Lecture 21: Minimax Theory

Akshay Krishnamurthy
akshay@cs.umass.edu

November 28, 2017

1 Recap

In the first part of the course, we spent the majority of our time studying *risk minimization*. We found many ways to upper bound the risk of various algorithms (mostly the ERM but also things like boosting). A natural question you might ask is whether these algorithms have optimal performance, and to do this we need to prove lower bounds. This is what we'll see how to do today.

2 Minimax theory

Recall that typically we have been studying risk functionals that have the form

$$R(f, D) = \mathbb{E}_{z \sim D}[\ell(f, z)].$$

For example, in classification problems, we have $z = (x, y)$ where $x \in \mathcal{X}, y \in \{\pm 1\}$ and maybe $\ell(f, z) = \mathbf{1}\{f(x) \neq y\}$ is the 0/1 loss. Today we will stay in the abstract for most of the time. In our definition of PAC learning (and also in online learning) we had an algorithm $\mathcal{A}$ which takes a samples S of size n and produced a function or predictor f. The risk this algorithm is

$$R_n(\mathcal{A}, D) = \mathbb{E}_{z_1^n \sim D, z \sim D}[\ell(\mathcal{A}(z_1^n), z)].$$

In the PAC-guarantee, we required that the algorithm has low risk for all distributions maybe with some assumptions. We capture these assumptions by defining a set $\mathcal{D}$ of distributions and defining the maximum risk

$$R_n(\mathcal{A}, \mathcal{D}) = \sup_{D \in \mathcal{D}} \mathbb{E}_D[\ell(\mathcal{A}(z_1^n), z)].$$

Finally, we are interested in understand what is possible and what is not possible here, so we want to consider the best algorithm. This leads to the **minimax risk**:

$$R_n(\mathcal{D}) = \inf_{\mathcal{A}} \sup_{D \in \mathcal{D}} \mathbb{E}_D[\ell(\mathcal{A}(z_1^n), z)].$$

Just to scrutinize the definition a bit, let us consider some alternatives. Replacing supremum with a pointwise definition is not interesting, since by taking infimum over estimators we could just output the best function for the specific distribution. If we took an expectation over a bunch of distribution in $\mathcal{D}$ then we get a useful characterization, but it is not clear what expectation we should take. This characterization will be useful, and is a *Bayesian* characterization of optimality.

The main ingredients here are (1) a loss function, (2) distribution family $\mathcal{D}$ (implicitly instance domain).

Upper bounds on minimax risk. To derive an upper bound on the minimax risk, we typically build an algorithm. For example, we saw that with 0/1 loss and if $\mathcal{D}$ corresponds to the set of distributions over labeled examples where $y = h(x)$ for some $h \in \mathcal{H}$, then ERM over $\mathcal{H}$ guarantees that $R(\mathcal{D}) \leq O(\log(|\mathcal{H}|)/\epsilon)$. Thus, analyzing algorithms leads to upper bounds on the minimax risk. Today we are working to derive lower bounds.

3 Lower bound recipe

I am going to re-write the minimax risk in the following way

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

Here Θ parametrizes the distributions, $\rho : \Theta \times \Theta \to \mathbb{R}$ is a metric and $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-decreasing function with $\Phi(0) = 0$. T is the algorithm and x^n is the training data.

Example 1 (Binary classification)**.** *In binary classification with realizability, let $\mathcal{H}$ be a hypothesis class and let $\Theta = \mathcal{H}$. The data distribution corresponding to h is uniform on $\mathcal{X}$ with labels $y = h(x)$. We set $\rho(h', h) = \mathbb{E}_x \mathbf{1}\{h(x) \neq h'(x)\}$ to be the disagreement metric and set $\Phi(t) = t$. This reproduces the setting before but now we are only considering one marginal distribution.*

Example 2 (Gaussian mean estimation)**.** *The setting is more general and also captures more classical statistical estimation problems. Let $\Theta = \mathbb{R}^d$ where for parameter θ the data is $X_1^n \sim \mathcal{N}(\theta, I_d)$ in d-dimensions. Let $\rho(\theta, \theta') = \|\theta - \theta'\|_2$ and let $\Phi(t) = t^2$. This is gaussian mean estimation in d dimensions, in mean-squared error.*

The recipe is as follows

Step 1: Discretization. Fix a $\delta > 0$ and find a large set of parameters $\Theta'\{\theta_j\}_{j=1}^M \subset \Theta$ such that

$$\rho(\theta_i, \theta_j) \geq 2\delta \qquad \forall i \neq j$$

This is called a *packing*, which is closely related to *coverings* that we have seen before.

Step 2: Reduction to Testing. Consider $j \sim \text{Uniform}([M])$ and $X \sim P_{\theta_j}$. Now if you cannot differentiate θ_j from some other θ you will certainly make error $\Phi(\delta)$ in the estimation problem. More formally

Proposition 1. *Let $\{\theta_j\}_{j=1}^M$ be a 2δ-packing in the ρ-metric. Then*

$$R_n(\Theta) \geq \Phi(\delta) \inf_\Psi \sup_{j \in [M]} \mathbb{P}_{x_1^n \sim P_{\theta_j}} [\Psi(x^n) \neq j] \geq \Phi(\delta) \inf_\Psi \mathbb{P}_{j \sim Uniform([M]), x^n \sim P_{\theta_j}} [\Psi(x^n) \neq j]$$

Proof. Fix an estimator T. Then for any fixed θ we have

$$\mathbb{E}[\Phi(\rho(T, \theta))] \geq \mathbb{E}\Phi(\delta)\mathbf{1}\{\rho(T, \theta) \geq \delta\} \geq \Phi(\delta)\mathbb{P}[\rho(T, \theta) \geq \delta].$$

Now define the "tester" $\Psi(T) = \text{argmin}_j \, \rho(T, \theta_j)$. If $\rho(T, \theta_j) \leq \delta$ then $\Psi(T) = j$ by the 2δ separation of the packing.

$$\rho(T, \theta_k) \geq \rho(\theta_j, \theta_k) - \rho(T, \theta_j) > 2\delta - \delta = \delta.$$

The converse of this statement is that if $\Psi(T) \neq j$ then $\rho(T, \theta_j) \geq \delta$ which means that the probability of the former event is smaller than the probability of the latter.

$$\sup_{\theta \in \Theta} \mathbb{P}[\rho(T, \theta) \geq \delta] \geq \sup_{j = [M]} \mathbb{P}_j[\rho(T, \theta_j) \geq \delta] \geq \sup_{j \in [M]} \mathbb{P}_j[\Psi(T) \neq j]$$

Now take infimum over all T, which is in correspondence with Ψ. Of course supremum is larger the expectation which proves the second inequality. □

Step 3. Use a testing lower bound. The next step, is more involved and there are many strategies here. But we have reduced the estimation problem to a testing problem, and now we have to use a testing lower bound.

4 Simple vs Simple testing lower bound (Le Cam's method)

We have reduced our problem to lowering bounding the following "testing" risk

$$\inf_{\Psi} \mathbb{P}_{j \sim \text{Uniform}([M]), x^n \sim P_{\theta_j}} [\Psi(x^n) \neq j]$$

Let's first study the simplest possible approach, where $M = 2$. In this case we have

$$\inf_{\Psi} \frac{1}{2}\mathbb{P}_0[\Psi \neq 0] + \frac{1}{2}\mathbb{P}_1[\Psi \neq 1]$$

The key step here is the Neyman-Pearson lemma. We need to define the total variation distance.

Definition 2 (Total variation distance). *For two measures P, Q over the same probability space $\mathcal{X}$,*

$$\|P - Q\|_{TV} = \sup_{A \subset \mathcal{X}} P(A) - Q(A) = \frac{1}{2}\int |p(x) - q(x)| dx$$

where p, q are the densities.

Lemma 3. *For any distributions P_0, P_1 over a space $\mathcal{X}$*

$$\inf_{\Psi} \mathbb{P}_0[\Psi \neq 0] + \mathbb{P}_1[\Psi \neq 1] = 1 - \|P_0 - P_1\|_{TV}.$$

Proof. Any deterministic test $\Psi : \mathcal{X} \to \{0, 1\}$ has an acceptance region $A = \{x \in \mathcal{X} \mid \Psi(x) = 1\}$. Then

$$\mathbb{P}_0[\Psi \neq 0] + \mathbb{P}_1[\Psi \neq 1] = \mathbb{P}_0[A] + \mathbb{P}_1[A^C] = 1 - \mathbb{P}_1[A] + \mathbb{P}_0[A]$$

And so

$$\inf_{\Psi} \mathbb{P}_0[\Psi \neq 1] + \mathbb{P}_1[\Psi \neq 0] = \inf_{A} 1 - \mathbb{P}_1[A] + \mathbb{P}_0[A] = 1 - \sup_{A} \mathbb{P}_0[A] - \mathbb{P}_1[A] = 1 - \|P_0 - P_1\|_{TV}$$

Any randomized test is just a distribution over deterministic tests, and so the bound also applies. □

For us, using Neyman-Pearson, we have

$$\inf_{\Psi} \mathbb{P}_{j \sim \text{Uniform}(\{0,1\}), x^n \sim P_{\theta_j}} [\Psi(x^n) \neq j] = \frac{1}{2} - \frac{1}{2}\|P_0^n - P_1^n\|_{TV}$$

Before turning to the examples, we need one more result that we will not prove

Lemma 4 (Pinsker's inequality). *For any distributions P, Q:*

$$\|P - Q\|_{TV}^2 \leq \frac{1}{2} KL(P||Q)$$

Actually we saw Pinsker's inequality in passing when we were studying mirror descent. Pinsker's inequality is how you show that the negative entropy regularizer is strongly convex in the ℓ_1-norm, which is the Mirror-descent way to analyze the Hedge algorithm.

Fact 5. *$KL(P^n||Q^n) = nKL(P||Q)$ where P^n is the n-fold product measure.*

This leads to the following theorem for testing-based lower bounds

Theorem 6.

$$\inf_{\Psi} \sup_{\theta \in \{0,1\}} \mathbb{P}_{X_1^n \sim \theta}[T(X_1^n) \neq \theta] \geq \frac{1}{2} - \frac{1}{2}\sqrt{\frac{n}{2} KL(P_0||P_1)}.$$

Example 3. *Consider the gaussian mean estimation problem in one-dimension.*

$$R_n = \inf_T \sup_{\mu \in \mathbb{R}} \mathbb{E}_{X_1^n \sim \mathcal{N}(\mu,1)}[(T(X_1^n) - \mu)^2]$$

If we choose $P_0 = \mathcal{N}(-\mu, 1)$ and $P_1 = \mathcal{N}(\mu, 1)$ then the reduction to testing gives us

$$R_n \geq \mu^2 \times \inf_\Psi \sup_{\theta \in \{0,1\}} \mathbb{P}_\theta[\Psi(X_1^n) \neq \theta]$$

Applying the Neyman-Pearson lemma and the KL translation

$$R_n \geq \mu^2 \left(\frac{1}{2} - \frac{1}{2} \sqrt{\frac{n}{2} KL(\mathcal{N}(-\mu, 1) || \mathcal{N}(\mu, 1))} \right)$$

The KL divergence between d-dimensional gaussians is available on wikipedia

$$KL(\mathcal{N}(\mu_0, \Sigma_0) || \mathcal{N}(\mu_1, \Sigma_1)) = \frac{1}{2} \left[\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1} (\mu_1 - \mu_0) - d + \log \frac{\det(\Sigma_1)}{\det(\Sigma_0)} \right]$$

In our case the KL is just $\mu^2/2$. In total the lower bound is

$$R_n \geq \mu^2 \left(\frac{1}{2} - \frac{1}{2} \sqrt{n\mu^2} \right).$$

We want to maximize with respect to μ. If $\mu = \sqrt{1/(4n)}$ then we get

$$\frac{1}{4n} \left(\frac{1}{2} - \frac{1}{2} \sqrt{\frac{n}{4n}} \right) = \frac{1}{16n}$$

So we have proved an $\Omega(1/n)$ lower bound for one-dimensional gaussian mean estimation.

As a sanity check, this is optimal, since if we collect n samples and use the empirical mean, by a gaussian tail bound, we guarantee

$$\mathbb{P}[|\hat{\mu} - \mu| \geq \epsilon] \leq 2\exp(-2n\epsilon^2)$$

Or that with probability at least $1 - \delta$ $(\hat{\mu} - \mu)^2 \leq O(\log(1/\delta)/n)$. So the empirical mean is an estimator that is minimax optimal (at least up to logarithmic factors).

Some remarks on testing. While we have arrived at a testing problem by reduction from estimation, testing problems are interesting in their own right. In some sense hypothesis testing is at the heart of most of the physical sciences. They also appear all over in computer science and statistics. Some examples are:

1. *Testing in the sciences.* In most scientific applications, we are often interested in collecting data to verify or disprove some scientific hypothesis. Often these hypotheses are formulated as distributions. Mathematically this often amounts to testing:

   $$H_0 : X_1^n \sim P_0 \qquad \text{versus} \qquad H_1 : X_1^n \sim Q \neq P_0.$$

   In words, we are asking whether some data we collect comes from some known distribution P_0 or not. When the alternative hypothesis H_1 is more separated from P_0 then we can use Le Cam's method to prove a lower bound here. Some of the terminology here may be unfamiliar but at a high level any time you see a p-value anywhere, some test like this is happening.

2. *Statistical Property Testing.* A recent and very active line of research involves testing properties of distributions. This typically operates in the discrete case where the distributions are supported on the elements $[d] = \{1, \ldots, d\}$. One example here is *uniformity testing*:

   $$H_0 : X_1^n \sim \mathrm{Unif}([d]) \qquad \text{versus} \qquad H_1 : X_1^n \sim P \text{ where } \|P - \mathrm{Unif}\|_{TV} > \epsilon.$$

> Here we are asking if the data comes from a uniform distribution or from some distribution that is far from uniform (in Total-Variation). Again Le Cam's method can derive a lower bound here (but it is not sharp).
>
> This idea can be generalized by changing the null hypothesis to include a family of distributions. For example, one might be interested in testing if the distribution is uni-modal or far from uni-modal. The point is that these testing problems are pervasive in applications and mathematically interesting in their own right, in addition to being the main way to prove lower bounds for other statistical problems.

Sharper testing lower bounds. In the above examples, we often have *composite* hypotheses, which consist of a family of distributions. For example in the uniformity testing example, the alternative is not a single distribution, but rather that the data comes from some distribution that is far from uniform, which is a set of distributions. On the other hand, when we used Le Cam's method we only considered *simple* hypotheses, with just a single distribution. Sometimes using a single distribution from the alternative doesn't produce a sharp lower bound and instead you have to mix over many distributions from the alternative. The idea is that by mixing over many distributions, you can make the KL to the null much closer. However doing this is often quite challenging.

Recap. Let's summarize where we are. We want to prove lower bounds on the minimax risk, which gives us a lower bound on the sample complexity we can hope to achieve for many learning problems. Tthe approach we have seen so far is to discretize the parameter space by taking two carefully chosen parameters and observing that if any algorithm is to have low risk, it must be able to test between these two parameters. Then using the Neyman-Pearson lemma, we can relate the performance of any tester to a notion of distance (or divergence) between the data distributions induced by the parameters. Calculating the divergence leads to the lower bound.

5 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.

The idea is to think about this as a channel decoding problem. The channel is $\Theta \to X$. The sender chooses $\theta \in [M]$ 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.

Lemma 7 (Fano)**.** *Consider a markov chain $\Theta \to X \to T$ 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 $\Theta = \{\theta_j\}_{j=1}^M$).

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

Next time we will prove Fano's lemma (which has a really cool proof) and see how to use multiple hypotheses to derive lower bounds for high-dimensional problems.