Theoretical Stats and Machine Learning (Homework 05)

Student Name: Trang Pham

UT EID: vp22673

Question 1: Practice with Gaussian processes and Gaussian width

For a bounded set $T \subset \mathbb{R}^d$, we have the definition of the Gaussian width:

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

a. Basic properties

a1. We have:

$$\begin{aligned}
&\langle g, t \rangle \leq \sup_{s \in S} \langle g, s \rangle \qquad \forall t \in T \ \text{ (because } T \subset S \text{ then every } t \in T \text{ implies } t \in S) \\
\implies &\sup_{t \in T} \langle g, t \rangle \leq \sup_{s \in S} \langle g, s \rangle \\
\implies &\mathbb{E} \sup_{t \in T} \langle g, t \rangle \leq \mathbb{E} \sup_{s \in S} \langle g, s \rangle \\
\implies &w(T) \leq w(S)
\end{aligned}$$

a2. By definition, we have:

$$\begin{aligned}
w(aT) &= \mathbb{E} \sup_{u \in aT} \langle g, u \rangle \\
&= \mathbb{E} \sup_{t \in T} \langle g, at \rangle \\
&= \mathbb{E} \sup_{t \in T} a \langle g, t \rangle \\
&= a \mathbb{E} \sup_{t \in T} \langle g, t \rangle \qquad \text{(because } a > 0\text{)} \\
&= a w(T)
\end{aligned}$$

a3. By definition, we have:

$$\begin{aligned} w(T+x_0) &= \mathbb{E} \sup_{u \in T+x_0} \langle g, u \rangle \\ &= \mathbb{E} \sup_{u \in T+x_0} \langle g, t + x_0 \rangle \\ &= \mathbb{E} \sup_{t \in T} \langle g, t \rangle + \mathbb{E}\langle g, x_0 \rangle \qquad \text{(because for fixed } g,\ \langle g, x_0 \rangle \text{ does not depend on } t) \end{aligned}$$

Moreover, $\mathbb{E}\langle g, x_0 \rangle = \langle \mathbb{E} g, x_0 \rangle = \langle 0, x_0 \rangle = 0$ because $g \sim \mathcal{N}(0, I_d)$ has mean zero. Therefore:

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

a4. Because $T \subset \text{conv}(T)$, from part (a1) we have:

$$w(T) \leq w(\text{conv}(T)) \tag{1}$$

We now prove the reverse inequality.

Fix $g \in \mathbb{R}^d$. Let $v \in \text{conv}(T)$. By definition of the convex hull, we have: $v = \sum_{i=1}^m \lambda_i t_i$ for some $t_1, \ldots, t_m \in T$, $\lambda_i \geq 0$ and $\sum_{i=1}^m \lambda_i = 1$. Therefore:

$$\langle g, v \rangle = \left\langle g, \sum_{i=1}^m \lambda_i t_i \right\rangle = \sum_{i=1}^m \lambda_i \langle g, t_i \rangle.$$

Because each $\langle g, t_i \rangle \leq \sup_{t \in T} \langle g, t \rangle$, we get:

$$\langle g, v \rangle \leq \sum_{i=1}^m \lambda_i \sup_{t \in T} \langle g, t \rangle = \left(\sum_{i=1}^m \lambda_i \right) \sup_{t \in T} \langle g, t \rangle = \sup_{t \in T} \langle g, t \rangle.$$

The above result holds for every $v \in \text{conv}(T)$, therefore $\sup_{v \in \text{conv}(T)} \langle g, v \rangle \leq \sup_{t \in T} \langle g, t \rangle$.

Take expectations, we obtain: $\mathbb{E} \sup_{v \in \text{conv}(T)} \langle g, v \rangle \leq \mathbb{E} \sup_{t \in T} \langle g, t \rangle$, or equivalently:

$$w(\text{conv}(T)) \leq w(T) \tag{2}$$

Combine 1 and 2, we conclude that: $\boxed{w(\text{conv}(T)) = w(T)}$.

b. Canonical examples

b1. For every $t \in B_2^d$, by Cauchy–Schwarz, we have:

$$\langle g, t\rangle \le |\langle g, t\rangle| \le \|g\|_2\|t\|_2 \le \|g\|_2.$$

Therefore:

$$\sup_{\|t\|_2\le 1} \langle g, t\rangle \le \|g\|_2 \tag{3}$$

If $g \neq 0$, choose $t = \frac{g}{\|g\|_2}$. Therefore $\|t\|_2 = 1$, so $t \in B_2^d$, and

$$\langle g, t\rangle = \left\langle g, \frac{g}{\|g\|_2}\right\rangle = \frac{\|g\|_2^2}{\|g\|_2} = \|g\|_2.$$

Therefore:

$$\sup_{\|t\|_2\le 1} \langle g, t\rangle \ge \|g\|_2 \tag{4}$$

From 3 and 4, we have:

$$\sup_{\|t\|_2\le 1} \langle g, t\rangle = \|g\|_2.$$

Take expectations, we conclude: $\boxed{w(B_2^d) = \mathbb{E}\|g\|_2}$.

Prove that: $c\sqrt{d} \le w(B_2^d) \le \sqrt{d}$.

For the upper bound, by Jensen's inequality, we obtain:

$$\begin{aligned}(\mathbb{E}\|g\|_2)^2 &\le \mathbb{E}\|g\|_2^2 \\ &= \mathbb{E}\sum_{i=1}^d g_i^2 \\ &= \sum_{i=1}^d \mathbb{E}g_i^2 = d\cdot\mathbb{E}g_1^2 = d\cdot 1 = d\end{aligned}$$

Therefore: $\boxed{w(B_2^d) = \mathbb{E}\|g\|_2 \le \sqrt{d}}$.

For the lower bound, let $\tilde{g} = (|g_1|, |g_2|, ..., |g_d|)$. By Cauchy–Schwarz, we have:

$$\langle \tilde{g}, \mathbf{1} \rangle = \sum_{i=1}^{d} |g_i| \leq \sqrt{d}\, \|g\|_2$$

Therefore: $\|g\|_2 \geq \frac{1}{\sqrt{d}} \sum_{i=1}^{d} |g_i|$.

Take expectations, we obtain: $\mathbb{E}\|g\|_2 \geq \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \mathbb{E}|g_i| = \sqrt{d}\, \mathbb{E}|g_1|$.

Because $g_1 \sim N(0,1)$, we have:

$$\mathbb{E}|g_1| = \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 2 \int_{0}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} x e^{-x^2/2}\, dx$$

$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-u}\, du \qquad (u = x^2/2,\ du = x\, dx)$$

$$= \frac{2}{\sqrt{2\pi}} \cdot 1 = \sqrt{\frac{2}{\pi}}.$$

Therefore: $\boxed{\mathbb{E}\|g\|_2 \geq \sqrt{\frac{2}{\pi}} \sqrt{d}}$.

b2. For any $t \in B_1^d$ (or equivalently, $\|t\|_1 \leq 1$), we have:

$$\langle g, t \rangle = \sum_{i=1}^{d} g_i t_i \leq \sum_{i=1}^{d} |g_i||t_i| \leq \|g\|_\infty \sum_{i=1}^{d} |t_i| = \|g\|_\infty \|t\|_1 \leq \|g\|_\infty.$$

Therefore:

$$\sup_{\|t\|_1 \leq 1} \langle g, t \rangle \leq \|g\|_\infty \tag{5}$$

Let $j \in \arg\max_{1 \leq i \leq d} |g_i|$ and $t = \operatorname{sgn}(g_j) e_j$.

Then $\|t\|_1 = 1$, so $t \in B_1^d$, and $\langle g, t \rangle = g_j \operatorname{sgn}(g_j) = |g_j| = \|g\|_\infty$. Therefore:

$$\sup_{\|t\|_1 \leq 1} \langle g, t \rangle \geq \|g\|_\infty \tag{6}$$

From 5 and 6, we obtain: $\sup_{\|t\|_1 \leq 1} \langle g, t \rangle = \|g\|_\infty$.

Take expectations, we conclude: $\boxed{w(B_1^d) = \mathbb{E}\|g\|_\infty}$.

Prove that $w(B_1^d) \simeq \sqrt{\log d}$.

To prove the upper bound, let $M := \|g\|_\infty = \max_{1\le i\le d} |g_i|$.

For any $\lambda > 0$, we have:

$$e^{\lambda M} = e^{\lambda \max_i |g_i|} \le \sum_{i=1}^{d} e^{\lambda |g_i|}$$
$$\implies \mathbb{E} e^{\lambda M} \le \sum_{i=1}^{d} \mathbb{E} e^{\lambda |g_i|}$$

Also, we have: $e^{\lambda|x|} \le e^{\lambda x} + e^{-\lambda x}$. Therefore, for $g_i \sim N(0,1)$, we have:

$$\mathbb{E} e^{\lambda |g_i|} \le \mathbb{E} e^{\lambda g_i} + \mathbb{E} e^{-\lambda g_i} = 2e^{\lambda^2/2}$$

Therefore: $\mathbb{E} e^{\lambda M} \le 2d\, e^{\lambda^2/2}$. By Jensen, in equality, we have:

$$e^{\lambda \mathbb{E} M} \le \mathbb{E} e^{\lambda M} \le 2d\, e^{\lambda^2/2}$$
$$\implies \lambda \mathbb{E} M \le \log(2d) + \frac{\lambda^2}{2}$$
$$\implies \mathbb{E} M \le \frac{\log(2d)}{\lambda} + \frac{\lambda}{2}$$

Optimize the right-hand side w.r.t λ to get tighter upper bound, we obtain: $\lambda = \sqrt{2\log(2d)}$ and $\mathbb{E} M \le \sqrt{2\log(2d)} \lesssim \sqrt{\log d}$.

For the lower bound, again let $M = \max_{1\le i\le d} |g_i|$.

For every $t > 0$, we have: $M \ge t\, \mathbf{1}_{\{M \ge t\}}$, so by taking expectation of both sides, we obtain: $\mathbb{E} M \ge t\, \mathbb{P}(M \ge t)$.

We also have:

$$\mathbb{P}(M \le t) = \prod_{i=1}^{d} \mathbb{P}(|g_i| \le t) = \big(1 - \mathbb{P}(|g_1| \ge t)\big)^d$$

Therefore: $\mathbb{P}(M \ge t) = 1 - \big(1 - \mathbb{P}(|g_1| \ge t)\big)^d$.

$$\mathbb{P}(M \ge t) = 1 - \big(1 - \mathbb{P}(|g_1| \ge t)\big)^d.$$

Using $1 - x \le e^{-x}$, we obtain: $\mathbb{P}(M \ge t) \ge 1 - \exp\big(-d\, \mathbb{P}(|g_1| \ge t)\big)$.

Choose $t = \frac{1}{2}\sqrt{\log d}$. We need a lower bound on $\mathbb{P}(|g_1| \geq t)$. We have:

$$\mathbb{P}(|g_1| \geq t) = \frac{2}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx \geq \frac{2}{\sqrt{2\pi}} \int_t^{t+1} e^{-x^2/2}\, dx$$

For $x \in [t, t+1]$, we have $x^2 \leq (t+1)^2$, so: $e^{-x^2/2} \geq e^{-(t+1)^2/2}$.

Therefore, $\mathbb{P}(|g_1| \geq t) \geq \frac{2}{\sqrt{2\pi}} e^{-(t+1)^2/2}$, which implies $d\,\mathbb{P}(|g_1| \geq t) \geq c\, d\, e^{-(t+1)^2/2}$.

With $t = \frac{1}{2}\sqrt{\log d}$, we have: $(t+1)^2 = \frac{1}{4}\log d + \sqrt{\log d} + 1$.

Therefore:

$$d\, e^{-(t+1)^2/2} = \exp\left(\log d - \frac{1}{8}\log d - \frac{1}{2}\sqrt{\log d} - \frac{1}{2}\right) = \exp\left(\frac{7}{8}\log d - \frac{1}{2}\sqrt{\log d} - \frac{1}{2}\right).$$

This tends to $+\infty$ as $d \to \infty$, so for all sufficiently large d, $d\,\mathbb{P}(|g_1| \geq t) \geq 1$.

Therefore, $\mathbb{P}(M \geq t) \geq 1 - e^{-1}$, which implies $\mathbb{E}M \geq t(1 - e^{-1}) \gtrsim \sqrt{\log d}$.

Combine the upper and lower bounds, we obtain: $\mathbb{E}\|g\|_\infty \asymp \sqrt{\log d}$.

Therefore, we conclude that: $\boxed{w(B_1^d) \asymp \sqrt{\log d}}$.

b3. Let $Z_j := \langle g, t_j \rangle$ and $\sigma := \max_{1 \leq j \leq M} \|t_j\|_2$.

$$Z_j := \langle g, t_j \rangle, \qquad \sigma := \max_{1 \leq j \leq M} \|t_j\|_2.$$

We have that each Z_j is centered Gaussian with variance $\mathrm{Var}(Z_j) = \|t_j\|_2^2 \leq \sigma^2$.

Because $w(T) = \mathbb{E}\max_{1 \leq j \leq M} Z_j$, we now bound $\mathbb{E}\max_j Z_j$.

For any $\lambda > 0$, we have: $\mathbb{E}e^{\lambda Z_j} = e^{\lambda^2 \|t_j\|_2^2/2} \leq e^{\lambda^2\sigma^2/2}$. We also have: $e^{\lambda \max_j Z_j} \leq \sum_{j=1}^M e^{\lambda Z_j}$.

Therefore:

$$\mathbb{E}e^{\lambda \max_j Z_j} \leq \sum_{j=1}^M \mathbb{E}e^{\lambda Z_j} \leq M e^{\lambda^2\sigma^2/2}.$$

By Jensen inequality, we have:

$$e^{\lambda \mathbb{E} \max_j Z_j} \leq \mathbb{E} e^{\lambda \max_j Z_j} \leq M e^{\lambda^2 \sigma^2 / 2}$$
$$\implies \lambda \, \mathbb{E} \max_j Z_j \leq \log M + \frac{\lambda^2 \sigma^2}{2}$$
$$\implies \mathbb{E} \max_j Z_j \leq \frac{\log M}{\lambda} + \frac{\lambda \sigma^2}{2}$$

Optimize the right-hand side w.r.t λ to get a tighter lower bound, we obtain $\lambda = \frac{\sqrt{2 \log M}}{\sigma}$ and $\mathbb{E} \max_j Z_j \leq \sigma \sqrt{2 \log M}$.

Therefore, we can conclude that: $\boxed{w(T) \leq \left(\max_{1 \leq j \leq M} \|t_j\|_2 \right) \sqrt{2 \log M}}$.

c. [Bonus] Width of polytopes and sparse sets.

c1. By part (a4), we have: $w(P) = w(\text{conv}\{v_1, \ldots, v_N\}) = w(\{v_1, \ldots, v_N\})$.

We apply part (b3) with $T = \{v_1, \ldots, v_N\}$. Because $\max_j \|v_j\|_2 \leq 1$, we obtain:

$$w(P) \leq \left(\max_{1 \leq j \leq M} \|v_j\|_2 \right) \sqrt{2 \log M} \leq \sqrt{2 \log N} \leq C \sqrt{\log N}.$$

c2. For each $S \subset [d]$ with $|S| = k$, define $Y_S := \sup\{\langle g, x \rangle : \text{supp}(x) \subset S, \ \|x\|_2 \leq 1\}$.

Every $x \in T_k$ has support of size at most k, so its support is contained in some set $S \subset [d]$ with $|S| = k$. Therefore: $\sup_{x \in T_k} \langle g, x \rangle \leq \max_{S: \ |S|=k} Y_S$. Take expectations, we obtain:

$$w(T_k) = \mathbb{E} \sup_{x \in T_k} \langle g, x \rangle \leq \mathbb{E} \max_{S: \ |S|=k} Y_S.$$

Fix $S \subset [d]$, $|S| = k$. Restricting to the coordinates in S, the problem becomes the Euclidean ball in $\mathbb{R}^k$, so by part (b1), we have $Y_S = \|g_S\|_2$, where g_S is the restriction of g to the coordinates in S. Because $g_S \sim N(0, I_k)$, part (b1) also gives $\mathbb{E} Y_S = \mathbb{E} \|g_S\|_2 \leq \sqrt{k}$.

We then show that $Y_S - \mathbb{E} Y_S$ is sub-Gaussian uniformly in S. Consider the function $f_S(g) := \|g_S\|_2$. For any $g, h \in \mathbb{R}^d$, we have: $|f_S(g) - f_S(h)| = \big| \|g_S\|_2 - \|h_S\|_2 \big| \leq \|g_S - h_S\|_2 \leq \|g - h\|_2$. Therefore f_S is 1-Lipschitz. Therefore, the Gaussian concentration yields absolute constants $c, C > 0$ such that for all $t \geq 0$,

$$\mathbb{P}\big(|Y_S - \mathbb{E} Y_S| \geq t\big) = \mathbb{P}\big(|f_S(g) - \mathbb{E} f_S(g)| \geq t\big) \leq 2 \exp(-c t^2)$$

and $\|Y_S - \mathbb{E}Y_S\|_{\psi_2} = \|f_S(g) - \mathbb{E}f_S(g)\|_{\psi_2} \le C$.

Therefore $Y_S - \mathbb{E}Y_S$ is sub-Gaussian with an absolute constant, uniformly in S.

The number of subsets $S \subset [d]$ with $|S| = k$ is $\binom{d}{k}$.

Apply the maximal inequality for finitely many sub-Gaussian random variables to the family $\{Y_S - \mathbb{E}Y_S : \ |S| = k\}$, we obtain: $\mathbb{E}\max_{|S|=k}(Y_S - \mathbb{E}Y_S) \le C\sqrt{\log\binom{d}{k}}$.

Therefore:

$$\mathbb{E}\max_{|S|=k} Y_S \le \max_{|S|=k} \mathbb{E}Y_S + \mathbb{E}\max_{|S|=k}(Y_S - \mathbb{E}Y_S) \le \sqrt{k} + C\sqrt{\log\binom{d}{k}}.$$

The above result implies that:

$$w(T_k) \le \sqrt{k} + C\sqrt{\log\binom{d}{k}}.$$

Finally, use the standard combinatorial bound we have $\binom{d}{k} = \frac{d!}{k!(d-k)!} = \frac{d(d-1)\cdots(d-k+1)}{k!} \le \frac{d^k}{k!} \le \frac{d^k}{(k/e)^k} = \left(\frac{ed}{k}\right)^k$, so $\log\binom{d}{k} \le k\log\left(\frac{ed}{k}\right)$. Therefore: $w(T_k) \le \sqrt{k} + C\sqrt{k\log\left(\frac{ed}{k}\right)}$. Because $d \ge k$ then $\log(ed/k) \ge 1$, which implies $\sqrt{k} \le \sqrt{k\log\left(\frac{ed}{k}\right)}$, the first term is absorbed by the second term. Mathematically, $\sqrt{k} + C\sqrt{k\log\left(\frac{ed}{k}\right)} \le (1+C)\sqrt{k\log\left(\frac{ed}{k}\right)}$. .Therefore, we can conclude $\boxed{w(T_k) \le C\sqrt{k\log\left(\frac{ed}{k}\right)}}$.

Question 2: Practice with Rademacher processes and Rademacher complexity

a1. We have:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i) \le \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i) \qquad \forall f \in \mathcal{F} \qquad \text{(because } \mathcal{F} \subset \mathcal{G}) \\
\implies &\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i) \le \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i) \\
\implies &\mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i) \le \mathbb{E}\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i) \\
\implies &\hat{\mathfrak{R}}_n(\mathcal{F}) \le \hat{\mathfrak{R}}_n(\mathcal{G})
\end{aligned}$$

a2. By definition, we have:

$$\begin{aligned}
\hat{\mathfrak{R}}_n(a\mathcal{F}) &= \mathbb{E} \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (a f(x_i)) \\
&= \mathbb{E} \sup_{f\in\mathcal{F}} a \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right) \\
&= a\mathbb{E} \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \qquad (\text{because } a \geq 0) \\
&= a\hat{\mathfrak{R}}_n(\mathcal{F})
\end{aligned}$$

a3. Because $\mathcal{F} \subset \text{conv}(\mathcal{F})$, from part (a1) we have:

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq \hat{\mathfrak{R}}_n(\text{conv}(\mathcal{F})) \tag{7}$$

We now prove the reverse inequality.

Fix ε. Let $h \in \text{conv}(\mathcal{F})$ then:

$$\begin{aligned}
h &= \sum_{i=1}^{m} \lambda_i f_i \quad (\lambda_i \geq 0, \sum_{i=1}^{m} \lambda_i = 1) \\
\implies \quad \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j h(x_j) &= \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j \sum_{i=1}^{m} \lambda_i f_i(x_j) \\
&= \sum_{i=1}^{m} \lambda_i \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j f_i(x_j) \\
&\leq \sum_{i=1}^{m} \lambda_i \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j f(x_j) \\
&= \left(\sum_{i=1}^{m} \lambda_i \right) \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j f(x_j) \\
&= \sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{j=1}^{n} \varepsilon_j f(x_j)
\end{aligned}$$

Therefore:

$$\sup_{h\in\text{conv}(\mathcal{F})} \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j h(x_j) \le \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j f(x_j)$$

$$\implies \quad \mathbb{E}\sup_{h\in\text{conv}(\mathcal{F})} \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j h(x_j) \le \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j f(x_j) \tag{8}$$

$$\implies \quad \hat{\mathfrak{R}}_n(\text{conv}(\mathcal{F})) \le \hat{\mathfrak{R}}_n(\mathcal{F})$$

Combine 7 and 8, we conclude that: $\hat{\mathfrak{R}}_n(\text{conv}(\mathcal{F})) = \hat{\mathfrak{R}}_n(\mathcal{F})$.

b. Finite-class bound (Massart-type bound)

Assume $\mathcal{F}$ is finite and for every $f \in \mathcal{F}$, $\frac{1}{n}\sum_{i=1}^{n} f(x_i)^2 \le r^2$.

We will prove that:

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

Step 1: Rewrite the quantity

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

Because $\mathcal{F}$ is finite, the supremum is a maximum. We define $Z_f := \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)$ and rewrite $\hat{\mathfrak{R}}_n(\mathcal{F}; x_1, \ldots, x_n)$ as follows:

$$\hat{\mathfrak{R}}_n(\mathcal{F}; x_1, \ldots, x_n) = \mathbb{E}_\varepsilon \max_{f\in\mathcal{F}} Z_f.$$

Now our problem is to bound the expectation of the maximum of finitely many random variables Z_f.

Step 2: Show that each Z_f is sub-Gaussian

Fix $f \in \mathcal{F}$. We have $Z_f = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)$.

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

Because each ε_i is Rademacher, we have:

$$\mathbb{E}e^{\lambda\varepsilon_i} = \frac{e^{\lambda} + e^{-\lambda}}{2} = \cosh(\lambda) \le e^{\lambda^2/2}.$$

Indeed, using the Taylor expansions $\cosh(\lambda) = \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!}$ and $e^{\lambda^2/2} = \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{2^k k!}$ and that $(2k)! \ge 2^k k!$ for all k, we have $\frac{1}{(2k)!} \le \frac{1}{2^k k!}$, therefore term-by-term $\frac{\lambda^{2k}}{(2k)!} \le \frac{\lambda^{2k}}{2^k k!}$, which implies $\cosh(\lambda) \le e^{\lambda^2/2}$. Therefore, ε_i is sub-Gaussian with proxy 1.

Therefore $\varepsilon_i f(x_i)$ is sub-Gaussian with proxy $f(x_i)^2$, because: $\mathbb{E}e^{\lambda \varepsilon_i f(x_i)} \leq e^{\lambda^2 f(x_i)^2/2}$.

We now compute the mgf of Z_f: $\mathbb{E}e^{\lambda Z_f} = \mathbb{E}\exp\left(\lambda \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right)$. By independence, we have:

$$\mathbb{E}e^{\lambda Z_f} = \prod_{i=1}^n \mathbb{E}e^{(\lambda/n)\varepsilon_i f(x_i)}.$$

Use the sub-Gaussian bound, we have:

$$\begin{aligned}
&\mathbb{E}e^{(\lambda/n)\varepsilon_i f(x_i)} \leq e^{\lambda^2 f(x_i)^2/(2n^2)} \\
&\implies \mathbb{E}e^{\lambda Z_f} \leq \exp\left(\frac{\lambda^2}{2n^2}\sum_{i=1}^n f(x_i)^2\right) \\
&\implies \mathbb{E}e^{\lambda Z_f} \leq \exp\left(\frac{\lambda^2 r^2}{2n}\right) \qquad \text{(use the assumption } \sum_{i=1}^n f(x_i)^2 \leq nr^2\text{)}
\end{aligned}$$

Therefore, Z_f is sub-Gaussian with variance proxy $\sigma^2 = \frac{r^2}{n}$.

Step 3: Bound $\mathbb{E}\max_{f\in\mathcal{F}} Z_f$

Let $Y := \max_{f\in\mathcal{F}} Z_f$.

For any $\lambda > 0$, we have

$$\begin{aligned}
e^{\lambda Y} = e^{\lambda \max_f Z_f} \qquad &= \max_f e^{\lambda Z_f} \leq \sum_{f\in\mathcal{F}} e^{\lambda Z_f} \\
\implies \mathbb{E}e^{\lambda Y} &\leq \sum_{f\in\mathcal{F}} \mathbb{E}e^{\lambda Z_f} \\
\implies \mathbb{E}e^{\lambda Y} &\leq |\mathcal{F}| e^{\lambda^2 r^2/(2n)} \quad \text{(step 2)}
\end{aligned}$$

By Jensen inequality, we have: $e^{\lambda \mathbb{E}Y} \leq \mathbb{E}e^{\lambda Y}$.

Therefore:

$$\begin{aligned}
e^{\lambda \mathbb{E}Y} &\leq |\mathcal{F}|\, e^{\lambda^2 r^2/(2n)} \\
\implies \lambda \mathbb{E}Y &\leq \log|\mathcal{F}| + \frac{\lambda^2 r^2}{2n} \\
\implies \mathbb{E}Y &\leq \frac{\log|\mathcal{F}|}{\lambda} + \frac{\lambda r^2}{2n}
\end{aligned}$$

Optimize over λ to get a tighter bound, we obtain $\lambda = \frac{\sqrt{2n\log|\mathcal{F}|}}{r}$ and $\frac{\log|\mathcal{F}|}{\lambda} + \frac{\lambda r^2}{2n} = r\sqrt{\frac{2\log|\mathcal{F}|}{n}}$.

Therefore: $\mathbb{E}Y \leq r\sqrt{\frac{2\log|\mathcal{F}|}{n}}$.

Because $Y = \max_{f \in \mathcal{F}} Z_f$, we conclude that:

$$\hat{\mathfrak{R}}_n(\mathcal{F}; x_1, \ldots, x_n) = \mathbb{E}_\varepsilon \max_{f \in \mathcal{F}} Z_f \leq r\sqrt{\frac{2 \log |\mathcal{F}|}{n}}.$$

c. Linear function classes

Let $\mathcal{F}_R := \{x \mapsto \langle w, x \rangle : \|w\|_2 \leq R\}$.

Therefore:

$$\begin{aligned}\hat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \ldots, x_n) &= \mathbb{E}_\varepsilon \sup_{\|w\|_2 \leq R} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle w, x_i \rangle \\ &= \frac{1}{n} \mathbb{E}_\varepsilon \sup_{\|w\|_2 \leq R} \Big\langle w, \sum_{i=1}^n \varepsilon_i x_i \Big\rangle\end{aligned}$$

By Cauchy–Schwarz, we have: $\sup_{\|w\|_2 \leq R} \langle w, z \rangle = R\|z\|_2$. Therefore:

$$\hat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \ldots, x_n) = \frac{R}{n} \mathbb{E}_\varepsilon \Big\| \sum_{i=1}^n \varepsilon_i x_i \Big\|_2$$

Apply $\mathbb{E}\|Z\|_2 \leq \sqrt{\mathbb{E}\|Z\|_2^2}$ (Jensen inequality for concave function) with $Z = \sum_{i=1}^n \varepsilon_i x_i$, we obtain:

$$\mathbb{E}_\varepsilon \Big\| \sum_{i=1}^n \varepsilon_i x_i \Big\|_2 \leq \Big(\mathbb{E}_\varepsilon \Big\| \sum_{i=1}^n \varepsilon_i x_i \Big\|_2^2 \Big)^{1/2}$$

We also have:

$$\begin{aligned}\mathbb{E}_\varepsilon \Big\| \sum_{i=1}^n \varepsilon_i x_i \Big\|_2^2 &= \mathbb{E}_\varepsilon \Big\langle \sum_{i=1}^n \varepsilon_i x_i, \sum_{j=1}^n \varepsilon_j x_j \Big\rangle \\ &= \mathbb{E}_\varepsilon \sum_{i=1}^n \sum_{j=1}^n \varepsilon_i \varepsilon_j \langle x_i, x_j \rangle \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}_\varepsilon[\varepsilon_i \varepsilon_j] \langle x_i, x_j \rangle \\ &= \sum_{i=1}^n \mathbb{E}_\varepsilon[\varepsilon_i^2] \|x_i\|_2^2 \; + \; \sum_{i \neq j} \mathbb{E}_\varepsilon[\varepsilon_i \varepsilon_j] \langle x_i, x_j \rangle \\ &= \sum_{i=1}^n \|x_i\|_2^2.\end{aligned}$$

Therefore: $\hat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \ldots, x_n) \leq \frac{R}{n} \Big(\sum_{i=1}^n \|x_i\|_2^2 \Big)^{1/2}$.

If $\|x_i\|_2 \leq 1$ for all i, then $\hat{\mathfrak{R}}_n(\mathcal{F}_R; x_1, \ldots, x_n) \leq \frac{R}{\sqrt{n}}$.

d. [Bonus] Sparse linear predictors

Let $\mathcal{G}_R := \{x \mapsto \langle w, x\rangle : \|w\|_1 \leq R\}$ with $\|x_i\|_\infty \leq 1$.

Therefore:

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

By duality, we have: $\sup_{\|w\|_1 \leq R}\langle w, z\rangle = R\|z\|_\infty$. Therefore:

$$\begin{aligned} \hat{\mathfrak{R}}_n(\mathcal{G}_R; x_1, \ldots, x_n) &= \frac{R}{n}\mathbb{E}_\varepsilon \Big\| \sum_{i=1}^n \varepsilon_i x_i \Big\|_\infty \\ &= \frac{R}{n}\mathbb{E}_\varepsilon \max_{1\leq j\leq d} \Big| \sum_{i=1}^n \varepsilon_i x_{ij} \Big| \end{aligned} \tag{9}$$

Let $Z_j := \sum_{i=1}^n \varepsilon_i x_{ij}$. Then each Z_j is sub-Gaussian with proxy $\leq n$. Indeed, we have proved from part (b) that ε_i is sub-Gaussian with proxy 1, which is equivalent to $\mathbb{E}e^{\lambda\varepsilon_i} \leq e^{\lambda^2/2}$. Replace $\lambda = \lambda x_{ij}$, we obtain: $\mathbb{E}e^{\lambda x_{ij}\varepsilon_i} \leq e^{\frac{\lambda^2 x_{ij}^2}{2}}$.

Then, for any $\lambda \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E}_\varepsilon e^{\lambda Z_j} &= \mathbb{E}_\varepsilon \exp\left(\lambda \sum_{i=1}^n \varepsilon_i x_{ij}\right) \\ &= \prod_{i=1}^n \mathbb{E}_\varepsilon e^{\lambda\varepsilon_i x_{ij}} \qquad \text{(by independence)} \\ &\leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 x_{ij}^2}{2}\right) \qquad \text{(sub-Gaussian bound)} \\ &= \exp\left(\frac{\lambda^2}{2}\sum_{i=1}^n x_{ij}^2\right). \end{aligned}$$

Because $|x_{ij}| \leq 1$, we have $\sum_{i=1}^n x_{ij}^2 \leq n$, then $\mathbb{E}_\varepsilon e^{\lambda Z_j} \leq \exp\Big(\frac{\lambda^2 n}{2}\Big)$, which implies that Z_j is sub-Gaussian with variance proxy $\leq n$.

Therefore, by maximal inequality, we have: $\mathbb{E}_\varepsilon \max_{1\leq j\leq d} |Z_j| \leq C\sqrt{n\log d}$.

Substitute the above result to 9, we conclude that:

$$\hat{\mathfrak{R}}_n(\mathcal{G}_R; x_1, \ldots, x_n) \leq \frac{R}{n} C\sqrt{n \log d} = CR\sqrt{\frac{\log d}{n}}$$

Question 3: Practice with VC dimension

a. One-dimensional class

a1. The class of half-line

Let $\mathcal{H}_{\text{half}} = \{\mathbf{1}_{(-\infty,a]} : a \in \mathbb{R}\}$, we prove that $\text{VC}(\mathcal{H}_{\text{half}}) = 1$.

First, we show that $\mathcal{H}_{\text{half}}$ shatters any single point. Let $x_1 \in \mathbb{R}$. For the two possible labelings, we choose $h(x_1) = 0$ if $a < x_1$ and choose $h(x_1) = 1$ if $a \geq x_1$. Therefore, all labelings of $\{x_1\}$ can be realized.

We then show that no set of two points can be shattered. Let $x_1 < x_2$. For any $a \in \mathbb{R}$, we have:

$$a < x_1 \Rightarrow (h(x_1), h(x_2)) = (0, 0), \qquad x_1 \leq a < x_2 \Rightarrow (1, 0), \qquad a \geq x_2 \Rightarrow (1, 1).$$

Therefore, the only realizable labelings are: $(0, 0)$, $(1, 0)$, $(1, 1)$ and the labeling $(0, 1)$ cannot be realized. Therefore $\mathcal{H}_{\text{half}}$ does not shatter any 2-point set.

We conclude that $\boxed{\text{VC}(\mathcal{H}_{\text{half}}) = 1}$.

a2. The class of intervals.

Let $\mathcal{H}_{\text{int}} = \{\mathbf{1}_{[a,b]} : a \leq b,\ a, b \in \mathbb{R}\}$, we prove that $\text{VC}(\mathcal{H}_{\text{int}}) = 2$.

First, we show that some 2-point set is shattered. Let $x_1 < x_2$. The four labels can all be realized as follows:

- $(0, 0)$ by choosing $[a, b]$ disjoint from $\{x_1, x_2\}$
- $(1, 0)$ by choosing an interval containing x_1 but not x_2
- $(0, 1)$ by choosing an interval containing x_2 but not x_1
- $(1, 1)$ by choosing $[a, b] \supset \{x_1, x_2\}$

Therefore, $\mathcal{H}_{\text{int}}$ shatters $\{x_1, x_2\}$, so $\text{VC}(\mathcal{H}_{\text{int}}) \geq 2$.

We then show that no 3-point set can be shattered. Let $x_1 < x_2 < x_3$. Consider the label $(1, 0, 1)$. If

an interval $[a, b]$ contained both x_1 and x_3, then because x_2 lies between them, it must also contain x_2. Therefore, the label $(1, 0, 1)$ cannot be realized by any function in $\mathcal{H}_{\text{int}}$, which implies that no 3-point can be shattered, or equivalently, $\text{VC}(\mathcal{H}_{\text{int}}) \leq 2$.

Combine two inequalities, we conclude that: $\boxed{\text{VC}(\mathcal{H}_{\text{int}}) = 2}$.

a3. The class of unions of at most k intervals.

Let $\mathcal{H}_k = \left\{ \mathbf{1}_{\bigcup_{j=1}^{k}[a_j, b_j]} : a_j \leq b_j \right\}$. We prove that $\mathcal{H}_k = \left\{ \mathbf{1}_{\bigcup_{j=1}^{k}[a_j, b_j]} : a_j \leq b_j \right\}$.

Lower bound ($\geq 2k$).

Let $x_1 < \cdots < x_{2k}$. We show that $\mathcal{H}_k$ shatters this set.

Consider an arbitrary labeling $y_1, \ldots, y_{2k} \in \{0, 1\}$. Define a *run of ones* as a maximal set of indices $\{i, \ldots, j\}$ such that $y_i = \cdots = y_j = 1$, and (if they exist) $y_{i-1} = 0$ and $y_{j+1} = 0$.

Each run corresponds to a contiguous segment of points $x_i, \ldots, x_j$ that must be covered by an interval. Because between any two runs there must be at least one index with label 0, the total number of runs is at most k when there are $2k$ points (the maximum is achieved by the alternating labeling $1, 0, 1, 0, \ldots$).

For each run $\{i, \ldots, j\}$, choose an interval $[x_i, x_j]$ that covers exactly those points. This uses at most k intervals and realizes the given labeling. Therefore $\mathcal{H}_k$ shatters $\{x_1, \ldots, x_{2k}\}$, so $\text{VC}(\mathcal{H}_k) \geq 2k$.

Upper bound ($\leq 2k$).

Let $x_1 < \cdots < x_{2k+1}$. Consider the alternating labeling

$$y_i = \begin{cases} 1, & \text{if } i \text{ is odd,} \\ 0, & \text{if } i \text{ is even.} \end{cases}$$

Therefore, there are $k + 1$ runs of ones at indices $i = 1, 3, 5, \ldots, 2k + 1$.

Each interval can cover points from at most one such run, because any interval containing two of these points would also contain an intermediate point labeled 0. Therefore, realizing this labeling requires at least $k + 1$ intervals, which is not allowed.

Therefore, this labeling cannot be realized, so no set of size $2k + 1$ is shattered, which implies $\text{VC}(\mathcal{H}_k) \leq 2k$.

Combine two bounds, we conclude $\boxed{\text{VC}(\mathcal{H}_k) = 2k}$.

b. Two-dimensional classes

b1. Axis-aligned rectangles.

Let $\mathcal{R} = \{[a, b] \times [c, d] : a \leq b,\ c \leq d\}$, we will prove that $\mathrm{VC}(\mathcal{R}) = 4$.

Lower bound (≥ 4). Consider the four points $(1, 0),\ (-1, 0),\ (0, 1),\ (0, -1)$, we show that they are shattered by axis-aligned rectangles.

Let S be the subset of points labeled 1. If $S = \varnothing$, choose a rectangle disjoint from all four points. If $S \neq \varnothing$, let

$$a = \min\{x_1 : (x_1, x_2) \in S\}, \qquad b = \max\{x_1 : (x_1, x_2) \in S\},$$

$$c = \min\{x_2 : (x_1, x_2) \in S\}, \qquad d = \max\{x_2 : (x_1, x_2) \in S\}.$$

We can see that a, b, c, d are the extreme coordinates of the points in S. More precisely, a is the smallest x-coordinate among all points in S, b is the largest x-coordinate among all points in S, c is the smallest y-coordinate among all points in S, and d is the largest y-coordinate among all points in S.

Therefore, the rectangle $[a, b] \times [c, d]$ is exactly the smallest axis-aligned rectangle containing all points of S, which also implies that every labeling of these four points can be realized, so $\mathrm{VC}(\mathcal{R}) \geq 4$.

Upper bound (≤ 4). Now consider any five points in $\mathbb{R}^2$. Among them, choose points with minimal and maximal x-coordinate and minimal and maximal y-coordinate. These account for at most four points, so there exists at least one remaining point z.

Label the four extreme points by 1 and label z by 0. Suppose there were an axis-aligned rectangle containing all four extreme points. Then its left side must lie to the left of the point with minimal x-coordinate, its right side must lie to the right of the point with maximal x-coordinate, its bottom side must lie below the point with minimal y-coordinate, and its top side must lie above the point with maximal y-coordinate. Therefore it must also contain every point whose coordinates lie between these extremes, in particular the point z. This contradicts the labeling.

Therefore no set of five points can be shattered, so $\mathrm{VC}(\mathcal{R}) \leq 4$.

Combine two bounds, we conclude that: $\boxed{\mathrm{VC}(\mathcal{R}) = 4}$.

b2. [Bonus] Convex sets

Let $\mathcal{C}_{\text{conv}}$ be the class of indicators of all convex subsets of $\mathbb{R}^2$. We will prove that $\text{VC}(\mathcal{C}_{\text{conv}}) = \infty$. Fix any $m \geq 1$. Choose m distinct points $x_1, \ldots, x_m$ on a circle. We show that this set is shattered. Consider any labeling of these points by $\{0, 1\}$. Let $S = \{x_i : y_i = 1\}$ be the subset of positively labeled points, and let $C = \text{conv}(S)$ be its convex hull. Therefore, C is convex and contains every positively labeled point.

We now show that C contains no negatively labeled point from the original set $\{x_1, \ldots, x_m\}$. Because all m points lie on a circle, they are in convex position: no point lies in the convex hull of the others. Therefore, if $x_j \notin S$, then $x_j \notin \text{conv}(S)$. Indeed, assume for contradiction that there exists a negatively labeled point $x_j \notin S$ such that $x_j \in \text{conv}(S)$. By by definition of convex hull, there exist coefficients $\lambda_i \geq 0$ for $i \in S$ with $\sum_{i \in S} \lambda_i = 1$ such that $x_j = \sum_{i \in S} \lambda_i x_i$.

Take inner product with x_j on both sides yield

$$\langle x_j, x_j \rangle = \sum_{i \in S} \lambda_i \langle x_j, x_i \rangle.$$

Because all points lie on a circle of radius R, we have: $\|x_j\|^2 = R^2$.

Also, for every $i \in S$, we have $x_i \neq x_j$ (because $x_j \notin S$), so $\langle x_j, x_i \rangle < R^2$. Therefore:

$$R^2 = \sum_{i \in S} \lambda_i \langle x_j, x_i \rangle < \sum_{i \in S} \lambda_i R^2 = R^2,$$

which is a contradiction. Therefore: $x_j \notin \text{conv}(S)$.

This result implies that C contains exactly the positively labeled points from the original m-point set. Therefore, every labeling can be realized by a convex set, so the m points are shattered. Because m was arbitrary, we can conclude that $\boxed{\text{VC}(\mathcal{C}_{\text{conv}}) = \infty}$.

c. [Bonus] Euclidean balls and combinatorial counting.

c1. Let $\mathcal{B}_d$ denote the class of Euclidean balls in $\mathbb{R}^d$. We will prove that $\text{VC}(\mathcal{B}_d) = d + 1$. Indeed:

$$x \in B(c, r) \iff \|x - c\|_2^2 \leq r^2 \iff \|x\|_2^2 - 2c^\top x + \|c\|_2^2 \leq r^2$$

$$\iff 2c^\top x - \|x\|_2^2 \geq \|c\|_2^2 - r^2.$$

If we define the lifting map $\phi(x) = (x, \|x\|_2^2) \in \mathbb{R}^{d+1}$, then:

$$x \in B(c, r) \iff (2c, -1)^\top \phi(x) \geq \|c\|_2^2 - r^2.$$

Therefore, Euclidean balls in $\mathbb{R}^d$ correspond to halfspaces in $\mathbb{R}^{d+1}$ acting on the lifted points $\phi(x)$. Because this class has VC dimension $d+1$, we have: $\mathrm{VC}(\mathcal{B}_d) = d+1$.

c2. If $\mathcal{H}$ is a Boolean class with $\mathrm{VC}(\mathcal{H}) = v$ then on any n-point set it induces at most $\sum_{j=0}^{v} \binom{n}{j}$ distinct labelings. By the standard binomial estimate, we have:

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

c3. From part (b1), we know that $\mathrm{VC}(\mathcal{R}) = 4$.

Applying part (c2) with $v = 4$, we conclude that on any n-point set in $\mathbb{R}^2$, the number of distinct subsets that can be cut out by axis-aligned rectangles is at most $\boxed{\sum_{j=0}^{4} \binom{n}{j}}$.

Question 4: Practice with statistical learning theory

a. Excess-risk lemma

Let $h^* \in \arg\min_{h \in \mathcal{H}} R(h)$ and let $\hat{h} \in \arg\min_{h \in \mathcal{H}} R_n(h)$. We have:

$$R(\hat{h}) - R(h^*) = \big(R(\hat{h}) - R_n(\hat{h})\big) + \big(R_n(\hat{h}) - R_n(h^*)\big) + \big(R_n(h^*) - R(h^*)\big)$$

Because $\hat{h}$ minimizes the empirical risk, we know that $R_n(\hat{h}) - R_n(h^*) \leq 0$.

Therefore:

$$\begin{aligned} R(\hat{h}) - R(h^*) &\leq \big(R(\hat{h}) - R_n(\hat{h})\big) + \big(R_n(h^*) - R(h^*)\big) \\ &\leq |\big(R(\hat{h}) - R_n(\hat{h})\big)| + |\big(R_n(h^*) - R(h^*)\big)| \\ &\leq \sup_{h \in \mathcal{H}} |R_n(h) - R(h)| + \sup_{h \in \mathcal{H}} |R_n(h) - R(h)| \\ &\leq 2 \sup_{h \in \mathcal{H}} |R_n(h) - R(h)| \end{aligned}$$

b. Symmetrization

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

Write $\ell_h(x,y) := \ell(h(x),y)$, so $R(h) = \mathbb{E}[\ell_h(X,Y)]$ and $R_n(h) = \frac{1}{n}\sum_{i=1}^n \ell_h(X_i,Y_i)$.

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

Let $(X_1',Y_1'),\ldots,(X_n',Y_n')$ be an independent ghost sample with the same law as (X_i,Y_i).

Because $\mathbb{E}[\ell_h(X,Y)] = \mathbb{E}[\ell_h(X_i',Y_i')]$, we have

$$\begin{aligned}\mathbb{E}\sup_{\ell_h\in\mathcal{L}_\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n \ell_h(X_i,Y_i) - \mathbb{E}[\ell_h(X,Y)]\right| &= \mathbb{E}\sup_{\ell_h\in\mathcal{L}_\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n \big(\ell_h(X_i,Y_i) - \mathbb{E}[\ell_h(X_i',Y_i')]\big)\right| \\ &\le \mathbb{E}\sup_{\ell_h\in\mathcal{L}_\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^n \big(\ell_h(X_i,Y_i) - \ell_h(X_i',Y_i')\big)\right| \qquad \text{(by Jensen inequality)}\end{aligned}$$

Let $\varepsilon_1,\ldots,\varepsilon_n$ be i.i.d. Rademacher variables, independent of both samples. Because the joint law of $\big((X_i,Y_i),(X_i',Y_i')\big)$ is invariant under swapping the two coordinates, we obtain:

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

By the triangle inequality, we have:

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

Take suprema over $\ell_h \in \mathcal{L}_\mathcal{H}$ and then expectations, and use the fact that that the two terms have the same distribution, we obtain:

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

By definition, the right-hand side is: $2\,\mathbb{E}\,\hat{\mathfrak{R}}_n(\mathcal{L}_\mathcal{H};(X_1,Y_1),\ldots,(X_n,Y_n))$. Therefore, we conclude that:

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

c. Finite hypothesis classes

Assume $\mathcal{H}$ is finite and the loss is bounded in $[0,1]$. By part (a), we have:

$$R(\hat{h}) - R(h^*) \leq 2 \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|.$$

Take expectations and then applying part (b), we obtain:

$$\mathbb{E}\big[R(\hat{h}) - R(h^*)\big] \leq 2\,\mathbb{E} \sup_{h \in \mathcal{H}} |R_n(h) - R(h)| \leq 4\,\mathbb{E}\,\hat{\mathfrak{R}}_n(\mathcal{L}_{\mathcal{H}}; (X_1, Y_1), \ldots, (X_n, Y_n)) \tag{10}$$

We then apply Problem 2(b) to the finite class $\mathcal{L}_{\mathcal{H}}$. Because the loss takes values in $[0,1]$, for every $\ell_h \in \mathcal{L}_{\mathcal{H}}$ we have:

$$\frac{1}{n}\sum_{i=1}^{n} \ell_h(X_i, Y_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n} 1^2 \leq 1$$

Therefore, let $r = 1$ to get:

$$\hat{\mathfrak{R}}_n(\mathcal{L}_{\mathcal{H}}; (X_1, Y_1), \ldots, (X_n, Y_n)) \leq \sqrt{\frac{2\log|\mathcal{L}_{\mathcal{H}}|}{n}}.$$

Substitute this into 10, we obtain:

$$\mathbb{E}\big[R(\hat{h}) - R(h^*)\big] \leq 4\sqrt{\frac{2\log|\mathcal{L}_{\mathcal{H}}|}{n}}.$$

Finally, because each $h \in \mathcal{H}$ determines one loss function ℓ_h, we have: $|\mathcal{L}_{\mathcal{H}}| \leq |\mathcal{H}|$. Therefore:

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

We conclude that:

$$\boxed{\mathbb{E}\big[R(\hat{h}) - R(h^*)\big] \leq C\sqrt{\frac{\log|\mathcal{H}|}{n}}}$$

d. Boolean class and VC dimension

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

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

Because $\mathcal{L}_\mathcal{H}$ is a Boolean class on $\mathcal{X} \times \mathcal{Y}$, we will apply the VC generalization bound (Theorem 5.2, Lecture 20) to $\mathcal{L}_\mathcal{H}$.

First, we verify that $\mathrm{VC}(\mathcal{L}_\mathcal{H}) \leq \mathrm{VC}(\mathcal{H}) = v$.

Indeed, to show that $\mathrm{VC}(\mathcal{L}_\mathcal{H}) \leq \mathrm{VC}(\mathcal{H})$, it suffices to prove that whenever $\mathcal{L}_\mathcal{H}$ shatters m labeled points $((x_i, y_i))_{i=1}^m$, the class $\mathcal{H}$ shatters the corresponding inputs $x_1, \ldots, x_m$.

Assume that $\mathcal{L}_\mathcal{H}$ shatters $((x_i, y_i))_{i=1}^m$. Then for every binary vector $(z_1, \ldots, z_m) \in \{0,1\}^m$ where $z_i = \ell(h(x_i), y_i)$, there exists $h \in \mathcal{H}$ such that:

$$\mathbf{1}\{h(x_i) \neq y_i\} = z_i, \qquad i = 1, \ldots, m.$$

Equivalently:

$$h(x_i) = z_i \oplus y_i, \qquad i = 1, \ldots, m,$$

where $\oplus$ denotes XOR. Because coordinatewise XOR with the fixed vector $(y_1, \ldots, y_m)$ is a bijection on $\{0,1\}^m$, as $(z_1, \ldots, z_m)$ ranges over all of $\{0,1\}^m$, so does $(z_1 \oplus y_1, \ldots, z_m \oplus y_m)$. Therefore, for every binary vector $(w_1, \ldots, w_m) \in \{0,1\}^m$, there exists $h \in \mathcal{H}$ such that

$$h(x_i) = w_i, \qquad i = 1, \ldots, m.$$

Therefore $\mathcal{H}$ shatters $x_1, \ldots, x_m$, which implies that $\mathrm{VC}(\mathcal{H}) \geq m$ for any m points that $L_\mathcal{H}$ can shatter. Therefore, $\mathrm{VC}(\mathcal{L}_\mathcal{H}) \leq \mathrm{VC}(\mathcal{H})$.

Apply the VC generalization bound, we obtain:

$$\mathbb{E}\big[R(\hat{h}) - R(h^*)\big] \leq C\sqrt{\frac{\mathrm{VC}(\mathcal{L}_\mathcal{H}) \log\big(en/\mathrm{VC}(\mathcal{L}_\mathcal{H})\big)}{n}}.$$

Because $\mathrm{VC}(\mathcal{L}_\mathcal{H}) \leq \mathrm{VC}(\mathcal{H}) = v$, we conclude that: $\boxed{\mathbb{E}\big[R(\hat{h}) - R(h^*)\big] \leq C\sqrt{\frac{v \log(en/v)}{n}}}$.

e. A VC-style sample complexity statement.

By part (a), we have:

$$R(\hat{h}) - R(h^\star) \leq 2 \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|$$

by part (a). Therefore, it we = obtain a high-probability bound for $\sup_{h\in\mathcal{H}} |R_n(h)-R(h)|$ (eg. we can prove that $\mathbb{P}(\sup_{h\in\mathcal{H}} |R_n(h)-R(h)| \lesssim \varepsilon) \geq 1-\delta$, then it will implies that $\mathbb{P}(R(\hat{h})-R(h^\star) \leq \varepsilon) \geq 1-\delta$. For the 0-1 loss, we have the loss class $\mathcal{L}_\mathcal{H} = \{(x,y) \mapsto \mathbf{1}\{h(x) \neq y\} : h \in \mathcal{H}\}$. Fix a sample $S = ((x_1,y_1),\ldots,(x_n,y_n))$. On this sample, the set of possible loss vectors is

$$\mathcal{L}_\mathcal{H}(S) = \Big\{\big(\mathbf{1}\{h(x_1)\neq y_1\},\ldots,\mathbf{1}\{h(x_n)\neq y_n\}\big) : h \in \mathcal{H}\Big\}.$$

By the XOR argument that the loss is obtained by applying XOR map with $\mathbf{y}$ on $\mathbf{h}(\mathbf{x})$, the map $(h(x_1),\ldots,h(x_n)) \longmapsto (\mathbf{1}\{h(x_1)\neq y_1\},\ldots,\mathbf{1}\{h(x_n)\neq y_n\})$ is a bijection on $\{0,1\}^n$. Therefore:

$$|\mathcal{L}_\mathcal{H}(S)| = |\mathcal{H}|_{x_1^n}|.$$

We then apply the finite-class generalization bound to the finite class of loss vectors on the fixed sample (Theorem 4.1, Lecture 20). Because the loss is bounded in $[0,1]$, we obtain that with probability at least $1-\delta$,

$$R_n(h) - R(h) \leq C\sqrt{\frac{\log\big(2|\mathcal{L}_\mathcal{H}(S)|/\delta\big)}{n}}$$

for some absolute constant C. Moreover:

$$\begin{aligned}\mathbb{P}\left(\sup_{h\in\mathcal{H}} |R_n(h)-R(h)| \leq C\sqrt{\frac{\log\big(2|\mathcal{L}_\mathcal{H}(S)|/\delta\big)}{n}}\right) &\leq \mathbb{P}\left(|R_n(h)-R(h)| \leq C\sqrt{\frac{\log\big(2|\mathcal{L}_\mathcal{H}(S)|/\delta\big)}{n}}\right)\\ &\leq \mathbb{P}\left(R_n(h)-R(h) \leq C\sqrt{\frac{\log\big(2|\mathcal{L}_\mathcal{H}(S)|/\delta\big)}{n}}\right)\end{aligned}$$

which implies $\mathbb{P}\left(\sup_{h\in\mathcal{H}} |R_n(h)-R(h)| \leq C\sqrt{\frac{\log\left(2|\mathcal{L}_\mathcal{H}(S)|/\delta\right)}{n}}\right) \geq 1-\delta$.

Therefore, $R(\hat{h}) - R(h^\star) \leq C\sqrt{\frac{\log\left(2|\mathcal{L}_\mathcal{H}(S)|/\delta\right)}{n}}$ with probability at least $1-\delta$.

Now we bound $|\mathcal{L}_\mathcal{H}(S)|$. Because $VC(\mathcal{H}) = v$, Sauer's lemma gives

$$|\mathcal{H}|_{x_1^n}| \leq \sum_{j=0}^{v}\binom{n}{j} \leq \left(\frac{en}{v}\right)^v$$

Therefore, $|\mathcal{L}_{\mathcal{H}}(S)| = |\mathcal{H}|_{x_1^n}| \leq \left(\frac{en}{v}\right)^v$ and so $\log\big(2|\mathcal{L}_{\mathcal{H}}(S)|/\delta\big) \leq v\log\left(\frac{en}{v}\right) + \log\left(\frac{2}{\delta}\right)$. Substituting into the previous bound yields

$$R(\hat{h}) - R(h^\star) \leq C\sqrt{\frac{v\log(en/v) + \log(2/\delta)}{n}}$$

with probability at least $1-\delta$.

Therefore, if $n \geq C\,\frac{v+\log(1/\delta)}{\varepsilon^2}$ up to the extra logarithmic factor $\log(en/v)$, then $R(\hat{h}) \leq R(h^\star) + \varepsilon$ with probability at least $1-\delta$.

Indeed, if we need to control $R(\hat{h}) - R(h^\star)$ such that $R(\hat{h}) - R(h^\star) \leq \epsilon$, we need $C\sqrt{\frac{v\log(en/v)+\log(2/\delta)}{n}} \leq \varepsilon$. Solve this to obtain n, we get $n \geq C\frac{v\log(en/v)+\log(1/\delta)}{\varepsilon^2}$.

Question 5: Practice with nonparametric regression

a. Loss class versus hypothesis class

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

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

$$\begin{aligned}
\|(f-T)^2 - (g-T)^2\|_\infty &= \|(f-g)(f+g-2T)\|_\infty \\
&= \sup_x |(f-g)(f+g-2T)| \\
&= \sup_x |(f-g)| \cdot |(f+g-2T)|
\end{aligned}$$

We have that for all x, $|f-g| \leq \sup_x |f-g| = \|f-g\|_\infty$ and $|f+g-2T| \leq \sup_x |f+g-2T| = \|f+g-2t\|_\infty$. Therefore, for all x, we have: $|f-g|\cdot|f+g-2T| \leq \|f-g\|_\infty \cdot \|f+g-2T\|_\infty$, which implies that $\sup_x |(f-g)|\cdot|(f+g-2T)| \leq \|f-g\|_\infty \|f+g-2T\|_\infty$ or $\|(f-T)^2 - (g-T)^2\|_\infty \leq \|f-g\|_\infty \|f+g-2T\|_\infty$.

Moreover, because f, g, T takes values in $[0,1]$ for all $x \in [0,1]^d$, we have $|f(x)+g(x)-2T(x)| \leq 2$, which implies $\boxed{\|(f-T)^2 - (g-T)^2\|_\infty \leq 2\|f-g\|_\infty}$.

Let $N := \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon/2)$, which means there exist $f_1, \ldots, f_N \in \mathcal{F}$ such that for every $f \in \mathcal{F}$ $\|f - f_j\|_\infty \leq \varepsilon/2$. Consider the corresponding loss functions

$$\ell_j := (f_j - T)^2, \qquad j = 1, \ldots, N$$

We can show that $\{\ell_1, \ldots, \ell_N\}$ is an ε-cover of $\mathcal{L}_\mathcal{F}$ under $\|\cdot\|_\infty$.

Indeed, take any $\ell \in \mathcal{L}_\mathcal{F}$. Then $\ell = (f-T)^2$ for $f \in \mathcal{F}$. Moreover, for every $f \in \mathcal{F}$, we have f_j such that $\|f - f_j\|_\infty \le \varepsilon/2$. By the inequality proved above, we have:

$$\|\ell - \ell_j\|_\infty = \|(f-T)^2 - (f_j - T)^2\|_\infty \le 2\|f - f_j\|_\infty \le 2 \cdot \frac{\varepsilon}{2} = \varepsilon$$

Therefore $\{\ell_1, \ldots, \ell_N\}$ is an ε-cover of $\mathcal{L}_\mathcal{F}$. We conclude that:

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

b. One-dimensional Lipschitz regression.

By the excess-risk lemma, we have:

$$\begin{aligned}\mathbb{E}\big[R(\hat{f}_n) - R(f^\star)\big] &\le 2\,\mathbb{E}\sup_{f\in\mathcal{F}} |R_n(f) - R(f)| \\ &= 2\,\mathbb{E}\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n (f(X_i) - T(X_i))^2 - \mathbb{E}[(f(X) - T(X))^2]|\end{aligned}$$

We are given the loss class $\mathcal{L}_\mathcal{F} := \{\,(f-T)^2 : f \in \mathcal{F}\,\}$. Therefore, for $g(x) = (f(x) - T(x))^2$, we have:

$$\begin{cases} R_n(f) = \frac{1}{n}\sum_{i=1}^n (f(X_i) - T(X_i))^2 = \frac{1}{n}\sum_{i=1}^n g(X_i) = \mu_n(g) \\ R(f) = \mathbb{E}[(f(X) - T(X))^2] = \mathbb{E}[g(X)] = \mu(g) \end{cases}$$

Therefore:

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

From part (a), we obtain: $\log \mathcal{N}(\mathcal{L}_\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le \log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon/2)$ and with $\mathcal{F} = \mathcal{F}_{L,1}$, we have:

$$\log \mathcal{N}(\mathcal{F}_{L,1}, \|\cdot\|_\infty, \varepsilon/2) \le C\frac{L}{\varepsilon/2} \le C\frac{L}{\varepsilon}.$$

Therefore, we conclude: $\mathbb{E}\big[R(\hat{f}_n) - R(f^\star)\big] \le \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\frac{L}{\varepsilon}}\, d\varepsilon = \frac{C\sqrt{L}}{\sqrt{n}} \int_0^1 \varepsilon^{-1/2}\, d\varepsilon \le C\sqrt{\frac{L}{n}}$.

c. Higher-dimensional Lipschitz regression.

c1. Under the entropy bound $\log \mathcal{N}(\mathcal{F}_{L,d}, \|\cdot\|_\infty, \varepsilon) \le C_d(L/\varepsilon)^d$, we have:

$$\sqrt{\log \mathcal{N}(\mathcal{F}_{L,d}, \|\cdot\|_\infty, \varepsilon)} \le C_d(L/\varepsilon)^{d/2}$$

Because $d \ge 2$, we have $d/2 \ge 1$ and therefore $\int_0^1 \varepsilon^{-d/2}\, d\varepsilon = \infty$. This implies that the naive Dudley bound from part (b) is no longer useful in general.

c2. Let $N_\varepsilon \subset \mathcal{L}_\mathcal{F}$ be an ε-net of $\mathcal{L}_\mathcal{F}$ in $\|\cdot\|_\infty$. For any $g \in \mathcal{L}_\mathcal{F}$, choose $h \in N_\varepsilon$ such that $\|g-h\|_\infty \le \varepsilon$.

Then $\mu_n(g) - \mu(g) = \big(\mu_n(g) - \mu_n(h)\big) + \big(\mu_n(h) - \mu(h)\big) + \big(\mu(h) - \mu(g)\big)$.

By triangle inequality, we have:

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

Moreover, we have:

$$\begin{cases} |\mu_n(g) - \mu_n(h)| = \left|\frac{1}{n}\sum_{i=1}^n (g(X_i) - h(X_i))\right| \le \frac{1}{n}\sum_{i=1}^n |g(X_i) - h(X_i)| \le \|g-h\|_\infty \le \varepsilon, \\ |\mu(h) - \mu(g)| = |\mathbb{E}[h(X) - g(X)]| \le \mathbb{E}|h(X) - g(X)| \le \|h-g\|_\infty \le \varepsilon. \end{cases}$$

Therefore,

$$|\mu_n(g) - \mu(g)| \le 2\varepsilon + |\mu_n(h) - \mu(h)| \le 2\varepsilon + \max_{h \in N_\varepsilon} |\mu_n(h) - \mu(h)|.$$

Take the supremum over $g \in \mathcal{L}_\mathcal{F}$, we conclude: $\boxed{\sup_{g \in \mathcal{L}_\mathcal{F}} |\mu_n(g) - \mu(g)| \le 2\varepsilon + \max_{h \in N_\varepsilon} |\mu_n(h) - \mu(h)|}$.

c3. By the excess-risk lemma:

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

For $g(x) = (f(x) - T(x))^2$, we have $R_n(f) = \mu_n(g)$ and $R(f) = \mu(g)$. Therefore:

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

Using part (c2), we obtain:

$$\mathbb{E}[R(\hat{f}_n) - R(f^\star)] \le C\left(\varepsilon + \mathbb{E}\max_{h\in N_\varepsilon} |\mu_n(h) - \mu(h)|\right)$$

Because N_ε is finite, the finite-class bound yield:

$$\mathbb{E}\max_{h\in N_\varepsilon} |\mu_n(h) - \mu(h)| \le C\sqrt{\frac{\log|N_\varepsilon|}{n}}$$

Choosing N_ε with $|N_\varepsilon| = \mathcal{N}(\mathcal{L}_\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$, we obtain:

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

By part (a), we have:

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

With $\mathcal{F} = \mathcal{F}_{L,d}$, this implies

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

Therefore, we conclude that $\boxed{\mathbb{E}[R(\hat{f}_n) - R(f^\star)] \le C_d\left(\varepsilon + \sqrt{\frac{(L/\varepsilon)^d}{n}}\right)}$.

c4. Choose $\varepsilon \asymp L^{d/(d+2)}n^{-1/(d+2)}$. Therefore:

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

We can see that now both terms are of the same order. Therefore, we can conclude that:

$$\boxed{\mathbb{E}[R(\hat{f}_n) - R(f^\star)] \le C_d\, L^{d/(d+2)}n^{-1/(d+2)}}$$

To make the excess risk at most δ, it suffices to require $C_d\, L^{d/(d+2)}n^{-1/(d+2)} \le \delta$. Equivalently, $n \gtrsim L^d\delta^{-(d+2)}$. We can see that the required sample size grows polynomially in $1/\delta$ with exponent $d+2$, which becomes rapidly worse as d increases. This is an example of the curse of dimensionality.

d. [Bonus] Smoother classes help.

In this part, we assume that $\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le A\varepsilon^{-p} \quad (0 < \varepsilon \le 1)$.

Using the same finite-net argument as in part (c4), we have

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

By part (a), we obtain: $\log \mathcal{N}(\mathcal{L}_\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le \log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon/2) \le A(\varepsilon/2)^{-p} \le CA\varepsilon^{-p}$.

Therefore, $\mathbb{E}[R(\hat{f}_n) - R(f^\star)] \le C\left(\varepsilon + \sqrt{\frac{A\varepsilon^{-p}}{n}}\right) = C\left(\varepsilon + A^{1/2}n^{-1/2}\varepsilon^{-p/2}\right)$.

Similar to part (c4), we choose ε so that the two terms are of the same order, which means:

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

Substitute this choice back into the bound , we have:

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

Suppose $p = d/s$. Therefore, the heuristic rate is: $\mathbb{E}[R(\hat{f}_n) - R(f^\star)] \lesssim n^{-1/(2+d/s)} = n^{-s/(2s+d)}$.

As s increases, the exponent $s/(2s+d)$ increases, , so the rate improves (this can be proved by taking the derivative of $s/(2s+d)$, we can see that the derivative is greater than 0, which implies that $s/(2s+d)$ is an increasing function w.r.t s). When $s \to \infty$, we have: $\frac{s}{2s+d} \to \frac{1}{2}$. In this case, the rate approaches the parametric rate $n^{-1/2}$. This shows that additional smoothness reduces the effective complexity of the class and helps mitigate the curse of dimensionality.