Statistics 210B Lecture 25 Notes

Daniel Raban

April 21, 2022

1 L^2 Prediction Error Bounds for Nonparametric Function Regression

1.1 Recap: prediction error bounds for $\|\cdot\|_n$ compared to $\|\cdot\|_{L^2}$.

We have been studying non-parametric function regression, where we observe $x_i, y_i \in \mathbb{R}$ with $y_i = f^*(x_i) + w_i$ for $i \in [n]$. We assume $f^* \in \mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$ for some specific function class $\mathcal{F}$ and take the noise to be $w_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

For the non-parametric least squares problem, we have the constrained form

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

Our goal is to bound the prediction error,

$$\|\widehat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 = \frac{1}{n} \sum_{i=1}^n (\widehat{f}(x_i) - f^*(x_i))^2.$$

We proved a prediction error bound that relies on a critical equation.

Theorem 1.1. *Let $\mathcal{F}^* = \mathcal{F} - \{f^*\}$ be star-shaped. Then $\mathbb{E}_w[\|\widehat{f}_n - f^*\|_n^*] \lesssim \delta_n^2$, where δ_n solves the critical equation $\mathcal{G}_n(\delta; \mathcal{F}^*) = \delta^2 .(2\sigma)$.*

What if we want to look at the behavior of $\widehat{f}$ on a new dataset $\widetilde{x} \sim \mathbb{P}$ instead of x_i in the original dataset? If we have $y_i = f^*(x_i) + \widetilde{w}_i$, where $\widetilde{w}_i \sim N(0, \sigma^2)$ and $(\widetilde{x}_i, \widetilde{y}_i) \overset{\text{iid}}{\sim} (x_i, y_i)$, we can see that

$$\mathbb{E}_{\widetilde{x}_i, \widetilde{y}_i}[(\widehat{f}(\widetilde{x}_i) - \widetilde{y}_i)]^2 = \sigma^2 + \|\widehat{f} - f^*\|_{L^2}^2.$$

So in many cases, we want to control the L^2 distance between $\widehat{f}$ and f^*.

1.2 Relation between $\|\cdot\|_n^2$ and $\|\cdot\|_{L^2}^2$

Let $f \in \mathcal{F}$. Then if the function f does not depend on our training data set,

$$\begin{aligned}\mathbb{E}_X[\|f\|_n^2] &= \mathbb{E}_x\left[\frac{1}{n}\sum_{i=1}^n f(x_i)^2\right] \\ &= \mathbb{E}[f(x)^2] \\ &= \|f\|_{L^2}.\end{aligned}$$

Now suppose that $\widehat{f}(x) = h(x; \{x_i, y_i\}_{i\in[n]})$ depends on our training data set. Then

$$\mathbb{E}_{x_i}[\|\widehat{f} - f^*\|_n^2] = \mathbb{E}_x\left[\frac{1}{n}\sum_{i=1}^n (\widehat{f}(x_i; \{x_i, y_i\}_{i\in[n]}) - f^*(\widetilde{x}))^2\right] \quad \neq \mathbb{E}_x[(\widehat{f}(\widetilde{x}; \{x_i, y_i\}_{i\in[n]}) - f^*(\widetilde{x}))^2]$$

We hope to show a result like

$$\|\widehat{f} - f^*\|_{L^2}^2 \lesssim \underbrace{\|\widehat{f} - f^*\|_n^2}_{\delta_n^2} + \varepsilon_n^2,$$

where $\varepsilon_n^2 \to 0$ as $n \to \infty$.

Today, we will show two bounds:

1. Naive bound: If we do not care about how fast $\varepsilon_n \to 0$, we can get a bound by using a global uniform bound, a global Rademacher complexity bound, and using bounded difference concentration.
2. Tighter bound: We will use
   (a) the local uniform bound
   (b) local Rademacher complexity
   (c) a tighter concentration inequality, known as the Talagrand concentration inequality.

1.3 Naive bound

Let $f = \widehat{f} - f^* \in \mathcal{F}^*$. Then

$$\begin{aligned}\left|\|\widehat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 - \|\widehat{f} - f^*\|_{L^2(\mathbb{P})}^2\|\right| &\le \sup_{g\in\mathcal{F}^*} |\|g\|_n^2 - \|g\|_{L^2}^2| \\ &= \sup_{g\in\mathcal{F}^*}\left|\frac{1}{n}\sum_{i=1}^n g(x_i)^2 - \mathbb{E}[g(x)^2]\right|\end{aligned}$$

$$=: Z$$

We first try to find a bound on the expectation of Z:

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}\left[\sup_{g\in\mathcal{F}^*}\left|\frac{1}{n}\sum_{i=1}^n g(x_i)^2 - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n g(\widetilde{x}_i)^2\right]\right|\right] \\ &\le \mathbb{E}\left[\sup_{g\in\mathcal{F}^*}\left|\frac{1}{n}\sum_{i=1}^n (g(x_i)^2 - g(\widetilde{x}_i)^2)\right|\right]\end{aligned}$$

Since the distribution of this is symmetric about 0,

$$\begin{aligned}&= \mathbb{E}\left[\sup_{g\in\mathcal{F}^*}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon^2(g(x_i)^2 - g(\widetilde{x}_i)^2)\right|\right] \\ &\le 2\,\mathbb{E}\left[\sup_{g\in\mathcal{F}^*}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon^2 g(x_i)^2\right|\right]\end{aligned}$$

If this just had g instead of g^2, this quantity would be the Rademacher complexity. So we want to bound this by the Rademacher complexity. Write $\phi(t) = t^2$, so

$$\le 2\,\mathbb{E}\left[\sup_{g\in\mathcal{F}^*}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon^2\phi(g(x_i))\right|\right]$$

The function ϕ is $2\|\mathcal{F}^*\|_\infty$-Lipschitz, where we can assume that $\|\mathcal{F}^*\|_\infty = 1$.

$$\le 4\ \underbrace{\mathbb{E}\left[\sup_{g\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)\right|\right]}_{\overline{\mathcal{R}_n}(\mathcal{F}^*)\text{ Rademacher complexity}}.$$

We can use chaining to bound this.

Now let's bound the distance from the mean. Using the bounded difference inequality, $|Z - \mathbb{E}[Z]| \sim \mathrm{sG}(\|g\|_\infty^2/n)$, so

$$\|\widehat{f} - f^*\|_{L^2}^2 \lesssim \underbrace{\|\widehat{f} - f^*\|_n^2}_{\delta_n^2} + \overline{\mathcal{R}_n}(\mathcal{F}^*) + O(1/\sqrt{n}).$$

If $\mathcal{F}^*$ is parametric with d parameters, then $\delta_n^2 \asymp \frac{d}{n}$ and $\overline{\mathcal{R}_n}(\mathcal{F}^*) \asymp \sqrt{\frac{d}{n}}$.

1.4 Using localization to get a faster rate

We will present some heuristics, rather than something completely rigorous. The rigoorous treatment is in Chapter 14 of Wainwright's textbook. Suppose we already know that $\|\widehat{f} - f^*\|_{L^2(\mathbb{P})} \le r$. We can think about r decaying to 0 as $n \to \infty$. It may seem strange to assume that the L^2 norm is bounded when this is what we want to prove, but the idea

is that we will get a more refined bound. So we can iterate this bound to get a nice final result

Letting $g = \widehat{f} - f^* \in \mathcal{F}^*$,

$$\left| \|\widehat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 - \|\widehat{f} - f^*\|_{L^2(\mathbb{P})}^2 \right| \leq \sup_{\substack{g \in \mathcal{F}^* \\ \|g\|_{L^2} \leq r}} \left| \frac{1}{n} \sum_{i=1}^n g(x_i)^2 - \mathbb{E}[g(x_i)^2] \right| =: Z(r).$$

Now we bound the expectation using the same line of argument as before:

$$\mathbb{E}[Z(r)] \leq 4 \underbrace{\mathbb{E}\left[\sup_{\substack{g \in \mathcal{F} \\ \|g\|_{L^2} \leq r}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(x_i) \right| \right]}_{\overline{\mathcal{R}_n}(r; \mathcal{F}^*) \text{ localized Rademacher complexity}}.$$

We now show that $\overline{\mathcal{R}_n}(r; \mathcal{F}^*) \lesssim \varepsilon_n \cdot r$, where $\varepsilon_n = \inf\{\varepsilon : \overline{\mathcal{R}_n}(\varepsilon; \mathcal{F}^*) \leq \frac{\varepsilon^2}{16b}\}$ and $b = \sup_{g \in \mathcal{F}^*} \|g\|_\infty = 1$. This is because for any $r \geq \varepsilon_n$, $\frac{\overline{\mathcal{R}_n}(r; \mathcal{F}^*)}{r}$ is non-increasing (as long as $\mathcal{F}^*$ is star-shaped). This tells us that

$$\frac{\overline{\mathcal{R}_n}(r; \mathcal{F}^*)}{r} \leq \frac{\overline{\mathcal{R}_n}(\varepsilon_n; \mathcal{F}^*)}{\varepsilon_n} = \frac{\varepsilon_n}{16}.$$

This means that $\overline{\mathcal{R}_n}(r; \mathcal{F}^*) \lesssim \varepsilon_n r$.

Now let's see how this implies an upper bound for the prediction error of the L^2 norm. Suppose that $Z(r) \approx \mathbb{E}[Z(r)]$ for any $r \in \mathbb{R}$ (this should be made quantitative with the Talagrand concentration inequality or a tighter concentration inequality). Then

$$\begin{aligned} \left| \underbrace{\|\widehat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2}_{a^2} - \underbrace{\|\widehat{f} - f^*\|_{L^2(\mathbb{P})}^2}_{b^2} \right| &\lesssim \overline{\mathcal{R}_n}(r; \mathcal{F}^*) \\ &\lesssim \varepsilon_n r \\ &= \varepsilon_n \underbrace{\|\widehat{f} - f^*\|_{L^2}}_{b} \end{aligned}$$

This is heuristic because the quantity $\|\widehat{f} - f^*\|_{L^2}$ is random and depends on the training data set. However, we can use the iterative argument to make sense of this argument. We now have

$$|a^2 - b^2| \leq \varepsilon_n b \leq \frac{b^2}{4} + 4\varepsilon_n^2,$$

which gives $b^2 \lesssim a^2 + \varepsilon_n^2$. So we get that

$$\|\widehat{f} - f^*\|_{L^2}^2 \lesssim \underbrace{\|\widehat{f} - f^*\|_n^2}_{\delta_n^2} + \varepsilon_n^2.$$

This tells us that the upper bound of the prediction error in terms of the L^2 norm is of the same order as the upper bound of the prediction error in terms of the $L^2(\mathbb{P}_n)$ norm. If T is parametric with d parameters, then $\delta_n^2 \asymp \varepsilon_n^2 \asymp \frac{d}{n}$.

Here, our proof is different in two ways from the treatment in the textbook.

1. The first way is that we have assumed that our concentration inequality does not destroy our bound. If we just use the bounded differences inequality, we get the naive bound
$$|Z(r) - \mathbb{E}[Z(r)]| \lesssim \sqrt{\frac{1}{n}} = \eta_n.$$
The issue with this is that $Z(r)$ is $O(1/n)$ and $\mathbb{E}[Z(r)]$ is $O(1/n)$. Instead, we need to use the Talagrand inequality.

2. The second difference is that we have assumed beforehand that $\|\widehat{f} - f^*\|_{L^2(\mathbb{P})} \le r$. The textbook instead uses a peeling argument. We actually want to find a bound on $\sup_r |Z(r) - \mathbb{E}[Z(r)]|$. To use a union bound, we need to discretize r, and a clever way to do so is to use a log scale, rather than a uniform grid.

In the end, we get the following theorem, which we state informally. This is Corollary 14.15 in the textbook.

Theorem 1.2. *Let*

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

Then

$$\|\widehat{f} - f^*\|_{L^2} \lesssim \varepsilon_n^2 + \delta_n^2,$$

where

$$\varepsilon_n = \inf\left\{ r : \overline{\mathcal{R}_n}(r; \mathcal{F}) \lesssim \frac{r^2}{b} \right\}, \qquad \delta_n = \inf\left\{ \delta : \mathcal{G}_n(\delta; \mathcal{F}^*) \lesssim \frac{\delta}{b} \right\}.$$

Here, ε_n is deterministic, as $\overline{\mathcal{R}_n}$ is averaged over $(x_i)_{i \in [n]}$. On the other hand, δ_n is random, as $\mathcal{G}_n$ depends on $(x_i)_{i \in [n]}$.

1.5 Uniform law for Lipschitz cost function

More generally, we may want to consider cost functions which are not the squared error. Suppose we have $(x_i, y_i)_{i\in[n]} \overset{\text{iid}}{\sim} \mathbb{P} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ with a function class $\mathcal{F} \subseteq \{f : \mathcal{X} \to \widehat{\mathcal{Y}}\}$. Let the loss be $\mathcal{L} : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$. Then we have the empirical risk

$$\mathbb{P}_n\mathcal{L}(f(x), y) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f(x_i), y_i),$$

with empirical risk minimizer

$$\widehat{f} = \arg\min_{f\in\mathcal{F}} \mathbb{P}_n\mathcal{L}(f(x), y)$$

and population risk minimizer

$$f^* = \arg\min_{f} \underbrace{\mathbb{P}\mathcal{L}(f(x), y)}_{\mathbb{E}_{x,y}[\mathcal{L}(f(x),y)]}.$$

Our goal is to bound $\|\widehat{f} - f^*\|_{L^2}^2$.

We assume the loss is L-Lipschitz:

$$|\mathcal{L}(z, y) - \mathcal{L}(z', y)| \le L|z - z'|.$$

Another assumption, which is harder to check, is that L is r-**strongly convex**: If we let $L_f(x, y) := L(f(x), y)$, then we require

$$\mathbb{P}\left(\mathcal{L}_f - L_{f^*} - \left.\frac{\partial\mathcal{L}}{\partial z}\right|_{f^*}(f - f^*)\right) \ge \frac{r}{2}\|f - f^*\|_{L^2}^2.$$

Example 1.1 (Logistic regression)**.** Let $\mathcal{Y} = \{\pm 1\}$, $\mathcal{L}(\widehat{y}, y) = \log(1 + e^{-2y\widehat{y}})$, and

$$\mathbb{P}(y \mid x) = \frac{1}{1 + e^{-2yf^*(x)}}.$$

Then $\mathcal{L}(\widehat{y}, y)$ is 1-Lipschitz in $\widehat{y}$. Under mild conditions, $\mathbb{P}\mathcal{L}_f$ is r-strongly convex.

Here is Theorem 14.20 in the textbook.

Theorem 1.3. *Assume that $\mathcal{F}$ is 1-uniformly bounded and star-shaped, with population minimizer f^*. Let $\delta_n = \inf\{\delta > \frac{c}{\sqrt{n}} : \overline{\mathcal{R}_n}(\delta; \mathcal{F}^*) \le \delta^2\}$.*

(a) If $\mathcal{L}$ is L-Lipschitz in $\widehat{\mathcal{Y}}$, then with high probability,

$$\sup_{f\in\mathcal{F}} \frac{|\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})|}{\|f - f^*\|_{L^2} + \delta_n} \le 10L \cdot \delta_n.$$

(b) If $\mathbb{P}\mathcal{L}_f$ is also r-strongly convex, then with high probability, for all $\widehat{f}$ such that $\mathbb{P}_n(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*}) \leq 0$, we have

$$\|\widehat{f} - f^*\|_2^2 \leq \left(\frac{20L}{r} + 1\right)^2 \delta_n^2$$

and

$$\mathbb{P}(\mathcal{L}_{\widehat{f}} - \mathcal{L}_{f^*}) \leq 10L \left(\frac{20L}{r} + 1\right)^2 \delta_n^2.$$

Remark 1.1. Statement (b) is a direct consequence of statement (a), using the r-strong convexity condition. The proof of (a) also relies on a local Rademacher complexity bound. We can bound $\sup_{f \in \mathcal{F}} |\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})|$ using the Rademacher complexity, and we can get a faster rate using local Rademacher complexity.

This concludes our discussion of nonparametric function estimation. Next time, we will move on to minimax lower bounds.