6.883: Online Methods in Machine Learning

Alexander Rakhlin

LECTURE 14

1. COLLABORATIVE FILTERING, FIRST ATTEMPT, CONTINUED

In the previous lecture, we had settled for

$$\phi(x_{1:n}, y_{1:n}) = \min_{\|M\|_* \le b, \|M\|_\infty \le 1} \frac{1}{n} \sum_{t=1}^{n} \frac{1}{2} |M(i_t, j_t) - y_t| + C_n \tag{1}$$

as the benchmark that captures our prior knowledge about the low-rank (low trace-norm) structure of the rating matrix. We have also evaluated C_n, so that we know how fast we approach the performance of the benchmark. What is left is to specify an optimization problem to evaluate ϕ. This amounts to finding

$$\sup_{\|M\|_* \le b, \|M\|_\infty \le 1} \sum_{t=1}^{n} y_t M(i_t, j_t) = \sup_{\|M\|_* \le b, \|M\|_\infty \le 1} Y \bullet M \tag{2}$$

The method is based on semidefinite programming, and students unfamiliar with this approach can skip the next section.

1.1 An algorithm for trace norm benchmark

We first pretend that the order of user-movie pairs is known. Then the algorithm is: draw $\epsilon_{t+1:n}$ and predict according to

$$\widehat{q}_t(x_{1:n}, y_{1:t-1}, \epsilon_{t+1:n}) = n\Big[\phi(x_{1:n}, y_{1:t-1}, -1, \epsilon_{t+1:n}) - \phi(x_{1:n}, y_{1:t-1}, +1, \epsilon_{t+1:n})\Big] \tag{3}$$

This mixed strategy can be evaluated efficiently whenever we can maximize the linear objective in (2) over the set $\{M : \|M\|_* \le b, \|M\|_\infty \le 1\}$. It turns out that

$$\|M\|_* = \min_{U,V} \; a \tag{4}$$

$$\text{s.t.} \quad \begin{bmatrix} U & M \\ M^\intercal & V \end{bmatrix} \succeq 0 \tag{5}$$

$$\text{trace}(U) + \text{trace}(V) = 2a \tag{6}$$

with optimization over matrices U, V. Then the maximization problem in (2) can be written as

$$\min \quad -Y \bullet M \tag{7}$$

$$-1 \le M_{i,j} \le 1 \tag{8}$$

$$\begin{bmatrix} U & M \\ M^\intercal & V \end{bmatrix} \succeq 0 \tag{9}$$

$$\text{trace}(U) + \text{trace}(V) = 2b \tag{10}$$

and the variables are $M \in \mathbb{R}^{k\times p}$, and positive semidefinite $U \in \mathbb{R}^{k\times k}$, $V \in \mathbb{R}^{p\times p}$. Let $X = [U, M; M^\intercal V]$ be the semidefinite matrix being optimized over, and pad Y with zeros in the corresponding blocks to write $Y \bullet M$ as $\tilde{Y} \bullet X$. Then the optimization problem is

$$\begin{aligned} \min \quad & -\tilde{Y} \bullet X && (11)\\ & -1 \le X_{i,j} \le 1, \quad i \in [k+1, \dots, k+p], \quad j = [k] && (12)\\ & X \succeq 0 && (13)\\ & \operatorname{trace}(X) = 2b && (14) \end{aligned}$$

This problem can be solved with some known optimization techniques, and we refer to [AHK05] and Chapter 5 of [GM12].

1.2 Max-norm benchmark

We briefly point out a related optimization problem

$$\begin{aligned} \min \quad & -\tilde{Y} \bullet X && (15)\\ & X \succeq 0 && (16)\\ & X_{i,i} \le b && (17) \end{aligned}$$

with $\tilde{Y} = \frac{1}{2}[\mathbf{0} \;\; Y; Y^\intercal \;\; \mathbf{0}]$. This optimization problem corresponds to the benchmark

$$\sup_{\|M\|_{\max}\le b} Y \bullet M \tag{18}$$

with a bound on the max-norm of M. The max-norm, defined as

$$\|Z\|_{\max} \triangleq \min_{LR^\intercal = Z} \|L\|_{2,\infty} \times \|R\|_{2,\infty} \tag{19}$$

with $L \in \mathbb{R}^{k\times r}$, $R \in \mathbb{R}^{p\times r}$, r unconstrained, and $\|L\|_{2,\infty} = \max_j \|L_j\|_2 = \max_j \left(\sum L_{j,i}^2\right)^{1/2}$, has been a popular substitute for the rank constraint. The max-norm has several nice properties, and admits a relatively efficient algorithm [LRS+10, Jag11, SRJ04].

While there are existing poly-time methods for max-norm-constrained optimization, we remark that the resulting potential function is b/n-smooth, but not $1/n$-smooth. Indeed, the bound on the diagonal of X does ensure boundedness of the off-diagonal entries, but not by 1 as required. This is very annoying: we are severely constrained by the smoothness requirement of Cover. If we think back to where this smoothness of ϕ was used, it was to get a closed-form solution for the next **Rel** function as we worked backwards from $t = n$ to $t = 1$. *Hopefully, there is now enough motivation to develop machinery beyond Cover's result, and we shall do this in the following lectures.*

2. COLLABORATIVE FILTERING, SECOND ATTEMPT

Let us go around the limitation imposed by the smoothness requirement of ϕ. This smoothness requirement arose from having to predict a mixed strategy from which to draw the prediction. Suppose we change the problem as follows. On round t, we predict $\widehat{y}_t \in \mathbb{R}$, the outcome $y_t \in \{\pm 1\}$ is revealed, and the loss our algorithm incurs is $-\widehat{y}_t y_t$, a real value. If we predicted the sign correctly, we have a negative loss, and the more confidence we put

into our prediction, the smaller is the loss. On the other hand, a highly confident wrong prediction will incur a large positive loss.

The proposed loss function is *linear*. Later, we shall develop another set of methods to deal with more natural loss functions (e.g. square, absolute, logistic, etc). For now let us demonstrate that the assumption of smoothness of ϕ is no longer needed.

Let us write down the desired statement. We would like to develop a method such that the average loss is no more than a benchmark:

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} -\widehat{y}_t y_t\right] \le \min_{\|M\|_* \le b}\left\{\frac{1}{n}\sum_{t=1}^{n} -M(i_t, j_t) y_t\right\} + C_n \tag{20}$$

for any $y_1, \dots, y_n$ and any order of presentation of entries.

It turns out that we no longer need the boundedness of entries of M. Also note that the prediction strategy need not be randomized (the counterexample in lecture 1 no longer holds), but we will take advantage of randomization once again by drawing random signs.

Lemma 1. *For the collaborative filtering problem with linear loss, there is a randomized method that attains*

$$\mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} -\widehat{y}_t y_t\right] \le \min_{\|M\|_* \le b}\left\{\frac{1}{n}\sum_{t=1}^{n} -M(i_t, j_t) y_t\right\} + c \cdot \frac{b(\sqrt{k} + \sqrt{p})}{n} \tag{21}$$

for any $y_1, \dots, y_n$ and any order of presentation of entries. The method draws $\epsilon_{t+1:n}$ i.i.d. Rademacher and then computes

$$\widehat{y}_t = \frac{b}{2}\left[\mathbb{E}_{\epsilon_{t+1:n}} \|Y_t^+\| - \mathbb{E}_{\epsilon_{t+1:n}} \|Y_t^-\|\right] \tag{22}$$

where Y_t^+ is a matrix with y_s in corresponding entries (i_s, j_s) for $s < t$, a $+1$ in the entry (i_t, j_t), and $\epsilon_{t+1:n}$ in the remaining entries.

Proof. Let us first do the proof for the case when the order $x_1, \dots, x_n$ of entries in the matrix is known ahead of time. Let

$$\mathbf{Rel}(x_{1:n}, y_{1:n}) = -\min_{\|M\|_* \le b}\left\{\frac{1}{n}\sum_{t=1}^{n} -M(i_t, j_t) y_t\right\}. \tag{23}$$

The optimization problem at the last time point n is

$$\begin{aligned}
&\min_{\widehat{y}_n \in \mathbb{R}} \max_{y_n \in \{\pm 1\}}\left\{-\frac{1}{n}\widehat{y}_n y_n + \mathbf{Rel}(x_{1:n}, y_{1:n})\right\} \\
&= \min_{\widehat{y}_n \in \mathbb{R}} \max\left\{-\frac{1}{n}\widehat{y}_n + \mathbf{Rel}(x_{1:n}, y_{1:n-1}, +1), \frac{1}{n}\widehat{y}_n + \mathbf{Rel}(x_{1:n}, y_{1:n-1}, -1)\right\}.
\end{aligned} \tag{24}$$

The unconstrained minimum is achieved at the intersection of the two lines, as argued before. Hence, the strategy is

$$\widehat{y}_n = \frac{n}{2}\left[\mathbf{Rel}(x_{1:n}, y_{1:n-1}, +1) - \mathbf{Rel}(x_{1:n}, y_{1:n-1}, -1)\right]$$

and (24) is equal to

$$\mathbb{E}_{\epsilon_n} \mathbf{Rel}(x_{1:n}, y_{1:n-1}, \epsilon_n), \tag{25}$$

which we define as $\mathbf{Rel}(x_{1:n-1}, y_{1:n-1})$. Continuing this process,

$$\mathbf{Rel}(x_{1:t}, y_{1:t}) = -\frac{1}{n}\mathbb{E}_{\epsilon_{t+1:n}} \min_{\|M\|_* \le b} \{-M \bullet Y_t\} \tag{26}$$

where Y_t contains $y_{1:t}$ in the positions given by $x_{1:t}$, and the rest of locations $x_{t+1:n}$ are filled with independent Rademacher random variables. Finally,

$$\mathbf{Rel}(\varnothing) = -\frac{1}{n}\mathbb{E}_{\epsilon_{1:n}} \min_{\|M\|_* \le b} \{M \bullet \mathcal{E}\}\,. \tag{27}$$

We now recall that maximum of a linear form over a unit ball is definition of the dual norm, which is the spectral norm (largest singular value) in our case. Hence,

$$-\min_{\|M\|_* \le b} \{-M \bullet Y_t\} = b \max_{\|M\|_* \le 1} \{M \bullet Y_t\} = b\,\|Y_t\| \tag{28}$$

The last norm is the spectral norm, and we leave it unadorned. We conclude that

$$\mathbf{Rel}(x_{1:t}, y_{1:t}) = \frac{b}{n}\mathbb{E}_{\epsilon_{t+1:n}} \|Y_t\|\,. \tag{29}$$

The algorithm now takes on the following form:

$$\widehat{y_t} = \frac{b}{2}\left[\mathbb{E}_{\epsilon_{t+1:n}} \|Y_t^+\| - \mathbb{E}_{\epsilon_{t+1:n}} \|Y_t^-\|\right] \tag{30}$$

where Y_t^+ and Y_t^-, respectively, put +1 and −1 at the position (i_t, j_t), values y_s in previously observed locations (i_s, j_s), and ϵ_s in the unseen locations. This strategy is not randomized, but requires averaging over random signs. Instead, we may draw $\epsilon_{t+1:n}$ and define prediction $\widehat{y_t}$ as

$$\widehat{y_t}(x_{1:n}, y_{1:t-1}, \epsilon_{t+1:n}) = \frac{b}{2}\big[\,\|Y_t^+\| - \|Y_t^-\|\,\big] \tag{31}$$

Finally,

$$\mathbf{Rel}(\varnothing) = \frac{b}{n}\mathbb{E}_{\epsilon_{1:n}} \|\mathcal{E}\| \le c \cdot \frac{b(\sqrt{k} + \sqrt{p})}{n}.$$

Homework: argue that the order of presentation need not be known in advance for the recursion to unfold. □

Two remarks. First, the computation essentially involves computing two spectral norms, which amounts to computing the largest singular values. This can be done by the power method. Second, if our initial target rank is r, then we should be thinking of $b = \min\{k, p\} \times r$. For simplicity, let's take $k = p = \sqrt{n}$. Then the constant C_n is proportional to

$$\frac{rk\sqrt{k}}{k^2} = k^{-1/2},$$

and thus for a large matrix the average error of the algorithm is close to the best model, as given by the benchmark.

References

[AHK05] Sanjeev Arora, Elad Hazan, and Satyen Kale. Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*, pages 339–348. IEEE, 2005.

[GM12] Bernd Gärtner and Jiri Matousek. *Approximation algorithms and semidefinite programming*. Springer Science & Business Media, 2012.

[Jag11] Martin Jaggi. Convex optimization without projection steps. *arXiv preprint arXiv:1108.1170*, 2011.

[LRS+10] Jason D Lee, Ben Recht, Nathan Srebro, Joel Tropp, and Ruslan R Salakhutdinov. Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems*, pages 1297–1305, 2010.

[SRJ04] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2004.