Statistics 210B Lecture 24 Notes

Daniel Raban

April 19, 2022

1 Examples of and Oracle Inequality for Non-Parametric Least Squares Regression

1.1 Recap: localized Gaussian complexity bound for non-parametric least squares

We are studying non-parametric regression. Our model is that we observe $x_i \in \mathscr{X}$ and $y_i \in \mathbb{R}$, where

$$y_i = f^*(z_i) + \sigma \cdot w_i, \qquad i \in [n]$$

and $f^* \in \mathcal{F} \subseteq \{f : \mathscr{X} \to \mathbb{R}\}$ is in a designated function class. The noise is $w_i \overset{\text{iid}}{\sim} N(0, 1)$.

We consider the non-parametric least squares problem, which has the constrained form

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

Our goal is to bound the prediction error

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

Last time, we proved the following localized Gaussian complexity bound.

Theorem 1.1. *Suppose that $\mathcal{F}* = \mathcal{F} - \{f^*\}$ is star shaped. Then*

$$\mathbb{E}_{w_i}[\|\widehat{f}_n - f^*\|_n^2] \lesssim \delta_n^2,$$

where δ_n^2 solves $\mathcal{G}_n(\delta; \mathcal{F}^*) = \delta^2/(2\sigma)$, which is

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

The chaining method gives us a bound

$$\mathcal{G}_n(\delta; \mathcal{F}^*) \lesssim \frac{\delta^2}{4\sigma} + \frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\delta}}^{\delta} \sqrt{\log N_n(t; B_n(\delta; \mathcal{F}^*))}\, dt.$$

Let's look at some concrete examples for this localized Gaussian complexity bound.

1.2 Applications of the localized Gaussian complexity bound

Example 1.1. Let $\mathcal{F}_{1:n} = \{f_\theta(\cdot) = \langle \cdot, \theta \rangle : \theta \in \mathbb{R}^d\}$, and let

$$y_i = \langle x_i, \theta^* \rangle + \sigma \cdot w_i, \qquad i \in [n],$$

where $\theta^* \in \mathbb{R}^d$. Our estimator is

$$\widehat{\theta} = \min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle x_i, \theta \rangle)^2,$$

so

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

We will show that

$$\begin{aligned} \|f_{\widehat{\theta}} - f_{\theta^*}\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \langle x_i, \widehat{\theta} - \theta^* \rangle^2 \\ &= \frac{\|X(\theta^* - \widehat{\theta})\|_2^2}{n} \\ &\lesssim \sigma^2 \cdot \frac{\operatorname{rank}(X)}{n} \\ &\lesssim \sigma^2 \frac{d}{n}. \end{aligned}$$

We have the upper bound proportional to $\frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{4\delta}}^{\delta} \sqrt{\log N_n(t; B_n(\delta; \mathcal{F}^*))}\, dt$, so we just need to calculate this covering number. This ball is

$$B_n(\delta; \mathcal{F}_{1:n}) = \left\| f_\theta(x) = \langle x, \theta \rangle : \sqrt{\frac{1}{n} \sum_{i=1}^{n} \langle x_i, \theta \rangle^2} \leq \delta \right\|,$$

which is isomorphic to the δ-ball in the range of X (where $\dim \operatorname{range}(X) = \operatorname{rank}(X)$. Using a volume argument, the covering number is

$$N_n(t; B_n(\delta; \mathcal{F}_{1:n})) \leq r \cdot \log\left(1 + \frac{2\delta}{t}\right), \qquad r = \operatorname{rank}(X).$$

So the metric entropy integral is upper bounded by

$$\frac{\sqrt{r}}{\sqrt{n}} \int_{\frac{\delta^2}{4\delta}}^{\delta} \sqrt{\log\left(1+\frac{2\delta}{t}\right)}\, dt \leq c \cdot \delta\sqrt{r} n.$$

We have $c\delta\sqrt{\frac{r}{n}} = \frac{\delta^2}{4\sigma}$, so solving gives $\delta_n = c\sigma\sqrt{\frac{r}{n}}$. So $\delta_n^2 = c\sigma\sqrt{\frac{r}{n}}$, and we get

$$\mathbb{E}_w[\|f_{\widehat{\theta}} - f_{\theta^*}\|_n^2] \lesssim \sigma\sqrt{\frac{r}{n}}.$$

Example 1.2 (Lipschitz function class)**.** Let $\mathcal{F}_{\mathrm{Lip}}(L) = \{f : [0,1] \to \mathbb{R} : f(0) = 0, f \text{ is } L-$ Lipschitz$\}$. Then

$$\mathcal{F}^* \subseteq \mathcal{F}_{\mathrm{Lip}}(L) - \mathcal{F}_{\mathrm{Lip}}(L) = \mathcal{F}_{\mathrm{Lip}}(2L).$$

We have upper bounded the metric entropy of this function class as

$$\log N(\varepsilon; \mathcal{F}(2L), \|\cdot\|_\infty) \lesssim \frac{L}{\varepsilon},$$

where $\|f\|_\infty = \sup_{x\in\mathscr{X}}$, so $\|f\|_n = (\frac{1}{n}\sum_{i=1}^n f(x_i)^2)^{1/2} \leq \|f\|_\infty$. This tells us that

$$\log N(\varepsilon; \mathcal{F}(2L), \|\cdot\|_n) \leq \log N(\varepsilon; \mathcal{F}(2L), \|\cdot\|_\infty) \lesssim \frac{L}{\varepsilon}.$$

So the metric entropy integral is

$$\begin{aligned}
\frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{4\delta}}^{\delta} \sqrt{\log N_n(t; \mathcal{F}(2L), \|\cdot\|_\infty)}\, dt &\leq \frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\frac{L}{t}}\, dt \\
&= \sqrt{\frac{L}{n}} \left(2\sqrt{t}\Big|_{\frac{\delta^2}{4\sigma}}^{\delta} \right) \\
&= c\sqrt{\frac{L}{n}}(\sqrt{\delta} - \sqrt{\delta^2/(4\sigma)}) \\
&\leq c\sqrt{\frac{L}{n}}\sqrt{\delta}.
\end{aligned}$$

Solving $\sqrt{\frac{L\delta}{n}} = \delta^2$ gives $\delta^2 \lesssim (\frac{L\sigma^2}{n})^{2/3}$.

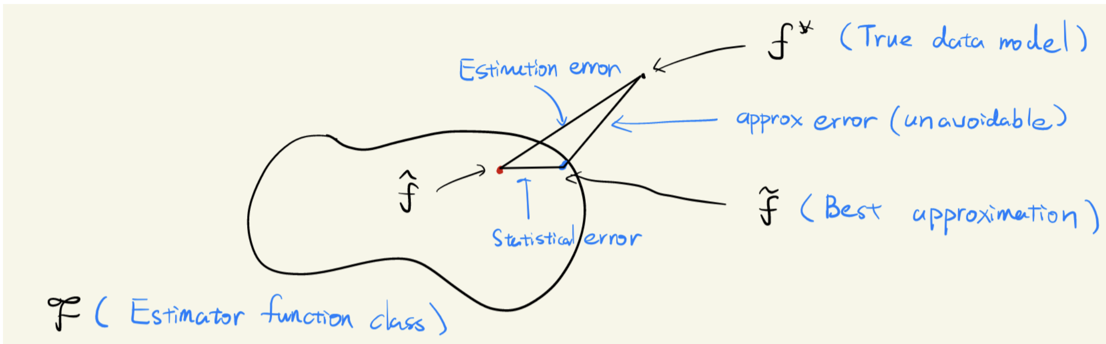
Example 1.3. What if $\log N \asymp \frac{1}{\varepsilon^d}$ for $d \geq 3$ (Lipschitz in d dimensions)? Then

$$\begin{aligned}
\frac{1}{\sqrt{n}} \int_{\varepsilon}^{\delta} \frac{1}{t^d/2}\, dt &= \frac{1}{\sqrt{n}} \frac{2}{d-2} \frac{-1}{t^{d/2-1}}\Big|_{\varepsilon}^{\delta} \\
&\leq \frac{1}{\sqrt{n}} \frac{2}{d-2} \frac{1}{\varepsilon^{d/2-1}}.
\end{aligned}$$

Take $\varepsilon = \frac{\delta^2}{4\sigma}$ and compare $\frac{1}{\sqrt{n}} \frac{2}{d-2} \frac{1}{\varepsilon^{d/2-1}} = \varepsilon$ to get $\varepsilon \lesssim \frac{1}{n^{4/d}}$. This gives $\delta^2 \lesssim \frac{1}{n^{4/d}}$.

1.3 Oracle inequalities

In practice, we may encounter the situation $f^* \notin \mathcal{F}$, like if we fit a linear model to something which is not exactly linear.



Suppose $\widetilde{f} \in \mathcal{F}$ is closest to f^*. We hope that $\widehat{f}$ is close to $\widetilde{f}$ when we have a lot of samples. That is, we hope that

$$\|\widehat{f} - f^*\| \lesssim \inf_{f \in \mathcal{F}} \|f - f^*\| + \varepsilon_n,$$

where $\varepsilon_n \to 0$ as $n \to \infty$. We would also like ε_n to decay as fast as possible. This kind of bound gives us a justification that our nonparametric regression gives us a best approximation to the function f^*.

Define $\partial \mathcal{F} = \mathcal{F} - \mathcal{F} = \{f - g : f, g \in \mathcal{F}\}$. Assume that $\partial \mathcal{F}$ is star-shaped; we can always take the star hull to make this true, so this is not a stringent assumption.

Theorem 1.2. *Let $\delta_n = \inf\{\delta > 0 : \mathcal{G}_n(\delta; \partial\mathcal{F}) \leq \frac{\delta^2}{2\sigma}\}$. Then there exist constants c_0, c_1, c_2 such that the event*

$$\{\widehat{f} - f^*\|_n^2 \leq \inf_{\gamma \in (0,1)} \left[\frac{1 + gamma}{1 - \gamma} \|f - f^*\|_n^2 + \frac{c_0}{\gamma(1-\gamma)} \delta_n t \right] \qquad \forall f \in \mathcal{F}$$

occurs with probability at least $1 - c_1 e^{-c_2 \frac{nt\delta_n}{\sigma^2}}$.

This says that

$$\|\widehat{f} - f^*\|_n^2 \lesssim \inf_{f \in \mathcal{F}} \|f - f^*\|_n^2 + \delta_n^2,$$

so we can integrate this probability bound to get an expectation bound:

$$\mathbb{E}[\|\widehat{f} - f^*\|_n^2] \lesssim \inf_{f \in \mathcal{F}} \|f - f^*\|_n^2 + \delta_n^2 + \frac{\sigma^2}{n}.$$

Note that if $f^* \in \mathcal{F}$, then the first term is 0, so this recovers the prediction error bound in the previous theorem.

Proof. We start from a basic inequality:

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2 \le \frac{1}{2n}\sum_{i=1}^{n}(y - i - f^*(x_i))^2.$$

This tells us that

$$\frac{1}{2}\|\widehat{f} - f^*\|_n^2 \le \frac{1}{2}\|\widetilde{f} - f^*\|_n^2 + \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n} w_i(\widehat{f}(x_i) - \widetilde{f}(x_i))\right|}_{(*)}.$$

We want to upper bound the right term; this is basically the same thing we did for the previous prediction error bound, but with $\widetilde{f}$ instead of f^*. Recall that by definition, $\mathcal{F}_n(\delta; \partial\mathcal{F}) = \mathbb{E}[\sup_{\substack{g\in\partial\mathcal{F}\\ \|g\|_n\|\le\delta}} |\frac{1}{n}\sum_{i=1}^n w_i g(x_i)|]$ and $\mathcal{G}_n(\delta.\partial\mathcal{F}) \asymp \delta_n^2$.

The simple case is when $\|\widehat{f} - \widetilde{f}\|_n \le \delta$. In this case,

$$(*) \lesssim \mathcal{G}_n(\delta_n; \partial\mathcal{F}) \asymp \delta_n^2.$$

The harder case is when $\|\widehat{f} - \widetilde{f}\|_n \ge \delta_n$. In this case, our goal is to show that $(*) \lesssim \delta_n\|\widehat{f} - \widetilde{f}\|_n$.

$$(*) = \left|\frac{1}{n}\sum_{i=1}^{n} w_i \underbrace{(\widehat{f}(x_i) - \widetilde{f}(x_i))\frac{\delta_n}{\|\widehat{f} - \widetilde{f}\|_n}}_{=:g(x_i)}\right| \frac{\|\widehat{f} - \widetilde{f}\|_n}{\delta_n}$$

Since $\partial\mathcal{F}$ is star-shaped, we have $g \in \partial\mathcal{F}$. Also observe that $\|g\|_n \le \delta_n$.

$$\lesssim \sup_{\substack{g\in\partial F\\ \|g\|_n\le\delta}} \left|\frac{1}{n}\sum_{i=1}^{n} w_i g(x_i)\right| \frac{\|\widehat{f} - \widetilde{f}\|_n}{\delta_n}$$

If we have an argument to show that this quantity concentrates around its mean, we get

$$\begin{aligned}
&\lesssim \mathcal{G}_n(\delta_n; \partial\mathcal{F})\frac{\|\widehat{f} - \widetilde{f}\|_n}{\delta_n} \\
&= \delta_n\|\widehat{f} - \widetilde{f}\|_n.
\end{aligned}$$

Using this line of argument, we can show that

$$\|\widehat{f} - f^*\|_n \le \|\widetilde{f} - f^*\|_n + 2\max\{\delta_n^2, \delta_n\|\widehat{f} - \widetilde{f}\|_n\}$$

The way to deal with the last term is to use the inequality

$$\begin{aligned}
\delta_n\|\widehat{f} - \widetilde{f}\|_n \le \delta_n(\|\widehat{f} - f^*\|_n + \|\widetilde{f} - f^*\|_n) &\le \frac{1}{\varepsilon}\delta_n^2 + \varepsilon(\|\widehat{f} - f^*\|_n + \|\widetilde{f} - f^*\|_n)^2 \\
&\le \frac{1}{\varepsilon}\delta_n^2 + 2\varepsilon\|\widehat{f} - f^*\|_n^2 + 2\varepsilon\|\widetilde{f} - f^*\|_n^2.
\end{aligned}$$

Here, we are using the Fenchel-Young inequality, $ab = (a/\sqrt{\varepsilon})(b\sqrt{\varepsilon}) \le (\frac{a}{\sqrt{\varepsilon}})^2 + (\sqrt{\varepsilon}b)^2$. $\square$

1.4 Applications of the oracle inequality

Example 1.4. Suppose $\{\phi_m\}_{m=1}^{\infty}$ is an orthogonal basis of $L^2(\mathbb{P})$, and let $\mathcal{F}_{\text{ortho}}(1,T) := \{f = \sum_{n=1}^{T} \beta_m \phi_m : \sum_{m=1}^{T} \beta_m^2 \leq 1$. If $f^* = \sum_{m=1}^{\infty} \theta_m^* \phi_m$, then $f^* \notin \mathcal{F}_{\text{ortho}}$. Using this oracle inequality, we can get

$$\|\widehat{f} - f^*\|_n^2 \lesssim \sum_{m>T}^{\infty} (\theta_m^*)^2 + \sigma^2 \frac{T}{n}.$$

The intuition is that if we have n samples, we can choose $T = \varepsilon n$ so that the right term is small. Then the error is roughly the contribution of the first term.

Example 1.5. Let $y_i = \langle x_i, \theta_* \rangle + \varepsilon_i$, and let $f_{\theta^*} = \langle \cdot, \theta_* \rangle$. Then consider the function class $\mathcal{F}_{\text{sparse}}(s) = \{f_\theta = \langle \cdot, \theta \rangle : \theta \in \mathbb{R}^d . \|\theta\|_0 \leq s\}$. Our estimator is then

$$\widehat{\theta} = \arg\min_{\|\theta\|_0 \leq s} \|y - X\theta\|_2^2.$$

This is the ℓ_0-variant of LASSO, which is not efficiently computable. Even if the model is not s-sparse, we get

$$\frac{\|X(\widetilde{\theta} - \theta^*)\|_2^2}{n} \leq \inf_{\|\theta\|_0 \leq s} \frac{\|X(\theta - \theta^*)\|_2^2}{n} + \frac{\delta_n^2}{n}.$$

Here, we know that

$$\delta_n^2 \lesssim \sigma^2 \frac{s \log(ed/s)}{n}.$$

In section 13.4.1 of Wainwright's book, there is a discussion of oracle inequalities for regularized estimators.