Lecture 21: Minimax Theory

Akshay Krishnamurthy
akshay@cs.umass.edu

December 19, 2017

1 Recap

Last time we discussed how to prove lower bounds on the minimax risk

$$R_n(\Theta) = \inf_T \sup_{\theta \in \Theta} \mathbb{E}_\theta[\Phi \circ \rho(T(x_1^n), \theta)]$$

The main requirements here are that ρ is a metric and that $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-decreasing function with $\Phi(0) = 0$. T is the algorithm and x_1^n is the training data.

Last time we saw the recipe

1. Find a 2δ packing $\Theta' = \{\theta_j\}_{j=1}^M \subset \Theta$.
2. Reduce to a testing problem

$$R_n(\Theta) \geq \Phi(\delta) \inf_\Psi \sup_{j \in [M]} \mathbb{P}_j[\Psi \neq j]$$

3. Use a testing lower bound.

We also saw how to use the Neyman-Pearson lemma to derive a lower bound when $M = 2$, which is the simple versus simple hypothesis testing case. This is known as Le Cam's method

At the end of last class, I mentioned how Le Cam's method doesn't work well for multi-dimensional estimation problems. To see why, let's revisit the gaussian mean estimation problem where $\rho = \|\cdot\|_2$ and $\Phi(t) = t^2$, but now in higher dimension. For simplicity let us consider two hypotheses, where $H_0 : \mathcal{N}(0, I_d)$ and $H_1 = \mathcal{N}(2v, I_d)$ and we will optimize for v at the end. All of the calculations we used last time apply here

$$R_n \geq \|v\|^2 \inf_\Psi \sup_{\theta \in \{0,1\}} \mathbb{P}_\theta[\Psi(x_1^n) \neq \theta] \geq \|v\|^2 \left(\frac{1}{2} - \frac{1}{2}\sqrt{\frac{n}{2} KL(\mathcal{N}(0, I_d) || \mathcal{N}(2v, I_d))} \right).$$

Unfortunately the KL here is $2\|v\|^2$ so we get

$$\|v\|_2^2 \left(\frac{1}{2} - \frac{1}{2}\sqrt{n\|v\|_2^2} \right)$$

This is exactly the same optimization we had before and it leads to $\Omega(1/n)$ lower bound. However, the upper bound for the empirical mean is $O(d/n)$ so something is loose!

The problem is that when we consider just two hypotheses, the estimator in the infimum in some sense knows that we are only considering two hypotheses. So from its perspective, we are just in a one-dimensional problem. As a result, once we reduce to two hypotheses here, we will never require the effect of multiple dimensions!

2 Multiple hypotheses and Fano's method

The above recipe produces tight lower bounds for simple problems, typically in one dimension, but it does not work well in higher dimension. In high dimension, we cannot reduce to a simple versus simple testing problem with just two hypotheses. Instead we need to consider many alternatives. This requires a more information theoretic approach. For now, let's just assume that n is one. Nothing really changes when we see more samples.

The idea is to think about this as a channel decoding problem. The channel is $\Theta \to X$. The sender samples $\theta \in [M]$ from some distribution P and the channel corrupts this to $x \sim P_\theta$. The reciever, seeing x wants to decode the original message, which amounts to recovering θ. Information theory studies the decoding error rates for such problems and the key result for us is Fano's lemma. Before we get there we need some definitions:

Definition 1 (Entropy, conditional entropy)**.** *For a random variable Z with distribution/density $p(Z)$, we write $H(Z) = -\sum_z p(z)\log(p(z)$. For two random variables Y, Z the conditional entropy is*

$$H(Y|Z) = \sum_z p(z)H(Y|Z=z) = -\sum_z p(z)\sum_y p(y|z)\log p(y|z)$$

Entropy describes the average uncertainty of a random variable. Roughly, it corresponds to how many bits from $z \sim P$ do you have to tell me (on average) before I can figure out z. If P is uniform on $[M]$, then $H(Z) = \log(M)$, since you must tell me M bits before I can figure out the sample.

In a similar way, the conditional entropy is the average uncertainty of a random variable Y after observing Z. In other words, if you sample $(Y, Z) \sim P$ and then show me Z, how many more bits (on average) would you need to tell me before I know Z.

Definition 2 (KL Divergence)**.** *For two distributions p, q on the same probability space*

$$KL(p||q) = \sum_z p(z)\log(p(z)/q(z)) = \sum_z p(z)\log(1/q(z)) - H(p)$$

Lemma 3 (Fano)**.** *Consider a markov chain $\theta \to x \to T$ (where θ is also a random variable) and let $P_e = \mathbb{P}[T \neq \theta]$. Then for any T*

$$h(P_e) + P_e \log(|\Theta| - 1) \geq H(\Theta|X).$$

Here $h(\cdot)$ is the bernoulli entropy $h(p) = -p\log p - (1-p)\log(1-p)$ which is at most $\log(2)$ and $H(\Theta|X)$ is the conditional entropy.

Another way to state the inequality is (with $\Theta = \{\theta_j\}_{j=1}^M$).

$$\inf_T \mathbb{P}_{\theta\sim\mathrm{Unif}, x\sim P_\theta}[T(x) \neq \theta] \geq 1 - \frac{\mathbb{E}_{\theta\sim\mathrm{Unif}(\Theta)} KL(P_\theta||P_\pi) + \log 2}{\log|\Theta|} \geq 1 - \frac{\frac{1}{M^2}\sum_{i,j} KL(P_{\theta_i}||P_{\theta_j}) + \log 2}{\log M}$$

Once we are here, we no longer need any information theory. There are two proofs of Fano's inequality. One is more intuitive and algorithmic, which I like

Proof . The idea here is that if you had a good way to reconstruct Θ from X then you could compress the distribution over Θ substantially. Let $T = g(X)$ be an estimator that has low probability of error. From this I will design a compression scheme for the source Θ.

1. Sample $\theta \sim P$, Sample $X \sim P(\cdot|\theta)$, Compute $T = g(X)$
2. Check if $T = \theta$. If so, output $X, 0$.
3. Otherwise, output $X, 1, \theta$.

Not counting X the entropy of the remainder of the string is at most $H(p_e) + p_e \log(|\Theta| - 1)$. This follows from the fact that the first extra bit is a bernoulli, with parameter p_e according to the error probability of T. With probabilty $1 - p_e$ this extra bit is all there is, and with probability p_e we append θ which has entropy at most $\log(|\Theta| - 1)$ (since we know that $\theta \neq T(X)$, we already have counted one possibility). Since this string is enough to perfectly recover θ, it must explain all of the remaining uncertainty in θ after seeing x. Thus

$$H(\theta|x) \leq H(p_e) + p_e \log(|\Theta| - 1)$$

which is the desired inequality. □

Using the re-statement of Fano's inequality, we obtain the following theorem

Theorem 4. *Let Θ be a parameter space and suppose there exists $\{\theta_j\}_{j=1}^M$ such that*

1. *$\rho(\theta_i, \theta_j) \geq 2\delta$ for all $i \neq j$*
2. *$KL(P_{\theta_i}^{(n)} || P_{\theta_j}^{(n)}) \leq \alpha_n$ for all $i \neq j$.*

Then

$$R_n(\Theta) \geq \Phi(\delta) \left(1 - \frac{\alpha_n + \log 2}{\log M}\right)$$

In other words, if $\log(M) \geq 2(\alpha_n + \log 2)$ we obtain $R_n(\Theta) \geq \Phi(\delta)/2$.

Example 1 (Gaussian mean estimation in ℓ_2)**.** *Let $x_1^n \sim \mathcal{N}(v, I)$, $v \in \mathbb{R}^d$. And consider*

$$R_n = \inf_T \sup_{v \in \mathbb{R}^d} \mathbb{E}_{x_1^n \sim \mathcal{N}(v,I)} \|T(x_1^n) - v\|_2^2$$

which generalizes the 1-dimensional problem we were studying last time. We want to use Fano's inequality to get a sharper lower bound. Let U be a $1/2$ packing of the unit ball in the ℓ_2 metric. We saw (awhile ago) that the covering number of this set is $\geq 2^d$. Covering numbers and packings are closely related, and the packing number here is also $\geq 2^d$. So we have $|U| = \Omega(2^d)$. Now write $V = \{4\delta u \mid u \in U\}$ to be a scaled down version of this set. We have

$$\|v - v'\|_2 = 4\delta \|u - u'\| \geq \delta$$

so we have met the first property. As for the KL note that by the triangle inequality

$$\|v - v'\| \leq \|v\| + \|v'\| \leq 4\delta(\|u\| + \|u'\|) \leq 8\delta$$

since $\|u\| \leq 1$. This means that

$$KL(\mathcal{N}(v)||\mathcal{N}(v')) = \frac{\|v - v'\|_2^2}{2} \leq 32\delta^2$$

which means we can set $\alpha_n = 32n\delta^2$. To get a good lower bound, we need

$$\log(2^d) \geq 2(32n\delta^2 + \log 2) \Rightarrow \delta \leq \sqrt{\frac{d \log(2)/2 - \log 2}{32n}}$$

So we get an $\Omega(d/n)$ lower bound!

Assouad's method. There is one more technique that we will not cover, which can produce sharp lower bounds in some specialized cases. Rather than reduce to a single testing problem (with two or more hypotheses), we can reduce to several binary hypothesis testing problems, usually one for each region of the space. For example, in non-parametric problems, we typically partition the space into many bins and ask the estimator to solve a binary hypothesis testing problem in each bin. The idea is that if it cannot solve many of these testing problems simultaneously, then it must have high estimation error. This method is most commonly used in nonparametric estimation problems like density estimation, etc.

3 Modern examples

Lower bounds arise all over the place and there are still many avenues for research. Let me just mention a few

1. Bandits and partial feedback settings. In the last several years a number of papers have produced more and more refined lower bounds for various bandit problems. These lower bounds are quite interesting and can in some cases lead to new algorithms. For example, one lower bound in some sense prescribes a sampling distribution, and and algorithm that tries to track this distribution has very nice properties.

2. Last time I mentioned results on statistical property testing. Proving these lower bounds is actually quite challenging. Since the problem itself is a hypothesis testing problem with just two hypotheses, you cannot use Fano's method and must resort to Le Cam-type arguments. To get sharp lower bounds, you must mix over distributions in the null and alternative, which can be quite challenging.

3. Lower bounds on iteration complexity for optimization methods. The techniques I have described here are essentially the main ways to prove lower bounds on the number of iterations required for optimization algorithms, especially for stochastic optimization methods. The paradigm is essentially the same, we can think of the algorithm as receiving noisy gradient (say) sampled from an unknown function, and we are asking that the algorithm estimate the optimum.

4. Lower bounds with application-specific constraints. These techniques can also be used to understand techniques where data privacy, distributed computation, compressive sensing, or some other computational considerations are important. The main idea is to show that under these constraints, the information about the parameter is significantly reduce (in the KL sense). For example you can do this in a compressive sensing setup to obtain very precise rates for covariance estimation.