Open Problem: Can Single-Shuffle SGD be Better than Reshuffling SGD and GD?

Chulhee Yun CHULHEEY@MIT.EDU
Suvrit Sra SUVRIT@MIT.EDU
Ali Jadbabaie JADBABAI@MIT.EDU
Massachusetts Institute of Technology

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We propose matrix norm inequalities that extend the Recht and Ré (2012) conjecture on a noncommutative AM-GM inequality, by supplementing it with another inequality that accounts for *single-shuffle* in stochastic finite-sum minimization. Single-shuffle is a popular without-replacement sampling scheme that shuffles only once in the beginning, but has not been studied in the Recht-Ré conjecture and the follow-up literature. Instead of focusing on general positive semidefinite matrices, we restrict our attention to positive definite matrices with small enough condition numbers, which are more relevant to matrices that arise in the analysis of SGD. For such matrices, we conjecture that the means of matrix products satisfy a series of spectral norm inequalities that imply "*single-shuffle SGD converges faster than random-reshuffle SGD, which is in turn faster than with-replacement SGD and GD*" in special cases.

1. Introduction: Recht-Ré matrix AM-GM inequality conjecture

Stochastic gradient descent (SGD) and its variants have become indispensable to solving finite-sum optimization problems that arise in modern machine learning. At each iteration, these methods evaluate the gradient of one component function sampled from the entire set of components and use it as a noisy estimate of the full gradient.

Depending on how the components are chosen, SGD-based methods can broadly be classified into two categories: *with-replacement* and *without-replacement*. In many theoretical studies the indices of component functions are assumed to be chosen with replacement, making the choice at each iteration independent of other iterations. In contrast, the vast majority of practical implementations use without-replacement sampling, where all the indices are randomly shuffled and are then visited exactly once per *epoch* (i.e., one pass through all the components). There are two popular variants of shuffling schemes: one that *reshuffles* the components at every epoch and another that *shuffles only once* at the beginning and reuses that order every epoch.

Practitioners opt for without-replacement sampling schemes not only because they are easier to implement, but also because they often result in faster convergence (Bottou, 2009). However, analyzing algorithms based on without-replacement sampling is considerably trickier than their with-replacement counterparts, because the component chosen at each iteration is dependent on the previous iterates of an epoch. Despite the difficulty, recent results have shown tight convergence bounds for without-replacement SGD that are *faster* than with-replacement SGD (Nagaraj et al., 2019; Safran and Shamir, 2020; Rajput et al., 2020; Ahn et al., 2020; Mishchenko et al., 2020).

In 2012, Recht and Ré (2012) proposed a conjecture that the mean of without-replacement products of positive semidefinite (PSD) matrices has spectral norm no larger than the mean of their with-replacement products. Formally, given n real PSD matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_n$, they conjectured:

$$\left\| \frac{1}{n!} \sum\nolimits_{\sigma \in \mathcal{S}_n} \prod\nolimits_{i=1}^{n} \boldsymbol{A}_{\sigma(i)} \right\| \leq \left\| \frac{1}{n} \sum\nolimits_{i=1}^{n} \boldsymbol{A}_i \right\|^n, \tag{1}$$

where $\mathcal{S}_n$ is the set of all permutations $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$. This so-called *matrix AM-GM inequality*[1] conjecture has drawn much attention because it can help explain why without-replacement algorithms, such as SGD and the randomized Kaczmarz algorithm (Karczmarz, 1937; Strohmer and Vershynin, 2009), may converge faster than with-replacement ones (see Section 2).

The conjecture (1) is true for $n = 2$ (Recht and Ré, 2012) and $n = 3$ (Zhang, 2018; Lai and Lim, 2020). However, it was recently proven to be false for $n \geq 5$. Lai and Lim (2020) disproved the AM-GM inequality (1) by formulating the Positivstellensatz of the matrix polynomials from an equivalent conjecture into a semidefinite program (SDP). By solving the SDP, they proved that (1) does not hold for $n = 5$. The conjecture for higher n's was settled by De Sa (2020), who constructed concrete counterexamples to disprove the conjecture (1) for any $n \geq 5$.

2. Motivations for a revised conjecture

Although this conjecture-and-disproof story seems inescapable, we point out important facts about the conjecture (1) and its disproof, reigniting new hope for a revised/extended conjecture.

Disproofs of Recht-Ré conjecture break for well-conditioned matrices. Both disproofs (Lai and Lim, 2020; De Sa, 2020) of the conjecture (1) leverage rank-deficient PSD matrices, and adding a positive multiple of the identity matrix to each matrix makes them satisfy (1). This leaves open the possibility that the conjecture may still be true for positive definite (PD) matrices with sufficiently small condition numbers. Readers can refer to Yun et al. (2021) for details on breaking the disproofs.

Well-conditioned matrices arise in analysis of SGD. Suppose we want to minimize a finite-sum function $F : \mathbb{R}^d \to \mathbb{R}$ of the form $F(\boldsymbol{z}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\boldsymbol{z})$. For simplicity of illustration, consider quadratic component functions $f_i(\boldsymbol{z}) := \frac{1}{2} \boldsymbol{z}^T \boldsymbol{M}_i \boldsymbol{z}$, where $\boldsymbol{M}_i$'s are d-by-d PSD matrices. We consider running SGD with a constant step-size $\eta > 0$, with the following updates:

$$\boldsymbol{z}_t := \boldsymbol{z}_{t-1} - \eta \nabla f_{i(t)}(\boldsymbol{z}_{t-1}) = \boldsymbol{z}_{t-1} - \eta \boldsymbol{M}_{i(t)} \boldsymbol{z}_{t-1} = (\boldsymbol{I} - \eta \boldsymbol{M}_{i(t)}) \boldsymbol{z}_{t-1}, \tag{2}$$

where $i(t) \in \{1, \dots, n\}$ is the index at iteration t chosen by the algorithm.

With-replacement SGD, which we denote as SGD, chooses $i(t)$ from the uniform distribution over all indices at each iteration. In contrast, without-replacement SGD shuffles the n indices and makes a complete pass through the shuffled indices, which we call an epoch. Two different variants are popular in practice: RANDOMSHUFFLE reshuffles the indices after each epoch, whereas SINGLESHUFFLE shuffles in the beginning and adheres to that order for all the epochs.

1. Developing a matrix counterpart of the well-known scalar arithmetic-geometric means inequality has been a long-standing research topic. Different geometric means and their corresponding AM-GM inequalities have been proposed and studied in the literature (Bhatia and Davis, 1993; Horn, 1995; Bhatia and Kittaneh, 2000; Ando et al., 2004; Bhatia and Holbrook, 2006; Bini et al., 2010; Bhatia and Karandikar, 2012).

Suppose we run nK iterations (K epochs) of the three variants of SGD initialized at $\boldsymbol{z}_0$. It is straightforward to derive that the expected iterates after nK iterations are written as follows:

$$\mathbb{E}[\boldsymbol{z}_{nK}^{\mathrm{SGD}}] = \mathbb{E}_{i\sim\mathrm{Unif}(\{1,\dots,n\})}[\boldsymbol{I}-\eta\boldsymbol{M}_i]^{nK}\boldsymbol{z}_0 = \left(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{I}-\eta\boldsymbol{M}_i)\right)^{nK}\boldsymbol{z}_0, \tag{3}$$

$$\mathbb{E}[\boldsymbol{z}_{nK}^{\mathrm{RS}}] = \mathbb{E}_{\sigma\sim\mathrm{Unif}(\mathcal{S}_n)}\left[\prod_{i=n}^{1}(\boldsymbol{I}-\eta\boldsymbol{M}_{\sigma(i)})\right]^{K}\boldsymbol{z}_0 = \left(\frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\prod_{i=1}^{n}(\boldsymbol{I}-\eta\boldsymbol{M}_{\sigma(i)})\right)^{K}\boldsymbol{z}_0, \tag{4}$$

$$\mathbb{E}[\boldsymbol{z}_{nK}^{\mathrm{SS}}] = \mathbb{E}_{\sigma\sim\mathrm{Unif}(\mathcal{S}_n)}\left[\left(\prod_{i=n}^{1}(\boldsymbol{I}-\eta\boldsymbol{M}_{\sigma(i)})\right)^{K}\right]\boldsymbol{z}_0 = \frac{1}{n!}\sum_{\sigma\in\mathcal{S}_n}\left(\prod_{i=1}^{n}(\boldsymbol{I}-\eta\boldsymbol{M}_{\sigma(i)})\right)^{K}\boldsymbol{z}_0. \tag{5}$$

Comparing (3) and (4), we can notice their connection to the original AM-GM inequality conjecture (1), with $\boldsymbol{A}_i = \boldsymbol{I}-\eta\boldsymbol{M}_i$. If true, inequality (1) would imply that in the special case $f_i(\boldsymbol{z}) = \frac{1}{2}\boldsymbol{z}^T\boldsymbol{M}_i\boldsymbol{z}$, the expected iterate of RANDOMSHUFFLE is closer to the global minimum than that of SGD or gradient descent (GD),[2] if we compare upper bounds.

In many recent theoretical advances on without-replacement SGD (Haochen and Sra, 2019; Nagaraj et al., 2019; Rajput et al., 2020; Ahn et al., 2020), $\eta = O(\frac{\log(nK)}{nK})$ is chosen as the step-size to prove tight convergence rates of RANDOMSHUFFLE and SINGLESHUFFLE that are faster than SGD. Note that such choices of η make the matrices $\boldsymbol{A}_i = \boldsymbol{I}-\eta\boldsymbol{M}_i$ close to identity; in other words, the matrices $\boldsymbol{A}_i$ that arise in the analysis of SGD are well-conditioned. Therefore, it is evident that understanding the conjecture for well-conditioned PD matrices is of great importance.

Recht-Ré conjecture (1) fails to account for SINGLESHUFFLE. It is important to note from (3), (4), and (5) that Recht-Ré conjecture (1) only implies faster convergence of RANDOMSHUFFLE than SGD. It does not provide any useful insights towards the analysis of SINGLESHUFFLE, an equally (if not more) popular scheme in practice. This begs the question: *is there an additional inequality that we can add to the conjecture (1) in order to account for* SINGLESHUFFLE*?*

SINGLESHUFFLE beats RANDOMSHUFFLE and SGD in linear regression. Consider solving an underdetermined linear regression problem $F(\boldsymbol{z}) = \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\boldsymbol{z}-y_i)^2$ with a random Gaussian dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ using the variants of SGD and GD. Perhaps surprisingly, experiments show a quite consistent trend that given the same initialization $\boldsymbol{z}_0$ and step-size η, SINGLESHUFFLE minimizes $F(\boldsymbol{z})$ faster than RANDOMSHUFFLE, which in turn outperforms SGD and GD. We observe a similar trend if we plot t vs. the spectral norm of matrix products $\prod_{j=t}^{1}(\boldsymbol{I}-\eta\boldsymbol{x}_{i(j)}\boldsymbol{x}_{i(j)}^T)$ (note from (2) that this matrix product multiplied with $\boldsymbol{z}_0$ appears in $\boldsymbol{z}_t$). See Yun et al. (2021) for more details.

This inspiring empirical observation motivates an extension of the AM-GM inequality conjecture (1), which corresponds to "RANDOMSHUFFLE $\leq$ SGD/GD," to a new conjecture by adding a new inequality "SINGLESHUFFLE $\leq$ RANDOMSHUFFLE."

3. New conjecture: SS-RS-GD inequalities

Based on the motivations discussed so far, we now formally state our new conjecture in Conjecture 1, which we refer to as the *SS-RS-GD inequalities* conjecture. Conjecture 1 extends the AM-GM inequality conjecture (1) by adding the SINGLESHUFFLE algorithm to the picture. It also refines the original conjecture to a setting that is more relevant to the convergence analysis of SGD.

2. Note that the RHS of (3) coincides with the nK-th iterate of GD: $\boldsymbol{z}_t := \boldsymbol{z}_{t-1}-\eta\nabla F(\boldsymbol{z}_{t-1}) = (\boldsymbol{I}-\frac{\eta}{n}\sum_i\boldsymbol{M}_i)\boldsymbol{z}_{t-1}$.

Conjecture 1 *For any $n \geq 2$, $K \geq 1$, and $d \geq 1$, there exists a step-size constant $\eta_{n,K} \in (0, 1]$ such that the following statement holds: Suppose d-by-d real symmetric matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$ satisfy $(1 - \eta_{n,K})\boldsymbol{I} \preceq \boldsymbol{A}_i \preceq \boldsymbol{I}$ for all $i \in \{1, \ldots, n\}$. Then, for matrices*

$$\boldsymbol{W}_{\mathrm{SS}} := \frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \left(\prod_{i=1}^{n} \boldsymbol{A}_{\sigma(i)} \right)^K, \boldsymbol{W}_{\mathrm{RS}} := \left(\frac{1}{n!} \sum_{\sigma \in \mathcal{S}_n} \prod_{i=1}^{n} \boldsymbol{A}_{\sigma(i)} \right)^K, \boldsymbol{W}_{\mathrm{GD}} := \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{A}_i \right)^{nK}, \quad (6)$$

the following spectral norm inequalities hold:[3] $\|\boldsymbol{W}_{\mathrm{SS}}\| \leq \|\boldsymbol{W}_{\mathrm{RS}}\| \leq \|\boldsymbol{W}_{\mathrm{GD}}\|$.

Requirements on "scaling" are not strict. Although we require $(1 - \eta_{n,K})\boldsymbol{I} \preceq \boldsymbol{A}_i \preceq \boldsymbol{I}$, the specific scaling is not strictly necessary, because the norm inequalities do not break when we scale all $\boldsymbol{A}_i$'s with the same factor. Also, for the first inequality $\|\boldsymbol{W}_{\mathrm{SS}}\| \leq \|\boldsymbol{W}_{\mathrm{RS}}\|$, individual scale of $\boldsymbol{A}_i$ is not important because scaling a single $\boldsymbol{A}_i$ does not change the sign of the inequality; rather, the actual requirement is that $\boldsymbol{A}_i$'s have condition numbers at most $\frac{1}{1-\eta_{n,K}}$.

Matrices that commute. For matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n$ that commute, Conjecture 1 is true with $\eta_{n,K} = 1$. The product $\prod_{i=1}^{n} \boldsymbol{A}_{\sigma(i)}$ is identical for all $\sigma \in \mathcal{S}_n$, so the first inequality of the conjecture holds with equality. The second inequality boils down to the scalar AM-GM inequality.

Doesn't Conjecture 1 contradict matrix convexity of $\boldsymbol{A} \mapsto \boldsymbol{A}^2$? While $\|\boldsymbol{W}_{\mathrm{RS}}\| \leq \|\boldsymbol{W}_{\mathrm{GD}}\|$ is the Recht-Ré conjecture (1) with restrictions on $\boldsymbol{A}_i$'s, the new inequality $\|\boldsymbol{W}_{\mathrm{SS}}\| \leq \|\boldsymbol{W}_{\mathrm{RS}}\|$ may look questionable at first glance. It is of the form $\|\mathbb{E}_\sigma[\boldsymbol{P}_\sigma^K]\| \leq \|\mathbb{E}_\sigma[\boldsymbol{P}_\sigma]^K\|$, whose sign is the ***opposite*** of Jensen's inequality. In particular, it is well-known that the map $\boldsymbol{A} \mapsto \boldsymbol{A}^2$ on symmetric matrices is matrix convex, which may seem to contradict $\|\boldsymbol{W}_{\mathrm{SS}}\| \leq \|\boldsymbol{W}_{\mathrm{RS}}\|$. Here, the key is that the product $\prod_{i=1}^{n} \boldsymbol{A}_{\sigma(i)}$ is not necessarily symmetric; hence, there is no contradiction.

Does Conjecture 1 alone imply that SINGLESHUFFLE always converges faster? The short answer is, in general, no. For the special case of $f_i(\boldsymbol{z}) = \frac{1}{2}\boldsymbol{z}^T\boldsymbol{M}_i\boldsymbol{z}$, the conjectured inequalities indeed imply faster convergence of SINGLESHUFFLE. However, for general quadratic functions $f_i(\boldsymbol{z}) = \frac{1}{2}\boldsymbol{z}^T\boldsymbol{M}_i\boldsymbol{z} + \boldsymbol{b}_i^T\boldsymbol{z} + c_i$, linear coefficients $\boldsymbol{b}_i$ introduce "noise" terms and SGD iterates read

$$\boldsymbol{z}_t = \Big(\prod\nolimits_{j=t}^{1} (\boldsymbol{I} - \eta \boldsymbol{M}_{i(j)}) \Big) \boldsymbol{z}_0 - \eta \boldsymbol{b}_{i(t)} - \eta \Big(\sum\nolimits_{l=1}^{t-1} \Big(\prod\nolimits_{j=t}^{l+1} (\boldsymbol{I} - \eta \boldsymbol{M}_{i(j)}) \Big) \boldsymbol{b}_{i(l)} \Big), \quad (7)$$

and the terms involving $\boldsymbol{b}_i$'s become a dominant factor that determines the convergence speed.[4] Although Conjecture 1 alone does not prove superior performance of SINGLESHUFFLE over other algorithms, proving it provides a versatile tool for the analysis of without-replacement sampling methods in general, since the matrix products in Conjecture 1 naturally arise in other setups too.

Preliminary progress. We invite readers to refer to Yun et al. (2021) for some initial progress. Yun et al. (2021) prove $\|\boldsymbol{W}_{\mathrm{SS}}\| \leq \|\boldsymbol{W}_{\mathrm{RS}}\|$ for $\boldsymbol{A}_i = \boldsymbol{I} - \eta \boldsymbol{M}_i$, when η is small enough (but *dependent* on $\boldsymbol{M}_i$'s). Together with Theorem 1 of De Sa (2020), this proves $\|\boldsymbol{W}_{\mathrm{SS}}\| \leq \|\boldsymbol{W}_{\mathrm{RS}}\| \leq \|\boldsymbol{W}_{\mathrm{GD}}\|$ for small enough η. Although this result provides some evidence, it does not prove Conjecture 1 itself because we conjecture that $\eta_{n,K}$'s are *independent* of d and $\boldsymbol{A}_i$'s. Other theorems in Yun et al. (2021) prove Conjecture 1 for special cases and suggest that $\eta_{n,K} = O(1/nK)$ may be *sufficient*. This may look like a strong restriction, but we emphasize that this choice of $\eta_{n,K}$ matches (up to log factors) the order of the step-sizes chosen in analyses on without-replacement SGD (see Section 2). We believe $\eta_{n,K} = O(1/nK)$ is not strictly *necessary*, although $\eta_{n,K}$ has to decay with n and K.

3. The inequalities may also hold for other matrix norms induced from some ellipsoidal norms defined with $\boldsymbol{A}_i$'s.
4. For strongly convex quadratic F, RANDOMSHUFFLE converges faster than SINGLESHUFFLE (Ahn et al., 2020) because the noise terms "cancel out" better. Still, this does not contradict Conjecture 1 which is independent of $\boldsymbol{b}_i$'s.

Acknowledgments

CY acknowledges Korea Foundation for Advanced Studies, NSF CAREER grant 1846088, and ONR grant N00014-20-1-2394 for financial support. SS acknowledges support from NSF BIGDATA grant 1741341, NSF CAREER grant 1846088, an MIT RSC award, and an Amazon Research Award. AJ acknowledges support from ONR grant N00014-20-1-2394, and from MIT-IBM Watson AI lab.

References

Kwangjun Ahn, Chulhee Yun, and Suvrit Sra. SGD with shuffling: optimal rates without component convexity and large epoch requirements. *Advances in Neural Information Processing Systems*, 33, 2020.

Tsuyoshi Ando, Chi-Kwong Li, and Roy Mathias. Geometric means. *Linear algebra and its applications*, 385:305–334, 2004.

Rajendra Bhatia and Chandler Davis. More matrix forms of the arithmetic-geometric mean inequality. *SIAM Journal on Matrix Analysis and Applications*, 14(1):132–136, 1993.

Rajendra Bhatia and John Holbrook. Riemannian geometry and matrix geometric means. *Linear algebra and its applications*, 413(2-3):594–618, 2006.

Rajendra Bhatia and Rajeeva L Karandikar. Monotonicity of the matrix geometric mean. *Mathematische Annalen*, 353(4):1453–1467, 2012.

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

Dario Bini, Beatrice Meini, and Federico Poloni. An effective matrix geometric mean satisfying the ando-li-mathias properties. *Mathematics of Computation*, 79(269):437–452, 2010.

Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, 2009.

Christopher M De Sa. Random reshuffling is not always better. *Advances in Neural Information Processing Systems*, 33, 2020.

Jeff Haochen and Suvrit Sra. Random shuffling beats SGD after finite epochs. In *International Conference on Machine Learning*, pages 2624–2633, 2019.

Roger A Horn. Norm bounds for hadamard products and an arithmetic-geometric mean inequality for unitarily invariant norms. *Linear algebra and its applications*, 223:355–361, 1995.

S Karczmarz. Angenaherte auflosung von systemen linearer glei-chungen. *Bull. Int. Acad. Pol. Sic. Let., Cl. Sci. Math. Nat.*, pages 355–357, 1937.

Zehua Lai and Lek-Heng Lim. Recht-Ré noncommutative arithmetic-geometric mean conjecture is false. In *International Conference on Machine Learning*, 2020.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *arXiv preprint arXiv:2006.05988*, 2020.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. SGD without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pages 4703–4711, 2019.

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of SGD without replacement. In *International Conference on Machine Learning*, 2020.

Benjamin Recht and Christopher Ré. Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. In *Conference on Learning Theory*, pages 11–1, 2012.

Itay Safran and Ohad Shamir. How good is SGD with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020.

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

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Can single-shuffle SGD be better than reshuffling SGD and GD? *arXiv preprint arXiv:2103.07079*, 2021.

Teng Zhang. A note on the matrix arithmetic-geometric mean inequality. *The Electronic Journal of Linear Algebra*, 34:283–287, 2018.