Structural Risk Minimization

CPSC 532S: Modern Statistical Learning Theory

7 February 2022

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

Admin

- On Zoom today (obviously)
- Also on Wednesday – mostly better now, but playing it safe
- Hybrid mode starts next week, in DMP 101
- Office hours still online-only this week

- A2 is up, due next Friday night
  - Groups of up to three, allowed separate per question
  - Piazza “search for teammates” thing if you want
- A1 grading: hopefully done this week (sorry)

The course so far

- We've talked about learning binary classifiers in a fixed hypothesis class $\mathscr{H}$
    - (agnostic|realizable) PAC learning
    - uniform convergence property
    - VC dimension of $\mathscr{H}$
    - Rademacher complexity of $\mathscr{H}$

The course so far

- We've talked about learning binary classifiers in a fixed hypothesis class $\mathscr{H}$
  - (agnostic|realizable) PAC learning
  - uniform convergence property
  - VC dimension of $\mathscr{H}$
  - Rademacher complexity of $\mathscr{H}$
- Also a little bit about regression, based on Rademacher complexity

The course so far

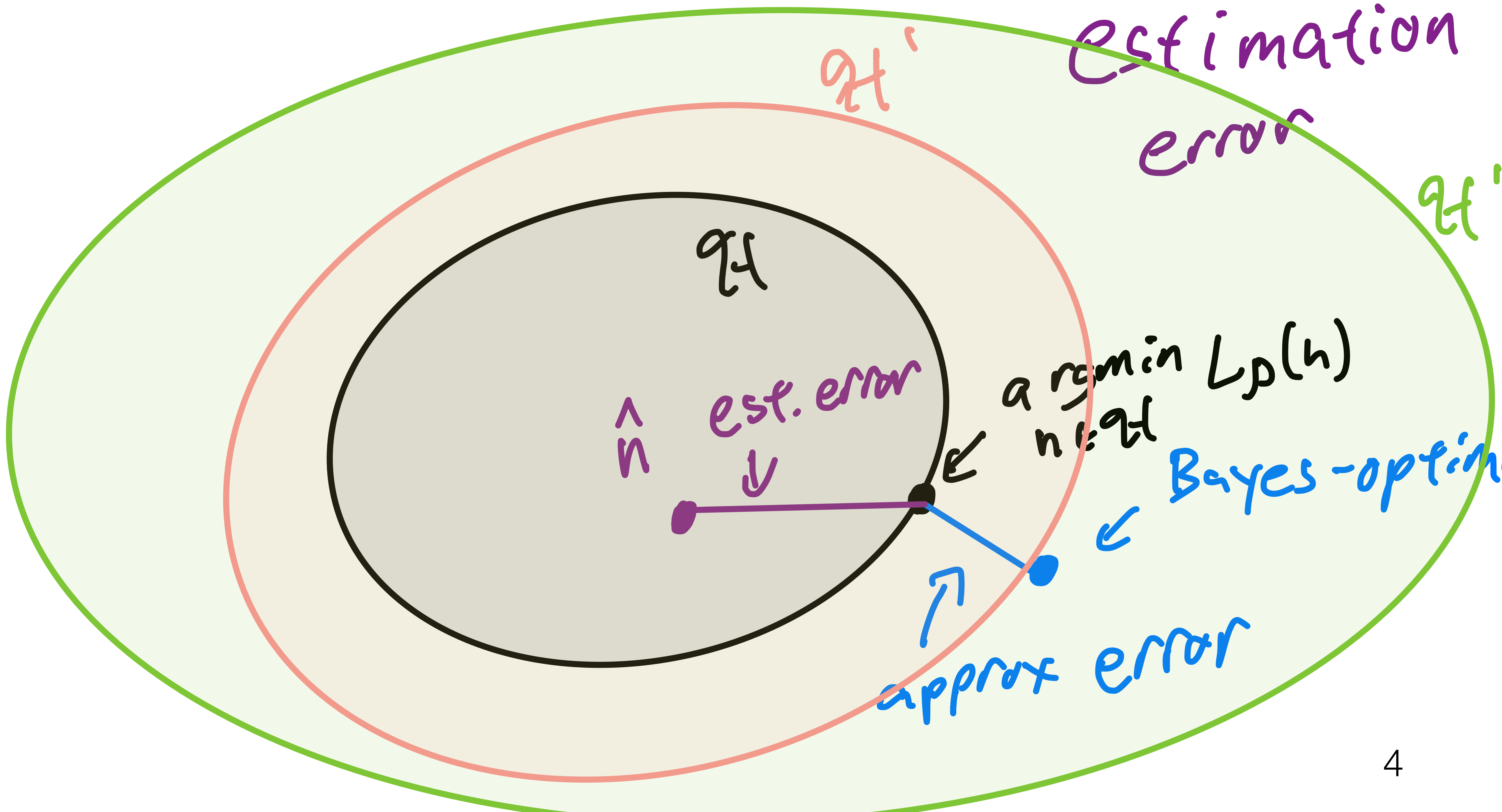
- We've talked about learning binary classifiers in a fixed hypothesis class $\mathscr{H}$
  - (agnostic|realizable) PAC learning
  - uniform convergence property
  - VC dimension of $\mathscr{H}$
  - Rademacher complexity of $\mathscr{H}$
- Also a little bit about regression, based on Rademacher complexity
- Proved bounds like $\Pr\left(\sup_{h \in \mathscr{H}} L_{\mathscr{D}}(h) - L_S(h) > \varepsilon\right) \leq \delta$
  - Imply ERM works: $L_{\mathscr{D}}(\hat{h}_S) \leq L_S(\hat{h}_S) + \varepsilon \leq L_S(h) + \varepsilon \leq L_{\mathscr{D}}(h) + 2\varepsilon$ for all h

realizable: $\exists h^* \in \mathcal{H}$ with $L_D(h^*) = 0$ well-specified: $\exists h^* \in \mathcal{H}$ with $L_D(h^*) = \inf_{\text{all predictors}} L_D(h) = L_D^*$

Importance of choosing $\mathscr{H}$

- Can't PAC-learn $\mathscr{H}$ if it has infinite VC dimension: no free lunch

- $$\underbrace{L_{\mathscr{D}}(h) - L^*_{\mathscr{D}}}_{\text{excess error}} = \underbrace{\left(L_{\mathscr{D}}(h) - \inf_{h' \in \mathscr{H}} L_{\mathscr{D}}(h') \right)}_{\text{estimation error}} + \underbrace{\left(\inf_{h' \in \mathscr{H}} L_{\mathscr{D}}(h') - L^*_{\mathscr{D}} \right)}_{\text{approximation error}}$$



Bayes-optimal $h_D(x) = \begin{cases} 1 & \text{if } \Pr(y=1 \mid x) \geq \frac{1}{2} \\ 0 & \text{o.w.} \end{cases}$

well-specified: $y \sim \mathcal{N}(w^T x, \sigma^2)$

$h^*(x) = w^T x$ $L_D(h^*) = \sigma^2$ (square loss)

Importance of choosing $\mathscr{H}$

- Can't PAC-learn $\mathscr{H}$ if it has infinite VC dimension: no free lunch

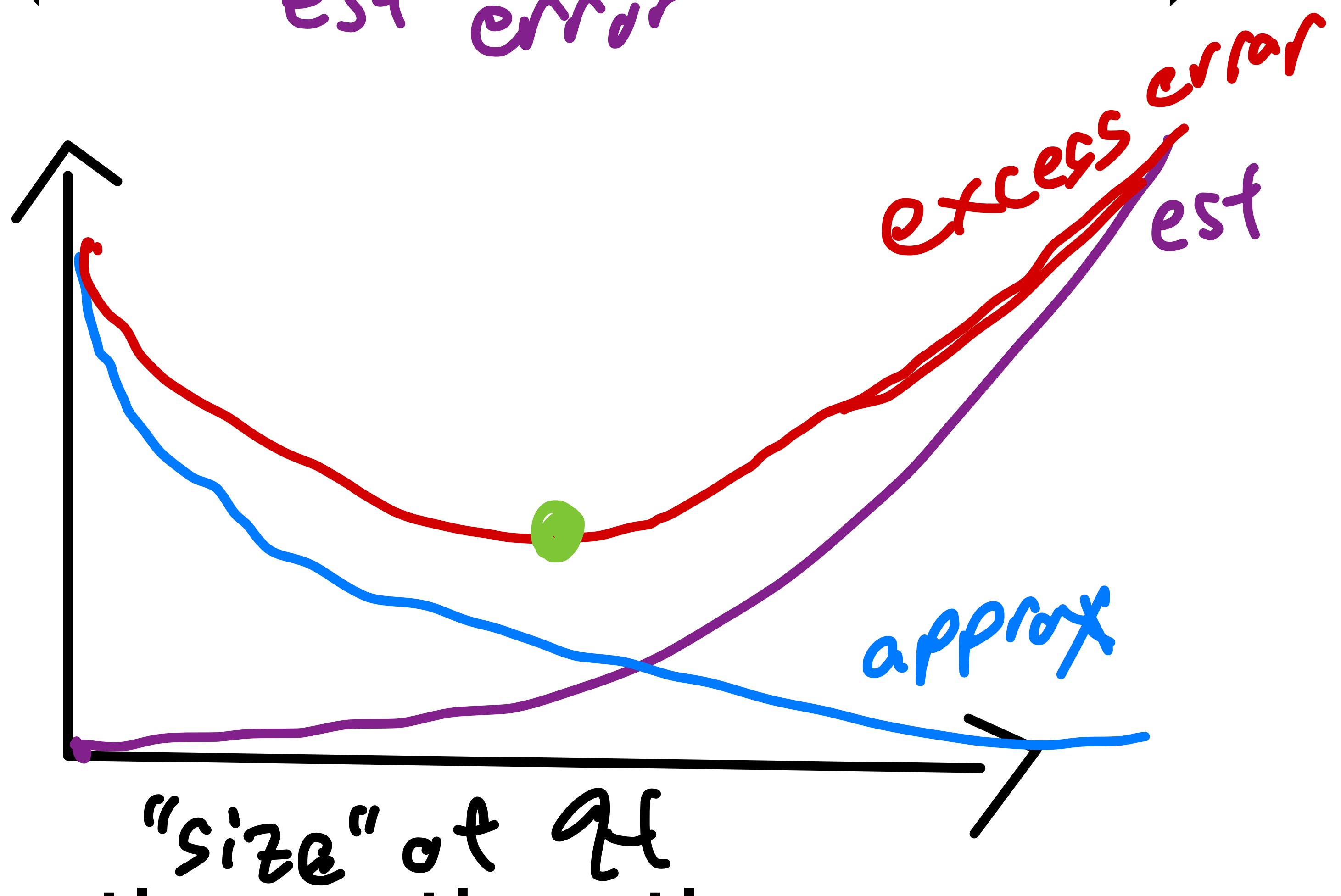
$\leq 2R_n(\mathscr{H}) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$

- $$L_{\mathscr{D}}(h) - L^*_{\mathscr{D}} = \left(L_{\mathscr{D}}(h) - \inf_{h' \in \mathscr{H}} L_{\mathscr{D}}(h') \right) + \left(\inf_{h' \in \mathscr{H}} L_{\mathscr{D}}(h') - L^*_{\mathscr{D}} \right)$$

est error

approx error

???



- Can bound/estimate the estimation error;
generally can't really estimate the approximation error

Importance of choosing $\mathscr{H}$

- Can't PAC-learn $\mathscr{H}$ if it has infinite VC dimension: no free lunch
- $$L_{\mathscr{D}}(h) - L^*_{\mathscr{D}} = \left(L_{\mathscr{D}}(h) - \inf_{h' \in \mathscr{H}} L_{\mathscr{D}}(h')\right) + \left(\inf_{h' \in \mathscr{H}} L_{\mathscr{D}}(h') - L^*_{\mathscr{D}}\right)$$
- Can bound/estimate the estimation error;
  generally can't really estimate the approximation error
- So…how to pick?

Structural Risk Minimization

- Idea: let $\mathscr{H}$ be really really big

Structural Risk Minimization

- Idea: let $\mathcal{H}$ be really really big
  - Approximation error $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L^*$ is small, maybe zero

Structural Risk Minimization

- Idea: let $\mathscr{H}$ be really really big
  - Approximation error $\inf_{h \in \mathscr{H}} L_{\mathscr{D}}(h) - L^*$ is small, maybe zero
  - So maybe $\mathrm{VCdim}(\mathscr{H}) = \infty$, $\mathfrak{R}_n(\mathscr{H})$ is big, etc: bad estimation error

Structural Risk Minimization

- Idea: let $\mathscr{H}$ be really really big
  - Approximation error $\inf_{h \in \mathscr{H}} L_{\mathscr{D}}(h) - L^*$ is small, maybe zero
  - So maybe $\mathrm{VCdim}(\mathscr{H}) = \infty$, $\mathfrak{R}_n(\mathscr{H})$ is big, etc: bad estimation error

- But decompose it into $\mathscr{H} = \mathscr{H}_1 \cup \mathscr{H}_2 \cup \cdots = \bigcup_{k \in \mathbb{N}} \mathscr{H}_k$

$\mathscr{H}$ decision trees — $\mathscr{H}_1$ w/ depth 1 — $\mathscr{H}_2$ w/ depth 2 — ... — $\mathscr{H}_k$ depth k

polynomial classifiers — linear $\mathbb{1}(w^T x + b \geq 0)$ — quadratics $\mathbb{1}(x^T A x + w^T x + b \geq 0)$ — degree-k polynomials

regularized SVM — SVMs $\|w\| \leq 10^{-4}$ — $\|w\| \leq 10^{-3}$ — $\|w\| \leq 10^{-5+k}$

Structural Risk Minimization

- Idea: let $\mathscr{H}$ be really really big
  - Approximation error $\inf_{h \in \mathscr{H}} L_{\mathscr{D}}(h) - L^*$ is small, maybe zero
  - So maybe $\mathrm{VCdim}(\mathscr{H}) = \infty$, $\mathfrak{R}_n(\mathscr{H})$ is big, etc: bad estimation error
- But decompose it into $\mathscr{H} = \mathscr{H}_1 \cup \mathscr{H}_2 \cup \cdots = \bigcup_{k \in \mathbb{N}} \mathscr{H}_k$
  - Assume **each** $\mathscr{H}_k$ has uniform convergence property: for all $\mathscr{D}$,
    $\sup_{h \in \mathscr{H}_k} \left| L_{\mathscr{D}}(h) - L_S(h) \right| \leq \varepsilon_k(n, \delta)$ with prob at least $1 - \delta$ over $S \sim \mathscr{D}^n$

Structural Risk Minimization

- Idea: let $\mathcal{H}$ be really really big
  - Approximation error $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L^*$ is small, maybe zero
  - So maybe $\mathrm{VCdim}(\mathcal{H}) = \infty$, $\mathfrak{R}_n(\mathcal{H})$ is big, etc: bad estimation error
- But decompose it into $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots = \bigcup_{k \in \mathbb{N}} \mathcal{H}_k$
  - Assume **each** $\mathcal{H}_k$ has uniform convergence property: for all $\mathcal{D}$, $\sup_{h \in \mathcal{H}_k} \left| L_{\mathcal{D}}(h) - L_S(h) \right| \leq \varepsilon_k(n, \delta)$ with prob at least $1 - \delta$ over $S \sim \mathcal{D}^n$
  - Choose **weights** $w_k \geq 0$ with $\sum_{k=1}^{\infty} w_k \leq 1$

$\sum_{k=1}^{\infty} \frac{6}{\pi^2 k^2} = 1$

$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$

Structural Risk

$S \sim D^n \Rightarrow \sup_{h \in \mathcal{H}_k} |L_D(h) - L_S(h)| \le \varepsilon_k(n, \delta)$ w/ prob $1-\delta$

- $\mathcal{H} = \bigcup_{k \in \mathbb{N}} \mathcal{H}_k$, each $\mathcal{H}_k$ has uniform convergence with $\varepsilon_k(n, \delta)$, weights $\sum_{k=1}^{\infty} w_k \le 1$

Structural Risk

- $\mathcal{H} = \bigcup_{k \in \mathbb{N}} \mathcal{H}_k$, each $\mathcal{H}_k$ has uniform convergence with $\varepsilon_k(n, \delta)$, weights $\sum_{k=1}^{\infty} w_k \leq 1$
- **Theorem**: For any $\mathscr{D}$, with probability at least $1 - \delta$ over choice of $S \sim \mathscr{D}^n$, we have

$$\min_{w} \underbrace{L_S(w)}_{\frac{1}{2}\|Xw - y\|^2} + \lambda \|w\|_1 \equiv \min_{w : \|w\|_1 \leq B} L_S(w)$$

Structural Risk

- $\mathcal{H} = \bigcup_{k \in \mathbb{N}} \mathcal{H}_k$, each $\mathcal{H}_k$ has uniform convergence with $\varepsilon_k(n, \delta)$, weights $\sum_{k=1}^{\infty} w_k \leq 1$
- **Theorem**: For any $\mathcal{D}$, with probability at least $1 - \delta$ over choice of $S \sim \mathcal{D}^n$, we have
  - For all k simultaneously, $\sup_{h \in \mathcal{H}_k} \left| L_{\mathcal{D}}(h) - L_S(h) \right| \leq \varepsilon_k(n, \delta w_k)$

Structural Risk

- $\mathscr{H} = \bigcup_{k \in \mathbb{N}} \mathscr{H}_k$, each $\mathscr{H}_k$ has uniform convergence with $\varepsilon_k(n, \delta)$, weights $\sum_{k=1}^{\infty} w_k \leq 1$
- **Theorem**: For any $\mathscr{D}$, with probability at least $1 - \delta$ over choice of $S \sim \mathscr{D}^n$, we have
  - For all k simultaneously, $\sup_{h \in \mathscr{H}_k} \left| L_{\mathscr{D}}(h) - L_S(h) \right| \leq \varepsilon_k(n, \delta w_k)$
  - Thus for all $h \in \mathscr{H}$ simultaneously,

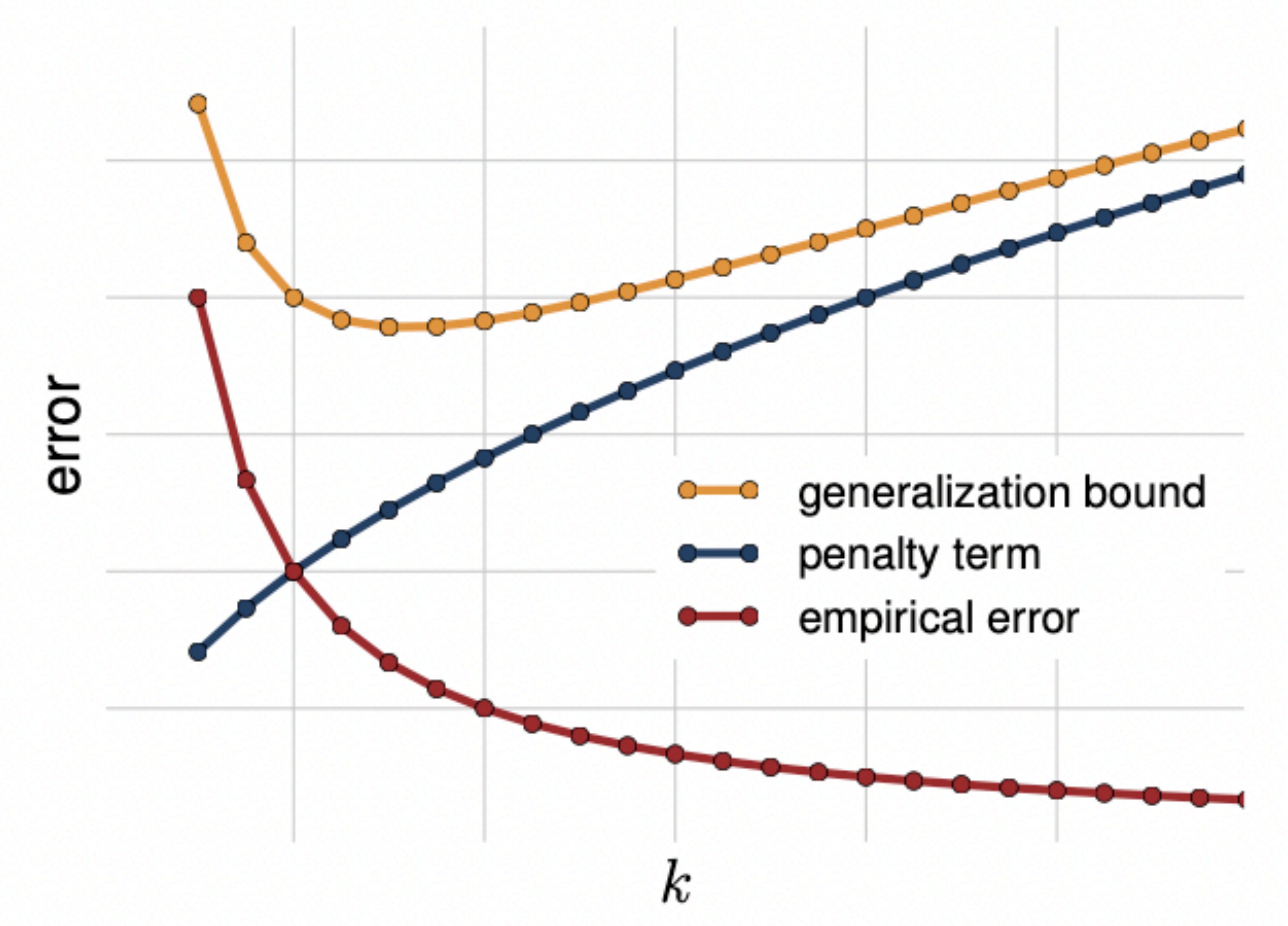
$$L_{\mathscr{D}}(h) \leq L_S(h) + \min_{k \,:\, h \in \mathscr{H}_k} \varepsilon_k(n, \delta w_k)$$

Structural Risk

- $\mathcal{H} = \bigcup_{k \in \mathbb{N}} \mathcal{H}_k$, each $\mathcal{H}_k$ has uniform convergence with $\varepsilon_k(n, \delta)$, weights $\sum_{k=1}^{\infty} w_k \le 1$
- **Theorem**: For any $\mathcal{D}$, with probability at least $1 - \delta$ over choice of $S \sim \mathcal{D}^n$, we have
  - For all k simultaneously, $\sup_{h \in \mathcal{H}_k} \left| L_{\mathcal{D}}(h) - L_S(h) \right| \le \varepsilon_k(n, \delta w_k)$
  - Thus for all $h \in \mathcal{H}$ simultaneously,
$$L_{\mathcal{D}}(h) \le L_S(h) + \min_{k \,:\, h \in \mathcal{H}_k} \varepsilon_k(n, \delta w_k)$$
- Proof: union bound over convergence in each $\mathcal{H}_k$, giving probability δw_k to each

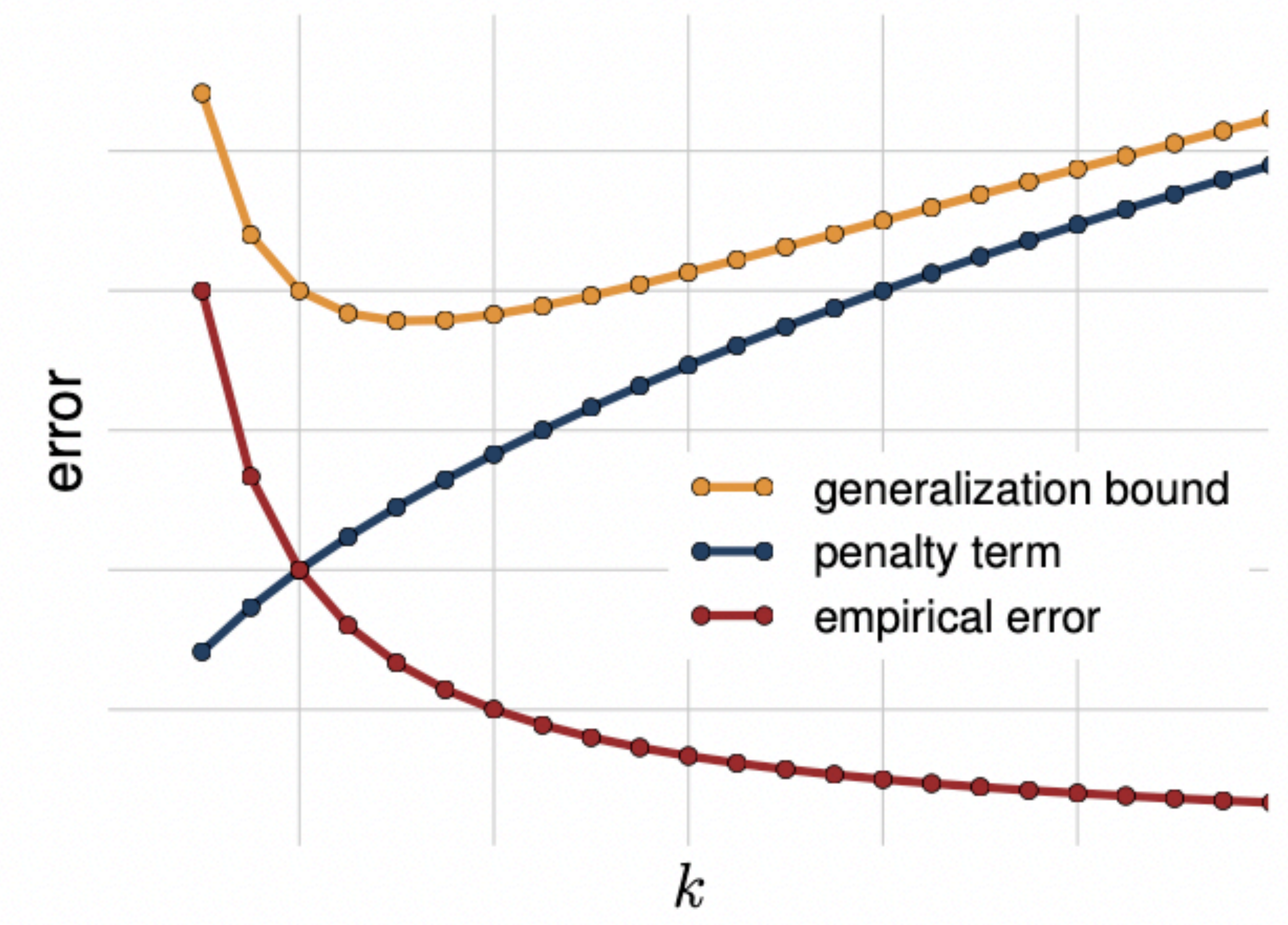
Bound Minimization

- What we really want is an h minimizing $L_{\mathscr{D}}(h)$, but we don't know $L_{\mathscr{D}}(h)$
- SRM algorithm minimizes an *upper bound* on $L_{\mathscr{D}}(h)$:



Bound Minimization

- What we really want is an h minimizing $L_{\mathscr{D}}(h)$, but we don't know $L_{\mathscr{D}}(h)$
- SRM algorithm minimizes an *upper bound* on $L_{\mathscr{D}}(h)$:
  - $h \in \text{argmin}_{h \in \mathscr{H}} \left[L_S(h) + \varepsilon_{k_h}\left(n, \delta w_{k_h}\right) \right]$ where $k_h = \min\{k : h \in \mathscr{H}_k\}$



Bound Minimization

- What we really want is an h minimizing $L_{\mathcal{D}}(h)$, but we don't know $L_{\mathcal{D}}(h)$
- SRM algorithm minimizes an *upper bound* on $L_{\mathcal{D}}(h)$:
  - $h \in \operatorname{argmin}_{h \in \mathcal{H}} \left[L_S(h) + \varepsilon_{k_h}\left(n, \delta w_{k_h}\right) \right]$ where $k_h = \min\{k : h \in \mathcal{H}_k\}$

Bound Minimization

- What we really want is an h minimizing $L_{\mathscr{D}}(h)$, but we don't know $L_{\mathscr{D}}(h)$
- SRM algorithm minimizes an *upper bound* on $L_{\mathscr{D}}(h)$:
- $h \in \mathrm{argmin}_{h \in \mathscr{H}} \left[L_S(h) + \varepsilon_{k_h}\left(n, \delta w_{k_h}\right) \right]$ where $k_h = \min\{k : h \in \mathscr{H}_k\}$
    - Can implement (with an "ERM oracle") as:

Bound Minimization

- What we really want is an h minimizing $L_{\mathscr{D}}(h)$, but we don't know $L_{\mathscr{D}}(h)$
- SRM algorithm minimizes an *upper bound* on $L_{\mathscr{D}}(h)$:
- $h \in \operatorname{argmin}_{h \in \mathscr{H}} \left[L_S(h) + \varepsilon_{k_h}\left(n, \delta w_{k_h}\right) \right]$ where $k_h = \min\{k : h \in \mathscr{H}_k\}$
  - Can implement (with an "ERM oracle") as:
    - best_loss = ∞
    - for $k = 1,2,\ldots$
      - cand = $\mathrm{ERM}(\mathscr{H}_k)$; cand_loss = $L_S(\text{cand}) + \varepsilon_k(n, w_k\delta)$

Bound Minimization

- What we really want is an h minimizing $L_{\mathscr{D}}(h)$, but we don't know $L_{\mathscr{D}}(h)$
- SRM algorithm minimizes an *upper bound* on $L_{\mathscr{D}}(h)$:
- $h \in \operatorname{argmin}_{h\in\mathscr{H}} \left[L_S(h) + \varepsilon_{k_h}\left(n, \delta w_{k_h}\right)\right]$ where $k_h = \min\{k : h \in \mathscr{H}_k\}$
  - Can implement (with an "ERM oracle") as:
    - best_loss = ∞
    - for $k = 1,2,\ldots$
      - cand = $\mathrm{ERM}(\mathscr{H}_k)$; cand_loss = $L_S(\text{cand}) + \varepsilon_k(n, w_k\delta)$
      - if (cand_loss < best_loss) { best = cand; best_loss = cand_loss; }

Bound Minimization

- What we really want is an h minimizing $L_{\mathscr{D}}(h)$, but we don't know $L_{\mathscr{D}}(h)$
- SRM algorithm minimizes an *upper bound* on $L_{\mathscr{D}}(h)$:
- $h \in \operatorname{argmin}_{h \in \mathscr{H}} \left[L_S(h) + \varepsilon_{k_h}\left(n, \delta w_{k_h}\right) \right]$ where $k_h = \min\{k : h \in \mathscr{H}_k\}$
  - Can implement (with an "ERM oracle") as:
    - best_loss = ∞
    - for $k = 1,2,\ldots$
      - cand = $\mathrm{ERM}(\mathscr{H}_k)$; cand_loss = $L_S(\text{cand}) + \varepsilon_k(n, w_k \delta)$
      - if (cand_loss < best_loss) { best = cand; best_loss = cand_loss; }
      - if ($\min_{k' > k} \varepsilon_k(n, \delta)$ > best_loss) { break; }

SRM ⊃ ERM

$\varepsilon_1(n,\delta) = \varepsilon_2(n,\delta)$

$\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$

$L_S(h) + \varepsilon_1(n, \frac{\delta}{2})$

$L_S(h) + \varepsilon_2(n, \frac{\delta}{2})$

- ERM is a special case of SRM with one k:
  - $\text{argmin}_{h\in\mathcal{H}} L_S(h) = \text{argmin}_{h\in\mathcal{H}} \left[L_S(h) + \varepsilon(n,\delta)\right]$

- If we split $\mathcal{H}$ into K parts of equal "size" (same ε function) and same weight, also the same as ERM

- What happens more generally?

- Pick $w_k = \frac{6}{\pi^2 k^2} \approx \frac{0.61}{k^2}$; have $\sum_k w_k = 1$

- Pick $w_k = \dfrac{6}{\pi^2 k^2} \approx \dfrac{0.61}{k^2}$; have $\sum_k w_k = 1$
    - By prev theorem, $L_{\mathscr{D}}(h) \leq L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for all $h \in \mathscr{H}$

- Pick $w_k = \frac{6}{\pi^2 k^2} \approx \frac{0.61}{k^2}$; have $\sum_k w_k = 1$
  - By prev theorem, $L_{\mathscr{D}}(h) \leq L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for all $h \in \mathscr{H}$
  - So $L_{\mathscr{D}}(\hat{h}) \leq L_S(\hat{h}) + \varepsilon_{k_{\hat{h}}}(n, w_{k_{\hat{h}}}\delta)$ for $\hat{h}$ the SRM solution

- Pick $w_k = \dfrac{6}{\pi^2 k^2} \approx \dfrac{0.61}{k^2}$; have $\sum_k w_k = 1$
  - By prev theorem, $L_{\mathcal{D}}(h) \leq L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for all $h \in \mathscr{H}$
  - So $L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + \varepsilon_{k_{\hat{h}}}(n, w_{k_{\hat{h}}}\delta)$ for $\hat{h}$ the SRM solution

    $\leq L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for *any* h $\in \mathcal{H}$, by def of SRM

- Pick $w_k = \dfrac{6}{\pi^2 k^2} \approx \dfrac{0.61}{k^2}$; have $\sum_k w_k = 1$
  - By prev theorem, $L_{\mathscr{D}}(h) \leq L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for all $h \in \mathscr{H}$
  - So $L_{\mathscr{D}}(\hat{h}) \leq L_S(\hat{h}) + \varepsilon_{k_{\hat{h}}}(n, w_{k_{\hat{h}}}\delta)$ for $\hat{h}$ the SRM solution

$\leq L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for *any* h, by def of SRM

$\leq L_{\mathscr{D}}(h) + 2\varepsilon_{k_h}(n, w_{k_h}\delta)$ using uniform convergence

In class, I said this was wrong and you needed $\varepsilon_{k_{\hat{h}}} + \varepsilon_{k_h}$.

That's <u>not true</u>: the slides as written are correct.

- Pick $w_k = \dfrac{6}{\pi^2 k^2} \approx \dfrac{0.61}{k^2}$; have $\sum_k w_k = 1$
  - By prev theorem, $L_{\mathscr{D}}(h) \le L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for all $h \in \mathscr{H}$
  - So $L_{\mathscr{D}}(\hat{h}) \le L_S(\hat{h}) + \varepsilon_{k_{\hat{h}}}(n, w_{k_{\hat{h}}}\delta)$ for $\hat{h}$ the SRM solution

$\le L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for *any* h, by def of SRM

$\le L_{\mathscr{D}}(h) + 2\varepsilon_{k_h}(n, w_{k_h}\delta)$ using uniform convergence

$\le L_{\mathscr{D}}(h) + \varepsilon$ if $n \ge n^{UC}_{\mathscr{H}_{k_h}}\left(\dfrac{\varepsilon}{2}, \dfrac{6\delta}{\pi^2 k_h^2}\right)$

- Pick $w_k = \dfrac{6}{\pi^2 k^2} \approx \dfrac{0.61}{k^2}$; have $\sum_k w_k = 1$
  - By prev theorem, $L_{\mathscr{D}}(h) \leq L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for all $h \in \mathscr{H}$
  - So $L_{\mathscr{D}}(\hat{h}) \leq L_S(\hat{h}) + \varepsilon_{k_{\hat{h}}}(n, w_{k_{\hat{h}}}\delta)$ for $\hat{h}$ the SRM solution

$\leq L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for *any* h, by def of SRM

$\leq L_{\mathscr{D}}(h) + 2\varepsilon_{k_h}(n, w_{k_h}\delta)$ using uniform convergence

$\leq L_{\mathscr{D}}(h) + \varepsilon$ if $n \geq n^{UC}_{\mathscr{H}_{k_h}}\left(\dfrac{\varepsilon}{2}, \dfrac{6\delta}{\pi^2 k_h^2}\right)$

- We say that $\hat{h}$ **(ε, δ)-competes with** h for $L_{\mathscr{D}}(\hat{h}) \geq L_{\mathscr{D}}(h) + \varepsilon$ w/ prob $1 - \delta$

- Pick $w_k = \dfrac{6}{\pi^2 k^2} \approx \dfrac{0.61}{k^2}$; have $\sum_k w_k = 1$
  - By prev theorem, $L_{\mathscr{D}}(h) \le L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for all $h \in \mathscr{H}$
  - So $L_{\mathscr{D}}(\hat{h}) \le L_S(\hat{h}) + \varepsilon_{k_{\hat{h}}}(n, w_{k_{\hat{h}}}\delta)$ for $\hat{h}$ the SRM solution

$$
\begin{aligned}
&\le L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta) && \text{for } \textit{any } h\text{, by def of SRM} \\
&\le L_{\mathscr{D}}(h) + 2\varepsilon_{k_h}(n, w_{k_h}\delta) && \text{using uniform convergence} \\
&\le L_{\mathscr{D}}(h) + \varepsilon && \text{if } n \ge n^{UC}_{\mathscr{H}_{k_h}}\left(\frac{\varepsilon}{2}, \frac{6\delta}{\pi^2 k_h^2}\right)
\end{aligned}
$$

- We say that $\hat{h}$ **(ε, δ)-competes with** h for $L_{\mathscr{D}}(\hat{h}) \ge L_{\mathscr{D}}(h) + \varepsilon$ w/ prob $1 - \delta$
- An algorithm $A(S)$ **nonuniformly learns** $\mathscr{H}$ if for all $\varepsilon, \delta \in (0,1)$ and $h \in \mathscr{H}$, for any $\mathscr{D}$, if $n \ge n^{NUL}_{\mathscr{H}}(\varepsilon, \delta, h)$, then $A(S)$ (ε, δ)-competes with h

- Pick $w_k = \dfrac{6}{\pi^2 k^2} \approx \dfrac{0.61}{k^2}$; have $\sum_k w_k = 1$
  - By prev theorem, $L_{\mathscr{D}}(h) \le L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for all $h \in \mathscr{H}$
  - So $L_{\mathscr{D}}(\hat{h}) \le L_S(\hat{h}) + \varepsilon_{k_{\hat{h}}}(n, w_{k_{\hat{h}}}\delta)$ for $\hat{h}$ the SRM solution

$\le L_S(h) + \varepsilon_{k_h}(n, w_{k_h}\delta)$ for *any* h, by def of SRM

$\le L_{\mathscr{D}}(h) + 2\varepsilon_{k_h}(n, w_{k_h}\delta)$ using uniform convergence

$\le L_{\mathscr{D}}(h) + \varepsilon$ if $n \ge n^{UC}_{\mathscr{H}_{k_h}}\left(\dfrac{\varepsilon}{2}, \dfrac{6\delta}{\pi^2 k_h^2}\right)$

- We say that $\hat{h}$ **(ε, δ)-competes with** h for $L_{\mathscr{D}}(\hat{h}) \ge L_{\mathscr{D}}(h) + \varepsilon$ w/ prob $1 - \delta$
- An algorithm $A(S)$ **nonuniformly learns** $\mathscr{H}$ if for all $\varepsilon, \delta \in (0,1)$ and $h \in \mathscr{H}$, for any $\mathscr{D}$, if $n \ge n^{NUL}_{\mathscr{H}}(\varepsilon, \delta, h)$, then $A(S)$ (ε, δ)-competes with h
- So: SRM with these weights nonuniformly learns any $\mathscr{H}$ that decomposes into a countable sum of things with finite VC dimension

Nonuniform learnability

- SRM nonuniformly learns any countable union of agnostic PAC-learnable $\mathscr{H}_k$

Nonuniform learnability

- SRM nonuniformly learns any countable union of agnostic PAC-learnable $\mathscr{H}_k$
- If $\mathscr{H}$ is nonuniformly learnable, it's a countable union of agnostic PAC-learnable $\mathscr{H}_k$:

Nonuniform learnability

- SRM nonuniformly learns any countable union of agnostic PAC-learnable $\mathscr{H}_k$
- If $\mathscr{H}$ is nonuniformly learnable, it's a countable union of agnostic PAC-learnable $\mathscr{H}_k$:
  - Let $\mathscr{H}_k = \left\{ h \in \mathscr{H} : n_{\mathscr{H}}^{NUL}\left(\frac{1}{8}, \frac{1}{7}, h\right) \leq k \right\}$

Nonuniform learnability

- SRM nonuniformly learns any countable union of agnostic PAC-learnable $\mathscr{H}_k$

- If $\mathscr{H}$ is nonuniformly learnable, it's a countable union of agnostic PAC-learnable $\mathscr{H}_k$:
  - Let $\mathscr{H}_k = \left\{ h \in \mathscr{H} : n_{\mathscr{H}}^{NUL}\left(\frac{1}{8}, \frac{1}{7}, h\right) \leq k \right\}$
    - Have $\mathscr{H} = \cup_{k \geq 1} \mathscr{H}_k$

Nonuniform learnability

- SRM nonuniformly learns any countable union of agnostic PAC-learnable $\mathscr{H}_k$
- If $\mathscr{H}$ is nonuniformly learnable, it's a countable union of agnostic PAC-learnable $\mathscr{H}_k$:
  - Let $\mathscr{H}_k = \left\{ h \in \mathscr{H} : n_{\mathscr{H}}^{NUL}\left(\frac{1}{8}, \frac{1}{7}, h\right) \leq k \right\}$
    - Have $\mathscr{H} = \cup_{k \geq 1} \mathscr{H}_k$
    - For any realizable $\mathscr{D}$ wrt $\mathscr{H}_k$, implies $\Pr_S\left(L_{\mathscr{D}}(\hat{h}_S) \leq \frac{1}{8}\right) \geq \frac{6}{7}$ (since $L_{\mathscr{D}}(h) = 0$)

Nonuniform learnability

- SRM nonuniformly learns any countable union of agnostic PAC-learnable $\mathscr{H}_k$
- If $\mathscr{H}$ is nonuniformly learnable, it's a countable union of agnostic PAC-learnable $\mathscr{H}_k$:
  - Let $\mathscr{H}_k = \left\{ h \in \mathscr{H} : n_{\mathscr{H}}^{NUL}\left(\frac{1}{8}, \frac{1}{7}, h\right) \leq k \right\}$
  - Have $\mathscr{H} = \cup_{k \geq 1} \mathscr{H}_k$
  - For any realizable $\mathscr{D}$ wrt $\mathscr{H}_k$, implies $\Pr_S \left(L_{\mathscr{D}}(\hat{h}_S) \leq \frac{1}{8} \right) \geq \frac{6}{7}$ (since $L_{\mathscr{D}}(h) = 0$)
  - But no free lunch theorem implies: if $\mathrm{VCdim}(\mathscr{H}_k) = \infty$, this would be impossible

Nonuniform learnability

- SRM nonuniformly learns any countable union of agnostic PAC-learnable $\mathscr{H}_k$
- If $\mathscr{H}$ is nonuniformly learnable, it's a countable union of agnostic PAC-learnable $\mathscr{H}_k$:
  - Let $\mathscr{H}_k = \left\{ h \in \mathscr{H} : n_{\mathscr{H}}^{NUL}\left(\frac{1}{8}, \frac{1}{7}, h\right) \leq k \right\}$
  - Have $\mathscr{H} = \cup_{k \geq 1} \mathscr{H}_k$
  - For any realizable $\mathscr{D}$ wrt $\mathscr{H}_k$, implies $\Pr_S \left(L_{\mathscr{D}}(\hat{h}_S) \leq \frac{1}{8} \right) \geq \frac{6}{7}$ (since $L_{\mathscr{D}}(h) = 0$)
  - But no free lunch theorem implies: if $\mathrm{VCdim}(\mathscr{H}_k) = \infty$, this would be impossible
  - So $\mathscr{H}_k$ has finite VC dim, so is agnostic PAC-learnable

Nonuniform learnability

- SRM nonuniformly learns any countable union of agnostic PAC-learnable $\mathscr{H}_k$
- If $\mathscr{H}$ is nonuniformly learnable, it's a countable union of agnostic PAC-learnable $\mathscr{H}_k$:
  - Let $\mathscr{H}_k = \left\{ h \in \mathscr{H} : n_{\mathscr{H}}^{NUL}\left(\frac{1}{8}, \frac{1}{7}, h\right) \leq k \right\}$
  - Have $\mathscr{H} = \cup_{k \geq 1} \mathscr{H}_k$
  - For any realizable $\mathscr{D}$ wrt $\mathscr{H}_k$, implies $\Pr_S\left(L_{\mathscr{D}}(\hat{h}_S) \leq \frac{1}{8}\right) \geq \frac{6}{7}$ (since $L_{\mathscr{D}}(h) = 0$)
  - But no free lunch theorem implies: if $\mathrm{VCdim}(\mathscr{H}_k) = \infty$, this would be impossible
  - So $\mathscr{H}_k$ has finite VC dim, so is agnostic PAC-learnable
- Set of all measurable $\mathscr{H}$ is **not** a countable union of finite-VC classes

SRM with Rademacher

- Recall that for 0-1 loss, $\mathscr{H}$ to ± 1, $\sup_{h \in \mathscr{H}_k} L_{\mathscr{D}}(h) - L_S(h) \leq \mathfrak{R}_n(\mathscr{H}_k) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$

SRM with Rademacher

- Recall that for 0-1 loss, $\mathscr{H}$ to ± 1, $\sup_{h \in \mathscr{H}_k} L_{\mathscr{D}}(h) - L_S(h) \leq \mathfrak{R}_n(\mathscr{H}_k) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$
- Implies (as before, but dropping abs value) that, simultaneously for all $h \in \mathscr{H}$,

$$L_{\mathscr{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{2n} \log \frac{1}{w_{k_h} \delta}}$$

SRM with Rademacher

- Recall that for 0-1 loss, $\mathscr{H}$ to ± 1, $\sup_{h \in \mathscr{H}_k} L_{\mathscr{D}}(h) - L_S(h) \leq \mathfrak{R}_n(\mathscr{H}_k) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$
- Implies (as before, but dropping abs value) that, simultaneously for all $h \in \mathscr{H}$,

$$L_{\mathscr{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{2n} \log \frac{1}{w_{k_h} \delta}}$$

  - Pick (as before) $w_k = 6/(\pi^2 k^2)$

SRM with Rademacher

- Recall that for 0-1 loss, $\mathscr{H}$ to ± 1, $\sup_{h \in \mathscr{H}_k} L_{\mathscr{D}}(h) - L_S(h) \leq \mathfrak{R}_n(\mathscr{H}_k) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$
- Implies (as before, but dropping abs value) that, simultaneously for all $h \in \mathscr{H}$,

$$L_{\mathscr{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{2n} \log \frac{1}{w_{k_h}\delta}}$$

- Pick (as before) $w_k = 6/(\pi^2 k^2)$
- $\log \frac{1}{w_k \delta} = \log \frac{1}{2w_k} + \log \frac{2}{\delta} = \log \frac{\pi^2 k^2}{12} + \log \frac{2}{\delta} \leq 2 \log k + \log \frac{2}{\delta}$

SRM with Rademacher

- Recall that for 0-1 loss, $\mathcal{H}$ to ± 1, $\sup_{h \in \mathcal{H}_k} L_{\mathcal{D}}(h) - L_S(h) \leq \mathfrak{R}_n(\mathcal{H}_k) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$
- Implies (as before, but dropping abs value) that, simultaneously for all $h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathcal{H}_{k_h}) + \sqrt{\frac{1}{2n} \log \frac{1}{w_{k_h}\delta}}$$

- Pick (as before) $w_k = 6/(\pi^2 k^2)$
- $\log \frac{1}{w_k \delta} = \log \frac{1}{2w_k} + \log \frac{2}{\delta} = \log \frac{\pi^2 k^2}{12} + \log \frac{2}{\delta} \leq 2 \log k + \log \frac{2}{\delta}$
- $\sqrt{\frac{1}{2n} \log \frac{1}{w_{k_h}\delta}} \leq \sqrt{\frac{1}{n} \log k_h + \frac{1}{2n} \log \frac{2}{\delta}} \leq \sqrt{\frac{1}{n} \log k_h} + \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$

SRM with Rademacher

- Recall that for 0-1 loss, $\mathscr{H}$ to ± 1, $\sup_{h \in \mathscr{H}_k} L_{\mathscr{D}}(h) - L_S(h) \leq \mathfrak{R}_n(\mathscr{H}_k) + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}}$
- Implies (as before, but dropping abs value) that, simultaneously for all $h \in \mathscr{H}$,

$$L_{\mathscr{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{2n} \log \frac{1}{w_{k_h}\delta}}$$

- Pick (as before) $w_k = 6/(\pi^2 k^2)$
- $\log \frac{1}{w_k \delta} = \log \frac{1}{2w_k} + \log \frac{2}{\delta} = \log \frac{\pi^2 k^2}{12} + \log \frac{2}{\delta} \leq 2 \log k + \log \frac{2}{\delta}$
- $\sqrt{\frac{1}{2n} \log \frac{1}{w_{k_h}\delta}} \leq \sqrt{\frac{1}{n} \log k_h + \frac{1}{2n} \log \frac{2}{\delta}} \leq \sqrt{\frac{1}{n} \log k_h} + \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$
- So $\quad L_{\mathscr{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{n} \log k_h} + \sqrt{\frac{1}{2n} \log \frac{2}{\delta}}$

SRM with Rademacher

- Have $L_{\mathscr{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$

SRM with Rademacher

- Have $L_{\mathscr{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$
- SRM algorithm minimizes $L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h}$

SRM with Rademacher

- Have $L_{\mathscr{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$
- SRM algorithm minimizes $L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h}$
- Plugging in uniform convergence twice, can get

$$L_{\mathscr{D}}(\hat{h}) \leq \inf_{h\in\mathscr{H}}\left[L_{\mathscr{D}}(h) + 2\mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h}\right] + \sqrt{\frac{2}{n}\log\frac{3}{\delta}}$$

SRM with Rademacher

- Have $L_{\mathcal{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathcal{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$
- SRM algorithm minimizes $L_S(h) + \mathfrak{R}_n(\mathcal{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h}$
- Plugging in uniform convergence twice, can get

$$L_{\mathcal{D}}(\hat{h}) \leq \inf_{h\in\mathcal{H}} \left[L_{\mathcal{D}}(h) + 2\mathfrak{R}_n(\mathcal{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h}\right] + \sqrt{\frac{2}{n}\log\frac{3}{\delta}}$$

- If there's an optimal h^*, then the $\sqrt{\frac{1}{n}\log k_{h^*}}$ term is the only thing worse than just learning directly in $\mathcal{H}_{k^*}$ in the first place

SRM with Rademacher

- Have $$L_{\mathscr{D}}(h) \leq L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h} + \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$$
- SRM algorithm minimizes $$L_S(h) + \mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h}$$
- Plugging in uniform convergence twice, can get
$$L_{\mathscr{D}}(\hat{h}) \leq \inf_{h\in\mathscr{H}} \left[L_{\mathscr{D}}(h) + 2\mathfrak{R}_n(\mathscr{H}_{k_h}) + \sqrt{\frac{1}{n}\log k_h} \right] + \sqrt{\frac{2}{n}\log\frac{3}{\delta}}$$
- If there's an optimal h^*, then the $\sqrt{\frac{1}{n}\log k_{h^*}}$ term is the only thing worse than just learning directly in $\mathscr{H}_{k^*}$ in the first place
    - Not usually a big deal, especially if we order the $\mathscr{H}_k$ reasonably!

SRM with singleton classes

- If $\mathscr{H}$ is countable, we can number the elements and take $\mathscr{H} = \cup_{n\in\mathbb{N}} \{h_n\}$
  - "Uniform" convergence on $\{h_n\}$ via Hoeffding: $\varepsilon_k(n,\delta) = \sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$
  - SRM is $\operatorname{argmin}_{h\in\mathscr{H}} \left[L_S(h) + \sqrt{\frac{1}{2n}\left[-\log w_h + \log\frac{2}{\delta}\right]}\right]$
  - Entirely determined by our choice of "prior" w_h
  - How to choose a prior?

Minimum Description Length

- Come up with a *prefix-free* binary language $\mathcal{S} \subseteq \{0,1\}^*$ describing each h

h = gzip(C++ code for a function implementing h)

0 1 0 0 0

0 1 0 0 0 1 1

Minimum Description Length

- Come up with a *prefix-free* binary language $\mathcal{S} \subseteq \{0,1\}^*$ describing each h
  - Kraft Inequality: $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} 2^{-|\boldsymbol{\sigma}|} \leq 1$

0 0 1 1 0 0 1

0 0 1 0

1 1 0 0 1 1 1 1 1

Minimum Description Length

- Come up with a *prefix-free* binary language $\mathcal{S} \subseteq \{0,1\}^*$ describing each h
  - Kraft Inequality: $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} 2^{-|\boldsymbol{\sigma}|} \leq 1$
- Let $|h|$ be the "description length" of h using $\mathcal{S}$

Minimum Description Length

- Come up with a *prefix-free* binary language $\mathcal{S} \subseteq \{0,1\}^*$ describing each h
  - Kraft Inequality: $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} 2^{-|\boldsymbol{\sigma}|} \leq 1$
- Let $|h|$ be the "description length" of h using $\mathcal{S}$
- Then MDL is SRM with weights $w_h = 1/2^{|h|}$:

  we know that $L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{1}{2n}\left[|h| \; \log 2 + \log \frac{2}{\delta}\right]}$ uniformly,

  and MDL principle minimizes the RHS

Minimum Description Length

- Come up with a *prefix-free* binary language $\mathcal{S} \subseteq \{0,1\}^*$ describing each h
  - Kraft Inequality: $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} 2^{-|\boldsymbol{\sigma}|} \leq 1$
- Let $|h|$ be the "description length" of h using $\mathcal{S}$
- Then MDL is SRM with weights $w_h = 1/2^{|h|}$:

  we know that $L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{1}{2n}\left[|h| \, \log 2 + \log \frac{2}{\delta}\right]}$ uniformly,

  and MDL principle minimizes the RHS

- One formalization of Occam's Razor

Minimum Description Length

- Come up with a *prefix-free* binary language $\mathcal{S} \subseteq \{0,1\}^*$ describing each h
  - Kraft Inequality: $\sum_{\boldsymbol{\sigma} \in \mathcal{S}} 2^{-|\boldsymbol{\sigma}|} \leq 1$
- Let $|h|$ be the "description length" of h using $\mathcal{S}$
- Then MDL is SRM with weights $w_h = 1/2^{|h|}$:

  we know that $L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{1}{2n}\left[|h|\, \log 2 + \log \frac{2}{\delta}\right]}$ uniformly,

  and MDL principle minimizes the RHS

- One formalization of Occam's Razor
- But "simplest" is *not* inherent; we're pre-committing to what we call "simple" based on our choice of $\mathcal{S}$

Problems with bound minimization

- Concentration inequalities are usually pretty conservative
  - Hold for *all* distributions that are, e.g., bounded
  - Symmetrization in Rademacher introduces a factor of 2 that's often not needed

- SRM is based on these worst-case assumptions
  - So, can't adapt to e.g. the fast $1/n$ rate if turns out to be realizable: will just operate assuming the slow $1/\sqrt{n}$ agnostic rate

- Performance of the algorithm fundamentally based on how good at analysis you are
  - We'd usually prefer the algorithm work whether we're smart or not

Summary

- SRM allows learning over infinite-VC $\mathscr{H}$
  - We just learn slower if h is harder
- Need to choose a countable decomposition into $\mathscr{H}_k$
- Often, little penalty vs if we knew which $\mathscr{H}_k$ the optimal solution is in beforehand
- Generic way to pick weights: $w_k = 6/(\pi^2 k^2)$; gives a $\log k$ term in Rademacher
- Minimum Description Length is another, semi-universal way to divide

- **Next time:** choosing h using a validation set