Lectures 14–15: Nonparametric Regression

Lecturer: Yudong Chen  Scribe: Sean Sinclair, Connor Lawless

1 Brief Review

Last class we discussed the basics of statistical learning theory framework, using a symmetrization and con- traction technique in order to upper bound the population risk by the Randemacher complexity. This week we focus on specializing the setting to non-parametric regression with noisy observations.

Reading: Sections 13.1 and 13.2 in the Wainwright textbook.

2 Problem Setup

Consider the general statistical learning theory set-up, where we observe datapoints \((x_i, y_i)_{i=1}^n\) where

\[ y_i = f^*(x_i) + \sigma w_i \]

and \(w_i\) are i.i.d. \(\mathcal{N}(0,1)\) random variables. Here \(\sigma^2\) is the noise variance, \(y_i \in \mathcal{Y}\) is the response variable, and \(x_i \in \mathcal{X}\) are the covariates or features.

Remark Notice that \(f^*\) minimizes the population risk or mean-squared error discussed last week, i.e.

\[ f^*(\cdot) = \arg\min_f \mathbb{E} \left[ (Y - f(X))^2 \right] = \mathbb{E} [Y | X = \cdot], \]

which is the Bayes optimal solution to minimize the expected mean squared error. Unfortunately the conditional distribution of \(y\) given \(x\) is not known, and so we settle for an approximation using the observed data.

We consider the constrained empirical risk minimizer, where we take our estimate to be

\[ \hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2, \]

where \(\mathcal{F}\) is an user-specified function class.

3 Examples

The main difficulty in non-parametric regression is deciding on a function class \(\mathcal{F}\) to optimize over.

<table>
<thead>
<tr>
<th>nonparametric</th>
<th>parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approximate by locally smoothing</td>
<td>Kernel methods</td>
</tr>
<tr>
<td>Approximate with neural network</td>
<td>Approximate with linear function</td>
</tr>
</tbody>
</table>
In general there is a spectrum of function classes that can be considered; see the figure above for an illustration. One side of the spectrum constitutes parametric models, where \( F \) can be described by finitely many parameters. These are strong assumptions on the underlying function \( f^\star \), but often lead to tighter guarantees which avoid the curse of dimensionality. The other side of the spectrum are nonparametric models, where \( F \) is more complex, thus encompassing more models, but the bounds are sometimes worse. (Note: The picture above should be taken as just an crude illustration. A large neural network, for example, may correspond to a function class more complex than a simpler non-parametric model.)

We will focus on the non-parametric assumption, and give a guarantee that scales on a local complexity instead of a global complexity. We start with some parametric examples.

### 3.1 Linear Regression

Here we take the function class as
\[
F_C = \{ x \mapsto \langle \theta, x \rangle : \theta \in C \subseteq \mathbb{R}^d \}.
\]

Some examples of this include ridge regression, where \( C = \{ \theta \in \mathbb{R}^d : \|\theta\|_2 \leq R_2 \} \), and \( \ell_1 \) regression/LASSO, where \( C = \{ \theta \in \mathbb{R}^d : \|\theta\|_1 \leq R_1 \} \). In general we can have \( \ell_q \) regression, where the set \( C \) is the \( \ell_q \) “ball”:
\[
C = \left\{ \theta \in \mathbb{R}^d : \sum_{j=1}^d |\theta_j|^q \leq R_q \right\}
\]
for some given number \( q \in [0, 2] \).

Next we will be looking at some nonparametric function classes. Some examples include the following.

### 3.2 Lipschitz Regression

In this setting we take the function class as
\[
F_{\text{Lip}}(L) = \{ f : [0, 1] \to \mathbb{R} | f(0) = 0, f \text{ is } L\text{-Lipschitz}\}.
\]

The optimal solution \( \hat{f} \) in this function class will be a piecewise linear approximation of the datapoints \( (x_i, y_i)_{i=1}^n \).

### 3.3 Convex Regression

In this setting we take
\[
F_{\text{conv}} = \{ f : [0, 1] \to \mathbb{R} | f \text{ is convex}\}.
\]

In this case we need to solve the following (apparently infinite dimensional) optimization problem
\[
\hat{f} = \arg \min_{f \in F_{\text{conv}}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2.
\]

The solution to this optimization problem can be found numerically as follows.

**Step 1:** Solve the quadratic program
\[
\min_{(\hat{y}_i, \hat{g}_i)_{i=1}^n} \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2
\]
\[
s.t. \; \hat{y}_j \geq \hat{y}_i + \langle \hat{g}_i, x_j - x_i \rangle, \forall i, j = 1, \ldots, n.
\]
These constraints arise because a convex function \( f \) satisfies the subgradient condition that 
\[
 f(x_j) \geq f(x_i) + \langle \nabla f(x_i), x_j - x_i \rangle.
\]

**Step 2:** Set the estimate
\[
 \hat{f}(x) = \max_{i=1, \ldots, n} \{ \hat{y}_i + \langle \hat{g}_i, x - x_i \rangle \}.
\]

Note that with this estimator we have that \( \hat{f}(x_i) = \hat{y}_i \).

The two-step procedure above is equivalent to the original optimization problem, because the objective function of the latter only depends on the \( n \) data points \( x_1, \ldots, x_n \).

### 3.4 Cubic Smoothing Spline

Here we take the function class as
\[
 \mathcal{F}(R) = \left\{ f: [0, 1] \to \mathbb{R} \mid \int_0^1 (f''(x))^2 dx \leq R \right\}.
\]

The solution \( \hat{f} \) is a natural cubic spline with knots at \( x_1, \ldots, x_n \). This solution can be found by representing the function as a linear combination of certain basis functions
\[
 \hat{f}(x) = \beta_0 + \beta_0 x + \sum_{i=1}^{n-1} \beta_i (\phi_i(x) - \phi_{n-1}(x)), \quad \text{where}
\]
\[
 \phi_i(x) = \frac{(x - x_i)^3 - (x - x_n)^3}{x_n - x_i}, \quad i = 1, \ldots, n - 1,
\]
and then solving for the parameters \( \beta_0, \beta_0, \beta_1, \ldots, \beta_n \) using standard least squares.

### 3.5 Kernel Ridge Regression

Here we solve the regularized ERM problem
\[
 \hat{f} = \arg \min_{f \in H} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \| f \|_H^2,
\]
where \( H \) is a Reproducing Kernel Hilbert Space (RKHS) with kernel \( K: \mathcal{X} \times \mathcal{X} \to \mathbb{R} \). We let \( \langle \cdot, \cdot \rangle_H \) denote the inner product in \( H \), which induces the norm \( \| f \|^2_H = \langle f, f \rangle_H \). If we define the empirical kernel matrix \( \hat{K} \in \mathbb{R}^{n \times n} \) with entries \( \hat{K}_{i,j} = K(x_i, x_j)/n \), then the solution to the above problem is
\[
 \hat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\alpha}_i K(\cdot, x_i), \quad \text{where}
\]
\[
 \hat{\alpha} = (\hat{K} + \lambda_n I_n)^{-1} \frac{y}{\sqrt{n}}.
\]

### 4 Assumptions

We will focus on trying to bound the empirical error,
\[
 \|f - f^*\|^2_n := \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - f^*(x_i))^2.
\]

Using techniques from the previous lectures you can convert this into bounds on the population error
\[
 \|f - f^*\|_{\mathcal{L}_2(\mu)} := \mathbb{E}_{x \sim \mu} [(f(x) - f^*(x))^2].
\]

Before we start, we will need some definitions and assumptions on the function class \( \mathcal{F} \).
Definition 1. The **shifted function class** is defined as $F^* := \{ f - f^* : f \in F \}$.

Assumption 1. We assume that the shifted function class $F^*$ is **star-shaped**, i.e.,

$$\forall h \in F^* \text{ and } \alpha \in [0, 1] \text{ we have that } \alpha h \in F^*.$$  

Notice that under this assumption we have that $0 \in F^*$, which means that $f^* \in F$. Because we are considering non-parametric function classes this is a relatively mild assumption on the underlying data-generation process. Moreover, it is easy to see that if $F$ is convex then $F^*$ is star-shaped; the converse is not true in general.

Definition 2. The **localized Gaussian complexity** of $F^*$ is

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

where $w_i$ are i.i.d. $\mathcal{N}(0, 1)$. The number $\delta > 0$ is said to be the radius you are measuring the Gaussian complexity of. The **critical radius** $\delta^*$ is defined as

$$\delta^* := \min_{\delta > 0} \left\{ \delta \mid \frac{G_n(\delta, F^*)}{\delta} \leq \frac{\delta}{2\sigma} \right\}.$$

With these notations, we have the following simple lemma:

Lemma 1. If $F^*$ is star-shaped, then the function

$$\delta \mapsto \frac{G_n(\delta, F^*)}{\delta}$$

is non-increasing on $(0, \infty)$. Consequently, the critical radius $\delta^*$ exists and is finite.

Proof. Consider any $0 < \delta < t$. We show that $G_n(t, F^*)/t \leq G_n(\delta, F^*)/\delta$. This proof will crucially use the fact that $F^*$ is star-shaped.

Consider any $h \in F^*$ such that $\|h\|_n \leq t$. Define the new function $\hat{h} = \frac{\delta}{t} h$. Note that $\hat{h} \in F^*$ as $\frac{\delta}{t} \leq 1$. Moreover, we have that

$$\|\hat{h}\|_n = \frac{\delta}{t} \|h\|_n \leq \delta.$$

We also have that

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

Combining these two things together and taking the supremum over all $h \in F^*$ shows that

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

The left hand side is $(\delta/t)G_n(t, F^*)$ and the right hand side is $G_n(\delta, F^*)$ and hence

$$\frac{G_n(t, F^*)}{t} \leq \frac{G_n(\delta, F^*)}{\delta}.$$

The existence of a finite critical radius $\delta^*$ then follows immediately from the fact that the function is non-increasing and $\lim_{\delta \to 0} G_n(\delta, F^*)/\delta = \infty$. 

\[\square\]
5 Error Bound

We are now ready to prove an error bound on our ERM $\hat{f}$ versus the true Bayes optimal solution $f^*$.

**Theorem 1.** Suppose that $\mathcal{F}^*$ is star-shaped. Then for each number $t \geq \delta^*$, we have

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

with probability at least $1 - e^{nt\delta^*/2\sigma^2}$.

**Proof** We start by noting that since $\hat{f}$ is optimal to ERM, and $f^*$ is feasible we get that

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

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

(Rearranging and using $y_i = f^*(x_i) + \sigma w_i$)

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

$$\frac{1}{2} \left\| \Delta \right\|_n^2 \leq \frac{\sigma}{n} \sum_{i=1}^{n} w_i \Delta(x_i),$$

which is often referred to as the “Basic Inequality”.

Since the left hand side is what we want to bound, we need to work on bounding the right-hand side. We start by defining the event

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

for each number $u \geq \delta^*$. Note that the complement of the event $A(u)$ is:

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

To finish our proof we make use of the following lemma which we’ll prove later.

**Lemma 2.** For all $u \geq \delta^*$ we have

$$\Pr[A(u)] \leq e^{-\frac{n u^2}{2\sigma^2}}.$$

With this lemma, we can set $u = \sqrt{t\delta^*}$, where $t \geq \delta^*$, and note that $\Pr[A(\sqrt{t\delta^*})^c] \geq 1 - e^{-\frac{nt\delta^*}{2\sigma^2}}$.

To conclude our proof we simply look at the two cases for $\|\Delta\|_n$. If $\|\Delta\|_n \leq \sqrt{t\delta^*}$, then we’re done as $\|\Delta\|_n^2 \leq t\delta^* \leq 16t\delta^*$. If $\|\Delta\|_n > \sqrt{t\delta^*}$, then on the event $A(\sqrt{t\delta^*})^c$ we have that

$$\frac{1}{2} \|\Delta\|_n^2 \leq 2 \|\Delta\|_n \sqrt{t\delta^*} \Rightarrow \|\Delta\|_n^2 \leq 16t\delta^*$$

as claimed. \(\square\)

Now for the more involved part, proving Lemma 2.

5
Proof We start by rewriting $\Pr[\mathcal{A}(u)]$ as follows:

$$
\Pr[\mathcal{A}(u)] = \Pr\left[ \sup_{g \in \mathcal{F}^* : \|g\|_n \geq u} \frac{1}{n} |\sigma w, g(x^n_1)| \geq 2u \right]
$$

$$
\leq \Pr\left[ \sup_{g \in \mathcal{F}^* : \|g\|_n = u} \frac{1}{n} |\sigma w, g(x^n_1)| \geq 2u^2 \right] \quad \text{(rescale by $\frac{u}{\|g\|_n}$, $\mathcal{F}^*$ is star-shaped)}
$$

$$
= \Pr[\mathcal{Z}_n(u) \geq 2u^2],
$$

where we define the random variable $\mathcal{Z}_n(u) := |\sigma w, g(x^n_1)|$.

Concentration: Start by noting that $\mathcal{Z}_n(u)$ is a function of $w$ with Lipschitz constant:

$$
L \leq \sup_{\|g\|_n = u} \sigma \|g(x^n_1)\|_2 = \frac{\sigma}{\sqrt{n}} \|g\|_n = \frac{\sigma u}{\sqrt{n}}.
$$

Using the Gaussian Lipschitz concentration inequality we get:

$$
\Pr \left[ \mathcal{Z}_n(u) \geq \mathbb{E}[\mathcal{Z}_n(u)] + u^2 \right] \leq e^{-\frac{u^2}{2 \sqrt{n} \mathbb{E}[\mathcal{Z}_n(u)]}} = e^{-\frac{u^2}{2 \sqrt{n} \sigma^2}}.
$$

Expectation Bound: We can see that $\mathbb{E}[\mathcal{Z}_n(u)] \leq \sigma G_n(u, \mathcal{F}^*)$. By Lemma 1 we know that the function $v \mapsto G_n(v, \mathcal{F}^*)$ is non-increasing and by assumption we have $u \geq \delta^*$. It follows that

$$
\frac{\sigma G_n(u, \mathcal{F}^*)}{u} \leq \frac{\sigma G_n(\delta^*, \mathcal{F}^*)}{\delta^*} \leq \frac{\delta^*}{2} \leq \delta^*
$$

and thus $\mathbb{E}[\mathcal{Z}_n(u)] \leq \delta^* u$.

Combining: we get

$$
\Pr \left[ \mathcal{Z}_n(u) \geq 2u^2 \right] \leq \Pr \left[ \mathcal{Z}_n(u) \geq u \delta^* + u^2 \right] \leq e^{-\frac{u^2}{2 \sqrt{n} \sigma^2}}
$$

as claimed. \qed

Now we have a way of bounding $\|\hat{f} - f^*\|_n$ with $\delta^*$; the next step is to find a way to upper bound $\delta^*$. To start we introduce some notation.

Definition 3. We denote by $B_n(\delta)$ as the unit ball with respect to the $\|\cdot\|_n$ norm, i.e.,

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

Definition 4. We let $N_\delta(t)$ denote the covering number of $B_n(\delta)$, i.e.,

$$
N_\delta(t) = N(t, B_n(\delta), \|\cdot\|_n).
$$

Using the above definitions we can get the following theorem:

Theorem 2. If $\mathcal{F}^*$ is star-shaped, and a number $\delta \in [0, \sigma]$ satisfies

$$
\frac{16}{\sqrt{n}} \int_{\delta^2}^{\delta^*} \log N_\delta(t) \, dt \leq \frac{\delta^2}{4 \sigma},
$$

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


Proof Since \( \delta \in [0, \sigma] \), we get that \( \frac{\delta^2}{4\sigma} < \delta \), where the RHS is the radius of \( B_n(\delta) \). Let \( \{g^1, ..., g^M\} \) be a minimal \( \frac{\delta^2}{4\sigma} \)-covering of \( B_n(\delta) \). So \( \forall g \in B_n(\delta), \exists j \text{ s.t. } \|g^j - g\|_n \leq \frac{\delta^2}{4\sigma} \). Consequently, we have

\[
\left| \frac{1}{n} \sum_{i=1}^{n} w_i g(x_i) \right| = \left| \frac{1}{n} \langle w, g(x^n) \rangle \right|
\leq \left| \frac{1}{n} \langle w, g'(x^n) \rangle \right| + \left| \frac{1}{n} \langle w, g(x^n) - g'(x^n) \rangle \right|
\leq \max_{j=1,...,M} \left| \frac{1}{n} \langle w, g'(x^n) \rangle \right| + \sqrt{\frac{\|w\|_2^2}{n}} \sqrt{\frac{\|g(x^n) - g'(x^n)\|_2^2}{n}}
\leq \max_{j=1,...,M} \left| \frac{1}{n} \langle w, g'(x^n) \rangle \right| + \frac{\|w\|_2}{\sqrt{n}} \frac{\delta^2}{4\sigma}.
\]

Taking the supremum over \( g \in B_n(\delta) \) and the expectation with respect to \( w \) we have that:

\[
G_n(\delta, \mathcal{F}^*) \leq \mathbb{E}_w \left[ \max_{j=1,...,M} \left| \frac{1}{n} \langle w, g'(x^n) \rangle \right| \right] + \frac{\delta^2}{4\sigma}
\leq \frac{\delta}{\sqrt{n}} \sqrt{\log N_\delta \left( \frac{\delta^2}{4\sigma} \right)} + \frac{\delta^2}{4\sigma},
\]

where the last step follows from the known bound on Gaussian maxima.

Actually using the chaining argument we are able to give a better bound:

**Lemma 3.**

\[
\mathbb{E} \left[ \max_{j=1,...,M} \left| \frac{1}{n} \langle w, g'(x^n) \rangle \right| \right] \leq \frac{16}{\sqrt{n}} \int_{\frac{\delta}{4\sigma}}^{\delta} \sqrt{\log N_\delta(t)} \, dt.
\]

**Proof** We prove this by using Dudley’s integral bound with a slightly smarter look at the bounds we use. Start by defining \( Z(g^j) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i g^j(x_i) \) for \( j = 1, ..., M \). Note that \( Z(g^j) \) is zero-mean and sub-Gaussian with metric \( \rho(g^j, g^k) = \|g^j - g^k\|_n \). Note that since \( \{g^1, ..., g^M\} \) is a minimal \( \frac{\delta^2}{4\sigma} \)-covering of \( B_n(\delta) \), we don’t need to extend the chaining smaller than a resolution of \( \frac{\delta^2}{4\sigma} \) since at that resolution we can uniquely identify each point. We also only need to start the chaining at a resolution of \( \delta \), as the set \( B_n(\delta) \) has a diameter of \( 2\delta \). Putting this together and working through the arithmetic of the chaining argument we get:

\[
\mathbb{E} \left[ \max_{j=1,...,M} \left| \frac{1}{n} \langle w, g'(x^n) \rangle \right| \right] = \mathbb{E} \left[ \max_{j=1,...,M} \frac{|Z(g^j)|}{\sqrt{n}} \right]
\leq \frac{1}{\sqrt{n}} \mathbb{E} \left[ \max_{j=1,...,M} |Z(g^j)| \right]
\leq \frac{16}{\sqrt{n}} \int_{\frac{\delta}{4\sigma}}^{\delta} \sqrt{\log N_\delta(t)} \, dt,
\]

where for the last inequality we used a version of Dudley’s integral bound that includes explicit constants.\(^1\)

Using Lemma 3 we have that:

\[
G_n(\delta, \mathcal{F}^*) \leq \frac{16}{\sqrt{n}} \int_{\frac{\delta}{4\sigma}}^{\delta} \sqrt{\log N_\delta(t)} \, dt + \frac{\delta^2}{4\sigma}
\]

\(^1http://www.stat.cmu.edu/~arinaldo/Teaching/36755/F16/Scribed_Lectures/36755_F16_Nov02.pdf\)
\[ \leq \frac{\delta^2}{2\sigma}, \]

where the last step follows from the assumption of Theorem 2. It follows that

\[ \frac{G_n(\delta, \mathcal{F}^*)}{\delta} \leq \frac{\delta}{2\sigma}, \]

\[ \Rightarrow \delta \geq \delta^* := \min \left\{ \delta' > 0 : \frac{G_n(\delta', \mathcal{F}^*)}{\delta'} \leq \frac{\delta'}{2\sigma} \right\} \]

as claimed.

Combining Theorems 1 and 2, we have the following convenient corollary.

**Corollary 1.** If

\[ \int_{\delta^*/4\sigma}^{\delta} \sqrt{\log N_\delta(s)} \, ds \lesssim \frac{\delta^2}{\sigma} \] and \( t \geq \delta \)

then \( \| \hat{f} - f^* \|_n^2 \lesssim t \delta \) with probability at least \( 1 - e^{-\frac{nt\delta^2}{2}} \).

### 6 Applications

We look at several concrete applications of the above bounds.

#### 6.1 Linear Regression (\( n \geq d \))

As a warm-up, we start by considering the classic linear regression case, where

\[ y_i = f^*(x_i) + w_i = \langle \theta, x_i \rangle + w_i, \]

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

Clearly \( \mathcal{F} = \mathcal{F}^* \) is convex and star-shaped. We also have that \( B_n(\delta) \) is isomorphic to the ball \( \left\{ X\theta : \frac{\|X\theta\|_2}{\sqrt{n}} \leq \delta, \theta \in \mathbb{R}^d \right\} \subset \text{range}(X) \), where \( \text{range}(X) \) has dimension at most \( d \). So

\[ \log N_\delta(s) \leq \log N(s, B^d_2(\delta, \|\cdot\|_2)) \leq d \log (1 + \frac{2\delta}{s}). \]

Hence

\[ \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log N_\delta(s)} \, ds \leq \frac{\sqrt{d}}{n} \int_0^\delta \sqrt{\log (1 + \frac{2\delta}{s})} \, ds \]

\[ \lesssim \delta \sqrt{\frac{d}{n}} \]

\[ \leq \delta^2 \text{ for } \delta = \sqrt{\frac{d}{n}}. \]

And by Corollary 1 we get

\[ \| \hat{f} - f^* \|_n^2 = \frac{1}{n} \| X(\hat{\theta} - \theta^*) \|_n^2 \lesssim \delta^2 = \frac{d}{n} \]

with probability \( \geq 1 - e^{-d/2} \). This bound is minimax optimal.
6.2 High-dimensional $\ell_q$ regression

We next consider a more complicated application, namely high-dimensions $\ell_q$ regression, where the function class is

$$\mathcal{F} = \{ f_\theta(\cdot) = \langle \theta, \cdot \rangle : \theta \in B^d_q(R) \}$$

with

$$B^d_q(R) := \{ \theta \in \mathbb{R}^d : \sum_{j=1}^d |\theta_j|^q \leq R \}.$$

First consider $q = 1$ (i.e., Lasso). We have that $\mathcal{F}^*$ is convex and star-shaped. When the columns of $X$ have a norm bounded by $\sqrt{n}$, we can also show that

$$\log N(s, \mathcal{F}^*, \| \cdot \|_2) \lesssim R^2 (\frac{1}{s})^2 \log d.$$

So

$$\frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log N(s)} \, ds \lesssim R \sqrt{\frac{\log d}{n}} \int_0^\delta \frac{1}{s} \, ds = R \sqrt{\frac{\log d}{n}} \log \frac{4}{\delta} \lesssim \delta^2 \quad \text{for } \delta^2 = R \sqrt{\frac{\log d}{n}}.$$

Hence by Corollary 1 we get $\| \hat{f} - f^* \|_2^2 \lesssim R (\frac{\log d}{n})$ with high probability. For general $q \in (0, 1)$, we can prove that $\| \hat{f} - f^* \|_2^2 \lesssim R (\frac{\log d}{n})^{1-q/2}$, which is minimax optimal.

6.3 Lipschitz Regression

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

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

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

$$\frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log N(s)} \, ds \leq \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log N(s, \mathcal{F}, \| \cdot \|_\infty)} \, ds \leq \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\frac{L}{s}} \, ds \lesssim \sqrt{\frac{L\delta}{n}} \lesssim \delta^2 \quad \text{for } \delta = \left( \frac{L}{n} \right)^{1/3}.$$

By Corollary 1 we get $\| \hat{f} - f^* \|_n^2 \leq (\frac{L}{n})^{2/3}$ with high probability, which is minimax optimal.
6.4 Convex Regression

Finally we look at the same set of functions as before with the added assumption of convexity:

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

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

Note that this bound is better than the \( (\frac{1}{n})^{2/3} \) 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.