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

Lecture 6: November 12

Lecturer: Alessandro Rinaldo *Scribe: Veeranjaneyulu Sadhanala*

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

Tsybakov's master thoerem for minimax bounds

Theorem 6.1 (Theorem 2.5 in Tsybakov's book) *Let $M \geq 2$ and $\theta_0, \theta_1, \cdots, \theta_M \in \Theta$ be such that*

(i) $d(\theta_i, \theta_j) \geq 2\delta$ *for all* $0 \leq i < j \leq M$

(ii) $P_i \ll P_0$ *for all* $i = 1, 2, \cdots M$ *and*

(iii) For an $\alpha \in (0, 1/8)$,

$$\frac{1}{M}\sum_{i=1}^{M} \mathrm{KL}(P_i, P_0) \leq \alpha \log M$$

Then

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\Big[w(d(\hat{\theta}, \theta(P)))\Big] \geq w(\delta) C(\alpha)$$

where

$$C(\alpha) = \frac{\sqrt{M}}{1+\sqrt{M}}\Big(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\Big)$$

We apply this theorem to obtain minimax lower bound in L_2 loss for nonparametric regression.

Let $Y_i = f(X_i) + \epsilon_i$ for $i = 1, 2, \cdots, n$ where $\epsilon_i \sim N(0, \sigma^2)$.

Assumption: Let p_ϵ be a density function. There exist $p^*, v_0 > 0$ such that

$$\int p_\epsilon(\mu) \log \frac{p_\epsilon(\mu)}{p_\epsilon(\mu + v)} d\mu \leq p^* v^2 \quad \text{ if } |\nu| \leq v_0. \tag{6.1}$$

In other words, the KL divergence between p_ϵ and its translated versions is bounded in terms of the amount of translation. Note that if p_ϵ is Gaussian, the bound is satisfied for all v and $p^* = 1/2$.

For $\beta, L > 0$, define the Holder class of functions

$$\Sigma(\beta, L) = \Big\{ f : [0,1] \to \mathbb{R} \mid \forall x, y \in [0,1],\ \big|f^{(\rho)}(x) - f^{(\rho)}(y)\big| \leq L|x-y|^{\beta - \rho} \Big\}$$

where $\rho = \lfloor \beta \rfloor$, the smallest integer strictly less than β.

We want to find

$$\inf_{\hat{f}} \sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f\left[\|f-\hat{f}\|_2\right]$$

where $\|g\|_2 = \int g^2(x)dx$. It can be shown that there exists $\hat{f}$ such that

$$\sup_{f\in\Sigma(\beta,L)} \mathbb{E}_f\left[\|f-\hat{f}\|_2\right] \asymp n^{-\beta/(2\beta+1)}.$$

As $\beta \to \infty$, the bound goes to $n^{-1/2}$ which is the parametric rate. Now we lower bound this rate using Theorem 6.1.

Proof: Let c_0 be a constant chosen later. Partition $[0,1]$ into $m = \lceil c_0 n^{1/(2\beta+1)} \rceil$ intervals of width $1/m$ and let $x_k = (k-1/2)/m$ for $k = 1,\cdots,m$ be the mid-points of those intervals. Also let $\Delta_k = (\frac{k-1}{m}, \frac{k}{m}]$ for $k = 1,\cdots,m$. Define the blip on the kth interval

$$\psi_k(x) = Lh^\beta K\Big(\frac{x-x_k}{h}\Big)$$

where the kernel $K \in \Sigma(\beta, L/2) \cap C^\infty, \text{supp}(K) = (-1/2, 1/2)$ and $K > 0$.

For example, $K(\mu)$ can be $aK_0(2\mu)$ where $K_0(z) = \exp\left(\frac{-1}{1-z^2}\mathbb{1}\{|z|<1\}\right)$ and $a > 0$.

By construction, ψ_k's have non-overlapping support.

Next let $\Omega = \{0,1\}^m$ and for any $\omega \in \Omega$, let ω_j denote the jth component of ω where $1 \le j \le m$.

Denote $f_\omega(x) = \sum_{k=1}^m w_k\psi_k(x)$. Then for any distinct $\omega, \omega' \in \Omega$,

$$\begin{aligned}\|f_\omega - f_{\omega'}\|_2^2 &= \sum_{k=1}^m (\omega_k - \omega_k')^2 \int_{\Delta_k} \psi_k^2(x)\,dx \\ &= d_H(\omega,\omega') \int_{\Delta_1} \psi_1^2(x)\,dx \\ &= d_H(\omega,\omega') L^2 h^{2\beta+1} \|K\|_2^2\end{aligned}$$

where $d_H(\omega,\omega') = \sum_{k=1}^m \mathbb{1}\{\omega_k \ne \omega_k'\}$ denotes the Hamming distance.

To show the lower bound using Theorem 6.1, it is sufficient to have a subset Ω' of Ω such that for all distinct $\omega, \omega' \in \Omega'$,

$$d_H(\omega,\omega') \gtrsim 1/h = m \quad \text{so that}$$
$$\|f_\omega - f_{\omega'}\|_2 \ge 2\delta_n \asymp n^{-\frac{-\beta}{2\beta+1}}$$

while still satsifying *(iii)* of Theorem 6.1. The following result gives such a subset Ω'.

Lemma 6.2 (Varshamov-Gilbert) *Let $m \ge 8$. There exists a subset $\{\omega^{(0)},\cdots,\omega^{(M)}\}$ of Ω with $M \ge 2^{m/8}$ such that $\omega^{(0)} = (0,0,\cdots,0)$ and*

$$d_H(\omega^{(i)},\omega^{(j)}) \ge m/8,$$

for $0 \le i \le j \le M$.

Continuing the proof of the lower bound, let $f_j = f_{\omega^{(j)}}$ for $j = 0, 1, \cdots, M$ where $\omega^{(j)}$ are chosen as in Varshamov-Gilbert's lemma above. To apply Theorem 6.1, we verify the hypothesis of the theorem.

It can be shown that $f_j \in \Sigma(\beta, L)$ from the fact that $\psi_k \in \Sigma(\beta, L/2)$.

To show *(i)*, that is, that the functions are apart, recall that we have

$$\|f_i - f_j\|_2^2 = L^2 h^{2\beta+1} \|K\|_2^2 d_H(\omega^{(i)}, \omega^{(j)}) \geq L^2 h^{2\beta+1} \|K\|_2^2 \frac{m}{8} = \frac{1}{8} L^2 m^{-2\beta} \|K\|_2^2$$

where we used $h = 1/m$ for the last equality. Recalling that $m = \lceil c_0 n^{1/(2\beta+1)} \rceil$, for $m \geq 8$ and sufficiently larger n, we have

$$\|f_i - f_j\|_2 \geq \frac{L\|K\|_2}{4} (2c_0)^{-\beta} \, n^{-\frac{\beta}{2\beta+1}}$$

Now we want to show that

$$\frac{1}{M} \sum_{j=1}^{M} \mathrm{KL}(P_j, P_0) \leq \alpha \log M$$

where P_j is the distribution of $Y_1, \cdots, Y_n$ under f_j for $j = 0, 1, \cdots, M$. P_j has Lebesgue density

$$(y_1, y_2, \cdots, y_n) \to \Pi_{i=1}^n p_\epsilon(y_i - f(X_i)$$

where p_ϵ is the distribution of the noise term. The KL divergence can be upper bounded as follows:

$$\begin{aligned}
\mathrm{KL}(P_j, P_0) &= \int_{\mathbb{R}^n} \log \Pi_{i=1}^n \frac{p_\epsilon(y_i - f_j(X_i))}{p_\epsilon(y_i)} \Pi_{i=1}^n p_\epsilon(y_i - f_j(X_i))\, dy_1 \cdots dy_n \\
&= \sum_{i=1}^n \int_{\mathbb{R}} p_\epsilon(y - f_j(X_i)) \log \frac{p_\epsilon(y - f_j(X_i))}{p_\epsilon(y)}\, dy \\
&\leq p^* \sum_{i=1}^n f_j^2(X_i) \qquad \text{by the assumption 6.1} \\
&\leq p^* \sum_{k=1}^m \sum_{i: X_i \in \Delta_k} \psi_k^2(X_i) \\
&\leq p^* L^2 K_{\max}^2 h^{2\beta} \sum_{k=1}^m \Big| \{i : X_i \in \Delta_k]\} \Big| \qquad \text{where } K_{\max} = \sup_\mu K(\mu) \\
&= p^* L^2 K_{\max}^2 h^{2\beta} n \\
&\leq p^* L^2 K_{\max}^2 c_0^{-(2\beta+1)} m
\end{aligned}$$

Observe that $m \leq 8 \log_2 M$ and choose

$$c_0 = \Big(\frac{8p^* L^2 K_{\max}^2}{\alpha \log 2} \Big)^{1/(2\beta+1)}$$

so that we have the desired bound

$$\mathrm{KL}(P_j, P_0) \leq \alpha \log M.$$

Thus we have verified the conditions required for theorem 6.1 to hold. Therefore,

$$\max_{f \in \{f_0, \cdots, f_M\}} \mathbb{P}_f \Big(\|\hat{f} - f\|_2 \geq A n^{\frac{-\beta}{2\beta+1}} \Big) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \Big(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}} \Big)$$

■

Note that if we use L_∞ norm instead of the L_2 norm, then the minimax rate is only slightly worse: $(\log n/n)^{\beta/(2\beta+1)}$.

Assouad's Method

It consists of many binary hypothesis testing problems in contrast to the previous methods which deal with a multiple hypothesis test in general. The method is not always applicable, but worth trying after the methods we discussed previously in the class.

Let S_m denote the hypercube $\{-1,1\}^m$ for positive integers m.

Assumption: There exists a sub-family $\Big\{P_v, v \in S_m\Big\} \subset \mathcal{P}$ and a function $v : \theta(\mathcal{P}) \to S_m$ such that $\forall v, v' \in S_m$,

$$w(d(\theta(P_v), \theta(P_{v'}))) \geq 2\delta d_H(v, v').$$

Think of the function v as something which maps θ to the closest corner of the hypercube S_m.

Let $v \in \text{Uniform}(S_m)$ and $\mathbb{P}_{\pm j}$ be the conditional distribution of (X, v) given $v_j = \pm 1$, then

$$\begin{aligned}
\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}[w(d(\hat{\theta}, \theta(P)))] &\geq 2\delta \frac{1}{2^m} \sum_{v \in S_m} \mathbb{E}_{P_v}[d_H(v(\hat{\theta}), v)] \\
&\geq \delta \sum_{j=1}^{m} \Big(1 - d_{\text{TV}}(\mathbb{P}_{+j}, \mathbb{P}_{-j})\Big) \\
&\geq m\delta \min_{v, v' \in S_m, d_H(v, v')=1} \Big(1 - d_{\text{TV}}(P_v, P_{v'})\Big)
\end{aligned}$$

We continue the discussion on Assouad's method in the next lecture.