More VC + Rademacher

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

Admin

- Reminder: no office hours this week (ICML…)
  - But feel free to Piazza / schedule a meeting if needed
- A1 grading: probably late next week

Admin

- Reminder: no office hours this week (ICML…)
  - But feel free to Piazza / schedule a meeting if needed
- A1 grading: probably late next week
- A2 will be released probably next week, due ~2-3 weeks after release. A weird plan:
  - Groups of up to 3 allowed
  - You can use a different group *per question* if you want
    - (e.g. do one problem alone, one with person A, one with B+C)
  - You’ll hand in questions as separate Gradescope assignments
    - (one per group per question, using Gradescope group feature)
  - Drop lowest: still for total assignment grade (or more advantageous, TBD)
  - Trying to encourage **actively participating in each question**
    - Please **don’t** just split assignment in thirds
  - Dropping lowest assignment grade is still per student
  - Will try to calibrate difficulty/length a bit, but you’ll have groups

Previously: “Fundamental Theorem of Learning”

For binary classification with 0-1 loss:

These are all equivalent:

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

Previously: “Fundamental Theorem of Learning”

For binary classification with 0-1 loss:

These are all equivalent:

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

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

Previously: "Fundamental Theorem of Learning"

For binary classification with 0-1 loss:

These are all equivalent:

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

saw today

immediate

no free lunch

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

Previously: "Fundamental Theorem of Learning"

For binary classification with 0-1 loss:

These are all equivalent:

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

saw today

immediate

no free lunch

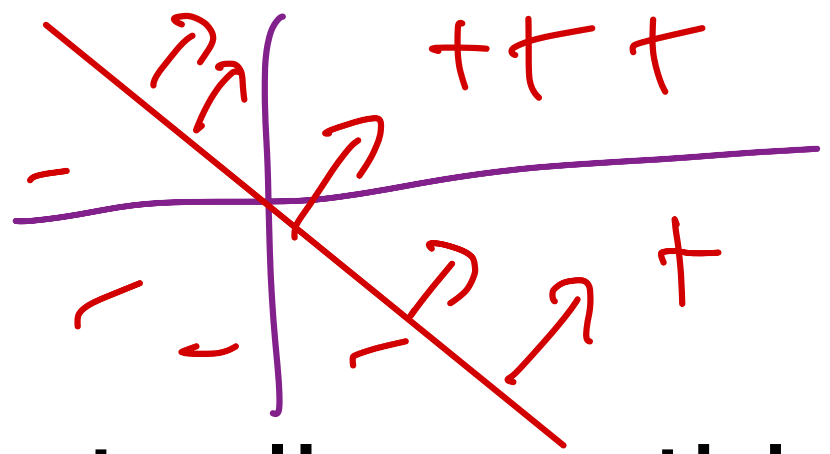
to show!

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

Last time: Finite VCdim implies uniform convergence

- **Uniform convergence**: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon$ with probability at least $1 - \delta$
- Will prove in terms of the **growth function**: $\tau_{\mathcal{H}}(n) = \max_{C \subseteq \mathcal{X}:\ |C|=n} |\mathcal{H}_C|$
  - How many *actually different* functions from $\mathcal{H}$ are there on sets of size n?
  - If $\mathrm{VCdim}(\mathcal{H}) = d$, then $\tau_{\mathcal{H}}(n) = 2^n$ for $n \leq d$
- **Theorem** (SSBD 6.11): $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}$ This is the part we didn't prove yet
- **Sauer-Shelah lemma**: when $n \geq d$, $\quad \tau_{\mathcal{H}}(n) \leq (en/d)^d = \mathcal{O}\left(n^d\right)$
- Plugging together: uniform convergence when $n \geq 4\dfrac{2d}{(\delta\varepsilon)^2} \log\left(\dfrac{2d}{(\delta\varepsilon)^2}\right) + \dfrac{4d\log(2e/d)}{(\delta\varepsilon)^2}$

VC dimension of linear classifiers



- $h_w(x) = \mathbb{I}\left(w^\top x \geq 0\right)$ on $\mathbb{R}^d$
  - Can shatter, e.g., $\{e_1, e_2, \ldots, e_d\}$ (actually, anything in "general position")

$(1, 0, \cdots, 0) \in \mathbb{R}^d$

$w = (2y_1 - 1, \cdots, 2y_d - 1)$

$w^T e_i = 2y_i - 1$

$\Sigma^{+}_{ii} = \begin{cases} 1/\Sigma_{ii} & \text{if } \Sigma_{ii} > 0 \\ 0 & \text{if } \Sigma_{ii} = 0 \end{cases}$

let $\underset{n\times d}{X} = U \Sigma V^T$ be its SVD

$\underset{d\times 1}{w} = \underbrace{\underset{d\times r}{V}\, \underset{r\times r}{\Sigma^{-1}}\, \underset{r\times n}{U^T}}_{X^{+}\ \text{(pseudo-inverse)}}\, \underset{n\times 1}{Y}$ $\leftarrow (y_1, \ldots, y_n)$, $y_i \in \{-1, 1\}$

$r \leq \min(n, d)$

$Xw = X V \Sigma^{-1} U^T y$

$= U \Sigma \underbrace{V^T V}_{I} \Sigma^{-1} U^T y = \overbrace{U U^T}^{\text{projection data rank-}r\text{ subspace}} \underset{n\times 1}{y} = y$ IF $r = n$

VC dimension of linear classifiers

- $h_w(x) = \mathbb{I}\left(w^\top x \geq 0\right)$ on $\mathbb{R}^d$
  - Can't shatter anything of size $d+1$

$x_1, \ldots, x_{d+1}$

not all $a_i = 0$

$$\sum_{i=1}^{d+1} a_i x_i = 0$$

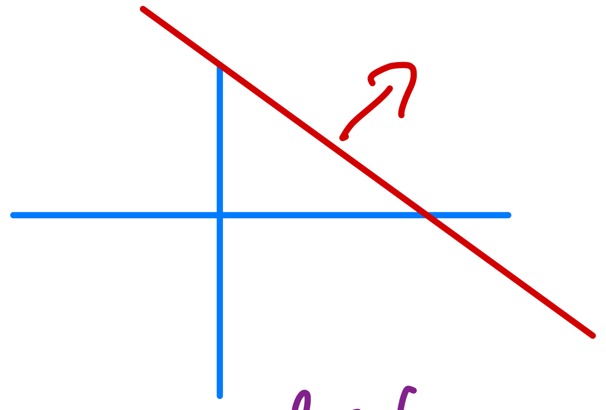
what about all $a_i \geq 0$, $w^\top x_i = 0$

$$\sum_{i:a_i>0} a_i x_i = \sum_{i:a_i<0} (-a_i) x_i$$

$w^\top x_i \geq 0$ $\qquad$ $w^\top x_i < 0$

$$0 < \sum_{i:a_i>0} \underbrace{a_i}_{>0} \underbrace{x_i^\top w}_{\geq 0} = \sum_{i:a_i<0} \underbrace{(-a_i)}_{>0} \underbrace{x_i^\top w}_{<0} < 0$$

VC dimension of non-homogenous linear classifiers



- $h_w(x) = \mathbb{1}\left(w_0 + w^\top x \geq 0\right)$

$\tilde{x} \Rightarrow (1, x) \in \mathbb{R}^{d+1}$

$\tilde{w} = (w_0, w) \in \mathbb{R}^{d+1}$

Can't shatter sets of size $d+2$

Can shatter $\{e_1, \dots, e_{d+1}\}$

VC dim $= d+1$

M-ing the ER for halfspaces

- Just showed that ERM will agnostically PAC-learn a linear classifier (halfspace) with $\Omega\left(\frac{1}{\varepsilon}\left[d+\log\frac{1}{\delta}\right]\right)$ samples (with 0-1 loss)

M-ing the ER for halfspaces

- Just showed that ERM will agnostically PAC-learn a linear classifier (halfspace) with $\Omega\left(\frac{1}{\varepsilon}\left[d+\log\frac{1}{\delta}\right]\right)$ samples (with 0-1 loss)

- But…turns out that finding the optimal halfspace (for 0-1 loss) is NP-hard to even *approximate* (i.e. to get loss $(1+\gamma)\,L^*$)

[Ben-David+Simon, NeurIPS 2000]

M-ing the ER for halfspaces

- Just showed that ERM will agnostically PAC-learn a linear classifier (halfspace) with $\Omega\left(\frac{1}{\varepsilon}\left[d + \log\frac{1}{\delta}\right]\right)$ samples (with 0-1 loss)

- But…turns out that finding the optimal halfspace (for 0-1 loss) is NP-hard to even *approximate* (i.e. to get loss $(1+\gamma)\,L^*$)
  [Ben-David+Simon, NeurIPS 2000]

- In the realizable (separable) case ($L_S(h_S) = 0$), easy algorithms in polynomial time
  - Perceptron
  - Linear programming
  - Logistic regression
  - SVMs
  - …

(pause)

Generalization bound from growth functions

- **SSBD's theorem 6.11**: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}$

  with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^n$, for any $\mathcal{D}$

Generalization bound from growth functions

- **SSBD's theorem 6.11**: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}$

  with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^n$, for any $\mathcal{D}$

- Follows from $\mathbb{E} \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\sqrt{2n}}$

Generalization bound from growth functions

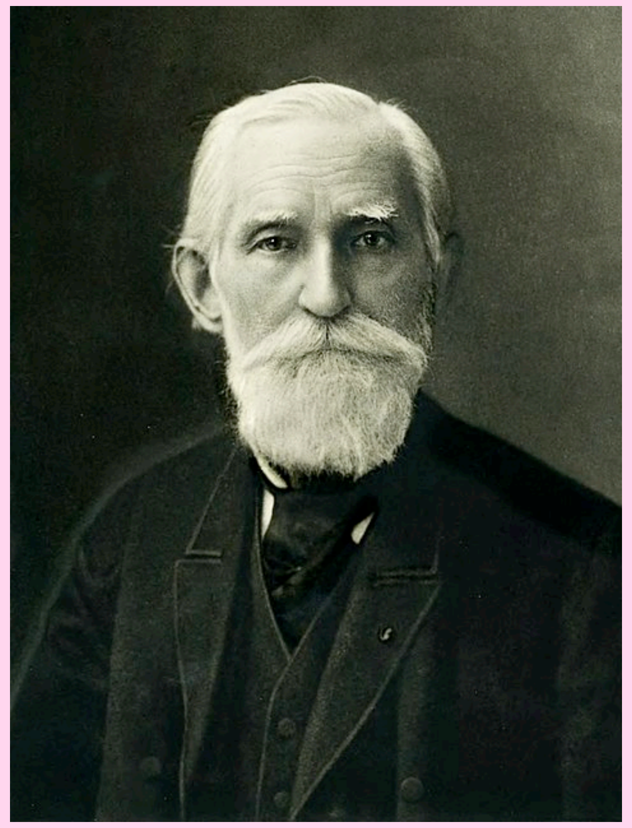
- **SSBD's theorem 6.11**: $\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathscr{H}}(2n))}}{\delta\sqrt{2n}}$

  with probability at least $1 - \delta$ over the choice of $S \sim \mathscr{D}^n$, for any $\mathscr{D}$

- Follows from $\mathbb{E} \sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathscr{H}}(2n))}}{\sqrt{2n}}$

Markov's inequality (~1860s)

Andrey Markov



Pafnuty Chebyshev

Generalization bound from growth functions

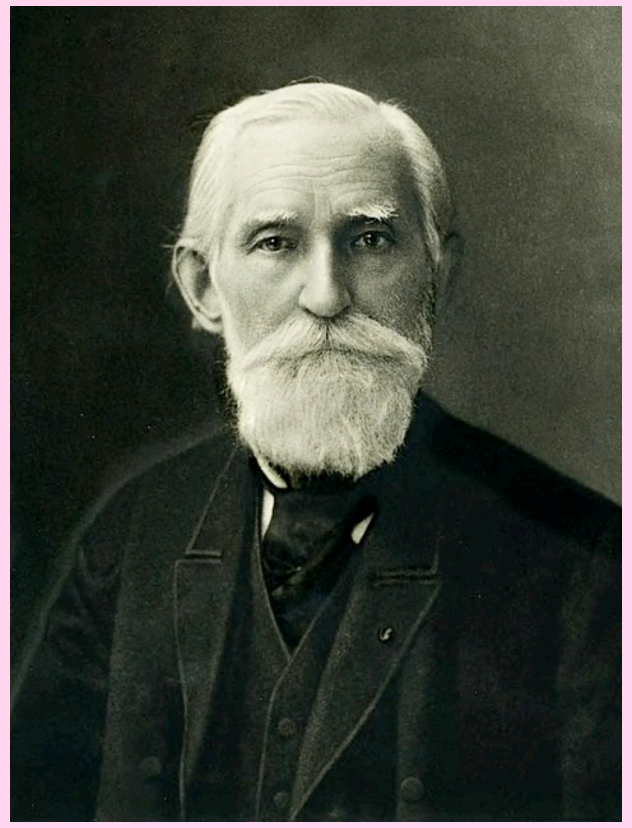
- **SSBD's theorem 6.11**: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}$

  with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^n$, for any $\mathcal{D}$

- Follows from $\mathbb{E} \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\sqrt{2n}}$

Markov's inequality (~1860s)

Andrey Markov



Pafnuty Chebyshev

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

Generalization bound from growth functions

- **SSBD's theorem 6.11**: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \le \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}$

  with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^n$, for any $\mathcal{D}$

- Follows from $\mathbb{E} \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \le \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\sqrt{2n}}$

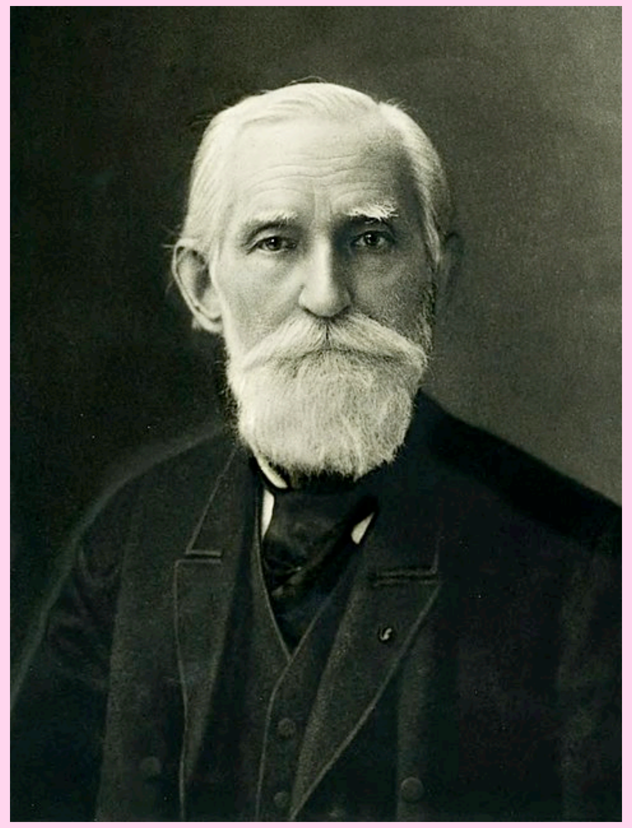
$\frac{1}{\delta} = e^{a}$

$\log \frac{1}{\delta} = 9$

$\sqrt{\log \frac{1}{\delta}} = 3$

Markov's inequality (~1860s)

Andrey Markov



Pafnuty Chebyshev

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

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

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

but, actually, we’re going to do better than that but, actually, we're going to do better than that

using (a) Rademacher complexity
and (b) McDiarmid's inequality rather than Markov's

but, actually, we're going to do better than that

using (a) Rademacher complexity
and (b) McDiarmid's inequality rather than Markov's

the overall proof uses the same core techniques,
and the basic version gets *much* closer to the optimal rate,
using machinery that we were going to do pretty soon anyway

most of this is in SSBD chapter 26,
but today's presentation will more or less follow MRT chapter 3

Rademacher complexities

- Measure the complexity of a set of functions $\mathcal{G}$ from $\mathcal{Z}$ to $\mathbb{R}$
  - For example, could have $\mathcal{Z} = \mathcal{X}$ and $\mathcal{G} = \mathcal{H}$ of binary classifiers
  - But (unlike VC), will be easy to use for much more general settings

Rademacher complexities

- Measure the complexity of a set of functions $\mathcal{G}$ from $\mathcal{Z}$ to $\mathbb{R}$
  - For example, could have $\mathcal{Z} = \mathcal{X}$ and $\mathcal{G} = \mathcal{H}$ of binary classifiers
  - But (unlike VC), will be easy to use for much more general settings

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

Rademacher complexities

- Measure the complexity of a set of functions $\mathcal{G}$ from $\mathcal{Z}$ to $\mathbb{R}$
  - For example, could have $\mathcal{Z} = \mathcal{X}$ and $\mathcal{G} = \mathcal{H}$ of binary classifiers
  - But (unlike VC), will be easy to use for much more general settings

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

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

Rademacher complexities

- Measure the complexity of a set of functions $\mathcal{G}$ from $\mathcal{Z}$ to $\mathbb{R}$
  - For example, could have $\mathcal{Z} = \mathcal{X}$ and $\mathcal{G} = \mathcal{H}$ of binary classifiers
  - But (unlike VC), will be easy to use for much more general settings

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

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

$\boldsymbol{\sigma} \sim \mathrm{Rad}^n$

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

Rademacher complexities

- Measure the complexity of a set of functions $\mathcal{G}$ from $\mathcal{Z}$ to $\mathbb{R}$
  - For example, could have $\mathcal{Z} = \mathcal{X}$ and $\mathcal{G} = \mathcal{H}$ of binary classifiers
  - But (unlike VC), will be easy to use for much more general settings

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

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

$\boldsymbol{\sigma} \sim \mathrm{Rad}^n$

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

Rademacher complexities

- Measure the complexity of a set of functions $\mathcal{G}$ from $\mathcal{Z}$ to $\mathbb{R}$
  - For example, could have $\mathcal{Z} = \mathcal{X}$ and $\mathcal{G} = \mathcal{H}$ of binary classifiers
  - But (unlike VC), will be easy to use for much more general settings

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

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

$\boldsymbol{\sigma} \sim \mathrm{Rad}^n$

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

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

Rademacher complexities

- Measure the complexity of a set of functions $\mathscr{G}$ from $\mathscr{Z}$ to $\mathbb{R}$
  - For example, could have $\mathscr{Z} = \mathscr{X}$ and $\mathscr{G} = \mathscr{H}$ of binary classifiers
  - But (unlike VC), will be easy to use for much more general settings

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

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

$\boldsymbol{\sigma} \sim \mathrm{Rad}^n$

$\sigma_i \sim \mathrm{Rad}$ 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 $\mathfrak{R}_n(\mathscr{G}) = \mathbb{E}_{S \sim \mathscr{D}^n}\left[\widehat{\mathfrak{R}}_S(\mathscr{G})\right]$
  - Distribution-dependent notion of complexity!

$\mathfrak{R}$ of singleton sets

- If $\mathscr{H} = \{h\}$

$$\hat{R}_S(\mathcal{H}) = \mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i \sigma_i h(z_i)$$

$$= \mathbb{E}_\sigma \frac{1}{n} \sum_i \sigma_i h(z_i)$$

$$= \frac{1}{n} \sum_i \underbrace{\mathbb{E}[\sigma_i]}_{0} h(z_i)$$

$$= 0$$

$\mathfrak{R}$ of linear functions

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

Turns out $\hat{R}_S(\mathscr{H}) \leq \dfrac{2B \sup_{x \in \mathcal{X}} \|x\|}{\sqrt{n}}$

Pascal Massart

Relating back to VC dim

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

Pascal Massart

Relating back to VC dim

- **Massart's lemma**: for $\mathscr{A} \subset \mathbb{R}^n$, if $\max_{a \in \mathscr{A}} \|a\| \leq r$, $\quad \mathbb{E}_{\boldsymbol{\sigma}}\Big[\max_{a \in \mathscr{A}} \frac{1}{n} \boldsymbol{\sigma}^\top a\Big] \leq \frac{1}{n} r \sqrt{2 \log |\mathscr{A}|}$
  - (Proof is a nice + not too complicated result on concentration of max of sums; we'll come back to it)

Pascal Massart

Relating back to VC dim

- **Massart's lemma**: for $\mathscr{A} \subset \mathbb{R}^n$, if $\max_{a \in \mathscr{A}} \|a\| \leq r$, $\quad \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{a \in \mathscr{A}} \frac{1}{n} \boldsymbol{\sigma}^\top a\right] \leq \frac{1}{n} r \sqrt{2 \log |\mathscr{A}|}$
  - (Proof is a nice + not too complicated result on concentration of max of sums; we'll come back to it)
- For binary classifiers, with output in $\{0,1\}$ or $\{-1,1\}$:

Pascal Massart

Relating back to VC dim

- **Massart's lemma**: for $\mathscr{A} \subset \mathbb{R}^n$, if $\max_{a\in\mathscr{A}} \|a\| \leq r$, $\quad \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{a\in\mathscr{A}} \frac{1}{n}\boldsymbol{\sigma}^\top a\right] \leq \frac{1}{n} r\sqrt{2\log|\mathscr{A}|}$
  - (Proof is a nice + not too complicated result on concentration of max of sums; we'll come back to it)

- For binary classifiers, with output in $\{0,1\}$ or $\{-1,1\}$:
  - Recall $\mathscr{H}_S = \left\{ \big(h(x_1), \ldots, h(x_n)\big) : h \in \mathscr{H} \right\} = \{\mathbf{h}_S : h \in \mathscr{H}\}$

Pascal Massart

Relating back to VC dim

- **Massart's lemma**: for $\mathscr{A} \subset \mathbb{R}^n$, if $\max_{a \in \mathscr{A}} \|a\| \leq r$, $\quad \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{a \in \mathscr{A}} \frac{1}{n}\boldsymbol{\sigma}^\top a\right] \leq \frac{1}{n} r \sqrt{2 \log |\mathscr{A}|}$
  - (Proof is a nice + not too complicated result on concentration of max of sums; we'll come back to it)

- For binary classifiers, with output in $\{0,1\}$ or $\{-1,1\}$:
  - Recall $\mathscr{H}_S = \left\{ \left(h(x_1), \ldots, h(x_n)\right) : h \in \mathscr{H} \right\} = \{\mathbf{h}_S : h \in \mathscr{H}\}$
  - But $\widehat{\mathfrak{R}}_S(\mathscr{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathscr{H}} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{h}_S\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{\mathbf{h}_S \in \mathscr{H}_S} \frac{1}{n}\boldsymbol{\sigma}^\top \mathbf{h}_S\right]$

Pascal Massart

Relating back to VC dim

- **Massart's lemma**: for $\mathscr{A} \subset \mathbb{R}^n$, if $\max_{a \in \mathscr{A}} \|a\| \leq r$, $\quad \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{a \in \mathscr{A}} \frac{1}{n} \boldsymbol{\sigma}^\top a \right] \leq \frac{1}{n} r \sqrt{2 \log |\mathscr{A}|}$
  - (Proof is a nice + not too complicated result on concentration of max of sums; we'll come back to it)

- For binary classifiers, with output in $\{0,1\}$ or $\{-1,1\}$:
  - Recall $\mathscr{H}_S = \left\{ \left(h(x_1), \ldots, h(x_n) \right) : h \in \mathscr{H} \right\} = \{ \mathbf{h}_S : h \in \mathscr{H} \}$
  - But $\widehat{\mathfrak{R}}_S(\mathscr{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathscr{H}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{h}_S \right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{\mathbf{h}_S \in \mathscr{H}_S} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{h}_S \right]$

$$\leq \frac{1}{n} \sqrt{n} \sqrt{2 \log |\mathscr{H}_S|}$$

Massart's lemma, using
$\|\mathbf{h}_S\| \leq \sqrt{1^2 + \cdots + 1^2} = \sqrt{n}$

Pascal Massart

Relating back to VC dim

- **Massart's lemma**: for $\mathscr{A} \subset \mathbb{R}^n$, if $\max_{a \in \mathscr{A}} \|a\| \le r$, $\quad \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{a \in \mathscr{A}} \frac{1}{n} \boldsymbol{\sigma}^\top a\right] \le \frac{1}{n} r \sqrt{2 \log |\mathscr{A}|}$
  - (Proof is a nice + not too complicated result on concentration of max of sums; we'll come back to it)

- For binary classifiers, with output in $\{0,1\}$ or $\{-1,1\}$:
  - Recall $\mathscr{H}_S = \left\{ \left(h(x_1), \ldots, h(x_n) \right) : h \in \mathscr{H} \right\} = \{ \mathbf{h}_S : h \in \mathscr{H} \}$
  - But $\widehat{\mathfrak{R}}_S(\mathscr{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathscr{H}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{h}_S\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{\mathbf{h}_S \in \mathscr{H}_S} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{h}_S\right]$

$$\le \frac{1}{n}\sqrt{n}\sqrt{2\log|\mathscr{H}_S|} \le \sqrt{\frac{2}{n}\log \tau_{\mathscr{H}}(n)}$$

Massart's lemma, using
$\|\mathbf{h}_S\| \le \sqrt{1^2 + \cdots + 1^2} = \sqrt{n}$

by definition of the growth function
$\tau_{\mathscr{H}}(n) = \sup_{|S|=n} |\mathscr{H}_S|$

Pascal Massart

Relating back to VC dim

- **Massart's lemma**: for $\mathscr{A} \subset \mathbb{R}^n$, if $\max_{a \in \mathscr{A}} \|a\| \leq r$, $\quad \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{a \in \mathscr{A}} \frac{1}{n} \boldsymbol{\sigma}^\top a\right] \leq \frac{1}{n} r \sqrt{2 \log |\mathscr{A}|}$
  - (Proof is a nice + not too complicated result on concentration of max of sums; we'll come back to it)

- For binary classifiers, with output in $\{0,1\}$ or $\{-1,1\}$:
  - Recall $\mathscr{H}_S = \left\{ \left(h(x_1), \ldots, h(x_n)\right) : h \in \mathscr{H} \right\} = \{\mathbf{h}_S : h \in \mathscr{H}\}$
  - But $\widehat{\mathfrak{R}}_S(\mathscr{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathscr{H}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{h}_S\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{\mathbf{h}_S \in \mathscr{H}_S} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{h}_S\right]$

$$\leq \frac{1}{n}\sqrt{n}\sqrt{2\log|\mathscr{H}_S|} \leq \sqrt{\frac{2}{n}\log \tau_{\mathscr{H}}(n)} \leq \sqrt{\frac{2}{n} d\left[1 + \log n - \log d\right]}$$

Massart's lemma, using $\|\mathbf{h}_S\| \leq \sqrt{1^2 + \cdots + 1^2} = \sqrt{n}$

by definition of the growth function $\tau_{\mathscr{H}}(n) = \sup_{|S|=n} |\mathscr{H}_S|$

where d is the VC dimension of $\mathscr{H}$ and $n \geq d$ (Sauer-Shelah)

Pascal Massart

Relating back to VC dim

- **Massart's lemma**: for $\mathscr{A} \subset \mathbb{R}^n$, if $\max_{a \in \mathscr{A}} \|a\| \leq r$, $\quad \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{a \in \mathscr{A}} \frac{1}{n} \boldsymbol{\sigma}^\top a\right] \leq \frac{1}{n} r \sqrt{2 \log |\mathscr{A}|}$
  - (Proof is a nice + not too complicated result on concentration of max of sums; we'll come back to it)

- For binary classifiers, with output in $\{0,1\}$ or $\{-1,1\}$:
  - Recall $\mathscr{H}_S = \left\{ \left(h(x_1), \ldots, h(x_n)\right) : h \in \mathscr{H} \right\} = \left\{ \mathbf{h}_S : h \in \mathscr{H} \right\}$
  - But $\widehat{\mathfrak{R}}_S(\mathscr{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in \mathscr{H}} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{h}_S\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{\mathbf{h}_S \in \mathscr{H}_S} \frac{1}{n} \boldsymbol{\sigma}^\top \mathbf{h}_S\right]$

where d is the VC dimension of $\mathscr{H}$ and $n \geq d$ (Sauer-Shelah)

$$\leq \frac{1}{n}\sqrt{n}\sqrt{2 \log |\mathscr{H}_S|} \quad \leq \sqrt{\frac{2}{n} \log \tau_{\mathscr{H}}(n)} \quad \leq \sqrt{\frac{2}{n} d \left[1 + \log n - \log d\right]}$$

Massart's lemma, using $\|\mathbf{h}_S\| \leq \sqrt{1^2 + \cdots + 1^2} = \sqrt{n}$

by definition of the growth function $\tau_{\mathscr{H}}(n) = \sup_{|S|=n} |\mathscr{H}_S|$

if $d \geq 3$ $\quad \leq \sqrt{\dfrac{2d \log n}{n}}$

- Our goal was to upper bound $\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)|$ in terms of VC dimension

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \ \mathrm{VCdim}(\mathcal{H}) \ \log(n) / n}$

- Our goal was to upper bound $\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathscr{H}) \le \sqrt{2 \text{ VCdim}(\mathscr{H}) \log(n)/n}$
- One more step: consider $\mathfrak{R}_n(\mathscr{G})$ for $\mathscr{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathscr{H} \right\}$

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n) / n}$
- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n) / n}$
- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathcal{H} \right\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n) / n}$
- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathcal{H} \right\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- Our goal was to upper bound $\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathscr{H}) \leq \sqrt{2 \text{ VCdim}(\mathscr{H}) \log(n)/n}$
- One more step: consider $\mathfrak{R}_n(\mathscr{G})$ for $\mathscr{G} = \left\{z \mapsto \ell(h, z) : h \in \mathscr{H}\right\}$
  - $g_h : \mathscr{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathscr{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathscr{H})$ (proof is next)

  or $\frac{1}{2}$ that, if $\mathscr{H}$ maps to ± 1
- **Theorem**: for $\mathscr{G}$ mapping to $[0,1]$, prob is $\geq 1 - \delta$ over $S \sim \mathscr{D}^n$ that for **all** $g \in \mathscr{G}$

- Our goal was to upper bound $\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathscr{H}) \le \sqrt{2 \ \mathrm{VCdim}(\mathscr{H}) \ \log(n) / n}$
- One more step: consider $\mathfrak{R}_n(\mathscr{G})$ for $\mathscr{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathscr{H} \right\}$
  - $g_h : \mathscr{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \ne y)$, and $\widehat{\mathfrak{R}}_S(\mathscr{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathscr{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathscr{H}$ maps to ± 1

- **Theorem**: for $\mathscr{G}$ mapping to $[0,1]$, prob is $\ge 1 - \delta$ over $S \sim \mathscr{D}^n$ that for **all** $g \in \mathscr{G}$
  - $\mathbb{E}[g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \le 2\mathfrak{R}_n(\mathscr{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n)/n}$

- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- **Theorem**: for $\mathcal{G}$ mapping to $[0,1]$, prob is $\geq 1 - \delta$ over $S \sim \mathcal{D}^n$ that for **all** $g \in \mathcal{G}$
  - $\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i) \leq 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$ $\leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\left(\sqrt{\text{VCdim}} + \sqrt{\log\frac{1}{\delta}}\right)\right)$

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \ \mathrm{VCdim}(\mathcal{H}) \ \log(n) / n}$
- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- **Theorem**: for $\mathcal{G}$ mapping to $[0,1]$, prob is $\geq 1 - \delta$ over $S \sim \mathcal{D}^n$ that for **all** $g \in \mathcal{G}$
  - $\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i) \leq 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$ $\leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\left(\sqrt{\mathrm{VCdim}} + \sqrt{\log\frac{1}{\delta}}\right)\right)$

$L_{\mathcal{D}}(h)$

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n) / n}$
- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathcal{H} \right\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- **Theorem**: for $\mathcal{G}$ mapping to $[0,1]$, prob is $\geq 1 - \delta$ over $S \sim \mathcal{D}^n$ that for **all** $g \in \mathcal{G}$
  - $$\underbrace{\mathbb{E}[g(z)]}_{L_{\mathcal{D}}(h)} - \underbrace{\frac{1}{n} \sum_{i=1}^{n} g(z_i)}_{L_S(h)} \leq 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}} \left(\sqrt{\text{VCdim}} + \sqrt{\log \frac{1}{\delta}} \right) \right)$$

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
    - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n)/n}$

- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\}$
    - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
    - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- **Theorem**: for $\mathcal{G}$ mapping to $[0,1]$, prob is $\geq 1 - \delta$ over $S \sim \mathcal{D}^n$ that for **all** $g \in \mathcal{G}$
    - $\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i) \leq 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$ $\leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\left(\sqrt{\text{VCdim}} + \sqrt{\log\frac{1}{\delta}}\right)\right)$

$L_{\mathcal{D}}(h)$ $L_S(h)$ bounds $\mathbb{E} \sup L_{\mathcal{D}}(h) - L_S(h)$

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n) / n}$

- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathcal{H} \right\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- **Theorem**: for $\mathcal{G}$ mapping to $[0,1]$, prob is $\geq 1 - \delta$ over $S \sim \mathcal{D}^n$ that for **all** $g \in \mathcal{G}$
  - $\mathbb{E}[g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \leq 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}} \left(\sqrt{\text{VCdim}} + \sqrt{\log \frac{1}{\delta}} \right) \right)$

$L_{\mathcal{D}}(h)$ $\quad$ $L_S(h)$ $\quad$ bounds $\mathbb{E} \sup L_{\mathcal{D}}(h) - L_S(h)$ $\quad$ how much bigger than its mean is likely

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n)/n}$
- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \left\{z \mapsto \ell(h, z) : h \in \mathcal{H}\right\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- **Theorem**: for $\mathcal{G}$ mapping to $[0,1]$, prob is $\geq 1 - \delta$ over $S \sim \mathcal{D}^n$ that for **all** $g \in \mathcal{G}$
  - $\left|\underbrace{\mathbb{E}[g(z)]}_{L_{\mathcal{D}}(h)} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} g(z_i)}_{L_S(h)}\right| \leq \underbrace{2\mathfrak{R}_n(\mathcal{G})}_{\text{bounds } \mathbb{E}\sup|L_{\mathcal{D}}(h) - L_S(h)|} + \underbrace{\sqrt{\frac{1}{2n}\log\frac{1}{\delta}}}_{\text{how much bigger than its mean is likely}} \leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\left(\sqrt{\text{VCdim}} + \sqrt{\log\frac{1}{\delta}}\right)\right)$

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \le \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n)/n}$
- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \left\{z \mapsto \ell(h, z) : h \in \mathcal{H}\right\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- **Theorem**: for $\mathcal{G}$ mapping to $[0,1]$, prob is $\ge 1 - \delta$ over $S \sim \mathcal{D}^n$ that for **all** $g \in \mathcal{G}$
  - $\left|\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right| \le 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$ $\le \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\left(\sqrt{\text{VCdim}} + \sqrt{\log\frac{1}{\delta}}\right)\right)$

    $L_{\mathcal{D}}(h)$ $L_S(h)$ bounds $\mathbb{E}\sup\left|L_{\mathcal{D}}(h) - L_S(h)\right|$ how much bigger than its mean is likely
  - $\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i) \le 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n)/n}$
- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathcal{H} \right\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- **Theorem**: for $\mathcal{G}$ mapping to $[0,1]$, prob is $\geq 1 - \delta$ over $S \sim \mathcal{D}^n$ that for **all** $g \in \mathcal{G}$
  - $\left| \mathbb{E}[g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right| \leq 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$ $\leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}} \left(\sqrt{\text{VCdim}} + \sqrt{\log \frac{1}{\delta}} \right) \right)$

    $L_{\mathcal{D}}(h)$ $\quad$ $L_S(h)$ $\quad$ bounds $\mathbb{E} \sup |L_{\mathcal{D}}(h) - L_S(h)|$ $\quad$ how much bigger than its mean is likely
  - $\left| \mathbb{E}[g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right| \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$

- Our goal was to upper bound $\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n)/n}$
- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \left\{z \mapsto \ell(h, z) : h \in \mathcal{H}\right\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- **Theorem**: for $\mathcal{G}$ mapping to $[0,1]$, prob is $\geq 1 - \delta$ over $S \sim \mathcal{D}^n$ that for **all** $g \in \mathcal{G}$
  - $\left|\underbrace{\mathbb{E}[g(z)]}_{L_{\mathcal{D}}(h)} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} g(z_i)}_{L_S(h)}\right| \leq \underbrace{2\mathfrak{R}_n(\mathcal{G})}_{\text{bounds } \mathbb{E}\sup|L_{\mathcal{D}}(h) - L_S(h)|} + \underbrace{\sqrt{\frac{1}{2n}\log\frac{1}{\delta}}}_{\text{how much bigger than its mean is likely}} \leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\left(\sqrt{\text{VCdim}} + \sqrt{\log\frac{1}{\delta}}\right)\right)$
  - $\left|\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right| \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$

because $\widehat{\mathfrak{R}}_S$ is close to $\mathfrak{R}_n$

- Our goal was to upper bound $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)|$ in terms of VC dimension
  - We've shown $\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2 \text{ VCdim}(\mathcal{H}) \log(n)/n}$
- One more step: consider $\mathfrak{R}_n(\mathcal{G})$ for $\mathcal{G} = \left\{ z \mapsto \ell(h, z) : h \in \mathcal{H} \right\}$
  - $g_h : \mathcal{Z} \to \mathbb{R}$ says, "what would the loss of predictor h be for a given (x, y)?"
  - For 0-1 loss, $g_h((x, y)) = \mathbb{I}(h(x) \neq y)$, and $\widehat{\mathfrak{R}}_S(\mathcal{G}) = \widehat{\mathfrak{R}}_{S|_x}(\mathcal{H})$ (proof is next)

or $\frac{1}{2}$ that, if $\mathcal{H}$ maps to ± 1

- **Theorem**: for $\mathcal{G}$ mapping to $[0,1]$, prob is $\geq 1 - \delta$ over $S \sim \mathcal{D}^n$ that for **all** $g \in \mathcal{G}$
  - $\underbrace{\left|\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right|}_{} \leq 2\mathfrak{R}_n(\mathcal{G}) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \leq \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\left(\sqrt{\text{VCdim}} + \sqrt{\log \frac{1}{\delta}}\right)\right)$

    $L_{\mathcal{D}}(h)$ $L_S(h)$ bounds $\mathbb{E} \sup \left|L_{\mathcal{D}}(h) - L_S(h)\right|$ how much bigger than its mean is likely
  - $\left|\mathbb{E}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right| \leq 2\widehat{\mathfrak{R}}_S(\mathcal{G}) + 3\sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$ (proof on Monday)

    because $\widehat{\mathfrak{R}}_S$ is close to $\mathfrak{R}_n$

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

$\mathfrak{R}$ 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 $\tilde{h}(x) = 2h(x) - 1$; $\tilde{y}_i = 2y_i - 1$

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

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

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

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

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

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

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

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

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

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

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

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

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

$\mathfrak{R}$ 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 $\tilde{y}_i$, $-\sigma_i\tilde{y}_i \sim \mathrm{Rad}$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\max_{a \in \mathscr{A}} \|a\| \leq r$, $\quad \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{a \in \mathscr{A}} \frac{1}{n} \boldsymbol{\sigma}^\top a\right] \leq \frac{1}{n} r \sqrt{2 \log |\mathscr{A}|}$

Summary

- VC dimension for linear classifiers:
  - d for homogenous, $d + 1$ for not (same as # of params)
  - So ERM works with 0-1 loss…except we can't do ERM in non-separable case!

- We *still* haven't quite proved the fundamental theorem
  - But we'll prove a much better bound when we do, on Monday
- **Empirical Rademacher complexity**: how much can $\mathscr{G}$ correlate with ± 1 noise on S?
- **Rademacher complexity**: its expectation over a random $S \sim \mathscr{D}^n$
- Upper bounded in terms of VC dimension
- Stated a generalization bound in terms of Rademacher complexity
  - Will imply (almost) the optimal sample complexity for agnostic case
- Will be easy(ish) to extend to things beyond 0-1 loss binary classification