ORIE 7790 High Dimensional Probability and Statistics Lecture 19 - 04/16/2020

Lecture 19: Applications of Local Fano's Technique

Lecturer: Yudong Chen *Scribe: Lucy Huo*

In last lecture, we developed the following "local" Fano's method (Corollary 1 therein) for lower boundign the minimax risk of an estimation problem:

$$\begin{aligned}
\inf_{\psi} \sup \mathbb{E}[\rho(\hat{\theta}, \theta)] &\geq \delta \cdot \inf_{\psi} \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}(\psi(X) \neq j \mathbb{P}_{\theta_j}) && \text{reduce estimation to testing} \\
&\geq \delta \cdot \inf_{\psi} \left\{ 1 - \frac{I(X;J) + \log 2}{\log M} \right\} && \text{Fano's Inequality} \\
&\geq \delta \cdot \inf_{\psi} \left\{ 1 - \frac{\frac{1}{M^2} \sum_{i,j} D(\mathbb{P}_{\theta_i} || \mathbb{P}_{\theta_j})}{\log M} \right\} && \text{Local Fano's} \qquad (1)
\end{aligned}$$

Here $\{\theta_j, j = 1, \ldots . M\}$ is a 2δ packing of the parameter space Θ, whose pairwise KL divergences $\mathbb{P}_{\theta_i} || \mathbb{P}_{\theta_j})$ can be uniformly upper bounded (hence a local packing of Θ).

In this lecture, we will discuss three applications of the local Fano's method in statistical problems.

1 Low-dimensional Linear Regression

Here, we consider the following regression problem set-up,

$$y = X\theta + e, \quad X \in \mathbb{R}^{n \times d}, \quad e \sim N(0, \sigma^2 I), \quad \theta \in \Theta,$$

where $X \in \mathbb{R}^{n \times d}$ is a *fixed* covariate, d the number of parameters, n the sample size, and e the additive noise. We focus on the low dimensional setting where $n > d$. Note that $\mathbb{P}_\theta = N(X\theta, \sigma^2 I)$.

To apply the local Fano's method, we consider the subset[1]

$$\Theta_0 := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 4\delta\sqrt{n}\}.$$

Let $\{\theta_1, \ldots, \theta_M\}$ be a $2\delta\sqrt{n}$-packing of Θ_0 w.r.t. the metric $\rho = \|\cdot\|_2$. We know that the packing number satisfies$M \equiv M(2\delta\sqrt{n}, \Theta_0, \ell_2) = M(\frac{1}{2}, \mathbb{B}_2^d, \ell_2) \geq 2^d$. Moreover, the pairwise KL can be bounded as

$$\begin{aligned}
D(\mathbb{P}_{\theta_i} \| \mathbb{P}_{\theta_j}) &= D\Big(N(X\theta_i, \sigma^2 I) \| N(X\theta_j, \sigma^2 I)\Big) \\
&= \frac{1}{2\sigma^2} \|X\theta_i - X\theta_j\|_2^2 \\
&\leq \frac{1}{2\sigma^2} \|X\|_{\text{op}}^2 \|\theta_i - \theta_j\|_2^2 \\
&\leq \frac{1}{2\sigma^2} \|X\|_{\text{op}}^2 (4\delta\sqrt{n} + 4\delta\sqrt{n})^2 = \frac{32}{\sigma^2} \|X\|_{\text{op}}^2 \cdot \delta^2 n.
\end{aligned}$$

Note that the inequality on the last line holds due to triangle inequality and the fact that θ_i, θ_j have norm upper bounded by $4\delta\sqrt{n}$ local packing Θ_0,

$$\|\theta_i - \theta_j\|_2 \leq \|\theta_i\|_2 + \|\theta_j\|_2 \leq 4\delta\sqrt{n} + 4\delta\sqrt{n}.$$

[1]Note that here, when determining the local subset to pack, $\|\theta\|_2 \leq D$, the value D is often chosen later in such a way that we the RHS of the local Fano's bound (1) is at least $\frac{1}{2}\delta$

Recall Corollary 1 from last lecture: if $\{\theta_i, i = 1, \dots M\}$ is a 2δ-packing of Θ and $\max_{i,j} D(\theta_i \| \theta_j) \le g(\delta)$, then we have

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_\theta}[\rho(\hat{\theta}(x), \theta)] \ge \delta \left(1 - \frac{g(\delta) + \log 2}{\log M}\right).$$

Here, we have $g(\delta) = \frac{32}{\sigma^2} \|X\|_{\text{op}}^2 \cdot \delta^2 n$, hence

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}\|\hat{\theta} - \theta\|_2 \ge 2\delta\sqrt{n} \left(1 - \frac{\frac{32}{\sigma^2}\|X\|_{\text{op}}^2 \cdot \delta^2 n}{\log 2^d}\right).$$

Choose $\delta^2 = \frac{\sigma^2}{64} \cdot \frac{d}{n} \cdot \frac{1}{\|X\|_{\text{op}}}$, we have

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}\|\hat{\theta} - \theta\|_2 \gtrsim \frac{\sigma}{\|X\|_{\text{op}}} \sqrt{d}.$$

The above bound holds for any fixed covariate matrix X. If in addition we assume that $X_{ij} \overset{\text{iid}}{\sim} N(0,1)$, then $\|X\|_{\text{op}} \lesssim \sqrt{d} + \sqrt{n} \le 2\sqrt{n}$ holds with high probability, as $n > d$ under current set up, we can further simplify RHS of above relationship and arrive at

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}\|\hat{\theta} - \theta\|_2 \gtrsim \sigma \sqrt{\frac{d}{n}}.$$

This lower bound shows that the standard least-squares estimator is minimax optimal (up to a multiplicative constant) under this setting.

2 Sparse Regression in High Dimension

Similarly to the previous section, here we still have $y = X\theta + e$, but now we restrict to the set of sparse vectors $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\|_0 \le s\}$, with s controlling the sparsity.

We consider the subset $\Theta_0 = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \le s, \|\theta\|_2 \le 1\}$, which is the ℓ_2 ball intersecting the set of sparse vectors. To locally pack this subset, we utilize following lemma, which often appears in Information/Coding Theory.

Lemma 1 (Sparse Gilbert-Varshamov)**.** *Suppose $s \le \frac{d}{8}$. There exist a set of binary vectors $w_1, \dots, w_M \in \{0,1\}^d$ such that*

1. $\|w_i - w_j\|_0 \ge \frac{s}{2}, \forall i \ne j$,
2. $\|w_i\|_0 = s, \forall i$,
3. $M \ge (\frac{d}{2s})^{\frac{s}{8}}$.

Proof Here is a sketch of proof utilizing the *probabilistic method.* Take $w_i \overset{\text{iid}}{\sim} \text{Unif}\{w \in \{0,1\}^d : \|w\|_0 = s\}$. Then $\mathbb{P}\big(\text{property } (i) \text{ and } (ii) \text{ hold}\big) > 0$ as long as $M = (\frac{d}{2s})^{\frac{s}{8}}$. This non-zero probability implies the existence of a set of vectors satisfying the above three properties. □

Now take $\theta_i = \frac{w_i}{\sqrt{s}}$, $i = 1, \dots, M$. One can verify that $\{\theta_1, \dots, \theta_M\}$ forms a $\frac{1}{2}$-packing of $\Theta_0 = \{\|\theta\|_0 \le s, \|\theta\|_2 \le 1\}$, with

1. $\|\theta_i - \theta_j\|_2 = \frac{1}{\sqrt{s}}\|w_i - w_j\|_2 = \frac{1}{\sqrt{s}}\sqrt{\|w_i - w_j\|_0} \ge \frac{1}{2}$, and
2. $\|\theta_i\|_2 = \frac{1}{\sqrt{s}}\|w_i\|_2 = \frac{1}{\sqrt{s}}\sqrt{\|w_i\|_0} = 1$.

Then by re-scaling, we have a 2δ-packing satisfying $\|\theta_i - \theta_j\| \le 8\delta$, whence

$$\begin{aligned} D(\mathbb{P}_{\theta_i} \| \mathbb{P}_{\theta_j}) &= \frac{1}{2\sigma^2} \|X(\theta_i - \theta_j)\|_2^2 \\ &\le \frac{1}{2\sigma^2} \max_{|T|=2s} \left\{ \frac{1}{n} \|X_T\|_{\text{op}}^2 \right\} \|\theta_i - \theta_j\|_2^2 \cdot n \\ &\le \frac{1}{2\sigma^2} \gamma^2 \cdot 64\delta^2 n, \qquad \text{for } \gamma = \max_{|T|=2s} \left\{ \frac{1}{n} \|X_T\|_{\text{op}}^2 \right\} \\ &= \frac{32\gamma^2}{\sigma^2} \delta^2 n. \end{aligned}$$

Applying Corollary 1 from previous Lecture, we obtain that

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E} \|\hat{\theta} - \theta\|_2 \ge 2\delta \cdot \left(1 - \frac{\frac{32\gamma^2}{\sigma^2} \delta^2 n}{\log(\frac{d}{2s})^{\frac{s}{8}}} \right).$$

Choosing $\delta^2 = \frac{1}{n} \cdot \frac{\sigma^2}{64\gamma^2} \cdot \frac{s}{8} \log(\frac{d}{2s})$, we see thta the above inequality becomes

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E} \|\hat{\theta} - \theta\|_2 \gtrsim \frac{\sigma}{\gamma} \sqrt{\frac{s \cdot \log \frac{d}{s}}{n}}$$

Again, the above bound holds for any fixed X. If we have Gaussian X, which satisfies $\gamma \le \frac{3}{2}$ with high probability due to RIP (see Lectures 16 and 17 on Sparse Regression), we can continue simplify RHS and arrive at

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E} \|\hat{\theta} - \theta\|_2 \gtrsim \sigma \sqrt{\frac{s \cdot \log \frac{d}{s}}{n}}.$$

Comparing with bounds for Lasso from previous lecture, that

$$\mathbb{E} \|\hat{\theta}_{\text{Lasso}} - \theta\| \lesssim \sigma \sqrt{\frac{s \log d}{n}},$$

we can conclude that Lasso is (almost) minimax optimal, especially when $s \ll d \Rightarrow \log d \asymp \log \frac{d}{s}$.

3 Matrix Completion

We have previously discussed matrix completion problem in the lecture on random matrix theory when we covered the matrix Bernstein's inequality. Here we revisit this problem again and we would like to examine whether the estimator developed therein is minimax optimal.

Here is the set up. Let $G \in [-1, 1]^{d \times d}$ be an unknown rank-1 matrix. For each $(i, j) \in [d] \times [d]$, we observe $Y_{ij} = X_{ij} G_{ij} + e_{ij}$, with $X_{ij} \overset{\text{iid}}{\sim} \text{Ber}(p)$[2] and $e_{ij} \overset{\text{iid}}{\sim} N(0, 1)$. Given Y, we would like to estimate G.

Claim 1. *There exists a $\frac{\delta d}{2}$-packing $\{G^1, \ldots, G^M\}$ of $[-\delta, \delta]^{d \times d}$ in $\|\cdot\|_F$ with $\|G^i\|_F \le \delta d$ and $\log M \gtrsim d$.*

Proof Take $\{w_1, \ldots, w_M\} \in \{0, 1\}^d$ from Gilbert-Varshamov Lemma, and then we shift the vector to obtain $u_i = w_i - \frac{1}{2} \in \{-\frac{1}{2}, \frac{1}{2}\}^d$, which are Rademacher-like vectors. Then let $G^i = u_i u_i^T$. We can check these vectors satisfy above properties. □

[2] X_{ij} is called the mask matrix, which means that it reveals the entry G_{ij} with probability p; otherwise, we observe 0 and do not see the entry.

By Theorem 1 from last lecture (estimation to testing), we have

$$\begin{aligned}\inf_{\hat{G}}\sup_{G}\mathbb{E}\|\hat{G}-G\|_F &\gtrsim \delta\cdot d\cdot\inf_{\psi}\frac{1}{M}\sum_{j=1}^{M}\mathbb{P}(\psi(Y)\neq j|G^j)\\ &=\delta\cdot d\cdot\inf_{\psi}\mathbb{P}(\psi(Y)\neq J),\quad J\sim\mathrm{Unif}\{1,\dots,M\}\\ &=\delta\cdot d\cdot\inf_{\psi}\mathbb{E}_X\left[\mathbb{P}(\psi(Y)\neq J)|X\right].\end{aligned}$$

Note that $Y|X,G^j\sim N(X\circ G^j,1)$ (d^2-dimensional Gaussian), with $\circ$ represents element-wise multiplication. Then, for each fixed X, by equation (1) (Fano's inequality followed by upper bound on mutual information using KL-divergence), we have

$$\begin{aligned}\mathbb{P}(\psi(Y)\neq J|X)&\geq 1-\frac{\frac{1}{M^2}\sum_{i,j=1}^{M}D\Big(N(X\circ G^i)\|N(X\circ G^j)\Big)}{\log M}\\ &=1-\frac{\frac{1}{M^2}\sum_{i,j=1}^{M}\|X\circ G^i-X\circ G^j\|_F^2}{\log M}.\end{aligned}$$

Taking expectation w.r.t. X and by linearity of expectation, we have

$$\mathbb{E}_X[\mathbb{P}(\psi(Y)\neq J|X)]\geq 1-\frac{\frac{1}{M^2}\sum_{i,j=1}^{M}\mathbb{E}_X[\|X\circ G^i-X\circ G^j\|_F^2]}{\log M}\tag{2}$$

Notice that

$$\mathbb{E}_x\left[\|X\circ G^i-X\circ G^j\|_F^2\right]=\mathbb{E}_x\sum_{u,v}X_{uv}^2(G_{uv}^i-G_{uv}^j)^2)=p\cdot\|G^i-G^i\|_F^2\leq p\cdot(2\delta d)^2=4p\delta^2d^2.$$

Substituting above relationship into equation (2), we obtain that

$$\mathbb{P}(\psi(Y)\neq J|X)\geq 1-\frac{\frac{p}{M^2}\sum_{i,j=1}^{M}\|G^i-G^j\|_F^2}{\log M}\geq 1-\frac{4p\delta^2d^2}{d}=1-4pd\delta^2.$$

Take $\delta=\sqrt{\frac{1}{8pd}}$ and we arrive at $\inf_{\hat{G}}\sup_G\|\hat{G}-G\|_F\gtrsim\delta d\asymp\sqrt{\frac{d}{p}}$.

The above result can be generalized to the rank-r, in which case we have the minimax lower bound

$$\inf_{\hat{G}}\sup_{G}\|\hat{G}-G\|_F\gtrsim\sqrt{\frac{rd}{p}}.$$

In Lecture 9, we used matrix Bernstein's inequality to show that the singular value thresholding estimator satisfies the error bound

$$\|\hat{G}-G\|_F\lesssim\sqrt{\frac{rd\log d}{p}}.$$

We see that this estimator is minimax optimal up to a $\log d$ factor.