht man die Aufgabe komplexe Phaseninformation aus Amplitudenmessungen zu gewinnen. Dieses Problem ist allgegenwärtig in vielen Disziplinen, zum Beispiel in Kristallographie, Astronomie und “Diffraction Imaging”. Im Laufe dieses Projektes wurde eine Reihe stetig besser werdender konvexer Rekonstruktionsgarantien hergeleitet. Auf der einen Seite haben wir bestehende Resultate verbessert, welche für Messmodelle gelten die typische experimentelle Prozeduren in “Diffraction Imaging” imitieren. Auf der anderen Seite, haben wir sphärische t-Designs als Allzweck-Werkzeug für das Derandomisieren von Datenrekonstruktionsverfahren identifiziert. Vereinfacht gesagt, ist ein t-Design eine Konfiguration endlich vieler Vektoren, welches “gleichverteilt ist” in dem Sinn, dass sie die ersten $2t$ Momente der uniformen Verteilung auf der Sphäre reproduziert. Derartige Konfigurationen wurden zum Beispiel im Rahmen der algebraischen Kombinatorik, der Kodierungstheorie und in der Quanteninformationstheorie untersucht. Wir haben gezeigt, dass bereits sphärische 4-Designs es erlauben, beinahe optimale konvexe Rekonstruktionsgarantien für Phase Retrieval herzuleiten.

Der Erfolg eines solchen Programms hängt stark von expliziten Konstruktionen sphärischer t-Designs ab. Um das zu erreichen, haben wir die Designeigenschaften von *Stabilsatorzuständen* untersucht. Diese sind eine in der Quanteninformationstheorie sehr wichtige Vektorkonfiguration. In mathematischer Hinsicht können sie mit diskreten symplektischen Vektorräumen in Verbindung gebracht werden—Eine Struktur die wir stark ausnützen. Wir haben gezeigt, dass diese Vektoren ein sphärisches 3-Design bilden, welche zudem einem 4-Design in gewisser

Weise nahekommen. In dem wir diese Errungenschaften mit den Obengenannten verbinden, leiten wir optimale Schranken für Phase Retrieval mittels Stabilisatorzuständen her.

Im Verlauf meiner Arbeit an der Derandomisierung von Phase Retrieval habe ich eine Anzahl an weiteren Resultaten im Rahmen der konvexen Signalanalyse erarbeitet. Diese beinhalten *Compressed Sensing von anisotropen Messungen*, *verrauschtes nicht-negatives Compressed Sensing*, und das *Identifizieren eines besseren konvexen Regularisierers für bestimmte Matrixrekonstruktionsprobleme*. Darüber hinaus, können die mathematischen Methoden, welche ich zum Bearbeiten schlecht-gestellter inverser Probleme verwendet habe, auf eine Vielzahl an Problemen der Quanteninformation angewendet werden. Konkret handelt es sich hierbei um *die kausale Struktur hinter Bellungleichungen*, neue *Möglichkeiten Experimente mit dem "Fault-Tolerance Threshold" zu vergleichen*, einen neuen Maßstab für *Quantenzustandstomographie durch Bayes'sche Schätztheorie*, und die Aufgabe *Quantenzustände zu unterscheiden*.

6.4 The author's contributions

For each publication and draft, the author's contribution to conception, realization and report writing is listed here.

[GKK15a]: *A partial derandomization of PhaseLift using spherical designs*
The author is the main contributor of this publication. In particular, he developed the mathematical proofs, conducted the numerical experiments and contributed to the presentation.

[KGK15]: *Spherical designs as a tool for derandomization: the case of PhaseLift*
The author is the main contributor of this publication. He proposed the problems, developed the mathematical proofs and contributed to the presentation.

[KRT15]: *Low rank matrix recovery from rank one measurements*
The author contributed important parts. In particular, he developed the proofs concerning spherical 4-designs, wrote the section about quantum state tomography and contributed to the presentation.

[KKRT16]: *Stable low-rank matrix recovery via null space properties*
The author is in large parts responsible for the results on positive-semidefinite matrix reconstruction, wrote the sections about phase retrieval and quantum information, and contributed to the presentation.

[GKK15b]: *Improved recovery guarantees for phase retrieval from coded diffraction patterns*
The author is the main contributor of this publication. In particular, he proposed the problem, developed the mathematical proofs and contributed to the presentation.

[KG15]: *Qubit stabilizer states are complex projective 3-designs*
The author is the main contributor of this publication. In particular, he developed the mathematical proofs and contributed to the presentation.

[Kue15]: *Low rank matrix recovery from few orthonormal bases*
The author is the sole contributor of this publication.

[KG14]: *RIPless compressed sensing from anisotropic measurements*
The author is the main contributor of this publication. In particular he developed the mathematical proofs and contributed to the presentation.

[KJ16]: *Robust nonnegative sparse recovery and the null space property of* $0/1$ *measurements*
The author is the main contributor of this publication. In particular, he is responsible for the technical aspects of this paper and contributed to the presentation.

[KKEG16]: *Improving compressed sensing with the diamond norm*

The author contributed important parts to the conception of the work, developed the mathematical proof of the main technical result and wrote the sections about matrix reconstruction.

[CKBG15]: *A unifying framework for relaxations of the causal assumptions in Bell's theorem*
The author is responsible for the technical aspects of this paper, in particular the linear programming techniques.

[KF15]: *Near optimal quantum tomography: estimators and bounds*
The author is the main contributor of this paper. In particular, he developed the proofs for the main technical statements and contributed to the presentation.

[KLDF16]: *Comparing experiments to the fault-tolerance threshold*
The author contributed important parts. In particular, he is responsible for developing the mathematical proofs and wrote the supplementary material.

[ZKGG16] *The Clifford group fails gracefully to be a unitary 4-design*
The author was instrumental in proposing the problem and pointing out potential applications.

[KZG16b] *Low rank matrix recovery from Clifford orbits*
The author is the main contributor of this draft. In particular, he proposed the problem, developed the mathematical proofs and is responsible for the presentation.

[KZG16a] *Distinguishability of quantum states under Clifford orbits*
The author is the main contributor of this draft. In particular, he proposed the problem, developed the mathematical proofs and is responsible for the presentation.

6.5 Eigenständigkeitserklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit einschließlich Tabellen, Karten und Abbildungen, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie abgesehen von den angegebenen Teilpublikationen noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. David Gross betreut worden.

Richard Küng