IEOR 265 – Lecture 8
Matrix Completion

1 Model and NP-Hard Formulation

We will use the notation $[r]$ to refer to the set $[r] = \{1, \ldots, r\}$. Consider a matrix $\psi \in \mathbb{R}^{r_1 \times r_2}$, and suppose some small subset of entries are measured. In particular, suppose our measurements consist of index-pairs $x\langle i \rangle \in [r_1] \times [r_2]$ for $i = 1, \ldots, n$ along with noisy measurements of the corresponding matrix entries $y\langle i \rangle = \psi_{x\langle i \rangle} + \epsilon_i$, where ϵ_i is bounded zero-mean noise.

The problem of matrix completion is to estimate the matrix ψ given the measurements $(x\langle i \rangle, y\langle i \rangle)$ for $i = 1, \ldots, n$. This problem is ill-posed when $n \ll r_1 \cdot r_2$, because a matrix $\psi \in \mathbb{R}^{r_1 \times r_2}$ has $r_1 \cdot r_2$ degrees of freedom. Without imposing additional structure, the problem cannot be solved. One reasonable structure that leads to an identifiable problem is to assume that ψ is low rank. For instance, if $\text{rank}(\psi) = k$, then ψ has $k \cdot (r_1 + r_2)$ degrees of freedom. (To intuitively see why, note that this is the degrees of freedom of the singular value decomposition of a rank-k matrix.)

Given a matrix completion problem with a rank-k constraint, we can pose the estimation problem as the following M-estimator:

$$\hat{\psi} = \min_{\psi} \Big\{ \sum_{i=1}^{n} (y\langle i \rangle - \psi_{x\langle i \rangle})^2 \;\Big|\; \text{rank}(\psi) \leq k \Big\}.$$

Unfortunately, including a rank constraint makes this an NP-hard optimization problem. The question we will examine is how to solve this problem using more tractable approaches.

2 Nuclear Norm Approach

Recall that if $\text{rank}(\psi) \leq k$, then $\sigma(\psi)_j = 0$ (i.e., the j-th singular value of ψ is zero) for $j > k$. If we collect the singular values into a vector $\sigma(\psi)$, then $\text{rank}(\psi) \leq k$ is equivalent to a constraint that $\|\sigma(\psi)\|_0 \leq k$. As in the case of Lasso regression, we can relax this using the ℓ_1-norm. Thus, one possible relaxation for a rank constraint is $\|\sigma(\psi)\|_1 \leq k \cdot \mu$, where $\mu = \max_j \sigma(\psi)_j$.

By definition, $\|\sigma(\psi)\|_1$ is known as the *nuclear norm* of the matrix ψ. The notation $\|\psi\|_* = \|\sigma(\psi)\|_1$ is often used to denote the nuclear norm. This means that that constraint $\|\sigma(\psi)\|_1 \leq k$ must be convex, since norms are always convex functions. However, the key question is whether the convex function $\|\psi\|_*$ is computable in polynomial time. Interestingly, there are many functions (including norms) that are convex but NP-hard to compute. Fortunately, the nuclear norm of a matrix can be computed in polynomial time. In particular, the nuclear norm of

ψ is given by the solution of a semidefinite program (SDP) with $r_1 \cdot r_2 + r_1 \cdot (r_1+1)/2 + r_2 \cdot (r_2+1)/2$ variables:

$$\begin{aligned} \|\psi\|_* = \min\ & \mathsf{trace}(U) + \mathsf{trace}(V) \\ \text{s.t.}\ & \begin{bmatrix} U & \psi \\ \psi' & V \end{bmatrix} \geq 0. \end{aligned}$$

As a result, we can formulate a convex relaxation of the rank-constrained matrix completion problem as the following SDP:

$$\begin{aligned} \hat{\psi} = \min\ & \sum_{i=1}^{n} (y\langle i \rangle - \psi_{x\langle i \rangle})^2 \\ \text{s.t.}\ & \mathsf{trace}(U) + \mathsf{trace}(V) \leq \lambda \\ & \begin{bmatrix} U & \psi \\ \psi' & V \end{bmatrix} \geq 0. \end{aligned}$$

This problem is often represented in more compact notation as:

$$\hat{\psi} = \min_{\psi} \Big\{ \sum_{i=1}^{n} (y\langle i \rangle - \psi_{x\langle i \rangle})^2 \ \Big|\ \|\psi\|_* \leq \lambda \Big\}.$$

This problem is also written in the following form:

$$\hat{\psi} = \min_{\psi} \sum_{i=1}^{n} (y\langle i \rangle - \psi_{x\langle i \rangle})^2 + \lambda \cdot \|\psi\|_*.$$

Unfortunately, the M^*-bound is not applicable to this setting. And so to show high-dimensional consistency, we require a different technique. One approach is to use an alternative (but related) complexity measure.

3 Rademacher Complexity

Let ϵ_i be iid Rademacher random variables (i.e., ϵ_i has the distribution such that $\mathbb{P}(\pm\epsilon_i) = \frac{1}{2}$), and suppose $x_i \in \mathbb{R}^d$ are iid. Then the Rademacher complexity of a set of functions $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R}\}$ is defined as

$$\mathsf{R}_n(\mathcal{F}) = \mathbb{E}\Big(\sup_{f \in \mathcal{F}} \Big| \frac{2}{n} \sum_{i=1}^{n} \epsilon_i \cdot f(x_i) \Big| \Big)$$

3.1 Gaussian Complexity

There is a similar complexity notion known as Gaussian compelxity, which is defined as

$$\mathsf{G}_n(\mathcal{F}) = \mathbb{E}\Big(\sup_{f \in \mathcal{F}} \Big| \frac{2}{n} \sum_{i=1}^{n} g_i \cdot f(x_i) \Big| \Big),$$

where $g_i \sim \mathcal{N}(0,1)$ are iid. The notions of Rademacher and Gaussian complexity are equivalent in the sense that

$$c\mathsf{R}_n(\mathcal{F}) \leq \mathsf{G}_n(\mathcal{F}) \leq C \log\big(n\mathsf{R}_n(\mathcal{F})\big).$$

3.2 Three Useful Results

We need three results before we are able to use Rademacher complexity to define risk bounds. The first result is that (i) if U is a symmetric random variable (e.g., its pdf is even $f_U(-u) = f_U(u)$), and (ii) ϵ is a Rademacher random variable (i.e., $\mathbb{P}(\epsilon = \pm 1) = \frac{1}{2}$) that is independent of U; then ϵU has the same distribution as U. (To see why this is the case, note that if we condition $\epsilon \cdot U$ on the two possible values $\epsilon \pm 1$, then the conditional distribution is the same as the distribution of U because of the symmetry of U.)

The second result is *McDiarmid's Inquality*: Suppose $x_1, \ldots, x_n \in \mathcal{X}$ are iid random variabes belonging to set $\mathcal{X}$. If $f : \mathcal{X}^n \to \mathbb{R}$ satisfies

$$\sup_{u_1,\ldots,u_n,\tilde{u}_j \in \mathcal{X}} \Big| f(u_1, \ldots, u_n) - f(u_1, \ldots, u_{j-1}, \tilde{u}_j, u_{j+1}, \ldots, u_n) \Big| \leq c_i,$$

for all $i = 1, \ldots, n$, then for every $t > 0$ we have

$$\mathbb{P}\Big(f(x_1, \ldots, x_n) - \mathbb{E}(f(x_1, \ldots, x_n)) \geq t \Big) \leq \exp(-2t^2 / \sum_{i=1}^{n} c_i^2).$$

The third useful result is that if $\varphi(\cdot)$ is a Lipschitz continuous function (with Lipschitz constant L) with $\varphi(0) = 0$, then we have

$$\mathbb{E} \sup_{t \in T} \Big| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot \psi(t_i) \Big| \leq 2L \cdot \mathbb{E} \sup_{t \in T} \Big| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \cdot t_i \Big|.$$

If $L = 1$, then the function $\varphi(\cdot)$ is also known as a *contraction*, and so this result is sometimes known as a contraction comparison theorem.

3.3 Risk Bounds

Though an M-estimator of the form

$$\hat{\theta} = \arg\min_{\theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i; \theta))^2 \;\Big|\; \theta \in \Theta \Big\}.$$

is typically understood as the minimizer of the above defined optimization problem, it can also be interpreted as finding a function $f \in \mathcal{F}$ within the function class $\mathcal{F} = \{g(x, \theta) : \theta \in \Theta\}$:

$$\hat{f}(x) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

An alternative interpretation of this estimator is that it is trying to minimize $\mathbb{E}((y - f(x))^2)$, and is using the sample average as an approximation to this expectation.

The sample average approximation is interesting. The quantity $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$ will converge to $\mathbb{E}((y - f(x))^2)$; however, because we are picking the f by minimizing over $\mathcal{F}$, the discrepency between $\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$ and $\mathbb{E}((y - f(x))^2)$ can be very large. The intuition is that we are not just looking for the discrepency between these two quantities for a single f, but instead we are looking for the discrepency between these two quantities over the entire set $\mathcal{F}$.

One of the strengths of Rademacher complexity is that it can be used to quantify the discrepency between these two quantities, which allows us to make conclusions the quality of the estimator. In particular, we are interested in bounding

$$\mathbb{E}((y - \hat{f}(x))^2) - \inf_{f \in \mathcal{F}} \mathbb{E}((y - f(x))^2.$$

The quantity $\mathbb{E}((y - \hat{f}(x))^2)$ is the true error of the estimate $\hat{f}(\cdot)$, and the quantity $\inf_{f\in\mathcal{F}} \mathbb{E}((y - f(x))^2$ is the true error of the best possible estimate taken from $\mathcal{F}$.

Suppose x_i, y_i are bounded random variables. The approach will be to first bound

$$\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 - \mathbb{E}((y - f(x))^2)\Big|.$$

We will achieve this using a symmetrization argument. Specifically, we have

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 - \mathbb{E}((y - f(x))^2)\Big| &\leq \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\Big((y_i - f(x_i))^2 - (\tilde{y}_i - f(\tilde{x}_i))^2\Big)\Big| \\
&\leq \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot \Big((y_i - f(x_i))^2 - (\tilde{y}_i - f(\tilde{x}_i))^2\Big)\Big| \\
&\leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot \Big((y_i - f(x_i))^2\Big)\Big|,
\end{aligned}$$

where (i) the first line follows by Jensen's inequality, (ii) the second line follows by the first useful result from above, and (iii) the third line follows from the triangle inequality.

Since x_i, y_i are assumed to be bounded random variables, the functions $h(u) = (y_i - u)^2 - y_i^2$ are Lipschitz continuous with some constant L and satisfy $h(0) = 0$. And so using the third useful result gives

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 - \mathbb{E}((y - f(x))^2)\Big| &\leq 4L \cdot \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot f(x_i)\Big| + \frac{C}{n} \\
&\leq 2\mathsf{R}_n(\mathcal{F}) + \frac{C}{n}.
\end{aligned}$$

where C is a constant. A tighter bound is possible, but requires a more careful argument. (Essentially, the $\frac{C}{n}$ term can be removed.)

Finally, we return to the quantity $\mathbb{E}((y - \hat{f}(x))^2) - \inf_{f \in \mathcal{F}} \mathbb{E}((y - f(x))^2$. Applying the triangle inequality gives

$$\mathbb{E}((y - \hat{f}(x))^2) - \inf_{f \in \mathcal{F}} \mathbb{E}((y - f(x))^2 \leq$$
$$\mathbb{E}\Big|\mathbb{E}((y - \hat{f}(x))^2) - \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2\Big| + \mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 - \inf_{f \in \mathcal{F}} \mathbb{E}((y - f(x))^2\Big|.$$

To bound the first term, we use the above Rademacher bound:

$$\mathbb{E}\Big|\mathbb{E}((y - \hat{f}(x))^2) - \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2\Big| \leq 2\mathsf{R}_n(\mathcal{F}) + \frac{C}{n}.$$

To bound the second term, note that since $\hat{f}$ and f^* are minimizers of their objective functions we must have

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(y_i - f^*(x_i))^2$$
$$\mathbb{E}((y - f^*(x))^2 \leq \mathbb{E}((y - \hat{f}(x))^2,$$

where $f^* = \arg\inf_{f \in \mathcal{F}} \mathbb{E}((y - f(x))^2$. These two inequalities imply that

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 - \mathbb{E}((y - f^*(x))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(y_i - f^*(x_i))^2 - \mathbb{E}((y - f^*(x))^2$$
$$\mathbb{E}((y - f^*(x))^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \leq \mathbb{E}((y - \hat{f}(x))^2 - \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

Thus, we must have that

$$\mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 - \inf_{f \in \mathcal{F}} \mathbb{E}((y - f(x))^2\Big| \leq 2\mathsf{R}_n(\mathcal{F}) + \frac{C}{n}.$$

Combining everything, we have

$$\mathbb{E}((y - \hat{f}(x))^2) - \inf_{f \in \mathcal{F}} \mathbb{E}((y - f(x))^2 \leq 4\mathsf{R}_n(\mathcal{F}) + \frac{C}{n}.$$

Or rewritten, this gives

$$\mathbb{E}((y - \hat{f}(x))^2) \leq \inf_{f \in \mathcal{F}} \mathbb{E}((y - f(x))^2 + 4\mathsf{R}_n(\mathcal{F}) + \frac{C}{n}.$$

A similar result that holds with high probability can be shown by using McDiarmid's inequality.

3.4 Nuclear Norm Approach Revisted

It turns out that the Rademacher complexity of the function class $\mathcal{F} = \{\psi_x : \|\psi\|_* \leq \mu\}$ is

$$\mathsf{R}(\mathcal{F}) = C\mu\sqrt{\frac{(r_1 + r_2)\log^{3/2}(\max(r_1, r_2))}{nr_1r_2}}.$$

This bound can be found in the paper:

N. Srebro and A. Shraibman (2005) "Rank, Trace-Norm and Max-Norm", in *Learning Theory*. Springer, pp. 545–560.

Thus, for the nuclear norm approach to matrix completion we have

$$\mathbb{E}((y - \hat{\psi}_x)^2) \leq \sigma^2 + C\mu\sqrt{\frac{(r_1 + r_2)\log^{3/2}(\max(r_1, r_2))}{nr_1r_2}} + \frac{C}{n},$$

where $\sigma^2 = \mathsf{var}(\epsilon)$.

4 Alternating Least Squares Approach

One major issue with the nuclear norm approach to matrix completion is that it is computationally difficult. Even though it can be solved in polynomial time, it requires solving a large SDP. And large SDP's are often difficult to solve on computers. Given these difficulties, a heuristic is often used that can provide good solutions but at the cost of lacking theoretical guarantees except in specific cases.

Alternating least squares (ALS) is a popular heuristic. This approach begins with the observation that we can represent a rank-k matrix $\psi \in \mathbb{R}^{r_1 \times r_2}$ as

$$\psi = \sum_{j=1}^{k} u^j v_{j'},$$

where $u^j \in \mathbb{R}^{r_1}$ and $v^j \in \mathbb{R}^{r_2}$. And the main idea of this approach is that if we look at the matrix completion optimization problem:

$$\hat{\psi} = \min_{\psi} \left\{ \sum_{i=1}^{n} (y\langle i\rangle - \psi_{x\langle i\rangle})^2 \;\middle|\; \mathsf{rank}(\psi) \leq k \right\} = \min_{\psi} \sum_{i=1}^{n} (y\langle i\rangle - \sum_{j=1}^{k} u^j_{x\langle i\rangle_1} v^{j'}_{x\langle i\rangle_2})^2,$$

then this is simply a least squares problem when either the u^j (as a group) or the v^j (as a group) are fixed. So the ALS algorithm is to fix u^j, solve for v^j, fix v^j at the minimizer of the previous optimization problem, solve for u^j, fix u^j at the minimizer of the previous optimization problem, etc. until the objective function stops decreasing. The benefit of this heuristic is that solving large-scale least squares problems is much more tractable on a computer than solving large-scale SDP's.