Homework V
Theoretical Statistics/Machine Learning

Sagar Ghosh
Department of Statistics and Data Sciences
The University of Texas at Austin
Email: sg63684@my.utexas.edu

April 10, 2026

1 Practice with Gaussian processes and Gaussian width

We studied suprema of Gaussian and sub-Gaussian processes. For a bounded set $T \subset \mathbb{R}^d$, an especially important quantity is the *Gaussian Width* $w(T) = \mathbb{E}\sup_{t\in T}\langle g, t\rangle$, where $g \sim \mathcal{N}(0, I_d)$. This quantity measures the size of T as seen by a random Gaussian direction. It plays the role of an effective dimension in many high-dimensional problems. This exercise develops some basic properties of Gaussian width and computes it for several canonical sets.

(a) **Basic Properties.** Throughout we would assume that $g \sim \mathcal{N}(0, I_d)$ and $c, C > 0$ are absolute constants.

(a_1) **Monotonicity of Gaussian Width.** For $T \subset S$, the set $\{\langle g, t\rangle : t \in T\} \subset \{\langle g, s\rangle : s \in S\}$, since sup is preserved under set inclusion, we automatically have $\sup_{t\in T}\langle g, t\rangle \leq \sup_{s\in S}\langle g, s\rangle$. Using the monotonicity of $\mathbb{E}$, we finally conclude that $\mathbb{E}\sup_{t\in T}\langle g, t\rangle \leq \mathbb{E}\sup_{s\in S}\langle g, s\rangle$, i.e., $w(T) \leq w(S)$.

(a_2) **Positive Homogeneity.** For every $a \geq 0$, $\langle g, at\rangle = a\langle g, t\rangle$. Hence, $\sup_{t\in T}\langle g, at\rangle = \sup_{t\in T} a\langle g, t\rangle = a\sup_{t\in T}\langle g, t\rangle$. Taking the expectation on both sides and using the homogeneity of $\mathbb{E}$, we conclude $w(aT) = \mathbb{E}\sup_{t\in T}\langle g, at\rangle = a\mathbb{E}\sup_{t\in T}\langle g, t\rangle = aw(T)$.

(a_3) **Translation Invariance.** Under any fixed translation of the set T by $x_0 \in \mathbb{R}^d$, we get $\sup_{t\in T}\langle g, t + x_0\rangle = \langle g, x_0\rangle + \sup_{t\in T}\langle g, t\rangle$, [since the sup does not depend on the translation]. Taking the expectation on both sides and noting that $\mathbb{E}\langle g, x_0\rangle = 0$ [by the law of transformation of normal variable, since $\langle g, x_0\rangle \sim \mathcal{N}(0, \|x_0\|^2)$]. Hence, $w(T + x_0) = \mathbb{E}\sup_{t\in T}\langle g, t + x_0\rangle = \mathbb{E}\sup_{t\in T}\langle g, t\rangle = w(T)$.

(a_4) **Dependency only on the convex hull.** The $\mathrm{Conv}(T) = \{\sum_{i=1}^k a_i t_i : t_i \in T, a_i \geq 0, \sum_{i=1}^k a_i = 1, k \in \mathbb{N}_+\}$. Let's denote the positive unit truncated cuboid $A = \{(a_1, ..., a_k) : a_i \geq 0, \sum_{i=1}^k a_i = 1\}$. Therefore,

$$
\begin{aligned}
w(\mathrm{Conv(T)}) &= \mathbb{E}\sup_{t\in \mathrm{Conv}(T)}\langle g, t\rangle \\
&= \mathbb{E}\sup_{\{a\}\in A}\sup_{\{t_i\}\in T}\langle g, \sum_{i=1}^k a_i t_i\rangle \\
&= \mathbb{E}\left[\sup_{\{a\}\in A}\sum_{i=1}^k a_i \sup_{t_i\in T}\langle g, t_i\rangle\right] \\
&= \mathbb{E}\left[\sup_{t\in T}\langle g, t\rangle \sup_{\{a\}\in A}\sum_{i=1}^k a_i\right] \quad \text{[Since the later sup is independent of the sup over } A\text{]} \\
&= \mathbb{E}\left[\sup_{t\in T}\langle g, t\rangle 1\right] \quad \text{[Since } \sum_{i=1}^k a_i = 1\text{]} \\
&= w(T).
\end{aligned}
$$

(b) **Canonical examples.**

(b_1) We have $w(B_2^d) = \mathbb{E}\sup_{\|t\|_2\leq 1}\langle g,t\rangle \leq \mathbb{E}\sup_{\|t\|_2\leq 1}\|g\|_2\|t\|_2$ [By Cauchy Schwarz inequality]. And the sup is attained at $t = \frac{g}{\|g\|_2}$ by the equality condition of the Cauchy-Schwarcz inequality. Hence, $w(B_2^d) = \mathbb{E}[\langle g, \frac{g}{\|g\|_2}\rangle] = \mathbb{E}\|g\|_2$.

Using the concavity of the $\sqrt{\cdot}$, and invoking Jensen's inequality, we get $\mathbb{E}[\|g\|_2] = \mathbb{E}\left[\sqrt{\sum_{i=1}^d g_i^2}\right] \leq \sqrt{\mathbb{E}\sum_{i=1}^d g_i^2} = \sqrt{d}$, since each $g_i \sim \mathcal{N}(0,1)$. To use the other side, we note that $\sum_{i=1}^d g_i^2 \sim \chi_d^2$. Therefore, $\mathbb{E}\left[\sqrt{\sum_{i=1}^d g_i^2}\right] \geq \mathbb{E}\left[\sqrt{\sum_{i=1}^d g_i^2 \mathbb{1}_{g_i^2\leq r}}\right] \geq \mathbb{E}\left[\sqrt{\sum_{i=1}^d r}\right] = r\sqrt{d}$, for some $r > 0$. Therefore, $r\sqrt{d} \leq w(B_2^d) \leq \sqrt{d}$.

(b_2) We note that for $\|t\|_1 \leq 1$, $\sup\langle g,t\rangle \leq \|g\|_\infty$ and the sup is attained at exactly $t = \{0,...,1,...,0\}$ for the coordinate where $\|g\|_\infty = g_i$. Therefore, $w(B_1^d) = \mathbb{E}\sup_{\|t\|_1\leq 1}\langle g,t\rangle = \mathbb{E}\|g\|_\infty$.

Let $\|g\|_\infty = |g_j|$ for some $j \in \{1,2,...,d\}$. Since g_j is Gaussian, $|g_j|$ is also Gaussian, hence we have $\|g_j\|_{\psi_2} \leq c$ for some absolute constant $c > 0$. Using Jensen's inequality and union bound, we get for any $\theta \in \mathbb{R}$, $\mathbb{E}|g_j| = \frac{1}{\theta}\mathbb{E}\log(e^{\theta|g_j|}) \leq \frac{1}{\theta}\log\mathbb{E}(\sum_{i=1}^d e^{\theta|g_i|}) \leq \frac{1}{\theta}[\log d + \frac{c^2\theta^2}{2}]$. Minimizing over θ on the RHS, we get $\mathbb{E}|g_j| \leq c'\sqrt{\log d}$. Hence, we have the upper bound.

To derive the lower bound we note that for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\{\max_{1\leq i\leq d}|g_i| \geq t\} &= 1 - \mathbb{P}\{\max_{1\leq i\leq d}|g_i| < t\} \\
&= 1 - \prod_{i=1}^d \mathbb{P}(|g_i| < t) \\
&= 1 - [\mathbb{P}(|h| < t)]^d \quad \text{[By independence, where } h \sim \mathcal{N}(0,1)] \\
&= 1 - [1 - \mathbb{P}(|h| \geq t)]^d .
\end{aligned}$$

And for $t = \sqrt{\log d}$, $1 - [1 - \mathbb{P}(|h| \geq t)]^d \geq 1 - e^{-c_1 d^2}$ for some constant $c_1 > 0$. Combining everything and using Markov's inequality, we get $\mathbb{E}\|g\|_\infty \geq t\mathbb{P}[\|g\|_\infty \geq t] \geq c_1'\sqrt{\log d}$, for $t = \sqrt{\log d}$, for some constant $c_1' > 0$. Therefore, $w(B_1^d) = \mathbb{E}\|g\|_\infty \asymp \sqrt{\log d}$.

(b_3) Let $T = \{t_1,\ldots,t_M\} \subset \mathbb{R}^d$ be a finite set. The random variables $\langle g,t_j\rangle \sim \mathcal{N}(0,\|t_j\|^2)$. Therefore, for all $j \in \{1,...,M\}$, $\|\langle g,t_j\rangle\|_{\psi_2} \leq (\max_{1\leq j\leq M}\|t_j\|_2) = L$ (say), hence we can use L as the variance proxy for all $\langle g,t_j\rangle$. Now, using the Maximal inequality for finite many sub-Gaussian random variables, $w(T) = \mathbb{E}\sup_{t\in T}\langle g,t\rangle \leq CL\sqrt{\log M}$. Putting the value of L, we conclude that $w(T) \lesssim (\max_{1\leq j\leq M}\|t_j\|_2)\sqrt{2\log M}$.

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

(c_1) Let $Q = \{v_1, v_2, \ldots, v_N\} \subset \mathbb{R}^d$. Then using ($a_4$), we get $w(P) = w(Q) = \mathbb{E}\sup_{q\in Q}\langle g,q\rangle$. Further using part

(b_3), we get $w(Q) \lesssim (\max_{1\le j\le M} \|v_j\|_2)\sqrt{\log N} \lesssim \sqrt{\log N}$, since $(\max_{1\le j\le M} \|v_j\|_2) \le 1$.

(c_2) For each support $S = \{a_1, \ldots, a_d\}$ with exactly k many of the a_i s are 1, rest are 0, i.e., with $|S| = k$, we define $Y_S = \sup\{\langle g, x\rangle, \text{Supp}(x) \subset S, \|x\|_2 \le 1\} = \|g_S\|_2$. Then, $\mathbb{E}[Y_S] \le \sqrt{EY_S^2} \le \sqrt{k}$, since exactly k many components from g participate in forming the random variables in Y_Ss. Therefore, $Y_s - \mathbb{E}Y_S$ s are sub-gaussian with $\|Y_S - \mathbb{E}Y_S\|_{\psi_2} \le (\sup_{x\in S^{d-1}} \|x\|_2\|g\|_{\psi_2}) \lesssim c$ for some absolute constant $c > 0$. Finally using the maximal inequality over at most $\binom{d}{k} \le \left(\frac{ed}{k}\right)^k$ pairs, we get $w(T_k) \le c\sqrt{\log \binom{d}{k}} \le c\sqrt{k \log\left(\frac{ed}{k}\right)}$.

2 Practice with Rademacher processes and Rademacher complexity

Let $x_1, ..., x_n \in \mathcal{X}$ be fixed sample points, and let $\epsilon_1, ..., \epsilon_n$ be independent Rademacher random variables. For a class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$, define the empirical Rademacher complexity $\hat{\mathcal{R}}_n(\mathcal{F}; x_1, ..., x_n) := \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)$. This exercise develops some of the standard bounds and examples.

(a) **Basic Properties.** Let $\mathcal{F}$ and $\mathcal{G}$ be two function classes and let $a \geq 0$.

(a_1) Let $\mathcal{F} \subset \mathcal{G}$. Then the set inclusion holds true : $\{\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) \mid f \in \mathcal{F}\} \subset \{\frac{1}{n}\sum_{i=1}^n \epsilon_i g(x_i) \mid g \in \mathcal{G}\}$. Using the monotonicity of the sup under set inclusion, we get $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) \leq \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(x_i)$. taking the expectation on both sides, we conclude that $\hat{\mathcal{R}}_n(\mathcal{F}; x_1, ..., x_n) = \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) \leq \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(x_i) = \hat{\mathcal{R}}_n(\mathcal{G}; x_1, ..., x_n)$.

(a_2) Let $a\mathcal{F} = \{af \mid f \in \mathcal{F}\}$. Then, $\hat{\mathcal{R}}_n(a\mathcal{F}) = \mathbb{E}_\epsilon \sup_{r\in a\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i r(x_i) = \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} a\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) = a\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) = a\hat{\mathcal{R}}_n(\mathcal{F})$, since every $r \in a\mathcal{F}$ is of the form $r = af$ for some $f \in \mathcal{F}$.

(a_3) We can write the convex hull of $\mathcal{F}$ as $\text{Conv}(\mathcal{F}) = \{\sum_{r=1}^k a_r f_r; f_r \in \mathcal{F}, a_r \geq 0, \sum_{r=1}^k a_r = 1; k \in \mathbb{N}_+\}$. Following the same notation from (a_4) [Qs 1], we get

$$\begin{aligned}
\hat{\mathcal{R}}_n(\text{Conv}(\mathcal{F})) &= \mathbb{E}_\epsilon \sup_{g\in\text{Conv}(\mathcal{F})} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(x_i) \\
&= \mathbb{E} \sup_{\{f_r\}\in\mathcal{F}} \sup_{\{a_r\}\in A} \frac{1}{n}\sum_{i=1}^n \sum_{r=1}^k \epsilon_i a_r f_r(x_i) \\
&= \sup_{\{a_r\}\in A} \sum_{r=1}^k a_r \mathbb{E}_\epsilon \sup_{\{f_r\}\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f_r(x_i) \quad \text{[Since the } \mathbb{E} \text{ is independent of the } a_i\text{s]} \\
&= \sup_{\{a_r\}\in A} \sum_{r=1}^k a_r \hat{\mathcal{R}}_n(\mathcal{F}) \\
&= \hat{\mathcal{R}}_n(\mathcal{F}) \sup_{\{a_r\}\in A} 1 \\
&= \hat{\mathcal{R}}_n(\mathcal{F}).
\end{aligned}$$

(b) **Finite-class bound (Massart-type bound).** We know that ϵ_is are sub-Gaussian [they are Rademacher variables], and $\mathbb{E}_{\epsilon_i}[\epsilon_i f(x_i)] = f(x_i)\mathbb{E}_{\epsilon_i}[\epsilon_i] = 0$, since x_is are fixed sample points. We also compute the variance of $\sum_{i=1}^n \epsilon_i f(x_i)$, which is $\text{Var}_\epsilon\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(\epsilon_i f(x_i)) = \frac{1}{n^2}\sum_{i=1}^n f(x_i)^2 \leq \frac{r^2}{n}$ [By the stated condition]. Now, we have that $\{\frac{1}{n}\sum_{\epsilon_i} f(x_i) \mid f \in \mathcal{F}\}$ are sub-gaussian random variables with $\|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\|_{\psi_2} \leq \frac{r}{\sqrt{n}}$. Therefore, using the maximal inequality for finite many sub-gaussian random vari-

ables over a finite class $\mathcal{F}$, we get

$$\hat{\mathcal{R}}_n(\mathcal{F}, x_1, ..., x_n) \leq c\frac{r}{\sqrt{n}}\sqrt{\log|\mathcal{F}|} \lesssim r\sqrt{\frac{\log|\mathcal{F}|}{n}}.$$

(c) **Linear function classes.** Let $\mathcal{X} = \mathbb{R}^d$, and consider the class $\mathcal{F}_R := \{x \to \langle w, x\rangle : \|w\|_2 \leq R\}$. By a straightforward computation, we can show that

$$\begin{aligned}
\hat{\mathcal{R}}_n(\mathcal{F}_R; x_1, ..., x_n) &= \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}_R} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i) \\
&= \mathbb{E}_\epsilon \sup_{w:\|w\|_2\leq R} \frac{1}{n}\sum_{i=1}^n \epsilon_i\langle w, x_i\rangle \\
&= \frac{1}{n}\mathbb{E}_\epsilon \sup_{w:\|w\|_2\leq R} \langle w, \sum \epsilon_i x_i\rangle \quad \text{[By the linearity of inner product]} \\
&= \frac{R}{n}\mathbb{E}_\epsilon \left\|\sum_{i=1}^n \epsilon_i x_i\right\|_2 \quad \text{[Since the sup is obtained at } w = R\frac{\sum_{i=1}^n \epsilon_i x_i}{\|\sum_{i=1}^n \epsilon_i x_i\|_2}\text{]} \\
&\leq \frac{R}{n}\sqrt{\mathbb{E}_\epsilon\|\sum_{i=1}^n \epsilon_i x_i\|_2^2} \quad \text{[By Cauchy-Schwarz inequality]} \\
&= \frac{R}{n}\sqrt{\sum_{i=1}^n \mathbb{E}_\epsilon\epsilon_i^2\|x_i\|_2^2} \quad \text{[By independence of the } \epsilon_i\text{s]} \\
&= \frac{R}{n}\sqrt{\sum_{i=1}^n \|x_i\|_2^2}.
\end{aligned}$$

(d) **[Bonus] Sparse linear predictors.** Let $\mathcal{G}_R := \{x \to \langle w, x\rangle : \|w\|_1 \leq R\}$ and we assume that $\|x_i\|_\infty \leq 1$ for all $i \in \{1, ..., n\}$. At first we note that the $\sup_{w:\|w\|_1\leq R}\langle w, z\rangle = R\|z\|_\infty$, and the sup is attained at setting the coordinates of w to be 0 except one coordinate at R, exactly where $\|z\|_\infty$ is attained. Therefore,

$$\begin{aligned}
\hat{\mathcal{R}}_n(\mathcal{G}_R; x_1, ..., x_n) &= \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}_R} \frac{1}{n}\sum_{i=1}^n \epsilon_i g(x_i) \\
&= \mathbb{E}_\epsilon \sup_{w:\|w\|_1\leq R} \frac{1}{n}\sum_{i=1}^n \epsilon_i\langle w, x_i\rangle \\
&= \mathbb{E}_\epsilon \sup_{w:\|w\|_1\leq R} \frac{1}{n}\langle w, \sum_{i=1}^n \epsilon_i x_i\rangle \\
&= \frac{R}{n}\mathbb{E}_\epsilon \left\|\sum_{i=1}^n \epsilon_i x_i\right\|_\infty
\end{aligned}$$

Now to control the last quantity, we note that each $\epsilon_i x_i$ is sub-gaussian with $\|\epsilon_i x_i\|_\infty \leq \|x_i\|_\infty \leq 1$, and $\|\epsilon_i x_i\|_{\psi_2} \leq \|x_i\|_2 \leq \sqrt{n}$, since $\|x\|_2 \leq \sqrt{n}\|x\|_\infty \leq \sqrt{n}$. Therefore, the expression $\mathbb{E}_\epsilon \|\sum_{i=1}^n \epsilon_i x_i\|_\infty$ is reduced to finding the expectation over d many sub-Gaussian random variables, each of which has a variance proxy of $\sqrt{n}$. Finally using the maximal inequality for finite many sub-Gaussian random variables, we get $\mathbb{E}_\epsilon \|\sum_{i=1}^n \epsilon_i x_i\|_\infty \leq$

$c\sqrt{n}\sqrt{\log d}$, for some constant $c > 0$. Putting everything together, we conclude that

$$\begin{aligned}
\hat{\mathcal{R}}_n(\mathcal{G}_R; x_1, ..., x_n) &= \frac{R}{n}\mathbb{E}_\epsilon \left\| \sum_{i=1}^{n} \epsilon_i x_i \right\|_\infty \\
&\leq \frac{R}{n} c\sqrt{n}\sqrt{\log d} \\
&= cR\sqrt{\frac{\log d}{n}}.
\end{aligned}$$

3 Practice with VC dimension

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

(a) **One-dimensional classes.**

(a_1) **The Class of Half Lines.** $\mathcal{H}_{\text{half}} = \{\mathbf{1}_{(-\infty,a]}; a \in \mathbb{R}\}$. It is clear that every 1 point set can be shattered by $\mathcal{H}_{\text{half}}$. Let's call the point x. If the label of x is 0, then we can take $a < x$, else we take $a > x$. So, $\text{vc}(\mathcal{H}_{\text{half}}) \geq 1$. Now take two points x_1, x_2, assume (WLOG) $x_1 < x_2$. Then the labellings $x_1 \mapsto 0$ and $x_2 \mapsto 1$, can not be realized by any labelling in $\mathcal{H}_{\text{half}}$, otherwise we need to find a such that $a < x_1$ and $a > x_2$, i.e., $x_2 < a < x_1$, contradicting the assumption that $x_1 < x_2$. Therefore, $\text{vc}(\mathcal{H}_{\text{half}}) = 1$.

(a_2) **The Class of Intervals.** $\mathcal{H}_{\text{int}} = \{\mathbf{1}_{[a,b]}, a, b \in \mathbb{R}, a \leq b\}$. Take two points $x_1 < x_2$. For the pairs of labellings, we can choose the following a, b such that the labels can be realized by $\mathcal{H}_{\text{int}}$:

(x_1, x_2)	$(0, 0)$	$a \leq b \leq x_1 < x_2$
(x_1, x_2)	$(1, 0)$	$a \leq x_1 \leq b < x_2$
(x_1, x_2)	$(0, 1)$	$x_1 < a \leq x_2 \leq b$
(x_1, x_2)	$(1, 1)$	$a \leq x_1 < x_2 \leq b$

Table 1: Choice of a, b for two labeling by $\mathcal{H}_{\text{int}}$

Now take three points $x_1 < x_2 < x_3$. Consider the labeling $x_1 \mapsto 1, x_2 \mapsto 0, x_3 \mapsto 1$. Then, we need to choose $a \leq x_1 \leq b < x_2$, but once we choose $b < x_2$, we can not realize the labeling 1 of x_3 by any means. Hence, $\text{vc}(\mathcal{H}_{\text{int}}) = 2$.

(a_3) **The Class of Union of at most k Intervals.** Suppose $\mathcal{H}_k = \left\{\mathbf{1}_{\cup_{i=1}^{k}[a_i,b_i]}; a_i \leq b_i, a_i, b_i \in \mathbb{R}\right\}$. In part ($a_2$), we showed that for $k = 1$, $\text{vc}(H_1) = 2 \times 1 = 2$. We will proceed by induction. Suppose the result holds for some $1 \leq r < k$, i.e., $\text{vc}(\mathcal{H}_r) = 2r$. Consider $\mathcal{H}_{r+1}$. For the labellings of any set of $2(r + 1)$ points $x_1 < ... < x_{2r+1}$, the points $x_1 < ... < x_{2r}$ can be realized by some function $h_r = \{\mathbf{1}_{\cup_{i=1}^{r}[a_i,b_i]}\}$ in $\mathcal{H}_r$. To realize the labels of $x_{2r+2} < x_{2r+2}$, we can again apply the technique in (a_2) to make it realize by some function $h_1 = \{\mathbf{1}_{[a_{r+1},b_{r+1}]}\}$ in $\mathcal{H}_1$. Now, if we make the intervals in h_r and h_1 disjoint [which we can easily do, by choosing $b_r < a_{r+1}$], then $h_r \cup h_1$ can realize the labeling of $\{x_1, ..., x_{2r+2}\}$. But for $2r + 3$ points $x_1 < ... < x_{2r+3}$, the labeling $\{1, 0, 1, 0, ..., 1\}$ can not be realized by any function in h_{r+1}, for this if the

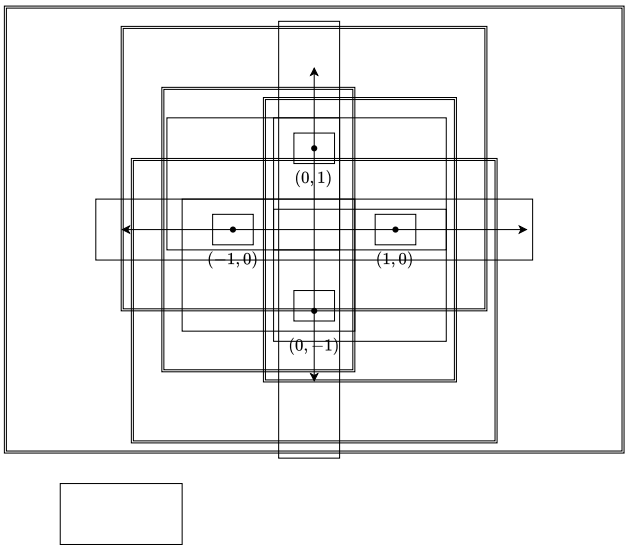

Figure 1: Visual Illustration of the VC Dimension of $\mathcal{H}_{\mathcal{R}}$

last disjoint interval is $[a_{r+1}, b_{r+1}]$, then $a_{r+1} \leq x_{2r+1} \leq b_{r+1} < x_{2r+2}$, but once we choose $b_{r+1} < x_{2r+2}$, we can not realize the labelling 1 of x_{2r+3} by any means, therefore, $\text{vc}(\mathcal{H}_{r+1}) = 2(r+1)$, or $\text{vc}(\mathcal{H}_k) = 2k$.

(b) **Two-dimensional classes.**

(b_1) **Axis-aligned rectangles.** We will give a visual illustration to show that $\text{vc}(\mathcal{H}_{\mathcal{R}}) = 4$. We take four points in Figure 1 namely $(1,0), (0,1), (-1,0), (0,-1)$. There will be at most $16 (= 2^4)$ different labeling possible for this set of four points. We enclose all combination of these 4 points using axis aligned rectangles, specifically, the single rectangles either enclose 1 or 2 points, indicating when they have labels 1 and the rest have 0. The double rectangles enclose every 3 combinations of these 4 points, indicating these 3 have labels 1, and the other one has label 0. The largest double rectangle will be chosen when all of them have labels 1, and the completely outside rectangle will be chosen when all of them have labels 0. The details are in Figure 1.

For the upper bound, we consider 5 points, p_1, p_2, p_3, p_4, p_5. Among these, WLOG we assume that p_1, p_2, p_3, p_4 are the points having the min and max of both x and y coordinates. Then the labels $p_1 \mapsto 1, p_2 \mapsto 1, p_3 \mapsto 1, p_4 \mapsto 1, p_5 \mapsto 0$, can not be realized by any function in $\mathcal{H}_{\mathcal{R}}$, since any axis-aligned rectangle enclosing p_1, p_2, p_3, p_4 will enclose p_5 as well, since both the x and y coordinates of p_5 lie within the ranges of the coordinates of p_1, p_2, p_3, and p_4.

(b_2) **[Bonus] Convex sets.** Let $\mathcal{C}_{\text{conv}}$ be the class of all indicators on convex subsets in $\mathbb{R}^2$. Choose any

$m \in \mathbb{N}_+$, and take m points on the unit circle in $\mathbb{R}^2$. Now, among these m points, let's say ℓ points have labels 1 and rest have labels 0. Since these points are on the circle, we can cyclically order these ℓ points, such as $p_1, ..., p_\ell$, where p_1 has the least tangent argument with the x axis and p_ℓ has the highest tangent argument with the x axis. If we now consider the polygon by constructing the sides $\overline{p_1, p_2}, \overline{p_2, p_3}, ..., \overline{p_\ell, p_1}$, then it is the convex hull of the points $p_1, ..., p_\ell$. And of course it excludes the points which has label 0, since any point having label 0 must lie on the circular chord joining p_i, p_{i+1}, and the line joining $\overline{p_i, p_{i+1}}$ lies closer to origin than the point having label 0. Therefore, the indicator on this polygon will realize the labeling of these m points, and the indicator on this polygon is of course a member in $\mathcal{C}_{\text{conv}}$. Since we have started with an arbitrary labeling, any labeling of these m points can be realized by the class $\mathcal{C}_{\text{conv}}$. Therefore, $\text{vc}(\mathcal{C}_{\text{conv}}) \geq m$ for any $m \in \mathbb{N}_+$. Since $m \in \mathbb{N}_+$ is arbitrary, we have $\text{vc}(\mathcal{C}_{\text{conv}}) = \infty$.

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

(c_1) For the ball $B(x_0, r)$, centered at x_0 with radius r, the points $x \in B(x_0, r)$ satisfies $\|x - x_0\|^2 \leq r^2 \implies -2x_0^\top x + \|x\|^2 \leq -\|x_0\|^2 + r^2$, so if we apply the lifting $x \mapsto (x, \|x\|^2) \in \mathbb{R}^{d+1}$, then the relationship in $\mathbb{R}^{d+1}$ becomes of the form $w^\top x \leq c$, where $w = (-2x_0, 1), c = r^2 - \|x_0\|^2$, which are the equations of half planes. Therefore, computing the VC dimension of balls in $\mathbb{R}^d$ is reduced to computing the VC dimension for half spaces in $\mathbb{R}^{d+1}$.

In the light of the above discussion, we will show that the indicators on the class of half spaces in $\mathbb{R}^d$ has VC dimension d. Let's call the class of half spaces as $\mathcal{H}_{\text{half}} = \{\mathbf{1}_{w^\top x + b \leq 0} \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$ [we can take $b = 1$ by normalizing if $b \neq 0$.]. Consider the set of points as the collection of unit vectors $\{e_1, ..., e_d\}$. Suppose $\{i_1, ..., i_k\} \in \{1, 2, ..., d\}$ are the indices for which the labels of $e_{i_r} \mapsto 1$ and the rest maps to 0. Then, we can choose $w_{i_r} = -1, b = 0$, and the other coordinates of w to be 0, then, we see that this hyperplane can realize any arbitrary labels of the points. Hence, $\text{vc}(\mathcal{H}_{\text{half}}) \geq d$.

Now we will show that there are a set of $d + 1$ points for which the indicators on this class $\mathcal{H}_{\text{half}}$ can not shatter the set of points. To this end, consider the set of points $\{e_1, ..., e_d, p\}$ where $p = \sum_{i=1}^d \frac{1}{d} e_i$. Consider the labeling: $e_i \mapsto 1 \forall i$ and $p \mapsto 0$. Therefore, $w^\top e_i + b \leq 0$ for all i, this means $\sum_{i=1}^d (w^\top e_i + b) \leq 0 \implies w^\top \sum_{i=1}^d \frac{1}{d} e_i + b \leq 0$ [By dividing both sides by d], i.e., $w^\top p + b \leq 0$. Hence, the label $p \mapsto 0$ can not be realized by the indicators on the class $\mathcal{H}_{\text{half}}$. Therefore, $\text{vc}(\mathcal{H}_{\text{half}}) = d$. As per our initial comments, the vc dimension of the indicators on class of all balls in $\mathbb{R}^d$ is $d + 1$.

(c_2) Suppose $\mathcal{H}$ is a Boolean class with $\text{vc}(\mathcal{H}) = v$. Then, any labeling of a set of v points can be realized by $\mathcal{H}$. So, if we are given n points, then we can realize the distinct labels of at most v many points by the class $\mathcal{H}$. Now, it is easy to see that the no of distinct labeling in $\{0, 1\}$ by any functions in $\mathcal{H}$ can be $\sum_{i=0}^{v} \binom{n}{i}$, since we can choose i points among n points and label them 1 and rest 0, but i can go up to at most v, since this is the maximum no of points that can be realized by $\mathcal{H}$. Finally using a simple binomial relation, we can get $\sum_{i=0}^{v} \binom{n}{i} \leq \left(\frac{en}{v}\right)^v$

(c_3) Using part (b_1), we know that the axis aligned rectangles in $\mathbb{R}^2$ has vc dimension 4. Plugging in 4 in part (c_2), we get the no of distinct labeling produced by n points in $\mathbb{R}^2$ that can be realized by axis-aligned rectangles is $\sum_{i=0}^{4} \binom{n}{i}$.

4 Practice with statistical learning theory

We now connect empirical processes to statistical learning. The setup is the standard supervised learning model. Let (X, Y) be a random pair taking values in $\mathcal{X} \times \mathcal{Y}$, and let $(X_1, Y_1), ..., (X_n, Y_n)$ 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 $h \in \mathcal{H}$, we define the population risk $R(h) = \mathbb{E}[\ell(h(X), Y)]$ and the empirical risk $R_n(h) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\ell(h(X_i), Y_i)]$. Let $\hat{h} \in \arg\min_{h\in\mathcal{H}} R_n(h), h^* = \arg\min_{h\in\mathcal{H}} R(h)$. This exercise derives the excess risk bound via symmetrization, Rademacher Complexity and VC Dimension.

(a) **Excess-risk lemma.** We let $\epsilon = \sup_{h\in\mathcal{H}} |R_n(h) - R(h)|$. Hence, for all $h \in \mathcal{H}$, $R_n(h) \leq R(h) + \epsilon ...(I)$, and $R(h) \leq R_n(h) + \epsilon ...(II)$

Using these two equations, we get

$$\begin{aligned} R(\hat{h}) &\leq R_n(\hat{h}) + \epsilon \quad [\text{Using } h = \hat{h} \text{ in } (II)] \\ &\leq R_n(h^*) + \epsilon \quad [\text{Since } R_n(\hat{h}) \leq R_n(h^*), \text{ as } \hat{h} \text{ is the empirical risk minimizer}] \\ &\leq R(h^*) + \epsilon + \epsilon \quad [\text{Putting } h = h^* \text{ in } (I)] \\ &= R(h^*) + 2\epsilon. \end{aligned}$$

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

(b) **Symmetrization.** We will start with a ghost sample $(X_1', Y_1'), ..., (X_n', Y_n')$ independent copies of $(X_1, Y_1), ..., (X_n, Y_n)$. Since $(X_i', Y_i') \overset{d}{=} (X_i, Y_i)$, we have $\mathbb{E}[\ell(h(X_i), Y_i)] = \mathbb{E}[\ell(h(X_i'), Y_i')]$. Therefore,

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

Now we apply Jensen's inequality conditionally on the pairs $(X_1, Y_1), ..., (X_n, Y_n)$ to get

$$(X_1, Y_1), ..., (X_n, Y_n) \mapsto \sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^n [\ell(h(X_i), Y_i) - \mathbb{E}\ell(h(X_i), Y_i)] \right|$$

is convex, hence $\mathbb{E}\sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n [\ell(h(X_i), Y_i) - \mathbb{E}\ell(h(X_i), Y_i)]\right| \leq \mathbb{E}\sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^n [\ell(h(X_i), Y_i) - \mathbb{E}\ell(h(X_i'), Y_i')]\right|$.

Now because of the identical distributions of the ghost sample, we get

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

Finally, one more triangle inequality gives us

$$\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i[\ell(h(X_i),Y_i)-\ell(h(X_i'),Y_i')]\right| \le \left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell(h(X_i),Y_i)\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ell(h(X_i'),Y_i')\right|.$$

Finally, taking the $\mathbb{E}$ on both sides, we get

$$\mathbb{E}\sup_{h\in\mathcal{H}}|R_n(h)-R(h)| = \mathbb{E}\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}[\ell(h(X_i),Y_i)-\mathbb{E}\ell(h(X_i),Y_i)]\right| \le 2\mathbb{E}\hat{\mathcal{R}}_n(\mathcal{L}_{\mathcal{H}};(X_1,Y_1),...,(X_n,Y_n)).$$

(c) **Finite hypothesis classes.** Combining part (a) and part (b) of Problem 4, we get

$$\begin{aligned}\mathbb{E}[R(\hat{h})-R(h^*)] &\le 2\mathbb{E}\sup_{h\in\mathcal{H}}|R_n(h)-R(h)| \\ &\le 4\mathbb{E}\hat{\mathcal{R}}_n(\mathcal{L}_{\mathcal{H}};(X_1,Y_1),...,(X_n,Y_n)).\end{aligned}$$

Since the loss function is bounded in $[0,1]$, we have $\frac{1}{n}\sum_{i=1}^{n}\ell(h(X_i),Y_i)^2 \le 1$, therefore, using part $2(b)$, we get

$$\begin{aligned}\mathbb{E}[R(\hat{h})-R(h^*)] &\le 2\mathbb{E}\sup_{h\in\mathcal{H}}|R_n(h)-R(h)| \\ &\le 4\mathbb{E}\hat{\mathcal{R}}_n(\mathcal{L}_{\mathcal{H}};(X_1,Y_1),...,(X_n,Y_n)) \\ &\le 4\cdot 1\cdot\sqrt{\frac{\log|\mathcal{L}_{\mathcal{H}}|}{n}} \\ &\le 4\sqrt{\frac{\log|\mathcal{H}|}{n}} \quad [\text{Since } |\mathcal{L}_{\mathcal{H}}|\le|\mathcal{H}|].\end{aligned}$$

(d) **Boolean classification and VC dimension.** Since the VC Dimension of $\mathcal{H}$ is v, the cardinality of $|\mathcal{H}|$ is essentially the no of distinct boolean labels that can be realized on the samples $(X_1,Y_1),...,(X_n,Y_n)$, which is at most $\left(\frac{en}{v}\right)^v$ [by part (c_2), problem 3]. Hence, $\log|\mathcal{H}| \le \log\left(\frac{en}{v}\right)^v = v\log(en/v)$. Plugging in this bound in part (c) of Problem 4, we get

$$\mathbb{E}[R(\hat{h})-R(h^*)] \le 4\sqrt{\frac{\log|\mathcal{H}|}{n}} \le 4\sqrt{\frac{v\log(en/v)}{n}}.$$

We can also argue it through the labels on the sample obtained by XOR with the fixed label vectors $(y_1,...,y_n)$, which will give us $\left|\mathcal{L}_{\mathcal{H}}\mid_{\{(X_i,Y_i)\}_{i=1}^n}\right| = \left|\mathcal{H}\mid_{\{X_i,Y_i)\}_{i=1}^n}\right|$, which essentially says that the loss classes have the same complexity in case of $0-1$ loss.

(e) **[Bonus] A VC-style sample complexity statement.** Applying the Hoeffding's inequality on the class of all loss functions give us

$$\mathbb{P}(|R_n(h)-R(h)|\ge t) \le 2e^{-cnt^2},$$

for some $c > 0$. Using the union bound of $\mathcal{H}$ and the vc dimension of $\mathcal{H}$ gives us

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |R_n(h) - R(h)| \geq t\right] \leq 2|\mathcal{H}|e^{-cnt^2} \leq 2\left(\frac{en}{v}\right)^v e^{-cnt^2}.$$

Putting $t = \epsilon$ and solving for ϵ in the relation $2\left(\frac{en}{v}\right)^v e^{-cn\epsilon^2} = \delta$, we get

$$\begin{aligned} \log 2 + v\log\left(\frac{en}{v}\right) - cn\epsilon^2 &= \log\delta \\ \implies n = \frac{v\log\frac{en}{v} + \log(1/\delta)}{\epsilon^2} &> \frac{v + \log(1/\delta)}{\epsilon^2}, \end{aligned}$$

since $en/v > c'$ for some constant $c' > 0$.

Rewriting in terms of the probability bound and using the fact that $R(\hat{h}) - R(h^*) \leq 2\sup_{h\in\mathcal{H}} |R_n(h) - R(h)|$, we get with probability at least $1 - \delta$,

$$R(\hat{h}) \leq R(h^*) + \epsilon, \quad [\text{for } n > \frac{v + \log(1/\delta)}{\epsilon^2}].$$

5 Practice with nonparametric regression

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 non-parametric regression model. We will see how excess-risk bounds can be derived from covering number. 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 $[0,1]^d$ following the law μ and let $((X_1, TX_1)), ..., (X_n, T(X_n)))$ be noiseless training data, where $X_1, ..., X_n$ 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 the empirical risk by

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

Let $f^* \in \arg\min_{f\in\mathcal{F}} R(f)$, and $\hat{f}_n \in \arg\min_{f\in\mathcal{F}} R_n(f)$ denote a population risk minimizer and an empirical risk minimizer. We will use the following facts proved earlier:

- **Excess Risk Lemma:**

$$R(\hat{f}_n) - R(f^*) \le 2\sup_{f\in\mathcal{F}} |R_n(f) - R(f)|.$$

- **Empriical Process by Dudley bound:** If $\mathcal{G}$ is a class of functions taking values in $[0,1]$, then

$$\mathbb{E}\sup_{g\in\mathcal{G}} |\mu_n(g) - \mu(g)| \le \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{G}, \|\cdot\|_\infty)}d\epsilon.$$

- **Finite Calss Bound:** If $\mathcal{G}$ is a finite class of functions taking values in $[0,1]$, then

$$\mathbb{E}\sup_{g\in\mathcal{G}} |\mu_n(g) - \mu(g)| \le C\sqrt{\frac{\log|\mathcal{G}|}{n}}.$$

For $L > 0$, define the Lipschitz class

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

We may also use the following covering number bounds without proof:

$$\log \mathcal{N}(\epsilon, \mathcal{F}_{L,1}, \|\cdot\|_\infty) \le C\frac{L}{\epsilon}, \quad 0 < \epsilon \le 1.$$
$$\log \mathcal{N}(\epsilon, \mathcal{F}_{L,d}, \|\cdot\|_\infty) \le C_d\left(\frac{L}{\epsilon}\right)^d, \quad 0 < \epsilon \le 1,$$

where C_d may depend on the ambient dimension d.

(a) **Loss class versus hypothesis class.** Let $\mathcal{L}_{\mathcal{F}} := \{x \mapsto (f(x) - T(x))^2 \mid f \in \mathcal{F}\}$. Therefore, for any $f, g \in \mathcal{F}$ and $x \in \mathcal{X}$, we have

$$(f-T)^2(x) - (g-T)^2(x) = (f(x) - g(x))(f(x) + g(x) - 2T(x)).$$

since f, g, T all take values in $[0,1]$, $|f(x) + g(x) - 2T(x)| = |(f(x) - T(x)) + (g(x) - T(x))| \leq |f(x) - T(x)| + |g(x) - T(x)| \leq 2$, since for $p, q \in [0,1]$, $|p - q| \leq 1$. Finally, taking the sup w.r.t. x on both sides we get the $\|\cdot\|_\infty$ norm bound:

$$\|(f-T)^2 - (g-T)^2\|_\infty = \sup_{x \in \mathcal{X}} \|(f-T)^2(x) - (g-T)^2(x)\| \leq 2 \sup_{x \in \mathcal{X}} \|f(x) - g(x)\| = 2\|f - g\|_\infty.$$

(b) **One-dimensional Lipschitz regression.** Using the excess-risk lemma, we have

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

Since the class of functions in $\mathcal{F}$ take values in $[0,1]$, using the empirical process by Dudley bound, we get

$$\mathbb{E} \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| \leq \frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)} d\epsilon.$$

and hence,

$$\mathbb{E}[R(\hat{f}_n) - R(f^*)] \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| \leq 2\frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)} d\epsilon.$$

Finally, the covering number bound for the Lipschitz class [for $d = 1$, $\mathcal{F} = \mathcal{F}_{L,1}$] gives us

$$\log \mathcal{N}(\epsilon, \mathcal{F}_{L,1}, \|\cdot\|_\infty) \leq C\frac{L}{\epsilon}.$$

Plugging in this bound into the Dudley entropy bound, we get

$$\begin{aligned}
\mathbb{E}[R(\hat{f}_n) - R(f^*)] &\leq 2\frac{C}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)} d\epsilon \\
&\leq 2\frac{C}{\sqrt{n}} \int_0^1 \sqrt{\frac{L}{\epsilon}} d\epsilon \\
&= 4C\sqrt{\frac{L}{n}}, \quad [\text{Since } \int_0^1 \frac{1}{\sqrt{\epsilon}} d\epsilon = 2].
\end{aligned}$$

Therefore,

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

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

(c_1) The Dudley bound from part (b) is no longer useful, since for $d > 2$, the integral diverges:

$$\begin{aligned}\int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}_{L,d}, \|\cdot\|_\infty)} d\epsilon &\leq \sqrt{C_d L^{d/2}} \int_0^1 \sqrt{\epsilon^d} d\epsilon \\ &= \sqrt{C_d L^{d/2}} \left[\frac{\epsilon^{-\frac{d}{2}+1}}{-\frac{d}{2}+1}\right]_0^1 \\ &\to \infty \quad [\text{for } d > 2].\end{aligned}$$

(c_2) Let $\mathcal{N}_\epsilon \subset \mathcal{L}_\mathcal{F}$ be an $\epsilon-$ net of $\mathcal{L}_\mathcal{F}$ in $\|\cdot\|_\infty$. For any $g \in \mathcal{L}_\mathcal{F}$ $\exists f \in \mathcal{F}$, such that $\|g - f\|_\infty < \epsilon$

$$\begin{aligned}|\mu_n(g) - \mu(g)| &= |(\mu_n(g) - \mu_n(f)) + (\mu_n(f) - \mu(f)) + (\mu(f) - \mu(g))| \\ &\leq |\mu_n(f) - \mu_n(g)| + |\mu(f) - \mu(g)| + |\mu_n(f) - \mu(f)| \\ &\leq \epsilon + \epsilon + \sup_{f \in \mathcal{N}_\epsilon} |\mu_n(f) - \mu(f)| \quad [\text{The first two inequality holds since } \|f - g\|_\infty \leq \epsilon].\end{aligned}$$

Since the RHS is independent of f, we can take the sup w.r.t. $g \in \mathcal{L}_\mathcal{F}$ on the LHS, doing so, we get

$$\sup_{g \in \mathcal{L}_\mathcal{F}} \leq 2\epsilon + \sup_{f \in \mathcal{N}_\epsilon} |\mu_n(f) - \mu(f)|.$$

(c_3) For each $f \in \mathcal{N}_\epsilon$, the random variable $\mu_n(f) - \mu(f)$ is a centered sub-gaussian [boundedness followed by Hoeffding] random variable, hence the maximal inequality for finitely many sub-gaussian random variables give us

$$\max_{f \in \mathcal{N}_\epsilon} |\mu_n(f) - \mu(f)| \leq c\sqrt{\frac{\log |\mathcal{N}_\epsilon|}{n}} = c\sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{L}_\mathcal{F}, \|\cdot\|_\infty)}{n}}.$$

Combining this with part (c_2) gives us

$$\mathbb{E}[R(\hat{f}_n) - R(f^*)] \leq c'_d \left(\epsilon + \sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{L}_\mathcal{F}, \|\cdot\|_\infty)}{n}}\right).$$

Finally, plugging in the covering number bound for $\mathcal{L}_\mathcal{F}$, we get

$$\mathbb{E}[R(\hat{f}_n) - R(f^*)] \leq c'_d \left(\epsilon + \sqrt{\frac{(L/\epsilon)^d}{n}}\right).$$

(c_4) Choosing $\epsilon \asymp L^{d/(d+2)} n^{-1/(d+2)}$, we get

$$\sqrt{\frac{(L/\epsilon)^d}{n}} = \sqrt{\frac{(Ln^{1/(d+2)}/L^{d/(d+2)})^d}{n}} = \sqrt{\frac{L^{2d/(2+d)}}{n^{2/(2+d)}}} = L^{d/(d+2)} n^{-1/(d+2)}.$$

Therefore,

$$\mathbb{E}[R(\hat{f}_n) - R(f^*)] \leq 2C_d L^{d/(d+2)} n^{-1/(d+2)}.$$

We note that the sample complexity rate on the RHS is much weaker than $\sqrt{1/n}$, since $n^{-1/2} \ll n^{-1/(d+2)}$ when $d = \Omega(n)$, i.e., we can not make the rate independent of the ambient dimension d, hence the *curse of dimensionality* can not be avoided here.

(d) **[Bonus] Smoother classes help.** We now assume that $\mathcal{F} \subset \{f : [0,1]^d \to [0,1]\}$ is a hypothesis class whose covering number satisfies $\log \mathcal{N}\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq A\epsilon^{-p}$ for all $0 < \epsilon \leq 1$ for some constant $A > 0$ and $p > 0$. Using part (c_3) [following the same argument in part (c_1)] and the choice of $\epsilon \asymp A^{1/(p+2)} n^{-1/(p+2)}$, we get

$$\mathbb{E}[R(\hat{f}_n) - R(f^*)] \leq C\left(\epsilon + \sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty)}{n}}\right) \leq 2CA^{1/(p+2)} n^{-1/(p+2)}.$$

Putting $p = d/s$, we get

$$\mathbb{E}[R(\hat{f}_n) - R(f^*)] \leq 2CA^{s/(2s+d)} n^{-s/(2s+d)} \lesssim n^{-s/(2s+d)}.$$

As $s \to \infty$, we can see that $\mathbb{E}[R(\hat{f}_n) - R(f^*)] \lesssim n^{-1/2}$, i.e., smoothness reduces the *effective complexity* of the function class, so even in high dimensions, machine learning problems can achieve near-parametric rates, escaping the curse of dimensionality.