# Nonparametric regression I

SDS 391P.6, Spring 2026

Pratik Patil



## 1 Motivation

So far in the course, we studied two main themes:

- concentration and random matrix tools for controlling random quantities;
- complexity measures such as Gaussian width, covering numbers, Dudley's entropy integral, Rademacher complexity, and VC dimension.

We now begin a new module on *nonparametric regression*. The basic statistical task is simple to state:

> Given noisy observations of an unknown function $f^\star$, how accurately can we estimate $f^\star$?

Unlike linear regression, we do not assume that $f^\star$ belongs to a low-dimensional parametric family. Instead, we only assume that it belongs to a larger class of functions $\mathcal{F}$, such as a Lipschitz class, a Sobolev class, or an RKHS ball. The class $\mathcal{F}$ may be infinite-dimensional.

A central message of this module is that nonparametric regression can often be reduced to a geometric estimation problem in Euclidean space. The prototype is the *normal means model*, where one observes

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

and estimates $\theta^\star$ under some structural assumption $\theta^\star \in \Theta \subset \mathbb{R}^n$. This model looks much simpler than regression, but it already contains the key geometry.

The plan for today is:

1. define the fixed-design nonparametric regression problem and the error criterion we care about;
2. reduce it to the normal means model;
3. prove a general theorem for projection estimators in the normal means model;
4. work out two examples: dense linear regression and sparse linear regression.

The next lecture will return to general function classes and show how Dudley's entropy integral turns metric entropy bounds into regression rates.

# 2 Fixed-design nonparametric regression

## 2.1 Observation model

Let $x_1, \dots, x_n \in \mathcal{X}$ be fixed design points. We observe responses

$$Y_i = f^\star(x_i) + \sigma W_i, \qquad W_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad i = 1, \dots, n, \tag{1}$$

where $f^\star : \mathcal{X} \to \mathbb{R}$ is the unknown regression function and $\sigma > 0$ is the noise level.

We assume that $f^\star$ belongs to a function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, and we estimate it by constrained least squares:

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

## 2.2 Error criterion

Since we are working with fixed design points, the most natural loss is the *empirical prediction error*

$$\|\widehat{f} - f^\star\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} \big(\widehat{f}(x_i) - f^\star(x_i)\big)^2.$$

This is exactly the squared $L^2$ error over the observed covariates. We will often write

$$\|f\|_n := \left( \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2 \right)^{1/2}$$

for the associated empirical norm.

**Remark 2.1.** There are several natural notions of regression error.

- $\|\widehat{f} - f^\star\|_n^2$: the *fixed-design prediction error*, which is our main focus in this lecture.
- $\mathbb{E}_X[(\widehat{f}(X) - f^\star(X))^2]$: the *population prediction error*, relevant for random-design regression.
- $\|\widehat{\theta} - \theta^\star\|_2^2$: the *parameter estimation error*, when the model is parameterized by some vector $\theta$.

These are related but distinct questions. Today we focus on the first one.

# 3 Reduction to the normal means model

## 3.1 Sampling operator

Define the linear sampling map

$$\Phi_n : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^n, \qquad \Phi_n(f) := \big(f(x_1), \dots, f(x_n)\big).$$

Let

$$\Theta := \Phi_n(\mathcal{F}) \subset \mathbb{R}^n, \qquad \theta^\star := \Phi_n(f^\star) \in \Theta, \qquad Y := (Y_1, \dots, Y_n) \in \mathbb{R}^n.$$

Then the model (1) becomes

$$Y = \theta^\star + \sigma W, \qquad W \sim \mathcal{N}(0, I_n). \tag{3}$$

Moreover, the least-squares estimator (2) is exactly the Euclidean projection estimator

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta} \|Y - \theta\|_2^2, \qquad \widehat{\theta} = \Phi_n(\widehat{f}). \tag{4}$$

Finally, note that the regression error becomes

$$\|\widehat{f} - f^\star\|_n^2 = \frac{1}{n}\|\widehat{\theta} - \theta^\star\|_2^2. \tag{5}$$

So the fixed-design regression problem has been reduced to studying Euclidean projection onto the set $\Theta \subset \mathbb{R}^n$.

### 3.2 Local geometry

Given $\theta^\star \in \Theta$, define the shifted set

$$\Theta_{\theta^\star} := \Theta - \theta^\star = \{\theta - \theta^\star : \theta \in \Theta\}.$$

For $u > 0$, define the localized set

$$\Theta_{\theta^\star}(u) := \Theta_{\theta^\star} \cap uB_2^n = \{\Delta \in \Theta_{\theta^\star} : \ \|\Delta\|_2 \leqslant u\}.$$

We will also use the *local Gaussian complexity*

$$\gamma\big(\Theta_{\theta^\star}(u)\big) := \mathbb{E} \sup_{\Delta \in \Theta_{\theta^\star}(u)} |\langle W, \Delta\rangle|, \qquad W \sim \mathcal{N}(0, I_n).$$

This is a localized version of Gaussian width. It measures how large the feasible perturbations around $\theta^\star$ look to a Gaussian noise vector.

**Definition 3.1** (Star-shaped sets)**.** A set $A \subset \mathbb{R}^n$ is called *star-shaped* if

$$\Delta \in A, \ \alpha \in [0,1] \qquad \Longrightarrow \qquad \alpha\Delta \in A.$$

We say that $\Theta$ is star-shaped around $\theta^\star$ if $\Theta - \theta^\star$ is star-shaped.

This condition will allow us to pass from local control at radius $u$ to global control at larger radii.

## 4 A general theorem for projection estimators

We now prove a general bound for the projection estimator (4).

**Theorem 4.1** (Projection estimator in the normal means model)**.** *Assume that $Y = \theta^\star + \sigma W$ with $W \sim \mathcal{N}(0, I_n)$, and that $\Theta_{\theta^\star}$ is star-shaped. Let $\widehat{\theta}$ be any solution of (4).*

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

$$\|\widehat{\theta} - \theta^\star\|_2 \ \leqslant \ \max\left\{ \frac{2\sigma}{u}\Big(\gamma(\Theta_{\theta^\star}(u)) + tu\Big), \sqrt{2\sigma\Big(\gamma(\Theta_{\theta^\star}(u)) + tu\Big)} \right\}.$$

The theorem has a clean interpretation: if the local Gaussian complexity near $\theta^\star$ is small, then the projection estimator must be close to $\theta^\star$.

### 4.1 Two lemmas

**Lemma 4.2** (Gaussian concentration for the local supremum)**.** *Fix $u > 0$. Then*

$$Z_u(W) := \sup_{\Delta \in \Theta_{\theta^\star}(u)} |\langle W, \Delta \rangle|$$

*is $u$-Lipschitz as a function of $W \in \mathbb{R}^n$. Consequently, for every $t \geqslant 0$,*

$$\mathbb{P}\left\{Z_u(W) \geqslant \mathbb{E}Z_u(W) + tu\right\} \leqslant e^{-t^2/2}.$$

*Proof.* For $W, W' \in \mathbb{R}^n$,

$$|Z_u(W) - Z_u(W')| \leqslant \sup_{\Delta \in \Theta_{\theta^\star}(u)} |\langle W - W', \Delta \rangle| \leqslant u\|W - W'\|_2$$

by Cauchy–Schwarz. Thus $Z_u$ is $u$-Lipschitz. The concentration bound is Gaussian concentration for Lipschitz functions. □

**Lemma 4.3** (Star-shaped scaling)**.** *Assume $\Theta_{\theta^\star}$ is star-shaped. Fix $u > 0$. Then for every $\Delta \in \Theta_{\theta^\star}$,*

$$|\langle W, \Delta \rangle| \leqslant \max\left\{\frac{\|\Delta\|_2}{u}, 1\right\} \sup_{\widetilde{\Delta} \in \Theta_{\theta^\star}(u)} |\langle W, \widetilde{\Delta} \rangle|.$$

*Proof.* If $\|\Delta\|_2 \leqslant u$, there is nothing to prove. If $\|\Delta\|_2 > u$, define

$$\widetilde{\Delta} := \frac{u}{\|\Delta\|_2}\Delta.$$

Since $\Theta_{\theta^\star}$ is star-shaped, $\widetilde{\Delta} \in \Theta_{\theta^\star}(u)$. Also,

$$|\langle W, \Delta \rangle| = \frac{\|\Delta\|_2}{u} |\langle W, \widetilde{\Delta} \rangle| \leqslant \frac{\|\Delta\|_2}{u} \sup_{\widetilde{\Delta} \in \Theta_{\theta^\star}(u)} |\langle W, \widetilde{\Delta} \rangle|.$$

□

### 4.2 Proof of Theorem 4.1

*Proof.* Let

$$\widehat{\Delta} := \widehat{\theta} - \theta^\star \in \Theta_{\theta^\star}.$$

Since $\widehat{\theta}$ minimizes $\|Y - \theta\|_2^2$ over $\Theta$, and $\theta^\star \in \Theta$, we have

$$\|Y - \widehat{\theta}\|_2^2 \leqslant \|Y - \theta^\star\|_2^2.$$

Substitute $Y = \theta^\star + \sigma W$:

$$\|\sigma W - \widehat{\Delta}\|_2^2 \leqslant \|\sigma W\|_2^2.$$

Expanding and canceling $\sigma^2\|W\|_2^2$ gives the basic inequality

$$\frac{1}{2}\|\widehat{\Delta}\|_2^2 \leqslant \sigma\langle W, \widehat{\Delta} \rangle.$$

Hence

$$\frac{1}{2}\|\widehat{\Delta}\|_2^2 \leqslant \sigma|\langle W, \widehat{\Delta} \rangle|.$$

Now apply Lemma 4.2 and Lemma 4.3. With probability at least $1 - e^{-t^2/2}$,

$$|\langle W, \widehat{\Delta} \rangle| \leqslant \max\left\{ \frac{\|\widehat{\Delta}\|_2}{u}, 1 \right\} \Big( \gamma(\Theta_{\theta^\star}(u)) + tu \Big).$$

Therefore

$$\frac{1}{2}\|\widehat{\Delta}\|_2^2 \leqslant \sigma \max\left\{ \frac{\|\widehat{\Delta}\|_2}{u}, 1 \right\} \Big( \gamma(\Theta_{\theta^\star}(u)) + tu \Big).$$

If $\|\widehat{\Delta}\|_2 \geqslant u$, then

$$\frac{1}{2}\|\widehat{\Delta}\|_2^2 \leqslant \sigma \frac{\|\widehat{\Delta}\|_2}{u} \Big( \gamma(\Theta_{\theta^\star}(u)) + tu \Big),$$

so

$$\|\widehat{\Delta}\|_2 \leqslant \frac{2\sigma}{u} \Big( \gamma(\Theta_{\theta^\star}(u)) + tu \Big).$$

If $\|\widehat{\Delta}\|_2 \leqslant u$, then

$$\frac{1}{2}\|\widehat{\Delta}\|_2^2 \leqslant \sigma \Big( \gamma(\Theta_{\theta^\star}(u)) + tu \Big),$$

so

$$\|\widehat{\Delta}\|_2 \leqslant \sqrt{2\sigma \Big( \gamma(\Theta_{\theta^\star}(u)) + tu \Big)}.$$

Combining the two cases completes the proof. □

## 5 Critical radius and a simpler corollary

The theorem becomes especially transparent if the local complexity obeys a quadratic bound.

**Corollary 5.1** (Critical radius bound)**.** *Suppose $u > 0$ satisfies*

$$\gamma(\Theta_{\theta^\star}(u)) \leqslant \frac{\kappa u^2}{2\sigma}$$

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

$$\|\widehat{\theta} - \theta^\star\|_2 \leqslant \kappa u + 2\sigma t.$$

*Proof.* Under the assumed bound,

$$\frac{2\sigma}{u} \Big( \gamma(\Theta_{\theta^\star}(u)) + tu \Big) \leqslant \kappa u + 2\sigma t.$$

Also,

$$2\sigma \Big( \gamma(\Theta_{\theta^\star}(u)) + tu \Big) \leqslant \kappa u^2 + 2\sigma t u \leqslant (\kappa u + 2\sigma t)^2$$

after enlarging constants slightly if desired. The result follows from Theorem 4.1. □

Using (5), the same statement translates to regression as

$$\|\widehat{f} - f^\star\|_n = \frac{1}{\sqrt{n}}\|\widehat{\theta} - \theta^\star\|_2 \leqslant \frac{\kappa u}{\sqrt{n}} + \frac{2\sigma t}{\sqrt{n}}.$$

## 6 Example 1: dense linear regression

Consider the linear function class

$$\mathcal{F}_{\text{lin}} := \{x \mapsto \langle \beta, x \rangle : \ \beta \in \mathbb{R}^d\},$$

with fixed design matrix $X \in \mathbb{R}^{n \times d}$, whose $i$-th row is $x_i^\top$. Then

$$\Theta = \{X\beta : \ \beta \in \mathbb{R}^d\} = \text{range}(X) \subset \mathbb{R}^n.$$

Since $\Theta$ is a linear subspace, it is star-shaped around every $\theta^\star \in \Theta$.

Let $r := \text{rank}(X)$. Then

$$\Theta_{\theta^\star}(u) = \text{range}(X) \cap uB_2^n.$$

If $\Pi_X$ denotes the orthogonal projector onto range$(X)$, then

$$\gamma(\Theta_{\theta^\star}(u)) = u\,\mathbb{E}\|\Pi_X W\|_2 \leqslant u\sqrt{\mathbb{E}\|\Pi_X W\|_2^2} = u\sqrt{\text{tr}(\Pi_X)} = u\sqrt{r}.$$

So the critical inequality holds as long as

$$u\sqrt{r} \leqslant \frac{\kappa u^2}{2\sigma}, \qquad \text{i.e.} \qquad u \gtrsim \sigma\frac{\sqrt{r}}{\kappa}.$$

Choosing $\kappa$ as an absolute constant, Corollary 5.1 yields

$$\|\widehat{\theta} - \theta^\star\|_2 \lesssim \sigma\sqrt{r}$$

with high probability, hence

$$\|\widehat{f} - f^\star\|_n^2 = \frac{1}{n}\|\widehat{\theta} - \theta^\star\|_2^2 \lesssim \frac{\sigma^2 r}{n}.$$

This recovers the familiar prediction-error rate for least squares: the effective dimension is the rank of the design matrix.

## 7 Example 2: sparse linear regression

Now consider the sparse linear class

$$\mathcal{F}_s := \{x \mapsto \langle \beta, x \rangle : \ \|\beta\|_0 \leqslant s\}.$$

Then

$$\Theta = \{X\beta : \ \|\beta\|_0 \leqslant s\} \subset \mathbb{R}^n.$$

Unlike the dense case, $\Theta$ is no longer a linear subspace. Still, its local geometry can be controlled.

Let $\theta^\star = X\beta^\star$ with $\|\beta^\star\|_0 \leqslant s$. Then

$$\Theta - \theta^\star \subseteq \{X\delta : \ \|\delta\|_0 \leqslant 2s\}.$$

So it suffices to study

$$\mathcal{U}_{2s}(u) := \{X\delta : \ \|\delta\|_0 \leqslant 2s, \ \|X\delta\|_2 \leqslant u\}.$$

Assume, for simplicity, that the design has been normalized so that each column satisfies

$$\frac{\|X_j\|_2}{\sqrt{n}} \leqslant 1.$$

Then one can show

$$\gamma(\mathcal{U}_{2s}(u)) \lesssim u\sqrt{s\log\Big(\frac{ed}{s}\Big)}. \tag{6}$$

The proof uses two ingredients: for each fixed support $S \subset [d]$ with $|S| = 2s$, the set of vectors $X\delta$ supported on $S$ lies in a $2s$-dimensional subspace, and there are $\binom{d}{2s}$ possible supports.

Plugging (6) into the critical inequality gives

$$u\sqrt{s\log(ed/s)} \leqslant \frac{\kappa u^2}{2\sigma},$$

so

$$u \gtrsim \sigma\frac{\sqrt{s\log(ed/s)}}{\kappa}.$$

Therefore

$$\|\widehat{\theta} - \theta^\star\|_2^2 \lesssim \sigma^2 s\log\Big(\frac{ed}{s}\Big),$$

and hence

$$\|\widehat{f} - f^\star\|_n^2 = \frac{1}{n}\|\widehat{\theta} - \theta^\star\|_2^2 \lesssim \frac{\sigma^2 s\log(ed/s)}{n}.$$

This is the usual sparse-regression prediction rate: the effective dimension is $s\log(ed/s)$, not $d$.

## 8 Look ahead

Today we reduced fixed-design nonparametric regression to the normal means model and proved a general projection-estimator theorem driven by local Gaussian complexity. The two examples of dense and sparse linear regression shows the main idea that the estimation error is controlled by localized geometric complexity:

- For dense linear regression, the local complexity is that of an $r$-dimensional subspace.
- For sparse linear regression, the local complexity is that of a union of many low-dimensional subspaces, which introduces the extra $\log(ed/s)$ factor.

So the question "how hard is the problem?" becomes a geometric question about the size of the feasible perturbation set near the truth.

In the next lecture, we return to general function classes $\mathcal{F}$. The key missing ingredient is how to bound the local Gaussian complexity in terms of a more explicit complexity measure. This is where Dudley's entropy integral enters:

$$\text{metric entropy} \quad \Longrightarrow \quad \text{local Gaussian complexity} \quad \Longrightarrow \quad \text{prediction-error rates.}$$

We will use this to analyze Lipschitz regression and to see, in a concrete way, the curse of dimensionality.

## 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.