Local Linear Convergence of Douglas-Rachford for Linear Programming: a Probabilistic Analysis

Oisín Faust [1 2] **Hamza Fawzi** [1 2]

Abstract

Douglas-Rachford splitting/ADMM (henceforth DRS) is a very popular algorithm for solving convex optimisation problems to low or moderate accuracy, and in particular for solving large-scale linear programs. Despite recent progress, obtaining highly accurate solutions to linear programs with DRS remains elusive. In this paper we analyze the local linear convergence rate r of the DRS method for random linear programs, and give explicit and tight bounds on r. We show that $1-r^2$ is typically of the order of $m^{-1}(n-m)^{-1}$, where n is the number of variables and m is the number of constraints. This provides a quantitative explanation for the very slow convergence of DRS/ADMM on random LPs. The proof of our result relies on an established characterisation of the linear rate of convergence as the cosine of the Friedrichs angle between two subspaces associated to the problem. We also show that the cosecant of this angle can be interpreted as a condition number for the LP.

1. Introduction

A *linear program* (LP) is an optimisation problem which can be expressed in the following *standard form*[1]

$$\begin{aligned} \text{minimise} \quad \langle c, x\rangle \quad \text{s.t.} \quad & Ax = A\bar{x} \\ & x \in \mathbb{R}^n_+. \end{aligned} \tag{1}$$

The data for this LP consist of vectors $\bar{x}, c \in \mathbb{R}^n$, and a matrix $A \in \mathbb{R}^{m\times n}$ with full row rank m, $1 \le m < n$. The *dual linear program* is

$$\begin{aligned} \text{maximise} \quad \langle A\bar{x}, y\rangle \quad \text{s.t.} \quad & s := c - A^\top y \in \mathbb{R}^n_+ \\ & y \in \mathbb{R}^m. \end{aligned} \tag{2}$$

Douglas-Rachford splitting (Lions & Mercier, 1979) is a fundamental algorithm for convex optimisation (and more generally, for finding the zeros of monotone operators). For convex optimisation problems, Douglas-Rachford splitting (DRS) is known to be equivalent to the equally well-known Alternating Direction Method of Multipliers (ADMM) applied to the dual problem (Gabay, 1983; Eckstein & Bertsekas, 1992). One area where DRS/ADMM has found much success is in solving large-scale conic programs – see e.g. (O'Donoghue et al., 2016). A direct application of DRS for solving (1) is listed in Algorithm 1. (See Appendix A.1 for the derivation of this algorithm.)

Algorithm 1 Douglas-Rachford for (1)

Input: Initial point $z^0 \in \mathbb{R}^n$
$k \leftarrow 0$
repeat
  $s^k \leftarrow \Pi_{\mathbb{R}^n_+}(z^k) - z^k$
  $x^k \leftarrow (z^k + 2s^k) - A^\top(AA^\top)^{-1}A\left(z^k + 2s^k - \bar{x}\right)$
  $z^{k+1} \leftarrow x^k - s^k$
  $k \leftarrow k + 1$
until convergence, or other termination criterion satisfied

Like most other first-order methods (and as opposed to, say, interior point methods), DRS is known to have an initial phase of "fast" convergence towards an approximate solution, followed by a phase of slow convergence towards the exact solution. For this reason, splitting algorithms have typically only been used to obtain solutions of low to moderate accuracy. However, there has been recent interest in using splitting methods to solve large-scale linear programs to high accuracy, see e.g., (Applegate et al., 2021; Lu & Yang, 2021). For linear programs it has been known (Boley, 2013) that the "slow" phase of convergence of DRS/ADMM is in fact linear convergence (i.e. the error decreases geometrically). The catch is that the rate of ultimate linear convergence may be very close to 1. This is illustrated in

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom [2]Cantab Capital Institute for the Mathematics of Information, University of Cambridge, Cambridge, United Kingdom. Correspondence to: Oisín Faust <opbf2@maths.cam.ac.uk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Our representation of the affine constraints as $Ax = A\bar{x}$ instead of the traditional $Ax = b$ encodes the implicit assumption that the corresponding affine space is nonempty.

Figure 1, which shows the behaviour of DRS on small random linear programs ($n \leq 40$), and where DRS may require more than 10^4 iterations to get within distance $\epsilon = 10^{-6}$ from the solution.

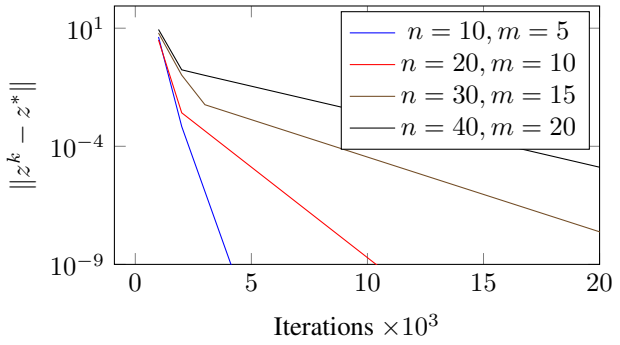


Figure 1. Convergence of Algorithm 1 for random linear programs with n variables and $m = n/2$ equality constraints. Note the linear convergence after an initial transient phase. Also note that the linear programs are fairly small in size, and yet DRS may require more than $\approx 10^4$ iterations to reach an accuracy of 10^{-6}.

1.1. Contributions

In this paper we perform an average-case analysis of Douglas-Rachford splitting for linear programs, and give tight and explicit bounds on the rate of eventual linear convergence for random linear programs generated from a natural distribution. Our main result, stated below, gives a quantitative explanation for the very slow convergence of DRS/ADMM on random LP instances.

Theorem 1.1. *Let $1 \leq m < n$. Let $c, \bar{x}$ be independent vectors drawn from spherically symmetric distributions on $\mathbb{R}^n$ which give zero probability to the point 0. Let the entries of $A \in \mathbb{R}^{m \times n}$ be i.i.d. $\mathcal{N}(0,1)$ random variables independent of c and $\bar{x}$.*

Let E be the event that (1) and (2) have unique optimal solutions x^* and (s^*, y^*), respectively. For any $(A, c, \bar{x}) \in E$ and $z^0 \in \mathbb{R}^n$, there are numbers $r = r(A, c, \bar{x}) < 1$ and $K = K(z^0, A, c, \bar{x}) \in \mathbb{N}$ such that the iterates of Algorithm 1 satisfy

$$\left\|z^k - z^*\right\| \leq r^{k-K} \left\|z^K - z^*\right\| \leq r^{k-K} \left\|z^0 - z^*\right\|$$

for every $k \geq K$. The distribution of r satisfies

$$\mathbb{P}\left(\frac{\delta^2}{m(n-m)} \leq 1 - r^2 \leq \frac{2\log(1/\delta)}{m(n-m)} \;\middle|\; E\right) \geq 1 - 2\delta \tag{3}$$

for every $\delta \in (0, \frac{1}{2})$.

Remark 1.2. It is well-known, going back at least to (Adler & Berenguer, 1981), that the event E defined in Theorem 1.1 occurs with probability exactly $\frac{1}{2^n}\binom{n}{m}$. See also (Amelunxen & Bürgisser, 2015b).

Figure 2 shows a scatter plot of the quantity $m(n-m)(1-r^2)$ for randomly generated LPs for $1 \leq n \leq 1000$ and $1 \leq m \leq n-1$. As expected, we observe that $m(n-m)(1-r^2)$ is bounded both above and away from zero in probability, as stated in Theorem 1.1.

The proof of Theorem 1.1 relies at its core on the notion of *angles* between subspaces. It can be shown that, when the LP is feasible and bounded with unique primal and dual solutions, the local linear rate of convergence of DRS is equal to $\sin\theta$, where θ is the *largest principal angle* between the subspaces $\ker(A)$ and the coordinate subspace spanned by the support of x^* (Bauschke et al., 2014; Demanet & Zhang, 2016; Aspelmeier et al., 2016; Liang et al., 2017). This angle is very difficult to estimate in general as it depends on the optimal solution x^* of the considered LP. The crux of Theorem 1.1 is to show that for randomly generated LPs, the distribution of this angle is the same as the one for two uniformly chosen random subspaces from $\mathrm{Gr}(n-m, m)$, the Grassmannian manifold of $(n-m)$-dimensional subspaces of $\mathbb{R}^n$. In (Absil et al., 2006), a formula for the density of this distribution was computed in terms of matrix hypergeometric functions. Using properties of the latter, we prove simple tail bounds which allow us to obtain the estimate (3), and which we believe can also be of independent interest.

1.2. Related work

Average-case analysis has a long history in the analysis of algorithms, and in particular for linear programming. Average-case analyses of the simplex method were performed in (Borgwardt, 1982; Smale, 1983) showing that the complexity is polynomial in average, in constrast to the exponential worst-case complexity. See also (Bürgisser & Cucker, 2013) where average-case analyses of various problems in numerical linear algebra and optimisation are carried out from the point of view of *condition numbers*. In fact, we will see in Definition 4.1 that the quantity $\frac{1}{\sqrt{1-r^2}}$ can be interpreted as a certain condition number for the primal-dual LP, i.e., it measures the sensitivity of the solution to perturbations in the input. However, unlike the many notions of condition numbers for LP which have been introduced (Renegar, 1995; Cheung & Cucker, 2001; Amelunxen & Bürgisser, 2011) which measure sensitivity to pertubations of the triple $(A, \bar{x}, c)$, the quantity $\frac{1}{\sqrt{1-r^2}}$ measures only the sensitivity with respect to the "right-hand sides" $(\bar{x}, c)$ while keeping the matrix A fixed. A byproduct of Theorem 1.1 will allow us to obtain a tight estimate on this condition number, showing that $\mathbb{E}\log\left(\frac{1}{\sqrt{1-r^2}}\right) \approx \log\left(\sqrt{m(n-m)}\right)$ (see Remark 4.3).

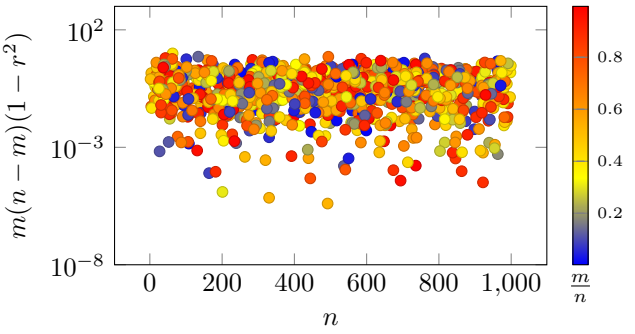


Figure 2. The defect of the eventual linear rate of convergence of DRS for random LPs. r is computed in each case by SVD. The ratio $\frac{m}{n}$ is indicated using colour.

1.3. Organisation

In Section 2 we recall some concepts related to linear programming, DRS, and angles between subspaces which will be used in the rest of the paper. We also characterise the rate of local linear convergence of Algorithm 1 for a primal-dual feasible LP geometrically, in terms of angles between subspaces. We prove Theorem 1.1 in Section 3. In Section 4, we discuss a connection to a condition number for LPs.

2. Preliminaries

2.1. Linear Programming

As we will see, Algorithm 1 only depends on the matrix A through its kernel. Moreover, the algorithm simultaneously solves (1) and its dual (2) in an entirely symmetric manner. Consequently, in this paper it will be both convenient and mathematically pleasing to recast the primal-dual pair of LPs (1) and (2) in a form which includes only the essential information about A, and emphasises the symmetry between the primal and dual LPs. To this end, let us write $L = \ker A$ (a subspace of $\mathbb{R}^n$ of codimension m). Define $\hat{s} = \Pi_L(c)$, $\hat{x} = \Pi_{L^\perp}(\bar{x})$, where, given a closed convex set $C \subseteq \mathbb{R}^n$ we denote by Π_C the Euclidean projection onto C. The primal is equivalent to

$$\text{minimise} \quad \langle x, \hat{s} \rangle \ \text{ s.t. } \ x \in (\hat{x} + L) \cap \mathbb{R}^n_+ \qquad \text{(LP)}$$

while the dual is equivalent to

$$\text{minimise} \quad \langle \hat{x}, s \rangle \ \text{ s.t. } \ s \in (\hat{s} + L^\perp) \cap \mathbb{R}^n_+. \qquad \text{(LP-D)}$$

This can be checked by substituting $s = c - A^\top y$ and observing that $L^\perp = \operatorname{Im} A^\top$; although the objective values of (1) and (2) differ from those of (LP) and (LP-D), the optimal solutions x^* and s^* (if they exist) are the same. In this formulation, both the primal *and the dual* are minimisation problems over the intersection of the nonnegative orthant with an affine space, and weak duality corresponds to the fact that for feasible x and s we have

$$\begin{aligned} -\langle \hat{x}, s \rangle \leq \langle x - \hat{x}, s \rangle &= \langle x - \hat{x}, \Pi_L s \rangle \\ &= \langle x - \hat{x}, \hat{s} \rangle = \langle x, \hat{s} \rangle, \end{aligned} \qquad (4)$$

i.e. $\langle x, \hat{s} \rangle + \langle \hat{x}, s \rangle \geq 0$. Standard duality theory for LPs states that if x is optimal for (LP) and s is optimal for (LP-D), then equality holds (*strong duality*). If one of the programs is infeasible, strong duality states that its dual program admits feasible points of arbitrarily low objective value.

From now on, we will only be concerned with the linear programming formulation (LP) and (LP-D). The data for the pair (LP), (LP-D) consists of the linear subspace L, and the vectors $(\hat{x}, \hat{s})$. Since $(\hat{x}, \hat{s})$ live in orthogonal spaces $(L^\perp, L)$ it is tempting to combine them into one vector $\hat{z} := \hat{x} - \hat{s} \in \mathbb{R}^n$. $(\hat{x}, \hat{s})$ can always be recovered from $\hat{z}$ by projecting onto $L^\perp$ or L. If x^* is optimal for (LP) and s^* is optimal for (LP-D), strong duality necessitates that they are complementary (their supports do not overlap), so that $x^* = \Pi_{\mathbb{R}^n_+}(x^* - s^*)$ and $-s^* = \Pi_{\mathbb{R}^n_-}(x^* - s^*)$. It is a fact that every primal-dual feasible LP has at least one *strictly complementary* optimal solution pair, i.e. such that $z^* := x^* - s^*$ has full support (Goldman & Tucker, 1957). Consequently, if both (LP) and (LP-D) have unique solutions, these must be strictly complementary.

2.2. DRS for Linear Programming

The iteration for solving the pair (LP) and (LP-D) which we will study is given in Algorithm 2. It is simply a fixed point iteration of the *Douglas-Rachford operator* (for LPs)

$$T^L_{\hat{z}} = \Pi_L \circ \Pi_{\mathbb{R}^n_+} + \Pi_{L^\perp} \circ \Pi_{\mathbb{R}^n_-} + \hat{z}. \qquad (5)$$

In Appendix A.1, we derive Algorithm 2 as Douglas-Rachford splitting applied to a reformulation of (LP) as the minimisation of the sum of two convex functions. It can be verified that Algorithm 2 is precisely equivalent to Algorithm 1, expressed in terms of L and $\hat{z}$ as introduced in Section 2.1. Indeed the matrix $A^\top(AA^\top)A$ represents the operator $\Pi_{L^\perp}$.

Algorithm 2 Douglas-Rachford for Linear Programs

Input: Initial point z^0, subspace L, vector $\hat{z}$
$k \leftarrow 0$
repeat
  $z^{k+1} \leftarrow T^L_{\hat{z}}(z^k)$
  $k \leftarrow k + 1$
until convergence, or other termination criterion satisfied
$s^k \leftarrow \Pi_{\mathbb{R}^n_+}(-z^k)$
$x^k \leftarrow \Pi_{\mathbb{R}^n_+}(z^k)$

By Theorem A.3, for any initial z^0, Algorithm 2 converges to a fixed point of the Douglas-Rachford operator $T_{\hat{z}}^L$, if one exists. The point $z \in \mathbb{R}^n$ is a fixed point of $T_{\hat{z}}^L$ if and only if

$$\Psi^L(z) = \hat{z} \tag{6}$$

where

$$\Psi^L \;=\; \Pi_{L^\perp} \circ \Pi_{\mathbb{R}^n_+} + \Pi_L \circ \Pi_{\mathbb{R}^n_-} \;=\; \mathrm{Id} - T_{\hat{z}}^L + \hat{z},$$

which depends only on L (not on $\hat{z}$), is the *forward map* for linear programming. Indeed, it is easy to check that if x is optimal for (LP) and s is optimal for (LP-D), then $z := x - s$ satisfies (6). The converse also holds.

Proposition 2.1. *$z \in \mathbb{R}^n$ satisfies (6) if and only if $x := \Pi_{\mathbb{R}^n_+}(z)$ is an optimal solution to (LP) and $s := \Pi_{\mathbb{R}^n_+}(-z)$ is an optimal solution to (LP-D).*

Proof. See Appendix B. □

Remark 2.2. Proposition 2.1 shows that (6) effectively encodes the *KKT conditions* for linear programming

$$\begin{cases} x \cdot s = 0 \\ x \in \hat{x} + L, x \geq 0 \\ s \in \hat{s} + L^\perp, s \geq 0 \end{cases} \tag{7}$$

(here $\cdot$ denote elementwise multiplication in $\mathbb{R}^n$). The celebrated class of interior-point methods work by solving these KKT equations using Newton's method. DRS/ADMM reformulates KKT optimality conditions as a fixed point equation $T_{\hat{z}}^L(x - s) = x - s$ and applies the fixed-point iteration.

Remark 2.3. Equation (6) suggests an interpretation of the dual pair of linear programs (LP, LP-D) as an inverse problem parameterised by L and with observed data $\hat{z}$.

2.3. Angles between subspaces

Given subspaces $L_1,\, L_2 \subset \mathbb{R}^n$ there is a notion of *minimal angle* between them. This is the angle in $[0, \frac{\pi}{2}]$ whose cosine equals

$$c_0(L_1,\, L_2) := \sup\{\langle u, v\rangle \mid u \in L_1 \cap B, v \in L_2 \cap B\},$$

where B is the unit ball for the Euclidean norm in $\mathbb{R}^n$. Equivalently,

$$c_0(L_1,\, L_2) = \|\Pi_{L_1}\Pi_{L_2}\|\,. \tag{8}$$

When this angle is zero (i.e. the subspaces share a straight line in common), we will often be more interested in the *Friedrichs angle*, obtained by quotienting out their intersection, which is positive. This angle has cosine given by

$$c(L_1,\, L_2) := c_0\big(L_1 \cap (L_1 \cap L_2)^\perp,\, L_2 \cap (L_1 \cap L_2)^\perp\big).$$

Equivalently, by Lemma 10 in (Deutsch, 1995),

$$c(L_1,\, L_2) = \|\Pi_{L_1}\Pi_{L_2} - \Pi_{L_1 \cap L_2}\|\,. \tag{9}$$

One can easily see that

Fact 2.4. $c(L_1,\, L_2) < 1$. Moreover, $c(L_1,\, L_2) = c_0(L_1, L_2) \iff c_0(L_1, L_2) < 1 \iff L_1 \cap L_2 = \{0\}$.

However, the following important result is not obvious:

Proposition 2.5. $c(L_1,\, L_2) = c(L_1^\perp, L_2^\perp)$.

Proof. See Theorem 16 in (Deutsch, 1995) for a proof of this result in the more general setting of a (possibly infinite-dimensional) Hilbert space. □

2.3.1. Relationship to Largest Principal Angle

In Section 3, we will need to study the distribution of the minimal angle between random subspaces. In this subsection, we show that this is equivalent to studying the distribution of the *largest principal angle* between random subspaces. This will allow us to apply the results of (Absil et al., 2006) later on, in Section 3.1.

Definition 2.6. $d_p(W_1, W_2) := \|\Pi_{W_1} - \Pi_{W_2}\|$ is a distance on $\mathrm{Gr}(n, m)$ known as the *projection distance*. The quanitity $\arcsin d_p(W_1, W_2)$ is the *largest principal angle* between W_1 and W_2.

Proposition 2.7. *For any m-dimensional subspaces $W_1, W_2 \subset \mathbb{R}^n$ such that $W_1^\perp \cap W_2 = \{0\}$,*

$$c_0(W_1^\perp, W_2) = d_p(W_1, W_2).$$

Proof. See Appendix B. □

2.4. Local Linear Convergence

Local linear convergence of the Douglas-Rachford splitting algorithm for linear programs was proved in (Boley, 2013). In this section we will provide an explicit expression for the local convergence rate in terms of the angle between certain subspaces. The next result can be derived from (Liang et al., 2017), which give several general conditions for local linear convergence of the Douglas-Rachford splitting algorithm using ideas of partial smoothness and metric subregularity. However, for the reader's convenience we provide an elementary proof in Appendix B.

Theorem 2.8. *Let (LP) and (LP-D) be feasible, so that the DRS iteration (Algorithm 2) converges to a solution z^* of (6). Define $W^+ = \mathrm{span}\{e_i \mid z_i^* > 0\}$ and $W^- = \mathrm{span}\{e_i \mid z_i^* < 0\}$. Suppose that z^* has full support, so that $W^+ \oplus W^- = \mathbb{R}^n$. Then there exists $K \in \mathbb{N}_0$ such that for all $k \geq K$*

$$\left\|z^{k+1} - z^*\right\| \leq r \left\|z^k - z^*\right\| \tag{10}$$

where $r = c(L, W^+) = c(L^\perp, W^-)$.

This rate r is best possible in the sense that, given a solution z^* to (6) with full support, we can choose z^0 so that $\left\|z^k - z^*\right\| = r^k \left\|z^0 - z^*\right\|$ for every $k \geq 0$.

The condition that z^* has full support in the theorem above is satisfied generically, and in particular when the pair of primal-dual LPs have a unique solution. In fact, in the latter situation, the quantity $c(L, W^+)$ is equal to $c_0(L, W^+)$. This is the content of the next corollary.

Corollary 2.9. *Suppose* (6) *has a unique solution* z^*. *Then* W^+, W^- *defined in Theorem 2.8 do not depend on the initial point* z^0 *of the DRS iteration. They are orthogonal complements in* $\mathbb{R}^n$, *so* (10) *holds. Moreover, in this case*

$$c_0(L, W^+) = c(L, W^+) = c(L^\perp, W^-) = c_0(L^\perp, W^-).$$

Proof. See Appendix B. □

3. Probabilistic Analysis

In this section we analyze the behavior of DRS for random LPs, and prove our main result, Theorem 1.1. We start by formally defining the random model, in a coordinate-free way. It can be easily shown that this random model contains the model of Theorem 1.1 as a special case, as we explain later.

Definition 3.1 (Symmetric random model)**.** Let $1 \leq m < n$. Let L be a random subspace drawn from the Haar measure on the Grassmannian $\mathrm{Gr}(n, n-m)$. Conditional on L, let $\hat{x} \in L^\perp$ and $\hat{s} \in L$ be random variables drawn from spherically symmetric distributions which give zero probability to 0.

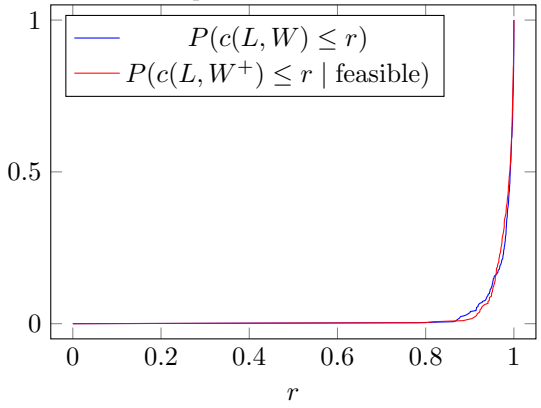
We will see that under such a model, conditional on primal-dual feasibility, (6) has a unique solution with (conditional) probability one, so Corollary 2.9 applies. Moreover, we show that the (conditional) probability distribution of the relevant convergence rate $r = c(L, W^+)$ has a particularly simple description. It is the same as the distribution of $c(L, W)$ where W is a *fixed* m-dimensional subspace. We emphasise that it is not *a priori* obvious that this should be the case. Indeed the optimal primal support subspace W^+ itself depends on L, as well as on $\hat{z} = \hat{x} - \hat{s}$.

Theorem 3.2. *Under a symmetric random model as in Definition 3.1, with probability* $2^{-n}\binom{n}{m}$, *there are unique solutions* x^* *and* s^* *for* (LP) *and* (LP-D), *respectively. In this case, let* W^+ *be as in Theorem 2.8, with* $z^* = x^* - s^*$. *The distribution of* $c(L, W^+)$ *(conditioned on the existence of unique solutions for* (LP) *and* (LP-D)*) is the same as that of* $c(L, W)$, *where* W *is an arbitrary fixed subspace of dimension* m.

In Figure 3 we plot the empirical distribution of both $c(L, W^+)$ [the convergence rate of DRS] and $c(L, W)$ where W is a fixed (arbitrary) subspace; we verify that both distributions match.

We now introduce an important definition that will be used in the proof of Theorem 3.2.

Figure 3. Empirical estimates of the probability distributions in Theorem 3.2, with $n = 10, m = 7$, each based on 300 samples. To obtain $P(c(L, W^+) \leq r \mid \text{feasible})$ we discarded sample LPs that were not primal-dual feasible until we had 300 feasible ones. Theorem 3.2 says these distributions should be the same, which is consistent with these empirical estimates.



Definition 3.3. We will use the notation

$$J_m := \{\beta \in \{\pm 1\}^n \mid \sharp\{i : \beta_i = +1\} = m\}.$$

Let L be a subspace of $\mathbb{R}^n$ of codimension m. For each $\beta \in J_m$, define the complementary orthogonal subspaces

$$W_\beta^+ := \mathrm{span}\{e_i \mid \beta_i = +1\}$$
$$W_\beta^- := \mathrm{span}\{e_i \mid \beta_i = -1\},$$

and the linear map

$$\Psi_\beta^L := \Pi_{L^\perp}\Pi_{W_\beta^+} + \Pi_L \Pi_{W_\beta^-}.$$

Note that Ψ_β^L coincides with the (nonlinear) map Ψ^L on $\{z : \mathrm{sign}(z) = \mathrm{sign}(\beta)\}$. Given a pair $(L, \hat{z})$,

$$\Psi_\beta^L(z) = \hat{z} \iff z = x - s \text{ and } \begin{cases} x \in W_\beta^+, s \in W_\beta^- \\ x \in \hat{x} + L \\ s \in \hat{s} + L^\perp. \end{cases} \tag{11}$$

In other words, and in light of Remark 2.2, Ψ_β^L encodes a linearisation of the KKT system, where the supports of x and s are fixed. From this, we see that $z = z^*$ is a solution of the LP if $\mathrm{sign}(z^*) = \beta$ and $\Psi_\beta^L(z^*) = \hat{z}$.

The key to the proof of Theorem 3.2 is to consider the action of the group $\{\pm 1\}^n$ on $(L, \hat{z})$ which flips signs. Both the quantity $c(L, W_\beta^+)$, and the law of $(L, \hat{z})$, are invariant under this action. However, the induced action on $\mathrm{sign}\big((\Psi_\beta^L)^{-1}(\hat{z})\big)$ is transitive (provided L and $\hat{z}$ are generic enough that Ψ_β^L is invertible and $(\Psi_\beta^L)^{-1}(\hat{z})$ has full support). This observation, formalised in Lemma 3.6, will allow

us to prove Theorem 3.2. First we need some technical lemmas – the main argument begins after Lemma 3.5.

Observe that, given $\beta \in J_m$, Ψ_β^L is invertible if and only if

$$\begin{aligned}\operatorname{rank}(\Pi_{L^\perp}\Pi_{W_\beta^+}) = m \text{ and } \operatorname{rank}(\Pi_L\Pi_{W_\beta^-}) = n-m \\ \iff L\cap W_\beta^+ = \{0\} \text{ and } L^\perp \cap W_\beta^- = \{0\} \\ \iff L\cap W_\beta^+ = \{0\}.\end{aligned} \tag{12}$$

The last equivalence can be seen as follows:

$$\begin{aligned}\dim(L^\perp \cap W_\beta^-) &= n - \dim(L + W_\beta^+) \\ &= n - (\dim L + \dim W_\beta^+ - \dim(L\cap W_\beta^+)) \\ &= \dim(L\cap W_\beta^+).\end{aligned}$$

Given $\beta \in J_m$ and $\lambda \in \{\pm 1\}^n$, let us define events

$$F_{\beta,\lambda} := \{L\cap W_\beta^+ = \{0\} \text{ and } \operatorname{sign}\big((\Psi_\beta^L)^{-1}(\hat{z})\big) = \lambda\}.$$

Note that $F_{\beta,\lambda}$ is the set of instances where the linearised KKT system (11) admits a unique solution (x,s) and $\operatorname{sign}(x-s) = \lambda$. In particular, $F_{\beta,\beta}$ is the event that (LP) and (LP-D) have unique solutions (x^*,s^*) and that $\operatorname{sign}(x^*-s^*) = \beta$.

Lemma 3.4. *The event that* (LP) *and* (LP-D) *both have unique optimal solutions is precisely the disjoint union*

$$\dot{\bigcup}_{\beta\in J_m} F_{\beta,\beta}. \tag{13}$$

Proof. See Appendix B. □

Lemma 3.5. *For each $\beta \in J_m$,*

$$\mathbb{P}\left(\dot{\bigcup}_{\lambda\in\{\pm1\}^n} F_{\beta,\lambda}\right) = 1. \tag{14}$$

Proof. See Appendix B. □

Given $\beta \in J_m$, and $t \in [0,1]$, define the event

$$G_{\beta,t} = \{c(L, W_\beta^+) \le t\}.$$

We need to show that

$$\mathbb{P}\left(\dot{\bigcup}_{\beta\in J_m}(F_{\beta,\beta}\cap G_{\beta,t})\right) = 2^{-n}\binom{n}{m}\mathbb{P}(G_{\beta_0,t}) \tag{15}$$

where β_0 is an arbitrary element of J_m. Indeed, the left-hand side is the probability that (LP) and (LP-D) are both feasible and that $c(L,W_\beta^+) \le t$, while $\mathbb{P}(G_{\beta_0,t})$ is the probability that $c(L,W_{\beta_0}^+) \le t$ where $W_{\beta_0}^+$ is a fixed subspace of dimension m. First we will need one more lemma.

Lemma 3.6. *$\mathbb{P}(F_{\beta,\lambda}\cap G_{\beta,t})$ does not depend on $\lambda \in \{\pm1\}^n$.*

Proof. In this proof, we use the notation $\cdot$ to denote elementwise multiplication in $\mathbb{R}^n$. Let $\delta \in \{\pm1\}^n$. The new random variables $(L', \hat{x}', \hat{s}') := (\delta\cdot L, \delta\cdot\hat{x}, \delta\cdot\hat{s})$ have the same law as $(L,\hat{x},\hat{s})$, by spherical symmetry of the laws of $\hat{x}$ and of $\hat{s}$, and since the map $z \mapsto \delta\cdot z$ is orthogonal. Define the events $F'_{\beta,\lambda}$ and $G'_{\beta,t}$ as usual, but replacing the random variables $(L,\hat{x},\hat{s})$ by $(L',\hat{x}',\hat{s}')$. We certainly have $\mathbb{P}(F_{\beta,\lambda}\cap G_{\beta,t}) = \mathbb{P}(F'_{\beta,\lambda}\cap G'_{\beta,t})$.

On the other hand, $G_{\beta,t} = G'_{\beta,t}$ because $c(L', W_\beta^+) = c(L, \delta\cdot W_\beta^+) = c(L, W_\beta^+)$.

Moreover, it can be verified from the definition of Ψ_β^L that for all z, $\Psi_\beta^{L'}(z) = \delta\cdot\Psi_\beta^L(\delta\cdot z)$. Thus $F'_{\beta,\lambda} = F_{\beta,\delta\cdot\lambda}$.

Therefore, $\mathbb{P}(F_{\beta,\lambda}\cap G_{\beta,t}) = \mathbb{P}(F'_{\beta,\lambda}\cap G'_{\beta,t}) = \mathbb{P}(F_{\beta,\delta\cdot\lambda}\cap G_{\beta,t})$. □

Proof of Theorem 3.2. Recall that (15) is what we need to prove. We have for any $\beta \in J_m$

$$\begin{aligned}\mathbb{P}(F_{\beta,\beta}\cap G_{\beta,t}) &= \frac{1}{2^n}\sum_{\lambda\in\{\pm1\}^n}\mathbb{P}(F_{\beta,\lambda}\cap G_{\beta,t}) \\ &= \frac{1}{2^n}\mathbb{P}\left(\left(\dot{\bigcup}_{\lambda\in\{\pm1\}^n} F_{\beta,\lambda}\right)\cap G_{\beta,t}\right) \\ &= \frac{1}{2^n}\mathbb{P}(G_{\beta,t})\end{aligned} \tag{16}$$

where in the first line we used Lemma 3.6, and in passing to the third line we used Lemma 3.5. Equation (15) now follows by summing over $\beta \in J_m$ (since $\mathbb{P}(G_{\beta,t})$ does not depend on $\beta \in J_m$). □

3.1. Angles between random subspaces

In view of Theorem 3.2, we want to study the distribution of $c(L,W)$, where W is a fixed m-dimensional subspace of $\mathbb{R}^n$ and L is uniformly distributed in $\operatorname{Gr}(n, n-m)$. Note that for generic L we have $L\cap W = \{0\}$, hence the distribution of $c(L,W)$ matches that of $c_0(L,W)$. Recall from Section 2.3.1 that

$$c_0(L,W) = d_p(L^\perp, W) = \sin\hat{\theta}$$

where $\hat{\theta}$ is the largest principal angle between $L^\perp$ and W.

Since $L^\perp$ is uniformly distributed in $\operatorname{Gr}(n,m)$, $c_0(L,W)$ has the same distribution as the *sine* of the largest principal angle $\hat{\theta}$ between a fixed and a random m-dimensional subspace of $\mathbb{R}^n$ (or equivalently, between two random m-dimensional subspaces of $\mathbb{R}^n$).

The distribution of $\hat{\theta}$ was studied in the paper (Absil et al., 2006), where an expression for the distribution in terms of the the Gaussian hypergeometric function of matrix argument ${}_2F_1$ is derived. See Appendix C for a definition of this

function. Assuming $m \leq n/2$, the probability density of $\hat{\theta}$ is

$$p(\theta) = m(n-m)\frac{\Gamma(\frac{m+1}{2})\Gamma(\frac{n-m+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n+1}{2})}(\sin\theta)^{m(n-m)-1} \times {}_2F_1(\frac{n-m-1}{2}, \frac{1}{2}; \frac{n+1}{2}; (\sin\theta)^2 I_{m-1}) \quad (17)$$

and its cumulative distribution function is

$$\mathbb{P}(\hat{\theta} \leq \theta) = (\sin\theta)^{m(n-m)} \times \left(\frac{{}_2F_1(\frac{n-m}{2}, \frac{1}{2}; \frac{n+1}{2}; (\sin\theta)^2 I_m)}{{}_2F_1(\frac{n-m}{2}, \frac{1}{2}; \frac{n+1}{2}; I_m)}\right). \quad (18)$$

Here, I_m denotes the $m \times m$ identity matrix. From these formulae we derive the following proposition, which is proved in Appendix B. This result may be of independent interest, as the formulae (17) and (18) are somewhat opaque. We have not been able to find it elsewhere in the literature.

Proposition 3.7. *Fix $0 < m < n$ and define the random variable $\hat{\theta}$ as the largest principal angle between two independent random subspaces drawn from the Haar measure on $\mathrm{Gr}(n, n-m)$. We have*

$$\mathbb{E}[\cos\hat{\theta}] < \frac{1}{\sqrt{m(n-m)}}. \quad (19)$$

Moreover, for any $\epsilon \geq 0$

$$1 - \sqrt{m(n-m)}\,\epsilon \leq \mathbb{P}\left(\sin^2\hat{\theta} \leq 1 - \epsilon^2\right) \leq e^{-\frac{m(n-m)}{2}\epsilon^2}. \quad (20)$$

Proof. See Appendix B. □

This tells us that each of the two following inequalities hold with individual probabilities $1 - \delta$:

$$\sqrt{1 - \frac{2\log(1/\delta)}{m(n-m)}} \leq \sin\hat{\theta} \leq \sqrt{1 - \frac{\delta^2}{m(n-m)}}.$$

By the union bound, both inequalities hold simultaneously with probability $1 - 2\delta$.

Corollary 3.8. *Let $m, n, \hat{\theta}$ be as in Proposition 3.7. Then $\cos\hat{\theta}$ has geometric mean bounded above and below*

$$\frac{e^{-1}}{\sqrt{m(n-m)}} \leq \exp\left(\mathbb{E}[\log\cos\hat{\theta}]\right) < \frac{1}{\sqrt{m(n-m)}}.$$

Proof. See Appendix B. □

3.2. Proof of Theorem 1.1

The conditions of Theorem 1.1 guarantee that $L = \ker A$, $\hat{x} = \Pi_{L^\perp}\bar{x}$, and $\hat{s} = \Pi_L c$ follow a symmetric random model as defined in Definition 3.1. Therefore Theorem 3.2 applies: the event E holds with probability $2^{-n}\binom{n}{m}$, and conditional on E, $c_0(L, W^+)$ has the same distribution as $\sin\hat{\theta}$ where $\hat{\theta}$ is the largest principal angle between two random subspaces of dimension m. Also, since Algorithm 1 is equivalent to Algorithm 2, Algorithm 1 converges linearly with rate $r = c(L, W^+) = c_0(L, W^+) < 1$ on E, by Corollary 2.9. The bounds in probability on r follow from the paragraph following Proposition 3.7. □

4. Condition Number Interpretation

Recall that the primal-dual pair of linear programs (LP) and (LP-D) are equivalent to the inverse problem (6) (we recall that (6) reads

$$\Psi^L(z) = \hat{x} - \hat{s} =: \hat{z},$$

where $\Psi^L = \Pi_{L^\perp}\Pi_{\mathbb{R}^n_+} + \Pi_L\Pi_{\mathbb{R}^n_-}$). This was the content of Proposition 2.1.

Suppose (6) has a unique solution z^* with $\operatorname{sign} z^* = \beta \in \{\pm 1\}^n$. Then the nonlinear forward map for LP Ψ^L agrees with the the linear map Ψ^L_β (defined in Definition 3.3) on a neighbourhood of z^*. Moreover, Ψ^L_β is invertible (since z^* is a unique solution). It makes sense to think of the spectral norm of $(\Psi^L_\beta)^{-1}$ as a *local condition number* for the inverse problem (6) (and hence for the primal-dual LP by Proposition 2.1). To see why, observe that if we perturb $\hat{z} \mapsto \hat{z} + \Delta$ by a small enough displacement Δ, the new primal-dual solution is given by $z' = z^* + (\Psi^L_\beta)^{-1}\Delta$, so $\|z' - z^*\| \leq \left\|(\Psi^L_\beta)^{-1}\right\| \|\Delta\|$.

Definition 4.1. Given data $(L, \hat{z})$ such that (6) has a unique solution z^* with $\operatorname{sign} z^* = \beta$, we define

$$\mathcal{C}(L, \hat{z}) = \left\|(\Psi^L_\beta)^{-1}\right\|.$$

Our next result shows this condition number is the *cosecant* of the minimal angle for which the eventual rate of linear convergence of DRS is the *cosine*.

Proposition 4.2. *Suppose (6) has a unique solution z^* with $\operatorname{sign} z^* = \beta$. Then*

$$\mathcal{C}(L, \hat{z})^2 = \frac{1}{1 - c(L, W^+_\beta)^2}.$$

Proof. See Appendix B. □

Remark 4.3. Consider a symmetric random model, as defined in 3.1. By Theorem 3.2 and Corollary 3.8, the expected

logarithm of this condition number (conditional on primal-dual feasibility of the LP) satisfies

$$\log\left(\sqrt{m(n-m)}\right) \leq \mathbb{E}\log\left(\mathcal{C}(L,\hat{z}) \mid E\right) \leq \log\left(\sqrt{m(n-m)}\right) + 1. \quad (21)$$

Here, as in Theorem 1.1, E is the event that (6) has a unique solution z^*.

4.1. Relationship with other condition numbers for LP

In this section, we describe how the local condition number $\mathcal{C}(L,\hat{z})$ (defined in Definition 4.1) for the problem (6) ties in with some other condition numbers for linear programming from the literature. Specifically, we will consider existing notions of condition for the *homogeneous feasibility problem*

$$A^\top y \leq 0, \quad (22)$$

where $A \in \mathbb{R}^{m\times n}$ is a matrix with $(n-m)$-dimensional kernel L. Since our condition number $\mathcal{C}(L,\hat{z})$ depends not only on L but also on $\hat{z}$, we instead consider a "worst-case" condition over $\hat{z}$:

$$\mathcal{C}_{\max}(L) := \sup_{\{\hat{z} \text{ s.t. (6) has unique solution}\}} \mathcal{C}(L,\hat{z}) = \sup_{\beta\in J_m} \left(1 - c_0(L, W_\beta^+)^2\right)^{-1/2}.$$

The following condition numbers are defined for strictly feasible instances of (22), i.e. instances for which there exists y with $A^\top y < 0$ (equivalently $(\operatorname{int}\mathbb{R}^n_-)\cap L^\perp \neq \varnothing$).

1. The *Grassmann condition number* (Belloni & Freund, 2009; Amelunxen & Bürgisser, 2011) is a geometric condition number defined by

$$\mathcal{C}_{\mathrm{Gr}}(L) := \frac{1}{\min\limits_{x\geq 0,\, \|x\|=1} \|\Pi_{L^\perp} x\|}.$$

   By geometric, we mean it depends on A only through its kernel L.

2. *Renegar's condition number* (Renegar, 1995) is

$$\mathcal{C}_{\mathcal{R}}(A) := \frac{\|A\|}{\min\limits_{x\geq 0,\, \|x\|=1} \|Ax\|}.$$

   It controls the number of iterations needed to solve (22) using interior point methods.

3. The *GCC condition number* (Cheung & Cucker, 2001) is

$$\mathcal{C}_{\mathrm{GCC}}(A) := \frac{1}{\max\limits_{\|y\|=1} \min\limits_i \langle y, \bar{a}_i\rangle},$$

   where $\bar{a}_i$ are the normalised columns of A. It controls the convergence rate of the *perceptron algorithm* – see Appendix B of (Cheung et al., 2003).

It has been shown (see (Cheung & Cucker, 2001; Amelunxen & Bürgisser, 2011)) that $\mathcal{C}_{\mathrm{GCC}}(A) \leq \sqrt{n}\,\mathcal{R}(A) \leq \sqrt{n}\,\kappa(A)\,\mathcal{C}_{\mathrm{Gr}}(L)$. Here $\kappa(A)$ is the matrix condition number of A, defined as $\|A\|\left\|A^\dagger\right\|$ where $A^\dagger$ is the Moore-Penrose pseudoinverse of A.

Proposition 4.4. *Let $L \in \mathrm{Gr}(n, n-m)$ be generic in the sense that $L \cap W_\beta^+ = \{0\}$ for any $\beta \in J_m$, and strictly feasible for (22) in the sense that $(\operatorname{int}\mathbb{R}^n_-)\cap L^\perp \neq \varnothing$. Then*

$$\mathcal{C}_{\max}(L) \geq \mathcal{C}_{\mathrm{Gr}}(L).$$

Proof. See Appendix B. □

Therefore, for generic matrices A such that (22) is strictly feasible, we have the chain of inequalities, which lower bounds $\mathcal{C}_{\max}(L)$ in terms of the other notions of condition numbers:

$$\begin{aligned}\mathcal{C}_{\mathrm{GCC}}(A) &\leq \sqrt{n}\,\mathcal{R}(A)\\ &\leq \sqrt{n}\,\kappa(A)\,\mathcal{C}_{\mathrm{Gr}}(L)\\ &\leq \sqrt{n}\,\kappa(A)\,\mathcal{C}_{\max}(L).\end{aligned}$$

We note however that $\mathcal{C}_{\max}(L) \geq \mathcal{C}(L,\hat{z})$, and in fact we have observed numerically that $\mathcal{C}_{\max}(L) \gg \mathcal{C}(L,\hat{z})$. Indeed, for a particular subspace L and $\hat{z} \in \mathbb{R}^n$ such that (6) has a unique solution z^*, we have $\mathcal{C}(L,\hat{z}) = \left(1 - c_0(L, W_{\beta^*}^+)\right)^{-1/2}$, where $\beta^* = \operatorname{sign} z^*$, by definition. On the other hand, $\mathcal{C}_{\max}(L)$ is the maximum of the $\binom{n}{m}$ quantities $\left\{\left(1 - c_0(L, W_\beta^+)^2\right)^{-1/2} : \beta \in J_m\right\}$.

We also mention that various probabilistic analyses of both the GCC condition and the Grassmann condition have been carried out (Cheung & Cucker, 2002; Bürgisser & Amelunxen, 2012; Amelunxen & Bürgisser, 2015a).

5. Further Work

Recently, several authors have considered the DRS algorithm applied to *infeasible* conic programs (Liu et al., 2019; Banjac et al., 2019; Bauschke & Moursi, 2020). It can be shown that in this case, the sequence of differences between subsequent iterates in Algorithm 2 converges to a certificate of infeasibility. Using techniques similar to the ones presented here, one can show that, for random LPs, the rate of convergence to a certificate of infeasibility is eventually linear, and that the distribution of the local linear rate is no worse than in the feasible case treated in this work. Due to space constraints, we omit the details here and plan to discuss these results in a future work.

Also of interest will be to study different random models than that introduced in Definition 3.1, and to study conic programs over more general convex cones than $\mathbb{R}^n_+$ (for example, over the cone of positive semidefinite matrices).

References

Absil, P.-A., Edelman, A., and Koev, P. On the largest principal angle between random subspaces. *Linear Algebra and its Applications*, 414(1):288–294, 2006.

Adler, I. and Berenguer, S. E. Random Linear Programs. Technical report, Operations Research Center, University of California, Berkeley, 01 1981.

Amelunxen, D. and Bürgisser, P. A Coordinate-Free Condition Number for Convex Programming. *SIAM Journal on Optimization*, 22, 05 2011.

Amelunxen, D. and Bürgisser, P. Probabilistic Analysis of the Grassmann Condition Number. *Foundations of Computational Mathematics*, 15:3–51, 2015a.

Amelunxen, D. and Bürgisser, P. Intrinsic volumes of symmetric cones and applications in convex programming. *Mathematical Programming*, 149:105–130, 2015b.

Applegate, D., Díaz, M., Hinder, O., Lu, H., Lubin, M., O'Donoghue, B., and Schudy, W. Practical large-scale linear programming using primal-dual hybrid gradient. *NeurIPS*, 2021.

Aspelmeier, T., Charitha, C., and Luke, D. R. Local Linear Convergence of the ADMM/Douglas-Rachford Algorithms without Strong Convexity and Application to Statistical Imaging. *SIAM J. Imaging Sci.*, 9:842–868, 2016.

Banjac, G., Goulart, P. J., Stellato, B., and Boyd, S. P. Infeasibility Detection in the Alternating Direction Method of Multipliers for Convex Optimization. *J. Optim. Theory Appl.*, 183:490–519, 2019.

Bauschke, H., Bello-Cruz, Y., Phan, H., and Wang, X. The rate of linear convergence of the Douglas-Rachford algorithm for subspaces is the cosine of the Friedrichs angle. *Journal of Approximation Theory*, 06 2014.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Incorporated, 2011.

Bauschke, H. H. and Moursi, W. M. On the Behavior of the Douglas-Rachford Algorithm for Minimizing a Convex Function Subject to a Linear Constraint. *SIAM J. Optim.*, 30:2559–2576, 2020.

Belloni, A. and Freund, R. M. A geometric analysis of Renegar's condition number, and its interplay with conic curvature. *Mathematical Programming*, 119:95–107, 2009.

Boley, D. Local Linear Convergence of the Alternating Direction Method of Multipliers on Quadratic or Linear Programs. *SIAM J. Optim.*, 23:2183–2207, 2013.

Borgwardt, K.-H. The average number of pivot steps required by the simplex-method is polynomial. *Zeitschrift für Operations Research*, 26(1):157–177, 1982.

Bürgisser, P. and Amelunxen, D. Robust smoothed analysis of a condition number for linear programming. *Mathematical Programming*, 131:221–251, 2012.

Bürgisser, P. and Cucker, F. *Condition: The geometry of numerical algorithms*, volume 349. Springer Science & Business Media, 2013.

Cheung, D. and Cucker, F. A new condition number for linear programming. *Mathematical Programming, Series B*, 91:163–174, 10 2001.

Cheung, D. and Cucker, F. Probabilistic Analysis of Condition Numbers for Linear Programming. *Journal of Optimization Theory and Applications*, 114:55–67, 2002.

Cheung, D., Cucker, F., and Hauser, R. On Tail Decay and Moment Estimates of a Condition Number for Random Linear Conic Systems. *SIAM Journal on Optimization*, 15, 11 2003.

Demanet, L. and Zhang, X. Eventual linear convergence of the Douglas-Rachford iteration for basis pursuit. *ArXiv*, abs/1301.0542, 2016.

Deutsch, F. *The Angle Between Subspaces of a Hilbert Space*, pp. 107–130. Springer Netherlands, Dordrecht, 1995.

DLMF. *NIST Digital Library of Mathematical Functions*. URL http://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Eckstein, J. and Bertsekas, D. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 04 1992.

Gabay, D. Applications of the Method of Multipliers to Variational Inequalities. In Fortin, M. and Glowinski, R. (eds.), *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, volume 15 of *Studies in Mathematics and Its Applications*, pp. 299–331. Elsevier, 1983.

Goldman, A. J. and Tucker, A. W. *Theory of Linear Programming*, pp. 53–98. Princeton University Press, 1957.

Liang, J., Fadili, J., and Peyré, G. Local Convergence Properties of Douglas-Rachford and Alternating Direction Method of Multipliers. *Journal of Optimization Theory and Applications*, 172, 03 2017.

Lions, P. L. and Mercier, B. Splitting Algorithms for the Sum of Two Nonlinear Operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

Liu, Y., Ryu, E. K., and Yin, W. A new use of Douglas-Rachford splitting for identifying infeasible, unbounded, and pathological conic programs. *Mathematical Programming*, pp. 1–29, 2019.

Lu, H. and Yang, J. Linear convergence of stochastic primal dual methods for linear programming using variance reduction and restarts. *arXiv preprint arXiv:2111.05530*, 2021.

O'Donoghue, B., Chu, E., Parikh, N., and Boyd, S. P. Conic Optimization via Operator Splitting and Homogeneous Self-Dual Embedding. *Journal of Optimization Theory and Applications*, 169:1042–1068, 2016.

Renegar, J. Incorporating Condition Measures into the Complexity Theory of Linear Programming. *SIAM J. Optim.*, 5:506–524, 1995.

Smale, S. On the average number of steps of the simplex method of linear programming. *Mathematical programming*, 27(3):241–262, 1983.

A. The Douglas-Rachford Splitting Algorithm

The Douglas-Rachford splitting algorithm (DRS) was introduced by Lions and Mercier in (Lions & Mercier, 1979). It is commonly applied to minimise the sum of two proper, lower-semicontinuous convex functions $f, g : \mathbb{R}^n \to (-\infty, +\infty]$,

$$\underset{x\in\mathbb{R}^n}{\text{minimise}}\, f(x) + g(x), \tag{23}$$

assuming that efficient methods for computing the *proximal operators* of f and g are available. The proximal operator of a proper, lower-semicontinuous convex function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is

$$\operatorname{prox}_f(x) := \arg\min_y \left(f(y) + \frac{1}{2}\|x - y\|^2 \right).$$

Let $z^0 \in \mathbb{R}^n$ be arbitrary. Then the general DRS algorithm consists of the iteration listed in Algorithm 3.

Algorithm 3 Douglas-Rachford

Input: Initial point z^0
$k \leftarrow 0$
repeat
  $s^k \leftarrow z^k - \operatorname{prox}_g(z^k)$
  $x^k \leftarrow \operatorname{prox}_f(z^k - 2s^k)$
  $z^{k+1} \leftarrow x^k + s^k$
  $k \leftarrow k + 1$
until convergence, or other termination criterion satisfied

We will need the following notions.

Definition A.1. Let $T : \mathbb{R}^n \to \mathbb{R}^n$. T is *nonexpansive* if $\forall x,\, y \in \mathbb{R}^n$

$$\|Tx - Ty\| \le \|x - y\|.$$

T is *firmly nonexpansive* if $\forall x,\, y \in \mathbb{R}^n$

$$\|Tx - Ty\|^2 + \|(\mathrm{Id} - T)(x) + (\mathrm{Id} - T)(y)\|^2 \le \|x - y\|^2.$$

Notice that the iteration in Algorithm 3 can be written in terms of a single operator

$$T = \mathrm{Id} - \operatorname{prox}_g + \operatorname{prox}_f \circ (2\operatorname{prox}_g - \mathrm{Id})$$

i.e. $z^{k+1} = Tz^k$. We have the following.

Proposition A.2. *The Douglas-Rachford operator T is firmly nonexpansive.*

Proof. See Propositions 4.31 and 12.28 in (Bauschke & Combettes, 2011). □

The following important result says that if T has fixed points, then the sequence z^k converges to one of them.

Theorem A.3. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be firmly nonexpansive, and suppose $\operatorname{Fix} T \neq \varnothing$ (equivalently if $0 \in \operatorname{range}(\operatorname{Id} - T)$). Let $z^0 \in \mathbb{R}^n$ and for $k \in \mathbb{N}$ define $z^k = T^k z^0$. Then the sequence $(z^k)_{k\geq 0}$ converges to some $z^* \in \operatorname{Fix} T$.*

Proof. See Example 5.18 in (Bauschke & Combettes, 2011). □

A.1. Derivation of Algorithm 2

There are multiple ways to cast an LP in the form (LP) into the form (23). One of the simplest is to set $f(x) = I_{L+\hat{x}}(x) + \langle \hat{s}, x\rangle$ and $g(x) = I_{\mathbb{R}^n_+}(x)$. Here, given a nonempty closed convex set C, the *indicator function* I_C is defined by $I_C(x) = 0$ if $x \in C$ and $+\infty$ otherwise. We write $\Pi_C = \operatorname{prox}_{I_C}$ for the metric projection onto C with respect to the Euclidean distance.

We have

$$\operatorname{prox}_f(z) = \Pi_L(z) + \hat{x} - \hat{s},$$

and

$$\operatorname{prox}_g(z) = \Pi_{\mathbb{R}^n_+}(z).$$

We now consider the application of Algorithm 3 (Douglas-Rachford) to this choice of f and g. Given an initial point $z^0 \in \mathbb{R}^n$, the Douglas-Rachford sequence (c.f. Algorithm 3) is controlled by the update rule

$$\begin{aligned} z^{k+1} &= \Pi_L(2\Pi_{\mathbb{R}^n_+}(z^k) - z^k) + \hat{x} - \hat{s} + \Pi_{\mathbb{R}^n_-}(z^k) \\ &= \Pi_L \Pi_{\mathbb{R}^n_+}(z^k) + \Pi_{L^\perp}\Pi_{\mathbb{R}^n_-}(z^k) + \hat{z}. \end{aligned} \tag{24}$$

(recall that $\hat{z} := \hat{x} - \hat{s}$). This is precisely Algorithm 2.

B. Proofs

Proposition 2.1. Observe that (6) can be rearranged to

$$\Pi_{L^\perp}(x) - \hat{x} = \Pi_L(s) - \hat{s}. \tag{25}$$

For the if direction, if x is feasible for (LP) and s is feasible for (LP-D), then they certainly satisfy $x - \hat{x} \in L$, $s - \hat{s} \in L^\perp$, so both sides of (25) are 0.

For the only if direction, assume (25) holds. Since the left hand side lies in L and the right hand side lies in $L^\perp$, both sides must vanish and hence the pair (x, s) are primal and dual feasible. Moreover $\langle x, s\rangle = \langle \Pi_{\mathbb{R}^n_+}(z), \Pi_{\mathbb{R}^n_+}(-z)\rangle = 0$, which means x and s are complementary, hence optimal. □

Proposition 2.7. First note that

$$\begin{aligned} \Pi_{W_1} - \Pi_{W_2} &= \Pi_{W_1}(\operatorname{Id} - \Pi_{W_2}) - (\operatorname{Id} - \Pi_{W_1})\Pi_{W_2} \\ &= \Pi_{W_1}\Pi_{W_2^\perp} - \Pi_{W_1^\perp}\Pi_{W_2}, \end{aligned}$$

so $d_p(W_1, W_2) = \left\|\Pi_{W_1}\Pi_{W_2^\perp} - \Pi_{W_1^\perp}\Pi_{W_2}\right\|$. Using Lemma B.1, we obtain $d_p(W_1, W_2) = \max\{c_0(W_1^\perp, W_2), c_0(W_1, W_2^\perp)\}$. We will show that $c_0(W_1^\perp, W_2) = c_0(W_1, W_2^\perp)$.

Using Fact 2.4 and $W_1^\perp \cap W_2 = \{0\}$, we have $c_0(W_1^\perp, W_2) = c(W_1^\perp, W_2) < 1$. By Proposition 2.5 and Fact 2.4, we have

$$\begin{aligned} c_0(W_1^\perp, W_2) &= c(W_1^\perp, W_2) \\ &= c(W_1, W_2^\perp) = c_0(W_1, W_2^\perp). \end{aligned}$$

It now follows that $c_0(W_1^\perp, W_2) = d_p(W_1, W_2)$. □

Lemma B.1. *For* any *subspaces W_1, W_2 of $\mathbb{R}^n$, we have*

$$\left\|\Pi_{W_1}\Pi_{W_2^\perp} \pm \Pi_{W_1^\perp}\Pi_{W_2}\right\| = \max\{c_0(W_1^\perp, W_2), c_0(W_1, W_2^\perp)\}.$$

Proof. Write $M = \Pi_{W_1}\Pi_{W_2^\perp} \pm \Pi_{W_1^\perp}\Pi_{W_2}$. Denote by M^* its adjoint operator. We have

$$MM^* = \Pi_{W_1}\Pi_{W_2^\perp}\Pi_{W_1} + \Pi_{W_1^\perp}\Pi_{W_2}\Pi_{W_1^\perp}.$$

The largest eigenvalue of this self-adjoint operator is clearly $\max\{c_0(W_1^\perp, W_2)^2, c_0(W_1, W_2^\perp)^2\}$. □

Theorem 2.8. Combined with Proposition 2.1, Theorem A.3 guarantees that if (LP) and (LP-D) are both feasible, then Algorithm 2 converges to some z^* satisfying (6).

Let

$$K > \max\{k \mid \exists\, i \ \operatorname{sign} z_i^k \neq \operatorname{sign} z_i^*\},$$

which is finite because $z^k \to z^*$. Then for any $k \geq K$

$$\begin{aligned} z^{k+1} - z^* &= T_{\hat{z}}^L(z^k) - T_{\hat{z}}^L(z^*) \\ &= (\Pi_L\Pi_{W^+} + \Pi_{L^\perp}\Pi_{W^-})(z^k - z^*) \\ &= M(z^k - z^*), \end{aligned}$$

where

$$M := \Pi_L\Pi_{W^+} + \Pi_{L^\perp}\Pi_{W^-} = DT_{\hat{z}}^L|_{z^*} \tag{26}$$

is the Jacobian of $T_{\hat{z}}^L$ at z^*. By Lemma B.4, M is normal – c.f. Lemma 6.2(iii) in (Liang et al., 2017) – so it can be diagonalised by a unitary matrix U:

$$M = U\Lambda U^\dagger,$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Also by Lemma B.4, M is firmly nonexpansive, so for any vector u

$$\|Mu\|^2 + \|Mu - u\|^2 \leq \|u\|^2.$$

It follows that $\|Mu\| \leq \|u\|$, and if $\|Mu\| = \|u\|$ then $Mu = u$. So we must have $|\lambda_i| \leq 1$ for each λ_i, and if $|\lambda_i| = 1$, then $\lambda_i = 1$.

We now show that $\max\{|\lambda_i|\,,\, \lambda_i \neq 1\} = r$ where $r = c(L, W^+) = c(L^\perp, W^-)$.

Observe, by the definition of M in (26), that $\{u \mid Mu = u\} \supseteq L \cap W^+ + L^\perp \cap W^-$. On the other hand, if $Mu = u$ then $\Pi_L u = \Pi_L \Pi_{W^+} u$, hence $\Pi_{W^+} u \in L$. Similarly $\Pi_{W^-} u \in L^\perp$, so in fact $\{u \mid Mu = u\} = L \cap W^+ + L^\perp \cap W^-$. Consequently, $\{u \mid Mu = u\}^\perp = (L \cap W^+)^\perp \cap (L^\perp \cap W^-)^\perp$. Let $\hat{u} \in \{u \mid Mu = u\}^\perp$, i.e. $\hat{u} \in (L \cap W^+)^\perp$ and $\hat{u} \in (L^\perp \cap W^-)^\perp$. Then

$$\begin{aligned}
\|M\hat{u}\|^2 &= \|M\Pi_{W^+}(\hat{u}) + M\Pi_{W^-}(\hat{u})\|^2 \\
&= \|\Pi_L \Pi_{W^+}(\hat{u}) + \Pi_{L^\perp}\Pi_{W^-}(\hat{u})\|^2 \\
&= \|\Pi_L \Pi_{W^+}(\hat{u})\|^2 + \|\Pi_{L^\perp}\Pi_{W^-}(\hat{u})\|^2 \\
&\leq c(L, W^+)^2 \|\Pi_{W^+}(\hat{u})\|^2 + \\
&\qquad\qquad c(L^\perp, W^-)^2 \|\Pi_{W^-}(\hat{u})\|^2 \\
&= r^2 \|\hat{u}\|^2 .
\end{aligned}$$

The inequality holds because

$$\Pi_{W^+}(\hat{u}) \in (L \cap W^+)^\perp + W^- = (L \cap W^+)^\perp$$

(using the fact that $\hat{u} \in (L \cap W^+)^\perp$) and

$$\Pi_{W^-}(\hat{u}) \in (L^\perp \cap W^-)^\perp + W^+ = (L^\perp \cap W^-)^\perp$$

(using the fact that $\hat{u} \in (L^\perp \cap W^-)^\perp$).

This proves that $\max\{|\lambda_i|\,,\, \lambda_i \neq 1\} \leq r$. To see that equality holds, choose $\bar{u} \in W^+ \cap (L \cap W^+)^\perp$ such that $\|\Pi_L \bar{u}\| = c(L, W^+)\,\|\bar{u}\|$. Then certainly $\bar{u} \in (L \cap W^+)^\perp$, and $\bar{u} \in W^+ \subseteq (L^\perp \cap W^-)^\perp$. So $\bar{u} \in \{u \mid Mu = u\}^\perp$ and $\|M\bar{u}\| = r\,\|\bar{u}\|$. It follows that for some i, $r \leq |\lambda_i| < 1$, hence $|\lambda_i| = r$.

We can now prove (10). Let $k \geq K$. From the diagonalisation of M, we have for all $l \geq 0$

$$z^{K+l} - z^* = U \Lambda^l U^\dagger (z^K - z^*). \tag{27}$$

We know that $z^k - z^* \to 0$, so it must be the case that $z^K - z^* \in \{u \mid Mu = u\}^\perp$. Then $z^k - z^* \in \{u \mid Mu = u\}^\perp$ for all $k \geq K$, and

$$\left\|z^{k+1} - z^*\right\| = \left\|M(z^k - z^*)\right\| \leq r \left\|z^k - z^*\right\|$$

which proves (10).

To see that r is best possible, first observe we have shown that r is the largest singular value of the linear map M which is strictly less than 1. Let U' be an eigenvector of M^*M with eigenvalue r^2. By Lemma B.4, M commutes with its adjoint, so we can write

$$\begin{aligned}
\left\|M^k U'\right\|^2 &= \langle U', (M^*)^k M^k U' \rangle \\
&= \langle U', (M^*M)^k U' \rangle \\
&= r^{2k}.
\end{aligned}$$

By setting $z^0 = z^* - \epsilon U'$ for sufficiently small ϵ, we can arrange that z^k has the same sign as z^* for every k. It suffices to take $\epsilon < \min_i |z_i^*|$. □

Remark B.2. The first part of our proof shows that from iteration K onward, the algorithm is nothing other than DRS for finding the intersection of two linear subspaces. This was shown to converge with linear rate equal to the cosine of the Friedrichs angle in (Bauschke et al., 2014). We proved this fact again above for completeness.

Remark B.3. Alternatively, Theorem 2.8 can be derived from results in (Liang et al., 2017).

Corollary 2.9. Any primal-dual feasible LP has a strictly complementary pair of solutions (Goldman & Tucker, 1957). Since z^* is unique, it follows from Proposition 2.1 that z^* has full support. Thus W^+ and W^- are uniquely defined and are orthogonal complements, hence (10) holds.

Again using the fact that z^* has full support, z^* has a neighbourhood on which Ψ^L agrees with its Jacobian $D\Psi^L|^*_{z^*} = \Pi_{L^\perp}\Pi_{W^+} + \Pi_L \Pi_{W^-}$. Since z^* is a unique solution to $\Psi^L(z^*) = \hat{z}$, it follows that $D\Psi^L|^*_{z^*}$ is invertible. This is only possible if $L \cap W^+ = \{0\} = L^\perp \cap W^-$, which implies $c(L, W^+) = c_0(L, W^+)$ and $c(L^\perp, W^-) = c_0(L^\perp, W^-)$. Finally, $c(L, W^+) = c(L^\perp, W^-)$ by Proposition 2.5. □

Lemma B.4. *Let L, W be subspaces of $\mathbb{R}^n$. The linear map $\Pi_L \Pi_W + \Pi_{L^\perp}\Pi_{W^\perp}$ is normal (commutes with its adjoint) and firmly nonexpansive.*

Lemma B.4. Write $M = \Pi_L \Pi_W + \Pi_{L^\perp}\Pi_{W^\perp}$. We have $MM^* = \Pi_L \Pi_W \Pi_L + \Pi_{L^\perp}\Pi_{W^\perp}\Pi_{L^\perp}$ and $M^*M = \Pi_W \Pi_L \Pi_W + \Pi_{W^\perp}\Pi_{L^\perp}\Pi_{W^\perp}$. It is enough to check that both operators have the same effect on each $w \in W$ and each $w' \in W^\perp$. Because of the symmetry of the problem it is enough to check this for $w \in W$.

On the one hand, $M^*Mw = \Pi_W \Pi_L w$. On the other,

$$\begin{aligned}
MM^*w &= \Pi_L \Pi_W \Pi_L w + \Pi_{L^\perp}\Pi_{W^\perp}(w - \Pi_L w) \\
&= \Pi_L(\Pi_W \Pi_L w) - \Pi_{L^\perp}\Pi_{W^\perp}\Pi_L w \\
&= \Pi_L(\Pi_W \Pi_L w) + \Pi_{L^\perp}(\Pi_W \Pi_L w - \Pi_L w) \\
&= \Pi_L(\Pi_W \Pi_L w) + \Pi_{L^\perp}(\Pi_W \Pi_L w) \\
&= \Pi_W \Pi_L w.
\end{aligned}$$

This proves that M is normal.

To see that M is firmly nonexpansive, observe that

$$\begin{aligned} MM^* + (\mathrm{Id} - M)(\mathrm{Id} - M^*) \\ = \Pi_L \Pi_W \Pi_L + \Pi_{L^\perp} \Pi_{W^\perp} \Pi_{L^\perp} + \\ \Pi_{L^\perp} \Pi_W \Pi_{L^\perp} + \Pi_L \Pi_{W^\perp} \Pi_L \\ = \mathrm{Id}\,. \end{aligned} \quad (28)$$

So

$$\|M\hat{u}\|^2 + \|(\mathrm{Id} - M)\hat{u}\|^2 = \|\hat{u}\|^2\,.$$

□

Lemma 3.4. Since $F_{\beta,\beta}$ is the event that (LP) and (LP-D) have unique solutions (x^*, s^*) and that $\mathrm{sign}(x^* - s^*) = \beta$, the events $F_{\beta,\beta}$ are certainly disjoint. Moreover, each $F_{\beta,\beta}$ is contained in the event that (LP) and (LP-D) both have unique optimal solutions. It remains to show that if (LP) and (LP-D) both have unique optimal solutions, then $\beta := \mathrm{sign}(x^* - s^*) \in J_m$ and $L \cap W_\beta^+ = \{0\}$.

Since the solutions are unique, they are strictly complementary, so $\beta \in \{\pm 1\}^n$. Since Ψ^L agrees with $\Pi_{L^\perp}\Pi_{W_\beta^+} + \Pi_L \Pi_{W_\beta^-}$ on a neighbourhood of $z^* := x^* - s^*$, the linear map $\Pi_{L^\perp}\Pi_{W_\beta^+} + \Pi_L \Pi_{W_\beta^-}$ must be invertible (otherwise z^* would not be an *unique* solution). Therefore

$$\mathrm{rank}(\Pi_{L^\perp}\Pi_{W_\beta^+}) = m \text{ and } \mathrm{rank}(\Pi_L \Pi_{W_\beta^-}) = n - m.$$

But this implies $\dim W_\beta^+ \geq m$ and $\dim W_\beta^- \geq n - m$, which means $|\beta| = m$, ie.e, $\beta \in J_m$. It now also follows from (12) that $L \cap W_\beta^+ = \{0\}$. □

Lemma 3.5. Fix $\beta \in J_m$. The union $\dot{\bigcup}_{\lambda \in \{\pm 1\}^n} F_{\beta,\lambda}$ is clearly disjoint, by definiton of $F_{\beta,\lambda}$.

Note that

$$\dot{\bigcup}_{\lambda \in \{\pm 1\}^n} F_{\beta,\lambda} = \{L \cap W_\beta^+ = 0\} \cap \{\forall i\, (\Psi_\beta^L)^{-1}(\hat{z})_i \neq 0\}.$$

Since $\dim W_\beta^+ = m$ and $\dim L = n - m$, L generically intersects W_β^+ only at 0. Therefore $\mathbb{P}(L \cap W_\beta^+ = 0) = 1$.

On the other hand, given L satisfying $L \cap W_\beta^+ = 0$, Ψ_β^L is invertible and for each $i \in [n]$,

$$\mathbb{P}(\hat{z} \in \Psi_\beta^L(\{z \,:\, z_i = 0\})) = 0,$$

since $\hat{z}$ has a spherically symmetric law giving no mass to 0.

By conditioning on L, we deduce that

$$\mathbb{P}\left(\dot{\bigcup}_{\lambda \in \{\pm 1\}^n} F_{\beta,\lambda}\right) = 1.$$

□

Proposition 3.7. First note that we may, without loss of generality, assume that $m \leq \frac{n}{2}$. Indeed it follows from Definition 2.6 that the largest principal angles between two subspaces equals the largest principal angles between their respective complements. So the distribution of $\hat{\theta}$ is unchanged by replacing m by $n - m$. On the other hand, all the other quantities in the statement are also invariant in this substitution. So we may as well assume that $m \leq \frac{n}{2}$, which allows us to use expressions for the distribution of $\hat{\theta}$ from (Absil et al., 2006).

Changing variables in (17), we see that the probability density of $\cos\hat{\theta}$ is

$$\begin{aligned} p_{\cos}(\cos\theta) = m(n-m)\frac{\Gamma(\frac{m+1}{2})\Gamma(\frac{n-m+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n+1}{2})} \times \\ (\sin\theta)^{m(n-m)} \times \\ {}_2F_1(\frac{n-m-1}{2}, \frac{1}{2}; \frac{n+1}{2}; (\sin\theta)^2 I_{m-1}). \end{aligned} \quad (29)$$

Observe that when $a, b, c > 0$ are integer multiples of $\frac{1}{2}$ with $c \geq m/2$, the function $t \mapsto {}_2F_1(a, b; c; tI_{m-1})$ equals a power series with nonnegative coefficients on the interval $[0, 1)$ (see Appendix C), so it is increasing on this interval. We will use this observation twice. First applying it to (18), we see that

$$\mathbb{P}(\sin\hat{\theta} \leq \sin\theta) \leq (\sin\theta)^{m(n-m)}. \quad (30)$$

Next applying it to (29), we see that

$$p_{\cos}(\cos\theta) \leq p_{\cos}(0)(\sin\theta)^{m(n-m)}. \quad (31)$$

It is easily shown, using Fact C.2, that

$$p_{\cos}(0) = \frac{m(n-m)}{2}\frac{\Gamma(\frac{m+1}{2})\Gamma(\frac{n-m+1}{2})}{\Gamma(\frac{m+2}{2})\Gamma(\frac{n-m+2}{2})}.$$

Gautschi's inequality says that for any positive real number x and any $t \in (0,1)$ $x^{1-t} < \frac{\Gamma(x+1)}{\Gamma(x+t)} < (x+1)^{1-t}$. We use it to bound

$$p_{\cos}(0) < \sqrt{m(n-m)}.$$

We can now estimate

$$\begin{aligned} \mathbb{E}[\cos\hat{\theta}] &= \int_0^1 t\, p_{\cos}(t) dt \\ &\leq \sqrt{m(n-m)} \int_0^1 t(1-t^2)^{\frac{m(n-m)}{2}} dt \\ &= \sqrt{m(n-m)} \left[-\frac{(1-t^2)^{\frac{m(n-m)+2}{2}}}{m(n-m)+2} \right]_0^1 \\ &= \frac{\sqrt{m(n-m)}}{m(n-m)+2} < \frac{1}{\sqrt{m(n-m)}}, \end{aligned}$$

which proves (19). The first inequality used (31) and $p_{\cos}(0) < \sqrt{m(n-m)}$.

For the lower bound of (20), we first deduce from (31) the weaker bound $p_{\cos}(\cos\theta) \le p_{\cos}(0) < \sqrt{m(n-m)}$. Integrating this, we have $\mathbb{P}(\cos\hat{\theta} \le \epsilon) \le \sqrt{m(n-m)}\,\epsilon$ which is equivalent to the lower of (20).

Now we prove the upper of (20). Starting from (30), we have

$$\begin{aligned}\mathbb{P}(\sin^2\hat{\theta} \le 1-\epsilon^2) &\le (1-\epsilon^2)^{\frac{m(n-m)}{2}} \\ &= \exp\left(\frac{m(n-m)}{2}\log(1-\epsilon^2)\right) \\ &\le e^{-\frac{m(n-m)}{2}\epsilon^2},\end{aligned}$$

where in the last line we used $\log(1-\epsilon^2) \le -\epsilon^2$. □

Corollary 3.8. The upper bound is just Jensen's inequality applied to (19) in Proposition 3.7.

For the lower bound, we will make use of the lower bound in Proposition 3.7.

$$\begin{aligned}\mathbb{E}[-\log\cos\hat{\theta}] &= \int_0^\infty \mathbb{P}(-\log\cos\hat{\theta} \ge t)dt \\ &\le \int_0^{\log\sqrt{m(n-m)}} dt \quad + \\ &\qquad \sqrt{m(n-m)}\int_{\log\sqrt{m(n-m)}}^\infty e^{-t}dt \\ &= \log\sqrt{m(n-m)} + 1.\end{aligned}$$

□

Proposition 4.2. We have

$$\begin{aligned}\mathcal{C}(L,\hat{z})^{-2} &= \min_{\|z\|=1}\left\|\Psi_\beta^L(z)\right\|^2 \\ &= \min_{\|z\|=1}\left\|\Pi_{L^\perp}\Pi_{W_\beta^+}z + \Pi_L\Pi_{W_\beta-}z\right\|^2 \\ &= \min_{\|z\|=1}\left(\left\|\Pi_{L^\perp}\Pi_{W_\beta^+}z\right\|^2 + \left\|\Pi_L\Pi_{W_\beta-}z\right\|^2\right) \\ &= \min_{\|z\|=1}\left(1 - \left\|\Pi_L\Pi_{W_\beta^+}z\right\|^2 - \left\|\Pi_{L^\perp}\Pi_{W_\beta-}\right\|^2\right) \\ &= 1 - \max_{\|z\|=1}\left(\left\|\Pi_L\Pi_{W_\beta^+}z\right\|^2 + \left\|\Pi_{L^\perp}\Pi_{W_\beta-}\right\|^2\right) \\ &= 1 - c(L,W_\beta^+)^2.\end{aligned}$$

In the last equality, we used the fact that

$$c_0(L,W_\beta^+) = c(L,W_\beta^+) = c(L^\perp,W_\beta^-) = c_0(L^\perp,W_\beta^-).$$

by Corollary 2.9.

□

Proposition 4.4. Let x minimise $\|\Pi_{L^\perp}x\|$ subject to $x \in \mathbb{R}^n_+$ and $\|x\| = 1$. We claim that $|\mathrm{supp}(x)| \le m$. Suppose instead that $|\mathrm{supp}(x)| > m$, and using the notation of Section 3, define $\bar{\beta} \in \{\pm1\}^n$ by $\bar{\beta}_i = +1 \iff i \in \mathrm{supp}(x)$. Then there exists $w \in L \cap W_{\bar{\beta}}^+ \setminus \{0\}$, since $\dim(L \cap W_{\bar{\beta}}^+) \ge \dim L + \dim W_{\bar{\beta}}^+ - n > 0$. Without loss of generality, $\langle w, x\rangle \ge 0$. For $\epsilon > 0$ small enough, we have $x + \epsilon w \in \mathbb{R}^n_+$, howewever

$$\frac{\|\Pi_{L^\perp}(x+\epsilon w)\|^2}{\|x+\epsilon w\|^2} = \frac{\|\Pi_{L^\perp}x\|^2}{1 + \epsilon\langle x,w\rangle + \|\epsilon w\|^2} < \|\Pi_{L^\perp}x\|^2$$

which contradicts our assumption that x minimizes $\|\Pi_{L^\perp}x\|$ subject to $x \in \mathbb{R}^n_+$ and $\|x\| = 1$. It follows that $|\mathrm{supp}(x)| \le m$.

Now choose $\beta \in \{\pm1\}^n$ such that the set $\{i \mid \beta_i = 1\}$ has cardinality exactly m and contains $\mathrm{supp}(x)$. By (12), Ψ_β^L is invertible. Choose any $z \in \mathbb{R}^n$ such that $\mathrm{sign}(z) = \beta$. Since Ψ_L agrees with Ψ_β^L on small enough neighbourhoods of z, and Ψ_β^L is linear and invertible, z is the unique solution to $\Psi^L(z)\hat{z}$ in a neighbourhood of z. Since the set of solutions to $\Psi^L(z)\hat{z}$ is convex, z is the unique solution to $\Psi^L(z)\hat{z}$ in general. So, $\mathcal{C}_{\max}(L) \ge \mathcal{C}(L,\hat{z}) = (1 - c_0(L,W_\beta^+)^2)^{-1/2}$. But $x \in W_\beta^+$ and $\|x\| = 1$, so

$$\begin{aligned}\mathcal{C}_{\max}(L)^{-1} &\le \sqrt{1 - c_0(L,W_\beta^+)^2)} \\ &\le \sqrt{1 - \|\Pi_L x\|)^2)} \\ &= \|\Pi_{L^\perp}x\| \\ &= \mathcal{C}_{\mathrm{Gr}}(L)^{-1}.\end{aligned}$$

□

C. The Gaussian hypergeometric function of matrix argument

We first introduce some notation/definitions.

- For $x \in \mathbb{C}$, $k \in \mathbb{N}$ we have the rising factorial $(x)_k = x(x+1)\dots(x+k-1)$ and $(x)_0 = 1$.
- Given a partition κ of length m (i.e. a nonincreasing sequence of nonnegative integers $\kappa = (k_1,\dots,k_m)$), and $x \in \mathbb{C}$, the *partition shifted factorial* is $[x]_\kappa = \prod_{j=1}^m (x - \frac{j-1}{2})_{k_j}$.
- Given a partition κ of length m, the zonal polynomial $Z_\kappa : \mathbf{S}_m \to \mathbb{R}$ is a homogeneous polynomial of degree m, which depends only (and symmetrically) on the eigenvalues of its argument. Here $\mathbf{S}_m$ is the set of real symmetric $m \times m$ matrices. For arguments which are

scalar multiples of the identity, we have

$$Z_\kappa(sI_m) = s^{|\kappa|}\,|\kappa|!4^{|\kappa|}[m/2]_\kappa \times \frac{\prod_{1\le j<k\le l(\kappa)}(2k_j - 2k_l + l - j)}{\prod_{j=1}^{l(\kappa)}(2k_j + l(j) - j)!}.$$

Here $|\kappa| = k_1 + \cdots + k_m$ and $l(\kappa) = \max\{i \mid k_i > 0\}$.

Definition C.1 (Gaussian hypergeometric function of matrix argument)**.** Let $m \in \mathbb{N}$, and let $a, b, c \in \mathbb{C}$ be such that for every $j \in [m] = \{1, \ldots, m\}$ $\frac{j+1}{2} - c \notin \mathbb{N}$. Then we define, for symmetric $m \times m$ matrices T with spectral norm $\|T\| < 1$,

$${}_2F_1(a, b; c; T) = \sum_{k=0}^{\infty} \sum_{|\kappa|=k} \frac{[a]_\kappa [b]_\kappa}{k!\,[c]_\kappa} Z_\kappa(T).$$

The condition on c ensures that the denominator does not vanish in any term. The series converges for $\|T\| < 1$.

Fact C.2 (see Eq. 35.7.7 in (DLMF))**.** Whenever $\mathrm{Re}(c), \mathrm{Re}(c - a - b) > \frac{m-1}{2}$,

$${}_2F_1(a, b; c; I_m) = \frac{\Gamma_m(c)\Gamma_m(c - a - b)}{\Gamma_m(c - a)\Gamma_m(c - b)}.$$