36-709: Advanced Statistical Theory I **Spring 2020**

Lecture 18: March 31

Lecturer: Siva Balakrishnan *Scribe: Dongyu Li*

Last time, we introduced the Modified Varshamov-Gilbert bound and we applied it to prove some high-dimensional lower bounds: (1) variable selection lower bound; (2) estimating sparse mean when θ is s-sparse; (3) lower bound for estimating θ when the vector is ℓ_1-sparse.

Today, we continue to look at Yang-Barron method for proving minimax lower bounds.

18.1 Yang-Barron Method

The high level goal here is to improve the upper bound on mutual information, to obviate the need to do local packing. Here is a mini lemma we proved in the last lecture that upper bounds the mutual information $I(Z; J)$.

Lemma 18.1

$$\begin{aligned} I(Z; J) &= \frac{1}{M} \sum_{i=1}^{M} KL(P_{\theta^i}, \overline{P}) \\ &\leq \frac{1}{M} \sum_{i=1}^{M} KL(P_{\theta^i}, Q) \end{aligned}$$

where $\overline{P} = \sum_{i=1}^{M} P_{\theta^i}$ *and* Q *is any other distribution.*

In general, there are two ways to upper bound mutual information: (1) upper bound by KL diameter, which is the maximum KL-divergence between any two distributions (2) upper bound by $\log(M)$ in cardinality based method. Yang-Barron method essentially makes a statement about the trade-offs between the above two upper bounds.

Theorem 18.2 *(Yang-Barron Method)*
Let $N_{KL}(\epsilon, \mathcal{P})$ *denote the* ϵ*-covering number of* $\mathcal{P}$ *in the square-root KL divergence. Then the mutual information is upper bounded as*

$$I(Z; J) \leq \inf_{\epsilon > 0} \{\epsilon^2 + \log N_{KL}(\epsilon, \mathcal{P})\}$$

where, N_{KL} *is the number of distributions to cover* $\mathcal{P}$ *in the semi-metric* $\sqrt{KL}$*.*

Note that one significance of the statement is that nothing on the RHS depends on what the packing is, namely the collection of θs, $\{\theta^1, ..., \theta^M\}$. It only depends on the class of distributions $\mathcal{P}$ from which we're selecting the packing from.

Proof: By Lemma 18.1, we have

$$I(Z;J) \leq \frac{1}{M}\sum_{i=1}^{M} \mathrm{KL}(P_{\theta^i}, Q)$$

We will now pick a distribution Q to get a bound that we want. Let $Q = \frac{1}{N_{\mathrm{KL}}}\sum_{i=1}^{N_{\mathrm{KL}}} P_{\gamma^i}$, where $\{\gamma^1, ..., \gamma^{N_{\mathrm{KL}}}\}$ is a $\sqrt{\mathrm{KL}}$ ϵ-covering of $\mathcal{P}$. Now, plugging in Q,

$$I(Z;J) \leq \frac{1}{M}\sum_{i=1}^{M} \mathrm{KL}(P_{\theta^i}, \frac{1}{N}\sum_{j=1}^{N} P_{\gamma^j})$$

$\forall \theta^i$, there is some γ that's closest to it, and we call it $\widetilde{\gamma}^i$.

$$\begin{aligned} I(Z;J) &\leq \frac{1}{M}\sum_{i=1}^{M} \int P_{\theta^i} \log \frac{P_{\theta^i}}{\frac{1}{N}\sum_{j=1}^{N} P_{\gamma^j}} \\ &\leq \frac{1}{M}\sum_{i=1}^{M} \int P_{\theta^i} \log \frac{P_{\theta^i}}{\frac{1}{N} P_{\widetilde{\gamma}^i}} \\ &\leq \frac{1}{M}\sum_{i=1}^{M} \mathrm{KL}(P_{\theta^i}, P_{\widetilde{\gamma}^i}) + \log N \\ &\leq \epsilon^2 + \log N \end{aligned}$$

where the first term is a diameter based bound, and the second term is a cardinality based bound. ■

We have a choice here to pick ϵ. If we pick ϵ to be large, then we pay a large diameter price and a small cardinality price; and, if we pick ϵ to be small, then we pay a small diameter price and a large cardinality price.

18.2 Yang-Barron Application

18.2.1 Application Procedure

Procedure for applying Yang-Barron method:

1. We would like to find an ϵ such that

$$\inf_{\epsilon>0}\{\epsilon^2 + \log N_{\mathrm{KL}}\}$$

Since as ϵ increases, ϵ^2 increases and $\log N_{\mathrm{KL}}$ decreases, one way to balance these two terms is to pick the smallest ϵ such that $\epsilon^2 \geq \log N_{\mathrm{KL}}(\mathcal{P}, \epsilon)$.

2. Choose the largest δ that we can find a 2δ-packing $\{\theta^1, ..., \theta^M\}$ that satisfies the lower bound

$$\log M(\delta) \geq 2\epsilon^2$$

3. Then, we get $M \geq \Phi(\delta)$

Yang-Barron essentially decoupled step 1 and step 2 for us so that we don't have to reason about them simultaneously.

18.2.2 Example: Non-parametric Regression with Sobolev Functions

In non-parametric regression with sobolev functions, $f = \sum_{j=1}^{\infty} \theta_j \phi_j$, $\sum_{j=1}^{\infty} \theta_i^2 \leq 1$, where θ_j's are coefficients and ϕ_j's are basis vectors. And, we have the Sobolev Ellpsoid constraint that $\sum_{j=1}^{\infty} \theta_j^2 j^{2\alpha} \leq 1$.

One way to think about the constraint is that if we plug in $\alpha = 1$, then the M-th coefficient must satisfy $M^{2\alpha}\theta_M^2 \leq 1$, meaning that θ_M must decay at some rate. Consequently, the higher-order coefficients must be really small and we can think of it as an ordered sparsity constraint.

We know from Chapter 5 that $\log N(\mathcal{F}, \|.\|_2, \delta) \asymp \left(\frac{1}{\delta}\right)^{1/\alpha}$, where $\mathcal{F}$ is the collection of Sobolev functions. Using Yang-Barron, we can prove a lower bound just using the knowledge of this metric entropy.

We follow the steps in 18.2:

1. Under the general regression setup, where X is from some distribution say $X \sim$ Unif$[0, 1]$ and $Y|X \sim N(f(x), \sigma^2)$, we've derived before that

$$\mathrm{KL}(P_{f_1}, P_{f_2}) = \tfrac{n}{2\sigma^2} \|f_1 - f_2\|_2^2$$

We want to ensure that $\sqrt{\mathrm{KL}} \leq \epsilon, \forall f_1, f_2$, namely constructing a ϵ-covering in the $\sqrt{\mathrm{KL}}$ semi-metric.

$$\sqrt{\mathrm{KL}} = \sqrt{\frac{n}{2\sigma^2} \|f_1 - f_2\|_2^2} \leq \epsilon$$

$$\|f_1 - f_2\|_2 \leq \frac{\epsilon\sqrt{2\sigma^2}}{\sqrt{n}}$$

which tells us that it's sufficient to construct a $\frac{\epsilon\sqrt{2\sigma^2}}{\sqrt{n}}$-covering in the ℓ_2 metric. Now, we want to ensure

$$\begin{aligned}\epsilon^2 &\geq \log N_{\mathrm{KL}}(\mathcal{F}, \epsilon) \\ \epsilon^2 &\geq (\frac{\sqrt{n}}{\epsilon\sqrt{2\sigma^2}})^{1/\alpha} \\ \epsilon^{2+1/\alpha} &\geq n^{\alpha/2} \\ \epsilon &\geq n^{\frac{1}{2(2\alpha+1)}}\end{aligned}$$

2. The second step is to pick the largest δ such that $\log M(\delta) \geq \epsilon^2$

$$\begin{aligned}\frac{1}{\delta}^{1/\alpha} &\geq n^{\frac{1}{2\alpha+1}} \\ \delta &\leq n^{-\frac{\alpha}{2\alpha+1}}\end{aligned}$$

3. $\mathcal{M} \geq c\Phi(\delta) = c\delta^2 \geq cn^{-\frac{2\alpha}{2\alpha+1}}$

Note that this is the usual non-parametric rate for Sobolev functions and we just needed to know the metric entropy to get this lower bound.

18.3 Non-parametric Maximum Likelihood Methods

The high level goal is to find out how Le Cam shows up in upper bound. Roughly, if we can find an ϵ such that $n\epsilon^2 \geq$ some metric entropy, then he can show a corresponding estimator that achieves ϵ as the rate of convergence.

Let's first define the non-parametric maximum likelihood problem. Observe $X_1, ..., X_n \sim P_0 \in \mathcal{P}$. We'll do maximum likelihood over the class $\mathcal{P}$, namely that

$$\widehat{P_n} = \arg\max_{P \in \mathcal{P}} \ell_n(P)$$

We should regard the above as some description of the estimator. An alternative method is called the method of sieves. The idea is that if the class $\mathcal{P}$ is too big, then we just pick a subset of it $\{P_1, ..., P_N\}$, and optimize over this small subset

$$\widehat{P_n} = \arg\max_{P \in \{P_1, ..., P_N\}} \ell_n(P)$$

The high level goal is to analyze $\widehat{P_n}$ by asking how far is $\widehat{P_n}$ from P_0 in some metric, and we'll use the Hellinger metric (i.e. $\mathbb{E}\, h(\widehat{P_n}, P_0)$ or $\mathbb{P}(h(\widehat{P_n}, P_0) \geq \delta)$).

Claim: (Informal) If we can find an ϵ^* such that $n\epsilon^{*^2} \geq \log H(\mathcal{P}, \epsilon^*)$, then

$$\mathbb{P}(R(\widehat{P_n}, P_0) \geq C\epsilon^*) \leq C_1 \exp\{-C_2 n\epsilon^{*^2}\}$$

where $\widehat{P_n}$ is the sieve MLE.

The above claim basically says that Le Cam determines the rate of convergence of sieve MLE in the Hellinger distance. Note that Hellinger and $\sqrt{\text{KL}}$ are close (recall that $H^2 \leq TV \leq H \leq \sqrt{\text{KL}}$).

To prove this claim, we need to understand the relation between Hellinger distance and likelihood. Intuitively, we want to say that if we take a distribution from the net $\{P_1, ..., P_n\}$ that is far away from P_0 in Hellinger (a distribution that's not within an ϵ ball around P_0 in Hellinger distance), then it's unlikely that the distribution picked is maximizing the likelihood. In particular, we want such likelihood to be exponentially small so that we can union bound over all such distributions.

Lemma 18.3 *(Wong-Shen: probability inequality for the likelihood ratio) If for some distribution P such that $h(P, P_0) \geq \delta$, then*

$$\mathbb{P}(\prod_{i=1}^{n} \frac{P(X_i)}{P_0(X_i)} \geq \exp\{-\frac{n\delta^2}{2}\}) \leq \exp\{-\frac{n\delta^2}{4}\}.$$

The above lemma is basically saying that the likelihood ratio is exponentially small in the Hellinger distance with high probability. If P is far away from P_0, namely that δ is large, then this likelihood ratio is small.

Proof: The importance of the proof is to see where Hellinger comes up in likelihood ratio.

$$\begin{aligned}
\mathbb{P}(\prod_{i=1}^{n} \frac{P(X_i)}{P_0(X_i)} \geq \exp\{-\frac{n\delta^2}{2}\}) &= \mathbb{P}(\prod_{i=1}^{n} \sqrt{\frac{P(X_i)}{P_0(X_i)}} \geq \exp\{-\frac{n\delta^2}{4}\}) \\
&\leq \mathbb{E}[\prod_{i=1}^{n} \sqrt{\frac{P(X_i)}{P_0(X_i)}}] \times \exp\{\frac{n\delta^2}{4}\} \\
&= (\int \sqrt{P \times P_0})^n \times \exp\{\frac{n\delta^2}{4}\} \\
&= (1 - \frac{H^2}{2})^n \times \exp\{\frac{n\delta^2}{4}\} \\
&= (1 - \frac{\delta^2}{2})^n \times \exp\{\frac{n\delta^2}{4}\} \\
&\leq \exp\{-\frac{n\delta^2}{4}\}
\end{aligned}$$

where $H^2 = 2 - 2\int \sqrt{P \times P_0}$ and the second step is by Markov Inequality. ■

Now, we are going to construct $\{P_1, ..., P_N\}$ by using a Hellinger cover of $\mathcal{P}$. We will just maximize the likelihood over this collection, but one difficulty is that P_0 may not be in the sieve.

If P_0 is a sieve member, then we know that for $P \in \mathcal{P}$,

$$\mathbb{P}(\frac{L(P)}{L(P_0)} \geq 1) \leq \exp\{-nh^2\}$$

By union bound over the net,

$$\mathbb{P}(\exists j : h(P_j, P_0) \geq \epsilon, \frac{L(P_j)}{L(P_0)} \geq 1) \leq N \exp\{-nh^2\}$$

Le Cam makes sure that we pick an ϵ large enough, namely $n\epsilon^2 \geq \log N$, so that probability $N \exp\{-nh^2\} = \exp\{-n\epsilon^2 + \log N\}$ goes to 0. On a high level, one way to think about where Le Cam comes from is a union bound combined with the Wong-Shen exponential inequality.

The more difficult case is when P_0 is not in the sieve, where we will analyze the likelihood ratio of some P_j closest to P_0, in the next lecture.