Homework 5

SDS 391P.6, Spring 2026

Pratik Patil

Due: Apr 20 (Monday)

This homework is (also) a work in progress and is provided as-is for instructional purposes. The problems are drawn from various sources and do not (yet) have sufficient references to the original material. Additionally, errors may be present. so some cauution is advised! The document will be updated if corrections are necessary. Last updated: 2026-04-05.

0 Guidelines

- Please start early. If you have questions about the statements, notation, or possible typos, email us as soon as possible. When emailing about the course, please begin the subject line with [SDS 391P.6].
- Please begin your answer to each *main* question on a separate page. If you use any code, include it in an appendix. Submit a single combined PDF to Canvas. (If you encounter any submission issues, please let us know.)
- The problems and motivations draw on multiple sources (past course material, textbooks, and other standard references). If you use any external resources that materially guide your solution (beyond routine lookups), please cite them in your write-up.
- These questions are designed to build intuition and technique. You are welcome to go beyond what is explicitly asked. If you introduce additional assumptions (while keeping the spirit of the problem), state them clearly. If you discover something interesting along the way, feel free to include it as a brief remark; we may share especially instructive observations with the class.
- Parts labeled [Bonus] are optional: they are not required for full credit. They are intended as extra practice. You may skip them without penalty. If you attempt them, please label your solutions clearly with [Bonus].
- Many parts include hints intended to help you get started. You are not required to follow the suggested route, and you are encouraged to try alternative approaches when appropriate.
- We will grade primarily for correctness and clear reasoning. Do not over-optimize for minor presentation details. The spirit of the homework is for you to learn something new!

1 Practice with Gaussian processes and Gaussian width

In the last two lectures, we studied suprema of Gaussian and sub-Gaussian processes. For a bounded set $T \subset \mathbb{R}^d$, an especially important quantity is its *Gaussian width*

$$w(T) := \mathbb{E} \sup_{t \in T} \langle g, t \rangle, \qquad g \sim \mathcal{N}(0, I_d).$$

This quantity measures the size of T as seen by a random Gaussian direction. It plays the role of an effective dimension in many high-dimensional problems.

This exercise develops some basic properties of Gaussian width and computes it for several canonical sets.

Throughout, $g \sim \mathcal{N}(0, I_d)$, and $C, c > 0$ denote absolute constants.

(a) **Basic properties.** Let $T, S \subset \mathbb{R}^d$ be bounded.

(a1) Show that Gaussian width is monotone: if $T \subset S$, then

$$w(T) \leqslant w(S).$$

(a2) Show that Gaussian width is positively homogeneous: for every $a \geqslant 0$,

$$w(aT) = a\, w(T), \qquad aT := \{at : \ t \in T\}.$$

(a3) Show that Gaussian width is translation invariant: for every $x_0 \in \mathbb{R}^d$,

$$w(T + x_0) = w(T), \qquad T + x_0 := \{t + x_0 : \ t \in T\}.$$

(a4) Show that Gaussian width only depends on the convex hull:

$$w(\mathrm{conv}(T)) = w(T).$$

(*Hint:* For fixed g, the map $t \mapsto \langle g, t\rangle$ is linear, so its supremum over a convex hull is attained already on the original set.)

(b) **Canonical examples.**

(b1) Show that

$$w(B_2^d) = \mathbb{E}\|g\|_2,$$

and deduce that

$$c\sqrt{d} \leqslant w(B_2^d) \leqslant \sqrt{d}.$$

(*Hint:* Use Cauchy–Schwarz to identify the supremum over the Euclidean ball.)

(b2) Show that

$$w(B_1^d) = \mathbb{E}\|g\|_\infty.$$

Deduce that

$$w(B_1^d) \asymp \sqrt{\log d}.$$

(*Hint:* Use the duality relation $\sup_{\|t\|_1 \leqslant 1} \langle g, t\rangle = \|g\|_\infty$, then bound the maximum of d standard Gaussians by a union bound and a lower-tail argument. More explicitly, let $M = \max_{1 \leqslant i \leqslant d} |g_i|$. Show first that for a suitable constant $c > 0$,

$$\mathbb{P}\{M \leqslant c\sqrt{\log d}\} \leqslant e^{-c'}$$

for some absolute constant $c' > 0$. For the lower bound, use $\mathbb{E}M \geqslant t\, \mathbb{P}\{M \geqslant t\}$.)

(b3) Let $T = \{t_1, \dots, t_M\} \subset \mathbb{R}^d$ be a finite set. Show that

$$w(T) \leqslant \Big(\max_{1 \leqslant j \leqslant M} \|t_j\|_2 \Big) \sqrt{2 \log M}.$$

(*Hint:* The random variables $\langle g, t_j \rangle$ are centered Gaussian with variance $\|t_j\|_2^2$. Use the maximal inequality for finitely many sub-Gaussian random variables.)

(c) **[Bonus] Width of polytopes and sparse sets.**

(c1) Let $P = \operatorname{conv}\{v_1, \dots, v_N\} \subset \mathbb{R}^d$. Show that if $\|v_j\|_2 \leqslant 1$ for all j, then

$$w(P) \leqslant C \sqrt{\log N}.$$

(*Hint:* Combine part (a4) with part (b3).)

(c2) Let

$$T_k := \{x \in S^{d-1} : \ \|x\|_0 \leqslant k\},$$

the set of k-sparse unit vectors. Show that

$$w(T_k) \leqslant C \sqrt{k \log\Big(\frac{ed}{k}\Big)}.$$

(*Hint:* For each support $S \subset [d]$ with $|S| = k$, define

$$Y_S := \sup\{\langle g, x \rangle : \ \operatorname{supp}(x) \subset S, \ \|x\|_2 \leqslant 1\} = \|g_S\|_2.$$

Use part (b1) to show $\mathbb{E} Y_S \leqslant \sqrt{k}$. Then use Gaussian concentration to show that $Y_S - \mathbb{E} Y_S$ is sub-Gaussian with an absolute constant, uniformly in S. Finally apply the maximal inequality over the $\binom{d}{k}$ possible supports.)

2 Practice with Rademacher processes and Rademacher complexity

In the lecture on empirical processes, we introduced symmetrization and Rademacher averages. These quantities are the empirical-process analogue of Gaussian width.

Let $x_1, \dots, x_n \in \mathcal{X}$ be fixed sample points, and let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables. For a class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$, define the empirical Rademacher complexity

$$\widehat{\mathfrak{R}}_n(\mathcal{F}; x_1, \dots, x_n) := \mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i).$$

This exercise develops several standard bounds and examples.

(a) **Basic properties.** Show that for function classes $\mathcal{F}, \mathcal{G}$ and $a \geqslant 0$,

(a1) if $\mathcal{F} \subset \mathcal{G}$, then

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) \leqslant \widehat{\mathfrak{R}}_n(\mathcal{G});$$

(a2)
$$\widehat{\mathfrak{R}}_n(a\mathcal{F}) = a\,\widehat{\mathfrak{R}}_n(\mathcal{F}), \qquad a\mathcal{F} := \{af : \ f \in \mathcal{F}\};$$

(a3)
$$\widehat{\mathfrak{R}}_n(\mathrm{conv}(\mathcal{F})) = \widehat{\mathfrak{R}}_n(\mathcal{F}).$$

(*Hint:* For fixed signs ε_i, the map
$$f \mapsto \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i)$$
is linear in f.)

(b) **Finite-class bound (Massart-type bound).** Assume $\mathcal{F}$ is finite and that for every $f \in \mathcal{F}$,
$$\frac{1}{n}\sum_{i=1}^{n} f(x_i)^2 \leqslant r^2.$$
Show that
$$\widehat{\mathfrak{R}}_n(\mathcal{F}; x_1, \dots, x_n) \leqslant r\sqrt{\frac{2\log|\mathcal{F}|}{n}}.$$

(*Hint:* For fixed f, the Rademacher sum
$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i)$$
is centered sub-Gaussian with variance proxy $\frac{1}{n^2}\sum_{i=1}^{n} f(x_i)^2$. Then apply the maximal inequality for finitely many sub-Gaussian random variables.)

(c) **Linear function classes.** Let $\mathcal{X} = \mathbb{R}^d$, and consider the class
$$\mathcal{F}_R := \{x \mapsto \langle w, x\rangle : \ \|w\|_2 \leqslant R\}.$$
Show that
$$\widehat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \dots, x_n) = \frac{R}{n}\,\mathbb{E}_\varepsilon\Big\|\sum_{i=1}^{n} \varepsilon_i x_i\Big\|_2 \leqslant \frac{R}{n}\Big(\sum_{i=1}^{n} \|x_i\|_2^2\Big)^{1/2}.$$
Deduce that if $\|x_i\|_2 \leqslant 1$ for all i, then
$$\widehat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \dots, x_n) \leqslant \frac{R}{\sqrt{n}}.$$

(*Hint:* First compute the supremum over w using Cauchy–Schwarz, then bound the expectation by Jensen or Cauchy–Schwarz.)

(d) **[Bonus] Sparse linear predictors.** Let
$$\mathcal{G}_R := \{x \mapsto \langle w, x\rangle : \ \|w\|_1 \leqslant R\},$$
and assume $\|x_i\|_\infty \leqslant 1$ for all i. Show that
$$\widehat{\mathfrak{R}}_n(\mathcal{G}_R; x_1, \dots, x_n) \leqslant CR\sqrt{\frac{\log d}{n}}.$$

(*Hint:* Use the duality $\sup_{\|w\|_1 \leqslant R} \langle w, z \rangle = R\|z\|_\infty$, then bound

$$\mathbb{E}_\varepsilon \Big\| \sum_{i=1}^n \varepsilon_i x_i \Big\|_\infty$$

by controlling the maximum of d sub-Gaussian coordinates.)

3 Practice with VC dimension

In the lecture on VC theory, we introduced the notion of shattering and VC dimension as a combinatorial measure of complexity for Boolean classes. This exercise gives practice computing VC dimensions for several standard examples.

Recall that a class $\mathcal{H}$ of Boolean functions on a domain Ω *shatters* a finite set $\{x_1, \dots, x_m\} \subset \Omega$ if every labeling of these m points by $\{0,1\}$ can be realized by some $h \in \mathcal{H}$. The VC dimension $\mathrm{vc}(\mathcal{H})$ is the largest m for which some m-point set is shattered.

(a) **One-dimensional classes.** Compute the VC dimension of the following classes on $\mathbb{R}$:

(a1) the class of half-lines

$$\mathcal{H}_{\text{half}} = \{\mathbf{1}_{(-\infty, a]} : \ a \in \mathbb{R}\};$$

(a2) the class of intervals

$$\mathcal{H}_{\text{int}} = \{\mathbf{1}_{[a,b]} : \ a \leqslant b, \ a, b \in \mathbb{R}\};$$

(a3) the class of unions of at most k intervals

$$\mathcal{H}_k = \Big\{ \mathbf{1}_{\bigcup_{j=1}^k [a_j, b_j]} : \ a_j \leqslant b_j \Big\}.$$

Show that

$$\mathrm{vc}(\mathcal{H}_{\text{half}}) = 1, \qquad \mathrm{vc}(\mathcal{H}_{\text{int}}) = 2, \qquad \mathrm{vc}(\mathcal{H}_k) = 2k.$$

(*Hint:* For the upper bound, order the points as $x_1 < \cdots < x_{2k+1}$ and consider the alternating labeling $1, 0, 1, 0, \dots, 1$. Why would realizing this labeling require at least $k+1$ disjoint intervals? For the lower bounds, choose points in increasing order and explicitly construct intervals realizing any labeling.)

(b) **Two-dimensional classes.** Derive the VC dimension of the following classes on $\mathbb{R}^2$:

(b1) **Axis-aligned rectangles.** Let $\mathcal{R}$ be the class of axis-aligned rectangles in $\mathbb{R}^2$:

$$\mathcal{R} := \big\{ [a,b] \times [c,d] : \ a \leqslant b, \ c \leqslant d \big\}.$$

Show that the class of indicators of axis-aligned rectangles has VC dimension

$$\mathrm{vc}(\mathcal{R}) = 4.$$

(*Hint:* For the lower bound, use four points in the plane placed in general position, for example $(\pm 1, 0)$ and $(0, \pm 1)$, and show they can be shattered. For the upper bound, consider any five points. Among them, choose points with minimal and maximal x-coordinate and minimal and maximal y-coordinate. These account for at most four points. Show that the remaining point cannot be singled out by an axis-aligned rectangle while excluding all four extrema.)

(b2) **[Bonus] Convex sets.** Let $\mathcal{C}_{\text{conv}}$ be the class of indicators of all convex subsets of $\mathbb{R}^2$. Show that

$$\text{vc}(\mathcal{C}_{\text{conv}}) = \infty.$$

(*Hint:* Put m points on a circle. Given any labeling of these points by $\{0,1\}$, take the convex hull of the points labeled 1. Why does this convex set contain exactly the positively labeled points from the original m-point set?)

(c) **[Bonus] Euclidean balls and combinatorial counting.**

(c1) Show that the class of Euclidean balls in $\mathbb{R}^d$ has VC dimension $d+1$.

(*Hint:* You may use the standard lifting trick $x \mapsto (x, \|x\|_2^2)$ to reduce balls in $\mathbb{R}^d$ to half-spaces in $\mathbb{R}^{d+1}$, or prove the result directly.)

(c2) Use the Sauer–Shelah lemma to show that if $\mathcal{H}$ is any Boolean class with $\text{vc}(\mathcal{H}) = v$, then on any n-point set it induces at most

$$\sum_{j=0}^{v} \binom{n}{j} \leqslant \left(\frac{en}{v}\right)^v$$

distinct labelings.

(c3) Apply this bound to conclude that the number of subsets of an n-point set in $\mathbb{R}^2$ that can be cut out by axis-aligned rectangles is at most

$$\sum_{j=0}^{4} \binom{n}{j}.$$

4 Practice with statistical learning theory

We now connect empirical processes to statistical learning. The setup is the standard supervised-learning model.

Let (X, Y) be a random pair taking values in $\mathcal{X} \times \mathcal{Y}$, and let

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

be i.i.d. copies of (X, Y). Let $\mathcal{H}$ be a hypothesis class, and let $\ell(h(X), Y) \in [0,1]$ be a bounded loss.

For $h \in \mathcal{H}$, define the population risk

$$R(h) := \mathbb{E}[\ell(h(X), Y)]$$

and the empirical risk

$$R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i).$$

Let

$$\widehat{h} \in \arg\min_{h \in \mathcal{H}} R_n(h), \qquad h^\star \in \arg\min_{h \in \mathcal{H}} R(h).$$

This exercise derives excess-risk bounds using symmetrization, Rademacher complexity, and VC dimension.

(a) **Excess-risk lemma.** Show that

$$R(\widehat{h}) - R(h^\star) \leqslant 2 \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|.$$

(*Hint:* Add and subtract $R_n(\widehat{h})$ and $R_n(h^\star)$, and use the fact that $\widehat{h}$ minimizes the empirical risk.)

(b) **Symmetrization.** Let

$$\mathcal{L}_{\mathcal{H}} := \{(x, y) \mapsto \ell(h(x), y) : \ h \in \mathcal{H}\}.$$

Show that

$$\mathbb{E} \sup_{h \in \mathcal{H}} |R_n(h) - R(h)| \leqslant 2 \, \mathbb{E} \, \widehat{\mathfrak{R}}_n(\mathcal{L}_{\mathcal{H}}; (X_1, Y_1), \ldots, (X_n, Y_n)).$$

(*Hint:* Use the standard empirical-process symmetrization argument with an independent ghost sample and Rademacher signs.)

(c) **Finite hypothesis classes.** Assume $\mathcal{H}$ is finite, and the loss is bounded in $[0, 1]$. Show that

$$\mathbb{E}\big[R(\widehat{h}) - R(h^\star)\big] \leqslant C \sqrt{\frac{\log |\mathcal{H}|}{n}}.$$

(*Hint:* Apply part (b) and then Problem 2(b) to the finite class $\mathcal{L}_{\mathcal{H}}$. Note that $|\mathcal{L}_{\mathcal{H}}| \leqslant |\mathcal{H}|$, and each loss function has empirical L_2 norm at most 1.)

(d) **Boolean classification and VC dimension.** Assume now that $\mathcal{H}$ is a Boolean class, $\mathcal{Y} = \{0, 1\}$, and

$$\ell(h(x), y) = \mathbf{1}\{h(x) \neq y\}$$

is the 0-1 loss.

Suppose $\mathrm{vc}(\mathcal{H}) = v < \infty$. Show that

$$\mathbb{E}\big[R(\widehat{h}) - R(h^\star)\big] \leqslant C \sqrt{\frac{v \log(en/v)}{n}}.$$

(*Hint:* On a fixed sample $(x_i, y_i)_{i=1}^n$, the map

$$(h(x_1), \ldots, h(x_n)) \mapsto (\mathbf{1}\{h(x_1) \neq y_1\}, \ldots, \mathbf{1}\{h(x_n) \neq y_n\})$$

is a bijection of $\{0, 1\}^n$, obtained by XOR with the fixed label vector $(y_1, \ldots, y_n)$.)

(e) **[Bonus] A VC-style sample complexity statement.** Show that there exists an absolute constant C such that if

$$n \geqslant C \, \frac{v + \log(1/\delta)}{\varepsilon^2}$$

(up to an extra logarithmic factor if you keep the bound from part (d) in its current form), then with probability at least $1 - \delta$,

$$R(\widehat{h}) \leqslant R(h^\star) + \varepsilon.$$

(*Hint:* Upgrade expectation bounds to high-probability bounds using the same finite-class argument, or quote the VC law of large numbers from lecture if you prefer.)

5 Practice with nonparametric regression

In the previous problem, we used VC dimension to control excess risk for *Boolean* hypothesis classes. For real-valued regression classes, VC dimension is no longer the right complexity measure. A natural replacement is *metric entropy*, i.e. covering numbers of the hypothesis class.

This problem studies a basic idealized nonparametric regression model. We will see how excess-risk bounds can be derived from covering numbers. We will also see how the *curse of dimensionality* appears for Lipschitz classes, and how smoother classes improve the rate.

Let X be a random point in $[0,1]^d$ with law μ, and let

$$(X_1, T(X_1)), \dots, (X_n, T(X_n))$$

be noiseless training data, where $X_1, \dots, X_n$ are i.i.d. copies of X and $T : [0,1]^d \to [0,1]$ is an unknown target function.

For a hypothesis class $\mathcal{F} \subset \{f : [0,1]^d \to [0,1]\}$, define the population risk and empirical risk by

$$R(f) := \mathbb{E}\big[(f(X) - T(X))^2\big], \qquad R_n(f) := \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - T(X_i))^2.$$

Let

$$f^\star \in \arg\min_{f\in\mathcal{F}} R(f), \qquad \widehat{f}_n \in \arg\min_{f\in\mathcal{F}} R_n(f)$$

denote a population risk minimizer and an empirical risk minimizer.

You may use the following facts proved earlier:

- **Excess-risk lemma:**
$$R(\widehat{f}_n) - R(f^\star) \leqslant 2\sup_{f\in\mathcal{F}} |R_n(f) - R(f)|.$$
- **Empirical-process Dudley bound:** if $\mathcal{G}$ is a class of functions taking values in $[0,1]$, then
$$\mathbb{E}\sup_{g\in\mathcal{G}} \big|\mu_n(g) - \mu(g)\big| \leqslant \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{G}, \|\cdot\|_\infty, \varepsilon)}\, d\varepsilon.$$
- **Finite-class bound:** if $\mathcal{G}$ is a finite class of functions taking values in $[0,1]$, then
$$\mathbb{E}\sup_{g\in\mathcal{G}} \big|\mu_n(g) - \mu(g)\big| \leqslant C\sqrt{\frac{\log|\mathcal{G}|}{n}}.$$

For $L > 0$, define the Lipschitz class

$$\mathcal{F}_{L,d} := \big\{f : [0,1]^d \to [0,1] :\ \|f\|_{\mathrm{Lip}} \leqslant L\big\}.$$

You may also use the following covering-number bounds without proof:

$$\log \mathcal{N}(\mathcal{F}_{L,1}, \|\cdot\|_\infty, \varepsilon) \leqslant C\,\frac{L}{\varepsilon}, \qquad 0 < \varepsilon \leqslant 1,$$

$$\log \mathcal{N}(\mathcal{F}_{L,d}, \|\cdot\|_\infty, \varepsilon) \leqslant C_d \left(\frac{L}{\varepsilon}\right)^d, \qquad 0 < \varepsilon \leqslant 1,$$

where C_d may depend on the ambient dimension d.

(a) **Loss class versus hypothesis class.** Let

$$\mathcal{L}_{\mathcal{F}} := \big\{(x \mapsto (f(x) - T(x))^2) : \ f \in \mathcal{F}\big\}.$$

Show that for any $f, g \in \mathcal{F}$,

$$\|(f - T)^2 - (g - T)^2\|_\infty \leqslant 2\|f - g\|_\infty.$$

Deduce that for every $\varepsilon > 0$,

$$\mathcal{N}(\mathcal{L}_{\mathcal{F}}, \|\cdot\|_\infty, \varepsilon) \leqslant \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon/2).$$

(*Hint:* Use the identity

$$(f - T)^2 - (g - T)^2 = (f - g)(f + g - 2T),$$

and the fact that f, g, T all take values in $[0, 1]$.)

(b) **One-dimensional Lipschitz regression.** Assume now that $d = 1$ and $\mathcal{F} = \mathcal{F}_{L,1}$. Show that

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant C\sqrt{\frac{L}{n}}.$$

Thus, in one dimension, ERM over the class of L-Lipschitz functions achieves the same $n^{-1/2}$ scale as many parametric problems.

(*Hint:* Combine the excess-risk lemma with the empirical-process Dudley bound, then use part (a) and the covering-number estimate $\log \mathcal{N}(\mathcal{F}_{L,1}, \|\cdot\|_\infty, \varepsilon) \leqslant CL/\varepsilon$.)

(c) **Higher-dimensional Lipschitz regression.** Now let $d \geqslant 2$ and $\mathcal{F} = \mathcal{F}_{L,d}$.

(c1) Explain why the naive Dudley bound from part (b) is no longer useful in general: show that the integral

$$\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}_{L,d}, \|\cdot\|_\infty, \varepsilon)}\, d\varepsilon$$

diverges under the entropy bound

$$\log \mathcal{N}(\mathcal{F}_{L,d}, \|\cdot\|_\infty, \varepsilon) \leqslant C_d (L/\varepsilon)^d.$$

(*Hint:* Compare the integrand to a constant multiple of $\varepsilon^{-d/2}$.)

(c2) Let $\mathcal{N}_\varepsilon \subset \mathcal{L}_{\mathcal{F}}$ be an ε-net of $\mathcal{L}_{\mathcal{F}}$ in $\|\cdot\|_\infty$. Show that

$$\sup_{g \in \mathcal{L}_{\mathcal{F}}} |\mu_n(g) - \mu(g)| \leqslant 2\varepsilon + \max_{h \in \mathcal{N}_\varepsilon} |\mu_n(h) - \mu(h)|.$$

(*Hint:* For each $g \in \mathcal{L}_{\mathcal{F}}$, choose $h \in \mathcal{N}_\varepsilon$ with $\|g - h\|_\infty \leqslant \varepsilon$, then add and subtract $\mu_n(h) - \mu(h)$.)

(c3) Deduce that for every $\varepsilon \in (0,1)$,

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant C\left(\varepsilon + \sqrt{\frac{\log \mathcal{N}(\mathcal{L}_{\mathcal{F}}, \|\cdot\|_\infty, \varepsilon)}{n}}\right).$$

Then use part (a) and the covering-number bound for $\mathcal{F}_{L,d}$ to conclude that

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant C_d\left(\varepsilon + \sqrt{\frac{(L/\varepsilon)^d}{n}}\right).$$

(c4) Choose

$$\varepsilon \asymp L^{d/(d+2)}\, n^{-1/(d+2)}$$

and deduce the excess-risk bound

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant C_d\, L^{d/(d+2)}\, n^{-1/(d+2)}.$$

Finally, explain why this is an instance of the *curse of dimensionality*.

(*Hint:* To make the excess risk at most δ, solve for the sample size n in terms of δ, d, and L.)

(d) **[Bonus] Smoother classes help.** Suppose now that $\mathcal{F} \subset \{f : [0,1]^d \to [0,1]\}$ is a hypothesis class whose covering numbers satisfy

$$\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leqslant A\,\varepsilon^{-p} \qquad \text{for all } 0 < \varepsilon \leqslant 1,$$

for some constants $A > 0$ and $p > 0$.

Show, using the same finite-net argument as in part (c), that

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant C\, A^{1/(p+2)}\, n^{-1/(p+2)}.$$

Now suppose moreover that $\mathcal{F}$ is a class of s-smooth functions for which $p = d/s$ (this is the entropy behavior of many Hölder/Sobolev-type classes). Deduce the heuristic rate

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \lesssim n^{-\,s/(2s+d)}.$$

Explain briefly why increasing the smoothness s improves the rate, and why this illustrates one way machine learning can escape the curse of dimensionality.

(*Hint:* Let $p = d/s$ in the bound above and simplify the exponent. What happens formally as $s \to \infty$?)

Source material

Parts of this homework were inspired by exercises from Vershynin (2018); Tropp (2023), in addition to the author's accumulated experience working on related topics.

References

Tropp, J. A. (2023). Probability in high dimensions. Caltech CMS Lecture Notes 2021-01.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.