# Even More Rademacher

CPSC 532S: Modern Statistical Learning Theory

2 February 2022

cs.ubc.ca/~dsuth/532S/22/

# Admin

- Keep an eye on Piazza for homework release (probably later this week)

- Next class will be in hybrid mode, in DMP 101
  - Those not officially enrolled are still totally welcome
    - Should be plenty of space
  - Will continue to livestream + record
    - Probably on Zoom (same link)
      - Will announce by this weekend on Piazza if something else
    - I’ll test out the setup in this room beforehand
      - but still probably higher odds of glitches than

- Office hours:
  - Tuesday 10-11 still Zoom-only
  - Thursday 4-5, I’ll be in my office, ICCS X563 (and also on Zoom)

# Recap: Rademacher Complexity

$S = \left(z_1, \ldots, z_n\right) \in \mathcal{Z}^n \qquad \mathcal{G} \ni g : \mathcal{Z} \to \mathbb{R} \qquad \mathbf{g}_S = \left(g(z_1), \ldots, g(z_n)\right)$ $\in \mathbb{R}^n$

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

$$\boldsymbol{\sigma} \sim \mathrm{Rad}^n = \mathrm{Uniform}(\{-1,1\})^n$$

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{S \sim \mathcal{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathcal{G})\right]$$

# Recap: Rademacher Complexity

$S = \left(z_1, \ldots, z_n\right) \in \mathcal{Z}^n \qquad \mathcal{G} \ni g : \mathcal{Z} \to \mathbb{R} \qquad \mathbf{g}_S = \left(g(z_1), \ldots, g(z_n)\right)$

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \boldsymbol{\sigma}^\mathsf{T} \mathbf{g}_S\right]$$

$\boldsymbol{\sigma} \sim \mathrm{Rad}^n = \mathrm{Uniform}(\{-1,1\})^n$

$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{S \sim \mathcal{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathcal{G})\right]$

$\hat{\mathfrak{R}}_S(\mathcal{H})$: "how well can classifiers in $\mathcal{H}$ fit random labels on $S$?"

$\mathfrak{R}_n(\mathcal{H})$: "how well can classifiers in $\mathcal{H}$ fit random labels on sets from $\mathcal{D}$ of size $n$?"

# Recap: Rademacher Complexity

$S = \left(z_1, \ldots, z_n\right) \in \mathcal{Z}^n \qquad \mathcal{G} \ni g : \mathcal{Z} \to \mathbb{R} \qquad \mathbf{g}_S = \left(g(z_1), \ldots, g(z_n)\right)$

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \boldsymbol{\sigma}^{\mathsf{T}} \mathbf{g}_S\right]$$

$\boldsymbol{\sigma} \sim \mathrm{Rad}^n = \mathrm{Uniform}(\{-1,1\})^n$

$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{S \sim \mathcal{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathcal{G})\right]$

$\hat{\mathfrak{R}}_S(\mathcal{H})$: “how well can classifiers in $\mathcal{H}$ fit random labels on $S$?”

$\mathfrak{R}_n(\mathcal{H})$: “how well can classifiers in $\mathcal{H}$ fit random labels on sets from $\mathcal{D}$ of size $n$?”

$\hat{\mathfrak{R}}_S(\mathcal{G})$: ~“how big can $L_{S_1} - L_{S_2}$ be?”

when $\mathcal{G} = \left\{z \mapsto \ell(h, z) : h \in \mathcal{H}\right\}$

if $h$ outputs in $\{-1,1\}$, $\hat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H})$

# Recap: Rademacher Complexity

$S = \left(z_1, \ldots, z_n\right) \in \mathcal{Z}^n \qquad \mathcal{G} \ni g : \mathcal{Z} \to \mathbb{R} \qquad \mathbf{g}_S = \left(g(z_1), \ldots, g(z_n)\right)$

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[ \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right] = \mathbb{E}_{\boldsymbol{\sigma}} \left[ \sup_{g \in \mathcal{G}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{g}_S \right]$$

$$\boldsymbol{\sigma} \sim \mathrm{Rad}^n = \mathrm{Uniform}(\{-1,1\})^n$$

$$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{S \sim \mathcal{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathcal{G})\right]$$

$\hat{\mathfrak{R}}_S(\mathcal{H})$: "how well can classifiers in $\mathcal{H}$ fit random labels on $S$?"

$\mathfrak{R}_n(\mathcal{H})$: "how well can classifiers in $\mathcal{H}$ fit random labels on sets from $\mathcal{D}$ of size $n$?"

$\hat{\mathfrak{R}}_S(\mathcal{G})$: ~"how big can $L_{S_1} - L_{S_2}$ be?"

when $\mathcal{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathcal{H} \right\}$

if $h$ outputs in $\{-1,1\}$, $\hat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2} \hat{\mathfrak{R}}_S(\mathcal{H})$

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

so $\hat{\mathfrak{R}}_S(\mathcal{H}) \leq \sqrt{\frac{2}{n} \log \tau_{\mathcal{H}}(n)}$, where $\tau_{\mathcal{H}} \leq (en/\mathrm{VCdim}(\mathcal{H}))^{\mathrm{VCdim}(\mathcal{H})}$ is the growth function

# Recap: Rademacher Complexity

$S = (z_1, \ldots, z_n) \in \mathcal{Z}^n \qquad \mathcal{G} \ni g : \mathcal{Z} \to \mathbb{R} \qquad \mathbf{g}_S = (g(z_1), \ldots, g(z_n))$

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(z_i)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

$\boldsymbol{\sigma} \sim \mathrm{Rad}^n = \mathrm{Uniform}(\{-1,1\})^n$

$\mathfrak{R}_n(\mathcal{G}) = \mathbb{E}_{S \sim \mathcal{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathcal{G})\right]$

$\hat{\mathfrak{R}}_S(\mathcal{H})$: "how well can classifiers in $\mathcal{H}$ fit random labels on $S$?"

$\mathfrak{R}_n(\mathcal{H})$: "how well can classifiers in $\mathcal{H}$ fit random labels on sets from $\mathcal{D}$ of size $n$?"

$\hat{\mathfrak{R}}_S(\mathcal{G})$: ~"how big can $L_{S_1} - L_{S_2}$ be?"

when $\mathcal{G} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}$

if $h$ outputs in $\{-1,1\}$, $\hat{\mathfrak{R}}_S(\mathcal{G}) = \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H})$

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

so $\hat{\mathfrak{R}}_S(\mathcal{H}) \le \sqrt{\frac{2}{n} \log \tau_{\mathcal{H}}(n)}$, where $\tau_{\mathcal{H}} \le (en/\mathrm{VCdim}(\mathcal{H}))^{\mathrm{VCdim}(\mathcal{H})}$ is the growth function

**Theorem**: if $\mathcal{G}$ maps to $[0,B]$, $\quad \sup_{g \in \mathcal{G}} \left[\mathbb{E}[g(z)] - \frac{1}{n} \sum_{i=1}^n g(z_i)\right] \le 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{B^2}{2n} \log \frac{1}{\delta}}$

# The Other Rademacher Complexity

$S = (z_1, \ldots, z_n) \in \mathcal{Z}^n$ $\quad \mathcal{G} \ni g : \mathcal{Z} \to \mathbb{R}$ $\quad \mathbf{g}_S = (g(z_1), \ldots, g(z_n))$

$\hat{\mathfrak{R}}'_S$ notation here isn't standard: most people use one or the other

$$\hat{\mathfrak{R}}'_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right| \right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{g}_S \right| \right]$$

$\boldsymbol{\sigma} \sim \mathrm{Rad}^n = \mathrm{Uniform}(\{-1,1\})^n$

$\mathfrak{R}'_n(\mathcal{G}) = \mathbb{E}_{S \sim \mathcal{D}^n}\left[\hat{\mathfrak{R}}'_S(\mathcal{G})\right]$

$\hat{\mathfrak{R}}'_S(\mathcal{H})$: "how well can classifiers in $\mathcal{H}$ fit (or opposite-fit) random labels on $S$?"

$\mathfrak{R}'_n(\mathcal{H})$: "how well can classifiers in $\mathcal{H}$ fit (or opposite-fit) random labels on sets from $\mathcal{D}$ of size $n$?"

$\hat{\mathfrak{R}}'_S(\mathcal{G})$: ~"how big can $|L_{S_1} - L_{S_2}|$ be?"

when $\mathcal{G} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}$

if $h$ outputs in $\{-1,1\}$, $\hat{\mathfrak{R}}'_S(\mathcal{G}) \le \frac{1}{2}\hat{\mathfrak{R}}'_S(\mathcal{H}) + 1/(2\sqrt{n})$

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

apply old lemma to $\mathcal{A} \cup (-\mathcal{A})$

so $\hat{\mathfrak{R}}'_S(\mathcal{H}) \le \sqrt{\frac{2}{n} \log 2\tau_{\mathcal{H}}(n)}$, where $\tau_{\mathcal{H}} \le (en/\mathrm{VCdim}(\mathcal{H}))^{\mathrm{VCdim}(\mathcal{H})}$ is the growth function

**Theorem**: if $\mathcal{G}$ maps to $[0,B]$, $\quad \sup_{g \in \mathcal{G}} \left| \mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i) \right| \le 2\mathfrak{R}'_n(\mathcal{G}) + \sqrt{\frac{B^2}{2n} \log \frac{1}{\delta}}$

# A Tale of Two Complexities

$$\hat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

$$\hat{\mathfrak{R}}'_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right]$$

# A Tale of Two Complexities

$$\hat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

$$\hat{\mathfrak{R}}'_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right]$$

$\{-g : g \in \mathcal{G}\}$

$$\hat{\mathfrak{R}}_S(\mathscr{G}) \le \hat{\mathfrak{R}}_S(\mathscr{G} \cup (-\mathscr{G})) = \hat{\mathfrak{R}}'_S(\mathscr{G})$$ (so if $\mathscr{G}$ is already symmetric, they're the same)

$\hat{R}_S(\mathcal{G}) \le \hat{R}_S(\mathcal{F})$ if $\mathcal{G} \subseteq \mathcal{F}$

# A Tale of Two Complexities

$$\hat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right] \qquad \hat{\mathfrak{R}}'_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right]$$

$$\hat{\mathfrak{R}}_S(\mathscr{G}) \le \hat{\mathfrak{R}}_S(\mathscr{G}\cup(-\mathscr{G})) = \hat{\mathfrak{R}}'_S(\mathscr{G})$$ (so if $\mathscr{G}$ is already symmetric, they're the same)

$$\hat{\mathfrak{R}}_S(c\mathscr{G}) = |c|\hat{\mathfrak{R}}_S(\mathscr{G}) \qquad \hat{\mathfrak{R}}'_S(c\mathscr{G}) = |c|\hat{\mathfrak{R}}'_S(\mathscr{G})$$

$c\mathscr{G} = \{cg : g \in \mathscr{G}\}$

$[cg](z) = c\,g(z)$

$$\mathbb{E}_\sigma \sup_{g \in \mathscr{G}} \left|\frac{1}{n}\sigma^\top(c\, g_S)\right| = |c|\, \underbrace{\mathbb{E}_\sigma \sup_g \left|\frac{1}{n}\overbrace{[\mathrm{sign}(c)\,\sigma]}^{\sim \mathrm{Rad}^n}{}^\top g_S\right|}_{|c|\,\hat{R}_S}$$

$c \to \mathrm{sign}(c)\,|c|$

# A Tale of Two Complexities

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

$$\hat{\mathfrak{R}}'_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right]$$

$$\hat{\mathfrak{R}}_S(\mathcal{G}) \le \hat{\mathfrak{R}}_S(\underbrace{\mathcal{G} \cup (-\mathcal{G})}) = \hat{\mathfrak{R}}'_S(\mathcal{G})$$ (so if $\mathcal{G}$ is already symmetric, they're the same)

$$\hat{\mathfrak{R}}_S(c\mathcal{G}) = |c|\hat{\mathfrak{R}}_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}'_S(c\mathcal{G}) = |c|\hat{\mathfrak{R}}'_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right] = 0$$

$$\frac{1}{n}\underbrace{\mathbb{E}[\sigma]}_{0}{}^\top g_S$$

# A Tale of Two Complexities

$$\hat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

$$\hat{\mathfrak{R}}'_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right]$$

$$\hat{\mathfrak{R}}_S(\mathscr{G}) \le \hat{\mathfrak{R}}_S(\mathscr{G} \cup (-\mathscr{G})) = \hat{\mathfrak{R}}'_S(\mathscr{G})$$ (so if $\mathscr{G}$ is already symmetric, they're the same)

$$\hat{\mathfrak{R}}_S(c\mathscr{G}) = |c|\hat{\mathfrak{R}}_S(\mathscr{G})$$

$$\hat{\mathfrak{R}}'_S(c\mathscr{G}) = |c|\hat{\mathfrak{R}}'_S(\mathscr{G})$$

$$\hat{\mathfrak{R}}_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right] = 0$$

$$\hat{\mathfrak{R}}'_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right] > 0 \text{ if } \mathbf{g}_S \neq 0$$

# A Tale of Two Complexities

$$\hat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right] \qquad \hat{\mathfrak{R}}'_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right]$$

$$\hat{\mathfrak{R}}_S(\mathscr{G}) \le \hat{\mathfrak{R}}_S(\mathscr{G} \cup (-\mathscr{G})) = \hat{\mathfrak{R}}'_S(\mathscr{G})$$ (so if $\mathscr{G}$ is already symmetric, they're the same)

$$\hat{\mathfrak{R}}_S(c\mathscr{G}) = |c|\hat{\mathfrak{R}}_S(\mathscr{G}) \qquad \hat{\mathfrak{R}}'_S(c\mathscr{G}) = |c|\hat{\mathfrak{R}}'_S(\mathscr{G})$$

$$\hat{\mathfrak{R}}_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right] = 0 \qquad \hat{\mathfrak{R}}'_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right] > 0 \text{ if } \mathbf{g}_S \neq 0$$

$$\hat{\mathfrak{R}}'_S(\{c\mathbf{1}\}) = |c|\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\underbrace{|\boldsymbol{\sigma}^\top\mathbf{1}|}_{X}\right] \le \frac{|c|}{n}\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[|\boldsymbol{\sigma}^\top\mathbf{1}|^2\right]} = \frac{|c|}{n}\sqrt{\sum_{i,j}\underbrace{\mathbb{E}\sigma_i\sigma_j}} = \frac{|c|}{\sqrt{n}}$$

$\mathbb{E}X \le \sqrt{\mathbb{E}X^2}$

$(\mathbb{E}X)^2 \le \mathbb{E}X^2$

$(\sum_i \sigma_i)^2$

$= \sum_i \sum_j \sigma_i \sigma_j$

$n \cdot \sigma_i\sigma_i = n \cdot 1$

# A Tale of Two Complexities

$$\hat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right] \qquad \hat{\mathfrak{R}}'_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right]$$

$$\hat{\mathfrak{R}}_S(\mathscr{G}) \le \hat{\mathfrak{R}}_S(\mathscr{G} \cup (-\mathscr{G})) = \hat{\mathfrak{R}}'_S(\mathscr{G})$$ (so if $\mathscr{G}$ is already symmetric, they're the same)

$$\hat{\mathfrak{R}}_S(c\mathscr{G}) = |c|\hat{\mathfrak{R}}_S(\mathscr{G}) \qquad \hat{\mathfrak{R}}'_S(c\mathscr{G}) = |c|\hat{\mathfrak{R}}'_S(\mathscr{G})$$

$$\hat{\mathfrak{R}}_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right] = 0 \qquad \hat{\mathfrak{R}}'_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right] > 0 \text{ if } \mathbf{g}_S \neq 0$$

$$\hat{\mathfrak{R}}'_S(\{c\mathbf{1}\}) = |c|\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{1}|\right] \le \frac{|c|}{n}\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[|\boldsymbol{\sigma}^\top \mathbf{1}|^2\right]} = \frac{|c|}{n}\sqrt{\sum_{i,j}\mathbb{E}\sigma_i\sigma_j} = \frac{|c|}{\sqrt{n}}$$

$$\hat{\mathfrak{R}}_S(\mathscr{F} + \mathscr{G}) = \hat{\mathfrak{R}}_S(\mathscr{F}) + \hat{\mathfrak{R}}_S(\mathscr{G}) \qquad \hat{\mathfrak{R}}'_S(\mathscr{F} + \mathscr{G}) \le \hat{\mathfrak{R}}'_S(\mathscr{F}) + \hat{\mathfrak{R}}'_S(\mathscr{G})$$

$= \{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$

# A Tale of Two Complexities

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

$$\hat{\mathfrak{R}}'_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right]$$

$$\hat{\mathfrak{R}}_S(\mathcal{G}) \leq \hat{\mathfrak{R}}_S(\mathcal{G} \cup (-\mathcal{G})) = \hat{\mathfrak{R}}'_S(\mathcal{G})$$ (so if $\mathcal{G}$ is already symmetric, they're the same)

$$\hat{\mathfrak{R}}_S(c\mathcal{G}) = |c|\hat{\mathfrak{R}}_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}'_S(c\mathcal{G}) = |c|\hat{\mathfrak{R}}'_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right] = 0$$

$$\hat{\mathfrak{R}}'_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right] > 0 \ \text{ if } \mathbf{g}_S \neq 0$$

$$\hat{\mathfrak{R}}'_S(\{c\mathbf{1}\}) = |c|\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{1}|\right] \leq \frac{|c|}{n}\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[|\boldsymbol{\sigma}^\top \mathbf{1}|^2\right]} = \frac{|c|}{n}\sqrt{\sum_{i,j} \mathbb{E}\sigma_i\sigma_j} = \frac{|c|}{\sqrt{n}}$$

$$\hat{\mathfrak{R}}_S(\mathcal{F} + \mathcal{G}) = \hat{\mathfrak{R}}_S(\mathcal{F}) + \hat{\mathfrak{R}}_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}'_S(\mathcal{F} + \mathcal{G}) \leq \hat{\mathfrak{R}}'_S(\mathcal{F}) + \hat{\mathfrak{R}}'_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}_S(a\mathcal{G} + b) = |a|\hat{\mathfrak{R}}_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}'_S(a\mathcal{G} + b) \leq |a|\hat{\mathfrak{R}}'_S(\mathcal{G}) + \frac{|b|}{\sqrt{n}}$$

# A Tale of Two Complexities

$$\hat{\mathfrak{R}}_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

$$\hat{\mathfrak{R}}'_S(\mathscr{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right]$$

$$\hat{\mathfrak{R}}_S(\mathscr{G}) \le \hat{\mathfrak{R}}_S(\mathscr{G} \cup (-\mathscr{G})) = \hat{\mathfrak{R}}'_S(\mathscr{G})$$ (so if $\mathscr{G}$ is already symmetric, they're the same)

$$\hat{\mathfrak{R}}_S(c\mathscr{G}) = |c|\hat{\mathfrak{R}}_S(\mathscr{G})$$

$$\hat{\mathfrak{R}}'_S(c\mathscr{G}) = |c|\hat{\mathfrak{R}}'_S(\mathscr{G})$$

$$\hat{\mathfrak{R}}_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right] = 0$$

$$\hat{\mathfrak{R}}'_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right] > 0 \text{ if } \mathbf{g}_S \ne 0$$

$$\hat{\mathfrak{R}}'_S(\{c\mathbf{1}\}) = |c|\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{1}|\right] \le \frac{|c|}{n}\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[|\boldsymbol{\sigma}^\top \mathbf{1}|^2\right]} = \frac{|c|}{n}\sqrt{\sum_{i,j}\mathbb{E}\sigma_i\sigma_j} = \frac{|c|}{\sqrt{n}}$$

$$\hat{\mathfrak{R}}_S(\mathscr{F} + \mathscr{G}) = \hat{\mathfrak{R}}_S(\mathscr{F}) + \hat{\mathfrak{R}}_S(\mathscr{G})$$

$$\hat{\mathfrak{R}}'_S(\mathscr{F} + \mathscr{G}) \le \hat{\mathfrak{R}}'_S(\mathscr{F}) + \hat{\mathfrak{R}}'_S(\mathscr{G})$$

$$\hat{\mathfrak{R}}_S(a\mathscr{G} + b) = |a|\hat{\mathfrak{R}}_S(\mathscr{G})$$

$$\hat{\mathfrak{R}}'_S(a\mathscr{G} + b) \le |a|\hat{\mathfrak{R}}'_S(\mathscr{G}) + \frac{|b|}{\sqrt{n}}$$

so $\hat{\mathfrak{R}}_S\left(\frac{1}{2}(y_S \circ \mathscr{H}) + \frac{1}{2}\right) = \frac{1}{2}\hat{\mathfrak{R}}_S(\mathscr{H})$

# A Tale of Two Complexities

$$\hat{\mathfrak{R}}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right]$$

$$\hat{\mathfrak{R}}'_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right]$$

$$\hat{\mathfrak{R}}_S(\mathcal{G}) \le \hat{\mathfrak{R}}_S(\mathcal{G} \cup (-\mathcal{G})) = \hat{\mathfrak{R}}'_S(\mathcal{G})$$ (so if $\mathcal{G}$ is already symmetric, they're the same)

$$\hat{\mathfrak{R}}_S(c\mathcal{G}) = |c|\hat{\mathfrak{R}}_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}'_S(c\mathcal{G}) = |c|\hat{\mathfrak{R}}'_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{g}_S\right] = 0$$

$$\hat{\mathfrak{R}}'_S(\{g\}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{g}_S|\right] > 0 \text{ if } \mathbf{g}_S \ne 0$$

$$\hat{\mathfrak{R}}'_S(\{c\mathbf{1}\}) = |c|\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}|\boldsymbol{\sigma}^\top \mathbf{1}|\right] \le \frac{|c|}{n}\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[|\boldsymbol{\sigma}^\top \mathbf{1}|^2\right]} = \frac{|c|}{n}\sqrt{\sum_{i,j}\mathbb{E}\sigma_i\sigma_j} = \frac{|c|}{\sqrt{n}}$$

$\Longrightarrow |c| \cdot \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\sqrt{n}}$

$\approx 0.8$

$$\hat{\mathfrak{R}}_S(\mathcal{F} + \mathcal{G}) = \hat{\mathfrak{R}}_S(\mathcal{F}) + \hat{\mathfrak{R}}_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}'_S(\mathcal{F} + \mathcal{G}) \le \hat{\mathfrak{R}}'_S(\mathcal{F}) + \hat{\mathfrak{R}}'_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}_S(a\mathcal{G} + b) = |a|\hat{\mathfrak{R}}_S(\mathcal{G})$$

$$\hat{\mathfrak{R}}'_S(a\mathcal{G} + b) \le |a|\hat{\mathfrak{R}}'_S(\mathcal{G}) + \frac{|b|}{\sqrt{n}}$$

so $\hat{\mathfrak{R}}_S\left(\frac{1}{2}(y_S \circ \mathcal{H}) + \frac{1}{2}\right) = \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H})$

so $\hat{\mathfrak{R}}'_S\left(\frac{1}{2}(y_S \circ \mathcal{H}) + \frac{1}{2}\right) \le \frac{1}{2}\hat{\mathfrak{R}}'_S(\mathcal{H}) + \frac{1}{2\sqrt{n}}$

**Theorems**: if $\mathscr{G}$ maps to $[0,B]$,

$$\sup_{g\in\mathscr{G}} \left[\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right] \leq 2\mathfrak{R}_n(\mathscr{G}) + \sqrt{\frac{B^2}{2n}\log\frac{1}{\delta}}$$

$$\sup_{g\in\mathscr{G}} \left|\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right| \leq 2\mathfrak{R}'_n(\mathscr{G}) + \sqrt{\frac{B^2}{2n}\log\frac{1}{\delta}}$$

**Theorems**: if $\mathcal{G}$ maps to $[0,B]$,

$$\sup_{g\in\mathcal{G}} \left[\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right] \le 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{B^2}{2n}\log\frac{1}{\delta}}$$

$$\sup_{g\in\mathcal{G}} \left|\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right| \le 2\mathfrak{R}'_n(\mathcal{G}) + \sqrt{\frac{B^2}{2n}\log\frac{1}{\delta}}$$

So, for 0-1 loss, $\mathcal{H}$ outputting in $\{-1,1\}$:

$$\sup_{h\in\mathcal{H}} \left[L_{\mathcal{D}}(h) - L_S(h)\right] \le \mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

$$\sup_{h\in\mathcal{H}} \left|L_{\mathcal{D}}(h) - L_S(h)\right| \le \mathfrak{R}'_n(\mathcal{H}) + \frac{1}{\sqrt{n}}\left(1 + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)$$

**Theorems**: if $\mathscr{G}$ maps to $[0,B]$,

$$\sup_{g\in\mathscr{G}} \left[\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right] \le 2\mathfrak{R}_n(\mathscr{G}) + \sqrt{\frac{B^2}{2n}\log\frac{1}{\delta}}$$

$$\sup_{g\in\mathscr{G}} \left|\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right| \le 2\mathfrak{R}'_n(\mathscr{G}) + \sqrt{\frac{B^2}{2n}\log\frac{1}{\delta}}$$

So, for 0-1 loss, $\mathscr{H}$ outputting in $\{-1,1\}$:

$$\sup_{h\in\mathscr{H}} \left[L_{\mathscr{D}}(h) - L_S(h)\right] \le \mathfrak{R}_n(\mathscr{H}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

$2\mathfrak{R}_n(\mathscr{H}) + 2\sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$

$$\sup_{h\in\mathscr{H}} \left|L_{\mathscr{D}}(h) - L_S(h)\right| \le \mathfrak{R}'_n(\mathscr{H}) + \frac{1}{\sqrt{n}}\left(1 + \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right)$$

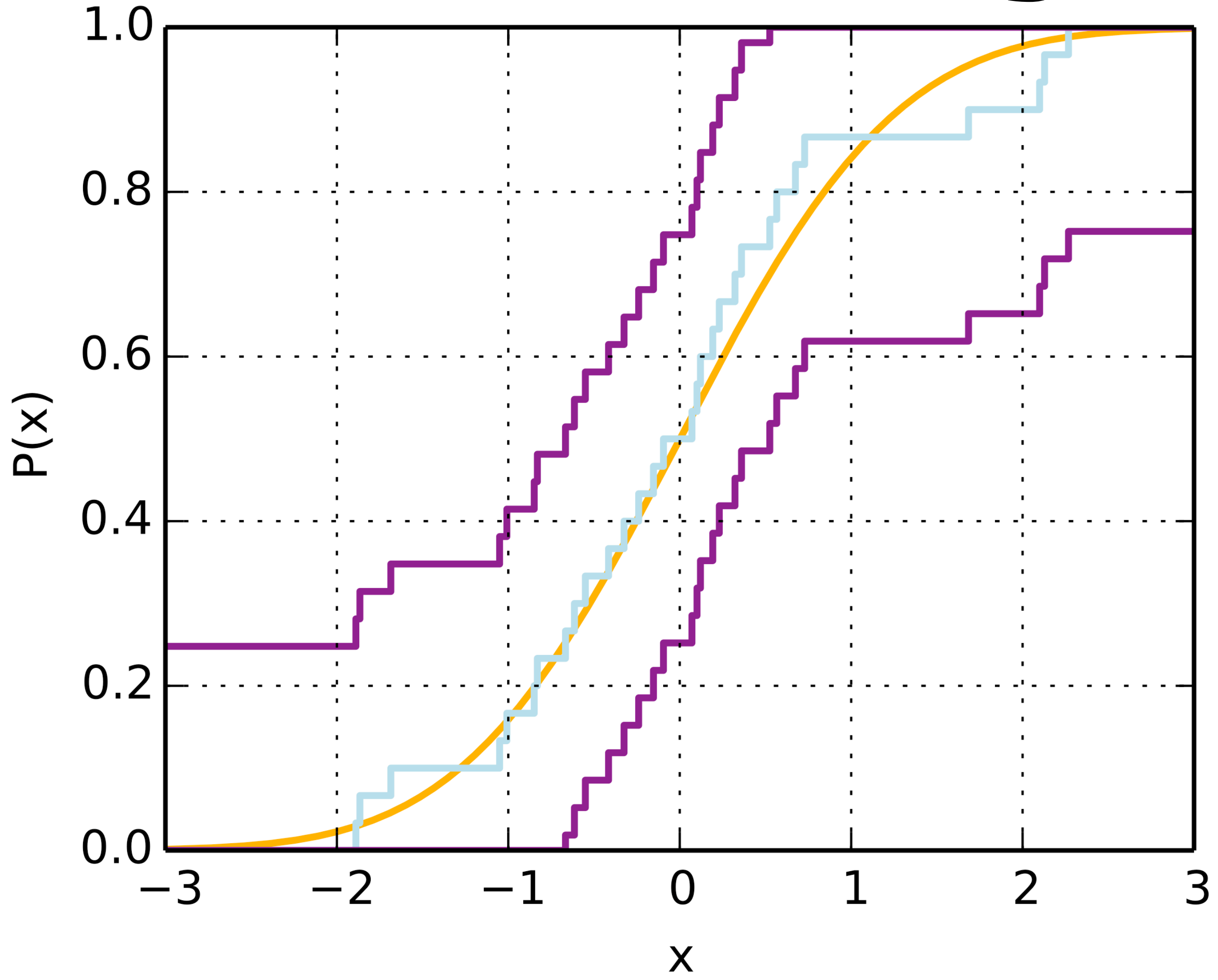
if we just want $L_{\mathscr{D}}(h) \le$ something,

then first bound is strictly smaller (same big-O rate);

but need the second one to prove "Fundamental Theorem of Statistical Learning"

# Glivenko-Cantelli / DKW inequality

- Let $F(t) = \Pr(X \leq t)$ be CDF of some random variable $X$
- $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq t)$ is empirical CDF of samples

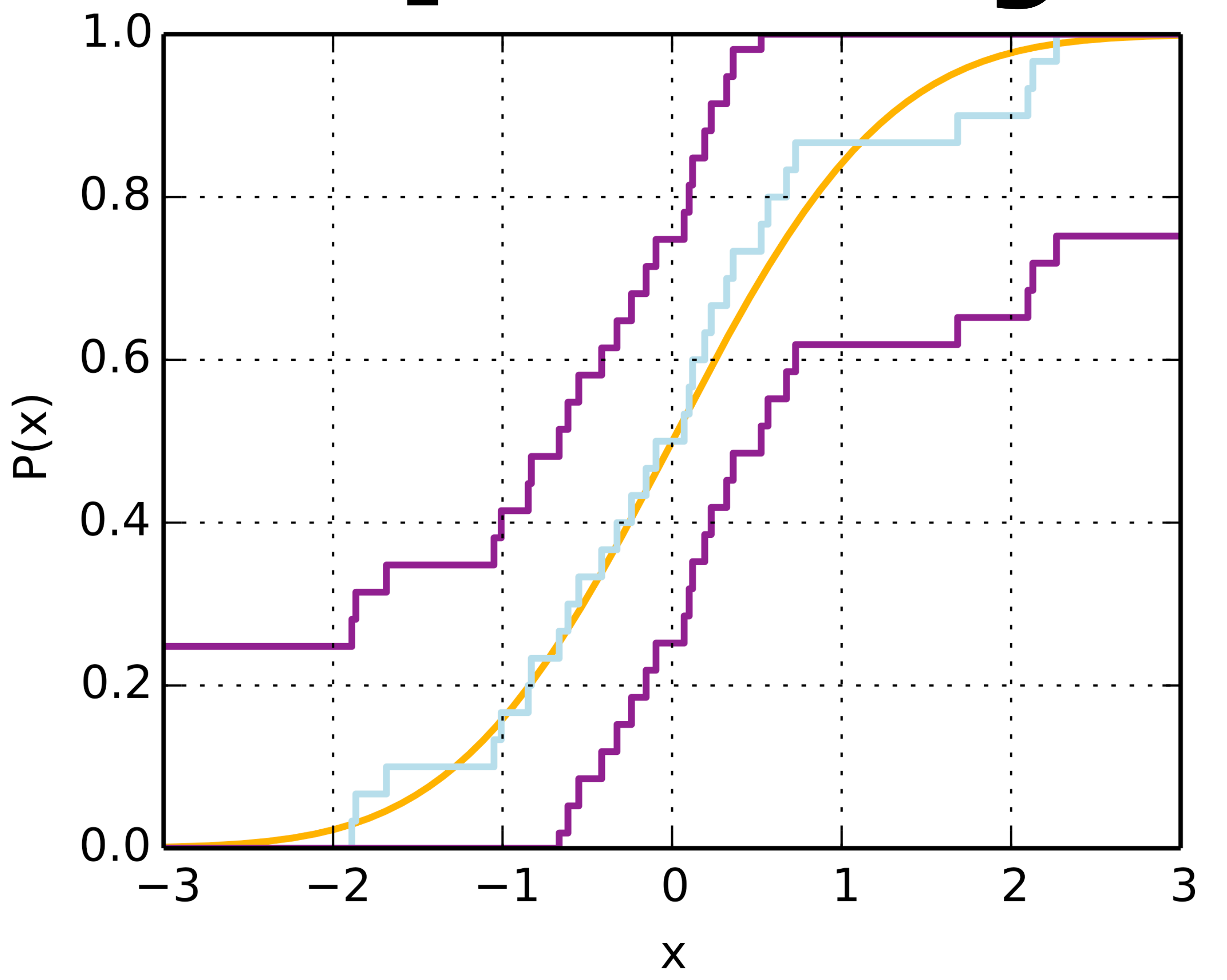



from Wikipedia on DKW inequality

# Glivenko-Cantelli / DKW inequality

- Let $F(t) = \Pr(X \leq t)$ be CDF of some random variable $X$
- $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \leq t)$ is empirical CDF of samples
- Worst-case error of $F_n$ is, letting $\mathscr{G} = \{x \mapsto \mathbb{I}(x \leq t) : t \in \mathbb{R}\}$,

$$\sup_t |F(t) - F_n(t)| \leq 2\hat{\mathfrak{R}}'_S(\mathscr{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$$



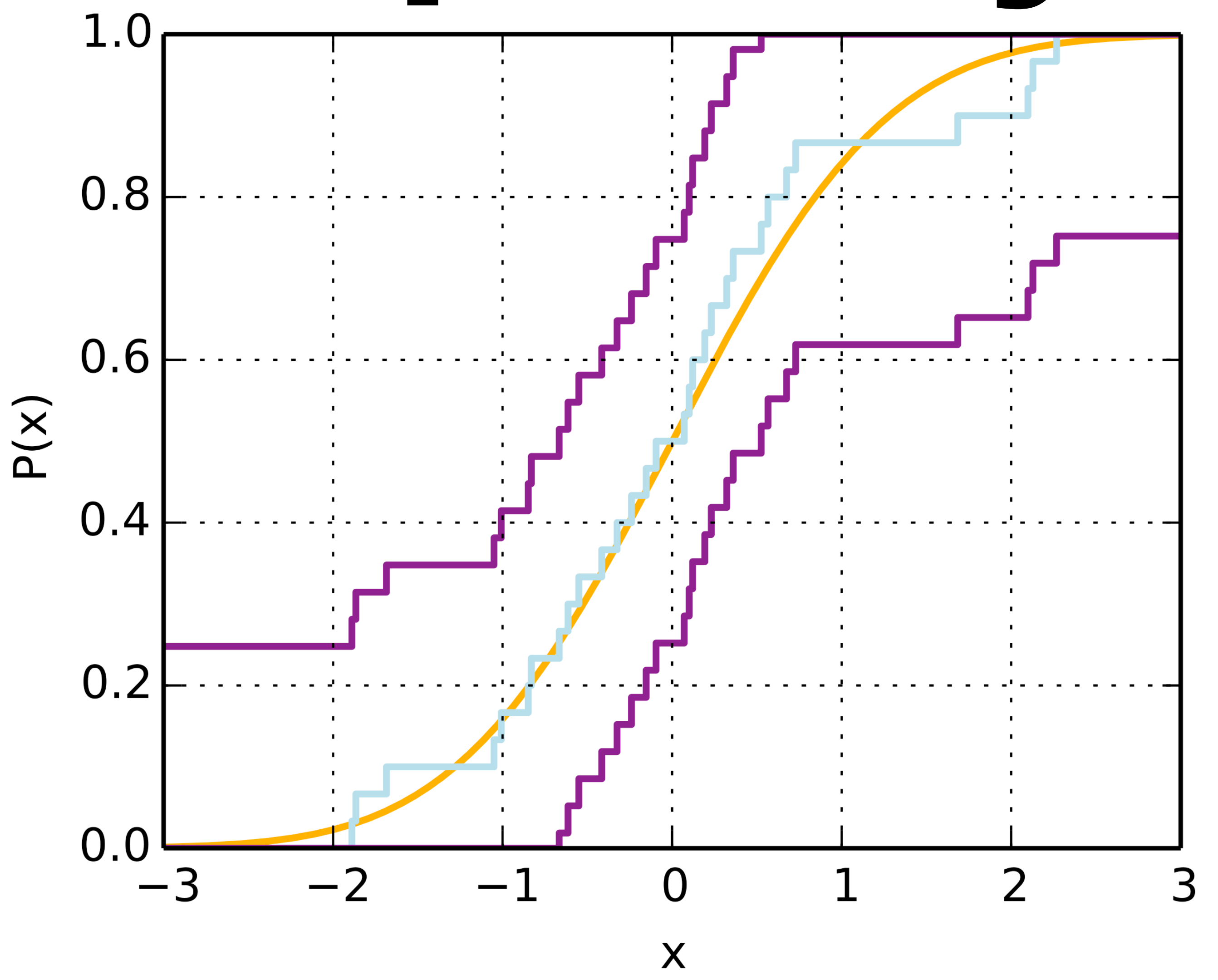

from Wikipedia on DKW inequality

# Glivenko-Cantelli / DKW inequality

- Let $F(t) = \Pr(X \le t)$ be CDF of some random variable $X$
- $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \le t)$ is empirical CDF of samples
- Worst-case error of $F_n$ is, letting $\mathcal{G} = \{x \mapsto \mathbb{I}(x \le t) : t \in \mathbb{R}\}$,

$$\sup_t |F(t) - F_n(t)| \le 2\hat{\mathfrak{R}}'_S(\mathcal{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$$

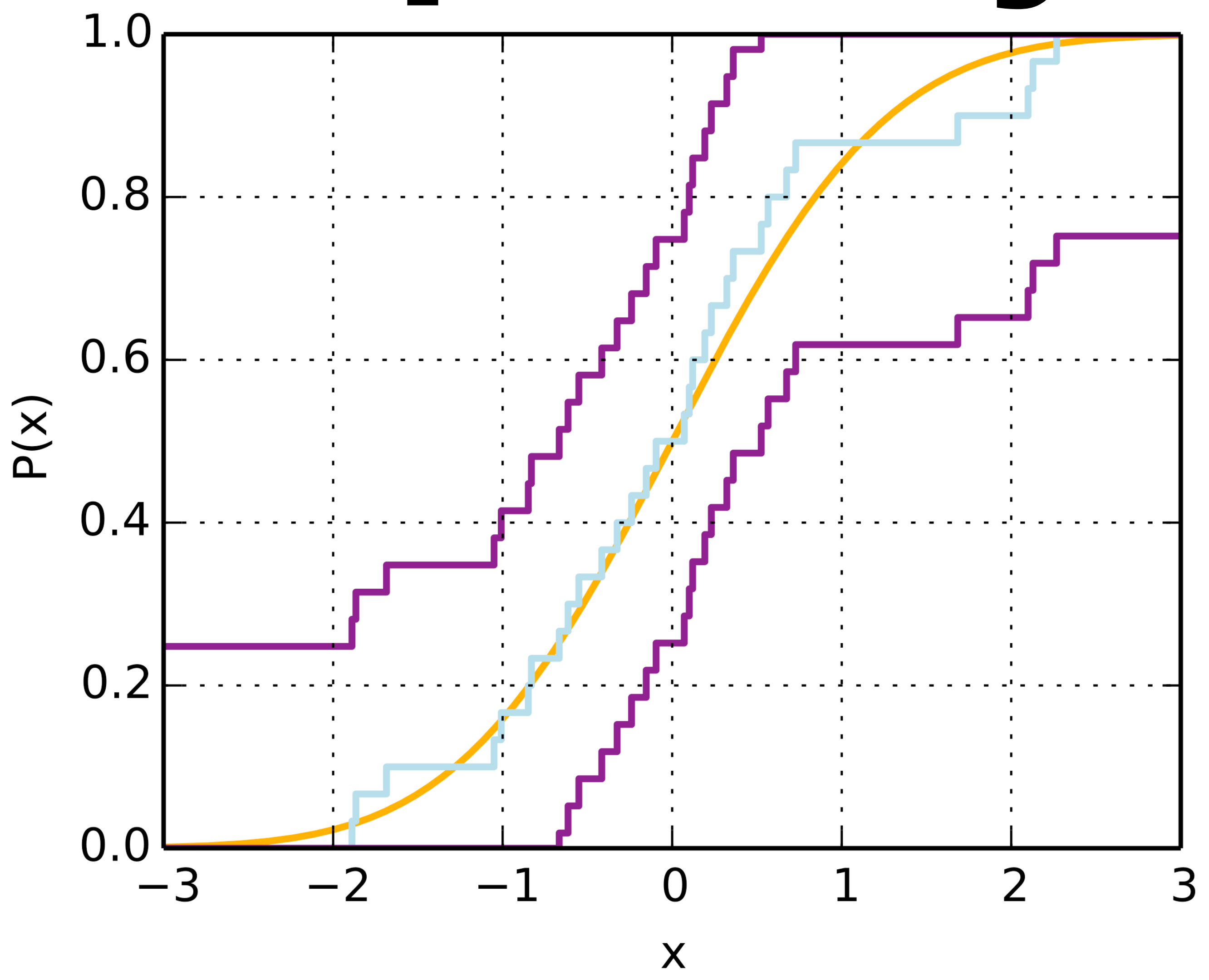
- The behaviour on $S$ is simple: $|\mathcal{G}_S| = n + 1$



from Wikipedia on DKW inequality

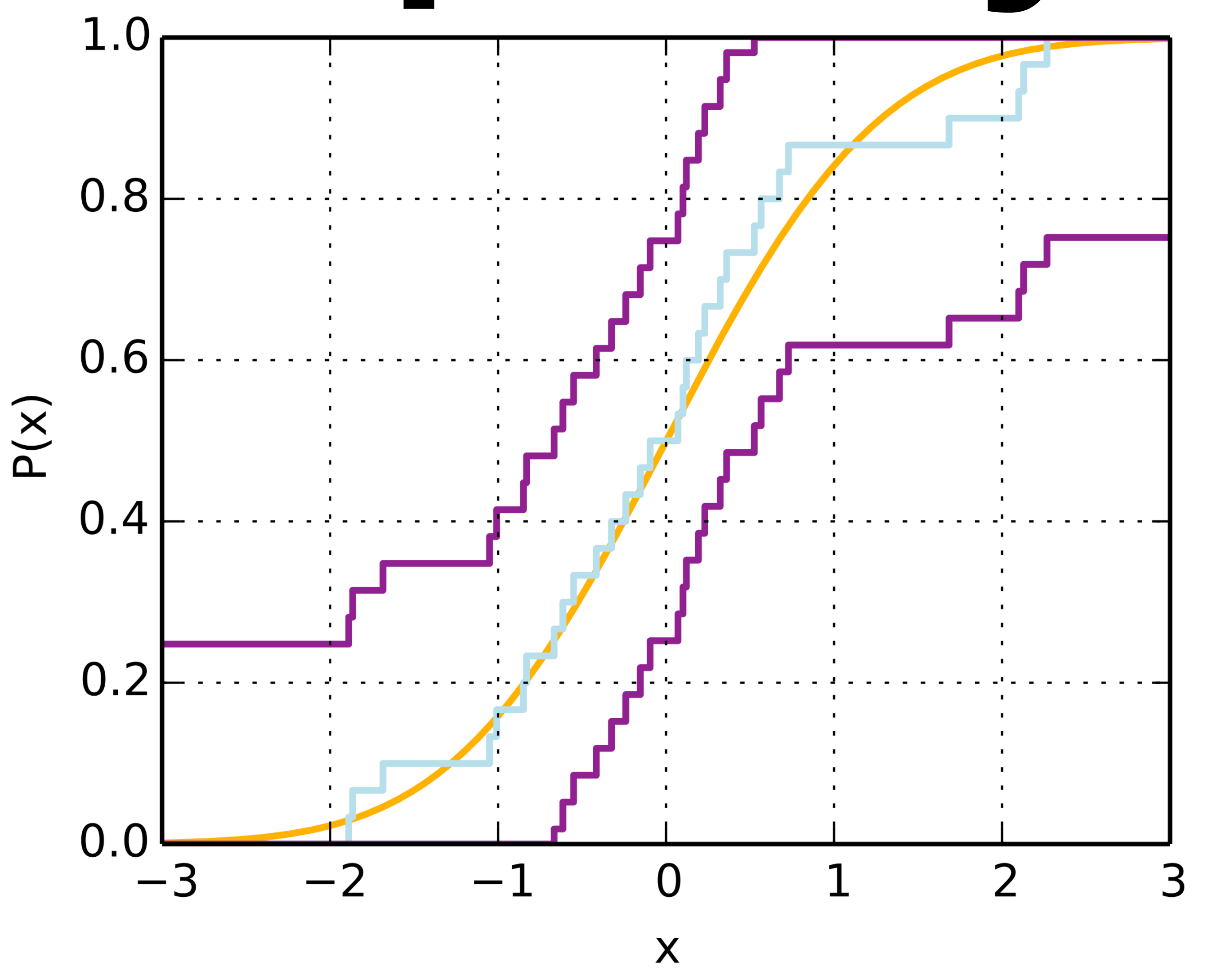
# Glivenko-Cantelli / DKW inequality

- Let $F(t) = \Pr(X \le t)$ be CDF of some random variable $X$
- $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \le t)$ is empirical CDF of samples
- Worst-case error of $F_n$ is, letting $\mathcal{G} = \{x \mapsto \mathbb{I}(x \le t) : t \in \mathbb{R}\}$,
  $\sup_t |F(t) - F_n(t)| \le 2\hat{\mathfrak{R}}'_S(\mathcal{G}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$
- The behaviour on $S$ is simple: $|\mathcal{G}_S| = n + 1$
  - Plugging in $\hat{\mathfrak{R}}'_S$ version of Massart's lemma, $\hat{\mathfrak{R}}'_S(\mathcal{G}) \le \sqrt{\frac{2}{n}\log(2(n+1))}$



from Wikipedia on DKW inequality

# Glivenko-Cantelli / DKW inequality

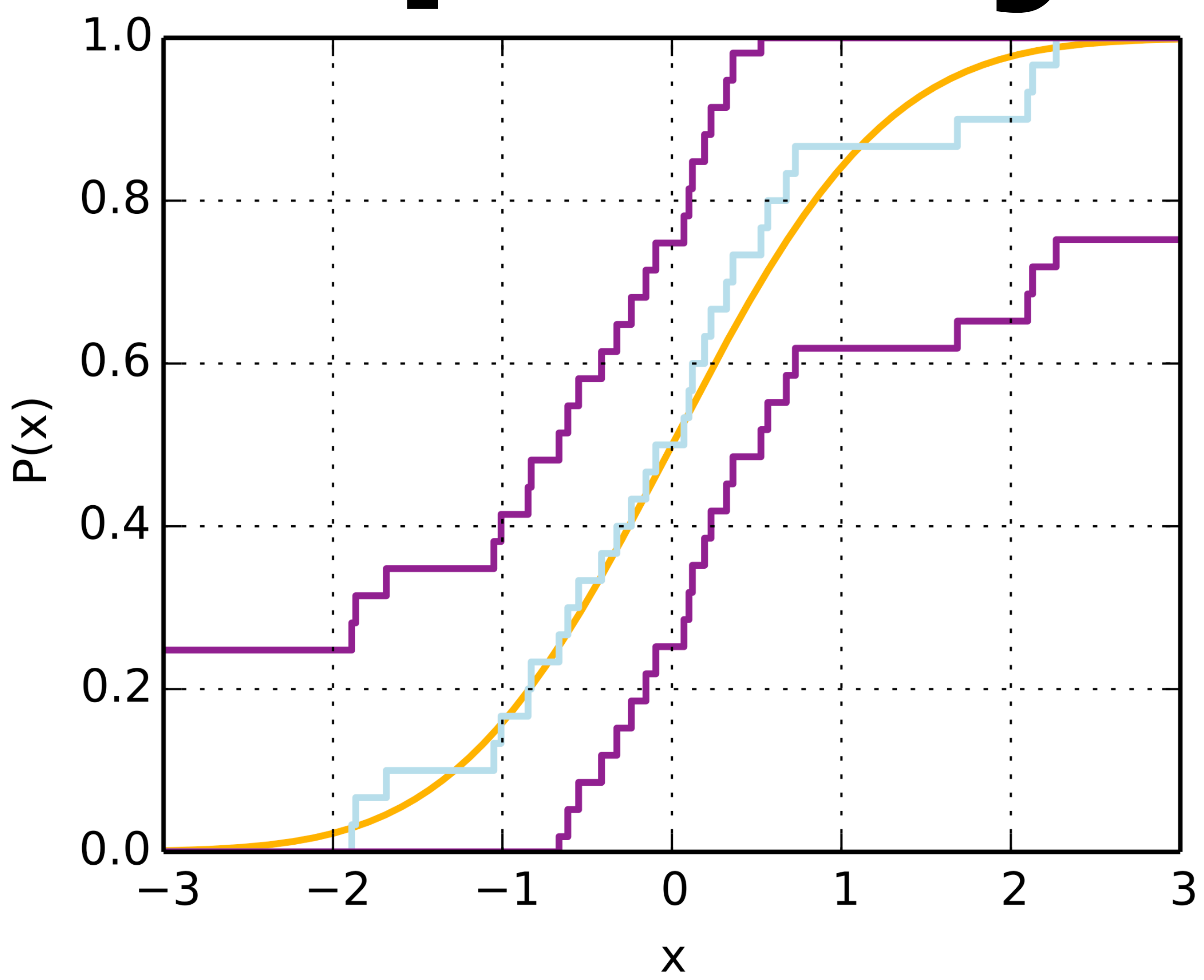


from Wikipedia on DKW inequality

- Let $F(t) = \Pr(X \leq t)$ be CDF of some random variable $X$
- $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \leq t)$ is empirical CDF of samples
- Worst-case error of $F_n$ is, letting $\mathcal{G} = \{x \mapsto \mathbb{I}(x \leq t) : t \in \mathbb{R}\}$,
  $\sup_t |F(t) - F_n(t)| \leq 2\hat{\mathfrak{R}}'_S(\mathcal{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$
- The behaviour on $S$ is simple: $|\mathcal{G}_S| = n + 1$
  - Plugging in $\hat{\mathfrak{R}}'_S$ version of Massart's lemma, $\hat{\mathfrak{R}}'_S(\mathcal{G}) \leq \sqrt{\frac{2}{n} \log(2(n+1))}$
  - $\sup_t |F(t) - F_n(t)| \leq \sqrt{\frac{8}{n} \log(2n+2)} + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$; known that $\sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$ is sharp

# Glivenko-Cantelli / DKW inequality

- Let $F(t) = \Pr(X \leq t)$ be CDF of some random variable $X$
- $F_n(t) = \dfrac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \leq t)$ is empirical CDF of samples
- Worst-case error of $F_n$ is, letting $\mathcal{G} = \{x \mapsto \mathbb{I}(x \leq t) : t \in \mathbb{R}\}$,
  $\sup_t |F(t) - F_n(t)| \leq 2\hat{\mathfrak{R}}'_S(\mathcal{G}) + \sqrt{\dfrac{1}{2n}\log\dfrac{1}{\delta}}$
- The behaviour on $S$ is simple: $|\mathcal{G}_S| = n + 1$
  - Plugging in $\hat{\mathfrak{R}}'_S$ version of Massart's lemma, $\hat{\mathfrak{R}}'_S(\mathcal{G}) \leq \sqrt{\dfrac{2}{n}\log(2(n+1))}$
  - $\sup_t |F(t) - F_n(t)| \leq \sqrt{\dfrac{8}{n}\log(2n+2)} + \sqrt{\dfrac{1}{2n}\log\dfrac{1}{\delta}}$; known that $\sqrt{\dfrac{1}{2n}\log\dfrac{2}{\delta}}$ is sharp
- For any fixed $\varepsilon$, $\Pr(\sup_t |F(t) - F_n(t)| \geq \varepsilon) \to 0$ as $n \to \infty$



from Wikipedia on DKW inequality

(pause)

# Lipschitz losses

if $\varphi : \mathbb{R} \to \mathbb{R}$ differentiable, then it is $\left(\sup_x |\varphi'(x)|\right)$-Lipschitz

- Recall a $\rho$-Lipschitz function $\phi$ has $\|\phi(x) - \phi(x')\| \leq \rho \|x - x'\|$

- **Contraction Lemma** (aka **Talagrand's Lemma**):
Let $\boldsymbol{\phi} \circ \mathscr{H} = \left\{ \left( \phi_1(h(z_1)), \ldots, \phi_n(h(z_n)) \right) : h \in \mathscr{H} \right\}$ for $\phi_i : \mathbb{R} \to \mathbb{R}$, $i \in [n]$.
If the $\phi_i$ are each $\rho$-Lipschitz, then $\hat{\mathfrak{R}}_S(\boldsymbol{\phi} \circ \mathscr{H}) \leq \rho \, \hat{\mathfrak{R}}_S(\mathscr{H})$.

$\varphi_1(\underbrace{h(x_1)}) = |\underbrace{h(x_1)} - y_1|$

$\varphi_1(\hat{y}_1) = |\hat{y}_1 - y_1|$

# Lipschitz losses

- Recall a $\rho$-Lipschitz function $\phi$ has $\|\phi(x) - \phi(x')\| \le \rho\|x - x'\|$

- **Contraction Lemma** (aka **Talagrand's Lemma**):
  Let $\boldsymbol{\phi} \circ \mathscr{H} = \left\{ \left( \phi_1(h(z_1)), \ldots, \phi_n(h(z_n)) \right) : h \in \mathscr{H} \right\}$ for $\phi_i : \mathbb{R} \to \mathbb{R}$, $i \in [n]$.
  If the $\phi_i$ are each $\rho$-Lipschitz, then $\hat{\mathfrak{R}}_S(\boldsymbol{\phi} \circ \mathscr{H}) \le \rho\, \hat{\mathfrak{R}}_S(\mathscr{H})$.

- If also $\phi_i(0) = 0$, then $\hat{\mathfrak{R}}'_S(\boldsymbol{\phi} \circ \mathscr{H}) \le 2\rho\, \hat{\mathfrak{R}}'_S(\mathscr{H})$.

# Lipschitz losses

- Recall a $\rho$-Lipschitz function $\phi$ has $\|\phi(x) - \phi(x')\| \leq \rho\|x - x'\|$
- **Contraction Lemma** (aka **Talagrand's Lemma**):
  Let $\boldsymbol{\phi} \circ \mathscr{H} = \left\{ \left( \phi_1(h(z_1)), \ldots, \phi_n(h(z_n)) \right) : h \in \mathscr{H} \right\}$ for $\phi_i : \mathbb{R} \to \mathbb{R}$, $i \in [n]$.
  If the $\phi_i$ are each $\rho$-Lipschitz, then $\hat{\mathfrak{R}}_S(\boldsymbol{\phi} \circ \mathscr{H}) \leq \rho\, \hat{\mathfrak{R}}_S(\mathscr{H})$.
- If also $\phi_i(0) = 0$, then $\hat{\mathfrak{R}}'_S(\boldsymbol{\phi} \circ \mathscr{H}) \leq 2\,\rho\, \hat{\mathfrak{R}}'_S(\mathscr{H})$.
  - If not: let $\boldsymbol{\phi}(0) = \left( \phi_1(0), \ldots, \phi_n(0) \right)$.
    Then $\hat{\mathfrak{R}}'_S(\boldsymbol{\phi} \circ \mathscr{H}) \leq 2\,\rho\, \hat{\mathfrak{R}}'_S(\mathscr{H}) + \hat{\mathfrak{R}}'_S\left( \left\{ \boldsymbol{\phi}(0) \right\} \right)$.

# Lipschitz losses

- Recall a $\rho$-Lipschitz function $\phi$ has $\|\phi(x) - \phi(x')\| \le \rho \|x - x'\|$

- **Contraction Lemma** (aka **Talagrand's Lemma**):
  Let $\boldsymbol{\phi} \circ \mathscr{H} = \left\{ \left( \phi_1(h(z_1)), \ldots, \phi_n(h(z_n)) \right) : h \in \mathscr{H} \right\}$ for $\phi_i : \mathbb{R} \to \mathbb{R}$, $i \in [n]$.
  If the $\phi_i$ are each $\rho$-Lipschitz, then $\hat{\mathfrak{R}}_S(\boldsymbol{\phi} \circ \mathscr{H}) \le \rho\, \hat{\mathfrak{R}}_S(\mathscr{H})$.

- If also $\phi_i(0) = 0$, then $\hat{\mathfrak{R}}'_S(\boldsymbol{\phi} \circ \mathscr{H}) \le 2\,\rho\, \hat{\mathfrak{R}}'_S(\mathscr{H})$.
  - If not: let $\boldsymbol{\phi}(0) = \left( \phi_1(0), \ldots, \phi_n(0) \right)$.
    Then $\hat{\mathfrak{R}}'_S(\boldsymbol{\phi} \circ \mathscr{H}) \le 2\,\rho\, \hat{\mathfrak{R}}'_S(\mathscr{H}) + \hat{\mathfrak{R}}'_S\left( \{ \boldsymbol{\phi}(0) \} \right)$.

- Proof is kind of annoying;
  MRT Lemma 5.7 or SSBD Lemma 26.9 for $\hat{\mathfrak{R}}_S$ (manageable),
  really long case analysis for $\hat{\mathfrak{R}}'_S$ (Theorem 4.12 of Ledoux & Talagrand [log in with UBC])

# Bounded regression bounds

- In (scalar) regression: $y \in \mathbb{R}$, use $h : \mathscr{X} \to \mathbb{R}$ instead of to a class label

# Bounded regression bounds

- In (scalar) regression: $y \in \mathbb{R}$, use $h : \mathcal{X} \to \mathbb{R}$ instead of to a class label
- Usual loss: $\ell(h, (x, y)) = (h(x) - y)^p$, usually for $p = 2$, sometimes $p = 1$
  - Get "partial credit" for being closer
  - Linear regression (ordinary least squares) uses $p = 2$

# Bounded regression bounds

- In (scalar) regression: $y \in \mathbb{R}$, use $h : \mathscr{X} \to \mathbb{R}$ instead of to a class label
- Usual loss: $\ell(h, (x, y)) = (h(x) - y)^p$, usually for $p = 2$, sometimes $p = 1$
  - Get "partial credit" for being closer
  - Linear regression (ordinary least squares) uses $p = 2$

- Assumption for today: $y$ and $h$ are each **bounded** in the interval $[A, A + B]$

# Bounded regression bounds

- In (scalar) regression: $y \in \mathbb{R}$, use $h : \mathcal{X} \to \mathbb{R}$ instead of to a class label
- Usual loss: $\ell(h, (x, y)) = (h(x) - y)^p$, usually for $p = 2$, sometimes $p = 1$
  - Get “partial credit” for being closer
  - Linear regression (ordinary least squares) uses $p = 2$

- Assumption for today: $y$ and $h$ are each **bounded** in the interval $[A, A + B]$
- Then $\hat{y} \mapsto |\hat{y} - y|^p$ is $(pB^{p-1})$-Lipschitz on relevant domain

$$\left| \frac{d}{d\hat{y}} \eta \right| = \left| \underbrace{p}_{p} \underbrace{|\hat{y} - y|^{p-1}}_{\leq B^{p-1}} \underbrace{\operatorname{sign}(\hat{y} - y)}_{1} \right|$$

# Bounded regression bounds

- In (scalar) regression: $y \in \mathbb{R}$, use $h : \mathcal{X} \to \mathbb{R}$ instead of to a class label
- Usual loss: $\ell(h, (x, y)) = (h(x) - y)^p$, usually for $p = 2$, sometimes $p = 1$
  - Get "partial credit" for being closer
  - Linear regression (ordinary least squares) uses $p = 2$

- Assumption for today: $y$ and $h$ are each **bounded** in the interval $[A, A + B]$
- Then $\hat{y} \mapsto |\hat{y} - y|^p$ is $(pB^{p-1})$-Lipschitz on relevant domain
- So $\mathcal{G} = \{(x, y) \mapsto |h(x) - y|^p : h \in \mathcal{H}\}$ has $\hat{\mathfrak{R}}_S(\mathcal{G}) \leq pB^{p-1}\hat{\mathfrak{R}}_S(\mathcal{H})$

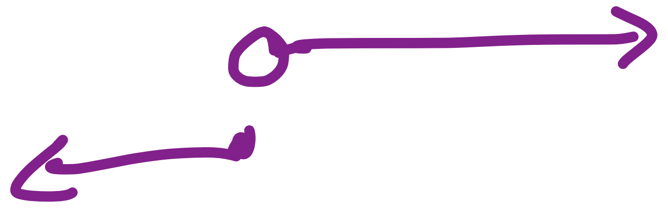
# Bounded regression bounds

- In (scalar) regression: $y \in \mathbb{R}$, use $h : \mathcal{X} \to \mathbb{R}$ instead of to a class label
- Usual loss: $\ell(h, (x, y)) = (h(x) - y)^p$, usually for $p = 2$, sometimes $p = 1$
  - Get "partial credit" for being closer
  - Linear regression (ordinary least squares) uses $p = 2$

- Assumption for today: $y$ and $h$ are each **bounded** in the interval $[A, A + B]$
- Then $\hat{y} \mapsto |\hat{y} - y|^p$ is $(pB^{p-1})$-Lipschitz on relevant domain
- So $\mathcal{G} = \{(x, y) \mapsto |h(x) - y|^p : h \in \mathcal{H}\}$ has $\hat{\mathfrak{R}}_S(\mathcal{G}) \leq pB^{p-1}\hat{\mathfrak{R}}_S(\mathcal{H})$
- Plugging into our theorem, since loss is in $[0,B^p]$:

$$\sup_{h \in \mathcal{H}} \left[L_{\mathcal{D}}(h) - L_S(h)\right] \overset{?}{\leq} pB^{p-1}\mathfrak{R}_n(\mathcal{H}) + B^p\sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

$h_{a,b,t}(x) = \begin{cases} a & \text{if } x \leq t \\ b & \text{if } x > t \end{cases}$



# $\mathfrak{R}$ of linear functions

- If $\mathscr{H} = \{x \mapsto w^\top x : \|w\| \le B\}$

$$\hat{R}_S(\mathcal{H}) = \frac{1}{n} \mathbb{E}_\sigma \sup_{\|w\| \le B} \sum_i \sigma_i \langle w, x_i \rangle$$

$$= \frac{1}{n} \mathbb{E} \sup \langle w, \sum_i \sigma_i x_i \rangle$$

$$\le \frac{1}{n} \mathbb{E} \sup_{\|w\| \le B} \underbrace{\|w\|}_{\le B} \|\sum_i \sigma_i x_i\|$$

$$\le \frac{B}{n} \mathbb{E}_\sigma \sqrt{\sum_i \sum_j \sigma_i \sigma_j \langle x_i, x_j \rangle}$$

$$\le \frac{B}{n} \sqrt{\mathbb{E}_\sigma \sum_i \sum_j \sigma_i \sigma_j \langle x_i, x_j \rangle} = \frac{B}{n} \sqrt{\sum_i \|x_i\|^2}$$

$$\mathbb{E}[\sigma_i \sigma_j] = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{o.w.} \end{cases}$$

$$\hat{R}_S(\mathcal{H}) \le \frac{B}{n} \sqrt{\sum_i \|x_i\|^2} = \frac{B}{n} \|X\|_F$$

$n \times d$ matrix

$$\therefore R_n(\mathcal{H}) \le \frac{B}{n} \mathbb{E} \|X\|_F$$

$$\le \frac{B}{n} \cdot \sqrt{n\, \mathbb{E}\|x_i\|^2}$$

any one random sample (in $\mathbb{R}^d$)

$$= \frac{B}{\sqrt{n}} \sqrt{\mathbb{E}\|x_i\|^2}$$

if $\|x_i\|^2 \le M^2$

$$\le \frac{B}{n} \sqrt{n \cdot M^2} = \frac{BM}{\sqrt{n}}$$

# Summary

- Sorted out the two definitions of Rademacher
  - One-sided is enough for upper bounds on $L_{\mathscr{D}}$
  - Need two-sided for uniform conv property, our not-quite-optimal DKW inequality

- Rademacher complexity of linear functions
- Contraction lemma
- Easy upper bounds for bounded regression problems with bounded data / $\|w\|$

- **Next time:** but how do we pick an $\mathscr{H}$ (e.g. a bound for $\|w\|$)???