IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lectures 14 & 15
Apr. 1, 2020

Goals: In this lecture, we motivate the study of the maxima of certain stochastic processes. The first motivation comes from the Kolmogorov-Smirnov test, and the second — from Statistical Learning. We present two approaches to analyzing the supremum of an empirical process: bracketing and symmetrization.

By now you have seen a number of finite-sample guarantees: estimation of a mean vector, matrix estimation, constrained and unconstrained linear regression. In all the examples, the key technical step was a control of the maximum of some collection of random variables. Over the next few lectures, we will extend the toolkit to arbitrary classes of functions and then apply it to questions of parametric and nonparametric estimation and statistical learning.

First, we present a couple of motivating examples.

1. KOLMOGOROV'S GOODNESS-OF-FIT TEST

Given n indepenent draws of a real-valued random variable X, you may want to ask whether it has a hypothesized distribution with cdf F_0. For instance, can you test the hypothesis that heights of people are $N(63, 3^2)$ (in inches)? Of course, we can try to see if the sample mean is "close" to the mean of the hypothesized distribution. We can also try the median, or some quantiles. In fact, we can try to compare all the quantiles at once and see if they match the quantiles of F_0. It turns out that comparing "all quantiles" is again a question about control of a maximum of a collection of correlated random variables. We will make this connection precise.

If you have taken a course on statistics, you might have seen several approaches to the hypothesis testing problem of whether X has a given distribution. One classical approach is the Kolmogorov-Smirnov test. Let

$$F(\theta) = P(X \le \theta)$$

be the cdf of X, and let

$$F_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \le \theta\}$$

be the empirical cdf obtained from n examples. The Glivenko-Cantelli Theorem (1933) states that

$$D_n = \sup_{\theta \in \mathbb{R}} |F_n(\theta) - F(\theta)| \to 0 \quad a.s.$$

Hence, given a candidate F, one can test whether X has distribution with cdf F, but for this we need to know the (asymptotic) distribution of D_n. Assuming continuity of F, Kolmogorov (1933) showed that the distribution of D_n does not depend on the law of X, and he calculated the asymptotic distribution (now known as the Kolmogorov distribution). Without going into details, we can observe that $F(X)$ has cdf of a uniform random variable

supported on $[0,1]$, and this transformation does not change the supremum. Hence, it is enough to calculate D_n for the uniform distribution on $[0,1]$. D_n fluctuates on the order of $1/\sqrt{n}$ and

$$\sqrt{n}D_n \longrightarrow \sup_{\theta\in\mathbb{R}} |B(F(\theta))|.$$

Here $B(x)$ is a Brownian bridge on $[0,1]$ (a continuous-time stochastic process with distribution being Wiener process conditioned on being pinned to 0 at the endpoints).

In particular, Kolmogorov in his 1933 paper calculates the asymptotic distribution, as well a table of a few values. For instance, he states that

$$P(D_n \le 2.4/\sqrt{n}) \longrightarrow \quad \text{approx} \quad 0.999973.$$

In the spirit of this course, we will take a non-asymptotic approach to this problem. While we might not obtain such sharp constants, the deviation inequalities will be valid for finite n.

We will now come to the same question of uniform deviations from a different angle – Statistical Learning Theory.

2. STATISTICAL LEARNING

2.1 Empirical Risk Minimization

Let $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be n i.i.d. copies of a random variable (X, Y) with distribution $P = P_X \times P_{Y|X}$, where the X variable lives in some abstract space $\mathcal{X}$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$. Fix a loss function $\boldsymbol{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

Fix a class of functions $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$. Given the dataset S, the empirical risk minimization (ERM) method is defined as

$$\widehat{f} \in \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\ell}(f(X_i), Y_i)$$

Examples:

- Linear regression: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, $\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathbb{R}^d\}$, $\boldsymbol{\ell}(a,b) = (a-b)^2$
- Linear classification: $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0,1\}$, $\mathcal{F} = \{x \mapsto (\text{sign}(\langle w, x\rangle) + 1)/2 : w \in \mathsf{B}_2\}$, $\boldsymbol{\ell}(a,b) = \mathbf{1}\{a \ne b\}$

We now define expected loss (error) as

$$\mathbf{L}(f) = \mathbb{E}_{(X,Y)}\boldsymbol{\ell}(f(X), Y)$$

and empirical loss (error) as

$$\widehat{\mathbf{L}}(f) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\ell}(f(X_i), Y_i)$$

For any $f^* \in \mathcal{F}$, The decomposition

$$\mathbf{L}(\widehat{f}) - \mathbf{L}(f^*) = \left[\mathbf{L}(\widehat{f}) - \widehat{\mathbf{L}}(\widehat{f})\right] + \left[\widehat{\mathbf{L}}(\widehat{f}) - \widehat{\mathbf{L}}(f^*)\right] + \left[\widehat{\mathbf{L}}(f^*) - \mathbf{L}(f^*)\right]$$

holds true. By definition of ERM, the second term is nonpositive. If f^* is independent of the random sample, the third term is a difference between an average of random variables $\boldsymbol{\ell}(f^*(X_i), Y_i)$ and their expectation. Hence, this term is zero-mean, and its fluctuations can be controlled with the tail bounds we have seen in class. The first term, however, is not zero in expectation (why?).

Let us proceed by taking expectation (with respect to S) of both sides:

$$\mathbb{E}\left[\mathbf{L}(\widehat{f})\right] - \mathbf{L}(f^*) \leq \mathbb{E}\left[\mathbf{L}(\widehat{f}) - \widehat{\mathbf{L}}(\widehat{f})\right] \leq \mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right] \tag{2.1}$$

Here we "removed the hat" on $\widehat{f}$ by "supping out" this data-dependent choice. We are only using the knowledge that $f \in \mathcal{F}$, and nothing else about the method. We will see later that for "curved" loss functions, such as square loss, the supremum can be further localized within $\mathcal{F}$.

2.2 Classification

We now specialize to the classification scenario with indicator loss $\boldsymbol{\ell}(a, b) = \mathbf{1}\{a \neq b\}$. Observe that $\mathbf{1}\{a \neq b\} = a + (1 - 2a)b$ for $a, b \in \{0, 1\}$. Hence, by taking $a = Y$ and $b = f(X)$,

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right] &= \mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}(Y + (1-2Y)f(X)) - \frac{1}{n}\sum_{i=1}^{n}(Y_i + (1-2Y_i)f(X_i))\right] \\ &= \mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}((1-2Y)f(X)) - \frac{1}{n}\sum_{i=1}^{n}(1-2Y_i)f(X_i)\right]\end{aligned}$$

Observe that $(1-2Y)$ is a random sign that is jointly distributed with X. Let us omit this random sign for a moment, and consider

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right]. \tag{2.2}$$

Over the next few lectures, we will develop upper bounds on the above expected supremum for any class $\mathcal{F}$. For now, let us gain a bit more intuition about this object by looking at a particular class of 1D thresholds:

$$\mathcal{F} = \{x \mapsto \mathbf{1}\{x \leq \theta\} : \theta \in \mathbb{R}\}.$$

Substituting this choice, (2.2) becomes

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[P(X \leq \theta) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq \theta\}\right] = \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[F(\theta) - F_n(\theta)\right]. \tag{2.3}$$

which is precisely the quantity from the beginning of the lecture (albeit without absolute values and in expectation). Again, (2.3) is the expected largest pointwise (and one-sided) distance between the CDF and empirical CDF. Does it go to zero as $n \to \infty$? How fast?

Let's introduce the shorthand

$$U_\theta = \mathbb{E}\mathbf{1}\{X \leq \theta\} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq \theta\}$$

$\{U_\theta\}_{\theta\in\mathbb{R}}$ is an uncountable collection of *correlated* random variables, so how does the maximum behave? We have already encountered the question (e.g. Lecture 5) in the context of linear forms $\langle X, \theta\rangle$, indexed by $\theta \in \mathsf{B}_2$ and we were able to use a covering argument to control the expected supremum. Recall the key step in that proof: we can introduce a cover $\theta_1, \ldots, \theta_N$ such that control of $\sup U_\theta$ can be reduced to control of $\max_{j=1,\ldots,N} U_{\theta_i}$. Does this idea work here? Problems with this approach start appearing immediately: how do we cover $\mathbb{R}$ by a finite collection?

We will now present two approaches for upper-bounding (2.3); both extend to the general case of (2.2).

2.2.1 The bracketing approach

While we cannot provide a finite ϵ-grid of $\mathbb{R}$ directly, we observe that we should be placing the covering elements according to the underlying measure P. Informally, U_θ is likely to be constant over regions of θ with small mass.

For simplicity assume that P does not have atoms, and let $\theta_1, \theta_1, \ldots, \theta_N$ (with $\theta_0 = -\infty, \theta_{N+1} = +\infty$) correspond to the quantiles: $P(\theta_i \le X \le \theta_{i+1}) = \frac{1}{N+1}$. For a given θ, let $u(\theta)$ and $\ell(\theta)$ denote, respectively, the upper and lower elements corresponding to the discrete collection $\theta_0, \ldots, \theta_{N+1}$. Then, trivially,

$$\begin{aligned}\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{X_i \le \theta\right\} &\le \mathbb{E}\mathbf{1}\left\{X \le u(\theta)\right\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{X_i \le \ell(\theta)\right\} \\ &\le \mathbb{E}\mathbf{1}\left\{X \le \ell(\theta)\right\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{X_i \le \ell(\theta)\right\} + \frac{1}{N+1}\end{aligned}$$

and thus

$$\begin{aligned}&\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{X_i \le \theta\right\}\right] \\ &\le \frac{1}{N+1} + \mathbb{E}\max_{j\in\{0,\ldots,N\}} \mathbb{E}\mathbf{1}\left\{X \le \theta_j\right\} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{X_i \le \theta_j\right\}\end{aligned}$$

Now, each random variable $\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\}$ is centered and $1/2$-subGaussian. Hence, for each j, U_{θ_j} is $\frac{1}{2\sqrt{n}}$-subGaussian, and the expected maximum is at most $\sqrt{\frac{2\log(N+1)}{2n}}$. The overall upper bound is then

$$\frac{1}{N+1} + \sqrt{\frac{\log(N+1)}{n}} = O\left(\sqrt{\frac{\log n}{n}}\right)$$

if we choose, for instance, $N = n$.

2.2.2 The symmetrization approach

An alternative is a powerful technique that replaces the expected value by a ghost sample. To motivate the technique, recall the following inequality for variance:

$$\mathbb{E}(X - \mathbb{E}X)^2 \le \mathbb{E}(X - X')^2 = 2\mathbb{E}(X - \mathbb{E}X)^2$$

where X' is an independent copy of X.

Observe that

$$\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i' \le \theta\right\}\right]$$

where $X_1', \ldots, X_n'$ are n independent copies of X. We have the following upper bound on (2.3):

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.4}$$

$$\le \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.5}$$

by convexity of the sup. Now, since distribution of $\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\}$ is the same as the distribution of $-\left(\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\}\right)$, we can insert arbitrary signs ϵ_i without changing the expected value:

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\})\right]. \tag{2.6}$$

Since the quantity is constant for all the choices of $\epsilon_1, \ldots, \epsilon_n$, we have the same value by taking an expectation. We have

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\mathbb{E}\mathbf{1}\left\{X \le \theta\right\} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.7}$$

$$\le \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(\mathbf{1}\left\{X_i' \le \theta\right\} - \mathbf{1}\left\{X_i \le \theta\right\})\right], \tag{2.8}$$

where ϵ_i's are now Rademacher random variables. Breaking up the supremum into two terms leads to an upper bound

$$\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i' \le \theta\right\}\right] + \mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}-\epsilon_i\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.9}$$

$$= 2\mathbb{E}\sup_{\theta\in\mathbb{R}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i \le \theta\right\}\right] \tag{2.10}$$

by symmetry of Rademacher random variables.

Now comes the key step. Let us condition on $X_1, \ldots, X_n$ and think of the random variables

$$V_\theta = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i \le \theta\right\}$$

as a function of the Rademacher random variables. How many truly distinct V_θ's do we have? Since $X_1, \ldots, X_n$ are now fixed, there are only at most $n+1$ choices (say, midpoints between datapoints), and so the last expression is

$$2\mathbb{E}\left[\mathbb{E}\left[\sup_{\theta\in\mathbb{R}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{1}\left\{X_i \le \theta\right\}\middle| X_{1:n}\right]\right] = 2\mathbb{E}\mathbb{E}\left[\max_{\theta\in\{\theta_1,\ldots,\theta_{n+1}\}}V_\theta\middle| X_{1:n}\right]$$

Since each V_θ is 1-subGaussian, and we get an overall upper bound

$$\sqrt{\frac{2\log(n+1)}{n}}$$

which, up to constants, matches the bound with the bracketing approach.

2.3 Discussion

The bracketing and symmetrization approaches produced similar upper bounds for the case of thresholds. We will see, however, that for more complex classes of functions, the two approaches can give different results.

In view of (2.1), the upper bounds we derived guarantee (modulo the fact that we omitted "$1-2Y$") that for empirical risk minimization,

$$\mathbb{E}\mathbf{L}(\widehat{f}) - \min_{f^* \in \mathcal{F}} \mathbf{L}(f^*) \lesssim \sqrt{\frac{\log(n+1)}{n}}$$

It is worth stating the symmetrization lemma more formally:

Lemma: Let $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$ be a class of real-valued functions. Let $X, X_1, \ldots, X_n$ be i.i.d. random variables with values in $\mathcal{X}$, and let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. Rademacher random variables. Then

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i)\right] \le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right].$$

Furthermore,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right] \le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^n f(X_i)\right| + \frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}|\mathbb{E}f|$$

Proof. We only prove the second part since the first statement was proved earlier (for indicators). Write

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right] \le \mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i) - \mathbb{E}f)\right] + \mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbb{E}f\right]$$

Consider the first term on the RHS:

$$\begin{aligned}
\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i) - \mathbb{E}f)\right] &\le \mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(X_i) - f(X_i'))\right] \\
&= \mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}f + \mathbb{E}f - f(X_i'))\right] \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n (\mathbb{E}f - f(X_i))\right] + \mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^n (f(X_i) - \mathbb{E}f)\right].
\end{aligned}$$

As for the second term,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbb{E}f\right] \leq \sup_{f\in\mathcal{F}}|\mathbb{E}f|\cdot\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right| \tag{2.11}$$

□

Of course, the symmetrization lemma can also be applied to the class of functions

$$\{(x,y)\mapsto(1-2y)f(x)\}.$$

Since $(1-2y)$ is $\{\pm 1\}$-valued, the distribution of $(1-2Y_i)\epsilon_i$ is also Rademacher. Hence,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(1-2Y_i)f(X_i)\right] = \mathbb{E}\sup_{f\in\mathcal{F}}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right].$$

This justifies omitting $(1-2Y)$ for binary classification, at least with the symmetrization approach.

2.4 Empirical Process

Let us also define an empirical process:

Definition: Let $\mathcal{F}=\{f:\mathcal{X}\to\mathbb{R}\}$ and $X, X_1,\ldots,X_n$ are i.i.d. The stochastic process

$$\nu_f = \mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n}f(X_i)$$

is called the *empirical process indexed by* $\mathcal{F}$.

We note that it is also customary to scale the empirical process as

$$\nu_f = \sqrt{n}\left(\mathbb{E}f(X) - \frac{1}{n}\sum_{i=1}^{n}f(X_i)\right)$$

Second, empirical process theory often employs the notation

$$\nu_f = \sqrt{n}(\mathbb{P}-\mathbb{P}_n)f$$

where $\mathbb{P}$ is the distribution of X and $\mathbb{P}_n$ is the empirical measure. You may also see the notation

$$\mathbb{E}\sup_{f\in\mathcal{F}}|\nu_f| = \|\mathbb{P}-\mathbb{P}_n\|_{\mathcal{F}}$$

Summary: We presented two approaches for analyzing the supremum of the difference of expected and empirical values: bracketing and symmetrization. We stated the symmetrization lemma in full generality.