More Rademacher

CPSC 532S: Modern Statistical Learning Theory
31 January 2022
cs.ubc.ca/~dsuth/532S/22/

Admin

- Back to normal office hours (Tuesday 10-11, Thursday 4-5)
- A1 grading: hopefully by end of the week
- A2 release: hopefully later this week

Admin

- Back to normal office hours (Tuesday 10-11, Thursday 4-5)
- A1 grading: hopefully by end of the week
- A2 release: hopefully later this week

- The website for SSBD is currently down (???)
  - If you don’t have the pdf saved, I put it on the Canvas Files section

Admin

- Back to normal office hours (Tuesday 10-11, Thursday 4-5)
- A1 grading: hopefully by end of the week
- A2 release: hopefully later this week

homogeneous

$\mathbb{1}(\langle w, x \rangle \geq 0)$

non-

$\mathbb{1}(\langle w, x \rangle + b \geq 0)$

- The website for SSBD is currently down (???)
  - If you don't have the pdf saved, I put it on the Canvas Files section

- Note on VC dim of homogenous linear classifiers: SSBD are being kinda sloppy
  - Radon's Theorem implies non-homogeneous halfspaces can't shatter size $d + 2$
    - (Basically the same proof as we talked about)
  - Can do a reduction between homogeneous in $\mathbb{R}^d$ and non-homog in $\mathbb{R}^{d-1}$

Forever ago: "Fundamental Theorem of Learning"

For binary classification with 0-1 loss:

These are all equivalent:

1. $\mathcal{H}$ has the uniform convergence property
2. Any ERM rule agnostically PAC learns $\mathcal{H}$
3. $\mathcal{H}$ is agnostic PAC learnable
4. Any ERM rule PAC learns $\mathcal{H}$
5. $\mathcal{H}$ is PAC learnable
6. $\mathrm{VCdim}(\mathcal{H}) < \infty$

saw

immediate

no free lunch

to show!

- If $\mathrm{VCdim}(\mathcal{H}) = d$:
  - $\mathcal{H}$ has uniform convergence property, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \leq \boxed{n_{\mathcal{H}}^{UC} \leq \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]}$
  - $\mathcal{H}$ is agnostic PAC learnable, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \leq n_{\mathcal{H}} \leq \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]$
  - $\mathcal{H}$ is PAC learnable, $\frac{C_1}{\varepsilon}\left[d + \log\frac{1}{\delta}\right] \leq n_{\mathcal{H}} \leq \frac{C_2}{\varepsilon}\left[d\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right]$

Last time: Rademacher complexity

- The **empirical Rademacher complexity** of $\mathscr{G}$ on a set $S = (z_1, \ldots, z_n)$ is

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\substack{\boldsymbol{\sigma} \\ \boldsymbol{\sigma} \sim \mathrm{Rad}^n}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right]$$

$\sigma_i \sim \mathrm{Rad}$ means $\Pr(\sigma_i = -1) = \frac{1}{2} = \Pr(\sigma_i = 1)$

Last time: Rademacher complexity

- The **empirical Rademacher complexity** of $\mathscr{G}$ on a set $S = (z_1, \ldots, z_n)$ is

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\substack{\boldsymbol{\sigma} \\ \boldsymbol{\sigma} \sim \mathrm{Rad}^n}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{g}_S \right] \qquad \mathbf{g}_S = \big(g(z_1), \ldots, g(z_n) \big)$$

"how well can functions from $\mathscr{G}$ correlate with random noise?"

$\sigma_i \sim \mathrm{Rad}$ means $\Pr(\sigma_i = -1) = \frac{1}{2} = \Pr(\sigma_i = 1)$

$\mathscr{G}$ has function $g : \mathcal{X} \to \{-1, 1\}$

$$\sup_{g \in \mathscr{G}} \frac{1}{n} \sum_i \sigma_i \, g(z_i)$$

σ_i | $g(z_i)$
+1 | +1
-1 | -1 → 1

+1 | -1 → -1
-1 | +1

Last time: Rademacher complexity

- The **empirical Rademacher complexity** of $\mathscr{G}$ on a set $S = (z_1, \ldots, z_n)$ is

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\substack{\boldsymbol{\sigma} \\ \boldsymbol{\sigma} \sim \mathrm{Rad}^n}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \boldsymbol{\sigma}^{\mathsf{T}} \mathbf{g}_S \right] \qquad \mathbf{g}_S = \big(g(z_1), \ldots, g(z_n)\big)$$

"how well can functions from $\mathscr{G}$ correlate with random noise?"

$\sigma_i \sim \mathrm{Rad}$ means $\Pr(\sigma_i = -1) = \frac{1}{2} = \Pr(\sigma_i = 1)$

- The ("average-case") **Rademacher complexity** is just $\mathfrak{R}_n(\mathscr{G}) = \mathbb{E}_{S \sim \mathscr{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathscr{G})\right]$

Last time: Rademacher complexity

- The **empirical Rademacher complexity** of $\mathscr{G}$ on a set $S = (z_1, \ldots, z_n)$ is

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\substack{\boldsymbol{\sigma} \\ \boldsymbol{\sigma} \sim \mathrm{Rad}^n}}\left[\sup_{g \in \mathscr{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathscr{G}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{g}_S\right] \qquad \mathbf{g}_S = \big(g(z_1), \ldots, g(z_n)\big)$$

$\sigma_i \sim \mathrm{Rad}$ means $\Pr(\sigma_i = -1) = \frac{1}{2} = \Pr(\sigma_i = 1)$

"how well can functions from $\mathscr{G}$ correlate with random noise?"

- The ("average-case") **Rademacher complexity** is just $\mathfrak{R}_n(\mathscr{G}) = \mathbb{E}_{S \sim \mathscr{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathscr{G})\right]$

- Take $\mathfrak{R}_n(\mathscr{G})$ for $\mathscr{G} = \left\{z \mapsto \ell(h, z) : h \in \mathscr{H}\right\}$
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathscr{G}) = \frac{1}{2}\widehat{\mathfrak{R}}_{S|_x}(\mathscr{H})$

$$\frac{1}{n}\sum_i \sigma_i g(z_i) = \frac{1}{n}\sum_{i:\sigma_i=1} \ell(h, z_i) - \frac{1}{n}\sum_{i:\sigma_i=-1} \ell(h, z_i)$$

$$\approx \frac{1}{2} L_{S_+}(h) - \frac{1}{2} L_{S_-}(h) = \frac{1}{2}\left[L_{S_+}(h) - L_{S_-}(h)\right]$$

$$\underset{ish}{\approx} \frac{1}{2}\left[L_{\mathcal{D}}(h) - L_S(h)\right]$$

$$S_+ = \{z_i \in S : \sigma_i = 1\}$$

$$\frac{1}{n}\sum \sigma_i g(z_i) = \frac{1}{n}\sum_{i:\sigma_i=1} \sigma_i g(z_i) + \frac{1}{n}\sum_{i:\sigma_i=-1} \overbrace{\sigma_i}^{-1} g(z_i)$$

Last time: Rademacher complexity

- The **empirical Rademacher complexity** of $\mathscr{G}$ on a set $S = (z_1, \ldots, z_n)$ is

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\substack{\boldsymbol{\sigma} \\ \boldsymbol{\sigma} \sim \mathrm{Rad}^n}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \boldsymbol{\sigma}^{\mathsf{T}} \mathbf{g}_S \right] \qquad \mathbf{g}_S = \big(g(z_1), \ldots, g(z_n)\big)$$

"how well can functions from $\mathscr{G}$ correlate with random noise?"

$\sigma_i \sim \mathrm{Rad}$ means $\Pr(\sigma_i = -1) = \frac{1}{2} = \Pr(\sigma_i = 1)$

- The ("average-case") **Rademacher complexity** is just $\mathfrak{R}_n(\mathscr{G}) = \mathbb{E}_{S \sim \mathscr{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathscr{G})\right]$

- Take $\mathfrak{R}_n(\mathscr{G})$ for $\mathscr{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathscr{H} \right\}$
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathscr{G}) = \frac{1}{4}\widehat{\mathfrak{R}}_{S|_x}(\mathscr{H})$
- Binary clf: $\mathfrak{R}_n(\mathscr{H}) \leq \sqrt{2 \ \mathrm{VCdim}(\mathscr{H}) \ \log(n) / n}$ $(n \geq d \geq 3$; **Massart, Sauer-Shelah**)

Last time: Rademacher complexity

- The **empirical Rademacher complexity** of $\mathscr{G}$ on a set $S = (z_1, \ldots, z_n)$ is

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathop{\mathbb{E}}_{\substack{\boldsymbol{\sigma} \\ \boldsymbol{\sigma} \sim \mathrm{Rad}^n}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \boldsymbol{\sigma}^\mathsf{T} \mathbf{g}_S \right] \qquad \mathbf{g}_S = \big(g(z_1), \ldots, g(z_n)\big)$$

$\sigma_i \sim \mathrm{Rad}$ means $\Pr(\sigma_i = -1) = \frac{1}{2} = \Pr(\sigma_i = 1)$

"how well can functions from $\mathscr{G}$ correlate with random noise?"

- The ("average-case") **Rademacher complexity** is just $\mathfrak{R}_n(\mathscr{G}) = \mathbb{E}_{S \sim \mathscr{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathscr{G})\right]$

- Take $\mathfrak{R}_n(\mathscr{G})$ for $\mathscr{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathscr{H} \right\}$
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathscr{G}) = \frac{1}{4}\widehat{\mathfrak{R}}_{S|_x}(\mathscr{H})$
- Binary clf: $\mathfrak{R}_n(\mathscr{H}) \leq \sqrt{2 \ \mathrm{VCdim}(\mathscr{H}) \ \log(n)/n}$ $\quad (n \geq d \geq 3;$ **Massart, Sauer-Shelah**)
- **Theorem**: if $\mathscr{G}$ maps to $[0,1]$, $\displaystyle \sup_{g \in \mathscr{G}} \left| \mathbb{E}[g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right| \leq 2\mathfrak{R}_n(\mathscr{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$

Last time: Rademacher complexity

- The **empirical Rademacher complexity** of $\mathscr{G}$ on a set $S = (z_1, \ldots, z_n)$ is

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\substack{\boldsymbol{\sigma} \\ \boldsymbol{\sigma} \sim \mathrm{Rad}^n}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \boldsymbol{\sigma}^\mathsf{T} \mathbf{g}_S \right] \qquad \mathbf{g}_S = \big(g(z_1), \ldots, g(z_n)\big)$$

"how well can functions from $\mathscr{G}$ correlate with random noise?"

$\sigma_i \sim \mathrm{Rad}$ means $\Pr(\sigma_i = -1) = \frac{1}{2} = \Pr(\sigma_i = 1)$

- The ("average-case") **Rademacher complexity** is just $\mathfrak{R}_n(\mathscr{G}) = \mathbb{E}_{S \sim \mathscr{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathscr{G})\right]$

- Take $\mathfrak{R}_n(\mathscr{G})$ for $\mathscr{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathscr{H} \right\}$
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathscr{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathscr{H})$
- Binary clf: $\mathfrak{R}_n(\mathscr{H}) \leq \sqrt{2 \ \mathrm{VCdim}(\mathscr{H}) \ \log(n) / n}$ ($n \geq d \geq 3$; **Massart, Sauer-Shelah**)
- **Theorem**: if $\mathscr{G}$ maps to $[0,1]$, $\displaystyle \sup_{g \in \mathscr{G}} \left| \underbrace{\mathbb{E}[g(z)]}_{L_{\mathscr{D}}(h)} - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right| \leq 2\mathfrak{R}_n(\mathscr{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$

Last time: Rademacher complexity

- The **empirical Rademacher complexity** of $\mathscr{G}$ on a set $S = (z_1, \ldots, z_n)$ is

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\substack{\boldsymbol{\sigma} \\ \boldsymbol{\sigma} \sim \mathrm{Rad}^n}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{g}_S \right] \qquad \mathbf{g}_S = \big(g(z_1), \ldots, g(z_n)\big)$$

"how well can functions from $\mathscr{G}$ correlate with random noise?"

$\sigma_i \sim \mathrm{Rad}$ means $\Pr(\sigma_i = -1) = \frac{1}{2} = \Pr(\sigma_i = 1)$

- The ("average-case") **Rademacher complexity** is just $\mathfrak{R}_n(\mathscr{G}) = \mathbb{E}_{S \sim \mathscr{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathscr{G})\right]$

- Take $\mathfrak{R}_n(\mathscr{G})$ for $\mathscr{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathscr{H} \right\}$
  - For 0-1 loss, $g_h((x, y)) = \mathbb{1}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathscr{G}) = \frac{1}{4} \widehat{\mathfrak{R}}_{S|_x}(\mathscr{H})$
- Binary clf: $\mathfrak{R}_n(\mathscr{H}) \leq \sqrt{2 \ \mathrm{VCdim}(\mathscr{H}) \ \log(n) / n}$ $(n \geq d \geq 3$; **Massart, Sauer-Shelah**)
- **Theorem**: if $\mathscr{G}$ maps to $[0,1]$, $\displaystyle \sup_{g \in \mathscr{G}} \Big| \underbrace{\mathbb{E}[g(z)]}_{L_{\mathscr{D}}(h)} - \underbrace{\frac{1}{n} \sum_{i=1}^{n} g(z_i)}_{L_S(h)} \Big| \leq 2\mathfrak{R}_n(\mathscr{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$

Last time: Rademacher complexity

- The **empirical Rademacher complexity** of $\mathcal{G}$ on a set $S = (z_1, \ldots, z_n)$ is

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\substack{\boldsymbol{\sigma} \\ \boldsymbol{\sigma} \sim \mathrm{Rad}^n}} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{g}_S \right] \qquad \mathbf{g}_S = \big(g(z_1), \ldots, g(z_n) \big)$$

"how well can functions from $\mathcal{G}$ correlate with random noise?"

$\sigma_i \sim \mathrm{Rad}$ means $\Pr(\sigma_i = -1) = \frac{1}{2} = \Pr(\sigma_i = 1)$

- The ("average-case") **Rademacher complexity** is just $\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{S \sim \mathcal{D}^n} \left[\widehat{\mathfrak{R}}_S(\mathcal{G}) \right]$

- Take $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathcal{H} \right\}$
  - For 0-1 loss, $g_h((x, y)) = \mathbb{1}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{4} \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$
- Binary clf: $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \ \mathrm{VCdim}(\mathcal{H}) \ \log(n) / n}$ ($n \geq d \geq 3$; **Massart, Sauer-Shelah**)
- **Theorem**: if $\mathcal{G}$ maps to $[0,1]$, $\displaystyle \sup_{g \in \mathcal{G}} \left| \underbrace{\mathbb{E}[g(z)]}_{L_{\mathcal{D}}(h)} - \underbrace{\frac{1}{n} \sum_{i=1}^{n} g(z_i)}_{L_S(h)} \right| \leq \underbrace{2 \mathfrak{R}_n(\mathcal{G})}_{\substack{\text{bounds} \\ \mathbb{E} \sup |L_{\mathcal{D}}(h) - L_S(h)|}} + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$

Last time: Rademacher complexity

- The **empirical Rademacher complexity** of $\mathscr{G}$ on a set $S = (z_1, \ldots, z_n)$ is

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\substack{\boldsymbol{\sigma} \\ \boldsymbol{\sigma} \sim \mathrm{Rad}^n}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(z_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g \in \mathscr{G}} \frac{1}{n} \boldsymbol{\sigma}^\mathsf{T} \mathbf{g}_S \right] \qquad \mathbf{g}_S = \big(g(z_1), \ldots, g(z_n)\big)$$

"how well can functions from $\mathscr{G}$ correlate with random noise?"

$\sigma_i \sim \mathrm{Rad}$ means $\Pr(\sigma_i = -1) = \frac{1}{2} = \Pr(\sigma_i = 1)$

- The ("average-case") **Rademacher complexity** is just $\mathfrak{R}_n(\mathscr{G}) = \mathbb{E}_{S \sim \mathscr{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathscr{G})\right]$

- Take $\mathfrak{R}_n(\mathscr{G})$ for $\mathscr{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathscr{H} \right\}$
  - For 0-1 loss, $g_h((x, y)) = \mathbb{1}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathscr{G}) = \frac{1}{4}\widehat{\mathfrak{R}}_{S|_x}(\mathscr{H})$
- Binary clf: $\mathfrak{R}_n(\mathscr{H}) \leq \sqrt{2 \ \mathrm{VCdim}(\mathscr{H}) \ \log(n) / n}$ $(n \geq d \geq 3$; **Massart, Sauer-Shelah**)
- **Theorem**: if $\mathscr{G}$ maps to $[0,1]$, $\sup_{g \in \mathscr{G}} \left| \underbrace{\mathbb{E}[g(z)]}_{L_{\mathscr{D}}(h)} - \underbrace{\frac{1}{n} \sum_{i=1}^n g(z_i)}_{L_S(h)} \right| \leq \underbrace{2\mathfrak{R}_n(\mathscr{G})}_{\text{bounds } \mathbb{E} \sup |L_{\mathscr{D}}(h) - L_S(h)|} + \underbrace{\sqrt{\frac{1}{2n} \log \frac{1}{\delta}}}_{\text{how much bigger than its mean is likely}}$

$\mathfrak{R}$ of binary classifiers vs loss

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in\mathscr{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\, \mathbb{I}(h(x_i) \neq y_i)\Big]$$

$\mathfrak{R}$ of binary classifiers vs loss

If $h : \mathscr{X} \to \{0,1\}$, define $\tilde{h} : \mathscr{X} \to \{-1,1\}$ by $\tilde{h}(x) = 2h(x) - 1$; $\quad \tilde{y}_i = 2y_i - 1$

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in\mathscr{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\, \mathbb{I}(h(x_i) \neq y_i)\Big]$$

$\mathfrak{R}$ of binary classifiers vs loss

If $h : \mathcal{X} \to \{0,1\}$, define $\tilde{h} : \mathcal{X} \to \{-1,1\}$ by $\tilde{h}(x) = 2h(x) - 1$; $\quad \tilde{y}_i = 2y_i - 1$

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\, \mathbb{1}(h(x_i) \neq y_i)\Big] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1 - \tilde{y}_i\tilde{h}(x_i)}{2}\right]$$

$\mathfrak{R}$ of binary classifiers vs loss

If $h : \mathcal{X} \to \{0,1\}$, define $\tilde{h} : \mathcal{X} \to \{-1,1\}$ by $\tilde{h}(x) = 2h(x) - 1$; $\tilde{y}_i = 2y_i - 1$

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{I}(h(x_i) \neq y_i)\Big] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1 - \tilde{y}_i\tilde{h}(x_i)}{2}\right]$$

Adding constants doesn't change $\widehat{\mathfrak{R}}_S$:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top(\mathbf{g}_S + c\mathbf{1})\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right) + c\,\boldsymbol{\sigma}^\top\mathbf{1}\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right]$$

$\mathfrak{R}$ of binary classifiers vs loss

If $h : \mathscr{X} \to \{0,1\}$, define $\tilde{h} : \mathscr{X} \to \{-1,1\}$ by $\tilde{h}(x) = 2h(x) - 1$; $\quad \tilde{y}_i = 2y_i - 1$

$$\widehat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in\mathscr{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\, \mathbb{I}(h(x_i) \neq y_i)\Big] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathscr{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1-\tilde{y}_i\tilde{h}(x_i)}{2}\right]$$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathscr{H}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i\tilde{y}_i\tilde{h}(x_i)\right]$$

Adding constants doesn't change $\widehat{\mathfrak{R}}_S$:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \boldsymbol{\sigma}^\top(\mathbf{g}_S + c\mathbf{1})\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\sup_{g\in\mathscr{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right) + c\,\boldsymbol{\sigma}^\top\mathbf{1}\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right]$$

$\mathfrak{R}$ of binary classifiers vs loss

If $h : \mathcal{X} \to \{0,1\}$, define $\tilde{h} : \mathcal{X} \to \{-1,1\}$ by $\tilde{h}(x) = 2h(x) - 1$; $\quad \tilde{y}_i = 2y_i - 1$

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \,\mathbb{I}(h(x_i) \neq y_i)\Big] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1 - \tilde{y}_i\tilde{h}(x_i)}{2}\right]$$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i\tilde{y}_i\tilde{h}(x_i)\right]$$

for fixed $\tilde{y}_i$: $\quad -\sigma_i\tilde{y}_i \sim \mathrm{Rad}$

Adding constants doesn't change $\widehat{\mathfrak{R}}_S$:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top(\mathbf{g}_S + c\mathbf{1})\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right) + c\,\boldsymbol{\sigma}^\top\mathbf{1}\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right]$$

$\mathfrak{R}$ of binary classifiers vs loss

If $h : \mathcal{X} \to \{0,1\}$, define $\tilde{h} : \mathcal{X} \to \{-1,1\}$ by $\tilde{h}(x) = 2h(x) - 1;\qquad \tilde{y}_i = 2y_i - 1$

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\, \mathbb{I}(h(x_i) \neq y_i)\Big] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1 - \tilde{y}_i\tilde{h}(x_i)}{2}\right]$$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i\tilde{y}_i\tilde{h}(x_i)\right]$$

for fixed $\tilde{y}_i$: $\quad -\sigma_i\tilde{y}_i \sim \mathrm{Rad}$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\tilde{h}(x_i)\right]$$

Adding constants doesn't change $\widehat{\mathfrak{R}}_S$:

$$\mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top(\mathbf{g}_S + c\mathbf{1})\right] = \mathbb{E}_{\sigma}\left[\left(\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right) + c\,\boldsymbol{\sigma}^\top\mathbf{1}\right] = \mathbb{E}_{\sigma}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right]$$

$\mathfrak{R}$ of binary classifiers vs loss

If $h : \mathcal{X} \to \{0,1\}$, define $\tilde{h} : \mathcal{X} \to \{-1,1\}$ by $\tilde{h}(x) = 2h(x) - 1$; $\quad \tilde{y}_i = 2y_i - 1$

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}(h(x_i) \neq y_i)\Big] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1 - \tilde{y}_i\tilde{h}(x_i)}{2}\right]$$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i\tilde{y}_i\tilde{h}(x_i)\right]$$

for fixed $\tilde{y}_i$: $\quad -\sigma_i\tilde{y}_i \sim \mathrm{Rad}$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\tilde{h}(x_i)\right] = \frac{1}{2}\widehat{\mathfrak{R}}_{S|_x}(\tilde{\mathcal{H}})$$

Adding constants doesn't change $\widehat{\mathfrak{R}}_S$:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top(\mathbf{g}_S + c\mathbf{1})\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right) + c\,\boldsymbol{\sigma}^\top\mathbf{1}\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right]$$

$\mathfrak{R}$ of binary classifiers vs loss

If $h : \mathcal{X} \to \{0,1\}$, define $\tilde{h} : \mathcal{X} \to \{-1,1\}$ by $\tilde{h}(x) = 2h(x) - 1$; $\quad \tilde{y}_i = 2y_i - 1$

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \, \mathbb{1}(h(x_i) \neq y_i)\Big] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1 - \tilde{y}_i\tilde{h}(x_i)}{2}\right]$$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i\tilde{y}_i\tilde{h}(x_i)\right]$$

for fixed $\tilde{y}_i$: $\quad -\sigma_i\tilde{y}_i \sim \mathrm{Rad}$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\tilde{h}(x_i)\right] = \frac{1}{2}\widehat{\mathfrak{R}}_{S|_x}(\tilde{\mathcal{H}})$$

Adding constants doesn't change $\widehat{\mathfrak{R}}_S$:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top(\mathbf{g}_S + c\mathbf{1})\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right) + c\,\boldsymbol{\sigma}^\top\mathbf{1}\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right]$$

Scaling by c gives $|c|\,\widehat{\mathfrak{R}}_S$:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top(c\mathbf{g}_S)\right] = |c|\,\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \operatorname{sign}(c)\,\boldsymbol{\sigma}^\top\mathbf{g}_S\right] = |c|\,\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top\mathbf{g}_S\right]$$

$\mathfrak{R}$ of binary classifiers vs loss

If $h : \mathcal{X} \to \{0,1\}$, define $\tilde{h} : \mathcal{X} \to \{-1,1\}$ by $\tilde{h}(x) = 2h(x) - 1$; $\quad \tilde{y}_i = 2y_i - 1$

$$\widehat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\, \mathbb{I}(h(x_i) \neq y_i)\Big] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \frac{1-\tilde{y}_i\tilde{h}(x_i)}{2}\right]$$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i\tilde{y}_i\tilde{h}(x_i)\right]$$

for fixed $\tilde{y}_i$: $\quad -\sigma_i\tilde{y}_i \sim \mathrm{Rad}$

$$= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i\tilde{h}(x_i)\right] = \frac{1}{2}\widehat{\mathfrak{R}}_{S|_x}(\tilde{\mathcal{H}}) = \frac{1}{4}\widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$$

Adding constants doesn't change $\widehat{\mathfrak{R}}_S$:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top(\mathbf{g}_S + c\mathbf{1})\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top \mathbf{g}_S\right) + c\,\boldsymbol{\sigma}^\top\mathbf{1}\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

Scaling by c gives $|c|\,\widehat{\mathfrak{R}}_S$:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top(c\mathbf{g}_S)\right] = |c|\,\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \operatorname{sign}(c)\,\boldsymbol{\sigma}^\top \mathbf{g}_S\right] = |c|\,\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

To start: a "low-probability" bound

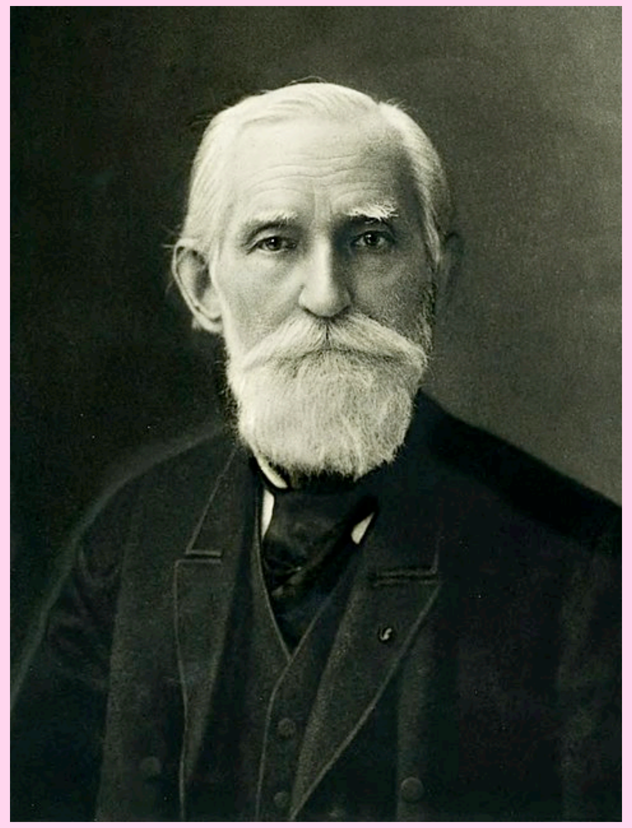
- We'll show $\mathbb{E} \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right| \leq 2\mathfrak{R}_n(\mathcal{G})$ (for any $\mathcal{G}$; don't need bounded)

To start: a "low-probability" bound

- We'll show $\mathbb{E} \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right| \leq 2\mathfrak{R}_n(\mathcal{G})$ (for any $\mathcal{G}$; don't need bounded)
- Immediately implies by Markov's inequality that $\sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right| \leq \frac{2}{\delta}\mathfrak{R}_n(\mathcal{G})$ with prob at least $1 - \delta$
  - Pretty similar to SSBD 6.11, but more general + better bound

Markov's inequality
(~1860s)

Andrey Markov



Pafnuty Chebyshev

If $\Pr(X \geq 0) = 1$, then $\Pr\left(X > \frac{1}{\delta}\,\mathbb{E}[X]\right) \leq \delta$

Proof: take $a = \frac{1}{\delta}\mathbb{E}X$ in:

$a\mathbb{1}_{[X \geq a]} \leq X$, so $\mathbb{E}[a\mathbb{1}_{[X \geq a]}] = a\Pr(X \geq a) \leq \mathbb{E}X$

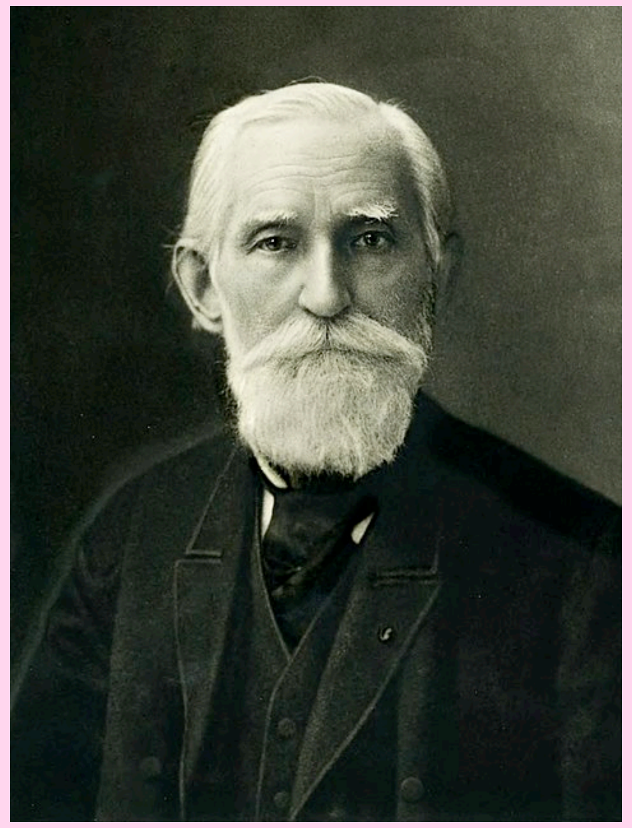
To start: a “low-probability” bound

- We’ll show $\mathbb{E} \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right| \leq 2\mathfrak{R}_n(\mathcal{G})$ (for any $\mathcal{G}$; don’t need bounded)
- Immediately implies by Markov’s inequality that $\sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right| \leq \frac{2}{\delta}\mathfrak{R}_n(\mathcal{G})$ with prob at least $1 - \delta$
  - Pretty similar to SSBD 6.11, but more general + better bound

- Will get a better bound with McDiarmid’s inequality after

Markov’s inequality
(~1860s)

Andrey Markov



Pafnuty Chebyshev

If $\Pr(X \geq 0) = 1$, then $\Pr\left(X > \frac{1}{\delta}\,\mathbb{E}[X]\right) \leq \delta$

Proof: take $a = \frac{1}{\delta}\mathbb{E}X$ in:

$a\mathbb{1}_{[X \geq a]} \leq X$, so $\mathbb{E}[a\mathbb{1}_{[X \geq a]}] = a\Pr(X \geq a) \leq \mathbb{E}X$

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\underbrace{\mathbb{E}[g]} - \underbrace{\hat{\mathbb{E}}_S[g]}| \leq 2\mathfrak{R}_n(\mathcal{G})$

$$\mathbb{E} g = \mathop{\mathbb{E}}_{z \sim D} g(z) \approx L_D(h)$$

$$\hat{\mathbb{E}}_S g = \frac{1}{n} \sum_{i=1}^{n} g(z_i) \approx L_S(h)$$

$$\hat{\mathbb{E}}_S g = \mathbb{E}_{\hat{D}} g$$

$$\int g(z)\, d(D - \hat{D})$$

$$= \int g(z)\, dD(z) - \int g(z)\, d\hat{D}(z)$$

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$

$S \sim \mathcal{D}^n$

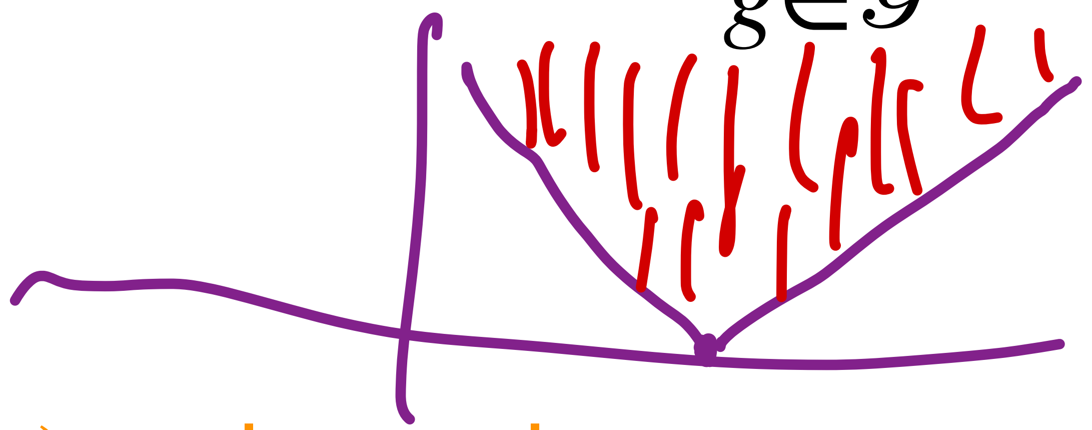
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
    - Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| = \mathbb{E}_S \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g]\right] - \hat{\mathbb{E}}_S[g] \right|$

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
  - Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| = \mathbb{E}_S \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g]\right] - \hat{\mathbb{E}}_S[g] \right|$
  $$\leq \mathbb{E}_S \sup_{g \in \mathcal{G}} \mathbb{E}_{S'} \left| \hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S[g] \right|$$

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
  - Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| = \mathbb{E}_S \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g]\right] - \hat{\mathbb{E}}_S[g] \right|$

$$\leq \mathbb{E}_S \sup_{g \in \mathcal{G}} \mathbb{E}_{S'} \left| \hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S[g] \right|$$



$f(x) = |x - y|$ is convex
so $|\mathbb{E}X - y| \leq \mathbb{E}|X - y|$

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
  - Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| = \mathbb{E}_S \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g]\right] - \hat{\mathbb{E}}_S[g] \right|$

$$\leq \mathbb{E}_S \sup_{g \in \mathcal{G}} \mathbb{E}_{S'} \left| \hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S[g] \right|$$

$f(x) = |x - y|$ is convex
so $|\mathbb{E}X - y| \leq \mathbb{E}|X - y|$

Jensen's inequality (1906)

If f is convex, $f(\mathbb{E}X) \leq \mathbb{E}f(X)$.

Johan Ludwig William Valdemar Jensen

- Want $\mathbb{E}\sup_{g\in\mathcal{G}}|\mathbb{E}[g]-\hat{\mathbb{E}}_S[g]| \le 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S'\sim\mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S'=(z_1',\ldots,z_n')$
  - Then $\mathbb{E}_S\sup_{g\in\mathcal{G}}|\mathbb{E}[g]-\hat{\mathbb{E}}_S[g]| = \mathbb{E}_S\sup_{g\in\mathcal{G}}\left|\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g]\right]-\hat{\mathbb{E}}_S[g]\right|$

$$\le \mathbb{E}_S\sup_{g\in\mathcal{G}}\mathbb{E}_{S'}\left|\hat{\mathbb{E}}_{S'}[g]-\hat{\mathbb{E}}_S[g]\right|$$

$f(x)=|x-y|$ is convex
so $|\mathbb{E}X-y| \le \mathbb{E}|X-y|$

Jensen's inequality (1906)

Johan Ludwig William Valdemar Jensen

If f is convex, $f(\mathbb{E}X) \le \mathbb{E}f(X)$.

To remember the direction:
X^2 is convex, and $[\mathbb{E}X]^2 \le \mathbb{E}[X^2]$

- Want $\mathbb{E} \sup_{g \in \mathscr{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathscr{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathscr{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
  - Then $\mathbb{E}_S \sup_{g \in \mathscr{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| = \mathbb{E}_S \sup_{g \in \mathscr{G}} \left| \mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g]\right] - \hat{\mathbb{E}}_S[g] \right|$

$$\leq \mathbb{E}_S \sup_{g \in \mathscr{G}} \mathbb{E}_{S'} \left| \hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S[g] \right|$$

$f(x) = |x - y|$ is convex
so $|\mathbb{E}X - y| \leq \mathbb{E}|X - y|$

Jensen's inequality (1906)

Johan Ludwig William Valdemar Jensen

If f is convex, $f(\mathbb{E}X) \leq \mathbb{E}f(X)$.

It is easily seen that the subdifferential is linear in y[citation needed] (that is false and the assertion requires Hahn-Banach theorem to be proved)

To remember the direction:
X^2 is convex, and $[\mathbb{E}X]^2 \leq \mathbb{E}[X^2]$

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
  - Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| = \mathbb{E}_S \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g]\right] - \hat{\mathbb{E}}_S[g] \right|$

$$\leq \mathbb{E}_S \sup_{g \in \mathcal{G}} \mathbb{E}_{S'} \left| \hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S[g] \right|$$

- Want $\mathbb{E}\sup_{g\in\mathscr{G}}|\mathbb{E}[g]-\hat{\mathbb{E}}_S[g]| \le 2\mathfrak{R}_n(\mathscr{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S'\sim\mathscr{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
  - Then $\mathbb{E}_S\sup_{g\in\mathscr{G}}|\mathbb{E}[g]-\hat{\mathbb{E}}_S[g]| = \mathbb{E}_S\sup_{g\in\mathscr{G}}\left|\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g]\right]-\hat{\mathbb{E}}_S[g]\right|$

$$\le \mathbb{E}_S\sup_{g\in\mathscr{G}}\mathbb{E}_{S'}\left|\hat{\mathbb{E}}_{S'}[g]-\hat{\mathbb{E}}_S[g]\right|$$

$$\le \mathbb{E}_S\mathbb{E}_{S'}\sup_{g\in\mathscr{G}}|\hat{\mathbb{E}}_{S'}[g]-\hat{\mathbb{E}}_S(g)|$$

Always have $\sup_y \mathbb{E}_X[f_y(X)] \le \mathbb{E}_X\left[\sup_y f_y(X)\right]$:

$f_y(X) \le \sup_y f_y(X)$ for each y

so $\mathbb{E}_X f_y(X) \le \mathbb{E}_X \sup_y f_y(X)$ for each y

so can take sup of LHS and still true

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
  - Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| = \mathbb{E}_S \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g]\right] - \hat{\mathbb{E}}_S[g] \right|$

$$\leq \mathbb{E}_S \sup_{g \in \mathcal{G}} \mathbb{E}_{S'} \left| \hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S[g] \right|$$

$$\leq \mathbb{E}_S \mathbb{E}_{S'} \sup_{g \in \mathcal{G}} |\hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_S(g)|$$

Always have $\sup_y \mathbb{E}_X[f_y(X)] \leq \mathbb{E}_X \left[\sup_y f_y(X) \right]$:

$$= \mathbb{E}_{S,S'} \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n \left[g(z_i') - g(z_i) \right] \right|$$

$f_y(X) \leq \sup_y f_y(X)$ for each y

so $\mathbb{E}_X f_y(X) \leq \mathbb{E}_X \sup_y f_y(X)$ for each y

so can take sup of LHS and still true

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
    - Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq \mathbb{E}_{S,S'} \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^{n} \left[g(z_i') - g(z_i)\right] \right|$

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
  - Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq \mathbb{E}_{S,S'} \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^{n} \left[g(z_i') - g(z_i)\right] \right|$

arbitrary $\sigma_i \in \{-1, 1\}$

$$= \mathbb{E}_{S,S'} \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \left[g(z_i') - g(z_i)\right] \right|$$

Trick called **symmetrization**:
it doesn't matter if we swap
z_i and z_i', since everything is iid
and only looked at once

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z_1', \ldots, z_n')$
  - Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq \mathbb{E}_{S,S'} \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^{n} \left[g(z_i') - g(z_i)\right] \right|$

arbitrary $\sigma_i \in \{-1,1\}$

$$\overset{\mathbb{E}_\sigma}{=} \mathbb{E}_{S,S'} \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \left[g(z_i') - g(z_i)\right] \right| = \mathbb{E}_{S,S'} \mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \left[g(z_i') - g(z_i)\right] \right|$$

Trick called **symmetrization**:
it doesn't matter if we swap
z_i and z_i', since everything is iid
and only looked at once

σ_i now iid $\mathrm{Uniform}(\{-1,1\})$
(**Rademacher** distribution)

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z'_1, \ldots, z'_n)$
- Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq \mathbb{E}_{S,S'} \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n \left[g(z'_i) - g(z_i)\right] \right|$

arbitrary $\sigma_i \in \{-1,1\}$

$$= \mathbb{E}_{S,S'} \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \left[g(z'_i) - g(z_i)\right] \right| = \mathbb{E}_{S,S'} \mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \left[g(z'_i) - g(z_i)\right] \right|$$

Trick called **symmetrization**:
it doesn't matter if we swap
z_i and z'_i, since everything is iid
and only looked at once

σ_i now iid $\text{Uniform}(\{-1,1\})$
(**Rademacher** distribution)

$\sup \left| \frac{1}{n} \sum \sigma_i g(z'_i) - \frac{1}{n} \sum \sigma_i g(z_i) \right|$
$\leq \sup \left| \frac{1}{n} \sum \sigma_i g(z'_i) \right| + \left| \frac{1}{n} \sum \sigma_i g(z_i) \right|$

$$\leq \mathbb{E}_{S,S'} \mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g(z'_i) \right| + \mathbb{E}_{S,S'} \mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n -\sigma_i g(z_i) \right|$$

- Want $\mathbb{E} \sup_{g \in \mathscr{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathscr{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathscr{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z'_1, \ldots, z'_n)$
  - Then $\mathbb{E}_S \sup_{g \in \mathscr{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq \mathbb{E}_{S,S'} \sup_{g \in \mathscr{G}} \frac{1}{n} \left| \sum_{i=1}^n \left[g(z'_i) - g(z_i)\right] \right|$

arbitrary $\sigma_i \in \{-1,1\}$

$$= \mathbb{E}_{S,S'} \sup_{g \in \mathscr{G}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \left[g(z'_i) - g(z_i)\right] \right| = \mathbb{E}_{S,S'} \mathbb{E}_\sigma \sup_{g \in \mathscr{G}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \left[g(z'_i) - g(z_i)\right] \right|$$

Trick called **symmetrization**: it doesn't matter if we swap z_i and z'_i, since everything is iid and only looked at once

σ_i now iid $\mathrm{Uniform}(\{-1,1\})$ (**Rademacher** distribution)

triangle inequality + $\sup_x [f(x) + g(x)] \leq \sup_x f(x) + \sup_x g(x)$

$$\leq \mathbb{E}_{S,S'} \mathbb{E}_\sigma \sup_{g \in \mathscr{G}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g(z'_i) \right| + \mathbb{E}_{S,S'} \mathbb{E}_\sigma \sup_{g \in \mathscr{G}} \frac{1}{n} \left| \sum_{i=1}^n -\sigma_i g(z_i) \right| = 2\mathfrak{R}_n(\mathscr{G})$$

if $\mathscr{G}$ is **symmetric**: $g \in \mathscr{G}$ implies $-g \in \mathscr{G}$

- Want $\mathbb{E} \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq 2\mathfrak{R}_n(\mathcal{G})$ always $\mathbb{E} \sup_{g \in \mathcal{G}} \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \leq 2\mathfrak{R}_n(\mathcal{G})$
- Rewrite using a **ghost sample**: $\mathbb{E}[g] = \mathbb{E}_{S' \sim \mathcal{D}^n}\left[\hat{\mathbb{E}}_{S'}[g]\right]$ with $S' = (z'_1, \ldots, z'_n)$
- Then $\mathbb{E}_S \sup_{g \in \mathcal{G}} |\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]| \leq \mathbb{E}_{S,S'} \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n \left[g(z'_i) - g(z_i)\right] \right|$

arbitrary $\sigma_i \in \{-1,1\}$

$$= \mathbb{E}_{S,S'} \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \left[g(z'_i) - g(z_i)\right] \right| = \mathbb{E}_{S,S'}\mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \left[g(z'_i) - g(z_i)\right] \right|$$

Trick called **symmetrization**: it doesn't matter if we swap z_i and z'_i, since everything is iid and only looked at once

σ_i now iid $\mathrm{Uniform}(\{-1,1\})$ (**Rademacher** distribution)

triangle inequality + $\sup_x [f(x) + g(x)] \leq \sup_x f(x) + \sup_x g(x)$

$$\leq \mathbb{E}_{S,S'}\mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g(z'_i) \right| + \mathbb{E}_{S,S'}\mathbb{E}_\sigma \sup_{g \in \mathcal{G}} \frac{1}{n} \left| \sum_{i=1}^n -\sigma_i g(z_i) \right| = 2\mathfrak{R}_n(\mathcal{G})$$

if $\mathcal{G}$ is **symmetric**: $g \in \mathcal{G}$ implies $-g \in \mathcal{G}$

So what does that mean?

- For $\{0,1\}$ classifiers in symmetric $\mathscr{H}$ with 0-1 loss, $n \geq d = \mathrm{VCdim}(\mathscr{H}) \geq 3$,

$$\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)| \leq \frac{2}{\delta} \mathfrak{R}_n(\mathscr{H}) \leq \frac{2}{\delta} \sqrt{\frac{2d \log n}{n}} = \varepsilon$$

$$\left(\frac{\delta \varepsilon}{2}\right)^2 \geq \frac{2d \log n}{n}$$

$$\frac{n}{\log n} \geq \underbrace{\frac{8d}{\delta^2 \varepsilon^2}}$$

$\mathscr{H}$ is binary classifiers

$\mathscr{G}$ is their 0-1 loss

if symmetric

$g \in \mathscr{G}$ then $1 - g \in \mathscr{G}$

So what does that mean?

- For $\{0,1\}$ classifiers in symmetric $\mathscr{H}$ with 0-1 loss, $n \geq d = \mathrm{VCdim}(\mathscr{H}) \geq 3$,

$$\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)| \leq \frac{2}{\delta} \mathfrak{R}_n(\mathscr{H}) \leq \frac{2}{\delta} \sqrt{\frac{2d \log n}{n}}$$

$n = \Omega\left(\frac{d}{\varepsilon^2 \delta^2}\right)$

- Showed $d < \infty$ implies $\mathscr{H}$ has the uniform convergence property
  - and so finally proved the Fundamental Theorem of Learning! (up to Massart's lemma)
- But optimal rate is $n_{\mathscr{H}}^{UC} \leq \frac{C_2}{\varepsilon^2} \left[d + \log \frac{1}{\delta} \right]$: *much* better in δ

One-sided vs two-sided convergence

One-sided vs two-sided convergence

- For $\{0,1\}$ classifiers, $\mathscr{H}$ not necessarily symmetric, with 0-1 loss, $n \geq d = \mathrm{VCdim}(\mathscr{H}) \geq 3$,

$$\sup_{h \in \mathscr{H}} L_{\mathscr{D}}(h) - L_S(h) \leq \frac{2}{\delta} \mathfrak{R}_n(\mathscr{H}) \leq \frac{2}{\delta} \sqrt{\frac{2d \log n}{n}}$$

One-sided vs two-sided convergence

- For $\{0,1\}$ classifiers, $\mathscr{H}$ not necessarily symmetric, with 0-1 loss, $n \geq d = \mathrm{VCdim}(\mathscr{H}) \geq 3$,

$$\sup_{h \in \mathscr{H}} L_{\mathscr{D}}(h) - L_S(h) \leq \frac{2}{\delta} \mathfrak{R}_n(\mathscr{H}) \leq \frac{2}{\delta} \sqrt{\frac{2d \log n}{n}}$$

- Is what we really care about: for any $h \in \mathscr{H}$, have $L_{\mathscr{D}}(h) \leq L_S(h) + \frac{2}{\delta} \mathfrak{R}_n(\mathscr{H})$

$L_{S'} - L_D$

$\mathbb{E}_{S'} L_{S'} = L_D$

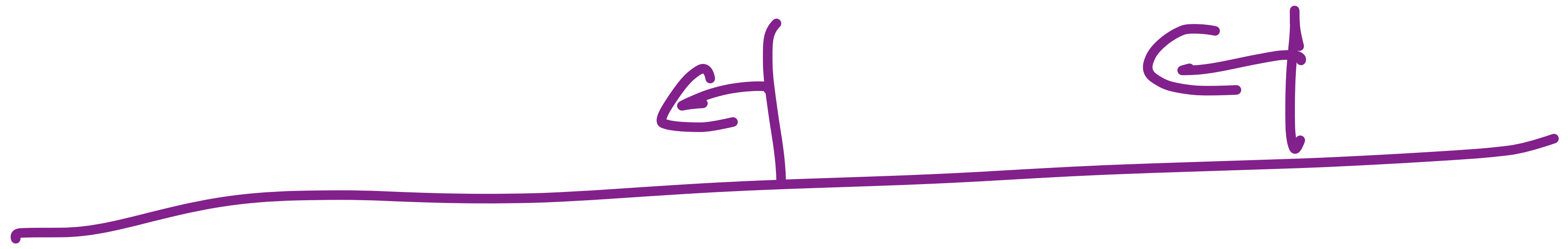
One-sided vs two-sided convergence

- For $\{0,1\}$ classifiers, $\mathscr{H}$ not necessarily symmetric, with 0-1 loss, $n \geq d = \mathrm{VCdim}(\mathscr{H}) \geq 3$,

$$\sup_{h \in \mathscr{H}} L_{\mathscr{D}}(h) - L_S(h) \leq \frac{2}{\delta} \mathfrak{R}_n(\mathscr{H}) \leq \frac{2}{\delta} \sqrt{\frac{2d \log n}{n}}$$

- Is what we really care about: for any $h \in \mathscr{H}$, have $L_{\mathscr{D}}(h) \leq L_S(h) + \frac{2}{\delta} \mathfrak{R}_n(\mathscr{H})$

- Rademacher results usually framed this way to avoid the symmetric condition
  - Some other bounds work this way too (e.g. PAC-Bayes)
  - Difference between two-sided and one-sided will be important later!



(pause)

McDiarmid’s inequality

McDiarmid's inequality

- Hoeffding's inequality:
  "the mean of independent, bounded RVs *concentrates* around its expectation"

McDiarmid's inequality

- Hoeffding's inequality:
  "the mean of independent, bounded RVs *concentrates* around its expectation"

- McDiarmid's inequality lets us take more general functions of independent RVs:

McDiarmid's inequality

- Hoeffding's inequality:
  "the mean of independent, bounded RVs *concentrates* around its expectation"

- McDiarmid's inequality lets us take more general functions of independent RVs:
  - Say f has **bounded differences** if for all $1 \leq i \leq n$,

$$\left| f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \leq c_i$$

$f(x_1, \ldots, x_n) = \frac{1}{n} \sum_i x_i$ $\qquad \frac{B}{n}$

$x_i \in [0,1]$ $\qquad \frac{1}{5}\left(0.3 + 0.2 + \underbrace{0.9}_{\in [A, A+B]} + 0.1 + 0.4 \right)$

0

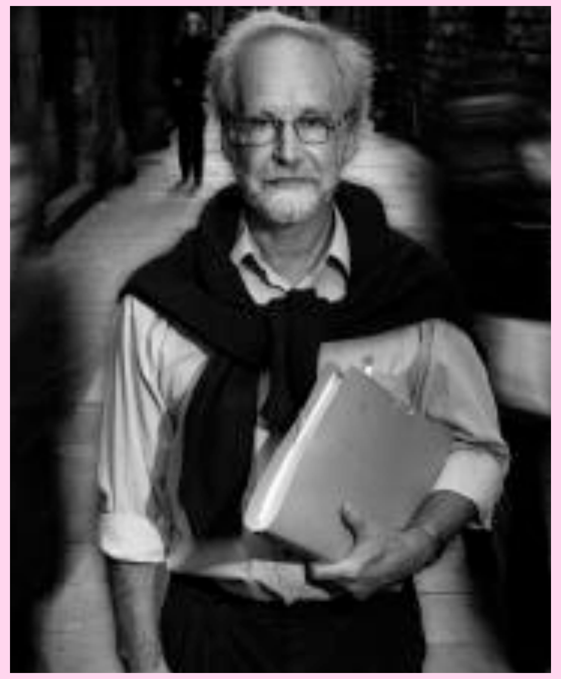
1

McDiarmid's inequality

- Hoeffding's inequality:
  "the mean of independent, bounded RVs *concentrates* around its expectation"

- McDiarmid's inequality lets us take more general functions of independent RVs:
  - Say f has **bounded differences** if for all $1 \leq i \leq n$,
    $$\left| f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \leq c_i$$

McDiarmid's inequality (1989)



Colin McDiarmid

If $X_1, \ldots, X_n \in \mathbb{R}$ independent, f has bounded diffs with c_i,

then $\Pr\left(f(S) > \mathbb{E}_{S'} f(S') + \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_i c_i^2}\right)$

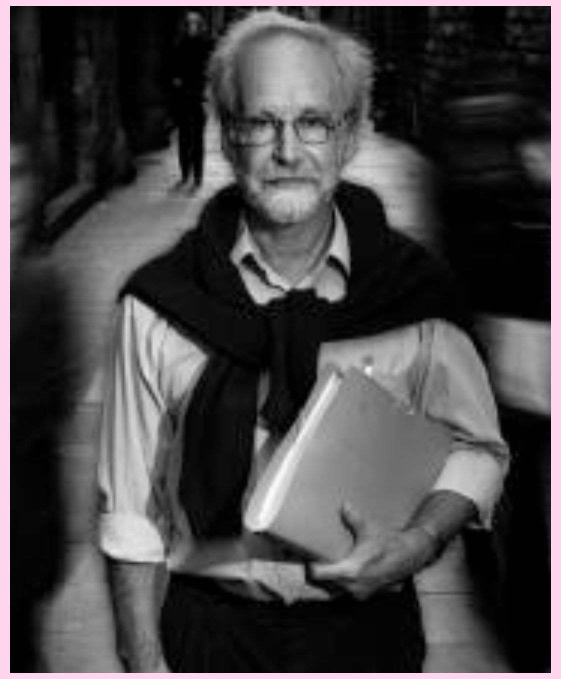
and same for $\Pr\left(f(S) < \mathbb{E}_{S'} f(S') - \varepsilon\right)$

More general than Hoeffding

$$\Pr\left(\left|\frac{1}{n}\sum x_i - \mathbb{E}x_i\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_i \left(\frac{B}{n}\right)^2}\right) = 2\exp\left(\frac{-2\varepsilon^2}{n\cdot\frac{B^2}{n^2}}\right) = 2\exp\left(\frac{-2n\varepsilon^2}{B^2}\right)$$

$$f(S) = f(X_1, \dots, X_n) = \frac{1}{n}\sum_i x_i$$

McDiarmid's inequality (1989)



Colin McDiarmid

If $X_1, \ldots, X_n \in \mathbb{R}$ independent, f has bounded diffs with c_i,

then $\Pr\left(f(S) > \mathbb{E}_{S'}f(S') + \varepsilon\right) \leq 2\exp\left(\frac{-2\varepsilon^2}{\sum_i c_i^2}\right)$

and same for $\Pr\left(f(S) < \mathbb{E}_{S'}f(S') - \varepsilon\right)$

Generalization gap has bounded differences

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$
- Let S' be S with z_i replaced by z_i' (for any *one* i)

Generalization gap has bounded differences

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$
- Let S' be S with z_i replaced by z_i' (for any *one* i)

$$\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$$

Generalization gap has bounded differences

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$
- Let S' be S with z_i replaced by z_i' (for any *one* i)

$$\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$$
$$= \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_{S'} g + \frac{1}{n} g(z_i') - \frac{1}{n} g(z_i) \right|$$

Generalization gap has bounded differences

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$
- Let S' be S with z_i replaced by z_i' (for any *one* i)

$$
\begin{aligned}
\Phi(S) &= \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right| \\
&= \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_{S'} g + \frac{1}{n} g(z_i') - \frac{1}{n} g(z_i) \right| \\
&\leq \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_{S'} g \right| + \frac{1}{n} \sup_{g \in \mathcal{G}} |g(z_i') - g(z_i)|
\end{aligned}
$$

Generalization gap has bounded differences

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$ and assume $g(z) \in [0,B]$
- Let S' be S with z_i replaced by z_i' (for any *one* i)

$$\begin{aligned}
\Phi(S) &= \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right| \\
&= \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_{S'} g + \frac{1}{n} g(z_i') - \frac{1}{n} g(z_i) \right| \\
&\leq \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_{S'} g \right| + \frac{1}{n} \sup_{g \in \mathcal{G}} |g(z_i') - g(z_i)|
\end{aligned}$$

Generalization gap has bounded differences

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$ and assume $g(z) \in [0,B]$
- Let S' be S with z_i replaced by z_i' (for any *one* i)

$$
\begin{aligned}
\Phi(S) &= \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right| \\
&= \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_{S'} g + \frac{1}{n} g(z_i') - \frac{1}{n} g(z_i) \right| \\
&\leq \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_{S'} g \right| + \frac{1}{n} \sup_{g \in \mathcal{G}} |g(z_i') - g(z_i)| \\
&\leq \Phi(S') + \frac{B}{n}
\end{aligned}
$$

Generalization gap has bounded differences

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$ and assume $g(z) \in [0,B]$
- Let S' be S with z_i replaced by z_i' (for any *one* i)

$$\begin{aligned}
\Phi(S) &= \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right| \\
&= \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_{S'}g + \frac{1}{n}g(z_i') - \frac{1}{n}g(z_i) \right| \\
&\leq \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_{S'}g \right| + \frac{1}{n}\sup_{g \in \mathcal{G}} |g(z_i') - g(z_i)| \\
&\leq \Phi(S') + \frac{B}{n} \qquad \text{so } |\Phi(S) - \Phi(S')| \leq \frac{B}{n}
\end{aligned}$$

A better generalization bound

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$ and assume $g(z) \in [0,B]$

A better generalization bound

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$ and assume $g(z) \in [0,B]$
- Φ satisfies bounded differences with $c_i = \frac{B}{n}$

A better generalization bound

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$ and assume $g(z) \in [0, B]$
- Φ satisfies bounded differences with $c_i = \frac{B}{n}$
- Already saw $\mathbb{E}_S \Phi(S) \leq 2\mathfrak{R}_n(\mathcal{G})$

A better generalization bound

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$ and assume $g(z) \in [0,B]$
- Φ satisfies bounded differences with $c_i = \dfrac{B}{n}$
- Already saw $\mathbb{E}_S \Phi(S) \leq 2\mathfrak{R}_n(\mathcal{G})$
- So we get

$$\Pr\left(\Phi(S) > 2\mathfrak{R}_n(\mathcal{G}) + \varepsilon\right) \leq \Pr\left(\Phi(S) > \mathbb{E}_{S'}\Phi(S') + \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{n \cdot (B/n)^2}\right)$$

A better generalization bound

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$ and assume $g(z) \in [0,B]$
- Φ satisfies bounded differences with $c_i = \dfrac{B}{n}$
- Already saw $\mathbb{E}_S \Phi(S) \leq 2\mathfrak{R}_n(\mathcal{G})$
- So we get

$$\Pr\left(\Phi(S) > 2\mathfrak{R}_n(\mathcal{G}) + \varepsilon\right) \leq \Pr\left(\Phi(S) > \mathbb{E}_{S'}\Phi(S') + \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{n \cdot (B/n)^2}\right)$$

$$\leq \exp\left(-2n\varepsilon^2/B^2\right)$$

A better generalization bound

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$ and assume $g(z) \in [0,B]$
- Φ satisfies bounded differences with $c_i = \dfrac{B}{n}$
- Already saw $\mathbb{E}_S \Phi(S) \leq 2\mathfrak{R}_n(\mathcal{G})$
- So we get

$$\Pr\left(\Phi(S) > 2\mathfrak{R}_n(\mathcal{G}) + \varepsilon\right) \leq \Pr\left(\Phi(S) > \mathbb{E}_{S'}\Phi(S') + \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{n \cdot (B/n)^2}\right)$$

$$\leq \exp\left(-2n\varepsilon^2/B^2\right)$$

Solving for ε, get $\Phi(S) \leq 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\dfrac{B^2}{2n} \log \dfrac{1}{\delta}}$

A better generalization bound

- Consider $\Phi(S) = \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right|$ and assume $g(z) \in [0,B]$
- Φ satisfies bounded differences with $c_i = \dfrac{B}{n}$
- Already saw $\mathbb{E}_S \Phi(S) \leq 2\mathfrak{R}_n(\mathcal{G})$

Still has pesky $\log n$ inside the $\mathfrak{R}_n$; can get optimal rate with *chaining*

- So we get

$$\Pr\left(\Phi(S) > 2\mathfrak{R}_n(\mathcal{G}) + \varepsilon\right) \leq \Pr\left(\Phi(S) > \mathbb{E}_{S'}\Phi(S') + \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{n \cdot (B/n)^2}\right)$$

$$\leq \exp\left(-2n\varepsilon^2/B^2\right)$$

Solving for ε, get $\Phi(S) \leq 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\dfrac{B^2}{2n} \log \frac{1}{\delta}}$

$\hat{\mathfrak{R}}_S$ converges to $\mathfrak{R}_n$

- $\hat{\mathfrak{R}}_S$ also satisfies bounded differences for bounded g
- So $|\hat{\mathfrak{R}}_S - \mathfrak{R}_n|$ concentrates, by McDiarmid
- Can split δ failure prob into $\delta/2$ for this, $\delta/2$ for previous theorem and bound $\Phi(S)$ in terms of $\hat{\mathfrak{R}}_S$

If there's time left, let's prove: (If not: we'll come back to it another time!)

Massart's lemma: for $\mathscr{A} \subset \mathbb{R}^n$, if $\max_{a \in \mathscr{A}} \|a\| \leq r$, $\quad \mathbb{E}_{\boldsymbol{\sigma}}\Big[\max_{a \in \mathscr{A}} \frac{1}{n}\boldsymbol{\sigma}^\top a\Big] \leq \frac{1}{n} r \sqrt{2 \log |\mathscr{A}|}$

Summary

- We *finally* proved the fundamental theorem:
  - Finite VCdim characterizes PAC learnability (realizable or agnostic) + ERM
- Proved a generalization bound via Rademacher complexity
  - Works for any bounded loss function