Empirical Process Theory for Nonparametric Analysis

Advanced Topics in Statistical Learning, Spring 2024

Ryan Tibshirani

1 Introduction

Sometimes nonparametric analyses can be carried out with "stone knives and bearskins", as was the case for k-nearest neighbors regression and kernel smoothing. Other times, we will require more sophisticated techniques, as with methods defined in terms of variational optimization, such as smoothing splines, thin plate splines, and RKHS regression. In the current lecture, we'll learn how to leverage such "sophisticated techniques" from empirical process theory in order analyze the smoothing spline. The smoothing spline is chosen by way of example, and is by no means the only estimator that can be analyzed with the tools you will learn.

Of course, empirical process theory is a vast subject and we cover this material in a utilitarian manner, that is, we'll mostly stick to the details needed to understand the example error analysis to come in the last section. For a broader perspective, two excellent references on the subject are van de Geer (2000); Wainwright (2019).

1.1 Problem setup

To get us started, as motivation, we'll develop a basic inequality for estimators that are defined by variational optimization. Assume that we observe data (x_i, y_i), $i = 1, \dots, n$ according to

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \dots, n. \tag{1}$$

We don't even need to specify anything else about the distribution yet: all calculations in this section will be deterministic. Consider defining an estimator by solving the optimization problem

$$\underset{f}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda J(f), \tag{2}$$

for some penalty functional J. The minimization is over all functions f for the which the criterion is well-defined and finite.

Throughout, we're going to assume that J is a seminorm acting on a vector space of functions. Recall, this means that it satisfies the following three properties, for all f, g in its domain and $a \in \mathbb{R}$:

1. $J(f + g) \leq J(f) + J(g)$;
2. $J(af) = |a| J(f)$;
3. $J(f) \geq 0$.

Note that this is weaker than a norm. For a norm, we would make an addendum to the third property to assert that equality holds iff $f = 0$. But a seminorm can have a nontrivial null space (we can have $J(f) = 0$ for $f \neq 0$).

A prominent example of a seminorm regularizer J is $J(f) = \int_a^b (D^m f)^2(x)\, dx$, which acts on functions $f : [a, b] \to \mathbb{R}$ that are m times weakly differentiable. Note that the null space of J is the space of polynomials of degree $m - 1$. Note also that for this choice of penalty functional, problem (2) gives rise to the smoothing spline estimator of polynomial degree $k = 2m - 1$. The most common choice, $m = 2$, yields the cubic smoothing spline.

Another interesting example is the seminorm $J(f) = \mathrm{TV}(D^k f)$, acting on functions $f : [a, b] \to \mathbb{R}$ that are k times weakly differentiable, where $\mathrm{TV}(\cdot)$ denotes the total variation functional. This can be seen as an L^1 analog of the last penalty functional; its null space is now the space of polynomials of degree k. Using this functional in problem (2) gives rise to an estimator that is called the *locally adaptive regression spline* (Mammen and van de Geer, 1997). We haven't studied this estimator yet, but it has powerful local adaptivity properties above and beyond the properties of the smoothing spline. It is not a linear smoother. It is also more difficult to fit computationally.

1.2 Basic inequality

Let $\hat{f}$ be a solution in (2). It does not need to be unique; the analysis that follows applies to any solution in (2). By virtue of optimality, note that we have, for any function f,

$$\frac{1}{n}\sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \lambda J(\hat{f}) \leq \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda J(f). \tag{3}$$

We're going to manipulate the above inequality and it's going to be convenient to write this using empirical norm and inner product notation. Recall we write P_n for the empirical distribution of $x_1 \ldots, x_n$, and define the $L^2(P_n)$ norm by

$$\|g\|_n^2 = \frac{1}{n}\sum_{i=1}^n g^2(x_i).$$

We can also define an $L^2(P_n)$ inner product by

$$\langle g, h\rangle_n = \frac{1}{n}\sum_{i=1}^n g(x_i)h(x_i),$$

so it is clear that $\|g\|_n^2 = \langle g, g\rangle_n$. For simplicity, we'll refer to these as the empirical L^2 norm and empirical L^2 inner product (and we'll often drop "L^2" when it is clear from the context). In a slight abuse of notation, we'll extend this notation to vectors in $\mathbb{R}^n$, so that if $v \in \mathbb{R}^n$, then $\langle g, v\rangle_n = \frac{1}{n}\sum_{i=1}^n g(x_i)v_i$.

With this notation, note that we can rewrite (3) compactly as

$$\|Y - \hat{f}\|_n^2 + \lambda J(\hat{f}) \leq \|Y - f\|_n^2 + \lambda J(f),$$

where $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ is the response vector. Rearranging,

$$\|Y - \hat{f}\|_n^2 - \|Y - f\|_n^2 \leq \lambda(J(f) - J(\hat{f})).$$

Adding and subtracting f in the leftmost term, and expanding, we get

$$\|\hat{f} - f\|_n^2 \leq 2\langle Y - f, \hat{f} - f\rangle_n + \lambda(J(f) - J(\hat{f}))$$

where we have moved the inner product term to the right-hand side.

This is true for any function f. Taking $f = f_0$ in particular, the regression function from (1), and noting that $Y - f_0 = \epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$, the noise vector, we get from the last display

$$\|\hat{f} - f_0\|_n^2 \leq 2\langle \epsilon, \hat{f} - f_0\rangle_n + \lambda(J(f_0) - J(\hat{f})), \tag{4}$$

This is often called *the basic inequality* for $\hat{f}$. We see that the empirical error of $\hat{f}$ is bounded by the sum of two terms. The second term is a difference of penalties between f_0 and $\hat{f}$, and if we regularize (choose λ) "appropriately", then we will see that this term can be controlled. The first term is an empirical inner product between the noise vector ϵ and the error vector $\hat{f} - f_0$. Dealing with this term will be the primary challenge.

1.3 Enter empirical processes

Our strategy will be to bound the first term in (4) by first rewriting it as

$$\langle \epsilon, \hat{f} - f_0 \rangle_n = (J(\hat{f}) + J(f_0)) \Big\langle \epsilon, \underbrace{\frac{\hat{f} - f_0}{J(\hat{f}) + J(f_0)}}_{g} \Big\rangle_n .$$

Since J is a seminorm, we can upper bound the empirical inner product by

$$\langle \epsilon, \hat{f} - f_0 \rangle_n \leq (J(\hat{f}) + J(f_0)) \Big(\sup_{J(g) \leq 1} \underbrace{|\langle \epsilon, g \rangle_n|}_{Z_g} \Big).$$

Thus we can reduce our problem to controlling the supremum of an empirical process Z_g indexed by g.

In general, a stochasic process is a collection of random variables $\{Z_t : \theta \in t \in T\}$ indexed by t over a set T. The most canonical examples are the discrete-time and continuous-time settings, $t \in \mathbb{Z}_+$ and $t \in \mathbb{R}_+$, respectively. But the index t and index set T can be very general. We usually reserve the term *empirical process* for the case when the index is a function $t = f$, and the index set a space of functions $T = \mathcal{F}$.

To make progress on bounding the supremum, $\sup_{J(g) \leq 1} Z_g$, we'll need several ingredients. First, we need some assumption about the distribution of the noise vairable ϵ_i, and sub-Gaussianity (which is weaker than Gaussianity) will do the job. Second, we need some conditions on J, in particular, a condition that quantifies the "richness" of its unit ball, $\{g : J(g) \leq 1\}$. Metric entropy will be our tool for this job.

We introduce these tools over the next several sections, before returning to our main analysis on the error of an estimator $\hat{f}$ defined by solving (2).

2 Entropy and Rademacher complexity

Given a class of functions $\mathcal{F}$, we define its *metric entropy* (or simply entropy) as the logarithm of smallest number of balls in a norm $\|\cdot\|$ of radius $\delta > 0$ needed to cover $\mathcal{F}$. This is denoted

$$\log N(\delta, \mathcal{F}, \|\cdot\|),$$

and also referred to as its log covering number. A visualization is given in Figure 1.

Seemingly unrelated, but actually connected, is Rademacher complexity. This itself comes in two flavors. The *empirical Rademacher complexity* based on a sample $x_1, \ldots, x_n$ is the expected largest absolute inner product achievable with i.i.d. Rademacher noise $\sigma_1, \ldots, \sigma_n$ (each of which take on the value ± 1 with equal probability). To be precise, this is defined as

$$\mathrm{Rad}(\mathcal{F}, x_{1:n}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \epsilon_i f(x_i) \right| \right].$$

The expectation above is over $\sigma_1, \ldots, \sigma_n$ (as signified by the subscript notation, $\mathbb{E}_\sigma$), with $x_1, \ldots, x_n$ fixed. The *population Rademacher complexity* is then defined as the expected value of the empirical Rademacher complexity over i.i.d. draws $x_1, \ldots, x_n$,

$$\mathrm{Rad}(\mathcal{F}) = \mathbb{E}_{x,\sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \right].$$

To emphasize, the expectation above is over both $\sigma_1, \ldots, \sigma_n$ and $x_1, \ldots, x_n$.

A general result connecting these two notions of complexity is called *Dudley's entropy integral*, which we can state as follows. Denote by $B_n(\rho)$ the ball in the norm $\|\cdot\|_n$ (the empirical norm based on $x_1, \ldots, x_n$) of radius $\rho > 0$ centered at the origin. Then there exists a constant $c > 0$ such that

$$\mathrm{Rad}(\mathcal{F} \cap B_n(\rho), x_{1:n}) \leq \frac{c}{\sqrt{n}} \int_0^\rho \sqrt{\log N(\delta, \mathcal{F}, \|\cdot\|_n)}\, d\delta. \tag{5}$$

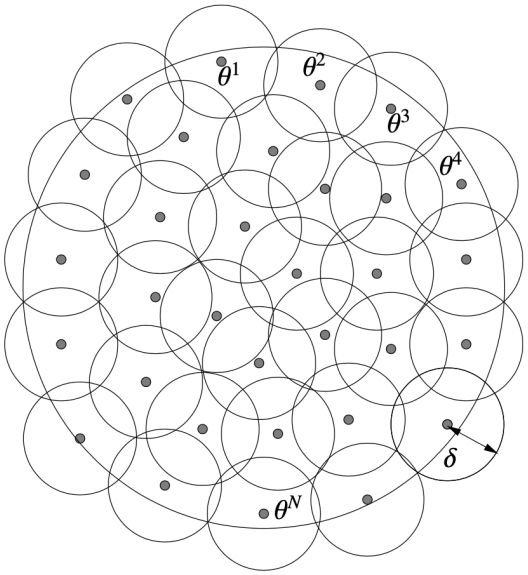

Figure 1: *Illustration of a covering using balls of radius δ. The entropy is the logarithm of the smallest possible number of balls in such a covering. Credit: Chapter 5.1 of Wainwright (2019).*

Entropy and Rademacher complexity are each interesting and important, and have several applications in probability and nonparametric analysis, but for our purposes, we can think of them as follows. Entropy arises naturally in a probabilistic technique known as chaining. This can be used to control sub-Gaussian processes; in particular, it can be used to bound what is known as the sub-Gaussian complexity

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. mean zero sub-Gaussian random variables (to be defined precisely below). We'll see this quantity arising naturally in our error analysis of smoothing splines, and we'll see the role entropy plays in Lemma 1 below.

On the other hand, Rademacher complexity appears when using a technique called symmetrization, and other techniques, to analyze the fluctuations of empirical averages around their population means. For us, we will see it can be used to control the difference between the empirical $\|\cdot\|_n$ and population $\|\cdot\|_2$ norms, with high probability (over the draws of $x_1, \ldots, x_n$). We'll see this in Lemma 2 below.

Lastly, we remark that we are commonly interested in (an upper bound on) how $\log N(\delta, \mathcal{F}, \|\cdot\|_n)$ scales with δ, as $\delta \to 0$, and since we are considering here the empirical norm $\|\cdot\|_n$, we ideally want log covering number bounds that hold for any arrangement of $x_1, \ldots, x_n$. Observe that this can be upper bounded by $\log N(\delta, \mathcal{F}, \|\cdot\|_\infty)$ where $\|\cdot\|_\infty$ is the sup norm (because $\|f - g\|_\infty \leq \delta$ implies $\|f - g\|_n \leq \delta$). Table 1 gives entropy rates for a few example function classes of interest.

3 Sub-Gaussian random variables

Sub-Gaussianity is a condition that captures some of the key properties of the Gaussian tail (but at the same time accomodates many distributions that are not Gaussian).

A random variable X is said to be sub-Gaussian with mean μ and variance proxy $\sigma^2 > 0$ if

$$\mathbb{E}[e^{t(X-\mu)}] \leq e^{\sigma^2 t^2/2}, \quad \text{for all } t \in \mathbb{R}.$$

Function class $\mathcal{F}$	Norm $\|\cdot\|$	Entropy rate
r-dimensional (e.g., class of natural splines with r knots) with $\mathrm{diam}(\mathcal{F}, \|\cdot\|_n) = \rho$	$\|\cdot\|_n$	$r \log(\rho/\delta)$
L-Lipschitz functions on $[0,1]^d$	$\|\cdot\|_\infty$	$(L/\delta)^d$
m times weakly differentiable functions on $[0,1]$ such that $\int_0^1 [D^m f(x)]^2 \, dx \leq L$ and $\|f\|_\infty \leq b$, for an integer $m \geq 1$	$\|\cdot\|_\infty$	$(L/\delta)^{1/m} + \log(b/\delta)$

Table 1: *Entropy rates for some function classes of interest in nonparametric analysis.*

This says that the moment generating function of X is dominated by the moment generating function of $N(\mu, \sigma^2)$ random variable. It turns out that an equivalent characterization of sub-Gaussianity is that

$$\mathbb{P}(|X - \mu| \geq t) \leq c \, \mathbb{P}(|Z| \geq t), \quad \text{for all } t \in \mathbb{R},$$

for some constant $c > 0$, where $Z \sim N(0,1)$. This is perhaps more intuitive and explains the name "sub-Gaussian": the tails must decay at least as fast as that of the Gaussian distribution.

Here are some examples of sub-Gaussian random variables.

- A $N(\mu, \sigma^2)$ random variable.
- A bounded random variable. Thus, e.g., Rademacher, Bernoulli, and binomial random variables are all sub-Gaussian.
- A random variable X for which all even moments exist and satisfy

  $$\mathbb{E}[X^{2k}] \leq \frac{(2k)!}{2^k k!} \theta^{2k}, \quad \text{for all } k = 1, 2, 3, \ldots,$$

  for some parameter $\theta \geq 0$. In fact this moment condition is equivalent to sub-Gaussianity.

An example of a random variable that is not sub-Gaussian is a Poisson random variable with any mean μ (this is instead *sub-exponential*, which is a weaker condition).

Next we give several useful bounds for sub-Gaussian random variables. We won't need all of them in our main analysis, but they are simple and important nonetheless (and we may use them in later lectures).

3.1 Tail bound for averages

An important fact about sub-Gaussian random variables is that they admit a Bernstein-type tail bound: if X_i, $i = 1, \ldots, n$ are independent sub-Gaussian random variables, with each X_i having mean zero and variance proxy σ_i^2, then for all $t > 0$,

$$\mathbb{P}(\bar{X}_n \geq t) \leq \exp\left(\frac{-nt^2}{\frac{2}{n}\sum_{i=1}^n \sigma_i^2}\right), \tag{6}$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. We can also get a two-sided bound where we multiply right-hand side above by 2 (since X is mean zero sub-Gaussian with variance proxy σ^2 if and only if $-X$ is).

3.2 Tail bound for maxima

Another useful fact about sub-Gaussian random variables is that maximum of a large number of them sharply concentrates: if X_i, $i = 1, \ldots, n$ are sub-Gaussian random variables, which need not be independent, and each X_i has mean zero and variance proxy σ^2, then for all $t > 0$,

$$\mathbb{P}\Big(\max_{i=1,\ldots,n} X_i \geq \sigma\sqrt{2(\log n + t)}\Big) \leq e^{-t}. \tag{7}$$

It also holds that

$$\mathbb{E}\Big[\max_{i=1,\ldots,n} X_i\Big] \leq \sigma\sqrt{2\log n}. \tag{8}$$

3.3 Tail bound for quadratic forms

A last useful fact we will cite about sub-Gaussian random variables concerns tail concentration for quadratic forms: if X_i, $i = 1, \ldots, n$ are independent sub-Gaussian random variables, with each having mean zero and variance proxy σ^2, then for any positive semidefinite matrix $Q \in \mathbb{R}^{n\times n}$ and all $t > 0$,

$$\mathbb{P}\Big(X^\mathsf{T} Q X \geq \sigma^2\big[\operatorname{tr}(Q) + 2\|Q\|_F\sqrt{t} + 2\|Q\|_{\mathrm{op}}\, t\big]\Big) \leq e^{-t}, \tag{9}$$

where $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is the vector of sub-Gaussian variates, so that $X^\mathsf{T} Q X = \sum_{i,j=1}^n Q_{ij} X_i X_j$, and $\|Q\|_F, \|Q\|_{\mathrm{op}}$ denote the Frobenius and operator norms of Q, respectively. The result we are stating here is from Hsu et al. (2012), which is similar to the Hanson-Wright inequality (but has simple explicit constants). Note in particular that when $Q = \frac{1}{n} I$, we get

$$\mathbb{P}\bigg(\frac{1}{n}\sum_{i=1}^n X_i^2 \geq \sigma^2\big[1 + 2\sqrt{t/n} + 2t/n\big]\bigg) \leq e^{-t}. \tag{10}$$

4 Sub-Gaussian complexity

We now present the first of two powerful tools we'll use from empirical process theory. It will be used to control the supremum of a sub-Gaussian process, indexed by functions $f \in \mathcal{F}$. We call such a quantity

$$\sup_{f\in\mathcal{F}} |\langle \epsilon, f\rangle_n|$$

the *sub-Gaussian complexity* associated with $\mathcal{F}$, based on a sample $x_1, \ldots, x_n$, used to define the empirical inner product. The next result controls this quantity, uniformly over all $x_1, \ldots, x_n$.

Lemma 1 (Adapted from Lemma 8.4 of van de Geer 2000)**.** *Let ϵ_i, $i = 1, \ldots, n$ denote independent sub-Gaussian random variables, each having mean zero variance proxy σ^2. Assume that there exist constants $0 < w < 2$ and $C > 0$ such that for some fixed $x_1, \ldots, x_n$ (which define the empirical norm $\|\cdot\|_n$),*

$$\log N\big(\delta, \mathcal{F}, \|\cdot\|_n\big) \leq C\delta^{-w}, \tag{11}$$

for sufficiently small $\delta > 0$. Then for any fixed $\rho > 0$, there exist constants $n_0, c_0, c_1 > 0$, depending only on σ, ρ, C, w, such that for all $n \geq n_0$ and $\gamma \geq c_0$,

$$\sup_{f\in\mathcal{F}\cap B_n(\rho)} \frac{|\langle \epsilon, f\rangle_n|}{\|f\|_n^{1-w/2}} \leq \frac{\gamma}{\sqrt{n}}, \tag{12}$$

with probability at least $1 - \exp(-c_1\gamma^2)$.

To get a sense of what the result in the lemma gives us, let's rewrite the conclusion in (12) a little differently: it says that uniformly over all $f \in \mathcal{F}$ with $\|f\|_n \leq \rho$, we have

$$|\langle \epsilon, f\rangle_n| \leq \gamma\|f\|_n \frac{\|f\|_n^{-w/2}}{\sqrt{n}},$$

with high probability. Now think about what a simple application of Cauchy-Schwarz would give us:

$$|\langle \epsilon, f\rangle_n| \leq \|\epsilon\|_n \|f\|_n \leq c\|f\|_n,$$

where the second inequality holds with high probability, for some constant $c > 0$, by the result (10) for a quadratic form of sub-Gaussians. If we ignore the leading factors of γ, c, then we see that the second-to-last display is better than the last display when

$$\frac{\|f\|_n^{-w/2}}{\sqrt{n}} \leq 1 \iff \|f\|_n \geq n^{-1/w}.$$

That is, except for functions f of really small empirical norm, we improve on Cauchy-Schwarz, and dramatically so when f is of larger empirical norm. For example, when $\|f\|_n = n^{-1/(2+w)}$ (this is not an arbitrarily chosen rate, you'll see why we're interested in this a bit later, when we do the error analyis), then the "speedup" we get from (12) is:

$$\frac{\|f\|_n^{-w/2}}{\sqrt{n}} = n^{-1/(2+w)}.$$

5 Empirical and population norm coupling

This section states the second of two powerful tools we'll use from empirical process theory. It will be used to control the supremum of difference between empirical and population norms, over all $f \in \mathcal{F}$ for some function class $\mathcal{F}$. Strictly speaking it won't be needed for the result in the main analysis, which bounds the empirical norm error, but we'll use it in the corollary given at the end, on the population norm error.

We need to cover a few more preliminary concepts before stating the result. We say that $\mathcal{F}$ is *star-shaped* if $f \in \mathcal{F}$ implies $\alpha f \in \mathcal{F}$ for all $\alpha \in [0,1]$. We say that $\mathcal{F}$ is *b-bounded* if $\mathcal{F} \subseteq B_\infty(b)$, the sup norm ball of radius b. Lastly and most importantly, we will introduce *localized* notions of Rademacher complexity. For $\delta > 0$, we define the localized empirical Rademacher complexity by

$$\hat{R}_n(\delta) = \mathrm{Rad}(\mathcal{F} \cap B_n(\delta), x_{1:n}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F} \cap B_n(\delta)} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \right],$$

where $B_n(\delta)$ is the empirical norm ball of radius δ. Similarly, we define the localized population Rademacher complexity by

$$R_n(\delta) = \mathrm{Rad}(\mathcal{F} \cap B_2(\delta)) = \mathbb{E}_{x,\sigma} \left[\sup_{f \in \mathcal{F} \cap B_2(\delta)} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(x_i) \right| \right],$$

where $B_2(\delta)$ is the population norm ball of radius δ. We are now ready to state the result, which we informally call a "coupling" between empirical and population norms.

Lemma 2 (Adapted from Theorem 14.1 and Proposition 14.25 of Wainwright 2019)**.** *Let $\mathcal{F}$ be a star-shaped and b-uniformly bounded class of functions for some $b > 0$. Denote by $\hat{\delta}_n$ the smallest positive solution to*

$$\hat{R}_n(\delta) \leq \delta^2/b.$$

and denote by δ_n the smallest positive solution to

$$R_n(\delta) \leq \delta^2/b.$$

Assume $n\delta_n^2 \geq c_0 \log\log(1/\delta_n)$ for a constant $c_0 > 0$. Then there exist constants $a, m_1, m_2, c_1 > 0$ such that with probability at least $1 - a\exp(-n\delta_n^2/(c_0 b))$, both of the following two statements hold:

$$m_1 \delta_n \leq \hat{\delta}_n \leq m_2 \delta_n, \tag{13}$$

$$\Big| \|f\|_n - \|f\|_2 \Big| \leq c_1 \delta_n, \quad \textit{for all } f \in \mathcal{F}. \tag{14}$$

The punchline in the lemma is really (14), which says that the empirical and population norms are uniformly close over all $f \in \mathcal{F}$. In traditional nonparametric applications, we should think of δ_n as being small, scaling as $n^{-\alpha}$ for some $\alpha < 1/2$, in which case the uniform coupling (14) gives us a strong result. Note that for such a scaling on δ_n, we will meet the required assumption $n\delta_n^2 \geq c_0 \log\log(1/\delta_n)$, since

$$n\delta_n^2 = n^{1-2\alpha} \quad \text{and} \quad \log\log(1/\delta_n) = \alpha \log\log n.$$

Backing up somewhat, we call δ_n the *population critical radius* of $\mathcal{F}$, and $\hat{\delta}_n$ the *empirical critical radius* of $\mathcal{F}$. The reason for introducing the latter (which, note, is a random quantity) is that it can be sometimes

easier to bound, using Dudley's entropy integral (5). Then, the result in (13) provides the link between the two—it says that the scaling of the empirical critical radius determines that of the population critical radius, with high probability.

The following calculation demonstrates this connection, and will be useful later on.

Lemma 3. *Assume that $\mathcal{F}$ satisfies the entropy bound (11) for sufficiently small $\delta > 0$ and all $x_1, \dots, x_n$, where $0 < w < 2$ and $C > 0$ are constants. Then the empirical critical radius of $\mathcal{F}$ satisfies $\hat{\delta}_n \leq c_1 n^{-1/(2+w)}$ for a constant $c_1 > 0$. By (13), assuming further that $\mathcal{F}$ is star-shaped and b-bounded, we hence also have $\delta_n \leq c_2 n^{-1/(2+w)}$ with probability at least $1 - \exp(-c_3 n^{2/(2+w)})$, for constants $c_2, c_3 > 0$.*

Proof. By (5), we have

$$\begin{aligned}
\mathrm{Rad}(\mathcal{F} \cap B_n(\delta), x_{1:n}) &\leq \frac{c}{\sqrt{n}} \int_0^\delta \sqrt{\log N(t, \mathcal{F}, \|\cdot\|_n)}\, dt \\
&\leq \frac{\sqrt{C}c}{\sqrt{n}} \int_0^\delta t^{-w/2}\, dt \\
&= \frac{c}{\sqrt{n}} \delta^{1-w/2}.
\end{aligned}$$

In the second line we applied the entropy bound (11), and in the third we simply computed the integral, redefining the constant $c > 0$ as necessary. The smallest positive solution $\hat{\delta}_n$ to $\hat{R}_n(\delta) \leq \delta^2/b$ can therefore by upper bounded by solving

$$\frac{c}{\sqrt{n}} \delta^{1-w/2} = \delta^2/b \iff \delta^{1+w/2} = \frac{(c/b)}{\sqrt{n}},$$

which gives $\hat{\delta}_n \leq c_1 n^{-1/(2+w)}$ for a constant $c_1 > 0$. The statement for δ_n is given by applying (13). □

6 Main analysis

Euipped with these tools, we are now ready to dive into the main analysis. In this section, we will prove the following theorem.

Theorem 1. *Let (x_i, y_i), $i = 1, \dots, n$ be i.i.d. satisfying (1), where each ϵ_i is sub-Gaussian with mean zero and variance proxy $\sigma^2 > 0$, each $x_i \sim Q$, an arbitrary continuous distribution supported on $[0, 1]$, and each $x_i \perp\!\!\!\perp \epsilon_i$. Let J be a seminorm acting on m times weakly differentiable functions, and assume that the following conditions hold for an integer $k \geq 0$, and constants $0 < w < 2$ and $M, C > 0$:*

A1. the null space of J consists of k^{th} order polynomials;

A2. $\sup_{x \in [0,1]} D^m f(x) - \inf_{x \in [0,1]} D^m f(t) \leq M$ for all functions $f \in B_J(1)$;

A3. $\mathcal{F} = B_J(1) \cap B_\infty(1)$ satisfies the entropy bound $\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \leq C\delta^{-w}$, for small enough $\delta > 0$.

To be clear, here we denote by $B_J(1)$ the unit ball in J, and by $B_\infty(1)$ the unit ball in sup norm. Finally, assume that the underlying regression function satisfies $1 \leq J(f_0) < \infty$.

Then there exists constants $c_0, c_1, c_2, c, n_0 > 0$ that depend only on σ, k, M, C, w such that for all $n \geq n_0$ and $\gamma \geq c_0$, the following holds. Fix any fixed exponent $v > 2w/(2+w)$. Any solution $\hat{f}$ to the problem

$$\underset{f}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda J^v(f), \tag{15}$$

with tuning parameter value $\gamma n^{-2/(2+w)} J(f_0)^{2w/(2+w)-v} \leq \lambda \leq \sigma^2 J(f_0)^{2-v}$, satisfies

$$\|\hat{f} - f_0\|_n^2 \leq 2\lambda J^v(f_0) \quad \text{and} \quad J(\hat{f}) \leq cJ(f_0), \tag{16}$$

with probability at least $1 - \exp(-c_1\gamma) - \exp(-c_2 n)$.

In particular, for $\lambda = \gamma n^{-2/(2+w)} J(f_0)^{2w/(2+w)-v}$*, it holds that*

$$\|\hat{f} - f_0\|_n^2 \leq 2\gamma n^{-2/(2+w)} J(f_0)^{2w/(2+w)}, \tag{17}$$

with the same probability.

The proof is similar to that for Theorem 10.2 in van de Geer (2000), except we make all the statements finite-sample, following a development similar to that in Theorem 1 of Sadhanala and Tibshirani (2019).

6.1 Remarks

Before delivering the proof in the next subsection, we make several remarks.

- The proof for the estimator defined by the analogous constrained problem

$$\underset{f}{\text{minimize}} \sum_{i=1}^n (y_i - f(x_i))^2 \quad \text{subject to} \quad J(f) \leq t,$$

  is actually much simpler. (We'll see this when we analyze the lasso, in the next lecture.) But penalized estimators are more common in practice.

- Theorem 1 is not the "pinnacle" of what can be achieved with this type of analysis, it's just supposed to be a (relatively simple and clean) example of what can be achieved. For example, we could also use similar techniques to analyze multivariate estimators, such as thin plate splines (in the supercritical regime $2m > d$; see Chapter 10.3 of van de Geer (2000)) and RKHS regression (see Chapter 13.4 of Wainwright 2019).

- Another important generalization that is important to mention is the following. We can extend the result in Theorem 1 to an estimator that is defined by minimization over a set $\mathcal{S}$:

$$\underset{f \in \mathcal{S}}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda J^v(f),$$

  for a set of functions $\mathcal{S}$ of our choosing. The set $\mathcal{S}$ need not contain any solutions to the unrestricted problem in (15). It need not contain the true regression function f_0. It's just something we choose for (say) computational convenience. For example, we may choose $\mathcal{S}$ to be a space of splines, since splines are nice to work with computationally—and again, we do not need it to be true that $\mathcal{S}$ actually contains a solution in (15). Then an extension of the analysis you'll see in the coming subsection will produce a guarantee of the form

$$\|\hat{f} - f_0\|_n^2 \leq \inf_{\bar{f} \in \mathcal{S}} \Big(\|\bar{f} - f_0\|_n^2 + c\lambda \max\{J(f_0), J(\bar{f})\}^v \Big),$$

  with high probability (for some constant $c > 0$). This is often called an *oracle inequality*. The first term in the above $\|\bar{f} - f_0\|_n^2$ can be made small by ensuring that $\mathcal{S}$ is chosen to have good approximation guarantees over $\{f : [0,1] \to \mathbb{R} : J(f) < \infty\}$. For example, this is true, with respect to various seminorms J, if we choose $\mathcal{S}$ to be a space of splines with suitably chosen knots. In such cases, the approximation error term $\|\bar{f} - f_0\|_n^2$ will be of much smaller order than the second term $2\lambda \max\{J(f_0), J(\bar{f})\}^v$, which will scale as $n^{-2/(2+w)}$ when $J(f_0) \asymp 1$ (as seen in (17)).

  This is very useful because it means that we can use splines to approximately solve (2) (regardless of whether or not splines actually solve (2) or not) and we will not incur any loss in the statistical error rate. We will not cover the details, but refer to Theorem 1 and Corollary 1 in Sadhanala and Tibshirani (2019) for an example of such a result. (We'll also derive explicit oracle inequalities in the next lecture on the lasso.)

- The rate in (17) is minimax optimal. This can be shown using Fano's inequality along with a construction that leverages a concept complementary to covering numbers (metric entropy) called *packing numbers*. See Theorem 4 in Sadhanala and Tibshirani (2019) for details.

6.2 Proof

We now prove Theorem 1. By the exact same arguments used earlier to produce (4) from problem (2), we have for our current problem (15) the basic inequality

$$\|\hat{f} - f_0\|_n^2 \leq 2\langle \epsilon, \hat{f} - f_0 \rangle_n + \lambda(J^v(f_0) - J^v(\hat{f})), \tag{18}$$

We break down the rest of the proof into parts, henceforth abbreviating $\hat{J} = J(\hat{f})$ and $J_0 = J(f_0)$.

Localization. In this part, we prove that $\|\hat{f} - f_0\|_n$ is bounded. This is important because it will be enable us to apply Lemma 1. By the sub-Gaussian tail bound in (10), taking $t = n$, we know that

$$\|\epsilon\|_n^2 \leq 5\sigma^2,$$

on an event Ω_1 with probability at least $1 - \exp(-n)$. Hence, returning to (18), using Cauchy-Schwarz and the above bound, we have on Ω_1,

$$\begin{aligned} \|\hat{f} - f_0\|_n^2 &\leq 2\sqrt{5}\sigma \|\hat{f} - f_0\|_n + \lambda(J_0^v - \hat{J}^v) \\ &\leq 2\sqrt{5}\sigma \|\hat{f} - f_0\|_n + \lambda J_0^v. \end{aligned}$$

This is a quadratic inequality of the form $x^2 \leq bx + c$ in $x = \|\hat{f} - f_0\|_n$, so we can upper bound x by the larger of the two roots, $x \leq (b + \sqrt{b^2 + 4c})/2 \leq b + \sqrt{c}$, which gives, on Ω_1,

$$\begin{aligned} \|\hat{f} - f_0\|_n &\leq 2\sqrt{5}\sigma + \sqrt{\lambda J_0^v} \\ &\leq (2\sqrt{5} + 1)\sigma J_0, \end{aligned} \tag{19}$$

where in the second line we use $J_0 \geq 1$ and $\lambda \leq \sigma^2 J_0^{2-v}$. This completes the desired "localization" step.

Bounding the sub-Gaussian complexity. In this part, we focus on bounding the first-term on the right-hand side in (18) using a sub-Gaussian complexity argument. The idea is the same as what we described in the introduction. Let

$$g = \frac{\hat{f} - f_0}{(\hat{J} + J_0)}.$$

By construction, we have $J(g) \leq 1$. Further, from (19), we have $\|g\|_n \leq (2\sqrt{5} + 1)\sigma$ on Ω_1.

We would like to apply Lemma 1 in order to bound $\langle \epsilon, g \rangle_n$, but we need one more step first. Assumption A3 in the theorem statement is about an entropy condition for the class $\mathcal{F} = B_J(1) \cap B_\infty(1)$. But at present, we do not know that g is bounded in sup norm, only in empirical norm. An argument (whose details we do not give) involving orthogonalization with respect to functions in the null space of J, which recall are polynomials of degree k, can be used to show that for constants $q_1, q_2, c_2 > 0$,

$$\|g\|_\infty \leq q_1 J(g) + q_2 \|g\|_n,$$

on an event Ω_2 with probability at least $1 - \exp(-c_2 n)$. See Lemma 7 (and Lemmas 4, 5, and 6 leading up to it) in Sadhanala and Tibshirani (2019). We note that this is why we need Assumptions A1 and A2, and the assumption that the input distribution Q is continous.

Thus, rescaling the definition of g by a constant $c > 0$ as needed,

$$g = \frac{\hat{f} - f_0}{c(\hat{J} + J_0)},$$

we have $J(g) \leq 1$ and $\|g\|_\infty \leq 1$ on $\Omega_1 \cap \Omega_2$. Lemma 1 then says that for constants $c_0, c_1 > 0$ and all $\gamma \geq c_0$ and sufficiently large n,

$$\langle \epsilon, g \rangle_n \leq \frac{\gamma \|g\|_n^{1-w/2}}{\sqrt{n}}$$

on an event $\Omega_1 \cap \Omega_2 \cap \Omega_3$ with probability at least with probability at least $1 - \exp(-c_1\gamma^2) - \exp(-c_2 n)$. Plugging this into the right-hand side of (18) gives

$$\begin{aligned}\|\hat{f} - f_0\|_n^2 &\leq 2c\gamma \frac{(\hat{J} + J_0)}{\sqrt{n}} \|g\|_n^{1-w/2} + \lambda(J_0^v - \hat{J}^v) \\ &= 2c\gamma \frac{(\hat{J} + J_0)^{w/2}}{\sqrt{n}} \|\hat{f} - f_0\|_n^{1-w/2} + \lambda(J_0^v - \hat{J}^v), \end{aligned} \tag{20}$$

on the event $\Omega = \Omega_1 \cap \Omega_2 \cap \Omega_3$, where we have redefined the constant c as needed.

Transforming to squared empirical norm. The next part is to transform (20) so that we remove the fractional exponent on the empirical norm error term $\|\hat{f} - f_0\|_n$, and end up with only squared empirical norm terms. First, we use the following inequality that holds for any $a, b \geq 0$, and any w,

$$ab^{1-w/2} \leq a^{1/(1+w/2)} b + a^{2/(1+w/2)}.$$

Applying this to the first term on the right-hand side in (20) with $a = (\hat{J} + J_0)^{w/2}/\sqrt{n}$ and $b = \|\hat{f} - f_0\|_n$, and abbreviating $r_n = n^{-1/(2+w)}$, yields

$$\|\hat{f} - f_0\|_n^2 \leq 2c\gamma r_n (\hat{J} + J_0)^{w/(2+w)} \|\hat{f} - f_0\|_n + 2c\gamma r_n^2 (\hat{J} + J_0)^{2w/(2+w)} + \lambda(J_0^v - \hat{J}^v),$$

on Ω. At this point, we could recognize the above as a quadratic of the form $x^2 \leq bx + c$ in $x = \|\hat{f} - f_0\|_n$, and proceed as we did in the localization step, but we take a different approach that leads to a slightly sharper dependence on various problem parameters. We apply $2ab \leq a^2 + b^2$ to the first term on the right-hand side of the above display, with $a = \sqrt{2}c\gamma r_n (\hat{J} + J_0)^{w/(2+w)}$ and $b = \|\hat{f} - f_0\|_n/\sqrt{2}$, which yields

$$\|\hat{f} - f_0\|_n^2 \leq \frac{1}{2}\|\hat{f} - f_0\|_n^2 + \gamma r_n^2 (\hat{J} + J_0)^{2w/(2+w)} + \lambda(J_0^v - \hat{J}^v),$$

on Ω, where we have redefined γ as needed. We have been careful to end up with a factor of the squared empirical norm on the right-hand side with a leading constant less than 1. Subtracting this term $\frac{1}{2}\|\hat{f} - f_0\|_n^2$ to the left-hand side gives, on Ω,

$$\frac{1}{2}\|\hat{f} - f_0\|_n^2 \leq \gamma r_n^2 (\hat{J} + J_0)^{2w/(2+w)} + \lambda(J_0^v - \hat{J}^v). \tag{21}$$

Bounding the achieved penalty term. The next part is to bound the achieved penalty term $\hat{J}$. Starting from (21), we can simply lower bound the left-hand side by zero, and rearrange, yielding

$$\hat{J}^v \leq \frac{\gamma r_n^2}{\lambda} (\hat{J} + J_0)^{2w/(2+w)} + J_0^v,$$

on Ω. Suppose $\hat{J} > 2J_0$. Then upper bounding the first term on right-hand side above,

$$\hat{J}^v \leq c\frac{\gamma r_n^2}{\lambda} \hat{J}^{2w/(2+w)} + \frac{1}{2}\hat{J}^v,$$

on Ω, for a constant $c > 0$. Subtracting $\frac{1}{2}\hat{J}^v$ to the left-hand side, we learn that provided $\lambda \geq \gamma r_n^2$,

$$\frac{1}{2}\hat{J}^v \leq c\hat{J}^{2w/(2+w)}, \tag{22}$$

on Ω. Since $v > 2w/(2+w)$, this means that $\hat{J} \leq c$ on Ω, where we redefine the constant c as necessary. Recall that this was established in the case $\hat{J} > 2J_0$. Hence, altogether, we learn that $\hat{J} \leq 2J_0 + c \leq cJ_0$ on Ω, redefining c once again as needed.

The home stretch: choosing λ. Returning to (21), and using our penalty bound from (22), we have

$$\|\hat{f} - f_0\|_n^2 \leq \gamma r_n^2 J_0^{2w/(2+w)} + \lambda J_0^v,$$

on Ω, redefining γ as needed. Choosing $\lambda \geq \gamma r_n^2 J_0^{2w/(2+w)-v}$ establishes (16), and completes the proof.

6.3 Corollary: population error

Using the coupling between empirical and population norms, we can also bound the population error as a corollary to Theorem 1.

Corollary 1. *Under the same conditions and notation as in Theorem 1, there exists constants $c_3, c_4 > 0$ such that for all $n \geq n_0$ and $\gamma \geq c_0$,*

$$\|\hat{f} - f_0\|_2^2 \leq 4\lambda J^v(f_0) + c_3 n^{-2/(2+w)} J^2(f_0), \tag{23}$$

with probability at least $1 - \exp(-c_1\gamma) - \exp(-c_4 n^{2/(2+w)})$.

Proof. Let $g = (\hat{f} - f_0)/((c+1)J_0)$. Then, as in the proof of Theorem 1, we have $J(g) \leq 1$ and $\|g\|_\infty \leq b$ for a constant $b > 0$, on the event Ω. By Lemma 3 applied to $\mathcal{F} = B_J(1) \cap B_\infty(b)$, the population critical radius satisfies $\delta_n \leq c_3 n^{-1/(2+w)}$ with probability at least $1 - \exp(-c_1\gamma) - \exp(-c_2 n) - \exp(-c_4 n^{2/(2+w)})$, for constants $c_3, c_4 > 0$. Applying Lemma 2—specifically, applying (14) to g, we see that

$$\|\hat{f} - f_0\|_2 \leq \|\hat{f} - f_0\|_n + c_3 n^{-1/(2+w)} J_0,$$

with probability at least $1 - \exp(-c_1\gamma) - \exp(-c_2 n) - \exp(-q_2 n^{2/(2+w)})$, where we adjust c_3 as necessary. The result in (23) follows by squaring both sides in the above display, using (16), and adjusting c_3, c_4 once again as needed. □

A remark: choosing $\lambda = \gamma n^{-2/(2+w)} J(f_0)^{2w/(2+w)-v}$, we see from (23) that when $J(f_0) \asymp 1$ we are able to bound the population error at the same rate $n^{-2/(2+w)}$ as the empirical error. But when $J(f_0)$ is growing with n, the population error rate established by the above corollary is worse. This is likely an artifact of the proof strategy.

6.4 Corollary: smoothing splines

Finally, we apply Theorem 1 and Corollary 1 to smoothing splines, where $J(f) = \int_0^1 (D^m f)^2(x)\,dx$, in order to get the following result.

Corollary 2. *Let (x_i, y_i), $i = 1, \ldots, n$ be i.i.d. satisfying (1), where each ϵ_i is sub-Gaussian with mean zero and variance proxy $\sigma^2 > 0$, each $x_i \sim Q$, an arbitrary continuous distribution supported on $[0,1]$, and each $x_i \perp\!\!\!\perp \epsilon_i$. Fix an integer $m \geq 1$ and assume that the underlying regression function satisfies*

$$1 \leq J_0^2 = \int_0^1 (D^m f_0)^2(x)\,dx < \infty.$$

Then there exists constants $c_0, c_1, c_2, c_3, c_4, c, n_0 > 0$ that depend only on σ, m such that for all $n \geq n_0$ and $\gamma \geq c_0$, the smoothing spline estimator $\hat{f}$, defined by solving

$$\underset{f}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int_0^1 (D^m f)^2(x)\,dx,$$

for any tuning parameter value $\gamma n^{-2m/(2m+1)} J_0^{-4m/(2m+1)} \lambda \leq \sigma^2$, satisfies

$$\|\hat{f} - f_0\|_n^2 \leq 2\lambda J_0^2 \quad \text{and} \quad \int_0^1 (D^m \hat{f})^2(x)\,dx \leq c J_0^2,$$

with probability at least $p_n = 1 - \exp(-c_1\gamma) - \exp(-c_2 n)$, and

$$\|\hat{f} - f_0\|_2^2 \leq 4\lambda J_0^2 + c_3 n^{-2m/(2m+1)} J_0^2,$$

with probability at least $q_n = 1 - \exp(-c_1\gamma) - \exp(-c_4 n^{2m/(2m+1)})$. In particular, for a choice of tuning parameter $\lambda = \gamma n^{-2m/(2m+1)} J_0^{-4m/(2m+1)}$, it holds that

$$\|\hat{f} - f_0\|_n^2 \leq 2\gamma n^{-2m/(2m+1)} J_0^{2/(2m+1)}, \quad \text{and} \quad \|\hat{f} - f_0\|_2^2 \leq n^{-2m/(2m+1)}(4\gamma J_0^{2/(2m+1)} + c_3 J_0^2),$$

with the same probabilities as above: p_n and q_n, respectively.

Proof. We must simply check the conditions A1–A3 in Theorem 1 for the choice $J(f) = \int_0^1 (D^m f)^2(x)\,dx$. It is straightforward to see that A1 and A2 are satisfied with $k = m - 1$. The condition A3 is satisfied for $w = 1/m$, as cited in Table 1, which is a result due to Birman and Solomyak (1967). The rest is just given by reading off the results of Theorem 1 and Corollary 1 for $w = 1/m$. □

References

M. S. Birman and M. Z. Solomyak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Mathematics of the USSR-Sbornik*, 73(115):331–335, 1967. In Russian.

Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012.

Enno Mammen and Sara van de Geer. Locally adaptive regression splines. *Annals of Statistics*, 25(1): 387–413, 1997.

Veeranjaneyulu Sadhanala and Ryan J. Tibshirani. Additive models via trend filtering. *Annals of Statistics*, 47(6):3032–3068, 2019.

Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.