STAT 582 Final Exam Guide

March 2023

Contents

1 Chapter 3: Concentration Inequalities

GOAL: Bound $\mathbb{P}\{f(X_1, ..., X_n) \geq t\}$ for $t > 0$ in the case of a finite sample

1.1 Bounds on Moments

Theorem 1.1.1: Markov's Inequality

If $X \geq 0$ and $\mathbb{E}[X] < \infty$, then $\forall t > 0$,

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

Corollary 1.1.1. *Suppose* $\mathbb{E}[X] < \infty$*,* $h : [0, \infty) \mapsto [0, \infty)$ *is non-decreasing, and* $\mathbb{E}[h(|X - \mathbb{E}[X]|)] < \infty$*, then* $\forall t > 0$*,*

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

Theorem 1.1.2: Chebyshev's Inequality

If $\mathbb{E}[X] < \infty$ and $\mathbb{E}[X^2] < \infty$, then $\forall t > 0$,

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geq t\} \leq \frac{Var(X)}{t^2}$$

1.2 Bounds on MGF

Theorem 1.2.1: Chernoff Bound

Suppose for RV X, $\exists b > 0$ s.t., $\forall |\lambda| \leq b$, $\mathbb{E}[e^{\lambda x}] < \infty$.
Then $\forall t > 0$

$$\mathbb{P}\{X - \mathbb{E}[X] \geq t\} \leq inf_{\lambda > 0} \frac{M_{X-\mu}(\lambda)}{e^{\lambda t}}$$

or, equivalently,

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

1.2.1 Sub-Gaussian RVs

Definition 1.2.2: Sub-Gaussian

An RV is **Sub-G** with parameter σ^2 iff, $\forall \lambda \in \mathbb{R}$

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

or, equivalently, $\exists c > 0, s > 0$ s.t. $\forall t > 0$

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

where $Z \sim N(0, 1)$

- *Result from Chernoff Bound*: If X is sub-G with parameter σ^2, then

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

Theorem 1.2.3: Hoeffding Theorem

If $X_1, ..., X_n$ are independent RVs with support in $[a, b]$, then $\bar{X}_n$ is sub-G with parameter $\sigma^2 = \frac{(b-a)^2}{4n}$.

- *Result from Chernoff Bound*: For $X_1, ..., X_n$ independent RVs with support in $[a, b]$,

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

1.2.2 Sub-Exponential RVs

Definition 1.2.4: Sub-Exponential

An RV X is **sub-E** with parameters (σ^2, b) if $\forall|\lambda| < 1/b$,

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

or, equivalently, $\exists c > 0, \ell > 0$ s.t. $\forall t > 0$

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geq t\} \leq c\mathbb{P}\{|\epsilon_\ell| \geq t\}(= ce^{-\ell t})$$

where $\epsilon_\ell \sim Exp(\ell)$

- *Result from Chernoff Bound*: If X is sub-E with parameters (σ^2, b), then $\forall t > 0$

$$log\mathbb{P}\{\bar{X}_n - \mathbb{E}[\bar{X}_n] \geq t\} \leq \begin{cases} \frac{-t^2}{2\sigma^2} & 0 \leq t \leq \sigma^2/b \\ \frac{-t}{2b} & t > \sigma^2/b \end{cases}$$

Theorem 1.2.5: Bernstein Theorem

If X is a bounded RV with variance σ^2 s.t. $|X - \mu| \leq b$ a.s., then X is sub-E with parameters (σ^2, b).

- *Result from Chernoff Bound*: If $X_1, ..., X_n$ are independent RVs with variances σ_i^2, then $\forall t > 0$

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

1.3 Bounded Differences Inequality

Definition 1.3.1: Bounded Differences Property

A function, f, satisfies the **Bounded Differences Property** if $\forall i\ \exists c_i < \infty$ s.t $\forall x_1, ..., x_n, \tilde{x}_i$

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

Proposition 1.3.2. *Let* $f(X_1, ..., X_n) := \sup_{g \in \mathscr{G}} |\frac{1}{n}\sum_i^n(g(X_i) - \mathbb{E}[g(X_i)])|$*, where* $\sup_{g \in \mathscr{G}} \sup_X |g(X)| \leq 1$*.*

Then f *is BDP with* $c_i = 2/n$ *for all* i*.*

Theorem 1.3.3: Bounded Differences Inequality

If $X = (X_1, ..., X_n)$ is a collection of independent RVs and f satisfies the BDP with constraints $c_1, ..., c_n$, then $\forall t > 0$

$$\mathbb{P}\{|f(X) - \mathbb{E}[f(X)]| \geq t\} \leq 2exp\{-\frac{2t^2}{\sum_i^n c_i^2}\}$$

Notes on proof: Makes use of Azuma-Hoeffding Lemma, which is proven by making repeated use of Hoeffding Theorem to show $f(X) - \mathbb{E}[f(X)]$ is sub-G with $\sigma^2 = \sum_i^n c_i^2/4$

2 Chapter 4: Bounding Regret of ERM

GOAL: Find $\hat{\theta} \in \Theta$ s.t. the risk, $\int \ell(X, \hat{\theta}) dP(X) =: P\ell(\cdot, \hat{\theta})$, approximates $inf_{\theta \in \Theta} P\ell(\cdot, \theta)$,
where $X_1, .., X_n \sim P$ and $\ell : X \mapsto \mathbb{R}$ is some loss function.

Because we do not know P, we estimate it with an empirical distribution, P_n, and so our empirical risk minimizer, $\hat{\theta} \in \Theta$, is that which minimizes $P_n \ell(\theta)$

2.1 Bounding the regret of an ERM

We can quantify how close we are to achieving this goal as, $Reg(\theta) := P\ell(\hat{\theta}) - inf_{\theta \in \Theta} P\ell(\theta)$.
Observe,

$$
\begin{aligned}
&0 \leq Reg(\hat{\theta}) = P\ell(\hat{\theta}) - P\ell(\theta_0) \\
&= (P_n - P)[\ell(\theta_0) - \ell(\hat{\theta})] \\
&\leq |(Pn - P)\ell(\theta_0)| - |(Pn - P)\ell(\hat{\theta})| \\
&\leq 2 \sup_{\theta \in \Theta} |(Pn - P)\ell(\theta)| \\
&= 2 \sup_{f \in \mathscr{F}} |(P_n - P)|_{\mathscr{F}} =: 2||P_n - P||_{\mathscr{F}}, \text{ where } \mathscr{F} = \{\ell(\theta) : \theta \in \Theta\}
\end{aligned}
$$

So we can upper bound $Reg(\hat{\theta})$ by upper bounding $2||P_n - P||_{\mathscr{F}}$

Proposition 2.1.1. *If $\mathscr{F}$ consists of $[0,1]-$ valued functions, then $||P_n - P||_{\mathscr{F}}$ satisfies the BDP with $c_i = 1/n$ for each i, and, by the Bounded Differences Inequality,*

$$\mathbb{P}\{|||P_n - P||_{\mathscr{F}} - \mathbb{E}||P_n - P||_{\mathscr{F}}| \geq t\} \leq 2exp\{-2nt^2\}$$

Theorem 2.1.2

If $\mathscr{F}$ consists of $[0,1]-$ valued functions, then with probability at least $1 - 2e^{-2nt^2}$, it holds that for $t > 0$,

$$
\begin{aligned}
\frac{1}{2}\mathbb{E}||R_n||_{\mathscr{F}} - \sqrt{\frac{log2}{n}} - t &\leq \mathbb{E}||P_n - P||_{\mathscr{F}} - t \\
&\leq ||P_n - P||_{\mathscr{F}} \\
&\leq \mathbb{E}||P_n - P||_{\mathscr{F}} + t \\
&\leq 2\mathbb{E}||R_n||_{\mathscr{F}} + t
\end{aligned}
$$

where the Rademacher complexity, $||R_n||_{\mathscr{F}} := \sup_{f \in \mathscr{F}} |R_n(f)|$, $R_n(f) := \frac{1}{n}\sum_i^n \epsilon_i f(x_i)$, and ϵ_i are Rademacher RVs.

Notes on Proof: Use of symmetrization and de-symmetrization arguments with a ghost sample, and Prop 2.1.1.

Remark 2.1.1. Properties of Rademacher Complexity (HW 3.2): *Let $\mathscr{F}$ and $\mathscr{G}$ denote collections of $X \mapsto R$ functions and f_0 denote a fixed and uniformly bounded function. Then,*

- $\mathbb{E}||R_n||_{\mathscr{F}+\mathscr{G}} \leq \mathbb{E}||R_n||_{\mathscr{F}} + \mathbb{E}||R_n||_{\mathscr{G}}$, *where* $\mathscr{F} + \mathscr{G} := \{f(x) + g(x) : f \in \mathscr{F}, g \in \mathscr{G}\}$.
- $\mathbb{E}||R_n||_{\mathscr{F}+f_0} \leq \mathbb{E}||R_n||_{\mathscr{F}} + \frac{||f_0||_\infty}{\sqrt{n}}$, *where* $\mathscr{F} + f_0 := \{f(x) + f_0(x) : f \in \mathscr{F}\}$.

Remark 2.1.2. Relating Rademacher complexity to regret: *Observe*

$$
\begin{aligned}
P\{Reg(\hat{\theta}_n) > K_n t\} &\leq P\{2||Pn - P||_{\mathscr{F}} > K_n t\} \\
&\leq \frac{2\mathbb{E}||P_n - P||_{\mathscr{F}}}{K_n t} \leq \frac{4\mathbb{E}||R_n||_{\mathscr{F}}}{K_n t}
\end{aligned}
$$

So if K_n converges to 0 faster than $\mathbb{E}||R_n||_{\mathscr{F}}$, then for all $\epsilon > 0$ and sufficiently large n, there exists some t such that $P\{Reg(\hat{\theta}_n) > K_n t\} \leq \epsilon$, so $Reg(\hat{\theta}_n) = O_p(K_n)$.

2.2 VC dimension

Definition 2.2.1: VC dimension

Let $\mathscr{F}$ be a class of functions mapping from $X \mapsto \{0,1\}$ and define the projection of $\mathscr{F}$ onto $x_1^n := (x_1, ..., x_n) \in X^n$ as $\mathscr{F}_{x_1^n} := \{(f(x_1), ..., f(x_n)) : f \in \mathscr{F}\}$

- Then we say $\mathscr{F}$ **shatters** x_1^n if $|\mathscr{F}_{x_1^n}| = 2^n$
- The **growth function**, of $\mathscr{F}$, $\Pi_{\mathscr{F}}(n) := \sup_{x_1^n} |\mathscr{F}_{x_1^n}|$
- The **VC dimension** of $\mathscr{F}$ is defined as $VC(\mathscr{F}) := \sup\{n \in N : \Pi_{\mathscr{F}}(n) = 2^n\}$
  - i.e. the largest n s.t $\mathscr{F}$ shatters x_1^n
- The **VC index** of $\mathscr{F}$ is defined as $VC(\mathscr{F}) := inf\{n \in N : \Pi_{\mathscr{F}}(n) < 2^n\}$
  - i.e. the smallest n s.t $\mathscr{F}$ does not shatter x_1^n

Remark 2.2.1. *If A is a collection of subsets of X then $VC(A) = VC(F_A)$ where $F_A := \{x \mapsto \mathbb{I}_B(x) : B \in A\}$*

Remark 2.2.2. *If $\mathscr{F}$ consists of mappings from $X \mapsto \mathbb{R}$, $VC(\mathscr{F})$ is equal to the VC dimension of the collection of subgraphs, $A := \{\{(x,t) \in X \times \mathbb{R} : t < f(x)\} : f \in \mathscr{F}\}$*

Theorem 2.2.2

Consider a family of boolean-valued functions, $\mathscr{F} = \{x \mapsto f(x,\theta) : \theta \in \mathbb{R}^p\}$, where each $f : \mathbb{R}^m \times \mathbb{R}^p \mapsto \{0,1\}$ and f can be computed using no more than t arithmetic or comparison operations. Then, $VC(\mathscr{F}) \leq 4p(t+2)$.

Example 2.2.1. *Let $A := \{(-\infty, b) : b \in \mathbb{R}\}$. Then the $VC(A) = 1$.*

Example 2.2.2. *Let $B := \{(a, b] : a, b \in \mathbb{R}\}$. Then the $VC(B) = 2$.*

Example 2.2.3. *Let $C = \{(-\infty, t_1] \times (-\infty, t_2] : (t_1, t_2) \in \mathbb{R}^2\}$. Then the $VC(C) = 2$.*

Example 2.2.4. *Let D be the collection of monotone increasing functions $f : \mathbb{R} \mapsto \mathbb{R}$. Then $VC(D) = \infty$.*

Example 2.2.5. *Let E be the collection of spheres in $\mathbb{R}^2$ with radius b and center (a_1, a_2). Then $VC(E) = 3$.*

Example 2.2.6. *Let $F = \{x \mapsto \mathbb{I}(x \in A) : A \subset \mathbb{R}^2$ and A is convex $\}$. Then $VC(F) = \infty$.*

Example 2.2.7. Permanence of the VC Property (HW 4.1)*: Let $\mathscr{F}$ be a VC class of functions and g be some fixed function. Then the following classes are also VC:*

- *$\{x : f(x) > 0\}$ as f ranges over $\mathscr{F}$*
- *$\{x \mapsto f(x) + g(x)\}$ as f ranges over $\mathscr{F}$*
- *$\{x \mapsto f(x)g(x)\}$ as f ranges over $\mathscr{F}$*

2.2.1 Bounding the Rademacher Complexity with VC dimension

Lemma 2.2.3: Finite Class Lemma

If $\mathscr{F}$ is a class of functions mapping to $[-1,1]$, then

$$\mathbb{E}||R_n||_{\mathscr{F}} \leq \sqrt{\frac{2log(2\mathbb{E}[|\mathscr{F}_{x_1^n}|])}{n}}$$

Corollary 2.2.1. *If $\mathscr{F}$ is a collection of boolean-valued functions then, $\mathbb{E}||R_n||_{\mathscr{F}} \leq \sqrt{\frac{2log(2\Pi_{\mathscr{F}}(n))}{n}}$*

Notes on proof: Derive analogous result for $\mathbb{E}[||R_n||_{\mathscr{F}}|x_1^n]$, use sub-Gaussianity of $\sum \epsilon_i z_i$, and take expectation of both sides.

Lemma 2.2.4: Sauer's Lemma

Let $d \geq VC(\mathscr{F})$ and $\mathscr{F}$ be a collection of boolean-valued functions. Then $\Pi_{\mathscr{F}}(n) \leq \sum_{k=0}^{d} \binom{N}{k}$, so it follows that

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

Corollary 2.2.2. *If $n > VC(\mathscr{F})$ and $\mathscr{F}$ is a collection of boolean-valued functions, then*

$$\mathbb{E}||R_n||_{\mathscr{F}} = O(\sqrt{\frac{logn}{n}})$$

2.3 Bracketing Numbers

Definition 2.3.1: $L^r(P)$ space

The $L^r(P)$ **space** is the space of function $f : X \mapsto \mathbb{R}$ s.t. $||f||_{L^r(P)} := [\int |f(x)|^r dP(x)]^{1/r} < \infty$
Also, $||f||_{L^\infty(P)} := \sup_{x \in X} |f(x)|$

Definition 2.3.2: Bracketing Numbers

Given 2 functions, $\ell : X \mapsto \mathbb{R}$ and $u : X \mapsto \mathbb{R}$,

- The **bracket**, $[\ell, u] := \{f \in L^r(P) : \ell \leq f \leq u \text{ pointwise}\}$
- We call $[\ell, u]$ an ϵ-**bracket** if $||u - \ell||_{L^r(P)} \leq \epsilon$
- The ϵ-**bracketing number**, $N_{[]}(\epsilon, \mathscr{F}, L^r(P) := inf\{m : \mathscr{F} \subseteq \cup_{i=1}^m [\ell_j, u_j]$ for a collection of ϵ-brackets, $[\ell_j, u_j]\}$, i.e. the minimal number of ϵ brackets needed to cover $\mathscr{F}$.

Example 2.3.1. (VdV 19.6): *Let $\mathscr{F} := \{f_t(x) : t \in \mathbb{R}\}$ where $f_t(x) = \mathbb{I}(x \leq t)$. Then,*

$$N_{[]}(\epsilon, \mathscr{F}, L_2(P)) \leq N_{[]}(\epsilon, \mathscr{F}, L_1(P)) \leq 1/\epsilon$$

Example 2.3.2. (HW 4.3): *Let $\mathscr{F}$ be a class of functions, $f : [0, 1] \mapsto [0, 1]$ s.t. $|f(x) - f(y)| \leq |x - y|$. Then $logN_{[]}(\epsilon, \mathscr{F}, L^2(P)) \leq C/\epsilon$*

Example 2.3.3. (Lipschitz parameterized function class): *Let $\mathscr{F} := \{f_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ where Θ is bounded and $|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m(x)||\theta_1 - \theta_2||$. Then,*

$$logN_{[]}(2\epsilon||m||_{L_2}, \mathscr{F}, L^2(P)) \leq dlog(diam(\Theta)/\epsilon)$$

2.3.1 Bounding $||P_n - P||_{\mathscr{F}}$ with bracketing numbers

Theorem 2.3.3: Glivenko-Cantelli

If $\mathscr{F}$ is a collection of functions s.t. $N_{[]}(\epsilon, \mathscr{F}, L^1(P) < \infty$ for all ϵ, then $\mathscr{F}$ is Glivenko-Cantelli, that is, $||P_n - P||_{\mathscr{F}} = o_p(1)$.

2.4 Covering and Packing Numbers

Definition 2.4.1: Covering Numbers

Let (S, d) denote a pseudometric space and $T \subseteq S$

- A set $T_1 \subseteq T$ is called an **ϵ-cover** of T if for each $\theta \in T$, $\exists \theta_1 \in T_1$ s.t. $d(\theta, \theta_1) \leq \epsilon$
- The **ϵ-covering number** for T, $N(\epsilon, T, d)$, is defined as the size of a minimal ϵ-cover for T
- The log covering number, $logN(\epsilon)$, is known as the **metric entropy** of T

Example 2.4.1. (Supremum norm on grid): $N(\epsilon, [0,1]^2, ||\cdot||_\infty) = O(\frac{1}{\epsilon^2})$

Example 2.4.2. (Lipschitz functions): *Let $\mathscr{F}$ denote a collection of functions mapping $[0,1] \mapsto [0,1]$ that are L-Lipschitz, i.e. $|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$ $\forall x_1, x_2 \in [0,1]$. Then,*

$$logN(\epsilon, \mathscr{F}, ||\cdot||_\infty) = O(\frac{L}{\epsilon})$$

.

Example 2.4.3. (Lipschitz functions with support in $[0,1]^d$): *Let $\mathscr{F}$ denote a collection of functions mapping $[0,1]^d \mapsto [0,1]$ that are L-Lipschitz, i.e. $||f(x_1) - f(x_2)|| \leq L||x_1 - x_2||_\infty$ $\forall x_1, x_2 \in [0,1]^d$. Then,*

$$logN(\epsilon, \mathscr{F}, ||\cdot||_\infty) = O((\frac{L}{\epsilon})^d)$$

.

Example 2.4.4. (Functions that are Lipschitz in indexing parameter): *Let $f : X \times B \mapsto \mathbb{R}$ be some function and $\mathscr{F} := \{x \mapsto f(x, \beta) : \beta \in B\}$. Suppose there exists $L > 0$ s.t. $\forall \beta_1, \beta_2 \in B$ $||f(\cdot, \beta_1) - f(\cdot, \beta_2)||_{\mathscr{F}} \leq L||\beta_1 - \beta_2||_B$. Then,*

$$N(\epsilon, \mathscr{F}, ||\cdot||_{\mathscr{F}}) \leq N(\epsilon/L, B, ||\cdot||_B)$$

where $||\cdot||_{\mathscr{F}} = ||g_1 - g_2||_{\mathscr{F}}$ and $||\cdot||_B = ||g_1 - g_2||_B$.

Example 2.4.5. (Covering numbers of VC class functions): *Let $\mathscr{F}$ denote a collection of functions mapping from $Z \mapsto [-1,1]$ that are VC. Then*

$$sup_Q N(\epsilon, \mathscr{F}, L^2(Q) \leq K \cdot VC(\mathscr{F})(16e)^{VC(\mathscr{F})} \frac{1}{\epsilon}^{2[VC(\mathscr{F})-1]}$$

Definition 2.4.2: Packing Numbers

Let (S, d) denote a pseudometric space and $T \subseteq S$

- A set $T_1 \subseteq T$ is called an **ϵ-packing** of T is for each $\theta_1, \theta_1' \in T_1$, $d(\theta_1, \theta_1') > \epsilon$
- The **ϵ-packing number** of T, $M(\epsilon, T, d)$, is defined as the size of the maximal ϵ-packing of T

Example 2.4.6. (Ball in $\mathbb{R}^d$): *Let $B(0, r)$ denote a ball of radius r in $\mathbb{R}^d$. Let $\{x_j : 1 \leq j \leq n\}$ and $\{y_j : 1 \leq j \leq m\}$ be an ϵ-covering and ϵ-packing, respectively. Then using,*

$$Vol(B(0,r) \leq Vol(\cup_i^n B(x_j, r))$$

we obtain $N(\epsilon, B(0,r), ||\cdot||_{L^p}) \geq (\frac{r}{\epsilon})^d$, and using

$$Vol(\cup_j^m B(y_j, \epsilon/2) \leq Vol(B(0, r + \epsilon/2))$$

we obtain $M(\epsilon, B(0,r), ||\cdot||_{L^p}) \leq (\frac{2r}{\epsilon} + 1)^d \leq (\frac{3r}{\epsilon})^d$

Theorem 2.4.3: Relation between and covering and packing numbers

$\forall \epsilon > 0$, $M(2\epsilon) \leq N(\epsilon) \leq M(\epsilon)$

Theorem 2.4.4: Relation between bracketing and covering numbers

Let $\mathscr{F} \subseteq L^r(P)$ for any $r \in \mathbb{N}$. Then $\forall \epsilon > 0$,

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

2.5 Sub-Gaussian Processes

Definition 2.5.1

- A **stochastic process** $\{X_\theta : \theta \in T\}$ is a collection of RVs
- A stochastic process is **zero mean** if $\mathbb{E}[X_\theta] = 0 \ \forall \theta \in T$
- A mean zero stochastic process is called **sub-Gaussian** wrt to a pseudometric d on T if, $\forall \theta, \theta' \in T$ and $\forall \lambda \in \mathbb{R}$,

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

  or, equivalently, $X_\theta - X_{\theta'}$ is sub-G with parameter $\sigma^2 = d(\theta, \theta')^2$

Definition 2.5.2: Canonical Rademacher Process

Let $S = \mathbb{R}^n$ and d denote the Euclidean metric on S. Let $T \subseteq S$ denote the index set and $r_1, ..., r_n$ denote iid Rademacher RVs. Then the **Canonical Rademacher Process**, $\{X_\theta : \theta \in T\}$ is defined s.t.

$$X_\theta = \sum_i^n \theta_i r_i = \langle \theta, r \rangle$$

Remark 2.5.1. *The canonical Rademacher Process is mean zero and sub-Gaussian w.r.t the Euclidean metric.*

2.5.1 Bounding Sub-Gaussian Processes

Lemma 2.5.3: Special case of FCL

If $\{X_\theta : \theta \in T\}$ is sub-G w.r.t. d and $A \subseteq T \times T$ then,

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

Theorem 2.5.4: One-step discretization bound

Let $\{X_\theta : \theta \in T\}$ denote a sub-Gaussian process w.r.t. d and let $D := \sup_{\theta,\theta' \in T} d(\theta, \theta')$ denote the diameter of T. Then for any $\epsilon > 0$,

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

Theorem 2.5.5: Dudley's entropy integral bound

Let $\{X_\theta : \theta \in T\}$ denote a sub-Gaussian process w.r.t. d and D denote the diameter of T. Then for any $\epsilon > 0$,

$$\mathbb{E}\sup_{\theta\in T} X_\theta \leq \mathbb{E}[\sup_{\theta,\theta'\in T: d(\theta,\theta')<\epsilon}(X_\theta - X_{\theta'})] + 8\int_{\epsilon/2}^{D}\sqrt{logN(\tilde{\epsilon}, T, d)}d\tilde{\epsilon}$$

and if $\{X_\theta : \theta \in T\}$ is a canonical Rademacher process, then

$$\mathbb{E}\sup_{\theta\in T} X_\theta \leq 8\int_{0}^{D}\sqrt{logN(\tilde{\epsilon}, T, d)}d\tilde{\epsilon}$$

2.5.2 Bounding Rademacher complexity via bounding of a sub-G process

Theorem 2.5.6: Bounding Rademacher complexity via one-step discretization bound

For a class of functions $\mathscr{F}$, it follows from the One-step Discretization bound that $\forall\delta > 0$,

$$\mathbb{E}||R_n||_{\mathscr{F}} \leq 2\delta + 2\mathbb{E}[D_{Z_1^n}]n^{-1}\sup_Q\sqrt{log2N(\delta, \mathscr{F}, L^2(Q))}$$

where $D_{Z_1^n}n^{-1/2} = \sup_{f_1,f_2\in\mathscr{F}}\sqrt{\frac{1}{n}\sum_i^n|f_1(x_i) - f_2(x_i)|^2}$ and the supremum is over all finitely supported probability measures, Q, whose support is contained in that of P.

Notes on proof: First, show from definition that $\mathbb{E}[||R_n||_{\mathscr{F}}|Z_1^n] = \frac{1}{n}\mathbb{E}[\sup_{\theta\in T} X_\theta]$ where $T = \mathscr{F} \cup -\mathscr{F}$. Then use one-step discretization bound and take expectation on both sides to show $\mathbb{E}[||R_n||_{\mathscr{F}}] \leq \frac{2\epsilon}{\sqrt{n}} + 2D\frac{1}{n}\sqrt{logN(\epsilon, T, ||\cdot||_2)}$. Finally, use the covering number of T to derive a covering number for $\mathscr{F}$.

Remark 2.5.2. *The one-step discretization bound diverges as δ approaches 0, because the covering number diverges, however it always finite for fixed δ.*

Theorem 2.5.7: Bounding Rademacher complexity via Dudley's entropy integral bound

If $\mathscr{F}$ is a class of functions mapping from $\mathscr{L}$ to $\mathbb{R}$ s.t. $f \in \mathscr{F} \iff -f \in \mathscr{F}$, then it follows from Dudley's entropy integral bound that

$$\mathbb{E}||R_n||_{\mathscr{F}} \leq \frac{8}{\sqrt{n}}\mathbb{E}[\int_0^\infty\sqrt{logN(\epsilon, \mathscr{F}, L^2(P_n))}d\epsilon]$$

$$\leq \frac{8}{\sqrt{n}}sup_Q\int_0^\infty\sqrt{logN(\epsilon, \mathscr{F}, L^2(Q))}d\epsilon$$

where the supremum is over all finitely supported probability measures, Q, whose support is contained in that of P.

Remark 2.5.3. *This bound is trivial in the case that the function class is so large that the integral diverges.*

Example 2.5.1. (Lipschitz function with support in $[0,1]$)*: Let $\mathscr{F}$ denote a class of L-Lipschitz functions mapping $[0,1] \mapsto [0,1]$. Then it follows from Example 2.4.2. and Dudley's entropy integral bound that $\mathbb{E}||R_n||_{\mathscr{F}} = O(\frac{1}{\sqrt{n}})$*

Remark 2.5.4. (Lipschitz function with support in $[0,1]^d$)*: When $\mathscr{F}$ denotes a class of L-Lipschitz functions mapping $[0,1]^d \mapsto [0,1]$, the supremum result of Dudley's entropy integral bound produces a trivial bound, so the expectation result much be used, which yields a rate slower than $n^{-1/2}$.*

Example 2.5.2. (Lipschitz parameterised functions)*: Let $\mathscr{F} := \{g_\beta : \beta \in \mathbb{R}^p, ||\beta||_2 \leq 1\}$ where $sup_X|g_{\beta_1}(x) - g_{\beta_2}(x)| \leq L||\beta_1 - \beta_2||_2$. Then we can bound $sup_Q logN(\epsilon, \mathscr{F}, L^2(Q)) \leq plog(\frac{2L}{\epsilon} + 1)$ using examples 2.4.4 and 2.4.6, and then it follows from Dudley's entropy integral bound that $\mathbb{E}||R_n||_{\mathscr{F}} = O(L\sqrt{\frac{p}{n}})$*

3 Appendix

Jensen's Inequality: For RV X and convex function f, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

Cauchy-Schwartz Inequality: $(\int P_1(w)P_2(w))^2 \leq (\int P_1^2(w)dw)(\int P_2^2(w)dw)$

Layer Cake representation: If Z is a non-negative RV, then $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}(Z \geq t)dt$

3.1 Summation and Limit Properties

- $\sum_{i=1}^n i = \frac{n(n+1)}{2}$
- $\sum_{k=0}^\infty r^k = \frac{1}{1-r}$ if $|r| < 1$
- $\sum_{k=0}^n r^k = \frac{1-r^{n+1}}{1-r}$ if $|r| < 1$
- $\sum_{n=0}^\infty \frac{1}{n!} = e$
- $\sum_{n=0}^\infty \frac{x^n}{n!} = e^x$
- $lim_{n\to\infty}(1 + 1/n)^n = e$
- $lim_{n\to\infty}(1 - 1/n)^n = 1/e$

3.2 Common Distributions

Beta

- PDF: $f_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$
- Support: $x \in [0, 1]$
- Parameters: $\alpha, \beta > 0$
- Mean: $\alpha/(\alpha + \beta$
- Variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- Mode: $\frac{\alpha-1}{\alpha+\beta-2}$
- MGF: $M_x(t) = 1 + \Sigma_k^\infty (\Pi_r^{k-1} \frac{\alpha+r}{\alpha+\beta+r}) \frac{t^k}{k!}$
- Relationship: If $X \sim Beta(\theta, 1)$, then $-logX \sim Exp(\theta)$

Binomial

- PMF: $f_p(k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Parameters: $p \in [0, 1]$
- Mean: np
- Variance: $np(1-p)$
- MGF: $((1-p) + pe^t)^n$

Chi-Squared

- PDF: $f_k(x) = \frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}$
- Support: $x \in [0, \infty)$
- Parameters: $k \in \mathbb{N}$
- Mean: k

- Variance: $2k$
- MGF: $M_x(t) = (1-2t)^{-k/2}$ for $t < 1/2$

Dirichlet

- PDF: $f_{\alpha_1,...,\alpha_K}(x_1, ..., x_K) = \Gamma(\sum_i^K \alpha_i) \times \Pi_i^K x_i^{\alpha_i - 1} / \Pi_i^K \Gamma(\alpha_i)$
- Support: $x_1, .., x_k \in (0,1)$ where $\sum_i^K x_i = 1$
- Parameters: $\alpha_1, ..., \alpha_K > 0$
- Mean: $\alpha_i / \sum_k^K \alpha_k$
- Variance: $\tilde{\alpha_i}(1-\tilde{\alpha_i})/(\alpha_0 + 1)$ where $\tilde{\alpha_i} = \alpha_i/\alpha_0$ and $\alpha_0 = \sum_k^K \alpha_k$
- Mode: $(\alpha_i - 1)/(\sum_k^K \alpha_k - K)$

Exponential

- PDF: $f_\lambda(x) = \lambda e^{-\lambda x} \mathbb{I}(x \geq 0)$
- CDF: $F_\lambda(x) = 1 - e^{-\lambda x}$
- Support: $x \in [0, \infty)$
- Parameters: $\lambda \in (0, \infty)$
- Mean: $1/\lambda$
- Variance: $1/\lambda^2$
- Mode: 0
- MGF: $M_X(t) = \frac{\lambda}{\lambda - t}$, for $t < \lambda$
- Relationship: $Exp(\lambda) \iff Gamma(1, \lambda)$

Gamma (shape/rate)

- PDF: $f_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
- Support: $x \in (0, \infty)$
- Parameters: $\alpha, \beta > 0$
- Mean: α/β
- Variance: α/β^2
- Mode: $\frac{\alpha-1}{\beta}$ for $\alpha \geq 1$
- MGF: $M_x(t) = (1 - \frac{t}{\beta})^{-\alpha}$ for $t < \beta$
- Relationship: The sum of independent $X_i \sim Gamma(\alpha_i, \beta)$ is distributed $Gamma(\sum_i \alpha_i, \beta)$
- Relationship: If $U \sim Gamma(\alpha, \lambda)$ and $V \sim Gamma(\beta, \lambda)$ then $\frac{U}{U+V} \sim Beta(\alpha, \beta)x$
- Relationship: $X \sim Gamma(k/2, 1/2) \iff X \sim Chisq(k)$

Gamma (shape/scale)

- PDF: $f_{k,\theta}(x) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-\frac{x}{\theta}}$
- Support: $x \in (0, \infty)$
- Parameters: $k, \theta > 0$

- Mean: $k\theta$
- Variance: $k\theta^2$
- Mode: $(k-1)\theta$ for $k \geq 1$
- MGF: $M_x(t) = (1-\theta t)^{-k}$ for $t < 1/\theta$
- Relationship: $X \sim Gamma(k/2, 2) \iff X \sim Chisq(k)$

Inv-Gamma (inverse of shape/rate)

- PDF: $f_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$
- Support: $x \in (0, \infty)$
- Parameters: $\alpha, \beta > 0$
- Mean: $\frac{\beta}{\alpha-1}$
- Variance: $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$
- Mode: $\frac{\beta}{\alpha+1}$
- Relationship: If $X \sim Inv-Gamma(\alpha, \beta)$, then $1/X \sim Gamma(\alpha, \beta)$

Normal

- PDF: $f_{\mu,\sigma^2}(x) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
- Support: $x \in \mathbb{R}$
- Parameters: $\mu \in \mathbb{R}, \sigma^2 > 0$
- Mean: μ
- Variance: σ^2
- MGF: $M_x(t) = e^{\mu t + \sigma^2 t^2/2}$
- Relationship: Sum of n iid standard normal variables is distributed χ^2_n

Multivariate Normal

- PDF: $f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{X}) = det(2\pi\boldsymbol{\Sigma})^{-1/2} exp\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\}$
- Support: $\mathbf{X} \in \mathbb{R}^k$
- Parameters: $\boldsymbol{\mu}, \boldsymbol{\Sigma}$
- Mean: $\boldsymbol{\mu}$
- Variance: $\boldsymbol{\Sigma}$
- MGF: $M_{\mathbf{X}}(\mathbf{t}) = e^{\boldsymbol{\mu}^T \mathbf{t} + \frac{1}{2}\mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}}$

Multinomial

- PMF: $f_{p_1,...,p_k}(x_1, ..., x_k) = \frac{n!}{x_1!...x_k!} p_1^{x_1} ... p_k^{x_k}$
- Support: $x_i \in \{0, ..., n\}$ with $\sum_i x_i = n$
- Parameters: $p_1, ..., p_k > 0; \sum_i^k p_i = 1$
- Mean: np_i
- Variance: $np_i(1-p_i)$

- MGF: $M_{x_1,...,x_k}(t_1,...,t_n) = (\sum_i^k p_i e^{t_i})^n$

Poisson

- PMF: $f_\lambda(k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- Support: $k \in \mathbb{N}$
- Parameters: $\lambda \in (0, \infty)$
- Mean: λ
- Variance: λ
- MGF: $M_x(t) = e^{\lambda(e^t - 1)}$
- Relationship: Sum of independent $X_i \sim Pois(\lambda_i)$ are distributed $Pois(\sum_i \lambda_i)$

Uniform

- PDF: $f_{a,b}(x) = \frac{1}{(b-a)}\mathbb{I}(x \in [a,b])$
- CDF: $F_{a,b}(x) = \frac{x-a}{b-a}$
- Support: $x \in [a,b]$
- Parameters: $a < b \in \mathbb{R}$
- Mean: $\frac{1}{2}(a+b)$
- Variance: $\frac{1}{12}(b-a)^2$
- MGF:

$$M_x(t) = \begin{cases} \dfrac{e^{tb} - e^{ta}}{t(b-a)} & : t \neq 0 \\ 1 & : t = 0 \end{cases}$$

Order Statistics

- $f_{X_{(1)}}(x) = n(1 - F_X(x))^{n-1} f_X(x)$
- $f_{X_{(n)}}(x) = n[F_X(x)]^{n-1} f_X(x)$
- $f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!}[F_X(x)]^{i-1}[1 - F_X(x)]^{n-i} f_X(x)$
- $f_{X_{(i)},X_{(j)}}(x,y) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F_X(x)]^{i-1}[F_X(y) - F_X(x)]^{j-i-1}[1 - F_X(x)]^{n-j} f_X(x) f_X(y)$
- $f_{X_{(1)},..,X_{(n)}}(y_1,..,y_n) = n! f_X(y_1)...f_X(y_n)$