10-704: Information Processing and Learning **Spring 2015**

Lecture 21: Examples of Lower Bounds and Assouad's Method

Lecturer: Akshay Krishnamurthy *Scribes: Soumya Batra*

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

21.1 Brief Review

Last lecture, we saw that for a parameter space Θ and $\{\theta_j\}_{j=1}^M$ as a 2δ packing in ρ-metric, then using Fano's method we can lower bound the Minimax risk as:

$$R(\Theta) = \inf_T \sup_{\theta \in \Theta} \mathbb{E}_{X_1^n \sim P_\theta} \left[\Phi \circ \rho\left(T, \theta\right)\right] \geq \Phi(\delta) \left[1 - \frac{I\left(\theta; X_1^n\right) + \log 2}{h\left(\pi\right)}\right] \tag{21.1}$$

where π is a prior on $\{\theta_j\}_{j=1}^M$, $T : \mathcal{X}^n \to \Theta$ is an estimator of Θ and $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-decreasing risk function between the true and estimated parameter distribution with $\Phi(0) = 0$ such as the squared function $\Phi(t) = t^2$.

We also have an upper bound to the mutual information term as:

$$I(\theta, X_1^n) \leq \frac{1}{M} \sum_{j=1}^M \pi_j \mathrm{KL}(P_{\theta_j} \| P_{\theta_1}) \tag{21.2}$$

where $\mathrm{KL}(P_{\theta_j} \| P_{\theta_1})$ is the Kullback-Leibler divergence between the distributions induced by the parameter θ_j and θ_1. This upper bound holds for all choices of "centering" parameter θ_1 here as long as it is in the packing.

In this lecture note, we will see two examples where this lower bound is used.

21.2 Example 1 : Non-parametric regression in Mean Integrated Square Error Density Estimation (L_2^2 risk)

21.2.1 Problem Setup

We are given data $\{(X_i, Y_i)\}_{i=1}^n$ such that:

1. $X_i \sim \mathrm{Unif}\left([0, 1]\right)$
2. $Y_i = f(X_i) + \epsilon_i$
3. $f : [0, 1] \to \mathbb{R}$ is a regression function, required to be Lipschitz with constant L. i.e.,

$$\forall\ x, x' \in [0, 1] :\ |f(x) - f(x')| \leq L|x - x'|$$

The above equation gives a notion of smoothness for f. Let $\Theta(1, L)$ denote the set of all L-Lipschitz functions from $[0, 1] \to \mathbb{R}$.

4. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Theorem 21.1 *For the nonparametric regression problem above, the minimax rate is $n^{-\frac{2}{3}}$:*

$$\inf_{\hat{f}} \sup_{f \in \Theta(1,L)} \mathbb{E}[\|\hat{f} - f\|_2^2] \asymp n^{-\frac{2}{3}}$$

Note that $\mathbb{E}[\|\hat{f} - f\|_2^2]$ can be written as:

$$\begin{aligned}\mathbb{E}[\|\hat{f} - f\|_2^2] &= \int_0^1 \left(\hat{f}(x) - f(x)\right)^2 dP(x) \\ &= \int_0^1 \left(\hat{f}(x) - f(x)\right)^2 dx \qquad (P(x) \text{ is uniform})\end{aligned}$$

Proof:

21.2.1.1 Upper Bound on MISE

Upper bound on the mean squared integrated error for Kernel density estimators with Lipschitz continuous regression functions f being in $\Theta(\beta, L) = O(n^{-\frac{2\beta}{2\beta+d}})$ as $n \to \infty$. For our case, $\beta = 1$ and number of dimensions $d = 1$. Hence, we have an upper bound of $O(n^{-\frac{2}{3}})$ on our MISE.

21.2.1.2 Lower Bound on MISE

Construction of Hypothesis

Intutively, we begin by constructing a set of hypothesis in our parameter space Θ. We start with hypothesis function $f_0 = 0,\ \forall x \in [0, 1]$. Next, we partition the domain $[0, 1]$ into m equal intervals of size $\frac{1}{m}$ each for some $m > 0$. Within each interval, we place a Lipschitz bump either pointing up or pointing down. From Figure 21.2.1.2), we can see that the height of each bump is $\frac{L}{2m}$. Now, we will parameterize each hypothesis function by a $\{0, 1\}^m$ sign vector indicating which direction the bumps point in (We assume 0 when bumps point downwards).

Formally, we define the following:

$$\varphi(x) = \frac{L}{2m} K\left((x - x_k)\, 2m\right),\ \forall\ k = 1 \text{ to } m,\ x_k \quad = \frac{k - \frac{1}{2}}{m},$$

$$K(x) = \begin{cases} 1 - x, & \text{if } x \in [0, 1] \\ x + 1, & \text{if } x \in [0, 1] \\ 0, & \text{otherwise} \end{cases}$$

where $K(x)$ is our bump function.

Using φ, we define the subset of the hypothesis space as the collection of functions:

$$\Omega = \left\{ f_\omega(x) = \sum_{k=1}^{m} \omega_k \varphi_k(x),\ \omega_k \in \{0, 1\} \right\}$$

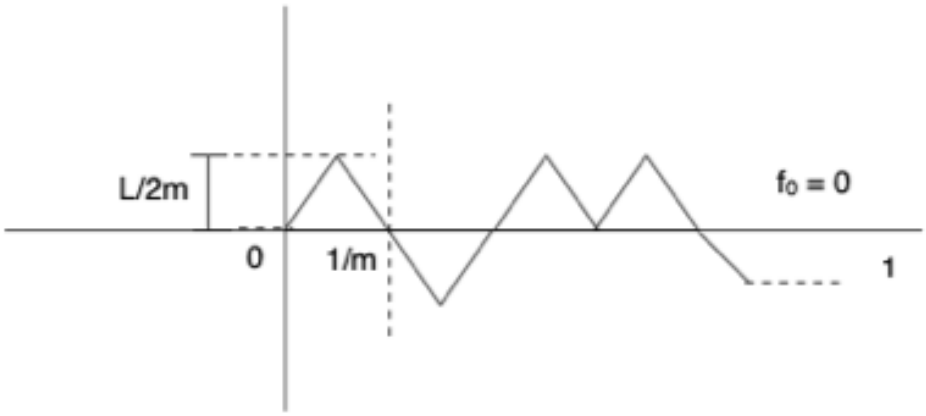

Figure 21.1: Set of hypothesis

Each f_ω has the Lipschitz bump in the bins for which ω is non-zero. It is flat in the bins where ω is zero.

L_2^2 distance between the hypothesis: L_2^2 distance between any 2 hypothesis f_ω and f'_ω can now be written as:

$$\|f_\omega - f'_\omega\|_2^2 = \int_0^1 (f_\omega(x) - f'_\omega(x))^2 = \sum_{k=1}^{m} (\omega_k - \omega'_k)^2 \int_{\Delta_k} \varphi_k^2(x) = \frac{cL^2}{m^3} \|\omega - \omega'\|_1$$

where Δ_k is the kth Lipschitz bin from our bump function and $\|\omega - \omega'\|_1$ is the Hamming distance between ω and ω' (number of disagreements in the character positions of the strings ω and ω'). This intuitively makes sense too as we pay the price of penalty whenever the bumps disagree.

Now, the KL divergence of the corresponding probability distribution function for each of these hypothesis functions with that of the hypothesis $f_0 = 0$ is given as:

$$\begin{aligned}
KL(P_\omega \| P_0) &= \int P_\omega(x,y) \log \frac{P_\omega(x,y)}{P_0(x,y)} = \int_x P(x) P_\omega(y|x) \log \frac{P(x) P_\omega(y|x)}{P(x) P_0(y|x} \\
&= \int_x P(x) \int P_\omega(y|x) \log \frac{P_\omega(y|x)}{P_0(y|x)} = \mathbb{E}_x \left[KL(\mathcal{N}(f_\omega(x), \sigma^2) \| \mathcal{N}(0, \sigma^2)) \right] \\
&= \mathbb{E}_x \left[\frac{1}{2\sigma^2} \underbrace{(f_w(x) - f_0(x))}_{f_0(x)=0}^2 \right] = \mathbb{E}_x \left[\frac{1}{2\sigma^2} f_w(x)^2 \right] \\
&= \frac{1}{2\sigma^2} \sum_{k=1}^{m} \omega_k \int_{\Delta_k} \varphi_k^2(x) = \frac{cL^2}{\sigma^2 m^3} \|\omega\|_1
\end{aligned}$$

From the above, we have the following facts that we will use to prove the lower bound:

- **Fact 1** : $\|f_\omega - f'_\omega\|^2 = \frac{cL^2}{m^3} \|\omega - \omega'\|_1$

- **Fact 2** : $KL(P_\omega \| P_0) = \frac{cL^2}{\sigma^2 m^3}\|\omega\|_1$
- **Fact 3** : $\forall \omega,\ f_\omega \in \Theta\,(1, L)$

Now we will create a 2δ packing in the Hamming metric. For that, we will use the following theorem:

Theorem 21.2 (Varshamov-Gilbert bound) *Let $m \geq 8$. Then there exists a subset $\{\omega_0, \ldots, \omega_M\}$ of $\{0,1\}^M$ such that $w_0 = (0, \ldots, 0)$ and*

$$\forall i \neq j,\ \|\omega_i - \omega_j\|_1 \geq \frac{m}{8} \qquad \text{and} \qquad M \geq 2^{\frac{m}{8}} \tag{21.3}$$

The above is a classic result in Coding Theory. It is a packing construction for the hypercube in the Hamming metric, precisely what we need for our problem.

Now we use this bound to create our 2δ packing.

Creating a 2δ packing in the Hamming metric

We reduce our hypothesis set Ω using Equation 21.3 of the Varshamov-Gilbert bound, we get:

- **Fact 1**

$$\|f_\omega - f'_\omega\|^2 = \frac{cL^2}{m^3}\|\omega - \omega'\|_1 \geq \frac{cL^2}{m^3}\frac{m}{8} = \frac{CL^2}{m^2}, \qquad C = \frac{c}{8} > 0$$

- **Fact 2**

$$KL(P_\omega^n \| P_0^n) = \frac{cL^2 n}{\sigma^2 m^3}\|\omega\|_1 \leq \frac{cL^2 n}{\sigma^2 m^3}\frac{m}{8} = \frac{CL^2}{\sigma^2 m^2}n, \qquad C = \frac{c}{8} > 0$$

- **Fact 3**

$$\forall \omega,\ f_\omega \in \Theta\,(1, L)$$

Now, we need the following condition to be true in order to construct our packing:

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

Using Fact 2 and the Varshamov-Gilbert bound from above it suffices to have:

$$\begin{aligned} &\frac{cL^2 n}{\sigma^2 m^2} \leq \frac{\alpha m}{8}\log 2 \\ \implies &m \asymp n^{\frac{1}{3}} \end{aligned}$$

From Fact 1, we get that L_2^2 risk is $\Omega(\frac{1}{m^2})$. Combining this with the above result we get $L_2^2 = \Omega(n^{-\frac{2}{3}})$.

Formally,

$$\inf_{\hat{f}} \sup_{f} \mathbb{E}[\|\hat{f} - f\|_2^2] \geq \frac{cL^2}{m^2} \underbrace{\inf_{T} \sup_{j=1,\ldots,M} \mathcal{P}[T \neq j]}_{\Omega(1)} \geq \Omega(n^{-\frac{2}{3}})$$

Combining Lower and Upper bounds, we finally prove that

$$\inf_{\hat{f}} \sup_{f \in \Theta(1,L)} \mathbb{E}[\|\hat{f} - f\|_2^2] \asymp n^{-\frac{2}{3}}$$

■

Here are a few takeaways from this example that are applicable to other non-parametric lower bounds:

1. Construct hypothesis based on a lot of bumps.
2. Compute error metric between hypothesis pairs (it might be a good idea to use Varshamov-Gilbert bound so that they can all be big).
3. Compute KL / Hellinger / Total Variation distance depending on the problem and use that to set the number of bumps.
4. We can follow this procedure for a higher order smoothness and in higher dimensions as well.
5. This can also be used for density estimation problems as well by using Hellinger distance instead of KL-Divergence as KL is particularly nice for the Gaussian noise model, which does not hold for density estimation.

21.3 Example 2: Adaptive Compressive Sensing

21.3.1 Problem Setup

This is a sparse linear regression problem where we choose the covariate vectors sequentially and adaptively.

More formally, we consider an input space $\theta \in \mathbb{R}^d$. We get to choose sensing vectors $a_i, i \in [n]$ and observe:

$$y_i = a_i^T \theta + z_i$$

where $x_i \sim \mathcal{N}(0, \sigma^2)$ is a noise term. We operate under a power constraint $\|a_i\| \leq 1$. Also, we can choose vector a_{i+1} after observing $(y_1, a_1, \ldots, y_i, a_i)$ leading to the sequential and adaptive selection of the vectors.

Theorem 21.3 ([2]) *Let $0 <= k < \frac{d}{2}$, and n be an arbitrary number of measurements. Assume that θ is sampled i.i.d. such that $\theta_j = 0$ with probability $1 - \frac{k}{d}$ and $\theta_j = \mu$ with probability $\frac{k}{d}$, $j = \{1, \ldots, d\}$ (so as to have k non-zero entries on average). Then, for $\mu - \frac{4}{3}\sqrt{\frac{d}{n}}$, any sensing and recovery procedure satisfies the following for an estimate $\hat{\theta}$:*

$$\frac{1}{d}\mathbb{E}\|\hat{\theta} - \theta\|_2^2 \geq \frac{1}{7}\frac{k}{n}\sigma^2$$

Some remarks on the result are in order:

1. We can write the above equation as

$$\inf_{\substack{a_1,\ldots,a_n \\ \hat{\theta}}} \mathop{\mathbb{E}}_{\theta \sim \text{i.i.d.Bernoulli}} \mathbb{E}_z\|\hat{\theta} - \theta\|_2^2 \geq \frac{d}{7}\frac{k}{n}\sigma^2$$

   This is not exactly a minimax statement as we are considering only the prior probability distribution. However, with a very high probability θ has $O(k)$ number of non-zeros thus making it quite close. This i.i.d distribution decouples the problem and makes the analysis much easier.

2. Say we choose the vectors a_i to be random unit vectors and use the Lasso (or Dantzig Selector) to find estimate $\hat{\theta}$, i.e.,
$$\hat{\theta} = \min_{\theta} \|y - A_\theta\|^2 + \lambda\|\theta\|_1$$
where λ is the regularization parameter, we get
$$\mathbb{E}\|\hat{\theta} - \theta\|_2^2 \leq C\frac{kd}{n}\log(d)\sigma^2$$
This is an important result as this shows that a passive (or random) procedure is almost as good as the lower bound obtained by the adaptive procedure, if we ignore logarithmic factors

Proof: We make the following claim and use it to prove the theorem. We then prove this claim.

Claim 21.4 *Given that we sample θ according to the Bernoulli prior, let S be the support of θ. Then, we claim that any estimate $\hat{S}$ for the support satisfies*
$$\mathbb{E}|\hat{S}\Delta S| \geq k\left(1 - \frac{\mu}{2}\sqrt{\frac{n}{d}}\right)$$
where $\hat{S}\Delta S = (S|\hat{S}) \cup (\hat{S}|S)$ is the symmetric set difference.

21.3.1.1 Proof of Theorem using Claim

Let set $\hat{S} = \{j : |\hat{\theta_j}| \geq \frac{\mu}{2}\}$. This gives the following:
$$\|\hat{\theta} - \theta\|_2^2 = \sum_{j\in S}(\hat{x_j} - x_j)^2 + \sum_{j\notin S}\hat{x}_j^2 \geq \frac{\mu^2}{4}\mathbb{E}|\hat{S}\Delta S| + \frac{\mu^2}{4}\mathbb{E}|S\Delta\hat{S}| = \frac{\mu^2}{4}|\hat{S}\Delta S|$$
$$\implies \mathbb{E}\|\hat{x} - x\|_2^2 \geq \frac{\mu^2}{4}\mathbb{E}|\hat{S}\Delta S| \geq \frac{\mu^2}{4}k\left(1 - \frac{\mu}{2}\sqrt{\frac{m}{n}}\right) \qquad \text{by the claim.}$$

Putting $\mu = \frac{4}{3}\sqrt{\frac{d}{n}}$ and simplifying proves the theorem.

21.3.1.2 Proof of Claim

Let $\pi_1 = \frac{k}{d}$ and $\pi_0 = 1 - \pi_1$. For each coordinate $j \in \{1, \ldots, d\}$, set $P_{0,j}(\cdot) = P(\cdot|\theta_j = 0)$ and $P_{1,j}(\cdot) = P(\cdot|\theta_j \neq 0)$. Now, by Neyman-Pearson:
$$\inf_T \sup_{i\in\{0,1\}} \mathbb{P}[T \neq i] \geq \min\{\pi_0, \pi_1\}(1 - \|P_0 - P_1\|_{\text{TV}})$$

Thus, we have:
$$\mathbb{E}|S\Delta\hat{S}| = \sum_{j=1}^{d}\mathbb{P}(\hat{S}_j \neq S_j) \geq \pi_1\sum_{j=1}^{d}(1 - \|P_{0,j} - P_{1,j}\|_{\text{TV}}) = k\left(1 - \sum_{j=1}^{d}\frac{1}{d}\|P_{0,j} - P_{1,j}\|_{\text{TV}}\right)$$
$$\geq k\left(1 - \frac{1}{\sqrt{d}}\sqrt{\left(\sum_{j=1}^{d}\|P_{0,j} - P_{1,j}\|_{\text{TV}}^2\right)}\right) \qquad \text{by Cauchy-Schwartz inequality} \tag{21.4}$$

Now we will compute the Total Variation sum: $\sum_{j=1}^{d} \|P_{0,j} - P_{1,j}\|_{\text{TV}}^2$. For a fixed j and using Pinsker's inequality we get

$$\|P_0 - P_1\|_{\text{TV}}^2 \leq \frac{\pi_0}{2}\text{KL}(P_0\|P_1) + \frac{\pi_1}{2}\text{KL}(P_1\|P_0) \tag{21.5}$$

For a sequence of observations: $y^n = \{y_1, \ldots, y_n\}$,

$$P_0(y^n) = \sum_{\theta'} P(\theta')P(y^n|\theta_j = 0, \theta') = \sum_{\theta':\theta'_j=0} P(\theta')P_{0,\theta'}(y^n)$$

Similarly,

$$P_1(y^n) = \sum_{\theta':\theta'_j\neq 0} P(\theta')P_{1,\theta'}(y^n)$$

By convexity of KL-divergence,

$$\text{KL}(P_0\|P_1) \leq \sum_{\theta'} P(\theta')\text{KL}(P_{0,\theta'}\|P_{1,\theta'}) \tag{21.6}$$

Now once all other coordinates are fixed, we have

$$\begin{aligned} y_i &= a_i^T\theta + z_i \\ &= c_i + z_i && \text{under } P_{0,\theta'}, \text{ and} \\ y_i &= a_{ij}\mu + c_i + z_i && \text{under } P_{1,\theta'} \end{aligned}$$

Thus, the KL-divergence can now be written as

$$\text{KL}(P_{0,\theta'}, P_{1,\theta'}) = \sum_{i=1}^{n} \mathbb{E}_{0,\theta'}\left(\frac{1}{2}(y_i - \mu a_{ij} - c_i)^2 - \frac{1}{2}(y_i - c_i)^2\right) = \frac{\mu^2}{2}\sum_{i=1}^{n}\mathbb{E}_{0,\theta'}(a_{ij}^2)$$

Plugging this in Equation 21.6 we get

$$\text{KL}(P_0\|P_1) \leq \frac{\mu^2}{2}\sum_{i=1}^{n}\mathbb{E}\left[a_{ij}^2|\theta_j = 0\right]$$

and

$$\text{KL}(P_1\|P_0) \leq \frac{\mu^2}{2}\sum_{i=1}^{n}\mathbb{E}\left[a_{ij}^2|\theta_j = \mu\right]$$

Plugging these in Equation 21.5 we get

$$\|P_{0,j} - P_{1,j}\|_{\text{TV}}^2 \leq \frac{\mu^2}{4}\sum_{i=1}^{n}\mathbb{E}\left[a_{ij}^2\right]$$

Hence, total variation sum now becomes:

$$\sum_{j=1} d\,\|P_{0,j} - P_{1,j}\|_{\text{TV}}^2 \leq \frac{\mu^2}{4}\sum_{j=1}^{d}\sum_{i=1}^{n}\mathbb{E}\left[a_{ij}^2\right] = \frac{\mu^2}{4}\sum_{i=1}^{n}\mathbb{E}[\sum_{j=1}^{d} a_{ij}^2] = \frac{\mu^2}{4}n, \qquad \text{since } \|a_i\| = 1$$

Finally, plugging this back in 21.4 proves the claim. ■

21.3.2 Takeaways

Following are a few tips to solve problems such as this:

- Break up the problem into many smaller independent ones.
- Claim that to do well on any individual problem, you must devote a lot of energy to that problem. In this case, each independent coordinate j required a lot of energy (or samples) in order to do well.
- If a lot of energy is not available, then many of the problems must have a high error.

This above technique is known as **Assouad's method**. It is a more abstract way to solve the problem.

21.4 Assouad's Method

There are a few tools such as Le Cam's method, Fano's method and Assouad's method that are widely used to prove lower bounds. We already saw Le Cam's and Fano's method. Here we develop Assouad's method.

To use Assuoad's lemma, we require some structure on the parameter space Θ. We want a discretization of Θ indexed by the hypercube: $\mathcal{V} = \{-1, 1\}^d$. For each $v \in \mathcal{V}$, we have a parameter θ_v and we want

$$\Phi\left(\rho\left(\theta, \theta_v\right)\right) \geq 2\Phi(\delta) \sum_{j=1}^{d} \mathbb{1}\{\hat{v}(\theta)_j \neq v_j\}$$

where $\hat{v}(\cdot)$ is some function that maps θ to $\{-1, +1\}^d$. This implies that the error depends on how many coordinates were missed (a refined approach!).

Error metrics such as L_1 and L_2 distances are typically able to meet this condition. Eg.:

$$\|\theta - \theta_v\|_1 \geq \sum_{i=1}^{d} |\theta_j - \theta_{v,j}|$$

Proposition 21.5 *Consider $v \sim \text{Unif}(\mathcal{V}$ and let*

$$P_{+j}(\cdot) = \frac{1}{2^{d-1}} \sum_{v: v_j = +1} \mathcal{P}(\cdot) \qquad \text{and} \qquad P_{-j}(\cdot) = \frac{1}{2^{d-1}} \sum_{v: v_j = -1} \mathcal{P}(\cdot)$$

If we have decomposability, then

$$\begin{aligned} R(\Theta, \Phi \circ p) &\geq \delta \sum_{j=1}^{d} \inf_{\psi} \left[\mathcal{P}_{+j}\left[\psi(x) \neq +1\right] + \mathcal{P}_{-j}\left[\psi(x) \neq -1\right]\right] \\ &= \delta \sum_{j=1}^{d} \left[1 - \|P_{+j} - P_{-j}\|_{\text{TV}}\right] \\ &\geq \delta d (1 - \max_{v,j} \|P_{v,+j} - P_{v,-j}\|_{\text{TV}}) \end{aligned}$$

Where the notation $P_{v,+j}$ denotes the distribution induced by the vector v but with the $v_j = +1$ and $P_{v,-j}$ is the same distribution but with $v_j = -1$. The maximization is over vectors v and coordinates to flip the sign.

All of these are version of Assouad's method. The last one is a weakening but it is commonly used in applications.

21.4.0.1 Brief Comparison with Le Cam's and Fano's method

Following are the methodology in brief for each of these three methods:

- **Le Cam's method** : Discretize → Reduce to binary hypothesis test → apply Neyman-Pearson lemma
- **Fano's method** : Discretize → Reduce to multiple hypothesis test → apply Fano's inequality
- **Assuoad's method** : Structured discretization → Apply above proposition

References

[1] Alexandre B. Tsybakov, "Ch.2 - Tsybakov: Introduction to Nonparametric Estimation," *Springer*, 2009, pp. 77 –136.

[2] Ery Arias-Castro, Emmanuel J. Candès and Mark A. Davenport, "On the Fundamental Limits of Adaptive Sensing," *ArXiv e-prints*, 2011(revised 2012).

[3] John Duchi, "Lecture 3 - Assuoad's method," 2014.