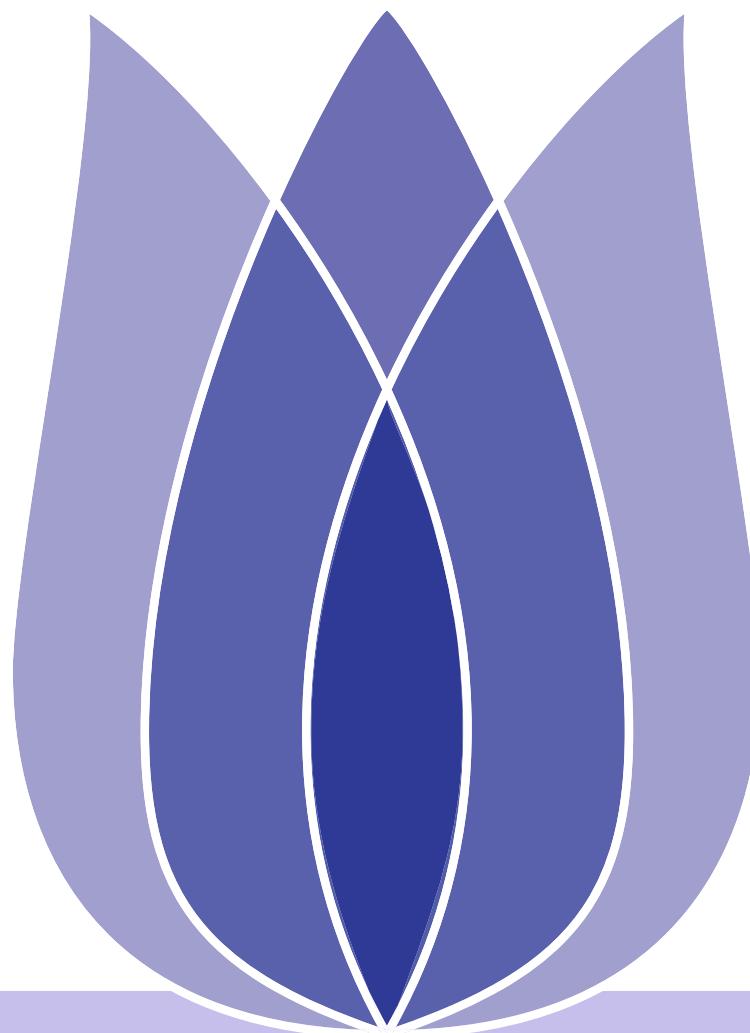


FUNDAMENTALS OF LEARNING AND INFORMATION PROCESSING

SESSION 13: STATISTICAL MACHINE LEARNING (III)



Gang Li

Deakin University, Australia

2021-08-09



Table of Content

[No Free Lunch Theorem](#)

[The VC Dimension](#)

[Quiz](#)

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Is *Infinite Class* PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz

No Free Lunch Theorem



PAC Learning

[No Free Lunch Theorem](#)

[PAC Learning](#)

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

[The VC Dimension](#)

[Quiz](#)

Leslie Valiant, Turing Award 2010



PAC Learning

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz

Leslie Valiant, Turing Award 2010

- For transformative contributions to the theory of computation, including
 - ◆ the theory of probably approximately correct (PAC) learning,



PAC Learning

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz

Leslie Valiant, Turing Award 2010

- For transformative contributions to the theory of computation, including
 - ◆ the theory of probably approximately correct (PAC) learning,
 - ◆ the complexity of enumeration and of algebraic computation,



PAC Learning

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz

Leslie Valiant, Turing Award 2010

- For transformative contributions to the theory of computation, including
 - ◆ the theory of **probably approximately correct** (PAC) learning,
 - ◆ the complexity of enumeration and of algebraic computation,
 - ◆ and the theory of parallel and distributed computing.





What is Learnable and How to Learn?

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz

For any finite hypothesis class \mathcal{H} , we have shown that:

- \mathcal{H} is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \rceil$$

- \mathcal{H} is Agnostic PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \rceil$$

- This sample complexity is obtained using the $ERM_{\mathcal{H}}$ learning rule





What is Learnable and How to Learn?

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz

For any finite hypothesis class \mathcal{H} , we have shown that:

- \mathcal{H} is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \rceil$$

- \mathcal{H} is Agnostic PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \rceil$$

- This sample complexity is obtained using the $ERM_{\mathcal{H}}$ learning rule

What is more?

- What about infinite hypothesis classes?



What is Learnable and How to Learn?

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz

For any finite hypothesis class \mathcal{H} , we have shown that:

- \mathcal{H} is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \rceil$$

- \mathcal{H} is Agnostic PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \rceil$$

- This sample complexity is obtained using the $ERM_{\mathcal{H}}$ learning rule

What is more?

- What about infinite hypothesis classes?
- What is the sample complexity of a given class?



What is Learnable and How to Learn?

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz

For any finite hypothesis class \mathcal{H} , we have shown that:

- \mathcal{H} is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \rceil$$

- \mathcal{H} is Agnostic PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \rceil$$

- This sample complexity is obtained using the $ERM_{\mathcal{H}}$ learning rule

What is more?

- What about infinite hypothesis classes?
- What is the sample complexity of a given class?
- Is there a generic learning algorithm that achieves the optimal sample complexity?





The No-Free-Lunch Theorem

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain \mathcal{X} , and m be any number representing a training set size: $m \leq |\mathcal{X}|/2$. Then, there exists a distribution \mathcal{D} over $\mathcal{X} \times \{0, 1\}$ such that

- There exists a function $f : \mathcal{X} \rightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$
- With probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8$



The No-Free-Lunch Theorem

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain \mathcal{X} , and m be any number representing a training set size: $m \leq |\mathcal{X}|/2$. Then, there exists a distribution \mathcal{D} over $\mathcal{X} \times \{0, 1\}$ such that

- There exists a function $f : \mathcal{X} \rightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$
- With probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8$

Intuition.

- It means that: *for every learner, there exists a task on which it fails, even though that task can be successfully learned by another learner.*



The No-Free-Lunch Theorem

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz



Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain \mathcal{X} , and m be any number representing a training set size: $m \leq |\mathcal{X}|/2$. Then, there exists a distribution \mathcal{D} over $\mathcal{X} \times \{0, 1\}$ such that

- There exists a function $f : \mathcal{X} \rightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$
- With probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8$

Intuition.

- It means that: *for every learner, there exists a task on which it fails, even though that task can be successfully learned by another learner.*
- We need to show
 1. For any algorithm A that receives a training set S of m examples from $\mathcal{X} \times \{0, 1\}$, there exists a function $f : \mathcal{X} \rightarrow \{0, 1\}$ such that $L_{\mathcal{D}}(f) = 0$ and $E_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq 1/4$
 2. From above, we derive that $P[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq 1/7$

□



The No-Free-Lunch Theorem

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain \mathcal{X} , and m be any number representing a training set size: $m \leq |\mathcal{X}|/2$. Then, there exists a distribution \mathcal{D} over $\mathcal{X} \times \{0, 1\}$ such that

- There exists a function $f : \mathcal{X} \rightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$
- With probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8$

Proof-1.

1. Estimate the probabilities that any given learner A will err on a random point: the probability that a random test point x , is not covered by S is at least $1/2$.

□



The No-Free-Lunch Theorem

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz

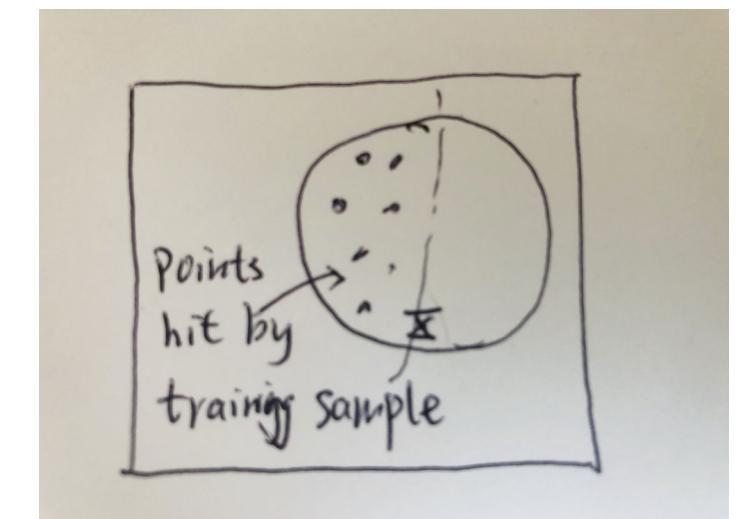


Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain \mathcal{X} , and m be any number representing a training set size: $m \leq |\mathcal{X}|/2$. Then, there exists a distribution \mathcal{D} over $\mathcal{X} \times \{0, 1\}$ such that

- There exists a function $f : \mathcal{X} \rightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$
- With probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8$

Proof-1.

1. Estimate the probabilities that any given learner A will err on a random point: the probability that a random test point x , is not covered by S is at least $1/2$.
2. Any hypothesis from A will have the probability 50 % of making error on x , so the expected error will be at least $1/4$.



□



The No-Free-Lunch Theorem

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

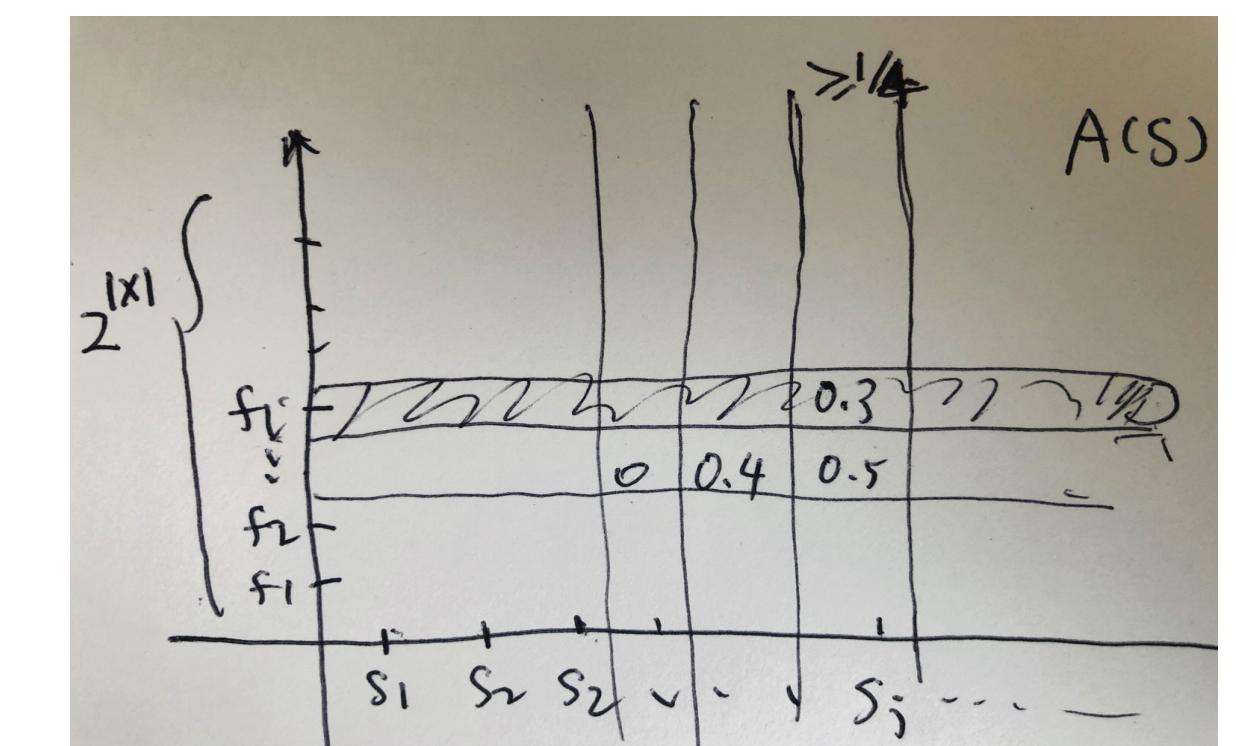
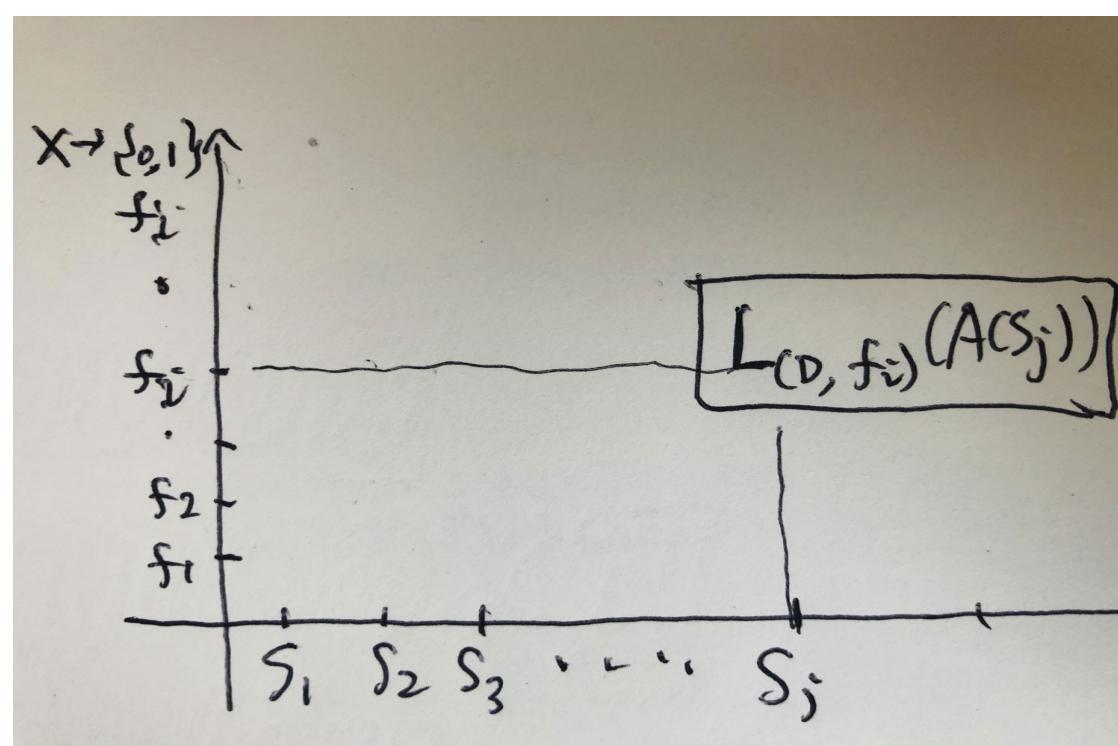
Quiz



Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain \mathcal{X} , and m be any number representing a training set size: $m \leq |\mathcal{X}|/2$. Then, there exists a distribution \mathcal{D} over $\mathcal{X} \times \{0, 1\}$ such that

- There exists a function $f : \mathcal{X} \rightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$
- With probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8$

Proof-1.





The No-Free-Lunch Theorem

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain \mathcal{X} , and m be any number representing a training set size: $m \leq |\mathcal{X}|/2$. Then, there exists a distribution \mathcal{D} over $\mathcal{X} \times \{0, 1\}$ such that

- There exists a function $f : \mathcal{X} \rightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$
- With probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8$

Proof-2.

1. Let θ be a random variable that receives values in $[0, 1]$ and $E[\theta] \geq 1/4$.



The No-Free-Lunch Theorem

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain \mathcal{X} , and m be any number representing a training set size: $m \leq |\mathcal{X}|/2$. Then, there exists a distribution \mathcal{D} over $\mathcal{X} \times \{0, 1\}$ such that

- There exists a function $f : \mathcal{X} \rightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$
- With probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$, we have $L_{\mathcal{D}}(A(S)) \geq 1/8$

Proof-2.

1. Let θ be a random variable that receives values in $[0, 1]$ and $E[\theta] \geq 1/4$.
2. From **Markov's Inequality**,

$$P[\theta \geq 1/8] \geq \frac{1/4 - 1/8}{1 - 1/8} = 1/7$$

□



NFL Theorem and Prior Knowledge

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let \mathcal{X} be an infinite domain set and \mathcal{H} be the set of **all functions** $\mathcal{X} \rightarrow \{0, 1\}$, then \mathcal{H} is not PAC learnable.



NFL Theorem and Prior Knowledge

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz



Let \mathcal{X} be an infinite domain set and \mathcal{H} be the set of **all functions** $\mathcal{X} \rightarrow \{0, 1\}$, then \mathcal{H} is not PAC learnable.

Proof-V1.

- By way of contradiction: if \mathcal{H} is learnable, choose $\epsilon < 1/8$ and $\delta < 1/7$.



NFL Theorem and Prior Knowledge

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz



Let \mathcal{X} be an infinite domain set and \mathcal{H} be the set of all functions $\mathcal{X} \rightarrow \{0, 1\}$, then \mathcal{H} is not PAC learnable.

Proof-V1.

- By way of contradiction: if \mathcal{H} is learnable, choose $\epsilon < 1/8$ and $\delta < 1/7$.
- By the definition of PAC learnability: there exists a learning algorithm A and an integer $m = m(\epsilon, \delta)$, such that, for any distribution \mathcal{D} over $\mathcal{X} \times \{0, 1\}$, if for some function $f : \mathcal{X} \rightarrow \{0, 1\}$, $L_{\mathcal{D}}(f) = 0$, then with probability greater than $1 - \delta$, $L_{\mathcal{D}}(A(S)) \leq \epsilon$.



NFL Theorem and Prior Knowledge

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz



Let \mathcal{X} be an infinite domain set and \mathcal{H} be the set of **all functions** $\mathcal{X} \rightarrow \{0, 1\}$, then \mathcal{H} is not PAC learnable.

Proof-V1.

- By way of contradiction: if \mathcal{H} is learnable, choose $\epsilon < 1/8$ and $\delta < 1/7$.
- By the definition of PAC learnability: there exists a learning algorithm A and an integer $m = m(\epsilon, \delta)$, such that, for **any distribution** \mathcal{D} over $\mathcal{X} \times \{0, 1\}$, if for some function $f : \mathcal{X} \rightarrow \{0, 1\}$, $L_{\mathcal{D}}(f) = 0$, then with probability greater than $1 - \delta$, $L_{\mathcal{D}}(A(S)) \leq \epsilon$.
- From NFL, since $|\mathcal{X}| > 2m$, for every learning algorithm, there **exists a distribution** \mathcal{D} such that with probability greater than $1/7 > \delta$, $L_{\mathcal{D}}(A(S)) > 1/8 > \epsilon$. contradiction!

□



NFL Theorem and Prior Knowledge

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz



Let \mathcal{X} be an infinite domain set and \mathcal{H} be the set of **all functions** $\mathcal{X} \rightarrow \{0, 1\}$, then \mathcal{H} is not PAC learnable.

Proof-V2.

- By way of contradiction: if \mathcal{H} is learnable, then for some ϵ and δ we have $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathcal{N}$, for every distribution \mathcal{D} over \mathcal{X} and function f :

$$P_{S \sim \mathcal{D}^m}[L_{\mathcal{D}, f}(A(S)) > \epsilon] < \delta$$

whenever $m > m_{\mathcal{H}}(\epsilon, \delta)$



NFL Theorem and Prior Knowledge

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz



Let \mathcal{X} be an infinite domain set and \mathcal{H} be the set of **all functions** $\mathcal{X} \rightarrow \{0, 1\}$, then \mathcal{H} is not PAC learnable.

Proof-V2.

- By way of contradiction: if \mathcal{H} is learnable, then for some ϵ and δ we have $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathcal{N}$, for every distribution \mathcal{D} over \mathcal{X} and function f :

$$P_{S \sim \mathcal{D}^m}[L_{\mathcal{D}, f}(A(S)) > \epsilon] < \delta$$

whenever $m > m_{\mathcal{H}}(\epsilon, \delta)$

- Let $m = 2m_{\mathcal{H}}(1/8, 1/7)$, \mathcal{H} includes every possible function from S to $\{0, 1\}$.



NFL Theorem and Prior Knowledge

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz



Let \mathcal{X} be an infinite domain set and \mathcal{H} be the set of all functions $\mathcal{X} \rightarrow \{0, 1\}$, then \mathcal{H} is not PAC learnable.

Proof-V2.

- By way of contradiction: if \mathcal{H} is learnable, then for some ϵ and δ we have $m_{\mathcal{H}} : (0, 1)^2 \rightarrow \mathcal{N}$, for every distribution \mathcal{D} over \mathcal{X} and function f :

$$P_{S \sim \mathcal{D}^m}[L_{\mathcal{D}, f}(A(S)) > \epsilon] < \delta$$

whenever $m > m_{\mathcal{H}}(\epsilon, \delta)$

- Let $m = 2m_{\mathcal{H}}(1/8, 1/7)$, \mathcal{H} includes every possible function from S to $\{0, 1\}$.
- From NFL, we have

$$m_{\mathcal{H}}(1/8, 1/7) \geq m/2 > m_{\mathcal{H}}(1/8, 1/7)$$

Contradiction!!!





NFL Theorem and Prior Knowledge

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#) **Current**

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let \mathcal{X} be an infinite domain set and \mathcal{H} be the set of **all functions** $\mathcal{X} \rightarrow \{0, 1\}$, then \mathcal{H} is not PAC learnable.



We can escape the hazards foreseen by the NFL theorem by **using our prior knowledge about a specific learning task**, to avoid the distributions that will cause us to fail when learning that task.

- Such prior knowledge can be expressed by restricting our hypothesis class.



Error Decomposition

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let h_S be an $ERM_{\mathcal{H}}$ hypothesis: $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$. Assume $f \in \mathcal{Y}^{\mathcal{X}}$ be the true hypothesis, and h^* is the best hypothesis in \mathcal{H} : $h^* = \min_{h^* \in \mathcal{H}} L_{\mathcal{D}}(h^*)$. Then we can have:

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f) = L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*) + L_{\mathcal{D}}(h^*) - L_{\mathcal{D}}(f) = \epsilon_{est} + \epsilon_{app}$$



Error Decomposition

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let h_S be an $ERM_{\mathcal{H}}$ hypothesis: $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$. Assume $f \in \mathcal{Y}^{\mathcal{X}}$ be the true hypothesis, and h^* is the best hypothesis in \mathcal{H} : $h^* = \min_{h^* \in \mathcal{H}} L_{\mathcal{D}}(h^*)$. Then we can have:

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f) = L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*) + L_{\mathcal{D}}(h^*) - L_{\mathcal{D}}(f) = \epsilon_{est} + \epsilon_{app}$$

The Approximation Error $\epsilon_{app} = L_{\mathcal{D}}(h^*) - L_{\mathcal{D}}(f)$

- How much risk do we have due to restricting to \mathcal{H}
- It does not depend on S
- It decreases with the complexity (size, or VC dimension) of \mathcal{H}



Error Decomposition

No Free Lunch Theorem

PAC Learning

The No-Free-Lunch Theorem

NFL Theorem and Prior Knowledge

Error Decomposition

The VC Dimension

Quiz



Let h_S be an $ERM_{\mathcal{H}}$ hypothesis: $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$. Assume $f \in \mathcal{Y}^{\mathcal{X}}$ be the true hypothesis, and h^* is the best hypothesis in \mathcal{H} : $h^* = \min_{h^* \in \mathcal{H}} L_{\mathcal{D}}(h^*)$. Then we can have:

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f) = L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*) + L_{\mathcal{D}}(h^*) - L_{\mathcal{D}}(f) = \epsilon_{est} + \epsilon_{app}$$

The Approximation Error $\epsilon_{app} = L_{\mathcal{D}}(h^*) - L_{\mathcal{D}}(f)$

- How much risk do we have due to restricting to \mathcal{H}
- It does not depend on S
- It decreases with the complexity (size, or VC dimension) of \mathcal{H}

The Estimation Error $\epsilon_{est} = L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*)$

- Result of L_S being only an estimate of $L_{\mathcal{D}}$
- It decreases with the size of S
- It increases with the complexity of \mathcal{H}
- We have $\epsilon_{est} = L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*) < 2 \sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)|$. From **Hoeffding's Inequality**, this provides an upper bound for ϵ_{est}



Error Decomposition

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let h_S be an $ERM_{\mathcal{H}}$ hypothesis: $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$. Assume $f \in \mathcal{Y}^{\mathcal{X}}$ be the true hypothesis, and h^* is the best hypothesis in \mathcal{H} : $h^* = \min_{h^* \in \mathcal{H}} L_{\mathcal{D}}(h^*)$. Then we can have:

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f) = L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*) + L_{\mathcal{D}}(h^*) - L_{\mathcal{D}}(f) = \epsilon_{est} + \epsilon_{app}$$

Bias-Complexity Tradeoff.

Overfitting Choosing \mathcal{H} to be a very rich class will decrease the ϵ_{app} , but at the same time might increase the ϵ_{est} , as a rich \mathcal{H} might lead to **overfitting**.



Error Decomposition

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let h_S be an $ERM_{\mathcal{H}}$ hypothesis: $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$. Assume $f \in \mathcal{Y}^{\mathcal{X}}$ be the true hypothesis, and h^* is the best hypothesis in \mathcal{H} : $h^* = \min_{h^* \in \mathcal{H}} L_{\mathcal{D}}(h^*)$. Then we can have:

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f) = L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*) + L_{\mathcal{D}}(h^*) - L_{\mathcal{D}}(f) = \epsilon_{est} + \epsilon_{app}$$

Bias-Complexity Tradeoff.

Overfitting Choosing \mathcal{H} to be a very rich class will decrease the ϵ_{app} , but at the same time might increase the ϵ_{est} , as a rich \mathcal{H} might lead to **overfitting**.

Underfitting Choosing \mathcal{H} to be a very small class will reduce the estimation error ϵ_{est} , but at the same time might increase the approximation error ϵ_{app} , as a small \mathcal{H} might lead to **underfitting**.



Error Decomposition

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

[Error Decomposition](#)

[The VC Dimension](#)

[Quiz](#)



Let h_S be an $ERM_{\mathcal{H}}$ hypothesis: $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$. Assume $f \in \mathcal{Y}^{\mathcal{X}}$ be the true hypothesis, and h^* is the best hypothesis in \mathcal{H} : $h^* = \min_{h^* \in \mathcal{H}} L_{\mathcal{D}}(h^*)$. Then we can have:

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f) = L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*) + L_{\mathcal{D}}(h^*) - L_{\mathcal{D}}(f) = \epsilon_{est} + \epsilon_{app}$$

Bias-Complexity Tradeoff.

Overfitting Choosing \mathcal{H} to be a very rich class will decrease the ϵ_{app} , but at the same time might increase the ϵ_{est} , as a rich \mathcal{H} might lead to **overfitting**.

Underfitting Choosing \mathcal{H} to be a very small class will reduce the estimation error ϵ_{est} , but at the same time might increase the approximation error ϵ_{app} , as a small \mathcal{H} might lead to **underfitting**.

Bayes Optimal However, Bayes optimal classifier depends on the underlying distribution \mathcal{D} , which we do not know.





Error Decomposition

[No Free Lunch Theorem](#)

[PAC Learning](#)

[The No-Free-Lunch Theorem](#)

[NFL Theorem and Prior Knowledge](#)

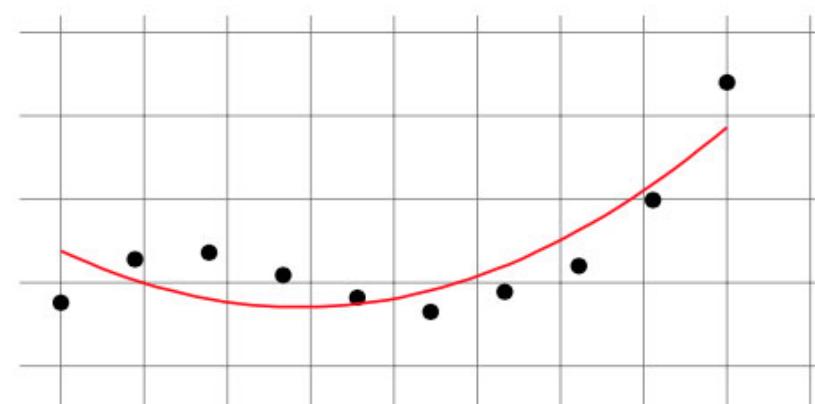
[Error Decomposition](#)



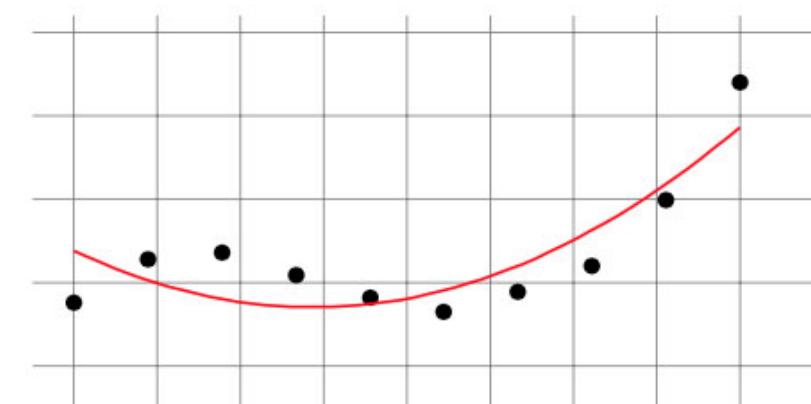
Let h_S be an $ERM_{\mathcal{H}}$ hypothesis: $h_S \in \operatorname{argmin}_{h \in \mathcal{H}} L_S(h)$. Assume $f \in \mathcal{Y}^{\mathcal{X}}$ be the true hypothesis, and h^* is the best hypothesis in \mathcal{H} : $h^* = \min_{h^* \in \mathcal{H}} L_{\mathcal{D}}(h^*)$. Then we can have:

$$L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(f) = L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h^*) + L_{\mathcal{D}}(h^*) - L_{\mathcal{D}}(f) = \epsilon_{est} + \epsilon_{app}$$

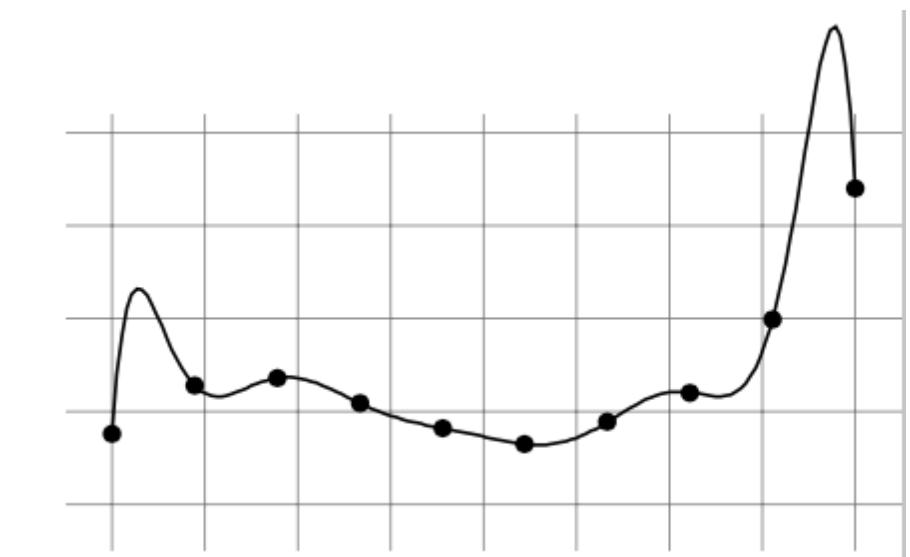
Bias-Complexity Tradeoff: Example.



(a) degree 2



(b) degree 3



(c) degree 10





No Free Lunch Theorem

The VC Dimension

Is *Infinite Class* PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

The VC Dimension



Is *Infinite Class* PAC learnable?

[No Free Lunch Theorem](#)

[The VC Dimension](#)

[Is *Infinite Class* PAC learnable?](#)

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

[Quiz](#)



Threshold functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_\theta = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathcal{R}\}$

\mathcal{H}_θ is PAC learnable using ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(2/\delta)}{\epsilon} \rceil$



Is Infinite Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

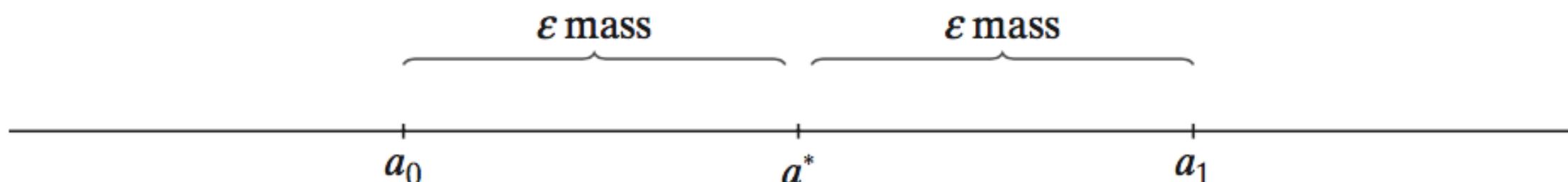


Threshold functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_\theta = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathcal{R}\}$

\mathcal{H}_θ is PAC learnable using ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(2/\delta)}{\epsilon} \rceil$

Proof.

1. Let a^* be the perfect threshold, namely $L_{\mathcal{D}}(h_{a^*}) = 0$. let $a_0 < a^* < a_1$ be such at $P[x \in (a_0, a^*)] = P[x \in (a^*, a_1)] = \epsilon$





Is Infinite Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

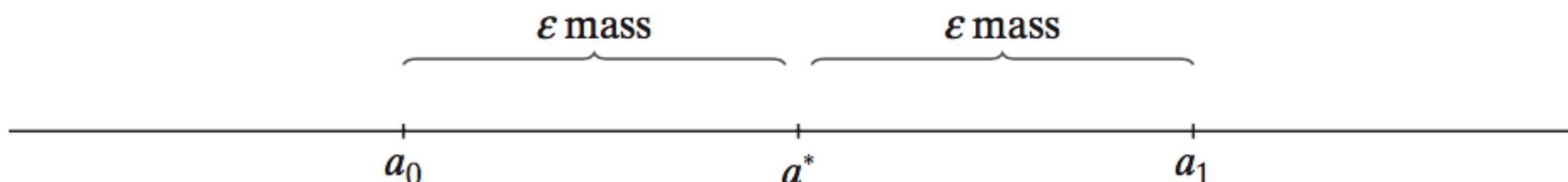


Threshold functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_\theta = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathcal{R}\}$

\mathcal{H}_θ is PAC learnable using ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(2/\delta)}{\epsilon} \rceil$

Proof.

1. Let a^* be the perfect threshold, namely $L_{\mathcal{D}}(h_{a^*}) = 0$. let $a_0 < a^* < a_1$ be such at $P[x \in (a_0, a^*)] = P[x \in (a^*, a_1)] = \epsilon$



2. Let b_S be the threshold for ERM result h_S , let $b_0 = \max\{x : (x, -1) \in S\}$ and $b_1 = \min\{x : (x, +1) \in S\}$.



Is Infinite Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

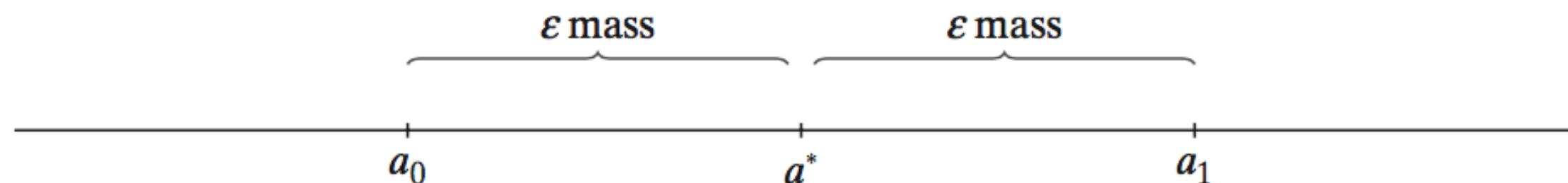


Threshold functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_\theta = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathcal{R}\}$

\mathcal{H}_θ is PAC learnable using ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(2/\delta)}{\epsilon} \rceil$

Proof.

1. Let a^* be the perfect threshold, namely $L_{\mathcal{D}}(h_{a^*}) = 0$. let $a_0 < a^* < a_1$ be such at $P[x \in (a_0, a^*)] = P[x \in (a^*, a_1)] = \epsilon$



2. Let b_S be the threshold for ERM result h_S , let $b_0 = \max\{x : (x, -1) \in S\}$ and $b_1 = \min\{x : (x, +1) \in S\}$.
3. A sufficient condition for $L_{\mathcal{D}}(h_s) > \epsilon$ can be rewritten as:

$$P[L_{\mathcal{D}}(h_s) > \epsilon] \leq P[b_0 < a_0 \vee b_1 > a_1] \leq P[b_0 < a_0] + P[b_1 > a_1]$$

□



Is Infinite Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Threshold functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_\theta = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathcal{R}\}$

\mathcal{H}_θ is PAC learnable using ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(2/\delta)}{\epsilon} \rceil$

Proof.

4. The event $b_0 < a_0$ happens if and only if all examples in S are not in the interval (a_0, a^*) , whose probability mass is defined to be ϵ :

$$P[b_0 < a_0] = P[\forall (x, y) \in S, x \notin (a_0, a^*)] = (1 - \epsilon)^m \leq e^{\epsilon m}$$



Is Infinite Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Threshold functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_\theta = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathcal{R}\}$

\mathcal{H}_θ is PAC learnable using ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(2/\delta)}{\epsilon} \rceil$

Proof.

4. The event $b_0 < a_0$ happens if and only if all examples in S are not in the interval (a_0, a^*) , whose probability mass is defined to be ϵ :

$$P[b_0 < a_0] = P[\forall (x, y) \in S, x \notin (a_0, a^*)] = (1 - \epsilon)^m \leq e^{\epsilon m}$$

5. When we have $m > \log \frac{2/\delta}{\epsilon}$, $P[b_0 < a_0]$ is at most $\frac{\delta}{2}$.



Is Infinite Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Threshold functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_\theta = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathcal{R}\}$

\mathcal{H}_θ is PAC learnable using ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(2/\delta)}{\epsilon} \rceil$

Proof.

4. The event $b_0 < a_0$ happens if and only if all examples in S are not in the interval (a_0, a^*) , whose probability mass is defined to be ϵ :

$$P[b_0 < a_0] = P[\forall (x, y) \in S, x \notin (a_0, a^*)] = (1 - \epsilon)^m \leq e^{-\epsilon m}$$

5. When we have $m > \log \frac{2/\delta}{\epsilon}$, $P[b_0 < a_0]$ is at most $\frac{\delta}{2}$.
6. Similarly, we have $P[b_1 > a_1] \leq \frac{\delta}{2}$.



Is Infinite Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Threshold functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_\theta = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathcal{R}\}$

\mathcal{H}_θ is PAC learnable using ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log(2/\delta)}{\epsilon} \rceil$

Proof.

4. The event $b_0 < a_0$ happens if and only if all examples in S are not in the interval (a_0, a^*) , whose probability mass is defined to be ϵ :

$$P[b_0 < a_0] = P[\forall (x, y) \in S, x \notin (a_0, a^*)] = (1 - \epsilon)^m \leq e^{-\epsilon m}$$

5. When we have $m > \log \frac{2/\delta}{\epsilon}$, $P[b_0 < a_0]$ is at most $\frac{\delta}{2}$.
6. Similarly, we have $P[b_1 > a_1] \leq \frac{\delta}{2}$.
7. Combining with above, we conclude the proof.





Is *Infinite Hypothesis Class* PAC learnable?

[No Free Lunch Theorem](#)

[The VC Dimension](#)

[Is Infinite Class PAC learnable?](#)

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

[Quiz](#)

Finite-Set functions Let $\mathcal{H}_F = \{x \mapsto h_F(x) : \text{finite set } F \subseteq \mathcal{R}\} \cup \{x \mapsto 1\}$



\mathcal{H}_F is **not** PAC learnable using ERM rule. $h_F(x) = \begin{cases} 1 & \text{if } x \in F \\ -1 & \text{if } x \notin F \end{cases}$



Is Infinite Hypothesis Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

Finite-Set functions Let $\mathcal{H}_F = \{x \mapsto h_F(x) : \text{finite set } F \subseteq \mathcal{R}\} \cup \{x \mapsto 1\}$



\mathcal{H}_F is not PAC learnable using ERM rule. $h_F(x) = \begin{cases} 1 & \text{if } x \in F \\ -1 & \text{if } x \notin F \end{cases}$

Proof.

1. Let \mathcal{D} be the uniform distribution over \mathcal{X} , and the labelling function is $f = \{x \mapsto 1\}$.



Is Infinite Hypothesis Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

Finite-Set functions Let $\mathcal{H}_F = \{x \mapsto h_F(x) : \text{finite set } F \subseteq \mathcal{R}\} \cup \{x \mapsto 1\}$



\mathcal{H}_F is not PAC learnable using ERM rule. $h_F(x) = \begin{cases} 1 & \text{if } x \in F \\ -1 & \text{if } x \notin F \end{cases}$

Proof.

1. Let \mathcal{D} be the uniform distribution over \mathcal{X} , and the labelling function is $f = \{x \mapsto 1\}$.
2. Let $S \sim \mathcal{D}^m$ be a sample with size m from distribution \mathcal{D} : $S = \{(x_1, 1), \dots, (x_m, 1)\}$



Is Infinite Hypothesis Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

Finite-Set functions Let $\mathcal{H}_F = \{x \mapsto h_F(x) : \text{finite set } F \subseteq \mathcal{R}\} \cup \{x \mapsto 1\}$



\mathcal{H}_F is not PAC learnable using ERM rule. $h_F(x) = \begin{cases} 1 & \text{if } x \in F \\ -1 & \text{if } x \notin F \end{cases}$

Proof.

1. Let \mathcal{D} be the uniform distribution over \mathcal{X} , and the labelling function is $f = \{x \mapsto 1\}$.
2. Let $S \sim \mathcal{D}^m$ be a sample with size m from distribution \mathcal{D} : $S = \{(x_1, 1), \dots, (x_m, 1)\}$
3. An ERM learner may pick a $h_F \in \mathcal{H}_F$ for $F = \{x_1, \dots, x_m\}$, and with $L_S(h_F) = 0$.



Is Infinite Hypothesis Class PAC learnable?

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

Finite-Set functions Let $\mathcal{H}_F = \{x \mapsto h_F(x) : \text{finite set } F \subseteq \mathcal{R}\} \cup \{x \mapsto 1\}$



\mathcal{H}_F is not PAC learnable using ERM rule. $h_F(x) = \begin{cases} 1 & \text{if } x \in F \\ -1 & \text{if } x \notin F \end{cases}$

Proof.

1. Let \mathcal{D} be the uniform distribution over \mathcal{X} , and the labelling function is $f = \{x \mapsto 1\}$.
2. Let $S \sim \mathcal{D}^m$ be a sample with size m from distribution \mathcal{D} : $S = \{(x_1, 1), \dots, (x_m, 1)\}$
3. An ERM learner may pick a $h_F \in \mathcal{H}_F$ for $F = \{x_1, \dots, x_m\}$, and with $L_S(h_F) = 0$.
4. What is $L_{\mathcal{D}}(h_F)$?
 - h_F fails on every test point which is outside F .
 - $L_{\mathcal{D}}(h_F) = 1$





The VC Dimension

[No Free Lunch Theorem](#)

[The VC Dimension](#)

Is Infinite Class PAC learnable?

[The VC Dimension](#)

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

[Quiz](#)

Let $C = \{x_1, \dots, x_m\} \subset \mathcal{X}$, and \mathcal{H} be a class of hypothesis from \mathcal{X} to $\{\pm 1\}$. The **restriction \mathcal{H}_C of \mathcal{H} to C** is the set of hypothesis from C to $\{\pm 1\}$ that can be derived from \mathcal{H} : $\forall x_i \in C, h_C(x_i) = h(x_i)$

■ we can represent \mathcal{H}_C as the vector $(h(x_1), \dots, h(x_m)) \in \{\pm 1\}^m$





The VC Dimension

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Let $C = \{x_1, \dots, x_m\} \subset \mathcal{X}$, and \mathcal{H} be a class of hypothesis from \mathcal{X} to $\{\pm 1\}$. The **restriction \mathcal{H}_C of \mathcal{H} to C** is the set of hypothesis from C to $\{\pm 1\}$ that can be derived from \mathcal{H} : $\forall x_i \in C, h_C(x_i) = h(x_i)$

- we can represent \mathcal{H}_C as the vector $(h(x_1), \dots, h(x_m)) \in \{\pm 1\}^m$

Defn

A hypothesis class \mathcal{H} **shatters** a finite set $C \subset \mathcal{X}$ if the restriction of \mathcal{H} to C is the set of all possible functions from C to $\{\pm 1\}$.

- in this case, we have $|\mathcal{H}_C| = 2^{|C|}$



The VC Dimension

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

Let $C = \{x_1, \dots, x_m\} \subset \mathcal{X}$, and \mathcal{H} be a class of hypothesis from \mathcal{X} to $\{\pm 1\}$. The **restriction \mathcal{H}_C of \mathcal{H} to C** is the set of hypothesis from C to $\{\pm 1\}$ that can be derived from \mathcal{H} : $\forall x_i \in C, h_C(x_i) = h(x_i)$

- we can represent \mathcal{H}_C as the vector $(h(x_1), \dots, h(x_m)) \in \{\pm 1\}^m$

Defn A hypothesis class \mathcal{H} **shatters** a finite set $C \subset \mathcal{X}$ if the restriction of \mathcal{H} to C is the set of all possible functions from C to $\{\pm 1\}$.

- in this case, we have $|\mathcal{H}_C| = 2^{|C|}$

Defn The **VC-dimension** of a hypothesis class \mathcal{H} , denoted $VCdim(\mathcal{H})$, the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by \mathcal{H} .

- If \mathcal{H} can **shatter** sets of arbitrary large size, we say that $VCdim(\mathcal{H}) = \infty$



The VC Dimension

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Let $C = \{x_1, \dots, x_m\} \subset \mathcal{X}$, and \mathcal{H} be a class of hypothesis from \mathcal{X} to $\{\pm 1\}$. The **restriction \mathcal{H}_C of \mathcal{H} to C** is the set of hypothesis from C to $\{\pm 1\}$ that can be derived from \mathcal{H} : $\forall x_i \in C, h_C(x_i) = h(x_i)$

- we can represent \mathcal{H}_C as the vector $(h(x_1), \dots, h(x_m)) \in \{\pm 1\}^m$

Defn

A hypothesis class \mathcal{H} **shatters** a finite set $C \subset \mathcal{X}$ if the restriction of \mathcal{H} to C is the set of all possible functions from C to $\{\pm 1\}$.

- in this case, we have $|\mathcal{H}_C| = 2^{|C|}$



The **VC-dimension** of a hypothesis class \mathcal{H} , denoted $VCdim(\mathcal{H})$, the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by \mathcal{H} .

- If \mathcal{H} can **shatter** sets of arbitrary large size, we say that $VCdim(\mathcal{H}) = \infty$



- To show that $VCdim(\mathcal{H}) = d$, we need to show

- ◆ There exists a set C of size d which is **shattered** by \mathcal{H}
 - ◆ Any set C of size $d + 1$ is **not shattered** by \mathcal{H}



The VC Dimension — Examples (1)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

Threshold functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_\theta = \{x \mapsto \text{sign}(x - \theta) : \theta \in \mathcal{R}\}$



$$VCdim(\mathcal{H}_\theta) = 1$$

Demo.

- Show that $\{0\}$ is shattered
- Show that any two points can not be shattered





The VC Dimension — Examples (2)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

Finite-Set functions

Let $\mathcal{H}_F = \{x \mapsto h_F(x) : \text{finite set } F \subseteq \mathcal{R}\} \cup \{x \mapsto 1\}$



$$VCdim(\mathcal{H}_F) = \infty$$

Demo.

- Show that any set can be shattered





The VC Dimension — Examples (3)

No Free Lunch Theorem

[The VC Dimension](#)

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

[Quiz](#)

Finite Hypothesis Classes



Let \mathcal{H} be a finite hypothesis class,

$$VCdim(\mathcal{H}) < \log_2(|\mathcal{H}|)$$

Demo.

- Show that any set with at most $\log_2(|\mathcal{H}|)$ size can be shattered





The VC Dimension — Examples (4)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4) **Current**

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

Intervals functions $\mathcal{X} = \mathcal{R}$, $\mathcal{Y} = \pm 1$, let $\mathcal{H}_{a,b} = \{h_{a,b}(x) : a < b \in \mathcal{R}\}$, where $h_{a,b}(x) = 1$ iff $x \in (a, b)$.



$$VCdim(\mathcal{H}_{a,b}) = 2$$

Demo.

- Show that $\{0, 1\}$ is shattered
- Show that any three points can not be shattered





The VC Dimension — Examples (5)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Axis aligned rectangles $\mathcal{X} = \mathbb{R}^2$, let $\mathcal{H}_{a_1, a_2, b_1, b_2} = \{h_{a_1, a_2, b_1, b_2}(x) : a_1 < a_2 \wedge b_1 < b_2\}$, where $h_{a_1, a_2, b_1, b_2}(x_1, x_2) = 1$ iff $x_1 \in (a_1, a_2)$ and $x_2 \in (b_1, b_2)$.

$$VCdim(\mathcal{H}_{a_1, a_2, b_1, b_2}) = 4$$

Demo.





The VC Dimension — Examples (6)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz

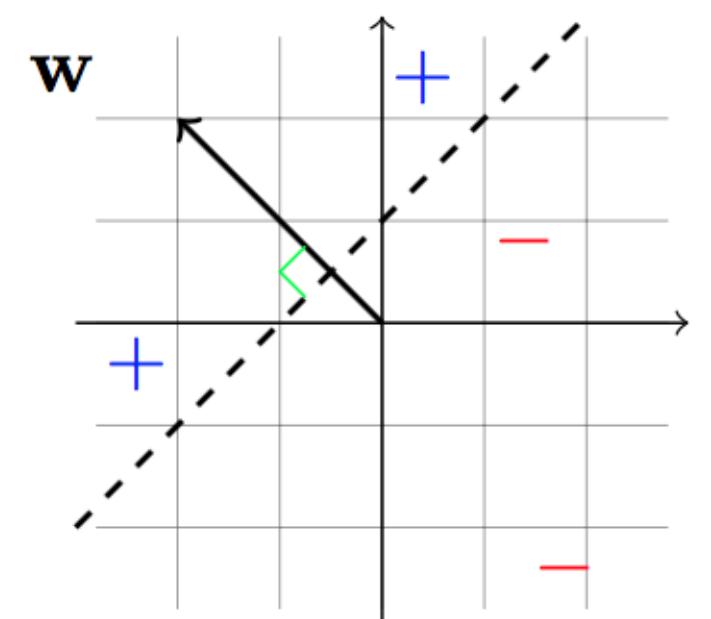


Halfspaces $\mathcal{X} = \mathbb{R}^d$, let $\mathcal{H}_S = \{\vec{x} \mapsto \text{sign}(\langle \vec{w}, \vec{x} \rangle) : \vec{w} \in \mathbb{R}^d\}$, where the inner product $\langle \vec{w}, \vec{x} \rangle = \vec{w} \cdot \vec{x} = \sum_{i=1}^d w_i x_i$.

$$VCdim(\mathcal{H}_S) = d$$

Demo.

- Show that $\{e_1, \dots, e_d\}$ is shattered
- Show that any $d + 1$ points can not be shattered





The VC Dimension — Examples (6)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Halfspaces $\mathcal{X} = \mathcal{R}^d$, let $\mathcal{H}_S = \{\vec{x} \mapsto \text{sign}(\langle \vec{w}, \vec{x} \rangle) : \vec{w} \in \mathcal{R}^d\}$, where the inner product $\langle \vec{w}, \vec{x} \rangle = \vec{w} \cdot \vec{x} = \sum_{i=1}^d w_i x_i$.

$$VCdim(\mathcal{H}_S) = d$$

Proof-1.

Show that the across zero half space $VCdim(\mathcal{H}_S^0) \geq d$

- Consider the points $\{e_1, \dots, e_d\}$, where $e_i = (0, \dots, 1, \dots, 0)$.



The VC Dimension — Examples (6)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Halfspaces $\mathcal{X} = \mathbb{R}^d$, let $\mathcal{H}_S = \{\vec{x} \mapsto \text{sign}(\langle \vec{w}, \vec{x} \rangle) : \vec{w} \in \mathbb{R}^d\}$, where the inner product $\langle \vec{w}, \vec{x} \rangle = \vec{w} \cdot \vec{x} = \sum_{i=1}^d w_i x_i$.

$$VCdim(\mathcal{H}_S) = d$$

Proof-1.

Show that the across zero half space $VCdim(\mathcal{H}_S^0) \geq d$

- Consider the points $\{e_1, \dots, e_d\}$, where $e_i = (0, \dots, 1, \dots, 0)$.
- Pick $B \subset \{e_1, \dots, e_d\}$, let $h_B = (w_1, \dots, w_d)$, with $w_i = \begin{cases} 1 & \text{if } e_i \in B \\ -1 & \text{if } e_i \notin B \end{cases}$



The VC Dimension — Examples (6)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Halfspaces $\mathcal{X} = \mathcal{R}^d$, let $\mathcal{H}_S = \{\vec{x} \mapsto \text{sign}(\langle \vec{w}, \vec{x} \rangle) : \vec{w} \in \mathcal{R}^d\}$, where the inner product $\langle \vec{w}, \vec{x} \rangle = \vec{w} \cdot \vec{x} = \sum_{i=1}^d w_i x_i$.

$$VCdim(\mathcal{H}_S) = d$$

Proof-1.

Show that the across zero half space $VCdim(\mathcal{H}_S^0) \geq d$

- Consider the points $\{e_1, \dots, e_d\}$, where $e_i = (0, \dots, 1, \dots, 0)$.
- Pick $B \subset \{e_1, \dots, e_d\}$, let $h_B = (w_1, \dots, w_d)$, with $w_i = \begin{cases} 1 & \text{if } e_i \in B \\ -1 & \text{if } e_i \notin B \end{cases}$
- $\forall e_i$, we have $h_B(e_i) = 1$ if $e_i \in B$, and -1 otherwise:

$$h_B(x) = \text{sign}(\langle \vec{w}, \vec{e} \rangle) = w_i$$





The VC Dimension — Examples (6)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Halfspaces $\mathcal{X} = \mathbb{R}^d$, let $\mathcal{H}_S = \{\vec{x} \mapsto \text{sign}(\langle \vec{w}, \vec{x} \rangle) : \vec{w} \in \mathbb{R}^d\}$, where the inner product $\langle \vec{w}, \vec{x} \rangle = \vec{w} \cdot \vec{x} = \sum_{i=1}^d w_i x_i$.

$$VCdim(\mathcal{H}_S) = d$$

Proof-2.

Show that the across zero half space $VCdim(\mathcal{H}_S^0) < d + 1$

- Aim: given any set $A = \{x_1, \dots, x_{n+1}\}$, which can not be shattered by \mathcal{H}_S .



The VC Dimension — Examples (6)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Halfspaces $\mathcal{X} = \mathbb{R}^d$, let $\mathcal{H}_S = \{\vec{x} \mapsto \text{sign}(\langle \vec{w}, \vec{x} \rangle) : \vec{w} \in \mathbb{R}^d\}$, where the inner product $\langle \vec{w}, \vec{x} \rangle = \vec{w} \cdot \vec{x} = \sum_{i=1}^d w_i x_i$.

$$VCdim(\mathcal{H}_S) = d$$

Proof-2.

Show that the across zero half space $VCdim(\mathcal{H}_S^0) < d + 1$

- Aim: given any set $A = \{x_1, \dots, x_{n+1}\}$, which can not be shattered by \mathcal{H}_S .
- From linear algebra, we have $\sum_{i=1}^{d+1} a_i x_i = 0$



The VC Dimension — Examples (6)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Halfspaces $\mathcal{X} = \mathbb{R}^d$, let $\mathcal{H}_S = \{\vec{x} \mapsto \text{sign}(\langle \vec{w}, \vec{x} \rangle) : \vec{w} \in \mathbb{R}^d\}$, where the inner product $\langle \vec{w}, \vec{x} \rangle = \vec{w} \cdot \vec{x} = \sum_{i=1}^d w_i x_i$.

$$VCdim(\mathcal{H}_S) = d$$

Proof-2.

Show that the across zero half space $VCdim(\mathcal{H}_S^0) < d + 1$

- Aim: given any set $A = \{x_1, \dots, x_{n+1}\}$, which can not be shattered by \mathcal{H}_S .
- From linear algebra, we have $\sum_{i=1}^{d+1} a_i x_i = 0$
- We decompose them into positive set P and negative set N , and then have
$$\sum_{i \in P} a_i x_i = \sum_{j \in N} |a_j| x_j$$



The VC Dimension — Examples (6)

No Free Lunch Theorem

The VC Dimension

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

Quiz



Halfspaces $\mathcal{X} = \mathbb{R}^d$, let $\mathcal{H}_S = \{\vec{x} \mapsto \text{sign}(\langle \vec{w}, \vec{x} \rangle) : \vec{w} \in \mathbb{R}^d\}$, where the inner product $\langle \vec{w}, \vec{x} \rangle = \vec{w} \cdot \vec{x} = \sum_{i=1}^d w_i x_i$.

$$VCdim(\mathcal{H}_S) = d$$

Proof-2.

Show that the across zero half space $VCdim(\mathcal{H}_S^0) < d + 1$

- Aim: given any set $A = \{x_1, \dots, x_{n+1}\}$, which can not be shattered by \mathcal{H}_S .
- From linear algebra, we have $\sum_{i=1}^{d+1} a_i x_i = 0$
- We decompose them into positive set P and negative set N , and then have $\sum_{i \in P} a_i x_i = \sum_{j \in N} |a_j| x_j$
- Let $B = \{x_i : i \in P\}$, let us apply \mathcal{H}_B on both sides:

$$\mathcal{H}_B\left(\sum_{i \in P} a_i x_i\right) = \sum_{i \in P} a_i \mathcal{H}_B(x_i) = \mathcal{H}_B\left(\sum_{j \in N} |a_j| x_j\right) = \sum_{j \in N} |a_j| \mathcal{H}_B(x_j)$$

the left hand side is positive, the right hand side is negative. Contradiction!!!





Radon's Lemma

[No Free Lunch Theorem](#)

[The VC Dimension](#)

Is *Infinite Class* PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

Radon's Lemma

[Quiz](#)

For every n , for every $\{x_1, x_2, \dots, x_{n+2}\} \subset \mathcal{R}^n$, there exists $B \subset \{x_1, x_2, \dots, x_{n+2}\}$,



$$CovHull(B) \cap CovHull(\{x_1, x_2, \dots, x_{n+2}\} - B) \neq \emptyset$$



Radon's Lemma

[No Free Lunch Theorem](#)

[The VC Dimension](#)

Is Infinite Class PAC learnable?

The VC Dimension

The VC Dimension — Examples (1)

The VC Dimension — Examples (2)

The VC Dimension — Examples (3)

The VC Dimension — Examples (4)

The VC Dimension — Examples (5)

The VC Dimension — Examples (6)

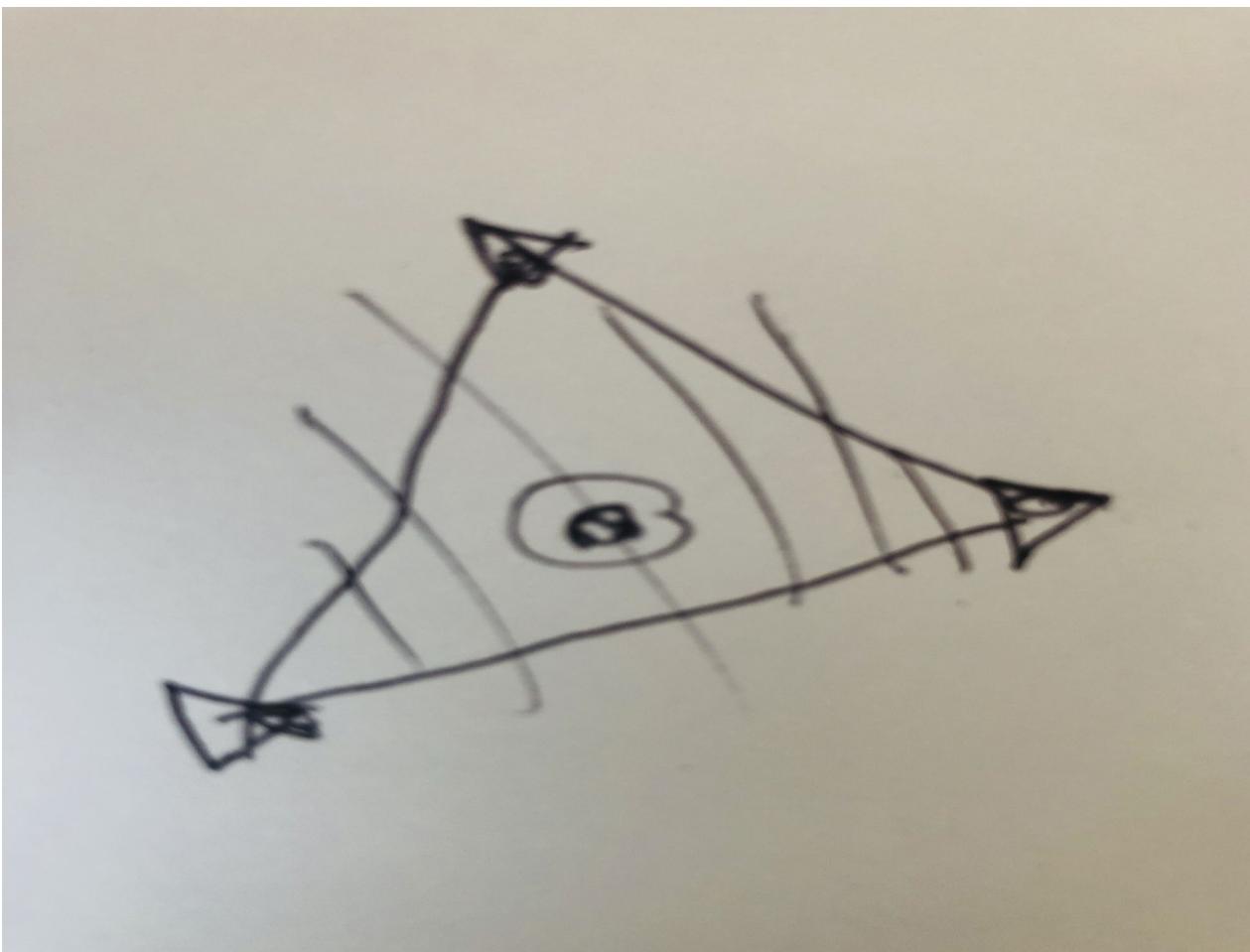
Radon's Lemma

[Quiz](#)

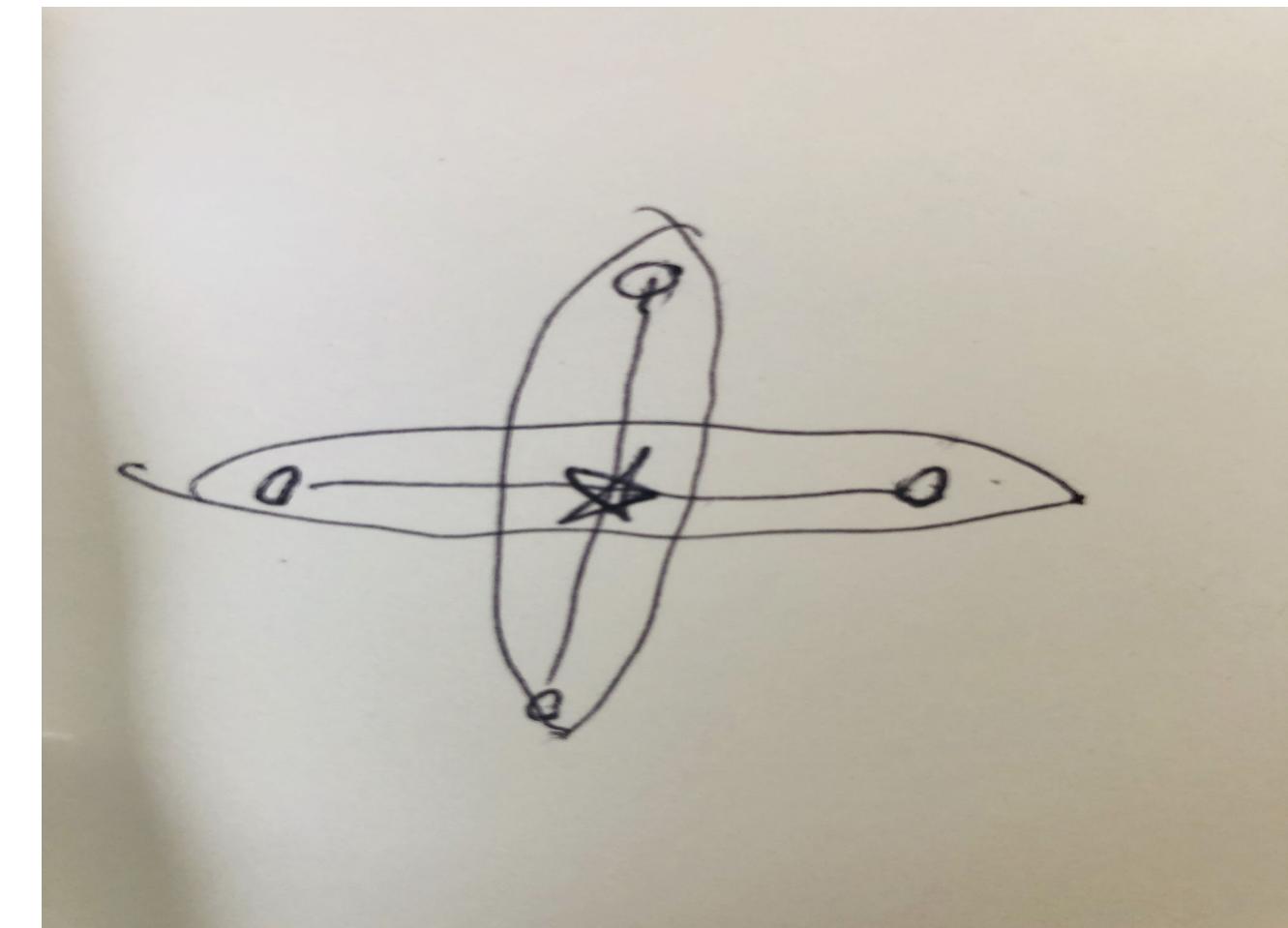


For every n , for every $\{x_1, x_2, \dots, x_{n+2}\} \subset \mathcal{R}^n$, there exists $B \subset \{x_1, x_2, \dots, x_{n+2}\}$,

$CovHull(B) \cap CovHull(\{x_1, x_2, \dots, x_{n+2}\} - B) \neq \emptyset$



(d)



(e)



[No Free Lunch Theorem](#)

[The VC Dimension](#)

[**Quiz**](#)

Quiz



No Free Lunch Theorem

[No Free Lunch Theorem](#)

[The VC Dimension](#)

[Quiz](#)



1. To predict whether a patient is going to suffer a heart attack based on features *blood pressure* (BP), *body-mass index* (BMI), *age* (A), *level of physical activity* (P), and *income* (I). You have to choose between two algorithms: the first picks an axis aligned rectangle in the 2-D space spanned by the features BP and BMI; and the other picks an axis aligned rectangle in the 5-D space spanned by all features.
 - Explain the pros and cons of each choice.
 - Explain how the size of training sample affects your choice.



No Free Lunch Theorem

[No Free Lunch Theorem](#)

[The VC Dimension](#)

[Quiz](#)



1. To predict whether a patient is going to suffer a heart attack based on features *blood pressure* (BP), *body-mass index* (BMI), *age* (A), *level of physical activity* (P), and *income* (I). You have to choose between two algorithms: the first picks an axis aligned rectangle in the 2-D space spanned by the features BP and BMI; and the other picks an axis aligned rectangle in the 5-D space spanned by all features.
 - Explain the pros and cons of each choice.
 - Explain how the size of training sample affects your choice.
2. Prove that if $|\mathcal{X}| \geq km$ for a positive integer $k \geq 2$, then we can replace the lower bound of $1/4$ in the NFL theorem with $\frac{k-1}{2k} = \frac{1}{2} - \frac{1}{2k}$. Namely, let A be a learning algorithm for the task of binary classification. Let m be any number smaller than $|\mathcal{X}|/k$, representing a training sample size. Then, there exists a distribution \mathcal{D} over $\mathcal{X} \times \{0, 1\}$ such that:
 - There exists a function f , which maps \mathcal{X} to $\{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$.
 - $E_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \frac{1}{2} - \frac{1}{2k}$.



The VC Dimension

[No Free Lunch Theorem](#)

[The VC Dimension](#)

[Quiz](#)



1. Show the following monotonicity property of VC-dimension: for every two hypothesis classes if $\mathcal{H}' \subseteq \mathcal{H}$, then $VCDim(\mathcal{H}') \leq VCDim(\mathcal{H})$.



The VC Dimension

[No Free Lunch Theorem](#)

[The VC Dimension](#)

[Quiz](#)



1. Show the following monotonicity property of VC-dimension: for every two hypothesis classes if $\mathcal{H}' \subseteq \mathcal{H}$, then $VCDim(\mathcal{H}') \leq VCDim(\mathcal{H})$.
2. For a finite hypothesis class \mathcal{H} , $VCDim(\mathcal{H}) \leq \lfloor \log(|\mathcal{H}|) \rfloor$. However, this is just an upper bound. The VC-dimension of a class can be much lower than that:
 - Find an example of a class \mathcal{H} of functions over the real interval $\mathbb{X} = [0, 1]$, such that \mathcal{H} is finite while $VCDim(\mathcal{H}) = 1$.
 - Give an example of a finite hypothesis class \mathcal{H} over the domain $\mathbb{X} = [0, 1]$, where $VCDim(\mathcal{H}) = \lfloor \log(|\mathcal{H}|) \rfloor$.



Questions?

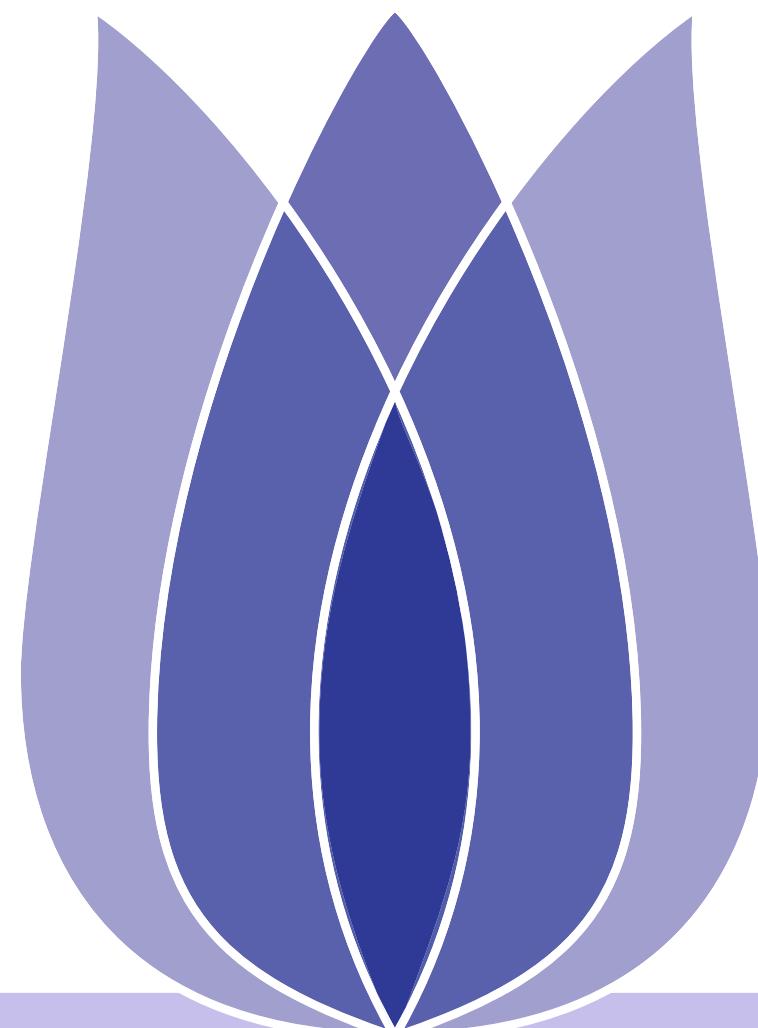
[No Free Lunch Theorem](#)

[The VC Dimension](#)

[Quiz](#)



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](#)