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

Lecturer: A Rakhlin
Scribe: A. Rakhlin

Lecture 22 & 23
Apr. 28 & 30, 2020

1. NONPARAMETRIC REGRESSION, CONTINUED

1.1 2nd approach to localization: offset

We start again with the basic inequality

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

and trivially write it as

$$\left\|\widehat{f} - f^*\right\|_n^2 \le 4\langle \eta, \widehat{f} - f^* \rangle_n - \left\|\widehat{f} - f^*\right\|_n^2$$

Now take the supremum on both sides:

$$\begin{aligned}\mathbb{E}\left\|\widehat{f} - f^*\right\|_n^2 &\le \mathbb{E}\sup_{f\in\mathcal{F}} 4\langle \eta, f - f^* \rangle_n - \|f - f^*\|_n^2 \\ &= \mathbb{E}\sup_{g\in\mathcal{F}-f^*} \frac{1}{n}\sum_{i=1}^n 4\eta_i g(x_i) - g(x_i)^2\end{aligned}$$

which we shall call *the offset Rademacher (or Gaussian) averages.*

Contrast this approach with the first approach where we divided both sides by the norm $\left\|\widehat{f} - f^*\right\|_n$ and then upper bounded by supremum over an appropriately localized subset, then squared both sides.

Surprisingly, this somewhat simpler approach yields correct upper bounds. Note that the negative quadratic term annihilates the fluctuations of the term $\eta_i g(x_i)$ when the magnitude of g becomes large enough (beyond some critical radius). Hence, the supremum is achieved in a finite radius, no larger than the critical radius:

Lemma: Let δ_n be the critical radius. Then for any $c \ge 1$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{F}^*} 2c\langle \eta, g\rangle_n - \|g\|_n^2 > 2c^2\delta_n^2 + \frac{2c^2 u}{n}\right) \le \exp\{-u/2\} \tag{1.1}$$

In particular,

$$\mathbb{E}\sup_{g\in\mathcal{F}^*} 2\langle \eta, g\rangle_n - \|g\|_n^2 \lesssim \delta_n^2 + \frac{1}{n}.$$

Proof. By Gaussian concentration,

$$\mathbb{P}\left(Z(\delta_n) \ge \mathbb{E}Z(\delta_n) + t\delta_n\right) \le \exp\left\{-\frac{nt^2}{2}\right\}. \tag{1.2}$$

We now condition on the complement of the above event. Take $g \in \mathcal{F}^*$. Consider two cases. First, if $\|g\|_n \le \delta_n$ then

$$2c\langle \eta, g\rangle_n - \|g\|_n^2 \le 2cZ(\delta_n) \le 2c\left(\mathbb{E}Z(\delta_n) + t\delta_n\right) \le 2c\left(\frac{\delta_n^2}{2} + t\delta_n\right) \le c(t+\delta_n)^2 \tag{1.3}$$

Second, if $\|g\|_n \ge \delta_n$, we set $r = \delta_n / \|g\|_n \le 1$. Then

$$2c\langle \eta, g\rangle_n - \|g\|_n^2 = \frac{2c}{r}\langle \eta, \frac{\delta_n}{\|g\|_n} g\rangle - \frac{\delta_n^2}{r^2} \le \frac{2c}{r} Z(\delta_n) - \frac{\delta_n^2}{r^2} = \frac{2\delta_n}{r}\frac{cZ(\delta_n)}{\delta_n} - \frac{\delta_n^2}{r^2}. \tag{1.4}$$

Using $2ab - b^2 \le a^2$, we get a further upper bound of

$$c^2\left(\frac{Z(\delta_n)}{\delta_n}\right)^2 \le c^2\left(\frac{\delta_n^2/2 + t\delta_n}{\delta_n}\right)^2 = c^2(\delta_n/2 + t)^2 \tag{1.5}$$

□

1.1.1 Example: linear regression

To get a sense of the behavior of the offset process, consider the linear class $\mathcal{F} = \{x \mapsto \langle w, x\rangle : w \in \mathbb{R}^d\}$. First, $\mathcal{F} - f^* = \mathcal{F}$. Second, note that functions are unbounded, and so Rademacher averages are unbounded too. However, offset averages are

$$\sup_{w\in\mathbb{R}^d} \sum_{i=1}^n \eta_i\langle w, x_i\rangle - c\langle w, x_i\rangle^2 = \sup_{w\in\mathbb{R}^d} \langle w, \sum_{i=1}^n \eta_i x_i\rangle - c\|w\|_\Sigma^2 \tag{1.6}$$

$$= \frac{1}{4c}\left\|\sum_{i=1}^n \eta_i x_i\right\|_{\Sigma^\dagger}^2 \tag{1.7}$$

where $\Sigma = \sum_{i=1}^n x_i x_i^\mathsf{T}$ and $\Sigma^\dagger$ is the pseudoinverse. Assuming $\mathbb{E}\eta_i^2 \le 1$,

$$\mathbb{E}\left\|\sum_{i=1}^n \eta_i x_i\right\|_{\Sigma^{-1}}^2 \le \sum_{i=1}^n x_i^\mathsf{T}\Sigma^\dagger x_i = \mathrm{tr}(\Sigma\Sigma^\dagger) = \mathrm{rank}(\Sigma)$$

We see that, these offset Rademacher/Gaussian averages have the right behavior: we already saw in the first part of the course that the fast rate for linear regression is $O\left(\frac{\mathrm{rank}(\Sigma)}{n}\right)$ without further assumptions.

We can view the negative term that extinguishes the fluctuations of the zero-mean process as coming from the curvature of the square loss. Without the curvature, the negative term is not there and we are left with the usual Rademacher/Gaussian averages.

2. LEAST SQUARES

2.0.1 Nonparametric

We would like to calculate the critical radius δ_n for some function clases of interest. Recall that δ_n is defined as the smallest number such that

$$\mathbb{E}\sup_{g\in\mathcal{F}^*:\|g\|_n\le\delta} \langle \eta, g\rangle_n \le \delta^2/2.$$

The strategy is to find upper bounds on the left-hand-side in terms of δ and then solve for the minimal δ. In particular, we know that for any $\alpha \geq 0$,

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim \alpha + \frac{1}{\sqrt{n}} \int_{\alpha/4}^{\delta} \sqrt{\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon)} d\varepsilon$$

Suppose we have

$$\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon) \lesssim \varepsilon^{-p}$$

for $p \in (0, 2)$. Then, taking $\alpha = 0$,

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim n^{-1/2} [\varepsilon^{1-p/2}]_0^{\delta} = n^{-1/2} \delta^{1-p/2}$$

Setting

$$n^{-1/2} \delta^{1-p/2} = \delta^2$$

yields

$$\delta_n = n^{-\frac{1}{2+p}}$$

and thus the rate of the least squares estimator is

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

It can be shown that minimax optimal rates of estimation (for any estimator) for fixed design are given by[1] the fixed point

$$\frac{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \delta_*)}{n} \asymp \delta_*^2 \tag{2.8}$$

If $\log \mathcal{N}(\mathcal{F}, L^2(P_n), \delta) \asymp \delta^{-p}$, the balance is

$$\delta_*^{-p} n^{-1} \asymp \delta_*^2$$

which gives the same rate of $\delta_*^2 = n^{-\frac{2}{2+p}}$. Hence, least squares are optimal in this minimax sense for $p \in (0, 2)$.

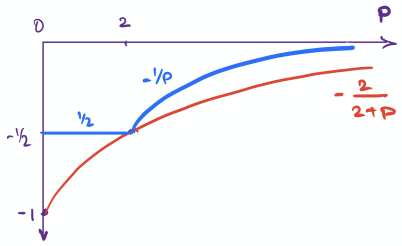


Figure 1: Optimal (in general) rates $n^{-\frac{2}{2+p}}$ (obtained with localization for $p \in (0, 2)$ by ERM) vs without localization (e.g. via global Rademacher averages)

[1] See Yang and Barron "Information-theoretic determination of minimax rates of convergence," 1999

Example: Convex L-Lipschitz functions on a compact domain in $\mathbb{R}^d$:

$$\log \mathcal{N}(\mathcal{F}_{\text{cvx,lip}}, L^2(P_n), \varepsilon) \leq (L/\varepsilon)^{d/2}$$

Example: L-Lipschitz functions on a compact domain in $\mathbb{R}^d$:

$$\log \mathcal{N}(\mathcal{F}_{\text{lip}}, L^2(P_n), \varepsilon) \leq (L/\varepsilon)^{d}$$

2.0.2 Parametric

Consider the parametric case,

$$\log \mathcal{N}(\mathcal{F}^*, L^2(P_n), \varepsilon) \lesssim d \log(1 + 2/\varepsilon)$$

Then

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{d \log(1 + 2/\varepsilon)} d\varepsilon \tag{2.9}$$

Change of variables gives an upper bound

$$\sqrt{\frac{d}{n}} \delta \cdot \int_0^1 \sqrt{\log(1 + 2/(u\delta))} du \tag{2.10}$$

Unfortunately, this gives a pesky logarithmic factor that should not be there. However, for some parametric cases one can, in fact, prove that *local covering numbers* behave as

$$\log \mathcal{N}(\mathcal{F}^* \cap \{g : \|g\|_n \leq \delta\}, L^2(P_n), \varepsilon) \lesssim d \log(1 + 2\delta/\varepsilon) \tag{2.11}$$

In this case, the change-of-variables leads to

$$\mathbb{E} \sup_{g \in \mathcal{F}^* : \|g\|_n \leq \delta} \langle \eta, g \rangle_n \lesssim \sqrt{\frac{d}{n}} \delta \cdot \int_0^1 \sqrt{\log(1 + 2/\varepsilon)} d\varepsilon \lesssim \sqrt{\frac{d}{n}} \delta \tag{2.12}$$

Equating

$$\delta \sqrt{\frac{d}{n}} \asymp \delta^2$$

yields

$$\delta_n^2 \asymp \frac{d}{n}$$

Note that local covering numbers (2.11) are available in some parametric cases (e.g. when we discretize the parameter space of linear functions) but may not be available for some other classes (e.g. for VC classes, except under additional conditions).

2.1 Remarks

- to bound metric entropy of $\mathcal{F}^* = \mathcal{F} - f^*$, instead consider $\mathcal{F} - \mathcal{F}$. This often leads to only mild increase in a constant. For instance, if $\mathcal{F}$ is a class of L-Lipschitz functions, then $\mathcal{F} - \mathcal{F}$ is a subset of $2L$-Lipschitz functions.

- Note that the rate δ_n^2 depends on local covering numbers (or, local complexity) around f^*. This gives a path to proving adaptivity results (e.g. if f^* is convex but has only k linear pieces, the rate of estimation is parametric because its neighborhood is "simple").

- A simple counting argument (see Yang & Barron 1999, Section 7) shows that for rich enough classes (e.g. nonparametric) worst-case local entropy (worst-case location in the class) and global entropies behave similarly. This implies, in particular, that instead of constructing a local packing for a lower bound (via hypothesis testing), one can instead use global entropy with Fano inequality, justifying the LHS of (2.8) as the lower bound for estimation. See also Mendelson's "local vs global parameters" paper for an in-depth discussion.

3. ORACLE INEQUALITIES

What if we do not assume the regression function f^* is in $\mathcal{F}$? How can we prove an oracle inequality

$$\mathbb{E}\left\|\widehat{f}-f^*\right\|_n^2 - \inf_{f\in\mathcal{F}}\|f-f^*\|_n^2 \le \phi(\mathcal{F},n)$$

Again, we will focus on fixed design.

3.1 Convex $\mathcal{F}$

Suppose $\mathcal{F}$ is convex (or, rather, $\mathcal{F}|_{x_1,\dots,x_n}$ is convex). Let $\widehat{f}$ be the constrained least squares:

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

For the basic inequality we used

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

but in the misspecified case this is no longer true. However, what is true is that

$$\left\|\widehat{f}-Y\right\|_n^2 \le \|f_{\mathcal{F}}-Y\|_n^2$$

Unfortunately, this inequality is not strong enough to get us the desired result. Fortunately, we can do better. Since $\widehat{f}$ is a projection of Y onto $F=\mathcal{F}_{x_1,\dots,x_n}$, it holds that

$$\left\|\widehat{f}-Y\right\|_n^2 \le \|f-Y\|_n^2 - \left\|\widehat{f}-f\right\|_n^2 \tag{3.13}$$

for any $f\in\mathcal{F}$, and in particular for $f_{\mathcal{F}}$. This is a simple consequence of convexity and pythagorean theorem. The negative quadratic will give us the extra juice we need.

Adding and subtracting f^* on both sides and expanding,

$$\left\|\widehat{f}-f^*\right\|_n^2 + \|f^*-Y\|_n^2 + 2\langle \widehat{f}-f^*, -\eta\rangle_n + \left\|f_{\mathcal{F}}-\widehat{f}\right\|_n^2 \le \|f_{\mathcal{F}}-f^*\|_n^2 + \|f^*-Y\|_n^2 + 2\langle f_{\mathcal{F}}-f^*, -\eta\rangle_n$$

which leads to

$$\left\|\widehat{f} - f^*\right\|_n^2 - \|f_{\mathcal{F}} - f^*\|_n^2 \le 2\langle \eta, \widehat{f} - f_{\mathcal{F}} \rangle_n - \left\|\widehat{f} - f_{\mathcal{F}}\right\|_n^2 \tag{3.14}$$

$$\le \sup_{h \in \mathcal{F} - f_{\mathcal{F}}} 2\langle \eta, h \rangle_n - \|h\|_n^2 \tag{3.15}$$

We conclude that for convex $\mathcal{F}$ and fixed design, the upper bounds we find for well-specified and misspecified cases match. Moreover, since the misspecified case is strictly more general and lower bounds for the well-specified case and polynomial entropy growth match the upper bounds, we conclude that constrained least squares are also minimax optimal for fixed design misspecified case.

Note: a crucial observation is that offset complexity would arise even if (3.13) had a different constant multiplier in front of $-\left\|f - \widehat{f}\right\|_n^2$. We will exploit this observation in a bit.

3.2 General $\mathcal{F}$

What if $\mathcal{F}$ is not convex? It turns out that least squares (ERM) can be suboptimal even if $\mathcal{F}$ is a finite class!

3.2.1 A lower bound for ERM

The suboptimality can be illustrated on a very simple example. Suppose $\mathcal{X} = \{x\}$, Y is $\{0,1\}$-valued, and $\mathcal{F} = \{f_0, f_1\}$ such that $f_0(x) = 0$ and $f_1(x) = 1$. The marginal distribution is the trivial $P_X = \delta_x$ and suppose we have two conditional distributions $P_0(Y = 1) = 1/2 - \alpha$ and $P_1(Y = 1) = 1/2 + \alpha$. Clearly, the population minimizer for P_j is f_j. Also, under P_0 the regression function is $f_0^* = 1/2 - \alpha$ while under P_1 it is $f_1^* = 1/2 + \alpha$. Finally, ERM is a method that goes after the most frequent observation in the data $Y_1, \ldots, Y_n$.

However, if $\alpha \propto 1/\sqrt{n}$, there is a constant probability of error in determining whether P_0 or P_1 generated the data. Note that the oracle risk is $\min_{f \in \{f_0, f_1\}} \|f - f_i^*\|^2 = (1/2 - \alpha)^2$ while the risk of the estimator $p(1/2 + \alpha)^2 + (1 - p)(1/2 - \alpha)^2$ where p is the probability of making a mistake and not selecting f_i under the distribution P_i. Hence, the overall comparison to the oracle is at least $p((1/2 + \alpha)^2 - (1/2 - \alpha)^2) = \Omega(\alpha)$ when p is constant.

Hence, ERM (or any "proper" method that selects from $\mathcal{F}$) cannot achieve excess loss smaller than $\Omega(n^{-1/2})$:

$$\max_{P_i \in \{P_0, P_1\}} \left\{ \mathbb{E} \left\|\widehat{f} - f_i^*\right\|^2 - \min_{f \in \{f_0, f_1\}} \|f - f_i^*\|^2 \right\} = \Omega(n^{-1/2})$$

Yet, an improper method that selects $\widehat{f}$ outside $\mathcal{F}$ can achieve an $O(n^{-1})$ rate.

A similar simple lower bound can be constructed for ERM with random design.[2]

3.2.2 How about ERM over Convex Hull?

Given that the procedure has to be "improper" (select from outside of $\mathcal{F}$), one can hypothesize that doing ERM over $\mathrm{conv}(\mathcal{F})$ may work. Interestingly, this procedure is also rate-suboptimal for a finite $\mathcal{F}$ since $\mathrm{conv}(\mathcal{F})$ is too expressive.[3]

[2] For more detailed discussion, we refer to "The importance of convexity in learning with squared loss" by Lee, Bartlett, Williamson, 1996.

[3] Proof can be found in Lecué & Mendelson

3.2.3 An improper procedure

Somewhat surprisingly, only a small modification of ERM is required to make it optimal for general classes. Consider the following two-step procedure[4] (*Star Estimator*):

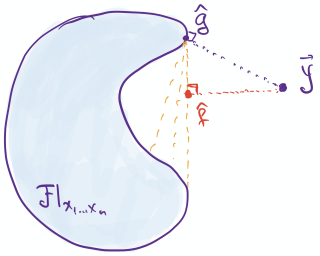
$$\widehat{g} = \operatorname*{argmin}_{f \in \mathcal{F}} \|f - Y\|_n^2 \tag{3.16}$$

$$\widehat{f} = \operatorname*{argmin}_{f \in \mathrm{star}(\mathcal{F}, \widehat{g})} \|f - Y\|_n^2 \tag{3.17}$$

where

$$\mathrm{star}(\mathcal{F}, g) = \{\alpha f + (1-\alpha) g : f \in \mathcal{F}, \alpha \in [0,1]\}.$$

Note that $\widehat{f}$ need not be in $\mathcal{F}$ but is an average of two elements of $\mathcal{F}$.



Note: the method is, in general, different from single ERM over a convex hull of $\mathcal{F}$, and so it is not clear that a version of (3.13) holds.[5]

Lemma: For any $f \in \mathcal{F}$,

$$\|f - Y\|_n^2 - \left\|\widehat{f} - Y\right\|_n^2 \geq \frac{1}{18} \left\|\widehat{f} - f\right\|_n^2. \tag{3.18}$$

The above inequality is an approximate version of (3.13), a generalization of the pythagorean relationship for convex sets.

As a consequence,

$$\left\|\widehat{f} - f^*\right\|_n^2 - \|f_{\mathcal{F}} - f^*\|_n^2 \leq 2\langle \eta, \widehat{f} - f_{\mathcal{F}} \rangle_n - \frac{1}{18} \left\|f_{\mathcal{F}} - \widehat{f}\right\|_n^2$$

and the same upper bounds hold as in the convex case, up to constants. The difference is that the supremum is now in $\mathrm{star}(\mathcal{F}, \widehat{f}) \subseteq \mathcal{F} - f^* + \mathrm{star}(\mathcal{F} - \mathcal{F})$ which is not significantly larger than $\mathcal{F}$ in terms of entropy (unless $\mathcal{F}$ is finite, which can be handled separately).

Remarks:

1. if the set is convex, $\widehat{f} = \widehat{g}$.

[4] For a finite class, the above estimator was analyzed by J-Y. Audibert in 2007 in "Progressive mixture rules are deviation suboptimal".

[5] See Liang-R-Sridharan 2015 for a proof.

2. the Star Estimator can be viewed as one step of Frank-Wolfe. More steps can improve the constant.

Exercise: for any $\varepsilon > 0$ and a set $F \subset \mathbb{R}^n$, the covering numbers satisfy

$$\log \mathcal{N}(F, \|\cdot\|, 2\varepsilon) \le \log \mathcal{N}(\mathrm{star}(F), \|\cdot\|, 2\varepsilon) \le \log(\mathrm{diam}(F)/\epsilon) + \log \mathcal{N}(F, \|\cdot\|, \varepsilon)$$

3.3 Offset Rademacher averages

For a set $V \subset \mathbb{R}^n$, the offset process indexed by V is defined as a stochastic process

$$v \mapsto \sum_{i=1}^n \epsilon_i v_i - c v_i^2 = \langle \epsilon, v \rangle - c \|v\|^2 .$$

Here ϵ_i are independent Rademacher, but the same results hold for any subGaussian random variables.

Lemma: Let $V \subset \mathbb{R}^n$ be a finite set of vectors, $\mathrm{card}(V) = N$. Then for any $c > 0$,

$$\mathbb{E}_\epsilon \max_{v \in V} \langle \epsilon, v \rangle - c \|v\|^2 \le \frac{\log N}{2c}.$$

Furthermore,

$$\mathbb{P}\left(\max_{v \in V} \langle \epsilon, v \rangle - c \|v\|^2 \ge \frac{1}{2c}(\log N + \log(1/\delta)) \right) \le \delta$$

Proof. Assuming the random variables are 1-subGaussian,

$$\begin{aligned}
\mathbb{E} \max_{v \in V} \langle \epsilon, v \rangle - c \|v\|^2 &= \frac{1}{\lambda} \mathbb{E} \log \exp \max \langle \epsilon, v \rangle - c \|v\|^2 \\
&\le \frac{1}{\lambda} \log \sum_{v \in V} \mathbb{E} \exp\{\lambda \langle \epsilon, v \rangle - \lambda c \|v\|^2\} \\
&\le \frac{1}{\lambda} \log \left(N \exp\{\lambda^2 \|v\|^2 / 2 - \lambda c \|v\|^2\} \right) \\
&= \frac{1}{2c} \log N
\end{aligned}$$

where we chose $\lambda = 2c$. □

Theorem: Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$. Then for any $x_1, \ldots, x_n \in \mathcal{X}$ and the corresponding empirical measure P_n,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) - c f(x_i)^2 \tag{3.19}$$

$$\le \inf_{\gamma \ge 0, \alpha \in [0, \gamma]} \left\{ \frac{(2/c) \log \mathcal{N}(\mathcal{F}, L^2(P_n), \gamma)}{n} + 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(P_n), \delta)} d\delta \right\} \tag{3.20}$$