JMLR: Workshop and Conference Proceedings vol 23 (2012) 38.14–38.17 25th Annual Conference on Learning Theory

Commentary on "Near-Optimal Algorithms for Online Matrix Prediction"

Rina Foygel RINA@GALTON.UCHICAGO.EDU
Department of Statistics, University of Chicago

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

1. Introduction

This piece is a commentary on the paper by Hazan et al. (2012b). In their paper, they introduce the class of (β, τ)-decomposable matrices, and show that well-known matrix regularizers and matrix classes (e.g. matrices with bounded trace norm) can be viewed as special cases of their construction. The β and τ terms can be related to the max norm and to the trace norm, respectively, as explored in the paper, which we discuss in detail below. The paper's main contribution is a powerful online learning guarantee when learning inside the (β, τ)-decomposable class, which scales with $\sqrt{\beta \cdot \tau}$, and an efficient algorithm for solving this learning problem. Crucially, the paper reframes the well-known problems of online max cut, learning a team ranking ("gambling"), and trace-norm regularized matrix completion (a.k.a. collaborative filtering) as special cases of learning inside (β, τ)-decomposable classes of matrices. This yields new algorithms for the three existing problems, with each algorithm giving a strong improvement over existing results in terms of either efficiency or error rate guarantees. In addition, the paper derives lower bounds on the error rates for each of the three problems that match (up to log factors) the upper bounds proved with the new algorithm—and in particular, for collaborative filtering with the trace norm, their lower bound solves an open problem posed by Shamir and Srebro in COLT 2011.

In this commentary, we explore the connections between the class of (β, τ)-decomposable matrices, introduced by Hazan et al. (2012b), and the matrix trace norm (a.k.a. nuclear norm) and max norm. Specifically, we are interested in the idea of a "trade-off" between the β and τ values for the class, and will consider how the resulting non-convex optimization question can be formulated as a series of convex optimization problems.

2. Connection to max norm and trace norm

Hazan et al. (2012a) show that the max norm and trace norm of $\mathbf{W}$ relate to (β, τ)-decomposability as follows (see their Appendix E):

$$\|\mathbf{W}\|_{\max} = 2 \min \{\beta : \exists \tau, \mathbf{W} \text{ is } (\beta, \tau)\text{-decomposable}\} \text{ and}$$
$$\|\mathbf{W}\|_* = \frac{1}{2} \min \{\tau : \exists \beta, \mathbf{W} \text{ is } (\beta, \tau)\text{-decomposable}\} .$$

These two norms have often been used as regularizers for many problems, including the three specific problems examined in Hazan et al. (2012b)'s paper. In general, for some loss

function Loss($\mathbf{W}$) we would compute

$$\widehat{\mathbf{W}} := \arg\min \{\|\mathbf{W}\| : \text{Loss}(\mathbf{W}) \leq c\} \ ,$$

where $\|\mathbf{W}\|$ is either the max or trace norm, and c is some constraint on the loss. Choosing either the max norm or the trace norm, we can think of this as special cases of regularizing with (β, τ)-decomposability, where we place all emphasis on β or on τ, respectively. For either norm, this is of course a convex optimization problem as long as the loss is convex.

Although the max norm and trace norm relate directly only to β or only to τ, respectively, they can each be used to give a loose bound on the other parameter. This is discussed in the decomposability lemmas of Hazan et al. (2012b) for the three specific problems, but can be summarized in general as follows: for a matrix $\mathbf{W} \in [-1, 1]^{n \times m}$,

$$\|\mathbf{W}\|_{\max} \leq 2\beta \ \Leftrightarrow \ \mathbf{W} \text{ is } (\beta, (n+m)\beta)\text{-decomposable} \ , \tag{1}$$

$$\|\mathbf{W}\|_{*} \leq \frac{1}{2}\tau \ \Leftrightarrow \ \mathbf{W} \text{ is } (\sqrt{n+m}, \tau)\text{-decomposable} \ . \tag{2}$$

For the three problems considered by Hazan et al. (2012b), the way that (β, τ)-decomposability is used in each problem reduces to (1) in the cases of the online max-cut and online gambling problems, and to (2) in the case of the online collaborative filtering problem.

In particular, since Hazan et al. (2012b)'s online learning guarantee scales with $\sqrt{\beta \cdot \tau}$ for the class of (β, τ)-decomposable matrices, we see that their results for the three applications can be derived from the fact that, for any matrix $\mathbf{W} \in [-1, 1]^{n \times m}$, this matrix is (β, τ)-decomposable for some (β, τ) satisfying

$$\sqrt{\beta \cdot \tau} \leq \min \left\{ \frac{1}{2} \|\mathbf{W}\|_{\max} \sqrt{n+m}, \sqrt{2 \|\mathbf{W}\|_{*} \sqrt{n+m}} \right\} \ ,$$

which we obtain simply by combining (1) and (2).

However, this bound might be very loose for other classes of matrices, and we are interested in the possibility of more accurate learning by balancing information from both β and τ, in a way that does not follow trivially from either a max norm bound or a trace norm bound like in (1) and (2).

3. The (β, τ) trade-off

In light of Hazan et al. (2012b)'s work, it can be valuable to consider regularized optimization problems that take both β and τ into account. In particular, since Hazan et al. (2012b)'s learning guarantees are a function of $\sqrt{\beta \cdot \tau}$, we might want to calculate the matrix

$$\widehat{\mathbf{W}} := \arg\min \left\{ \sqrt{\beta \cdot \tau} : \text{Loss}(\mathbf{W}) \leq c \text{ and } \mathbf{W} \text{ is } (\beta, \tau)\text{-decomposable} \right\} \ , \tag{3}$$

where Loss($\mathbf{W}$) is a convex loss function specified by the problem of interest. (For simplicity in this discussion, we consider the batch learning setting rather than the online learning setting.) Optimization with the regularizer $\sqrt{\beta \cdot \tau}$, or more generally with any function that combines information from both β and τ, may be computationally difficult. In particular, we believe that the optimization problem (3) is not convex (although we do not have an example). It is possible, though, to approach this question by considering a small number of convex optimization problems, as we describe in the next section.

4. The (β, τ) Pareto-optimal frontier

Consider the set of (β, τ) pairs that can be obtained under a bound on the loss,

$$\mathcal{L} = \{(\beta, \tau) : \exists \mathbf{W} \text{ s.t. } \mathrm{Loss}(\mathbf{W}) \le c \text{ and } \mathbf{W} \text{ is } (\beta, \tau)\text{-decomposable}\} \ ,$$

and the Pareto-optimal frontier $\mathcal{L}_{\mathrm{par}}$ of this set, i.e. the (β, τ) pairs in $\mathcal{L}$ where β and τ cannot be simultaneously improved (lowered) inside of the set:

$$\mathcal{L}_{\mathrm{par}} = \left\{(\beta, \tau) \in \mathcal{L} : \forall (\beta', \tau') \in \mathcal{L},\ \beta' < \beta \Rightarrow \tau' > \tau \text{ and } \tau' < \tau \Rightarrow \beta' > \beta\right\} \ .$$

We also consider the set of matrices $\mathcal{W}_{\mathrm{par}}$ that attain this Pareto-optimal frontier,

$$\mathcal{W}_{\mathrm{par}} = \{\mathbf{W} : \mathrm{Loss}(\mathbf{W}) \le c, \text{ and } \mathbf{W} \text{ is } (\beta, \tau)\text{-decomposable for some } (\beta, \tau) \in \mathcal{L}_{\mathrm{par}}\} \ .$$

It is clear that if we want to minimize $\sqrt{\beta \cdot \tau}$ as in (3), or more generally any increasing function of β and τ, subject to the constraint on loss, then our solution $\widehat{\mathbf{W}}$ must lie in the set $\mathcal{W}_{\mathrm{par}}$.

This immediately suggests an approach to calculating $\widehat{\mathbf{W}}$. To explain the intuition informally, although $\sqrt{\beta \cdot \tau}$ is not a convex regularizer, β and τ are each convex themselves. Therefore we can compute $\mathcal{W}_{\mathrm{par}}$ by constraining both the loss and the β value while optimizing τ, and allowing the constraint on β to vary (or vice versa). To write this precisely, we use the SDP formulation given in Appendix E of Hazan et al. (2012a): $\mathbf{W}$ is (β, τ)-decomposable if for some matrices $\mathbf{A}$ and $\mathbf{B}$,

$$\max\left\{\max_i \mathbf{A}_{ii}, \max_j \mathbf{B}_{jj}\right\} \le \beta, \ \mathrm{trace}(\mathbf{A}) + \mathrm{trace}(\mathbf{B}) \le \tau, \text{ and } \begin{pmatrix} \mathbf{A} & \mathbf{W} \\ \mathbf{W}^\top & \mathbf{B} \end{pmatrix} \succeq 0 \ .$$

Using this formulation, we see that

$$\mathcal{W}_{\mathrm{par}} = \bigcup_{b \ge 0} \arg\min_{\mathbf{W}} \Bigg\{ \mathrm{trace}(\mathbf{A}) + \mathrm{trace}(\mathbf{B}) : \mathrm{Loss}(\mathbf{W}) \le c, \\ \max\left\{\max_i \mathbf{A}_{ii}, \max_j \mathbf{B}_{jj}\right\} \le b, \ \begin{pmatrix} \mathbf{A} & \mathbf{W} \\ \mathbf{W}^\top & \mathbf{B} \end{pmatrix} \succeq 0 \Bigg\} \ .$$

By taking a sufficiently fine grid of values of the bound b on β, we can then get a reasonable approximation of the set $\mathcal{W}_{\mathrm{par}}$, and choose some approximation to $\widehat{\mathbf{W}}$ from this set.

5. An example of the Pareto-optimal frontier

We would now like to ask, how much is gained by considering the entire Pareto-optimal frontier of solutions $\mathcal{W}_{\mathrm{par}}$, instead of simply considering the "endpoints", that is, the solutions we get by regularizing only the max norm (the β) or the trace norm (the τ) rather than considering β and τ simultaneously:

$$\mathcal{W}_{\mathrm{endpoints}} = \arg\min\{\|\mathbf{W}\|_{\max} : \mathrm{Loss}(\mathbf{W}) \le c\} \cup \arg\min\{\|\mathbf{W}\|_* : \mathrm{Loss}(\mathbf{W}) \le c\} \subsetneq \mathcal{W}_{\mathrm{par}} \ .$$

To help understand the role of the Pareto-optimal frontier, we compute this frontier for a matrix completion problem, where for a partially-observed matrix $\mathbf{Y}$, the loss is given by

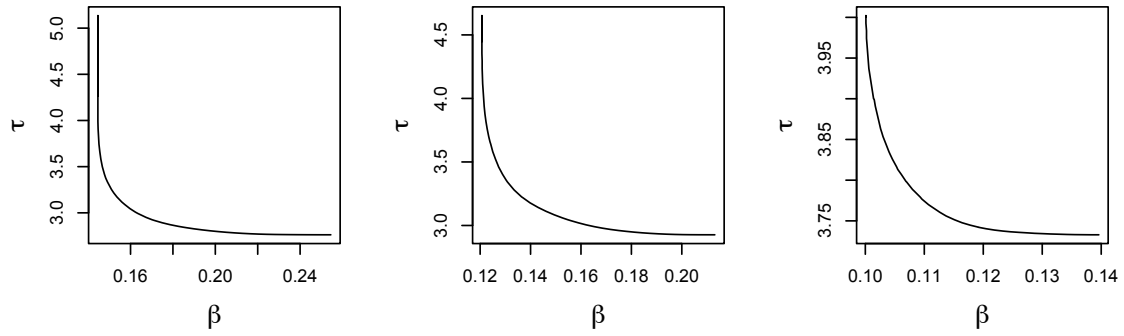

Figure 1: The Pareto-optimal frontier $\mathcal{L}_{\text{par}}$ for the (β, τ) trade-off for matrix reconstruction with a low-rank, approximately low-rank, and full-rank signal (from left to right).

$\sum_{(i,j) \text{ observed}} (\mathbf{Y}_{ij} - \mathbf{W}_{ij})^2$. In Figure 1, we plot $\mathcal{L}_{\text{par}}$ for three matrix completion problems, each with a partially-observed matrix $\mathbf{Y}$ of size 20×20. These matrices consist of a signal plus very low i.i.d. standard normal noise on each entry, with approximately half of the entries observed, where the signal is given by either a rank-3 matrix, an approximately low-rank matrix with singular values $\propto (1, 1/2, 1/3, 1/4, \dots,)$, or a full-rank matrix with all singular values equal. (In each case, the singular vectors are chosen randomly.)

We see that for the low-rank matrix, at each point on the frontier, either the β value is very close to the lowest β value attained for any τ, or the τ value is close to the lowest τ value obtained for any β—that is, the curve lies very close to a union of a horizontal segment and a vertical segment. Intuitively, this suggests that the set $\mathcal{W}_{\text{par}}$ might not be very different from its subset $\mathcal{W}_{\text{endpoints}}$. This tells us that we do not gain much by considering the solutions along this entire frontier, relative to what we would get by considering only the max norm regularized solution and the trace norm regularized solution. For the full-rank matrix, this is not the case at all, while the curve for the approximately low-rank matrix is somewhere between the two. Overall, for data that is not extremely close to low rank, studying the (β, τ) trade-off may be much more informative than simply testing both of the previously studied regularizers.

References

E. Hazan, S. Kale, and S. Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. *Arxiv preprint* `arXiv:1204.0136`, 2012a.

E. Hazan, S. Kale, and S. Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. *25th Annual Conference on Learning Theory (COLT)*, 2012b.