ECON 7377 Microeconometrics **Spring 2026**

Nonparametric Statistics

Contents

Overview. These notes cover the core theory of nonparametric estimation at a PhD-statistics level: (i) upper bounds for kernel density estimation and nonparametric least squares, built on concentration and empirical-process tools; (ii) minimax *lower* bounds via reduction to testing, using Le Cam's two-point method, Fano's method, and the Yang–Barron mutual-information method; and (iii) generalization error for empirical risk minimization in supervised learning. The material is largely self-contained and follows closely Tsybakov (2009), van der Vaart and Wellner (2023) and Wainwright (2019), with the high-level organization mirroring a graduate course on nonparametric statistics.

Notation. Throughout, $\{X_i\}_{i=1}^n$ are i.i.d. random variables with law P on a sample space $\mathcal{X}$ (typically $\mathcal{X} \subseteq \mathbb{R}^d$), with density $f = dP/d\nu$ when P is absolutely continuous with respect to a σ-finite dominating measure ν (usually Lebesgue). The product measure for n i.i.d. draws is $P^{\otimes n}$ or, when no confusion arises, P^n. The empirical measure is $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, and $\|g\|_n^2 := \frac{1}{n}\sum_{i=1}^n g(X_i)^2$ is the $L^2(\mathbb{P}_n)$-semi-norm. We use $\mathbb{E}$ (resp. $\mathbb{E}_\theta$, $\mathbb{E}_f$) for expectation under P (resp.

P_θ, P_f), Var for variance, and $\mathbb{1}$ for the indicator. Function classes are denoted $\mathcal{F}, \mathcal{H}, \mathcal{S}, \ldots$; Hölder classes $\mathcal{H}(\beta, L)$ and Sobolev classes $\mathcal{S}(\beta, L)$ are defined below. The symbol $a_n \lesssim b_n$ means $a_n \le C b_n$ for some absolute constant C not depending on n; $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We write $[m] = \{1, \ldots, m\}$.

Part I

Upper Bounds

1 Density Estimation via Kernels

1.1 From empirical CDF to the Rosenblatt estimator

Density estimation is the problem of recovering the probability density function $f = dP/d\nu$ from an i.i.d. sample $X_1, \ldots, X_n \sim P$. For $d = 1$, the cdf $F(x) = P(X_1 \le x)$ satisfies $f = F'$, so for small $h > 0$

$$f(x) \;\approx\; \frac{F(x+h) - F(x-h)}{2h}. \tag{1}$$

A natural estimator replaces F by its empirical counterpart $\widehat{F}_n(x) = n^{-1} \sum_{i=1}^n \mathbb{1}\{X_i \le x\}$. This produces the *Rosenblatt estimator*

$$\widehat{f}_h^n(x) := \frac{\widehat{F}_n(x+h) - \widehat{F}_n(x-h)}{2h} \;=\; \frac{1}{2nh} \sum_{i=1}^n \mathbb{1}\left\{-1 < \frac{X_i - x}{h} \le 1\right\}.$$

Writing $K_0(u) = \frac{1}{2}\mathbb{1}\{-1 < u \le 1\}$ (the *uniform*, or box-car, kernel), we recognize this as

$$\widehat{f}_h^n(x) = \frac{1}{nh} \sum_{i=1}^n K_0\left(\frac{X_i - x}{h}\right). \tag{2}$$

Replacing K_0 by a general integrable function with $\int K = 1$ yields the Parzen–Rosenblatt kernel density estimator.

1.2 Kernel density estimator

Definition 1 (Kernel density estimator)**.** Let $K : \mathbb{R}^d \to \mathbb{R}$ be integrable with $\int_{\mathbb{R}^d} K(u)\,du = 1$, and let $h > 0$ be a bandwidth. The kernel density estimator (KDE) of f at $x \in \mathbb{R}^d$ is

$$\widehat{f}_n(x) \;:=\; \frac{1}{nh^d} \sum_{i=1}^{n} K\!\left(\frac{X_i - x}{h}\right). \tag{3}$$

Example 1 (Classical kernels on $\mathbb{R}$)**.** • *Uniform:* $K(u) = \frac{1}{2}\mathbb{1}\{|u| \le 1\}$.

- *Triangular:* $K(u) = (1 - |u|)\mathbb{1}\{|u| \le 1\}$.
- *Epanechnikov (parabolic):* $K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}\{|u| \le 1\}$.
- *Biweight:* $K(u) = \frac{15}{16}(1 - u^2)^2\,\mathbb{1}\{|u| \le 1\}$.
- *Gaussian:* $K(u) = (2\pi)^{-1/2} e^{-u^2/2}$.

Remark 1 (Multivariate KDE)**.** For $d \ge 2$, the simplest construction is the *product kernel* $K(u_1, \dots, u_d) = \prod_{j=1}^{d} K_1(u_j)$ built from a univariate kernel K_1. The estimator (3) becomes

$$\widehat{f}_n(x) \;=\; \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{d} K_1\big((X_{ij} - x_j)/h\big).$$

More generally, one may allow a positive-definite bandwidth matrix H and use $K\big(H^{-1/2}(X_i - x)\big)/\det(H)^{1/2}$.

1.3 Pointwise bias–variance decomposition

For a fixed target point $x_0 \in \mathbb{R}^d$, the pointwise mean squared error decomposes in the usual way:

$$\begin{aligned} \mathrm{MSE}(x_0) &:= \mathbb{E}_P\big[\widehat{f}_n(x_0) - f(x_0)\big]^2 \\ &= \underbrace{\mathbb{E}_P\big[\widehat{f}_n(x_0) - \mathbb{E}_P\widehat{f}_n(x_0)\big]^2}_{=:\,\sigma^2(x_0)\ \text{(variance)}} + \underbrace{\big[\mathbb{E}_P\widehat{f}_n(x_0) - f(x_0)\big]^2}_{=:\,b^2(x_0)\ \text{(squared bias)}}. \end{aligned} \tag{4}$$

We control variance and bias separately.

Lemma 1 (Variance of KDE)**.** *Assume $d = 1$, $\sup_x f(x) \le \bar{f} < \infty$, and $\int_{\mathbb{R}} K^2(u)\,du < \infty$. Then*

$$\sigma^2(x_0) \;\le\; \frac{\bar{f}}{nh} \int_{\mathbb{R}} K^2(u)\,du.$$

In particular, if $h = h_n$ satisfies $nh_n \to \infty$, then $\sigma^2(x_0) \to 0$.

Proof. Set $\delta_i := K((X_i - x_0)/h) - \mathbb{E}K((X_i - x_0)/h)$; these are i.i.d. mean-zero and

$$\sigma^2(x_0) = \operatorname{Var}\left(\frac{1}{nh}\sum_{i=1}^{n}\delta_i\right) = \frac{1}{nh^2}\operatorname{Var}(\delta_1) \le \frac{1}{nh^2}\,\mathbb{E}\left[K^2\left(\frac{X_1 - x_0}{h}\right)\right]$$
$$= \frac{1}{nh^2}\int_{\mathbb{R}} K^2\left(\frac{u - x_0}{h}\right) f(u)\,du \;\le\; \frac{\bar{f}}{nh}\int_{\mathbb{R}} K^2(v)\,dv,$$

where the last step uses the substitution $v = (u - x_0)/h$. □

To control the bias we need smoothness assumptions on f.

Definition 2 (Hölder class on $\mathbb{R}$)**.** For $\beta, L > 0$, set

$$\ell := \lfloor\beta\rfloor_< \;:=\; \text{the largest integer } \textit{strictly} \text{ less than } \beta.$$

The Hölder class $\mathcal{H}(\beta, L)$ consists of all ℓ-times differentiable $f : \mathbb{R} \to \mathbb{R}$ such that

$$\left|f^{(\ell)}(x) - f^{(\ell)}(x')\right| \;\le\; L\,|x - x'|^{\beta-\ell} \qquad \forall\, x, x' \in \mathbb{R}.$$

Let $\mathcal{P}(\beta, L) := \{f \in \mathcal{H}(\beta, L) \,:\, f \ge 0,\ \int_{\mathbb{R}} f(x)\,dx = 1\}$ denote the induced class of densities.

Remark 2 (Floor convention)**.** We follow Tsybakov (2009, Def. 1.2): when β is not an integer, $\lfloor\beta\rfloor_<$ agrees with the standard floor $\lfloor\beta\rfloor$. When $\beta \in \{1, 2, 3, \dots\}$ is an integer, however, $\lfloor\beta\rfloor_< = \beta - 1$ (not β), so that the exponent $\beta - \ell > 0$ is always positive and the Hölder condition is non-trivial. For example, $\beta = 2$ gives $\ell = 1$: f is once-differentiable with f' Lipschitz, i.e., $f \in C^{1,1}$. Throughout these notes we write $\lfloor\beta\rfloor$ for $\lfloor\beta\rfloor_<$ with the understanding fixed here.

Definition 3 (Kernel of order ℓ)**.** A kernel $K : \mathbb{R} \to \mathbb{R}$ is of *order* $\ell \ge 1$ if $u \mapsto u^j K(u)$ is integrable for $j = 0, 1, \dots, \ell$ and

$$\int K(u)\,du = 1, \qquad \int u^j K(u)\,du = 0 \quad \text{for all } j = 1, \dots, \ell.$$

Classical symmetric kernels (uniform, triangular, Epanechnikov, Gaussian) are of order 1. Kernels of higher order are sometimes called *bias-reducing*; they necessarily take negative values, as shown in Section 1.4.

Lemma 2 (Pointwise bias of KDE on the Hölder class; cf. Tsybakov, 2009, Prop. 1.2)**.** *Suppose $f \in \mathcal{P}(\beta, L)$, K is a kernel of order $\ell = \lfloor\beta\rfloor$, and $\int_{\mathbb{R}} |u|^\beta |K(u)|\,du < \infty$. Then, for every $x_0 \in \mathbb{R}$*

and every $h > 0$,

$$\big|b(x_0)\big| \;=\; \big|\mathbb{E}_P \widehat{f}_n(x_0) - f(x_0)\big| \;\le\; \frac{L}{\ell!}\int_{\mathbb{R}} |u|^\beta |K(u)|\, du \cdot h^\beta.$$

Proof. A direct substitution gives

$$\mathbb{E}_P \widehat{f}_n(x_0) - f(x_0) = \int_{\mathbb{R}} K(u)\,\big[f(x_0 + hu) - f(x_0)\big]\, du.$$

By the integral form of Taylor's theorem with remainder (and since $f \in \mathcal{H}(\beta, L)$), for some $\tau = \tau(u) \in (0,1)$,

$$f(x_0 + hu) = \sum_{j=0}^{\ell-1} \frac{(hu)^j}{j!}\, f^{(j)}(x_0) + \frac{(hu)^\ell}{\ell!}\, f^{(\ell)}(x_0 + \tau hu).$$

Using $\int u^j K(u)du = 0$ for $j = 1, \dots, \ell - 1$ and the trivial identity $0 = \int u^\ell K(u)du \cdot f^{(\ell)}(x_0)/\ell!$ (since K has order ℓ), we obtain

$$\mathbb{E}_P \widehat{f}_n(x_0) - f(x_0) = \frac{h^\ell}{\ell!}\int_{\mathbb{R}} u^\ell K(u)\,\big[f^{(\ell)}(x_0 + \tau hu) - f^{(\ell)}(x_0)\big]\, du.$$

Applying the Hölder bound $|f^{(\ell)}(x_0 + \tau hu) - f^{(\ell)}(x_0)| \le L|\tau hu|^{\beta-\ell} \le L|hu|^{\beta-\ell}$ and taking absolute values yields the claim. □

Combining Lemmas 1 and 2,

$$\sup_{f \in \mathcal{P}(\beta, L)} \mathrm{MSE}(x_0) \;\le\; \frac{C_1}{nh} + C_2\, h^{2\beta}. \tag{5}$$

Optimizing over $h > 0$ gives the *pointwise minimax upper bound*

$$\inf_{h>0}\left[\frac{C_1}{nh} + C_2 h^{2\beta}\right] \asymp n^{-\frac{2\beta}{2\beta+1}}, \qquad h_n^* \asymp n^{-\frac{1}{2\beta+1}}. \tag{6}$$

For $\mathbb{R}^d$ with a product kernel of order ℓ, the same argument with h^d replacing h in the variance gives

$$\mathrm{MSE}(x_0) \lesssim \frac{1}{nh^d} + h^{2\beta} \;\;\Longrightarrow\;\; \mathrm{MSE}(x_0) \asymp n^{-\frac{2\beta}{2\beta+d}} \quad \text{at} \quad h_n^* \asymp n^{-\frac{1}{2\beta+d}},$$

which exhibits the *curse of dimensionality*: the rate deteriorates as d grows. The matching lower bound $n^{-2\beta/(2\beta+1)}$ is proved in Section 2.5 via Le Cam's method.

1.4 Construction of higher-order kernels via Legendre polynomials

For $\beta > 1$, Lemma 2 requires a kernel of order $\ell = \lfloor \beta \rfloor \geq 1$ that also has bounded moment $\int |u|^\beta |K(u)| du < \infty$. We now show such kernels exist and construct them explicitly on $[-1, 1]$.

Consider $L^2([-1, 1], \lambda)$. The *Legendre polynomials*

$$\varphi_0(x) := \frac{1}{\sqrt{2}}, \qquad \varphi_m(x) := \sqrt{\frac{2m+1}{2}} \cdot \frac{1}{2^m m!} \frac{d^m}{dx^m}\left[(x^2-1)^m\right] \quad (m \geq 1) \tag{7}$$

form an orthonormal basis of this space. The polynomial φ_m has degree m; in particular $\langle u^p, \varphi_q \rangle = 0$ whenever $p < q$.

Proposition 1 (Order-ℓ kernel via Legendre polynomials; Tsybakov, 2009, Prop. 1.3)**.** Define

$$K(u) \;:=\; \left(\sum_{m=0}^{\ell} \varphi_m(0)\, \varphi_m(u) \right) \mathbb{1}\{|u| \leq 1\}.$$

Then K is a kernel of order ℓ.

Proof. Integrability of $u \mapsto u^j K(u)$ on $[-1, 1]$ is immediate since each φ_m is a bounded polynomial. For any $j \in \{0, 1, \ldots, \ell\}$, express u^j in the Legendre basis as

$$u^j = \sum_{q=0}^{j} b_{j,q}\, \varphi_q(u), \qquad b_{j,q} := \langle u^j, \varphi_q \rangle_{L^2[-1,1]},$$

where the sum truncates at $q = j$ because $\langle u^p, \varphi_q \rangle = 0$ for $p < q$. Using orthonormality $\langle \varphi_q, \varphi_m \rangle = \delta_{qm}$,

$$\begin{aligned} \int_{-1}^{1} u^j K(u)\, du &= \int_{-1}^{1} \left(\sum_{q=0}^{j} b_{j,q}\, \varphi_q(u) \right) \left(\sum_{m=0}^{\ell} \varphi_m(0)\, \varphi_m(u) \right) du \\ &= \sum_{q=0}^{j} \sum_{m=0}^{\ell} b_{j,q}\, \varphi_m(0)\, \delta_{qm} \;=\; \sum_{q=0}^{j} b_{j,q}\, \varphi_q(0) \;=\; u^j\big|_{u=0}, \end{aligned}$$

where the last equality evaluates the Legendre expansion of u^j at $u = 0$. Hence $\int u^j K(u) du = 1$ for $j = 0$ and $\int u^j K(u) du = 0$ for $j = 1, \ldots, \ell$, which are the two conditions of Definition 3. □

Remark 3 (Negative-value caveat)**.** Since $\varphi_m(0) = 0$ for odd m, the construction yields an even (symmetric) kernel; in fact it is nontrivial only for ℓ even. For $\ell \geq 2$, the constraint $\int u^\ell K(u) du = 0$ forces K to take negative values (as $u^\ell \geq 0$ a.e. for even ℓ). Hence the KDE $\widehat{f}_n$ may be negative. One can always project onto the nonnegative functions via $\widehat{f}_n^+ := \max(\widehat{f}_n, 0)$ without degrading

the sup-norm risk:

$$\|\widehat{f}_n - f\|_\infty \;\geq\; \|\widehat{f}_n^{+} - f\|_\infty \qquad \text{(almost surely)}.$$

1.5 Global L^2 risk (MISE) on the Sobolev class

Integrating the pointwise MSE over $\mathbb{R}$ gives the *mean integrated squared error*,

$$\mathrm{MISE} := \mathbb{E}\Big[\int_{\mathbb{R}} \big(\widehat{f}_n(x) - f(x)\big)^2\, dx\Big] = \int_{\mathbb{R}} \mathrm{Var}\big[\widehat{f}_n(x)\big]\, dx + \int_{\mathbb{R}} \big(\mathbb{E}\widehat{f}_n(x) - f(x)\big)^2\, dx,$$

where the bias–variance split uses Fubini–Tonelli.

Lemma 3 (Global variance control)**.** *If $\int_{\mathbb{R}} K^2(u)\, du < \infty$, then $\int_{\mathbb{R}} \mathrm{Var}[\widehat{f}_n(x)]\, dx \leq \frac{1}{nh} \int_{\mathbb{R}} K^2(u)\, du$.*

Proof. By Fubini–Tonelli,

$$\begin{aligned}
\int_{\mathbb{R}} \mathrm{Var}[\widehat{f}_n(x)]dx &\leq \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{1}{nh^2} K^2\big(\tfrac{u-x}{h}\big)\, f(u)\, du\, dx \\
&= \frac{1}{nh^2} \int_{\mathbb{R}} f(u) \left[\int_{\mathbb{R}} K^2\big(\tfrac{u-x}{h}\big)\, dx\right] du \\
&= \frac{1}{nh} \int_{\mathbb{R}} K^2(v)\, dv \cdot \int_{\mathbb{R}} f(u)\, du = \frac{1}{nh} \int K^2.
\end{aligned}$$

□

For global bias control, replace the Hölder class by the *Sobolev class*.

Definition 4 (Sobolev class)**.** For integer $\beta \geq 1$ and $L > 0$, define

$$\mathcal{S}(\beta, L) := \Big\{ f : \mathbb{R} \to \mathbb{R} \;\Big|\; f \text{ is } (\beta-1)\text{-times diff. with } f^{(\beta-1)} \text{ abs. continuous}, \int_{\mathbb{R}} (f^{(\beta)})^2\, dx \leq L^2 \Big\}.$$

Let $\mathcal{P}_{\mathcal{S}}(\beta, L) := \{ f \in \mathcal{S}(\beta, L) \,:\, f \geq 0,\ \int f = 1 \}$.

Lemma 4 (Global bias control; Tsybakov, 2009, Prop. 1.5)**.** *Let $f \in \mathcal{P}_{\mathcal{S}}(\beta, L)$ and let K be a kernel of order $\ell = \beta - 1$ with $\int |u|^\beta |K(u)|\, du < \infty$. Then*

$$\int_{\mathbb{R}} \big(\mathbb{E}\widehat{f}_n(x) - f(x)\big)^2\, dx \;\leq\; C\, h^{2\beta}$$

for a constant $C = C(L, \beta, K)$.

Proof sketch. Following the pointwise computation of Lemma 2, for any $x \in \mathbb{R}$,

$$\mathbb{E}\widehat{f}_n(x) - f(x) \;=\; \frac{h^\ell}{\ell!} \int_{\mathbb{R}} u^\ell K(u) \left[f^{(\ell)}(x + \tau h u) - f^{(\ell)}(x)\right] du \qquad \text{(some } \tau = \tau(x,u) \in (0,1)).$$

By the integral form of the remainder for a β-times weakly differentiable f,

$$f^{(\ell)}(x+hu) - f^{(\ell)}(x) \;=\; \int_0^{hu} f^{(\ell+1)}(x+s)\,ds \;=\; hu \int_0^1 f^{(\beta)}(x+thu)\,dt,$$

since $\ell + 1 = \beta$ (integer-β Sobolev case). Hence

$$\mathbb{E}\widehat{f}_n(x) - f(x) \;=\; \frac{h^\beta}{\ell!} \int K(u)\, u^{\ell+1} \int_0^1 f^{(\beta)}(x+thu)\,dt\,du.$$

Squaring and integrating over $x \in \mathbb{R}$, applying Cauchy–Schwarz to the inner $dt\,du$ integral, and then exchanging order with Fubini–Tonelli,

$$\begin{aligned}\int \big(\mathbb{E}\widehat{f}_n(x) - f(x)\big)^2 dx &\le \frac{h^{2\beta}}{(\ell!)^2} \int_0^1 \int |u|^{2\beta} K(u)^2\,du \cdot \int \big|f^{(\beta)}(x+thu)\big|^2\,dx\,dt \\ &= \frac{h^{2\beta}}{(\ell!)^2} \cdot \|K\|_{\beta,2}^2 \cdot \int \big(f^{(\beta)}\big)^2 \;\le\; C\,L^2\,h^{2\beta},\end{aligned}$$

using $\int (f^{(\beta)})^2 \le L^2$ from the Sobolev definition and translation invariance of the Lebesgue integral. The constant C depends only on β and the moments of K. □

Combining the two lemmas and optimizing over h,

Theorem 1 (Global risk of KDE on the Sobolev class; Tsybakov, 2009, Thm. 1.2)**.** *Under the assumptions of Lemmas 3 and 4,*

$$\sup_{f \in \mathcal{P}_{\mathcal{S}}(\beta,L)} \mathbb{E}\left[\int_{\mathbb{R}} \big(\widehat{f}_n(x) - f(x)\big)^2\,dx\right] \;\le\; \frac{C_1}{nh} + C_2 h^{2\beta}.$$

The minimizing bandwidth is $h_n^* \asymp n^{-1/(2\beta+1)}$, yielding MISE $\asymp n^{-2\beta/(2\beta+1)}$.

2 Nonparametric Least Squares (Fixed Design)

2.1 Setup and motivating examples

Consider the *fixed-design regression* model: we observe $(x_i, Y_i)_{i=1}^n$ with $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ deterministic and

$$Y_i = f^*(x_i) + \varepsilon_i, \qquad \varepsilon_1, \dots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{8}$$

where $f^* : \mathcal{X} \to \mathbb{R}$ is the unknown regression function. A natural estimator is the *constrained least-squares* (CLS) estimator

$$\widehat{f}_n \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \big(Y_i - f(x_i)\big)^2, \tag{9}$$

where $\mathcal{F}$ is a function class controlling approximation error and complexity. We assume throughout that $f^* \in \mathcal{F}$. Normalizing the noise variance to 1 costs no generality since rescaling Y_i and f^* by σ^{-1} preserves the form of (9).

Example 2 (OLS, ridge, Lasso)**.** Take $\mathcal{F} = \{x \mapsto \langle \theta, x \rangle : \theta \in \mathbb{R}^d\}$. Writing $X = (x_1, \dots, x_n) \in \mathbb{R}^{d \times n}$ and $Y = (Y_1, \dots, Y_n)^\top$, (9) becomes $\min_\theta \|X^\top \theta - Y\|_2^2$ with unconstrained solution $\widehat{\theta}_{\text{ols}} = (XX^\top)^{-1} XY$ (when the inverse exists). Restricting $\mathcal{F}$ to $\{\theta : \|\theta\|_2^2 \le t\}$ or $\{\theta : \|\theta\|_1 \le t\}$ gives ridge and Lasso, respectively. See Hastie et al. (2009, Ch. 3) for details.

Example 3 (Cubic smoothing spline)**.** Let $\mathcal{X} = [0,1]$ and $\mathcal{F} = \{f : [0,1] \to \mathbb{R} : \int_0^1 (f'')^2 \le R\}$. The penalized form is

$$\min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \int_0^1 (f'')^2 \, dx \right\}.$$

A representer-theorem argument shows the minimizer is a *natural cubic spline* with knots at the design points $\{x_i\}$: piecewise cubic, C^2-continuous with jumps in the third derivative at each knot, and linear outside $[\min_i x_i, \max_i x_i]$ (Hastie et al., 2009, Ch. 5). The infinite-dimensional optimization reduces to a finite-dimensional problem of ridge-regression form.

Example 4 (Convex regression)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be convex and $\mathcal{F}$ the set of all convex functions $\mathcal{C} \to \mathbb{R}$. Setting $\widetilde{y} = (f(x_1), \dots, f(x_n))$, convexity is equivalent to the existence of sub-gradients $z_1, \dots, z_n \in \mathbb{R}^d$ such that

$$\widetilde{y}_j \ge \widetilde{y}_i + \langle z_i, \, x_j - x_i \rangle \qquad \forall \, i \ne j.$$

Hence (9) becomes the finite-dimensional quadratic program

$$\min_{\widetilde{y},\, \{z_i\}} \ \|Y - \widetilde{y}\|_2^2 \quad \text{s.t.} \ \ \widetilde{y}_j \ge \widetilde{y}_i + \langle z_i, x_j - x_i \rangle, \ \forall i \ne j.$$

This has $(d+1)n$ variables and $n(n-1)$ linear constraints; see Seijo and Sen (2011).

The goal is to bound the empirical L^2-error

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

2.2 Basic inequality and localized Gaussian complexity

Assume $f^* \in \mathcal{F}$. By definition of $\widehat{f}_n$ as the minimizer of the empirical squared-error over $\mathcal{F}$, and since f^* is a feasible competitor,

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \widehat{f}_n(x_i)\big)^2 \;\le\; \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - f^*(x_i)\big)^2.$$

Substituting $Y_i - \widehat{f}_n(x_i) = \big[Y_i - f^*(x_i)\big] + \big[f^*(x_i) - \widehat{f}_n(x_i)\big] = \varepsilon_i - \big[\widehat{f}_n(x_i) - f^*(x_i)\big]$ on the left, expanding the square, and cancelling the common $n^{-1}\sum \varepsilon_i^2$ term on both sides,

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{\big(\widehat{f}_n(x_i) - f^*(x_i)\big)^2 - 2\,\varepsilon_i\,\big(\widehat{f}_n(x_i) - f^*(x_i)\big)\Big\} \;\le\; 0,$$

which rearranges to the *basic inequality*

$$\frac{1}{2}\big\|\widehat{f}_n - f^*\big\|_n^2 \;\le\; \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\,\big(\widehat{f}_n(x_i) - f^*(x_i)\big). \tag{10}$$

The RHS is a Gaussian random variable linear in the data-dependent increment $\widehat{f}_n - f^*$, which is the object the remainder of this section controls.

Definition 5 (Shifted function class)**.** For fixed f^*, the f^*-shifted class is $\mathcal{F}^* := \{f - f^* : f \in \mathcal{F}\}$. It contains 0 and $\widehat{\Delta} := \widehat{f}_n - f^*$.

Definition 6 (Star-shaped)**.** A function class $\mathcal{H}$ is *star-shaped* (around 0) if $g \in \mathcal{H}$ and $\alpha \in [0,1]$ imply $\alpha g \in \mathcal{H}$. Every convex class containing 0 is star-shaped.

Since $\widehat{\Delta} \in \mathcal{F}^*$ and $\|\widehat{\Delta}\|_n^2 = \|\widehat{f}_n - f^*\|_n^2$, (10) gives

$$\tfrac{1}{2}\|\widehat{\Delta}\|_n^2 \;\le\; \sup\left\{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\,g(x_i) \;:\; g \in \mathcal{F}^*,\; \|g\|_n \le \|\widehat{\Delta}\|_n\right\}.$$

Definition 7 (Localized Gaussian complexity)**.** For a class $\mathcal{H}$ and $\delta > 0$, let

$$G_n(\delta;\mathcal{H}) := \mathbb{E}\left[\sup_{g\in\mathcal{H},\,\|g\|_n\le\delta}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\,g(x_i)\right], \qquad \varepsilon_i \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

This is the *Gaussian complexity localized at scale* δ.

Taking $\delta = \|\widehat{\Delta}\|_n$ and expectations in (10) (and using Jensen's inequality on the LHS and the

localization on the RHS) yields

$$\tfrac{1}{2}\,\mathbb{E}\|\widehat{\Delta}\|_n^2 \;\le\; \mathbb{E}\left[\sup_{g\in\mathcal{F}^*,\,\|g\|_n\le\|\widehat{\Delta}\|_n} \frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)\right] \;\le\; G_n\big(\|\widehat{\Delta}\|_n;\mathcal{F}^*\big),$$

where the second inequality relaxes the data-dependent localization radius to a deterministic one and uses that $\mathcal{F}^*$ is star-shaped so that the map $\delta \mapsto G_n(\delta;\mathcal{F}^*)/\delta$ is non-increasing (Lemma 5). Consequently, any δ with $\delta^2 \ge \mathbb{E}\|\widehat{\Delta}\|_n^2$ satisfies $\delta/2 \le G_n(\delta;\mathcal{F}^*)/\delta$, which motivates the following definition.

2.3 Critical radius and the main oracle inequality

Definition 8 (Critical radius)**.** When $\mathcal{H}$ is star-shaped, the map $\delta \mapsto G_n(\delta;\mathcal{H})/\delta$ is non-increasing (Lemma 5 below). The *critical radius* $\delta_n > 0$ is the smallest positive solution to the *critical inequality*

$$\frac{G_n(\delta;\mathcal{H})}{\delta} \;\le\; \frac{\delta}{2}. \tag{11}$$

Equivalently, $\delta_n^2 \ge 2G_n(\delta_n;\mathcal{H})$.

Lemma 5 (Monotonicity)**.** *If $\mathcal{H}$ is star-shaped, then $\delta \mapsto G_n(\delta;\mathcal{H})/\delta$ is non-increasing on $(0,\infty)$.*

Proof. Let $0 < \delta \le t$. For any $h \in \mathcal{H}$ with $\|h\|_n \le t$, set $\widetilde{h} := (\delta/t)\,h$. Since $\delta/t \in (0,1]$ and $\mathcal{H}$ is star-shaped, $\widetilde{h} \in \mathcal{H}$; and clearly $\|\widetilde{h}\|_n = (\delta/t)\|h\|_n \le \delta$. The map $h \mapsto \widetilde{h}$ therefore sends the constraint set $\{h \in \mathcal{H} : \|h\|_n \le t\}$ into $\{\widetilde{h} \in \mathcal{H} : \|\widetilde{h}\|_n \le \delta\}$, and $n^{-1}\sum_i \varepsilon_i \widetilde{h}(x_i) = (\delta/t)\cdot n^{-1}\sum_i \varepsilon_i h(x_i)$. Hence

$$\frac{\delta}{t}\,G_n(t;\mathcal{H}) \;=\; \mathbb{E}\sup_{\substack{h\in\mathcal{H}\\ \|h\|_n\le t}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i\,\widetilde{h}(x_i) \;\le\; \mathbb{E}\sup_{\substack{\widetilde{h}\in\mathcal{H}\\ \|\widetilde{h}\|_n\le \delta}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i\,\widetilde{h}(x_i) \;=\; G_n(\delta;\mathcal{H}),$$

the inequality because the second sup is over a superset of $\{\widetilde{h} : h \in \mathcal{H}, \|h\|_n \le t\}$. Rearranging gives $G_n(\delta)/\delta \ge G_n(t)/t$, i.e., the map $\delta \mapsto G_n(\delta;\mathcal{H})/\delta$ is non-increasing. □

Theorem 2 (Main NPLS oracle inequality; Wainwright, 2019, Thm. 13.5)**.** *Assume $f^* \in \mathcal{F}$ and that the shifted class $\mathcal{F}^* = \mathcal{F} - \{f^*\}$ is star-shaped. Let δ_n be the critical radius of $\mathcal{F}^*$. Then, for any $t \ge \delta_n$,*

$$\Pr\Big[\big\|\widehat{f}_n - f^*\big\|_n^2 \ge 16\,t\,\delta_n\Big] \;\le\; \exp\!\left(-\frac{n\,t\,\delta_n}{2}\right). \tag{12}$$

Consequently

$$\mathbb{E}\Big[\big\|\widehat{f}_n - f^*\big\|_n^2\Big] \;\lesssim\; \delta_n^2 + \frac{1}{n}.$$

The proof (sketched in Section 2.8) proceeds by combining the basic inequality (10) with a peeling-and-concentration argument that bounds the Gaussian process $g \mapsto n^{-1}\sum_i \varepsilon_i g(x_i)$ on localized sub-balls. The rate δ_n^2 reflects the complexity of $\mathcal{F}$; it can be tiny (even parametric) when $\mathcal{F}$ is rich enough to contain f^* but sufficiently low-complexity elsewhere.

To apply the theorem we need to compute δ_n, which requires controlling $G_n(\delta; \mathcal{F}^*)$. The standard tool is *metric entropy*, which we develop next.

2.4 Covering and packing numbers

Let (T, ρ) be a (pseudo)-metric space.

Definition 9 (Covering / packing)**.** A subset $\{\theta_1, \dots, \theta_N\} \subseteq T$ is a *δ-cover* of T if for every $\theta \in T$ there exists $i \in [N]$ with $\rho(\theta, \theta_i) \le \delta$. The δ-covering number $N(\delta; T, \rho)$ is the cardinality of a smallest δ-cover. A subset $\{\theta_1, \dots, \theta_M\} \subseteq T$ is a *δ-packing* if $\rho(\theta_i, \theta_j) > \delta$ for all $i \neq j$; the δ-packing number $M(\delta; T, \rho)$ is the cardinality of a largest δ-packing. The quantity $\log N(\delta; T, \rho)$ is called the *metric entropy*.

Lemma 6 (Covering–packing duality)**.** *For every $\delta > 0$,*

$$M(2\delta; T, \rho) \;\le\; N(\delta; T, \rho) \;\le\; M(\delta; T, \rho).$$

Proof. Upper bound. If $\{\theta_i\}_{i=1}^M$ is a maximal δ-packing, it must also be a δ-cover: otherwise some $\theta \in T$ would have $\rho(\theta, \theta_i) > \delta$ for all i, contradicting maximality. *Lower bound.* A 2δ-packing $\{\theta_i\}_{i=1}^{M'}$ has the property that no single ball $B(\theta, \delta)$ contains two packing points (triangle inequality), so a δ-cover needs at least M' elements. □

Example 5 (Unit cube)**.** For $T = [-1, 1]^d$ with $\rho = \|\cdot\|_\infty$, one has $d \log(1/\delta) \le \log N(\delta; T, \rho) \le d \log(1 + 1/\delta)$, so $\log N(\delta; T, \rho) \asymp d \log(1/\delta)$ for small δ.

Lemma 7 (Volume bounds)**.** *Let $\|\cdot\|$ and $\|\cdot\|'$ be norms on $\mathbb{R}^d$ with unit balls B, B'. For any $\delta > 0$,*

$$\delta^{-d}\,\frac{\mathrm{vol}(B)}{\mathrm{vol}(B')} \;\le\; N(\delta; B, \|\cdot\|') \;\le\; \frac{\mathrm{vol}\big(\frac{2}{\delta}B + B'\big)}{\mathrm{vol}(B')}, \tag{13}$$

where $A + B = \{a + b : a \in A, b \in B\}$ is the Minkowski sum.

Proof. Lower bound. Let $\{\theta_i\}_{i=1}^N$ be a δ-cover of B in $\|\cdot\|'$. Then $B \subseteq \bigcup_{i=1}^N (\theta_i + \delta B')$, giving $\mathrm{vol}(B) \le N\,\delta^d \mathrm{vol}(B')$ and (13). *Upper bound.* Let $\{\theta_j\}_{j=1}^M$ be a maximal δ-packing of B in $\|\cdot\|'$. The balls $\theta_j + (\delta/2)B'$ are disjoint and contained in $B + (\delta/2)B'$; hence

$$M \cdot (\delta/2)^d \,\mathrm{vol}(B') \;\le\; \mathrm{vol}(B + (\delta/2)B') = (\delta/2)^d \,\mathrm{vol}\big(\tfrac{2}{\delta}B + B'\big)\,.$$

Apply Lemma 6. □

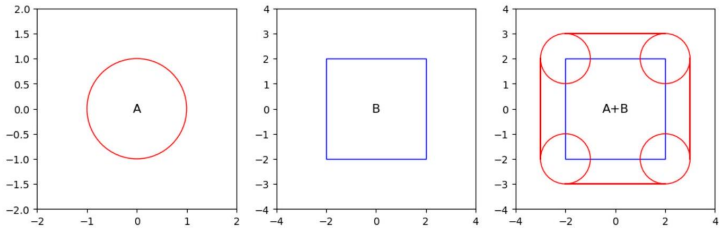

Figure 1: The Minkowski sum $A + B = \{a + b : a \in A,\, b \in B\}$ used in Lemma 7.

2.5 Metric entropy of Lipschitz and Hölder classes

Proposition 2 (Entropy of Lipschitz functions)**.** Let $\mathcal{F}_L := \{g : [0,1] \to \mathbb{R} \,:\, g(0) = 0,\ |g(x) - g(y)| \le L|x - y|\}$. Then

$$\log N(\delta; \mathcal{F}_L, \|\cdot\|_\infty) \asymp \frac{L}{\delta}. \tag{14}$$

Proof sketch. Lower bound. Fix $\varepsilon > 0$ and let $M = 1/\varepsilon$. Define knots $x_i = (i-1)\varepsilon$ for $i \in [M]$, let $\phi(u) = 0$ for $u < 0$, $\phi(u) = u$ for $u \in [0,1]$, $\phi(u) = 1$ for $u > 1$, and set

$$f_\beta(y) = \sum_{i=1}^{M} \beta_i \, \varepsilon L \, \phi\!\left(\frac{y - x_i}{\varepsilon}\right), \qquad \beta \in \{\pm 1\}^M.$$

Each f_β is L-Lipschitz and, for $\beta \neq \beta'$, differs on some interval of length ε by $\pm L$ on each side, so $\|f_\beta - f_{\beta'}\|_\infty \ge 2L\varepsilon$. Thus $\{f_\beta\}$ is a $(2L\varepsilon)$-packing of size 2^M, giving $\log N(L\varepsilon; \mathcal{F}_L, \|\cdot\|_\infty) \ge \log M(2L\varepsilon; \mathcal{F}_L, \|\cdot\|_\infty) \gtrsim 1/\varepsilon$. Setting $\delta = L\varepsilon$ yields $\log N(\delta) \gtrsim L/\delta$.

Upper bound. A classical piecewise-constant approximation on a grid of mesh δ/L gives the matching upper bound; see van der Vaart and Wellner (2023, Ex. 2.7.4). □

Proposition 3 (Entropy of Lipschitz functions in $\mathbb{R}^d$)**.** For $\mathcal{F}_L([0,1]^d) := \{g : [0,1]^d \to \mathbb{R} : g(0) = 0,\ |g(x) - g(y)| \le L\|x - y\|_\infty\}$,

$$\log N\big(\delta; \mathcal{F}_L([0,1]^d), \|\cdot\|_\infty\big) \asymp \left(\frac{L}{\delta}\right)^d,$$

which exhibits the curse of dimensionality.

Proposition 4 (Entropy of Hölder classes; van der Vaart and Wellner, 2023, Thm. 2.7.1)**.** Let $\mathcal{H}_\alpha([0,1]^d)$ be the class of functions $f : [0,1]^d \to \mathbb{R}$ with partial derivatives of all orders $|j| \le \lfloor \alpha \rfloor$

bounded by c_j and $\max_{|j|=\lfloor\alpha\rfloor} |D^j f(x) - D^j f(y)| \le L\|x-y\|_2^{\alpha-\lfloor\alpha\rfloor}$. Then

$$\log N(\delta; \mathcal{H}_\alpha([0,1]^d), \|\cdot\|_\infty) \;\lesssim\; \left(\frac{1}{\delta}\right)^{d/\alpha}.$$

For $d = 1$, $\log N(\delta; \mathcal{H}_\alpha, \|\cdot\|_\infty) \asymp (L/\delta)^{1/\alpha}$.

2.6 Sub-Gaussian processes, discretization, and chaining

Definition 10 (Sub-Gaussian process)**.** A collection $\{X_\theta\}_{\theta\in T}$ of mean-zero random variables is a *sub-Gaussian process* with respect to the semi-metric ρ on T if

$$\mathbb{E}\exp\big(\lambda(X_\theta - X_{\theta'})\big) \;\le\; \exp\Big(\tfrac{\lambda^2\rho(\theta,\theta')^2}{2}\Big) \qquad \forall\, \theta,\theta' \in T,\ \lambda\in\mathbb{R}. \tag{15}$$

A standard Chernoff argument gives $\Pr(|X_\theta - X_{\theta'}| \ge \epsilon) \le 2\exp(-\epsilon^2/(2\rho(\theta,\theta')^2))$.

Lemma 8 (Maxima of sub-Gaussian random variables)**.** *Let $X_1,\dots,X_N$ be (not necessarily independent) sub-Gaussian with parameter σ^2: $\mathbb{E}e^{\lambda X_i} \le e^{\lambda^2\sigma^2/2}$. Then*

$$\mathbb{E}\max_{i\in[N]} X_i \;\le\; \sqrt{2\sigma^2\log N}, \qquad \mathbb{E}\max_{i\in[N]} |X_i| \;\le\; \sqrt{2\sigma^2\log(2N)}.$$

Proof. By Jensen and the MGF bound, $e^{\lambda\mathbb{E}\max_i X_i} \le \mathbb{E}e^{\lambda\max_i X_i} \le Ne^{\lambda^2\sigma^2/2}$. Taking logs and optimizing over $\lambda > 0$ gives $\mathbb{E}\max_i X_i \le \lambda^{-1}\log N + \lambda\sigma^2/2$, minimized at $\lambda^* = \sqrt{2\log N/\sigma^2}$. □

Theorem 3 (One-step discretization bound)**.** *Let $\{X_\theta\}_{\theta\in T}$ be a mean-zero sub-Gaussian process on (T,ρ) with diameter $D := \sup_{\theta,\theta'}\rho(\theta,\theta') < \infty$. For every $\delta \in [0, D]$,*

$$\mathbb{E}\sup_{\theta,\theta'\in T}\big(X_\theta - X_{\theta'}\big) \;\le\; 2\,\mathbb{E}\sup_{\substack{\gamma,\gamma'\in T\\ \rho(\gamma,\gamma')\le\delta}}\big(X_\gamma - X_{\gamma'}\big) + 2\sqrt{2D^2\log N(\delta;T,\rho)}. \tag{16}$$

Consequently, fixing any $\theta_0 \in T$, $\mathbb{E}\sup_{\theta\in T} X_\theta \;=\; \mathbb{E}\sup_\theta(X_\theta - X_{\theta_0}) \le \mathbb{E}\sup_{\theta,\theta'}(X_\theta - X_{\theta'})$, so the same bound applies to $\mathbb{E}\sup_\theta X_\theta$.

Proof. Let $\{\theta_1,\dots,\theta_N\}$ be a minimal δ-cover of T under ρ; $N = N(\delta;T,\rho)$. For every $\theta\in T$ choose $\pi(\theta)\in\{\theta_1,\dots,\theta_N\}$ with $\rho(\theta,\pi(\theta)) \le \delta$. For any $\theta,\theta'\in T$, decompose

$$X_\theta - X_{\theta'} \;=\; \underbrace{(X_\theta - X_{\pi(\theta)})}_{\rho\le\delta} + \underbrace{(X_{\pi(\theta)} - X_{\pi(\theta')})}_{\text{in } N\times N \text{ grid}} + \underbrace{(X_{\pi(\theta')} - X_{\theta'})}_{\rho\le\delta}.$$

Taking the sup over $(\theta, \theta') \in T \times T$ and then expectations,

$$\mathbb{E} \sup_{\theta,\theta'} (X_\theta - X_{\theta'}) \;\le\; 2\,\mathbb{E} \sup_{\rho(\gamma,\gamma')\le\delta} (X_\gamma - X_{\gamma'}) + \mathbb{E} \max_{i,j\in[N]} (X_{\theta_i} - X_{\theta_j}). \tag{17}$$

The last term is a maximum over N^2 sub-Gaussian random variables with variance parameter at most $\rho(\theta_i, \theta_j)^2 \le D^2$; Lemma 8 gives

$$\mathbb{E} \max_{i,j\in[N]} (X_{\theta_i} - X_{\theta_j}) \;\le\; \sqrt{2\,D^2\,\log(N^2)} \;=\; 2\sqrt{D^2 \log N} \;\le\; 2\sqrt{2D^2 \log N(\delta; T, \rho)}.$$

Plugging back into (17) yields (16). The addendum on $\mathbb{E}\sup_\theta X_\theta$ follows from $0 = X_{\theta_0} - X_{\theta_0}$: the single-sup is bounded by the pair-sup shifted by X_{θ_0}. □

The one-step bound is sub-optimal: it pays the *full* diameter D of T per cover step. Chaining uses a sequence of progressively finer covers.

Theorem 4 (Dudley's entropy integral; Wainwright, 2019, Thm. 5.22)**.** *Under the hypotheses of Theorem 3, for every $\delta \in [0, D]$,*

$$\mathbb{E} \sup_{\theta,\theta'\in T} (X_\theta - X_{\theta'}) \;\lesssim\; \mathbb{E} \sup_{\rho(\gamma,\gamma')\le\delta} (X_\gamma - X_{\gamma'}) + \int_{\delta/4}^{D} \sqrt{\log N(u; T, \rho)}\, du. \tag{18}$$

Taking $\delta \to 0$ (assuming the first term vanishes, e.g., for separable processes with continuous trajectories) yields the classical bound

$$\mathbb{E} \sup_{\theta\in T} X_\theta \;\lesssim\; \int_0^D \sqrt{\log N(u; T, \rho)}\, du. \tag{19}$$

Proof sketch. Set $\varepsilon_m := D/2^m$ for $m = 1, \dots, L$, where L is chosen so that $\varepsilon_L < \delta \le \varepsilon_{L-1}$. For each m, let $\mathcal{N}_m$ be a minimal ε_m-cover of T; in particular $|\mathcal{N}_1| \le N(D/2; T, \rho)$ and $|\mathcal{N}_m| \le N(\varepsilon_m; T, \rho)$. For $\theta \in T$, build the *chain* $\gamma_L = \theta$ and $\gamma_{m-1} = \pi_{m-1}(\gamma_m) := \arg\min_{\beta\in\mathcal{N}_{m-1}} \rho(\gamma_m, \beta)$. By telescoping,

$$X_\theta - X_{\gamma_1} = \sum_{m=2}^{L} (X_{\gamma_m} - X_{\gamma_{m-1}}) \le \sum_{m=2}^{L} \max_{\beta\in\mathcal{N}_m} |X_\beta - X_{\pi_{m-1}(\beta)}|.$$

Each increment has sub-Gaussian parameter at most $\varepsilon_{m-1} = D/2^{m-1}$, and there are at most $|\mathcal{N}_m| \cdot |\mathcal{N}_{m-1}| \le N(\varepsilon_m)^2$ pairs, so Lemma 8 gives

$$\mathbb{E} \max_{\beta\in\mathcal{N}_m} |X_\beta - X_{\pi_{m-1}(\beta)}| \le 2\,\varepsilon_{m-1} \sqrt{\log N(\varepsilon_m; T, \rho)}.$$

Summing over m and comparing with the Riemann integral yields (18). □

2.7 Applying chaining to NPLS: rates for Lipschitz and convex regression

Applying Dudley's bound to the Gaussian process $g \mapsto n^{-1/2}\sum_i \varepsilon_i g(x_i)$ indexed by the local ball $\mathbb{B}_n(\delta;\mathcal{F}^*) := \{g \in \mathcal{F}^* : \|g\|_n \le \delta\}$, we obtain the following sufficient condition for the critical radius.

Lemma 9 (Entropy-integral control of the critical radius; Wainwright, 2019, Prop. 14.1)**.** *If δ satisfies*

$$\frac{16}{\sqrt{n}}\int_{\delta^2/4}^{\delta}\sqrt{\log N(u;\mathbb{B}_n(\delta;\mathcal{F}^*),\|\cdot\|_n)}\,du \;\le\; \frac{\delta^2}{4}, \tag{20}$$

then δ satisfies the critical inequality (11), and therefore $\delta_n \le \delta$.

Proof sketch. Take a minimal $(\delta^2/4)$-cover $\{g_1,\dots,g_m\}$ of $\mathbb{B}_n(\delta;\mathcal{F}^*)$ in the $\|\cdot\|_n$ norm. For any $g \in \mathbb{B}_n(\delta;\mathcal{F}^*)$ there is an index j with $\|g-g_j\|_n \le \delta^2/4$. Then

$$\left|\tfrac{1}{n}\sum_i \varepsilon_i g(x_i)\right| \le \left|\tfrac{1}{n}\sum_i \varepsilon_i g_j(x_i)\right| + \tfrac{1}{n}\|\varepsilon\|_2\,\|g-g_j\|_n,$$

and taking expectation $\mathbb{E}\frac{1}{n}\|\varepsilon\|_2^2 \le 1$, the second term is at most $\delta^2/4$. The first term is bounded by the chaining argument applied to the Gaussian process $Z_n(g_j) := n^{-1/2}\sum_i \varepsilon_i g_j(x_i)$; its sub-Gaussian metric is $\|g-g'\|_n$. Dudley's theorem applied to the finite set $\{g_1,\dots,g_m\}$ gives

$$\mathbb{E}\max_{j\in[m]}\tfrac{1}{\sqrt{n}}|Z_n(g_j)| \le \frac{16}{\sqrt{n}}\int_{\delta^2/4}^{\delta}\sqrt{\log N(u;\mathbb{B}_n(\delta;\mathcal{F}^*),\|\cdot\|_n)}\,du.$$

Combining the two terms gives $G_n(\delta;\mathcal{F}^*) \le \delta^2/4 + \delta^2/4 = \delta^2/2$, i.e. the critical inequality. □

We can now read off rates.

Corollary 5 (Lipschitz regression, rate $n^{-2/3}$)**.** *Let $\mathcal{F}_L$ be the Lipschitz class of Proposition 2 and assume $f^* \in \mathcal{F}_L$. Then $\mathcal{F}_L^* \subseteq \mathcal{F}_{2L}$, and (14) gives*

$$\frac{1}{\sqrt{n}}\int_{\delta^2/4}^{\delta}\sqrt{\log N(u;\mathbb{B}_n(\delta;\mathcal{F}_{2L}),\|\cdot\|_n)}\,du \;\lesssim\; \sqrt{\frac{L\,\delta}{n}}.$$

Solving $\sqrt{L\delta_n/n} \lesssim \delta_n^2$ yields $\delta_n \asymp (L/n)^{1/3}$, so by Theorem 2,

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

Corollary 6 (Convex Lipschitz regression, rate $n^{-4/5}$). *Let $\mathcal{F}_{LC} := \mathcal{F}_L \cap \{f \text{ convex}\}$. Its metric entropy satisfies $\log N(u; \mathcal{F}_{LC}, \|\cdot\|_\infty) \lesssim (L/u)^{1/2}$ (Seijo and Sen, 2011, §4). A similar computation gives*

$$\frac{1}{\sqrt{n}} \int_{\delta^2/4}^{\delta} \sqrt{\log N(u; \mathcal{F}_{LC}, \|\cdot\|_\infty)}\, du \;\lesssim\; \frac{L^{1/4}}{\sqrt{n}}\, \delta^{3/4},$$

so $\delta_n \asymp L^{1/5}/n^{2/5}$ and

$$\mathbb{E}\|\widehat{f}_n - f^*\|_n^2 \;\lesssim\; \frac{L^{2/5}}{n^{4/5}}.$$

2.8 Sketch of the proof of Theorem 2

Write $\widehat{\Delta} := \widehat{f}_n - f^*$ and recall from (10) that $\frac{1}{2}\|\widehat{\Delta}\|_n^2 \le n^{-1}\sum_i \varepsilon_i \widehat{\Delta}(x_i)$. The key technical device is the following *peeling lemma*, which converts the critical inequality into a uniform tail bound on the Gaussian process $g \mapsto n^{-1}\sum_i \varepsilon_i g(x_i)$ over sub-balls of $\mathcal{F}^*$.

Lemma 10 (Peeling lemma; Wainwright, 2019, Lem. 13.22). *Let $\mathcal{H}$ be star-shaped and $\delta_n > 0$ satisfy the critical inequality. Define*

$$\mathcal{A}(u) := \Big\{ g \in \mathcal{H} :\; \|g\|_n \ge u,\;\; \Big|\tfrac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)\Big| \ge 2\|g\|_n\, u \Big\}.$$

Then for every $u \ge \delta_n$,

$$\Pr\big[\mathcal{A}(u)\big] \;\le\; e^{-nu^2/2}.$$

Apply Lemma 10 with $\mathcal{H} = \mathcal{F}^*$ and $u = \sqrt{t\delta_n}$ for $t \ge \delta_n$: with probability at least $1 - e^{-nt\delta_n/2}$, the event $\mathcal{A}^c(\sqrt{t\delta_n})$ holds. On this event, if $\|\widehat{\Delta}\|_n \ge \sqrt{t\delta_n}$, then $\widehat{\Delta} \in \mathcal{F}^*$ satisfies

$$\big|\tfrac{1}{n}\sum_i \varepsilon_i \widehat{\Delta}(x_i)\big| \le 2\|\widehat{\Delta}\|_n \sqrt{t\delta_n}.$$

Combined with (10): $\frac{1}{2}\|\widehat{\Delta}\|_n^2 \le 2\|\widehat{\Delta}\|_n\sqrt{t\delta_n}$, so $\|\widehat{\Delta}\|_n \le 4\sqrt{t\delta_n}$, i.e. $\|\widehat{\Delta}\|_n^2 \le 16\, t\, \delta_n$. Hence

$$\Pr\Big[\|\widehat{\Delta}\|_n^2 \ge 16\, t\, \delta_n\Big] \le \Pr[\mathcal{A}(\sqrt{t\delta_n})] \le e^{-nt\delta_n/2},$$

which is (12). The in-expectation bound follows by integrating the tail: $\mathbb{E}\|\widehat{\Delta}\|_n^2 = \int_0^\infty \Pr[\|\widehat{\Delta}\|_n^2 > s]\, ds \lesssim \delta_n^2 + 1/n$.

Part II

Minimax Lower Bounds

1 Reduction from Estimation to Testing

1.1 The minimax framework

Let Θ be a parameter space (the "hypothesis class"; often infinite-dimensional, e.g., Hölder or Sobolev classes) and $\{P_\theta : \theta \in \Theta\}$ an associated family of probability measures on $\mathcal{X}$. Each θ defines a distribution of the data (densities, regression functions, etc.). An *estimator* $\widehat{\theta}_n = \widehat{\theta}_n(X_1, \dots, X_n)$ is any measurable function of the sample.

Fix a semi-metric d on Θ (the *loss*): a symmetric function with $d(\theta, \theta) = 0$ satisfying the triangle inequality (but not necessarily positive definite). The *risk* of $\widehat{\theta}_n$ at θ is $\mathbb{E}_\theta[d^2(\widehat{\theta}_n, \theta)]$ and the *worst-case risk* is

$$R(\widehat{\theta}_n) \;:=\; \sup_{\theta \in \Theta} \mathbb{E}_\theta \Big[d^2(\widehat{\theta}_n, \theta) \Big] .$$

The *minimax risk* is

$$R^* \;:=\; \inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \Big[d^2(\widehat{\theta}_n, \theta) \Big] ,$$

where the infimum ranges over all measurable estimators. A rate $\Phi_n \to 0$ is a *minimax upper bound* if $\limsup R^*/\Phi_n \le C$, and a *lower bound* if $\liminf R^*/\Phi_n \ge c > 0$. When the two rates match, the estimator is *minimax rate-optimal*. Obtaining matching lower bounds is the subject of this Part.

1.2 The three-step reduction to testing

Fix a target rate Φ_n. Lower bounds are obtained by reducing risk control to a testing problem.

Step 1 (Markov's inequality). For any $\Phi_n > 0$,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\frac{d(\widehat{\theta}_n, \theta)}{\Phi_n} \right] \;\ge\; \sup_{\theta \in \Theta} \mathbb{E}_\theta \Big[\mathbb{1}\{ d(\widehat{\theta}_n, \theta) \ge \Phi_n \} \Big] = \sup_{\theta \in \Theta} P_\theta \Big(d(\widehat{\theta}_n, \theta) \ge \Phi_n \Big) . \tag{21}$$

Step 2 (restriction to a finite subset). For any finite $\{\theta_1, \dots, \theta_M\} \subseteq \Theta$,

$$\sup_{\theta \in \Theta} P_\theta(d(\widehat{\theta}_n, \theta) \ge \Phi_n) \;\ge\; \max_{j \in [M]} P_{\theta_j}\big(d(\widehat{\theta}_n, \theta_j) \ge \Phi_n\big). \tag{22}$$

Step 3 (minimum-distance test + triangle inequality). A *test function* is a measurable map $\phi : \mathcal{X}^n \to [M]$. Define the *minimum-distance test*

$$\phi^* := \arg\min_{k \in [M]} d(\widehat{\theta}_n, \theta_k).$$

If $\{\theta_1, \ldots, \theta_M\}$ form a $2\Phi_n$-*packing* of Θ (i.e., $d(\theta_j, \theta_k) \geq 2\Phi_n$ for all $j \neq k$), then the event $\{d(\widehat{\theta}_n, \theta_j) \leq \Phi_n\}$ implies $\phi^* = j$ by the reverse triangle inequality:

$$d(\widehat{\theta}_n, \theta_k) \geq d(\theta_j, \theta_k) - d(\widehat{\theta}_n, \theta_j) \geq 2\Phi_n - \Phi_n = \Phi_n \geq d(\widehat{\theta}_n, \theta_j) \quad \forall k \neq j.$$

Hence $P_{\theta_j}(d(\widehat{\theta}_n, \theta_j) \geq \Phi_n) \geq P_{\theta_j}(\phi^* \neq j)$, so

$$\max_{j \in [M]} P_{\theta_j}(d(\widehat{\theta}_n, \theta_j) \geq \Phi_n) \;\geq\; \inf_{\phi} \max_{j \in [M]} P_{\theta_j}(\phi \neq j), \tag{23}$$

with the inf over all tests $\phi : \mathcal{X}^n \to [M]$.

Putting it together. Combining (21)–(23):

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\frac{d(\widehat{\theta}_n, \theta)}{\Phi_n} \right] \;\geq\; \inf_{\phi} \max_{j \in [M]} P_{\theta_j}(\phi \neq j), \tag{24}$$

and, since $d^2 \geq 0$ is monotone,

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[d^2(\widehat{\theta}_n, \theta) \right] \geq \Phi_n^2 \inf_{\phi} \max_{j \in [M]} P_{\theta_j}(\phi \neq j).$$

The rate Φ_n is proved to be a minimax lower bound once $\inf_\phi \max_j P_{\theta_j}(\phi \neq j) \geq c > 0$ for a constant c independent of n.

Often we upper-bound the max by the average,

$$\inf_{\phi} \max_{j \in [M]} P_{\theta_j}(\phi \neq j) \;\geq\; \inf_{\phi} \tfrac{1}{M} \sum_{j=1}^{M} P_{\theta_j}(\phi \neq j), \tag{25}$$

since Fano's and related information-theoretic inequalities are stated in terms of *averaged* error. The averaged error has a clean probabilistic interpretation: sample $J \sim \mathrm{Unif}([M])$ and, given $J = j$, draw $X \sim P_{\theta_j}$. Let Q denote the joint law of (X, J). Then

$$\inf_{\phi} \tfrac{1}{M} \sum_{j} P_{\theta_j}(\phi \neq j) = \inf_{\phi} Q(\phi(X) \neq J).$$

The marginal law of X under Q is the *mixture* $\bar{Q} = \frac{1}{M}\sum_{j=1}^{M} P_{\theta_j}$.

This reduction is the common backbone of the methods in the remainder of Part II: Le Cam ($M = 2$), Fano (M large), and Yang–Barron (also M large, but using a discretization of the distribution space rather than the parameter space).

2 Le Cam's Two-Point Method

2.1 f-divergences and key inequalities

Definition 11 (f-divergence)**.** Let $f : (0,\infty) \to \mathbb{R}$ be convex with $f(1) = 0$. For probability measures P, Q with densities p, q w.r.t. a σ-finite measure ν,

$$D_f(P \,\|\, Q) := \int q(x)\, f\!\left(\frac{p(x)}{q(x)}\right) d\nu(x) \qquad (= +\infty \text{ if } P \not\ll Q).$$

Four classical cases:

(i) Total variation ($f(x) = |x-1|/2$): $\|P - Q\|_{\mathrm{TV}} = \frac{1}{2}\int |p - q|\, d\nu = \sup_A |P(A) - Q(A)|$.

(ii) Hellinger squared ($f(x) = (\sqrt{x}-1)^2$): $H^2(P,Q) = \int (\sqrt{p} - \sqrt{q})^2\, d\nu$.

(iii) Kullback–Leibler ($f(x) = x\log x$): $\mathrm{KL}(P \,\|\, Q) = \int p \log(p/q)\, d\nu$.

(iv) Chi-squared ($f(x) = (x-1)^2$): $\chi^2(P \,\|\, Q) = \int \frac{(p-q)^2}{q}\, d\nu = \int \frac{p^2}{q}\, d\nu - 1$.

Lemma 11 (Classical divergence inequalities)**.** *For any probability measures P, Q:*

(i) Pinsker: $\|P - Q\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}\,\mathrm{KL}(P \,\|\, Q)}$.

(ii) Le Cam: $\frac{1}{2}H^2(P,Q) \le \|P - Q\|_{\mathrm{TV}} \le H(P,Q)\sqrt{1 - H^2(P,Q)/4}$.

(iii) χ^2 bound on TV: $\|P - Q\|_{\mathrm{TV}} \le \frac{1}{2}\sqrt{\chi^2(P \,\|\, Q)}$.

(iv) Ordering: $\|P - Q\|_{\mathrm{TV}} \le H(P,Q) \le \sqrt{\mathrm{KL}(P \,\|\, Q)} \le \sqrt{\chi^2(P \,\|\, Q)}$.

The proofs follow from Cauchy–Schwarz and $\log x \le x - 1$; see Tsybakov (2009, §2.4).

Lemma 12 (Tensorization)**.** *For i.i.d. samples and product measures $P^{\otimes n}, Q^{\otimes n}$,*

$$\begin{aligned} \mathrm{KL}(P^{\otimes n} \,\|\, Q^{\otimes n}) &= n\,\mathrm{KL}(P \,\|\, Q), \\ H^2(P^{\otimes n}, Q^{\otimes n}) &\le n\, H^2(P,Q), \\ \chi^2(P^{\otimes n} \,\|\, Q^{\otimes n}) &= \big(1 + \chi^2(P \,\|\, Q)\big)^n - 1. \end{aligned}$$

The TV distance does not *tensorize in a useful way.*

Proof for Hellinger. $1-\frac{1}{2}H^2(P^{\otimes n},Q^{\otimes n}) = \int\sqrt{dP^n\,dQ^n} = \big(\int\sqrt{dP\,dQ}\big)^n = (1-\frac{1}{2}H^2(P,Q))^n$. The inequality $1-(1-x)^n \le nx$ for $x\in[0,1]$ gives $H^2(P^n,Q^n)\le nH^2(P,Q)$. □

2.2 Binary testing and the TV lower bound

Specializing the three-step reduction to $M=2$ hypotheses θ_0,θ_1, the averaged error probability is

$$Q(\phi(X)\neq J) = \tfrac{1}{2}P_0(\phi(X)\neq 0) + \tfrac{1}{2}P_1(\phi(X)\neq 1),$$

where $P_j := P_{\theta_j}$. For any test ϕ, let $A := \{x : \phi(x)=1\}$. Then

$$Q(\phi = J) = \tfrac{1}{2}P_0(A^c) + \tfrac{1}{2}P_1(A) = \tfrac{1}{2} + \tfrac{1}{2}(P_1(A)-P_0(A)).$$

Taking the supremum over A gives the *Bayes risk*

$$\inf_{\phi} Q(\phi\neq J) = \tfrac{1}{2} - \tfrac{1}{2}\sup_A |P_1(A)-P_0(A)| = \tfrac{1}{2}\big(1-\|P_0-P_1\|_{\mathrm{TV}}\big).$$

Combining with the three-step reduction:

Theorem 7 (Le Cam's two-point lower bound)**.** *Let $\theta_0,\theta_1\in\Theta$ satisfy $d(\theta_0,\theta_1)\ge 2\Phi_n$. Then*

$$\inf_{\widehat{\theta}_n}\sup_{\theta\in\Theta}\mathbb{E}_\theta\Big[d(\widehat{\theta}_n,\theta)\Big] \;\ge\; \Phi_n\cdot\tfrac{1}{2}\big(1-\|P_{\theta_0}^n - P_{\theta_1}^n\|_{\mathrm{TV}}\big), \tag{26}$$

where $P_{\theta_j}^n = P_{\theta_j}^{\otimes n}$. In particular, if $\|P_{\theta_0}^n - P_{\theta_1}^n\|_{\mathrm{TV}}\le\alpha<1$, then the minimax lower bound scales like Φ_n^2 in squared loss.

In practice, TV is bounded via H^2 or χ^2 using Lemmas 11–12; a common workflow is: (i) choose two hypotheses separated by $2\Phi_n$ in parameter space; (ii) tensorize Hellinger/KL/χ^2 to get n-sample bounds; (iii) apply Lemma 11.

2.3 Example: Gaussian location (parametric benchmark)

Let $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta,\sigma^2)$ with $\theta\in\mathbb{R}$. The loss is $d(\theta,\theta') = |\theta-\theta'|$. Take $\theta_0=0$, $\theta_1 = 2\Phi_n$, so $d(\theta_0,\theta_1) = 2\Phi_n$.

For two univariate Gaussians $P=\mathcal{N}(\mu_0,\sigma^2)$ and $Q=\mathcal{N}(\mu_1,\sigma^2)$ with common variance, a direct computation gives

$$1+\chi^2(P\,\|\,Q) \;=\; \int\frac{p(x)^2}{q(x)}\,dx \;=\; \exp\!\left(\frac{(\mu_0-\mu_1)^2}{\sigma^2}\right). \tag{27}$$

Indeed, completing the square in $-2(x-\mu_0)^2+(x-\mu_1)^2$ shows that $p(x)^2/q(x)$ equals the density of $\mathcal{N}(2\mu_0-\mu_1,\,\sigma^2)$ multiplied by the constant $\exp\big((\mu_0-\mu_1)^2/\sigma^2\big)$, and the former integrates to 1. Combining with the tensorization identity $1+\chi^2(P^{\otimes n}\,\|\,Q^{\otimes n})=(1+\chi^2(P\,\|\,Q))^n$ from Lemma 12,

$$\chi^2(P_0^n\,\|\,P_1^n)\;=\;\exp\!\left(\frac{n(\theta_1-\theta_0)^2}{\sigma^2}\right)-1\;=\;\exp\!\left(\frac{4n\Phi_n^2}{\sigma^2}\right)-1.$$

Choosing $\Phi_n=\sigma/(2\sqrt{n})$ gives $\chi^2\le e-1$, hence $\|P_0^n-P_1^n\|_{\mathrm{TV}}\le\frac{1}{2}\sqrt{e-1}<1$. Theorem 7 yields $\inf_{\widehat\theta}\sup_\theta\mathbb{E}_\theta|\widehat\theta-\theta|\gtrsim\sigma/\sqrt{n}$, and by monotonicity of $x\mapsto x^2$, also $\gtrsim\sigma^2/n$ in squared loss. The sample mean $\bar X_n\sim\mathcal{N}(\theta,\sigma^2/n)$ achieves $\mathbb{E}_\theta(\bar X_n-\theta)^2=\sigma^2/n$, so the bound is sharp.

2.4 Modulus of continuity and pointwise Lipschitz density

Le Cam's two-point method is often phrased through the *modulus of continuity*. Let $\theta:\mathcal{F}\to\mathbb{R}$ be a real-valued functional on a class $\mathcal{F}$ of densities, with the Hellinger distance on $\mathcal{F}$. Define

$$\omega(\varepsilon;\theta,\mathcal{F}):=\sup\{|\theta(f)-\theta(g)|:H(f,g)\le\varepsilon\}.$$

Choosing $f,g\in\mathcal{F}$ with $H^2(f,g)\le 1/(4n)$ (so $H^2(P_f^n,P_g^n)\le 1/4$ and $\|P_f^n-P_g^n\|_{\mathrm{TV}}\le 1/2$ by Lemma 11(ii)) gives

$$\inf_{\widehat\theta_n}\sup_{f\in\mathcal{F}}\mathbb{E}\Big[|\widehat\theta_n-\theta(f)|^2\Big]\;\gtrsim\;\omega\Big(\tfrac{1}{\sqrt{4n}};\theta,\mathcal{F}\Big)^2.$$

Example 6 (Pointwise estimation of a Lipschitz density, rate $n^{-2/3}$)**.** Let $\mathcal{F}=\{f:[-1/2,1/2]\to\mathbb{R}_+:|f(x)-f(y)|\le|x-y|,\ \int f=1\}$ (Lipschitz densities on an interval). The target functional is $\theta(f)=f(0)$, with pointwise loss $d(f,g)=|f(0)-g(0)|$. Take $f\equiv 1$ and the two-lobe perturbation

$$g(x):=f(x)+\varphi(x),\qquad \varphi(x):=\begin{cases}\delta-|x|, & |x|\le\delta,\\ |x-2\delta|-\delta, & x\in(\delta,\,3\delta],\\ 0, & \text{otherwise},\end{cases}\tag{28}$$

for $\delta\in(0,\,1/6]$ (so $\mathrm{supp}(\varphi)\subseteq[-\delta,3\delta]\subseteq[-1/2,1/2]$). The function φ consists of a positive tent of height δ on $[-\delta,\delta]$ and a negative tent of depth δ on $[\delta,3\delta]$, each of base 2δ. Hence $\int\varphi=\delta^2-\delta^2=0$, so g integrates to 1; moreover $g\ge 1-\delta>0$ and g is 1-Lipschitz (the tents have slope ±1), so $g\in\mathcal{F}$.

A second-order Taylor expansion of $t\mapsto(\sqrt{1+t}-1)^2=t^2/4+O(t^3)$ around $t=0$ gives

$$H^2(f,g)=\int_{-1/2}^{1/2}\big(\sqrt{1}-\sqrt{1+\varphi(x)}\big)^2\,dx=\tfrac{1}{4}\int\varphi^2\,dx+O(\delta^4).$$

Since $\int \varphi^2 = 2 \cdot 2\int_0^\delta (\delta - u)^2\,du = \frac{4\delta^3}{3}$ (two tents, each with L^2-mass $2\delta^3/3$), we get $H^2(f,g) \le \frac{\delta^3}{3} + O(\delta^4) \lesssim \delta^3$. Choose $\delta = c\,n^{-1/3}$ with $c > 0$ small enough that $H^2(f,g) \le 1/(4n)$. Then the modulus lower bound $\omega(1/\sqrt{4n}) \ge |f(0) - g(0)| = |\varphi(0)| = \delta$ combined with Theorem 7 (via the Hellinger form of Section 2.4) gives

$$\inf_{\widehat{\theta}_n} \sup_{f \in \mathcal{F}} \mathbb{E}|\widehat{\theta}_n - f(0)|^2 \;\gtrsim\; \delta^2 \;\asymp\; n^{-2/3}.$$

The matching upper bound is obtained by a KDE with order-1 kernel (Lemmas 1–2 with $\beta = 1$). For the Hölder extension $\mathcal{H}(\beta, L)$, replacing φ by a scaled bump of height $L\delta^\beta$ yields $H^2 \lesssim \delta^{2\beta+1}$ and the rate $n^{-2\beta/(2\beta+1)}$; see Tsybakov (2009, §2.5).

2.5 Minimax lower bound for Hölder nonparametric regression

Consider fixed-design regression $Y_i = f(x_i) + \varepsilon_i$ with $x_i \in [0,1]$, $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, and $f \in \mathcal{H}(\beta, L)$ (Hölder class). The pointwise loss is $d(f,g) = |f(x_0) - g(x_0)|$.

Assume a mild *design regularity* condition: there exists $a > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i \in A\} \;\le\; a \cdot \max\big(\mathrm{Leb}(A),\, 1/n\big) \qquad \text{for every interval } A \subseteq [0,1], \tag{29}$$

i.e., the empirical design measure is dominated by Lebesgue measure plus a $1/n$ bulk term. Let

$$K_0(u) := \exp\!\left(-\frac{1}{1-u^2}\right) \mathbb{1}\{|u| < 1\}, \qquad K(u) := a_* \cdot K_0(2u),$$

where $a_* > 0$ is chosen so that $K \in \mathcal{H}(\beta,\, 1/2) \cap C^\infty$ and K is supported in $(-1/2,\, 1/2)$.

Choose

$$f_0 \equiv 0, \qquad f_1(x) = L\,h^\beta\,K\!\left(\frac{x - x_0}{h}\right), \qquad h = C_0\,n^{-1/(2\beta+1)},$$

with $C_0 > 0$ small.

(i) **Both hypotheses lie in the class.** $f_0 \in \mathcal{H}(\beta, L)$ trivially. Writing $u = (x - x_0)/h$ and $u' = (x' - x_0)/h$, one computes $f_1^{(\ell)}(x) = L\,h^{\beta-\ell}\,K^{(\ell)}(u)$, so

$$\big|f_1^{(\ell)}(x) - f_1^{(\ell)}(x')\big| \le Lh^{\beta-\ell} \cdot \tfrac{1}{2}\,|u - u'|^{\beta-\ell} = \tfrac{L}{2}|x - x'|^{\beta-\ell} \le L|x - x'|^{\beta-\ell},$$

using $K \in \mathcal{H}(\beta,\, 1/2)$ for the first inequality.

(ii) **Separation.** $d(f_0, f_1) = |f_1(x_0)| = Lh^\beta K(0) = LC_0^\beta\,K(0)\,n^{-\beta/(2\beta+1)} =: 2\Phi_n$.

(iii) **KL control.** Under f_1, $Y_i \sim \mathcal{N}(f_1(x_i), \sigma^2)$ (independently), so

$$\mathrm{KL}(P_0^n \,\|\, P_1^n) = \sum_{i=1}^n \frac{f_1^2(x_i)}{2\sigma^2} \le \frac{L^2 h^{2\beta} K_{\max}^2}{2\sigma^2} \cdot \sum_{i=1}^n \mathbb{1}\{\left|\tfrac{x_i - x_0}{h}\right| \le \tfrac{1}{2}\} \le \frac{L^2 K_{\max}^2}{2\sigma^2}\, n\, a\, h^{2\beta+1}.$$

Since $h^{2\beta+1} \asymp n^{-1}$, this equals $\frac{L^2 K_{\max}^2 a}{2\sigma^2} C_0^{2\beta+1} =: \alpha$, a constant. Choosing C_0 small enough makes $\alpha < \log 2$; Pinsker's inequality then gives $\|P_0^n - P_1^n\|_{\mathrm{TV}} \le \sqrt{\alpha/2} < 1/2$.

Theorem 7 yields

$$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{H}(\beta, L)} \mathbb{E}_f\Big[|\widehat{f}_n(x_0) - f(x_0)|^2\Big] \gtrsim n^{-\frac{2\beta}{2\beta+1}}. \tag{30}$$

This matches the kernel upper bound from Section 1.3: the KDE with bandwidth $h_n^* \asymp n^{-1/(2\beta+1)}$ and kernel of order $\lfloor \beta \rfloor$ is *minimax rate-optimal* for pointwise estimation on the Hölder class.

3 Fano's Method

The two-point method suffices when the rate is dominated by a "small" separation in the distribution space. When the rate is determined by the *complexity* of Θ—e.g., global L^2 risk on the Sobolev class—we need many hypotheses. This is handled by Fano's inequality.

3.1 Mutual information

With the mixture setup from Section 1.2: $J \sim \mathrm{Unif}([M])$, $X | J = j \sim P_{\theta_j}^n$. The marginal of X is $\bar{Q} = M^{-1} \sum_j P_{\theta_j}^n$.

Definition 12 (Mutual information)**.** $I(X; J) := \mathrm{KL}(Q_{X,J} \,\|\, Q_X \otimes Q_J) = \frac{1}{M} \sum_{j=1}^M \mathrm{KL}(P_{\theta_j}^n \,\|\, \bar{Q})$.

Note $I(X; J) \ge 0$, with equality iff $X \perp J$. The key convexity identity (see Lemma 13) relates $\bar{Q}$ to the KL-barycenter of $\{P_{\theta_j}^n\}$.

Lemma 13 (KL-barycenter)**.** $\bar{Q} = M^{-1} \sum_{j=1}^M P_{\theta_j}^n = \arg\min_Q \sum_{j=1}^M \mathrm{KL}(P_{\theta_j}^n \,\|\, Q)$. *In particular,* $I(X; J) \le \frac{1}{M} \sum_j \mathrm{KL}(P_{\theta_j}^n \,\|\, Q)$ *for any distribution* Q.

Proof. Using $\mathrm{KL}(P \,\|\, Q) = \mathbb{E}_P \log dP - \mathbb{E}_P \log dQ$ and dropping the term not depending on Q, the minimization reduces to $\max_Q \frac{1}{M} \sum_j \int \log dQ \, dP_{\theta_j}^n = \max_Q \int \log(dQ)\, d\bar{Q}$, which by Gibbs' inequality is maximized at $Q = \bar{Q}$. □

A Jensen-type corollary that is easier to compute in practice:

$$I(X; J) \;\le\; \tfrac{1}{M^2} \sum_{j,k=1}^M \mathrm{KL}(P_{\theta_j}^n \,\|\, P_{\theta_k}^n) \;\le\; \max_{j \ne k} \mathrm{KL}(P_{\theta_j}^n \,\|\, P_{\theta_k}^n).$$

3.2 Fano's inequality

Theorem 8 (Fano's inequality). *Let $\{\theta_1, \dots, \theta_M\}$ be a $2\Phi_n$-packing of (Θ, d). Then*

$$\inf_{\widehat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta \Big[d^2(\widehat{\theta}_n, \theta) \Big] \;\geq\; \Phi_n^2 \left[1 - \frac{I(X;J) + \log 2}{\log M} \right].$$

Proof sketch. Apply (24) and (25) to reduce to the averaged error $\inf_\phi Q(\phi(X) \neq J)$. The classical Fano inequality (see, e.g., Cover and Thomas, 2006, Thm. 2.10.1) states

$$H(J \mid X) \;\leq\; H_b\big(Q(\phi(X) \neq J)\big) \;+\; Q(\phi(X) \neq J) \cdot \log(M-1),$$

where $H_b(p) := -p \log p - (1-p)\log(1-p) \leq \log 2$. Since $J \sim \mathrm{Unif}([M])$ has entropy $H(J) = \log M$ and mutual information satisfies $I(X;J) = H(J) - H(J \mid X)$, we get $H(J \mid X) = \log M - I(X;J)$. Combined with Fano's inequality, using $H_b \leq \log 2$ and $\log(M-1) \leq \log M$,

$$\log M - I(X;J) \;\leq\; \log 2 + Q(\phi \neq J)\, \log M,$$

and rearranging gives $Q(\phi \neq J) \geq 1 - (I(X;J) + \log 2)/\log M$. Plugging into (24) completes the proof. □

To use Fano's inequality one needs a packing $\{\theta_j\}$ that is (a) well-separated in d and (b) has small KL divergence.

3.3 Varshamov–Gilbert packing on the hypercube

Lemma 14 (Varshamov–Gilbert). *Let $H^m := \{\pm 1\}^m$ and $d_H(\alpha, \beta) := m^{-1} \sum_{j=1}^m \mathbb{1}\{\alpha_j \neq \beta_j\}$ be the rescaled Hamming distance. There exists a subset $\Omega \subseteq H^m$ with*

$$|\Omega| \geq \exp(m/8) \quad \textit{and} \quad d_H(\alpha, \beta) \geq \tfrac{1}{4} \quad \textit{for all } \alpha \neq \beta \in \Omega.$$

Proof. Let $s = \lfloor m/4 \rfloor$. The number of points within Hamming distance s of a given α is $\sum_{j=0}^s \binom{m}{j}$. If $\{\alpha_1, \dots, \alpha_N\} \subseteq H^m$ is a maximal $(1/4)$-covering of H^m under d_H, the balls of Hamming radius s around the α_i cover H^m, so $N \cdot \sum_{j=0}^s \binom{m}{j} \geq 2^m$. Letting $\xi_1, \dots, \xi_m$ be i.i.d. Bernoulli(1/2),

$$2^{-m} \sum_{j=0}^{s} \binom{m}{j} = \Pr\left(\sum_{i=1}^m \xi_i \leq m/4 \right) \leq e^{-m/8}$$

by Hoeffding's inequality (since $\mathbb{E} \sum_i \xi_i = m/2$). Hence $N \geq e^{m/8}$. By duality, the packing number $M(1/4; H^m, d_H) \geq N \geq e^{m/8}$. □

3.4 Example: density estimation in C^2, rate $n^{-4/5}$

Let $\mathcal{F} = \{f : [0,1] \to [c_0, c_1] : \|f''\|_\infty \le c_2,\ \int_0^1 f = 1\}$ for constants $0 < c_0 < 1 < c_1,\ c_2 > 0$. We establish

$$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}\Big[H^2(\widehat{f}_n, f)\Big] \ \gtrsim\ n^{-4/5}.$$

Equivalently, since $H(f,g) \sim \|f - g\|_{L^2[0,1]} \sim \sqrt{\mathrm{KL}(f \,\|\, g)}$ on $\mathcal{F}$ (Theorem 9 below applied to a class of densities uniformly bounded away from 0), the same rate holds in L^2 and $\sqrt{\mathrm{KL}}$.

Theorem 9 (Equivalence of metrics on bounded-density classes)**.** *If P, Q have densities p, q with $0 < c_1 \le p(x), q(x) \le c_2 < \infty$ on $[0,1]$, then there exist constants k_1, k_2, k_3 depending only on (c_1, c_2) such that*

$$k_1 \int (p-q)^2 \le \mathrm{KL}(P \,\|\, Q)\ ,\ \chi^2(P \,\|\, Q)\ \le k_2 \int (p-q)^2, \qquad \mathrm{KL}(P \,\|\, Q) \le k_3 H^2(P, Q).$$

Step 1 (construct packing). Let $\phi : [0,1] \to \mathbb{R}$ be a fixed C^2 bump with $\|\phi\|_\infty \le 1/2$ and $\int_0^1 \phi = 0$. For $\alpha \in \{\pm 1\}^m$ and $x_j = j/m$ $(j = 0, \dots, m-1)$, set

$$f_\alpha(x) := 1 + \sum_{j=0}^{m-1} \alpha_j \phi_j(x), \qquad \phi_j(x) := \frac{C_*}{m^2}\, \phi(m(x - x_j))\, \mathbb{1}\{x \in [x_j, x_{j+1}]\}.$$

With $C_* > 0$ small enough, each $f_\alpha \in \mathcal{F}$. The choice of the prefactor m^{-2} (and not m^{-1}) is driven by the smoothness $\beta = 2$ implicit in the class $\mathcal{F}$: the second derivative of ϕ_j scales as $\|\phi_j''\|_\infty = C_* \|\phi''\|_\infty \cdot m^0$, so $\|f_\alpha''\|_\infty \le c_2$ provided $C_* \le c_2 / \|\phi''\|_\infty$. More generally, on $\mathcal{H}(\beta, L)$ one would use $\phi_j = (C_*/m^\beta)\phi(m(x - x_j))\mathbb{1}\{x \in [x_j, x_{j+1}]\}$, see the remark at the end of this section.

Step 2 (separation in Hellinger). For $\alpha \ne \beta$ in a Varshamov–Gilbert subset Ω with $d_H(\alpha, \beta) \ge 1/4$ (Lemma 14),

$$\begin{aligned} H^2(f_\alpha, f_\beta) &= \int_0^1 (\sqrt{f_\alpha} - \sqrt{f_\beta})^2 \asymp \int_0^1 (f_\alpha - f_\beta)^2 \\ &\gtrsim \sum_{j=0}^{m-1} (\alpha_j - \beta_j)^2 \cdot \frac{C_*^2}{m^4} \int_{x_j}^{x_{j+1}} \phi^2(m(x - x_j))\, dx \\ &\asymp \frac{m}{4} \cdot \frac{C_*^2}{m^5}\, \|\phi\|_{L^2[0,1]}^2 \ \gtrsim\ \frac{1}{m^4}. \end{aligned}$$

Hence the packing has Hellinger separation $\Phi_n \asymp m^{-2}$.

Step 3 (bound mutual information). By Theorem 9, $\mathrm{KL}(f_\alpha \,\|\, f_\beta) \lesssim \|f_\alpha - f_\beta\|_{L^2}^2 \lesssim m^{-4}$. Tensoriza-

tion gives

$$\mathrm{KL}(P_{f_\alpha}^n \,\|\, P_{f_\beta}^n) = n\,\mathrm{KL}(f_\alpha \,\|\, f_\beta) \lesssim \frac{n}{m^4}.$$

Combined with Lemma 13, $I(X;J) \lesssim n/m^4$.

Step 4 (apply Fano). From Lemma 14, $\log M \geq m/8$. Theorem 8 requires

$$\frac{I(X;J)+\log 2}{\log M} \leq \frac{Cn/m^4+\log 2}{m/8} < \frac{1}{2} \quad \Longleftrightarrow \quad n \lesssim m^5.$$

Taking $m \asymp n^{1/5}$ gives $\Phi_n \asymp m^{-2} \asymp n^{-2/5}$, yielding the $n^{-4/5}$ Hellinger-squared lower bound.

Remark 4 (Regression extension)**.** By essentially the same construction, one obtains the lower bound $n^{-4/5}$ for fixed-design Gaussian regression on the C^2-class, and more generally

$$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{H}(\beta,L)} \mathbb{E}\|\widehat{f}_n - f\|_{L^2}^2 \gtrsim n^{-2\beta/(2\beta+1)}.$$

The proof replaces the perturbations ϕ_j by bumps of magnitude C_*/m^β (matching the smoothness β) and repeats the four steps; see Tsybakov (2009, §2.6). This matches the upper bound achievable by the KDE of Section 1.5 on the Sobolev class.

Exercise 1. Prove: there exists $c > 0$ such that

$$\liminf_{n\to\infty} \inf_{\widehat{f}_n} \sup_{f \in \mathcal{H}(\beta,L)} \mathbb{E}_f\left[\|\widehat{f}_n - f\|_{L^2}^2 \cdot n^{\frac{2\beta}{2\beta+1}}\right] \geq c.$$

4 The Yang–Barron Method

Fano's method requires a packing of Θ *and* a uniform bound on pairwise KL. The Yang–Barron method (Yang and Barron, 1999) replaces the pairwise bound by a *covering of the distribution space* $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ in $\sqrt{\mathrm{KL}}$, which often simplifies computation.

Theorem 10 (Yang–Barron mutual information bound)**.** *Let $N_{\mathrm{KL}}(\epsilon;\mathcal{P})$ denote the ϵ-covering number of $\mathcal{P}$ in $\sqrt{\mathrm{KL}}$-distance ($\rho(P,Q) = \sqrt{\mathrm{KL}(P\,\|\,Q)}$). With the mixture setup of Section 3.1,*

$$I(X;J) \leq \inf_{\epsilon>0}\left[\epsilon^2 + \log N_{\mathrm{KL}}(\epsilon;\mathcal{P})\right].$$

Proof. By Lemma 13 with Q chosen freely, $I(X;J) \leq \frac{1}{M}\sum_j \mathrm{KL}(P_{\theta_j}\,\|\,Q) \leq \max_j \mathrm{KL}(P_{\theta_j}\,\|\,Q)$. Fix $\epsilon > 0$ and let $\{\gamma_1,\dots,\gamma_N\}$ be a $\sqrt{\mathrm{KL}}$-ϵ-cover of $\mathcal{P}$: for each j there is k_j with $\mathrm{KL}(P_{\theta_j}\,\|\,P_{\gamma_{k_j}}) \leq \epsilon^2$.

Take $Q = N^{-1}\sum_k P_{\gamma_k}$. Then

$$\begin{aligned}\mathrm{KL}(P_{\theta_j} \,\|\, Q) &= \mathbb{E}_{P_{\theta_j}} \log \frac{dP_{\theta_j}}{N^{-1}\sum_k dP_{\gamma_k}} \le \mathbb{E}_{P_{\theta_j}} \log \frac{dP_{\theta_j}}{N^{-1} dP_{\gamma_{k_j}}} \\ &= \mathrm{KL}(P_{\theta_j} \,\|\, P_{\gamma_{k_j}}) + \log N \le \epsilon^2 + \log N_{\mathrm{KL}}(\epsilon; \mathcal{P}).\end{aligned}$$

□

Proposition 5 (Yang–Barron method)**.** Suppose one can find sequences (ϵ_n, Φ_n) such that

(i) $\epsilon_n^2 \ge \log N_{\mathrm{KL}}(\epsilon_n/\sqrt{n}; \mathcal{P})$,

(ii) $\log M(2\Phi_n; \Theta, d) \ge 4\epsilon_n^2 + 2\log 2$.

Then the minimax d-risk is at least $\Phi_n^2/2$ up to constants.

Proof. Apply Theorem 10 to the n-sample family $\mathcal{P}^n$. The KL-tensorization $\mathrm{KL}(P^n\|Q^n) = n\,\mathrm{KL}(P\|Q)$ (Lemma 12) yields the covering-number identity

$$N_{\sqrt{\mathrm{KL}}}(\epsilon; \mathcal{P}^n) = N_{\sqrt{\mathrm{KL}}}(\epsilon/\sqrt{n}; \mathcal{P}),$$

so condition (i) implies $I(X; J) \le 2\epsilon_n^2$. Combined with (ii),

$$\frac{I(X;J) + \log 2}{\log M(2\Phi_n; \Theta, d)} \le \frac{2\epsilon_n^2 + \log 2}{4\epsilon_n^2 + 2\log 2} = \tfrac{1}{2}.$$

Fano's Theorem 8 then yields the claimed lower bound. □

Example 7 (Density estimation in C^2, revisited)**.** Take $\mathcal{F}$ as in Section 3.4. On this class, $\|f - g\|_{L^2} \sim H(f, g) \sim \sqrt{\mathrm{KL}(f \,\|\, g)}$ (Theorem 9), and it is standard that

$$\log N(\delta; \mathcal{F}, \|\cdot\|_2) \asymp \delta^{-1/2} \quad \text{as } \delta \to 0.$$

Hence the corresponding KL-covering number is $\log N_{\mathrm{KL}}(\epsilon; \mathcal{F}) \asymp \epsilon^{-1/2}$.

(i) Condition (i): $\epsilon_n^2 \ge \log N_{\mathrm{KL}}(\epsilon_n/\sqrt{n}; \mathcal{F}) \asymp (\sqrt{n}/\epsilon_n)^{1/2}$, giving $\epsilon_n^5 \asymp \sqrt{n}$, i.e. $\epsilon_n \asymp n^{1/10}$.

(ii) Condition (ii): by duality, $\log M(2\Phi_n; \mathcal{F}, \|\cdot\|_2) \asymp \log N(2\Phi_n) \asymp \Phi_n^{-1/2}$. So $\Phi_n^{-1/2} \gtrsim \epsilon_n^2 \asymp n^{1/5}$, giving $\Phi_n \asymp n^{-2/5}$.

Therefore $\Phi_n^2 \asymp n^{-4/5}$, matching the Fano computation in Section 3.4.

Part III

Generalization in Machine Learning

1 Empirical Risk Minimization, Risk, and Excess Risk

We now turn to supervised learning. Let $\mathcal{X}$ be a feature space, $\mathcal{Y}$ a label space, and suppose $(X_i, Y_i) \overset{\text{iid}}{\sim} P$ on $\mathcal{X} \times \mathcal{Y}$, for $i = 1, \dots, n$. Fix a hypothesis class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathcal{Y}$ (or $\mathcal{X} \to \mathbb{R}$ for regression) and a loss $\ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Example 8. • Regression, squared loss: $\ell(f; x, y) = (f(x) - y)^2$.

- Binary classification ($\mathcal{Y} = \{0, 1\}$), 0–1 loss: $\ell(f; x, y) = \mathbb{1}\{f(x) \neq y\}$.
- Binary classification, cross-entropy: $\ell(f; x, y) = -y \log \sigma(f(x)) - (1 - y) \log(1 - \sigma(f(x)))$ with $\sigma(z) = 1/(1 + e^{-z})$.

Define population and empirical risks

$$R(f) := \mathbb{E}_{(x,y)\sim P}\big[\ell(f; x, y)\big], \qquad \widehat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(f; X_i, Y_i).$$

The *empirical risk minimizer* is

$$\widehat{f}_n := \arg\min_{f \in \mathcal{F}} \widehat{R}_n(f).$$

Two quantities evaluate $\widehat{f}_n$:

- **Generalization error**: $\text{gen}(f) := R(f) - \widehat{R}_n(f)$. A random quantity measuring how well the in-sample risk predicts out-of-sample risk.
- **Excess risk**: $\Delta(f) := R(f) - R(f^*)$ where $f^* := \arg\min_{f \in \mathcal{F}} R(f)$ (often called the Bayes classifier in $\mathcal{F}$). The population-level benchmark.

The two are related (see Section 3) but not equal. In general, excess risk admits *fast rates* that generalization error alone cannot deliver.

2 Generalization Error for ERM

Standing assumptions for this section:

(A1) **Domain.** $\mathcal{X} = [0, 1]^d$ (technical convenience for concentration).

(A2) **Loss.** The loss is uniformly bounded, $\|\ell\|_\infty \le M$, and uniformly Lipschitz in its first argument,

$$|\ell(f;x,y) - \ell(f';x,y)| \le L\,\|f - f'\|_2 \qquad \forall\, x \in [0,1]^d, y \in \mathcal{Y}.$$

(A3) **VC-type hypothesis class.** There exist constants $A, V > 0$ such that the covering numbers of $\mathcal{F}$ in $\|\cdot\|_2$ satisfy

$$N(\epsilon\,\|\mathcal{F}\|_2; \mathcal{F}, \|\cdot\|_2) \;\le\; (A/\epsilon)^V \qquad \forall\, \epsilon \in (0,1]. \tag{31}$$

W.l.o.g. $\|\mathcal{F}\|_2 = 1$ (rescaling $\mathcal{F}$). We write $\mathcal{F} \in \mathrm{VC}(V, A)$.

Remark 5 (Uniform covering over all probability measures)**.** In the empirical-process literature (e.g., van der Vaart and Wellner, 2023, Ch. 2.5), the VC assumption is often formulated uniformly:

$$\sup_Q N(\varepsilon\|\mathcal{F}\|_{Q,2}; \mathcal{F}, \|\cdot\|_{Q,2}) \le (A/\varepsilon)^V,$$

where the supremum is over all finitely discrete probability measures Q on $\mathcal{X}$. This implies (31) for any fixed P with $\|\mathcal{F}\|_{P,2} < \infty$ (VdV–Wellner, Ch. 2.5, Exercise 1).

2.1 Crude bound via covering and union

Theorem 11 (Crude generalization-error bound)**.** *Under (A1)–(A3), for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\mathrm{gen}(\widehat{f}_n) \;\lesssim_{A,L,M}\; \sqrt{\frac{V \log n}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}. \tag{32}$$

Proof. Fix $f \in \mathcal{F}$. Since $\{\ell(f; X_i, Y_i)\}_{i=1}^n$ are i.i.d. in $[-M, M]$, Hoeffding gives

$$\Pr\left(\widehat{R}_n(f) - R(f) \ge \tau\right) \le \exp\left(-\frac{n\tau^2}{2M^2}\right). \tag{33}$$

Take a minimal ϵ-cover $\mathcal{N}_\epsilon$ of $\mathcal{F}$ in $\|\cdot\|_2$; by (A3), $|\mathcal{N}_\epsilon| \le (A/\epsilon)^V$. For the ERM $\widehat{f}_n$, there exists $g \in \mathcal{N}_\epsilon$ with $\|\widehat{f}_n - g\|_2 < \epsilon$. By (A2),

$$|\ell(\widehat{f}_n; x, y) - \ell(g; x, y)| \le L\epsilon,$$

and Jensen's inequality gives $|R(\widehat{f}_n) - R(g)| \le L\epsilon$ and $|\widehat{R}_n(\widehat{f}_n) - \widehat{R}_n(g)| \le L\epsilon$. Hence

$$\mathrm{gen}(\widehat{f}_n) = R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n) \le 2L\epsilon + \big[R(g) - \widehat{R}_n(g)\big]. \tag{34}$$

Taking a union bound over $\mathcal{N}_\epsilon$ in (33),

$$\Pr\left[\max_{g\in\mathcal{N}_\epsilon}(\widehat{R}_n(g)-R(g))\geq\tau\right]\leq(A/\epsilon)^V e^{-n\tau^2/(2M^2)}.$$

Set the RHS to δ: $\tau^2=\frac{2M^2}{n}[V\log(A/\epsilon)+\log(1/\delta)]$. Choose $\epsilon=1/\sqrt{n}$ to balance the Lipschitz error $2L\epsilon$ against τ; this yields (32). The $\sqrt{V\log n/n}$ arises from $V\log(A\sqrt{n})=V\log A+\frac{1}{2}V\log n$. □

2.2 Chaining: removing the $\log n$

Lemma 15 (McDiarmid's inequality)**.** *Let $X_1,\dots,X_n$ be independent random variables with $X_i\in\mathcal{X}_i$, and $f:\prod_i\mathcal{X}_i\to\mathbb{R}$ a measurable function satisfying the bounded-differences property*

$$\sup_{x_1,\dots,x_n,x_i'}|f(x)-f(x_1,\dots,x_i',\dots,x_n)|\leq c_i\qquad(i=1,\dots,n).$$

Then, for $Z=f(X_1,\dots,X_n)$ with $\mu=\mathbb{E}Z$, and every $t>0$,

$$\Pr(|Z-\mu|\geq t)\ \leq\ 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Theorem 12 (Chaining-based generalization bound)**.** *Under (A1)–(A3), for any $\delta\in(0,1)$, with probability at least $1-\delta$,*

$$\mathrm{gen}(\widehat{f}_n)\ \lesssim_{A,L,M}\ \sqrt{\frac{V}{n}}+\sqrt{\frac{\log(1/\delta)}{n}}.\tag{35}$$

Proof. Define $\varphi(\mathcal{S}):=\sup_{f\in\mathcal{F}}\left(R(f)-\widehat{R}_n(f)\right)$ viewed as a function of the sample $\mathcal{S}=\{(X_i,Y_i)\}_{i=1}^n$. Changing one data point changes $\widehat{R}_n(f)$ by at most $2M/n$ for any f, so the bounded-differences constant is $c_i=2M/n$. McDiarmid gives

$$\Pr(\varphi(\mathcal{S})-\mathbb{E}\varphi(\mathcal{S})\geq\tau)\leq\exp\left(-\frac{n\tau^2}{2M^2}\right),$$

so with probability at least $1-\delta$,

$$\varphi(\mathcal{S})\leq\mathbb{E}\varphi(\mathcal{S})+M\sqrt{\tfrac{2\log(1/\delta)}{n}}.\tag{36}$$

It remains to bound $\mathbb{E}\varphi(\mathcal{S})$. Define the centered empirical process

$$Q(f):=\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[\ell(f;X_i,Y_i)-\mathbb{E}\ell(f;X_i,Y_i)\right].$$

Then $\varphi(\mathcal{S}) = \frac{1}{\sqrt{n}} \sup_{f\in\mathcal{F}}(-Q(f))$ has mean $\frac{1}{\sqrt{n}}\mathbb{E}\sup_f Q(f)$ (by symmetry; see van der Vaart and Wellner, 2023, §2.3). Fix $f, g \in \mathcal{F}$. By Hoeffding applied to $\frac{1}{\sqrt{n}}\sum_i \big[\ell(f;\cdot) - \ell(g;\cdot)\big]$ (centered, bounded by $L\|f-g\|_2$ in magnitude),

$$\Pr(Q(f) - Q(g) \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2L^2\|f-g\|_2^2}\right).$$

Hence $\{Q(f)\}_{f\in\mathcal{F}}$ is a mean-zero sub-Gaussian process on $(\mathcal{F}, \frac{L}{\sqrt{2}}\|\cdot\|_2)$. The diameter of $\mathcal{F}$ is $D \lesssim L\sqrt{M}$ (by $\|f\|_\infty \leq M$). Theorem 4 (Dudley) combined with (31) gives

$$\begin{aligned}\mathbb{E}\sup_{f\in\mathcal{F}} Q(f) &\lesssim \int_0^D \sqrt{\log N(u;\mathcal{F}, L\|\cdot\|_2)}\, du \\ &\lesssim \int_0^{L\sqrt{M}/2} \sqrt{V\log(A/u)}\, du \;\lesssim\; \sqrt{V},\end{aligned}$$

since $\int_0^c \sqrt{\log(A/u)}\, du$ is a universal constant for bounded c. Therefore $\mathbb{E}\varphi(\mathcal{S}) \lesssim \sqrt{V/n}$, and combined with (36) yields (35). □

Remark 6 (Sharpness)**.** The rate $\sqrt{V/n}$ matches the parametric rate on VC-type classes (modulo the VC index V), and is known to be tight for typical classes; see Wainwright (2019, §4.3) and Bartlett et al. (2005).

2.3 Gaussian location: sharpness of $(V + \log(1/\delta))/n$

Consider $Y_i = w^* + \varepsilon_i$ with $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $w^* \in \mathbb{R}$, and squared loss $\ell(w;y) = (y-w)^2$. The ERM is $\widehat{w}_n = \bar{Y}_n = n^{-1}\sum_i Y_i$. Elementary computation:

$$\begin{aligned}R(\widehat{w}_n) &= \mathbb{E}(Y - \widehat{w}_n)^2 = (\widehat{w}_n - w^*)^2 + 1, \\ \widehat{R}_n(\widehat{w}_n) &= (\widehat{w}_n - w^*)^2 + \frac{1}{n}\sum_i \varepsilon_i^2 + \frac{2}{n}(w^* - \widehat{w}_n)\sum_i \varepsilon_i.\end{aligned}$$

Hence

$$\operatorname{gen}(\widehat{w}_n) = 1 - \tfrac{1}{n}\sum_i \varepsilon_i^2 - \tfrac{2(\widehat{w}_n - w^*)}{n}\sum_i \varepsilon_i.$$

The third term is lower order; the second is a chi-squared fluctuation. By standard concentration (Laurent and Massart, 2000, Lem. 1), $\Pr\left(\frac{1}{n}\sum_i(1-\varepsilon_i^2) \geq \tau\right) \leq e^{-n\tau^2/4}$, so

$$\operatorname{gen}(\widehat{w}_n) \asymp \sqrt{\tfrac{\log(1/\delta)}{n}} \qquad \text{with probability at least } 1-\delta.$$

The rate in Theorem 12 is thus sharp up to the VC constant, even in the simplest parametric example.

3 Excess Risk via Localization

The generalization-error bound controls the worst-case deviation between empirical and population risks. The *excess risk* $\Delta(\widehat{f}_n) = R(\widehat{f}_n) - R(f^*)$ is a weaker benchmark, but typically admits *fast* rates $\sim (V + \log(1/\delta))/n$ rather than $\sqrt{(V + \log(1/\delta))/n}$.

3.1 Link to generalization error

For any $\widehat{f}_n \in \mathcal{F}$ with $\widehat{R}_n(\widehat{f}_n) \le \widehat{R}_n(f^*)$ (proper ERM),

$$\begin{aligned}
\Delta(\widehat{f}_n) &= R(\widehat{f}_n) - R(f^*) \\
&= \underbrace{R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n)}_{=\mathrm{gen}(\widehat{f}_n)} + \underbrace{\widehat{R}_n(\widehat{f}_n) - \widehat{R}_n(f^*)}_{\le 0 \ \text{(ERM)}} + \underbrace{\widehat{R}_n(f^*) - R(f^*)}_{=-\mathrm{gen}(f^*)} \\
&\le \mathrm{gen}(\widehat{f}_n) - \mathrm{gen}(f^*) \\
&\le 2 \sup_{f \in \mathcal{F}} |R(f) - \widehat{R}_n(f)|.
\end{aligned}$$

Hence Theorem 12 gives $\Delta(\widehat{f}_n) \lesssim \sqrt{V/n}$; but a curvature argument yields $\Delta(\widehat{f}_n) \lesssim V/n$, as we now explain.

3.2 Curvature assumption and localization

Add to the assumptions of Section 2:

(A4) **Curvature.** There exists $\tau > 0$ such that for all $f \in \mathcal{F}, x \in \mathcal{X}, y \in \mathcal{Y}$,

$$\tfrac{1}{2}\tau \, \|f - f^*\|_2^2 \ \le \ \ell(f; x, y) - \ell(f^*; x, y) \ \le \ \tau \, \|f - f^*\|_2^2.$$

(A5) **Convexity.** $\mathcal{F}$ is convex.

The convexity assumption allows us to move along the line segment from $\widehat{f}_n$ to f^*. Define the interpolation

$$\widetilde{f}_t := t\widehat{f}_n + (1-t)f^*, \qquad t := \frac{\epsilon}{\epsilon + \|\widehat{f}_n - f^*\|_2} \in (0, 1],$$

for a tuning parameter $\epsilon > 0$. Then

$$\|\widetilde{f}_t - f^*\|_2 = t\|\widehat{f}_n - f^*\|_2 = \frac{\epsilon\|\widehat{f}_n - f^*\|_2}{\epsilon + \|\widehat{f}_n - f^*\|_2} \le \epsilon,$$

so $\widetilde{f}_t$ is always within ϵ of f^* in L^2. Moreover, the event $\{\|\widehat{f}_n - f^*\|_2 \ge \epsilon\}$ is equivalent (up to a factor of 2) to $\{\|\widetilde{f}_t - f^*\|_2 \ge \epsilon/2\}$, so proving a tail bound for the localized error suffices.

Define the *local class* at level $\eta := \tau\epsilon^2$:

$$\mathcal{F}(\eta) := \big\{f \in \mathcal{F} : R(f) - R(f^*) \le \eta\big\}.$$

By (A4), $R(\widetilde{f}_t) - R(f^*) \le \tau\epsilon^2 = \eta$, so $\widetilde{f}_t \in \mathcal{F}(\eta)$; also $f^* \in \mathcal{F}(\eta)$. By convexity of ℓ (which follows from (A4) with a minor adjustment) and ERM,

$$\widehat{R}_n(\widetilde{f}_t) \le t\widehat{R}_n(\widehat{f}_n) + (1-t)\widehat{R}_n(f^*) \le \widehat{R}_n(f^*).$$

Writing $R_0(f) := R(f) - R(f^*)$, this rearranges to

$$R_0(\widetilde{f}_t) = R_0(\widetilde{f}_t) - R_0(f^*) = \big[R_0(\widetilde{f}_t) - \widehat{R}_n(\widetilde{f}_t)\big] + \underbrace{\big[\widehat{R}_n(\widetilde{f}_t) - \widehat{R}_n(f^*)\big]}_{\le 0} + \big[\widehat{R}_n(f^*) - R_0(f^*)\big].$$

Applying $\widetilde{f}_t \in \mathcal{F}(\eta)$ and taking suprema,

$$R_0(\widetilde{f}_t) \;\le\; 2 \sup_{f \in \mathcal{F}(\eta)} \big|R_0(f) - \widehat{R}_n(f)\big|.$$

The lower curvature bound (A4) translates this to L^2-error:

$$\|\widetilde{f}_t - f^*\|_2^2 \;\le\; \tfrac{2}{\tau} \sup_{f \in \mathcal{F}(\eta)} \big|R_0(f) - \widehat{R}_n(f)\big|.$$

3.3 Talagrand's inequality and localized Rademacher complexity

To bound the local supremum, we use the following concentration result; the one-sided form with explicit constants is due to Massart (2000) and Bousquet (2002); see Wainwright (2019, §4.3) and Bartlett et al. (2005, Thm. 5.1) for the variance-dependent version used below.

Theorem 13 (Talagrand's concentration inequality; one-sided form)**.** *Let $\mathcal{G}$ be a countable class of functions $g : \mathcal{X} \to \mathbb{R}$ with $\sup_g \|g\|_\infty \le B$ and $\sup_g \mathbb{E}g^2 \le \sigma^2$. Let $\bar{g}(\mathcal{S}) := \frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}g(X)$.*

Then, for every $s > 0$,

$$\Pr\left\{\sup_{g\in\mathcal{G}}|\bar{g}(\mathcal{S})| \geq \mathbb{E}\sup_{g\in\mathcal{G}}|\bar{g}(\mathcal{S})| + \sqrt{\tfrac{2s\sigma^2}{n}} + \tfrac{2sB}{n}\right\} \leq e^{-s}.$$

Apply Theorem 13 to the centered excess-loss class $\mathcal{G} := \{g_f := \ell(f;\cdot) - \ell(f^*;\cdot) : f \in \mathcal{F}(\eta)\}$. Condition (A4) gives $\mathbb{E}g_f^2 \lesssim \tau\eta$, so we may take $\sigma^2 \lesssim \tau\eta$, and $B \leq 2M$ from (A2). Hence, with probability at least $1 - e^{-s}$,

$$\sup_{f\in\mathcal{F}(\eta)}\left|R_0(f) - \widehat{R}_n(f)\right| \;\leq\; \underbrace{\mathbb{E}\sup_{f\in\mathcal{F}(\eta)}\left|\tfrac{1}{n}\textstyle\sum_i g_f(X_i,Y_i) - \mathbb{E}g_f\right|}_{\text{localized Rademacher/Gaussian complexity}} + \sqrt{\tfrac{2s\tau\eta}{n}} + \tfrac{4sM}{n}.$$

The expectation can be bounded by chaining (as in Theorem 12) restricted to the local class $\mathcal{F}(\eta)$; on VC-type classes it scales as $\sqrt{\eta V/n}$. Solving the resulting fixed-point equation $\eta \asymp \sqrt{\eta V/n} + \tau\eta/\sqrt{n}\cdot\sqrt{s} + sM/n$ gives $\eta \asymp (V+s)/n$.

Theorem 14 (Fast rate for excess risk)**.** *Under (A1)–(A3) and (A4)–(A5), for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\Delta(\widehat{f}_n) \;\lesssim_{A,L,M,\tau}\; \frac{V}{n} + \frac{\log(1/\delta)}{n}.$$

Remark 7. The essential ingredients are (a) curvature, which converts parameter error into excess risk; (b) convexity (or star-shapedness) of $\mathcal{F}$, which allows the interpolation $\widetilde{f}_t$; and (c) Talagrand's inequality, which gives variance-dependent concentration. Relaxing (b) to star-shaped is generally non-trivial. See Bartlett et al. (2005) and Koltchinskii (2006) for a thorough treatment of localization and oracle inequalities.

Remark 8 (Scope and further reading)**.** The three parts of these notes are complementary: Part I proves rate-optimal upper bounds for KDE (pointwise) and NPLS (empirical L^2) under smoothness constraints; Part II shows these rates are minimax using three information-theoretic methods; and Part III extends the theory to general bounded-loss ERM on VC-type classes. For further material, see Tsybakov (2009), van der Vaart and Wellner (2023), Wainwright (2019), and Hastie et al. (2009).

`Zhan Gao, April 29, 2026` *Adapted from Xiaohui Chen's lecture notes of MATH647 at USC.*

Bibliography

References

Bartlett, P. L., O. Bousquet, and S. Mendelson (2005). Local Rademacher complexities. *The Annals of Statistics 33*(4), 1497–1537.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathématique 334*(6), 495–500.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory* (2nd ed.). John Wiley & Sons.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. Springer New York.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics 34*(6), 2593–2656.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics 28*(5), 1302–1338.

Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability 28*(2), 863–884.

Seijo, E. and B. Sen (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics 39*(3), 1633–1657.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. New York: Springer.

van der Vaart, A. W. and J. A. Wellner (2023). *Weak Convergence and Empirical Processes: With Applications to Statistics* (2 ed.). Springer Series in Statistics. Cham: Springer.

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

Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics 27*(5), 1564–1599.