# Lecture 25: concentration / Rademacher

Announcement: HW3 out today

* Concentration
  * "Modern concerns"
  * Hoeffding / sub-Gaussian
  * Rademacher complexity
  * Rademacher for DL
  * Rademacher basics
* Interpolation / double descent

---

**Theorem** (Hoeffding). Suppose $(Z_1 \cdots Z_n)$ given, independent, $Z_i \in [a_i, b_i]$ a.s.,

w/ pr $\geq 1-\delta$, $\quad \mathbb{E}\frac{1}{n}\sum_i Z_i \leq \frac{1}{n}\sum_i Z_i + \sqrt{\frac{\sum_i (b_i - a_i)^2}{2n^2}\ln\frac{1}{\delta}}$.

**Rem.** write $F(Z_1 \cdots Z_n) = \frac{1}{n}\sum_i Z_i$; note this convex & lipschitz in $\vec{Z} = (Z_1, \cdots Z_n)$; a special case of convex Lip theorem.

**Defn.** rv $X$ is $\sigma^2$-sub-Gaussian if $\forall t$ $\ln \mathbb{E}\exp(tX) \leq \frac{t^2\sigma^2}{2}$.

$\boxed{\exists \text{ 4 equiv defns of subG}}$

**Examples.** (last time, but $(X_1 \cdots X_n)$, $(\sigma_1^2 \cdots \sigma_n^2)$-sub-G

$\Rightarrow \frac{1}{n}\sum_i X_i$ is $\frac{\sum_i \sigma_i^2}{n^2}$

"concentration"

Chernoff bounding technique:

**Lemma** (Markov). rv $X \geq 0$ a.s., $a > 0$,

$$\Pr[X \geq a] \leq \frac{\mathbb{E}X}{a}.$$

**Proof.** Note $a\mathbb{1}[X \geq a] \leq X$, apply $\mathbb{E}(\cdot)$ to both sides //

Chernoff bounding technique ($X$ sub-G r.v., possibly negative) $\forall t > 0$

$$\Pr[X \geq \varepsilon] = \Pr[\exp(tX) \geq \exp(t\varepsilon)]$$

$$\Rightarrow \Pr[X \geq \varepsilon] \leq \inf_{t \geq 0} \Pr[\exp(tX) \geq \exp(t\varepsilon)]$$

$$\leq \inf_{t \geq 0} \frac{\mathbb{E}\exp(tX)}{\exp(t\varepsilon)} \leq \inf_{t \geq 0} \exp\left(\frac{t^2\sigma^2}{2} - t\varepsilon\right)$$

$$= \exp\left(\inf_{t \geq 0}\left[\frac{t^2\sigma^2}{2} - t\varepsilon\right]\right) = \exp\left(\frac{\varepsilon^2\sigma^2}{2\sigma^4} - \frac{\varepsilon^2}{\sigma^2}\right)$$

$$= \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

**Rem**

$$\frac{\sum_i X_i}{n} \xrightarrow{a.s.} \mathbb{E}X, \quad \text{SLLN}$$

$$\frac{\sum_i X_i}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{CLT}$$

$$\limsup_{n \to \infty} \left|\frac{\sum_i X_i}{\sqrt{2n \ln\ln n}}\right| = 1 \quad \text{a.s. LIL}$$

One more tail inequality

**Theorem** (McDiarmid). Suppose $(Z_1, \cdots Z_n)$ independent, $Z_i \in S_i$ a.s., let $F: \mathbb{R}^n \to \mathbb{R}$ be given, suppose $\exists (c_1, \cdots c_n)$ s.t. bounded differences hold: $\forall i$

$$\sup_{\substack{Z_1 \cdots Z_n \in \otimes S_i \\ Z_i'}} \left|F(Z_1, \cdots Z_n) - F(Z_1, \cdots Z_i', \cdots Z_n)\right| \leq c_i.$$

Then $F(Z_1 \cdots Z_n)$ is $\frac{\sum c_i^2}{4}$-subGaussian

and w/ pr $\geq 1-\delta$

$$\mathbb{E}F(Z_1, \cdots Z_n) \leq F(Z_1, \cdots Z_n) + \sqrt{\frac{\sum c_i^2}{2n^2}\ln\frac{1}{\delta}}$$ //

# Rademacher complexity

Abstract goal: w/ pr $\geq 1-\delta$, $\underbrace{\forall f \in \mathcal{F}}_{\text{brute force way to handle alg output}}$ $\quad R(f) \leq \hat{R}(f) + \sqrt{\frac{\ln 1/\delta}{n}} + \sqrt{\frac{\text{compl}(\mathcal{F})}{n}}$

We'll get this shortly

<u>Define.</u> Given $V \subseteq \mathbb{R}^n$, $\quad \text{URad}(V) = \mathbb{E}_{\varepsilon} \sup_{u \in V} \langle u, \varepsilon \rangle$

$\varepsilon \overset{iid}{\sim} \bigotimes_{i=1}^{n} \text{Rad}\ ; \quad \varepsilon_i \sim \text{Unif}(\{-1,+1\})$

<u>Rem.</u> ① Others use $\text{Rad}(V) = \frac{1}{n}\text{URad}(V)$

② Typical usage $\quad V = \mathcal{G}_{|Z} = \{ (g(z_1), \dots, g(z_n)) : g \in \mathcal{G} \}$

or $\quad \underbrace{(\ell \circ \mathcal{F})_{|S}}_{\text{"loss class}} = \{ (\ell(y_1, f(x_1)), \dots, \ell(y_n, f(x_n))) : f \in \mathcal{F} \} \subseteq \mathbb{R}^n$

## Theorem (core Rademacher bound).

Let $z_1, \dots, z_n$ iid. Let $\mathcal{G}$ be given with $\mathcal{G}(z_i) \in [a,b]$ a.s.. Then w/ pr $\geq 1-\delta$,

$$\sup_{g \in \mathcal{G}} \left( \mathbb{E}\, g(Z_1) - \frac{1}{n}\sum_i g(Z_i) \right) \leq \frac{2}{n}\text{URad}(\mathcal{G}_{|Z}) + 3(b-a)\sqrt{\frac{\ln 2/\delta}{2n}}.$$

<u>Remarks.</u>

① (History) term coined by Bartlett-Mendelson (2002), but is much older (tied "symmetrization")

② Many weaknesses due to where that proof was cut, e.g., $1/n$ rates

③ Not oldest technique (covering Kolmogorov-Tikhomirov)

④ Non-weakness: note $\text{URad}(\{\pm 1\}^n) = \mathbb{E}_{\varepsilon} \sup_{u \in \{-1,+1\}^n} \langle \varepsilon, u \rangle = n$

Recall DL is a universal approximator

(cf. "rethinking generalization")

(spirit of criticism: DL theorists need to think harder)

easy nonfix: norm bounded'

---

## Rademacher calculus

Sanity checks

$\text{URad}(\{u_0\}) = \mathbb{E}_{\varepsilon} \langle u_0, \varepsilon \rangle = 0$

$\text{URad}(V + \{u_0\}) = \mathbb{E}_{\varepsilon} \sup_{v \in V} \langle \varepsilon, v + u_0 \rangle$

$= \mathbb{E}_{\varepsilon} \langle u_0, \varepsilon \rangle + \mathbb{E}_{\varepsilon} \sup_{v \in V} \langle v, \varepsilon \rangle$

$= \text{URad}(V)$

$V \subseteq V' \Rightarrow \text{URad}(V) \leq \text{URad}(V')$

$\text{URad}(\{\pm 1\}^n) = n, \quad \text{URad}(\{(-1, \dots, -1), (+1, \dots, +1)\})$

$= \mathbb{E}_{\varepsilon} \sqrt{\left| \sum_i \varepsilon_i \right|^2} \in \left[\sqrt{\frac{n}{2}}, \sqrt{n}\right]$

$\text{URad}(V) \geq 0.$

Khintchine inequality

<u>Lemma.</u>

① $\text{URad}(cV + \{u_0\}) = |c| \cdot \text{URad}(V).$

② $\text{URad}(\text{conv}(V)) = \text{URad}(V).$

③ Let $(V_i)_{i \geq 1}$ given w/ $\sup_{u \in V_i} \langle \varepsilon, u \rangle \geq 0 \quad \forall \varepsilon \in \{-1,+1\}^n$

then $\quad \text{URad}\left(\bigcup_{i \geq 1} V_i\right) \leq \sum_{i \geq 1} \text{URad}(V_i).$

<u>Rem.</u> Ⓐ union rule often loose (even exponentially)

Ⓑ Another definition: $\mathbb{E}_{\varepsilon} \sup_{u \in V} |\langle \varepsilon, u \rangle| = \text{URad}(V \cup -V).$

---

<u>Lemma.</u> $\text{URad}(\{ x \mapsto w^T x : \|w - w_0\| \leq R \}_{|S}) \leq \|X\|_F \cdot R.$

<u>Proof.</u> ~~for any $\varepsilon \in \{\pm 1\}^n$,~~ $\cancel{\sup_{\|w - w_0\| \leq R} \langle \varepsilon, w \rangle} = \langle w_0, \varepsilon \rangle + \sup_{\|w - w_0\| \leq R} \langle \varepsilon, w - w_0 \rangle$

~~So~~ $\cancel{\sup_{\|v\| \leq R} \langle \varepsilon, v \rangle}$ $\quad$ Oops, forgot data.

$$\sup_{\|v\| \leq R} \sum_i \varepsilon_i v^T x_i = \sup_{\|v\| \leq R} \left\langle v, \sum_i \varepsilon_i x_i \right\rangle$$

$$= R \left\| \sum_i \varepsilon_i x_i \right\|.$$

$R \frac{\sum_i \varepsilon_i x_i}{\|\sum_i \varepsilon_i x_i\|}$

$$\mathbb{E}_{\varepsilon} \sup_{\|v\| \leq R} \sum_i \varepsilon_i v^T x_i = \mathbb{E}_{\varepsilon} R \left\| \sum_i \varepsilon_i x_i \right\|$$

$$\leq R \sqrt{\mathbb{E}_{\varepsilon} \left\langle \sum_i \varepsilon_i x_i, \sum_i \varepsilon_i x_i \right\rangle}$$

$$= R \sqrt{\mathbb{E}_{\varepsilon} \sum_i \varepsilon_i \|x_i\|^2} = R \sqrt{\sum_i \|x_i\|^2} = R \|X\|_F.$$

//