Contents

1 Minimax Lower Bounds

1.1 Decision theoretic foundations

We first provide the basic scaffolding of decision theory. Suppose we observe $W \in \mathcal{W}$ drawn from a distribution belonging to a statistical model: $P \in \mathcal{P}$. In most cases, we are interested in $P = Q^n$, which denotes the n-fold product measure (iid draws) on Q.

Based on our data realization W, we can take an action in the action space $\mathcal{A}$. The choice of action is determined by a decision rule, which maps from the data to action space:

$$T : \mathcal{W} \to \mathcal{A}$$

The quality of an action is judged by the loss:

$$L : \mathcal{A} \times \mathcal{P} \to \mathbb{R}$$

The quality of a decision rule is judged bt the risk, i.e., the expected loss:

$$R(T, P) = \int L(T, P) dP(w)$$

Another quantity of interest is the Bayes risk, which averages the risk over a prior on the statistical model:

$$r(T, \Pi) = \int R(T, P) d\Pi(P) \tag{1}$$

Shown below are a few estimands and their associated risks under a choice of loss.

Example 1 (Point estimation under squared error loss)**.**
Suppose our objective is to estimate some functional of P, $\psi(P) \in \mathbb{R}$.

Under squared error loss, $L(a, P) = (a - \psi(P))^2$, the risk is the mean-squared error:

$$R(T, P) = \int (T(w) - \psi(P))^2 dP(w)$$

Example 2 (Estimating a regression function)**.**
Suppose we observe n iid copies $(X, Y) \sim Q$ on support $\mathcal{X} \times \mathbb{R}$ and our goal is to estimate the regression function:

$$f_Q : x \to \mathbb{E}_Q(Y|X = x) \tag{2}$$

In this case, the action space is a collection fo functions mapping from $\mathcal{X}$ to $\mathbb{R}$. A common choice of loss is the integrated squared error loss:

$$L(a, P) = \int [a(x) - f_a(x)]^2 d\nu(x)$$

which yields a corresponding risk known as the mean integrated squared error.

Note that the risk depends on the particular choice of data generating distribution. Thus, we can judge the performance of a decision rule based on its maximal (worst case) risk with respect to the statistical model $\mathcal{P}$

Definition 1 (Minimax risk)**.** The maximal risk of an estimator over a statistical model is defined as so:

$$\sup_{P\in\mathcal{P}} R(T,P)$$

The minimax rule is optimal with respect to the maximal risk (i.e., minimizes maximal risk):

$$T^* := \operatorname*{argmin}_{T\in\mathcal{T}} \sup_{P\in\mathcal{P}} R(T,P) \iff \sup_{P\in\mathcal{P}} R(T^*,P) = \inf_{T\in\mathcal{T}} \sup_{P\in\mathcal{P}} R(T,P) \tag{3}$$

In some parametric settings, we can derive a closed form for the minimax estimator, but this is not always tractable. The alternative becomes finding estimators that achieve a minimax optimal rate (with respect to sample size).

Definition 2 (Minimax rate optimality)**.** In short, we desire a sequence of decision rules T_n that has a maximum risk that does not asymptotically dominate the minimax risk, i.e., is minimax rate optimal:

$$\liminf_{n\to\infty} \frac{\inf_{T\in\mathcal{T}} \sup_{Q\in\mathcal{Q}} R(T,Q^n)}{\sup_{Q\in\mathcal{Q}} R(T_n,Q^n)} > 0 \tag{4}$$

In other words, we need the minimax risk and the maximum risk of our estimator sequence to be of the same order with respect to n.

However, in practice, finding the form of the numerator and denominate of equation 4 is often difficult, so we settle for identifying bounds. Later on we'll develop bounds on the maximum risk, but in the next subsection, we develop bounds on the minimax risk.

1.2 Bounding the minimax risk

We'll focus on three main strategies for generating a lower bound on the minimax risk:

1. Bayes risk under least favorable prior
2. Le Cam's method
3. Fano method

Theorem 1 (Bayes risk bound)**.**
For any decision rule/estimator, we can bound the minimax rate by the Bayes risk under the least favorable prior:

$$\sup_{\Pi} \inf_{T\in\mathcal{T}} r(T,\Pi) \leq \inf_{T\in\mathcal{T}} \sup_{P\in\mathcal{P}} R(T,P) \tag{5}$$

Note the following bound holds for any choice of prior, by the least favorable prior gives the tightest bound

Proof:

$$\begin{aligned}
\sup_{\Pi} r(T,\Pi) &\leq \sup_{P} R(T,P) \quad \text{(Expectation} \leq \text{Maximum)} \\
\inf_{T\in\mathcal{T}} \sup_{\Pi} r(T,\Pi) &\leq \inf_{T\in\mathcal{T}} \sup_{P} R(T,P) \quad \text{(Inf both sides)} \\
\sup_{\Pi} \inf_{T\in\mathcal{T}} r(T,\Pi) &\leq \inf_{T\in\mathcal{T}} \sup_{\Pi} r(T,\Pi) \quad \text{(Max-min inequal)} \\
\implies \sup_{\Pi} \inf_{T\in\mathcal{T}} r(T,\Pi) &\leq \inf_{T\in\mathcal{T}} \sup_{P} R(T,P)
\end{aligned}$$

Next, we explore Le Cam's bound, which lower bounds the minimax risk for pairs of distributions in the statistical model. Clever choices of P_1, P_2 will yield the tightest bounds.

Theorem 2 (Le Cam's Method)**.**
Let R be a risk function defined according to a loss L. For any P_1, P_2:

$$\begin{aligned}
\inf_{T\in\mathcal{T}} \sup_{P\in\mathcal{P}} R(T,P) &\geq \frac{1}{2} d(P_1,P_2)||p_1 \wedge p_2||_1 \\
&\geq \frac{1}{4} d(P_1,P_2)\exp(-KL(P_1,P_2))
\end{aligned}$$

Where

1. Discrepancy: $d(P_1,P_2) = \inf_{a\in\mathcal{A}}[L(a,P_1) + L(a,P_2)]$ measures how different the estimation procedures are.
   - Point estimation under squared error loss: a little calculus shows that the discrepancy:
   $$\begin{aligned}
   d(P_1,P_2) &= \inf_{a\in\mathbb{R}}[L(a,P_1) + L(a,P_2)] \\
   &= \frac{1}{2}[\psi(P_1) - \psi(P_2)]^2
   \end{aligned}$$
   - Estimating a function with integrated squared error loss: suppose $P = Q^n$ and the action space is a convex subset of functions mapping from $\mathcal{X} \to \mathbb{R}$:
   $$d(P_1,P_2) = \frac{1}{2}\int [f_{Q_1}(x) - f_{Q_2}(x)]^2 d\nu(x)$$
2. Testing affinity: $||p_1 \wedge p_2||_1 = \int \min\left(\frac{dP_1}{d\nu}(w), \frac{dP_2}{d\nu}(w)\right) d\nu(w) \equiv \int \min(p_1,p_2) d\nu$ measures the overlap the between distributions P_1, P_2. Also note
   $$\begin{aligned}
   ||p_1 \wedge p_2||_1 &= 1 - \text{TV}(P_1,P_2) \\
   &= 1 - \sup_A |P_1(A) - P_2(A)|
   \end{aligned}$$
3. K-L divergence: $KL(P_1,P_2) := \begin{cases} \int \log\left(\frac{dP_1}{dP_2}(w)\right) dP_1(w) & \text{if } P_1 << P_2 \\ +\infty & \text{else} \end{cases}$
   quantifies the "distance" between the distributions

Loosely speaking, in order to obtain a tight bound, we desire the loss of the decision problem to be large (large discrepancy) while it is difficult to determine whether a sample came from P_1 or P_2 (small KL).

Proof: Bound 1: Let π denote a uniform prior over $\{P_1, P_2\}$, i.e., $\pi(P_1) = \pi(P_2) = \frac{1}{2}$.

$$\begin{aligned}
r(T,\pi) &= \sum_{j=1}^{2}\left[\int L(T(w),P_j)p_j(w)d\nu(w)\right]\pi(P_j) \\
&= \frac{1}{2}\int\left[\sum_{j=1}^{2} L(T(w),P_j)p_j(w)\right]d\nu(w) \quad \text{(Linearity and uniform prior)} \\
&\geq \frac{1}{2}\int \min(p_1(w),p_2(w))\left[\sum_{j=1}^{2} L(T(w),P_j)\right]d\nu(w) \quad \text{(Lower bound by min)} \\
&\geq \frac{1}{2}\int \min(p_1(w),p_2(w))\underbrace{\left[\inf_a \sum_{j=1}^{2} L(a,P_j)\right]}_{d(P_1,P_2)}d\nu(w) \quad \text{(Lower bound by inf)} \\
&= \frac{1}{2}||p_1 \wedge p_2||d(P_1,P_2)
\end{aligned}$$

Since T was arbitrary:

$$\frac{1}{2}||p_1 \wedge p_2||d(P_1,P_2) \leq \inf_T r(T,\pi) \underset{\text{sup}}{\leq} \sup_P \inf_T r(T,\pi) \underset{\text{Max-min}}{\leq} \inf_T \sup_P r(T,\pi) \underset{\text{Eq 5}}{\leq} \inf_{T\in\mathcal{T}} \sup_{P\in\mathcal{P}} R(T,P)$$

Bound 2:

$$\begin{aligned}
\exp(-KL(P_1,P_2)) &= \exp\left(-\int \log\left(\frac{p_1}{p_2}\right)dP_1(x)\right) \\
&= \exp\left(\int \log\left(\frac{\sqrt{p_2}^2}{\sqrt{p_1}^2}\right)dP_1(x)\right) \\
&= \exp\left(2\int \log\left(\frac{\sqrt{p_2}}{\sqrt{p_1}}\right)p_1 d\nu(x)\right) \\
&\leq \exp\left(2\log\left(\int\left(\frac{\sqrt{p_2}}{\sqrt{p_1}}\right)p_1 d\nu(x)\right)\right) \quad \text{(Jensens inequal)} \\
&= \exp\left(\log\left[\left(\int\sqrt{p_1p_2}d\nu(x)\right)^2\right]\right) \\
&= \left(\int\sqrt{p_1p_2}d\nu(x)\right)^2 \\
&= \left(\int\sqrt{\min(p_1,p_2),\max(p_1,p_2)}d\nu(x)\right)^2 \\
&= \left(\int\sqrt{\min(p_1,p_2),(p_1+p_2-\min(p_1,p_2))}d\nu(x)\right)^2 \\
&\leq \int(p_1+p_2)\min(p_1,p_2) - \min(p_1,p_2)^2 d\nu(x) \quad \text{(Jensen)} \\
&= 2\int\min(p_1,p_2)d\nu(x) - \int\min(p_1,p_2)^2 d\nu(x) \quad \text{(Pdfs integrate to 1)} \\
&= 2||p_1\wedge p_2||_1 - \int\min(p_1,p_2)^2 d\nu(x) \quad \text{(By definition)}
\end{aligned}$$

$$\implies \frac{1}{2}\exp(-KL(P_1,P_2)) \leq \frac{1}{2}\left(2||p_1\wedge p_2||_1 - \int\min(p_1,p_2)^2 d\nu(x)\right) = ||p_1\wedge p_2||_1 + C$$

Thus,

$$\frac{1}{4}d(P_1,P_2)\exp(-KL(P_1,P_2)) \leq \frac{1}{2}d(P_1,P_2)||p_1 \wedge p_2||_1 \leq \inf_{T\in\mathcal{T}}\sup_{P\in\mathcal{P}} R(T,P)$$

Note that the tightness of Le Cam's bound depends on choice of P_1, P_2

Strategy 1 (Clever choices of P_1, P_2 in Le Cam bound)**.**
Le Cam's method produces the tightest bounds when we compare two distributions with very different estimands of interest (large discrepancy) but are globally very similar (small KL divergence). A general way of constructing such densities is by considering a known distribution and perturbing it locally about the value of interest.

Definition 3 (Holder continuous functions)**.**
A Holder class imposes a smoothness condition on the orders of derivatives of a regression function. It is just a generalization of a Lipschitz condition. A $\Sigma(\beta, L)$ Holder class is defined as:

$$\mathcal{F} \equiv \left\{f : |f^{(\beta-1)}(x_1) - f^{(\beta-1)}(x_2)| \leq L|x_1 - x_2| \; \forall \; x_1, x_2 \in [0,1]\right\}$$

A sufficient condition for f to belong to the Holder class is f be β times differentiable and satisfy:

$$\sup_x |f^{(\beta)}(x)| \leq L$$

In certain cases, Le Cam's method will provide rate optimal lower bounds. A good case is on HW 1 Problem 2 for estimating a smooth density at a point.

Example 3 (Le Cam's minimax bound: estimating a smooth density at a point)**.**
Let $\Sigma(\beta, L)$ denote a Holder class where $\Sigma(\beta, L) \equiv \{f \; \forall x_1, x_2, |f^{\beta-1}(x_1) - f^{\beta-1}(x_2)| \leq L|x_1 - x_2|\}$

$$\mathcal{P}(\beta, L) = \left\{q | q \geq 0, \int q(x)dx = 1, q \in \Sigma(\beta, L)\right\}$$

Suppose we are interested in estimating the density at a point: $p(x_0)$. We can find a lower bound on the minimax risk of order $O(n^{-\frac{2\beta+1}{2\beta}})$.

Step 1: we are using Le Cam's method, so we propose two candidate distributions:

$$p_1 : x \to \sigma^{-1}\phi\left(\frac{x - x_0}{\sigma}\right)$$
$$p_2 : x \to p_1 + Lh_n^{\beta}\left[K\left(\frac{x - x_0}{h_n}\right) - K\left(\frac{x - h_n - x_0}{h_n}\right)\right]$$

And for sufficiently small choice of $a > 0$, $K : x \to a\exp\left(-\frac{1}{1-4x^2}\right)\mathbb{I}(|x| \leq 1/2)$.

Step 2: verify that these distributions are indeed in $\mathcal{P}$

1. p_1: is clearly infinitely differentiable and is a density by definition. To show it is Holder, we can show:

$$\left|\frac{d^\beta}{dx^\beta}p_1(x)\right| \leq L \text{ uniformly in } x \in \mathbb{R}$$

Turns out:

$$\frac{d^\beta}{dx^\beta}p_1(x) = (-1)^\beta H_\beta(x)\phi(x)$$

Where $H_\beta(x)$ is the β-th Hermite polynomial. Since $\lim_{|x|\to\infty} \frac{1}{\sqrt{2\pi}}H_\beta(x)e^{-x^2/2} = 0$ and the derivative is continuous, $\left|\frac{d^\beta}{dx^\beta}p_1(x)\right|$ is bounded uniformly by a constant. We can make this constant $\leq L$ by choosing σ large enough.

2. p_2: clearly p_2 integrates to 1 because the integrals of the K terms cancel. For p_2 to be positive, we need $p_1(x) - Lh_n^\beta K\left(\frac{x-h_n-x_0}{h_n}\right) > 0$ over the support of the bump, i.e.:

$$\begin{aligned}
&0 < p_1(x) - Lh_n^\beta K\left(\frac{x-h_n-x_0}{h_n}\right) \\
\equiv &0 < p_1(x) - Lh_n^\beta a \exp\left(-\frac{1}{1-4\left(\frac{x-h_n-x_0}{h_n}\right)^2}\right)\mathbb{I}\left(\left|\frac{x-h_n-x_0}{h_n}\right| \leq \frac{1}{2}\right) \\
\equiv &0 < p_1(x) - Lh_n^\beta a \exp\left(-\frac{1}{1-4\left(\frac{x-h_n-x_0}{h_n}\right)^2}\right)\mathbb{I}\left(x \in \left[x_0+1-\frac{h_n}{2}, x_0+1+\frac{h_n}{2}\right]\right)
\end{aligned}$$

Choose $a^* < \inf_{x\in\left[x_0+1-\frac{h_n}{2}, x_0+1+\frac{h_n}{2}\right]} \frac{p_1(x)}{Lh_n^\beta \exp\left(-\frac{1}{1-4\left(\frac{x-h_n-x_0}{h_n}\right)^2}\right)}$. This ensures positivity. To ensure the $(\beta-1)$-th derivative is bounded, we see that K is just a scaled bump function on $[-0.5, 0.5]$ and continuous functions on compact support attain their maximum and minimum, meaning the β-th derivative is bounded by a constant. We can force this constant to be less than L by choosing $\sigma, a > 0$ small enough.

Step 3: study the KL divergence:

$$\begin{aligned}
-KL(P_1, P_2) &= \int \log\left(\frac{p_2}{p_1}\right)p_1 d\nu \\
&= \int \log\left(\frac{p_1 + \text{bump}}{p_1}\right)p_1 d\nu \\
&= \int \left(\sum_{i=1}^\infty (-1)^{i+1}\frac{\left(\frac{\text{bump}}{p_1}\right)^i}{i}\right)p_1 d\nu
\end{aligned}$$

Let's inspect these terms individually:

$$\text{1st order term} = \int \text{bump } d\nu = 0$$

$$\begin{aligned}
\text{2nd order term} &= \frac{1}{2}\int L^2 h^{2\beta} p^{-1}(x)\left[K\left(\frac{x-x_0}{h_n}\right) - K\left(\frac{x-x_0-1}{h_n}\right)\right]^2 d\nu \\
&\overset{h_n \text{ small}}{=} \int c_1 h^{2\beta} p^{-1}(x)\left[K\left(\frac{x-x_0}{h_n}\right)^2 + K\left(\frac{x-x_0-1}{h_n}\right)^2\right] d\nu \quad \text{(When } h_n \text{ small, bumps orthog)} \\
&= c_1 h_n^{2\beta+1}\int p^{-1}(h_n U + x_0)\left(K^2(U) + K^2\left(U - \frac{1}{h_n}\right)\right) d\nu \quad \text{(U-sub)} \\
&= c_2 h_n^{2\beta+1} \\
\text{3rd order terms} &= o(h^{3\beta})
\end{aligned}$$

Thus, $-KL(P_1, P_2) \geq c h_n^{2\beta+1} + o(h^{3\beta})$. Under iid draws:

$$-KL(P_1^n, P_2^n) \geq cnh_n^{2\beta+1} + no(h^{3\beta})$$

To get a stable bound on the KL, $h_n = O\left(n^{-\frac{1}{2\beta+1}}\right)$. **Step 4**: study the discrepancy:

$$\begin{aligned}
d(P_1, P_2) &= \frac{1}{2}(p_1(x_0) - p_2(x_0))^2 \\
&= \frac{1}{2}(p_1(x_0) - p_1(x_0) - Lh_n^\beta(K(0) - K(-1/h_n)))^2 \\
&= Ch_n^{2\beta}
\end{aligned}$$

Thus, by Le Cam:

$$\begin{aligned}
\inf_{T\in\mathcal{T}}\sup_{P\in\mathcal{P}} R(T,P) &\geq \frac{1}{4} d(P_1, P_2)\exp(-KL(P_1, P_2)) \geq ch_n^{2\beta} \\
&= c^* n^{-\frac{2\beta}{2\beta+1}} \quad \text{(Subbing in from KL deriv)}
\end{aligned}$$

Another use case of Le Cam's method: estimating the derivative of a density at a point!

Example 4 (Estimating the derivative of a density at a point)**.**
Suppose we observe n iid draws from a density $f \in \mathcal{P}(\beta, L)$ for $L > 0, \beta > 1$. Our objective is to estimate $f'(x_0)$, the derivative at a fixed point. We quantify performance in terms of MSE. Given the KDE:

$$\hat{f}_h : x \to \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right)$$

For a kernel K satisyfing:

1. $K(u) = 0$ for all $u \notin [-1, 1]$
2. $\int K(u)du = 1$, and $\forall\, j = 1, \ldots, beta - 2\text{m}\int u^j K(u)du = 0$
3. K differentiable on whole real line with uniformly bounded derivative.

Turns out the KDE achieves $\inf_{T\in\mathcal{T}}\sup_{P\in\mathcal{P}} R(T,P) = O(n^{-\frac{2[\beta-1]}{2\beta+1}})$. See HW solutions for more details.

In other cases, Le Cam's method fails to provide a rate optimal lower bound, such as in the case of estimating a smooth regression function in terms of mean integrated squared error. In this setting, we need an alternative approach, provided by Fano's method.

Theorem 3 (Fano's method)**.**
Let R be a risk function defined according to nonnegative loss L. For $N \geq 3$, let $P_1, \dots, P_N \in \mathcal{P}$ and define η as the minimum discrepency and $\bar{P}$ as the uniform mixture of $P_1, \dots, P_N$:

$$\eta := \min_{j \neq k} d(P_j, P_k)$$
$$\bar{P} := \frac{1}{N} \sum_{j=1}^{N} P_j$$

We obtain the following lower bound on thje minimax risk:

$$\begin{aligned} \inf_{T \in \mathcal{T}} \sup_{P \in \mathcal{P}} R(T, P) &\geq \frac{\eta}{2} \left[1 - \frac{\log 2 + \frac{1}{N} \sum_{j=1}^{N} KL(P_j, \bar{P})}{\log(N)} \right] \\ &\geq \frac{\eta}{2} \left[1 - \frac{\log 2 + \max_{j \neq k} KL(P_j, \bar{P})}{\log(N)} \right] \end{aligned} \tag{6}$$

Proof: omitted for the sake of brevity, but can be found in the Chapter 1 lecture notes. Starts by lower bounding the bayes risk, then taking the maximum density, then applying Jensen's inequality to bound the integral of the max density.

Now we derive the lower bound for the minimax risk for estimating a Holder-continuous regression function!

Example 5 (Minimax lower bound for estimating a smooth regression function)**.**
Suppose we observe $(X_1, Y_1), \dots, (X_n, Y_n) \overset{iid}{\sim} Q \in \mathcal{Q}$, where $X \sim U[0,1]$ and $Y|X = x \sim N(f_Q(x), 1)$ where $f_Q(x) \in \mathcal{F}(\beta, L)$ where $\mathcal{F}(\beta, L)$. Suppose our objective is to estimate $f_Q(x)$, with performance quantified by the mean integrated squared error:

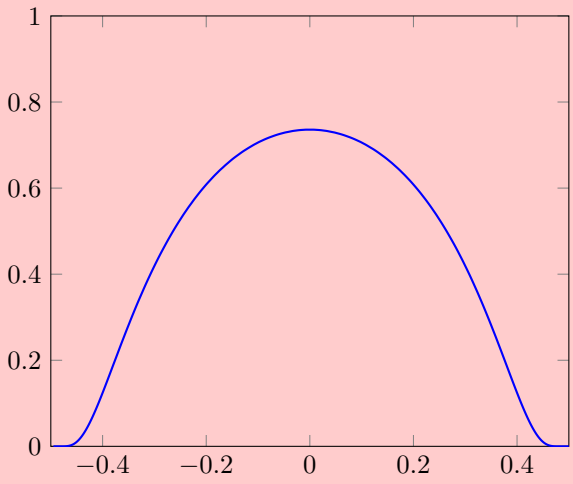
$$L(a, Q^n) = \int_0^1 [a(x) - f_Q(x)]^2 dx$$

Let our candidate class of functions be a convex combination of orthonormal basis functions, where the basis is just a scaled version of a smooth bump function:

$$\mathcal{F}_1 \equiv \left\{ x \to \sum_{j=1}^{m} w_j \phi_j(x) : w \in \{0,1\}^m \right\}$$

Where $\phi_j(x) = Lh^\beta K\left(\frac{x - \frac{j}{m+1}}{h}\right)$ and for fixed $h > 0$ we have $m \in \left[8, \frac{1}{h} - 1\right]$ where for sufficiently small $a > 0$:

$$K(x) = a \exp\left(-\frac{1}{4x^2}\right) \mathbb{I}(|x| < 1/2)$$



Thus, $f_W(x)$ is written as a sum of bump functions centered at $\frac{j}{m+1}$ for j up to m, scaled by binary w_j, and the condition that $m \leq \frac{1}{h} - 1 \implies h \leq \frac{1}{m+1}$ ensures the bumps do not overlap (are orthogonal).

Thus, we've created a class of regression functions s.t., $|\mathcal{F}_1| = 2^m$. Let's define our collection of distributions according to $\{P_w := Q_w^n : w \in \tilde{\Omega} \subset \{0,1\}^m\}$, since the collection of distributions is fully determined by $w \in \{0,1\}^m$.

Using Fano's method, we obtain the general lower bound on the minimax risk:

$$\inf_{T \in \mathcal{T}} \sup_{P \in \mathcal{P}} R(T,P) \geq \frac{\min_{w\nu} d(P_w, P_\nu)}{2} \left[1 - \frac{\log 2 + \max_{w \neq \nu} KL(P_w, P_\nu)}{\log |\tilde{\Omega}|}\right]$$

Recall from earlier in the chapter:

$$\begin{aligned}
d(P_w, P_\nu) &= \frac{1}{2}\int [f_w(x) - f_\nu(x)]^2 dx \\
&= \frac{1}{2}\sum_{j=1}^m [w_j - \nu_j]^2 \int \phi_j(x)^2 dx \quad \text{(Bases orthogonal so cross terms cancel)} \\
&= \frac{1}{2}\sum_{j=1}^m [w_j - \nu_j]^2 L^2 h^{2\beta+1} \underbrace{\int K(u)^2 du}_{c_2} \quad \text{(U-sub)} \\
&= c_2 L^2 h^{2\beta+1} \underbrace{\sum_{j=1}^m [w_j - \nu_j]^2}_{\text{Hamming dist}} \\
&= c_3 h^{2\beta+1} H(w,\nu) \quad \left(c_3 := \frac{c_2 L^2}{2}\right)
\end{aligned}$$

Also we can show:

$$\begin{aligned}
KL(P_w, P_\nu) &= \frac{n}{2}\int_0^1 [f_w(x) - f_\nu(x)]^2 dx \\
&= c_3 n h^{2\beta+1} H(w,\nu) \quad \text{(By same logic)} \\
&\leq c_3 n h^{2\beta+1} m \quad \text{since } H(w,\nu) \leq m)
\end{aligned}$$

Plugging into Fano's bound, we obtain:

$$\begin{aligned}\inf_{T\in\mathcal{T}}\sup_{P\in\mathcal{P}} R(T,P) &\geq \frac{c_3 h^{2\beta+1}}{2}\left(1-\frac{\log 2 + c_3 n h^{2\beta+1}}{\log|\tilde{\Omega}|}\right)\\ &= \frac{c_3 h^{2\beta+1}}{2}\left(1-\frac{\log 2 + c_3 n h^{2\beta+1}}{m\log 2}\right)\end{aligned}$$

In order for this bound to be useful, the RHS must be greater than 0, therefore, $h = O(n^{-\frac{1}{2\beta+1}})$, which indicates that the LHS converges to 0 rate no faster than n^{-1}.

Recall that in a general MLE problem:

$$\begin{aligned}&\sqrt{n}(\hat{\beta}-\beta_0) \Rightarrow N(0,\sigma^2)\\ \implies &\frac{n(\hat{\beta}-\beta_0)^2}{\sigma^2} \Rightarrow O_p(1)\\ \implies &(\hat{\beta}-\beta_0)^2 \Rightarrow O_p\left(\frac{1}{n}\right)\\ \implies &\mathbb{E}\left[\int(\hat{\beta}-\beta)^2 dx\right] = O_p\left(\frac{1}{n}\right)\end{aligned}$$

This means that our minimax lower bound problem is at least as hard as a parametric problem. This doesn't tell us much.

Can we do better? Can we find a set $\tilde{\Omega}\subset\{0,1\}^m$ for which $|\tilde{\Omega}|$ is large and the Hamming distance is also large, maximizing the RHS?

Turns out we can! The Varshamov-Gilbert Lemma guarantees that for $m\geq 8$, there exists a subset Ω s.t. $|\Omega|\geq 2^{m/8}$ (large cardinality) and $\min_{w\neq\nu} H(w,\nu)\geq\frac{m}{8}$. If we choose, the subset guaranteed to exist according to the lemma:

$$\begin{aligned}\inf_{T\in\mathcal{T}}\sup_{P\in\mathcal{P}} R(T,P) &\geq \frac{c_3 h^{2\beta+1} m}{16}\left(1-\frac{\log 2 + c_3 n h^{2\beta+1} m}{m\log(2)/8}\right)\\ &= \frac{c_3 h^{2\beta+1} m}{16}\left(1-\frac{8}{m}-\frac{8c_3 n h^{2\beta+1}}{\log(2)}\right)\end{aligned}$$

We want to choose m as large as possible to provide the tightest bound. Recall that earlier in the proof we assumed $m\leq\frac{1}{h}-1$. Choose $m=\lfloor\frac{1}{h}-1\rfloor$ (smallest integer less than $\frac{1}{h}-1$). Noting $\frac{1}{2h}<m<\frac{1}{h}$:

$$\begin{aligned}\inf_{T\in\mathcal{T}}\sup_{P\in\mathcal{P}} R(T,P) &\geq \frac{c_3 h^{2\beta+1} m}{16}\left(1-\frac{8}{m}-\frac{8c_3 n h^{2\beta+1}}{\log(2)}\right)\\ &\geq \frac{c_3 h^{2\beta}}{32}\left(1-8h-\frac{8c_3 n h^{2\beta+1}}{\log(2)}\right)\end{aligned}$$

To ensure the RHS is nonzero, $h = O_p(n^{-\frac{1}{2\beta+1}})$. This implies that the lower bound on the minimax risk is on the order of $\boxed{n^{-2\beta/(2\beta+1)}}$.

2 Kernel Density Estimation

This section deals with estimating a density. More information can be found in Chapter 24 of Van der Vaart's *Asymptotic Statistics* and Section 6.3 of Wasserman's *All of Nonparametric Statistics*.

Setup: Suppose we draw $X_1, \dots, X_n \overset{iid}{\sim} Q$, and let $F : x \to Q(X \leq x)$ denote the CDF of Q. We assume Q is continuous and the goal is to estimate f, its derivative.

A naive estimator of f might be built on the Empirical Distribution Function, $\hat{F} : x \to \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$, which is a known good estimator of F. However, the derivative of the EDF is not well-defined because it is a step function and is not differentiable.

Another estimator considers the limiting definition of a derivative:

$$\begin{aligned} f(x_0) &= \lim_{h \to 0} \frac{F(x_0 + h) - F(x_0)}{2h} \\ &\approx \frac{F(x_0 + h) - F(x_0)}{2h} \quad \text{for small } h > 0 \\ \implies \hat{f}_h(x_0) &= \frac{1}{2nh} \sum \mathbb{I}\left(\frac{|x_i - x_0|}{h} \leq 1 \right) \end{aligned}$$

This estimator takes the form of a KDE by placing mass at each point over with diameter $2h$. However, this estimate is not smooth! Can we achieve a smoother approximation?

Definition 4 (Kernel (s-order), KDE)**.**
A kernel is a function satisfying $\int K(u)du = 1$.

An s-order kernel satisfies $\int u^r K(u)du = 0$ for all $r = 1, 2, \dots, s-1$ and $|\int u^s K(u)du| < \infty$.

1. If K is symmetric, it is at least 2nd order.
2. Higher order kernels can lead to estimator with less bias.

A Kernel Density Estimator takes the form:

$$\hat{f}_h : x \to \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h} \right)$$

Examples of Kernels:

1. Uniform: $\frac{1}{2}\mathbb{I}(|u| \leq 1)$
2. Epanechnikov: $\frac{3}{4}(1 - u^2)\mathbb{I}(|u| \leq 1)$
3. Biweight: $\frac{15}{16}(1 - u^2)^2\mathbb{I}(|u| \leq 1)$
4. Triweight: $\frac{35}{32}(1 - u^2)^3\mathbb{I}(|u| \leq 1)$
5. Gaussian: $\frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$

Our goal is to study the performance of a KDE to estimate the density at a point $f(x_0)$, with performance quantified according to MSE.

Example 6 (Estimating density at a point with KDE in 1-D)**.**
Suppose $f \in \Sigma(\beta = 2, L)$ a Holder class. That is, suppose:

$$|f'(x_1) - f'(x_2)| \leq L|x_1 - x_2| \; \forall \; x_1, x_2$$

We focus on the situation where K is bounded, second order, and has bounded support. We are interested in the mean squared error:

$$\mathbb{E}[(\hat{f}_h(x_0) - f(x_0))^2] = \text{Bias}^2 + \text{Variance}$$

Step 1: Let's study the bias.

$$\begin{aligned}
\text{Bias} &= \mathbb{E}[\hat{f}_h(x_0)] - f(x_0) \\
&= \frac{1}{nh} \sum \mathbb{E}\left(K\left(\frac{X_i - x_0}{h}\right)\right) - f(x_0) \\
&= \frac{1}{h}\mathbb{E}\left(K\left(\frac{X_i - x_0}{h}\right)\right) - f(x_0) \quad \text{(iid)} \\
&= \frac{1}{h}\int K\left(\frac{X_i - x_0}{h}\right) f(x_i) dx_i - f(x_0) \\
&= \int K(u) f(x_0 + uh) du - f(x_0) \\
&= \int K(u) \underbrace{[f(x_0 + uh) - f(x_0)]}_{\star} du \quad \text{(Bc kernel integrates to 1)}
\end{aligned}$$

By mean value theorem, we know that $\star = uhf'(\tilde{x}_u h)$. Thus:

$$\begin{aligned}
\text{Bias} &= \int K(u) uh f'(\tilde{x}_u h) du \\
&= \underbrace{\int K(u) uh f'(x_0) du}_{\int uK(u)du = 0} + \int K(u) uh [f'(\tilde{x}_u h) - f'(x_0)] du \\
&= \int K(u) uh [f'(\tilde{x}_u h) - f'(x_0)] du
\end{aligned}$$

Now we can write:

$$\begin{aligned}
|\text{Bias}| &= |\int K(u) uh [f'(\tilde{x}_u h) - f'(x_0)] du| \\
&\leq h \int K(u)|u||f'(\tilde{x}_u h) - f'(x_0)| du \quad \text{(Jensen)} \\
&\leq Lh \int K(u)|u||\tilde{x}_{uh} - x_0| du \quad \text{(f is L2 Holder, deriv is Lipschitz)} \\
&\leq Lh^2 \underbrace{\int K(u) u^2 du}_{:=\sigma_k^2} \quad (\tilde{x}_{uh} \text{ is at most } uh \text{ from } x_0) \\
&= Lh^2 \sigma_k^2 \\
\implies \text{Bias}^2 &\leq L^2 h^4 \sigma_k^4
\end{aligned}$$

Step 2: Let's study the variance:

$$\begin{aligned}
\text{Var}(\hat{f}_h(x_0)) &= \text{Var}\left[\frac{1}{nh}\sum K\left(\frac{X_i - x_0}{h}\right)\right] \\
&= \frac{1}{n^2h^2}\sum \text{var}\left(K\left(\frac{X_i - x_0}{h}\right)\right) \quad \text{(independent)} \\
&= \frac{1}{nh^2}\text{var}\left(K\left(\frac{X_i - x_0}{h}\right)\right) \quad \text{(identical)} \\
&\leq \frac{1}{nh^2}\mathbb{E}\left[K\left(\frac{X_i - x_0}{h}\right)^2\right] \text{(var} < \text{2nd moment)} \\
&= \frac{1}{nh}\underbrace{\int K(u)^2 f(x_0 + uh)du}_{\star}
\end{aligned}$$

Recalling that f is holder and K has bdd support: let $k_1 = \inf(u : K(u) > 0)$ and $k_2 = \sup(u : K(u) > 0)$:

$$\begin{aligned}
\star &= \int_{k_1}^{k_2} K(u)^2 f(x_0 + uh)du \\
&\leq \left[\sup_{u\in[k_1,k_2]} f(x_0 + uh)\right]\int_{k_1}^{k_2} K(u)^2 du \\
&\leq \underbrace{\left[\sup_{t\in[k_1,k_2]} f(x_0 + t)\right]}_{\text{finite}} \underbrace{\int_{k_1}^{k_2} K(u)^2 du}_{\text{2nd order kern bounded}} \quad := c \quad \text{(If } h < 1\text{)}
\end{aligned}$$

Thus, we show that:

$$\text{Var}(\hat{f}_h(x_0)) = \frac{c}{nh}$$

Step 3: bring it all together:

$$\text{MSE} \leq L^2\sigma_k^4 h^4 + \frac{c}{nh}$$

Setting the two terms equal to each other ensures the rate that h should take to minimize the bound:

$$\begin{aligned}
L^2\sigma_k^4 h^4 &= \frac{c}{nh} \\
\implies h &= Cn^{-1/5}
\end{aligned}$$

Implying that the bound on the MSE is of the form $o(n^{-4/5})$.

Note: if we use a β-th order kernel, then the MSE of the order $n^{-2\beta/[2\beta+1]}$ and we need to evaluate a $(\ell - 1)$-th order taylor expansion of f at x_0 before applying mean value theorem.

Note: if we estimate a density in $d > 1$ dimesions, the rate is of the form $n^{-2\beta/[2\beta+d]}$.

3 Concentration Inequalities

You can find more information in Ch 2 of Wainwright text.

Suppose $X_1, \ldots, X_n$ are independent random variables, for which we want to bound the tail probability of the function of them:

$$P(f(X_1, \ldots, X_n) \geq t) \tag{7}$$

In many cases, $f : (X_1, \ldots, X_n) \to \frac{1}{n}\sum x_i$

We could consider bounding the expression in 7 using asymptotics:

$$P(\bar{X}_n \geq \mu + \frac{\sigma t}{\sqrt{n}}) \to 1 - \Phi(t)$$

Where Φ is a normal CDF. But we usually have finite sample, so the asymptotic guarantees are inexact. What can we say about finite samples.

In this section, we present three kinds of bounds:

1. Moment-based bounds: Markov, Chebyshev
2. MGF-based bounds: Chernoff, Hoeffding, sub-Gaussian random variables, sub-exponential random variables, Bernstein's inequality.
3. Martingale-based bounds: Azume-Hoeffding, bounded differences

3.1 Moment-based bounds

Theorem 4 (Markov Inequality)**.**
If $X \geq 0, \mathbb{E}(X) < \infty$ and $t > 0$, then:

$$P(X \geq t) \leq \frac{\mathbb{E}(X)}{t} \tag{8}$$

Proof:

$$\begin{aligned} P(X \geq t) &= \int_t^\infty dP(x) \\ &\leq \int_t^\infty \frac{X}{t} dP(x) \quad (1 \leq x/t \text{ when } x \geq t) \\ &\leq \int_0^\infty \frac{X}{t} dP(x) = \frac{\mathbb{E}(x)}{t} \end{aligned}$$

Theorem 5 (Applying Markov to transformations of $|X - \mathbb{E}(X)|$)**.**
Suppose $\mathbb{E}(X) < \infty$, $h : [0, \infty) \to [0, \infty)$ is a nondecreasing function and $\mathbb{E}[h(|X - \mathbb{E}(X)|)] < \infty$. Then:

$$P(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathbb{E}(h(|X - \mathbb{E}(X)|))}{h(t)}$$

A special case of this inequality is Chebyshev inequality: for $k \in \mathbb{N}$

$$P(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathbb{E}(|X - \mathbb{E}(X)|^k)}{t^k}$$

Proof: Because h is non-decreasing.

$$\{|X - \mathbb{E}(X)| \geq t\} \subset \{h(|X - \mathbb{E}(X)|) \geq h(t)\}$$

Hence:

$$\begin{aligned} P\{|X - \mathbb{E}(X)| \geq t\} &\leq P\{h(|X - \mathbb{E}(X)|) \geq h(t)\} \\ &\leq\subset \{h(|X - \mathbb{E}(X)|) \geq h(t)\end{aligned}$$

3.2 MGF-based bounds

We can attain even sharper tail bounds once we assume more about the distribution of X, such as making assumptions about the moment generating function! The fundamental MGF-based bound is the Chernoff bound.

Theorem 6 (Chernoff)**.**
Suppose X has an MGF in the neighborhood of 0, meaning there exists $b > 0$ such that $\mathbb{E}(\exp(\lambda X)) < \infty$ for all $|\lambda| \leq b$. Then for all $t > 0$ and $\lambda \in (0, b]$, it is true that:

$$\begin{aligned} P\{X - \mathbb{E}(X) \geq t\} = P\left\{e^{\lambda(X - \mathbb{E}(X))} \geq e^{\lambda t}\right\} &\leq \frac{\mathbb{E}\left[e^{\lambda(X - \mathbb{E}(X))}\right]}{e^{\lambda t}} \\ &\equiv \frac{M_{X-\mu}(\lambda)}{e^{\lambda t}}\end{aligned}$$

Hence for any $t > 0$:

$$\begin{aligned} P\{X - \mathbb{E}(X) \geq t\} &\leq \inf_{\lambda > 0} \frac{M_{X-\mu}(\lambda)}{e^{\lambda t}} \\ \log P\{X - \mathbb{E}(X) \geq t\} &\leq -\sup_{\lambda > 0}\{\lambda t - \log M_{X-\mu}(\lambda)\}\end{aligned} \tag{9}$$

Proof: follows directly from Markov's inequality and properties of MGF.

Example 7 (Chernoff bound on Gaussian RV)**.**
Suppose $X \sim N(\mu, \sigma^2)$. In this case:

$$\begin{aligned} M_{X-\mu}(\lambda) &= \mathbb{E}[\exp(\lambda(X - \mu))] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int \exp\left\{\lambda(x - \mu) - \frac{(x - \mu)^2}{2\sigma^2}\right\} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int \exp\left\{\lambda(z) - \frac{(z)^2}{2\sigma^2}\right\} dx \quad z = x - \mu \\ &= \exp\left\{\frac{\lambda^2\sigma^2}{2}\right\} \frac{1}{\sqrt{2\pi}\sigma} \int \exp\left\{-\frac{(z/\sigma - \lambda\sigma)^2}{2}\right\} dx \quad \text{(Complete the square)} \\ &= \exp\left\{\frac{\lambda^2\sigma^2}{2}\right\}\end{aligned}$$

The Chenoff bound gives us:

$$\begin{aligned}\log P\{X - \mathbb{E}(X) \geq t\} &\leq -\sup_{\lambda>0}\{\lambda t - \log M_{X-\mu}(\lambda)\} \\ &= -\sup_{\lambda>0}\left\{\lambda t - \frac{\lambda^2\sigma^2}{2}\right\}\end{aligned}$$

Solving for λ^* that maximizes yields $\lambda^* = t/\sigma^2$, yielding the following bound:

$$\log P\{X - \mathbb{E}(X) \geq t\} \leq -\frac{t^2}{2\sigma^2}$$

We can define the concentrations of other variables with respect to variables that we know and love! An obvious case is the Gaussian!

Definition 5 (Sub-Gaussian random variables)**.**
A random variable is sub-Gaussian if its cumulant generating function (log MGF) is less than the Gaussian:

$$\log M_{X-\mu}(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$$

By Chernoff, a sub-Gaussian random variable also satisfies the tail probability inequality:

$$\begin{aligned}\log P(X - \mu \geq t) &\leq -\frac{t^2}{2\sigma^2} \\ \equiv P(X - \mu \geq t) &\leq \exp\left(-\frac{t^2}{2\sigma^2}\right)\end{aligned}$$

Meaning that *the tails of a sub-Gaussian random variable with parameter σ^2 cannot be thicker than those of a $N(0, \sigma^2)$ random variable.*

We can apply the sub-Gaussian framework to the setting of random variables with bounded support. It is clear that they are sub-Gaussian bc they have 0 valued tails.

Theorem 7 (Hoeffding Inequality)**.** If the support of a random variable $X \sim P$ is bounded in $[a, b]$, then X is sub-Gaussian with parameter $\sigma^2 = (b-a)^2/4$. This yields the sub-Gaussian/Chernoff tail bound of:

$$\log P(X - \mu \geq t) \leq -\frac{2t^2}{(b-a)^2} \tag{10}$$

More generally, when $X_1, \ldots, X_n$ are independent with bounded support on $[a, b]$, then $\log M_{\sum X_i - \mathbb{E}(X_i)}(\lambda) = \sum \log M_{X_i - \mathbb{E}(X_i)}$, hence the bound on the $\bar{X}_n$ is just a sum of the upper bounds on individual X_i:

$$\log P\left\{\bar{X}_n - \mathbb{E}(\bar{X}_n) \geq t\right\} \leq -\frac{2nt^2}{(b-a)^2}$$

This proves the useful fact that sums of sub-G variables are sub-G.

Proof: WLOG suppose $\mathbb{E}(X) = 0$ (if not we can just mean center X). Let f: $\lambda \to M_{X-\mu}(\lambda)$ denote the

cumulant generating function X. Note $f'(\lambda = 0) = \frac{\mathbb{E}(X)}{\mathbb{E}(\exp(t \cdot 0))} = 0$. By the fundamental theorem of calculus:

$$f(\lambda) = \int_0^\lambda f'(r)dr = \int_0^\lambda \int_0^r f''(s)dsdr$$

So we can study the second derivative, bound it, and then integrate! Note:

$$\begin{aligned} f'(\lambda) &= \frac{\mathbb{E}(Xe^{\lambda X})}{\mathbb{E}(e^{\lambda X})} \\ f''(\lambda) &= \frac{\mathbb{E}(X^2e^{\lambda X})\mathbb{E}(e^{\lambda X})}{(\mathbb{E}(e^{\lambda X}))^2} - \frac{\mathbb{E}(Xe^{\lambda X})\mathbb{E}(Xe^{\lambda X})}{(\mathbb{E}(e^{\lambda X}))^2} \\ &= \frac{\mathbb{E}(X^2e^{\lambda X})}{\mathbb{E}(e^{\lambda X})} - \left(\frac{\mathbb{E}(Xe^{\lambda X})}{\mathbb{E}(e^{\lambda X})}\right)^2 \end{aligned}$$

Thus, f'' is the variance of a random variable Z_λ with density equal to $\frac{e^{\lambda z}}{\mathbb{E}(e^{\lambda z})}p(z)$, which has support bounded in $[a, b]$. Thus, Z_λ can be at most $\frac{b-a}{2}$ away from the midpoint $\frac{a+b}{2}$. Taking this fact:

$$\begin{aligned} f''(\lambda) = \text{Var}(Z_\lambda) \\ &= \text{Var}\left(Z_\lambda - \frac{a+b}{2}\right) \\ &\leq \mathbb{E}\left[\left(Z_\lambda - \frac{a+b}{2}\right)^2\right] \quad \text{(Var bdd by 2nd moment)} \\ &\leq \frac{(b-a)^2}{4} \end{aligned}$$

Plugging into the earlier equation:

$$\begin{aligned} f(\lambda) &\leq \frac{(b-a)^2}{4}\int_0^\lambda \int_0^r dsdr \\ &= \frac{(b-a)^2}{4}\frac{\lambda^2}{2} \end{aligned}$$

Hence, since $f(\lambda)$ was defined as the cumulant generating function we obtain that, X is sub-G with parameter $\sigma^2 = \frac{(b-a)^2}{4}$.

Now by the Chernoff bound on a sub-G random variable, we obtain Hoeffding's inequality:

$$P(X - \mu \geq t) \leq \exp\left(-\frac{2t^2}{(b-a)^2}\right)$$

We can apply a similar framework to Exponential random variables!

Definition 6 (Subexponential random variables)**.**
A random variable X is sub-exponential with parameters (σ^2, b) if for all $|\lambda| < \frac{1}{b}$,

$$\log M_{x-\mu}(\lambda) \leq \frac{\lambda^2\sigma^2}{2} \tag{11}$$

Alternatively, a nonnegative random variable is sub-exponential with parameter a if $\forall 0 < t < a$:

$$M_X(t) \equiv \mathbb{E}(\exp(tX)) \leq \frac{a}{a-t} \tag{12}$$

By Chernoff, a sub-Exponential random variable also satisfies the following tail probability inequality:

$$\log P(X \geq \mu + t) \leq \begin{cases} -\frac{t^2}{2\sigma^2}, & \text{if } 0 \leq t \leq \sigma^2/b \\ -\frac{t}{2b}, & \text{if } t > \sigma^2/b \end{cases} \tag{13}$$

Proof: By Chernoff

$$\begin{aligned} \log P(X - \mu \geq t) &\leq - \sup_{\lambda \in (0,1/b)} [\lambda t - \log M_{X-\mu}(\lambda)] \\ &\leq - \sup_{\lambda \in (0,1/b)} \left[\lambda t - \frac{\lambda^2 \sigma^2}{2}\right] \end{aligned}$$

Case 1: Over $\lambda \in \mathbb{R}$, the maximizing $\lambda^* = \frac{t}{\sigma^2}$. If $t < \sigma^2/b \implies \lambda^* < 1/b$, yielding:

$$- \sup_{\lambda \in (0,1/b)} \left[\lambda t - \frac{\lambda^2 \sigma^2}{2}\right] = - \left[\lambda^* t - \frac{(\lambda^*)^2 \sigma^2}{2}\right] = -\frac{t^2}{2\sigma^2}$$

Case 2: suppose $t \geq \sigma^2/b$. Since $f(\lambda) := \lambda t - \frac{(\lambda)^2 \sigma^2}{2}$ is monotonically increasing over $\lambda \in (0, 1/b)$, then:

$$\begin{aligned} \sup_{\lambda \in (0,1/b)} f(\lambda) = f(1/b) = \frac{t}{b} - \frac{\sigma^2}{2b^2} \geq \frac{t}{2b} \\ \implies - \sup_{\lambda \in (0,1/b)} f(\lambda) \leq -\frac{t}{2b} \end{aligned}$$

Together, cases 1 and 2 complete the proof!

Strategy 2 (Showing sub-E)**.**
Take the following approaches to show a random variable is sub-Exponential:

1. Show a random variable is sub-G and therefore sub=exponential with parameters (σ^2, b)
2. Show the random variable is a sum of sub-G/sub-E random variables.
3. If the following is true for some constant $a > 0$:

$$\mathbb{E}[X^b] \leq \frac{b!}{a^b}$$

   then X is sub-exponential.

4. A random variable X is sub-Gaussian iff X^2 is sub-exponential.

The following are approaches to showing a random variable is NOT sub-E:

1. Use contrapositive: if there exists some $a > 0$ such that for every natural number b:

$$\mathbb{E}[X^b] \leq 2^{b+1} \frac{b!}{a^b}$$

   then X is sub-E with parameter a. Show there does not exist a constant a satisfying as $b \to \infty$.

2. Show a particular moment of a random variable is infinite/unbounded.

Example 8 (Bounded random variable)**.**
Suppose X is concentrated about its mean $|X-\mu| \leq b$ and also that $\text{Var}(X) = \sigma^2$. We can show that X is sub-E with params $(2\sigma^2, 2b)$. Note:

$$\begin{aligned}
M_{X-\mu}(\lambda) &= \mathbb{E}[\exp(\lambda(X-\mu))] \\
&= 1 + 0 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}[(X-\mu)^k]}{k!} \\
&= 1 + \frac{\lambda^2\sigma^2}{2} + \sum \lambda^2 \lambda^{k-2} \frac{\mathbb{E}[(X-\mu)^2(X-\mu)^{k-2}]}{k!} \\
&\leq 1 + \frac{\lambda^2\sigma^2}{2} + \lambda^2\sigma^2 \sum_{k=3}^{\infty} \frac{(|\lambda|b)^{k-2}}{k!} \qquad (|X-\mu| \leq b) \\
&\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda|b)^{k-2} \\
&= 1 + \frac{\lambda^2\sigma^2}{2} \sum_{k=0}^{\infty} (|\lambda|b)^k \qquad \text{(recenter sum)}
\end{aligned}$$

If $|\lambda| < 1/b$, the geometric series converges:

$$\text{"} = 1 + \frac{\lambda^2\sigma^2}{2[1-|\lambda|b]} \leq \exp\left(\frac{\lambda^2\sigma^2}{2[1-|\lambda|b]}\right) \qquad \text{FACT } : 1 + x \leq exp(x)$$

Thus whenever $|\lambda| < \frac{1}{2b}$, $1 - |\lambda|b > 1/2$ meaning

$$M_{X-\mu}(\lambda) \leq \exp(\lambda^2\sigma^2)$$

implying that X is sub-E with params $(2\sigma^2, 2b)$

As in the previous example for a bounded random variable, we can establish a different concentration inequality, which we will subsequently extend to sample means.

Theorem 8 (Bernstein's inequality for bdd RVs)**.**
Suppose X is a RV that is bounded, $|X-\mu| \leq b$ and let $\sigma^2 = \text{Var}(X)$. For all $t > 0$;

$$P\{X - \mu \geq t\} \leq \exp\left(-\frac{t^2}{2[\sigma^2 + bt]}\right) \tag{14}$$

Proof: recall from the previous example for a bounded random variable that

$$M_{X-\mu}(\lambda) \leq \exp\left(\frac{\lambda^2\sigma^2}{2[1-|\lambda|b]}\right)$$

And by Chernoff for $|\lambda| < 1/b$

$$\begin{aligned}
P(X - \mu \geq t) &\leq \frac{M_{X-\mu}(\lambda)}{e^{\lambda t}} \\
\implies \log P(X - \mu \geq t) &\leq -[\lambda t - \log M_{X-\mu}(\lambda)] \\
&\leq -\lambda t + \frac{\lambda^2\sigma^2}{2[1-|\lambda|b]}
\end{aligned}$$

Plugging in $\lambda = \frac{t}{(bt+\sigma^2)}$ and reducing gives us the desired result.

Theorem 9 (Bernstein's inequality for sample means)**.**
Suppose $X_1, \ldots, X_n$ are independent RVs satisfying $|X_i - \mu_i| \leq b$ and let $\sigma_i^2 := \text{Var}(X_i)$ and $\bar{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n \sigma_i^2$:

$$P(\bar{X}_n - \mathbb{E}(\bar{X}_n) \geq t) \leq \exp\left(-\frac{nt^2}{2[\bar{\sigma}_n^2 + bt]}\right) \tag{15}$$

We can compare the performance of the Bernstein and Hoeffding concentration inequalities for sample means. Suppose $|X_i - \mu_i| \leq b$ and let $\bar{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. The following are true:

$$P(\bar{X}_n - \mathbb{E}(\bar{X}_n) \geq t) \leq \exp\left(-\frac{nt^2}{2[\bar{\sigma}_n^2 + bt]}\right) \quad \text{(Bernstein)}$$
$$P(\bar{X}_n - \mathbb{E}(X_n) \geq t) \leq \exp\left(-\frac{nt^2}{2b^2}\right) \quad \text{(Hoeffding)}$$

If $\bar{\sigma}_n^2$ is small (small average variance) and t is small, then Bernstein's inequality gives us sharper bounds. Thus, Bernstein beats Hoeffding in small variance cases.

3.3 Martingale-based bounds

We introduce the bounded differences inequality, which applies to functions of $X_1, \ldots, X_n$ that are not sample means! First, we must introduce the bounded differences property:

Definition 7 (Bounded differences property)**.**
A function $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the bounded differences property if for all i there exists a finite $c_i < \infty$ s.t. the following holds $\forall x_1, \ldots, x_n, x_i' \in \mathcal{X}$:

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i \tag{16}$$

Essentially, the BDP ensures that f cannot depend too heavily on one of its inputs.

Example 9 (f that satisifies BDP)**.**
Let $f : \mathcal{X}^n \to \mathbb{R}$ be defined as:

$$f(X) := \sup_{g \in G} \left| \frac{1}{n}\sum_{i=1}^n g(X_i) - \mathbb{E}(g(X_i)) \right|$$

If we require that the functions in G are uniformly bounded like so:

$$\sup_{g \in G} \sup_{x \in \mathcal{X}} |g(x)| \leq 1$$

f satisfes the BDP! To show this, we inspect:

$$\begin{aligned} f(X_1,\dots,X_n) - f(X_1,\dots,X_{i-1},X_i',X_{i+1},\dots,X_n) &= \sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^n [g(X_i)-\mathbb{E}(g(X_i))]\right| - \sup_{g\in G}\left|\frac{1}{n}\sum_{i=1}^n (g(X_i)-\mathbb{E}(g(X_i))) + \frac{1}{n}(g(X_i')-g(X_i))\right| \\ &\le \frac{1}{n}\sup_{g\in G}|g(X_i)-g(X_i')| \quad \text{(Triangle inequal)} \\ &\le \frac{2}{n} \end{aligned}$$

By symmetry, $|f(X) - f(X_1,\dots,X_{i-1},X_i',X_{i+1},\dots,X_n)| \le \frac{2}{n}$

Theorem 10 (Bounded differences inequality; McDiarmind's Inequality)**.**
If $X = (X_1,\dots,X_n)$ is a collection of independent random variables and arbitrary f satisfies the BDP with constants $c_1,\dots,c_n$, $\forall t > 0$:

$$P(|f(X) - \mathbb{E}(f(X))| \ge t) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \tag{17}$$

Proof: we can prove this by making use of a telescoping sum. Define:

$$\begin{aligned} D_1 &:= \mathbb{E}(f(X)|X_1) - \mathbb{E}(f(X)) \\ D_j &:= \mathbb{E}(f(X)|X_1,\dots,X_j) - \mathbb{E}(f(X)|X_1,\dots,X_{j-1}) \end{aligned}$$

Roughly speaking, D_j denotes the change in expectation by conditioning on the additional X_j. Note:

$$f(X) - \mathbb{E}(f(X)) = \mathbb{E}(f(X)|X_1,\dots,X_n) - \mathbb{E}(f(X)) \equiv \sum_{i=1}^n D_j$$

To establish 17, we show:

$$P\left(|\sum_{i=1}^n D_j| \le t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right) \tag{18}$$

Supposing the following about D_j:

1. D_j is a function of $(X_1,\dots,X_j)$ and is independent of $(X_{j+1},\dots,X_n)$
2. $\mathbb{E}[D_j|X_1,\dots,X_{j-1}] = \mathbb{E}[\mathbb{E}[f(X)|X_1,\dots,X_j] - \mathbb{E}[f(X)|X_1,\dots,X_{j-1}]|X_1,\dots,X_{j-1}] = 0$
3. $\mathbb{E}|D_j| < \infty$

Then $\{D_j\}_{j=1}^n$ is a martingale difference sequence. By the Azuma-Hoeffding Lemma, if the above enumerated conditions hold, then 18 holds as well! Then we have achieved McDiarmind's inequality!
Proving Azuma-Hoeffding makes repeated use of the sub-Gaussianity of bounded random variables. In short, let $A_j \le D_j \le B_j$ where:

$$\begin{aligned} A_j &:= \inf_{x_j} \mathbb{E}[f(X)|X_1,\dots,X_{j-1},X_j = x_j] - \mathbb{E}[f(X)|X_1,\dots,X_{j-1}] \\ B_j &:= \sup_{x_j} \mathbb{E}[f(X)|X_1,\dots,X_{j-1},X_j = x_j] - \mathbb{E}[f(X)|X_1,\dots,X_{j-1}] \end{aligned}$$

Also note that:

$$\begin{aligned} B_j - A_j &= \sup_{x_j,x_j'} \left(\mathbb{E}(f(X)|X_1,\dots,X_j = x_j) - \mathbb{E}(f(X)|X_1,\dots,X_j = x_j')\right) \\ &= \sup_{x_j,x_j'} \mathbb{E}(f(X_1,\dots,X_{j-1},x_j,X_{j+1},\dots,X_n)|X_1,\dots,X_{j-1}) \\ &\quad - \mathbb{E}[f(X_1,\dots,X_{j-1},x_j',X_{j+1},\dots,X_n|X_1,\dots,X_{j-1})] \\ &\le c_j \quad \text{(BDP)} \end{aligned}$$

Since $A_j \leq D_j \leq B_j$, then $D_j|X_1, \ldots, X_{j-1}$ is sub-G with parameter $\sigma_j^2 = \frac{c_j^2}{4}$. By definition of sub-G RV:

$$\mathbb{E}[\exp(\lambda D_j)|X_1, \ldots, X_{j-1}] \leq \exp\left(\frac{\lambda^2 c_j^2}{8}\right)$$

Now let's return to our quantity of interest $\sum_{j=1}^n D_j$:

$$\begin{aligned}
\mathbb{E}\left(\exp\left(\lambda \sum_{j=1}^n D_j\right)\right) &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda \sum_{j=1}^n D_j\right) \Big| X_1, \ldots, X_{n-1}\right]\right] \qquad \text{(Tower)} \\
&= \mathbb{E}\left[\exp\left(\lambda \sum_{j=1}^{n-1} D_j\right) \mathbb{E}(\exp(\lambda D_n)|X_1, \ldots, X_{n-1})\right] \\
&\leq \exp\left(\frac{\lambda^2 c_n^2}{8}\right) \mathbb{E}\left[\exp\left(\lambda \sum_{j=1}^{n-1} D_j\right)\right] \qquad \text{(Fact above)} \\
&\text{iterate} \ldots \\
&\leq \exp\left(\frac{\lambda^2 \sum_{i=1}^n c_j^2}{8}\right)
\end{aligned}$$

Hence, we've shown $\sum_{j=1}^n D_j$ is sub-G with parameter $\frac{\sum_{j=1}^n c_j^2}{4}$. Our desired result in 17 falls from Chernoff and recognizing:

$$\begin{aligned}
P(f(X) - \mathbb{E}(f(X)) \geq t) &\leq \exp\left(-\frac{2t^2}{\sum_{j=1}^n c_j^2}\right) \\
P(f(X) - \mathbb{E}(f(X)) \leq -t) &\leq \exp\left(-\frac{2t^2}{\sum_{j=1}^n c_j^2}\right) \\
\implies P(|f(X) - \mathbb{E}(f(X))| \geq t) &\leq 2\exp\left(-\frac{2t^2}{\sum_{j=1}^n c_j^2}\right)
\end{aligned}$$

4 Empirical Risk Minimization, VC Dimension, Rademacher Complexity

4.1 Empirical Risk Minimization

Many statistical estimation tasks can be cast as minimizers of a risk function averaged over the empirical distribution of your data.

As a quick review of empirical process notation, we define the true expectation and empirical expectations of a function f with respect to measures like so:

$$Pf := \int f(x)dP(x)$$

$$P_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

Suppose we have $X_1, \dots, X_n \overset{iid}{\sim} P$ and let $\ell : \mathcal{X} \times \Theta \to \mathbb{R}$ denote a loss function. The goal of estimation should be to find $\hat{\theta} \in \Theta$ such that the risk, defined as:

$$P\ell(\cdot, \hat{\theta}) := \int \ell(x, \hat{\theta})dP(x)$$

is nearly equal to $\inf_{\theta\in\Theta} P\ell(\cdot, \theta)$. Suppose that the infimum is achieved for some $\theta_0 \in \Theta$.

The regret quantifies how close we are to our goal:

$$\text{Reg}(\hat{\theta}) := P\ell(\hat{\theta}) - \inf_{\theta\in\Theta} P\ell(\theta) \tag{19}$$

If we had access to the true distribution function, P, we could solve for θ_0 exactly simply by solving a minimization problem: minimize $P\ell(\theta)$ subject to $\theta \in \Theta$. However, do don't know the distribution, however, we have a good approximation of it in the empirical distribution, P_n. Thus, the empirical risk minimizer which is the solution to: minimize $P_n\ell(\theta)$ subject to $\theta \in \Theta$.

Example 10 (Regression and Classification)**.**
Setting 1 (Regression): Suppose $X = (W, Y)$ where W is a feature and Y is a $\mathbb{R}$-valued outcome. The goal is to predict Y by W. Suppose $\Theta \subset L^2(P_w) := \{f : \mathcal{W} \to \mathbb{R}; \int f(w)^2 dP_w(w) < \infty\}$. Here we use the squared-error loss: $\ell(x, \theta) = [y - \theta(w)]^2$.
In this case, if Θ contains the regression function, $f^* : w \to \mathbb{E}_P(Y|W = w)$, then $\theta_0 = f^*$. If $f^* \notin \Theta$, then θ_0 is just the L-2 projection of the regression function onto Θ. The regret is defined as:

$$\text{Regret}(\hat{\theta}) = P[(\hat{\theta} - \theta)^2] = \int [\hat{\theta}(w) - \theta_0(w)]^2 dP_w(w)$$

Setting 2 (Classification): Suppose $X = (W, Y)$ where W is a feature and $Y \in \{0, 1\}$ and the goal is to classify Y based on W. $\Theta := \{f; \mathcal{W} :\to \{0, 1\}\}$. The loss is the 0-1 loss: $\ell(x, \theta) = \mathbb{I}(\theta(w) \neq y)$. If Θ contains $f^* : w \to \mathbb{I}(\mathbb{E}_P(Y|W = w) > 1/2)$, then $f^* = \theta_0$. The regret is defined as:

$$\text{Regret}(\theta) = \mathbb{E}_P[\ |2\mathbb{E}_P(Y|W = w) - 1|\ \mathbb{I}(\theta(w) \neq \theta_0(w))]$$

Thus, we pay the biggest price when we misclassify something easy, i.e., with $\mathbb{E}_P(Y|W = w)$ close to 0 or 1.

The following strategy provides a template for analyzing ERMs. It is useful in many contexts!

Strategy 3 (Analyzing ERMs)**.**
Note we can define the regret of an ERM as:

$$\begin{aligned}
\text{Regret}(\hat{\theta}) &:= P\ell(\hat{\theta}) - P\ell(\theta_0) \\
&\leq P\ell(\hat{\theta}) - P\ell(\theta_0) + \underbrace{P_n\ell(\theta_0) - P_n\ell(\hat{\theta})}_{\geq 0 \text{ bc } \hat{\theta} \text{ is ERM}} \\
&= (P_n - P)[\ell(\theta_0) - \ell(\hat{\theta})] \\
&\leq |(P_n - P)\ell(\theta_0)| + |(P_n - P)\ell(\hat{\theta})| \\
&\leq 2 \sup_{\theta\in\Theta} |(P_n - P)\ell(\theta)| \\
&= 2 \sup_{f\in\mathcal{F}} |(P_n - P)f| \quad \text{where } \mathcal{F} := \{\ell(\theta) : \theta \in \Theta\}
\end{aligned}$$

We define: $\sup |(P_n - P)f| =: ||P_n - P||_{\mathcal{F}}$, the Glivenko-Cantelli norm.

4.2 Rademacher Complexity

Rademacher complexity offers a way to upper bound the Glivenko-Cantelli norm, and hence the regret, of $[0,1]$-valued functions. Suppose $\mathcal{F}$ consists of $[0,1]$-valued functions.

Note that $||P_n - P||_{\mathcal{F}}$ satisfies the bounded differences property with $c_i = \frac{1}{n}$ because, as in Example 9, the functions in $\mathcal{F}$ are uniformly bounded. By the bounded differences inequality in equation 17:

$$P(|||P_n - P||_{\mathcal{F}} - \mathbb{E}||P_n - P||_{\mathcal{F}}| \geq t) \leq 2\exp(-\frac{2t^2}{\sum_{i=1}^n c_i^2}) = 2\exp(-2nt^2)$$

So with high probability, the Glivenko-Cantelli norm $||P_n - P||_{\mathcal{F}}$ is close to its mean, so it suffices to study it's mean $\mathbb{E}||P_n - P||_{\mathcal{F}}$. So bounding $\mathbb{E}||P_n - P||_{\mathcal{F}}$ also provides a bound on $||P_n - P||_{\mathcal{F}}$.

Before we provide bounds on this quantity, we define a Rademacher process and Rademacher complexity, quantities that will appear in our resulting bounds.

Definition 8 (Rademacher Process, Rademacher Complexity)**.**
The Rademacher Process: $R_n : \mathcal{F} \to \mathbb{R}$ is defined as:

$$R_n(f) := \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)$$

Where $\epsilon_i \overset{iid}{\sim}$ Rademacher, meaning:

$$\epsilon_i = \begin{cases} +1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2} \end{cases}$$

The Rademacher Complexity is just the expetcation of the supnorm of a Rademacher process:

$$\mathbb{E}||R_n||_{\mathcal{F}} := \mathbb{E}\left[\sup_{f\in\mathcal{F}} |R_n(f)|\right]$$

Theorem 11 (Bounding the GC norm via Rademacher Complexity)**.** Suppose $\mathcal{F}$ is a collection of $[0,1]$-valued functions. Then with probability at least $1-2\exp(-2nt^2)$, it holds that

$$\begin{aligned}\frac{1}{2}\mathbb{E}||R_n||_{\mathcal{F}} - \sqrt{\frac{\log 2}{2n}} - t &\leq \mathbb{E}||P_n - P||_{\mathcal{F}} - t \\ &\leq ||P_n - P||_{\mathcal{F}} \\ &\leq \mathbb{E}||P_n - P||_{\mathcal{F}} + t \\ &\leq 2\mathbb{E}||R_n||_{\mathcal{F}} + t\end{aligned} \tag{20}$$

Proof: The upper bound we achieve via a symmetrization argument using a *ghost sample*. Suppose $X_1,\dots,X_n \overset{iid}{\sim} P$ be the sample and $X_1',\dots,X_n' \overset{iid}{\sim} P$ be the ghost sample.

$$\begin{aligned}\mathbb{E}[||P_n - P||_{\mathcal{F}}] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_P(f(X_i'))\right|\right] \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(X_i) - f(X_i')\middle| X_1,\dots,X_n\right]\right|\right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(X_i')\right|\middle| X_1,\dots,X_n\right]\right] \quad \text{(Jensen)} \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(X_i'))\right|\right]\end{aligned}$$

To get rid of the ghost sample, introduce independent Rademacher noise into the sum. Since X_i' and X_i are exchangeable, we can flip the sign on their difference and still preserve equality of expectation:

$$\begin{aligned}\mathbb{E}[||P_n - P||_{\mathcal{F}}] &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(X_i'))\right|\right] \\ &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(X_i) - f(X_i'))\right|\right] \\ &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i')\right|\right] \quad \text{(triangle ineq)} \\ &\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right|\right] \quad \text{(b/c iid)} \\ &= 2\mathbb{E}||R_n||_{\mathcal{F}}\end{aligned}$$

We achieve the lower bound via desymmetrization, i.e., removing the Rademacher RVs.

$$\mathbb{E}||R_n||_{\mathcal{F}} \leq \underbrace{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(X_i)-Pf)\right|\right]}_{(i)} + \underbrace{\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i Pf\right|\right]}_{(ii)}$$

$$\begin{aligned}
(i) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(X_i)-Pf)\right|\right] \\
&\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum \epsilon_i(f(X_i)-f(X_i'))\right|\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum (f(X_i)-f(X_i'))\right|\right] \\
&\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-Pf)\right|\right] \quad \text{(Triangle)} \\
&= 2\mathbb{E}||P_n - P||_{\mathcal{F}} \\
(ii) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i Pf\right|\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}|Pf|\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right|\right] \\
&= \underbrace{||P||_{\mathcal{F}}}_{\leq 1}\cdot\mathbb{E}\left[\underbrace{\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\right|}_{\text{Sub-G}}\right] \\
&\leq \sqrt{\frac{2\log 2}{n}}
\end{aligned}$$

Therefore:

$$\frac{1}{2}\mathbb{E}||R_n||_{\mathcal{F}} - \sqrt{\frac{\log 2}{2n}} \leq \mathbb{E}||P_n - P||_{\mathcal{F}}$$

Then given the bounded differences inequality, $||P_n - P||_{\mathcal{F}}$ deviates absolutely from it's expectation by more than t with probability at most $1-2\exp(-2nt^2)$. Plugging in the bounds on the expectation term gives the desired result.

4.3 VC dimension

Rademacher Complexity provided some bounds on the regret of functions that map to $[0,1]$. But what about more general classes of functions? VC dimension provides an alternative framework to bound the regret of classes of functions.

Definition 9 (Projection, Shattering, Growth Function, VC dim/index)**.**
Let $\mathcal{F}$ be a class of functions that map $\mathcal{X} \to \{0,1\}$. For $(x_1,\ldots,x_n) \in \mathcal{X}^n$, the projection of $\mathcal{F}$ onto

$x_1^n := (x_1, \ldots, x_n)$ is denoted as:

$$\mathcal{F}_{x_1^n} := \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$$

We say $\mathcal{F}$ shatters x_1^n if it can accommodate every possible 0-1 labelling of the n points, i.e., $|\mathcal{F}_{x_1^n}| = 2^n$. The growth function of $\mathcal{F}$ is defined as the maximum cardinlity that the projection can take across all x_1^n in the domain:

$$\Pi_{\mathcal{F}}(n) := \sup_{x_1^n \in \mathcal{X}^n} |\mathcal{F}_{x_1^n}|$$

The VC dimension is simply the largest natural number input n for which x_1^n is shattered by $\mathcal{F}$. The VC index is the smallest natural number n for which x_1^n cannot be shattered by $\mathcal{F}$:

$$VC_{dim}(\mathcal{F}) := \sup\{n \in \mathcal{N} : \Pi_{\mathcal{F}}(n) = 2^n\}$$
$$VC_{ind}(\mathcal{F}) := \sup\{n \in \mathcal{N} : \Pi_{\mathcal{F}}(n) < 2^n\}$$

Definition 10 (VC dim for $\mathbb{R}$-valued functions)**.**
Suppose $\mathcal{F}$ consist of $\mathcal{X} \to \mathbb{R}$ functions. The VC dimension of $\mathcal{F}$ consists of the collection of subgraphs:

$$A := \{\{(x,t) \in \mathcal{X} \times \mathbb{R} : t < f(x)\} : f \in \mathcal{F}\}$$

We can also upper bound the VC dimension via the number of operations needed to compute f:

Theorem 12 (Upper bound VC dim by no. operations)**.**
Consider a parameterized family of functions $\mathcal{F}$ where $f(x,\theta)$ for $\theta \in \mathbb{R}^p$ and $f : \mathbb{R}^m \times \mathbb{R}^p \to \{0,1\}$. Suppose f can be computed using no more than t operations that are either arithmetic $(+,-,\div,\times)$ or comparisons $(>,\geq,<,\leq,=,\neq)$. Then:

$$VC(\mathcal{F}) \leq 4p(t+2)$$

Let's return to the case of boolean-valued functions. We know the empirical process term which bounds the regret is upper bounded by the Rademacher complexity. Turns out, we can upper bound the Rademacher complexity via the growth function, and the order of the growth function allows us to stochastically bound the empirical process term and therefore the regret!

Theorem 13 (Finite Class Lemma)**.**
If $\mathcal{F}$ is a class of functions $x \to [-1,1]$, then:

$$\mathbb{E}||R_n||_{\mathcal{F}} \leq \sqrt{\frac{2\log(2\mathbb{E}|\mathcal{F}_{x_1^n}|)}{n}} \tag{21}$$

Proof: To establish the desired result, we instead establish the result condition on the data $X_1^n = x_1^n$:

$$\mathbb{E}\left[||R_n||_{\mathcal{F}} | X_1^n = x_1^n\right] \leq \sqrt{\frac{2\log(2|\mathcal{F}_{x_1^n}|)}{n}}$$

Suppose $z \in [-1,1]^n$, $\epsilon \in \text{Rademacher}^n$, and $\langle \epsilon, z \rangle = \sum_{i=1}^n \epsilon_i z_i$. Let:

$$\mathcal{Z}_{x_1^n} = \mathcal{F}_{x_1^n} \cup \left(-\mathcal{F}_{x_1^n}\right)$$

Define:

$$\begin{aligned}
\exp\left\{\lambda \mathbb{E}\left[\sup_{f\in\mathcal{F}} |\langle \epsilon, f(x_1^n)\rangle|\right]\right\} &= \exp\left\{\lambda \mathbb{E}\left[\sup_{z\in\mathcal{Z}_{x_1^n}} \langle \epsilon, z\rangle\right]\right\} \\
&\leq \mathbb{E}\left[\exp\{\lambda \sup_{z\in\mathcal{Z}_{x_1^n}} \langle \epsilon, z\rangle\}\right] \quad \text{(Jensen)} \\
&= \mathbb{E}\left[\sup_{z\in\mathcal{Z}_{x_1^n}} \exp\{\lambda\langle \epsilon, z\rangle\}\right] \quad \text{(Monotonicity)} \\
&\leq \sum_{z\in\mathcal{Z}_{x_1^n}} \mathbb{E}\left[\exp(\lambda\langle \epsilon, z\rangle)\right]
\end{aligned}$$

Note that since $\epsilon_i z_i$ is bounded in $[-1,1]$, the random variable is sub-G with parameter 1 by Hoeffding. Thus, the iid sum $\langle \epsilon, z\rangle$ is sub-G with parameter n. Also note that the above expression is a sum of MGFs of $\langle \epsilon, z\rangle$. By the definition of a sub-G random variable:

$$\begin{aligned}
\mathbb{E}\left[\exp(\lambda\langle \epsilon, z\rangle)\right] &\leq \exp\left(\frac{\lambda^2 n}{2}\right) \\
\implies \sum_{z\in\mathcal{Z}_{x_1^n}} \mathbb{E}\left[\exp(\lambda\langle \epsilon, z\rangle)\right] &\leq \sum_{z\in\mathcal{Z}_{x_1^n}} \exp\left(\frac{\lambda^2 n}{2}\right) \\
&= |\mathcal{Z}_{x_1^n}| \exp\left(\frac{\lambda^2 n}{2}\right)
\end{aligned}$$

hence by taking logs we obtain:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\langle \epsilon, f(x_1^n)\rangle|\right] \leq \frac{\log|\mathcal{Z}_{x_1^n}|}{\lambda} + \frac{\lambda n}{2}$$

Plugging in $\lambda = \sqrt{2\log|\mathcal{Z}_{x_1^n}|/n}$ gives us:

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\langle \epsilon, f(x_1^n)\rangle|\right] &\leq n\lambda\left(\frac{\log|\mathcal{Z}_{x_1^n}|}{n\lambda^2} + \frac{1}{2}\right) \\
&\leq n\lambda = \sqrt{2n\log 2|\mathcal{F}_{x_1^n}|} \\
\mathbb{E}\left[||R_n||_{\mathcal{F}} | X_1^n = x_1^n\right] &= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\langle \epsilon, f(x_1^n)\rangle|\right] = \sqrt{\frac{2\log 2|\mathcal{F}_{x_1^n}|}{n}} \quad \square
\end{aligned}$$

A very simple extension of the finite class lemma gives us an upper bound on the expectation of the Glivenko-Cantelli norm in terms of the growth function:

Theorem 14 (Upper bounding GC norm via growth function)**.**

Suppose $\mathcal{F}$ consists of a class of Boolean-valued functions:

$$\mathbb{E}||P_n - P||_{\mathcal{F}} \leq 2\sqrt{\frac{2\log(2\Pi_{\mathcal{F}}(n))}{n}} \tag{22}$$

This result is achieved by upper bounding $\mathbb{E}|\mathcal{F}_{x_1^n}|$ with $\Pi_{\mathcal{F}}(n) = \sup_{x_1^n} |\mathcal{F}_{x_1^n}|$ in the preceding theorem. Note that this bound is trivial, i.e., does not shrink with $n \to \infty$, when $\Pi_{\mathcal{F}}(n) = 2^n$. But if $n > VC(\mathcal{F})$, then $\Pi_{\mathcal{F}}(n) < 2^n$.

This theorem illustrates that the bound on the expectation of the empirical process term when $n \leq VC(\mathcal{F})$. The bound is useful when $n > VC(\mathcal{F})$. But the question of how much smaller $\Pi_{\mathcal{F}}(n)$ is than 2^n is formalized by Sauer's Lemma.

Theorem 15 (Sauer's Lemma)**.**
Let $d \geq VC(\mathcal{F})$. It holds that $\Pi_{\mathcal{F}}(n) \leq \sum_{k=0}^{d} \binom{n}{k}$, hence:

$$\Pi_{\mathcal{F}}(n) \leq \begin{cases} 2^n & n \leq d \\ \left(\frac{e}{d}\right)^d \cdot n^d & n > d \end{cases} \tag{23}$$

Thus, when $n > VC(\mathcal{F})$, the growth function goes from exponential order to polynomial order.

This implies that if $VC(\mathcal{F}) \leq d \leq n$, then by combining with equation 22:

$$\begin{aligned} \mathbb{E}||P_n - P||_{\mathcal{F}} &\leq 2\sqrt{\frac{2\log 2 + 2d\log(\frac{e}{d}n)}{n}} \\ &= O\left(\sqrt{\frac{\log n}{n}}\right) \end{aligned} \tag{24}$$

4.4 Bracketing Numbers

In the previous section, we motivated bounding $\mathbb{E}||P_n - P||_{\mathcal{F}}$ by controlling the regret of Empirical Risk Minimizers. The main bound given by Equation 24 (Sauer's Lemma) focuses on the case where $\mathcal{F}$ is a Boolean-valued function, allowing us to provide regret guarantees for classification problems. How about in the more general cases: such as maximum likelihood or regression problems! This motivates the need for a new framework to bound $\mathbb{E}||P_n - P||_{\mathcal{F}}$ for non-boolean valued functions!

Definition 11 ($L^r(P)$ space)**.**
Let $\mathcal{F}$ be a subset of $L^r(P)$ space. For $r \geq 1$, $L^r(P)$ is a space of functions $f : \mathcal{X} \to \mathbb{R}$ s.t.

$$||f||_{L^r(P)} := \left[\int |f(x)|^r dP(x)\right]^{1/r} < \infty$$

When $r = \infty$, $L^r(P)$ consists of $f : \mathcal{X} \to \mathbb{R}$ s.t.

$$||f||_{L^\infty(P)} := \inf\{a \in \mathbb{R} : P(|f(x)| > a) = 0\}$$

Also note that $||f||_{L^r(p)}$ is non decreasing in r.

Definition 12 (Bracketing numbers)**.**
Given two functions $\ell : \mathcal{X} \to \mathbb{R}$, $u : \mathcal{X} \to \mathbb{R}$ in $L^r(P)$, a bracket, $[\ell, u]$ is the set of functions f with $\ell < f < u$ pointwise.

We call $[\ell, u]$ an ϵ-bracket if $||u - \ell||_{L^r(P)} \leq \epsilon$.

We define the Bracketing number of $\mathcal{F}$, $N_{[]}(\epsilon, \mathcal{F}, L^r(P))$ to be the minimal number of ϵ-brackets needed to cover $\mathcal{F}$:

$$N_{[]}(\epsilon, \mathcal{F}, L^r(P)) := \inf\{m : \mathcal{F} \subset \cup_{j=1}^m [\ell_j, u_j] \text{ s.t. } ||\ell_j - u_j||_{L^r(P)} \leq \epsilon \; \forall \; j = 1, \dots, m\} \tag{25}$$

Why are we interested in bracketing numbers? Turns out, finite bracketing numbers give us the asymptotic behavior of the empirical process/GC-norm via the Glivenko-Cantelli theorem:

Theorem 16 (Glivenko-Cantelli)**.**
If $\mathcal{F}$ is a function class with finite bracketing number, $N_{[]}(\epsilon, \mathcal{F}, L^r(P)) < \infty$ for all $\epsilon > 0$, then $\mathcal{F}$ is P-Glivenko-Cantelli meaning:

$$||P_n - P||_{\mathcal{F}} = o_P(1) \tag{26}$$

Proof: to be continued

4.5 Covering and Packing Numbers

Now we introduce the concepts of covering and packing numbers. We first introduce them, define the relationships between them, and then illustrate their connection to bounding the empirical process term of interest, $\mathbb{E}||P_n - P||_{\mathcal{F}}$.

Definition 13 (Pseudometric/Pseudometric Space)**.**
A function $d : S \times S \to [0, \infty)$ is called a pseudometric on S if:

1. $d(x, x) = 0$ for all $x \in S$
2. $d(x, y) = d(y, x)$ for all $x, y \in S$
3. $d(y, z) = d(x, y) + d(x, z)$ for all $x, y, z \in S$

In contrast to a metric, $d(x, y) = 0$ even if $x \neq y$.

A pseudometric space (S, d) is the pairing of set S with a pseudometric d.

A useful example of a psuedometric space is $(\mathcal{F}, d)$ where $d(f, g) := ||f - g||_{L^r(P)}$ (a pseudometric because f, g can disagree outside the support P.

Definition 14 (Covering number)**.**
Let (S, d) denote a pseudometric space and $T \subset S$. A set T_1 is called an ϵ-cover of T if for each $\theta \in T$, there exists $\theta_1 \in T_1$ such that $d(\theta, \theta_1) \leq \epsilon$. I.e., for every element in T, we can find a corresponding element in T_1 that is at most ϵ away. I.e., T_1 is a colection of points such that if we drew ϵ-balls about them, the balls would cover all of T.

An ϵ-covering number is defined as the size of the minimal ϵ-cover:

$$N(\epsilon, T, d) = \{|T_1| : T_1 \text{ is an } \epsilon\text{-cover of } T\} \tag{27}$$

Definition 15 (Packing numbers)**.**
A set $T_1 \subset T$ is called an ϵ-packing of T if for each distinct $\theta_1, \theta_1^{'} \in T$:

$$d(\theta_1, \theta_1') > \epsilon$$

I.e., if we draw balls centered at the elements of T_1, the balls will contain the other points in T_1 (each ball contains exactly one element of T_1). An ϵ-packing number of T is defined as:

$$M(\epsilon, T, d) := \sup\{|T_1| : T_1 \text{ is an } \epsilon\text{-packing of } T\} \tag{28}$$

Theorem 17 (Relationship between covering and packing numbers)**.**
Let N denote a covering number and M denote a packing number:

$$M(2\epsilon) \leq N(\epsilon) \leq M(\epsilon) \qquad \forall \epsilon > 0 \tag{29}$$

Proof

1. Prove $M(2\epsilon) \leq N(\epsilon)$: Consider T_1, an ϵ-cover, and T_2, a 2ϵ-packing of T. Because T_2 is a 2ϵ-packing, any two elements $\theta_2, \theta_2' \in T_2$ cannot be within ϵ of a common point $\theta \in T$, because then:

   $$d(\theta_2, \theta_2') \leq d(\theta_2, \theta) + d(\theta_2', \theta) \leq 2\epsilon$$

   Contradicting that T_2 is a 2ϵ-packing. Hence $\forall\ \theta_1 \in T_1$, there can be no more than one $\theta_2 \in T_2$ that is within ϵ of θ_1 (otherwise, we would have two elements of T_2 within ϵ of a point in T, which we just showed is not possible). Hence, $|T_2| \leq |T_1|$. Since choices of T_1 and T_2 were arbitrary, pick the smallest T_1 and largest T_2 to obtain:

   $$M(2\epsilon) \leq N(\epsilon)$$

2. Prove $N(\epsilon) \leq M(\epsilon)$: Consider an ϵ-packing of size $M(\epsilon)$, T_2. The goal will be to show that this ϵ-packing is also an ϵ-cover. Fix $\theta \in T$ and the goal is to show there exists a $\theta_2 \in T_2$ s.t., $d(\theta_2, \theta) \leq \epsilon$, implying T_2 is a cover. In case 1, suppose $\theta \in T_2$, in this case, let $\theta_2 = \theta \implies d(\theta_2, \theta) = 0$. In case 2, suppose $\theta \notin T_2$. Since T_2 is maximal, it must be the case that $T_2 \cup \{\theta\}$ is not an ϵ-packing of T because T_2 is the maximal packing. Hence, there exists distinct $\theta_2, \theta_2' \in T \cup \{\theta\}$ s.t. $d(\theta_2, \theta_2') \leq \epsilon$. Because T_2 is an ϵ-packing, one of θ_2 and θ_2' must equal θ. Hence, there exists $\theta_2 \in T_2$ s.t. for general $\theta \in T$:

   $$d(\theta_2, \theta) \leq \epsilon \implies N(\epsilon) \leq M(\epsilon)$$

Example 11 (Functions Lipschitz in Indexing parameter)**.**
Let $f : \mathcal{X} \times B \to \mathbb{R}$ be a function and:

$$\mathcal{F} := \{x \to f(x, \beta) : \beta \in B\}$$

And let $||\cdot||_\beta$, $||\cdot||_{\mathcal{F}}$ denote the norms on B and $\mathcal{F}$. Suppose the following Lipschitz condition holds, there exists an $L > 0$ s.t. $\forall \beta_1, \beta_2 \in B$:

$$||f(\cdot, \beta_1) - f(\cdot, \beta_2)||_{\mathcal{F}} \leq L||\beta_1 - \beta_2||_B$$

Then we can bound the covering number on $\mathcal{F}$ (a hard quantity) by the covering number for the index set (an easier thing):

$$N(\epsilon, \mathcal{F}, ||\cdot||_{\mathcal{F}}) \leq N\left(\frac{\epsilon}{L}, B, ||\cdot||_B\right)$$

Proof: Start on the right and and come up with a minimal cover! Let $\{\beta_j\}_{j=1}^n$ denotes a minimal ϵ/L-cover of B. Let $n = N\left(\frac{\epsilon}{L}, B, ||\cdot||_B\right)$ denote the covering number. We'll show that $\{f(\cdot, \beta_j)\}_{j=1}^n$ is an ϵ-cover for $\mathcal{F}$, completing the proof. For some $g \in \mathcal{F}, g = f(\cdot, \beta)$. Since $\{\beta_j\}_{j=1}^n$ is a cover, $||\beta - \beta_j||_B \leq \frac{\epsilon}{L}$. Now via the Lipschitz condition:

$$\begin{aligned} ||f(\cdot, \beta) - f(\cdot, \beta_j)||_{\mathcal{F}} &\leq L||\beta - \beta_j||_B \\ &\leq L \cdot \frac{\epsilon}{L} \\ &\leq \epsilon \end{aligned}$$

Thus, $\{f(\cdot, \beta_j)\}_{j=1}^n$ is an ϵ-cover for $\mathcal{F}$, so the equality holds.

Theorem 18 (Relation between Bracketing and Covering Numbers)**.**
Let $\mathcal{F} \subset L^r(P)$, $r \in [1, \infty]$. For $\epsilon > 0$, the following bound holds:

$$N_{[]}(2\epsilon, \mathcal{F}, L^r(P)) \leq N(\epsilon, \mathcal{F}, ||\cdot||_\infty) \tag{30}$$

Proof: Start with the thing on the left and create a minimal ϵ-cover. Then exhibit that it produces a bracketing. Let $\{f_j\}_{j=1}^n$ denote a minimal ϵ-cover of $\mathcal{F}$ s.t., $n = N(\epsilon, \mathcal{F}, ||\cdot||_\infty)$. For each $j = \{1, \ldots, n\}$ define a 2ϵ-bracket:

$$[f_j - \epsilon, f_j + \epsilon]$$

B/c $\{f_j\}_{j=1}^n$ covers $\mathcal{F}$, it holds that:

$$\mathcal{F} \subset \bigcup_{i=1}^n [f_j - \epsilon, f_j + \epsilon]$$

Thus, $\{[f_j - \epsilon, f_j + \epsilon]\}_{j=1}^n$ is a 2ϵ-bracket for $\mathcal{F}$, thus, the minimal bracket must have size smaller than n.

Example 12 (Class of Lipschitz functions have stochastically bounded covering number/metric entropy)**.**
Let $\mathcal{F}$ denote a collection of functions $\{f : [0, 1] \to [0, 1]\}$ for which there exists $L > 0$ s.t. for all $x_1, x_2 \in [0, 1]$:

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$$

Let $||\cdot||_\infty$ denote the supremum norm s.t. $||f||_\infty := \sup_x |f(x)|$. Turns out the metric entropy is stochastically bounded:

$$\log N(\epsilon, \mathcal{F}, ||\cdot||_\infty) = O(L/\epsilon) \tag{31}$$

4.6 Upper bounding the Rademacher Complexity

How do we connect the notion of covering and packing numbers to estimation tasks of interest? Recall that to control the regret of an Empirical Risk Minizer, it's enough to control:

$$\text{Regret}(\hat{\theta}) \leq 2\sup_{f\in\mathcal{F}}|(P_n - P)f| := 2\mathbb{E}||P_n - P||_{\mathcal{F}} \leq 2(2||R_n||_{\mathcal{F}}) \quad \text{(By Eq: 20)}$$

Thus, the goal is that upper bounding the Rademacher complexity is tantamount to upper bounding the GC-norm which is tantamount to upper bounding the regret of an ERM. To link the covering/packing numbers to these quantities of interest, we must invoke some stochastic process results.

Definition 16 (Stochastic Process, sub-G process)**.**
A stochastic process $\{X_\theta : \theta \in T\}$ is a collection of random variables. A stochastic process is a sub-Gaussian process with respect to pseudometric d if:

1. Mean zero: $\mathbb{E}[X_\theta] = 0$.
2. Gaussian bound on Cumulant generating function of differences: for all $\theta, \theta' \in \mathbb{R}$ and $\lambda \in \mathbb{R}$:

$$\log \mathbb{E}[\exp(\lambda(X_\theta - X_{\theta'}))] \leq \frac{\lambda^2 d(\theta,\theta')^2}{2}$$

i.e., $\forall \theta, \theta' \in T$, $(X_\theta - X'_\theta)$ is sub-G with parameter $d(\theta,\theta')^2$

Example 13 (Canonical Rademacher Process is sub-G)**.**
Let $S = \mathbb{R}^n$ and $d := ||\cdot||_2$ be the Euclidean norm. Let $T \subset S$ denote the index set and $r := (r_1, \ldots, r_n)$ denote iid Rademacher random variables. The canonical Rademacher process, $\{X_\theta : \theta \in T\}$, is defined as:

$$X_\theta = \sum_{i=1}^n \theta_i r_i = \langle \theta, r \rangle \tag{32}$$

The canonical Rademacher process is sub-G because Rademacher random variables are sub-G (bounded), and the sum of sub-G variables are also sub-G.

The following Finite class lemma upper bounds the maximum deviation of a sub-G process. Turns out, the maximum deviation of a sub-G process only scales logarithmically with the size of the index set A. The Finite Class Lemma will be used to provide an upper bound on the expected supremum of a sub-G process, which we will later connect to the Rademacher complexity.

Theorem 19 (Finite Class Lemma (for sub-G processes))**.**
If $\{X_\theta : \theta \in T\}$ is sub-G wrt d, and $A \subset T \times T$ is the index set:

$$\begin{aligned}\mathbb{E}\left[\max_{(\theta,\theta')\in A}(X_\theta - X_{\theta'})\right] &\leq \sqrt{2\log|A|}\cdot \max_{(\theta,\theta')\in A} d(\theta,\theta') \\ &\leq \sqrt{2\log|A|}\cdot \max_{(\theta,\theta')\in T} d(\theta,\theta') \\ &= \sqrt{2\log|A|}\,\text{Diameter}(T)\end{aligned} \tag{33}$$

We now can bound the supremum of a sub-G process using a one discretization bound, which depends on the covering number!

Theorem 20 (One step discretization bound)**.**
Let $\{X_\theta : \theta \in T\}$ denote a mean 0, sub-G process with respect to d. Let $D := \max_{(\theta,\theta')\in T} d(\theta,\theta')$ denote the diameter of T. For all $\epsilon > 0$:

$$\mathbb{E}[\sup_{\theta\in T} X_\theta] \leq \underbrace{2\,\mathbb{E}\left[\sup_{\theta,\theta'\in T: d(\theta,\theta')\leq\epsilon} (X_\theta - X'_\theta)\right]}_{1} + \underbrace{2D\sqrt{\log N(\epsilon, T, d)}}_{2} \tag{34}$$

Proof: recall that $\{X_\theta : \theta \in T\}$ is a mean 0 process:

$$\begin{aligned}\mathbb{E}[\sup_{\theta\in T} X_\theta] &= \mathbb{E}[\sup_{\theta\in T}(X_\theta - X_{\theta'})] \quad \text{(mean 0)}\\ &\leq \mathbb{E}\left[\sup_{(\theta,\theta')\in T}(X_\theta - X'_\theta)\right]\end{aligned}$$

Let T_1 denote the minimal ϵ-cover for T. Fix $\theta, \theta' \in T$ and let $\theta_1, \theta'_1 \in T_1$, s.t. $d(\theta,\theta_1) \leq \epsilon$, $d(\theta',\theta'_1) \leq \epsilon$. By an add-subtract trick we obatin:

$$\begin{aligned}X_\theta - X_{\theta'} &= (X_\theta - X_{\theta_1}) - (X_\theta - X_{\theta'_1}) + (X_{\theta_1} - X_{\theta'_1})\\ &\leq 2 \sup_{(\theta_2,\theta_3)\in T: d(\theta_2,\theta_3)\leq\epsilon} (X_{\theta_2} - X_{\theta_3}) + \max_{(\theta_4,\theta_5)\in T} X_{\theta_4} - X_{\theta_5}\end{aligned}$$

Combining these two displays yields:

$$\mathbb{E}[\sup_{\theta\in T} X_\theta] \leq 2\mathbb{E}\left[\sup_{(\theta_2,\theta_3)\in T: d(\theta_2,\theta_3)\leq\epsilon} (X_{\theta_2} - X_{\theta_3})\right] + \mathbb{E}\left[\max_{(\theta_4,\theta_5)\in T} X_{\theta_4} - X_{\theta_5}\right]$$

Notice that the first term matches 1. To show that the second term is upper bounded by 2, we use finite class lemma:

$$\begin{aligned}\mathbb{E}\left[\max_{(\theta_4,\theta_5)\in T} X_{\theta_4} - X_{\theta_5}\right] &\leq \sqrt{2\log|T_1 \times T_1|}\; D\\ &\leq \sqrt{2\log|T_1|^2}\; D\\ &\leq 2\sqrt{N(\epsilon,T,d)}D \quad \text{(Bc } T_1 \text{ is } \epsilon\text{-cover)}\end{aligned}$$

Now we have all the tools to bound the Rademacher complexity of a class with bounded range symmetric about zero.

Theorem 21 (Bounding Rademacher Complexity via 1 step discretization bound)**.** Suppose $\mathcal{F}$ is a class of functions with range $[-M, M]$ then $\forall\delta > 0$:

$$\mathbb{E}||R_n||_{\mathcal{F}} \leq 2\delta + 4Mn^{-1/2}\sup_Q\sqrt{\log(2N(\delta,\mathcal{F},L^2(Q)))} \tag{35}$$

Proof: In 3 steps:

1. Relate Rademacher complexity to supremum of sub-G process
2. Applying the one-step discretization bound
3. Find a bound in terms of the covering number of $\mathcal{F}$.

Step 1, Relate Rademacher Process to Supremum of Sub-G Process: Let Z_1^n denote an iid sample from DGD P. $T := \mathcal{F}_{Z_1^n} \cup -\mathcal{F}_{Z_1^n}$ where $\mathcal{F}_{z_1^n} = \{(f(z_1), \dots, f(Z_n)) : f \in \mathcal{F}\}$ is the projection of Z_1^n onto $\mathcal{F}$. Recall the canonical Rademacher process:

$$\left\{ X_\theta = \sum_{i=1}^n \theta_i r_i \equiv \langle \theta, r \rangle : \theta \in T \subset \mathbb{R}^n \right\}$$

Note that we can also define the Rademacher complexity as the expectation of the empirical rademacher complexity:

$$\mathbb{E}||R_n||_{\mathcal{F}} = \mathbb{E}\left[\mathbb{E}[||R_n||_{\mathcal{F}}|Z_n^1]\right]$$

hence it is enough to bound the empirical Rademacher complexity for a generic realization of the data: $\mathbb{E}[||R_n||_{\mathcal{F}}|Z_1^n = z_1^n]$:

$$\begin{aligned}
\mathbb{E}[||R_n||_{\mathcal{F}}|Z_1^n = z_1^n] &:= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(Z_i) r_i \right| \middle| Z_1^n = z_1^n \right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\langle f(z_i), r_i \rangle|\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{f\in\mathcal{F}\cup-\mathcal{F}} \langle f(z_i), r_i \rangle\right] \qquad (\text{B/c } |a| = \max(a, -a)) \\
&= \frac{1}{n}\mathbb{E}\left[\sup_{\theta\in T} X_\theta\right] \qquad (\text{By defn } T \text{ and } X_\theta \text{ is canonical Rademacher process})
\end{aligned}$$

Step 2, Apply one-step discretization bound: since X_θ is the canonical rademacher process, it is sub-G with mean 0, so we can apply the bound in Equation 34.

$$\mathbb{E}\left[\sup_{\theta\in T} X_\theta\right] \leq 2\mathbb{E}\left[\sup_{\theta,\theta'\in T: d(\theta,\theta')\leq\epsilon} (X_\theta - X'_\theta)\right] + 2D\sqrt{\log N(\epsilon, T, d)}$$

In our cases, d is the Euclidean metric because the Rademacher complexity is sub-G wrt Euclidean metric. The first term can be written:

$$\begin{aligned}
2\mathbb{E}\left[\sup_{\theta,\theta'\in T: d(\theta,\theta')\leq\epsilon} (X_\theta - X'_\theta)\right] &:= 2\mathbb{E}\left[\sup_{\theta,\theta'\in T: ||\theta-\theta'||_2\leq\epsilon} \langle \theta - \theta', r\rangle\right] \\
&\leq \mathbb{E}\left[\sup_{\nu\in\mathbb{R}^n: ||\nu||_2\leq\epsilon} \langle \nu, r\rangle\right] \\
&= \mathbb{E}\left[\langle \nu := \frac{\epsilon r}{||r||}, r\rangle\right] \qquad (\text{Dot product maximized by same direction}) \\
&= \epsilon\mathbb{E}[||r||_2] \\
&= \epsilon\sqrt{n}
\end{aligned}$$

Implying when combined with step 1:

$$\mathbb{E}[||R_n||_{\mathcal{F}}|Z_1^n = z_1^n] \leq \frac{1}{n}\left(2\epsilon\sqrt{n} + 2D\sqrt{\log N(\epsilon, T, ||\cdot||_2)}\right)$$

Step 3: find covering number wrt $\mathcal{F}$, not T. To do this we need to relate the covering number of $T := \mathcal{F} \cup -\mathcal{F}$ to the covering number of $\mathcal{F}$. Fix two $\theta_1, \theta_2 \in T$, then there exist $f_1, f_2 \in \mathcal{F} \cup -\mathcal{F}$ s.t. $\theta_1 = (f_1(z_1), \dots, f_1(z_n))$ and $\theta_2 = (f_2(z_1), \dots, f_2(z_n))$. Hence,

$$\begin{aligned}
||\theta_1 - \theta_2||_2 &= \sqrt{\sum_{i=1}^n [\theta_{1i} - \theta_{2i}]^2} \\
&= \sqrt{\sum_{i=1}^n [f_1(z_i) - f_2(z_i)]^2} \\
&= n^{1/2}\sqrt{\frac{1}{n}\sum_{i=1}^n [f_1(z_i) - f_2(z_i)]^2} \\
&= n^{1/2}||f_1 - f_2||_{L^2(P_n)} < \epsilon \\
&\implies ||f_1 - f_2||_{L^2(P_n)} < \epsilon n^{-1/2}
\end{aligned}$$

Where P_n denotes the empirical distribution of Z_1^n. Thus,

$$N(\epsilon, T, ||\cdot||_2) = N(\epsilon\, n^{-1/2}, \mathcal{F} \cup -\mathcal{F}, L^2(P_n))$$

Plugging into the result from step 2 and sup-ing out the random quantity P_n:

$$\begin{aligned}
\mathbb{E}[||R_n||_{\mathcal{F}}|Z_1^n = z_1^n] &\leq 2\epsilon n^{-1/2} + 2Dn^{-1}\sqrt{\log N(\epsilon\, n^{-1/2}, \mathcal{F} \cup -\mathcal{F}, L^2(P_n))} \\
&\leq 2\epsilon n^{-1/2} + 2Dn^{-1}\sup_Q \sqrt{\log N(\epsilon\, n^{-1/2}, \mathcal{F} \cup -\mathcal{F}, L^2(Q))}
\end{aligned}$$

Now we focus on the diameter! Since in $\mathcal{F}$, the range is contained in $[-M, M]$:

$$\begin{aligned}
D_{z_1^n} &= \sup_{(\theta_1,\theta_2) \in \mathcal{F}_{z_1^n} \cup -\mathcal{F}_{z_1^n}} ||\theta_1 - \theta_2||_2 \\
&= \sup_{f_1, f_2 \in \mathcal{F} \cup -\mathcal{F}} ||f_1(z_1^n) - f_2(z_1^n)||_2 \\
&= \sup_{f_1, f_2 \in \mathcal{F} \cup -\mathcal{F}} \sqrt{\sum_{i=1}^n [f_1(z_i) - f_2(z_i)]^2} \\
&\leq \sqrt{4M^2 n} = 2Mn^{1/2}
\end{aligned}$$

Now we focus on getting the covering number in terms of $\mathcal{F}$. Suppose T_1 is a minimal ϵ-cover for $\mathcal{F}$. Let $|T_1| = N(\epsilon, \mathcal{F}, L^2(P_n))$.
Propose the following set $T_2 := T_1 \cup -T_1$ s.t. $|T_2| \leq 2N(\epsilon, \mathcal{F}, L^2(P_n))$. Clearly, since $T_1 \subset T_2$, T_2 is an ϵ-cover for $\mathcal{F}$. Now we show it's also an ϵ-cover for $-\mathcal{F}$, hence a cover for the union. For any $f \in \mathcal{F}$, $\exists\ t \in T_1$ s.t. $||t - f||_{L^2(P_n)} < \epsilon$. For any general $-f \in -\mathcal{F}$, there also exists $t \in -T_1 \subset T_1 \cup -T_1$ s.t. $||(-t) - (-f)||_{L^2(P_n)} \equiv ||t - f||_{L^2(P_n)} < \epsilon$. Thus, T_2 is an ϵ-cover for $\mathcal{F} \cup -\mathcal{F}$:

$$N(\epsilon, \mathcal{F} \cup -\mathcal{F}, L^2(P_n)) \leq 2N(\epsilon, \mathcal{F}, L^2(P_n))$$

Letting $\delta := \epsilon n^{-1/2}$ and replacing the covering number, we obtain our result:

$$\mathbb{E}[||R_n||_{\mathcal{F}}] = \mathbb{E}[||R_n||_{\mathcal{F}}|Z_1^n = z_1^n] \leq 2\delta + 4Mn^{-1/2}\sup_Q \sqrt{\log 2N(\delta, \mathcal{F}, L^2(Q))}$$

Dudley observed that in the 1-step discretization bound, the latter term could be replaced by a term that did not diverge as $\delta \to 0$ under some appropriate conditions. This result has a special application to studying the Rademacher complexity:

Theorem 22 (Dudley's Entropy Integral)**.**
Let $\{X_\theta : \theta \in T\}$ denote a mean-0 sub-G process with respect to pseudometric d. Let D denote the diameter. For any $\epsilon > 0$:

$$\mathbb{E}\left[\sup_{\theta \in T} X_\theta\right] \leq \mathbb{E}\left[\sup_{\theta, \theta' : d(\theta, \theta') \leq \epsilon} (X_\theta - X'_\theta)\right] + 8 \int_{\epsilon/2}^{D} \sqrt{\log(N(\tilde{\epsilon}, T, d))} d\tilde{\epsilon} \tag{36}$$

If $\{X_\theta : \theta \in T\}$ is the canonical Rademacher process:

$$\mathbb{E}\left[\sup_{\theta \in T} X_\theta\right] \leq 8 \int_{0}^{D} \sqrt{\log(N(\tilde{\epsilon}, T, d))} d\tilde{\epsilon} \tag{37}$$

And the bound is not vacuous when the integral is finite, i.e., $\log N(\epsilon) = C\epsilon^{-r}$ for $r < 2$.

Proof: Let T_1 denote a minimal ϵ-cover of T. For a general $\theta \in T$, let θ_1 be s.t $d(\theta_1, \theta) \leq \epsilon$. Then we have teh following:

$$\begin{aligned} X_\theta &= (X_\theta - X_{\theta_1}) + X_{\theta_1} \\ &\leq \sup_{\theta', \tilde{\theta}' : d(\theta', \tilde{\theta}') \leq \epsilon} (X_{\theta'} - X_{\tilde{\theta}'}) + \sup_{\tilde{\theta}_1 \in T_1} X_{\tilde{\theta}_1} \\ \implies \mathbb{E}\left[\sup_{\theta \in T} X_\theta\right] &\leq \mathbb{E}\left[\sup_{\theta', \tilde{\theta}' : d(\theta', \tilde{\theta}') \leq \epsilon} (X_{\theta'} - X_{\tilde{\theta}'})\right] + \mathbb{E}\left[\sup_{\tilde{\theta}_1 \in T_1} X_{\tilde{\theta}_1}\right] \quad \text{(Expectations and sup on LHS)} \end{aligned}$$

Note that LHS and first term on RHS are equivalent to Dudley's entropy. We just need to show:

$$\mathbb{E}\left[\sup_{\tilde{\theta}_1 \in T_1} X_{\tilde{\theta}_1}\right] \leq 8 \int_{\epsilon/2}^{D} \sqrt{N(\tilde{\epsilon}, T, d)} d\tilde{\epsilon}$$

Recall the finite class lemma from Equation 33: for $A \subset T \times T$

$$\begin{aligned} \mathbb{E}\left[\max_{(\theta, \theta') \in A} (X_\theta - X_{\theta'})\right] &\leq \sqrt{2 \log |A|} \cdot \max_{(\theta, \theta') \in A} d(\theta, \theta') \\ &\leq \sqrt{2 \log |A|} \cdot \underbrace{\max_{(\theta, \theta') \in T} d(\theta, \theta')}_{\text{Diameter}} \end{aligned}$$

Let's leverage the FCL more by constructing an A to work with such that the maximum is much smaller than D. Consider the following telescoping sum for $k \in \mathbb{N}$:

$$X_{\theta_1} = X_{\theta_k} + \sum_{i=1}^{k-1} (X_{\theta_j} - X_{\theta_{j+1}})$$

(j=1 display): Let T_1 and T_2 denote ϵ and 2ϵ-covers of T respectively. For general elements $\theta_1 \in T_1$ and $\theta_2 \in T_2$, $d(\theta_1, \theta_2) \leq 2\epsilon$, and consider $f_2 : T_1 \to T_2$, i.e., $\theta_2 = f_2(\theta_1) \in T_2$ s.t. $d(\theta_1, \theta_2) \leq 2\epsilon$. Define the set:

$$A_1 = \{(f_2(\theta_1), \theta_1) : \theta_1 \in T_1\} \quad \text{s.t.} \quad \max_{(\theta, \theta') \in A_1} d(\theta, \theta') \leq 2\epsilon$$

And since $A_1 \subset T_2 \times T_1$:

$$\begin{aligned} \log |A_1| &\leq \log |T_2 \times T_1| \\ &\leq \log(N(2\epsilon) N(\epsilon)) \\ &\leq 2 \log N(\epsilon) \end{aligned}$$

By FCL:

$$\mathbb{E}\left[\sup_{(\theta,\theta')\in A_1}(X_\theta - X_{\theta'})\right] \leq 2\epsilon\sqrt{2\cdot 2\log(N(\epsilon))} = 4\epsilon\sqrt{\log(N(\epsilon))}$$

(general j display): Recall the telescoping sum:

$$\begin{aligned} X_{\theta_1} &= X_{\theta_k} + \sum_{i=1}^{k-1}(X_{\theta_j} - X_{\theta_{j+1}}) \\ &\leq X_{\theta_k} + \max_{(\theta,\theta')\in A_1}(X_{\theta'} - X_\theta) + \sum_{j=2}^{k-1}(X_{\theta_j} - X_{\theta_{j+1}}) \end{aligned}$$

Then take an expectation and apply the FCL lemma to control the second term. To control the third term, we iterate! We consider elements in the ϵ-cover, 2ϵ-cover, 4ϵ-cover, ….
Formally, for $j = 1, \ldots, k-1$, let $f_{j+1} : T_j \to T_{j+1}$ where T_j is the $2^{j-1}\epsilon$-cover and T_{j+1} is the $2^j\epsilon$-cover. Letting $\theta_{j+1} = f_{j+1}(\theta_j)$ and:

$$d(\theta_{j+1}, \theta_j) \leq 2^j\epsilon$$

We define:

$$A_j := \{(f_{j+1}\circ\ldots\circ f_2(\theta_1), f_j\circ\ldots\circ f_2\theta_1) : \theta_1 \in T_1\}$$

By identical arguments to the $j = 1$ case:

$$\mathbb{E}\left[\sup_{(\theta,\theta')\in A_j}(X_\theta - X_{\theta'})\right] \leq 2\cdot 2^j\epsilon\sqrt{\log N(2^{j-1}\epsilon)}$$

Note that once $2^j\epsilon \geq D := \sup_{(\theta,\theta'):d(\theta,\theta')\in T} d(\theta,\theta')$, a minimal $2^j\epsilon$-cover conatins one element. Now choose k to be minimal s.t. $2^{k-1}\epsilon \geq D$. Thus, $\theta_k = f_k\circ\ldots f_2(\theta_1)$ does not depend on the value of θ_1, since θ_k can only take one value. Returning to the telescoping sum, we can obtain an upper bound:

$$\begin{aligned} X_{\theta_1} &= X_{\theta_k} + \sum_{i=1}^{k-1}(X_{\theta_j} - X_{\theta_{j+1}}) \\ &\leq X_{\theta_k} + \sum_{j=1}^{k}\max_{(\theta,\theta')\in A_j}[X_{\theta'} - X_\theta] \end{aligned}$$

And via the FCL:

$$\begin{aligned} \mathbb{E}[X_{\theta_1}] &\leq \mathbb{E}[X_{\theta_k}] + \sum_{j=1}^{k}\mathbb{E}\left[\max_{(\theta,\theta')\in A_j} X_{\theta'} - X_\theta\right] \\ &\leq 0 + \sum_{j=1}^{k} 2\cdot 2^j\epsilon\sqrt{\log N(2^{j-1}\epsilon)} \\ &= 8\sum_{j=1}^{k-1} 2^{j-2}\epsilon\sqrt{\log N(2^{j-1}\epsilon)} \end{aligned}$$

Heuristic: consider $N(\epsilon)$ as function of ϵ. $N(\epsilon)$ is a monotontically decreasing step function. Consider the step between $2^{j-2}\epsilon$ and $2^{j-1}\epsilon$. The right hand area under the curve is equal to the width of the step $(2^{j-2}\epsilon)$

times the height ($\sqrt{\log N(2^{j-1}\epsilon)}$). This quantity lower bounds the integral from these two steps. Thus,

$$8\sum_{j=1}^{k-1} 2^{j-2}\epsilon\sqrt{\log N(2^{j-1}\epsilon)} \leq 8\int_{\epsilon/2}^{D}\sqrt{\log N(\tilde{\epsilon})}d\tilde{\epsilon}$$

Where the lower bound of the integral ius derived from the $j=1$ case and D denotes the largest distance permitted in T bc $\log N(u)=0$ for $u \geq D$. Thus, we are done and we have proved Equation 36.

To prove Equation 37 for the Canonical rademacher process, we rely on the simple fact that term 2 can be written as:

$$\mathbb{E}\left[(\theta,\theta') : \sup_{d(\theta,\theta')\leq\epsilon}(X_\theta - X_{\theta'})\right] \leq \epsilon\sqrt{n}$$

So combining with the previous part, as we let $\epsilon \to 0$, we obtain:

$$\mathbb{E}\left[\sup_{\theta\in T} X_\theta\right] \leq 8\int_0^D \sqrt{\log N(\tilde{\epsilon})}d\tilde{\epsilon}$$

Turns out, we can use Dudley's entropy integral to control the Rademacher Complexity.

Theorem 23 (Controlling Rademacher Complexity via Dudley)**.** Suppose $\mathcal{F}$ is a function class from $Z \to \mathbb{R}$ and $\mathcal{F} = -\mathcal{F}$ (closed under negations). Then:

$$||R_n||_{\mathcal{F}} \leq \frac{8}{\sqrt{n}}\mathbb{E}_{P_n}\left[\int_0^\infty \sqrt{\log N(\epsilon,\mathcal{F},L^2(P_n))}d\epsilon\right] \leq \frac{8}{\sqrt{n}}\sup_Q\left[\int_0^\infty \sqrt{\log N(\epsilon,\mathcal{F},L^2(Q))}d\epsilon\right] \tag{38}$$

Where the sup is over all finitely-supported probability distributions with support in P.
Implying Regret of ERM $= \mathcal{O}(n^{-1/2})$.

Proof:

$$\begin{aligned}
n\mathbb{E}||R_n||_{\mathcal{F}} &= n\mathbb{E}_{P_n}[\mathbb{E}[||R_n||_{\mathcal{F}}|Z_1^n = z_1^n]] \\
&= \mathbb{E}_{P_n}\left[\mathbb{E}\left[\sup_{\theta\in\mathcal{F}_{z_1^n}}\langle r,\theta\rangle\Big|Z_1^n = z_1^n\right]\right] \qquad \text{(Rad complex to Rad process)} \\
&\leq \mathbb{E}_{P_n}\left[8\int_0^\infty \sqrt{\log N(\epsilon,\mathcal{F},||\cdot||_2)}d\epsilon\right] \qquad \text{(Dudley)}
\end{aligned}$$

Based on a step in the proof of Equation 35, we replace $N(\epsilon,\mathcal{F}_{z_1^n},||\cdot||_2) = N(\epsilon/\sqrt{n},\mathcal{F},L^2(P_n))$.

$$\begin{aligned}
n\mathbb{E}||R_n||_{\mathcal{F}} &\leq \mathbb{E}_{P_n}\left[8\int_0^\infty \sqrt{\log N(\epsilon/\sqrt{n},\mathcal{F},L^2(P_n))}d\epsilon\right] \\
&= 8\sqrt{n}\mathbb{E}_{P_n}\left[\int_0^\infty \sqrt{\log N(u,\mathcal{F},L^2(P_n))}du\right] \\
\implies \mathbb{E}||R_n||_{\mathcal{F}} &\leq 8n^{-1/2}\mathbb{E}_{P_n}\left[\int_0^\infty \sqrt{\log N(u,\mathcal{F},L^2(P_n))}du\right]
\end{aligned}$$

Thus, by Dudley's entropy integral, we obtain:

$$\begin{aligned}
\mathbb{E}||R_n||_{\mathcal{F}} &\leq \frac{8}{n}\sup_Q\int_0^\infty \sqrt{\log N(u,\mathcal{F},L^2(Q))}du \\
&\leq \frac{8}{n}\sup_Q\int_0^\infty \sqrt{\log N(u,\mathcal{F},L^2(Q))}du
\end{aligned}$$

Below, we'll give some useful examples:

Example 14 (Rademacher Complexity of Lipschitz Functions)**.**
Let $\mathcal{F}$ denote a class of $[0,1] \to [0,1]$ functions s.t. for all $z_1, z_1 \in [0,1]$

$$|f(z_1) - f(z_2)| \leq L|z_1 - z_2|$$

We previously saw that $\log N(\epsilon, \mathcal{F}, ||\cdot||_\infty) = \mathcal{O}(L/\epsilon)$ and since $||\cdot||_{L^2(Q)} \leq ||\cdot||_\infty$ for all Q:

$$\sup_Q \log N(\epsilon, \mathcal{F}, L^2(Q)) = \mathcal{O}(L/\epsilon)$$

$$\mathbb{E}||R_n||_{\mathcal{F}} \leq 8n^{-1/2} \sup_Q \left[\int_0^\infty \sqrt{\log N(u, \mathcal{F}, L^2(Q))} du\right]$$

$$\leq 8n^{-1/2} \underbrace{\int_0^\infty \mathcal{O}(L/\epsilon) d\epsilon}_{<\infty}$$

$$\implies \mathbb{E}||R_n||_{\mathcal{F}} = \mathcal{O}(n^{-1/2})$$

Now consider the case where $d \geq 2$. $\mathcal{F}$ is the collection of L-lipschitz functions satisfying:

$$|f(z_1) - f(z_2)| \leq L||z_1 - z_2||_\infty \forall z_1, z_2 \in [0,1]^d$$

In this case,

$$\sup_Q \log N(\epsilon, \mathcal{F}, L^2(Q)) = \mathcal{O}\left[\left(\frac{L}{\epsilon}\right)^d\right]$$

Turns out, Equation 38 will not give us finite value for the uniform entropy integral when $d \geq 2$. So instead, we rely on the bound in Equation 36, which gives us slower convergence than $n^{-1/2}$ rate.

4.7 Upper bounding the empirical process term via bracketing integrals

We start with the definition of an envelope function:

Definition 17 (Envelope function)**.**
An *envelope function* F for a function class $\mathcal{F}$, is a function that pointwise dominates the absolute value of every function in the function class: i.e., $|f(z)| \leq F(z)\ \forall\ z\ \forall\ f \in \mathcal{F}$.
Also let $||F||_{Q,r} := [QF^r]^{1/r}$

Turns out we can bound the covering number of VC classes of functions, and the upper bound is polynomial in $1/\epsilon$, meaning that VC classes of functions are relatively small!

Theorem 24 (Covering No. of VC function classes (Lemma 19.15 VdV))**.**
Special case ($r = 2$, Envelope function $F = 1$): Let F denote a collection of VC functions that map from $\mathcal{Z} \to [-1,1]$ and with $V_i(\mathcal{F})$ denoting the VC index. Then it holds that:

$$\sup_Q N(\epsilon, \mathcal{F}, L^2(Q)) \leq kV_i(\mathcal{F})(16e)^{V_i(\mathcal{F})} \cdot \left(\frac{1}{\epsilon}\right)^{2(V_i(\mathcal{F})-1)} \tag{39}$$

General case: Let $r \geq 1$. Given an envelope function F on F, a VC class of functions, the following inequality holds:

$$\sup_Q N(\epsilon||F||_{Q,r}, \mathcal{F}, L^r(Q)) \leq kV_i(\mathcal{F})(16e)^{V_i(\mathcal{F})} \cdot \left(\frac{1}{\epsilon}\right)^{r(V_i(\mathcal{F})-1)} \tag{40}$$

We can also upper bound the empirical process term (which upper bounds the regret) via the bracketing integral:

Theorem 25 (Bracketing integral bound on the empirical process term)**.**
For any class of functions mapping from $\mathcal{Z}$ to $[-1, 1]$, it holds that

$$\mathbb{E}||P_n - P||_{\mathcal{F}} \leq \frac{c}{\sqrt{n}} \int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)}d\epsilon$$

For any class with envelope function F:

$$\mathbb{E}||P_n - P||_{\mathcal{F}} \leq \frac{c}{\sqrt{n}}||F||_{P,2} \int_0^1 \sqrt{\log N_{[]}(\epsilon||F||_{P,2}, \mathcal{F}, L^2(P)}d\epsilon$$

If the bracketing integral is finite, then the bound is not vacuous, and $\mathbb{E}||P_n - P||_{\mathcal{F}} = \mathcal{O}(n^{-1/2})$. This also means that the expected value of the empirical empirical process $\mathbb{E}\mathbb{G}_n f := \mathbb{E}\sqrt{n}(P_n - P)f = \mathcal{O}(1)$ for all $f \in \mathcal{F}$, meaning the empirical process convergence to a tight limit process in $\mathcal{F}$, implying that $\mathcal{F}$ is P-Donsker (see VdV 19.2 (pg 269)).

5 Useful facts

5.1 Useful inequalities

1. Jensen: if f is convex, $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$. If f is concave, $\mathbb{E}(f(X)) \leq f(\mathbb{E}(X))$.
2. Triangle: $|a+b| \leq |a| + |b|$
3. Reverse triangle: $||a| - |b|| \leq |a-b|$
4. Kolmogorov: tail bound on maximum partial sum. For independent $Z_1, \ldots, Z_n$ with $\mathbb{E}(Z) = 0, \mathbb{E}(Z_i^2) < \infty$:

$$P\left\{\max_{1\leq m\leq n}\left|\sum_{i=1}^{m} Z_i\right| > t\right\} \leq \frac{\sum_{i=1}^{n}\mathbb{E}(Z_i^2)}{t^2}$$

5.2 Useful analysis results

1. Continuous functions on compact supports are bounded.
2. $1 + x \leq \exp(x)$
3. b-th moment for non-negative random variable:

$$\mathbb{E}[X^b] = b\int_0^\infty x^b P(X > x)dx$$

4. Infinite series sum: $\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$
5. Taylor series for $e^{tX} = \sum_{b=0}^{\infty}\frac{t^b}{b!}X^b$
6. Taylor expansion for f about x_0:

$$f(x) = \sum_{n=0}^{\infty}\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$$

5.3 Useful concentration inequality results

1. Markov inequality is the foundational inequality, but most MGF-based results are based on Chernoff.
2. Sums of sub-G random variables are sub-G.
3. Suppose $X_1, \ldots, X_n$ are independent variables s.t. $\mathbb{E}[X_i] = \mu_i$ and $X_i \in SE(\sigma_i^2, a_i)$, then:

$$\sum_{i=1}^{n}(X_i - \mu_i) \in \text{SE}(\sum_{i=1}^{n}\sigma_i^2, \max_i a_i)$$

4. Any sub-G random variable with parameter σ^2 is also sub-exponential with parameters (σ^2, b) for any $b > 0$.