Chapter 3

Uniform Laws:

- Glivenko-Cantelli
- sample complexity of learning
- uniform law w\ Rademacher Complexity
- Upper Bounds on Rademacher Complexity:
  - Massart Lemma
  - Finite sample Glivenko-Cantelli
  - VC dimension
  - Dimension Independent generalization bounds
    - linear models
    - convex losses

Central Set-up:

Let $\mathcal{F}$ be a family of integrable functions on some probability space $(\mathcal{X}, \Sigma, \mathbb{P})$.

Goal: Estimate

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathop{\mathbb{E}}_{X \sim \mathbb{P}} f(X) \right|$$

where $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}$.

Ex: (Emperical CDF)

Suppose we want to estimate the CDF

$$g(t) := \mathbb{P}[X \leq t].$$

We can use the empirical CDF

$$g_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, t]}(X_i)$$

Want to control

$$\|g_n - g\|_\infty := \sup_t |g_n(t) - g(t)|$$

This is exactly

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

where $\mathcal{F} = \{ \mathbb{1}_{(-\infty, t]}(\cdot) : t \in \mathbb{R} \}$.

Thm: (Glivenko-Cantelli) For any distribution

$$\| g_n - g \|_\infty \xrightarrow{a.s.} 0$$

[We'll prove a finite sample version soon]

Cor (Plug in Estimator) [HW]
Let $\gamma(\cdot)$ be a $\| \cdot \|_\infty$-continuous functional of CDF's. Then

$$\gamma(g_n) \to \gamma(g) \text{ in probability}$$

Ex: γ quantiles, median, mean
uniform distance to hypothesis' CDF.

Ex: (Statistical Learning)

Consider a family of probability measures

$$\{P_\theta : \theta \in \Omega\}$$

where Ω is some set.

Suppose we are given

$$X_1, \dots, X_n \overset{iid}{\sim} P_{\theta^*}$$

for some θ^* [may not be in Ω]

Fix a "goodness of fit" function

$$\theta \mapsto L_\theta(X)$$

Goal is to minimize <u>population risk</u>

$$\min_{\theta \in \Omega} R(\theta) := \mathbb{E}_{X \sim P_{\theta^*}}[L_\theta(X)]$$

One approach is to instead minimize the <u>Emperical Risk</u>:

$$\min_{\theta \in \Omega} R_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} L_\theta(X_i) \quad (\star)$$

Goal: Bound the excess risk

$$E(\hat{\theta}_n) := R(\hat{\theta}_n) - \inf_{\theta \in \Omega} R(\theta)$$

where $\hat{\theta}_n$ is the minimizer of (A)

Ex: (Maximum Likelyhood)

Let p_θ be the density of P_θ

Define

$$L_\theta(x) = \log\left(\frac{p_{\theta^*}(x)}{p_\theta(x)}\right)$$

Then

$$\hat{\theta}_n \in \operatorname{argmax}_{\theta \in \Omega} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log(p_\theta(X_i)) \right\}$$

Ex: (Binary Classification)

Data $(X_i, Y_i) \in \mathbb{R}^d \times \{-1, +1\}$ according to P.

We want to estimate $f: \mathbb{R}^d \to \{-1, +1\}$ solving

$$\min_{f \in T} \underbrace{P_{(X,Y)}[f(X) \neq Y]}_{R(f)} = \min_{f \in T} \underbrace{E_{(X,Y)\sim P}\, \mathbb{1}\{f(X) \neq Y\}}_{R(f)}$$

T: some function class

Try instead to solve

$$\min_{f \in T} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[f(X_i) \neq y_i]}_{R_n(f)}$$

Decomposition of Access Risk

$$E(\hat{\theta}_n, \theta^*) = R(\hat{\theta}_n) - \inf_{\theta \in \Omega} R(\theta)$$

$$= R(\hat{\theta}_n) - R_n(\hat{\theta}_n)$$

$$+ R_n(\hat{\theta}_n) - R_n(\theta_0)$$

$$+ R_n(\theta_0) - R(\theta_0)$$

where $\theta_0 = \operatorname{argmin}_{\theta \in \Omega} R(\theta)$

is approximation error. How to control?

is optimization error. If $\hat{\theta}_n$ is true minimizer of $R_n(\theta)$, then ≤ 0

easy to control approximation error:

$$\mathbb{E}_{\theta^*} R_n(\theta_0) = R(\theta_0)$$

$= R(\hat{\theta}_n) - R_n(\hat{\theta}_n)$

Main Difficulty: $\hat{\theta}_n$ is not fixed!

But we can estimate

$\leq \sup_{\theta \in \Omega} |R_n(\theta) - R(\theta)|$

$= \sup_{\theta \in \Omega} \left| \frac{1}{n} \sum_{i=1}^{n} L_\theta(X_i) - \mathbb{E}_{X \sim P_\theta} L_\theta(X) \right|$

So

$$\boxed{\begin{aligned} &E(\hat{\theta}_n, \theta^*) \leq 2 \| P_n - P \|_{\mathcal{F}} \\ &\text{where } \mathcal{F} := \{ x \mapsto L_\theta(x) : \theta \in \Omega \} \end{aligned}}$$

Remark:
- Function class $\mathcal{F}$ s.t. $\| P_n - P \|_{\mathcal{F}} \to 0$ in probability is called Glivenko-Cantelli
- If $\mathcal{F}$ is too big it may not be Glivenko-Cantelli (see Ex 4.7 in W)

Our first approach will be based on Rademacher Complexity.

Defn:

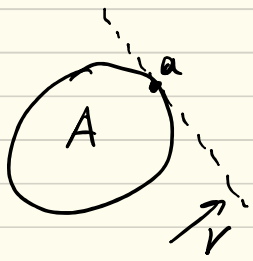
The Rademacher Complexity of a set $A \subseteq \mathbb{R}^n$, is the quantity

$$\mathcal{R}(A) := \mathbb{E}_\varepsilon \sup_{a \in A} \langle a, \varepsilon \rangle$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)$ are i.i.d. Rademacher random variables

Recall that support function of a set A is

$$\sigma_A(v) := \sup_{a \in A} \langle v, a \rangle$$



So

$$\mathcal{R}(A) = \mathbb{E}_\varepsilon \sigma_A(\varepsilon)$$

Defn: Consider a sequence of random (independent) $(x_1, \dots, x_n) \subseteq \mathcal{X}^n$ and a class of functions $\mathcal{F}$ on $\mathcal{X}$. The **Rademacher complexity** of $\mathcal{F}$ is

$$R_n(\mathcal{F}) := \mathbb{E}_{x,\epsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right|$$

$$= \mathbb{E}_x \, R(\mathcal{F}^I(x)/n)$$

$$\leq 2 \mathbb{E}_x \, R(\mathcal{F}(x)/n)$$

where $\mathcal{F}(x) = \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\}$

Thm: $\mathbb{E} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2 R_n(\mathcal{F})$

pf: Let $y_1, \dots, y_n \overset{iid}{\sim} \mathbb{P}$ and independent of $x_1, \dots, x_n$. Then

$$\mathbb{E} \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \mathbb{E}_x \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E}_{y_i}[f(y_i)] \right| \right]$$

$$= \mathbb{E}_x \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}_y \left[\frac{1}{n} \sum (f(x_i) - f(y_i)) \right] \right| \right]$$

$$= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}_Y \left[\frac{1}{n} \sum (f(x_i) - f(y_i)) \right] \right| \right]$$

$$\leq \mathbb{E}_X \mathbb{E}_Y \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum (f(x_i) - f(y_i)) \right|$$

$$= \mathbb{E}_{X,Y,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (f(x_i) - f(y_i)) \right|$$

$$\leq 2 \mathbb{E}_{X,\varepsilon} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(x_i) \right| = 2 \mathcal{R}_n(\mathcal{F})$$ ∎

Thm: Suppose $\mathcal{F}$ is b-uniformly bounded, meaning

$$\|f\|_\infty = \sup_{x \in X} |f(x)| \leq b \quad \forall f \in \mathcal{F}.$$

Then for any $n \in \mathbb{N}$, $t \geq 0$, it holds

$$\Pr\left(\|P_n - P\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + t \right) \geq 1 - e^{-\frac{nt^2}{2b^2}}$$

pf: All we have to do is show the bounded difference property for the function $\|P_n - P\|_{\mathcal{F}}$ with $L_i = \frac{2b}{n}$.

Define $\bar{f}(x) = f(x) - Ef(x)$.

Then $\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right|$

As before, let x' differ from x only in i'th entry. Then

$$\left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right| - \sup_{g \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{g}(x_i') \right|$$

$$\leq \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right| - \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i') \right| \leq \frac{2b}{n}$$

Take $\sup_{f \in \mathcal{F}}$ and exchange x, x' ☐

Rademacher complexity characterizes the assymptotics of $\|P_n - P\|_{\mathcal{F}}$.

Prop $\frac{1}{2}\mathcal{R}_n(\bar{\mathcal{F}}) \le \mathbb{E}_X\left[\|P_n - P\|_{\mathcal{F}}\right] \le 2\mathcal{R}_n(\mathcal{F})$

where $\bar{\mathcal{F}} = \{f - \mathbb{E}f : f \in \mathcal{F}\}$.

[See Prop 4.1.1 in Wainwright]

Prop (HW) Suppose $\mathcal{F}$ is b-bounded. Then $\forall n \in \mathbb{N}, t > 0$, it holds

$$\Pr\left(\|P_n - P\|_{\mathcal{F}} \ge \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}} |\mathbb{E}f(x)|}{2\sqrt{n}} - t\right)$$

$$\ge 1 - e^{-\frac{nt^2}{2b^2}}$$

Goal: Bound the Rademacher Complexity of interesting sets.

Lemma (Basic Calculus)
For any $A, B \subseteq \mathbb{R}^n$, $c \in \mathbb{R}$, $a_0 \in \mathbb{R}^n$ it holds

- $\mathcal{R}(cA) = |c| \mathcal{R}(A)$
- $\mathcal{R}(A+B) \leq \mathcal{R}(A) + \mathcal{R}(B)$
  with equality if A and B are convex
- $\mathcal{R}(A+a_0) = \mathcal{R}(A)$
- $\mathcal{R}(A) = \mathcal{R}(\text{conv}(A))$

Lemma: (Massart) Let $A = \{a_1, \dots, a_N\} \subseteq \mathbb{R}^n$.
Define $\bar{a} = \frac{1}{N} \sum_{i=1}^{N} a_i$. Then

$$\mathcal{R}(A) \leq \max_{a \in A} \|a - \bar{a}\| \cdot \sqrt{2 \log(N)}$$

pf: WLOG assume $\bar{a}=0$. Let $\lambda>0$ and $A'=\{\lambda a_1,\dots,\lambda a_N\}$. Then

$$R(A') = \mathbb{E}_\varepsilon \max_{a\in A'} \langle \varepsilon, a\rangle$$

$$= \mathbb{E}_\varepsilon \log\left(\max_{a\in A'} e^{\langle \varepsilon, a\rangle}\right)$$

$$\le \mathbb{E}_\varepsilon \log\left(\sum_{a\in A'} e^{\langle \varepsilon, a\rangle}\right)$$

Jensen

$$\le \log\left(\sum_{a\in A'} \mathbb{E}_\varepsilon e^{\langle \varepsilon, a\rangle}\right)$$

Observe

$$\mathbb{E}_\varepsilon e^{\langle \varepsilon, a\rangle} = \prod_{i=1}^{d} \mathbb{E} e^{\varepsilon_i a_i} \le \prod_{i=1} e^{\frac{a_i^2}{2}} = e^{\|a\|_2^2/2}$$

$$\le \log\left(\sum_{a\in A'} e^{\|a\|_2^2/2}\right)$$

$$\le \log\left(|A'| \max_{a\in A'} \left(e^{\|a\|_2^2/2}\right)\right)$$

$$= \log(|A'|) + \max_{a\in A'} \frac{\|a\|^2}{2}$$

Thus

$$\mathcal{R}(A) = \frac{1}{\lambda}\mathcal{R}(A') \le \frac{\log(|A|) + \max_{a \in A} \lambda^2 \|a\|^2 / 2}{\lambda}$$

Optimize over $\lambda > 0$. ∎

Defn: $\mathcal{F}$ has <u>polynomial discrimination of order $v \ge 1$</u> if $\forall n \in \mathbb{N}$ and $x_1, \dots, x_n \in \mathcal{X}$, we have

$$\text{card}\left(\mathcal{F}(x_1, \dots, x_n)\right) \le (n+1)^v.$$

Cor: Suppose $\mathcal{F}$ has polynomial discrimination of order v. Then

$$\mathcal{R}_n(\mathcal{F}) \le 4\, \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}\right] \cdot \sqrt{\frac{v \log(n+1)}{n}}$$

pf: Fix $x = (x_1, \dots, x_n)$. Then Massart implies

$$\mathcal{R}(\mathcal{F}(x)/n) \le 4 \cdot \frac{\max_{f \in \mathcal{F}} \sqrt{\sum_{i=1}^n f_i^2(x)}}{n} \cdot \sqrt{v \log(n+1)}$$

∎

In particular, if $\mathcal{F}$ is b-bounded, then

$$\boxed{R_n(\mathcal{F}) \le 4b\sqrt{\frac{v\log(n+1)}{n}}}$$

Cor: (Glivenko-Cantelli)

Let $g(t) = \mathbb{P}[X \le t]$ be the CDF of X, and let $g_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \le t]$ where $X_1, \dots, X_n \overset{i.i.d}{\sim} \mathbb{P}$. Then

$$\mathbb{P}\left[\|g_n - g\|_\infty \ge 8\sqrt{\frac{\log(n+1)}{n}} + \delta\right] \le \exp\left(-\frac{n\delta^2}{2}\right)$$

pf: Let $\mathcal{F} = \{\mathbb{1}_{(-\infty, t]} : t \in \mathbb{R}\}$.

Then

$$\|g_n - g\|_\infty = \sup_{t \in \mathbb{R}} \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \le t] - \mathbb{P}[X \le t]\right|$$

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

Since $\mathcal{F}$ is 1-bounded

$$\mathbb{P}\left[\|g_n - g\|_\infty \le 2R_n(\mathcal{F}) + \delta\right] \ge 1 - \exp\left(-\frac{n\delta^2}{2}\right)$$

Observe $\mathcal{F}$ has poly discrimination with $v = 1$ ☒

Vapnik-Chervonenkis (VC) Theory

Method for bounding polynomial discrimination of $\{0,1\}$-valued $\mathcal{F}$.

Consider a class $\mathcal{F}$ of binary valued functions on $\mathcal{X}$.

Def: We say that $x=(x_1,\dots,x_n)$ is *shattered* by $\mathcal{F}$ if $\text{card}(\mathcal{F}(x))=2^n$.

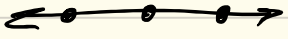
The *VC dimension* is

$$VC(\mathcal{F})=\sup\{n\in\mathbb{N}: \exists x\in\mathcal{X}^n \text{ shattered by } \mathcal{F}\}$$

Notation: If $\mathcal{F}=\{\mathbb{1}_S : S \text{ some sets}\}$ set $\mathcal{S}(x):=\mathcal{F}(x)$ and $VC(\mathcal{S}):=VC(\mathcal{F})$

Ex: $\mathcal{S}_{left}=\{(-\infty,a]: a\in\mathbb{R}\} \Rightarrow VC(\mathcal{S}_{left})=1$

$\mathcal{S}_{two}=\{(a,b]: a,b\in\mathbb{R}, a<b\} \Rightarrow VC(\mathcal{S}_{two})=2$



Thm (Sauer and Shelah)

For any $x = (x_1, \ldots, x_n)$ with $n \geq VC(S)$

we have

$$\text{card}(S(x)) \leq \sum_{i=0}^{VC(S)} \binom{n}{i} \leq (n+1)^{VC(S)}$$

Therefore F has polynomial discrimination of order $VC(S)$ and

$$\mathcal{R}_n(F) \leq 2\sqrt{\frac{VC(S) \cdot \log(n+1)}{n}}$$

$\log(n+1)$ can be removed.

[See Prop 4.18 in Wainwright for a proof.]

Examples: Let G be a class of functions.

For any $g: \mathcal{X} \to \mathbb{R}$ define

$$S_g = \{x \in \mathcal{X} \; g(x) \leq 0\}$$

$$S(G) = \{S_g : g \in G\}$$

Prop: Let G be a vector space of functions $g: \mathbb{R}^d \to \mathbb{R}$ with $\dim(G) < \infty$. Then

$$VC(S(G)) \leq \dim(G)$$

pf: Set $n = \dim(G) + 1$ and fix $x = \{x_1, \ldots, x_n\}$ with $x_i \in K$. Define $L: G \to \mathbb{R}^n$ by

$$L(g) = (g(x_1), \ldots, g(x_n))$$

Since $n > \dim(G)$, there exists $0 \neq \gamma \in \mathbb{R}^n$ s.t. $\langle \gamma, L(g) \rangle = 0 \quad \forall g \in G$.

$$\Rightarrow \sum_{\{i : \gamma_i \leq 0\}} (-\gamma_i) g(x_i) = \sum_{\{i : \gamma_i > 0\}} \gamma_i g(x_i) \quad \forall g \in G$$

WLOG suppose $\gamma_i > 0$ for some i.

Suppose there were $g \in G$ such that S_g includes only $\{x_i : \gamma_i \leq 0\}$. Then

$$0 \geq LHS = RHS > 0$$

Contradiction ∎

Ex: (Halfspaces)

Define $S_{a,b} = \{x \in \mathbb{R}^d : \langle a, x \rangle + b \leq 0\}$

$$\mathcal{S} = \{S_{a,b} : a, b \in \mathbb{R}\}$$

$$G = \{x \mapsto \langle a, x \rangle + b : a, b \in \mathbb{R}\}$$

Then

$$VC(\mathcal{S}) \leq \dim(G) = d+1.$$

actually equality

Ex: (Balls)

Define $S_{a,b} = \{x \in \mathbb{R}^d : \|x - a\|_2 \leq b\}$

$$\mathcal{S} = \{S_{a,b} : a \in \mathbb{R}^d, b \geq 0\}$$

Define $g_{a,b}(x) = \|x\|_2^2 - 2\langle a, x \rangle + \|a\|^2 - b^2$

$$= \|x - a\|^2 - b^2$$

Trick: Define $\Phi : \mathbb{R}^d \to \mathbb{R}^{d+2}$ by

$\Phi(x) = (1, x_1, \ldots, x_d, \|x\|_2^2)$

$g_c(x) = \langle c, \Phi(x) \rangle$ where $c \in \mathbb{R}^{d+2}$

Then

$$\left\{ g_{a,b} : a \in \mathbb{R},\ b \geq 0 \right\} \subseteq \left\{ g_c : c \in \mathbb{R}^d \right\}$$

vector space of dimension $d+2$.

$\Rightarrow VC(S) \leq d+2$

[Exact bound is $d+1$: harder to prove]

Rademacher Complexity and VC-dim often scale with the dimension of the ambient space $x \in \mathcal{X}$.

Thm: Consider $\min_{w \in W \subseteq \mathbb{R}^d} f(w) = \mathbb{E}_z f(w,z)$ where $\max_{w \in W} \|w\| \leq B$ and $f(\cdot, z)$ is L-Lipschitz

Then $\mathbb{E}\left[\sup_{w \in W} \left| \frac{1}{n}\sum_{i=1}^n f(w,z_i) - \underbrace{\mathbb{E} f(w,z)}_{= f(w)} \right|\right] \leq O\left(\sqrt{\frac{L^2 B^2 d \log(n)}{n}}\right)$ (tight)

[We'll prove this later!]

Dimension Independent Bound for Generalization

- linear models
- convexity

Linear Models:

Consider the problem

$$\min_{w \in W} \mathbb{E}_{(a,b)\sim P} \ell(\langle w, a \rangle, b)$$

To get generalization bounds we need to compute $\mathcal{R}_n(F)$ where

$$F = \{(a,b) \mapsto \ell(\langle w, a \rangle, b) : w \in W\}$$

Lemma: (Contraction)
Consider a set $A \subseteq \mathbb{R}^n$ and let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function.
Let $A' = \{(\phi_1(a_1), \ldots, \phi_n(a_n)) : a \in A\}$
Then $R(A') \leq \rho R(A)$

pf: WLOG, assume $p=1$. It suffices to assume $A' = \{(\phi_1(a_1), a_2, \dots, a_m) : a \in A\}$

Then

$$\mathcal{R}(A') = \mathbb{E}_\varepsilon\left[\sup_{a \in A'} \sum_{i=1}^n \varepsilon_i a_i\right]$$

$$= \frac{1}{2}\mathbb{E}_{\varepsilon_2,\dots,\varepsilon_n}\left[\sup_{a \in A}\left\{\phi_1(a_1) + \sum_{i=2}^n \varepsilon_i a_i\right\} + \sup_{\hat{a} \in A}\left\{-\phi_1(\hat{a}_1) + \sum_{i=2}^n \varepsilon_i \hat{a}_i\right\}\right]$$

$$= \frac{1}{2}\mathbb{E}_{\varepsilon_2,\dots,\varepsilon_n}\left[\sup_{a,\hat{a} \in A} \phi_1(a_1) - \phi_1(\hat{a}_1) + \sum_{i=2}^n \varepsilon_i (a_i + \hat{a}_i)\right]$$

$$\leq \frac{1}{2}\mathbb{E}_{\varepsilon_2,\dots,\varepsilon_n}\left[\sup_{a,\hat{a} \in A} a_1 - \hat{a}_1 + \sum_{i=2}^n \varepsilon_i (a_i + \hat{a}_i)\right]$$

$$= \mathcal{R}(A)$$

∎

Rademacher Complexity of linear classes

Lemma: Consider a set of vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ and define

$$A = \left\{ (\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle) : \|w\|_2 \leq 1 \right\}$$

Then $$\mathcal{R}(A) \leq \sqrt{\sum_{i=1}^{n} \|x_i\|_2^2}$$

and therefore

$$\mathcal{R}_n\left(\left\{ x \mapsto \langle w, x \rangle : \|w\|_2 \leq 1 \right\} \right) \leq 2 \frac{\mathbb{E} \max_{i=1,\ldots,n} \|x_i\|_2}{\sqrt{n}}$$

Pf:

$$\begin{aligned}
\mathcal{R}(A) &= \mathbb{E}_\varepsilon \sup_{a \in A} \sum_{i=1}^{m} \varepsilon_i a_i \\
&= \mathbb{E}_\varepsilon \sup_{\|w\|_2 \leq 1} \sum_{i=1}^{m} \varepsilon_i \langle w, x_i \rangle \\
&= \mathbb{E}_\varepsilon \sup_{\|w\|_2 \leq 1} \left\langle w, \sum_{i=1}^{m} \varepsilon_i x_i \right\rangle \\
&= \mathbb{E}_\varepsilon \left\| \sum_{i=1}^{m} \varepsilon_i x_i \right\|_2 \\
&\leq \sqrt{\mathbb{E}_\varepsilon \left\| \sum_{i=1}^{m} \varepsilon_i x_i \right\|_2^2} \leq \sqrt{\sum_{i=1}^{m} \|x_i\|_2^2}
\end{aligned}$$

□

Lemma: Consider a set of vectors $x_1, \ldots, x_m \in \mathbb{R}^d$ and define

$$A = \left\{ (\langle w, x_1 \rangle, \ldots, \langle w, x_m \rangle) : \|w\|_1 \le 1 \right\}$$

Then

$$\mathcal{R}(A) \le \sqrt{2n \log(2d)} \cdot \max_{i=1,\ldots,n} \|x_i\|_\infty$$

and therefore

$$\mathcal{R}_n\left(\left\{ x \mapsto \langle w, x \rangle : \|w\|_1 \le 1 \right\} \right) \le \sqrt{8 \ln(d)} \cdot \frac{\mathbb{E} \max_{i=1,\ldots,n} \|X_i\|_\infty}{\sqrt{n}}$$

Pf:

$$\begin{aligned}
\mathcal{R}(A) &= \mathbb{E}_\varepsilon \sup_{a \in A} \sum_{i=1}^m \varepsilon_i a_i \\
&= \mathbb{E}_\varepsilon \sup_{\|w\|_1 \le 1} \sum_{i=1}^m \varepsilon_i \langle w, x_i \rangle \\
&= \mathbb{E}_\varepsilon \sup_{\|w\|_1 \le 1} \left\langle w, \sum_{i=1}^m \varepsilon_i x_i \right\rangle \\
&= \mathbb{E}_\varepsilon \left\| \sum_{i=1}^m \varepsilon_i x_i \right\|_\infty \\
&= \mathbb{E}_\varepsilon \sup_{v \in \{\pm e_j\}_{j=1}^d} \sum_{i=1}^m \varepsilon_i \langle x_i, v \rangle
\end{aligned}$$

$$\mathbb{E}_\varepsilon \sup_{v \in \{\pm e_j\}_{j=1}^d} \sum_{i=1}^n \varepsilon_i \langle x_i, v \rangle$$

$$= \mathcal{R}\left((\langle x_1, v\rangle, \dots, \langle x_n, v\rangle) : v \in \{\pm e_j\} \right)$$

$$\le \sqrt{2 \log(2d)} \cdot \max_{v \in \{\pm e_j\}} \| (\langle x_1, v\rangle, \dots, \langle x_n, v\rangle) \|_2$$

$$\le \sqrt{2n \log(2d)} \cdot \max_{i=1,\dots,n} \|x_i\|_\infty \quad \blacksquare$$

Back to $\min_{w \in W} \mathbb{E}_{(a,b)\sim P} \ell(\langle w, a\rangle, b)$ where $W = B_2$ or B_1

So if $\ell(\cdot, b)$ is ρ-Lipschitz $\forall b$, then the Rademacher bounds are

ℓ_2-case: $\rho \cdot \dfrac{\mathbb{E} \max_{i=1,\dots,n} \|a\|_2}{\sqrt{n}}$

ℓ_1-case: $\rho \cdot \sqrt{\log(d)} \cdot \dfrac{\mathbb{E} \max_{i=1,\dots,n} \|a_i\|_\infty}{\sqrt{n}}$

Convexity: Regret without uniform laws

Suppose we want to solve

$$\text{(*)} \quad \min_{w \in W} f(w) = \mathbb{E}_{x \sim P} f(w, x)$$

Let $S = \{x_1, \ldots, x_n\}$ be i.i.d from P, and let $A(S)$ be an output of an algorithm which aims to approximately solve (*). Let $S^i = (x_1, \ldots, x_{i-1}, x', x_{i+1}, \ldots, x_n)$ where $x' \sim P$ independent of $x_1, \ldots, x_n$

Which $A(S)$ generalizes?

Intuition: $f(A(S^i), x_i) - f(A(S), x_i)$ should not be big, otherwise overfitting

Thm:

$$\mathbb{E}_S\left[f(A(S)) - \frac{1}{n}\sum_{i=1}^{n} f(A(S), x_i)\right]$$

$$= \mathbb{E}_{\substack{(S, x') \sim P^{n+1} \\ i \sim U(n)}}\left[f(A(S^i), x_i) - f(A(S), x_i)\right]$$

Pf: For every i,

$$\mathbb{E}_S f(A(S)) = \mathbb{E}_{S,z'} f(A(S), x') = \mathbb{E}_{S,z'} f(A(S^i), x_i)$$

Observe

$$\mathbb{E}_S \frac{1}{n}\sum_{i=1}^n f(A(S), x_i) = \mathbb{E}_{S,i}\left[f(A(S), x_i)\right]$$

Defn: $A(\cdot)$ is leave-one-out stable with rate $\varepsilon(n)$ if

$$\mathbb{E}_{(S,x')\sim p^{n+1}, i\sim U(n)}\left[f(A(S^i), x_i) - f(A(S), x_i)\right] \le \varepsilon(n).$$

Henceforth, fix $\lambda > 0$ and we'll analyze

$$A(S) := \arg\min_{w\in W} \frac{1}{n}\sum_{i=1}^n f(w, x_i) + \frac{\lambda}{2}\|w\|^2.$$

Also suppose W and $f(\cdot, x_i)$ are convex.

Defn: A function $f: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is α-strongly convex if $f - \frac{\alpha}{2}\|\cdot\|^2$ is convex.

Lemma: If f is α-strongly convex, then it has a unique minimizer $\bar{w}$ and

$$f(w) - f(\bar{w}) \geq \frac{\alpha}{2}\|w - \bar{w}\|^2 \quad \forall w$$

Thm: Suppose W is convex, and $f(\cdot; x)$ is convex and ρ-Lipschitz. Then the rule

$$A(S) = \operatorname*{argmin}_{w \in W} \frac{1}{n}\sum_{i=1}^{n} f(w, x_i) + \frac{\lambda}{2}\|w\|^2$$

is leave-one-out-stable with rate $\frac{2\rho^2}{\lambda n}$

Therefore

$$\mathbb{E}_S\left[f(A(S)) - \frac{1}{n}\sum_{i=1}^{n} f(A(S), x_i)\right] \leq \frac{2\rho^2}{\lambda n}$$

pf: Define $f_S(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, x_i) + \frac{\lambda}{2}\|w\|^2$.

$\Rightarrow f_S(w) - f_S(A(S)) \geq \frac{\lambda}{2}\|w - A(S)\|^2 \quad \forall w$

For all $w, v \in \mathcal{W}$ it holds

$$f_S(w) - f_S(v) = \frac{1}{n}\sum_{x \in S} f(w,x) + \frac{\lambda}{2}\|w\|^2 - \frac{1}{n}\sum_{x \in S} f(v,x) - \frac{\lambda}{2}\|v\|^2$$

$$= \frac{1}{n}\sum_{x \in S^i} f(w,x) + \frac{\lambda}{2}\|w\|^2 - \frac{1}{n}\sum_{x \in S^i} f(v,x) - \frac{\lambda}{2}\|v\|^2 + \frac{f(w,x_i) - f(v,x_i)}{n} + \frac{f(v,x') - f(w,x')}{n}$$

Setting $w = A(S^i)$, $v = A(S)$ get

$$f_S(A(S^i)) - f_S(A(S)) \leq f_{S^i}(A(S^i)) - f_{S^i}(A(S)) + \leq -\frac{\lambda}{2}\|A(S^i) - A(S)\|^2 + \frac{2\rho\|A(S^i) - A(S)\|}{n}$$

$$\Rightarrow \|A(S^i) - A(S)\| \leq \frac{2\rho}{\lambda n} \quad \square$$

Cor: $\mathbb{E}_S f(A(S)) \le \min_{w} f + \frac{\lambda}{2}\|\bar{w}\|^2 + \frac{2\rho^2}{\lambda n}$

where $\bar{w}$ is any minimizer of f on W.

Therefore under optimal choice $\lambda = \sqrt{\frac{4\rho^2}{n\|\bar{w}\|^2}}$

get

$$\mathbb{E}_S(f(A(S)) \le \min f + 2\sqrt{\frac{\rho^2\|\bar{w}\|^2}{n}}$$

pf: $\mathbb{E}_S f(A(S)) = \mathbb{E}_S\left[\frac{1}{n}\sum_{x\in S} f(A(S),x)\right]$

$$+ \mathbb{E}_S\left[f(A(S)) - \frac{1}{n}\sum_{x\in S} f(A(S),x)\right]$$

$$\le \mathbb{E}_S\left[\frac{1}{n}\sum_{x\in S} f(A(S),x)\right] + \frac{2\rho^2}{\lambda n}$$

For any w, we have

$$\mathbb{E}_S\left[\frac{1}{n}\sum_{x\in S} f(A(S),x) + \frac{\lambda}{2}\|A(S)\|^2\right] \le \mathbb{E}_S f_S(w)$$

$$\le f(w) + \frac{\lambda}{2}\|w\|^2$$

$$\Rightarrow \mathbb{E}_S\left[\frac{1}{n}\sum_{z\in S} f(A(S),z)\right] \le \min f + \frac{\lambda}{2}\|\bar{w}\|^2$$

□