Generalization & uniform convergence

Def A (centered) RV X is σ-subGaussian (or σ-SG; variance proxy σ^2) if

$$\mathbb{E}\exp(\lambda X) \le \exp(\lambda^2\sigma^2/2), \quad \forall \lambda \in \mathbb{R}.$$

Lemma If X is σ-subGaussian, then for any $\varepsilon > 0$,

$$\mathbb{P}(X \ge \varepsilon) \le \exp(-\varepsilon/2\sigma^2).$$

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

- 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 $X_1, \dots, X_n$ 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 $\frac{1}{n^2}\sum_1^n \sigma_i^2$.

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

$\longrightarrow$ 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,

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

$\longrightarrow$ 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,

$Z_i := \mathbb{1}(y_i \neq f_n(x_i))$ 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:

$$L(f) = L(f) - \hat{L}(f) + \hat{L}(f)$$

$$\leq \hat{L}(f) + \sup_{f \in \mathcal{F}} \{ L(f) - \hat{L}(f) \}.$$

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 $(x_i, y_i)_i^n$ are iid, $f_i : \mathbb{R}^d \to \{\pm 1\}$, then SG concentration as before gives for fixed f_ℓ,

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

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

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. wp $\geq 1-\delta$, 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

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

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 $\mathcal{F}$,

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

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

- $\mathrm{URad}(\mathcal{F}_{|S})$ is large if, for any $\varepsilon_i \in \{\pm 1\}$, there is some $f \in \mathcal{F}$ st $f(z_i) \doteq \varepsilon_i$.
- 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 $\mathrm{URad}(\{\pm 1\}^n) = \mathbb{E}_{\varepsilon} \sup_{x \in \{\pm1\}^n} \langle \varepsilon, x\rangle = \mathbb{E}_{\varepsilon} \|\varepsilon\|^2 = n$.

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

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

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

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

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 $> 1-\delta$,

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

③ wp $> 1-\delta$,

$$\sup_{f \in \mathcal{F}} \left\{ \mathbb{E} f(z) - \frac{1}{n} \sum_1^n f(z_i) \right\} \le \frac{2}{n} \mathrm{URad}(\mathcal{F}_{|S}) + 3(b-a) \sqrt{\frac{\log \frac{2}{\delta}}{n}}$$

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

Thm (MacDiarmid). Suppose $F: \mathbb{R}^n \to \mathbb{R}$ satisfies bounded differences: $\forall i \in [n]$, $\exists c_i$ 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 $> 1-\delta$,

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

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

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

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

<u>Pf</u>.

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. $\sup_{f\in\mathcal{F}}\{\mathbb{E}f-\hat{\mathbb{E}}_n f\} \le \mathbb{E}f_\varepsilon-\hat{\mathbb{E}}_n f_\varepsilon+\varepsilon$.

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

$$=\mathbb{E}_n\left[\mathbb{E}_n'\hat{\mathbb{E}}_n' f_\varepsilon-\hat{\mathbb{E}}_n f_\varepsilon+\varepsilon\right]$$

$$=\mathbb{E}_n'\mathbb{E}_n\left[\hat{\mathbb{E}}_n' f_\varepsilon-\hat{\mathbb{E}}_n f_\varepsilon\right]+\varepsilon$$

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

Since $\varepsilon>0$ is arbitrary, this completes the proof. $\boxed{\varepsilon}$