Nonparametric regression II

SDS 391P.6, Spring 2026

Pratik Patil

1 Motivation

In the previous lecture, we reduced fixed-design nonparametric regression to the normal means model

$$Y = \theta^\star + \sigma W, \qquad W \sim \mathcal{N}(0, I_n),$$

and we showed that the projection estimator is controlled by the *local Gaussian complexity*

$$\gamma\big(\Theta_{\theta^\star}(u)\big) = \mathbb{E} \sup_{\Delta \in \Theta_{\theta^\star}(u)} |\langle W, \Delta \rangle|.$$

This was conceptually satisfying, but not yet operational: to get an actual rate, we need to compute or bound that local Gaussian complexity.

For linear models, this was possible by direct geometric arguments. For general nonparametric classes, such as Lipschitz or Sobolev classes, the right tool is *metric entropy*. The principle here is to cover the localized function class at many scales, then chain.

The result is an entropy bound on local Gaussian complexity, and once combined with the theorem from the previous lecture, it yields prediction-error bounds for nonparametric least squares.

Today we will do three things:

1. define the localized Gaussian complexity of a function class;
2. bound it by Dudley's entropy integral;
3. apply this to Lipschitz regression and discuss the curse of dimensionality.

2 Localized Gaussian complexity for function classes

Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a function class, and let $f^\star \in \mathcal{F}$. Define the shifted class

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

For $\delta > 0$, define the localized empirical ball

$$\mathcal{F}^\star(\delta) := \{g \in \mathcal{F}^\star : \ \|g\|_n \leqslant \delta\}.$$

The relevant Gaussian complexity is

$$\mathcal{G}_n(\delta; \mathcal{F}^\star) := \mathbb{E}_W \sup_{g \in \mathcal{F}^\star(\delta)} \left| \frac{1}{n} \sum_{i=1}^n W_i g(x_i) \right|, \qquad W_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{1}$$

This is exactly the local Gaussian complexity of the induced set

$$\Theta_{\theta^\star}(\sqrt{n}\,\delta) = \{\Phi_n(g) : \ g \in \mathcal{F}^\star, \ \|g\|_n \leqslant \delta\},$$

scaled by $1/n$.

Indeed, since $\Phi_n(g) = (g(x_1), \ldots, g(x_n))$,

$$\mathcal{G}_n(\delta; \mathcal{F}^\star) = \frac{1}{n}\, \gamma\big(\Theta_{\theta^\star}(\sqrt{n}\,\delta)\big).$$

Therefore, the critical inequality from the previous lecture becomes: if

$$\mathcal{G}_n(\delta; \mathcal{F}^\star) \leqslant \frac{\kappa \delta^2}{2\sigma},$$

then the least-squares estimator $\widehat{f}$ satisfies

$$\|\widehat{f} - f^\star\|_n \leqslant \kappa\delta + \frac{2\sigma t}{\sqrt{n}}$$

with high probability.

So our next task is clear: upper bound $\mathcal{G}_n(\delta; \mathcal{F}^\star)$.

3 Dudley's entropy integral for local Gaussian complexity

3.1 Covering numbers in the empirical norm

For a function class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$, let

$$N_n(\varepsilon, \mathcal{H})$$

denote the covering number of $\mathcal{H}$ with respect to the empirical norm $\|\cdot\|_n$.

We also write

$$N_\infty(\varepsilon, \mathcal{H})$$

for the covering number in the sup norm

$$\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|.$$

Since $\|f\|_n \leqslant \|f\|_\infty$, any ε-cover in sup norm is also an ε-cover in empirical norm, so

$$N_n(\varepsilon, \mathcal{H}) \leqslant N_\infty(\varepsilon, \mathcal{H}). \tag{2}$$

3.2 Entropy bound

Proposition 3.1 (Localized Dudley bound)**.** *For any function class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ and any $\delta > 0$,*

$$\mathcal{G}_n(\delta; \mathcal{H}) \leqslant \frac{C}{\sqrt{n}} \int_0^\delta \sqrt{\log N_n(\varepsilon, \mathcal{H}(\delta))}\, d\varepsilon,$$

where

$$\mathcal{H}(\delta) := \{g \in \mathcal{H} : \ \|g\|_n \leqslant \delta\}.$$

Proof. Consider the Gaussian process indexed by $g \in \mathcal{H}(\delta)$:

$$X_g := \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i g(x_i).$$

It is centered Gaussian, and its canonical metric is

$$d(g,h)^2 = \mathbb{E}(X_g - X_h)^2 = \frac{1}{n} \sum_{i=1}^n (g(x_i) - h(x_i))^2 = \|g - h\|_n^2.$$

Therefore Dudley's inequality yields

$$\mathbb{E} \sup_{g \in \mathcal{H}(\delta)} |X_g| \leqslant C \int_0^\delta \sqrt{\log N_n(\varepsilon, \mathcal{H}(\delta))}\, d\varepsilon.$$

Finally,

$$\mathcal{G}_n(\delta; \mathcal{H}) = \frac{1}{\sqrt{n}} \mathbb{E} \sup_{g \in \mathcal{H}(\delta)} |X_g|,$$

which proves the claim. □

Combining Proposition 3.1 with the projection-estimator theorem from the previous lecture gives:

Corollary 3.2 (Entropy criterion for fixed-design regression)**.** *Assume that $\mathcal{F}^\star = \mathcal{F} - f^\star$ is star-shaped, and let $\widehat{f}$ be the constrained least-squares estimator*

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

Suppose $\delta > 0$ satisfies

$$\frac{C}{\sqrt{n}} \int_0^\delta \sqrt{\log N_n(\varepsilon, \mathcal{F}^\star(\delta))}\, d\varepsilon \leqslant \frac{\delta^2}{2\sigma}. \tag{3}$$

Then, for every $t \geqslant 0$, with probability at least $1 - e^{-t^2/2}$,

$$\|\widehat{f} - f^\star\|_n \leqslant C\delta + \frac{2\sigma t}{\sqrt{n}}.$$

Consequently,

$$\mathbb{E}\|\widehat{f} - f^\star\|_n^2 \lesssim \delta^2 + \frac{\sigma^2}{n}.$$

Remark 3.3. The most convenient way to use (3) is often to upper bound N_n by N_∞ using (2). This is sometimes crude, but it is easy and often already informative.

4 Lipschitz regression

We now apply the entropy criterion to a concrete class.

4.1 Function class

Fix $L > 0$, and consider

$$\mathcal{F}_L := \{f : [0,1] \to \mathbb{R} : \ f(0) = 0, \ \|f\|_{\mathrm{Lip}} \leqslant L\}.$$

Assume that $f^\star \in \mathcal{F}_L$, and let $\widehat{f}$ be the constrained least-squares estimator over $\mathcal{F}_L$.

Since

$$\mathcal{F}_L - f^\star \subseteq \mathcal{F}_{2L},$$

and $\mathcal{F}_{2L}$ is star-shaped, it suffices to control the entropy of $\mathcal{F}_{2L}$.

4.2 Covering numbers

A standard discretization argument gives

$$\log N_\infty(\varepsilon, \mathcal{F}_L) \leqslant C\,\frac{L}{\varepsilon}, \qquad \varepsilon \in (0,1]. \tag{4}$$

The idea is simple: sample the function values on an ε-grid of the interval $[0,1]$. Because the function is L-Lipschitz, once we know its value at one grid point, there are only $O(1)$ admissible choices at the next grid point. Thus the total number of approximants grows like $\exp(CL/\varepsilon)$.

4.3 Prediction rate

Using (4) and $N_n \leqslant N_\infty$, Proposition 3.1 gives

$$\mathcal{G}_n(\delta; \mathcal{F}_L - f^\star) \lesssim \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\frac{L}{\varepsilon}}\, d\varepsilon \lesssim \sqrt{\frac{L\delta}{n}}.$$

So the critical inequality $\mathcal{G}_n(\delta) \leqslant \delta^2/(2\sigma)$ holds when

$$\sqrt{\frac{L\delta}{n}} \lesssim \frac{\delta^2}{\sigma}.$$

Solving for δ yields

$$\delta \asymp \left(\frac{L\sigma^2}{n}\right)^{1/3}.$$

Therefore Corollary 3.2 implies

$$\|\widehat{f} - f^\star\|_n \lesssim \left(\frac{L\sigma^2}{n}\right)^{1/3}$$

with high probability, and hence

$$\mathbb{E}\|\widehat{f} - f^\star\|_n^2 \lesssim \left(\frac{L\sigma^2}{n}\right)^{2/3}. \tag{5}$$

Remark 4.1. This is the classical one-dimensional Lipschitz regression rate for fixed-design prediction error. It is much slower than the parametric rate n^{-1}, because the class $\mathcal{F}_L$ is infinite-dimensional.

5 A useful comparison: smoother classes give faster rates

The Lipschitz class is only one example. Suppose more generally that a function class obeys an entropy bound of the form

$$\log N_\infty(\varepsilon, \mathcal{F}) \lesssim \varepsilon^{-1/\alpha}$$

for some $\alpha > 0$. Then the same calculation gives

$$\mathcal{G}_n(\delta) \lesssim \frac{1}{\sqrt{n}} \int_0^\delta \varepsilon^{-1/(2\alpha)}\, d\varepsilon \lesssim \frac{\delta^{1-\frac{1}{2\alpha}}}{\sqrt{n}},$$

provided $\alpha > 1/2$. The critical inequality then yields

$$\delta^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}}.$$

So more regular function classes have smaller entropy, and therefore faster regression rates.

This is one precise mathematical way to say that extra smoothness reduces statistical complexity.

6 The curse of dimensionality

What happens for Lipschitz functions on $[0,1]^d$ instead of $[0,1]$?

Very roughly, the metric entropy becomes much larger:

$$\log N_\infty(\varepsilon, \mathcal{F}_L([0,1]^d)) \asymp \varepsilon^{-d}.$$

So the size of the class grows much faster as the ambient dimension d increases. This means the entropy integral becomes much larger, and the resulting regression rate deteriorates rapidly with d.

This phenomenon is called the *curse of dimensionality*: if all we assume is global Lipschitz regularity in a high-dimensional ambient space, then consistent estimation requires a huge amount of data.

Modern machine learning works precisely by exploiting *structure* beyond ambient dimension. Examples include:

- sparsity;
- additive structure;
- low-rank structure;
- kernel methods and RKHS constraints;
- compositional structure, such as that encoded by neural networks.

In each case, the point is the same: the *effective complexity* can be much smaller than what the ambient dimension alone would suggest.

7 Summary

The full pipeline for fixed-design nonparametric regression is now visible:

1. Start with the regression model
$$Y_i = f^\star(x_i) + \sigma W_i.$$
2. Reduce it to a normal means problem over the sampled function values.
3. Control the least-squares estimator by a local Gaussian complexity.
4. Bound that local Gaussian complexity using Dudley's entropy integral.
5. Solve the resulting critical inequality to obtain a prediction rate.

For one-dimensional Lipschitz regression, this gives

$$\mathbb{E}\|\widehat{f} - f^\star\|_n^2 \lesssim \left(\frac{L\sigma^2}{n}\right)^{2/3}.$$

The same blueprint applies much more broadly. What changes from one problem to the next is the complexity calculation: sometimes it is a direct Gaussian-width argument, sometimes a metric-entropy argument, and sometimes something more refined.

8 Look ahead

In the next step of the course, we can go in two directions.

One direction is to keep developing nonparametric regression: study smoother classes such as Sobolev or RKHS balls, derive sharper oracle inequalities, and compare constrained versus penalized estimators.

The other direction is to move toward minimax lower bounds: once we know how to prove upper bounds, we should ask whether they are optimal. That question leads to information-theoretic methods and packing arguments, which form the lower-bound counterpart to the entropy tools we used today.

Most likely, we will do minimax lower bounds as we have only one more week until the end!

Source material

Parts of this lecture are based on references: Wainwright (2019), in addition to the author's accumulated experience working on related topics.

References

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.