Machine Learning Theory (CS 6783)

1 Matrix Prediction and Collaborative Filtering

Consider the task of predicting if a given user will like of dislike a particular movie. Specifically, we would like to predict if user i would like movie j. As past data, to aid our prediction we have samples of user movie pairs and the corresponding ratings for that pair. If there are M users and M movies over all, we can view the set of user movie pair ratings as entries in a matrix. We would like to predict entries in a matrix. The key modeling assumption we would like to capture is something like, similar users might like similar movies.

This problem we write down in the online framework as follows:

For $t = 1$ to n

Adversary picks user movie pair (i_t, j_t) to predict

Learner predicts rating $\hat{y}_t$

True rating y_t for that user movie pair is revealed.

Learner suffers convex loss $\ell(y_t, \hat{y}_t)$

End For

Now given a bench mark set of matrices $\mathcal{F} \in \mathbb{R}^{M \times N}$ we would like to measure regret against, we can phrase our goal as minimizing regret given by

$$\text{Reg}_n = \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{F \in \mathcal{F}} \sum_{t=1}^{n} \ell(F[i_t, j_t], y_t)$$

if $\mathcal{F}$ is all possible set of matrices, then no meaningful regret bound is possible. A commonly used modeling choice is that $\mathcal{F}$ is a set of matrices with rank at most r, and entries bounded by 1. Such a modeling choice captures the idea that each user can be represented by an r dimensional vector and rating for each movie can be represented by a fixed linear combination of the r co-oridinates for the corresponding user. That is, we use

$$\mathcal{F} = \{F = UV : \text{ where } U \in \mathbb{R}^{M \times r} \text{ and } V = \mathbb{R}^{r \times N}, F \in [-1, 1]^{M \times N}\}$$

It turns out that the benchmark $\mathcal{F}$ is a computationally infeasible benchmark. However one can instead use the larger set

$$\bar{\mathcal{F}} = \{F : \sum_{i} |\lambda_i(F)| \leq d \times \sqrt{r}\}$$

where we will use $d = M+N$. That is, $\bar{\mathcal{F}} \supset \mathcal{F}$ is the set of all matrices with trace norm bounded by $d\sqrt{r}$. Whys is this? Well note that if we have a rank r matrix F, then since only r of the singular values are non-zero, we have that:

$$\sum_i |\lambda_i(F)| \leq \sqrt{r}\sqrt{\max_i |\lambda_i(F)|^2} = \sqrt{r}\|F\|_{\text{Fr}} \leq \sqrt{r}\sqrt{M \times N} \leq \sqrt{r}(M+N)$$

Hence, $\sum_i |\lambda_i(F)| \leq d \times \sqrt{r}$ and so $\bar{\mathcal{F}} \supset \mathcal{F}$. Hence obtaining diminishing regret against $\bar{\mathcal{F}}$ yields regret bound against $\mathcal{F}$ as well.

2 First Cut: Lets Try and Use Mirror Descent

One can formulate the above problem as a online convex optimization problem. Especially one with linear predictors. To see this, let us encode user movie entry to predict (i_t, j_t) by the indicator matrix $X_t = e_{i_t} e_{j_t}^\top$. That is, a matrix with one in the entry (i_t, j_t) and 0 elsewhere. In this case, we can write regret against $\bar{\mathcal{F}}$ as:

$$\text{Reg}_n = \sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{F \in \bar{\mathcal{F}}} \sum_{t=1}^n \ell(\langle F, X_t \rangle, y_t)$$

where $\langle F, X_t \rangle$ is the generalized inner product and this instance gives $F[i_t, j_t]$. Now since this is a convex loss with linear predictor, we can try and use mirror descent type algorithms. In this case, note first that $\|\nabla_t\|^* \propto \|X_t\|^*$ (in fact for absolute loss its equality). Now irrespective of which norm we pick, $\|X_t\|^* = 1$. Hence we get a bound of form,

$$\text{Reg}_n \leq O\left(\sqrt{\frac{\sup_{F \in \bar{\mathcal{F}}} D_R(F|\hat{y}_1)}{n}}\right)$$

But notice that by strong convexity w.r.t. whatever norm we pick, $D_R(F|\hat{y}_1) \geq \|F - \hat{y}_1\|^2 \approx \|F\|^2$. At this point, note that since $\bar{\mathcal{F}}$ is all trace norm bounded matrices, the only real option for the norm we can pick in mirror descent to be strong convex w.r.t. has to be the trace norm and we know that $\sup_{F \in \bar{\mathcal{F}}} \|F\|_{\text{tr}}^2 = d^2 r$. Hence. the mirror descent bound can never be better than order

$$\sqrt{\frac{rd^2}{n}}$$

But this means that we need n number of samples to be larger than number of entries in the entire matrix. But at this point the prediction problem is mute.

Can we even hope to do better?

In general, it turns out that this is the worst case regret rate when competing with $\bar{\mathcal{F}}$.

3 Burkholder Method

It turns out however that one can go for an adaptive bound that can yield improvement. If for instance, user movie pairs are picked uniformly at random, once can improve the bound on regret

for instance. In fact what we will show is an adaptive bound that is same in the worst case as the mirror descent one but can be much smaller when empirical distributions of user movie pairs are not very "peaky". Specifically, in the Burkholder termminology, we will go for the bound

$$\phi(X_1, y_1, \ldots, X_n, y_n) = \inf_{F \in \bar{\mathcal{F}}} \sum_{t=1}^{n} \ell(\langle F, X_t \rangle, y_t) + O\left(R \sqrt{\max\left\{ \left\| \sum_{t=1}^{n} X_t X_t^\top \right\|_\sigma, \left\| \sum_{t=1}^{n} X_t^\top X_t \right\|_\sigma \right\} \log(M+N)} \right)$$

where in the above $\|\cdot\|_\sigma$ is the spectral norm (magnitude of largest eigenvalue) and R is bound on trace norm which in our case is $d\sqrt{r}$. Again as we have been doing in previous lectures (which we fix via doubling trick), we go for the following edited ϕ^η (assuming loss is 1-Lipschitz):

$$\phi^\eta(X_1, y_1, \ldots, X_n, y_n) = \inf_{F \in \bar{\mathcal{F}}} \sum_{t=1}^{n} \ell(\langle F, X_t \rangle, y_t) + \frac{\eta}{2} R \max\left\{ \left\| \sum_{t=1}^{n} X_t X_t^\top \right\|, \left\| \sum_{t=1}^{n} X_t^\top X_t \right\| \right\} + \frac{R \log(M+N)}{\eta}$$

We now give an algorithm via the Burkholder method. To do this, let us first introduce the so called Hermitian dilation of a matrix given by

$$H(X) = \begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix}$$

The key thing about the hermitian dilation is that it makes the $M \times N$ matrix into a square matrix of size $M+N \times M+N$. Further, the eigen values of this matrix are $\pm\lambda(X)$, that is plus and minus the singular values of matrix X.

Lemma 1. *The mapping* $\mathbf{T}(X, \alpha) = (\alpha H(X), H(X)^2)$ *along with*

$$U(H, M) = \frac{R}{\eta} \log \operatorname{tr} \exp\left(\eta H - \frac{1}{2}\eta^2 M \right) - \frac{R \log(M+N)}{\eta}$$

is both a valid sufficient statistic for ϕ^η *described above and is a Burkholder mapping.*

Proof. First, note that $U(0) \leq 0$ because:

$$\begin{aligned} U(0) &= \frac{R}{\eta} \log \operatorname{tr} \exp\left(\eta 0 - \frac{1}{2}\eta^2 0 \right) - \frac{R \log(M+N)}{\eta} \\ &= \frac{R}{\eta} \log \operatorname{tr} \exp(0) - \frac{R \log(M+N)}{\eta} \\ &= \frac{R}{\eta} \log \operatorname{tr} I_{M+N \times M+N} - \frac{R \log(M+N)}{\eta} = 0 \\ &= \frac{R}{\eta} \log(M+N) - \frac{R \log(M+N)}{\eta} = 0 \end{aligned}$$

Next, to show that U is a valid sufficient statistic, note that first of all,

$$\max\left\{ \left\| \sum_{t=1}^{n} X_t X_t^\top \right\|_\sigma, \left\| \sum_{t=1}^{n} X_t^\top X_t \right\|_\sigma \right\} = \left\| \sum_{t=1}^{n} H(X_t)^2 \right\|_\sigma$$

and so,

$$
\begin{aligned}
\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{F\in\bar{\mathcal{F}}} \sum_{t=1}^n \ell(\langle F, X_t\rangle, y_t) - \frac{\eta}{2} R \max\left\{ \left\| \sum_{t=1}^n X_t X_t^\top \right\|_\sigma, \left\| \sum_{t=1}^n X_t^\top X_t \right\|_\sigma \right\} - \frac{R\log(M+N)}{\eta} \\
= \sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{F\in\bar{\mathcal{F}}} \sum_{t=1}^n \ell(\langle F, X_t\rangle, y_t) - \frac{\eta}{2} R \left\| \sum_{t=1}^n H(X_t)^2 \right\|_\sigma - \frac{R\log(M+N)}{\eta} \\
\le \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)\hat{y}_t - \inf_{F\in\bar{\mathcal{F}}} \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)\langle F, X_t\rangle - \frac{\eta}{2} R \left\| \sum_{t=1}^n H(X_t)^2 \right\|_\sigma - \frac{R\log(M+N)}{\eta} \\
= \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)\hat{y}_t + \sup_{F:\|F\|_{\mathrm{tr}}\le R} \left\langle F, \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) X_t \right\rangle - \frac{\eta}{2} R \left\| \sum_{t=1}^n H(X_t)^2 \right\|_\sigma - \frac{R\log(M+N)}{\eta} \\
= \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)\hat{y}_t + R \left\| \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) X_t \right\|_\sigma - \frac{\eta}{2} R \left\| \sum_{t=1}^n H(X_t)^2 \right\|_\sigma - \frac{R\log(M+N)}{\eta} \\
= \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)\hat{y}_t + R \left\| \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) H(X_t) \right\|_\sigma - \frac{\eta}{2} R \left\| \sum_{t=1}^n H(X_t)^2 \right\|_\sigma - \frac{R\log(M+N)}{\eta} \\
= \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)\hat{y}_t + R\lambda_1 \left(\sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) H(X_t) \right) - \frac{\eta}{2} R\lambda_1 \left(\sum_{t=1}^n H(X_t)^2 \right) - \frac{R\log(M+N)}{\eta} \\
\le \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)\hat{y}_t + R\lambda_1 \left(\sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) H(X_t) - \frac{\eta}{2} R \sum_{t=1}^n H(X_t)^2 \right) - \frac{R\log(M+N)}{\eta} \\
\le \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)\hat{y}_t + R\lambda_1 \left(\sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) H(X_t) - \frac{\eta}{2} R \sum_{t=1}^n H(X_t)^2 \right) - \frac{R\log(M+N)}{\eta} \\
\le \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)\hat{y}_t + \frac{R}{\eta} \log\left(\sum_{i=1}^{M+N} \exp\left(\eta \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) H(X_t) - \frac{\eta^2}{2} R \sum_{t=1}^n H(X_t)^2 \right) \right) - \frac{R\log(M+N)}{\eta} \\
= \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)\hat{y}_t + U(\tau_n)
\end{aligned}
$$

Finally to prove the restricted concavity condition more generally we use the fact that $\alpha \mapsto U(\tau + \mathbf{T}(x,\alpha))$ is convex in α and hence we only need to prove the condition over uniform distribution over $\{\pm 1\}$ for α. But instead of trying to prove this, let us just assume that the loss is the absolute loss and so $\alpha \in \{\pm 1\}$. Hence the only 0 mean distribution is in fact $\alpha = \epsilon$ which is a Rademacher random variable. Now to prove the restricted concavity assumption, note that:

$$
\begin{aligned}
\mathbb{E}_\epsilon U((H,M) + \mathbf{T}(x,\epsilon)) &= \mathbb{E}_\epsilon U(H + \epsilon H(X), M + H(X)^2) \\
&= \frac{R}{\eta} \mathbb{E}_\epsilon \log \mathrm{tr} \exp\left(\eta H + \eta\epsilon H(X) - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2 \right) - \frac{R\log(M+N)}{\eta}
\end{aligned}
$$

Using concavity of (scalar) logarithm function:

$$\leq \frac{R}{\eta} \log \mathbb{E}_\epsilon \mathrm{tr} \exp\left(\eta H + \eta\epsilon H(X) - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2\right) - \frac{R\log(M+N)}{\eta}$$

Think of how you would have proceeded in the scalar case. We would have gone with saying $\exp\left(\eta H + \eta\epsilon H(X) - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2\right) = \exp\left(\eta H - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2\right) \times \exp(\eta\epsilon H(X))$ and then pushed expectation inside trace and bounded $E_\epsilon \exp(\eta\epsilon H(X))$. This inequality is false for the matrix case!

Here is the proof for the matrix case:

$$\begin{aligned}
\mathbb{E}_\epsilon U((H,M) + \mathbf{T}(x,\epsilon)) &\leq \frac{R}{\eta} \log \mathbb{E}_\epsilon \mathrm{tr} \exp\left(\eta H + \eta\epsilon H(X) - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2\right) - \frac{R\log(M+N)}{\eta} \\
&= \frac{R}{\eta} \log \mathbb{E}_\epsilon \mathrm{tr} \exp\left(\eta H + \log\left(\exp(\eta\epsilon H(X))\right) - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2\right) - \frac{R\log(M+N)}{\eta}
\end{aligned}$$

Now we use (without proof) a theorem from matrix analysis called Leib's Concavity theorem which states that for any square matrix B and any positive definite matrix A, the mapping $A \mapsto \mathrm{tr}\exp(B + \log A)$ is concave. Now in the above use $B = \eta H - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2$ which is a square matrix since we only ever use sums of hermitian dilations. and use $A = \exp(\eta\epsilon H(X))$ which is positive definite since its exponential of a matrix. Hence using concavity to push expectation inside we get,

$$\begin{aligned}
\mathbb{E}_\epsilon U((H,M) + \mathbf{T}(x,\epsilon)) &\leq \frac{R}{\eta} \log \mathbb{E}_\epsilon \mathrm{tr} \exp\left(\eta H + \log\left(\exp(\eta\epsilon H(X))\right) - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2\right) - \frac{R\log(M+N)}{\eta} \\
&\leq \frac{R}{\eta} \log \mathrm{tr} \exp\left(\eta H + \log\left(\mathbb{E}_\epsilon \exp(\eta\epsilon H(X))\right) - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2\right) - \frac{R\log(M+N)}{\eta}
\end{aligned}$$

Now note that $\mathbb{E}_\epsilon \exp(\eta\epsilon H(X)) \preceq \exp(\eta^2 H(X)^2/2)$. Further, $\log(X)$ the matrix logarithm is what is called operator monotone, meaning that if $X \preceq Y$ then $\log(A) \preceq \log(B)$. Hence we have that $\log(\mathbb{E}_\epsilon \exp(\eta\epsilon H(X))) \preceq \log(\exp(\eta^2 H(X)^2/2)) = \eta^2 H(X)^2/2$. Finally, the scalar function $\mathrm{tr}\exp(X)$ of a matrix is monotone and so

$$\mathrm{tr}\exp\left(\eta H + \log\left(\mathbb{E}_\epsilon \exp(\eta\epsilon H(X))\right) - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2\right) \leq \mathrm{tr}\exp\left(\eta H + \frac{\eta^2}{2} H(X))^2 - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2\right)$$

Putting all this together, we conclude that,

$$\begin{aligned}
\mathbb{E}_\epsilon U((H,M) + \mathbf{T}(x,\epsilon)) &\leq \frac{R}{\eta} \log \mathrm{tr} \exp\left(\eta H + \log\left(\mathbb{E}_\epsilon \exp(\eta\epsilon H(X))\right) - \frac{1}{2}\eta^2 M - \eta^2 H(X)^2\right) - \frac{R\log(M+N)}{\eta} \\
&\leq \frac{R}{\eta} \log \mathrm{tr} \exp\left(\eta H + -\frac{1}{2}\eta^2 M\right) - \frac{R\log(M+N)}{\eta} \\
&= U(H,M)
\end{aligned}$$

This proves the restricted concavity of U. But be vary that while some of the inequalities look exactly like the scalar ones, the results of leib's concavity and operator monotonicity of $\log(\cdot)$ are highly non-trivial results unlike the scalar case.

□

Algorithm: As previously mentioned, the the mapping $\alpha \mapsto U(\tau + \mathbf{T}(x, \alpha))$ is convex and so the algorithm at round t is simply to predict

$$\begin{aligned}
\hat{y}_t &= \frac{1}{2}\left(U(\tau_{t-1} + \mathbf{T}(X, -1)) - U(\tau_{t-1} + \mathbf{T}(X, -1))\right) \\
&= \frac{R}{2\eta}\left(\log \operatorname{tr} \exp\left(\eta \sum_{j=1}^{t-1} H(X_j) - \eta H(X) - \frac{1}{2}\eta^2 \sum_{j=1}^{t-1} H(X_j)^2 - \eta^2 H(X_t)^2\right)\right. \\
&\qquad \left. - \log \operatorname{tr} \exp\left(\eta \sum_{j=1}^{t-1} H(X_j) + \eta H(X) - \frac{1}{2}\eta^2 \sum_{j=1}^{t-1} H(X_j)^2 - \eta^2 H(X_t)^2\right)\right)
\end{aligned}$$