Generalization & uniform convergence

Def A (centered) RV X is σ-subGaussian (or σ-SG; variance proxy σ^2) if
$$\mathbb{E}\exp(\lambda X) \leq \exp(\lambda^2\sigma^2/2), \quad \forall \lambda \in \mathbb{R}.$$

Lemma If X is σ-subGaussian, then for any $\varepsilon > 0$,
$$\mathbb{P}(X \geq \varepsilon) \leq \exp(-\varepsilon/2\sigma^2).$$

Pf: Exercise. Hint: note that $\mathbb{P}(X \geq \varepsilon) = \inf_{t \geq 0} \mathbb{P}(\exp(tX) \geq \exp(t\varepsilon))$.

- Bounded RV's are SG. Exercise: if $X \in [a,b]$ a.s., X is SG w/ variance proxy $(b-a)^2/4$.
- Gaussians are SG.
- Sums of SG are SG.

Homework 0: if $X_1, \ldots, X_n$ are indep. σ_i-SG RV's, then $Z := \sum_1^n X_i$ is SG with variance proxy $\sum_1^n \sigma_i^2$; and if $\alpha > 0$ then αX_i is $\alpha\sigma_i$-SG (variance proxy: $\alpha^2\sigma_i^2$).

Thus, if

$\bar{X}_n := \frac{1}{n}\sum_1^n X_i$ / then $\bar{X}$ is SG w/ variance proxy $\frac{1}{n^2}\sum_1^n \sigma_i^2$.

$\to$ Lemma says $\mathbb{P}(\bar{X}_n \geq \varepsilon) \leq \exp\left(\frac{-\varepsilon}{2\frac{1}{n^2}\sum_1^n \sigma_i^2}\right) = \exp\left(\frac{-n^2\varepsilon}{2\sum_1^n \sigma_i^2}\right)$.
for indep σ-SG X_i,

Similarly, $\mathbb{P}(\bar{X}_n \leq -\varepsilon) \leq \exp\left(-n^2\varepsilon / 2\sum_1^n \sigma_i^2\right)$

Thus, for indep σ-SG X_i, (each have $\mathbb{E}X_i = 0$)

$$\mathbb{P}(|\bar{X}_n| > \varepsilon) \leq 2\exp\left(-n^2\varepsilon / 2\sum_1^n \sigma_i^2\right).$$

Generalizing to non-mean zero, we get

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_1^n (X_i - \mu_i)\right| > \varepsilon\right) \leq 2\exp\left(-n^2\varepsilon / 2\sum_1^n \sigma_i^2\right).$$

Letting $\sigma^2 := \frac{1}{n}\sum_1^n \sigma_i^2$, we get $\leq 2\exp\left(\frac{-n\varepsilon}{2\sigma^2}\right)$.

For $\varepsilon := \sigma \cdot \sqrt{\frac{2\log(2/\delta)}{n}}$ we get $2\exp(-n\varepsilon/2\sigma^2) = \delta$, so

wp $> 1-\delta$, $\left|\frac{1}{n}\sum_1^n (X_i - \mu_i)\right| \leq \sqrt{\frac{1}{n}\sum_1^n \sigma_i^2} \cdot \sqrt{\frac{2\log(2/\delta)}{n}}$.

If X_i are iid, says wp $> 1-\delta$, $\left|\mu - \frac{1}{n}\sum_1^n X_i\right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}}$.
each σ-SG,

$\longrightarrow$ as n gets larger, sample mean closer to pop mean.

EX. Let $(X_i, Y_i) \overset{iid}{\sim} P$, $X \in \mathbb{R}^d$, $Y \in \{\pm 1\}$, $f: \mathbb{R}^d \to \{\pm 1\}$, and let $Z_i := \mathbb{1}(f(X_i) \neq y_i)$. Then each Z_i is iid, bounded in $[0,1]$ (hence SG: with variance proxy $\frac{1}{4}$). So by above,

w.p. $> 1-\delta$, $\mathbb{P}(y \neq f(x)) \leq \frac{1}{n} \sum_1^n \mathbb{1}(y_i \neq f(x_i)) + \sqrt{\frac{\log(2/\delta)}{2n}}$.

$\longrightarrow$ test error is bounded by train error $+ \tilde{O}(\sqrt{\frac{1}{n}})$.

Example. Suppose $(X_i, Y_i)_1^n$ are iid. For any $n \in \mathbb{N}$, define:

$$f_n(x) := \begin{cases} y_i : & x \in \{x_1, \dots, x_n\}, \\ -10 : & \text{otherwise} \end{cases}$$

Consider two situations:

① X has finite support. Then $\frac{1}{n}\sum_1^n \mathbb{1}(y_i \neq f_n(x_i)) = 0$ for all n by def^n.
and $\mathbb{P}(y \neq f_n(x)) \to 0$ as well, since we recover all pts.

② X has continuous distn. Then $\frac{1}{n}\sum_1^n \mathbb{1}(y \neq f(x_i)) = 0$ by construction,
but $\mathbb{P}(y \neq f_n(x)) = 1$ $\forall n$.

What broke subG concentration?

f_n is a random variable. Although (x_i, y_i) are iid,

$Z_i := \mathbb{1}(y_i f_n(x_i))$ are not independent.

(2) is overfitting: $\hat{L}(f) = 0$ but $L(f) = 1$.

How can we guarantee test error is small when looking at training error?

We'll see how via uniform convergence:

For iid z_i, loss $\ell(z_i)$,

$$L(f) := \mathbb{E} f(z), \quad \hat{L}(f) = \frac{1}{n}\sum_i^n f(z_i),$$

Goal: bound $L(f)$. Suppose $f \in \mathcal{F}$, some fcn class $\mathcal{F}$.

And suppose we use $S = \{z_i\}_i^n$ to fit $f = f(S)$. Then we typically lose indep. of $f(z_i; S)$.

Approach is then:

$$\begin{aligned} L(f) &= L(f) - \hat{L}(f) + \hat{L}(f) \\ &\leq \hat{L}(f) + \sup_{f \in \mathcal{F}} \{L(f) - \hat{L}(f)\}. \end{aligned}$$

Seems very silly, but we will see very fruitful to do so.

We'll prove deviation bounds that hold uniformly over $f \in \mathcal{F}$.

Example. Let $\mathcal{F} = \{f_1, \dots, f_k\}$, $|\mathcal{F}| = k$. If $(x_i, y_i)_1^n$ are iid, $f_i : \mathbb{R}^d \to \{\pm 1\}$, then SG concentration as before gives for fixed f_ℓ,

$$\mathbb{P}_{(x_i,y_i)_1^n}\left(\left| \mathbb{P}(f_\ell(X) \neq y) - \frac{1}{n}\sum_1^n \mathbb{1}(y_i \neq f_\ell(x_i)) \right| > \sqrt{\frac{\log 2/\delta}{2n}} \right) \leq \delta.$$

for fixed ℓ, w.p. $\geq 1-\delta$, $\left| \mathbb{P}(f_\ell(X) \neq y) - \hat{\mathbb{P}}(f_\ell(X) \neq y) \right| \leq \sqrt{\frac{\log \frac{2}{\delta}}{n}}$.

Union bound:

$$\mathbb{P}\left(\exists \ell \in [k] : \left| \mathbb{P}(f_\ell(X) \neq y) - \hat{\mathbb{P}}(f_\ell(X) \neq y) \right| \leq \sqrt{\frac{\log 2k/\delta}{2n}} \right) \leq k \cdot \frac{\delta}{k} = k.$$

i.e. w.p. $\geq 1-\delta$, for all $\ell \in [k]$, $\left| \mathbb{P}(y \neq f_\ell(X)) - \hat{\mathbb{P}}(f_\ell(X) \neq y) \right| \leq \sqrt{\frac{\log(2k/\delta)}{2n}}$.

$$\leq \sqrt{\frac{\log|\mathcal{F}|}{2n}} + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

For finite classes, get $\sqrt{\frac{\log|\mathcal{F}|}{2n}}$ extra term. We'll see next that Rademacher complexity allows for dealing w/ $|\mathcal{F}| = \infty$.

Def. For $V \subset \mathbb{R}^n$, the unnormalized/normalized Rademacher complexity is

$$\mathrm{URad}(V) := \mathop{\mathbb{E}}_{\varepsilon} \sup_{u \in V} \langle \varepsilon, u \rangle, \qquad \mathrm{Rad}(V) = \tfrac{1}{n} \mathrm{URad}(V),$$

where $\varepsilon \in \mathbb{R}^n$ is iid Rademacher: $\varepsilon_i \sim \mathrm{Unif}(\{\pm 1\})$.

We will typically apply this to outputs of a function class over training data.

E.g. for $z_i = (x_i, y_i)$, $S = \{z_i\}_1^n$, fn class $\mathcal{F}$,

$$\mathcal{F}_{|S} := \{ (f(z_1), \dots, f(z_n)) : f \in \mathcal{F} \}.$$

$$\rightarrow \mathrm{URad}(\mathcal{F}_{|S}) = \mathop{\mathbb{E}}_{\varepsilon} \sup_{f \in \mathcal{F}} \langle \varepsilon, u \rangle = \mathop{\mathbb{E}}_{\varepsilon} \sup_{f \in \mathcal{F}} \sum_1^n \varepsilon_i f(z_i).$$

- $\mathrm{URad}(\mathcal{F}_{|S})$ is large if, for any $\varepsilon_i \in \{\pm 1\}$, there is some $f \in \mathcal{F}$ st $f(z_i) \approx \varepsilon_i$.
- If we think of $f(z_i) \in \{\pm 1\}$, then this corresponds to $\mathcal{F}$ fitting "random labels" ε_i.
- We'll often look at URad for losses, i.e. for ℓ,

$$\mathrm{URad}((\ell \circ \mathcal{F})_{|S}) = \mathrm{URad}\big((\ell(y_1 f(x_1)), \dots, \ell(y_n f(x_n))) : f \in \mathcal{F} \big).$$

- $\mathrm{Rad}(V)$ roughly measures how large/complicated V is.

Properties: ① $\mathrm{URad}(\{u\}) = \mathbb{E}\langle \varepsilon, u\rangle = 0$.

② $\mathrm{URad}(V + \{u\}) = \mathrm{URad}(\{v+u : v \in V\}) = \mathrm{URad}(V)$.

③ If $V \subset V'$, $\mathrm{URad}(V) \subset \mathrm{URad}(V')$.

④ $\mathrm{URad}(\{\pm 1\}^n) = \mathbb{E}_{\varepsilon} \sup_{x \in \{\pm 1\}^n} \langle \varepsilon, x\rangle = \mathbb{E}_{\varepsilon} \|\varepsilon\|^2 = n$.

→ $\{\pm 1\}^n$ is as large as possible among vectors taking vals in ± 1.

⑤ $\mathrm{URad}(\{(-1,-1,\dots,-1),(1,\dots,1)\}) = \mathbb{E}_{\varepsilon} \max\{\sum \varepsilon_i, -\sum \varepsilon_i\} = \mathbb{E}_{\varepsilon} \left|\sum_1^n \varepsilon_i\right|$.

$\left|\sum_1^n \varepsilon_i\right| = \left|\sum_1^n (2 \cdot \mathrm{Ber}(\tfrac{1}{2}) - 1)\right| = \left|2 \cdot \mathrm{Bin}(n, \tfrac{1}{2}) - n\right|$.

Anti-concentration of Binomial shows $\left|2\mathrm{Bin}(n,\tfrac{1}{2}) - n\right| = \Theta(\sqrt{n})$.

You will also sometimes see an absolute value version of Rad. complexity,

$$\widetilde{\mathrm{URad}}(V) := \mathbb{E}_{\varepsilon} \sup_{u \in V} |\langle \varepsilon, v\rangle|.$$

Similar idea, but a bit less nice for reasons we won't get into.

Theorem. Let $\mathcal{F}$ be a fcn class w/ $f(z) \in [a,b]$ $\forall z$, $\forall f \in \mathcal{F}$. Let $\mathbb{P}$: distr over z.

① For any $\delta \in (0,1)$, w.p. $> 1-\delta$,

$$\sup_{f \in \mathcal{F}} \left\{ \mathbb{E} f(z) - \frac{1}{n} \sum_1^n f(z_i) \right\} \le \mathbb{E}_{z_i} \left(\sup_{f \in \mathcal{F}} \left\{ \mathbb{E} f(z) - \frac{1}{n} \sum_1^n f(z_i) \right\} \right) + (b-a) \sqrt{\frac{\log 1/\delta}{2n}}.$$

② wp $> 1-\delta$,

$$\mathbb{E}_{z_i} \, \mathrm{Rad}(\mathcal{F}_{|S}) \le \mathrm{Rad}(\mathcal{F}_{|S}) + (b-a) \sqrt{\frac{\log(1/\delta)}{2n}}$$

③ wp $> 1-\delta$,

$$\sup_{f \in \mathcal{F}} \left\{ \mathbb{E} f(z) - \frac{1}{n} \sum_i^n f(z_i) \right\} \le 2 \, \mathrm{Rad}(\mathcal{F}_{|S}) + 3(b-a) \sqrt{\frac{\log \frac{2}{\delta}}{n}}$$

To prove this, we'll use McDiarmid's ineq:

<u>Thm</u> (MacDiarmid). Suppose $F: \mathbb{R}^n \to \mathbb{R}$ satisfies <u>bounded differences</u>: $\forall i \in [n]$, $\exists c_i$ st

$$\sup_{z_1, \dots, z_n, z_i'} |F(z_1, \dots, z_i, z_{i+1}, \dots, z_n) - F(z_1, \dots, z_i', \dots, z_n)| \le c_i.$$

Then, wp $> 1-\delta$,

$$\mathbb{E}_{z_i} F(z_1, \dots, z_n) \le F(z_1, \dots, z_n) + \sqrt{\frac{\sum_i c_i^2 \log \frac{1}{\delta}}{2}}.$$

Lemma. Let $(z_1, \dots, z_n), (z_1', \dots, z_n')$ be iid from $\mathbb{P}$.

Let $\hat{\mathbb{P}}_n$: uniform on $(z_1, \dots, z_n)$; $\hat{\mathbb{P}}_n'$: uniform on $(z_1', \dots, z_n')$. Same for $\mathbb{P}_n, \mathbb{P}_n'$.

Then $\mathbb{E}_n \left[\sup_{f \in \mathcal{F}} \{ \mathbb{E}f - \hat{\mathbb{E}}_n f \} \right] \le \mathbb{E}_n \left[\mathbb{E}_n' \left(\sup_{f \in \mathcal{F}} \{ \hat{\mathbb{E}}_n' f - \hat{\mathbb{E}}_n f \} \right) \right]$.

Pf.

First note that since $z_i' \overset{d}{=} z$,

$$\mathbb{E} f_\varepsilon = \mathbb{E}_{z \sim \mathbb{P}} f_\varepsilon(z) = \mathbb{E}_n' \hat{\mathbb{E}}_n' f_\varepsilon, \text{ since } z_i' \sim \mathbb{P} \text{ so } \mathbb{E}_{z_i'} f_\varepsilon(z_i') = \mathbb{E} f_\varepsilon.$$

Let $\varepsilon > 0$. Then $\exists f_\varepsilon \in \mathcal{F}$ s.t. $\sup_{f \in \mathcal{F}} \{ \mathbb{E}f - \hat{\mathbb{E}}_n f \} \le \mathbb{E} f_\varepsilon - \hat{\mathbb{E}}_n f_\varepsilon + \varepsilon$.

$$\begin{aligned}
\Rightarrow \mathbb{E}_n \left[\sup_{f \in \mathcal{F}} \{ \mathbb{E}f - \hat{\mathbb{E}}_n f \} \right] &\le \mathbb{E}_n \left[\mathbb{E} f_\varepsilon - \hat{\mathbb{E}}_n f_\varepsilon + \varepsilon \right] \\
&= \mathbb{E}_n \left[\mathbb{E}_n' \hat{\mathbb{E}}_n' f_\varepsilon - \hat{\mathbb{E}}_n f_\varepsilon + \varepsilon \right] \\
&= \mathbb{E}_n' \mathbb{E}_n \left[\hat{\mathbb{E}}_n' f_\varepsilon - \hat{\mathbb{E}}_n f_\varepsilon \right] + \varepsilon \\
&\le \mathbb{E}_n' \mathbb{E}_n \left[\sup_{f \in \mathcal{F}} \{ \hat{\mathbb{E}}_n' f_\varepsilon - \hat{\mathbb{E}}_n f_\varepsilon \} \right] + \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, this completes the proof. $\blacksquare$

Lemma. $\mathbb{E}_n\left[\mathbb{E}_n' \sup_{f\in\mathcal{F}} \{\hat{\mathbb{E}}_n' f - \hat{\mathbb{E}}_n f\}\right] \le 2\mathbb{E}_n \mathrm{Rad}(\mathcal{F}|_S)$.

Pf For fixed $\varepsilon \in \{\pm 1\}^n$, let RV $\xi_i := (u_i, u_i') := \begin{cases} (z_i, z_i'), & \varepsilon = 1; \\ (z_i', z_i), & \varepsilon = -1. \end{cases}$

By def$\underline{n}$,

$$\mathbb{E}_n \mathbb{E}_n'\left[\sup_{f\in\mathcal{F}}\{\hat{\mathbb{E}}_n' f - \hat{\mathbb{E}}_n f\}\right] = \mathbb{E}_n \mathbb{E}_n'\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_1^n (f(z_i') - f(z_i))\right\}\right]$$
$$= \mathbb{E}_n \mathbb{E}_n'\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_1^n \varepsilon_i \cdot (f(u_i') - f(u_i))\right\}\right]$$

Since $\{z_i, z_i'\}$ are iid, if ε_i are iid Rademacher, so are $\{u_i, u_i'\}$, and in particular $(z_1, \dots, z_n, z_1', \dots, z_n') = (u_1, \dots, u_n, u_1', \dots, u_n')$. Thus,

$$\mathbb{E}_\varepsilon \mathbb{E}_n \mathbb{E}_n'\left[\sup_{f\in\mathcal{F}}\{\hat{\mathbb{E}}_n' f - \hat{\mathbb{E}}_n f\}\right] = \mathbb{E}_\varepsilon \mathbb{E}_n \mathbb{E}_n'\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_1^n \varepsilon_i (f(u_i') - f(u_i))\right\}\right]$$
$$= \mathbb{E}_\varepsilon \mathbb{E}_n \mathbb{E}_n'\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_1^n \varepsilon_i (f(z_i') - f(z_i))\right\}\right]$$
$$\le \mathbb{E}_\varepsilon \mathbb{E}_n \mathbb{E}_n'\left[\sup_{f, f'\in\mathcal{F}}\left\{\frac{1}{n}\sum_1^n \varepsilon_i (f(z_i') - f'(z_i))\right\}\right]$$
$$= \mathbb{E}_\varepsilon \mathbb{E}_n'\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_1^n \varepsilon_i f(z_i')\right] + \mathbb{E}_\varepsilon \mathbb{E}_n\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_1^n (-\varepsilon_i) f'(z_i)\right]$$
$$= 2\mathbb{E}_\varepsilon \mathbb{E}_n\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_1^n \varepsilon_i f(z_i)\right] \quad \text{since } z_i \stackrel{d}{=} z_i',\ \varepsilon_i \stackrel{d}{=} -\varepsilon_i$$
$$= 2\mathbb{E}_n\left[\frac{1}{n}\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \sum_1^n \varepsilon_i f(z_i)\right] = 2\mathbb{E}_n \mathrm{Rad}(\mathcal{F}|_S). \quad \blacksquare$$

This shows that

$$\mathbb{E}_{\{z_i\}_1^n}\left[\sup_{f\in\mathcal{F}}\left\{\mathbb{E}f-\hat{\mathbb{E}}_n f\right\}\right]\le 2\,\mathbb{E}_n\,\mathrm{Rad}(\mathcal{F}|_S).$$

We'll now work on making a high-probability version of this

<u>Theorem</u> (McDiarmid bounded differences):

Suppose $g:\mathbb{R}^n\to\mathbb{R}$ is s.t. $\forall i\in\{1,\dots,n\}$, $\exists c_i$ s.t.

$$\sup_{z_1,\dots,z_n,z_i'}\left|g(z_1,\dots,z_i,\dots,z_n)-g(z_1,\dots,z_{i-1},z_i',z_{i+1},\dots,z_n)\right|\le c_i.$$

Then, w.p. $\geq 1-\delta$,

$$\mathbb{E}_n\, g(z_1,\dots,z_n)\le g(z_1,\dots,z_n)+\sqrt{\frac{\sum_i c_i^2}{2}\log(1/\delta)}$$

Pf omitted, see linked notes from Daniel Hsu

We'll now prove Thm XX.

① We will verify that $\sup_{f\in\mathcal{F}}\{\mathbb{E} f(Z) - \hat{\mathbb{E}}_n f\}$ satisfies bounded difference with constant $\frac{b-a}{n}$.

Consider $z_1,\dots,z_n,z_i'$. For $j\neq i$, call $z_j'=z_i$. Then

$$\left|\sup_{f\in\mathcal{F}}\{\mathbb{E}f-\hat{\mathbb{E}}_n f\} - \sup_{g\in\mathcal{F}}\{\mathbb{E}g-\hat{\mathbb{E}}_n' g\}\right| \qquad \sup\{f+g$$

$$= \left|\sup_{f\in\mathcal{F}}\{\mathbb{E}f-\hat{\mathbb{E}}_n f\} - \sup_{g\in\mathcal{F}}\{\mathbb{E}g-\tfrac{1}{n}\textstyle\sum_1^n g(z_i') + \tfrac{1}{n}g(z_i) - \tfrac{1}{n}g(z_i)\}\right|$$

$$= \left|\sup_{f\in\mathcal{F}}\{\mathbb{E}f-\hat{\mathbb{E}}_n f\} - \sup_{g\in\mathcal{F}}\{\mathbb{E}g-\tfrac{1}{n}\textstyle\sum_1^n g(z_i) + \frac{g(z_i)}{n} - \frac{g(z_i')}{n}\}\right|$$

$$\leq \sup_{h\in\mathcal{F}}\left\{\left|\sup_{f\in\mathcal{F}}\{\mathbb{E}f-\hat{\mathbb{E}}_n f\} - \sup_{g\in\mathcal{F}}\{\mathbb{E}g-\hat{\mathbb{E}}_n g + \frac{h(z_i)}{n} - \frac{h(z_i')}{n}\}\right|\right\}$$

$$= \sup_{h\in\mathcal{F}}\left|\frac{h(z_i)-h(z_i')}{n}\right| \leq \frac{b-a}{n}.$$

$\Rightarrow$ satisfies bounded diff. w/ $c_i=\frac{b-a}{n}$ $\forall i$. $\sum c_i^2 = \frac{n(b-a)^2}{n^2}$ so

Thus, w.p. $\geq 1-\delta$,

$$\sup_{f\in\mathcal{F}}\{\mathbb{E}f-\hat{\mathbb{E}}_n f\} \leq \mathbb{E}_n\left[\sup_{f\in\mathcal{F}}\{\mathbb{E}f-\hat{\mathbb{E}}_n f\}\right] + \sqrt{\frac{(b-a)^2\log(2/\delta)}{n}}.$$

Let $S=\{z_i\}$, $S'=\{z_i'\}$.

② $\left| URad(\mathcal{F}|_S) - URad(\mathcal{F}|_{S'}) \right|$

$$= \left| URad(\mathcal{F}|_S) - \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \sum_1^n \varepsilon_i f(z_i') \right|$$

$$= \left| URad(\mathcal{F}|_S) - \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left\{ \sum_1^n \varepsilon_i f(z_i) - \varepsilon_i f(z_i) + \varepsilon_i f(z_i) \right\} \right|$$

$$= \left| URad(\mathcal{F}|_S) - \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left\{ \sum_1^n \varepsilon_i f(z_i) - \varepsilon_i f(z_i) + \varepsilon_i f(z_i') \right\} \right|$$

$$\le \sup_{h\in\mathcal{F}} \left| URad(\mathcal{F}|_S) - \mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}} \left\{ \sum_1^n \varepsilon_i f(z_i) - \varepsilon_i h(z_i) + \varepsilon_i h(z_i') \right\} \right|$$

$$= \sup_{h\in\mathcal{F}} \left| \mathbb{E}_\varepsilon \left[\varepsilon_i h(z_i) - \varepsilon_i h(z_i') \right] \right|$$

$$\le \sup_{h\in\mathcal{F}} \mathbb{E}_\varepsilon \left| \varepsilon_i h(z_i) - \varepsilon_i h(z_i') \right| \le (b-a).$$

Satisfies bounded differences w/ $c_i = b-a$, so $\sum c_i^2 = (b-a)^2 n$; dividing by n to get normalized Rad complexity we get

wp $> 1-\delta$,

$$\mathbb{E}_n Rad(\mathcal{F}|_S) \le \quad Rad(\mathcal{F}|_S) + \sqrt{\frac{(b-a)^2 \log 2/\delta}{n}}.$$

Putting everything together,

Thus, w.p. $\geq 1-2\delta$,

$$\sup_{f\in\mathcal{F}}\{\mathbb{E}f-\hat{\mathbb{E}}_n f\} \leq \mathbb{E}_n\left[\sup_{f\in\mathcal{F}}\{\mathbb{E}f-\hat{\mathbb{E}}_n f\}\right] + \sqrt{\frac{(b-a)^2\log(2/\delta)}{n}}.$$

$$\leq \mathbb{E}_n\left[\mathbb{E}_n'\left(\sup_{f\in\mathcal{F}}\{\hat{\mathbb{E}}_n' f-\hat{\mathbb{E}}_n f\}\right)\right] + (b-a)\sqrt{\frac{\log 2/\delta}{n}}$$

$$\leq 2\,\mathbb{E}_n \operatorname{Rad}(\mathcal{F}_{|S}) + (b-a)\sqrt{\frac{\log 2/\delta}{n}}$$

$$\leq 2\operatorname{Rad}(\mathcal{F}_{|S}) + 3(b-a)\sqrt{\frac{\log 2/\delta}{n}}. \qquad \boxtimes$$

Thus Rademacher complexity provides a distribution-dependent (via $\mathcal{F}_{|S}$; S depends on $\mathbb{P}$) way to guarantee uniform convergence.

We'll now instantiate for particular function classes.

<u>Example</u> logistic regression with bounded weights.

$$\ell(y f(x)) := \log(1+\exp(-y f(x)));$$

$$\mathcal{F} = \{w\in\mathbb{R}^d : \|w\|\leq B\};$$

$$(\ell\circ\mathcal{F})_{|S} := \{(\ell(y_1 w^T x_1),\dots,\ell(y_n w^T x_n)) : \|w\|\leq B\},$$

$$R(w) := \mathbb{E}\,\ell(y\langle w,x\rangle),\quad \hat{R}(w) = \frac{1}{n}\sum_1^n \ell(y_i\langle w,x_i\rangle).$$

Via prev theorem, suffices to bound $\operatorname{Rad}((\ell\circ\mathcal{F})_{|S})$.

Lemma Let $l : \mathbb{R}^n \to \mathbb{R}^n$ have components l_i which are univariate & L-Lip.

Then $\mathrm{Rad}(l \circ V) \le L \cdot \mathrm{Rad}(V)$.

Pf
$$\mathrm{URad}(l \circ V) = \mathbb{E}_\varepsilon \sup_{u \in V} \sum_1^n \varepsilon_i l_i(u_i)$$
$$= \mathbb{E}_\varepsilon \Big[\sup_{u \in V} \Big\{ \varepsilon_1 l_1(u_1) + \sum_2^n \varepsilon_i l_i(u_i) \Big\} \Big]$$
$$= \frac{1}{2} \mathbb{E}_{\varepsilon_{2:n}} \Big[\sup_{u \in V} \Big\{ l_1(u_1) + \sum_2^n \varepsilon_i l_i(u_i) \Big\} + \sup_{u \in V} \Big\{ -l_1(u_1) + \sum_2^n \varepsilon_i l_i(u_i) \Big\}$$
$$= \frac{1}{2} \mathbb{E}_{\varepsilon_{2:n}} \Big[\sup_{u, w \in V} \Big\{ l_1(u_1) - l_1(w_1) + \sum_{i=2}^n \varepsilon_i (l_i(u_i) + l_i(w_i)) \Big\} \Big]$$
$$\le \frac{1}{2} \mathbb{E}_{\varepsilon_{2:n}} \Big[\sup_{u, w \in V} \Big\{ L|u_1 - w_1| + \sum_2^n \varepsilon_i (l_i(u_i) + l_i(w_i)) \Big] $$
$$= \frac{1}{2} \mathbb{E}_{\varepsilon_{2:n}} \Big[\sup_{u, w \in V} \Big\{ L(u_1 - w_1) + \sum_2^n \varepsilon_i (l_i(u_i) + l_i(w_i)) \Big\} \Big]$$
$$= \frac{1}{2} \mathbb{E}_{\varepsilon_{2:n}} \Big[\sup_{u \in V} \Big\{ L u_1 + \sum_2^n \varepsilon_i l_i(u_i) \Big\} + \sup_{w \in V} \Big\{ -L w_1 + \sum_2^n \varepsilon_i l_i(w_i) \Big\} \Big]$$
$$= \mathbb{E} \sup_{u \in V} \Big\{ L \varepsilon_1 u_1 + \sum_2^n \varepsilon_i l_i(u_i) \Big\}$$
$$= \dots = \mathbb{E} \sup_{u \in V} L \langle u, \varepsilon \rangle = \mathrm{URad}(L \cdot V) = L \cdot \mathrm{URad}(V).$$

Corollary If ℓ is L-Lip. & $\ell \circ f \in [a,b]$ a.s., then w.p. $\geq 1-\delta$, $\forall f \in \mathcal{F}$,

$$R_\ell(f) \le \hat{R}_\ell(f) + 2L\,\mathrm{Rad}(\mathcal{F}|_S) + 3(b-a)\sqrt{\frac{\log\frac{2}{\delta}}{n}}.$$

Pf:
$$|\ell(-y_i f(x_i)) - \ell(-y_i f'(x_i))| \le L|-y_i f(x_i) - y_i f'(x_i)|$$
$$\le L|f(x_i) - f'(x_0)|. \quad \blacksquare$$

Theorem Given $S = (x_1, \dots, x_n)$, $X \in \mathbb{R}^{n \times d}$ with rows x_i^T,

$$\mathrm{Rad}(\{x \mapsto \langle w, x\rangle : \|w\|_2 \le B\}|_S) \le B\|X\|_F / n.$$

Pf Let $\varepsilon \in \{\pm 1\}^n$. Then,

$$\sup_{\|w\| \le B} \sum_i \varepsilon_i \langle w, x_i \rangle = \sup_{\|w\| \le B} \langle w, \sum_i \varepsilon_i x_i \rangle$$
$$= \sup_{\|w/B\| \le 1} \langle w, \sum_i \varepsilon_i x_i \rangle$$
$$= \|\sum_i \varepsilon_i x_i\|_2$$

By Jensen's inequality (for convex φ, $\varphi(\mathbb{E}X) \le \mathbb{E}\,\varphi(X)$; reversed for concave)

$$\mathbb{E}\|\textstyle\sum_i \varepsilon_i x_i\|_2 = \mathbb{E}\sqrt{\|\sum_i \varepsilon_i x_i\|_2^2} \le \sqrt{\mathbb{E}\|\sum_i x_i \varepsilon_i\|_2^2}.$$

$$\mathbb{E}\|\textstyle\sum_i \varepsilon_i x_i\|_2^2 = \mathbb{E}\left[\sum_i \varepsilon_i^2 \|x_i\|^2 + \sum_{i\neq j} \varepsilon_i \varepsilon_j \langle x_i, x_j\rangle\right]$$

$$= \mathbb{E}\textstyle\sum_i \|x_i\|^2 + 0$$

$$= \textstyle\sum_i \|x_i\|^2 = \|X\|_F^2.$$

$$\Rightarrow \quad \mathrm{Rad}\left(\{x \mapsto \langle w, x\rangle : \|w\|_2 \le B\} \,|\, S\right) \le \frac{1}{n}\mathbb{E}\|\textstyle\sum_i \varepsilon_i x_i\| \le \frac{\|X\|_F}{n}. \quad \boxed{\xi}$$

EX. Logistic regression: Suppose $\|w\| \le B$, $\|x\| \le 1$, $\ell(t) = \log(1+\exp(-t))$. $|\ell'(t)| \le 1$, 1-Lip. loss. Note $\ell(t) \ge 0$ & $\ell(\langle w, yx\rangle) \le \log(2) + \langle w, yx\rangle$
$\le \log 2 + \|w\|$. Then,

w.p. $> 1-\delta$, $\forall w \in \mathbb{R}^d$ with $\|w\| \le B$,

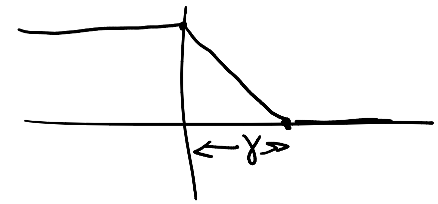
$$R_\ell(w) \le \hat{R}_\ell(w) + 2\mathrm{Rad}\left((\ell \circ f)|S\right) + 3(\log 2 + B)\sqrt{\frac{\log 2/\delta}{n}}$$

$$\le \hat{R}_\ell(w) + \frac{2B\|X\|_F}{n} + 3(\log 2 + B)\sqrt{\frac{\log 2/\delta}{n}}$$

$$\le \hat{R}_\ell(w) + \frac{2B + 3(B + \log 2)\sqrt{\frac{\log 2/\delta}{n}}}{\sqrt{n}} \quad \text{since } \|X\|_F^2 = \textstyle\sum_1^n \|x_i\|^2 \le n.$$

Margin bounds:

Let $\ell_\gamma(z) := \max(0, \min(1 - z/\gamma, 1))$



($z \geq \gamma$: $\ell_\gamma(z) = \max(0, 1 - z/\gamma) = 0$.

$z \leq 0$: $\ell_\gamma(z) = \max(0, 1) = 1$.

"ramp loss". Call $R_\gamma(f) = \mathbb{E}\, \ell_\gamma(y f(x))$, $R_{0\text{-}1}(f) = \mathbb{P}(y \neq \mathrm{sgn}(f(x)))$.

<u>Theorem</u>. For any $\gamma > 0$, wp $\geq 1 - \delta$, $\forall f \in \mathcal{F}$,

$$R_{01}(f) \leq R_\gamma(f) \leq \hat{R}_\gamma(f) + \frac{2}{\gamma} \mathrm{Rad}(\mathcal{F}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{n}}.$$

<u>Pf.</u> $\mathbb{1}(y \neq \mathrm{sgn}(f(x))) = \mathbb{1}(y \cdot f(x) < 0) \leq \ell_\gamma(y f(x))$, thus

$R_{01}(f) \leq R_\gamma(f)$. Next, note that ℓ_γ is γ^{-1}-Lipschitz,

so previous results imply the theorem, as $\ell_\gamma(z) \in [0, 1]$. ∎

If margins are large, & # samples $\gg \gamma^{-1}$, then have uniform convergence.

We'll now prove the following bound on Rademacher complexity of deep NNs.

Notation: for matrix $M \in \mathbb{R}^{m \times d}$, $\|M\|_{b,c} := \|(\|M_{:,1}\|_b, \ldots, \|M_{:,d}\|_b)\|_c$.

Take $\|\cdot\|_b$ norm of each column, then take c norm.

Theorem (Bartlett & Mendelson, '02) If $W^{(1)} \in \mathbb{R}^{m \times d}$, $W^{(i)} \in \mathbb{R}^{m \times m}$, $i = 2, \ldots, L-1$, $W^{(L)} \in \mathbb{R}^{1 \times m}$,

Let $\mathcal{F} = \{x \mapsto \varphi_L(W^{(L)} \varphi_{L-1}(\cdots \varphi_1(W^{(1)} x) \cdots)) :$

$$\|(W^{(i)})^T\|_{1,\infty} = \max_j \|w_j^{(i)}\|_1 \leq B \}$$ ($w_j^{(i)}$: j^{th} row of $W^{(i)}$)

where φ_i are ρ-Lipschitz & $\varphi_i(0) = 0$ $\forall i$.

then:
$$\mathrm{Rad}(\mathcal{F}|S) \leq \frac{1}{n} \|X\|_{2,\infty} \cdot (2\rho B)^L \sqrt{2 \log d}.$$
$$= \frac{1}{n} (2\rho B)^L \sqrt{2 \log d} \cdot \max_j \|([x_1]_j, \ldots, [x_n]_j)\|_2$$

Exp dependence on 2^L is not ideal.

No dependence on number of neurons in the network! Only log dependence on ambient dimension.

Note that we expect $\mathrm{Rad}(\cdot) \to 0$ if ρ is very small: composing many contracting functions collapses to 0.

We'll prove a simpler version for 2 layer nets. Let $\varphi = \mathrm{ReLU}$,

<u>Thm</u> Let $\mathcal{F}_{m,R} = \{ x \mapsto \sum_{j=1}^m a_j \varphi(\langle w_j, x\rangle) : a_j \in \{\pm 1/\sqrt{m}\}, \|W\|_F \le R \}$.

Then
$$\mathrm{Rad}(\mathcal{F}_{m,R}) \le \frac{2R\|X\|_F}{n}.$$
($X \in \mathbb{R}^{n\times d}$ has rows x_i^T)

<u>No</u> dependence on # neurons <u>or</u> ambient dimension!

We first prove the following auxiliary lemma.

<u>Lemma</u> Let $\mathcal{F}$ be a function class s.t. $0 \in \mathcal{F}$. Then,
$$\mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \left| \sum_1^n \varepsilon_i f(z_i) \right| \right] \le 2n\,\mathrm{Rad}(\mathcal{F}|_S).$$

Pf. By def$^{\underline{n}}$,
$$\mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \left| \sum_1^n \varepsilon_i f(z_i) \right| \right] = \mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \max\left\{ \sum_1^n \varepsilon_i f(z_i), -\sum_1^n \varepsilon_i f(z_i) \right\} \right]$$
$$\le \mathbb{E}_\varepsilon \left[\max\left\{ \sup_{f\in\mathcal{F}} \sum_1^n \varepsilon_i f(z_i), \sup_{f\in\mathcal{F}} \sum_1^n -\varepsilon_i f(z_i) \right\} \right]$$

$0 \in \mathcal{F}$ implies each $\sup\{\cdot\}$ is ≥ 0, so $\max(a,b) \le a+b$.

$$\le \mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \sum_1^n \varepsilon_i f(z_i) + \sup_{f\in\mathcal{F}} \sum_1^n \varepsilon_i f(z_i) \right]$$
$$= 2\,\mathrm{Rad}(\mathcal{F}|_S) \quad \text{since } -\varepsilon_i \overset{d}{=} \varepsilon_i.$$

$\boxed{\varepsilon}$

we now return to proof of thm on Rad(F|) for 2-layer ReLU.

Pf Let $f(x;W) = \sum_{j=1}^{m} a_j \varphi(\langle w_j, x\rangle)$.

By defn,

$$\mathrm{URad}(\mathcal{F}_{m,R}) = \mathbb{E}_{\varepsilon}\left[\sup_{\|W\|_F \le R} \sum_{i=1}^{n} \varepsilon_i f(x_i; W)\right]$$

$$= \mathbb{E}_{\varepsilon}\left[\sup_{\|W\|_F \le R} \sum_{i=1}^{n} \varepsilon_i \sum_{j=1}^{m} a_j \varphi(\langle w_j, x_i\rangle)\right]$$

homogeneity

$$= \mathbb{E}_{\varepsilon}\left[\sup_{\|W\|_F \le R} \sum_{j=1}^{m} a_j \|w_j\|_2 \cdot \sum_{i=1}^{n} \varepsilon_i \varphi\left(\left\langle \frac{w_j}{\|w_j\|_2}, x_i\right\rangle\right)\right]$$

$$\le \mathbb{E}_{\varepsilon}\left[\left(\sup_{\|W\|_F \le R} \sum_{j=1}^{m} |a_j|\,\|w_j\|_2\right) \cdot \max_{\substack{j\in[m] \\ \|W\|_F \le R}} \left|\sum_{i=1}^{n} \varepsilon_i \varphi\left(\left\langle \frac{w_j}{\|w_j\|_2}, x_i\right\rangle\right)\right|\right]$$

Cauchy-Schwartz: $\sum_{j=1}^{m} |a_j|\,\|w_j\|_2 = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \|w_j\|_2 \le \sum_{j=1}^{m} \|w_j\|_2^2 = \|W\|_F^2$.

$$\le R \cdot \mathbb{E}_{\varepsilon} \max_{\substack{j\in[m], \\ \|W\|_F \le R}} \left|\sum_{1}^{n} \varepsilon_i \varphi\left(\left\langle \frac{w_j}{\|w_j\|_2}, x_i\right\rangle\right)\right|$$

$$\le R \cdot \mathbb{E}_{\varepsilon} \sup_{\|\bar{w}\|_2 \le 1} \left|\sum_{1}^{n} \varepsilon_i \varphi(\langle \bar{w}, x_i\rangle)\right|$$

$$\le 2R\, \mathbb{E}_{\varepsilon}\left[\sup_{\|\bar{w}\|_2 \le 1} \sum_{1}^{n} \varepsilon_i \varphi(\langle \bar{w}, x_i\rangle)\right]$$

$$= 2R\, \mathrm{URad}\left(\{x \mapsto \varphi(\langle \bar{w}, x\rangle) : \|\bar{w}\|_2 \le 1\}|_S\right)$$

φ is 1-Lip. $\longrightarrow$

$$\le 2R\, \mathrm{URad}\left(\{x \mapsto \langle \bar{w}, x\rangle : \|\bar{w}\|_2 \le 1\}|_S\right)$$

$$\le 2R\, \|X\|_F .$$

$\boxtimes$

So, if we look at 2-layer ReLUs, as long as $\|W\|_F$ is small relative to number of training samples, we can guarantee uniform convergence: empirical risk $\approx$ population