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

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 20 & 21
Apr. 21 & 23, 2020

Goals:

1. REGRESSION. PREDICTION VS ESTIMATION

As before, let $\mathcal{S} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be a set of i.i.d. pairs with distribution $P = P_X \times P_{Y|X}$ on $\mathcal{X} \times \mathcal{Y}$. Let $f^*(x) = \mathbb{E}[Y|X = x]$ be the *regression function*. One can show that

$$f^* \in \operatorname*{argmin}_f \mathbb{E}(f(X) - Y)^2$$

where minimization is over all measurable functions.

Given a class $\mathcal{F}$ of functions $\mathcal{X} \to \mathcal{Y}$, we also define

$$f_{\mathcal{F}} \in \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2$$

to be the best predictor within the class $\mathcal{F}$.

Risk of a function f is defined as

$$\mathbb{E}(f(X) - f^*(X))^2 = \|f - f^*\|^2_{L^2(P)} = \|f - f^*\|^2$$

We will be interested in analyzing estimators $\widehat{f}$ constructed on the basis of n datapoints. The hat on $\widehat{f}$ reminds us about the dependence on $\mathcal{S}$.

Note that for any function f,

$$\begin{aligned}
\mathbb{E}(f(X) - Y)^2 - \min_h \mathbb{E}(h(X) - Y)^2 &= \mathbb{E}(f(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 \\
&= \mathbb{E}(f(X) - f^*(X) + f^*(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2 \\
&= \mathbb{E}(f(X) - f^*(X))^2
\end{aligned}$$

Question: given i.i.d. data $\mathcal{S}$, can we select estimator $\widehat{f}$ such that risk

$$\left\| \widehat{f} - f^* \right\|^2$$

is small in expectation or high-probability (with respect to the draw of $\mathcal{S}$)? Without further assumptions this is not possible.

Two standard scenarios:

- Well-specified case: given some class $\mathcal{F}$, assume $f^* \in \mathcal{F}$. More precisely, P is such that the regression function is in the class $\mathcal{F}$.

- Misspecified case (agnostic learning in CS community): Redefine goal as

$$\begin{aligned} &\left\|\widehat{f} - f^*\right\|^2 - \min_{f\in\mathcal{F}} \|f - f^*\|^2 \\ &= \mathbb{E}(\widehat{f}(X) - Y)^2 - \min_{f\in\mathcal{F}} \mathbb{E}(f(X) - Y)^2 \end{aligned} \tag{1.1}$$

  but do not insist that $f^* \in \mathcal{F}$. Upper bounds on (1.1) are called Oracle Inequalities in statistics, while the prediction form has been also studied in statistical learning theory.

We see that the problem of prediction and the problem of estimation naturally coincide for square loss. Moreover, the misspecified problem arises naturally as a relaxation of an assumption on the form of the distribution.

Here, the road naturally forks into at least several paths: analyze the well-specified case, analyze the misspecified case, or change the loss function altogether. Let us briefly consider the last generalization.

2. PREDICTION WITH OTHER LOSS FUNCTIONS

This will be a brief but useful detour. Consider changing the loss function in the prediction problem (1.1) on the previous page:

$$\mathbb{E}\boldsymbol{\ell}(f(X), Y) - \min_{f\in\mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f(X), Y) \tag{2.2}$$

for some $\boldsymbol{\ell} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. In Lecture 14 we already showed that ERM

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

enjoys

$$\mathbb{E}\boldsymbol{\ell}(\widehat{f}(X), Y) - \min_{f\in\mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f(X), Y) \le \mathbb{E} \sup_{f\in\mathcal{F}} \mathbb{E}\boldsymbol{\ell}(f(X), Y) - \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\ell}(f(X_i), Y_i).$$

The latter is at most

$$2\mathbb{E} \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \boldsymbol{\ell}(f(X_i), Y_i) \tag{2.3}$$

by symmetrization, which is Rademacher averages of the loss class

$$\boldsymbol{\ell} \circ \mathcal{F}|_{(X_1,Y_1),\ldots,(X_n,Y_n)}$$

We would like to further upper bound this with Rademacher averages of the function class itself. This can be done if $\boldsymbol{\ell}$ is Lipschitz in the first argument.

Lemma (Contraction): Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz, $i = 1, \ldots, n$. Let $\Theta \subset \mathbb{R}^n$ and $\phi \circ \theta = (\phi_1(\theta_1), \ldots, \phi_n(\theta_n))$ for $\theta \in \Theta$. Denote $\phi \circ \Theta = \{\phi \circ \theta : \theta \in \Theta\}$. Then

$$\widehat{\mathcal{R}}(\phi \circ \Theta) \le \widehat{\mathcal{R}}(\Theta).$$

Proof. Conditionally on $\epsilon_1, \dots, \epsilon_{n-1}$,

$$\begin{aligned}
\mathbb{E}_{\epsilon_n} \sup_{\theta\in\Theta} \langle \phi\circ\theta, \epsilon\rangle &= \frac{1}{2}\left(\sup_{\theta\in\Theta}\left\{\langle \phi\circ\theta_{1:n-1}, \epsilon_{1:n-1}\rangle + \phi_n(\theta_n)\right\} + \sup_{\theta'\in\Theta}\left\{\langle \phi\circ\theta'_{1:n-1}, \epsilon_{1:n-1}\rangle - \phi_n(\theta'_n)\right\}\right) \\
&\le \frac{1}{2}\sup_{\theta,\theta'\in\Theta} \langle \phi\circ\theta_{1:n-1}, \epsilon_{1:n-1}\rangle + \langle \phi\circ\theta'_{1:n-1}, \epsilon_{1:n-1}\rangle + |\theta_n - \theta'_n| \\
&= \frac{1}{2}\sup_{\theta,\theta'\in\Theta} \langle \phi\circ\theta_{1:n-1}, \epsilon_{1:n-1}\rangle + \langle \phi\circ\theta'_{1:n-1}, \epsilon_{1:n-1}\rangle + \theta_n - \theta'_n \\
&= \frac{1}{2}\left(\sup_{\theta\in\Theta}\left\{\langle \phi\circ\theta_{1:n-1}, \epsilon_{1:n-1}\rangle + \theta_n\right\} + \sup_{\theta'\in\Theta}\left\{\langle \phi\circ\theta'_{1:n-1}, \epsilon_{1:n-1}\rangle - \theta'_n\right\}\right) \\
&= \mathbb{E}_{\epsilon_n} \sup_{\theta\in\Theta} \langle \phi\circ\theta_{1:n-1}, \epsilon_{1:n-1}\rangle + \epsilon_n\theta_n
\end{aligned}$$

The inequality follows from the Lipschitz condition and the following equality is justified because of the symmetry of the other two terms with respect to renaming θ and θ'. Proceeding to remove the other signs concludes the proof. □

We now apply this lemma to functions $\phi_i(\cdot) = \ell(\cdot, Y_i)$. As long as these functions are L-Lipschitz, contraction lemma gives

$$\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \ell(f(X_i), Y_i) \le L\cdot \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) = L\cdot \frac{1}{n}\mathbb{E}\widehat{\mathcal{R}}(\mathcal{F}|_{X_1,\dots,X_n}), \tag{2.4}$$

the (expected) Rademacher averages of $\mathcal{F}$. The argument can be seen as a generalization of the argument we did in Lecture 14 for classification where we "erased" multipliers $(1-2Y_i)$.

The simple analysis we just performed applies to any Lipschitz loss function. For uniformly bounded $\mathcal{F}$ and $\mathcal{Y}$, square loss is Lipschitz, but that is no longer true for unbounded $\mathcal{Y}$ (e.g. for real-value prediction with Gaussian noise). Hence, such an analysis only goes so far.

Second, observe that one would only obtain rates $n^{-1/2}$ or worse with such an analysis, while we might hope to have faster decrease. For instance, in finite-dimensional regression, one can recall the classical $d\cdot n^{-1}$ rates for Least Squares.

A quick inspection tells us that the second step (see Lecture 14) in the sequence of inequalities

$$\mathbb{E}\left[\mathbf{L}(\widehat{f})\right] - \mathbf{L}(f^*) \le \mathbb{E}\left[\mathbf{L}(\widehat{f}) - \widehat{\mathbf{L}}(\widehat{f})\right] \le \mathbb{E}\sup_{f\in\mathcal{F}}\left[\mathbf{L}(f) - \widehat{\mathbf{L}}(f)\right] \tag{2.5}$$

for ERM $\widehat{f}$ may be too loose. The second step only used the fact that $\widehat{f}$ belongs to $\mathcal{F}$. It turns out one can localize its place in $\mathcal{F}$ better than that.

Next few lectures will be on nonparametric regression. We will start with well-specified models.

3. NONPARAMETRIC REGRESSION: WELL-SPECIFIED CASE

We will start with "fixed design": $x_1, \dots, x_n \in \mathcal{X}$ are fixed. Let

$$Y_i = f^*(x_i) + \eta_i$$

where η_i are zero-mean independent subGaussian. Suppose $f^* \in \mathcal{F}$. Goal: estimate f^* on the points $x_1, \ldots, x_n$ (denoise the observed values). That is, the goal is to provide nonasymptotic bounds on

$$\mathbb{E}_\eta \left\| \widehat{f} - f^* \right\|_{L^2(P_n)}^2,$$

where $\widehat{f}$ is the least squares (ERM) constrained to $\mathcal{F}$. In constrast, in random design the goal is w.r.t. $L^2(P)$ with P unknown, while here P_n is known. We write the $L^2(P_n)$ norm more succinctly as $\mathbb{E} \left\| \widehat{f} - f^* \right\|_n^2$.

Since

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(x_i) - Y_i)^2 = \|f - Y\|_n^2$$

we have

$$\|f^* - Y\|_n^2 \geq \left\| \widehat{f} - Y \right\|_n^2 = \left\| \widehat{f} - f^* + f^* - Y \right\|_n^2 = \left\| \widehat{f} - f^* \right\|_n^2 + \|f^* - Y\|_n^2 + 2\langle \widehat{f} - f^*, f^* - Y \rangle_n$$

where $\langle a, b \rangle_n = \frac{1}{n} \langle a, b \rangle$. Thus,

$$\left\| \widehat{f} - f^* \right\|_n^2 \leq 2\langle \eta, \widehat{f} - f^* \rangle_n \tag{3.6}$$

which we will call *the basic inequality*.

3.1 Informal intuition for localization

Before developing the localization approach, we provide some intuition. The first intuition comes from viewing (3.6) as a fixed point.

Let's assume for simplicity that η_i are 1-subGaussian. For $a \in \mathbb{R}^n$, we have that with high probability

$$\langle \eta, a \rangle \lesssim \|a\|$$

Hence, if it holds that

$$\|a\|^2 \leq \langle \eta, a \rangle,$$

then $\|a\| \lesssim 1$.

We can try to repeat this argument with a being the values of $\widehat{f} - f^*$ on the data. However, since $\widehat{f}$ depends on η, we do not have the averaging that we need. Still, we can do the mental experiment of assuming that the dependence is "weak" (e.g. we fit linear regression in small d and large n). Then a bound on the size of $\left\| \widehat{f} - f^* \right\|_n$ would lead to an improved bound on the RHS of the basic inequality, which would in turn tighten the bound on the LHS of the basic inequality, suggesting some kind of a fixed point. It also seems intuitive that this fixed point likely depends on $\mathcal{F}$ and its richness.

3.2 1st approach to localization: ratio-type inequalities

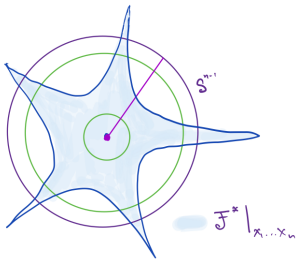
To simplify the proof somewhat, we will assume that $\eta_1, \ldots, \eta_n$ are independent standard normal $N(0, 1)$.

We proceed as in the linear case earlier in the course. First, we divide both sides of the Basic Inequality (3.6) by $\left\|\widehat{f} - f^*\right\|_n$ and further upper bound the right-hand side by a supremum over f, removing the dependence of the algorithm on the data:

$$\left\|\widehat{f} - f^*\right\|_n \leq 2 \sup_{f \in \mathcal{F}} \langle \eta, \frac{f - f^*}{\|f - f^*\|_n} \rangle_n \tag{3.7}$$

By squaring both sides, we would get an upper bound on the estimation error (in probability or in expectation).

Let us use the shorthand $\mathcal{F}^* = \mathcal{F} - f^*$. The rest of the discussion will be about complexity of the neighborhood around f^* in $\mathcal{F}$, or, equivalently, complexity of the neighborhood of 0 in $\mathcal{F}^*$. Observe that we only care about values of functions on the data $x_1, \ldots, x_n$, so the discussion is really about the set $\mathcal{F}^*|_{x_1,\ldots,x_n}$, drawn in blue below.



At this point, one can say that there is no difference from the linear case, and we should just go ahead and analyze

$$\sup_{g \in \mathcal{F}^*} \langle \eta, \frac{g}{\|g\|_n} \rangle_n$$

After all, this is just the Gaussian width (normalized by $\sqrt{n}$) of the subset of the sphere obtained by rescaling all the functions:

$$K = \{v \in \mathsf{S}^{n-1} : \exists g \in \mathcal{F}^* \text{ s.t. } v = (g(x_1), \ldots, g(x_n))/(\sqrt{n}\,\|g\|_n)\}.$$

(here the normalization is because $\|g\|_n$ is scaled as $1/\sqrt{n}$ times the ℓ_2 norm.) How big is this subset of the sphere? Note: if the set is all of S^{n-1}, we are doomed since in that case

$$\sup_{g \in \mathcal{F}^*} \langle \eta, \frac{g}{\|g\|_n} \rangle_n = \sup_{v \in \mathsf{S}^{n-1}} \frac{1}{\sqrt{n}} \langle \eta, v \rangle = \frac{1}{\sqrt{n}} \|\eta\| \sim 1$$

and does not converge to zero. What we would need is that K is a *significantly smaller* subset of the sphere. In the linear case, this was easy: we simply used the fact that the subset is d-dimensional. However, for nonlinear functions, it is not easy to see what the set is.

There is a bigger problem, however. Upon rescaling every vector to the sphere, all the functions are treated equally even if their unscaled versions are very close to being zero (that is, close to f^* in the original class $\mathcal{F}$). In other words, the quantity

$$\sup_{g \in \mathcal{F}^* : \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n$$

can be potentially much smaller than the unrestricted supremum. This is depicted in the above figure. If we look at functions within the smaller green sphere, its rescaled version is

the whole sphere. However, at larger scales (e.g. the larger green sphere), the set can be much smaller. Understanding the map

$$u \mapsto \sup_{g \in \mathcal{F}^* : \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n$$

will be key. In particular, we can break up the balance at scale u and instead have a better upper bound

$$\left\| \widehat{f} - f^* \right\|_n \leq u + 2 \sup_{g \in \mathcal{F}^* : \|g\|_n \geq u} \langle \eta, \frac{g}{\|g\|_n} \rangle_n \tag{3.8}$$

Consider the following assumption:

Definition: A class $\mathcal{H}$ is *star-shaped* (around 0) if $h \in \mathcal{H}$ implies $\lambda h \in \mathcal{H}$ for $h \in [0, 1]$. In particular, if $\mathcal{H}$ is convex and contains 0, it is star-shaped.

We will assume that $\mathcal{F}^*$ is star-shaped. In particular, if $\mathcal{F}$ is convex, then $\mathcal{F}^*$ is star-shaped. The key property of a star-shaped class is that by increasing the radius, the sets cannot become more complex, as for any function there is a scaled copy of it at a smaller magnitude.

In light of this last remark, we claim that the inequality $\|g\|_n \geq u$ in the supremum in (3.8) can be replaced with an *equality* if the class is star-shaped. Indeed, for any $g \in \mathcal{F}^*$ with $\|g\|_n \geq u$, there is a corresponding function $h = u \frac{g}{\|g\|_n}$ with norm $\|h\|_n = u$ and

$$\langle \eta, \frac{g}{\|g\|_n} \rangle_n = \langle \eta, \frac{h}{u} \rangle_n$$

Hence,

$$\langle \eta, \frac{g}{\|g\|_n} \rangle_n \leq \frac{1}{u} \sup_{h \in \mathcal{F}^* : \|h\|_n = u} \langle \eta, h \rangle_n$$

Taking a supremum on the LHS over g with $\|g\|_n \geq u$ gives an upper bound on (3.8) as

$$\begin{aligned} \left\| \widehat{f} - f^* \right\|_n &\leq u + \frac{2}{u} \sup_{g \in \mathcal{F}^* : \|g\|_n = u} \langle \eta, g \rangle_n \\ &\leq u + \frac{2}{u} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq u} \langle \eta, g \rangle_n \end{aligned} \tag{3.9}$$

where in the last step we included all the functions below level u. We will use concentration to replace the second term with its expectation. In particular, define

$$Z(u) = \sup_{g \in \mathcal{F}^* : \|g\|_n \leq u} \langle \eta, g \rangle_n$$

and

$$G(u) = \mathbb{E} Z(u).$$

If we were to replace $Z(u)$ on the RHS of (3.9) with $G(u)$, the natural balance between the two terms would be

$$u = \frac{2}{u} G(u)$$

Definition: The *critical radius* δ_n will be the minimum δ satisfying

$$G(\delta) \leq \delta^2/2$$

One can ask if this critical radius is actually well-defined. This follows from the following:

Lemma: If $\mathcal{F}^*$ is star-shaped, the function $u \mapsto G(u)/u$ is non-increasing.

Proof. Let $\delta' < \delta$. Take any $h \in \mathcal{F}^*$ with $\delta' < \|h\|_n \leq \delta$. By star-shapedness,

$$h' = \left(\frac{\delta'}{\delta}\right) h \in \mathcal{F}^*$$

and $\|h'\|_n = \frac{\delta'}{\delta} \|h\|_n \leq \delta'$. Hence,

$$\langle \eta, h \rangle_n = \frac{\delta}{\delta'} \langle \eta, h' \rangle_n \leq \frac{\delta}{\delta'} Z(\delta')$$

Taking supremum on the left-hand side over h with $\|h\|_n \leq \delta$, as well as expectation on both sides, finishes the proof. □

In particular, for any $u \geq \delta_n$,

$$G(u) \leq u^2/2$$

Indeed,

$$G(u) = u\frac{G(u)}{u} \leq u\frac{G(\delta_n)}{\delta_n} \leq u\delta_n/2 \leq u^2/2. \tag{3.10}$$

To formally replace $Z(u)$ with $G(u)$ in the balancing equation, we need a concentration result.

Lemma (Gaussian Concentration): Let $\eta = (\eta_1, \ldots, \eta_n)$ be a vector of independent standard normals. Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz (w.r.t. Euclidean norm). Then for all $t > 0$

$$\mathbb{P}\left(\phi(\eta) - \mathbb{E}\phi \geq t\right) \leq \exp\left\{-\frac{t^2}{2L^2}\right\}$$

First, observe that $Z(u)$ is $(u/\sqrt{n})$-Lipschitz function of η. Omitting the argument u,

$$Z[\eta] - Z[\eta'] \leq \sup_{g \in \mathcal{F}^*, \|g\|_n \leq u} \langle \eta, g \rangle_n - \langle \eta', g \rangle_n \leq \left\|\eta - \eta'\right\|_n \sup_{g \in \mathcal{F}^*, \|g\|_n \leq u} \|g\|_n \leq \frac{u}{\sqrt{n}} \left\|\eta - \eta'\right\|$$

Hence, for any $u > 0$,

$$\mathbb{P}\left(Z(u) - \mathbb{E}Z(u) \geq t\right) \leq \exp\left\{-\frac{nt^2}{2u^2}\right\} \tag{3.11}$$

In particular, by setting $t = u^2$,

$$\mathbb{P}\left(Z(u) \geq G(u) + u^2\right) \leq \exp\left\{-\frac{nu^2}{2}\right\} \tag{3.12}$$

In light of (3.10), we have proved

Lemma: Assuming $\mathcal{F}^*$ is star-shaped, with probability at least $1 - \exp\left\{-\frac{nu^2}{2}\right\}$,

$$Z(u) \leq 1.5u^2 \tag{3.13}$$

for any $u \geq \delta_n$.

Thus, from (3.9), we have

$$\left\|\widehat{f} - f^*\right\|_n \leq 4u \tag{3.14}$$

with probability at least $1 - \exp\left\{-\frac{nu^2}{2}\right\}$, for any $u \geq \delta_n$. Squaring both sides, yields

Theorem: Assume $x_1, \ldots, x_n$ are fixed, $\eta_1, \ldots, \eta_n$ are i.i.d. standard normal, and $Y_i = f^*(x_i) + \eta_i$ with $f^* \in \mathcal{F}$. Assume $\mathcal{F} - f^*$ is star-shaped and δ_n the corresponding critical radius. Then constrained least squares $\widehat{f}$ satisfies

$$\mathbb{P}\left(\left\|\widehat{f} - f^*\right\|_n^2 \geq 16s\delta_n^2\right) \leq \exp\left\{-\frac{ns\delta_n^2}{2}\right\} \tag{3.15}$$

for any $s \geq 1$. In particular, this implies

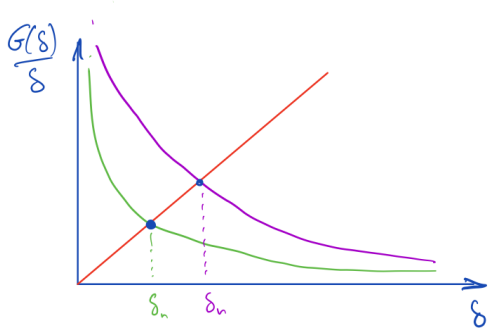
$$\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 \lesssim \delta_n^2 + \frac{1}{n}.$$

Note: in the literature, you will find a slightly different parametrization. Write $\psi(r) = \mathbb{E}Z(\sqrt{r})$. In other words, $\psi(u^2) = G(u)$. Then ψ has the *subroot* property:

$$\psi(ra) \leq \sqrt{a}\psi(r)$$

using the same type of proof as above. The fixed point then reads as the smallest r such that $\psi(r) \leq r$ (ignoring the constant).

Let's quickly discuss the behavior of $G(\delta)/\delta$.



The above sketch shows the function $\delta \mapsto G(\delta)/\delta$ for two classes of functions. The purple curve corresponds to a more complex class, since the Gaussian width (normalized by δ) grows faster as $\delta \to 0$. The corresponding fixed point is larger for a more rich class.