Statistical Learning Theory Cookbook

Ronak Mehta

Contents

1 Concentration Inequalities

Theorem 1.1 (Markov's inequality). *If* $\mathbb{E}[X] < \infty$*,* $t > 0$*,* $h : [0, \infty) \to [0, \infty)$ *is non-decreasing, and* $\mathbb{E}[h(|X - \mathbb{E}[X]|)] < \infty$*, then*

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

Theorem 1.2 (Chernoff bound). *Let* X *have a moment generating function in a neighborhood of zero, meaning that there is some constant* $b > 0$ *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*

$$\log \mathbb{P}[X - \mathbb{E}[X] \geq t] \leq -\sup_{\lambda > 0} [\lambda t - \log M_{X-\mu}(\lambda)].$$

Definition 1.1 (Sub-Gaussianity). A random variable X is called *sub-Gaussian* with parameter σ^2 if, for all $\lambda \in \mathbb{R}$,

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

Proposition 1.1 (Equivalent characterization sub-Gaussianity). *Let* X *be a mean-*μ *random variable.* X *is sub-Gaussian if and only if there exist* $c, s > 0$ *such that*

$$\mathbb{P}[|X - \mu| > t] \leq c\mathbb{P}[|sZ| \geq t]$$

for all $t > 0$ *(where* $Z \sim N(0, 1)$*).*

Proposition 1.2 (Tail inequality for sub-Gaussian variables). *Let* X *be a mean-*μ *sub-Gaussian variable with parameter* σ^2*. Then,*

$$\log \mathbb{P}[X - \mu \geq t] \leq -\frac{t^2}{2\sigma^2}.$$

Return to Appendix A.

Theorem 1.3 (Hoeffding's inequality)**.** *If the support of a random variable X is bounded in $[a, b]$, then*

1. *X is sub-Gaussian with parameter $\sigma^2 = \frac{(b-a)^2}{4}$.*
2. *It holds that*

$$\log \mathbb{P}\left[X - \mu \geq t\right] \leq -\frac{2t^2}{(b-a)^2}.$$

When $X_1, ..., X_n$ are independent with support contained in $[a, b]$, then

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

Definition 1.2 (Sub-exponentiality)**.** A random variable X is sub-exponential with parameters (σ^2, b) if, for all $|\lambda| \leq \frac{1}{b}$,

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

Proposition 1.3 (Equivalent characterization sub-Gaussianity)**.** *Let X be a mean-μ random variable. X is sub-exponential if and only if there exist $c, \ell > 0$ such that*

$$\mathbb{P}\left[|X - \mu| \leq t\right] \leq c\mathbb{P}\left[|\mathcal{E}_\ell| \geq t\right]$$

for all $t > 0$ (where $\mathcal{E}_\ell \sim Exp(-\ell t)$).

Proposition 1.4 (Tail inequality for sub-exponential variables)**.** *Let X be a mean-μ sub-exponential with parameter (σ^2, b). Then,*

$$\log \mathbb{P}\left[X - \mu \geq t\right] \leq \begin{cases} -\frac{t^2}{2\sigma^2} & \text{if } 0 \leq t \leq \frac{\sigma^2}{b}, \\ -\frac{t}{2b} & \text{if } t > \frac{\sigma^2}{b} \end{cases}.$$

Theorem 1.4 (Bernstein's inequality)**.** *If a mean-μ random variable X is bounded in $[\mu - b, \mu + b]$ with variance σ^2, then for all $t > 0$,*

$$\log \mathbb{P}\left[X - \mu \geq t\right] \leq -\frac{t^2}{2(\sigma^2 + bt)}.$$

For independent random variables $X_1, ..., X_n$ with $|X_i - \mu_i| \leq b$ and variances σ_i^2, it holds that for all $t > 0$,

$$\log \mathbb{P}\left[X - \mu \geq t\right] \leq -\frac{nt^2}{2(\frac{1}{n}\sum_{i=1}^n \sigma_i^2 + bt)}.$$

Return to Appendix A.

Definition 1.3 (Bounded differences property)**.** A function $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the *bounded differences property* if for all i, there exists a constant $c_i < \infty$ so that the following inequality holds for all $x_1, ..., x_n, x_i' \in \mathcal{X}$:

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

Theorem 1.5 (McDiarmid's inequality)**.** *Let $X = (X_1, ..., X_n)$ be a collection of independent random variables with f satisfying the bounded differences inequality with bounds $c_1, ..., c_n$ and $\mathbb{E}[|f(X)|] < \infty$. Then, for all $t > 0$,*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

2 Complexity of Function Classes

2.1 Motivation, VC Dimension, and Rademacher Complexity

Proposition 2.1 (Uniform convergence bound)**.** *It holds that*

$$\begin{aligned} \textit{Reg}\left(\hat{\theta}\right) &\leq 2\sup_{\theta\in\Theta}|(P_n - P)\ell(\theta)| \\ &= 2\sup_{f\in\mathcal{F}}|(P_n - P)f| \\ &=: 2\,||P_n - P||_{\mathcal{F}}, \end{aligned}$$

Proposition 2.2 (Tail bound for GC-norm)**.** *When $\mathcal{F}$ is a collection of $[0,1]$-valued functions, it holds that*

$$\mathbb{P}[|||P_n - P||_{\mathcal{F}} - \mathbb{E}[||P_n - P||_{\mathcal{F}}]| > t] \leq 2\exp\left(-2nt^2\right).$$

So, to control the tail behavior of $||P_n - P||_{\mathcal{F}}$ it is sufficient to bound its expectation.

Remark 2.1. We needed to put the restriction on $\mathcal{F}$ above so that the function

$$g(x_1, ..., x_n) \equiv \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(x_i) - \mathbb{E}[f(X_1)]\right|$$

satisfies the bounded differences property. This can also be assumed to yield the same result.

Proposition 2.3 (Ghost sample trick/symmetrization)**.** *Letting $X' = (X_1', ..., X_n') \sim P$ iid and P_n' be the corresponding sample mean functional, we have that*

$$\mathbb{E}_X[||P_n - P||_{\mathcal{F}}] \leq \mathbb{E}_{X,X'}[||P_n - P_n'||_{\mathcal{F}}].$$

Return to Appendix A.

Definition 2.1 (Rademacher complexity)**.** The Rademacher process $R_n : \mathcal{F} \to \mathbb{R}$ for sample $X = (X_1, ..., X_n) \sim P$ and mutually independent $\epsilon = (\epsilon_1, ..., \epsilon_n) \sim$ Unif($\{-1, 1\}$) is given by

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

Let $||R_n||_{\mathcal{F}} \equiv \sup_{f\in\mathcal{F}} |R_n(f)|$. Then, the *Rademacher complexity* is $\mathbb{E}_{\epsilon,X}\left[||R_n||_{\mathcal{F}}\right]$

Proposition 2.4 (Rademacher complexity bounds GC-norm)**.** *for any nondecreasing convex function ϕ, we have that*

$$\mathbb{E}\left[\phi\left(||P_n - P||_{\mathcal{F}}\right)\right] \leq \mathbb{E}\left[\phi\left(2\mathbb{E}\left[||R_n||_{\mathcal{F}}\right]\right)\right].$$

In particular, for ϕ as the identity function,

$$\mathbb{E}\left[||P_n - P||_{\mathcal{F}}\right] \leq 2\mathbb{E}\left[||R_n||_{\mathcal{F}}\right]$$

Proposition 2.5 (Desymmetrization)**.** *If $\mathcal{F}$ is a class of $[0,1]$ functions, it also holds by a desymmetrization argument that*

$$\mathbb{E}\left[||P_n - P||_{\mathcal{F}}\right] \geq \frac{1}{2}\mathbb{E}\left[||R_n||_{\mathcal{F}}\right] - \sqrt{\frac{\log 2}{2n}}.$$

Definition 2.2 (Projection)**.** Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $\{0,1\}$. For $(x_1, ..., x_n) \in \mathcal{X}^n$, the *projection of $\mathcal{F}$ onto* $(x_1, ..., x_n)$ is given by

$$\mathcal{F}_{x_1,\ldots,x_n} \equiv \{(f(x_1), ..., f(x_n) : f \in \mathcal{F}\}.$$

Definition 2.3 (Growth function)**.** The *growth function* or *shattering number* of $\mathcal{F}$ at n is given by

$$\Pi_{\mathcal{F}}(n) = \sup_{x_1,\ldots,x_n} \left|\mathcal{F}_{x_1,\ldots,x_n}\right|.$$

It can also be defined for a collections of sets $\mathcal{A}$ of sets, but letting $\mathcal{F} = \{x \mapsto \mathbb{1}_A[x] : A \in \mathcal{A}\}$. This can be thought of as the number of labellings that can be realized by functions from $\mathcal{F}$, maximized over the input points.

Return to Appendix A.

Proposition 2.6 (Properties of growth functions)**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be two families of sets. The growth function satisfies the following.*

- $\Pi_{\mathcal{A}}(n+m) \leq \Pi_{\mathcal{A}}(n)\Pi_{\mathcal{A}}(m)$.
- $\Pi_{\mathcal{A}\cup\mathcal{B}}(n) \leq \Pi_{\mathcal{A}}(n) + \Pi_{\mathcal{B}}(n)$.
- $\Pi_{\{A\cup B: A\in\mathcal{A}, B\in\mathcal{B}\}}(n) \leq \Pi_{\mathcal{A}}(n)\Pi_{\mathcal{B}}(n)$.
- $\Pi_{\{A\cap B: A\in\mathcal{A}, B\in\mathcal{B}\}}(n) \leq \Pi_{\mathcal{A}}(n)\Pi_{\mathcal{B}}(n)$.
- $\Pi_{\mathcal{A}}(n) = \Pi_{\{A^c: A\in\mathcal{A}\}}(n)$.
- $\Pi_{\{A\}}(n) = 1$ *for all* n.
- *If* $\mathcal{A} \subseteq \mathcal{B}$*, then* $\Pi_{\mathcal{A}}(n) \leq \Pi_{\mathcal{B}}(n)$ *for all* n.

Definition 2.4 (VC dimension)**.** The *VC dimension* of a class of sets $\mathcal{A}$ is

$$\text{VC}\,(\mathcal{A}) \equiv \sup\left\{n : \Pi_{\mathcal{A}}(n) = 2^n\right\}.$$

The *VC dimension* of a class of function $\mathcal{F}$ is

$$\text{VC}\,(\mathcal{F}) \equiv \sup\left\{n : \Pi_{\mathcal{F}}(n) = 2^n\right\}.$$

The *VC index* is VC $(\mathcal{F}) + 1$, representing the smallest n at which no set of n points can be shattered by $\mathcal{F}$.

For real-valued functions, the VC dimension is given by the VC dimension of the collection of subgraphs, or

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

Theorem 2.1 (VC dimension bound)**.** *Consider a family of boolean-valued functions*

$$\mathcal{F} = \{x \mapsto f(x,\theta) : \theta \in \mathbb{R}^p\},$$

where each $f : \mathbb{R}^m \times \mathbb{R}^p \to \{0,1\}$*. Suppose that each* f *can be computed using no more than* t *operations of the following type:*

- *arithmetic* ($+$. $-$, $\times$, $/$).
- *comparisons of real numbers* ($>$, $\geq$, $=$, $\neq$, $\leq$, $<$).

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

Return to Appendix A.

Lemma 2.1 (Finite-class lemma). *If $\mathcal{F}$ is a set of functions satisfying $|f(x)| \leq 1$, then*

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

where $X_1^n = (X_1, ..., X_n)$ is a random set of points in $\mathcal{X}$.

Lemma 2.2 (Sauer's lemma). *If VC $(\mathcal{F}) \leq d$, then*

$$\Pi_{\mathcal{F}}(n) \leq \sum_{k=0}^{d} \binom{n}{k}.$$

Consequently,

$$\Pi_{\mathcal{F}}(n) \leq \begin{cases} 2^n & \text{if } n \leq d \\ (\frac{e}{d})^d n^d & \text{if } n > d. \end{cases}$$

In summary, if $\mathcal{F}$ is VC, then $\Pi_{\mathcal{F}}(n) = O(n^d)$.

Proposition 2.7 (Sufficient condition for VC class). *A sufficient but not necessary condition for $\mathcal{F}$ being VC is if $\Pi_{\mathcal{F}}(n) = o(2^n)$.*

Proposition 2.8 (Learning bound for VC). *If VC $(F) \leq d < n \in \mathbb{N}$, then*

$$\mathbb{E}\left[||P_n - P||_{\mathcal{F}}\right] \leq 2\sqrt{\frac{2\log 2 + 2d\log(en/d)}{n}}.$$

2.2 Bounds via Bracketing and Covering Numbers

Definition 2.5 (Bracketing number). Let $\mathcal{F} \subseteq L^r(P)$.

- For $\ell, u \in L^r(P)$, the *bracket* $[\ell, u]$ is the set $\{f : \ell \leq f \leq u$ pointwise$\}$.
- An ϵ-bracket is a bracket for which $||u - \ell||_{L^r(P)} \leq \epsilon$.
- The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, L^r(P))$ of $\mathcal{F}$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$. That is, the minimal m such that there is a collection of brackets $\{[\ell_j, u_j] : j = 1, ..., m\}$ for which $\mathcal{F} \subseteq \cup_{i=1}^m [\ell_j, u_j]$.

Note: The ℓ_j and u_j functions need not belong to $\mathcal{F}$, just $L^r(P)$.

Return to Appendix A.

Theorem 2.2 (Bracketing number GC theorem). *If $\mathcal{F}$ is a class of functions for which $N_{[]}(\epsilon, \mathcal{F}, L^1(P)) < \infty$ for every $\epsilon > 0$, then $\mathcal{F}$ is P-Glivenko-Cantelli, that is,*

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

Definition 2.6 (Covering number and metric entropy). Let (S, d) be a pseudometric space and let $T \subseteq S$.

- A set $T_1 \subseteq T$ is called an ϵ-*cover* of T if, for each $\theta \in T$, there is a $\theta_1 \in T_1$ such that $d(\theta, \theta_1) \leq \epsilon$.
- The ϵ-*covering number* of T is

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

- The function $\epsilon \mapsto \log N(\epsilon, T, d)$ is called the *metric entropy* of T.

Definition 2.7 (Totally bounded). T is called *totally bounded* if, for all $\epsilon > 0$, $N(\epsilon, T, d) < \infty$.

Definition 2.8 (Packing number). Let (S, d) be a pseudometric space and let $T \subseteq S$.

- A set $T_1 \subseteq T$ is called an ϵ-*packing* of T if, for each $\theta, \theta' \in T_1$, $d(\theta, \theta') > \epsilon$.
- The ϵ-*packing number* of T is

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

Theorem 2.3 (Relationship between covering and packing number). *For all ϵ,*

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

Return to Appendix A.

Proposition 2.9 (Covering number examples). *The following are known covering number examples.*

- **Euclidean ball:** *Let $||\cdot||$ be an ℓ^p norm on $\mathbb{R}^d$, $p \geq 1$, and let $B(a,r)$ denote a ball centered at a of radius r. For all $\epsilon \in (0,r]$,*

$$\left(\frac{r}{\epsilon}\right)^d \leq N\left(\epsilon, B(0,r), ||\cdot||\right) \leq \left(\frac{2r}{\epsilon} + 1\right)^d.$$

- **Functions Lipschitz in index:** *Let $f : \mathcal{X} \times B \to \mathbb{R}$ be some function and let*

$$\mathcal{F} \equiv \{x \mapsto f(x, \beta) : \beta \in B\}.$$

Let $||\cdot||_B$ and $||\cdot||_{\mathcal{F}}$ denote norms on B and $\mathcal{F}$, respectively. Suppose the Lipschitz condition holds:

$$||f(\cdot, \beta_1) - f(\cdot, \beta)2)||_{\mathcal{F}} \leq L\, ||\beta_1 - \beta_2||_B$$

for all $\beta_1, \beta_2 \in B$. Then,

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

- **Lipschitz functions in Euclidean space:** *Let $\mathcal{F}$ be the L-Lipschitz $[0,1]^d \to [0,1]$ functions (w.r.t. the* $\sup$*-norms on the domain and range). Then,*

$$\log N\left(\epsilon, \mathcal{F}, ||\cdot||_\infty\right) = \Theta\left(\left(\frac{L}{\epsilon}\right)^d\right).$$

Theorem 2.4 (Relationship between bracketing and cover number). *Let $\mathcal{F} \subset L^r(P)$, $r \in [1,\infty]$. For all $\epsilon > 0$, it is true that*

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

Definition 2.9 (Zero-mean stochastic process). A *stochastic process* $\{X_\theta : \theta \in T\}$ is a collection of random variables. It is called *zero-mean* of $\mathbb{E}\left[X_\theta\right] = 0$ for all $\theta \in T$.

Definition 2.10 (Sub-Gaussian process). In a pseudometric space (S,d), a stochastic process $\{X_\theta : \theta \in T \subseteq S\}$ is called *sub-Gaussian with respect to d* if for all $\theta, \theta' \in T$, and $\lambda \in \mathbb{R}$,

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

Return to Appendix A.

Lemma 2.3 (General finite class lemma). *If $\{X_\theta : \theta \in T\}$ is sub-Gaussian with respect to d, and $A \subseteq T \times T$, then*

$$\mathbb{E}\left[\max_{(\theta,\theta')\in A} (X_\theta - X_{\theta'})\right] \leq \sqrt{2\log|A|} \max_{(\theta,\theta')\in A} d(\theta,\theta').$$

Theorem 2.5 (Supremum bound on sub-Gaussian process). *Let $\{X_\theta : \theta \in T\}$ be a zero-mean sub-Gaussian process. Let $D \equiv \sup_{\theta,\theta'\in T} d(\theta,\theta')$ denote the diameter of T. For any $\epsilon > 0$,*

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

Proposition 2.10 (One-step discretization bound). *The Rademacher complexity satisfies*

$$\mathbb{E}\left[||R_n||_{\mathcal{F}}\right] \leq 2\delta + 2\mathbb{E}\left[D_{Z_1^n}\right] n^{-1} \sup_Q \sqrt{\log 2N(\delta, \mathcal{F}, L^2(Q)}.$$

Theorem 2.6 (Covering number G-C theorem). *Suppose functions in $\mathcal{F}$ have range in $[-M, M]$, and*

$$\sup_Q \log N\left(\delta, \mathcal{F}, L^2(Q)\right) < \infty \text{ for all } \delta.$$

Then, $\mathcal{F}$ is P-Glivenko-Cantelli for all distributions P, that is, for all P,

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

Definition 2.11 (Separable stochastic process). A process $\{X_\theta : \theta\}$ is said to be separable if there exists a countable dense subset $\tilde{T}$ of (T, d) such that the following is almost surely true: for all $\theta \in T$, there exists a sequence $\{\theta_j\}_{j=1}^{\infty}$ in $\tilde{T}$ such that $d(\theta_j, \theta) \to 0$ and $X_{\theta_j} \to X_\theta$ as $j \to \infty$.

Theorem 2.7 (Dudley's entropy integral). *Let $\{X_\theta : \theta \in T\}$ be a zero-mean stochastic process that is sub-Gaussian w.r.t. pseudometric space (T, d). Let D be the diameter of T. Then for any ϵ,*

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

If, moreover, $\{X_\theta\}$ is seperable, then

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

Remark 2.2. If $\log N(\epsilon) = C\epsilon^{-r}$, then the integral exists if $r < 2$, and does not exist otherwise. In the latter case, we could use the first bound.

Return to Appendix A.

Theorem 2.8 (Dudley bound on Rademacher complexity)**.** *If $\mathcal{F}$ is class of functions from $\mathcal{Z}$ to $\mathbb{R}$ that satisfies $\mathcal{F} = -\mathcal{F}$, then*

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

In the first display, P_n is the empirical distribution of the sample Z_1^n (to which the expectation corresponds), and in the second display, the supremum on Q is taken over all discrete probability measures.

Theorem 2.9 (Bracketing integral bound)**.** *There is a universal constant $C > 0$ so that, for any class of functions $\mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}$ with envelope function F ($f(z) \leq F(z)$ for all $f \in \mathcal{F}$, $z \in \mathcal{Z}$):*

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

where $||F|| = \sqrt{\int F^2(z)dz}$.

Return to Appendix A.

A Problem Solving Algorithms

A.1 Bounding the deviation of a random variable from its mean.

1. Apply the Chernoff bound (Theorem 1.2). Don't forget to multiply by 2!
2. Show that the variable is sub-Gaussian, and then apply Proposition 1.2.
3. If it is bounded, use Hoeffding's inequality (Theorem 1.3) or Bernstein's inequality (Theorem 1.4). Bernstein will be tighter if the variance is small and t is small.
4. Show that the variable is sub-exponential, and then apply Proposition 1.4. Don't forget to check both cases for the bound.
5. Show that the variable is a function of independent random variables, where the function satisfies the bounded differences property, and apply McDiarmid's inequality (Theorem 1.5).

A.2 Show that a random variable is sub-Gaussian.

1. By Definition 1.1.
2. By Proposition 1.1.
3. Show that it is the sum of sub-Gaussian variables.

A.3 Show that a random variable is sub-exponential.

1. By Definition 1.2.
2. By Proposition 1.3.

A.4 Bound the regret of an ERM (or show its convergence to zero).

1. Apply the uniform convergence bound (Proposition 2.1).
2. Apply Proposition 2.1 to bound by $||P_n - P||_{\mathcal{F}}$, McDiarmid's to bound its deviation from its mean (Theorem 1.5), then bound $\mathbb{E}\left[||P_n - P||_{\mathcal{F}}\right]$ by the Rademacher complexity (Proposition 2.4). Then, bound the Rademacher complexity.
3. Apply Proposition 2.1 to bound by $||P_n - P||_{\mathcal{F}}$, McDiarmid's to bound its deviation from its mean (Theorem 1.5), then bound $\mathbb{E}\left[||P_n - P||_{\mathcal{F}}\right]$ by the bracketing integral bound (Theorem 2.9).
4. Apply the bracket number G-C theorem (Theorem 2.2).
5. Apply the covering number G-C theorem (Theorem 2.6).

A.5 Compute or bound the Rademacher complexity of function class.

1. By Definition 2.1.
2. By the finite class lemma (Lemma 2.1).
3. Use Dudley's bound (Theorem 2.8). Change the upper bound from ∞ to the highest point after which the covering number is 1.
4. Use the one-step discretization bound (Proposition 2.10).

A.6 Compute or bound the growth function of a function class.

1. By Definition 2.3.
2. Bound using Proposition 2.6.
3. If $\mathcal{F}$ is a VC class, then use Sauer's lemma (Lemma 2.2).

A.7 Compute or bound the VC-dimension of a function class.

1. By Definition 2.4. That is, show
   - For some n, propose a set of points $x_1, ..., x_n$ such that for all $y_1, ..., y_n \in \{0, 1\}$, there is an $f \in \mathcal{F}$ such that $y_i = f(x_i)$ for all i.
   - Prove that for any set of $n+1$ points $x_1, ..., x_{n+1}$, there exists a labeling $y_1, ..., y_{n+1}$ such that for any $f \in \mathcal{F}$, $y_i \neq f(x_i)$ for some i.
2. Bound using Theorem 2.1.
3. Bound the growth function and use Proposition 2.7.

A.8 Compute or bound the bracketing number of a function class.

1. By Definition 2.5.
2. Computing the $\frac{\epsilon}{2}$-covering number (for the $\sup$-norm) for an upper bound (Theorem 2.4).
3. Use the fact that if $||\cdot||$ and $||\cdot||'$ are two norms, $f, g \in \mathcal{F}$, and

   $$||f - g|| \leq \phi\left(||f - g||'\right)$$

   for some monotonically increasing function, then

   $$N_{[]}\left(\epsilon, \mathcal{G}, ||\cdot||\right) \leq N_{[]}\left(\phi^{-1}\left(\epsilon\right), \mathcal{G}, ||\cdot||'\right).$$

   Solve using $||\cdot||'$ instead.

A.9 Compute or bound the covering number/metric entropy of a function class.

1. By Definition 2.6.
2. Computing the ϵ-packing number and using Theorem 2.3 for an upper bound.
3. Computing the 2ϵ-packing number and using Theorem 2.3 for a lower bound.
4. Computing the 2ϵ-bracketing number for a lower bound (Theorem 2.4).
5. Notice that the set is a ball or a special type of Lipschitz class and apply Proposition 2.9.
6. Remember to pass the covering number bound for a set of loss functions to a covering number bound on the parameter space (as in Homework 5 Problem 1a).

A.10 Compute or bound the packing number of a function class.

1. By Definition 2.8.
2. Computing the $\frac{\epsilon}{2}$-covering number and using Theorem 2.3 for an upper bound.
3. Computing the ϵ-covering number and using Theorem 2.3 for a lower bound.

A.11 Show that a stochastic process is sub-Gaussian.

1. By Definition 2.10.

A.12 Show that a stochastic process is separable.

1. By Definition 2.11.

A.13 Bounding the supremum of a sub-Gaussian process.

1. If it is a supremum of differences, use the second finite class lemma (Lemma 2.3).
2. Use Dudley's entropy integral (Theorem 2.7). Especially if the covering number is known to satisfy $\log N(\epsilon) = C\epsilon^{-r}$ for $r < 2$.
3. Use the the "first pass" method (Theorem 2.5). Dudley's is preferred, however.

B Generalities

B.1 Taylor series approximations

1. **Taylor's theorem:** for any function f that is k-times differentiable function at a point a, there exists a function h_k such that

$$f(x) = f(a) + f^{(1)}(a)(x-a) + \frac{1}{2}f^{(2)}(a)(x-a)^2 + ... + \frac{1}{k!}f^{(k)}(a)(x-a)^k + h_k(x)(x-a)^k,$$

and

$$\lim_{x \to a} h_k(x) = 0.$$

Equivalently, $h_k(x)(x-a)^k = o\left((x-a)^k\right)$.

2. $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} - ....$

B.2 Identities and Inequalities

1. For any $x \in \mathbb{R}$,

$$x + 1 \leq e^x.$$

2. For any $\{a_t\}_t$, $\{b_t\}_t$,

$$\sup_t |a_t| - \sup_t |b_t| \leq \sup_t |a_t - b_t|.$$

3. For any f, if $r_1 < r_2$, then

$$||f||_{L^{r_1}(P)} \leq ||f||_{L^{r_2}(P)}.$$

4. For non-negative integers a, b,

$$\binom{a}{b} \leq \left(\frac{ae}{b}\right)^b.$$

5. For any functions $f, g : \mathcal{X} \to [-1, 1]$,

$$||f-g||_2^2 = P[f-g]^2 \leq 2P|f-g| = 2||f-g||_1.$$

6. For $a, b \in \mathbb{R}$,

$$2ab \leq \frac{a^2+b^2}{2}.$$

7. For any $x, y, z \in \mathbb{R}$,

$$(y-x)^2 - (y-z)^2 = (2y-x-z)(x-z)^2$$

B.3 Notions of convergence and stochastic order notation

The following definitions concern a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{B}\left(\mathbb{R}^d\right)$ be the Borel sets on $\mathbb{R}^d$. The sequence $(X_n)_n$ and X are $\mathcal{F}$-$\mathcal{B}\left(\mathbb{R}^d\right)$-measurable random variables on Ω.

Definition B.1 (Convergence almost surely). $(X_n)_n \to X$ *almost surely* if

$$\mathbb{P}\left[\{\omega \in \Omega : X_n(\omega) \overset{n\to\infty}{\to} X(\omega)\}\right] = 1.$$

Definition B.2 (Convergence in probability). $(X_n)_n \to X$ *in probability* if for any $\epsilon > 0$,

$$\mathbb{P}\left[||X_n - X||\epsilon\right] \overset{n\to\infty}{\to} 0.$$

Definition B.3 (Convergence in distribution). $(X_n)_n \to X$ *weakly* or *in distribution* if for every bounded, continuous function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}\left[f(X_n)\right] \to \mathbb{E}\left[f(X)\right] \text{ as } n \to \infty.$$

Notation: $X_n \implies X$.

Definition B.4 (Big-O and little-o notation). Let $(x_n)_n$ and $(r_n)_n$ be real-valued sequences, with $r_n \neq 0$ for large n.

1. **Big-O**: the following are equivalent.
   (a) $x_n = O(r_n)$.
   (b) $\limsup_{n\to\infty} \left|\frac{x_n}{r_n}\right| < \infty$.
   (c) There exists $M > 0$ such that $\mathbb{1}\left[|x_n| \leq M|r_n|\right] \to_n 1$.
2. **Little-o**: the following are equivalent.
   (a) $x_n = o(r_n)$.
   (b) $\limsup_{n\to\infty} \left|\frac{x_n}{r_n}\right| = 0$.
   (c) For any $M > 0$, $\mathbb{1}\left[|x_n| \leq M|r_n|\right] \to_n 1$.

Definition B.5 (Big-O and little-o in probability notation). Let $(X_n)_n$ and $(R_n)_n$ be sequences of $\mathbb{R}^d$-valued random variables on the same probability space.

1. **Big-O-P**: We say $X_n = O_P(R_n)$ if for any $\delta > 0$, there exists some $M = M_\delta > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left[||X_n|| \leq M\,||R_n||\right] > 1 - \delta.$$

2. **Little-o-P**: We say $X_n = o_P(R_n)$ if for any constant $M > 0$,

$$\lim_{n\to\infty} \mathbb{R}||X_n|| \leq M\,||R_n|| = 1.$$

Theorem B.1 (Continuous mapping)**.** *Let $g : \mathbb{R}^d \to \mathbb{R}^m$ be continuous at every point of a probability 1 set. The following hold.*

1. *If $X_n \implies X$, then $g(X_n) \implies X$.*
2. *If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} X$.*
3. *If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} X$.*

Theorem B.2 (Slutsky's)**.** *Let $X_n \implies X$, all realized in $\mathbb{R}^d$. Then,*

1. *If $Y_n \xrightarrow{p} c \in \mathbb{R}^d$, then $X_n + Y_n \implies X + c$.*
2. *If $Y_n \xrightarrow{p} c \in \mathbb{R}^d$, then $Y_n X_n \implies cX$.*
3. *If $Y_n \xrightarrow{p} c \in \mathbb{R}^d$ and $c \neq 0$, then $\frac{X_n}{Y_n} \implies \frac{X}{c}$.*