36-789: Topics in High Dimensional Statistics II	**Fall 2015**

Lecture 3: November 3

Lecturer: Alessandro Rinaldo *Scribes: Chun-Liang Li*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

3.1 Examples of Le cam Lemma

3.1.1

Let $x_1, \cdots, x_n$ are i.i.d. samples from $\{-1, 1\}$, $\mathbb{E}(x_i) = \theta$ and $w\left(d(\hat{\theta}, \theta)\right) = |\hat{\theta}, \theta|^2$. By the two points arguments,

$$P_{\theta_1}(1) = \tfrac{1+\delta}{2}, \quad P_{\theta_{-1}}(1) = \tfrac{1-\delta}{2}, \quad P_{\theta_1}(-1) = \tfrac{1-\delta}{2}, \quad P_{\theta_{-1}}(-1) = \tfrac{1+\delta}{2}.$$

We then have $\mathbb{E}_{\theta_1}(x) = \delta$, $\mathbb{E}_{\theta_{-1}}(x) = -\delta$ and $d(\theta_1, \theta_{-1}) = 2\delta$.

By Le cam Lemma, the minimax risk is $\delta^2 \dfrac{1 - d_{TV}(P^n_{\theta_1}, P^n_{\theta_{-1}})}{2}$. If $d_{TV}(P^n_{\theta_1}, P^n_{\theta_{-1}}) \leq 1/2$, the minimax risk is $\delta^2/4$. To find δ such that the above is true, we have

$$d_{TV}(P^n_{\theta_1}, P^n_{\theta_{-1}})^2 \leq \frac{n}{2} KL(P_{\theta_1}, P_{\theta_{-1}}) = \frac{n}{2}\delta \log \frac{1+\delta}{1-\delta} \leq \frac{n}{2} \times 3\delta.$$

Then $d_{TV}(P^n_{\theta_1}, P^n_{\theta_{-1}})^2 \leq \delta\sqrt{\frac{3n}{2}}$, which is 1/2 if $\delta = \sqrt{\frac{1}{6n}}$. So the minimax lower bound is $1/24n$.

3.1.2

Assume we have $\theta_1, \cdots, \theta_n$, $d(\theta_i, \theta_j) \geq 2\delta, \forall i \neq j$, and let $\bar{P} = \frac{1}{m}\sum_{i=1}^m P_{\theta_i}$. By La cam,

$$\inf_{\hat{\theta}} \sup_{\theta} \mathbb{E}_\theta(d(\hat{\theta}, \theta)) \geq \frac{\delta}{2}\left(1 - d_{TV}(P_\theta, \bar{P})\right).$$

Use above in the following problem, $y_i = \theta_i + \epsilon_i/\sqrt{n}$, where $\epsilon \sim N(0,1)$ and $1 \leq i \leq p$. Then $P_{\theta_0} = N(0, I_p/\sqrt{n})$ and $P_{\theta_i} = N(\theta_i), I_p/\sqrt{n}$, where $\theta_i \in \mathbb{R}^p$, all zeros except for the i-th coordinate, which is equal to $\delta = \sqrt{\frac{a \log p}{n}}$, where $0 < a < 1$.

Let f_i be density of P_{θ_i} and let's look at χ^2 divergence between P_{θ_0} and $\bar{P} = \frac{1}{m}\sum_{i=1}^m P_{\theta_i}$. Then

$$\int \frac{\frac{1}{p}\sum_{i=1}^p (f_i - f_0)^2}{f_0} dx = \int \frac{\frac{1}{p}\sum_{i=1}^p f_i^2}{f_0} dx - 1 = \frac{1}{p^2}\sum_{i,j}\left(\frac{f_i f_j}{f_0} dx - 1\right) = \frac{1}{p^2}\sum_{i=1}^p \left(\frac{f_i^2}{f_0} dx - 1\right) = \frac{1}{p} e^{a \log p} - \frac{1}{p} \to 0,$$

as $p = \to \infty$.

So $\exists c > 0$ such that $1 - d_{TV}(P_0, \bar{P}) \geq c > 0$. Now we have $\delta = \sqrt{\frac{a \log p}{n}}$, then the lower bound is up to constants.

3.2 Fano Method

- Very popular to get minimax rates in high dimensions
- Choose $P_0, \cdots, P_m$ such that $d\left(\theta(P_i), \theta(P_j)\right) \geq 2\delta, \forall i \neq j$. We write $\theta_i = \theta(P_i)$.

Let V is sampled from $unifor(m\{0, \cdots, m\})$, and $\hat{V} = \phi^*(x)$, where ϕ^* is the minimum distance test $\phi^*(x) = \arg\min d(\hat{\theta}, \theta_j)$. Then $\hat{V} = j$ if $d(\hat{\theta}, \theta_j) \leq \delta$, so

$$\begin{aligned} \max \mathbb{E}_{\theta_j}\left(w(d(\hat{\theta}, \theta_j))\right) &\geq w(\delta) \max_j P_{\theta_j}(d(\hat{\theta}, \theta_j) > \delta) \\ &\geq \tfrac{w(\delta)}{m+1} \textstyle\sum_{j=0}^m P(d(\hat{\theta}, \theta_j) > \delta | V = j) \\ &\geq \tfrac{w(\delta)}{m+1} \textstyle\sum_{j=0}^m P(\hat{V} \neq V | V = j) \\ &\geq w(\delta) P(\hat{V} \neq V) \end{aligned}$$

By Fano inequality, $P(V \neq \hat{V}) \geq 1 - \frac{I(V;X) + \log 2}{\log(m+1)}$, where $I(V;X) = KL(P_{(X,V), P_X \times P_V})$ is the mutual information between X and V. So a minimax lower bound is $w(\delta)\left(1 - \frac{I(V;X)+\log 2}{\log(m+1)}\right)$. All we have to do is find $P_0, \cdots, P_m$ and compute $I(X;V)$. In information theoretical setting, $V \to X \to \hat{V}$, which is a Markov chain.

Theorem 3.1 *(Fano inequality)*

Let $P_e = P(V \neq \hat{V})$, then we have

$$h(P_e) + P_e \log(m+1) \geq H(V) - I(V;X),$$

where h is the entropy and $H(V) = -\sum_{j=0}^m P(V_j) \log P(V_j)$.

For us, $H(V) = \log(m+1)$, then

$$P_e = P(V \neq \hat{V}) \geq \frac{\log(m+1) - I(V;X) - h(P_e)}{\log m} \geq \frac{\log(m+1) - I(V;X) - h(1/2)}{\log(m+1)} = 1 - \frac{I(X;V) + \log 2}{\log(m+1)}.$$

So we need to ensure that $1 - \frac{I(X;V)+\log 2}{\log(m+1)} \geq c > 0$ for some c, which means we need to upper bound $I(V;X)$. Recall that $X|V_j \sim P_j$. If V has probability $\pi(0), \cdots, \pi(m)$, where $P(V = i) = \pi_i$, then $I(V;X) = \sum_{j=0}^m \pi(j) KL(P_0 | \bar{P})$, where $\bar{P} = \sum_{j=0}^m \pi(j) P_j$. In our case, $\pi(j) = \frac{1}{m+1}$, so $I(V;X) = \frac{1}{m+1} \sum_{i=0}^m KL(P_i | \frac{1}{m+1} \sum_{i=0}^m P_i)$.

By concavity of log, $I(V;X) \leq \frac{1}{(m+1)^2} \sum_{i,j} KL(P_i, P_j)$. If $\max KL(P_i, P_j) \leq \beta(M, \delta)$, then $I(V;X) \leq \beta(\delta, m)$. So a minimax lower bound is $w(\delta)\left(1 - \frac{\beta(M,\delta)+\log 2}{\log(m+1)}\right)$.