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.
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

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

Solution:
Because $T \subset S$, the supremum of $\langle g, s \rangle$ over $s \in S$ is at least the supremum of $\langle g, t \rangle$ over $t \in T$, and finally taking the expectation gives

$$\mathbb{E} \sup_{t \in T} \langle g, t \rangle \leqslant \mathbb{E} \sup_{s \in S} \langle g, s \rangle \implies w(T) \leqslant w(S). \qquad \square$$

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

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

Solution:

$$w(aT) = \mathbb{E} \sup_{t \in T} \langle g, at \rangle = \mathbb{E} \sup_{t \in T} a \langle g, t \rangle = a \left(\mathbb{E} \sup_{t \in T} \langle g, t \rangle \right) = a\, w(T), \qquad \text{for all } a \geqslant 0. \qquad \square$$

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

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

Solution:
Using the property of inner product $\langle x, y + \alpha \rangle = \langle x, y \rangle + \langle x, \alpha \rangle$, for every $x_0 \in \mathbb{R}^d$,

$$w(T + x_0) = \mathbb{E} \sup_{t \in T} \langle g, t + x_0 \rangle = \mathbb{E} \sup_{t \in T} \langle g, t \rangle + \mathbb{E} \langle g, x_0 \rangle \overset{(i)}{=} \mathbb{E} \sup_{t \in T} \langle g, t \rangle = w(T),$$

where (i) follows from the fact that $g^\top x_0$ is a weighted sum of independent mean-zero random variables and thus $\mathbb{E}[g^\top x_0] = 0$. Hence, the Gaussian width is translation invariant. $\square$

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

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

Solution:
The convex hull of a set T is defined as:

$$\operatorname{conv}(T) = \left\{ \sum_{i=1}^{m} \lambda_i t_i : m \in \mathbb{N},\, t_i \in T,\, \lambda_i \geqslant 0,\, \sum_{i=1}^{m} \lambda_i = 1 \right\}.$$

Fix g. Then for any $x \in \operatorname{conv}(T)$, using linearity of the inner products,

$$\langle g, x \rangle = \sum_{i=1}^{m} \lambda_i \langle g, t_i \rangle \leqslant \sup_{t \in T} \langle g, t \rangle \sum_{i=1}^{m} \lambda_i = \sup_{t \in T} \langle g, t \rangle.$$

This holds for every $x \in \operatorname{conv}(T)$, so

$$\sup_{x \in \operatorname{conv}(T)} \langle g, x \rangle \leqslant \sup_{t \in T} \langle g, t \rangle.$$

Taking expectation on both side yields $w(\operatorname{conv}(T)) \leqslant w(T)$. We know, $T \subset \operatorname{conv}(T)$, so using part (a1), $w(T) \leqslant w(\operatorname{conv}(T))$. Hence, combining both we get

$$w(\operatorname{conv}(T)) = w(T). \qquad \square$$

(b) **Canonical examples.**

(b1) Show that

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

and deduce that

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

Solution:
Here, the Euclidean ball in $\mathbb{R}^d$ is defined as

$$B_2^d = \{x : \|x\|_2 \leqslant 1,\, x \in \mathbb{R}^d\}.$$

For a fixed realization of g, by Cauchy-Schwarz:

$$\langle g, x \rangle \leqslant \|g\|_2 \|x\|_2 \leqslant \|g\|_2, \qquad \text{for every } x \in B_2^d.$$

This gives

$$\sup_{x \in B_2^d} \langle g, x \rangle \leqslant \|g\|_2.$$

If $g \neq 0$, choose $x = \dfrac{g}{\|g\|_2} \in B_2^d$, then $\left\langle g, \dfrac{g}{\|g\|_2} \right\rangle = \|g\|_2$. If $g = 0$, then the equality holds trivially.
Thus, $\sup_{x \in B_2^d} \langle g, x \rangle = \|g\|_2$ a.s., and taking expectation gives $w(B_2^d) = \mathbb{E}\,\|g\|_2$. $\square$
Now,
The mapping $x \mapsto \sqrt{x}$ is concave, so using Jensen's inequality

$$\mathbb{E}\,\|g\|_2 = \mathbb{E}\sqrt{\sum_{i=1}^{d} g_i^2} \leqslant \sqrt{\mathbb{E}\sum_{i=1}^{d} g_i^2} = \sqrt{\sum_{i=1}^{d} \mathbb{E}[g_i^2]} = \sqrt{d}.$$

Let $X = \|g\|_2^2 = \sum_{i=1}^{d} g_i^2$, then $X \sim \chi_d^2$, so $\mathbb{E}X = d$ and $\mathbb{E}X^2 = d^2 + 2d$. Applying Paley-Zygmund inequality with $\lambda = 1/2$,

$$\mathbb{P}\left\{X \geqslant \frac{d}{2}\right\} \geqslant \frac{1}{4}\frac{d^2}{d^2+2d} \geqslant \frac{1}{12},$$

because $d^2 + 2d \leqslant 3d^2$ for $d \geqslant 1$. We know, for a non-negative random variable Y and $t \geqslant 0$, $Y \geqslant t\mathbb{1}_{\{Y \geqslant t\}} \implies \mathbb{E}Y \geqslant t\mathbb{P}\{Y \geqslant t\}$. Using this fact

$$\mathbb{E}\,\|g\|_2 \geqslant \sqrt{\frac{d}{2}}\,\mathbb{P}\left\{\|g\|_2 \geqslant \sqrt{\frac{d}{2}}\right\} \geqslant \frac{1}{12}\sqrt{\frac{d}{2}} \gtrsim \sqrt{d}.$$

Hence, for some absolute constant $c > 0$,

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

(b2) Show that

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

Deduce that

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

Solution:
Here, the unit ball associated with the ℓ_1 norm in $\mathbb{R}^d$ is defined as

$$B_1^d = \{x : \|x\|_1 \leqslant 1,\, x \in \mathbb{R}^d\}.$$

For a fixed realization of g, using Hölder's inequality:

$$\langle g, x\rangle \leqslant \|g\|_\infty \|x\|_1 \leqslant \|g\|_\infty, \qquad \text{for every } x \in B_1^d.$$

This gives

$$\sup_{x \in B_1^d} \langle g, x\rangle \leqslant \|g\|_\infty.$$

To get equality, choose an index i_* such that

$$|g_{i_*}| = \|g\|_\infty,$$

and set $t = \operatorname{sign}(g_{i_*})e_{i_*}$. Then $\|t\|_1 = 1$, so $t \in B_1^d$, and

$$\langle g, t\rangle = \operatorname{sign}(g_{i_*})g_{i_*} = |g_{i_*}| = \|g\|_\infty.$$

Hence

$$\sup_{x \in B_1^d} \langle g, x\rangle = \|g\|_\infty.$$

Taking expectations,

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

Now, define $2d$ centered Gaussian variables

$$Z_1, \ldots, Z_{2d} := g_1, \ldots, g_d, -g_1, \ldots, -g_d.$$

Here, each Z_k is a standard Normal distribution, hence sub-Gaussian with ψ_2-norm bounded by an absolute constant. Also,

$$\max_{1 \leqslant k \leqslant 2d} Z_k = \max_{1 \leqslant j \leqslant d} |g_j| = \|g\|_\infty.$$

Therefore, using the Finite maximal inequality (Lemma 4.1 from Lecture 17):

$$\mathbb{E}\,\|g\|_\infty = \mathbb{E}\max_{1\leqslant k\leqslant 2d} Z_k \leqslant C\sqrt{\log(2d)} \leqslant C\sqrt{\log d},$$

which gives

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

First, we prove a Gaussian tail lower bound. Let $G \sim \mathcal{N}(0,1)$. Then,

$$\mathbb{P}\{|G| > t\} = 2\mathbb{P}\{G > t\} = \frac{2}{\sqrt{2\pi}}\int_t^\infty e^{-x^2/2}\,dx.$$

For $t \geqslant 1$,

$$\int_t^\infty e^{-x^2/2}\,dx \geqslant \int_t^{t+1/t} e^{-x^2/2}\,dx.$$

Now, if $x \in [t, t+1/t]$, then

$$x^2 \leqslant \left(t+\frac{1}{t}\right)^2 = t^2 + 2 + \frac{1}{t^2} \leqslant t^2 + 3,$$

because $t \geqslant 1$. Therefore, $e^{-x^2/2} \geqslant e^{-(t^2+3)/2} = e^{-3/2}e^{-t^2/2}$ for all $x \in [t, t+1/t]$. This gives

$$\int_t^{t+1/t} e^{-x^2/2}\,dx \geqslant \frac{1}{t}e^{-3/2}e^{-t^2/2},$$

and multiplying by $2/\sqrt{2\pi}$, we obtain

$$\mathbb{P}\{|G| > t\} \geqslant \frac{2e^{-3/2}}{\sqrt{2\pi}}\frac{e^{-t^2/2}}{t}, \qquad (t \geqslant 1).$$

So there is an absolute constant $c_0 > 0$ such that,

$$\mathbb{P}\{|G| > t\} \geqslant c_0\frac{e^{-t^2/2}}{t}, \qquad t \geqslant 1.$$

Now, since the coordinates of g are independent

$$\mathbb{P}\{\|g\|_\infty \leqslant t\} = \prod_{i=1}^d \mathbb{P}\{|g_i| \leqslant t\} = \left(1 - \mathbb{P}\{|G| > t\}\right)^d.$$

Using the identity $1 - x \leqslant e^{-x}$,

$$\mathbb{P}\{\|g\|_\infty \leqslant t\} \leqslant \exp\left(-d\,\mathbb{P}\{|G| > t\}\right).$$

Choose $t = c\sqrt{\log d}$, where $c > 0$ is a small absolute constant to be chosen. For d large enough, $t = c\sqrt{\log d} \geqslant 1$, so using the Gaussian tail lower bound derived earlier,

$$d\,\mathbb{P}\{|G| > t\} \geqslant c_0\,d\,\frac{e^{-t^2/2}}{t}.$$

Substituting $t = c\sqrt{\log d}$, $e^{-t^2/2} = d^{-c^2/2}$. This gives

$$d\,\mathbb{P}\left\{|G| > c\sqrt{\log d}\right\} \geqslant \frac{c_0}{c\sqrt{\log d}}d^{1-c^2/2}.$$

Now choose $c > 0$ so that $1 - c^2/2 > 0$, then the right-hand side tends to $+\infty$ as $d \to \infty$. In particular, for such a choice of c, there exists an absolute constant $a > 0$ such that for all sufficiently large d,

$$d\,\mathbb{P}\left\{|G| > c\sqrt{\log d}\right\} \geqslant a.$$

Putting this bound back,

$$\mathbb{P}\left\{\|g\|_\infty \leqslant c\sqrt{\log d}\right\} \leqslant e^{-a}.$$

Equivalently,

$$\mathbb{P}\left\{\|g\|_\infty \geqslant c\sqrt{\log d}\right\} \geqslant 1 - e^{-a}.$$

Using the inequality $\mathbb{E}\,\|g\|_\infty \geqslant t\,\mathbb{P}\{\|g\|_\infty \geqslant t\}$ with $t = c\sqrt{\log d}$, we get

$$\mathbb{E}\,\|g\|_\infty \geqslant c\sqrt{\log d}\,\mathbb{P}\left\{\|g\|_\infty \geqslant c\sqrt{\log d}\right\} \geqslant c\sqrt{\log d}(1 - e^{-a}) \gtrsim \sqrt{\log d}.$$

Hence, combining both lower and upper bounds, we get

$$w(B_1^d) = \mathbb{E}\,\|g\|_\infty \asymp \sqrt{\log d}. \qquad \square$$

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

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

Solution:
Since T is a finite set, by the definition of Gaussian width

$$w(T) = \mathbb{E}\max_{1\leqslant j\leqslant M}\langle g, t_j\rangle, \qquad g \in \mathcal{N}(0, I_d).$$

Define $Z_j := \langle g, t_j\rangle$, for $j = 1, \ldots, M$. Each Z_j is Gaussian with mean 0 and variance $\|t_j\|_2^2$, i.e., $Z_j \sim \mathcal{N}(0, \|t_j\|_2^2)$. Thus, for all j,

$$\mathrm{Var}(Z_j) \leqslant R^2, \qquad R := \max_{1\leqslant j\leqslant M} \|t_j\|_2 .$$

Fix $\lambda > 0$. Since exponential is increasing $e^{\lambda \max_j Z_j} = \max_j e^{\lambda Z_j} \leqslant \sum_{j=1}^M e^{\lambda Z_j}$. So, taking logs and dividing by λ,

$$\max_j Z_j \leqslant \frac{1}{\lambda}\log\left(\sum_{j=1}^M e^{\lambda Z_j}\right).$$

Since log is concave, taking expectations and using Jensen gives

$$\mathbb{E}\max_{1\leqslant j\leqslant M} Z_j \leqslant \frac{1}{\lambda}\log\left(\sum_{j=1}^M \mathbb{E}e^{\lambda Z_j}\right).$$

Because $Z_j \sim \mathcal{N}(0, \|t_j\|_2^2)$, using the sub-Gaussian MGF bound,

$$\mathbb{E}e^{\lambda Z_j} \leqslant e^{\lambda^2\|t_j\|_2^2/2} \leqslant e^{\lambda^2 R^2/2} \implies \sum_{j=1}^M \mathbb{E}e^{\lambda Z_j} \leqslant Me^{\lambda^2 R^2/2}.$$

This gives

$$\mathbb{E}\max_{1\leqslant j\leqslant M} Z_j \leqslant \frac{1}{\lambda}\log\left(Me^{\lambda^2 R^2/2}\right) = \frac{\log M}{\lambda} + \frac{\lambda R^2}{2}.$$

Now, optimize in λ by minimizing $f(\lambda) = \dfrac{\log M}{\lambda} + \dfrac{\lambda R^2}{2}$, which gives the optimal choice is $\lambda^* = \dfrac{\sqrt{2\log M}}{R}$. Substituting λ^*,

$$\frac{\log M}{\lambda} + \frac{\lambda R^2}{2} = \frac{R\log M}{\sqrt{2\log M}} + \frac{R\sqrt{2\log M}}{2} = R\sqrt{2\log M}.$$

Hence, we get,

$$w(T) = \mathbb{E} \max_{1 \leqslant j \leqslant M} Z_j \leqslant \left(\max_{1 \leqslant j \leqslant M} \|t_j\|_2 \right) \sqrt{2 \log M}. \qquad \square$$

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

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

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

[Bonus] Solution:
From part (a4), Gaussian width depends only on the convex hull, so

$$w(P) = w(\{v_1, \ldots, v_N\}).$$

Applying part (b3), to finite set $T = \{v_1, \ldots, v_N\}$ gives,

$$w(P) = w(T) \leqslant \left(\max_{1 \leqslant j \leqslant N} \|v_j\|_2 \right) \sqrt{2 \log N} \leqslant \sqrt{2 \log N},$$

because $\|v_j\|_2 \leqslant 1$ for all j. Hence, for some absolute constant $C > 0$,

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

(c2) Let

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

the set of k-sparse unit vectors. Show that

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

[Bonus] Solution:
First, for each support $S \subset [d]$ with $|S| = k$, define

$$E_S := \{x \in \mathbb{R}^d \ : \ \operatorname{supp}(x) \subset S\}.$$

Then E_S is a $k-$dimensional coordinate subspace, and

$$T_k \subset \bigcup_{|S|=k} (B_2^d \cap E_S).$$

So on each support, we just see a k-dimensional Euclidean ball. Also, E_S is isometric to $\mathbb{R}^k$, so $B_2^d \cap E_S$ is isometric to B_2^k. Fix such an S. Using the covering number property for Euclidean balls (from Lecture 11), B_2^k admits a 1/2-net of size at most 6^k, and certainly at most C^k for an absolute constant C. Let $N_S \subset B_2^d \cap E_S$ be a 1/2-net. Then $|N_S| \leqslant C^k$. Let $K = \operatorname{conv}(N_S)$. Fix $g \in E_S$. Choose

$$x_* = \frac{g}{\|g\|_2} \in B_2^d \cap E_S$$

if $g \neq 0$. Since N_S is a 1/2-net, there exists $y \in N_S$ such that

$$\|x_* - y\|_2 \leqslant \frac{1}{2}.$$

Using Cauchy-Schwarz $\langle g, y - x_* \rangle \geqslant - \|g\|_2 \|y - x_*\|_2$, so

$$\langle g, y \rangle = \langle g, x_* \rangle + \langle g, y - x_* \rangle \geqslant \|g\|_2 - \|g\|_2 \|y - x_*\|_2 \geqslant 1/2 \|g\|_2 .$$

Hence,

$$\sup_{z \in K} \langle g, z \rangle \geqslant \frac{1}{2} \|g\|_2 .$$

But,

$$\sup_{x \in B_2^d \cap E_S} \langle g, x \rangle = \|g\|_2 ,$$

by Cauchy-Schwarz and taking $x = \dfrac{g}{\|g\|_2}$ if $g \neq 0$. Therefore,

$$\sup_{x \in B_2^d \cap E_S} \langle g, x \rangle \leqslant 2 \sup_{z \in K} \langle g, z \rangle = \sup_{z \in 2K} \langle g, z \rangle,$$

because scaling a set by 2 scales its support functions by 2. Since both $B_2^d \cap E_S$ and $2K$ are closed convex subsets of E_S, and above inequality holds for every $g \in E_S$,

$$B_2^d \cap E_S \subset 2K = 2 \operatorname{conv}(N_S).$$

Now, define

$$V := \bigcup_{|S|=k} N_S.$$

Because every point of T_k lies in some $B_2^d \cap E_S$,

$$T_k \subset 2 \operatorname{conv}(V).$$

Also every $v \in V$ satisfies $\|v\|_2 \leqslant 1$ and

$$|V| \leqslant \binom{d}{k} C^k.$$

Using monotonicity, positive homogeneity, and convex-hull invariance of Gaussian width

$$w(T_k) \leqslant 2w(\operatorname{conv}(V)) = 2w(V).$$

Applying part (c1) to the polytope $\operatorname{conv}(V)$ gives

$$w(T_k) \leqslant 2C_0 \sqrt{\log |V|}.$$

Since

$$\log |V| \leqslant \log \binom{d}{k} + k \log C \leqslant k \log \left(\frac{ed}{k} \right) + C'k,$$

using $\binom{d}{k} \leqslant \left(\dfrac{ed}{k} \right)^k$, we obtain

$$\log |V| \leqslant C'' k \log \left(\frac{ed}{k} \right).$$

Hence, we get

$$w(T_k) \leqslant C \sqrt{k \log \left(\frac{ed}{k} \right)}. \qquad \square$$

2 Practice with Rademacher processes and Rademacher complexity

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

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

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

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

Solution:
For each fixed realization of the Rademacher r.v. $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$, define

$$A_\varepsilon(f) := \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i).$$

Since $\mathcal{F} \subset \mathcal{G}$, the supremum over the larger class is at least the supremum over the smaller:

$$\sup_{f \in \mathcal{F}} A_\varepsilon(f) \leqslant \sup_{g \in \mathcal{G}} A_\varepsilon(g).$$

That is,

$$\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \leqslant \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(x_i).$$

Since expectation preserves inequalities, taking expectation w.r.t. ε gives

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right] \leqslant \mathbb{E}_\varepsilon \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(x_i) \right].$$

Therefore,

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) \leqslant \widehat{\mathfrak{R}}_n(\mathcal{G}). \qquad \square$$

(a2)

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

Solution:
Using the definition of Rademacher complexity:

$$\begin{aligned} \widehat{\mathfrak{R}}_n(a\mathcal{F}) &= \mathbb{E}_\varepsilon \left[\sup_{g \in a\mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(x_i) \right] \\ &= \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i\, af(x_i) \right] \\ &= a \cdot \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right] = a\, \widehat{\mathfrak{R}}_n(\mathcal{F}). \qquad \square \end{aligned}$$

(a3)

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

Solution:
For a fixed realization $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ define the linear functional

$$A_\varepsilon(f) := \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i) \quad \text{such that} \quad \widehat{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}} A_\varepsilon(f)\right].$$

Take any $g \in \mathrm{conv}(\mathcal{F})$. By the definition of convex hull, there exists $f_1, \ldots, f_m$ and non-negative coefficients $\lambda_1, \ldots, \lambda_m$ such that

$$g = \sum_{j=1}^{m} \lambda_j f_j \quad \text{with} \quad \sum_{j=1}^{m} \lambda_j = 1.$$

Since $A_\varepsilon(g)$ is linear,

$$A_\varepsilon(g) = A_\varepsilon\left(\sum_{j=1}^{m} \lambda_j f_j\right) = \sum_{j=1}^{m} \lambda_j A_\varepsilon(f_j).$$

Because $\lambda_j \geqslant 0$, bounding the sum by the sum of the maximum gives,

$$A_\varepsilon(g) \leqslant \left(\max_{1\leqslant j\leqslant m} A_\varepsilon(f_j)\right)\sum_{j=1}^{m}\lambda_j = \max_{1\leqslant j\leqslant m} A_\varepsilon(f_j) \leqslant \sup_{f\in\mathcal{F}} A_\varepsilon(f),$$

for every $g \in \mathrm{conv}(\mathcal{F})$. Thus taking the supremum over $g \in \mathrm{conv}(\mathcal{F})$, we get

$$\sup_{g\in\mathrm{conv}(\mathcal{F})} A_\varepsilon(g) \leqslant \sup_{f\in\mathcal{F}} A_\varepsilon(f).$$

Finally taking expectation w.r.t. ε yields:

$$\mathbb{E}_\varepsilon\left[\sup_{g\in\mathrm{conv}(\mathcal{F})} A_\varepsilon(g)\right] \leqslant \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}} A_\varepsilon(f)\right] \quad \text{i.e.,} \quad \widehat{\mathfrak{R}}_n(\mathrm{conv}(\mathcal{F})) \leqslant \widehat{\mathfrak{R}}_n(\mathcal{F}).$$

We know $\mathcal{F} \subset \mathrm{conv}(\mathcal{F})$, so using part (a1), $\widehat{\mathfrak{R}}_n(\mathcal{F}) \leqslant \widehat{\mathfrak{R}}_n(\mathrm{conv}(\mathcal{F}))$. Hence, combining both

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

(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

$$\widehat{\mathfrak{R}}_n(\mathcal{F}; x_1, \ldots, x_n) \leqslant r\sqrt{\frac{2\log|\mathcal{F}|}{n}}.$$

Solution:
For each $f \in \mathcal{F}$ define

$$Z_f := \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(x_i) \quad \text{so that} \quad \widehat{\mathfrak{R}}_n(\mathcal{F}; x_1, \ldots, x_n) = \mathbb{E}_\varepsilon\left[\max_{f\in\mathcal{F}} Z_f\right].$$

Fix $f \in \mathcal{F}$. Since the ε_i's are independent Rademacher variables, $\mathbb{E}_\varepsilon Z_f = 0$. Moreover,

$$\mathbb{E}_\varepsilon e^{\lambda Z_f} = \prod_{i=1}^{n} \mathbb{E}_\varepsilon \exp\left(\lambda \frac{\varepsilon_i f(x_i)}{n}\right).$$

A Rademacher random variable ε_i is sub-Gaussian and satisfies

$$\mathbb{E}_\varepsilon e^{t\varepsilon_i} \leqslant e^{t^2/2}, \qquad \forall t \in \mathbb{R}.$$

This gives,

$$\mathbb{E}_\varepsilon \exp\left(\lambda \frac{\varepsilon_i f(x_i)}{n}\right) \leqslant \exp\left(\frac{\lambda^2 f(x_i)^2}{2n^2}\right).$$

Putting it back gives,

$$\mathbb{E}_\varepsilon e^{\lambda Z_f} \leqslant \exp\left(\sum_{i=1}^{n} \frac{\lambda^2 f(x_i)^2}{2n^2}\right) = \exp\left(\frac{\lambda^2}{2n} \cdot \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2\right) \overset{(i)}{\leqslant} \exp\left(\frac{\lambda^2 r^2}{2n}\right),$$

where (i) uses the given assumption. Thus each Z_f is centered sub-Gaussian with variance proxy r^2/n. Now, using the log-sum-exp trick to bound the maximum (same as used in Problem 1b3 above), for any $\lambda > 0$,

$$\max_{f \in \mathcal{F}} Z_f \leqslant \frac{1}{\lambda} \log\left(\sum_{f \in \mathcal{F}} e^{\lambda Z_f}\right).$$

Since the log is concave, taking expectations and using Jensen gives

$$\mathbb{E}_\varepsilon\left[\max_{f \in \mathcal{F}} Z_f\right] \leqslant \frac{1}{\lambda} \log\left(\sum_{f \in \mathcal{F}} \mathbb{E}_\varepsilon\left[e^{\lambda Z_f}\right]\right).$$

Define $M := |\mathcal{F}|$. Using the sub-Gaussian MGF bound for Z_f derived earlier gives:

$$\sum_{f \in \mathcal{F}} \mathbb{E}_\varepsilon\left[e^{\lambda Z_f}\right] \leqslant \sum_{f \in \mathcal{F}} \exp\left(\frac{\lambda^2 r^2}{2n}\right) = M \exp\left(\frac{\lambda^2 r^2}{2n}\right).$$

This gives

$$\mathbb{E}_\varepsilon\left[\max_{f \in \mathcal{F}} Z_f\right] \leqslant \frac{1}{\lambda} \log\left(M e^{\lambda^2 r^2/(2n)}\right) = \frac{\log M}{\lambda} + \frac{\lambda r^2}{2n}.$$

Now, optimize in λ by minimizing $f(\lambda) = \dfrac{\log M}{\lambda} + \dfrac{\lambda r^2}{2n}$, which gives the optimal choice is $\lambda^* = \dfrac{\sqrt{2n \log M}}{r}$. Substituting λ^*,

$$\frac{\log M}{\lambda} + \frac{\lambda r^2}{2n} = \frac{r \log M}{\sqrt{2n \log M}} + \frac{r\sqrt{2n \log M}}{2n} = r\sqrt{\frac{2 \log M}{n}}.$$

Hence, using $|\mathcal{F}| = M$, we get,

$$\widehat{\mathfrak{R}}_n(\mathcal{F}; x_1, \ldots, x_n) = \mathbb{E}_\varepsilon\left[\max_{f \in \mathcal{F}} Z_f\right] \leqslant r\sqrt{\frac{2 \log |\mathcal{F}|}{n}}. \qquad \square$$

(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

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \dots, x_n) \leqslant \frac{R}{\sqrt{n}}.$$

Solution:
By the definition of empirical Rademacher complexity,

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \dots, x_n) = \mathbb{E}_\varepsilon\left[\sup_{\|w\|_2\leqslant R} \frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle w, x_i\rangle\right].$$

Using the linearity of the inner product:

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \dots, x_n) = \frac{1}{n}\,\mathbb{E}_\varepsilon\left[\sup_{\|w\|_2\leqslant R}\left\langle w, \sum_{i=1}^n \varepsilon_i x_i\right\rangle\right].$$

Now, the Cauchy-Schwarz inequality gives

$$\sup_{\|w\|_2\leqslant R}\langle w, v\rangle \leqslant \sup_{\|w\|_2\leqslant R}\|w\|_2\|v\|_2 \leqslant R\|v\|_2, \qquad \text{for all } v \in \mathbb{R}^d.$$

The equality is attained by choosing $w = R\dfrac{v}{\|v\|_2}$ when $v \neq 0$, and for $v = 0$, the equality is attained trivially. Therefore, putting the supremum back gives,

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \dots, x_n) = \frac{R}{n}\,\mathbb{E}_\varepsilon\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2. \qquad \square$$

For the upper bound, applying Jensen to the concave map $u \mapsto \sqrt{u}$ gives:

$$\mathbb{E}_\varepsilon\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2 \leqslant \left(\mathbb{E}_\varepsilon\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2^2\right)^{1/2},$$

where

$$\mathbb{E}_\varepsilon\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2^2 = \mathbb{E}_\varepsilon\left\langle\sum_{i=1}^n \varepsilon_i x_i, \sum_{j=1}^n \varepsilon_j x_j\right\rangle = \sum_{i,j=1}^n \mathbb{E}_\varepsilon[\varepsilon_i\varepsilon_j]\langle x_i, x_j\rangle.$$

Since the Rademacher variables are independent and centered,

$$\mathbb{E}_\varepsilon[\varepsilon_i\varepsilon_j] = 0 \quad (i \neq j), \quad \text{and} \quad \mathbb{E}_\varepsilon[\varepsilon_i^2] = 1.$$

So all the cross terms vanish, and we get

$$\mathbb{E}_\varepsilon\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2^2 = \sum_{i=1}^n \|x_i\|_2^2.$$

Hence, substituting everything back

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \dots, x_n) \leqslant \frac{R}{n}\left(\mathbb{E}_\varepsilon\left\|\sum_{i=1}^n \varepsilon_i x_i\right\|_2^2\right)^{1/2} = \frac{R}{n}\left(\sum_{i=1}^n \|x_i\|_2^2\right)^{1/2}. \qquad \square$$

Finally, if $\|x_i\|_2 \leqslant 1$ for all i, then

$$\sum_{i=1}^{n} \|x_i\|_2^2 \leqslant n.$$

Hence, we obtain

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \ldots, x_n) \leqslant \frac{R}{n}(n)^{1/2} = \frac{R}{\sqrt{n}}. \qquad \square$$

(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

$$\widehat{\mathfrak{R}}_n(\mathcal{G}_R; x_1, \ldots, x_n) \leqslant CR\sqrt{\frac{\log d}{n}}.$$

[Bonus] Solution:
By the definition of Rademacher complexity and linearity of inner product:

$$\widehat{\mathfrak{R}}_n(\mathcal{G}_R; x_1, \ldots, x_n) = \frac{1}{n}\,\mathbb{E}_\varepsilon\left[\sup_{\|w\|_1 \leqslant R} \left\langle w, \sum_{i=1}^{n} \varepsilon_i x_i \right\rangle\right].$$

Using Hölder's inequality for the pair (ℓ_1, ℓ_∞) gives,

$$\sup_{\|w\|_1 \leqslant R} \langle w, v\rangle \leqslant \sup_{\|w\|_1 \leqslant R} \|w\|_1 \|v\|_\infty \leqslant R\|v\|_\infty, \qquad \text{for all } v \in \mathbb{R}^d.$$

The equality is attained by choosing $w = R\,\mathrm{sign}(v_j)e_j$, where j is such that $|v_j| = \|v\|_\infty$, and e_j is the j-th basis vector for $\mathbb{R}^d$. Therefore, putting the supremum back yields,

$$\widehat{\mathfrak{R}}_n(\mathcal{G}_R; x_1, \ldots, x_n) = \frac{R}{n}\,\mathbb{E}_\varepsilon \left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|_\infty .$$

Now, define the random vector

$$S := \sum_{i=1}^{n} \varepsilon_i x_i \in \mathbb{R}^d,$$

and write its coordinates as

$$S_j = \sum_{i=1}^{n} \varepsilon_i x_{ij}, \qquad j = 1, \ldots, d.$$

Fix j. Since the ε_i's are independent Rademacher variables, $\mathbb{E}_\varepsilon S_j = 0$. Moreover,

$$\mathbb{E}_\varepsilon e^{\lambda S_j} = \prod_{i=1}^{n} \mathbb{E}_\varepsilon \exp\left(\lambda \varepsilon_i x_{ij}\right).$$

A Rademacher random variable ε_i is sub-Gaussian and satisfies

$$\mathbb{E}_\varepsilon e^{t\varepsilon_i} \leqslant e^{t^2/2}, \qquad \forall t \in \mathbb{R}.$$

This gives,

$$\mathbb{E}_\varepsilon \exp\left(\lambda \varepsilon_i x_{ij}\right) \leqslant \exp\left(\frac{\lambda^2 x_{ij}^2}{2}\right) \leqslant \exp\left(\frac{\lambda^2}{2}\right),$$

because $|x_{ij}| \leqslant \|x_i\|_\infty \leqslant 1$. Putting it back gives,

$$\mathbb{E}_\varepsilon e^{\lambda S_j} \leqslant \exp\left(\sum_{i=1}^n \frac{\lambda^2}{2}\right) = \exp\left(\frac{n\lambda^2}{2}\right).$$

Thus, each S_j is centered sub-Gaussian with variance proxy n, i.e., $\|S_j\|_{\psi_2} \lesssim \sqrt{n}$ for all j. Now, define $2d$ random variables

$$Z_1, \ldots, Z_{2d} \coloneqq S_1, \ldots, S_d, -S_1, \ldots, -S_d,$$

which are all centered sub-Gaussian with $\|Z_k\|_{\psi_2} \lesssim \sqrt{n}$, and

$$\|S\|_\infty = \max_{1 \leqslant j \leqslant d} |S_j| = \max_{1 \leqslant k \leqslant 2d} Z_k.$$

Therefore, using the Finite maximal inequality (Lemma 4.1 from Lecture 17):

$$\mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|_\infty = \mathbb{E}_\varepsilon \|S\|_\infty \leqslant C\sqrt{n}\sqrt{\log(2d)} \leqslant C\sqrt{n \log d},$$

for some absolute constant $C > 0$. Hence,

$$\widehat{\mathfrak{R}}_n(\mathcal{G}_R; x_1, \ldots, x_n) = \frac{R}{n} \mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|_\infty \leqslant \frac{R}{n} \cdot C\sqrt{n \log d} = CR\sqrt{\frac{\log d}{n}}. \qquad \square$$

3 Practice with VC dimension

Recall that a class $\mathcal{H}$ of Boolean functions on a domain Ω *shatters* a finite set $\{x_1, \ldots, x_m\} \subset \Omega$ 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

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

Solution:
Take any single point $x \in \mathbb{R}$. We can realize both possible labelings of $\{x\}$:

- label 0: choose $a < x$, then $\mathbf{1}_{(-\infty,a]}(x) = 0$;
- label 1: choose $a \geqslant x$, then $\mathbf{1}_{(-\infty,a]}(x) = 1$.

So one point can be shattered, and thus $\mathrm{vc}(\mathcal{H}_{\text{half}}) \geqslant 1$. Now, take any two points in $\mathbb{R}$ and order them $x_1 < x_2$. The labeling $(x_1 = 0, x_2 = 1)$ is impossible to realize using $\mathbf{1}_{(-\infty,a]}$ because if $x_2 \in (-\infty, a]$, then automatically $x_1 \in (-\infty, a]$ as well since $x_1 < x_2$. Therefore, no two-point set can be shattered, and thus $\mathrm{vc}(\mathcal{H}_{\text{half}}) \leqslant 1$. Hence, combining both

$$\mathrm{vc}(\mathcal{H}_{\text{half}}) = 1. \qquad \square$$

(a2) the class of intervals

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

Solution:
Take any two points $x_1 < x_2$ in $\mathbb{R}$. We can realize all four labelings:

- $(0,0)$: choose an interval disjoint from both points, for example $\mathbf{1}_{[x_2+1, x_2+1]}$;
- $(1,0)$: choose $\mathbf{1}_{[x_1,x_1]}$;
- $(0,1)$: choose $\mathbf{1}_{[x_2,x_2]}$;
- $(1,1)$: choose $\mathbf{1}_{[x_1,x_2]}$.

So some two-point set is shattered, and thus $\mathrm{vc}(\mathcal{H}_{\text{int}}) \geqslant 2$. Now take any three points $x_1 < x_2 < x_3$ in $\mathbb{R}$. Consider the labeling:

$$(x_1 = 1, \ x_2 = 0, \ x_3 = 1).$$

This cannot be realized by any interval $[a, b]$ because if the interval contains both x_1 and x_3, then automatically x_2 belongs to the interval since $x_1 < x_2 < x_3$. So no three-point set can be shattered, and thus $\mathrm{vc}(\mathcal{H}_{\text{int}}) \leqslant 2$. Hence, combining both

$$\mathrm{vc}(\mathcal{H}_{\text{int}}) = 2. \qquad \square$$

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

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

Solution:
Take $2k$ distinct points in increasing order:

$$x_1 < x_2 < \ldots < x_{2k}.$$

Group them into k consecutive pairs:

$$\{x_1, x_2\}, \{x_3, x_4\}, \ldots, \{x_{2k-1}, x_{2k}\}.$$

For each pair $\{x_{2j-1}, x_{2j}\}$, part (a2) shows that the class of single intervals can realize all four labelings on two points:

$$(0,0), (1,0), (0,1), (1,1).$$

So for each $j = 1, \ldots, k$ there exists an interval I_j whose indicator produces exactly the desired labels on the two points $\{x_{2j-1}, x_{2j}\}$. Now define

$$I_{\text{union}} := \bigcup_{j=1}^{k} I_j.$$

Because each I_j controls the labels on its own pair, and we are allowed to use up to k intervals, this realizes the prescribed labeling on all $2k$ points. Also $\mathbf{1}_{I_{\text{union}}} \in \mathcal{H}_k$ because I_{union} is a union of at most k intervals. Therefore the set $\{x_1, \ldots, x_{2k}\}$ is shattered, and hence

$$\text{vc}(\mathcal{H}_k) \geqslant 2k.$$

Now take any $2k + 1$ distinct points in increasing order:

$$x_1 < x_2 < \ldots < x_{2k} < x_{2k+1}.$$

Consider the alternating labeling

$$1, 0, 1, 0, \ldots, 1, 0, 1.$$

This labeling has exactly $(k + 1)$ separated runs of 1's:

$$\{x_1\}, \{x_3\}, \ldots, \{x_{2k+1}\}.$$

A single interval can cover at most one such run if it is to avoid including a 0-labeled point. If one interval contained two distinct points x_{2r-1} and x_{2s-1} with $r < s$, then it would contain at least one 0-labeled point. So one interval cannot realize two separated 1-runs. Thus, to realize this alternating labeling, one would need at least $k + 1$ intervals. But every function in $\mathcal{H}_k$ is a union of at most k intervals. Therefore, this labeling cannot be realized by $\mathcal{H}_k$. Hence, no set of $2k + 1$ points can be shattered, so $\text{vc}(\mathcal{H}_k) \leqslant 2k$. Finally, combining both

$$\text{vc}(\mathcal{H}_k) = 2k. \qquad \square$$

(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 $\mathbb{R}^2$:

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

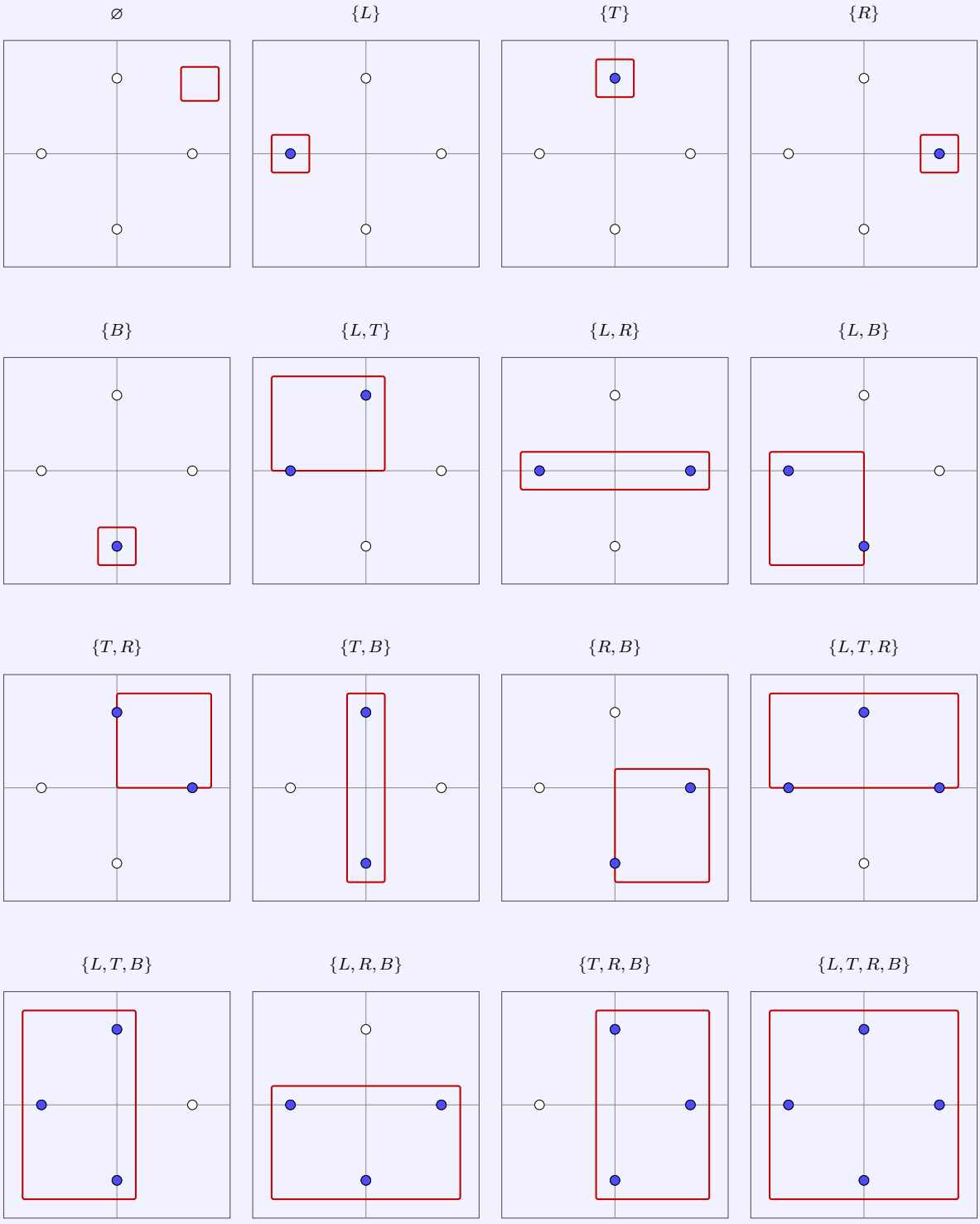
Show that the class of indicators of axis-aligned rectangles has VC dimension

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

Solution:

Four points in diamond position: $L = (-1, 0),\ T = (0, 1),\ R = (1, 0),\ B = (0, -1)$

Each panel shows one labeling and an axis-aligned rectangle realizing it.



Thus,

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

Now, take any 5 points in $\mathbb{R}^2$. Among them, choose:

- a point x_L with smallest x-coordinate,
- a point x_R with largest x-coordinate,
- a point x_B with smallest y-coordinate, and
- a point x_T with largest y-coordinate.

These are the leftmost, rightmost, bottommost, and topmost points. Since there are at most 4 such extreme points, there exists at least one point x_* among those 5 points that is not among those extremes. Now consider the labeling that assigns:

- label 1 to each of the extreme points x_L, x_R, x_B, x_T, and
- label 0 to the remaining point x_*.

Any axis-aligned rectangle containing all four extreme points must have:

$$a \leqslant x_L^{(1)} \leqslant x_R^{(1)} \leqslant b \qquad \text{and} \qquad c \leqslant x_B^{(2)} \leqslant x_T^{(2)} \leqslant d,$$

so its horizontal span must include the minimum and maximum x-coordinates of the sample, and its vertical span must include the minimum and maximum y-coordinates of the sample. Therefore, it must contain the entire bounding box determined by the sample extremes. Since x_* is one of the sample points, it lies inside the bounding box, and hence must also belong to the rectangle. Thus any rectangle that labels all four extreme points by 1 must also label x_* by 1, so the labeling above cannot be realized by $\mathcal{R}$. Therefore no set of 5 points can be shattered by axis-aligned rectangles, and so

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

Hence, combining both

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

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

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

[Bonus] Solution:
Take any $m \geqslant 1$. Consider m distinct points

$$x_1, \ldots, x_m \in \mathbb{R}^2$$

lying on a circle, for example, equally spaced on a circle of radius $\delta > 0$. Since these points lie on a circle, they are in strictly convex position: each x_j is a vertex of a convex polygon

$$\mathrm{conv}(\{x_1, \ldots, x_m\}).$$

Choose any labeling for those m points and let $E \subset \{x_1, \ldots, x_m\}$ be the points labeled 1. Now, define the convex hull

$$V := \mathrm{conv}(E).$$

This is a convex subset of $\mathbb{R}^2$, so $\mathbf{1}_V \in \mathcal{C}_{\mathrm{conv}}$. The claim is V contains exactly the points of E among the sample points $\{x_1, \ldots, x_m\}$. We have two cases:

- if $x_i \in E$ then clearly $x_i \in V$, because $E \subset \text{conv}(E)$.
- if $x_j \notin E$ then $x_j \notin V$. Assume towards contradiction that $x_j \in \text{conv}(E)$. By the definition of convex hull, there exist points $y_1, \ldots, y_r \in E$ and coefficients $\lambda_1, \ldots, \lambda_r \geqslant 0$ such that

$$x_j = \sum_{i=1}^{r} \lambda_i y_i, \qquad \text{with} \qquad \sum_{i=1}^{r} \lambda_i = 1.$$

  Since $x_j \notin E$, all points $y_1, \ldots, y_r$ are different from x_j. Thus, this expresses x_j as a convex combination of other sample points. But this is not possible, because each of the points $x_1, \ldots, x_m$ is a vertex of the convex polygon $\text{conv}(\{x_1, \ldots, x_m\})$, i.e., an extreme point of a convex polygon, and an extreme point cannot be written as a convex combination of the other points of the set. This is a contradiction, so $x_j \notin \text{conv}(E)$, i.e., $x_j \notin V$.

Thus, we have

$$V \cap \{x_1, \ldots, x_m\} = E,$$

which means that $\mathbf{1}_V$ realizes exactly the chosen labeling. Since the labeling was arbitrary, every labeling of $\{x_1, \ldots, x_m\}$ can be realized by some convex set. Therefore $\{x_1, \ldots, x_m\}$ is shattered by $\mathcal{C}_{\text{conv}}$. Because this works for every $m \geqslant 1$, the class $\mathcal{C}_{\text{conv}}$ shatters arbitrarily large finite sets. Hence,

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

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

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

[Bonus] Solution:
Let $\mathcal{B}_d$ be the class of indicators of Euclidean balls in $\mathbb{R}^d$. We want to show:

$$\text{vc}(\mathcal{B}_d) = d + 1.$$

Consider $d+1$ points in $\mathbb{R}^d$:

$$S = \{0, e_1, \ldots, e_d\}$$

where e_i's are the basis vectors in $\mathbb{R}^d$. We want to show that for any arbitrary subset $A \subset S$, we can find a ball $B(c, r)$ such that $S \cap B(c, r) = A$. Let's construct the center $c = (c_1, \ldots, c_d)$ and radius r based on the subset A:

1. The origin is in the subset $(0 \in A)$:
   - set $c_i = 1/2$ if $e_i \in A$ and $c_i = 0$ if $e_i \notin A$, and set $r^2 = \|c\|_2^2$.
   - Check for 0 : $\|0 - c\|_2^2 = \|c\|_2^2 \leqslant r^2$, so $0 \in B(c, r)$.
   - Check for $e_i \in A$: $\|e_i - c\|_2^2 = \left(1 - \frac{1}{2}\right)^2 + \sum_{j \neq i} c_j^2 = \|c\|_2^2 \leqslant r^2$, so $e_i \in B(c, r)$.
   - Check for $e_i \notin A$: $\|e_i - c\|_2^2 = (1-0)^2 + \sum_{j \neq i} c_j^2 = 1 + \|c\|_2^2 > r^2$, so $e_i \notin B(c, r)$.
2. The origin is not in the subset $(0 \notin A)$ and A is non-empty $(A \neq \varnothing)$:
   - set $c_i = 1$ if $e_i \in A$ and $c_i = 0$ if $e_i \notin A$, and set $r^2 = |A| - 1$.
   - Check for 0 : $\|0 - c\|_2^2 = \|c\|_2^2 = |A|$. Since $|A| > |A| - 1 = r^2$, $0 \notin B(c, r)$.

- Check for $e_i \in A$: $\|e_i - c\|_2^2 = (1-1)^2 + \sum_{j \neq i} c_j^2 = |A| - 1 \leqslant r^2$, so $e_i \in B(c,r)$.
- Check for $e_i \notin A$: $\|e_i - c\|_2^2 = (1-0)^2 + \sum_{j \neq i} c_j^2 = 1 + \|c\|_2^2 = 1 + |A| > r^2$, so $e_i \notin B(c,r)$.

3. The subset is empty ($A = \varnothing$):

- Set $c = (2, 2, \ldots, 2)$ and $r = 0$.
- For any $x \in S$, $\|x - c\|_2^2 > 0$, so the ball captures none of them.

Thus, the labeling of 1 for the subset A and 0 everywhere else can be realized by some $\mathbf{1}_{B(c,r)}$. Since A was arbitrary, every labeling of S can be realized by some Euclidean ball in $\mathbb{R}^d$, which means S is shattered by $\mathcal{B}_d$. Thus, we get

$$\mathrm{vc}(\mathcal{B}_d) \geqslant d + 1.$$

Now, take any $d+2$ points $\{x_1, \ldots, x_{d+2}\} \in \mathbb{R}^d$. By Radon's theorem, there exist two disjoint nonempty subsets S_1, S_2 of $\{x_1, \ldots, x_{d+2}\}$ such that

$$\mathrm{conv}(S_1) \cap \mathrm{conv}(S_2) \neq \varnothing.$$

Choose x^* in the intersection, i.e.,

$$x^* \in \mathrm{conv}(S_1) \cap \mathrm{conv}(S_2).$$

Then there exist coefficients $\alpha_i \geqslant 0$ for $x_i \in S_1$ and $\beta_j \geqslant 0$ for $x_j \in S_2$ such that

$$\sum_{i\,:\,x_i \in S_1} \alpha_i = 1, \qquad \sum_{j\,:\,x_j \in S_2} \beta_j = 1,$$

and

$$x^* = \sum_{x_i \in S_1} \alpha_i x_i = \sum_{x_j \in S_2} \beta_j x_j.$$

Assume towards contradiction that the set $\{x_1, \ldots, x_{d+2}\}$ is shattered by some Euclidean balls. Then there exists a ball

$$B_1 = B(c,r) = \{x \in \mathbb{R}^d : \|x - c\|_2^2 \leqslant r^2\}$$

such that

$$x_i \in B_1 \text{ for all } x_i \in S_1, \qquad \text{and} \qquad x_j \notin B_1 \text{ for all } x_j \in S_2.$$

For each $x_i \in S_1$,

$$\|x_i - c\|_2^2 \leqslant r^2 \iff \|x_i\|_2^2 - 2\langle c, x_i \rangle + \|c\|_2^2 \leqslant r^2.$$

Multiply by α_i and sum over S_1:

$$\sum_{x_i \in S_1} \alpha_i \|x_i\|_2^2 - 2\left\langle c, \sum_{x_i \in S_1} \alpha_i x_i \right\rangle + \|c\|_2^2 \cdot \sum_{x_i \in S_1} \alpha_i \leqslant r^2 \cdot \sum_{x_i \in S_1} \alpha_i.$$

Using $\sum_{x_i \in S_1} \alpha_i = 1$ and $\sum_{x_i \in S_1} \alpha_i x_i = x^*$, we get

$$\sum_{x_i \in S_1} \alpha_i \|x_i\|_2^2 - 2\langle c, x^* \rangle + \|c\|_2^2 \leqslant r^2. \tag{$*$}$$

Similarly, for each $x_j \in S_2$,

$$\|x_j - c\|_2^2 > r^2 \iff \|x_j\|_2^2 - 2\langle c, x_j \rangle + \|c\|_2^2 > r^2.$$

Multiply by β_j and sum over S_2. Using $\sum_{x_j \in S_2} \beta_j = 1$ and $\sum_{x_j \in S_2} \beta_j x_j = x^*$, we get

$$\sum_{x_j \in S_2} \beta_j \|x_j\|_2^2 - 2\langle c, x^* \rangle + \|c\|_2^2 > r^2. \qquad (**)$$

Comparing inequalities $(*)$ and $(**)$ gives

$$\sum_{x_i \in S_1} \alpha_i \|x_i\|_2^2 < \sum_{x_j \in S_2} \beta_j \|x_j\|_2^2. \qquad (1)$$

But shattering also means that the opposite labeling is realizable. So there exists another ball B_2 such that

$$x_j \in B_2 \text{ for all } x_j \in S_2, \qquad \text{and} \qquad x_i \notin B_2 \text{ for all } x_i \in S_1.$$

Repeating the same argument with S_1 and S_2 interchanged yields

$$\sum_{x_j \in S_2} \beta_j \|x_j\|_2^2 < \sum_{x_i \in S_1} \alpha_i \|x_i\|_2^2. \qquad (2)$$

The inequalities (1) and (2) contradict each other. Therefore, $\{x_1, \ldots, x_{d+2}\}$ cannot be shattered by Euclidean balls. Thus, no set of $d+2$ points can be shattered, and so

$$\mathrm{vc}(\mathcal{B}_d) \leqslant d + 1.$$

Hence, combining the lower and upper bounds,

$$\mathrm{vc}(\mathcal{B}_d) = d + 1. \qquad \square$$

(c2) Use the Sauer–Shelah lemma to show that if $\mathcal{H}$ is any Boolean class with $\mathrm{vc}(\mathcal{H}) = v$, 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.

[Bonus] Solution:
Using the Sauer-Shelah lemma,

$$\Pi_{\mathcal{H}}(n) \leqslant \sum_{j=0}^{v} \binom{n}{j}.$$

Now, for $n \geqslant v$, we have $0 \leqslant \dfrac{v}{n} \leqslant 1$. Therefore,

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

Using the Binomial theorem,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k,$$

we get

$$\left(\frac{v}{n}\right)^v \sum_{j=0}^{v} \binom{n}{j} \leqslant \left(1 + \frac{v}{n}\right)^n \overset{(i)}{\leqslant} e^v,$$

where (i) uses the identity $(1 + x/n)^n \leqslant e^x$ for all $x \geqslant 0$. Hence, if $\mathcal{H}$ is any Boolean class with $\mathrm{vc}(\mathcal{H}) = v$, then on any n-point set it induces at most

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

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

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

[Bonus] Solution:
From part (b1), we know that for axis-aligned rectangles

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

The subsets cut out by axis-aligned rectangles are exactly the labelings induced by the class $\mathcal{R}$ on the n-point set. Hence, applying part (c2) with $v = 4$, we obtain 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

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

4 Practice with statistical learning theory

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

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

Solution:
Let

$$\Delta := \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|.$$

By definition of Δ, for every $h \in \mathcal{H}$,

$$R(h) \leqslant R_n(h) + \Delta \qquad \text{and} \qquad R_n(h) \leqslant R(h) + \Delta. \tag{$*$}$$

Now apply the first inequality in $(*)$ with $h = \widehat{h}$:

$$R(\widehat{h}) \leqslant R_n(\widehat{h}) + \Delta.$$

Since $\widehat{h}$ minimizes the empirical risk over $\mathcal{H}$, $R_n(\widehat{h}) \leqslant R_n(h^\star)$, so

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

Similarly, applying the second inequality of $(*)$ with $h = h^\star$:

$$R_n(h^\star) \leqslant R(h^\star) + \Delta.$$

Combining the last two inequalities gives

$$R(\widehat{h}) \leqslant R(h^\star) + 2\Delta.$$

Subtracting $R(h^\star)$ from both sides yields

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

(b) **Symmetrization.** Let

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

Show that

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

Solution:
Let $f_h(x,y) := \ell(h(x),y)$ for $h \in \mathcal{H}$, so that

$$\mathcal{L}_{\mathcal{H}} = \{f_h : h \in \mathcal{H}\}.$$

Then

$$R_n(h) = \frac{1}{n}\sum_{i=1}^n f_h(X_i,Y_i), \qquad R(h) = \mathbb{E}\, f_h(X,Y),$$

and

$$\sup_{h\in\mathcal{H}} |R_n(h) - R(h)| = \sup_{f\in\mathcal{L}_{\mathcal{H}}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i,Y_i) - \mathbb{E}f(X,Y) \right|.$$

Now let $(X_1',Y_1'),\dots,(X_n',Y_n')$ be an independent copy of $(X_1,Y_1),\dots,(X_n,Y_n)$, independent of everything else. Since (X_i',Y_i') has the same law as (X_i,Y_i),

$$R(h) = \mathbb{E}f_h(X_i,Y_i) = \mathbb{E}f_h(X_i',Y_i')$$

for every i. Therefore

$$\mathbb{E}\sup_{h\in\mathcal{H}} |R_n(h) - R(h)| = \mathbb{E}\sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^n \big(f_h(X_i,Y_i) - \mathbb{E}f_h(X_i',Y_i')\big) \right|.$$

Now condition on $(X_1,Y_1),\dots,(X_n,Y_n)$, the map

$$(y_1,\dots,y_n) \mapsto \sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^n \big(f_h(X_i,Y_i) - y_i\big) \right|$$

is convex, since it is a supremum of convex functions of $(y_1,\dots,y_n)$. Thus, by Jensen's inequality,

$$\mathbb{E}\sup_{h\in\mathcal{H}} |R_n(h) - R(h)| \leqslant \mathbb{E}\sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^n \big(f_h(X_i,Y_i) - f_h(X_i',Y_i')\big) \right|.$$

Now introduce i.i.d. Rademacher variables $\varepsilon_1,\dots,\varepsilon_n$, independent of everything else. Since the joint law of $\big((X_i,Y_i),(X_i',Y_i')\big)$ is invariant under swapping the two coordinates, we have

$$\left(\frac{1}{n}\sum_{i=1}^n \big(f_h(X_i,Y_i) - f_h(X_i',Y_i')\big) \right)_{h\in\mathcal{H}} \overset{d}{=} \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(f_h(X_i,Y_i) - f_h(X_i',Y_i')\big) \right)_{h\in\mathcal{H}}.$$

Therefore

$$\mathbb{E}\sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^n \big(f_h(X_i,Y_i) - f_h(X_i',Y_i')\big) \right| = \mathbb{E}\sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(f_h(X_i,Y_i) - f_h(X_i',Y_i')\big) \right|.$$

Now apply the triangle inequality:

$$\left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(f_h(X_i,Y_i) - f_h(X_i',Y_i')\big) \right| \leqslant \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f_h(X_i,Y_i) \right| + \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f_h(X_i',Y_i') \right|.$$

Taking the supremum over $h \in \mathcal{H}$, then expectation, gives

$$\mathbb{E} \sup_{h\in\mathcal{H}} |R_n(h) - R(h)| \leqslant \mathbb{E} \sup_{h\in\mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_h(X_i, Y_i) \right| + \mathbb{E} \sup_{h\in\mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_h(X_i', Y_i') \right|.$$

The two terms have the same distribution, so

$$\mathbb{E} \sup_{h\in\mathcal{H}} |R_n(h) - R(h)| \leqslant 2\, \mathbb{E} \sup_{h\in\mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_h(X_i, Y_i) \right|.$$

Finally, by the definition of empirical Rademacher complexity of the (non-symmetric) loss class,

$$\widehat{\mathfrak{R}}_n^{\mathrm{abs}}\big(\mathcal{L}_{\mathcal{H}}; (X_1, Y_1), \ldots, (X_n, Y_n)\big) = \mathbb{E}_{\varepsilon} \sup_{f\in\mathcal{L}_{\mathcal{H}}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i, Y_i) \right|,$$

so taking expectation over the sample yields

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

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

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

Solution:
Combining part (a) and part (b) gives

$$\mathbb{E}\big[R(\widehat{h}) - R(h^\star)\big] \leqslant 4\, \mathbb{E}\, \widehat{\mathfrak{R}}_n^{\mathrm{abs}}\big(\mathcal{L}_{\mathcal{H}}; (X_1, Y_1), \ldots, (X_n, Y_n)\big).$$

Define $\mathcal{L}_{\mathcal{H}}^{\pm} := \mathcal{L}_{\mathcal{H}} \cup (-\mathcal{L}_{\mathcal{H}})$. Since $\mathcal{H}$ is finite, $\mathcal{L}_{\mathcal{H}}^{\pm}$ is also finite and

$$|\mathcal{L}_{\mathcal{H}}^{\pm}| \leqslant 2|\mathcal{L}_{\mathcal{H}}| \leqslant 2|\mathcal{H}|.$$

Since the loss is bounded in $[0, 1]$, for every $f \in \mathcal{L}_{\mathcal{H}}^{\pm}$,

$$\|f\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} f(X_i, Y_i)^2 \leqslant 1.$$

Therefore applying the result from Problem 2(b) with $r = 1$ yields

$$\widehat{\mathfrak{R}}_n^{\mathrm{abs}}\big(\mathcal{L}_{\mathcal{H}}; S\big) = \widehat{\mathfrak{R}}_n(\mathcal{L}_{\mathcal{H}}^{\pm}; S) \leqslant \sqrt{\frac{2 \log |\mathcal{L}_{\mathcal{H}}^{\pm}|}{n}} \leqslant \sqrt{\frac{2 \log(2|\mathcal{H}|)}{n}},$$

where $S = (X_1, Y_1), \ldots, (X_n, Y_n)$ are the samples. Taking expectation preserves this bound, so

$$\mathbb{E}\, \widehat{\mathfrak{R}}_n^{\mathrm{abs}}\big(\mathcal{L}_{\mathcal{H}}; (X_1, Y_1), \ldots, (X_n, Y_n)\big) \leqslant \sqrt{\frac{2 \log(2|\mathcal{H}|)}{n}}.$$

Combining the inequalities above, we get

$$\mathbb{E}\big[R(\widehat{h}) - R(h^\star)\big] \leqslant 4 \sqrt{\frac{2 \log(2|\mathcal{H}|)}{n}}.$$

Thus

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

for $|\mathcal{H}| \geqslant 2$ and an absolute constant $C > 0$. $\square$

(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

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

Solution:

Let

$$\mathcal{L}_{\mathcal{H}} = \{(x,y) \mapsto \mathbf{1}\{h(x) \neq y\} : h \in \mathcal{H}\}$$

be the 0-1 loss class. Combining part (a) and part (b) gives

$$\mathbb{E}\big[R(\widehat{h}) - R(h^\star)\big] \leqslant 4\,\mathbb{E}\,\widehat{\mathfrak{R}}_n^{\mathrm{abs}}\big(\mathcal{L}_{\mathcal{H}}; (X_1,Y_1),\dots,(X_n,Y_n)\big). \tag{1}$$

Now fix a sample

$$(x_1,y_1),\dots,(x_n,y_n).$$

For each $h \in \mathcal{H}$, its prediction vector on the sample is

$$(h(x_1),\dots,h(x_n)) \in \{0,1\}^n,$$

and its loss vector is

$$\big(\mathbf{1}\{h(x_1) \neq y_1\},\dots,\mathbf{1}\{h(x_n) \neq y_n\}\big) \in \{0,1\}^n.$$

This map from prediction vector to loss vector is a bijection of $\{0,1\}^n$, obtained by coordinatewise XOR with the fixed label vector $(y_1,\dots,y_n)$. Therefore the number of distinct loss vectors induced by $\mathcal{L}_{\mathcal{H}}$ on the sample is exactly the same as the number of distinct prediction vectors induced by $\mathcal{H}$ on $x_1,\dots,x_n$. Let N_n denote this number of distinct loss vectors. Since $\mathrm{vc}(\mathcal{H}) = v < \infty$, using the result from Problem 3 part (c2), we get

$$N_n \leqslant \sum_{j=0}^{v}\binom{n}{j} \leqslant \left(\frac{en}{v}\right)^v. \tag{2}$$

On the fixed sample, let $\mathcal{F}_{\mathrm{res}} \subset [0,1]^n$ be the set of distinct loss vectors induced by $\mathcal{L}_{\mathcal{H}}$. Then $|\mathcal{F}_{\mathrm{res}}| = N_n$. Define $\mathcal{F}_{\mathrm{res}}^{\pm} := \mathcal{F}_{\mathrm{res}} \cup (-\mathcal{F}_{\mathrm{res}})$, then $|\mathcal{F}_{\mathrm{res}}^{\pm}| \leqslant 2N_n$. The Rademacher complexity is

$$\widehat{\mathfrak{R}}_n^{\mathrm{abs}}\big(\mathcal{L}_{\mathcal{H}}; S\big) = \mathbb{E}_\varepsilon \sup_{u \in \mathcal{F}_{\mathrm{res}}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i u_i\right| = \mathbb{E}_\varepsilon \sup_{w \in \mathcal{F}_{\mathrm{res}}^{\pm}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i w_i,$$

where $S = (x_1,y_1),\dots,(x_n,y_n)$ is the sample vector. Since $\mathcal{F}_{\mathrm{res}}^{\pm}$ is finite and each $w \in \mathcal{F}_{\mathrm{res}}^{\pm}$ satisfies $\frac{1}{n}\sum_{i=1}^n w_i^2 \leqslant 1$, applying the result from Problem 2(b) with $r = 1$ yields

$$\widehat{\mathfrak{R}}_n^{\mathrm{abs}}\big(\mathcal{L}_{\mathcal{H}}; (x_1,y_1),\dots,(x_n,y_n)\big) \leqslant \sqrt{\frac{2\log(2N_n)}{n}}.$$

Using (2), we obtain

$$\widehat{\mathfrak{R}}_n^{\mathrm{abs}}\big(\mathcal{L}_{\mathcal{H}}; (x_1,y_1),\dots,(x_n,y_n)\big) \leqslant \sqrt{\frac{2\{\log 2 + v\log(en/v)\}}{n}}.$$

Since this bound is deterministic, taking expectation gives

$$\mathbb{E}\,\widehat{\mathfrak{R}}_n^{\mathrm{abs}}\big(\mathcal{L}_{\mathcal{H}}; (X_1,Y_1),\dots,(X_n,Y_n)\big) \leqslant \sqrt{\frac{2\{\log 2 + v\log(en/v)\}}{n}}. \tag{3}$$

Finally, combining (1) and (3), we get

$$\mathbb{E}\big[R(\widehat{h}) - R(h^\star)\big] \leqslant 4\sqrt{\frac{2\{\log 2 + v\log(en/v)\}}{n}}.$$

Hence, for some absolute constant $C > 0$,

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

(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 $1 - \delta$,

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

[Bonus] Solution:
Let $(X_1', Y_1'), \ldots, (X_n', Y_n')$ be an independent ghost sample with the same law as $(X_i, Y_i)_{i=1}^n$, and define the ghost empirical risk

$$R_n'(h) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{h(X_i') \neq Y_i'\}.$$

Fix $t > 0$. Since $\mathbb{E}[R_n'(h) \mid (X_i, Y_i)_{i=1}^n] = R(h)$, a standard symmetrization argument gives

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}} |R_n(h) - R(h)| > t\right\} \leqslant 2\,\mathbb{P}\left\{\sup_{h\in\mathcal{H}} |R_n(h) - R_n'(h)| > \frac{t}{2}\right\}. \tag{1}$$

Now condition on the combined sample

$$Z := \big((X_1, Y_1), \ldots, (X_n, Y_n), (X_1', Y_1'), \ldots, (X_n', Y_n')\big).$$

For each $h \in \mathcal{H}$, define the two loss vectors

$$u(h) := \big(\mathbf{1}\{h(X_1) \neq Y_1\}, \ldots, \mathbf{1}\{h(X_n) \neq Y_n\}\big) \in \{0,1\}^n,$$

$$u'(h) := \big(\mathbf{1}\{h(X_1') \neq Y_1'\}, \ldots, \mathbf{1}\{h(X_n') \neq Y_n'\}\big) \in \{0,1\}^n.$$

Then $u(h)$ and $u'(h)$ are fixed once Z is fixed, and

$$R_n(h) - R_n'(h) = \frac{1}{n}\sum_{i=1}^n \big(u_i(h) - u_i'(h)\big).$$

Now, introduce i.i.d. Rademacher variables $\varepsilon_1, \ldots, \varepsilon_n$, independent of everything. Since the joint law of $\big((X_i, Y_i), (X_i', Y_i')\big)$ is invariant under swapping the two coordinates, we have

$$\left(\frac{1}{n}\sum_{i=1}^n \big(u_i(h) - u_i'(h)\big)\right)_{h\in\mathcal{H}} \overset{d}{=} \left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(u_i(h) - u_i'(h)\big)\right)_{h\in\mathcal{H}}.$$

Hence, conditional on Z, we get

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}} |R_n(h) - R_n'(h)| > \frac{t}{2} \;\middle|\; Z\right\} = \mathbb{P}_\varepsilon\left\{\sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(u_i(h) - u_i'(h)\big)\right| > \frac{t}{2}\right\}. \tag{2}$$

On the fixed combined sample, define the full $2n$-dimensional loss vector

$$w(h) := \big(u(h), u'(h)\big) \in \{0,1\}^{2n}.$$

The number of distinct such loss vectors induced by $\mathcal{L}_{\mathcal{H}}$ on the $2n$ points is finite. Since the loss vector is obtained from the prediction vector by coordinatewise XOR with the fixed labels, this number equals the number of distinct prediction vectors induced by $\mathcal{H}$ on the $2n$ sample points. Since $\mathrm{vc}(\mathcal{H}) = v$, Sauer–Shelah gives

$$N_{2n} \leqslant \sum_{j=0}^{v} \binom{2n}{j} \leqslant \left(\frac{2en}{v}\right)^{v}. \tag{3}$$

Now the difference vector

$$d(h) := u(h) - u'(h) \in \{-1,0,1\}^n$$

is completely determined by the full loss vector $w(h)$. Therefore the number of distinct difference vectors is at most N_{2n}. Thus the supremum in (2) is actually over a fixed finite class of size at most N_{2n}. Fix one such difference vector $d = (d_1, \ldots, d_n) \in \{-1,0,1\}^n$. Then

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i d_i$$

is an average of independent mean-zero bounded random variables with $|\varepsilon_i d_i| \leqslant 1$. So using Hoeffding's inequality, we get

$$\mathbb{P}_{\varepsilon}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i d_i\right| > \frac{t}{2}\right\} \leqslant 2e^{-nt^2/8}.$$

Now conditional on Z, applying the union bound over at most N_{2n} possible difference vectors,

$$\mathbb{P}_{\varepsilon}\left\{\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\big(u_i(h) - u_i'(h)\big)\right| > \frac{t}{2}\right\} \leqslant 2N_{2n}e^{-nt^2/8}.$$

Since this bound does not depend on Z, combining with (2) and taking expectation over Z yields

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}}|R_n(h) - R_n'(h)| > \frac{t}{2}\right\} \leqslant 2N_{2n}e^{-nt^2/8}.$$

Using (1) and (3), we get

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}}|R_n(h) - R(h)| > t\right\} \leqslant 4\left(\frac{2en}{v}\right)^{v} e^{-nt^2/8}.$$

Set $t = \varepsilon/2$. Then

$$\mathbb{P}\left\{\sup_{h\in\mathcal{H}}|R_n(h) - R(h)| > \frac{\varepsilon}{2}\right\} \leqslant 4\left(\frac{2en}{v}\right)^{v} e^{-n\varepsilon^2/32}.$$

Now choose n such that

$$4\left(\frac{2en}{v}\right)^{v} e^{-n\varepsilon^2/32} \leqslant \delta \implies n \geqslant \frac{v\log(2en/v) + \log(4/\delta)}{\varepsilon^2/32}.$$

Hence, there exists an absolute constant C such that if

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

then

$$\sup_{h\in\mathcal{H}}|R_n(h) - R(h)| \leqslant \frac{\varepsilon}{2}$$

with probability at least $1-\delta$. By part (a),

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

with probability at least $1-\delta$. $\square$

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
$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \big|\mu_n(g) - \mu(g)\big|\right] \leqslant \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{G}, \|\cdot\|_\infty, \varepsilon)}\, d\varepsilon.$$
- **Finite-class bound:** if $\mathcal{G}$ is a finite class of functions taking values in $[0,1]$, then
$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \big|\mu_n(g) - \mu(g)\big|\right] \leqslant C\sqrt{\frac{\log|\mathcal{G}|}{n}}.$$

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 $\varepsilon > 0$,

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

Solution:
For any $f, g \in \mathcal{F}$,

$$\begin{aligned}\big\|(f-T)^2 - (g-T)^2\big\|_\infty &= \big\|f^2 - 2fT + T^2 - g^2 + 2gT - T^2\big\|_\infty \\ &= \|(f-g)(f+g) - 2T(f-g)\|_\infty \\ &= \|(f-g)(f+g-2T)\|_\infty \\ &\leqslant \|f-g\|_\infty \|f+g-2T\|_\infty .\end{aligned}$$

Since f, g, T take values in $[0,1]$

$$|f(x) + g(x) - 2T(x)| \leqslant |f(x) - T(x)| + |g(x) - T(x)| \leqslant 2.$$

Hence, we get

$$\big\|(f-T)^2 - (g-T)^2\big\|_\infty \leqslant 2\,\|f-g\|_\infty . \qquad \square$$

Now, let $\{f_1, \ldots, f_m\}$ be the minimal $\varepsilon/2$-net of $\mathcal{F}$ in $\|\cdot\|_\infty$, then $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon/2) = m$. Consider the loss functions $\{(f_1 - T)^2, \ldots, (f_m - T)^2\} \subseteq \mathcal{L}_{\mathcal{F}}$. For any $(f-T)^2 \in \mathcal{L}_{\mathcal{F}}$, using the $\varepsilon/2$-net of $\mathcal{F}$, there exists $f_k \in \{f_1, \ldots, f_m\}$ such that

$$\|f - f_k\|_\infty \leqslant \varepsilon/2.$$

By the inequality proved above,

$$\big\|(f-T)^2 - (f_k - T)^2\big\|_\infty \leqslant 2\,\|f - f_k\|_\infty \leqslant 2 \cdot \varepsilon/2 = \varepsilon.$$

Thus $\{(f_1 - T)^2, \ldots, (f_m - T)^2\}$ is an ε-net of $\mathcal{L}_{\mathcal{F}}$. Hence, for every $\varepsilon > 0$,

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

(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.

Solution:
Using the excess-risk lemma stated in the problem,

$$R(\widehat{f}_n) - R(f^\star) \leqslant 2 \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|.$$

Taking the expectation gives

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant 2\,\mathbb{E}\sup_{f\in\mathcal{F}}|R_n(f) - R(f)|.$$

Now the loss class is defined as

$$\mathcal{L}_{\mathcal{F}} := \{x \mapsto (f(x) - T(x))^2 : \ f \in \mathcal{F}_{L,1}\}.$$

Then for each $f \in \mathcal{F}$,

$$R(f) = \mu(\ell_f), \qquad R_n(f) = \mu_n(\ell_f),$$

where $\ell_f(x) = (f(x) - T(x))^2$. Therefore

$$\sup_{f\in\mathcal{F}}|R_n(f) - R(f)| = \sup_{g\in\mathcal{L}_{\mathcal{F}}}|\mu_n(g) - \mu(g)|.$$

Since every $g \in \mathcal{L}_{\mathcal{F}}$ takes values in $[0,1]$, the empirical-process Dudley bound gives

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant 2\,\mathbb{E}\sup_{g\in\mathcal{L}_{\mathcal{F}}}|\mu_n(g) - \mu(g)| \leqslant \frac{C}{\sqrt{n}}\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{L}_{\mathcal{F}}, \|\cdot\|_\infty, \varepsilon)}\, d\varepsilon.$$

Using the result from part (a) and the given covering-number bound for Lipschitz functions

$$\mathcal{N}(\mathcal{L}_{\mathcal{F}}, \|\cdot\|_\infty, \varepsilon) \leqslant \mathcal{N}(\mathcal{F}_{L,1}, \|\cdot\|_\infty, \varepsilon/2) \leqslant \exp\left(\frac{2CL}{\varepsilon}\right).$$

Substituting it back, we get

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant \frac{C}{\sqrt{n}}\int_0^1 \sqrt{\frac{L}{\varepsilon}}\, d\varepsilon = C\sqrt{\frac{L}{n}}\int_0^1 \varepsilon^{-1/2}\, d\varepsilon \leqslant C\sqrt{\frac{L}{n}}. \qquad \square$$

(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

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

Solution:
Using the given entropy bound,

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

the Dudley integral is bounded above by

$$\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}_{L,d}, \|\cdot\|_\infty, \varepsilon)}\, d\varepsilon \ \leqslant\ \sqrt{C_d}\, L^{d/2}\int_0^1 \varepsilon^{-d/2}\, d\varepsilon.$$

Now, $\int_0^1 \varepsilon^{-d/2}\, d\varepsilon$ converges near 0 if and only if $d/2 < 1$, i.e. $d < 2$. Hence, for every $d \geqslant 2$, this integral diverges, and so the naive Dudley bound is not useful in general. $\square$

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

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

Solution:
Let $g \in \mathcal{L}_\mathcal{F}$. Since $\mathcal{N}_\varepsilon$ is an ε-net of $\mathcal{L}_\mathcal{F}$ in $\|\cdot\|_\infty$, there exists $h \in \mathcal{N}_\varepsilon$ such that

$$\|g - h\|_\infty \leqslant \varepsilon.$$

Then, using the triangle inequality

$$|\mu_n(g) - \mu(g)| \leqslant |\mu_n(g) - \mu_n(h)| + |\mu_n(h) - \mu(h)| + |\mu(h) - \mu(g)|.$$

Now

$$|\mu_n(g) - \mu_n(h)| = \left|\frac{1}{n}\sum_{i=1}^n (g(X_i) - h(X_i))\right| \leqslant \|g - h\|_\infty \leqslant \varepsilon,$$

and similarly using Jensen's

$$|\mu(h) - \mu(g)| = |\mathbb{E}[h(X) - g(X)]| \leqslant \mathbb{E}[|h(X) - g(X)|] \leqslant \|h - g\|_\infty \leqslant \varepsilon.$$

Therefore

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

Taking the supremum over $g \in \mathcal{L}_\mathcal{F}$ yields

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

(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

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

Solution:
Taking expectation in the excess-risk lemma stated in the problem,

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant 2\,\mathbb{E}\sup_{g\in\mathcal{L}_\mathcal{F}} |\mu_n(g) - \mu(g)|.$$

Fix $\varepsilon \in (0, 1)$, and let $\mathcal{N}_\varepsilon \subset \mathcal{L}_\mathcal{F}$ be an ε-net in $\|\cdot\|_\infty$. Using the result from part (c2),

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant 4\varepsilon + 2\,\mathbb{E}\max_{h\in\mathcal{N}_\varepsilon} |\mu_n(h) - \mu(h)|.$$

Using the finite-class bound stated in the problem,

$$\mathbb{E}\big[R(\widehat{f}_n) - R(f^\star)\big] \leqslant 4\varepsilon + 2C\sqrt{\frac{\log|\mathcal{N}_\varepsilon|}{n}} = 4\varepsilon + 2C\sqrt{\frac{\log \mathcal{N}(\mathcal{L}_\mathcal{F}, \|\cdot\|_\infty, \varepsilon)}{n}}.$$

Therefore, after absorbing the constants,

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

Now let $\mathcal{F} = \mathcal{F}_{L,d}$. Using the result from part (a),

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

and by the given covering-number bound for Lipschitz functions

$$\log \mathcal{N}(\mathcal{F}_{L,d}, \|\cdot\|_\infty, \varepsilon/2) \leqslant C_d \left(\frac{L}{\varepsilon}\right)^d,$$

after absorbing constants, we obtain

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

(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*.

Solution:
From part (c3), for every $\varepsilon \in (0, 1)$,

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

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

$$\begin{aligned}
\varepsilon + \sqrt{\frac{(L/\varepsilon)^d}{n}} &\asymp L^{d/(d+2)}\, n^{-1/(d+2)} + L^{d/2} n^{-1/2} L^{-d^2/2(d+2)}\, n^{d/2(d+2)} \\
&\asymp L^{d/(d+2)}\, n^{-1/(d+2)} + L^{d/(d+2)}\, n^{-1/(d+2)} \\
&\asymp L^{d/(d+2)}\, n^{-1/(d+2)}.
\end{aligned}$$

This gives

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

This is an instance of the curse of dimensionality because the exponent $1/(d+2)$ decreases as d increases, so the rate becomes slower in higher dimensions. Equivalently, to achieve excess risk at most δ, one needs

$$C_d\, L^{d/(d+2)} n^{-1/(d+2)} \lesssim \delta \implies n \gtrsim C_d' L^d \delta^{-(d+2)}.$$

Thus, the required sample size grows polynomially in $1/\delta$ with exponent $d+2$, which becomes very large when d is large. This is exactly the curse of dimensionality.

(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.

[Bonus] Solution:
Let

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

By the same argument as in part (c3), for every $\varepsilon \in (0,1)$,

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

Using part (a) and the given entropy assumption

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

Substituting and after absorbing constants, we get

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

Choose $\varepsilon \asymp A^{1/(p+2)}\,n^{-1/(p+2)}$, which gives

$$\begin{aligned}
\varepsilon + A^{1/2} n^{-1/2} \varepsilon^{-p/2} &\asymp A^{1/(p+2)}\,n^{-1/(p+2)} + A^{1/2} n^{-1/2} A^{-p/2(p+2)}\,n^{p/2(p+2)} \\
&\asymp A^{1/(p+2)} n^{-1/(p+2)} + A^{1/(p+2)} n^{-1/(p+2)} \\
&\asymp A^{1/(p+2)} n^{-1/(p+2)}.
\end{aligned}$$

Thus we obtain

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

Now let $p = d/s$. Then the bound becomes

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

As s increases, the exponent $s/(2s+d)$ increases, so the rate improves. In the formal limit as $s \to \infty$, $\dfrac{s}{2s+d} \to \dfrac{1}{2}$ so the rate approaches the parametric $n^{-1/2}$ scale. Thus, additional smoothness reduces the complexity of the class and helps mitigate the curse of dimensionality.

* * * End of Solutions * * *