JMLR: Workshop and Conference Proceedings vol 23 (2012) 11.25–11.27 25th Annual Conference on Learning Theory

Commentary on "Towards a Noncommutative Arithmetic-Geometric Mean Inequality" by B. Recht and C. Ré

John C. Duchi JDUCHI@EECS.BERKELEY.EDU
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720 USA

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

In their paper, Recht and Ré have presented conjectures and consequences of non-commutative variants of the arithmetic mean-geometric mean (AM-GM) inequality for positive definite matrices. Let $A_1, \ldots, A_n$ be a collection of positive semidefinite matrices and $i_1, \ldots, i_k$ be random indices in $\{1, \ldots, n\}$. To avoid symmetrization issues that arise since matrix products are non-commutative, Recht and Ré define the expectation operators

$$\mathbb{E}_{\text{wo}}[f(A_{i_1}, \ldots, A_{i_k})] := \frac{(n-k)!}{n!} \sum_{\substack{\{j_1,\ldots,j_k\} \\ \text{distinct}}} f(A_{j_1}, \ldots, A_{j_k})$$

and

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

Let $\|\!|\cdot|\!\|$ denote the operator norm of a matrix. The authors' main conjecture in the paper is that for any $k \leq n$, the following arithmetic-geometric mean inequality holds:

$$\left\|\!\left| \mathbb{E}_{\text{wo}}\left[\prod_{j=1}^{k} A_{i_j}\right] \right|\!\right\| \leq \left\|\!\left| \mathbb{E}_{\text{wr}}\left[\prod_{j=1}^{k} A_{i_j}\right] \right|\!\right\|. \tag{1}$$

Even formalizing appropriate non-commutative variants of AM-GM-type inequalities is difficult (e.g. Bhatia, 2007, Chapter 6), so the suggestion of useful inequalities is important.

A more general question that naturally asserts itself is for which (scalar or matrix-valued) functions f can one obtain inequalities of the form (1), that is, identifying situations when

$$\left\|\!\left| \mathbb{E}_{\text{wo}}\left[f(A_{i_1}, \ldots, A_{i_k})\right] \right|\!\right\| \leq \left\|\!\left| \mathbb{E}_{\text{wr}}\left[f(A_{i_1}, \ldots, A_{i_k})\right] \right|\!\right\|. \tag{2}$$

As a natural starting point, taking f to be the norm of its arguments' products,

$$f : \mathbb{R}^{m\times m} \times \cdots \times \mathbb{R}^{m\times m} \to \mathbb{R}_+, \quad (A_1, \ldots, A_k) \mapsto f(A_1, \ldots, A_k) = \left\|\!\left| A_1 A_2 \cdots A_k \right|\!\right\|, \tag{3}$$

we obtain a slightly different form of of the initial conjecture (1). To provide motivation for studying such alternate inequalities, I provide two examples for which the setting (3) appears more natural than the proposed AM-GM inequality (1).

I begin with one of Recht and Ré's examples, the Kaczmarz algorithm. The algorithm is an iterative algorithm for solving the linear system $\Phi x = y$, where $\Phi \in \mathbb{R}^{n \times d}$ with $n > d$ and there exists an x_* such that $\Phi x_* = y$. Let $\phi_i^\top$ denote the ith row of the matrix Φ. In its kth iteration, the Kaczmarz algorithm selects an index $i_k \in [n]$ and iterates

$$x_k = x_{k-1} + \frac{y_{i_k} - \phi_{i_k}^\top x_{k-1}}{\|\phi_{i_k}\|_2^2} \phi_{i_k}.$$

Defining $A_i = I - \phi_i \phi_i^\top / \|\phi_i\|_2^2$, it is not difficult to see that since $\phi_i^\top x_* = y_i$,

$$x_k - x_* = \prod_{j=1}^{k} A_{i_j} (x_0 - x_*).$$

The Kaczmarz algorithm's convergence rate is then given by taking the function f to be

$$f(A_1, \ldots, A_k) = \|A_1 A_2 \cdots A_k (x_0 - x_*)\|_2 \le |\!|\!| A_1 A_2 \cdots A_k |\!|\!| \, \|x_0 - x_*\|_2$$

in the conjecture (2). This is distinct from the norm inequality (1) conjectured by the authors, and it will require different analysis than that the authors present using random vectors in the incremental gradient method. The distinction in the setting here is that Φ is fixed—so we must condition on it (i.e. on the matrices A_i)—and we wish to understand the relative convergence rates when only the indices i are random.

The second example, whose analysis may be somewhat easier, arises out of work on distributed consensus and averaging algorithms. In consensus algorithms, a network of m nodes, each node $i \in [m]$ owning a real-valued parameter $x^i \in \mathbb{R}$, wish to compute the average $\bar{x} = (1/m) \sum_{i=1}^m x^i$ efficiently using local message-passing. Gossip algorithms (see, e.g., Boyd et al., 2006) are robust, low-communication iterative schemes to achieve this averaging and proceed as follows. Let $x_k \in \mathbb{R}^m$ denote the nodes' parameters at iteration k of the algorithm. The kth iteration consists of selecting a random pair (i, j) of connected nodes in the network and averaging the values of the selected pair, while every other node in the network does nothing. If we define the matrix $A_{i,j}$ to be the identity, except that entries (i, i), (j, j), (i, j), and (j, i) are equal to $\frac{1}{2}$, we see that $x_k = \prod_{l=1}^k A_{i_l, j_l} x_0$. Defining $\mathbb{1} \in \mathbb{R}^m$ to be the all-ones vector, we have $A_{i,j} \mathbb{1} = \mathbb{1}$ and $(1/m) \mathbb{1}\mathbb{1}^\top x_0 = \bar{x}$, and we can measure the convergence of a gossip-style method by

$$f(A_{i_1, j_1}, \ldots, A_{i_k, j_k}) = \|x_k - \bar{x}\mathbb{1}\| = \left\| \prod_{l=1}^k A_{i_l, j_l} (x_0 - \bar{x}\mathbb{1}) \right\| = \left\| \prod_{l=1}^k \left(A_{i_l, j_l} - \frac{1}{m} \mathbb{1}\mathbb{1}^\top \right) x_0 \right\|. \quad (4)$$

Boyd et al. (2006) give convergence rates of randomized gossip algorithms (edges are sampled with replacement) that are optimal for certain types of networks, such as expander graphs. However, for structured networks such as cycles and grids, Dimakis et al. (2008) (and others following) demonstrate that directed and slightly less random communication yields substantial improvements in convergence rate. Can we obtain similar improvements by using sampling without replacement in gossip algorithms? Such improvements would also yield improvements in distributed optimization algorithms based on local message passing (e.g. Nedić and Ozdaglar, 2009; Duchi et al., 2012). Given the special structure of

	Iterations k	$k = 2m$	$k = 4m$	$k = 10m$	$k = 20m$
Ratio $\|x_k^{\text{wr}} - \bar{x}\mathbb{1}\| \,/\, \|x^{\text{wo}} - \bar{x}\mathbb{1}\|$	$m = 25$	2.010	3.125	11.678	97.787
	$m = 100$	1.337	1.408	1.901	2.965
	$m = 400$	1.288	1.234	1.277	1.411

Table 1: With-replacement versus without-replacement sampling for gossip.

the matrices in the product (4) (they have two off-diagonal entries and are idempotent), it may be easier to prove an inequality of the form (2) when we take f as in (4).

A simple simulation suggests—as Recht and Ré conjecture—that without replacement sampling is better. Consider a toroidal grid network with $m = \{25, 100, 400\}$ nodes (such a network has $2m$ edges). Let $x_k^{\text{wo}} \in \mathbb{R}^m$ denote the node values in the network after k iterations of without replacement gossip and $x_k^{\text{wr}} \in \mathbb{R}^m$ denote node values using with-replacement sampling. We may study the relative convergence rates by computing the ratio $\|x_k^{\text{wr}} - \bar{x}\mathbb{1}\| \,/\, \|x_k^{\text{wo}} - \bar{x}\mathbb{1}\|$ as a function of the number of iterations k and the network size m; in the simulation I allow k to be larger than the number of edges by choosing a permutation of all the edges in the network after each edge has been selected once. Table 1 shows the mean of the ratios over 200 experiments for each of the different network sizes; the ratios are always positive, suggesting the benefits of without replacement sampling.

Recht and Ré have opened an interesting avenue of research with their progress toward noncommutative arithmetic-geometric mean inequalities. It will be interesting to see whether quantitative versions of the inequalities (1) or (2) are true; I also look forward to the new algorithmic and statistical insights such inequalities will provide.

Acknowledgments

I thank Lester Mackey for stimulating discussions on matrix inequalities and Recht and Ré's paper in particular, including questions of the placement of the norm in the inequality (1).

References

R. Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, 2007.

S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

A. G. Dimakis, A. Sarwate, and M. J. Wainwright. Geographic gossip: Efficient averaging for sensor networks. *IEEE Transactions on Signal Processing*, 53:1205–1216, March 2008.

J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57 (3):592–606, 2012.

A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54:48–61, 2009.