Fundamentals of Learning and Information Processing

Session 14: Statistical Machine Learning (IV)

Gang Li

Deakin University, Australia

2021-09-19

Table of Content

The Fundamental Theorem of PAC Learning

The Fundamental Theorem of PAC Learning

Let $\mathscr{H}$ be a hypothesis class of functions from a domain $\mathscr{X}$ to $\{0,1\}$, and let the loss function be the $0-1$ loss. Then the following statements are equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule is a successful agnostic PAC learner for $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. $\mathscr{H}$ is PAC learnable
5. Any ERM rule is a successful PAC learner for $\mathscr{H}$
6. $\mathscr{H}$ has a finite VC-dimension

The Fundamental Theorem of PAC Learning

Let $\mathscr{H}$ be a hypothesis class of functions from a domain $\mathscr{X}$ to $\{0,1\}$, and let the loss function be the $0-1$ loss. Then the following statements are equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule is a successful agnostic PAC learner for $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. $\mathscr{H}$ is PAC learnable
5. Any ERM rule is a successful PAC learner for $\mathscr{H}$
6. $\mathscr{H}$ has a finite VC-dimension

Intuition.

- ① → ② was proved in UC property

The Fundamental Theorem of PAC Learning

Let $\mathscr{H}$ be a hypothesis class of functions from a domain $\mathscr{X}$ to $\{0,1\}$, and let the loss function be the $0-1$ loss. Then the following statements are equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule is a successful agnostic PAC learner for $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. $\mathscr{H}$ is PAC learnable
5. Any ERM rule is a successful PAC learner for $\mathscr{H}$
6. $\mathscr{H}$ has a finite VC-dimension

Intuition.

- ① → ② was proved in UC property
- ② → ③, ② → ⑤ and ③ → ④ are trivial

The Fundamental Theorem of PAC Learning

Let $\mathscr{H}$ be a hypothesis class of functions from a domain $\mathscr{X}$ to $\{0,1\}$, and let the loss function be the $0-1$ loss. Then the following statements are equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule is a successful agnostic PAC learner for $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. $\mathscr{H}$ is PAC learnable
5. Any ERM rule is a successful PAC learner for $\mathscr{H}$
6. $\mathscr{H}$ has a finite VC-dimension

Intuition.

- ① → ② was proved in UC property
- ② → ③, ② → ⑤ and ③ → ④ are trivial
- ④ → ⑥ and ⑤ → ⑥ follow from the NFL theorem.

The Fundamental Theorem of PAC Learning

Let $\mathscr{H}$ be a hypothesis class of functions from a domain $\mathscr{X}$ to $\{0,1\}$, and let the loss function be the $0-1$ loss. Then the following statements are equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule is a successful agnostic PAC learner for $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. $\mathscr{H}$ is PAC learnable
5. Any ERM rule is a successful PAC learner for $\mathscr{H}$
6. $\mathscr{H}$ has a finite VC-dimension

Intuition.

- ① → ② was proved in UC property
- ② → ③, ② → ⑤ and ③ → ④ are trivial
- ④ → ⑥ and ⑤ → ⑥ follow from the NFL theorem.
- ⑥ → ①: proven by *Vapnik* and *Chervonenkis* (1971), and related to PAC by *Blumer*, *Ehrenfeucht*, *Haussler* and *Warmuth* (1989)

The Fundamental Theorem of PAC Learning

Let $\mathscr{H}$ be a hypothesis class of functions from a domain $\mathscr{X}$ to $\{0,1\}$, and let the loss function be the $0-1$ loss. Then the following statements are equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule is a successful agnostic PAC learner for $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. $\mathscr{H}$ is PAC learnable
5. Any ERM rule is a successful PAC learner for $\mathscr{H}$
6. $\mathscr{H}$ has a finite VC-dimension

Intuition.

- ① → ② was proved in UC property
- ② → ③, ② → ⑤ and ③ → ④ are trivial
- ④ → ⑥ and ⑤ → ⑥ follow from the NFL theorem.
- ⑥ → ①: proven by *Vapnik* and *Chervonenkis* (1971), and related to PAC by *Blumer*, *Ehrenfeucht*, *Haussler* and *Warmuth* (1989)
  - If $VCDim(\mathscr{H}) = d$, then even though $\mathscr{H}$ might be infinite, when restricting to a finite set $C \subseteq \mathscr{X}$, the size of $|\mathscr{H}_C|$ grows polynomially with $|C|$

The Fundamental Theorem of PAC Learning

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$, and let the loss function be the $0-1$ loss. Then the following statements are equivalent:

1. $\mathcal{H}$ has the uniform convergence property
2. Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$
3. $\mathcal{H}$ is agnostic PAC learnable
4. $\mathcal{H}$ is PAC learnable
5. Any ERM rule is a successful PAC learner for $\mathcal{H}$
6. $\mathcal{H}$ has a finite VC-dimension

Intuition.

- ① → ② was proved in UC property
- ② → ③, ② → ⑤ and ③ → ④ are trivial
- ④ → ⑥ and ⑤ → ⑥ follow from the NFL theorem.
- ⑥ → ①: proven by *Vapnik* and *Chervonenkis* (1971), and related to PAC by *Blumer*, *Ehrenfeucht*, *Haussler* and *Warmuth* (1989)
  - If $VCDim(\mathcal{H}) = d$, then even though $\mathcal{H}$ might be infinite, when restricting to a finite set $C \subseteq \mathcal{X}$, the size of $|\mathcal{H}_C|$ grows polynomially with $|C|$
  - Uniform convergence holds for the "small effective size" class $\mathcal{H}$.

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$, and let the loss function be the $0-1$ loss. Assume that $VCDim(\mathcal{H}) = d < \infty$. Then, there are absolute constants C_1 and C_2 such that

1. $\mathcal{H}$ has the uniform convergence property with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \le m_{\mathcal{H}}^{UC}(\epsilon,\delta) \le C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

2. $\mathcal{H}$ is agnostic PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \le m_{\mathcal{H}}^{APAC}(\epsilon,\delta) \le C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

3. $\mathcal{H}$ is PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \le m_{\mathcal{H}}^{PAC}(\epsilon,\delta) \le C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

The Fundamental Theorem of PAC Learning (Quantitative)

Let $\mathscr{H}$ be a hypothesis class of functions from a domain $\mathscr{X}$ to $\{0,1\}$, and let the loss function be the $0-1$ loss. Assume that $VCDim(\mathscr{H}) = d < \infty$. Then, there are absolute constants C_1 and C_2 such that

1. $\mathscr{H}$ has the uniform convergence property with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathscr{H}}^{UC}(\epsilon,\delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

2. $\mathscr{H}$ is agnostic PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathscr{H}}^{APAC}(\epsilon,\delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$

3. $\mathscr{H}$ is PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathscr{H}}^{PAC}(\epsilon,\delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

Intuition.

- *NFL* is on the lower bound; while *Hoffding's Inequality* is on the upper bound.

Proof of the Lower Bounds

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}^{PAC}(\epsilon, \delta)$$

Proof of the Lower Bounds

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}^{PAC}(\epsilon, \delta)$$

Intuition.

- From NFL theorem, we have $\frac{d}{2} \leq m_{\mathcal{H}}^{PAC}(1/8, 1/7)$

Proof of the Lower Bounds

$C_1 \frac{d+\log(1/\delta)}{\epsilon} \le m_{\mathcal{H}}^{PAC}(\epsilon,\delta)$

Intuition.

- From NFL theorem, we have $\frac{d}{2} \le m_{\mathcal{H}}^{PAC}(1/8, 1/7)$
- Pick a set $S = \{x_1, \cdots, x_d\}$ of size d that is shattered by $\mathcal{H}$. Let $\mathcal{D}$ be:

$$p(x_i) = \begin{cases} 1-\epsilon & \text{if } i = 1 \\ \frac{\epsilon}{d-1} & i \in [2, d] \\ 0 & \text{otherwise} \end{cases}$$

Proof of the Lower Bounds

$C_1 \frac{d+\log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}^{PAC}(\epsilon, \delta)$

Intuition.

- From NFL theorem, we have $\frac{d}{2} \leq m_{\mathcal{H}}^{PAC}(1/8, 1/7)$
- Pick a set $S = \{x_1, \cdots, x_d\}$ of size d that is shattered by $\mathcal{H}$. Let $\mathcal{D}$ be:

$$p(x_i) = \begin{cases} 1-\epsilon & \text{if } i = 1 \\ \frac{\epsilon}{d-1} & i \in [2, d] \\ 0 & \text{otherwise} \end{cases}$$

- By the NFL, as long as the sample hits $\{x_2, \cdots, x_d\}$ at most $\frac{d-1}{2}$ times, the probability of making an error over $\{x_2, \cdots, x_d\}$ is $\geq \frac{1}{4}$.

Proof of the Lower Bounds

$C_1 \frac{d+\log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}^{PAC}(\epsilon, \delta)$

Intuition.

- From NFL theorem, we have $\frac{d}{2} \leq m_{\mathcal{H}}^{PAC}(1/8, 1/7)$
- Pick a set $S = \{x_1, \cdots, x_d\}$ of size d that is shattered by $\mathcal{H}$. Let $\mathcal{D}$ be:

$$p(x_i) = \begin{cases} 1-\epsilon & \text{if } i = 1 \\ \frac{\epsilon}{d-1} & i \in [2, d] \\ 0 & \text{otherwise} \end{cases}$$

- By the NFL, as long as the sample hits $\{x_2, \cdots, x_d\}$ at most $\frac{d-1}{2}$ times, the probability of making an error over $\{x_2, \cdots, x_d\}$ is $\geq \frac{1}{4}$.
- So the overall expected error will be $\frac{\epsilon}{4}$.

Proof of the Lower Bounds

$C_1 \frac{d+\log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}^{PAC}(\epsilon,\delta)$

Intuition.

- From NFL theorem, we have $\frac{d}{2} \leq m_{\mathcal{H}}^{PAC}(1/8, 1/7)$
- Pick a set $S = \{x_1, \cdots, x_d\}$ of size d that is shattered by $\mathcal{H}$. Let $\mathcal{D}$ be:

$$p(x_i) = \{ \begin{array}{ll} 1-\epsilon & \text{if } i = 1 \\ \frac{\epsilon}{d-1} & i \in [2, d] \\ 0 & \text{otherwise} \end{array}$$

- By the NFL, as long as the sample hits $\{x_2, \cdots, x_d\}$ at most $\frac{d-1}{2}$ times, the probability of making an error over $\{x_2, \cdots, x_d\}$ is $\geq \frac{1}{4}$.
- So the overall expected error will be $\frac{\epsilon}{4}$.
- Only roughly $m\epsilon$ of a sample S of size m will hit $\{x_2, \cdots, x_d\}$.

Proof of the Lower Bounds

$C_1 \frac{d+\log(1/\delta)}{\epsilon} \le m_{\mathcal{H}}^{PAC}(\epsilon,\delta)$

Intuition.

- From NFL theorem, we have $\frac{d}{2} \le m_{\mathcal{H}}^{PAC}(1/8, 1/7)$
- Pick a set $S = \{x_1, \cdots, x_d\}$ of size d that is shattered by $\mathcal{H}$. Let $\mathcal{D}$ be:

$$p(x_i) = \begin{cases} 1-\epsilon & \text{if } i = 1 \\ \frac{\epsilon}{d-1} & i \in [2, d] \\ 0 & \text{otherwise} \end{cases}$$

- By the NFL, as long as the sample hits $\{x_2, \cdots, x_d\}$ at most $\frac{d-1}{2}$ times, the probability of making an error over $\{x_2, \cdots, x_d\}$ is $\ge \frac{1}{4}$.
- So the overall expected error will be $\frac{\epsilon}{4}$.
- Only roughly $m\epsilon$ of a sample S of size m will hit $\{x_2, \cdots, x_d\}$.
- So in order to get the expected error below $\frac{\epsilon}{4}$, we need the $m\epsilon > \frac{d-1}{2}$, so $m > \frac{d-1}{2\epsilon}$

Proof of the Lower Bounds

$C_1 \frac{d+\log(1/\delta)}{\epsilon} \le m_{\mathscr{H}}^{PAC}(\epsilon,\delta)$

Intuition.

- From NFL theorem, we have $\frac{d}{2} \le m_{\mathscr{H}}^{PAC}(1/8, 1/7)$
- Pick a set $S = \{x_1, \cdots, x_d\}$ of size d that is shattered by $\mathscr{H}$. Let $\mathscr{D}$ be:

$$p(x_i) = \begin{cases} 1-\epsilon & \text{if } i = 1 \\ \frac{\epsilon}{d-1} & i \in [2,d] \\ 0 & \text{otherwise} \end{cases}$$

- By the NFL, as long as the sample hits $\{x_2, \cdots, x_d\}$ at most $\frac{d-1}{2}$ times, the probability of making an error over $\{x_2, \cdots, x_d\}$ is $\ge \frac{1}{4}$.
- So the overall expected error will be $\frac{\epsilon}{4}$.
- Only roughly $m\epsilon$ of a sample S of size m will hit $\{x_2, \cdots, x_d\}$.
- So in order to get the expected error below $\frac{\epsilon}{4}$, we need the $m\epsilon > \frac{d-1}{2}$, so $m > \frac{d-1}{2\epsilon}$
- Now we have $\forall \epsilon$, $E_{S\sim\mathscr{D}^m}[L_{\mathscr{D}}(A(S))] < \frac{\epsilon}{4}$, then $m > \frac{d-1}{2\epsilon}$

□

Proof of the Lower Bounds

$C_1 \frac{d+\log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}^{APAC}$

Proof of the Lower Bounds

$C_1 \frac{d+\log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}^{APAC}$

Intuition.

- Why for agnostic case, we need so many more examples?

Proof of the Lower Bounds

$C_1 \frac{d+\log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}^{APAC}$

Intuition.

- Why for agnostic case, we need so many more examples?
- Consider one sample with two points $S = \{x, y\}$ both points with probability $\frac{1}{2}$ to be selected. $P(1|X) = P(0|Y) = \frac{1}{2} + \epsilon$ and $P(0|X) = P(1|Y) = \frac{1}{2} - \epsilon$

Proof of the Lower Bounds

$$C_1 \frac{d+\log(1/\delta)}{\epsilon^2} \le m_{\mathcal{H}}^{APAC}$$

Intuition.

- Why for agnostic case, we need so many more examples?
- Consider one sample with two points $S = \{x, y\}$ both points with probability $\frac{1}{2}$ to be selected. $P(1|X) = P(0|Y) = \frac{1}{2} + \epsilon$ and $P(0|X) = P(1|Y) = \frac{1}{2} - \epsilon$
- $\mathcal{H} = \{(1,1),(1,0),(0,1),(0,0)\}$ contains all functions over $\{x, y\}$. Bayes predictor (best one) is $(1,0)$.

Proof of the Lower Bounds

$C_1 \frac{d+\log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}^{APAC}$

Intuition.

- Why for agnostic case, we need so many more examples?
- Consider one sample with two points $S = \{x, y\}$ both points with probability $\frac{1}{2}$ to be selected. $P(1|X) = P(0|Y) = \frac{1}{2} + \epsilon$ and $P(0|X) = P(1|Y) = \frac{1}{2} - \epsilon$
- $\mathcal{H} = \{(1,1),(1,0),(0,1),(0,0)\}$ contains all functions over $\{x, y\}$. Bayes predictor (best one) is $(1,0)$.
- In order to detect if a coin has $\frac{1}{2} + \epsilon$ towards tails or heads, one needs $\frac{1}{\epsilon^2}$ coin tosses, as ϵ gets small

$\square$

The Growth Function

Let $\mathscr{H}$ be a hypothesis class, then the growth function of $\mathscr{H}$, denoted $\tau_{\mathscr{H}} : \mathscr{N} \to \mathscr{N}$, is defined as:

$$\tau_{\mathscr{H}}(m) = \max_{C \subseteq \mathscr{X} : |C| = m} |\mathscr{H}_C|$$

The Growth Function

Let $\mathcal{H}$ be a hypothesis class, then the growth function of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathcal{N} \to \mathcal{N}$, is defined as:

$$\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X} : |C| = m} |\mathcal{H}_C|$$

Observations.

- For every $\mathcal{H}$ and every m, $\tau_{\mathcal{H}}(m) \leq 2^m$

The Growth Function

Let $\mathcal{H}$ be a hypothesis class, then the growth function of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathcal{N} \to \mathcal{N}$, is defined as:

$$\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X} : |C| = m} |\mathcal{H}_C|$$

Observations.

- For every $\mathcal{H}$ and every m, $\tau_{\mathcal{H}}(m) \leq 2^m$
- If $\mathcal{H}$ shatters a set of size m, then $\tau_{\mathcal{H}}(m) = 2^m$

The Growth Function

Let $\mathcal{H}$ be a hypothesis class, then the growth function of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathcal{N} \to \mathcal{N}$, is defined as:

$$\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X} : |C| = m} |\mathcal{H}_C|$$

Observations.

- For every $\mathcal{H}$ and every m, $\tau_{\mathcal{H}}(m) \leq 2^m$
- If $\mathcal{H}$ shatters a set of size m, then $\tau_{\mathcal{H}}(m) = 2^m$
- If $VCDim(\mathcal{H}) < m$, then $\tau_{\mathcal{H}}(m) < 2^m$

$\square$

The Growth Function

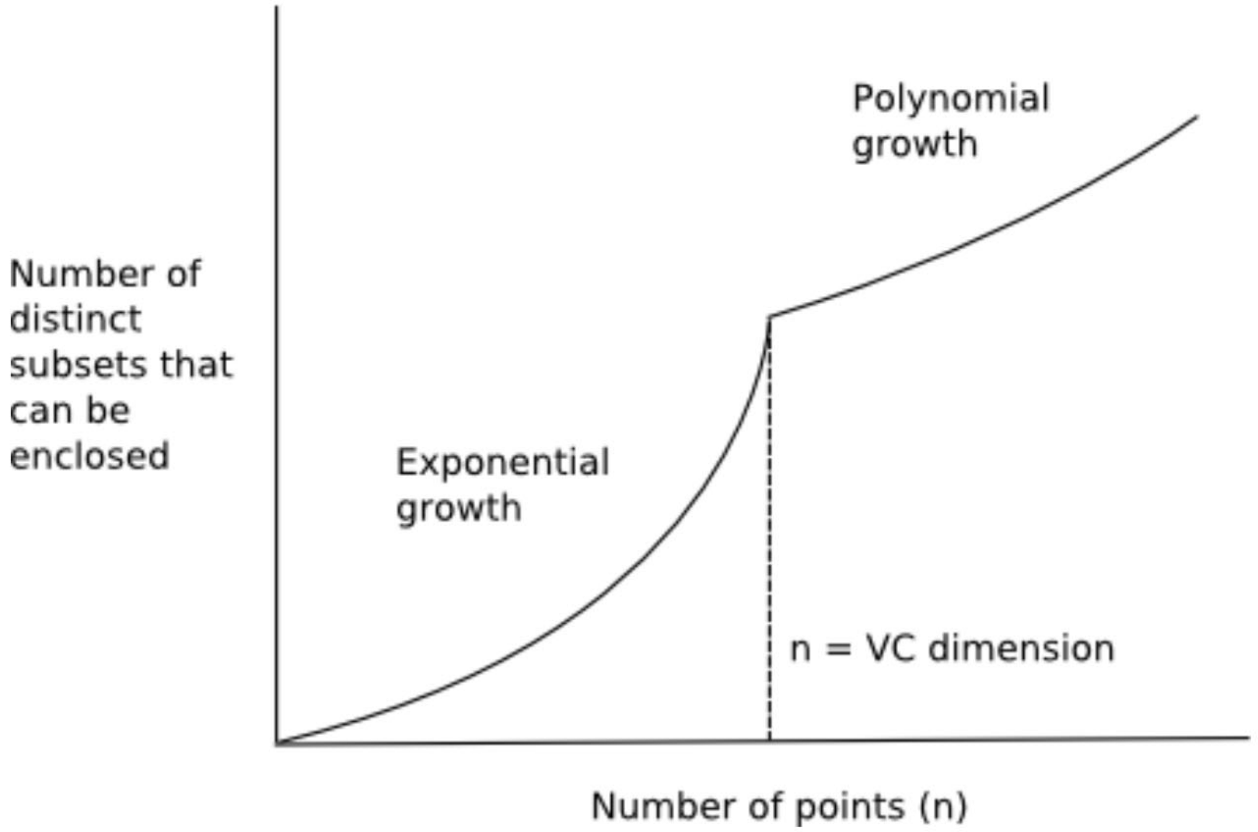
Let $\mathcal{H}$ be a hypothesis class, then the growth function of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathcal{N} \to \mathcal{N}$, is defined as:

$$\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X} : |C| = m} |\mathcal{H}_C|$$

Intuition

$$\begin{aligned}\tau_{\mathcal{H}}(m) &= \max_{|A| \le m} |h \cap A : h \in \mathcal{H}| \\ &= |\{g : A \to \{0,1\} : \\ &\quad \exists h \in \mathcal{H}, \forall x \in A, g(x) = h(x)\}|\end{aligned}$$

Exponential $2^m, 3^m, \cdots$
... $m^{\log m}$, etc.
Polynomial $m^2, m^3, \cdots$



Sauer's Lemma

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \leq (em/d)^d$

Sauer's Lemma

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \leq (em/d)^d$

- The number of partitions realizable by a linear half-space:
  - $VCDim(\mathscr{H}_{HS^2}) = 3$

Sauer's Lemma

Let $\mathcal{H}$ be a hypothesis class with $VCDim(\mathcal{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$

- The number of partitions realizable by a linear half-space:
  - $VCDim(\mathcal{H}_{HS^2}) = 3$
  - For 1000 points with binary labels, there are less than 1000^3 partitions by $\mathcal{H}_{HS^2}$

Sauer's Lemma

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \leq (em/d)^d$

- The number of partitions realizable by a linear half-space:
  - $VCDim(\mathscr{H}_{HS^2}) = 3$
  - For 1000 points with binary labels, there are less than 1000^3 partitions by $\mathscr{H}_{HS^2}$
  - There are 2^{1000} colouring of those 1000 points

Sauer's Lemma

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \leq (em/d)^d$

- The number of partitions realizable by a linear half-space:
  - $VCDim(\mathscr{H}_{HS^2}) = 3$
  - For 1000 points with binary labels, there are less than 1000^3 partitions by $\mathscr{H}_{HS^2}$
  - There are 2^{1000} colouring of those 1000 points
  - $1000^3 << 2^{1000}$

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \le d \le \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i} \le m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \le (em/d)^d$

Intuition.

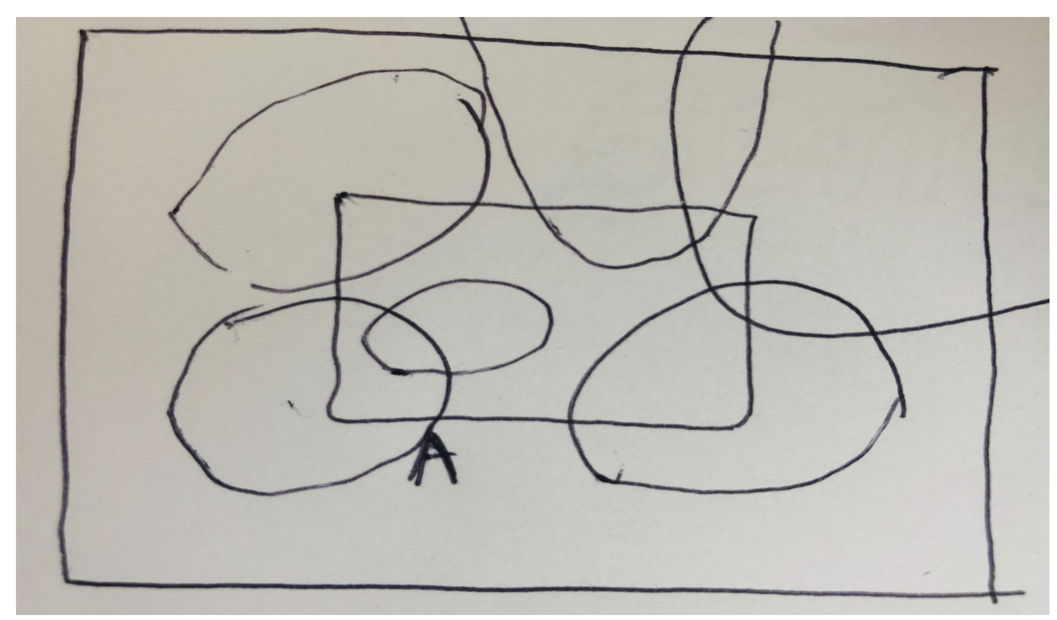
1. We prove a stronger claim: $\forall C = \{c_1, \cdots, c_m\}$, we have $\forall \mathscr{H}$, $|\mathscr{H}_C| \le |\{B \subseteq C : \mathscr{H} \text{ shatters } B\}|$.

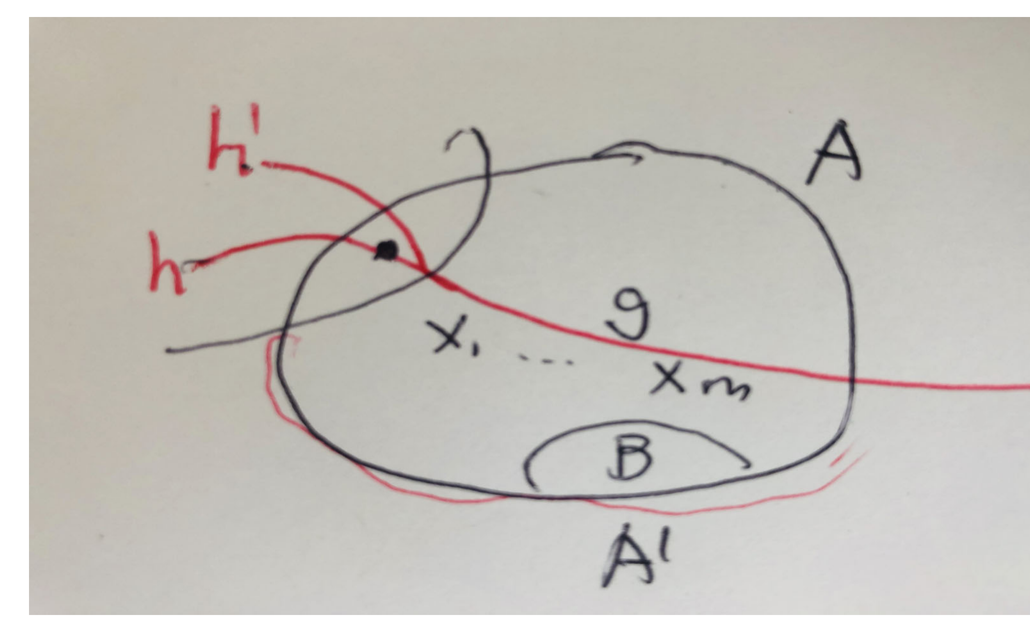
Sauer's Lemma

Let $\mathcal{H}$ be a hypothesis class with $VCDim(\mathcal{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$

Intuition.

1. We prove a stronger claim: $\forall C = \{c_1, \cdots, c_m\}$, we have $\forall \mathcal{H}$, $|\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|$.
2. If $VCDim(\mathcal{H}) \leq d$, then no set with size larger than d is shattered by $\mathcal{H}$: $|\{B \subseteq C : \mathcal{H} \text{ shatters } B\}| \leq \sum_{i=0}^{d} \binom{m}{i}$





□

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \leq (em/d)^d$

$\forall C = \{c_1, \cdots, c_m\}$, we have $\forall \mathscr{H}$, $|\mathscr{H}_C| \leq |\{B \subseteq C : \mathscr{H} \text{ shatters } B\}|$.

Inductive Argument.

1. The empty set ϕ is always shattered by $\mathscr{H}$.

Sauer's Lemma

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \leq (em/d)^d$

$\forall C = \{c_1, \cdots, c_m\}$, we have $\forall \mathscr{H}$, $|\mathscr{H}_C| \leq |\{B \subseteq C : \mathscr{H} \text{ shatters } B\}|$.

Inductive Argument.

1. The empty set ϕ is always shattered by $\mathscr{H}$.
2. Assume it holds for any size $m-1$, let us prove it for the size m. Denote $C' = \{c_2, \cdots, c_m\}$ and define:

$$Y_0 = \{(y_2, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \mathscr{H}_C \vee (1, y_2, \cdots, y_m) \in \mathscr{H}_C\}$$

and

$$Y_1 = \{(y_2, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \mathscr{H}_C \wedge (1, y_2, \cdots, y_m) \in \mathscr{H}_C\}$$

We have $|\mathscr{H}_C| = |Y_0| + |Y_1|$.

Let $\mathcal{H}$ be a hypothesis class with $VCDim(\mathcal{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$

$\forall C = \{c_1, \cdots, c_m\}$, we have $\forall \mathcal{H}$, $|\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \text{ shatters } B\}|$.

Inductive Argument.

1. The empty set ϕ is always shattered by $\mathcal{H}$.
2. Assume it holds for any size $m-1$, let us prove it for the size m. Denote $C' = \{c_2, \cdots, c_m\}$ and define:

   $$Y_0 = \{(y_2, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \mathcal{H}_C \vee (1, y_2, \cdots, y_m) \in \mathcal{H}_C\}$$

   and

   $$Y_1 = \{(y_2, \cdots, y_m) : (0, y_2, \cdots, y_m) \in \mathcal{H}_C \wedge (1, y_2, \cdots, y_m) \in \mathcal{H}_C\}$$

   We have $|\mathcal{H}_C| = |Y_0| + |Y_1|$.
3. Since $Y_0 = \mathcal{H}_{C'}$, using induction assumption, we have $|Y_0| = |\mathcal{H}_{C'}| \leq |\{B \subseteq C' : \mathcal{H} \text{ shatters } B\}| = |\{B \subseteq C : c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}|$

□

Sauer's Lemma

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \le d \le \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i} \le m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \le (em/d)^d$

$\forall C = \{c_1, \cdots, c_m\}$, we have $\forall \mathscr{H}$, $|\mathscr{H}_C| \le |\{B \subseteq C : \mathscr{H}$ shatters $B\}|$.

Inductive Argument.

4. Let $\mathscr{H}' \subseteq \mathscr{H}$,
   $\mathscr{H}' = \{h \in \mathscr{H} : \exists h' \in \mathscr{H}(1-h'(c_1), h'(c_2), \cdots, h'(c_m)) = (h(c_1), h(c_2), \cdots, h(c_m))\}$
   - $\mathscr{H}'$ contains pairs of hypotheses that agree on C' and differ on c_1.
   - If $\mathscr{H}'$ shatters a set $B \subseteq C'$, then it also shatters the set $B \cup c_1$ and vice versa

Sauer's Lemma

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \le d \le \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i} \le m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \le (em/d)^d$

$\forall C = \{c_1, \cdots, c_m\}$, we have $\forall \mathscr{H}$, $|\mathscr{H}_C| \le |\{B \subseteq C : \mathscr{H} \text{ shatters } B\}|$.

Inductive Argument.

4. Let $\mathscr{H}' \subseteq \mathscr{H}$,
   $\mathscr{H}' = \{h \in \mathscr{H} : \exists h' \in \mathscr{H} (1 - h'(c_1), h'(c_2), \cdots, h'(c_m)) = (h(c_1), h(c_2), \cdots, h(c_m))\}$
   - $\mathscr{H}'$ contains pairs of hypotheses that agree on C' and differ on c_1.
   - If $\mathscr{H}'$ shatters a set $B \subseteq C'$, then it also shatters the set $B \cup c_1$ and vice versa
5. Since $Y_1 = \mathscr{H}'_{C'}$, using induction assumption, we have

$$|Y_1| = |\mathscr{H}'_{C'}| \le |\{B \subseteq C' : \mathscr{H}' \text{ shatters } B \cup \{c_1\}\}| = |\{B \subseteq C : c_1 \in B \wedge \mathscr{H}' \text{ shatters } B\}|$$
$$\le |\{B \subseteq C : c_1 \in B \wedge \mathscr{H} \text{ shatters } B\}|$$

Sauer's Lemma

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \le d \le \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i} \le m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \le (em/d)^d$

$\forall C = \{c_1, \cdots, c_m\}$, we have $\forall \mathscr{H}$, $|\mathscr{H}_C| \le |\{B \subseteq C : \mathscr{H} \text{ shatters } B\}|$.

Inductive Argument.

4. Let $\mathscr{H}' \subseteq \mathscr{H}$,
   $\mathscr{H}' = \{h \in \mathscr{H} : \exists h' \in \mathscr{H}(1-h'(c_1), h'(c_2), \cdots, h'(c_m)) = (h(c_1), h(c_2), \cdots, h(c_m))\}$
   - $\mathscr{H}'$ contains pairs of hypotheses that agree on C' and differ on c_1.
   - If $\mathscr{H}'$ shatters a set $B \subseteq C'$, then it also shatters the set $B \cup c_1$ and vice versa
5. Since $Y_1 = \mathscr{H}'_{C'}$, using induction assumption, we have

$$|Y_1| = |\mathscr{H}'_{C'}| \le |\{B \subseteq C' : \mathscr{H}' \text{ shatters } B \cup \{c_1\}\}| = |\{B \subseteq C : c_1 \in B \wedge \mathscr{H}' \text{ shatters } B\}| \\ \le |\{B \subseteq C : c_1 \in B \wedge \mathscr{H} \text{ shatters } B\}|$$

6. Overall we have shown that:

$$|\mathscr{H}_C| = |Y_0| + |Y_1| \le |\{B \subseteq C : c_1 \notin B \wedge \mathscr{H} \text{ shatters } B\}| + |\{B \subseteq C : c_1 \in B \wedge \mathscr{H} \text{ shatters } B\}| \\ = |\{B \subseteq C : \mathscr{H} \text{ shatters } B\}|$$

□

Sauer's Lemma

Let $\mathcal{H}$ be a hypothesis class with $VCDim(\mathcal{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$

Consequences of Sauer's Lemma.

- The vast majority of partitions of a subset of m points from $\mathcal{R}^n$ can not be realized by a linear separator: $m^{n+1} << 2^m$

Sauer's Lemma

Let $\mathcal{H}$ be a hypothesis class with $VCDim(\mathcal{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$

Consequences of Sauer's Lemma.

- The vast majority of partitions of a subset of m points from $\mathcal{R}^n$ can not be realized by a linear separator: $m^{n+1} << 2^m$
- Upper bounding $m_{\mathcal{H}}^{PAC}(\epsilon, \delta)$

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \leq (em/d)^d$

Consequences of Sauer's Lemma.

- The vast majority of partitions of a subset of m points from $\mathscr{R}^n$ can not be realized by a linear separator: $m^{n+1} << 2^m$
- Upper bounding $m_{\mathscr{H}}^{PAC}(\epsilon, \delta)$
- Upper bounding $VCdim(\mathscr{H}_1 \cup \mathscr{H}_2)$ by $VCDim(\mathscr{H}_1)$ and $VCDim(\mathscr{H}_2)$.

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \leq (em/d)^d$

Consequences of Sauer's Lemma.

- The vast majority of partitions of a subset of m points from $\mathscr{R}^n$ can not be realized by a linear separator: $m^{n+1} << 2^m$
- Upper bounding $m_{\mathscr{H}}^{PAC}(\epsilon, \delta)$
- Upper bounding $VCdim(\mathscr{H}_1 \cup \mathscr{H}_2)$ by $VCDim(\mathscr{H}_1)$ and $VCDim(\mathscr{H}_2)$.
  - Assume $\mathscr{H}_1 \cup \mathscr{H}_2$ shatters a set S of size m, $|\{h \cap A : h \in \mathscr{H}_1 \vee h \in \mathscr{H}_2\}| \geq 2^m$

Sauer's Lemma

Let $\mathscr{H}$ be a hypothesis class with $VCDim(\mathscr{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathscr{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathscr{H}}(m) \leq (em/d)^d$

Consequences of Sauer's Lemma.

- The vast majority of partitions of a subset of m points from $\mathscr{R}^n$ can not be realized by a linear separator: $m^{n+1} << 2^m$
- Upper bounding $m_{\mathscr{H}}^{PAC}(\epsilon, \delta)$
- Upper bounding $VCdim(\mathscr{H}_1 \cup \mathscr{H}_2)$ by $VCDim(\mathscr{H}_1)$ and $VCDim(\mathscr{H}_2)$.
  - Assume $\mathscr{H}_1 \cup \mathscr{H}_2$ shatters a set S of size m, $|\{h \cap A : h \in \mathscr{H}_1 \vee h \in \mathscr{H}_2\}| \geq 2^m$
  - On the other hand,

$$
\begin{aligned}
&|\{h \cap A : h \in \mathscr{H}_1 \vee h \in \mathscr{H}_2\}| \\
&\quad \leq |\{h \cap A : h \in \mathscr{H}_1| + |\{h \cap A : h \in \mathscr{H}_2| \\
&\quad \leq m^{VCDim(\mathscr{H}_1)} + m^{VCDim(\mathscr{H}_2)}
\end{aligned}
$$

Sauer's Lemma

Let $\mathcal{H}$ be a hypothesis class with $VCDim(\mathcal{H}) \leq d \leq \infty$, then $\forall m$, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$. In particular, if $m > d+1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$

Consequences of Sauer's Lemma.

- The vast majority of partitions of a subset of m points from $\mathcal{R}^n$ can not be realized by a linear separator: $m^{n+1} << 2^m$
- Upper bounding $m_{\mathcal{H}}^{PAC}(\epsilon, \delta)$
- Upper bounding $VCdim(\mathcal{H}_1 \cup \mathcal{H}_2)$ by $VCDim(\mathcal{H}_1)$ and $VCDim(\mathcal{H}_2)$.
  - Assume $\mathcal{H}_1 \cup \mathcal{H}_2$ shatters a set S of size m, $|\{h \cap A : h \in \mathcal{H}_1 \vee h \in \mathcal{H}_2\}| \geq 2^m$
  - On the other hand,

$$
\begin{aligned}
&|\{h \cap A : h \in \mathcal{H}_1 \vee h \in \mathcal{H}_2\}| \\
&\quad \leq |\{h \cap A : h \in \mathcal{H}_1| + |\{h \cap A : h \in \mathcal{H}_2| \\
&\quad \leq m^{VCDim(\mathcal{H}_1)} + m^{VCDim(\mathcal{H}_2)}
\end{aligned}
$$

  - m can not be too big...

$\square$

Proof of the Upper Bounds

$$m_{\mathcal{H}}^{PAC} \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

Proof of the Upper Bounds

$m_{\mathcal{H}}^{PAC} \le C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$

Intuition.

- Recall our proof for the finite class $\mathcal{H}$
  - For a single hypothesis h, we have shown that the probability of the event $L_S(h) = 0$ given that $L_{(\mathcal{D},f)} > \epsilon$ is at most $e^{-\epsilon m}$
  - If we applied the union bound over all "bad" hypotheses, to obtain the bound on ERM failure: $|\mathcal{H}| e^{-\epsilon m}$

Proof of the Upper Bounds

$m_{\mathcal{H}}^{PAC} \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$

Intuition.

- Recall our proof for the finite class $\mathcal{H}$
  - For a single hypothesis h, we have shown that the probability of the event $L_S(h) = 0$ given that $L_{(\mathcal{D},f)} > \epsilon$ is at most $e^{-\epsilon m}$
  - If we applied the union bound over all "bad" hypotheses, to obtain the bound on ERM failure: $|\mathcal{H}|e^{-\epsilon m}$
- If $\mathcal{H}$ is infinite or large, the union bound yields a meaningless bound

□

Proof of the Upper Bounds

$m_{\mathcal{H}}^{PAC} \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$

Main Idea.

Two Samples Trick We evaluate the error of any hypothesis $h \in \mathcal{H}_B$ on a fresh sample T, a.k.a. the ghost sample.

- Since S and T are i.i.d., we can think of first sampling $2m$ samples, and then splitting them to S and T at random
- On this evaluation, it suffices to consider $\mathcal{H}_{S \cup T}$, the set of possible behaviors of $S \cup T$
- If we fix $S \cup T$, we can take a union bound only over $\mathcal{H}_{S \cup T}$: $P_{S,T \sim \mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_T(h)| > \epsilon/2] \leq (2m)^d 2e^{-2m(\frac{\epsilon}{2})^2}$

$m_{\mathcal{H}}^{PAC} \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$

Main Idea.

Two Samples Trick We evaluate the error of any hypothesis $h \in \mathcal{H}_B$ on a fresh sample T, a.k.a. the ghost sample.

- Since S and T are i.i.d., we can think of first sampling $2m$ samples, and then splitting them to S and T at random
- On this evaluation, it suffices to consider $\mathcal{H}_{S \cup T}$, the set of possible behaviors of $S \cup T$
- If we fix $S \cup T$, we can take a union bound only over $\mathcal{H}_{S \cup T}$: $P_{S,T \sim \mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_T(h)| > \epsilon/2] \leq (2m)^d 2e^{-2m(\frac{\epsilon}{2})^2}$

Symmetrization For any $\epsilon > 0$, and $m\epsilon^2 \geq 2$, we have:

$$P_{S \sim \mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq 2P_{S,T \sim \mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_T(h)| > \epsilon/2]$$

Proof of the Upper Bounds

$m_{\mathcal{H}}^{PAC} \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$

Main Idea.

Two Samples Trick We evaluate the error of any hypothesis $h \in \mathcal{H}_B$ on a fresh sample T, a.k.a. the ghost sample.

- Since S and T are i.i.d., we can think of first sampling $2m$ samples, and then splitting them to S and T at random
- On this evaluation, it suffices to consider $\mathcal{H}_{S \cup T}$, the set of possible behaviors of $S \cup T$
- If we fix $S \cup T$, we can take a union bound only over $\mathcal{H}_{S \cup T}$: $P_{S,T \sim \mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_T(h)| > \epsilon/2] \leq (2m)^d 2e^{-2m(\frac{\epsilon}{2})^2}$

Symmetrization For any $\epsilon > 0$, and $m\epsilon^2 \geq 2$, we have:

$$P_{S \sim \mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq 2 P_{S,T \sim \mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_T(h)| > \epsilon/2]$$

Combination When m is sufficiently large, the probability of bad hypothesis will $< \epsilon$:

$$P_{S \sim \mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] \leq 2(2m)^d 2e^{-2m(\frac{\epsilon}{2})^2}$$

Symmetrization

For any $\epsilon > 0$, and $m\epsilon^2 \geq 2$, we have:

$$P_{S\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_{\mathscr{D}}(h)| > \epsilon] = P_{S\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_{\mathscr{D}}(h)| \rceil) > \epsilon]$$
$$\leq 2P_{S,T\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_T(h)| > \epsilon/2] = 2P_{S,T\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_T(h)| \rceil) > \epsilon/2]$$

Symmetrization

For any $\epsilon > 0$, and $m\epsilon^2 \geq 2$, we have:

$$P_{S\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_{\mathscr{D}}(h)| > \epsilon] = P_{S\sim\mathscr{D}^m}[(\lceil|L_S(h) - L_{\mathscr{D}}(h)|\rceil) > \epsilon]$$
$$\leq 2P_{S,T\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_T(h)| > \epsilon/2] = 2P_{S,T\sim\mathscr{D}^m}[(\lceil|L_S(h) - L_T(h)|\rceil) > \epsilon/2]$$

Intuition.

1. Assume the left hand side is maximized at $h_S \in \mathscr{H}$. Here h_S is dependent on the sample S.

Symmetrization

For any $\epsilon > 0$, and $m\epsilon^2 \geq 2$, we have:

$$P_{S\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_{\mathscr{D}}(h)| > \epsilon] = P_{S\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_{\mathscr{D}}(h)| \rceil) > \epsilon]$$
$$\leq 2P_{S,T\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_T(h)| > \epsilon/2] = 2P_{S,T\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_T(h)| \rceil) > \epsilon/2]$$

Intuition.

1. Assume the left hand side is maximized at $h_S \in \mathscr{H}$. Here h_S is dependent on the sample S.
2. We have $(|L_{\mathscr{D}}(h_S) - L_S(h_S)| > \epsilon) \wedge (|L_{\mathscr{D}}(h_S) - L_T(h_S)| < \epsilon/2) \leq (|L_S(h_S) - L_T(h_S)| > \epsilon/2)$

Symmetrization

For any $\epsilon > 0$, and $m\epsilon^2 \geq 2$, we have:

$$P_{S\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_{\mathscr{D}}(h)| > \epsilon] = P_{S\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_{\mathscr{D}}(h)| \rceil) > \epsilon]$$
$$\leq 2P_{S,T\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_T(h)| > \epsilon/2] = 2P_{S,T\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_T(h)| \rceil) > \epsilon/2]$$

Intuition.

1. Assume the left hand side is maximized at $h_S \in \mathscr{H}$. Here h_S is dependent on the sample S.
2. We have $(|L_{\mathscr{D}}(h_S) - L_S(h_S)| > \epsilon) \wedge (|L_{\mathscr{D}}(h_S) - L_T(h_S)| < \epsilon/2) \leq (|L_S(h_S) - L_T(h_S)| > \epsilon/2)$
3. Take the expectation on the ghost sample:

$$(|L_{\mathscr{D}}(h_S) - L_S(h_S)| > \epsilon) \wedge P(|L_{\mathscr{D}}(h_S) - L_T(h_S)| < \epsilon/2) \leq P(|L_S(h_S) - L_T(h_S)| > \epsilon/2)$$

Symmetrization

For any $\epsilon > 0$, and $m\epsilon^2 \geq 2$, we have:

$$P_{S\sim\mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon] = P_{S\sim\mathcal{D}^m}[(\lceil |L_S(h) - L_{\mathcal{D}}(h)| \rceil) > \epsilon]$$
$$\leq 2P_{S,T\sim\mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_T(h)| > \epsilon/2] = 2P_{S,T\sim\mathcal{D}^m}[(\lceil |L_S(h) - L_T(h)| \rceil) > \epsilon/2]$$

Intuition.

1. Assume the left hand side is maximized at $h_S \in \mathcal{H}$. Here h_S is dependent on the sample S.
2. We have $(|L_{\mathcal{D}}(h_S) - L_S(h_S)| > \epsilon) \wedge (|L_{\mathcal{D}}(h_S) - L_T(h_S)| < \epsilon/2) \leq (|L_S(h_S) - L_T(h_S)| > \epsilon/2)$
3. Take the expectation on the ghost sample:

$$(|L_{\mathcal{D}}(h_S) - L_S(h_S)| > \epsilon) \wedge P(|L_{\mathcal{D}}(h_S) - L_T(h_S)| < \epsilon/2) \leq P(|L_S(h_S) - L_T(h_S)| > \epsilon/2)$$

4. As h_S is independent of the ghost sample, we can use Chebyshev's Inequality:

$$P(|L_{\mathcal{D}}(h_S) - L_T(h_S)| \geq \epsilon/2) \leq \frac{4var(h_S)}{m\epsilon^2} \leq \frac{1}{m\epsilon^2}, \quad \text{because } var(h_S) \leq 1/4$$
$$P(|L_{\mathcal{D}}(h_S) - L_T(h_S)| < \epsilon/2) \geq 1 - \frac{1}{m\epsilon^2} \geq 1/2, \quad \text{when } 1 - \frac{1}{m\epsilon^2} \geq 1/2$$

Symmetrization

For any $\epsilon > 0$, and $m\epsilon^2 \geq 2$, we have:

$$P_{S\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_{\mathscr{D}}(h)| > \epsilon] = P_{S\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_{\mathscr{D}}(h)| \rceil) > \epsilon]$$
$$\leq 2P_{S,T\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_T(h)| > \epsilon/2] = 2P_{S,T\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_T(h)| \rceil) > \epsilon/2]$$

Intuition.

5. Replace it and then we have: $\frac{1}{2}(|L_{\mathscr{D}}(h_S) - L_S(h_S)| > \epsilon) \leq P(|L_T(h_S) - L_S(h_S)| > \epsilon/2)$

Symmetrization

For any $\epsilon > 0$, and $m\epsilon^2 \geq 2$, we have:

$$P_{S\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_{\mathscr{D}}(h)| > \epsilon] = P_{S\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_{\mathscr{D}}(h)| \rceil) > \epsilon]$$
$$\leq 2P_{S,T\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_T(h)| > \epsilon/2] = 2P_{S,T\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_T(h)| \rceil) > \epsilon/2]$$

Intuition.

5. Replace it and then we have: $\frac{1}{2}(|L_{\mathscr{D}}(h_S) - L_S(h_S)| > \epsilon) \leq P(|L_T(h_S) - L_S(h_S)| > \epsilon/2)$
6. Take the expectation on S, we have:

$$\begin{aligned} P(\lceil L_{\mathscr{D}}(h_S) - L_S(h_S) \rceil) &= P(|L_{\mathscr{D}}(h_S) - L_S(h_S)| > \epsilon) \\ &\leq 2P(|L_T(h_S) - L_S(h_S)| > \epsilon/2) \\ &\leq 2P(\lceil L_T(h_S) - L_S(h_S) \rceil > \epsilon/2) \end{aligned}$$

$\square$

Symmetrization

For any $\epsilon > 0$, and $m\epsilon^2 \geq 2$, we have:

$$P_{S\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_{\mathscr{D}}(h)| > \epsilon] = P_{S\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_{\mathscr{D}}(h)| \rceil) > \epsilon]$$
$$\leq 2P_{S,T\sim\mathscr{D}^m}[\exists h \in \mathscr{H} : |L_S(h) - L_T(h)| > \epsilon/2] = 2P_{S,T\sim\mathscr{D}^m}[(\lceil |L_S(h) - L_T(h)| \rceil) > \epsilon/2]$$

Intuition.

5. Replace it and then we have: $\frac{1}{2}(|L_{\mathscr{D}}(h_S) - L_S(h_S)| > \epsilon) \leq P(|L_T(h_S) - L_S(h_S)| > \epsilon/2)$
6. Take the expectation on S, we have:

$$\begin{aligned} P(\lceil L_{\mathscr{D}}(h_S) - L_S(h_S) \rceil) &= P(|L_{\mathscr{D}}(h_S) - L_S(h_S)| > \epsilon) \\ &\leq 2P(|L_T(h_S) - L_S(h_S)| > \epsilon/2) \\ &\leq 2P(\lceil L_T(h_S) - L_S(h_S) \rceil > \epsilon/2) \end{aligned}$$

□

- With Symmetrization, hypothesis classes with finite VC dimension can be PAC learnable. Hence ⑥ → ① in the The Fundamental Theorem of PAC Learning.

The fundamental theorem of statistic learning

- A class $\mathscr{H}$ is PAC learnable if and only if it has finite VC dimension.
- The sample complexity is characterized by the VC dimension.
- The $ERM_{\mathscr{H}}$ is a generic (near) optimal learner

What is Learnable and *How to Learn*?

The fundamental theorem of statistic learning

- A class $\mathscr{H}$ is PAC learnable if and only if it has finite VC dimension.
- The sample complexity is characterized by the VC dimension.
- The $ERM_{\mathscr{H}}$ is a generic (near) optimal learner

(Left to right) J. Rissanen, V. Vapnik, A. Gammerman, A. Chervonenkis, C. Wallace and R. Solomonoff

Learning Beyond Classes with Finite VC Dimension

Nonuniform Learnability

A hypothesis class $\mathcal{H}$ is nonuniformly learnable (NUL) if there exist a learning algorithm A, and a function $m_{\mathcal{H}}^{NUL} : (0,1)^2 \times \mathcal{H} \to \mathcal{N}$ such that:

- For every $\epsilon, \delta \in (0,1)$, for every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$, then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathcal{D}^m$, it holds that:

  $$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ is nonuniformly learnable (NUL) if there exist a learning algorithm A, and a function $m_{\mathscr{H}}^{NUL} : (0,1)^2 \times \mathscr{H} \rightarrow \mathscr{N}$ such that:

- For every $\epsilon, \delta \in (0,1)$, for every $h \in \mathscr{H}$, if $m \geq m_{\mathscr{H}}^{NUL}(\epsilon, \delta, h)$, then for every distribution $\mathscr{D}$ over $\mathscr{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathscr{D}^m$, it holds that:

$$L_{\mathscr{D}}(A(S)) \leq L_{\mathscr{D}}(h) + \epsilon$$

A hypothesis class $\mathscr{H}$ is agnostically PAC learnable if there exist a learning algorithm A, and a function $m_{\mathscr{H}} : (0,1)^2 \rightarrow \mathscr{N}$ such that:

- For every $\epsilon, \delta \in (0,1)$, if $m \geq m_{\mathscr{H}}(\epsilon, \delta)$, then for every distribution $\mathscr{D}$ over $\mathscr{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathscr{D}^m$, it holds that:

$$L_{\mathscr{D}}(A(S)) \leq \min_{h' \in \mathscr{H}} L_{\mathscr{D}}(h') + \epsilon$$

Nonuniform Learnability

A hypothesis class $\mathcal{H}$ is nonuniformly learnable (NUL) if there exist a learning algorithm A, and a function $m_{\mathcal{H}}^{NUL} : (0,1)^2 \times \mathcal{H} \to \mathcal{N}$ such that:

- For every $\epsilon, \delta \in (0,1)$, for every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$, then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathcal{D}^m$, it holds that:

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

A hypothesis class $\mathcal{H}$ is agnostically PAC learnable if there exist a learning algorithm A, and a function $m_{\mathcal{H}} : (0,1)^2 \to \mathcal{N}$ such that:

- For every $\epsilon, \delta \in (0,1)$, if $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathcal{D}^m$, it holds that:

$$L_{\mathcal{D}}(A(S)) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

- Nonuniformly learnability (NUL) is a relaxation of agnostic PAC learnability

Nonuniform Learnability

A hypothesis class $\mathcal{H}$ is nonuniformly learnable (NUL) if there exist a learning algorithm A, and a function $m_{\mathcal{H}}^{NUL} : (0,1)^2 \times \mathcal{H} \to \mathcal{N}$ such that:

- For every $\epsilon, \delta \in (0,1)$, for every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$, then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathcal{D}^m$, it holds that:

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

A hypothesis class $\mathcal{H}$ is agnostically PAC learnable if there exist a learning algorithm A, and a function $m_{\mathcal{H}} : (0,1)^2 \to \mathcal{N}$ such that:

- For every $\epsilon, \delta \in (0,1)$, if $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathcal{D}^m$, it holds that:

$$L_{\mathcal{D}}(A(S)) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

- Nonuniformly learnability (NUL) is a relaxation of agnostic PAC learnability
- A class that is agnostic PAC learnable is also nonuniformly learnable

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n \in \mathcal{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n \in \mathscr{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Finite functions Let $\mathscr{X} = \mathscr{R}$, $\mathscr{Y} = \{0,1\}$, for each k, let $\mathscr{H}_k = \{h : h(x) = 0 \text{ for at most } k \text{ many } x\text{s}\}$

- $\mathscr{H}_1$ is the class of singletons

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n\in\mathcal{N}}^{\infty}\mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Finite functions Let $\mathscr{X} = \mathscr{R}$, $\mathscr{Y} = \{0,1\}$, for each k, let $\mathscr{H}_k = \{h : h(x) = 0 \text{ for at most } k \text{ many } x\text{s}\}$

- $\mathscr{H}_1$ is the class of singletons
- $VCDim(\mathscr{H}_k) = ?$

Nonuniform Learnability

A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathcal{H}$:

- let $\mathcal{H} = \cup_{n\in\mathcal{N}}^{\infty} \mathcal{H}_n$ be a hypothesis class
- each $\mathcal{H}_n$ having a finite VC dimension.

Finite functions Let $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \{0,1\}$, for each k, let $\mathcal{H}_k = \{h : h(x) = 0 \text{ for at most } k \text{ many } x\text{s}\}$

- $\mathcal{H}_1$ is the class of singletons
- $VCDim(\mathcal{H}_k) =?$
- $VCDim(\mathcal{H}) =?$

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n \in \mathcal{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Two preliminary questions:

- Is there a class that is nonuniformly learnable but not PAC learnable?

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n \in \mathcal{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Two preliminary questions:

- Is there a class that is nonuniformly learnable but not PAC learnable?
- Is there a class that is not nonuniformly learnable?

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n \in \mathcal{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Threshold Polynomial functions Let $\mathscr{H}$ be the class of all threshold polynomial functions: For each polynomial $P(x) = a_0 + a_1 x + \cdots + a_n x^n + \cdots$, $h_P(x) = sign(P(x))$. Let $\mathscr{H} = \cup_{n=0}^{\infty} \mathscr{H}_n$ be the class of all such polynomial predictors h_P, where $\mathscr{H}_n = \{h_P : P \text{ is a polynomial with degree} \leq n\}$

- Each $\mathscr{H}_n$ is PAC learnable and has finite VC dimension, so $\mathscr{H}$ is NUL

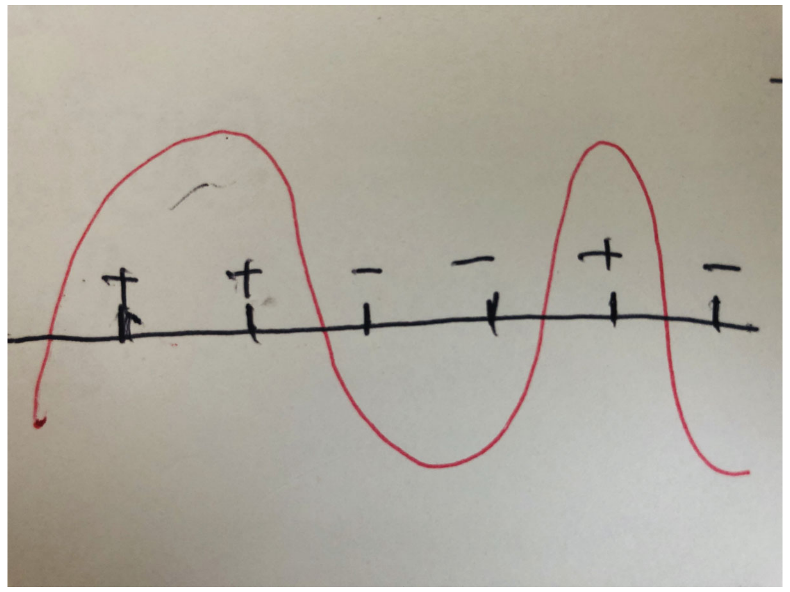
Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n\in\mathcal{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Threshold Polynomial functions Let $\mathscr{H}$ be the class of all threshold polynomial functions: For each polynomial $P(x) = a_0 + a_1 x + \cdots + a_n x^n + \cdots$, $h_P(x) = sign(P(x))$. Let $\mathscr{H} = \cup_{n=0}^{\infty} \mathscr{H}_n$ be the class of all such polynomial predictors h_P, where $\mathscr{H}_n = \{h_P : P \text{ is a polynomial with degree} \leq n\}$

- Each $\mathscr{H}_n$ is PAC learnable and has finite VC dimension, so $\mathscr{H}$ is NUL
- $\mathscr{H}$ is not PAC learnable, because it has infinite VC dimension



Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n\in\mathscr{N}}^{\infty}\mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Infinite functions For every infinite domain $\mathscr{X}$, the class of all binary function over $\mathscr{X}$ can not be covered by any union of $\cup_{n=0}^{\infty}\mathscr{H}_n$ where $\mathscr{H}_n$ is with finite VC dimension.

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n\in\mathcal{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Infinite functions For every infinite domain $\mathscr{X}$, the class of all binary function over $\mathscr{X}$ can not be covered by any union of $\cup_{n=0}^{\infty} \mathscr{H}_n$ where $\mathscr{H}_n$ is with finite VC dimension.

- This implies that there exist some classes that are not NUL

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n \in \mathcal{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Proof-1. NUL $\Rightarrow$ Countable Union

- Assume that $\mathscr{H}$ is NUL using A with sample complexity $m_{\mathscr{H}}^{NUL}$

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n \in \mathscr{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Proof-1. NUL $\Rightarrow$ Countable Union

- Assume that $\mathscr{H}$ is NUL using A with sample complexity $m_{\mathscr{H}}^{NUL}$
- For every $n \in \mathscr{N}$, let $\mathscr{H}_n = \{h \in \mathscr{H} : m_{\mathscr{H}}^{NUL}(1/8, 1/7, h) \leq n\}$

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n\in\mathscr{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Proof-1. NUL $\Rightarrow$ Countable Union

- Assume that $\mathscr{H}$ is NUL using A with sample complexity $m_{\mathscr{H}}^{NUL}$
- For every $n \in \mathscr{N}$, let $\mathscr{H}_n = \{h \in \mathscr{H} : m_{\mathscr{H}}^{NUL}(1/8, 1/7, h) \le n\}$
- Clearly, we have $\mathscr{H} = \cup_{n\in\mathscr{N}} \mathscr{H}_n$

A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathcal{H}$:

- let $\mathcal{H} = \cup_{n\in\mathcal{N}}^{\infty} \mathcal{H}_n$ be a hypothesis class
- each $\mathcal{H}_n$ having a finite VC dimension.

Proof-1. NUL $\Rightarrow$ Countable Union

- Assume that $\mathcal{H}$ is NUL using A with sample complexity $m_{\mathcal{H}}^{NUL}$
- For every $n \in \mathcal{N}$, let $\mathcal{H}_n = \{h \in \mathcal{H} : m_{\mathcal{H}}^{NUL}(1/8, 1/7, h) \leq n\}$
- Clearly, we have $\mathcal{H} = \cup_{n\in\mathcal{N}} \mathcal{H}_n$
- For every n, $VCDim(\mathcal{H}_n) \leq 2n$ is finite

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n\in\mathscr{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Proof-1. NUL $\Rightarrow$ Countable Union

- Assume that $\mathscr{H}$ is NUL using A with sample complexity $m_{\mathscr{H}}^{NUL}$
- For every $n \in \mathscr{N}$, let $\mathscr{H}_n = \{h \in \mathscr{H} : m_{\mathscr{H}}^{NUL}(1/8, 1/7, h) \leq n\}$
- Clearly, we have $\mathscr{H} = \cup_{n\in\mathscr{N}} \mathscr{H}_n$
- For every n, $VCDim(\mathscr{H}_n) \leq 2n$ is finite
  - Otherwise, pick a set $A \sim \mathscr{D}^{2n}$ such that $\mathscr{H}_n$ shatters A

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n\in\mathscr{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Proof-1. NUL ⇒ Countable Union

- Assume that $\mathscr{H}$ is NUL using A with sample complexity $m_{\mathscr{H}}^{NUL}$
- For every $n \in \mathscr{N}$, let $\mathscr{H}_n = \{h \in \mathscr{H} : m_{\mathscr{H}}^{NUL}(1/8, 1/7, h) \le n\}$
- Clearly, we have $\mathscr{H} = \cup_{n\in\mathscr{N}} \mathscr{H}_n$
- For every n, $VCDim(\mathscr{H}_n) \le 2n$ is finite
  - Otherwise, pick a set $A \sim \mathscr{D}^{2n}$ such that $\mathscr{H}_n$ shatters A
  - NFL theorem tells us that we need n size of training sample to learn the class of all functions over A with accuracy at least ϵ and confidence 6/7.

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n\in\mathcal{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Proof-1. NUL $\Rightarrow$ Countable Union

- Assume that $\mathscr{H}$ is NUL using A with sample complexity $m_{\mathscr{H}}^{NUL}$
- For every $n \in \mathcal{N}$, let $\mathscr{H}_n = \{h \in \mathscr{H} : m_{\mathscr{H}}^{NUL}(1/8, 1/7, h) \leq n\}$
- Clearly, we have $\mathscr{H} = \cup_{n\in\mathcal{N}} \mathscr{H}_n$
- For every n, $VCDim(\mathscr{H}_n) \leq 2n$ is finite
  - Otherwise, pick a set $A \sim \mathscr{D}^{2n}$ such that $\mathscr{H}_n$ shatters A
  - NFL theorem tells us that we need n size of training sample to learn the class of all functions over A with accuracy at least ϵ and confidence 6/7.
  - If A is shattered by $\mathscr{H}_n$, $\mathscr{H}_n$ contains all functions over A.

A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathcal{H}$:

- let $\mathcal{H} = \cup_{n\in\mathcal{N}}^{\infty}\mathcal{H}_n$ be a hypothesis class
- each $\mathcal{H}_n$ having a finite VC dimension.

Proof-1. NUL $\Rightarrow$ Countable Union

- Assume that $\mathcal{H}$ is NUL using A with sample complexity $m_{\mathcal{H}}^{NUL}$
- For every $n \in \mathcal{N}$, let $\mathcal{H}_n = \{h \in \mathcal{H} : m_{\mathcal{H}}^{NUL}(1/8, 1/7, h) \leq n\}$
- Clearly, we have $\mathcal{H} = \cup_{n\in\mathcal{N}}\mathcal{H}_n$
- For every n, $VCDim(\mathcal{H}_n) \leq 2n$ is finite
  - Otherwise, pick a set $A \sim \mathcal{D}^{2n}$ such that $\mathcal{H}_n$ shatters A
  - NFL theorem tells us that we need n size of training sample to learn the class of all functions over A with accuracy at least ϵ and confidence 6/7.
  - If A is shattered by $\mathcal{H}_n$, $\mathcal{H}_n$ contains all functions over A.
  - By the definition of $\mathcal{H}_n$, it can be learned by n sample size.

Nonuniform Learnability

A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathcal{H}$:

- let $\mathcal{H} = \cup_{n\in\mathcal{N}}^{\infty} \mathcal{H}_n$ be a hypothesis class
- each $\mathcal{H}_n$ having a finite VC dimension.

Proof-1. NUL ⇒ Countable Union

- Assume that $\mathcal{H}$ is NUL using A with sample complexity $m_{\mathcal{H}}^{NUL}$
- For every $n \in \mathcal{N}$, let $\mathcal{H}_n = \{h \in \mathcal{H} : m_{\mathcal{H}}^{NUL}(1/8, 1/7, h) \leq n\}$
- Clearly, we have $\mathcal{H} = \cup_{n\in\mathcal{N}} \mathcal{H}_n$
- For every n, $VCDim(\mathcal{H}_n) \leq 2n$ is finite
  - Otherwise, pick a set $A \sim \mathcal{D}^{2n}$ such that $\mathcal{H}_n$ shatters A
  - NFL theorem tells us that we need n size of training sample to learn the class of all functions over A with accuracy at least ϵ and confidence 6/7.
  - If A is shattered by $\mathcal{H}_n$, $\mathcal{H}_n$ contains all functions over A.
  - By the definition of $\mathcal{H}_n$, it can be learned by n sample size.
  - Contradiction!

□

Nonuniform Learnability

A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathcal{H}$:

- let $\mathcal{H} = \cup_{n\in\mathcal{N}}^{\infty} \mathcal{H}_n$ be a hypothesis class
- each $\mathcal{H}_n$ having a finite VC dimension.

Proof-2. NUL $\Leftarrow$ Countable Union

- Assume that $\mathcal{H} = \cup_{n\in\mathcal{N}} \mathcal{H}_n$, and $VCdim(\mathcal{H}_n) = d_n < \infty$.

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n \in \mathcal{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Proof-2. NUL $\Leftarrow$ Countable Union

- Assume that $\mathscr{H} = \cup_{n \in \mathcal{N}} \mathscr{H}_n$, and $VCdim(\mathscr{H}_n) = d_n < \infty$.
- Choose a weight function $\omega : \mathcal{N} \mapsto [0, 1]$ subject to $\sum_n \omega(n) \leq 1$

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n \in \mathscr{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Proof-2. NUL $\Leftarrow$ Countable Union

- Assume that $\mathscr{H} = \cup_{n \in \mathscr{N}} \mathscr{H}_n$, and $VCdim(\mathscr{H}_n) = d_n < \infty$.
- Choose a weight function $\omega : \mathscr{N} \mapsto [0,1]$ subject to $\sum_n \omega(n) \leq 1$
- Choose $\delta_n = \delta \times \omega(n)$, and $\epsilon_n(m, \delta_n) = \min\{\epsilon \in (0,1) : m^{UC}_{\mathscr{H}_n}(\epsilon, \delta_n) \leq m\}$

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n \in \mathscr{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Proof-2. NUL $\Leftarrow$ Countable Union

- Assume that $\mathscr{H} = \cup_{n \in \mathscr{N}} \mathscr{H}_n$, and $VCdim(\mathscr{H}_n) = d_n < \infty$.
- Choose a weight function $\omega : \mathscr{N} \mapsto [0,1]$ subject to $\sum_n \omega(n) \leq 1$
- Choose $\delta_n = \delta \times \omega(n)$, and $\epsilon_n(m, \delta_n) = \min\{\epsilon \in (0,1) : m_{\mathscr{H}_n}^{UC}(\epsilon, \delta_n) \leq m\}$
  - ϵ_n represents the best approximation I can have with a m size sample

A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathcal{H}$:

- let $\mathcal{H} = \cup_{n \in \mathcal{N}}^{\infty} \mathcal{H}_n$ be a hypothesis class
- each $\mathcal{H}_n$ having a finite VC dimension.

Proof-2. NUL $\Leftarrow$ Countable Union

- Assume that $\mathcal{H} = \cup_{n \in \mathcal{N}} \mathcal{H}_n$, and $VCdim(\mathcal{H}_n) = d_n < \infty$.
- Choose a weight function $\omega : \mathcal{N} \mapsto [0,1]$ subject to $\sum_n \omega(n) \leq 1$
- Choose $\delta_n = \delta \times \omega(n)$, and $\epsilon_n(m, \delta_n) = \min\{\epsilon \in (0,1) : m_{\mathcal{H}_n}^{UC}(\epsilon, \delta_n) \leq m\}$
  - ϵ_n represents the best approximation I can have with a m size sample
- From the uniform convergence and ϵ_n, it follows that for every m and δ_n, with at most probability δ_n over the choice of $S \sim \mathcal{D}^m$ we have

$$\forall h \in \mathcal{H}_n, |L_{\mathcal{D}}(h) - L_S(h)| \geq \epsilon_n(m, \delta_n)$$

Nonuniform Learnability

A hypothesis class $\mathscr{H}$ of binary classifiers is nonuniformly learnable (NUL) if and only if it is a countable union of agnostic PAC learnable hypothesis classes $\mathscr{H}$:

- let $\mathscr{H} = \cup_{n\in\mathscr{N}}^{\infty} \mathscr{H}_n$ be a hypothesis class
- each $\mathscr{H}_n$ having a finite VC dimension.

Proof-2. NUL $\Leftarrow$ Countable Union

- Assume that $\mathscr{H} = \cup_{n\in\mathscr{N}} \mathscr{H}_n$, and $VCdim(\mathscr{H}_n) = d_n < \infty$.
- Choose a weight function $\omega : \mathscr{N} \mapsto [0,1]$ subject to $\sum_n \omega(n) \leq 1$
- Choose $\delta_n = \delta \times \omega(n)$, and $\epsilon_n(m, \delta_n) = \min\{\epsilon \in (0,1) : m^{UC}_{\mathscr{H}_n}(\epsilon, \delta_n) \leq m\}$
  - ϵ_n represents the best approximation I can have with a m size sample
- From the uniform convergence and ϵ_n, it follows that for every m and δ_n, with at most probability δ_n over the choice of $S \sim \mathscr{D}^m$ we have

$$\forall h \in \mathscr{H}_n, |L_{\mathscr{D}}(h) - L_S(h)| \geq \epsilon_n(m, \delta_n)$$

- Apply the union bound over n, we have at most $\sum_n \delta_n = \delta$,

$$\forall h \in \mathscr{H}, |L_{\mathscr{D}}(h) - L_S(h)| \geq \min_{n: h\in\mathscr{H}_n} \epsilon_n(m, \delta_n)$$

NUL versus A-PAC

Table 1: Comparison of NUL and Agnostic PAC (I)

Learnability	Conditions
Agnostic PAC	$\forall h$, $\forall \mathscr{D}$, $\forall \delta$ and $\forall \epsilon$, if $m \geq \ldots$, we have $$P_{S\sim\mathscr{D}^m}[\exists h : \lvert L_{\mathscr{D}}(h) - L_S(h)\rvert > \epsilon] < \delta$$
NUL	$\mathscr{H} = \cup_{n\in\mathscr{N}} \mathscr{H}_n$, each $\mathscr{H}_n$ with $m_n^{UC}(\epsilon,\delta)$, $\omega : \mathscr{N} \mapsto [0,1]$ and $\sum \omega(n) \leq 1$. $\forall \mathscr{D}$, $\forall \delta$ and $\forall \epsilon$, $\forall h$, we have $$P_{S\sim\mathscr{D}^m}[\exists h : \lvert L_{\mathscr{D}}(h) - L_S(h)\rvert > \epsilon_{n(h)}(m,\omega(n(h))\delta)] < \delta$$ This implies that $$L_{\mathscr{D}}(h) \leq L_S(h) + \epsilon_{n(h)}(m,\omega(n(h))\delta)$$

Convergence bounds on S: $\lvert L_S(h) - L_{\mathscr{D}}(h)\rvert$

NUL versus A-PAC

Table 2: Comparison of NUL and Agnostic PAC (II)

Learnability	Learner
Agnostic PAC	ERM learner: $\forall h$, $\forall \mathscr{D}$, $\forall \delta$ and $\forall \epsilon$, if $m \geq \ldots$, we have $$L_{\mathscr{D}}(A(S)) \leq L_S(h) + \epsilon$$ for $m \geq m_{\mathscr{H}}^{UC}(\epsilon/2, \delta)$ with probability at least $1-\delta$.
NUL	SRM learner: let $A(S)$ be any hypothesis $h \in \mathscr{H}$ that minimizes $$L_{\mathscr{D}}(A(S)) \leq L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\delta)$$

Learner bounds its error on S: $|L_{\mathscr{D}}(A(S)) - L_{\mathscr{D}}(h)|$

Structural Risk Minimization

A Structural Risk Minimization (SRM) algorithm is given some training sample S and a confidence parameter δ, and picks h that minimizes

$$L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\delta)$$

Structural Risk Minimization

A Structural Risk Minimization (SRM) algorithm is given some training sample S and a confidence parameter δ, and picks h that minimizes

$$L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\delta)$$

Prior knowledge. In PAC learning, learner expresses prior knowledge by specifying the hypothesis class $\mathscr{H}$. But there are other ways to express prior knowledge:

- a union of classes of hypothesis $\mathscr{H} = \cup_{n\in\mathscr{N}} \mathscr{H}_n$

Structural Risk Minimization

A Structural Risk Minimization (SRM) algorithm is given some training sample S and a confidence parameter δ, and picks h that minimizes

$$L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\delta)$$

Prior knowledge. In PAC learning, learner expresses prior knowledge by specifying the hypothesis class $\mathscr{H}$. But there are other ways to express prior knowledge:

- a union of classes of hypothesis $\mathscr{H} = \cup_{n\in\mathscr{N}} \mathscr{H}_n$
- and prior weight $\omega(n)$ for each $\mathscr{H}_n$, with $n(h) = \min\{n : h \in \mathscr{H}_n\}$

$\square$

Structural Risk Minimization

Given $\mathscr{H} = \cup_{n \in \mathscr{N}} \mathscr{H}_n$, where each $\mathscr{H}_n$ satisfies the uniform convergence property with a sample complexity function $m^{UC}_{\mathscr{H}_n}$. Given any weight function $\omega : \mathscr{N} \mapsto [0,1]$ and $\sum \omega(n) \leq 1$. Any SRM algorithm is a successful Non-Uniform Learner with sample

$$m^{NUL}_{\mathscr{H}}(\epsilon, \delta, h) \leq m^{UC}_{\mathscr{H}_{n(h)}}(\epsilon/2, \omega(n(h))\delta)$$

Given $\mathscr{H} = \cup_{n\in\mathscr{N}} \mathscr{H}_n$, where each $\mathscr{H}_n$ satisfies the uniform convergence property with a sample complexity function $m^{UC}_{\mathscr{H}_n}$. Given any weight function $\omega : \mathscr{N} \mapsto [0,1]$ and $\sum \omega(n) \le 1$. Any SRM algorithm is a successful Non-Uniform Learner with sample

$$m^{NUL}_{\mathscr{H}}(\epsilon,\delta,h) \le m^{UC}_{\mathscr{H}_{n(h)}}(\epsilon/2, \omega(n(h))\delta)$$

Prior knowledge.

- By the SRM algorithm, we have

$$L_{\mathscr{D}}(A(S)) \le L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\delta)$$

Structural Risk Minimization

Given $\mathcal{H} = \cup_{n\in\mathcal{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ satisfies the uniform convergence property with a sample complexity function $m^{UC}_{\mathcal{H}_n}$. Given any weight function $\omega : \mathcal{N} \mapsto [0,1]$ and $\sum \omega(n) \le 1$. Any SRM algorithm is a successful Non-Uniform Learner with sample

$$m^{NUL}_{\mathcal{H}}(\epsilon,\delta,h) \le m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2,\omega(n(h))\delta)$$

Prior knowledge.

- By the SRM algorithm, we have

$$L_{\mathcal{D}}(A(S)) \le L_S(h) + \epsilon_{n(h)}(m,\omega(n(h))\delta)$$

- If $m \ge m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2,\omega(n(h))\delta)$, then clearly we have $\epsilon_{n(h)}(m,\omega(n(h))\delta) \le \epsilon/2$.

Structural Risk Minimization

Given $\mathcal{H} = \cup_{n\in\mathcal{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ satisfies the uniform convergence property with a sample complexity function $m^{UC}_{\mathcal{H}_n}$. Given any weight function $\omega : \mathcal{N} \mapsto [0,1]$ and $\sum \omega(n) \le 1$. Any SRM algorithm is a successful Non-Uniform Learner with sample

$$m^{NUL}_{\mathcal{H}}(\epsilon,\delta,h) \le m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2, \omega(n(h))\delta)$$

Prior knowledge.

- By the SRM algorithm, we have

$$L_{\mathcal{D}}(A(S)) \le L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\delta)$$

- If $m \ge m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2, \omega(n(h))\delta)$, then clearly we have $\epsilon_{n(h)}(m, \omega(n(h))\delta) \le \epsilon/2$.
- From UC property of each $\mathcal{H}_n$, we have that with probability of at most δ_n,

$$L_S(h) > L_{\mathcal{D}}(h) + \epsilon/2$$

Structural Risk Minimization

Given $\mathcal{H} = \cup_{n\in\mathcal{N}}\mathcal{H}_n$, where each $\mathcal{H}_n$ satisfies the uniform convergence property with a sample complexity function $m^{UC}_{\mathcal{H}_n}$. Given any weight function $\omega : \mathcal{N} \mapsto [0,1]$ and $\sum\omega(n) \leq 1$. Any SRM algorithm is a successful Non-Uniform Learner with sample

$$m^{NUL}_{\mathcal{H}}(\epsilon,\delta,h) \leq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2,\omega(n(h))\delta)$$

Prior knowledge.

- By the SRM algorithm, we have

$$L_{\mathcal{D}}(A(S)) \leq L_S(h) + \epsilon_{n(h)}(m,\omega(n(h))\delta)$$

- If $m \geq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2,\omega(n(h))\delta)$, then clearly we have $\epsilon_{n(h)}(m,\omega(n(h))\delta) \leq \epsilon/2$.
- From UC property of each $\mathcal{H}_n$, we have that with probability of at most δ_n,

$$L_S(h) > L_{\mathcal{D}}(h) + \epsilon/2$$

- Combining together we obtain that at most δ we have

$$L_{\mathcal{D}}(A(S)) > L_{\mathcal{D}}(h) + \epsilon$$

Minimum Description Length

Bias to Shorter Description:

- Let $\mathscr{H}$ be a countable hypothesis class
- Let $\omega : \mathscr{H} \mapsto \mathscr{R}$ be such that $\sum_{h \in \mathscr{H}} \omega(h) \leq 1$
  - The function ω reflects prior knowledge on how important h is

Bias to Shorter Description:

- Let $\mathscr{H}$ be a countable hypothesis class
- Let $\omega : \mathscr{H} \mapsto \mathscr{R}$ be such that $\sum_{h \in \mathscr{H}} \omega(h) \leq 1$
  - The function ω reflects prior knowledge on how important h is

Description Length.

- Suppose that each $h \in \mathscr{H}$ is described by some word $d(h) \in \{0,1\}^*$, such as a binary representation.

Minimum Description Length

Bias to Shorter Description:

- Let $\mathcal{H}$ be a countable hypothesis class
- Let $\omega : \mathcal{H} \mapsto \mathcal{R}$ be such that $\sum_{h \in \mathcal{H}} \omega(h) \leq 1$
  - The function ω reflects prior knowledge on how important h is

Description Length.

- Suppose that each $h \in \mathcal{H}$ is described by some word $d(h) \in \{0,1\}^*$, such as a binary representation.
- Suppose the binary representation is prefix-free, namely, for every $h \neq h'$, $d(h)$ is not a prefix of $d(h')$.

Minimum Description Length

Bias to Shorter Description:

- Let $\mathscr{H}$ be a countable hypothesis class
- Let $\omega : \mathscr{H} \mapsto \mathscr{R}$ be such that $\sum_{h \in \mathscr{H}} \omega(h) \leq 1$
  - The function ω reflects prior knowledge on how important h is

Description Length.

- Suppose that each $h \in \mathscr{H}$ is described by some word $d(h) \in \{0,1\}^*$, such as a binary representation.
- Suppose the binary representation is prefix-free, namely, for every $h \neq h'$, $d(h)$ is not a prefix of $d(h')$.
- Let $|h|$ be the length of $d(h)$, then set $\omega(h) = \frac{1}{2^{|h|}}$.

Minimum Description Length

Bias to Shorter Description:

- Let $\mathscr{H}$ be a countable hypothesis class
- Let $\omega : \mathscr{H} \mapsto \mathscr{R}$ be such that $\sum_{h\in\mathscr{H}} \omega(h) \le 1$
  - The function ω reflects prior knowledge on how important h is

Description Length.

- Suppose that each $h \in \mathscr{H}$ is described by some word $d(h) \in \{0,1\}^*$, such as a binary representation.
- Suppose the binary representation is prefix-free, namely, for every $h \neq h'$, $d(h)$ is not a prefix of $d(h')$.
- Let $|h|$ be the length of $d(h)$, then set $\omega(h) = \frac{1}{2^{|h|}}$.
- From Kraft's inequality, we have $\sum_h \omega(h) \le 1$
  - define probability over words in $d(H)$ as follows: repeatedly toss an unbiased coin, until the sequence of outcomes is a member of $d(H)$, and then stop.
  - Since $d(H)$ is prefix-free, this is a valid probability over $d(H)$, and the probability to get $d(h)$ is $\omega(h)$.

□

Minimum Description Length

Let $\omega : \mathcal{H} \mapsto \mathcal{R}$ be such that $\sum_{h \in \mathcal{H}} \omega(h) \leq 1$. Then, with probability of at least $1 - \delta$ over $S \sim \mathcal{D}^m$, we have $\forall h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{-\log(\omega(h)) + \log(2/\delta)}{2m}}$$

Let $\omega : \mathcal{H} \mapsto \mathcal{R}$ be such that $\sum_{h\in\mathcal{H}} \omega(h) \leq 1$. Then, with probability of at least $1-\delta$ over $S \sim \mathcal{D}^m$, we have $\forall h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{-\log(\omega(h)) + \log(2/\delta)}{2m}}$$

Proof.

- For every h, define $\delta_h = \omega(h) \cdot \delta$.

Minimum Description Length

Let $\omega : \mathscr{H} \mapsto \mathscr{R}$ be such that $\sum_{h\in\mathscr{H}} \omega(h) \le 1$. Then, with probability of at least $1-\delta$ over $S \sim \mathscr{D}^m$, we have $\forall h \in \mathscr{H}$,

$$L_{\mathscr{D}}(h) \le L_S(h) + \sqrt{\frac{-\log(\omega(h)) + \log(2/\delta)}{2m}}$$

Proof.

- For every h, define $\delta_h = \omega(h)\cdot\delta$.
- By Hoeffding's bound, for every h, $\mathscr{D}^m(\{S : L_{\mathscr{D}}(h) > L_S(h) + \sqrt{\frac{\log(2/\delta_h)}{2m}}\}) \le \delta_h$

Minimum Description Length

Let $\omega : \mathcal{H} \mapsto \mathcal{R}$ be such that $\sum_{h\in\mathcal{H}} \omega(h) \leq 1$. Then, with probability of at least $1-\delta$ over $S \sim \mathcal{D}^m$, we have $\forall h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{-\log(\omega(h)) + \log(2/\delta)}{2m}}$$

Proof.

- For every h, define $\delta_h = \omega(h)\cdot\delta$.
- By Hoeffding's bound, for every h, $\mathcal{D}^m(\{S : L_{\mathcal{D}}(h) > L_S(h) + \sqrt{\frac{\log(2/\delta_h)}{2m}}\}) \leq \delta_h$
- Applying the union bound,

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, L_{\mathcal{D}}(h) > L_S(h) + \sqrt{\frac{\log(2/\delta_h)}{2m}}\})$$

$$= \mathcal{D}^m(\cup_{h\in\mathcal{H}}\{S : L_{\mathcal{D}}(h) > L_S(h) + \sqrt{\frac{\log(2/\delta_h)}{2m}}\})$$

$$\leq \sum_{h\in\mathcal{H}} \delta_h \leq \delta$$

Let $\omega : \mathcal{H} \mapsto \mathcal{R}$ be such that $\sum_{h \in \mathcal{H}} \omega(h) \le 1$. Then, with probability of at least $1 - \delta$ over $S \sim \mathcal{D}^m$, we have $\forall h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h) \le L_S(h) + \sqrt{\frac{-\log(\omega(h)) + \log(2/\delta)}{2m}}$$

Bounds Minimization.

Learning Objective to minimize $L_{\mathcal{D}}(h)$ over $h \in \mathcal{H}$

Minimum Description Length

Let $\omega : \mathcal{H} \mapsto \mathcal{R}$ be such that $\sum_{h\in\mathcal{H}} \omega(h) \leq 1$. Then, with probability of at least $1-\delta$ over $S \sim \mathcal{D}^m$, we have $\forall h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{-\log(\omega(h)) + \log(2/\delta)}{2m}}$$

Bounds Minimization.

Learning Objective to minimize $L_{\mathcal{D}}(h)$ over $h \in \mathcal{H}$

VC Bound $\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{-VCdim(\mathcal{H}) + \log(2/\delta)}{2m}}$

Minimum Description Length

Let $\omega : \mathcal{H} \mapsto \mathcal{R}$ be such that $\sum_{h\in\mathcal{H}} \omega(h) \le 1$. Then, with probability of at least $1-\delta$ over $S \sim \mathcal{D}^m$, we have $\forall h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h) \le L_S(h) + \sqrt{\frac{-\log(\omega(h)) + \log(2/\delta)}{2m}}$$

Bounds Minimization.

Learning Objective to minimize $L_{\mathcal{D}}(h)$ over $h \in \mathcal{H}$

VC Bound $\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) \le L_S(h) + \sqrt{\frac{-VCdim(\mathcal{H}) + \log(2/\delta)}{2m}}$

Algorithms

ERM minimizes the VC bound
MDL minimizes the MDL bound

□

Consistency Learnability

Let $\mathcal{Z}$ be a domain set, let $\mathcal{P}$ be a set of probability over $\mathcal{Z}$, and let $\mathcal{H}$ be a hypothesis class. A learning rule A is consistent with respect to $\mathcal{H}$ and $\mathcal{P}$ if there exists a function $m_{\mathcal{H}}^{CON} : (0,1)^2 \times \mathcal{H} \times \mathcal{P} \to \mathcal{N}$ such that, for every $\epsilon, \delta \in (0,1)$, for every $h \in \mathcal{H}$, and every $\mathcal{D} \in \mathcal{P}$, if $m \geq m_{\mathcal{H}}^{CON}(\epsilon, \delta, h, \mathcal{D})$, then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathcal{D}^m$, it holds that:

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

If $\mathcal{P}$ is the set of all distributions, we say that A is universally consistent with respect to $\mathcal{H}$.

Consistency Learnability

Let $\mathcal{Z}$ be a domain set, let $\mathcal{P}$ be a set of probability over $\mathcal{Z}$, and let $\mathcal{H}$ be a hypothesis class. A learning rule A is consistent with respect to $\mathcal{H}$ and $\mathcal{P}$ if there exists a function $m_{\mathcal{H}}^{CON} : (0,1)^2 \times \mathcal{H} \times \mathcal{P} \to \mathcal{N}$ such that, for every $\epsilon, \delta \in (0,1)$, for every $h \in \mathcal{H}$, and every $\mathcal{D} \in \mathcal{P}$, if $m \geq m_{\mathcal{H}}^{CON}(\epsilon, \delta, h, \mathcal{D})$, then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathcal{D}^m$, it holds that:

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

If $\mathcal{P}$ is the set of all distributions, we say that A is universally consistent with respect to $\mathcal{H}$.

- *Consistency* is a relaxation of *Nonuniformly Learnability* (NUL) learnability.

Consistency Learnability

Let $\mathcal{Z}$ be a domain set, let $\mathcal{P}$ be a set of probability over $\mathcal{Z}$, and let $\mathcal{H}$ be a hypothesis class. A learning rule A is consistent with respect to $\mathcal{H}$ and $\mathcal{P}$ if there exists a function $m_{\mathcal{H}}^{CON} : (0,1)^2 \times \mathcal{H} \times \mathcal{P} \to \mathcal{N}$ such that, for every $\epsilon, \delta \in (0,1)$, for every $h \in \mathcal{H}$, and every $\mathcal{D} \in \mathcal{P}$, if $m \geq m_{\mathcal{H}}^{CON}(\epsilon, \delta, h, \mathcal{D})$, then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathcal{D}^m$, it holds that:

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

If $\mathcal{P}$ is the set of all distributions, we say that A is universally consistent with respect to $\mathcal{H}$.

- *Consistency* is a relaxation of *Nonuniformly Learnability* (NUL) learnability.
- Both PAC learning and NUL learning provide an upper bound on the true risk of the learned hypothesis based on its empirical risk, but *consistency guarantees* do not provide such a bound.

Consistency Learnability

Let $\mathscr{Z}$ be a domain set, let $\mathscr{P}$ be a set of probability over $\mathscr{Z}$, and let $\mathscr{H}$ be a hypothesis class. A learning rule A is consistent with respect to $\mathscr{H}$ and $\mathscr{P}$ if there exists a function $m_{\mathscr{H}}^{CON} : (0,1)^2 \times \mathscr{H} \times \mathscr{P} \rightarrow \mathscr{N}$ such that, for every $\epsilon, \delta \in (0,1)$, for every $h \in \mathscr{H}$, and every $\mathscr{D} \in \mathscr{P}$, if $m \geq m_{\mathscr{H}}^{CON}(\epsilon, \delta, h, \mathscr{D})$, then for every distribution $\mathscr{D}$ over $\mathscr{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathscr{D}^m$, it holds that:

$$L_{\mathscr{D}}(A(S)) \leq L_{\mathscr{D}}(h) + \epsilon$$

If $\mathscr{P}$ is the set of all distributions, we say that A is universally consistent with respect to $\mathscr{H}$.

- *Consistency* is a relaxation of *Nonuniformly Learnability* (NUL) learnability.
- Both PAC learning and NUL learning provide an upper bound on the true risk of the learned hypothesis based on its empirical risk, but *consistency guarantees* do not provide such a bound.
  - We can estimate the risk of the output predictor using a validation set
- How many examples are required to be as good as the best hypothesis in $\mathscr{H}$?

Consistency Learnability

Let $\mathcal{Z}$ be a domain set, let $\mathcal{P}$ be a set of probability over $\mathcal{Z}$, and let $\mathcal{H}$ be a hypothesis class. A learning rule A is consistent with respect to $\mathcal{H}$ and $\mathcal{P}$ if there exists a function $m_{\mathcal{H}}^{CON} : (0,1)^2 \times \mathcal{H} \times \mathcal{P} \to \mathcal{N}$ such that, for every $\epsilon, \delta \in (0,1)$, for every $h \in \mathcal{H}$, and every $\mathcal{D} \in \mathcal{P}$, if $m \geq m_{\mathcal{H}}^{CON}(\epsilon, \delta, h, \mathcal{D})$, then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathcal{D}^m$, it holds that:

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

If $\mathcal{P}$ is the set of all distributions, we say that A is universally consistent with respect to $\mathcal{H}$.

- *Consistency* is a relaxation of *Nonuniformly Learnability* (NUL) learnability.
- Both PAC learning and NUL learning provide an upper bound on the true risk of the learned hypothesis based on its empirical risk, but *consistency guarantees* do not provide such a bound.
  - We can estimate the risk of the output predictor using a validation set
- How many examples are required to be as good as the best hypothesis in $\mathcal{H}$?
  - PAC gives a crisp answer; NUL and consistency provide no answer: it depends on the best hypothesis in $\mathcal{H}$ or the underlying distribution in NUL or CON.

Consistency Learnability

Let $\mathcal{Z}$ be a domain set, let $\mathcal{P}$ be a set of probability over $\mathcal{Z}$, and let $\mathcal{H}$ be a hypothesis class. A learning rule A is consistent with respect to $\mathcal{H}$ and $\mathcal{P}$ if there exists a function $m_{\mathcal{H}}^{CON} : (0,1)^2 \times \mathcal{H} \times \mathcal{P} \to \mathcal{N}$ such that, for every $\epsilon, \delta \in (0,1)$, for every $h \in \mathcal{H}$, and every $\mathcal{D} \in \mathcal{P}$, if $m \geq m_{\mathcal{H}}^{CON}(\epsilon, \delta, h, \mathcal{D})$, then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability of at least $(1-\delta)$ over the choice of $S \sim \mathcal{D}^m$, it holds that:

$$L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon$$

If $\mathcal{P}$ is the set of all distributions, we say that A is universally consistent with respect to $\mathcal{H}$.

- *Consistency* is a relaxation of *Nonuniformly Learnability* (NUL) learnability.
- Both PAC learning and NUL learning provide an upper bound on the true risk of the learned hypothesis based on its empirical risk, but *consistency guarantees* do not provide such a bound.
  - We can estimate the risk of the output predictor using a validation set
- How many examples are required to be as good as the best hypothesis in $\mathcal{H}$?
  - PAC gives a crisp answer; NUL and consistency provide no answer: it depends on the best hypothesis in $\mathcal{H}$ or the underlying distribution in NUL or CON.
  - What to do if the returned hypothesis with a large risk?

Consistency Learnability

A `Memorize` algorithm is universally consistent for every countable domain $\mathscr{X}$ and a finite label set $\mathscr{Y}$, with respect to the zero-one loss.

Intuition. $\square$

Consistency Learnability

A `Memorize` algorithm is universally consistent for every countable domain $\mathscr{X}$ and a finite label set $\mathscr{Y}$, with respect to the zero-one loss.

Intuition.

- Let $\{x_i : i \in \mathcal{N}\}$ be an enumeration of the elements of $\mathscr{X}$ so that for all $i \leq j$, we have $\mathscr{D}(\{x_i\}) \leq \mathscr{D}(\{x_j\})$. Also $\lim_{n\to\infty} \sum_{i\geq n} \mathscr{D}(\{x_i\}) = 0$

Consistency Learnability

A `Memorize` algorithm is universally consistent for every countable domain $\mathscr{X}$ and a finite label set $\mathscr{Y}$, with respect to the zero-one loss.

Intuition.

- Let $\{x_i : i \in \mathscr{N}\}$ be an enumeration of the elements of $\mathscr{X}$ so that for all $i \leq j$, we have $\mathscr{D}(\{x_i\}) \leq \mathscr{D}(\{x_j\})$. Also $\lim_{n\to\infty} \sum_{i \geq n} \mathscr{D}(\{x_i\}) = 0$
- Given any $\epsilon > 0$, there exists $\epsilon_D > 0$ such that $\mathscr{D}(\{x \in \mathscr{X} : \mathscr{D}(\{x\}) < \epsilon_D\}) < \epsilon$

Consistency Learnability

A `Memorize` algorithm is universally consistent for every countable domain $\mathscr{X}$ and a finite label set $\mathscr{Y}$, with respect to the zero-one loss.

Intuition.

- Let $\{x_i : i \in \mathscr{N}\}$ be an enumeration of the elements of $\mathscr{X}$ so that for all $i \leq j$, we have $\mathscr{D}(\{x_i\}) \leq \mathscr{D}(\{x_j\})$. Also $\lim_{n\to\infty} \sum_{i \geq n} \mathscr{D}(\{x_i\}) = 0$
- Given any $\epsilon > 0$, there exists $\epsilon_D > 0$ such that $\mathscr{D}(\{x \in \mathscr{X} : \mathscr{D}(\{x\}) < \epsilon_D\}) < \epsilon$
- For every $\eta > 0$, if n is such that $\mathscr{D}(\{x_i\}) < \eta$ for all $i > n$, then for every $m \in \mathscr{N}$, we have $\mathscr{P}_{S\sim\mathscr{D}^m}[\exists x_i : (\mathscr{D}(\{x_i\}) > \eta)$ and $x_i \notin S] \leq n e^{-\eta m}$

Consistency Learnability

A `Memorize` algorithm is universally consistent for every countable domain $\mathcal{X}$ and a finite label set $\mathcal{Y}$, with respect to the zero-one loss.

Intuition.

- Let $\{x_i : i \in \mathcal{N}\}$ be an enumeration of the elements of $\mathcal{X}$ so that for all $i \le j$, we have $\mathcal{D}(\{x_i\}) \le \mathcal{D}(\{x_j\})$. Also $\lim_{n\to\infty} \sum_{i \ge n} \mathcal{D}(\{x_i\}) = 0$
- Given any $\epsilon > 0$, there exists $\epsilon_D > 0$ such that $\mathcal{D}(\{x \in \mathcal{X} : \mathcal{D}(\{x\}) < \epsilon_D\}) < \epsilon$
- For every $\eta > 0$, if n is such that $\mathcal{D}(\{x_i\}) < \eta$ for all $i > n$, then for every $m \in \mathcal{N}$, we have $\mathcal{P}_{S \sim \mathcal{D}^m}[\exists x_i : (\mathcal{D}(\{x_i\}) > \eta)$ and $x_i \notin S] \le n e^{-\eta m}$
- If $\mathcal{X}$ is countable then for every probability distribution $\mathcal{D}$ over $\mathcal{X}$ there exists a function $m_{\mathcal{D}} : (0,1) \times (0,1) \to \mathcal{N}$ such that for every $\epsilon, \delta > 0$ if $m > m_{\mathcal{D}}(\epsilon, \delta)$ then $\mathcal{P}_{S \sim \mathcal{D}^m}[\mathcal{D}(\{x : x \notin S\}) > \epsilon] < \delta$

Consistency Learnability

A `Memorize` algorithm is universally consistent for every countable domain $\mathscr{X}$ and a finite label set $\mathscr{Y}$, with respect to the zero-one loss.

Intuition.

- Let $\{x_i : i \in \mathscr{N}\}$ be an enumeration of the elements of $\mathscr{X}$ so that for all $i \leq j$, we have $\mathscr{D}(\{x_i\}) \leq \mathscr{D}(\{x_j\})$. Also $\lim_{n\to\infty} \sum_{i \geq n} \mathscr{D}(\{x_i\}) = 0$
- Given any $\epsilon > 0$, there exists $\epsilon_D > 0$ such that $\mathscr{D}(\{x \in \mathscr{X} : \mathscr{D}(\{x\}) < \epsilon_D\}) < \epsilon$
- For every $\eta > 0$, if n is such that $\mathscr{D}(\{x_i\}) < \eta$ for all $i > n$, then for every $m \in \mathscr{N}$, we have $\mathscr{P}_{S\sim\mathscr{D}^m}[\exists x_i : (\mathscr{D}(\{x_i\}) > \eta)$ and $x_i \notin S] \leq ne^{-\eta m}$
- If $\mathscr{X}$ is countable then for every probability distribution $\mathscr{D}$ over $\mathscr{X}$ there exists a function $m_{\mathscr{D}} : (0,1) \times (0,1) \to \mathscr{N}$ such that for every $\epsilon, \delta > 0$ if $m > m_{\mathscr{D}}(\epsilon, \delta)$ then $\mathscr{P}_{S\sim\mathscr{D}^m}[\mathscr{D}(\{x : x \notin S\}) > \epsilon] < \delta$
- `Memorize` is a consistent learner for every class of binary valued functions over any countable domain.

□

Quiz

The VC Dimension

- Let $\mathscr{H}_1, \ldots, \mathscr{H}_r$ be hypothesis classes over some fixed domain set $\mathscr{X}$. Let $d = max_i VCDim(\mathscr{H}_i)$ and assume for simplicity that $d \geq 3$.

- Let $\mathcal{H}_1, \ldots, \mathcal{H}_r$ be hypothesis classes over some fixed domain set $\mathcal{X}$. Let $d = max_i VCDim(\mathcal{H}_i)$ and assume for simplicity that $d \geq 3$.

  1. Prove that for $r = 2$, it holds that

  $$VCDim(\mathcal{H}_\infty \cup \mathcal{H}_\epsilon) \leq 2d + 1$$

- Let $\mathscr{H}_1,\ldots,\mathscr{H}_r$ be hypothesis classes over some fixed domain set $\mathscr{X}$. Let $d = max_i VCDim(\mathscr{H}_i)$ and assume for simplicity that $d \geq 3$.
  1. Prove that for $r = 2$, it holds that

  $$VCDim(\mathscr{H}_\infty \cup \mathscr{H}_\epsilon) \leq 2d + 1$$

  2. Prove that

  $$VCDim(\cup_{i=1}^{r}\mathscr{H}_i \leq 4d\log(2d) + 2\log(r)$$

Non-uniform Learnability

- Let $\mathscr{H} = h_n \colon n \in \mathscr{N}$ be an infinite countable hypothesis class for binary classification. Show that it is impossible to assign weights to the hypotheses in $\mathscr{H}$ such that

Non-uniform Learnability

- Let $\mathcal{H} = h_n \colon n \in \mathcal{N}$ be an infinite countable hypothesis class for binary classification. Show that it is impossible to assign weights to the hypotheses in $\mathcal{H}$ such that
  1. $\mathcal{H}$ could be learnted nonuniformly using these weights. That is, the weighting function $\omega \colon \mathcal{H} \mapsto [0,1]$ should satisfy the condition $\sum_{h \in \mathcal{H}} \omega(\mathcal{H}) \leq 1$.

Non-uniform Learnability

- Let $\mathscr{H} = h_n \colon n \in \mathscr{N}$ be an infinite countable hypothesis class for binary classification. Show that it is impossible to assign weights to the hypotheses in $\mathscr{H}$ such that
  1. $\mathscr{H}$ could be learnted nonuniformly using these weights. That is, the weighting function $\omega \colon \mathscr{H} \mapsto [0,1]$ should satisfy the condition $\sum_{h \in \mathscr{H}} \omega(\mathscr{H}) \leq 1$.
  2. The weights could be monotonically nondecreaasing, namely, if $i < j$, then $\omega(\mathscr{H}_i) \leq \omega(\mathscr{H}_j)$.

Machine Learning Project

You are free to do either of the following projects

1. [Theoretical Analysis of Predictivity] Check the reference [ZLM+21], and think about how to evaluate the learnability or predictivity for time series data, or the sample complexity.

Machine Learning Project

You are free to do either of the following projects

1. [Theoretical Analysis of Predictivity] Check the reference [ZLM+21], and think about how to evaluate the learnability or predictivity for time series data, or the sample complexity.
2. [Tourism Demand Forecasting] Check the following references [LLFH19, ZLM+20, ZLML20] on the topic of tourism demand forecasting, and try to design/implement method to beat them in the prediction performance. The related source code and data sets are publicly available at:

Machine Learning Project

You are free to do either of the following projects

1. [Theoretical Analysis of Predictivity] Check the reference [ZLM+21], and think about how to evaluate the learnability or predictivity for time series data, or the sample complexity.
2. [Tourism Demand Forecasting] Check the following references [LLFH19, ZLM+20, ZLML20] on the topic of tourism demand forecasting, and try to design/implement method to beat them in the prediction performance. The related source code and data sets are publicly available at:
   - [TULIP Open-Code] `https://github.com/tulip-lab/open-code`
   - [TULIP Open-Data] `https://github.com/tulip-lab/open-data`

References

[LLFH19] Rob Law, Gang Li, Davis Fong, and Xin Han. Tourism demand forecasting: A deep learning approach. *Annals of Tourism Research*, January 2019.

[ZLM+20] Yishuo Zhang, Gang Li, Birgit Muskat, Rob Law, and Yating Yang. Group pooling for deep tourism demand forecasting. *Annals of Tourism Research*, 82:102899, 2020.

[ZLM+21] Chris Zhang, Gang Li, Birgit Muskat, Quan Vu, and Rob Law. Predictivity of tourism demand data. *Annals of Tourism Research*, 2021.

[ZLML20] Yishuo Zhang, Gang Li, Birgit Muskat, and Rob Law. Tourism demand forecasting: A decomposed deep learning approach. *Journal of Travel Research*, 2020.

Questions?

Contact Information

Associate Professor **Gang Li**
School of Information Technology
Deakin University
Geelong, Victoria 3216, Australia

gangli@tulip.org.au
Open Resources of TULIP-Lab
Team for Universal Learning and Intelligent Processing