Statistical Learning Theory

Omar Montasser

Lecture 5

Non-Uniform Learning

Fundamental Theorem of Statistical Learning

Theorem. For any hypothesis class H with finite VC dimension:

- H is (realizably)-PAC-learnable with sample complexity
$$m(\epsilon, \delta) = \Theta\left(\frac{\mathrm{vc}(H) + \ln(1/\delta)}{\epsilon}\right).$$
ERM incurs a multiplicative $\log(1/\epsilon)$.
- H is (agnostically)-PAC-learnable with sample complexity
$$m(\epsilon, \delta) = \Theta\left(\frac{\mathrm{vc}(H) + \ln(1/\delta)}{\epsilon^2}\right).$$
Achieved by ERM.
- H satisfies the uniform convergence property with sample complexity
$$m(\epsilon, \delta) = \Theta\left(\frac{\mathrm{vc}(H) + \ln(1/\delta)}{\epsilon^2}\right).$$

Key Takeaways / Implications

- What is learnable? VC classes.
- How to learn? ERM.
- Tight quantitative understanding of sample complexity.

(Efficient) Proper Learning

Definition. A family of classes $\{H_n\}_{n\in\mathbb{N}}$ is efficiently-properly-PAC-learnable in the realizable setting if there exists a learning algorithm A such that
$\forall n \forall(\epsilon,\delta)\in(0,1)^2, \exists m(n,\epsilon,\delta)\in\mathbb{N}, \forall D$ s.t. $\inf_{h\in H} L_D(h)=0$,

$$\mathbb{P}_{S\sim D^{m(\epsilon,\delta)}}\left\{L_D(A(S))\leq\epsilon\right\}\geq 1-\delta,$$

and A runs in time polynomial in $n, 1/\epsilon, \log(1/\delta)$, and A always outputs a predictor in H_n.

We assume that RP $\neq$ NP. Then, for any family $\{H_n\}_{n\in\mathbb{N}}$:

- If $\mathrm{vc}(H_n)\leq\mathrm{poly}(n)$ and there is a poly-time algorithm implementing Consistent Learning, then family H_n is efficiently-properly-PAC learnable.
- If solving $\mathrm{CONS}_{H_n}(S)$ is NP-hard, then family H_n is *not* efficiently-properly-PAC learnable.

What about *improper* learning?

Hardness of Improper Learning

- Based on Cryptographic Assumptions:
  - “If crypto is possible, then efficient learning is impossible.”
- General Recipe:
  - Take a cryptographic problem that is assumed to be computationally intractable (in an average-case sense).
  - Define an appropriate hypothesis class family, and show that an efficient-PAC-learning algorithm for this family can be used to efficiently solve the cryptographic problem.
- Examples:
  - Assuming “Discrete Cube Root” is computationally intractable (the RSA public-key crypto system is based on this assumption), then
    - the class of log-depth polynomial-size circuits (AND/OR networks) is not efficiently-PAC-learnable (even improperly).
- Suggested Reading:
  - M. Kearns and U. Vazirani, An Introduction to Computational Learning Theory
    - Chapter 1 (Sections 1.3 — 1.5), and Chapter 6.

Relizable vs. Agnostic Learning

Definition. A family of classes $\{H_n\}_{n\in\mathbb{N}}$ is efficiently-properly-PAC-learnable in the realizable setting if there exists a learning algorithm A such that
$\forall n \forall(\epsilon,\delta)\in(0,1)^2, \exists m(n,\epsilon,\delta)\in\mathbb{N}, \forall D$ s.t. $\inf_{h\in H} L_D(h)=0$,

$$\mathbb{P}_{S\sim D^{m(\epsilon,\delta)}}\left\{L_D(A(S))\le\epsilon\right\}\ge 1-\delta,$$

and A runs in time polynomial in $n, 1/\epsilon, \log(1/\delta)$, and A always outputs a predictor in H_n.

Definition. A family of classes $\{H_n\}_{n\in\mathbb{N}}$ is efficiently-properly-PAC-learnable in the agnostic setting if there exists a learning algorithm A such that
$\forall n \forall(\epsilon,\delta)\in(0,1)^2, \exists m(n,\epsilon,\delta)\in\mathbb{N}, \forall D$

$$\mathbb{P}_{S\sim D^{m(\epsilon,\delta)}}\left\{L_D(A(S))\le \inf_{h\in H_n} L_D(h)+\epsilon\right\}\ge 1-\delta,$$

and A runs in time polynomial in $n, 1/\epsilon, \log(1/\delta)$, and A always outputs a predictor in H_n.

Conditions for Efficient Agnostic Learning

$$\mathrm{ERM}_{H_n}(S) = \arg\min_{h \in H_n} \frac{1}{|S|} \sum_{(x,y)\in S} \mathbf{1}\{h(x_i) \neq y_i\}$$

Claim.
- If $\mathrm{vc}(H_n) \leq \mathrm{poly}(n)$, and
- there is a poly-time algorithm implementing ERM for $\{H_n\}_{n\in\mathbb{N}}$,

then $\{H_n\}_{n\in\mathbb{N}}$ is efficiently-agnostically-properly-PAC-Learnable.

For a family $\{H_n\}_{n\in\mathbb{N}}$ consider the decision problem:

$$\mathrm{AGREEMENT}_{H_n}(S,k) = 1 \text{ iff } \exists h \in H_n, L_S(h) \leq 1 - \frac{k}{|S|}.$$

Claim. If H_n is efficiently-agnostically-properly-PAC-learnable then $\mathrm{AGREEMENT}_{H_n} \in \mathrm{RP}$.

Corollary. If $\mathrm{RP} \neq \mathrm{NP}$ and $\mathrm{AGREEMENT}_{H_n}$ is NP-hard, then H_n is not efficiently-agnostically-properly-PAC-learnable.

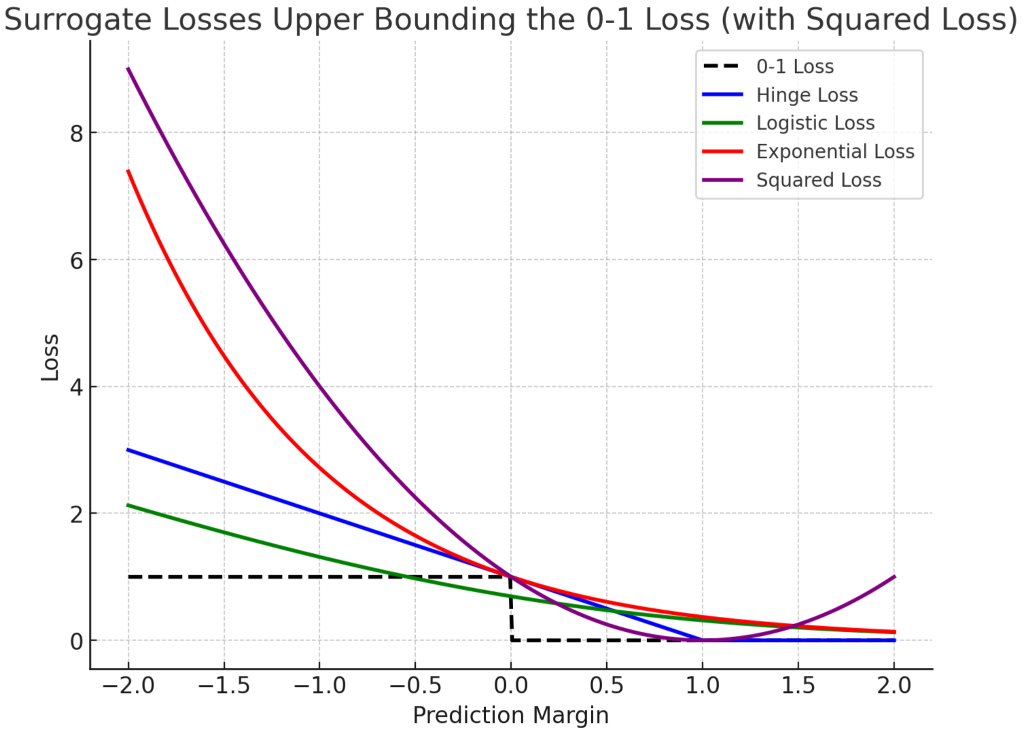
What is Efficiently Properly Agnostically PAC Learnable?

- Poly-time functions? No! (Not even in the realizable case)
- Poly-size depth-2 neural networks? No! (Not even in the realizable case)
- Halfspaces (linear predictors)?
  - $X_n = \{0,1\}^n, H_n = \{x \mapsto \mathbf{1}[\langle w, x\rangle > 0] : w \in \mathbb{R}^n\}$.
  - Claim: AGREEMENT$_{H_n}$ is NP-hard.
  - No!
- Conjunctions? No!
- Unions of segments on the line?
  - $X_n = [0,1], H_n = \left\{x \mapsto \vee_{i=1}^n \mathbf{1}[a_i \leq x \leq b_i] \mid a_i, b_i \in [0,1]\right\}$.
  - Yes! Efficiently Properly Agnostically PAC Learnable.

Surrogate Losses

$$\min_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i)$$

- For example, instead of the 0-1 loss $\ell^{01}(z, y) = \mathbf{1}[yz \leq 0]$, use:
  - Squared Loss: $\ell(z, y) = (z - y)^2$
  - Hinge Loss: $\ell(z, y) = \max\{0, 1 - yz\}$.
  - Logistic Loss: $\ell(z, y) = \log(1 + \exp(-yz))$
  - Exponential Loss: $\ell(z, y) = \exp(-yz)$.



High-Level Picture

- Computational efficiency is a major challenge in Machine Learning.
- In the face of worst-case hardness results, we sometimes need to use heuristics 启发的 (e.g., surrogate losses).
- Hardness results help illuminate what is not possible computationally, so we should direct our efforts elsewhere.
- One major challenge in machine learning theory is to reconcile the success of deep learning methods (based on local search procedures, e.g. Stochastic Gradient Descent) with worst-case hardness results.
    - It must be that learning problems where deep learning succeeds are not worst-case in nature, but understanding what makes these problems efficiently learnable is a major open research direction.

Non-Uniform Bias

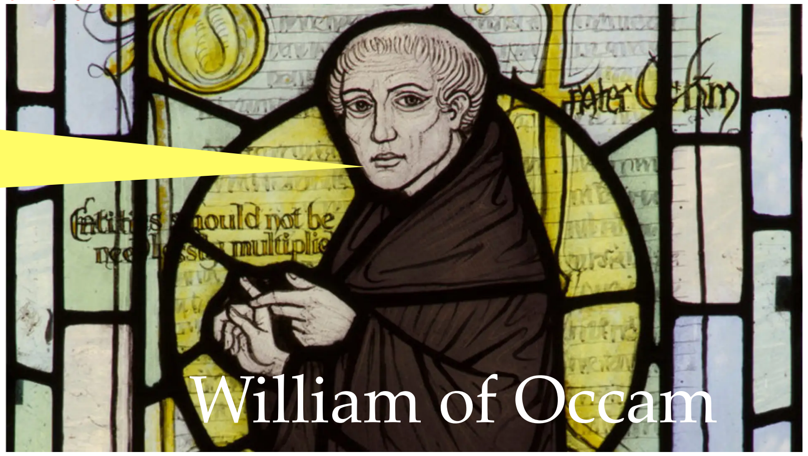
- So far: a uniform prior over H, each $h \in H$ is equally likely.
- Instead: prior $p(h)$ that encodes "preference" or "bias",
  - $p : H \to [0,1], \sum_{h \in H} p(h) \leq 1.$
- A more general way of encoding our prior knowledge.
- Bias towards simple predictors, $p(h)$ encodes "simplicity".
- Bias towards shorter explanations, $p(h)$ encodes "description length".

Non-Uniform Bias

- So far: a uniform prior over H, each $h \in H$ is equally likely.
- Instead: prior $p(h)$ that encodes "preference" or "bias",
  - $p : H \to [0,1], \sum_{h \in H} p(h) \leq 1.$
- A more general way of encoding our prior knowledge.
- Bias towards simple predictors, $p(h)$ encodes "simplicity".
- Bias towards shorter explanations, $p(h)$ encodes "description length".

偏向简单的预测器：p(h) 编码了“简单性”。这意味着，模型会赋予那些结构更简单、参数更少的假设更高的先验概率。
偏向更短的解释：p(h) 编码了“描述长度”。这意味着，那些能够用更简洁的方式描述的假设被认为更好。

Occam's Razor: A short explanation is preferred over a longer one.



Bias to Shorter Description

- Let H be a countable union of hypotheses.
- Let $d : H \to \{0,1\}^\star$ be a description language for H that is *prefix-free*: for any $h, h' \in H$, $d(h)$ is not a prefix of $d(h')$.
  - Define prior $p(h) = 2^{-|d(h)|}$.
  - Kraft's Inequality: $\sum_h 2^{-|d(h)|} \leq 1$.

Bias to Shorter Description

- Let H be a countable union of hypotheses.
- Let $d : H \to \{0,1\}^\star$ be a description language for H that is *prefix-free*: for any $h, h' \in H$, $d(h)$ is not a prefix of $d(h')$.
  - Define prior $p(h) = 2^{-|d(h)|}$.
  - Kraft's Inequality: $\sum_h 2^{-|d(h)|} \leq 1$.

$$\boxed{\mathrm{MDL}_p(S) = \arg\max_{L_S(h)=0} p(h) = \arg\min_{L_S(h)=0} |d(h)|\,.}$$

Bias to Shorter Description

$$\boxed{\mathrm{MDL}_p(S) = \arg\max_{L_S(h)=0} p(h) = \arg\min_{L_S(h)=0} |d(h)|\,.}$$

Theorem. For prior p over a countable H s.t. $\sum_h p(h) \le 1$ (e.g., $p(h) = 2^{-|d(h)|}$ for a prefix-free d), any $\delta \in (0,1)$, $m \in \mathbb{N}$, and any distribution D s.t. $\exists h^\star \in H, L_D(h^\star) = 0$, with probability $\ge 1 - \delta$ over $S \sim D^m$,

$$L_D(\mathrm{MDL}_p(S)) \le \frac{\ln\left(\frac{1}{p(h^\star)}\right) + \ln(1/\delta)}{m} = \frac{|d(h^\star)|\ln(2) + \ln(1/\delta)}{m}.$$

Proof. For any $h \in H$ such that $L_D(h) > \epsilon_h := \dfrac{\ln(1/p(h)) + \ln(1/\delta)}{m}$. Then, $\mathbb{P}_{S\sim D^m}\{L_S(h) = 0\} \le (1-\epsilon_h)^m \le e^{-\epsilon_h m} = p(h)\cdot\delta$. By a union bound, $\mathbb{P}_{S\sim D^m}\{\exists h, L_D(h) > \epsilon_h : L_S(h) = 0\} \le \sum_{h\in H} \mathbb{P}_{S\sim D^m}\{L_S(h) = 0\} \le \sum_{h\in H} p(h)\delta \le \delta$. Finally, since $\exists h^\star \in H, L_D(h^\star) = 0$, it follows that $p\left(\mathrm{MDL}_p(S)\right) \ge p(h^\star)$.

MDL and Universal Learning

Theorem. For prior p over a countable H s.t. $\sum_h p(h) \leq 1$ (e.g., $p(h) = 2^{-|d(h)|}$ for a prefix-free d), any $\delta \in (0,1)$, $m \in \mathbb{N}$, and any distribution D s.t. $\exists h^\star \in H, L_D(h^\star) = 0$, with probability $\geq 1 - \delta$ over $S \sim D^m$,

$$L_D(\mathrm{MDL}_p(S)) \leq \frac{\ln\left(\frac{1}{p(h^\star)}\right) + \ln(1/\delta)}{m} = \frac{|\,d(h^\star)\,|\ln(2) + \ln(1/\delta)}{m}.$$

- Can learn any countable class!
  - Class of all computable functions.
  - Class numerable with $n : H \to \mathbb{N}$ with $p(h) = 2^{-n(h)}$.
- But the VC dimension of all computable functions is infinite!
- Why no contradiction to Fundamental Theorem?
  - PAC Learning: sample complexity $m(\epsilon, \delta)$ is uniform over all $h \in H$. Depends on H, but not on any specific $h^\star \in H$
  - MDL: sample complexity $m(\epsilon, \delta, h)$ depends on $h \in H$.

Uniform and Non-Uniform Learning

Definition. A hypothesis class H is agnostically-PAC-learnable if there exists a learning rule A such that $\forall(\epsilon, \delta) \in (0,1)^2$, $\exists m(\epsilon, \delta) \in \mathbb{N}$, $\forall h \in H$, $\forall D$,

$$\mathbb{P}_{S \sim D^{m(\epsilon,\delta)}} \left\{ L_D(A(S)) \leq L_D(h) + \epsilon \right\} \geq 1 - \delta \, .$$

Definition. A hypothesis class H is non-uniformly-learnable if there exists a learning rule A such that $\forall(\epsilon, \delta) \in (0,1)^2$, $\forall h \in H$, $\exists m(h, \epsilon, \delta) \in \mathbb{N}$, $\forall D$,

$$\mathbb{P}_{S \sim D^{m(h,\epsilon,\delta)}} \left\{ L_D(A(S)) \leq L_D(h) + \epsilon \right\} \geq 1 - \delta \, .$$

Corollary. For any prior p over a countable H s.t. $\sum_h p(h) \leq 1$ and any $h^\star \in H$. With sample complexity

$$m(h^\star, \epsilon, \delta) = \frac{\ln\left(\frac{1}{p(h^\star)}\right) + \ln(1/\delta)}{\epsilon} \, ,$$

for any distribution D s.t. $L_D(h^\star) = 0$,

$$\mathbb{P}_{S \sim D^{m(h^\star,\epsilon,\delta)}} \left\{ L_D(\mathrm{MDL}_p(S)) \leq \epsilon \right\} \geq 1 - \delta \, .$$

So far: guarantee in the realizable setting. What about general non-uniform learning?

Allowing Errors: Structural Risk Minimization

Given a prior p over H, with probability $\geq 1 - \delta$ over $S \sim D^m$:

$$L_D(h) \leq \underbrace{L_S(h)}_{\text{Minimized by ERM}} + \underbrace{\sqrt{\frac{\ln(1/p(h)) + \ln(2/\delta)}{2m}}}_{\text{Minimized by MDL}}.$$

Allowing Errors: Structural Risk Minimization

Given a prior p over H, with probability $\geq 1-\delta$ over $S \sim D^m$:

$$L_D(h) \leq \underbrace{L_S(h)}_{\text{Minimized by ERM}} + \underbrace{\sqrt{\frac{\ln(1/p(h)) + \ln(2/\delta)}{2m}}}_{\text{Minimized by MDL}}.$$

$$\boxed{\mathrm{SRM}_p(S) = \arg\min_h \underbrace{L_S(h)}_{\text{Fit data}} + \underbrace{\sqrt{\frac{\ln(1/p(h))}{2m}}}_{\text{Match the prior / simple / short description}}}$$

Theorem. For prior p over a countable H s.t. $\sum_h p(h) \leq 1$, any distribution D, any $\delta \in (0,1), m \in \mathbb{N}$, with probability $\geq 1-\delta$ over $S \sim D^m$:

$$L_D(\mathrm{SRM}_p(S)) \leq \inf_h \left(L_D(h) + 2\sqrt{\frac{\ln(1/p(h)) + \ln(2/\delta)}{2m}} \right).$$

Non-Uniform Learning: Beyond Cardinality

- So far: we considered countable classes H.
- Essentially, we generalized a cardinality-based bound using a prior $p : H \to [0,1]$.
- What about *uncountable* classes?
  - Example: $H = \{x \mapsto \mathrm{sign}(f(x)) \mid f : \mathbb{R}^d \to \mathbb{R} \text{ is a polynomial}\}$.
  - $\mathrm{vc}(H) = \infty$.
  - H is uncountable, and there is no prior p such that $\forall_{h \in H}\, p(h) > 0$.
  - What if we bias towards lower order polynomials?
- Answer 1: use a prior over *hypothesis classes*.
  - Describe $H = \cup_{r \in \mathbb{N}} H_r$ (e.g., H_r is degree-r polynomials).
  - Use prior $p(H_r)$ over hypothesis classes.

Prior Over Hypothesis Classes

- VC bound: $\forall_r \mathbb{P}_{S \sim D^m}\left[\forall h \in H_r : L_D(h) \leq L_S(h) + \epsilon_r\right] \geq 1 - \delta_r$, where $\epsilon_r = O\left(\sqrt{\frac{\mathrm{vc}(H_r) + \ln(1/\delta_r)}{m}}\right)$.
- Setting $\delta_r = p(H_r) \cdot \delta$ and taking a union bound over r implies:
  $\mathbb{P}_{S \sim D^m}\left[\forall r \forall h \in H_r : L_D(h) \leq L_S(h) + \epsilon_r\right] \geq 1 - \delta$, where $\epsilon_r = O\left(\sqrt{\frac{\mathrm{vc}(H_r) + \ln(1/p(H_r)) + \ln(1/\delta)}{m}}\right)$.

$$\boxed{\mathrm{SRM}_p(S) = \arg\min_{r, h \in H_r} L_S(h) + c\sqrt{\frac{\mathrm{vc}(H_r) + \ln(1/p(H_r))}{m}}}$$

Theorem. For $H = \cup_{r \in \mathbb{N}} H_r$ and prior $p(H_r)$ s.t. $\sum_r p(H_r) \leq 1$, any distribution D, any $\delta \in (0,1), m \in \mathbb{N}$, with probability $\geq 1 - \delta$ over $S \sim D^m$:

$$L_D(\mathrm{SRM}_p(S)) \leq \inf_{r, h \in H_r}\left(L_D(h) + c\sqrt{\frac{\mathrm{vc}(H_r) + \ln(1/p(H_r)) + \ln(1/\delta)}{m}}\right).$$

Prior Over Hypothesis Classes

Theorem. For $H = \cup_{r \in \mathbb{N}} H_r$ and prior $p(H_r)$ s.t. $\sum_r p(H_r) \le 1$, any distribution D, any $\delta \in (0,1), m \in \mathbb{N}$, with probability $\ge 1 - \delta$ over $S \sim D^m$:

$$L_D(\mathrm{SRM}_p(S)) \le \inf_{r, h \in H_r} \left(L_D(h) + c\sqrt{\frac{\mathrm{vc}(H_r) + \ln(1/p(H_r)) + \ln(1/\delta)}{m}} \right).$$

- When $H_r = \{h_r\}$, $\mathrm{vc}(H_r) = 0$, reduces to "standard" SRM guarantee over countable class.
- When there is r_0 such that $p(H_{r_0}) = 1$, reduces to ERM over H_r.
- $H = \{x \mapsto \mathrm{sign}(f(x)) \mid f : \mathbb{R}^d \to \mathbb{R} \text{ is a polynomial}\}$
  - $H = \cup_{r \in \mathbb{N}} H_r$ (H_r is degree-r polynomials), and prior $p(H_r) = 2^{-r}$.
  - $m(h, \epsilon, \delta) = O\left(\dfrac{d^{\deg(h)} + \deg(h) + \ln(1/\delta)}{\epsilon^2} \right).$

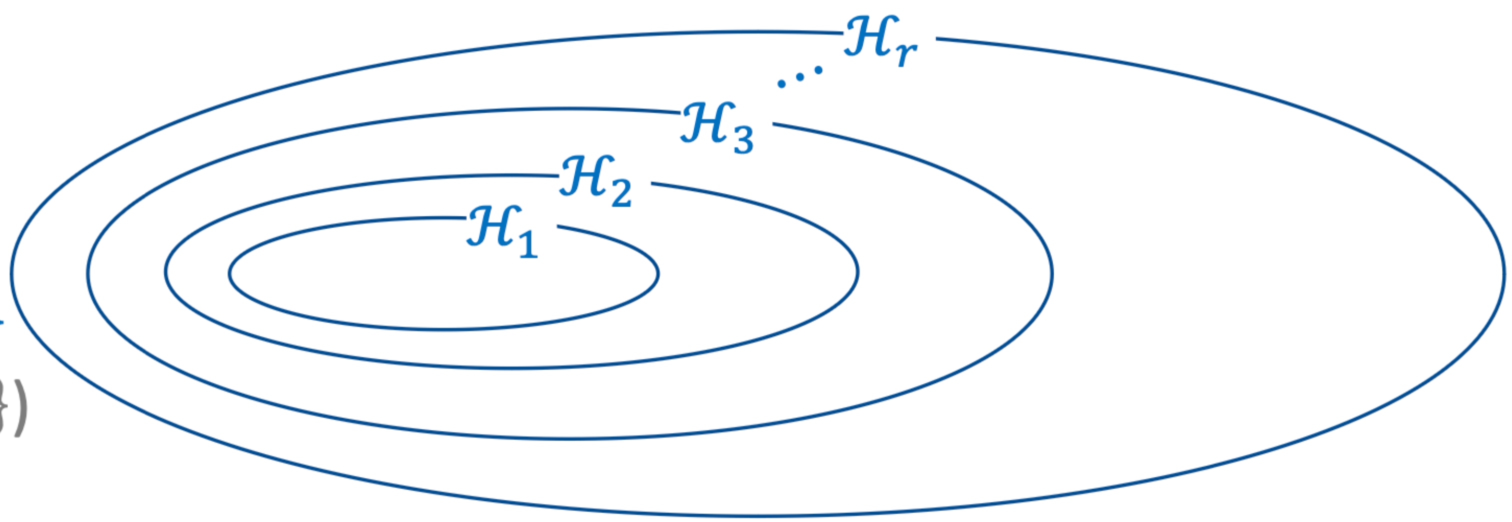
SRM in Practice

$$\mathrm{SRM}_p(S) = \arg \min_{r, h \in H_r} L_S(h) + c\sqrt{\frac{\mathrm{vc}(H_r) + \ln(1/p(H_r))}{m}}$$

- Typically, $\mathrm{vc}(H_r)$ and $\ln(1/p(H_r))$ are monotone in "complexity" $r : H \to \mathbb{N}$.

$\mathcal{H}_r = \{h | r(h) \leq r\}$

(or $\mathcal{H}_r = \{h | r(h) = r\}$)

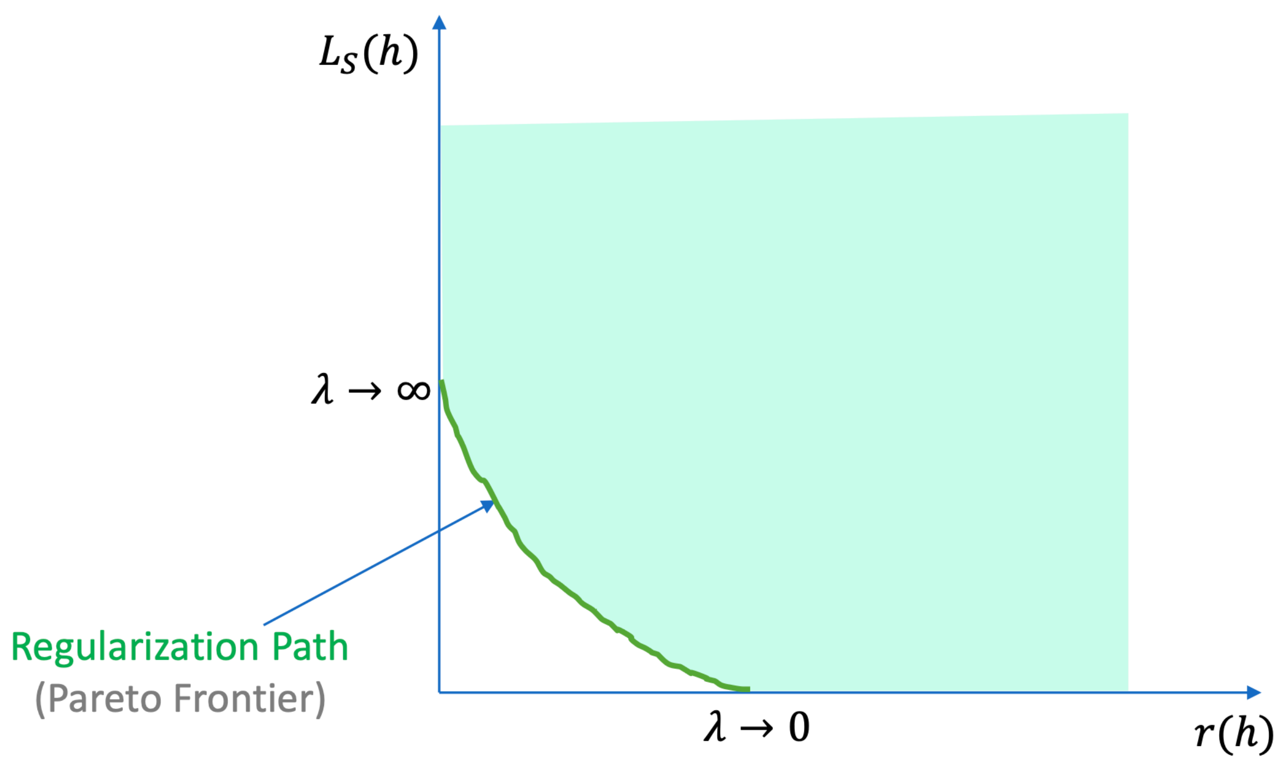


- View as a bi-criterion optimization problem:
  - $\arg\min\{L_S(h)$ and $r(h)\}$, where $r(h) = \min\{r \in \mathbb{N} \mid h \in H_r\}$.

SRM in Practice

$$\arg\min L_S(h) \text{ and } r(h)$$

$$\text{Regularization-Path} = \{\arg\min L_S(h) + \lambda r(h) \mid 0 \leq \lambda < \infty\}$$
$$\text{, or: } \{\arg\min L_S(h) \text{ s.t. } r(h) \leq r \mid 0 \leq r < \infty\}$$



Uniform and Non-Uniform Learning

Definition. A hypothesis class H is agnostically-PAC-learnable if there exists a learning rule A such that $\forall(\epsilon,\delta) \in (0,1)^2$, $\exists m(\epsilon,\delta) \in \mathbb{N}$, $\forall h \in H$, $\forall D$,

$$\mathbb{P}_{S\sim D^{m(\epsilon,\delta)}}\left\{L_D(A(S)) \leq L_D(h) + \epsilon\right\} \geq 1 - \delta\,.$$

Definition. A hypothesis class H is non-uniformly-learnable if there exists a learning rule A such that $\forall(\epsilon,\delta) \in (0,1)^2$, $\forall h \in H$, $\exists m(h,\epsilon,\delta) \in \mathbb{N}$, $\forall D$,

$$\mathbb{P}_{S\sim D^{m(h,\epsilon,\delta)}}\left\{L_D(A(S)) \leq L_D(h) + \epsilon\right\} \geq 1 - \delta\,.$$

Theorem. A hypothesis class H is non-uniformly-learnable if and only if H is a countable union of finite VC classes ($H = \cup_{r\in\mathbb{N}} H_r$ and $\forall_r \mathrm{vc}(H_r) < \infty$).

Definition. A hypothesis class H is consistently-learnable if there exists a learning rule A such that $\forall(\epsilon,\delta) \in (0,1)^2$, $\forall h \in H$, $\forall D$, $\exists m(h,D,\epsilon,\delta) \in \mathbb{N}$,

$$\mathbb{P}_{S\sim D^{m(h,\epsilon,\delta)}}\left\{L_D(A(S)) \leq L_D(h) + \epsilon\right\} \geq 1 - \delta\,.$$

Claim. There exists domain X and a class H that is consistently-learnable but **not** non-uniformly-learnable.

Non-Uniform Learning: Beyond Cardinality

- So far: we considered countable classes H.
- Essentially, we generalized a cardinality-based bound using a prior $p : H \to [0,1]$.
- What about *uncountable* classes?
- Answer 1: use a prior over *hypothesis classes*.
- Answer 2: PAC-Bayes Theory.
  - Prior P (not necessarily discrete) over H.

David McAllister

PAC-Bayes

- So far we have used a discrete prior/distribution over hypotheses, or discrete prior over hypothesis classes (in MDL and SRM).

- What about arbitrary distributions/priors over uncountable H?

- Consider randomized (average) predictor h_Q defined as:

  - $h_Q(x) = y$ w.p. $\mathbb{P}_{h \sim Q}(h(x) = y)$.

  - $L_D(h_Q) = \mathbb{E}_{(x,y) \sim D} \mathbb{E}_{h \sim Q} \mathbf{1}[h(x) \neq y] = \mathbb{E}_{h \sim Q} L_D(h)$.

Theorem. For any class H and any prior distribution P over H, any distribution D over $X \times Y$, any $\delta \in (0,1), m \in \mathbb{N}$, with probability $\geq 1 - \delta$ over $S \sim D^m$:

$$\forall \text{ posterior dist'ns } Q \text{ over } H : |\, L_D(h_Q) - L_S(h_Q) \,| \leq \sqrt{\frac{\mathrm{KL}(Q\,||\,P) + \log(2m/\delta)}{2(m-1)}}.$$

PAC-Bayes

Theorem. For any class H and any prior distribution P over H, any distribution D over $X \times Y$, any $\delta \in (0,1), m \in \mathbb{N}$, with probability $\geq 1 - \delta$ over $S \sim D^m$:

$$\forall \text{ posterior dist'ns } Q \text{ over } H : |\, L_D(h_Q) - L_S(h_Q)\, | \leq \sqrt{\frac{\mathrm{KL}(Q\,||\,P) + \log(2m/\delta)}{2(m-1)}}.$$

- Non-vacuous only when $\mathrm{supp}(Q) \subseteq \mathrm{supp}(P)$.
- Finite H with P being uniform over H.
    - For Q being point mass on some $h \in H$, $\mathrm{KL}(Q\,||\,P) = \log|H|$.
- More generally, for discrete distributions P and point-mass Q,
    - $\mathrm{KL}(Q\,||\,P) = \log(1/P(h))$.
- For continuous P (e.g., over linear predictors or polynomials)
    - If Q is point mass, then $\mathrm{KL}(Q\,||\,P)$ is infinite.

PAC-Bayes

Theorem. For any class H and any prior distribution P over H, any distribution D over $X \times Y$, any $\delta \in (0,1)$, $m \in \mathbb{N}$, with probability $\geq 1 - \delta$ over $S \sim D^m$:

$$\forall \text{ posterior dist'ns } Q \text{ over } H: \left| L_D(h_Q) - L_S(h_Q) \right| \leq \sqrt{\frac{\mathrm{KL}(Q \,||\, P) + \log(2m/\delta)}{2(m-1)}}.$$

$$\text{SRM-style:} \quad \arg\min_Q L_S(h_Q) + \lambda \mathrm{KL}(Q \,||\, P)$$

Claim. Solution is $Q_\lambda(h) \propto P(h) e^{-\eta L_S(h)}$, for some "inverse temperature" η.

Summary

- Non-uniform Learning.

- Occam's Razor.

- Minimum-Description-Length and Structural-Risk-Minimization.

- Takeaway: any target concept is learnable with sample complexity depending on its description length.