CS 839 Probability and Learning in High Dimension Lecture 16–17 - 03/30, 04/04/2022

Lecture 16–17 : Non-parametric Least Squares

Lecturer: Yudong Chen *Scribe: Chenyang Jiang, Dekun Zhou*

In this lecture,[1] we will introduce the problem setting and examples of non-parametric least squares. We will also focus on bounding the denoised error.

1 Recap

In the previous lecture, we shows that when we compute an estimator using ERM, the test error can be bound by the sum of two terms: the approximation error, which is the best function in your function class that minimizes the test error, and the generalization gap, which can be further upper bounded by some complexity measure of the function class (iIn the last lecture we consider the Rademacher complexity).

$$\begin{array}{lllll} \text{test error} & \leq & \text{approximation error} & + & R_n(\mathcal{F}) \\ \mathbb{E}L(\hat{f}) & & L(f_0) & & \mathbb{E}\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)| \end{array} \tag{1}$$

Usually there is a trade-off between these two terms, if the function class is large, then the approximation error will be smaller, but a large function class will lead to large complexity.

We remark that the test error $\mathbb{E}L(\hat{f})$ involves two expectations. The outer expectation is with respect to $\hat{f}$. This is the expectation with respect to randomness in the training data. On the other hand, recall that $L(\hat{f}) = \mathbb{E}_{(x,y)\sim\mu}[(\hat{f}(x) - y)^2]$. This is the expectation with respect to a new test data point. We have two sources of randomness here.

The bound (1) is sometimes not tight. In particular, when we measure the complexity of the function class, we take the supreme over the entire function class. This is sometimes pretty loose, because what we really care about is a particular function $\hat{f}$ in the function class. Today, we will talk about a refinement of this kind of bound, which involves a more refined notion of complexity measure. In particular, we take the supreme over functions only in a neighborhood of f^*:

$$\mathbb{E} \sup_{\substack{f\in\mathcal{F} \\ \|f-f^*\|\leq\delta}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right|.$$

This is called the localized Rademacher complexity.

In many problems, using this localized complexity will give a tighter control on the generalization gap. The high level intuition is that we expect $\hat{f}$ to be not too far away from f^*, so we only need to take the supremum over functions that are close to f^*.

We mention in passing that the above bound may still be insufficient for some more complicated problems (e.g., modern, overparametrized neural networks). Obtaining good bounds for these problems requires understanding (explicit/implicit) regularization as well as the optimization aspect. The previous bound in (1) assumes that we can find the ERM $\hat{f}$; for many problems, solving for $\hat{f}$ involves a highly nontrivial optimization problem. For these more challenging problems, we probably cannot separately look at the approximation error, generalization gap and optimization error. Rather, we would need to jointly study these three components and understand the interplay between them.

[1] *Reading:* Section 13.1 and 13.2 in [Wainwright, 2019].

2 Setup

In today's lecture, we consider the following nonparametric least square problem. Suppose we observe (x_i, y_i), $i = 1, 2..., n$, where $y_i = f^*(x_i) + \sigma w_i$. Here:

1. f^* is the unknown ground-truth regression function.
2. $x_i \in \mathcal{X}$ is feature/covariate vector;
3. $y_i \in \mathbb{R}$: is the response;
4. $w_i \overset{i.i.d}{\sim} N(0,1)$ is the additive noise;
5. σ^2 is the noise variance

The setting is almost the same as in the previous lecture, with the only difference that we have noise w_i, whose magnitude is control by the noise variance σ^2.

To estimate the unknown regression function f^*, we consider the empirical risk minimizer (ERM), which is given by

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2. \tag{2}$$

2.1 Examples

We discuss some concrete examples of the above problem with different choices of the function class $\mathcal{F}$.

2.1.1 Linear Regression

Consider the function class

$$\mathcal{F}_C = \{x \mapsto \langle \theta, x \rangle : \theta \in C \subseteq \mathbb{R}^d\}.$$

Here the regression vector θ takes on values from the set C.

- If we take $C = \{\theta \in \mathcal{R}^d : \|\theta\|_2 \leq R\}$, then the ERM problem (2) becomes Ridge Regression, which constrains the ℓ_2-norm of θ.
- More generally, we may take $C = \{\theta \in \mathbb{R}^d : \sum_{j=1}^d |\theta_j|^q \leq R_q\}$, where $0 \leq q \leq 2$ and R_q is a given number. This is called ℓ_q-regression. If $q \in [1, 2]$, then the set C is convex and corresponds to the ball of the ℓ_q norm. If $q \in (0, 1]$, then the sec C is non-convex. When $q = 0$, then $\sum_{j=1}^d |\theta_j|^q$ equals the number of non-zero coordinate in the vector θ, so C is the set of R_0 sparse vectors in $\mathbb{R}^d$. In this case, the ERM problem (2) becomes the sparse regression problem.

This is an example where $\mathcal{F}$ is a parametric function class that f is defined by a finite dimensional parameter θ.

In the next few examples, we consider non-parametric function classes.

2.1.2 Lipschitz Regression

Consider the function class

$$\mathcal{F}_{\text{Lip}}(L) = \{f : [0, 1] \mapsto \mathbb{R}, f(0) = 0, f \text{ is } L\text{-Lipschitz}\}.$$

This is a non-parametric function class. With this function class, the ERM (2) problem appears to be an infinite dimensional optimization problem. It turns out that an optimal solution $\hat{f}$ for this problem can be taken to be a piece-wise linear function.

2.1.3 Convex Regression

Consider the function class

$$\mathcal{F}_{\text{conv}} = \{f : f \text{ is convex on } \mathbb{R}^d\}.$$

With this function class, the apparently infinite-dimensional optimization problem (2) can be converted to a finite dimension problem as follows.

Step 1: Solve the following Quadratic Program (QP)

$$\min_{\{\hat{y}_i, \hat{g}_i\}_{i=1}^n \in \mathbb{R}\times\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2 \qquad \textbf{s.t.} \quad \hat{y}_j \geq \hat{y}_i + \langle \hat{g}_i, x_j - x_i \rangle \ \forall i, j.$$

One can interpret the optimization variables above as follows: $\hat{y}_i = \hat{f}(x_i)$ and $\hat{g}_i$ is similar to $\nabla \hat{f}(x_i)$. In the above QP, we try to minimize the squared error for fitting the training data, subject to constraint that the function values must come from a convex function.

Step 2: Given the solution $\{\hat{y}_i, \hat{g}_i\}_{i=1}^n$ obtained above, we define the function $\hat{f}$ over the entire domain via interpolation: for each x, set

$$\hat{f}(x) = \max_{i=1,2\ldots n} \{\hat{y}_i + \langle \hat{g}_i, x - x_i \rangle\}.$$

Note that if we evaluate the function at the data point x_i, the function value satisfies $\hat{f}(x_i) = \hat{y}_i$.

2.1.4 Cubic Smoothing Splines

Consider the function class

$$\mathcal{F}(R) = \left\{ f : [0,1] \mapsto \mathbb{R}, \int_0^1 (f''(x))^2 \, \mathrm{d}x \leq R \right\}.$$

The constraint on integral of second derivative encourages the function to be smooth, so this is a form of smoothness constraint.

With this function class, one can show that an optimal solution of the ERM problem (2) can be taken to be a natural cubic spline, which is a piecewise cubic polynomial function, where different pieces are connected in a smooth way (i.e., the adjacent pieces have matching function value and first two derivatives at the connection point).

3 Error Bounds

In this section, we establish general error bounds for the ERM $\hat{f}$ for the non-parametric least-squares problems.

3.1 Notations and Assumptions

Define the shifted function class as follow:

$$\mathcal{F}^* := \mathcal{F} - f^* = \{f - f^* : f \in \mathcal{F}\}.$$

We assume that $\mathcal{F}^*$ is *star-shaped*, meaning that

$$\forall h \in \mathcal{F}^*, \ \forall \alpha \in [0,1] : \quad \alpha h \in \mathcal{F}^*.$$

If $\mathcal{F}^*$ is star-shaped, we must have $0 \in \mathcal{F}^*$, which is equivalent to $f^* \in \mathcal{F}$. One observation is that if the function class $\mathcal{F}^*$ is convex, then it is star-shaped. The converse is not true in general.

Recall that $(x_i, y_i), i = 1, \dots, n$ are the data points, where $y_i = f^*(x_i) + \sigma w_i$. In the sequel, we consider the feature vectors $x_i, i = 1, \dots, n$ as fixed (this is called the *fixed design* setting). In this case, the only randomness comes from noise w_i. For a function g, define the empirical ℓ_2 norm as

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

Our goal is to control the following quantity

$$\|\hat{f} - f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^n \left(\hat{f}(x_i) - f^*(x_i)\right)^2.$$

Note that this quantity can be viewed as a denoised version of training error. (If we replace $f^*(x_i)$ above by y_i, then the quantity becomes the training error.)

Remark Using techniques from last lecture, we can bound the difference between the denoised training error and the test error $\mathbb{E}_{x\sim\mu}(\hat{f}(x) - f^*(x))^2$, which is what we ultimately care about. Today we will focus on bounding the denoised training error.

3.2 Localized Gaussian Complexity

For a given radius parameter $\delta > 0$. define localized Gaussian complexity of $\mathcal{F}^*$ as follow:

$$G_n(\delta; \mathcal{F}^*) := \mathbb{E}_w \left[\sup_{g\in\mathcal{F}^* \ \|g\|_n \le \delta} \left|\frac{1}{n}\sum_{i=1}^n w_i g(x_i)\right|\right],$$

where $w_i \overset{i.i.d}{\sim} N(0,1)$.

Define critical radius as follow:

$$\delta^* := \min_{\delta > 0}\left\{\delta : \frac{G_n(\delta; \mathcal{F}^*)}{\delta} \le \frac{\delta}{2\sigma}\right\}.$$

We will later show that δ^* provides an upper bound for the denoised training error. Before doing so, we first show that δ^* is well-defined.

Lemma 1. *If $\mathcal{F}^*$ is star-shaped, then the function $\delta \mapsto \frac{G_n(\delta;\mathcal{F}^*)}{\delta}$ is non-increasing on $(0, \infty)$. Hence δ^* exists and is finite.*

Proof For any $0 < \delta < t$, we want to show that $\frac{G_n(t,\mathcal{F}^*)}{t} \le \frac{G_n(\delta;\mathcal{F}^*)}{\delta}$.

Given $h \in \mathcal{F}^*$ with $\|h\|_n \le t$, define the rescaled function $\tilde{h} = \frac{\delta}{t}h$. We have $\tilde{h} \in \mathcal{F}^*$ by definition with $\|h\|_n \le \delta$. It is easy to see that

$$\frac{1}{n}\left(\frac{\delta}{t}\sum_{i=1}^n w_i h(x_i)\right) = \frac{1}{n}\sum_{i=1}^n w_i \tilde{h}(x_i).$$

Taking the supreme and expectation on both side over h, we obtain that

$$\frac{\delta}{t}\mathbb{E}\left[\sup_{h\in\mathcal{F}^*:\|h\|_n\le t}\frac{1}{n}\sum_{i=1}^n w_i h(x_i)\right] \le \mathbb{E}\left[\sup_{\tilde{h}\in\mathcal{F}^*:\|\tilde{h}\|_n\le\delta}\frac{1}{n}\sum_{i=1}^n w_i\tilde{h}(x_i)\right].$$

This is equivalent to desired inequality

$$\frac{G_n(t,\mathcal{F}^*)}{t} \le \frac{G_n(\delta,\mathcal{F}^*)}{\delta}$$

by definition of the localized Gaussian complexity. □

Remark If we take $\delta = t/2$ in the above proof, then Lemma 1 gives

$$G_n\big(t/2, \mathcal{F}^*\big) \geq \frac{1}{2} G_n(t, \mathcal{F}^*).$$

Therefore, Lemma 1 can be interpreted as saying that when the radius goes to zero, the localized Gaussian complexity decays more slowly than (or as slow as) linear.

3.3 Master Error Bound

With the above notations, we are ready to state our main theorem, which bounds the error $\|\hat{f} - f^*\|_n^2$ in terms of the critical radius of the localized Gaussian complexity.

Theorem 1. *Suppose $\mathcal{F}^*$ is star-shaped and δ^* is defined above. For each $t \geq \delta^*$, with probability at least $1 - \exp\big(-\frac{nt\delta^*}{2\delta^2}\big)$, it holds that*

$$\|\hat{f} - f^*\|_n^2 \leq 16t\delta^*.$$

In the above theorem, one usually takes the smallest possible t, which is δ^*. In this case, we have $\|\hat{f} - f^*\|_n^2 \leq 16\delta^{*2}$. One thing we need to check is that the probability $1 - e^{\frac{-nt\delta^*}{2\delta^2}}$ is indeed close to 1.

The proof of Theorem 1 depends on the following lemma:

Lemma 2. *For each $u \geq \delta^*$, define the following "bad" event $A(u)$ as follows:*

$$A(u) := \left\{ \exists g \in \mathcal{F}^* \cap \{\|g\|_n \geq u\} : \left| \frac{\sigma}{n} \sum_{i=1}^n w_i g(x_i) \right| \geq 2\|g\|_n u \right\}.$$

Then, $\mathbb{P}(A(u)) \leq \exp\{-nu^2/(2\sigma^2)\}$.

The proof of this lemma makes use of several techniques we learned in the previous lectures.

Proof of Lemma 2 We see that

$$\mathbb{P}(A(u)) = \mathbb{P}\left(\sup_{g \in \mathcal{F}^*, \|g\|_n \geq u} \frac{1}{\|g\|_n} \left| \frac{\sigma}{n} \sum_{i=1}^n w_i g(x_i) \right| \geq 2u \right) \leq \mathbb{P}\left(\sup_{g \in \mathcal{F}^*, \|g\|_n = u} \left| \frac{\sigma}{n} \sum_{i=1}^n w_i g(x_i) \right| \geq 2u^2 \right),$$

where the inequality is due to the fact that we further restrict to $\|g\|_n = u$. Define the random variable $Z(u) := \sup_{g \in \mathcal{F}^*, \|g\|_n = u} |\frac{\sigma}{n} \sum_{i=1}^n w_i g(x_i)|$. We upper bound the probability of the event $\{Z(u) \geq 2u^2\}$ as follows:

- (Concentration) We observe that $Z(u)$ is a Lipschitz function of $w_1, \cdots, w_n$, with Lipschitz constant upper bounded by $\frac{\sigma}{n} \sup_{\|g\|_n = u} \|(g(x_1), \cdots, g(x_n))\|_2 = \frac{\sigma}{n} \cdot u\sqrt{n} = \frac{\sigma u}{\sqrt{n}}$. Hence, by Gaussian Lipschitz concentration inequality in Lecture 12, we obtain that

$$\mathbb{P}(Z(u) \geq \mathbb{E}Z(u) + u^2) \leq \exp\{-u^4 n/(2\sigma^2 u^2)\} = \exp\{-nu^2/(2\sigma^2)\}.$$

- (Expectation) By definition of $G_n(\delta; \mathcal{F}^*)$, we see that

$$\mathbb{E}\, Z(u) \overset{\text{(i)}}{\leq} \sigma G_n(u; \mathcal{F}^*) = \sigma u \cdot \frac{G_n(u; \mathcal{F}^*)}{u} \overset{\text{(ii)}}{\leq} \sigma u \cdot \frac{G_n(\delta^*; \mathcal{F}^*)}{\delta^*} \overset{\text{(iii)}}{\leq} \sigma u \cdot \frac{\delta^*}{2\sigma} \leq u\delta^*,$$

where inequality (i) is due to the fact that we constrain on $\|g\|_n = u$ in the definition of $Z(u)$, but $\|g\|_n \geq u$ in the definition of $G_n(u; \mathcal{F}^*)$, inequality (ii) holds since $u \geq \delta^*$ and $\delta \mapsto \frac{G_n(\delta; \mathcal{F}^*)}{\delta}$ is a non-increasing function, and inequality (iii) follows from the definition of δ^*.

Combining these two steps, we obtain that

$$\begin{aligned}\mathbb{P}(Z(u) \geq u^2 + u^2) &\leq \mathbb{P}(Z(u) \geq u\delta^* + u^2)\\ &= \mathbb{P}(Z(u) - \mathbb{E}\, Z(u) + \mathbb{E}\, Z(u) \geq u^2 + u\delta^*)\\ &\leq \mathbb{P}(Z(u) - \mathbb{E}\, Z(u) \geq u^2)\\ &\leq \exp\{-nu^2/(2\sigma^2)\}\end{aligned}$$

as claimed. □

Now we are ready to prove Theorem 1.

Proof of Theorem 1: Since $\hat{f}$ is optimal to the ERM problem (2) and $f^* \in \mathcal{F}$ is feasible, we have

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

Also recall that

$$y_i = f^*(x_i) + \sigma w_i, \quad 1 \leq i \leq n.$$

We plug this expression into y_i's in equation (3), open the squares and rearrange terms. Doing so gives the "basic inequality"

$$\frac{1}{2}\|\hat{f} - f^*\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^n w_i(\hat{f}(x_i) - f^*(x_i)) \tag{4}$$

Introducing the shorthand $\Delta := \hat{f} - f^* \in \mathcal{F}^*$, we rewrite the above basic inequality compactly as

$$\frac{1}{2}\|\Delta\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^n w_i \Delta(x_i). \tag{5}$$

Then from Lemma 2, we obtain that for each $t \geq \delta^*$,

$$\begin{aligned}\mathbb{P}(A(\sqrt{t\delta^*})^c) &= \mathbb{P}\left(\forall g \in \mathcal{F}^* \cap \{\|g\|_n \geq \sqrt{t\delta^*}\} : \left|\frac{\sigma}{n}\sum_{i=1}^n w_i g(x_i)\right| \leq 2\|g\|_n\sqrt{t\delta^*}\right)\\ &\geq 1 - \exp\{-nt\delta^*/(2\sigma^2)\}.\end{aligned}$$

We can then conclude our proof by discussing the magnitude of $\|\Delta\|_n$:

1. If $\|\Delta\|_n \leq \sqrt{t\delta^*}$, then we are done because $\|\Delta\|_n^2 = \|\hat{f} - f^*\|_n^2 \leq t\delta^* < 16t\delta^*$.
2. Otherwise, if $\|\Delta\|_n > \sqrt{t\delta^*}$, then conditioning on the "good" event $A(\sqrt{t\delta^*})^c$, we see that

   $$\left|\frac{\sigma}{n}\sum_{i=1}^n w_i\Delta(x_i)\right| \leq 2\|\Delta\|_n\sqrt{t\delta^*}.$$

   Combining with basic inequality (5) proves that $\|\Delta\|_n^2 \leq 16t\delta^*$.

□

We briefly explain why the localized Gaussian complexity provides tighter bounds. The simplest way to bound the last RHS of the basic inequality (4) is by using a worst case bound that involves the supremum over the entire function class $\mathcal{F}^*$:

$$\frac{\sigma}{n}\sum_{i=1}^n w_i(\hat{f}(x_i) - f^*(x_i)) \leq \sigma \sup_{g \in \mathcal{F}^*} \frac{1}{n}\sum_{i=1}^n w_i g(x_i).$$

Taking the expectation of the RHS gives the (global) Gaussian complexity. This is, however, not tight. What we really care about is the ERM $\hat{f}$. We expect that $\hat{f}$ is be close to f^*, so we only need to take supreme over a smaller ball that contains $\hat{f}$. The proof of Theorem 1 essentially establishes that for any δ that satisfies $\frac{G_n(\delta;\mathcal{F})}{\delta} \leq \frac{\delta}{2\sigma}$, the radius-$\delta$ ball centered at f^* is large enough to contain $\hat{f}$. The critical radius δ^* is the smallest radius of such a ball, so we use δ^* to get the best bound.

3.4 Controlling the Critical Radius

In this subsection, we provide a way to upper bound the critical radius δ^*. For simplicity, we define a ball in $\mathcal{F}^*$ and its covering number, both with respect to $\|\cdot\|_n$:

Definition 1. $B_n(\delta) := \{h \in \mathcal{F}^* : \|h\|_n \leq \delta\}$.

Definition 2. *The covering number of $B_n(\delta)$ with accuracy t and metric $\|\cdot\|_n$ is denoted as $N_\delta(t) := N(t, B_n(\delta), \|\cdot\|_n)$.*

The following theorem translates the problem of bounding δ^* into that of controlling the covering number/metric entropy.

Theorem 2. *If $\mathcal{F}^*$ is star-shaped, and $\delta \in (0, \sigma]$ satisfies the following condition:*

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_\delta(t)}\,\mathrm{d}t \leq \frac{\delta^2}{4\sigma}, \tag{6}$$

then we have $\delta^* \leq \delta$.

Proof Fix a $\delta \in (0, \sigma]$ that satisfies (6). Note that $\frac{\delta^2}{4\sigma} < \delta$ for all $\delta \in (0, \sigma]$. Let $\{g^1, \cdots, g^M\}$ be a minimal $\frac{\delta^2}{4\sigma}$-covering of $B_n(\delta)$. Therefore, for any $g \in B_n(\delta)$, there exists some $j \in [M]$ such that $\|g^j - g\|_n \leq \frac{\delta^2}{4\sigma}$. Introduce that shorthand $g(x_1^n) := (g(x_1), \ldots, g(x_n)) \in \mathbb{R}^n$. We have

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n w_i g(x_i)\right| &= \left|\frac{1}{n}\langle w, g(x_1^n)\rangle\right| \\
&\leq \left|\frac{1}{n}\langle w, g^j(x_1^n)\rangle\right| + \left|\frac{1}{n}\langle w, g(x_1^n) - g^j(x_1^n)\rangle\right| \\
&\leq \max_{j\in[M]} \left|\frac{1}{n}\langle w, g^j(x_1^n)\rangle\right| + \sqrt{\frac{\|w\|_2^2}{n}} \cdot \sqrt{\frac{\|g(x_1^n) - g^j(x_1^n)\|_2^2}{n}} \\
&\leq \max_{j\in[M]} \left|\frac{1}{n}\langle w, g^j(x_1^n)\rangle\right| + \frac{\|w\|_2}{\sqrt{n}} \cdot \frac{\delta^2}{4\sigma}.
\end{aligned} \tag{7}$$

Next, we upper bound the expectation in the first RHS term by using Dudley's integral bound with a slightly smarter look at the bounds we previously had. Define the random variable

$$Z(g^j) := \frac{1}{\sqrt{n}} \sum_{i=1}^n w_i g^j(x_i)$$

for $j \in [M]$. Note that $Z(g^j)$ is zero-mean and sub-Gaussian with metric $\rho(g^j, g^k) = \|g^j - g^k\|_n$. Since $\{g^1, \cdots, g^M\}$ is a minimal $\frac{\delta^2}{4\sigma}$-covering of $B_n(\delta)$, we do not need to extend the chaining to smaller than a resolution of $\frac{\delta^2}{4\sigma}$, as at that resolution we can uniquely identify each point. Furthermore, we also only need to start the chaining at a resolution of δ, as the set $B_n(\delta)$ has a diameter of 2δ. Combining all these and working through the arithmetic of the chaining argument, we get that, for every $\delta \in (0, \sigma]$ satisfying (6),

$$\begin{aligned}\mathbb{E} \max_{j \in [M]} \left| \frac{1}{n} \langle w, g^j(x_1^n) \rangle \right| &= \mathbb{E} \max_{j \in [M]} \left| \frac{Z(g^j)}{\sqrt{n}} \right| \\ &\leq \frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_\delta(t)} \mathrm{d}t \qquad (8) \\ &\leq \frac{\delta^2}{4\sigma}, \qquad (9)\end{aligned}$$

where for the first inequality we used a version of Dudley's integral bound that includes explicit constants, and in the last inequality we used the assumption (6). Combining with the inequality (7), we obtain

$$\begin{aligned}G_n(\delta; \mathcal{F}^*) &= \mathbb{E} \sup_{g \in \mathcal{F}^*, \|g\|_n \leq \delta} \left| \frac{1}{n} \sum_{i=1}^{n} w_i g(x_i) \right| \\ &\leq \mathbb{E} \max_{j \in [M]} \left| \frac{1}{n} \langle w, g^j(x_1^n) \rangle \right| + \mathbb{E} \frac{\|w\|_2}{\sqrt{n}} \cdot \frac{\delta^2}{4\sigma} \\ &\leq \frac{\delta^2}{4\sigma} + \frac{\delta^2}{4\sigma} = \frac{\delta^2}{2\sigma},\end{aligned}$$

where the last inequality follows from (9) and the fact that $\mathbb{E} \|w\|_2 \leq \sqrt{\mathbb{E} \|w\|_2^2} = \sqrt{n}$. Since δ^* is the smallest δ that satisfies the above inequality, we obtain that $\delta^* \leq \delta$. □

Remark In this remark we point out that, for the term $\mathbb{E} \max_{j \in [M]} |\frac{1}{n} \langle w, g^j(x_1^n) \rangle|$, Gaussian maxima bound is worse than the Dudley's integral bound. Indeed, from Gaussian maxima, we obtain that

$$\mathbb{E} \max_{j \in [M]} \left| \frac{1}{n} \langle w, g^j(x_1^n) \rangle \right| \lesssim \frac{\delta}{\sqrt{n}} \sqrt{\log N_\delta\Big(\frac{\delta^2}{4\sigma}\Big)},$$

where we use the fact that for each $j \in [M]$, the random variable $\langle w, g^j(x_1^n) \rangle$ is Gaussian with variance upper bounded by $n \|g^j\|_n^2 \leq n\delta^2$. The above bound is worse than the bound (8) obtained using Dudley's integral, as illustrated in Figure 1.

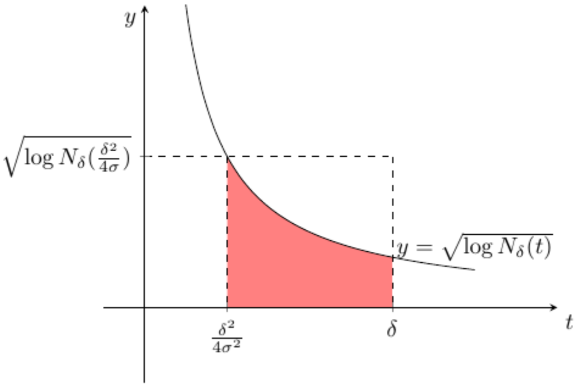
y
$\sqrt{\log N_\delta(\frac{\delta^2}{4\sigma})}$
$y = \sqrt{\log N_\delta(t)}$
$\frac{\delta^2}{4\sigma^2}$
δ
t

Figure 1: Illustration of Gaussian maxima bound versus Dudley's integral bound. The shaded area in red represents the Dudley integral bound, while the area of the rectangle defined by the t-axis, y-axis, $y = \sqrt{\log N_\delta(\frac{\delta^2}{4\sigma})}$ and $t = \delta$ represents the Gaussian maxima bound.

Combining Theorem 1 and Theorem 2, we obtain the following corollary:

Corollary 1. *Suppose that $\mathcal{F}^*$ is star-shaped, and $\delta \in (0, \sigma)$ satisfies*

$$\frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_\delta(t)}\, \mathrm{d}t \lesssim \frac{\delta^2}{\sigma}.$$

Then for each $t \geq \delta$, we have $\|\hat{f} - f^*\|_n^2 \leq t\delta$ with probability at least $1 - \exp\{-nt\delta/(2\sigma^2)\}$.

4 Applications

We look at several concrete applications of the above bounds.

4.1 Linear Regression ($n \geq d$)

As a warm-up, we start by considering the classic linear regression case, where the data points are generated from the ground-truth model

$$y_i = f^\star(x_i) + w_i = \langle \theta^*, x_i \rangle + w_i, \quad i = 1, \dots, n,$$

This model can be written compactly in vector form as

$$y = X\theta^* + w,$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$, $\theta^* \in \mathbb{R}^d$ and $w \in \mathbb{R}^n$. We assume that $n \geq d$. Consider the class of linear functions on $\mathbb{R}^d$:

$$\mathcal{F} = \{f_\theta(\cdot) = \langle \theta, \cdot \rangle : \theta \in \mathbb{R}^d\}.$$

Clearly $\mathcal{F} = \mathcal{F}^\star$ is convex and star-shaped. We also have that $B_n(\delta)$ is isomorphic to the ball

$$\left\{ X\theta : \frac{\|X\theta\|_2}{\sqrt{n}} \leq \delta, \theta \in \mathbb{R}^d \right\} \subset \operatorname{range}(X),$$

where range(X) has dimension at most d. So

$$\log N_\delta(s) \leq \log N(s, B_2^d(\delta), \|\cdot\|_2)) \leq d \log\left(1 + \frac{2\delta}{s}\right).$$

Hence

$$\begin{aligned} \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log N_\delta(s)}\, \mathrm{d}s &\leq \sqrt{\frac{d}{n}} \int_0^\delta \sqrt{\log\left(1 + \frac{2\delta}{s}\right)}\, \mathrm{d}s \\ &\lesssim \delta \sqrt{\frac{d}{n}} \\ &\leq \delta^2 \qquad \text{for } \delta = \sqrt{\frac{d}{n}}. \end{aligned}$$

By Corollary 1 we get

$$\left\| \hat{f} - f^\star \right\|_n^2 = \frac{1}{n} \left\| X(\hat{\theta} - \theta^\star) \right\|_n^2 \lesssim \delta^2 = \frac{d}{n}$$

with probability $\geq 1 - e^{-d/2}$. This bound is minimax optimal.

4.2 High-dimensional ℓ_q regression

We next consider an extension of the above setting, namely high-dimensional ℓ_q regression, where the function class is

$$\mathcal{F} = \left\{ f_\theta(\cdot) = \langle \theta, \cdot \rangle : \theta \in B_q^d(R) \right\} \quad \text{with}$$
$$B_q^d(R) := \left\{ \theta \in \mathbb{R}^d : \sum_{j=1}^d |\theta_j|^q \le R \right\}.$$

Here the dimension d is allowed to be larger than the sample size n.

First consider $q = 1$ (i.e., ℓ_1-constrained linear regression, or Lasso). We have that $\mathcal{F}^\star$ is convex and star-shaped. Using the same vector notations as in the previous subsection, we assume that the columns of X are normalized to have ℓ_2 norm bounded by $\sqrt{n}$. We can also show that

$$\log N_\delta(s) \lesssim \log N(s, B_1^d(R), \|\cdot\|_2) \lesssim R^2 \left(\frac{1}{s}\right)^2 \log d.$$

So

$$\begin{aligned}
\frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{4}}^{\delta} \sqrt{\log N_\delta(s)}\, \mathrm{d}s &\lesssim R\sqrt{\frac{\log d}{n}} \int_{\frac{\delta^2}{4}}^{\delta} \frac{1}{s}\, \mathrm{d}s \\
&= R\sqrt{\frac{\log d}{n}} \log \frac{4}{\delta} \\
&\lesssim \delta^2 \qquad \text{for } \delta^2 = R\sqrt{\frac{\log d}{n}}.
\end{aligned}$$

Hence by Corollary 1 we get $\left\|\hat{f} - f^\star\right\|_n^2 \lesssim R\left(\frac{\log d}{n}\right)^{1/2}$ with high probability. For general $q \in (0,1)$, we can prove that $\left\|\hat{f} - f^\star\right\|_n^2 \lesssim R\left(\frac{\log d}{n}\right)^{1-q/2}$, which is minimax optimal.

4.3 Lipschitz Regression

The next class of functions we consider is a subset of Lipschitz functions:

$$\mathcal{F} = \{ f : [0,1] \to \mathbb{R} : f(0) = 0, f \text{ is } L\text{-Lipschitz} \}.$$

We have that $\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \lesssim \frac{L}{\epsilon}$ as proved in Homework 1, and thus

$$\begin{aligned}
\frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log N_\delta(s)}\, \mathrm{d}s &\le \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log N(s, \mathcal{F}, \|\cdot\|_\infty)}\, \mathrm{d}s \\
&\lesssim \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\frac{L}{s}}\, \mathrm{d}s \\
&\lesssim \sqrt{\frac{L\delta}{n}} \\
&\lesssim \delta^2 \quad \text{for } \delta = \left(\frac{L}{n}\right)^{1/3}.
\end{aligned}$$

By Corollary 1 we get $\left\|\hat{f} - f^\star\right\|_n^2 \le \left(\frac{L}{n}\right)^{2/3}$ with high probability, which is minimax optimal.

The result above can be generalized to higher dimensions. Consider the following class of Lipschitz functions on the d-dimensional space:

$$\mathcal{F} = \left\{ f : [0,1]^d \to \mathbb{R} : f(0) = 0, f \text{ is } L\text{-Lipschitz} \right\}.$$

It can be shown that $\left\|\hat{f} - f^\star\right\|_n^2 \leq \left(\frac{L}{n}\right)^{2/(2+d)}$. Note the exponential dependence on the dimension. This is an example of the curse of dimensionality.

4.4 Convex Regression

Finally we look at the same set of one-dimensional Lipschitz functions as before but with the additional assumption of convexity:

$$\mathcal{F} = \{f : [0,1] \to \mathbb{R} : f(0) = 0, f \text{ is 1-Lipschitz and convex}\}$$

It can be shown that $\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \lesssim \sqrt{\frac{1}{\epsilon}}$. Then, by a similar argument as above we can take $\delta = \left(\frac{1}{n}\right)^{2/5}$. Corollary 1 we get $\left\|\hat{f} - f^\star\right\|_n^2 \lesssim \left(\frac{1}{n}\right)^{4/5}$, which is minimax optimal.

Note that this bound is better than the $\left(\frac{1}{n}\right)^{2/3}$ bound for Lipschitz functions. This makes sense because the additional convexity assumption puts a constraint on the second derivative, whereas Lipschitz-ness just bounds the first derivative.

References

[Wainwright, 2019] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.