Generalization Bounds

Yevgeny Seldin

Generalization Bounds

- Generalization bound for a single hypothesis
- Learning by selection
- Generalization lower bound
- Generalization bound for a finite $\mathcal{H}$ (finite selection)
- Approximation-Estimation (bias-variance) trade-off
- Occam's razor: generalization bound for a countable $\mathcal{H}$
- Application example (Occam): binary decision trees

Generalization bound: single hypothesis

$$\boxed{S = \{(X_1, Y_1), \dots, (X_n, Y_n)\}}$$

$$h \longrightarrow \hat{L}(h, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i)$$

Hoeffding:

$$\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right] - \frac{1}{n}\sum_{i=1}^{n} Z_i \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leq \delta$$

- Hoeffding:

$$\mathbb{P}\left(L(h) - \hat{L}(h, S) \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \leq \delta$$

$$\Rightarrow \mathbb{P}\left(L(h) - \hat{L}(h, S) \leq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \geq 1 - \delta$$

- In words: with probability at least $1 - \delta$:

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}$$

(the probability is over observing $\hat{L}(h, S)$, not over $L(h)$)

Learning by Selection

$$S = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

$\mathcal{H}$

$$h_1 \rightarrow \hat{L}(h_1, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h_1(X_i), Y_i)$$

$$h_2 \rightarrow \hat{L}(h_2, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h_2(X_i), Y_i)$$

$$h_3 \rightarrow \hat{L}(h_3, S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h_3(X_i), Y_i)$$

$$\vdots$$

$$\hat{h}_S^* = \arg\min_{\mathrm{h}\in\mathcal{H}} \hat{L}(h, S) \longrightarrow L\left(\hat{h}_S^*\right) = \mathbb{E}\left[\ell\left(\hat{h}_S^*(X), Y\right)\right]$$

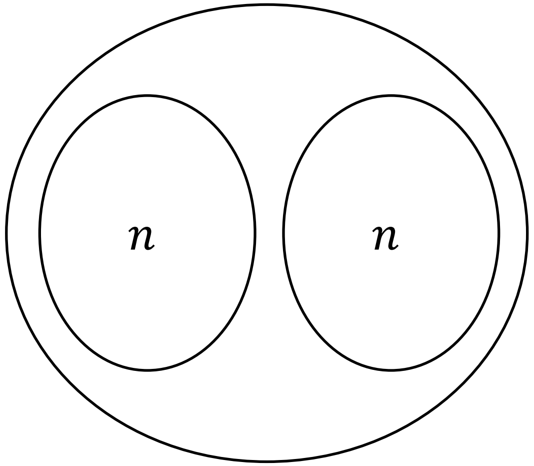
$$(X, Y)$$

$$\boxed{\mathbb{E}_S\left[\hat{L}\left(\hat{h}_S^*, S\right)\right] \neq \mathbb{E}_S\left[L\left(\hat{h}_S^*\right)\right]}$$

Lower bound for learning by selection from finite $\mathcal{H}$

- Lower bound

- $|\mathcal{X}| = 2n$
- $|\mathcal{H}| = 2^{2n}$
- $p(x)$ – uniform
- y – random w.p. $\frac{1}{2}$



- $\mathbb{E}\left[\hat{L}\left(\hat{h}_S^*, S\right)\right] = 0$
- $\mathbb{E}\left[L\left(\hat{h}_S^*\right)\right] \geq \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} \geq \frac{1}{4}$

- Corollary:

$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) - \hat{L}(h, S) \geq \frac{1}{8}\right) \geq \frac{1}{8}$$

- Proof by contradiction:

Assume that

$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) - \hat{L}(h, S) \geq \frac{1}{8}\right) < \frac{1}{8}$$

Then

$$\mathbb{E}\left[L\left(\hat{h}_S^*\right)\right] \leq \frac{1}{8} \cdot 1 + \left(1 - \frac{1}{8}\right)\left(\underbrace{\hat{L}\left(\hat{h}_S^*, S\right)}_{=0} + \frac{1}{8}\right) < \frac{1}{4}$$

Generalization bound for finite $\mathcal{H}$

- Theorem: Let $\mathcal{H}$ be finite with $|\mathcal{H}| = M$. Then
$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}\right) \leq \delta$$

- Corollary: $\mathbb{P}\left(L(\hat{h}_S^*) \geq \hat{L}(\hat{h}_S^*, S) + \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}\right) \leq \delta$
- Equivalently: $\mathbb{P}\left(L(\hat{h}_S^*) \leq \hat{L}(\hat{h}_S^*, S) + \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}\right) \geq 1 - \delta$
- In words: with probability at least $1-\delta$:
$$L(\hat{h}_S^*) \leq \hat{L}(\hat{h}_S^*, S) + \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}$$

- For single h we had:
$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}$$
- Price of selection: $\ln M$

- Lower bound: for $|\mathcal{H}| = M = 2^{2n}$
$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) - \hat{L}(h,S) \geq \frac{1}{8}\right) \geq \frac{1}{8}$$
  - No contradiction: $\sqrt{\frac{\ln\frac{M}{\delta}}{2n}} \approx 1$

- For $M \ll e^n$ we get a meaningful bound

Proof

$\mathbb{E}_S[\hat{L}(\hat{h}_S^*, S)] \neq \mathbb{E}_S[L(\hat{h}_S^*)]$
We cannot apply Hoeffding!

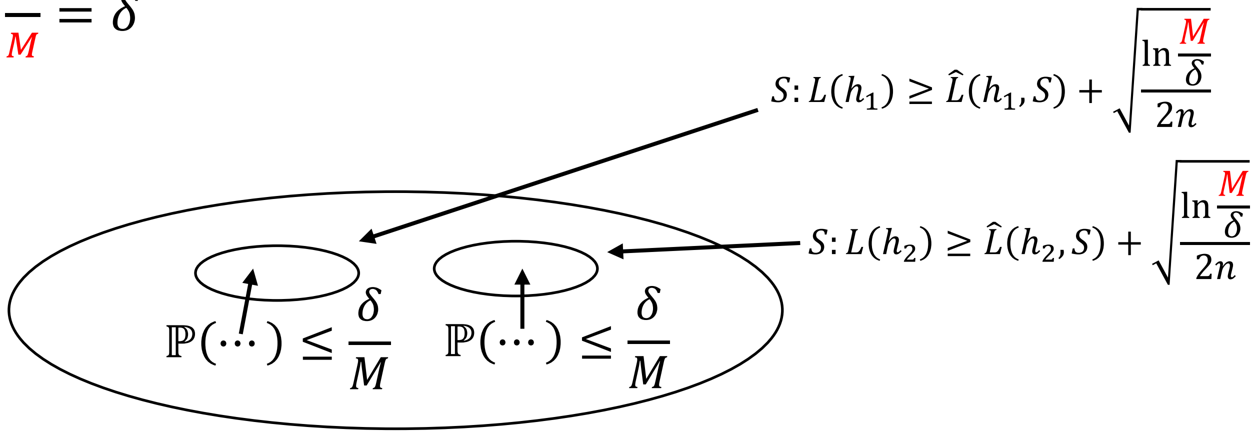
We break the dependence

$$\mathbb{P}\left(L(\hat{h}_S^*) \geq \hat{L}(\hat{h}_S^*, S) + \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}\right) \leq \mathbb{P}\left(\exists h \in \mathcal{H}:\ L(h) \geq \hat{L}(h, S) + \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}\right)$$

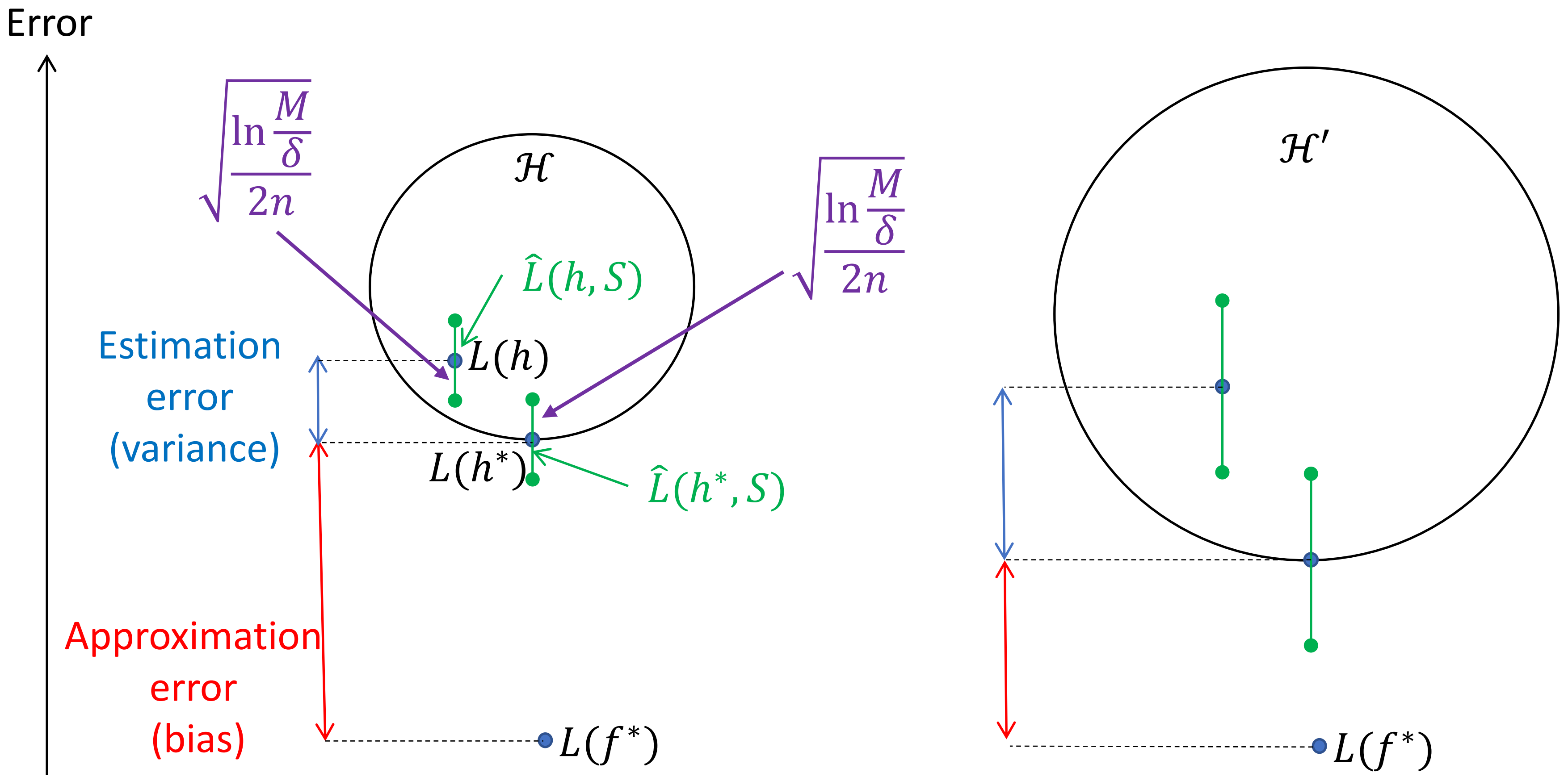
(Union bound) $$\leq \sum_{h\in\mathcal{H}} \mathbb{P}\left(L(h) \geq \hat{L}(h, S) + \sqrt{\frac{\ln\frac{M}{\delta}}{2n}}\right)$$

(Hoeffding with $\delta' = \frac{\delta}{M}$) $$\leq \sum_{h\in\mathcal{H}} \frac{\delta}{M} = \delta$$

In the background:
The space $(\mathcal{X} \times \mathcal{Y})^n$ of all possible samples S of size n
Each h gets $\frac{1}{M}$ share of the confidence budget δ
The total probability mass of violations of the inequality is bounded by δ



Approximation-Estimation (bias-variance) trade-off



Estimation error $L(h) - L(h^*)$ can be up to $2\sqrt{\frac{\ln\frac{M}{\delta}}{2n}}$

Selection from a small $\mathcal{H}$ Selection from a large $\mathcal{H}$

Occam's razor – Generalization bound for countable $\mathcal{H}$

- Theorem (Occam's razor): Let $\pi(h)$ be nonnegative and **independent of S** and satisfy $\sum_{h\in\mathcal{H}} \pi(h) \le 1$. Then:

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \ge \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}}\right) \le \delta.$$

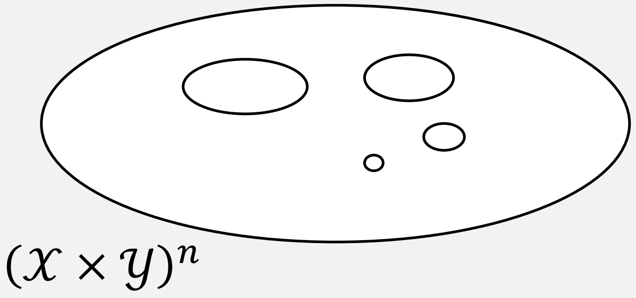
- Proof:

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \ge \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}}\right)$$

(Union bound) $\le \sum_{h\in\mathcal{H}} \mathbb{P}\left(L(h) \ge \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}}\right)$

(Hoeffding, π is independent of S!) $\le \sum_{h\in\mathcal{H}} \pi(h)\delta \le \delta$

In the background: uneven distribution of the confidence budget δ according to $\pi(h)$



Occam's razor selection

$$\mathbb{P}\left(\exists h \in \mathcal{H}: L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}\right) \leq \delta$$

$$h^* = \arg\min_{\mathrm{h}} \underbrace{\hat{L}(h,S)}_{\substack{\text{Empirical} \\ \text{Performance}}} + \underbrace{\sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}}_{\text{Complexity}}$$

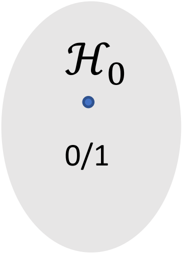
With probability at least $1-\delta$: $L(\hat{h}_S^*) \leq \hat{L}(\hat{h}_S^*, S) + \sqrt{\frac{\ln \frac{1}{\pi(h)\delta}}{2n}}$

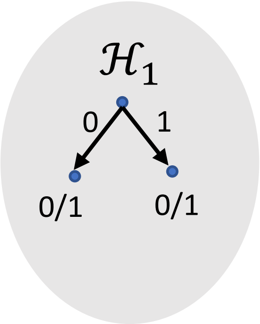
Application example: binary decision trees

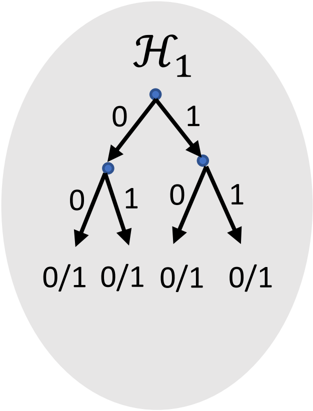
$$\pi(\mathcal{H}_0) = \frac{1}{2}$$
$$|\mathcal{H}_0| = 2 = 2^{2^0}$$

$$\pi(\mathcal{H}_1) = \frac{1}{2^2} = \frac{1}{4}$$
$$|\mathcal{H}_1| = 4 = 2^{2^1}$$

$$\pi(\mathcal{H}_2) = \frac{1}{2^3} = \frac{1}{8}$$
$$|\mathcal{H}_2| = 2^{2^2}$$







…

- $d(h)$ - depth of tree h
- $\pi(h) = \pi\left(\mathcal{H}_{d(h)}\right)\frac{1}{|\mathcal{H}_{d(h)}|} = \frac{1}{2^{d(h)+1}}\frac{1}{2^{2^{d(h)}}}$
- $\sum_{h\in\mathcal{H}}\pi(h) = \sum_{d=0}^{\infty}\sum_{h\in\mathcal{H}_d}\pi(h) = \sum_{d=0}^{\infty}\sum_{h\in\mathcal{H}}\frac{1}{2^{d(h)+1}}\frac{1}{2^{2^{d(h)}}} = \sum_{d=0}^{\infty}\frac{1}{2^{d(h)+1}}\underbrace{\sum_{h\in\mathcal{H}_d}\frac{1}{2^{2^{d(h)}}}}_{=1} = \sum_{d=0}^{\infty}\frac{1}{2^{d(h)+1}} = 1$

dominant

- With probability $\geq 1-\delta$, for all $h \in \mathcal{H}$: $L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln\frac{1}{\pi(h)\delta}}{2n}} = \hat{L}(h,S) + \sqrt{\frac{\ln(2)\left(2^{d(h)}+d(h)+1\right)+\ln\frac{1}{\delta}}{2n}}$

dominant

Alternative $\pi(\mathcal{H}_d) = \frac{1}{(d+1)(d+2)}$

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln(2)2^{d(h)} + \ln\frac{(d+1)(d+2)}{\delta}}{2n}}$$

- Permutation-symmetric trees get the same prior
- In absence of prior knowledge, no reason to discriminate (structurally symmetric prior)
- The size of $\mathcal{H}_d$ gives the dominant term

- Why no contradiction with the lower bound?