JMLR: Workshop and Conference Proceedings vol 23 (2012) 11.1–11.24 25th Annual Conference on Learning Theory

Toward a Noncommutative Arithmetic-geometric Mean Inequality: Conjectures, Case-studies, and Consequences

Benjamin Recht BRECHT@CS.WISC.EDU
Christopher Ré CHRISRE@CS.WISC.EDU
Department of Computer Sciences, University of Wisconsin-Madison

Editor: Shie Mannor, Nathan Srebro, Robert C. Williamson

Abstract

Randomized algorithms that base iteration-level decisions on samples from some pool are ubiquitous in machine learning and optimization. Examples include stochastic gradient descent and randomized coordinate descent. This paper makes progress at theoretically evaluating the difference in performance between sampling with- and without-replacement in such algorithms. Focusing on least means squares optimization, we formulate a *noncommutative arithmetic-geometric mean inequality* that would prove that the expected convergence rate of without-replacement sampling is faster than that of with-replacement sampling. We demonstrate that this inequality holds for many classes of random matrices and for some pathological examples as well. We provide a deterministic worst-case bound on the gap between the discrepancy between the two sampling models, and explore some of the impediments to proving this inequality in full generality. We detail the consequences of this inequality for stochastic gradient descent and the randomized Kaczmarz algorithm for solving linear systems.

Keywords: Positive definite matrices. Matrix Inequalities. Randomized algorithms. Random matrices. Optimization. Stochastic gradient descent.

1. Introduction

Randomized sequential algorithms abound in machine learning and optimization. The most famous is the stochastic gradient method (see Bottou, 1998; Bertsekas, 2012; Nemirovski et al., 2009; Shalev-Shwartz and Srebro, 2008), but other popular methods include algorithms for alternating projections (see Strohmer and Vershynin, 2009; Leventhal and Lewis, 2010), proximal point methods (see Bertsekas, 2011), coordinate descent (see Nesterov, 2010) and derivative free optimization (see Nesterov, 2011; Nemirovski and Yudin, 1983). In all of these cases, an iterative procedure is derived where, at each iteration, an *independent* sample from some distribution determines the action at the next stage. This sample is selected *with-replacement* from a pool of possible options.

In implementations of many of these methods, however, practitioners often choose to break the independence assumption. For instance, in stochastic gradient descent, many implementations pass through each item exactly once in a random order (i.e., according to a random permutation). In randomized coordinate descent, one can cycle over the coordinates in a random order. These strategies, employing *without-replacement* sampling, are often easier to implement efficiently, guarantee that every item in the data set is touched at least once, and often have better empirical performance than their with-replacement counterparts (see Bottou, 2009).

Unfortunately, the analysis of without-replacement sampling schemes are quite difficult. The independence assumption underlying with-replacement sampling provides an elegant Markovian framework for analyzing incremental algorithms. The iterates in without-replacement sampling are correlated, and studying them requires sophisticated probabilistic tools. Consequently, most of the analyses without-replacement optimization assume that the iterations are assigned deterministically. Such deterministic orders might incur exponentially worse convergence rates than randomized methods (Nedic and Bertsekas, 2000), and deterministic orders still require careful estimation of accumulated errors (see Luo, 1991; Tseng, 1998). The goal of this paper is to make progress towards patching the discrepancy between theory and practice of without-replacement sampling in randomized algorithms.

In particular, in many cases, we demonstrate that without-replacement sampling outperforms with-replacement sampling provided a *noncommutative* version of the arithmetic-geometric mean inequality holds. Namely, if $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$ are a collection of $d \times d$ positive semidefinite matrices, we define the arithmetic and (symmetrized) geometric means to be

$$\boldsymbol{M}_A := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{A}_i\,, \quad \text{and} \quad \boldsymbol{M}_G := \frac{1}{n!} \sum_{\sigma \in S_n} \boldsymbol{A}_{\sigma(1)} \times \cdots \times \boldsymbol{A}_{\sigma(n)}$$

where S_n denotes the group of permutations. Our conjecture is that the norm of $\boldsymbol{M}_G$ is always less than the norm of $(\boldsymbol{M}_A)^n$. Assuming this inequality, we show that without-replacement sampling leads to faster convergence for both the least mean squares and randomized Kaczmarz algorithms of Strohmer and Vershynin (2009).

Using established work in matrix analysis, we show that these noncommutative arithmetic-geometric mean inequalities hold when there are only two matrices in the pool. We also prove that the inequality is true when all of the matrices commute. We demonstrate that if we don't symmetrize, there are deterministically ordered products of n matrices whose norm exceeds $\|\boldsymbol{M}_A\|^n$ by an exponential factor. That is, symmetrization is necessary for the noncommutative arithmetic-geometric mean inequality to hold.

While we are unable to prove the noncommutative arithmetic-geometric mean inequality in full generality, we verify that it holds for many classes of *random* matrices. Random matrices are, in some sense, the most interesting case for machine learning applications as in empirical risk minimization, online learning, and many other settings, we assume that our data is generated by some i.i.d random process. In Section 4, we show that if $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$ are generated i.i.d. from certain distributions, then the noncommutative arithmetic-geometric mean inequality holds in expectation with respect to the $\boldsymbol{A}_i$. Section 4.1 assumes that $\boldsymbol{A}_i = \boldsymbol{Z}_i \boldsymbol{Z}_i^T$ where $\boldsymbol{Z}_i$ have independent entries, identically sampled from some symmetric distribution. In Section 4.2, we analyze the random matrices that commonly arise in stochastic gradient descent and related algorithms, again proving that without-replacement sampling exhibits faster convergence than with-replacement sampling. We close with a discussion of other open conjectures that could impact machine learning theory, algorithms, and software.

2. Sampling in incremental gradient descent

To illustrate how with- and without-replacement sampling methods differ in randomized optimization algorithms, we focus on one core algorithm, the Incremental Gradient Method (IGM). Recall that the IGM minimizes the function

$$\underset{\boldsymbol{x}}{\text{minimize}}\ f(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(\boldsymbol{x}) \tag{1}$$

via the iteration

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \gamma_k \nabla f_{i_k}(\boldsymbol{x}_{k-1})\,. \tag{2}$$

Here, $\boldsymbol{x}_0$ is an initial starting vector, γ_k are a sequence of nonnegative step sizes, and the indices i_k are chosen using some (possibly deterministic) sampling scheme. When f is strongly convex, the IGM iteration converges to a near-optimal solution of (1) for any $\boldsymbol{x}_0$ under a variety of step-sizes protocols and sampling schemes including constant and diminishing step-sizes (see Anstreicher and Wolsey, 2000; Bertsekas, 2012; Nemirovski et al., 2009). In the next examples, we study the specialized case where the f_i are quadratic and the IGM is equivalent to the *least mean squares* algorithm of Widrow and Hoff (1960).

2.1. One-dimensional Examples

First consider the following toy one-dimensional least-squares problem

$$\underset{x}{\text{minimize}}\ \frac{1}{2}\sum_{i=1}^{n}(x-y_i)^2\,. \tag{3}$$

where y_i is a sequence of scalars with mean μ_y and variance σ^2. Applying (2) to (3) results in the iteration $x_k = x_{k-1} - \gamma_k(x_{k-1} - y_{i_k})$. If we initialize the method with $x_0 = 0$ and take n steps of incremental gradient with stepsize $\gamma_k = 1/k$, we have $x_n = \frac{1}{n}\sum_{j=1}^{n} y_{i_j}$, where i_j is the index drawn at iteration j. If the steps are chosen using a without-replacement sampling scheme, $x_n = \mu_y$, the global minimum. On the other hand, using with-replacement sampling, we will have $\mathbb{E}[(x_n-\mu_y)^2] = \frac{\sigma^2}{n}$, which is positive mean square error.

Another toy example that further illustrates the discrepancy is the least-squares problem

$$\underset{x}{\text{minimize}}\ \frac{1}{2}\sum_{i=1}^{n}\beta_i(x-y)^2$$

where β_i are positive weights. Here, y is a scalar, and the global minimum is clearly y. Let's consider the incremental gradient method with constant stepsize $\gamma_k = \gamma < \min \beta_i^{-1}$. Then after n iterations we will have

$$|x_n - y| = |y|\prod_{j=1}^{n}(1-\gamma\beta_{i_j})$$

If we perform without-replacement sampling, this error is given by

$$|x_n - y| = |y|\prod_{i=1}^{n}(1-\gamma\beta_i)\,.$$

On the other hand, using with-replacement sampling yields

$$\mathbb{E}[|x_n - y|] = |y|\left(1-\frac{\gamma}{n}\sum_{i=1}^{n}\beta_i\right)^n\,.$$

By the arithmetic-geometric mean inequality, we then have that the without-replacement sample is always closer to the optimal value in expectation. This sort of discrepancy is not simply a feature of these toy examples. We now demonstrate that similar behavior arises in multi-dimensional examples.

2.2. IGM in more than one dimension

Now consider IGM in higher dimensions. Let $\boldsymbol{x}_\star$ be a vector in $\mathbb{R}^d$ and set

$$y_i = \boldsymbol{a}_i^T \boldsymbol{x}_\star + \omega_i \quad \text{for } i = 1, \dots, n$$

where $\boldsymbol{a}_i \in \mathbb{R}^d$ are some test vectors and ω_i are i.i.d. Gaussian random variables with mean zero and variance ρ^2.

We want to compare with- vs without-replacement sampling for IGD on the cost function

$$\underset{\boldsymbol{x}}{\text{minimize}} \ \sum_{i=1}^{n} (\boldsymbol{a}_i^T \boldsymbol{x} - y_i)^2 \,. \tag{4}$$

Suppose we walk over k steps of IGD with constant stepsize γ and we access the terms $i_1, \dots, i_k$ in that order. Then we have

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \gamma \boldsymbol{a}_{i_k} (\boldsymbol{a}_{i_k}^T \boldsymbol{x}_{i_{k-1}} - y_{i_k}) = \left(I - \gamma \boldsymbol{a}_{i_k} \boldsymbol{a}_{i_k}^T\right) \boldsymbol{x}_{i_{k-1}} + \gamma \boldsymbol{a}_{i_k} y_{i_k} \,.$$

Subtracting $x_\star$ from both sides of this equation then gives

$$\begin{aligned} \boldsymbol{x}_k - \boldsymbol{x}_\star &= \left(\boldsymbol{I} - \gamma \boldsymbol{a}_{i_k} \boldsymbol{a}_{i_k}^T\right) (\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star) + \gamma \boldsymbol{a}_{i_k} \omega_{i_k} \\ &= \prod_{j=1}^{k} \left(\boldsymbol{I} - \gamma \boldsymbol{a}_{i_j} \boldsymbol{a}_{i_j}^T\right) (\boldsymbol{x}_0 - \boldsymbol{x}_\star) + \sum_{\ell=1}^{k} \prod_{k \geq j > \ell} \left(\boldsymbol{I} - \gamma \boldsymbol{a}_{i_j} \boldsymbol{a}_{i_j}^T\right) \gamma \boldsymbol{a}_{i_\ell} \omega_{i_\ell} \,. \end{aligned} \tag{5}$$

Here, the product notation means we multiply by the matrix with smallest index first, then left multiply by the matrix with the next index and so on up to the largest index.

Our goal is to estimate the risk after k steps, namely $\mathbb{E}[\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2]$, and demonstrate that this error is smaller for the without-replacement model. The expectation is with respect to the IGM ordering and the noise sequence ω_i. To simplify things a bit, we take a partial expectation with respect to ω_i:

$$\mathbb{E}[\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2] = \mathbb{E}\left[\left\|\prod_{j=1}^{k} \left(\boldsymbol{I} - \gamma \boldsymbol{a}_{i_j} \boldsymbol{a}_{i_j}^T\right) (\boldsymbol{x}_0 - \boldsymbol{x}_\star)\right\|^2\right] + \rho^2 \gamma^2 \sum_{\ell=1}^{k} \mathbb{E}\left[\left\|\prod_{k \geq j > \ell} \left(\boldsymbol{I} - \gamma \boldsymbol{a}_{i_j} \boldsymbol{a}_{i_j}^T\right) \boldsymbol{a}_{i_\ell}\right\|^2\right] \tag{6}$$

In this case, we need to compare the expected value of matrix products under with or without-replacement sampling schemes in order to conclude which is better. Is there a simple conjecture, analogous to the arithmetic-geometric mean inequality, that would guarantee without-replacement sampling is always better?

2.3. The Randomized Kaczmarz algorithm

As another high-dimensional example in the same spirit, we consider the randomized Kaczmarz algorithm of Strohmer and Vershynin (2009). The Kaczmarz algorithm is used to solve the over-determined linear system $\boldsymbol{\Phi x} = \boldsymbol{y}$. Here $\boldsymbol{\Phi}$ is an $n \times d$ matrix with $n > d$ and we assume there exists an exact solution $\boldsymbol{x}_\star$ satisfying $\boldsymbol{\Phi x}_\star = \boldsymbol{y}$. Kaczmarz's method solves this system by alternating projections (Kaczmarz, 1937) and was implemented in the earliest medical scanning devices (Hounsfield, 1973). In computer tomography, this method is called the *Algebraic Reconstruction Technique* (Herman, 1980; Natterer, 1986) or *Projection onto Convex Sets* (Sezan and Stark, 1987).

Kaczmarz's algorithm consists of iterations of the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \frac{y_i - \boldsymbol{\phi}_{i_k}^T \boldsymbol{x}_k}{\|\boldsymbol{\phi}_{i_k}\|^2} \boldsymbol{\phi}_{\boldsymbol{i_k}} . \tag{7}$$

where the rows of Φ are accessed in some deterministic order. This sequence can be interpreted as an incremental variant of Newton's method on the least squares cost function

$$\underset{\boldsymbol{x}}{\text{minimize}} \sum_{i=1}^{n} (\boldsymbol{\phi}_i^T \boldsymbol{x}_k - y_i)^2$$

with step size equal to 1 (see Bertsekas, 1999).

Establishing the convergence rate of this method proved difficult in imaging science. On the other hand, Strohmer and Vershynin (2009) proposed a randomized variant of the Kaczmarz method, choosing the next iterate with-replacement with probability proportional to the norm of $\boldsymbol{\phi}_i$. Strohmer and Vershynin established linear convergence rates for their iterative scheme. Expanding out (7) for k iterations, we see that

$$\boldsymbol{x}_k - \boldsymbol{x}_\star = \prod_{j=1}^{k} \left(\boldsymbol{I} - \frac{\boldsymbol{\phi}_{i_j} \boldsymbol{\phi}_{i_j}^T}{\|\boldsymbol{\phi}_{i_j}\|^2} \right) (\boldsymbol{x}_0 - \boldsymbol{x}_\star) .$$

Let us suppose that we modify Strohmer and Vershynin's procedure to employ without-replacement sampling. After k steps is the with-replacement or without-replacement model closer to the optimal solution?

3. Conjectures concerning the norm of geometric and arithmetic means of positive definite matrices

To formulate a sufficient conjecture which would guarantee that without-replacement sampling outperforms with-replacement, let us first formalize some notation. Throughout, $[n]$ denotes the set of integers from 1 to n, and $\|\cdot\|$ represents the operator norm for matrices and ℓ_2 norm for vectors unless explicitly stated otherwise. Let $\mathbb{D}$ be some domain, $f : \mathbb{D}^k \to \mathbb{R}$, and $(x_1, \ldots, x_n)$ a set of n elements from $\mathbb{D}$. We define the without-replacement expectation as

$$\mathbb{E}_{\text{wo}}[f(x_{i_1}, \ldots, x_{i_k})] = \tfrac{(n-k)!}{n!} \sum_{j_1 \neq j_2 \neq \ldots \neq j_k} f(x_{j_1}, \ldots, x_{j_k}) .$$

That is, we average the value of f over all ordered tuples of elements from $(x_1, \ldots, x_n)$. Similarly, the with-replacement expectation is defined as

$$\mathbb{E}_{\text{wr}}[f(x_{i_1}, \ldots, x_{i_k})] = n^{-k} \sum_{(j_1, \ldots, j_k)=1}^{n} f(x_{j_1}, \ldots, x_{j_k}) .$$

With these conventions, we can list our main conjectures as follows:

Conjecture 1 (Operator Inequality of Noncommutative Arithmetic and Geometric Means) *Let $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$ be a collection of positive semidefinite matrices. Then we conjecture that the following two inequalities always hold:*

$$\left\| \mathbb{E}_{\mathrm{wo}} \left[\prod_{j=1}^{k} \boldsymbol{A}_{i_j} \right] \right\| \leq \left\| \mathbb{E}_{\mathrm{wr}} \left[\prod_{j=1}^{k} \boldsymbol{A}_{i_j} \right] \right\| \tag{8}$$

$$\left\| \mathbb{E}_{\mathrm{wo}} \left[\prod_{j=1}^{k} \boldsymbol{A}_{i_{k-j+1}} \prod_{j=1}^{k} \boldsymbol{A}_{i_j} \right] \right\| \leq \left\| \mathbb{E}_{\mathrm{wr}} \left[\prod_{j=1}^{k} \boldsymbol{A}_{i_{k-j+1}} \prod_{j=1}^{k} \boldsymbol{A}_{i_j} \right] \right\| \tag{9}$$

Note that in (8), we have $\mathbb{E}_{\mathrm{wr}}[\prod_j \boldsymbol{A}_{i_j}] = (\frac{1}{n}\sum_i \boldsymbol{A}_i)^k = (\boldsymbol{M}_A)^k$.

Assuming this conjecture holds, let us return to the analysis of the IGM (6). Assuming that $\boldsymbol{x}_0 - \boldsymbol{x}_\star$ is an arbitrary starting vector and that (9) holds, we have that each term in this summation is smaller for the without-replacement sampling model than for the with-replacement sampling model. In turn, we expect the without-replacement sampling implementation will return lower risk after one pass over the data-set. Similarly, for the randomized Kaczmarz iteration (7), Conjecture 1 implies that a without-replacement sample will have lower error after $k < n$ iterations.

In the remainder of this document we provide several case studies illustrating that these noncommutative variants of the arithmetic-geometric mean inequality hold in a variety of settings, establishing along the way tools and techniques that may be useful for proving Conjecture 1 in full generality.

3.1. Prior Art: Two matrices and a search for the geometric mean

Both of the inequalities (8) and (9) are true when $n = 2$. These inequalities all follow from an well-estabilished line of research in estimating the norms of products of matrices, started by the seminal work of Bhatia and Kittaneh (1990). First, the symmetrized geometric mean actually precedes the square of the arithmetic mean in the positive definite order. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be positive semidefinite. Then we have

$$\left(\tfrac{1}{2}\boldsymbol{A} + \tfrac{1}{2}\boldsymbol{B}\right)^2 - \left(\tfrac{1}{2}\boldsymbol{A}\boldsymbol{B} + \tfrac{1}{2}\boldsymbol{B}\boldsymbol{A}\right) = \tfrac{1}{4}\boldsymbol{A}^2 + \tfrac{1}{4}\boldsymbol{B}^2 - \tfrac{1}{4}\boldsymbol{A}\boldsymbol{B} - \tfrac{1}{4}\boldsymbol{B}\boldsymbol{A} = \left(\tfrac{1}{2}\boldsymbol{A} - \tfrac{1}{2}\boldsymbol{B}\right)^2 \succeq 0$$

implying (8). For two matrices, considerably stronger inequalities apply. Bhatia and Kittaneh (2000) showed that

$$\|\boldsymbol{A}\boldsymbol{B}\| \leq \left\|\tfrac{1}{2}\boldsymbol{A} + \tfrac{1}{2}\boldsymbol{B}\right\|^2 \tag{10}$$

demonstrating the arithmetic-geometric mean inequality holds for deterministic orderings of two matrices. (9) is a consequence of this inequality. We provide a proof of this fact in Appendix B. The interested reader should consult Bhatia and Kittaneh (2008) for a comprehensive list of similar inequalities concerning pairs of matrices.

Unfortunately, these techniques are specialized to the case of two matrices, and no proof currently exists for the inequalities when $n \geq 3$. There have been a varied set of attempts to extend the noncommutative arithmetic-geometric mean inequalities to more than two matrices. Much of the work in this space has focused on how to properly define the geometric mean of a collection of positive semidefinite matrices. For instance, Ando et al. (2004) demarcate a list of properties desirable by any geometric mean, with one of the properties being that the geometric mean must precede the arithmetic mean in the positive-definite ordering. Ando *et al* derive a geometric mean satisfying all of these properties, but the resulting mean in no way resembles the means of matrices discussed in this paper. Instead, their geometric mean is defined as a fixed point of a nonlinear map on matrix tuples. Bhatia and Holbrook (2006) and Bonnabel and Sepulchre (2009) propose geometric means based on geodesic flows on the Riemannian manifold of positive definite matrices, however these means also do not correspond to the averaged matrix products that we study in this paper.

3.2. When is it necessary to symmetrize the order?

When the matrices commute, Conjecture 1 is a consequence of the standard arithmetic-geometric mean inequality (more precisely, a consequence of Maclaurin's inequalities). A discussion of the commutative case and Maclaurin's inequalities can be found in Appendix A. In fact, in this case, any order of the matrix products will satisfy the desired arithmetic-geometric mean inequalities.

In contrast, symmetrizing over the order of the product is necessary for noncommutative operators. The following example in fact provides deterministic without-replacement orderings that have exponentially larger norm than the with-replacement expectation. Let $\omega_n = \pi/n$. For $n \geq 3$, define the collection of vectors

$$\boldsymbol{a}_{k;n} = \begin{bmatrix} \cos(k\omega_n) \\ \sin(k\omega_n) \end{bmatrix} . \tag{11}$$

Note that all of the $\boldsymbol{a}_{k;n}$ have norm 1 and, for $1 \leq k < n$, $\langle \boldsymbol{a}_{k;n}, \boldsymbol{a}_{k+1;n} \rangle = \cos(\omega_n)$. The matrices $\boldsymbol{A}_k := \boldsymbol{a}_{k;n}\boldsymbol{a}_{k;n}^T$ are all positive semidefinite for $1 \leq k \leq n$, and we have the identity

$$\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{A}_k = \tfrac{1}{2} I \, . \tag{12}$$

Any set of unit vectors satisfying (12) is called a *normalized tight frame*, and the vectors (11) form a *harmonic frame* due to their trigonometric origin (see Hassibi et al., 2001; Goyal et al., 2001). The product of the $\boldsymbol{A}_i$ is given by

$$\prod_{i=1}^{k} \boldsymbol{A}_i = \boldsymbol{a}_{k;n}\boldsymbol{a}_{1;n}^T \prod_{j=1}^{k-1} \langle \boldsymbol{a}_{j;n}, \boldsymbol{a}_{j+1;n} \rangle = \boldsymbol{a}_{k;n}\boldsymbol{a}_{1;n}^T \cos^{k-1}(\omega_n) \ ,$$

and hence

$$\left\| \prod_{i=1}^{k} \boldsymbol{A}_i \right\| = \cos^{k-1}(\omega_n) \geq 2^k \cos^{k-1}(\omega_n) \left\| \left(\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{A}_k \right)^k \right\| .$$

Therefore, the arithmetic mean is less than the (deterministic) geometric mean for all $n \geq 3$.

It turns out that this harmonic frame example is in some sense the worst case. The following proposition shows that the geometric mean is always within a factor of d^k of the arithmetic mean for any ordering of the without-replacement matrix product.

Proposition 2 *Let $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$ be $d \times d$ positive semidefinite matrices. Then*

$$\left\| \mathbb{E}_{\text{wo}} \left[\prod_{i=1}^{k} \boldsymbol{A}_{j_i} \right] \right\| \leq d^k \left\| \mathbb{E}_{\text{wr}} \left[\prod_{i=1}^{k} \boldsymbol{A}_{j_i} \right] \right\| .$$

Proof If we sample $j_1, \ldots, j_k$ uniformly from $[n]$, then we have

$$\left\| \mathbb{E}_{\text{wo}} \left[\prod_{i=1}^{k} \boldsymbol{A}_{j_i} \right] \right\| \le \mathbb{E}_{\text{wo}} \left[\left\| \prod_{i=1}^{k} \boldsymbol{A}_{j_i} \right\| \right] \le \mathbb{E}_{\text{wo}} \left[\prod_{i=1}^{k} \|\boldsymbol{A}_{j_i}\| \right] \le \mathbb{E}_{\text{wo}} \left[\prod_{i=1}^{k} \operatorname{trace}(\boldsymbol{A}_{j_i}) \right]$$
$$\le \left(\frac{1}{n} \sum_{i=1}^{n} \operatorname{trace}(\boldsymbol{A}_i) \right)^k = \operatorname{trace} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{A}_i \right)^k \le \left\| \frac{d}{n} \sum_{i=1}^{n} \boldsymbol{A}_i \right\|^k .$$

Here, the first inequality follows from the triangle inequality. The second, because the operator norm is submultiplicative. The third inequality follows because the trace dominates the operator norm. The fourth inequality is Maclaurin's. The fifth inequality follows because the trace of a $d \times d$ positive semidefinite matrix is upper bounded by d times the operator norm. ■

For the interested reader, we construct examples saturating this worst-case bound in higher dimensions using harmonic frames in Appendix D.

At first glance, the harmonic frames example appears to cast doubt on the validity of Conjecture 1. However, after symmetrizing over the symmetric group, the $2d$ harmonic frames do obey (8).

Theorem 3 *Let* $\lambda(n) = {}_2F_3 \left[\begin{matrix} 1 & -n/2 + 1/2 & -n/2 \\ 1/2 & -n+1 & \end{matrix} ; 1 \right]$. *With the* $\boldsymbol{a}_{k;n}$ *defined in (11),*

$$\tfrac{1}{n!} \sum_{\sigma \in S_n} \prod_{i=1}^{n} \boldsymbol{a}_{\sigma(i);n} \boldsymbol{a}_{\sigma(i+1);n}^T = -\lambda(n) 2^{-n} \boldsymbol{I} \,, \quad \textit{and} \quad 1 \ge \lambda(n) = \mathcal{O}(n^{-1}) \,.$$

This theorem additionally verifies that there is an asymptotic gap between the arithmetic and geometric means of the harmonic frames example after symmetrization. We include a full proof of this result in Appendix E. Our first step is to treat the norm variationally using the identity that $\|\boldsymbol{X}\|_2$ is the maximum of $\boldsymbol{v}^T \boldsymbol{X} \boldsymbol{v}$ over all unit vectors $\boldsymbol{v}$. Our computation in this stage is effectively computing a Fourier transform of the function of $\boldsymbol{v}$ in an appropriately defined finite group. We show that the Fourier coefficients can be viewed as enumerating sets, and we compute them exactly using generating functions.

The combinatorial argument that we use to prove Theorem 3 is very specialized. To provide a broader set of examples, we now turn to show that Conjecture 1 does in fact hold for many classes of *random* matrices.

4. Random matrices

In this section, we show that if $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$ are generated i.i.d. from certain distributions, then Conjecture 1 holds in expectation with respect to the $\boldsymbol{A}_i$. Section 4.1 assumes that $\boldsymbol{A}_i = \boldsymbol{Z}_i \boldsymbol{Z}_i^T$ where $\boldsymbol{Z}_i$ have independent entries, identically sampled from some symmetric distribution. In Section 4.2, we explore when the matrices $\boldsymbol{A}_i$ are random rank-one perturbations of the identity as was the case in the IGM and Kaczmarz examples.

4.1. Random matrices satisfy the noncommutative arithmetic-geometric mean inequality

In this section, we prove the following

Theorem 4 *For each $i = 1, \ldots, n$, suppose $\boldsymbol{A}_i = \boldsymbol{Z}_i \boldsymbol{Z}_i^T$ with $\boldsymbol{Z}_i$ a $d \times r$ random matrix whose entries are i.i.d. samples from some symmetric distribution. Then Conjecture 1 holds in expectation.*

Proof Suppose the entries of each $\boldsymbol{Z}_i$ have finite variance σ^2 (the theorem would be otherwise vacuous if we assumed infinite variance). Let the (a, b) entry of $\boldsymbol{Z}_i$ be denoted by $Z_{a,b}^{(i)}$. Also, denote by $\boldsymbol{W}$ the matrix with all of the $\boldsymbol{Z}_i$ stacked as columns: $\boldsymbol{W} = \sigma^{-1}[\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n]$.

Let's first prove that (8) holds in expectation for these matrices. First, consider the without-replacement samples, which are considerably easy to analyze. Let $(j_1, \ldots, j_k)$ be a without-replacement sample from $[n]$. Then

$$\left\| \mathbb{E}\left[\prod_{i=1}^{k} \boldsymbol{A}_{j_i}\right] \right\| = \| \mathbb{E}[\boldsymbol{A}_1]^k \| = r^k \sigma^{2k} .$$

For the arithmetic mean, we can compute

$$r^{-k}\sigma^{-2k} \left\| \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}_i\right)^k\right] \right\| \geq \frac{1}{r^k \sigma^{2k} d} \operatorname{trace}\left(\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}_i\right)^k\right]\right)$$

$$= \mathbb{E}\left[d^{-1} \operatorname{trace}\left(\left(\frac{1}{nr\sigma^2}\sum_{i=1}^{n} \boldsymbol{A}_i\right)^k\right)\right] = \mathbb{E}\left[d^{-1} \operatorname{trace}\left(\left(\frac{1}{nr}\boldsymbol{W}\boldsymbol{W}^T\right)^k\right)\right] \tag{13}$$

$$= d^{-1}(nr)^{-k} \sum_{\{a_1,\ldots,a_k\}=1}^{d} \sum_{\{b_1,\ldots,b_k\}=1}^{nr} \mathbb{E}[W_{a_1,b_1} W_{a_2,b_1} W_{a_2,b_2} W_{a_3,b_2} \ldots W_{a_k,b_k} W_{a_1,b_k}]$$

$$= (nr)^{-k} \sum_{\{a_2,\ldots,a_k\}=1}^{d} \sum_{\{b_1,\ldots,b_k\}=1}^{nr} \mathbb{E}[W_{1,b_1} W_{a_2,b_1} W_{a_2,b_2} W_{a_3,b_2} \ldots W_{a_k,b_k} W_{1,b_k}] . \tag{14}$$

Note that since W_{ij} are iid, symmetric random variables, each term in this sum is zero if it contains an odd power of W_{ij} for some i and j. If all of the powers in a summand are even, its expected value is bounded below by 1. A simple lower bound for this final term (14) thus looks only at the contribution from when all of the indices a_i are set equal to 1.

$$(nr)^{-k} \sum_{\{b_1,\ldots,b_k\}=1}^{nr} \mathbb{E}[W_{1,b_1}^2 W_{1,b_2}^2 \ldots W_{1,b_k}^2] = \mathbb{E}\left[\left(\frac{1}{nr}\sum_{b=1}^{nr} W_{1,b}^2\right)^k\right] \geq \left(\mathbb{E}\left[\frac{1}{nr}\sum_{b=1}^{nr} W_{1,b}^2\right]\right)^k = 1 .$$

Here the inequality is Jensen's. This calculation proves the noncommutative arithmetic-geometric mean inequality for our family of random matrices. That is, we have demonstrated that the expected value of the with-replacement sample has greater norm than the expected value of the without-replacement sample.

To prove (9), we need to control the quantity $\boldsymbol{V}(i_1, \ldots, i_n) := \boldsymbol{A}_{i_k} \cdots \boldsymbol{A}_{i_2} \boldsymbol{A}_{i_1}^2 \boldsymbol{A}_{i_2} \cdots \boldsymbol{A}_{i_k}$ under the two different sampling models. Essentially the calculation parallels the above proof strategy. We write out the expectation of $\boldsymbol{V}$ exactly in the case of the without-replacement sampling. Then we lower bound the arithmetic mean by dropping some terms, all of which must be positive because the Z_{ij} are symmetrically distributed. The remaining terms sum to the geometric mean, proving (9).

■

The arguments used to prove Theorem 4 grossly undercount the number of terms that contribute to the expectation. Bounds on the quantity (13) commonly arise in the theory of random matrices (see the survey by Bai, 1999, for more details and an extensive list of references). Indeed, if we let $d = \delta n$ and assume that W_{ij} have bounded fourth moment, we have that (13) tends to $(1+\sqrt{\delta})^{2k}$ almost surely a $n \to \infty$. That is, the gap between the with- and without-replacement sampling grows exponentially with k in this scaling regime. Similarly, there is an asymptotic, exponential gap between the with and without-replacement expectations in (9). In the appendix, we specialize to the case where the $\boldsymbol{Z}_i$ are Gaussian (and hence the $\boldsymbol{A}_i$ are Wishart) and demonstrate that the ratio of the expectation is bounded below by $re^{\frac{1}{4k(k+1)}}\left(\frac{16k}{e^2 r(r+d+1)}\right)^k$.

4.2. Random vectors and the incremental gradient method

We can also use a random analysis to demonstrate that for the least-squares problem (4), without-replacement sampling outperforms with-replacement sampling if the data is randomly generated.

Let's look at one step of the recursion (5) and assume that the $\boldsymbol{a}_i$ are sampled i.i.d. from some distribution. Assume that the moments $\boldsymbol{\Lambda} := \mathbb{E}[\boldsymbol{a}_i\boldsymbol{a}_i^T]$ and $\boldsymbol{\Delta} := \mathbb{E}[\|\boldsymbol{a}_i\|^2\boldsymbol{a}_i\boldsymbol{a}_i^T]$ exist. Then we see immediately that

$$\mathbb{E}_{\text{wo}}[\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2] = \mathbb{E}_{\text{wo}}[\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star]^T(\boldsymbol{I} - 2\gamma\boldsymbol{\Lambda} + \gamma^2\boldsymbol{\Delta})\,\mathbb{E}_{\text{wo}}[\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star] + \rho^2\gamma^2\operatorname{trace}(\boldsymbol{\Lambda})$$

because $\boldsymbol{a}_{j_k}$ is chosen independently from $(\boldsymbol{a}_{j_1}, \ldots, \boldsymbol{a}_{j_{k-1}})$ On the other hand, in the with-replacement model, we have

$$\mathbb{E}_{\text{wr}}[\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2] = \mathbb{E}_{\text{wr}}\left[(\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star)^T(\boldsymbol{I} - 2\gamma\boldsymbol{\Lambda}_n + \gamma^2\boldsymbol{\Delta}_n)(\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star)\right] + \rho^2\gamma^2\operatorname{trace}(\boldsymbol{\Lambda})$$

where

$$\boldsymbol{\Lambda}_n := \frac{1}{n}\sum_{i=1}^n \boldsymbol{a}_i\boldsymbol{a}_i^T \quad \text{and} \quad \boldsymbol{\Delta}_n := \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{a}_i\|^2\boldsymbol{a}_i\boldsymbol{a}_i^T\,.$$

In this case, we cannot distribute the expected value because the vector $\boldsymbol{x} - \boldsymbol{x}_\star$ depends on all $\boldsymbol{a}_i$ for $1 \le i \le n$. To get a flavor for how these differ, consider the conditional expectation

$$\begin{aligned}\mathbb{E}_{\text{wr}}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2 \mid \{\boldsymbol{a}_i\}\right] \le \left(1 - 2\gamma\lambda_{\min}(\boldsymbol{\Lambda}_n) + \gamma^2\lambda_{\max}(\boldsymbol{\Delta}_n)\right)\mathbb{E}_{\text{wr}}\left[\|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|^2 \mid \{\boldsymbol{a}_i\}\right]\\ + \rho^2\gamma^2\operatorname{trace}(\boldsymbol{\Lambda})\end{aligned}$$

Similarly,

$$\mathbb{E}_{\text{wo}}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2\right] \le \left(1 - 2\gamma\lambda_{\min}(\boldsymbol{\Lambda}) + \gamma^2\lambda_{\max}(\boldsymbol{\Delta})\right)\mathbb{E}_{\text{wo}}\left[\|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|^2\right] + \rho^2\gamma^2\operatorname{trace}(\boldsymbol{\Lambda})\,.$$

Expanding out these recursions, we have

$$\begin{aligned}\mathbb{E}_{\text{wo}}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2\right] &\le \left(1 - 2\gamma\lambda_{\min}(\boldsymbol{\Lambda}) + \gamma^2\lambda_{\max}(\boldsymbol{\Delta})\right)^k\mathbb{E}_{\text{wo}}\left[\|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|^2\right] + \tfrac{\rho^2\gamma\operatorname{trace}(\boldsymbol{\Lambda})}{2\lambda_{\min}(\boldsymbol{\Lambda}) - \gamma\lambda_{\max}(\boldsymbol{\Delta})}\\ \mathbb{E}_{\text{wr}}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2 \mid \{\boldsymbol{a}_i\}\right] &\le \left(1 - 2\gamma\lambda_{\min}(\boldsymbol{\Lambda}_n) + \gamma^2\lambda_{\max}(\boldsymbol{\Delta}_n)\right)^k\mathbb{E}_{\text{wr}}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2 \mid \{\boldsymbol{a}_i\}\right]\\ &\quad + \tfrac{\rho^2\gamma\operatorname{trace}(\boldsymbol{\Lambda}_n)}{2\lambda_{\min}(\boldsymbol{\Lambda}_n) - \gamma\lambda_{\max}(\boldsymbol{\Delta}_n)}\end{aligned}$$

Now, since $\sum_i \boldsymbol{a}_i \boldsymbol{a}_i^T$ is positive definite and since $\lambda_{\min}$ is concave on Hermitian matrices, we have by Jensen's inequalty that

$$\mathbb{E}[\lambda_{\min}(\boldsymbol{\Lambda}_n)] = \mathbb{E}\left[\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{a}_i^T\right)\right] \leq \lambda_{\min}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{a}_i^T\right]\right) = \lambda_{\min}(\boldsymbol{\Lambda}) \ ,$$

and, since $\lambda_{\max}$ is convex for symmetric matrices,

$$\mathbb{E}\left[\lambda_{\max}(\boldsymbol{\Delta}_n)\right] = \mathbb{E}\left[\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{a}_i\|^2 \boldsymbol{a}_i \boldsymbol{a}_i^T\right)\right] \geq \lambda_{\max}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{a}_i\|^2 \boldsymbol{a}_i \boldsymbol{a}_i^T\right]\right) = \lambda_{\max}(\boldsymbol{\Delta}) \, .$$

This means that the with-replacement upper bound is worse than the without-replacement estimate with reasonably high probability on most models of $\boldsymbol{a}_i$. Under mild conditions on $\boldsymbol{a}_i$ (including Gaussianity, bounded entries, subgaussian moments, or bounded Orlicz norm), we can estimate tail bounds for the eigenvalues of Λ_n and Δ_n (by applying the techniques of Tropp, 2011, for example). These large deviation inequalities provide quantitative estimates of the gap between with- and without-replacement sampling for the least mean squares and randomized Kaczmarz algorithms. Similar, but more tedious analysis, would reveal that with-replacement sampling fares worse with diminishing step sizes as well.

5. Discussion and open problems

While i.i.d. matrices are of significant importance in machine learning, the major piece of open work is proving Conjecture 1 for all positive semidefinite matrix tuples or finding a counterexample for either of the assertions. As demonstrated by the harmonic frames example, symmetrized products of deterministic matrices become quickly tedious and difficult to study. Some sort of combinatorial structure might need to be exploited for a short proof to arise in general. It remains to be seen if this sort of combinatorics employed in proving Theorem 3 generalizes beyond this particular example, but we expect these techniques will be useful in future studies of Conjecture 1. In particular, it would be interesting to see if we could reduce the proof of the conjecture to verifying the conjecture on frames that arise as the orbit of the representation of some finite group. These frames have been fully classified by Hassibi et al. (2001), and would reduce Conjecture 1 to a finite list of cases.

A further conjecture and its consequences The generalization of (10) to $n \geq 3$ asserts a stronger version of (8)

$$\mathbb{E}_{\mathrm{wo}}\left[\left\|\prod_{j=1}^{k} \boldsymbol{A}_{i_j}\right\|^2\right] \leq \left\|\tfrac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}_i\right\|^{2k} . \tag{15}$$

Certainly, (8) follows from (15) by Jensen's inequality the triangle inequality. Moreover, we showed that for two matrices, (9) also followed from (10). When $n \geq 3$, is it the case that (15) holds? It could be that for general matrices, it is easier to analyze (15) rather than (9) because the left hand side of the inequality is not quadratic in the matrices $\boldsymbol{A}_i$. A further question is if we can generalize the two matrix case to show that when $n \geq 3$, (15) always implies (9).

Effect of biased orderings. Another possible technique for solving incremental algorithms is to choose the best ordering of the increments to reach the cost function. In terms of matrices, can we find the ordering of the matrices $\boldsymbol{A}_i$ that achieves the minimum norm. At first glance this seems daunting. Suppose $\boldsymbol{A}_i = \boldsymbol{a}_i \boldsymbol{a}_i^T$ where the $\boldsymbol{a}_i$ are all unit vectors. Then for $\sigma \in S_n$

$$\left\| \prod_{i=1}^{n} \boldsymbol{A}_{\sigma(i)} \right\| = \prod_{i=1}^{n-1} |\langle \boldsymbol{a}_{\sigma(i)}, \boldsymbol{a}_{\sigma(i+1)} \rangle|$$

minimizing this expression with respect to σ amounts to finding the minimum weight traveling salesman path in the graph with weights $\log |\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle|$. Are there simple heuristics that can get within a small constant of the optimal tour for these graphs? How do greedy heuristics fare? This sort of approach was explored with some success for the Kaczmarz method by Eldar and Needell (2011).

Nonlinear extensions Extending even the random results in this paper to nonlinear algorithms such as the general incremental gradient descent algorithm or randomized coordinate descent would require modifying the analyses used here. However, it would be of interest to see which of the randomization tools employed in this work can be extended to the nonlinear case. For example, if we assume that the cost function (1) has summands which are sampled i.i.d., can we use similar tools (e.g., Jensen's inequality, moment bounds) to show that without-replacement sampling works even in the nonlinear case?

Acknowledgments

The authors would like to thank Dimitri Bertsekas, Aram Harrow, Pablo Parrilo for many helpful conversations and suggestions. BR is generously supported by ONR award N00014-11-1-0723 and NSF award CCF-1139953. CR is generously supported by the Air Force Research Laboratory (AFRL) under prime contract no. FA8750-09-C-0181, the NSF CAREER award under IIS-1054009, ONR award N000141210041, and gifts or research awards from Google, Greenplum, Johnson Controls, Inc., LogicBlox, and Oracle. Any opinions, findings, and conclusion or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of any of the above sponsors including DARPA, AFRL, or the US government.

References

T. Ando, C.-K. Li, and R. Mathias. Geometric means. *Linear Algebra and Its Applications*, 385: 305–334, 2004.

K. M. Anstreicher and L. A. Wolsey. Two "well-known" properties of subgradient optimization. *Mathematical Programming (Series B)*, 120(1):213–220, 2000.

Z. D. Bai. Methodologies in spectral analysis of large dimensional random matrices. *Statistica Sinica*, 9(3):611–661, 1999.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 2nd edition, 1999.

D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129:163–195, 2011.

D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization in Machine Learning*, pages 85–119. MIT Press, 2012.

R. Bhatia and J. Holbrook. Riemannian geometry and matrix geometric means. *Linear Algebra and Its Applications*, 413(2-3):594–618, 2006.

R. Bhatia and F. Kittaneh. On the singular values of a product of operators. *SIAM Journal on Matrix Analysis and Applications*, 11:272, 1990.

R. Bhatia and F. Kittaneh. Notes on matrix arithmetic-geometric mean inequalities. *Linear Algebra and Its Applications*, 308(1-3):203–211, 2000.

R. Bhatia and F. Kittaneh. The matrix arithmetic-geometric mean inequality revisited. *Linear Algebra and Its Applications*, 428(8-9):2177–2191, 2008.

S. Bonnabel and R. Sepulchre. Geometric distance and mean for positive semi-definite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1077, 2009.

L. Bottou. Online algorithms and stochastic approximations. In D. Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998.

L. Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science*, Paris, April 2009. Available at http://www.ceremade.dauphine.fr/~rahal/SLDS/LearningIII/bottou/sgdproblem.pdf.

K. Cafuta, I. Klep, and J. Povh. NCSOStools: a computer algebra system for symbolic and numerical computation with noncommutative polynomials. *Optimization Methods and Software*, 26(3): 363–380, 2011. Software available at http://ncsostools.fis.unm.si/download.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 3rd edition, 2009.

Y. C. Eldar and D. Needell. Acceleration of randomized Kaczmarz method via the Johnson-Lindenstrauss Lemma. *Numerical Algorithms*, 58(2):163–177, 2011.

V. Goyal, J. Kovacevic, and J. Kelner. Quantized frame expansions with erasures. *Applied and Computational Harmonic Analysis*, 10(3):203–233, 2001.

G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, 2nd edition, 1952.

B. Hassibi, B. Hochwald, A. Shokrollahi, and W. Sweldens. Representation theory for high-rate multiple-antenna code design. *IEEE Trans. Inform. Theory*, 47(6):2335–2367, 2001.

G. T. Herman. *Fundamentals of Computer Tomography: Image reconstruction from projections*. Springer, 1st edition, 1980.

G. N. Hounsfield. Computerized transverse axial scanning (tomography): Part I. description of the system. *British Journal of Radiology*, 46:1016–1022, 1973.

S. Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *International Bulletin of the Polish Academy of Sciences, Letters A*, pages 335–357, 1937.

J. Konvalina. Roots of unity and circular subsets without consecutive elements. *The Fibonacci Quarterly*, 33(5):412–415, 1995.

D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: Convergence rates and conditioning. *Mathematics of Operations Research*, 35(3):641–654, 2010.

Z.-Q. Luo. On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks. *Neural Computation*, 3(2):226–245, 1991.

F. Natterer. *The mathematics of computerized tomography*. Wiley, 1986.

A. Nedic and D. P. Bertsekas. Convergence rate of incremental subgradient algorithms. In S. Uryasev and P. M. Pardalos, editors, *Stochastic Optimization: Algorithms and Applications*, pages 263–304. Kluwer Academic Publishers, 2000.

A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, New York, 1983.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. CORE Discussion Paper 2010/2, 2010.

Y. Nesterov. Random gradient-free minimization of convex functions. CORE Discussion Paper 2011/16, 2011.

K. M. Sezan and H. Stark. Applications of convex projection theory to image recovery in tomography and related areas. In H. Stark, editor, *Image Recovery: Theory and Application*, pages 415–462. Academic Press, 1987.

S. Shalev-Shwartz and N. Srebro. SVM Optimization: Inverse dependence on training set size. In *Proceedings of the 25th Internation Conference on Machine Learning (ICML)*, 2008.

T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 2011. Springer Online First doi:10.1007/s10208-011-9099-z.

P. Tseng. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Joural on Optimization*, 8(2):506–531, 1998.

B. Widrow and M. E. Hoff. Adaptive switching circuits. In *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record*, pages 96–104, 1960.

Appendix A. Commutative matrices

One case which follows from simple, classical analysis is when the matrices do in fact commute. Consider the case where $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$ are positive semidefinite matrices that mutually commute with one another: $\boldsymbol{A}_i \boldsymbol{A}_j = \boldsymbol{A}_j \boldsymbol{A}_i$. In this case, we shall see that Conjecture 1 will follow from an application of Maclaurin's Inequalities, which generalize the scalar arithmetic-geometric mean inequality:

Theorem 5 (Maclaurin's Inequalities) *Let $x_1, \ldots, x_n$ be positive scalars. Let*

$$s_k = \binom{n}{k}^{-1} \sum_{\substack{\Omega \subset [n] \\ |\Omega|=k}} \prod_{i \in \Omega} x_i$$

be the normalized kth symmetric sum. Then we have

$$s_1 \geq \sqrt{s_2} \geq \ldots \geq \sqrt[n-1]{s_{n-1}} \geq \sqrt[n]{s_n}$$

Note that $s_1 \geq \sqrt[n]{s_n}$ is the standard form of the arithmetic-geometric mean inequality. See Hardy et al. (1952) for a discussion and proof of this chain of inequalities.

To see that these inequalities immediately imply Conjecture 1 when the matrices $\boldsymbol{A}_i$ are mutually commutative, note first that when $d = 1$, we have

$$\mathbb{E}_{\mathrm{wo}}\left[\prod_{i=1}^{k} a_i\right] = \binom{n}{k}^{-1} \sum_{\substack{\Omega \subset [n] \\ |\Omega|=k}} \prod_{i \in \Omega} a_i \leq \left(\frac{1}{n}\sum_{i=1}^{n} a_i\right)^k = \mathbb{E}_{\mathrm{wr}}\left[\prod_{i=1}^{k} a_i\right].$$

The higher dimensional analogs follow from similarly. Indeed, if all of the $\boldsymbol{A}_i$ commute, then the matrices are mutually diagonalizable. That is, we can write $\boldsymbol{A}_i = \boldsymbol{U}\Lambda_i\boldsymbol{U}^T$ where $\boldsymbol{U}$ is an orthogonal matrix, and the $\Lambda_i = \mathrm{diag}(\lambda_1^{(i)}, \ldots, \lambda_d^{(i)})$ are all diagonal matrices of the eigenvalues. Then we have

$$\left\|\mathbb{E}_{\mathrm{wo}}\left[\prod_{i=1}^{k} \boldsymbol{A}_i\right]\right\| = \left\|\mathbb{E}_{\mathrm{wo}}\left[\prod_{i=1}^{k} \boldsymbol{\Lambda}_i\right]\right\| = \max_j \prod_{i=1}^{k} \lambda_j^{(i)} \leq \max_j \mathbb{E}_{\mathrm{wr}} \prod_{i=1}^{k} \lambda_j^{(i)} = \left\|\mathbb{E}_{\mathrm{wr}}\left[\prod_{i=1}^{k} \boldsymbol{A}_i\right]\right\|$$

verifying our conjecture.

Appendix B. Proof of (9) for $n = 2$

Let $\boldsymbol{A}_1 = \boldsymbol{A}$ and $\boldsymbol{A}_2 = \boldsymbol{B}$ be two positive semindefinite matrices. Note that

$$\mathbb{E}_{\mathrm{wo}}[\boldsymbol{A}_{i_1}\boldsymbol{A}_{i_2}^2\boldsymbol{A}_{i_1}] = \tfrac{1}{2}\boldsymbol{A}\boldsymbol{B}^2\boldsymbol{A} + \tfrac{1}{2}\boldsymbol{B}\boldsymbol{A}^2\boldsymbol{B} \tag{16}$$

and

$$\mathbb{E}_{\mathrm{wr}}[\boldsymbol{A}_{i_1}\boldsymbol{A}_{i_2}^2\boldsymbol{A}_{i_1}] = \tfrac{1}{4}\boldsymbol{A}^4 + \tfrac{1}{4}\boldsymbol{A}\boldsymbol{B}^2\boldsymbol{A} + \tfrac{1}{4}\boldsymbol{B}\boldsymbol{A}^2\boldsymbol{B} + \tfrac{1}{4}\boldsymbol{B}^4.$$

We can bound (16) by

$$\tfrac{1}{2}\|\boldsymbol{A}\boldsymbol{B}^2\boldsymbol{A} + \boldsymbol{B}\boldsymbol{A}^2\boldsymbol{B}\| \leq \|\boldsymbol{A}\boldsymbol{B}^2\boldsymbol{A}\| = \|\boldsymbol{A}\boldsymbol{B}\|^2 \leq \|\tfrac{1}{2}\boldsymbol{A} + \tfrac{1}{2}\boldsymbol{B}\|^4 = \|(\tfrac{1}{2}\boldsymbol{A} + \tfrac{1}{2}\boldsymbol{B})^4\|.$$

Here, the first inequality is the triangle inequality and the subsequent equality follows because the norm of $\boldsymbol{X}^T\boldsymbol{X}$ is equal to the squared norm of $\boldsymbol{X}$. The second inequality is (10).

To complete the proof we show

$$\boldsymbol{X}_L := (\tfrac{1}{2}\boldsymbol{A} + \tfrac{1}{2}\boldsymbol{B})^4 \preceq \tfrac{1}{4}\boldsymbol{A}^4 + \tfrac{1}{4}\boldsymbol{A}\boldsymbol{B}^2\boldsymbol{A} + \tfrac{1}{4}\boldsymbol{B}\boldsymbol{A}^2\boldsymbol{B} + \tfrac{1}{4}\boldsymbol{B}^4 := \boldsymbol{X}_R$$

in the semidefinite ordering. But this follows by observing

$$\begin{aligned} \boldsymbol{X}_R - \boldsymbol{X}_L = \tfrac{1}{48}&(3\boldsymbol{A}^2 - \boldsymbol{A}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A} - \boldsymbol{B}^2)^2 \\ &+ \tfrac{1}{24}(2\boldsymbol{B}\boldsymbol{A} - \boldsymbol{A}\boldsymbol{B} - \boldsymbol{B}^2)(2\boldsymbol{A}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A} - \boldsymbol{B}^2) \\ &+ \tfrac{1}{8}(\boldsymbol{A}\boldsymbol{B} - \boldsymbol{B}^2)(\boldsymbol{B}\boldsymbol{A} - \boldsymbol{B}^2)\,, \end{aligned} \tag{17}$$

which means that $\boldsymbol{X}_R - \boldsymbol{X}_L$ is a sum of products of the form $\boldsymbol{Y}\boldsymbol{Y}^T$ and hence must be positive semidefinite. This means, in particular, that $\|\boldsymbol{X}_R\| \geq \|\boldsymbol{X}_L\|$, completing the proof.

For the interested reader, the decomposition (17) was found using the software NCSOSTools by Cafuta et al. (2011). This software finds decompositions of matrix polynomials into sums of Hermitian squares.

Appendix C. Additional calculations for random matrices

To verify that (9) holds for our random matrix model, we first record the following property about the fourth moments of the entries of the $\boldsymbol{A}_i$. Let $\xi := \mathbb{E}[G_{ij}^4]^{1/4}$. Then we have

Lemma 6 *For any fixed $u = 1, \ldots, n$:*

$$\mathbb{E}[A^u_{i_1,j_1}A^u_{i_2,j_2}] = \begin{cases} r(r-1)\sigma^4 + r\xi^4 & \{i_1,j_2\} = \{i_2,j_2\} \text{ and } i_1 = i_2 \\ r\sigma^4 & \{i_1,j_2\} = \{i_2,j_2\} \text{ and } i_1 \neq i_2 \\ 0 & \text{otherwise} \end{cases}$$

Proof [of Lemma 6] $A^u_{i,j} = \sum_{k=1}^r g_{ik}g_{ik}$ so that if $i \neq j$ then

$$\mathbb{E}_{\mathcal{G}}[(A^u_{ii})^2] = \mathbb{E}_{\mathcal{G}}\left[\sum_{k=1,k'=1}^{r} g_{ik}^2 g_{ik'}^2\right] = \mathbb{E}_{\mathcal{G}}[\sum_k g_{ik}^4] + \sum_{k,k'=1:k\neq k'}^{r} \mathbb{E}_{\mathcal{G}}[g_{ik}^2]\,\mathbb{E}_{\mathcal{G}}[g_{ik'}^2] = r\xi^4 + r(r+1)\sigma^4\,.$$

$$\mathbb{E}_{\mathcal{G}}[(A^u_{ij})^2] = \mathbb{E}_{\mathcal{G}}\left[\sum_{k=1,k'=1}^{r} g_{ik}g_{jk}g_{ik'}g_{jk'}\right] = \sum_{k=1}^{r}\mathbb{E}_{\mathcal{G}}[g_{ik}^2 g_{jk}^2] = r\sigma^4\,.$$

■

A consequence of this lemma is that $\mathbb{E}[\boldsymbol{A}^2] = (r(r+d-1)\sigma^4 + r\xi^4)\boldsymbol{I}_d$. Using this identity, we can set $\zeta := r(r+d-1)\sigma^4 + r\xi^4$ and we then have

$$\mathbb{E}[\boldsymbol{V}(i_1,\ldots,i_k)] = \mathbb{E}[\boldsymbol{A}_{i_k}\ldots\boldsymbol{A}_{i_1}^2\ldots\boldsymbol{A}_{i_k}] = \zeta\,\mathbb{E}[\boldsymbol{A}_{i_k}\ldots\boldsymbol{A}_{i_2}^2\ldots\boldsymbol{A}_{i_k}] = \cdots = \zeta^k\boldsymbol{I}_d\,.$$

We compute this identity in a second way that describes its combinatorics more explicitly, which we will use as to derive our lower bound.

$$\begin{aligned}\mathbb{E}[V_{u,v}(i_1,\ldots,i_k)] &= \mathbb{E}\left[\sum_{p\in[d]^{2k}} A^{(i_1)}_{u,p_1} A^{(i_{2n})}_{p_2,v} \prod_{j=2}^{k-1} A^{(i_j)}_{p_{i_{j-1}},p_{i_j}} A^{(i_{2k-j+1})}_{p_{i_{2k-j}},p_{i_{2k-j+1}}}\right] \\ &= \sum_{p\in[d]^{2k}} \mathbb{E}[A^{(i_1)}_{u,p_1} A^{(i_{2n})}_{p_2,v}] \prod_{j=2}^{k-1} \mathbb{E}[A^{(i_j)}_{p_{i_{j-1}},p_{i_j}} A^{(i_{2k-j+1})}_{p_{i_{2k-j}},p_{i_{2k-j+1}}}]\,.\end{aligned}$$

The second equality uses linearity coupled with the fact that $i_1,\ldots,i_k$ are distinct, hence $\mathbb{E}[A^{(i_j)}_{u,v} A^{(i_l)}_{u',v'}] = \mathbb{E}[A^{(i_j)}_{u,v}]\,\mathbb{E}[A^{(i_l)}_{u',v'}]$ since elements from distinct matrices are independent. Many of the terms in this sum contain odd powers which are zero. Using the fact that $\boldsymbol{A} = \boldsymbol{A}^T$, we see that all terms that are non-zero must contain only products of two forms: A^2_{uu} or A^2_{uv}. Then, we can write the sum:

$$V_{u,v}(i_1,\ldots,i_k) = \sum_{p\in[d]^k} \mathbb{E}[(A^{(i_1)}_{u,p_1})^2] \prod_{j=2}^{k} \mathbb{E}[(A^{(i_j)}_{p_{j-1},p_j})^2]\,.$$

Now consider the case that some index may be repeated (i.e., there exist k,l such that $i_j = i_l$ for $j \neq l$). The key observation is the following. Let w be a real-valued random variable with a finite second moment. Then,

$$\mathbb{E}[w^{2p}] \geq \mathbb{E}[w^2]^p \text{ for } p = 0,1,\ldots,n\,, \tag{18}$$

with equality only for $p = 0, 1$. This is Jensen's inequality applied to x^p for $x \geq 0$ (since w is real then w^2 is positive, and x^p is convex on $[0,\infty)$ for $p = 0,1,2,\ldots$.)

Now consider the case when $i_1,\ldots,i_k$ may be repeated. Let n_i be the number of times index i is repeated.

$$\begin{aligned}\mathbb{E}[\boldsymbol{V}(i_1,\ldots,i_k)] &= \mathbb{E}[\sum_{\bar{p}\in[d]^2}\sum_{u,v} x_u x_v A^{(i_1)}_{u,p(1)} A^{(i_{2N})}_{p(2),v} \prod_{j=2}^{k-1} A^{(i_j)}_{p(i_{j-1}),p(i_j)} A^{(i_{2k-j+1})}_{p(i_{2k-j}),p(i_{2k-j+1})}] \\ &\geq \sum_{p\in[d]^k}\sum_{u,v} x_u x_v\, \mathbb{E}[(A^{(i_1)}_{u,p(1)})^2 \prod_{j=2}^{k} (A^{(i_j)}_{p(j-1),p(j)})^2] \\ &\geq \sum_{p\in[d]^k}\sum_{u,v} x_u x_v\, \mathbb{E}[(A^{(i_1)}_{u,p(1)})^2] \prod_{j=2}^{k} \mathbb{E}[(A^{(i_j)}_{p(j-1),p(j)})^2]\end{aligned}$$

The first inequality follows from Lemma 6, since all terms are non-negative. The second inequality is repeated application of (18). The final equality is the calculation we performed for our second proof of the equality condition.

This last expression is precisely equal to the without-replacement average. On the other hand, observe that (18) is strict for $p \geq 2$ for Wishart random variables. If there is even a single repeated value, i.e., $i_j = i_l$ for $j \neq l$, the second inequality above is also strict. Thus, equality is achieved if and only if all indicies are distinct.

For the special case of Wishart matrices, we can show that this gap is quite large.

Lemma 7 *For $i = 1, \ldots, k$, we have*

$$\mathbb{E}_{\mathcal{G}}[\boldsymbol{V}(i, \ldots, i)] \geq \|x\|^2 r 2^{-2k} \frac{4k!}{2k!} \sigma^{4k}$$

Proof

$$\begin{aligned}\mathbb{E}[\boldsymbol{V}(xi, \ldots, i)] &\geq \sum_{u=1}^{d} x_u^2 \, \mathbb{E}_{\mathcal{G}}[A_{u,u}^{(i)} A_{u,u}^{(i)} \prod_{j=2}^{k-1} A_{u,u}^{(i)} A_{u,u}^{(i)}] \\ &= \|x\|^2 \, \mathbb{E}[(A_{u,u}^{(i)})^{2k}] \,.\end{aligned}$$

The first inequality is because all the terms are positive and we are selecting out only the self loops. The equality just groups terms. The following lower bound completes the proof.

$$\mathbb{E}[(A_{uu}^{i})^{2k}] = \mathbb{E}\left[\sum_{l_1,\ldots,l_r} \binom{2k}{l_1, \ldots, l_r} \prod_{l=1}^{r} g_{il}^2 l_i\right] \geq \sum_{l=1}^{r} \mathbb{E}[g_{il}^{4k}] = r 2^{-2k} \frac{4k!}{2k!} \sigma^{4k} \,.$$

■

A simple corollary is the following lower bound on the AM

$$\mathbb{E}_{wr} \, \mathbb{E}[\boldsymbol{V}(i_1, \ldots, i_k)] \geq k^{-k} \|x\|^2 r 2^{-2k} \cdot \frac{4k!}{2k!} \sigma^{4k}$$

We examine the following ratio $\rho(r, k, d)$

$$\rho(k, r, d) = \frac{\mathbb{E}_{wr} \, \mathbb{E}_{\mathcal{G}}[\boldsymbol{G}(x, i_1, \ldots, i_k)]}{\mathbb{E}_{wo} \, \mathbb{E}_{\mathcal{G}}[\boldsymbol{G}(x, i_1, \ldots, i_k)]} \geq r \frac{4k!}{2k!} (4kr(r + d + 1))^{-k} \,.$$

For fixed r, d, ρ grows exponentially with k.

Lemma 8 *For $k, r, d \geq 0$ then*

$$\rho(k, r, d) \geq r e^{\frac{1}{4k(k+1)}} \left(\frac{16k}{e^2 r(r + d + 1)}\right)^k$$

Proof We use a very crude lower and upper bound pair that holds for all k (Cormen et al., 2009, p. 55).

$$\sqrt{2\pi k} \left(\frac{k}{e}\right)^k e^{\frac{1}{2k+1}} \leq k! \leq \sqrt{2\pi k} \left(\frac{k}{e}\right)^k e^{1/2k} \,.$$

Using this, we can write:

$$\begin{aligned}\rho(k, r, d) &\geq r \exp\{k \ln(4k/e)^4 - k \ln(2k/e)^2 - k \ln(4kr(r + d + 1)) + \frac{1}{2k} - \frac{1}{2k + 1}\} \\ &= r \exp\left\{k \ln \frac{4^2 k}{e^2 r(r + d + 1)} - \frac{1}{4k(k + 1)}\right\} = r \left(\frac{16k}{e^2 r(r + d + 1)}\right)^k e^{\frac{1}{4k(k+1)}} \,.\end{aligned}$$

■

Appendix D. Harmonic frames in higher dimensions

The generalization to higher dimensions uses a collection of harmonic frames described by Goyal et al. (2001). Again, let $\omega_n = \pi/n$.

If d even, set

$$\boldsymbol{f}_{k+1} = \sqrt{\frac{2}{d}} \begin{bmatrix} \boldsymbol{a}_{k;n} \\ \boldsymbol{a}_{3k;n} \\ \vdots \\ \boldsymbol{a}_{(d-1)k;n} \end{bmatrix} \quad \text{for } k = 0, 1, \ldots, n-1.$$

One can verify again using standard trigonometric identities that

$$\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{f}_k \boldsymbol{f}_k^T = \frac{1}{d}\boldsymbol{I}$$

Note that the inner products of adjacent $\boldsymbol{f}_i$ can be expressed in terms of a Fejer kernel

$$\boldsymbol{f}_i^T \boldsymbol{f}_{i+1} = \frac{2}{d}\sum_{t=1}^{d/2} \cos\left((2t-1)\omega_n\right) = \frac{2}{d}\cos\left((d/2-1)\omega_n\right)\frac{\sin\left((d/2+1)\omega_n\right)}{\sin\left(\omega_n\right)} - \frac{2}{d}\cos\left(\omega_n\right)$$

Setting $\boldsymbol{A}_k = \boldsymbol{f}_k \boldsymbol{f}_k^T$, we have

$$\left\| \prod_{k=1}^{n} \boldsymbol{A}_k \right\| = \left(\frac{2}{d}\cos\left((d/2-1)\omega_n\right)\frac{\sin\left((d/2+1)\omega_n\right)}{\sin\left(\omega_n\right)} - \frac{2}{d}\cos\left(\omega_n\right) \right)^{d-1}$$

showing that this order violates the desired arithmetic-geometric mean inequality.

For d odd

$$\boldsymbol{f}_{k+1} = \sqrt{\tfrac{2}{d}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \boldsymbol{a}_{2k;n} \\ \boldsymbol{a}_{4k;n} \\ \vdots \\ \boldsymbol{a}_{(d-1)k;n} \end{bmatrix} \quad \text{for } k = 0, 1, \ldots, n-1\,.$$

We can again check that

$$\boldsymbol{f}_i^T \boldsymbol{f}_{i+1} = -\frac{1}{d} + \frac{2}{d}\sum_{t=0}^{(d-1)/2} \cos\left(2t\omega_n\right) = \frac{2}{d}\cos\left((d-1)/2\omega_n\right)\frac{\sin\left((d+1)/2\omega_n\right)}{\sin\left(\omega_n\right)} - \frac{1}{d}$$

which will again violate the arithmetic-geometric mean inequality.

Appendix E. Harmonic frames satisfy the noncommutative arithmetic-geometric mean inequality

Problem Let S be the SGM of a set of rank 1, idempotent matrices that are parametrized by angles $\phi_1, \ldots, \phi_n$ (This is slightly more general than we need for our theorem above). Our goal is to compute the 2-norm of S:

$$\max_{v:\|v\|=1} v^T S v = \max_{\phi_v \in [0,2\pi]} \frac{1}{n!}\sum_{\sigma \in S_n} \cos(\phi_v - \phi_{\sigma(1)}) \times \prod_{i=1}^{n-1}\cos(\phi_{\sigma(i)} - \phi_{\sigma(i+1)}) \times \cos(\phi_{\sigma(n)} - \phi_v)$$

E.1. Cosine combinatorics

We will write this function as fourier transform (we pull out the 2^n for convenience):

$$2^n \max_{v:\|v\|=1} v^T S v = \sum_k c_k \cos(k\pi n^{-1}) + \max_{\phi_v} \sum_l d_l \cos(\phi_v + l\pi n^{-1}) \tag{19}$$

To find the c_k and d_k, we repeatedly apply the following identity:

$$\cos x \cos y = 2^{-1}(\cos(x+y) + \cos(x-y))$$

Fix $\phi_1, \dots, \phi_2, \dots, \phi_n, \dots \in [0, 2\pi]$. We first consider a related form, T_n, for $n = 1, 2, 3, \dots$, defined by the following recurrence

$$T_1 = 1 \text{ and } T_{n+1} = T_n \cos(\psi_n - \psi_{n+1})$$

We compute T_n using the above transformation. But, first, we show the pattern by example:

Example 1

$$\begin{aligned}
&T_1 = 1\\
&T_2 = \cos(\psi_1 - \psi_2)\\
&T_3 = \cos(\psi_1 - \psi_3) + \cos(\psi_1 - 2\psi_2 + \psi_3)\\
&T_4 = \cos(\psi_1 - \psi_4) + \cos(\psi_1 - 2\psi_3 + \psi_4) + \cos(\psi_1 - 2\psi_2 + 2\psi_3 - \psi_4)\cos(\psi_1 - 2\psi_2 + \psi_4)\\
&T_4 = \cos(\psi_1 - \psi_4) + \cos(\psi_1 - 2\psi_2 + 2\psi_3 - \psi_4) + \cos(\psi_1 - 2\psi_3 + \psi_4) + \cos(\psi_1 - 2\psi_2 + \psi_4)\\
&T_5 = \cos(\psi_1 - \psi_5) + \cos(\psi_1 - 2\psi_2 + 2\psi_3 - \psi_5) + \cos(\psi_1 - 2\psi_3 + \psi_5) + \cos(\psi_1 - 2\psi_2 + \psi_5)\\
&\quad = \cos(\psi_1 - 2\psi_4 + \psi_5) + \cos(\psi_1 - 2\psi_2 + 2\psi_3 - 2\psi_4 + \psi_5) + \cos(\psi_1 - 2\psi_3 + 2\psi_4 - \psi_5)\\
&\qquad\qquad + \cos(\psi_1 - 2\psi_2 + 2\psi_4 - \psi_5)
\end{aligned}$$

In our computation above, $\psi_1 = \phi_v = \psi_n$. And so, after writing this out, we will get two kinds of terms: even terms (corresponding to c_k) that do not depend on ψ_v (they cancel) and odd terms that do contain $2\psi_v$.

We encapsulate this example in a lemma:

Lemma 9 *With T_n as defined above, we have for $n \geq 2$*

$$T_n = 2^{-n} \sum_k \sum_{\substack{i_1,\dots,i_k \\ 1<i_1<i_2<\dots<i_k<n}} \cos(\psi_1 - 2\psi_{i_1} + 2\psi_{i_2} - \dots + (-1)^k 2\psi_{i_k} + (-1)^{k+1}\psi_n)$$

Proof By induction, we have:

$$T_{n+1} = 2^{-n} \sum_k \sum_{\substack{i_1,\dots,i_k \\ 1<i_1<i_2<\dots<i_k<n}} \cos(\psi_1 - 2\psi_{i_1} + 2\psi_{i_2} - \dots + (-1)^k \psi_{i_k} + (-1)^{k+1}\psi_n) \cos(\psi_n - \psi_{n+1})$$

$$= 2^{-(n+1)} \sum_k \sum_{\substack{i_1,\dots,i_k \\ 1<i_1<i_2<\dots<i_k<n}} \cos(\psi_1 - 2\psi_{i_1} + 2\psi_{i_2} - \dots + (-1)^k \psi_{i_k} + (-1)^{k+2}\psi_{n+1})$$

$$+ \cos(\psi_1 - 2\psi_{i_1} + 2\psi_{i_2} - \dots + (-1)^k \psi_{i_k} + 2(-1)^{k+1}\psi_n + (-1)^{k+2}\psi_{n+1})$$

$$= 2^{-(n+1)} \sum_k \sum_{\substack{i_1,\dots,i_k \\ 1<i_1<i_2<\dots<i_k<n+1}} \cos(\psi_1 - 2\psi_{i_1} + 2\psi_{i_2} - \dots + (-1)^k \psi_{i_k} + (-1)^{k+2}\psi_{n+1})$$

■

Fix an n. We now count a symmetrized version of T_n defined as follows: For $\sigma \in S_n$:

$$S_n = \frac{1}{n!} \sum_{\sigma \in S_n} \prod_{i=1}^{n-1} \cos(\sigma(i) - \sigma(i+1))$$

We now show that S_n can be written in a form that removes the permtuation. We also assume some structure here that mimics our product above, namely that $\phi_1 = \phi_n$.

Lemma 10 *Let $\phi_1, \dots, \phi_n \in [0, 2\pi]$ such that $\phi_1 = \phi_n$. Then,*

$$S_n = \sum_{X,Y \subseteq [n] : |X| = |Y| \text{ or } |X| = |Y|+1} \binom{|X|+|Y|}{|Y|}^{-1} \cos\left(2\left(\sum_{i \in X} \phi_i - \sum_{j \in Y} \phi_j\right)\right)$$

Proof To see this formula, Consider a pair of sets $X, Y \subseteq [n]$. In how many permutations $\sigma \in S_n$ does (X, Y) contribute a term? We need to choose $|X| + |Y|$ positions for these terms to appear out of n possible places in the order. Thus there are $\binom{n}{|X|+|Y|}$ permutations to choose the slots for (X, Y). An $(|X|, |Y|)$ pair only appears in a permutation in σ if the elements of X and Y can be alternated starting with X. This implies that $|X| = 2|Y| + z_i$ where $z_i \in \prime, \infty$. Moreover, there are $|X|!|Y|!(n - (|X| + |Y||)!$ permutations that respect this structure (for any choice of $|X| + |Y|$ slots, any ordering of X and Y and the elements outside can occur).

$$\binom{n}{|X|+|Y|} |X|!|Y|!(n - (|X| + |Y|)! = n! \binom{|X|+|Y|}{|Y|}^{-1}$$

Pushing the $1/n!$ factor inside completes the proof. ■

E.2. Counting on harmonic, finite groups

In the case we care about, the ϕ_i have more structure: the set $\in \phi \rangle_{i=1}^{n}$ forms a cyclic group under addition modulo 2π. Let n denote the number of elements in the frame. Fix n. Let ζ denote a nth root of unity. Define a (harmonic) generating function f

$$f(\zeta, y, z) = \prod_{i=1}^{n}(1 + \zeta^i x + \zeta^{-i} y)$$

We give a shorthand for its coefficients $q_{k,m}$ and $r_{k,m}$ as follows

$$q_{k,m} := [\zeta^k x^m y^m]f \text{ and } r_{k,m} := [\zeta^k x^{m+1} y^m]f$$

We observe that $q_{k,m}$ computes the number of sets (X, Y) where $X, Y \subseteq \mathbb{Z}_n$ such that:

1. $\sum_{i \in X} i - \sum_{j \in Y} j = k \mod n$ (since we inspect ζ^k),
2. $|X| = |Y| = m$ (since we inpect $x^m y^m$),

For $r_{k,m}$ the only change is that $|X| = |Y| + 1$ (since $x^{m+1}y^m$). With this notation, we can express the coefficients from Eq. 19.

$$c_k = \sum_m q_{k,m} \binom{2m}{m}^{-1} \text{ and } d_k = \sum_m r_{k,m} \binom{2m+1}{m}^{-1}$$

We use this representation to prove that the SGM is rotationally invariant (i.e., $d_k = 0$ for $k = 0, 1, \ldots, n-1$). First, we show that all d_k are equal.

Lemma 11 *Consider a frame of size n. For any m and $k, l = 0, \ldots, n-1$, $\sum_k d_k \cos(\phi_v + 2\pi k) = 0$.*

Proof This follow by examining the generating function above. First observe that we have congruence $f(x, x^j y, z) = f(x, y, z)$ for $j = 0, \ldots, n-1$ tells us that $[x^j y z^m]f = [y z^m]f$. And, the congruence that $[x^j y z^m]f = [x^j y^{-1} z^m]f$. Combining these facts, we have that $r_{k,m} = r_{l,m}$. Since this holds for all k, l, we can conclude that $d_k = d_l$ by summing over m. Finally, since $\sum_{l=0}^{n} \cos(\phi_v + 2l\pi n^{-1}) = 0$ for any fixed ϕ_v we conclude the lemma. ∎

Since the SGM does not depend on ϕ_v, we conclude it must be of the form αI for some α. The remainder of this note is to compute that α.

E.3. Computing the coefficients

The argument of this subsection is a generalization of that of Konvalina (1995).

Lemma 12

$$q_{k,m} = (-1)^k \binom{n-k}{k} \frac{n}{n-k}$$

Proof Define R_n as:

$$R_n(x, y) = x^n + (-y)^n - \sum_k \binom{n-k}{k} \frac{n}{n-k} (xy)^k$$

Since $f(\zeta, x, y) = f(\zeta^i, x, y)$ for any integer n, $F_n(x, y) = f(\zeta, x, y)$ is a function of n alone. That is, we can write

$$F_n(x, y) = \prod_{\zeta \in U_n} (1 + \zeta x + \zeta^{-i} y)$$

Thus, claim boils down to $F_n(x, -y) = R_n(x, y)$.

We show that the zero sets of $F_n(x, -y)$ and R_n are equal. The zero set of $F_n(x, -y)$ is the set of lines described by

$$\{(x, y) \mid y = \zeta + \zeta^2 x\} \text{ for } \zeta \in U_n$$

where ζ is any n-th root of unity. Substituting y at the root equation, we get that $xy = x\zeta + \zeta^2 x^2$.

Now, we check that the following is zero:

$$x^n + (-y)^n - \sum_k \binom{n-k}{k} \frac{n}{n-k} (\zeta x + \zeta^2 x^2)^k$$

Here, we use the generating function:

$$\sum_k \frac{n}{n-k} \binom{n-k}{k} y^k = \left(\frac{1 - \sqrt{1+4y}}{2} \right)^n + \left(\frac{1 + \sqrt{1+4y}}{2} \right)^n$$

using this sum, we have:

$$\begin{aligned}
&\sum_{k=0}^{n} \binom{n-k}{k} \frac{n}{n-k} (\zeta x + \zeta^2 x^2)^k \\
=&\left(\frac{1 - (2\zeta x + 1)}{2} \right)^n + \left(\frac{1 + (2\zeta x + 1)}{2} \right)^n \\
=&(\zeta x)^n + (1 + \zeta x)^n \\
=&x^n + (-y)^n
\end{aligned}$$

The first equality follows from $1 + 4\zeta x + 4\zeta^2 x^2 = (2x\zeta + 1)^2$. The second is just algebra. Finaly, we use on each term that $\zeta^n = 1$ and that $\zeta + \zeta^2 x = -y$. This claim holds for all ζ that are roots of unity, and so the function is identically zero.

To conclude the proof, observe that the zero set described above is the union of n lines of the form $(1 + \zeta x + \zeta^{-1} y)$. These lines are unique in $\mathbb{C}$: if $(1 + \zeta x + \zeta^{-1} y) = (1 + \omega x + \omega^{-1} y)$ then since the x coefficients are the same $\zeta = \omega$ and so they must be the same. By direct inspection, this R_n can only have these factors (else the total degree would be higher). Hence, $R_n = Q_n$. ■

E.4. Finally, to a hypergeometric series

It is possible to get an explicit formula for λ that is related to ${}_3F_2$. We consider the following series and show that it is hypergeometric in k:

$$\sum_k T(n, k) \binom{2k}{k}^{-1} x^k = \sum_k v(k) x^k$$

Consider the ratio:

$$\begin{aligned}\frac{v(k+1)}{v(k)} &= -\frac{(n-k-1)!((k+1)!)^2}{k+1!(n-2k-2)!(n-k-1)(2k+2)!}\frac{2k!(n-2k)!k!(n-k)}{(n-k)!(k!)^2} \\ &= -\frac{(n-2k)(n-2k-1)(k+1)}{(n-k-1)(2k+2)(2k+1)} \\ &= \frac{(k-n/2)(k-n/2+1/2)(k+1)}{(k-n+1))(k+1/2)(k+1)}\end{aligned}$$

And so, this is a hypergeometric:

$${}_2F_3\left[\begin{array}{ccc} 1 & -n/2+1/2 & -n/2 \\ 1/2 & -n+1 & \end{array};1\right] = \mathcal{O}(n^{-1})$$

This completes the proof.