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

Lecture 17: March 26

Lecturer: Siva Balakrishnan *Scribe: Ankur Mallick*

In this lecture we will see some more applications of Fano's Method for deriving minimax lower bounds in high-dimensional settings. We will also see an alternate upper bound for Mutual Information using Yang-Barron's Lemma that improves on the upper bounds we have been using so far.

17.1 More Examples of Fano's Method

17.1.1 Lower Bound for Support Recovery

Consider a collection of random vectors $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2 I_d)$. The problem of support recovery seeks to recover the support of θ given $X_1, \ldots, X_n$.

Recall that if $\theta_{\min} = \min_{i:\theta_i \neq 0} |\theta_i| \gg \sigma\sqrt{\frac{\log d}{n}}$, then the hard thresholding estimator $\mathrm{HT}(\bar{X})$ defined as

$$\mathrm{HT}(\bar{X})_i = \begin{cases} \bar{X}_i, & \text{if } \bar{X}_i \gtrsim \sigma\sqrt{\frac{\log d}{n}} \\ 0, & \text{otherwise} \end{cases} \tag{17.1}$$

where $\bar{X} = \frac{1}{n}\sum_i X_i$, recovers the support of θ with high probability.

To apply Fano's inequality, consider the set of distributions $\{\mathbb{P}_{\theta^1}, \ldots, \mathbb{P}_{\theta^d}\}$, where

$$\theta^i = [0, \ldots, \theta_{\min}, \ldots, 0]^T$$

is the vector with zeros at all positions except the i^{th} positions where it has the value $\theta_{\min}$ and $\mathbb{P}_{\theta^i}$ is a Gaussian centered at θ^i with covariance $\sigma^2 I_d$. We will use Fano's inequality on the test function $\Psi(X_1, \ldots, X_n) \in \{1, \ldots, d\}$ which seeks to identify the data generating distribution from among $\{\mathbb{P}_{\theta^1}, \ldots, \mathbb{P}_{\theta^d}\}$ since if we cannot solve the testing problem, we definitely cannot solve the support recovery problem as all $\mathbb{P}_{\theta^i}$'s have different support.

We know from Fano's inequality that the testing error is lower bounded as

$$Q(\Psi(X_1, \ldots, X_n) \neq J) \geq 1 - \frac{I(X_1, \ldots, X_n; J) + \log 2}{\log d} \tag{17.2}$$

where $I(X_1, \ldots, X_n; J) \leq \frac{1}{d^2}\sum_{i,j} \mathrm{KL}(\mathbb{P}_{\theta^i} \| \mathbb{P}_{\theta^j})$ and the KL divergence between any pair $\mathbb{P}_{\theta^j}, \mathbb{P}_{\theta^j}$ is given by

$$\mathrm{KL}(\mathbb{P}_{\theta^i} \| \mathbb{P}_{\theta^j}) = \frac{2n\theta_{\min}^2}{2\sigma^2} = \frac{n\theta_{\min}^2}{\sigma^2}. \tag{17.3}$$

Since the KL divergence is the same for any such pair, $I(X_1, \ldots, X_n; J) \leq \frac{n\theta_{\min}^2}{\sigma^2}$ and from (17.2) we can see that if $I(X_1, \ldots, X_n; J) \lesssim \log d$ then $Q(\Psi(X_1, \ldots, X_n) \neq J) \geq c$ for some constant c. Consequently, the minimax support recovery error $\inf_{\widehat{S}} P(\widehat{S} \neq S)$ is lower bounded by some constant. In other words, if

$$\frac{n\theta_{\min}^2}{\sigma^2} \lesssim \log d \tag{17.4}$$

$$\text{i.e. } \theta_{\min} \lesssim \sigma\sqrt{\frac{\log d}{n}} \tag{17.5}$$

then support recovery is hard. This matches the condition on $\theta_{\min}$ for the hard thresholding estimator in (17.1).

Note that, the above results are specific to the support recovery problem. We would get different conditions on $\theta_{\min}$ if we used a different loss like the Hamming distance between estimated and true parameters, or if only cared about recovering elements of θ greater than a certain threshold value etc.

17.1.2 Lower Bounds for Sparse Estimation

To derive lower bounds for sparse estimation we will need to modify the Varshamov-Gilbert constrution that we used earlier to construct packing sets. The modified bound is given by the following Lemma from [1]

Lemma 17.1 *Define $\widetilde{\Omega} = \{\boldsymbol{\omega} \in \{0,1\}^d \colon \|\boldsymbol{\omega}\|_0 \leq S\}$. Then there exists $\Omega' \subseteq \widetilde{\Omega}$ such that:*

- *every $\boldsymbol{\omega} \in \Omega'$ has $\|\boldsymbol{\omega}\|_0 = S$,*
- *every pair $\boldsymbol{\omega}_i, \boldsymbol{\omega}_j \in \Omega'$ has Hamming distance $d_H(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) \geq S/2$,*
- *$|\Omega'| \geq c\left(\frac{de}{S}\right)^S$.*

This is important because even though the ℓ_0-ball $\widetilde{\Omega}$ has cardinality $|\widetilde{\Omega}| \leq \sum_{k=1}^{S}\binom{d}{k} \sim \left(\frac{de}{S}\right)^S$, all its elements are not well separated. However the above lemma tells us that we can find a packing of well separated vectors that is a subset of $|\widetilde{\Omega}|$ and has the same cardinality (order-wise).

17.1.2.1 Lower Bound for ℓ_0-Sparsity

Consider a collections of random vectors $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2 I_d)$ where $\|\theta\|_0 \leq S$ (ℓ_0-Sparsity). The hard thresholding estimator as defined in (17.1) has the following upper bound on the expected ℓ_2-error

$$\mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \lesssim \frac{\sigma^2 S \log d}{n} \tag{17.6}$$

where we use $\widehat{\theta}$ denote the hard thresholding estimate of θ.

To find a lower bound, we first construct a packing set of $M := \left(\frac{de}{S}\right)^S$, $S-$sparse vectors with the Hamming distance between any two elements $\boldsymbol{\omega}_i, \boldsymbol{\omega}_j$ lower bounded as $d_H(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) \geq S/2$ (as per Lemma 17.1).

Consider θ_i's generated as $\theta_i = \boldsymbol{\omega}_i \times \theta_{\min}$ where $\theta_{\min} = \min_{i:\theta_i \neq 0} |\theta_i|$. Once again, to apply Fano's inequality, we will consider the set of distributions $\{\mathbb{P}_{\theta^1}, \ldots, \mathbb{P}_{\theta^M}\}$, ($\mathbb{P}_{\theta^i}$ is a Gaussian centered at θ^i with covariance $\sigma^2 I_d$) and the corresponding testing function $\Psi(X_1, \ldots, X_n) \in \{1, \ldots, M\}$.

Observe that,

$$\mathrm{KL}(\mathbb{P}_{\theta^i} \| \mathbb{P}_{\theta^j}) = \frac{n}{2\sigma^2}\|\theta^i - \theta^j\|_2^2 \leq \frac{2nS\theta_{\min}^2}{2\sigma^2} = \frac{nS\theta_{\min}^2}{\sigma^2} \tag{17.7}$$

Since $\log|\Omega'| \gtrsim S \log \frac{de}{S}$, we can get a lower bound from Fano if $\mathrm{KL}(\mathbb{P}_{\theta^i} \| \mathbb{P}_{\theta^j}) \lesssim S \log \frac{de}{S}$ i.e.

$$\frac{nS\theta_{\min}^2}{\sigma^2} \lesssim S \log \frac{de}{S} \tag{17.8}$$

$$\text{or } \theta_{\min} \lesssim \sigma \sqrt{\frac{1}{n} \log \frac{de}{S}} \tag{17.9}$$

If $\theta_{\min}$ satisfies the above condition the lower bound then the minimax risk is lower bounded by the separation (δ) between two elements of the packing. From Lemma 17.1 we know that the separation satisfies,

$$\|\theta^i - \theta^j\|_2^2 \geq d_H(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j)\theta_{\min}^2 \geq \frac{S}{2}\theta_{\min}^2 \tag{17.10}$$

Combining (17.9) and (17.10) we get the lower bound on the minimax risk,

$$\mathcal{M}(\Phi \circ \rho) \geq c\Phi(\delta) = \frac{cS}{2}\frac{\sigma^2}{n} \log\left(\frac{ed}{S}\right) \simeq \sigma^2 \frac{S \log(d/S)}{n} \tag{17.11}$$

which matches the upper bound for sub-linear sparsity.

17.1.2.2 Lower Bound for ℓ_1-Sparsity

Consider a collections of random vectors $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2 I_d)$ where $\|\theta\|_1 \leq R$ (ℓ_1-Sparsity). The estimator $\widehat{\theta}$ obtained by solving the corresponding ℓ_1-minimization problem has expected ℓ_2 error (when $\sqrt{n} \ll d$) of:

$$\mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] \lesssim \sigma R \sqrt{\frac{\log d}{n}}. \tag{17.12}$$

Now we show a corresponding lower bound.

To find a lower bound, we once again construct a packing set of $M := \left(\frac{de}{K}\right)^K$, $K-$sparse vectors with the Hamming distance between any two elements $\boldsymbol{\omega}_i, \boldsymbol{\omega}_j$ lower bounded as $d_H(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) \geq K/2$ (as per Lemma 17.1). The crucial difference from the ℓ_0-case is that here K is a free parameter which we can choose (since the constraint is on the ℓ_1-norm and not on the ℓ_0-norm).

Consider θ_i's generated as $\theta_i = \boldsymbol{\omega}_i R/K$ which ensures that $\|\theta\|_1 \leq R$ and $\{\mathbb{P}_{\theta^1}, \ldots, \mathbb{P}_{\theta^M}\}$ as earlier.

$$\mathrm{KL}(\mathbb{P}_{\theta^i}\|\mathbb{P}_{\theta^j}) = \frac{n}{2\sigma^2}\|\theta^i - \theta^j\|_2^2 \leq \frac{n}{2\sigma^2} \times 2K \times \frac{R^2}{K^2} = \frac{nR^2}{K\sigma^2} \tag{17.13}$$

Therefore we need to ensure that $\frac{nR^2}{K\sigma^2} \lesssim K \log\left(\frac{ed}{K}\right)$ i.e.

$$K \gtrsim \frac{1}{\sigma}\sqrt{\frac{nR^2}{\log(ed/K)}} \tag{17.14}$$

The above condition can be satisfied if $\sqrt{n} \ll d$, otherwise we should pick K as the minimum of the RHS of (17.14) and d (recall that K is a parameter which we are free to pick to maximize the lower bound).

To simplify the rest of the analysis we suppose that $\sqrt{n} \ll d^{1-\epsilon}$ for some small constant $\epsilon > 0$. In this case, we can see that the choice,

$$K \gtrsim \frac{1}{\sigma}\sqrt{\frac{nR^2}{\log d}}, \tag{17.15}$$

works for our purposes.

This in turn yields,

$$\Phi(\delta) \geq \frac{K}{2} \times \frac{R^2}{K^2} = \frac{R^2}{K} \tag{17.16}$$

$$= \frac{R^2\sigma}{\sqrt{nR^2}}\sqrt{\log d} = \sigma R \sqrt{\frac{\log d}{n}} \tag{17.17}$$

which is the lower bound on the minimax risk $\mathcal{M}$ (since by substituting our choice of K in (17.14) into Fano's Inequality we can lower bound the Mutual Information term by a constant) and matches the upper bound in (17.12).

17.2 Alternate Upper Bound for Mutual Information

So far we have only considered local packings to bound the mutual information as

$$I(Z;J) \leq \sup_{\theta^i,\theta^j} \mathrm{KL}(\mathbb{P}_{\theta^i}\|\mathbb{P}_{\theta^j}) \tag{17.18}$$

where $\mathrm{KL}(\mathbb{P}_{\theta^i}\|\mathbb{P}_{\theta^j})$ is the KL diameter of the packing $\{\theta^1,\ldots,\theta^M\}$

Consider θ^1,θ^2 with disjoint support. The KL diameter is infinite due to disjoint support but the mutual information is always upper bounded by the log of the cardinality of the set, as long as the set has finite cardinality, because

$$I(Z;J) = \frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_{\theta^j}\|\frac{1}{M}\sum_{i=1}^{M}\mathbb{P}_{\theta^i}) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{E}_{\mathbb{P}_{\theta^j}}\left[\log\frac{\mathbb{P}_{\theta^j}}{\frac{1}{M}\sum_{i=1}^{M}\mathbb{P}_{\theta^i}}\right] \leq \log M \tag{17.19}$$

Thus in this case $I(Z;J) \leq \log 2$.

Thus we see two upper bounds on the Mutual Information - one which uses a local packing, and another which relies on the cardinality of the packing set. The following lemma enables us to combine these two bounds through a global backing to improve on the above two bounds

Lemma 17.2 (Yang-Barron) *If $\mathcal{N}_{KL}(\mathcal{P},\epsilon)$ is the covering number of $\mathcal{P}$ in the $\sqrt{KL}$ metric then*

$$I(Z;J) \leq \inf_{\epsilon>0}\left\{\epsilon^2 + \log\mathcal{N}_{KL}(\mathcal{P},\epsilon)\right\} \tag{17.20}$$

Proof: Consider a packing $\{\theta^1,\ldots,\theta^M\}$ of $\mathcal{P}$. We know that

$$I(Z;J) = \frac{1}{M}\sum_{j=1}^{M}\mathrm{KL}(\mathbb{P}_{\theta^j}\|\bar{\mathbb{P}}) \tag{17.21}$$

where $\bar{\mathbb{P}} = \frac{1}{M}\sum_{i=1}^{M}\mathbb{P}_{\theta^i}$. Also observe that

$$\frac{1}{M}\sum_{j=1}^{M}\mathrm{KL}(\mathbb{P}_{\theta^j}\|\mathbb{Q}) = \frac{1}{M}\sum_{j=1}^{M}\mathbb{E}_{\mathbb{P}_{\theta^j}}\left[\log\left(\frac{\mathbb{P}_{\theta^j}}{\mathbb{Q}}\times\frac{\bar{\mathbb{P}}}{\bar{\mathbb{P}}}\right)\right] \tag{17.22}$$

$$= \frac{1}{M}\sum_{j=1}^{M}KL(\mathbb{P}_{\theta^j}\|\bar{\mathbb{P}}) + \mathrm{KL}(\bar{\mathbb{P}}\|\mathbb{Q}) \tag{17.23}$$

Combining (17.21) and (17.23) and observing that $\mathrm{KL}(\bar{\mathbb{P}}\|\mathbb{Q}) \geq 0$ (always), we see that

$$I(Z;J) \leq \frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_{\theta^j}\|\mathbb{Q}) \tag{17.24}$$

for any distribution $\mathbb{Q}$. We choose $\mathbb{Q}$ to be the uniform distribution over an ϵ cover in the $\sqrt{\mathrm{KL}}$ metric, $\{\gamma_1, \ldots, \gamma_{\mathcal{N}_{\mathrm{KL}}}\}$ i.e.

$$\mathbb{Q} = \frac{1}{\mathcal{N}_{\mathrm{KL}}}\sum_{j=1}^{\mathcal{N}_{\mathrm{KL}}} \mathbb{P}_{\gamma_j} \tag{17.25}$$

■ Thus,

$$\begin{aligned} I(Z;J) &\leq \frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_{\theta^j}\|\mathbb{Q}) && (17.26)\\ &= \frac{1}{M}\sum_{j=1}^{M} \mathbb{E}_{\mathbb{P}_{\theta^j}}\left[\log\left(\frac{\mathbb{P}_{\theta^j}}{\frac{1}{\mathcal{N}_{\mathrm{KL}}}\sum_{j=1}^{\mathcal{N}_{\mathrm{KL}}}\mathbb{P}_{\gamma_j}}\right)\right] && (17.27)\\ &\leq \frac{1}{M}\sum_{j=1}^{M} \mathbb{E}_{\mathbb{P}_{\theta^j}}\left[\log\left(\frac{\mathbb{P}_{\theta^j}}{\mathbb{P}_{\widetilde{\gamma_j}}/\mathcal{N}_{\mathrm{KL}}}\right)\right] && (17.28) \end{aligned}$$

where $\mathbb{P}_{\widetilde{\gamma_j}}$ is the element of the cover that is closest to $\mathbb{P}_{\theta^j}$ in the $\sqrt{\mathrm{KL}}$ metric. Therefore

$$I(Z;J) \leq \frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(\mathbb{P}_{\theta^j}\|\mathbb{P}_{\widetilde{\gamma_j}}) + \log \mathcal{N}_{\mathrm{KL}} \tag{17.29}$$

$\mathrm{KL}(\mathbb{P}_{\theta^j}\|\mathbb{P}_{\widetilde{\gamma_j}}) \leq \epsilon^2$ since $\{\gamma_1, \ldots, \gamma_{\mathcal{N}_{\mathrm{KL}}}\}$ is an ϵ cover in the $\sqrt{\mathrm{KL}}$ metric. Therefore

$$I(Z;J) \leq \inf_{\epsilon > 0}\left\{\epsilon^2 + \log \mathcal{N}_{\mathrm{KL}}(\mathcal{P}, \epsilon)\right\} \tag{17.30}$$

The Yang-Barron Lemma is applied to upper bound the mutual information in Fano's Inequality in the following steps

1. Pick the smallest ϵ such that $\epsilon^2 \geq \log \mathcal{N}_{\mathrm{KL}}(\mathcal{P}, \epsilon)$ which balances the two terms on the right-hand side of the Yang-Barron Lemma and implies that $I(Z;J) \leq 2\epsilon^2$.
2. Find a 2δ packing $\{\theta^1, \ldots, \theta^M\}$ of $\mathcal{P}$ with the largest possible δ such that $\log M \geq 2\epsilon^2$.
3. The above two points implies that $I(Z:J) \leq \log M$ for this choice of δ and by Fano's Inequality the testing error $Q(\Psi \neq J)$ is lower bounded by a constant due to which the minimax risk is lower bounded as $\mathcal{M} \geq \Phi(\delta)$

The main advantage of using the Yang-Barron Lemma is that we can decouple picking ϵ and picking the picking whereas earlier we had to do both together. Now, we have abstracted the problem of computing an upper bound on the mutual information to finding the KL metric entropy and the packing number of the set.

References

[1] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.