CS5339 Machine Learning

Estimation I

Lee Wee Sun
School of Computing
National University of Singapore
leews@comp.nus.edu.sg

Semester 2, 2019/20

Estimation

In this part of the course, we will study how much data is required to learn.

Outline

Finite Class

- We start with the simpler case of *empirical risk minimization* on a finite hypothesis class $\mathcal{H}$ in the *realizable* case.
  - By realizable, we mean that there exists a hypothesis $h^* \in \mathcal{H}$ with zero expected error, $L_{(D,f)}(h^*) = 0$, where f gives the target labeling.
  - For empirical risk minimization, we will be minimizing the training set error

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m},$$

or equivalently the 0-1 loss. We denote the hypothesis that minimizes the empirical risk as

$$h_S \in \arg\min_{h \in \mathcal{H}} L_S(h).$$

- We assume that the training set S is selected i.i.d. from a distribution $\mathcal{D}$. Hence $S \sim \mathcal{D}^m$ where m is the sample size, and $\mathcal{D}^m$ denotes the probability over m-tuples induced by applying $\mathcal{D}$ to pick each element of the tuple independently.
- $L_{(D,f)}(h_S)$ depends on the training set S which is randomly selected, hence it is a random variable.

Output of ~~the~~ $L_{Df}(h_S)$ depends on which random S ~~the~~ is sampled

- There is a probability that a "bad" sample S is selected such that $L_{(D,f)}(h_S)$ is larger than a desired value ϵ.

Desired have small ϵ

  - We denote the probability of getting a bad sample δ and call it the *confidence parameter*.
  - We call the desired accuracy ϵ the *accuracy parameter*.

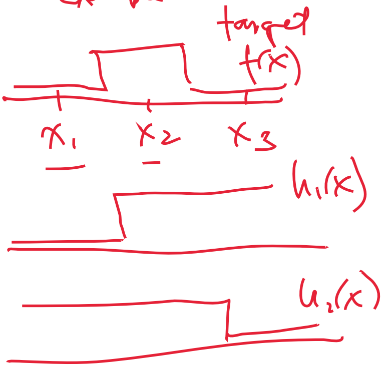
- Let $S_{|x} = (x_1, \ldots, x_m)$ be instances in the training set. We would like to upper bound

$$\mathcal{D}^m(\{S_{|x} : L_{(D,f)}(h_S) > \epsilon\}).$$

- Let $\mathcal{H}_B$ be the set of "bad" hypothesis:

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{(D,f)}(h) > \epsilon\}.$$

Example :



Let x_1, x_2, x_3 be equally likely

Let $\mathcal{H} = \{f, h_1, h_2\}$

For $\epsilon = 1/6$

h_1 is bad because expected err $\geq 1/3$

h_2 is also bad err $\geq 1/3$

- Let M be the misleading set of samples,

$$M = \{S_{|x} : \exists h \in \mathcal{H}_B, L_S(h) = 0\},$$

that is, for every $S_{|x} \in M$, there is a bad hypothesis that looks good on $S_{|x}$.

In example

$S_1 = \{(x_1, 0), (x_2, 1)\}$

is misleading because

h_1 is bad but

has zero err on S_1

- With empirical risk minimization $L_S(h_S) = 0$: h_S must have been selected from among h with zero empirical error.
  - Sufficient to bound the probability of M.
- We can rewrite M as

$$M = \cup_{h \in \mathcal{H}_b} \{S_{|x} : L_S(h) = 0\}, \text{ hence}$$

$$\begin{aligned}\mathcal{D}^m(\{S_{|x} : L_{(D,f)}(h_S) > \epsilon\}) &\leq \mathcal{D}^m(M) \\ &= \mathcal{D}^m(\cup_{h \in \mathcal{H}_b}\{S_{|x} : L_S(h) = 0\})\end{aligned}$$

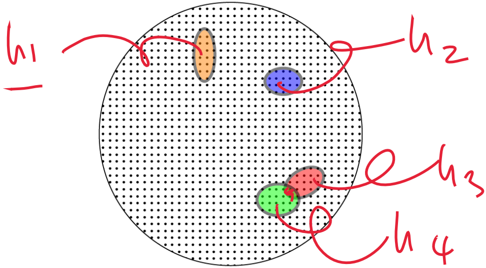


Figure: From SSBD. Each point represents a m-tuple of instances. Each oval represents a set of misleading m-tuples for a bad hypothesis. The total probability of the misleading n-tuples is bounded using the union bound.

- Applying the union bound, we get

$$\mathcal{D}^m(\{S_{|x} : L_{(D,f)}(h_S) > \epsilon\}) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S_{|x} : L_S(h) = 0\}).$$

- As the training set is i.i.d.

$$\mathcal{D}^m(\{S_{|x} : L_S(h) = 0\}) = \mathcal{D}^m(\{S_{|x} : \forall i, h(x_i) = f(x_i)\})$$
$$= \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\}).$$

- For each individual sampling of x_i, we have

$$\mathcal{D}(\{x_i : h(x_i) = f(x_i)\}) = 1 - L_{(D,f)}(h) \leq 1 - \epsilon.$$

- Using $1 - \epsilon \leq e^{-\epsilon}$,

$$\mathcal{D}^m(\{S_{|x} : L_S(h) = 0\}) \leq (1-\epsilon)^m \leq e^{-\epsilon m}.$$

- Combining with the union bound, we get

$$\mathcal{D}^m(\{S_{|x} : L_{(D,f)}(h_S) > \epsilon\}) \leq |\mathcal{H}_B| e^{-\epsilon m} \leq |\mathcal{H}| e^{-\epsilon m}.$$

Set prob $\leq \delta$

$$|\mathcal{H}| e^{-\epsilon m} \leq \delta$$

$$\log|\mathcal{H}| - \epsilon m \leq \log \delta$$

$$m \geq \frac{1}{\epsilon}\left(\log|\mathcal{H}| + \log 1/\delta\right)$$

Setting the right hand side to δ, we have proven the following:

Theorem: (SSBD Corollary 2.3) Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$ and $\epsilon > 0$, and let m be an integer that satisfies

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}.$$

Then, for any labeling function f, and for any distribution $\mathcal{D}$ for which the realizability assumption holds, with probability at least $1-\delta$ over the choice of an i.i.d. sample S of size m, we have that for every empirical risk minimization hypothesis h_S, it holds that

$$L_{(D,f)}(h_S) \leq \epsilon.$$

Exercise 1:

Consider learning a finite hypothesis class. Which of the following requires a larger sample size?

A. Halving the accuracy parameter from ϵ to $\epsilon/2$.

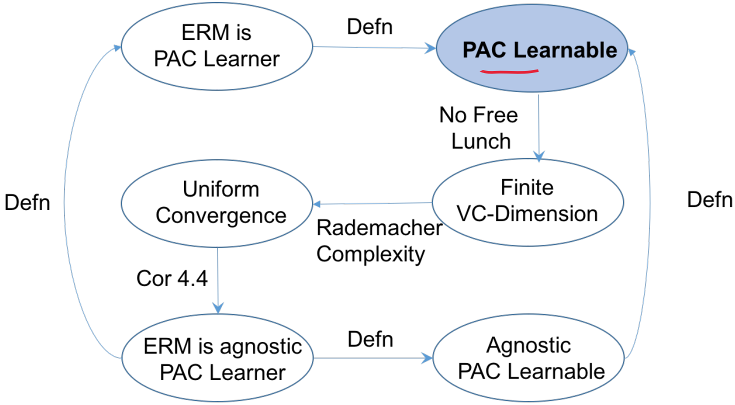
B. Doubling the number of hypotheses in the function class.

A. Double sample size

B. $\frac{m'}{m} = \frac{\log 2|H|/\delta}{\log |H|/\delta} = \left(\frac{\log 2}{\log \frac{|H|}{\delta}} + 1\right)$

If $\frac{|H|}{\delta} = 1000$, $\frac{\log 2}{\log \frac{|H|}{\delta}} \approx 0.1$ times the sample size

Outline



Fundamental Theorem: These are equivalent

PAC Learning

Definition (PAC Learnability): (SSBD Defn 3.1) A hypothesis class $\mathcal{H}$ is Probably Approximately Correct (PAC) learnable if there exists a function $m_{\mathcal{H}}(\epsilon, \delta)$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0, 1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function f, if the realizability assumption holds with respect to $\mathcal{H}$, $\mathcal{D}$, f, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability at least $1 - \delta$ (over the choice of examples), $L_{(D,f)}(h) \leq \epsilon$.

- The accuracy parameter ϵ determines how far the classifier is allowed to be from optimal (*approximately correct*).
- The confidence parameter δ indicates how likely the classifier is to meet the accuracy requirement (*probably* part of PAC).

Sample Complexity

The function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{R}$ determines the sample complexity of learning $\mathcal{H}$.

Corollary: Every finite hypothesis class is PAC learnable with sample complexity

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

Strong req: has to work for all
ϵ, δ, D, realizable f

General Loss Functions

The definition of PAC learning is too restrictive in practice. We relax it in the following ways:

- **Realizability assumption:** Real data is often noisy. We generalize the data generating distribution $\mathcal{D}$ to a distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, i.e. a joint distribution over the domain points and the labels.
  - For binary classification, given any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$, the best predictor is called the *Bayes optimal predictor*:

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } \Pr[y = 1|x] \geq 1/2 \\ 0 & \text{otherwise} \end{cases}.$$

Assuming that $\mathcal{H}$ contains the Bayes optimal predictor, and trying to learn the predictor is sometimes a reasonable alternative to assuming realizability.

- **Agnostic learning:** Assuming that $\mathcal{H}$ contains the Bayes optimal predictor is not always reasonable.
  - An alternative is to ask the learning algorithm to produce a predictor whose error is not much larger than the error of the best predictor in a benchmark class $\mathcal{H}$.
  - The predictor will do well if the benchmark class contains a good approximator of the Bayes optimal predictor.
  - This is sometimes called agnostic learning.

- **Loss function:** We would like to go beyond binary classification to other learning problems such as multiclass classification, regression, and even unsupervised learning. To do that we use a loss function in learning.
  - Given a set $\mathcal{H}$ of hypotheses of models, and a domain $\mathcal{Z}$ ($\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ for supervised learning), let ℓ be a function from $\mathcal{H} \times \mathcal{Z}$ to non-negative real numbers $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$. We call such a function a **loss function**.
    - The **0-1 loss** measures the misclassification error in classification

$$\ell_{0-1}(h, (x, y)) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y. \end{cases}$$

    - The **square loss** is commonly used for regression

$$\ell_{sq}(h, (x, y)) = (h(x) - y)^2.$$

- Continued ...
  - The **risk function** is the expected loss of the hypothesis,

    $$L_{\mathcal{D}}(h) = E_{z\sim\mathcal{D}}[\ell(h, z)].$$

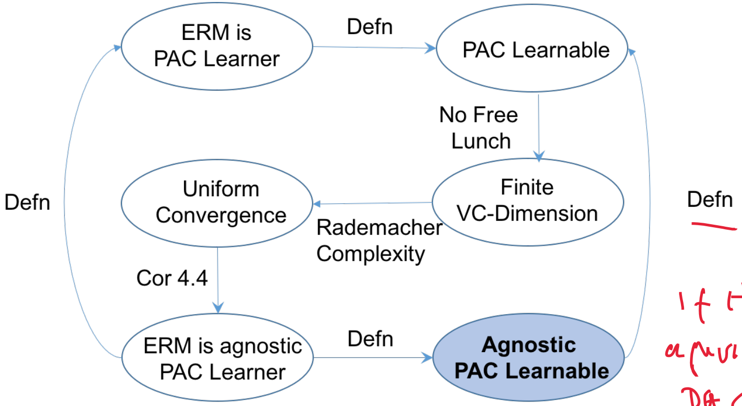
    We are interested in finding a hypothesis h that has small risk, or expected loss.
  - We also define the empirical risk as

    $$L_S(h) = \sum_{i=1}^{m} \ell(h, z_i).$$

Definition (Agnostic PAC Learnability for General Loss Functions): (SSBD Defn 3.4) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set $\mathcal{Z}$ and a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$, if there exists a function $m(\epsilon, \delta)$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0, 1)$, for every distribution $\mathcal{D}$ over $\mathcal{Z}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns $h \in \mathcal{H}$ such that, with probability at least $1 - \delta$ (over the choice of the m training examples),

$$L_{\mathcal{D}}(h) \leq \underbrace{\min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')}_{= 0 \text{ for PAC learning}} + \epsilon, \quad ✓ ✓$$

where $L_{\mathcal{D}}(h) = E_{z \sim \mathcal{D}}[\ell(h, z)]$.



Fundamental Theorem: These are equivalent

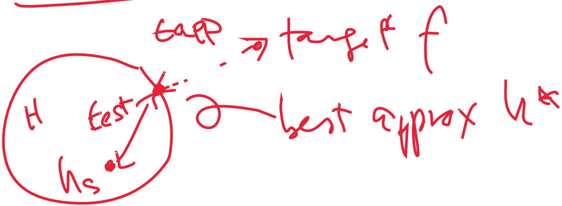
If H is agnostically PAC learnable then H is PAC learnable

Error Decomposition

- Let h_S be a $ERM_{\mathcal{H}}$ hypothesis. By using agnostic learning, we can decompose the error into two components:

$$L_{\mathcal{D}}(h_S) \leq \epsilon_{app} + \epsilon_{est},$$

- The approximation error $\epsilon_{app} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ is the minimum risk achievable by hypotheses in the class.
- The estimation error $\epsilon_{est} \geq L_{\mathcal{D}}(h_S) - \epsilon_{app}$ is an upper bound on the difference between the error achieved by the $ERM_{\mathcal{H}}$ predictor and the minimum risk achievable by hypotheses in the class.



- By choosing a rich class $\mathcal{H}$, we can often reduce the approximation error.
- However, a richer class often has higher estimation error and can lead to *overfitting*.
- Choosing $\mathcal{H}$ too small, on the other hand, may lead to *underfitting*.

Exercise 2:

The function class $\mathcal{H}$ is known to be agnostically PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$. Assume that A is the agnostic learning algorithm. Which of the following is false? Modify the statement to make it correct.

A. A achieves expected loss of no more than ϵ when $y = h(x)$ for some $h \in \mathcal{H}$. True

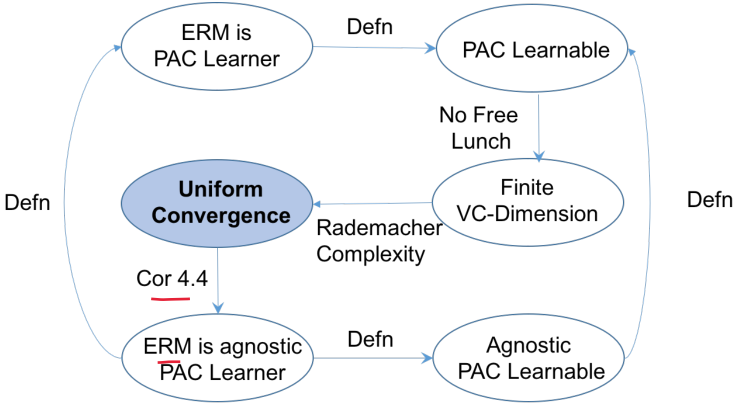
B. A does not need to know the distribution $\mathcal{D}_x$ of x and works for every distribution. True

C. A achieves the Bayes error. False if Bayes opt classifier $\in$ H

D. A does not require the target function to be in $\mathcal{H}$. True

E. A can be used even when y is a random variable drawn from $p(y|x)$. True

Outline



Fundamental Theorem: These are equivalent

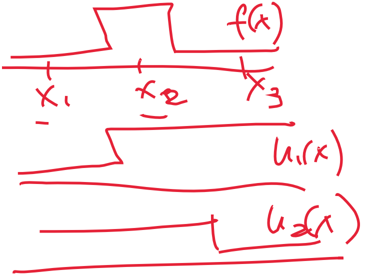
Uniform convergence is a general tool for showing learnability, including agnostic learning.

Definition (ϵ-representative sample): (SSBD Defn 4.1) A training sample S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

Examples

$H = \{f, h_1, h_2\}$



For $\epsilon = 1/6$, 0-1 loss

$S_1 = \{(x_1, 0), (x_2, 1)\}$ is not ϵ-rep.

$|L_S(h_1) - L_D(h_1)| = |0 - 1/3| = 1/3$ ✗

$|L_S(h_2) - L_D(h_2)| = |1/2 - 1/3| = 1/6$

$|L_S(f) - L_D(f)| = 0$

Definition (Uniform Convergence): (SSBD Defn 4.3) We say that a hypothesis class $\mathcal{H}$ has the uniform convergence property (w.r.t. domain Z and loss function ℓ) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ examples drawn i.i.d. from $\mathcal{D}$, then, with probability at least $1 - \delta$, S is ϵ-representative.

Uniform refers to having a fixed sample size for all $h \in \mathcal{H}$ and all distributions.

Lemma: (SSBD Lemma 4.2) Assume that training set S is $\epsilon/2$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$). Then any output of $ERM_{\mathcal{H}}(S)$ (empirical risk minimizer), namely, any $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$, satisfies

$$L_{\mathcal{D}}(h_S) \le \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

– agnostic learning

Proof:

For every $h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h_S) \overset{\text{rep}}{\le} L_S(h_S) + \epsilon/2 \overset{\text{ERM}}{\le} L_S(h) + \epsilon/2 \overset{\text{rep}}{\le} L_{\mathcal{D}}(h) + \epsilon/2 + \epsilon/2 = L_{\mathcal{D}}(h) + \epsilon,$$

where the first and third inequalities are because S is $\epsilon/2$-representative and the second inequality holds because h_S is an ERM predictor. $\square$

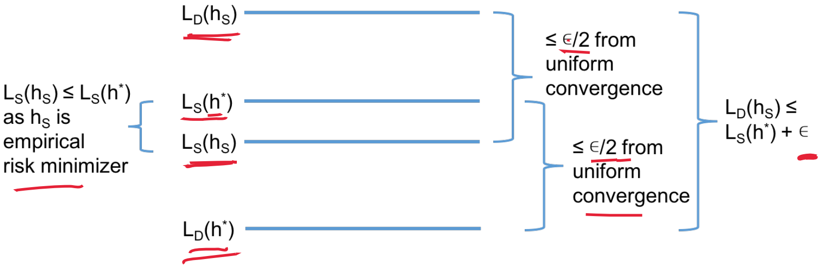
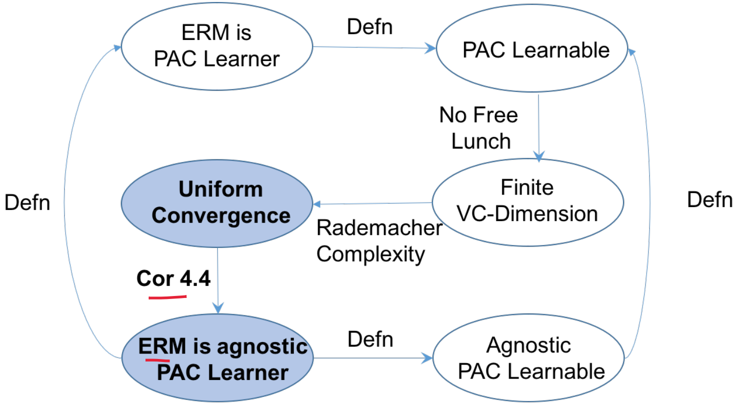


Figure: Illustration of how uniform convergence plus empirical risk minimization implies agnostic learning.

Corollary: (SSBD Corollary 4.4) If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$. Furthermore, in that case, the $ERM_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.



Fundamental Theorem: These are equivalent

Finite Classes are Agnostic PAC Learnable

Recall **Hoeffding's Inequality**:
Let $Z_1, \dots, Z_m$ be a sequence of i.i.d. random variables and let $\bar{Z} = \frac{1}{m}\sum_{i=1}^m Z_i$. Assume that $E[\bar{Z}] = \mu$ and $\Pr[a \leq Z_i \leq b] = 1$ for every i. Then for any $\epsilon > 0$,

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon\right] \leq 2\exp(-2m\epsilon^2/(b-a)^2).$$

Letting $Z_i = \ell(h, z_i)$, we get $L_S(h) = \sum_{i=1}^m Z_i$ and $L_{\mathcal{D}}(h) = \mu$.

Assuming that the range of ℓ is $[0,1]$ and applying the union bound, we get

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} 2\exp(-2m\epsilon^2)$$
$$= \underbrace{2|\mathcal{H}|}_{\text{size of } |H| \text{ from union bound}} \underbrace{\exp(-2m\epsilon^2)}_{\text{Hoeffding}}.$$

Set $2|H| \exp(-2m\epsilon^2) = \delta$

$\log 2|H| - 2m\epsilon^2 = \log \delta$

$2m\epsilon^2 = \log 2|H|/\delta$

$m = \dfrac{\log 2|H|/\delta}{2\epsilon^2}$

Solving for m so that the right hand side is no more than δ, we get the following.

Corollary: (SSBD Corollary 4.6) Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and let $\ell : \mathcal{H} \times Z \to [0, 1]$ be a loss function. Then $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

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

Furthermore, the class is agnostically PAC learnable using ERM with sample complexity

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

Discretization:

- If we use a 64-bit computer, each parameter can take at most 2^{64} possible values.
- With d parameters, the hypothesis class size is at most 2^{64d}.
- Applying the corollary, the sample complexity is bounded by

$$\frac{128d + 2\log(2/\delta)}{\epsilon^2}.$$

Exercise 3:

Which of the following statements is false? Modify it so that it is correct.

A. A hypothesis class $\mathcal{H}$ satisfies the uniform convergence property if the training error is close to the expected error for ~~some~~ all $h \in \mathcal{H}$ when the training set has size at least $m_{\mathcal{H}}^{UC}(\epsilon, \delta)$. False

B. $\mathcal{H}$ is PAC learnable if it satisfies the uniform convergence property. True

C. $\mathcal{H}$ is agnostically learnable if it satisfies the uniform convergence property. True

Exercise 4:

From the bounds we have so far, which of these requires a smaller number of samples as ϵ goes to 0.

A. PAC learning. $O(1/\epsilon)$

B. Agnostic PAC learning. $O(1/\epsilon^2)$

For PAC learning, variance reduces as we get close to 0 error, e.g. Bernoulli r.v.

$var = p(1-p)$

For agnostic, p may be 0.5

var may not reduce.

Exercise 5:

In this experiment, we look at the effect of the number of functions tested on the selecting the best function using a validation set. We use the digits dataset with Gaussian SVM. We test over different values of the variance parameter γ. We test 4, 8, and 12 values in sets 0, 1, and 2 respectively.

Does increasing the number of functions tested increase the probability of selecting a suboptimal choice? Are the results in the experiment consistent with what theory suggests?

- sample size to achieve ϵ increases logarithmically with $|H|$
- okay as long as $|H|$ is not too large.

Outline

Theorem (No-Free-Lunch): (SSBD Theorem 5.1) Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

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

Implication: For every learner, there exists a task on which it fails.

How might another learner learn a task where another algorithm A fails in?

- The no-free-lunch theorem says that learning is impossible without some form of *prior knowledge*.
- A learner can learn if it has prior knowledge. In an extreme case, the learner knows enough to use hypothesis class $\mathcal{H} = \{f\}$ where f is the target to be learned.
- More realistically, the learner may know that the target f belongs to some "small" hypothesis class $\mathcal{H}$, e.g. a finite $\mathcal{H}$. This is often called *inductive bias*.
- By using a richer hypothesis class, we can often increase the chance that we have a hypothesis that does well. However, this often comes at the cost of increased estimation error – in the worst case, an unconstrained hypothesis class cannot be learned.

Proof Sketch (No-Free-Lunch):

- Let C be a subset of $\mathcal{X}$ of size $2m$.
- We consider all $T = 2^{2m}$ possible functions from C to $\{0, 1\}$ denoted $f_1, \ldots, f_T$.

Example: $m = 1$ $|C| = 2$ $C = \{x_1, x_2\}$

$T = 2^2 = 4$ possible fun

f_1 : $f_1(x_1) = 0$, $f_1(x_2) = 0$

f_2 : $f_2(x_1) = 0$ $f_2(x_2) = 1$

f_3 : $f_3(x_1) = 1$, $f_3(x_2) = 0$

f_4 : $f_4(x_1) = 1$, $f_4(x_2) = 1$

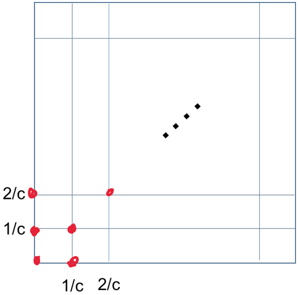
- For each possible target function f_i, we set the distribution of x to be uniform on C and labels to be $f_i(x)$.
- The intuition is that observing the training set (no more than half of C) tells us nothing about the labels of the unobserved instances since all functions are possible.

- Hence the expected error over C is at least $1/4$ (correct on observed instances, $1/2$ on each unobserved instance).
- According to Markov's inequality, $P(Z \geq a) \leq E[Z]/a$ for a non-negative random variable Z.
- Applying that, we have $P((1 - L_D(A(S)) \geq 7/8) \leq (3/4)/(7/8) = 6/7$. Hence, with probability at least $1/7$, $L_D(A(S)) \geq 1/8$.

□

We saw the following result when we discussed the curse of dimensionality and how it affects the nearest neighbour algorithm.

Theorem: (SSBE Theorem 19.4) For any $c > 1$, and every learning rule, L, there exists a distribution over $[0,1]^d \times \{0,1\}$, such that $p(y|x)$ is c-Lipschitz, the Bayes error of the distribution is 0, but for sample sizes $m \leq (c+1)^d/2$, the true error of the rule L is greater than $1/4$.



$(0, \frac{1}{c})$
$(\frac{1}{c}, \frac{2}{c})$ . . .
. . .
$(\frac{a_1}{c}, \frac{a_2}{c})$

Figure: 2D grid for construction of Lipschitz functions.

Proof Sketch:

- Fix any values of c and d.
- Let G_c^d be the grid on $[0, 1]^d$ with distance $1/c$ between points on the grid:
  - Each point is of the form $(a_1/c, \ldots, a_d/c)$ where a_i is in $\{0, \ldots, c-1, c\}$.

- Any two points on the grid is at least $1/c$ apart.
- Any function $p(y|x) : G_c^d \mapsto [0, 1]$ is a c-Lipschitz function.
- Hence, the set of c-Lipschitz function contain all binary functions over G_c^d.
- The number of grid points is $(c+1)^d$.
- Using the same ideas as in the proof of SSBD Theorem 5.1, if $m < (c+1)^d/2$, it is not possible to predict the labels on the unseen examples.
- Hence there is a target where the true error is greater than $1/4$. $\square$

Exercise 6:
Assume that $\mathcal{X}$ is finite. Then for any sample size m, with probability at least $1/7$, the expected loss of any algorithm is at least $1/8$. True or False, and why?

False. X is finite. Total # of fns $\le 2^{|X|}$

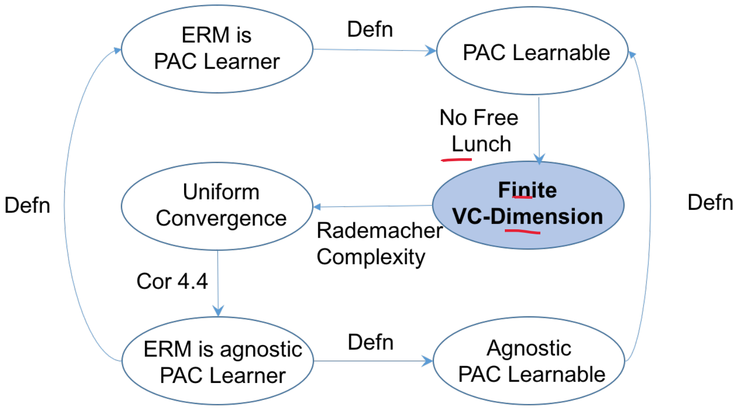
PAC learning shows

$$\epsilon \le \frac{\log |H| / \delta}{m}$$

In proof of no free lunch

$$|X| \ge 2m$$

Outline



Fundamental Theorem: These are equivalent

VC-Dimension and Infinite Function Classes

It turns out that some infinite function classes are also PAC learnable. For binary classification, learnability is characterized by the VC-dimension: finite VC-dimension is a necessary and sufficient condition for PAC learnability.

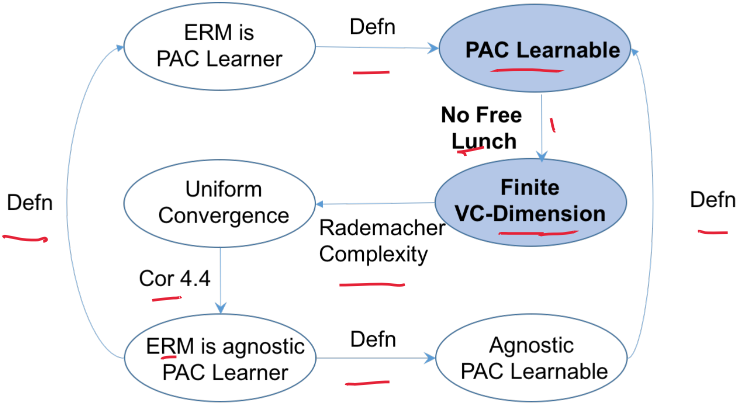
Recall the following:

- **Shattering:** A hypothesis class $\mathcal{H}$ *shatters* a set $C \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{-1, 1\}$, i.e. $|\mathcal{H}_C| = 2^{|C|}$.
- **VC-dimension:** The VC-dimension of hypothesis class $\mathcal{H}$ is the size of the largest set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$.

The following is a corollary of the no-free-lunch theorem.

Corollary: (SSBD Corollary 6.4) Let C be a hypothesis class from $\mathcal{X}$ to $\{0,1\}$. Let m be the training set size. Assume that there exists a set $C \subset \mathcal{X}$ of size $2m$ that is shattered by $\mathcal{H}$. Then for any learning algorithm A, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a predictor $h \in \mathcal{H}$ such that $L_D(h) = 0$ but with probability at least $1/7$, over the choices of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.

Consequently, if a class $\mathcal{H}$ has infinite VC-dimension, it is not PAC learnable.



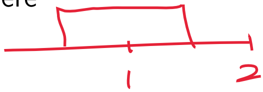
Fundamental Theorem: These are equivalent

More on the VC-dimension

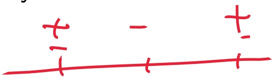
To show that the VC-dim$(\mathcal{H}) = d$, we need to show

1. There exists a set C of size d that is shattered by $\mathcal{H}$.
2. Every set of size $d+1$ is not shattered by $\mathcal{H}$.

Intervals: Let $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ where $h_{a,b}(x) = \mathbb{1}_{x \in [a,b]}$.



- $C = \{1, 2\}$ is shattered.
- Consider any set $\{c_1, c_2, c_3\}$ where $c_1 \leq c_2 \leq c_3$. Then the labeling $(1, 0, 1)$ cannot be obtained by any interval.
- Therefore, VC-dim$(\mathcal{H}) = 2$.



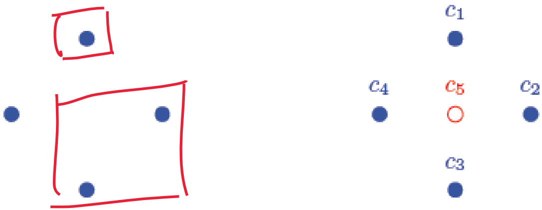


Figure: From SSBD Fig 6.1. Shattered set on the left. No axis aligned rectangle can classify c_5 as 0 while classifying the rest of the points as 1.

Axis Aligned Rectangles: Let $\mathcal{H} = \{h_{a_1,a_2,b_1,b_2} : a_1 \leq a_2 \text{ and } b_1 \leq b_2\}$ where

$$h_{a_1,a_2,b_1,b_2}(x_1,x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq a_2 \text{ and } b_1 \leq x_2 \leq b_2 \\ 0 & \text{otherwise.} \end{cases}$$

VC-dim($\mathcal{H}$) $= 4$ for axis aligned rectangles.

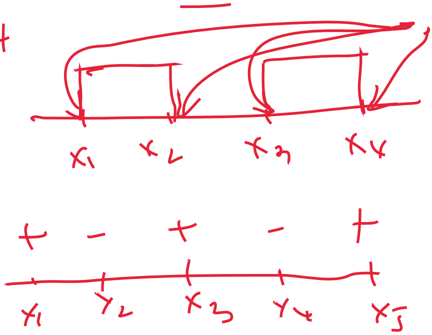
- The figure on the left shows a set of 4 points that is shattered.
- For any 5 points, select the leftmost, rightmost, topmost, bottommmost points. Label them as 1 and label the remaining point (which must be in the interior of the rectangle) with 0. This labeling cannot be represented using a rectangle. The figure on the right gives an example.
    - Must be true for *any* 5 points, not just the one shown in the figure.

Exercise 7 :

Intervals: Let $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ where $h_{a,b}(x) = \mathbb{1}_{x \in [a,b]}$.

What is the VC-dimension of the union of two intervals (disjunction of the indicator functions of two intervals)?

VC Dim = 4



shatter by perturbing intervals slightly to give all possible classifications

Not possible using 2 intervals

Fundamental Theorem of PAC Learning

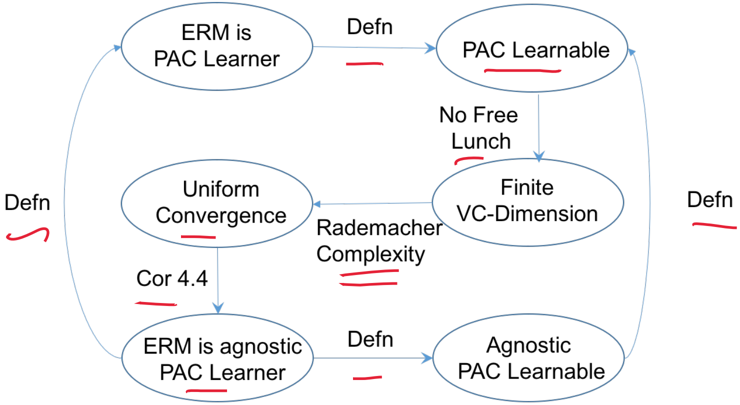
Theorem (Fundamental Theorem): (SSBD Theorem 6.7) Let $\mathcal{H}$ be hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function be the $0 - 1$ loss. Then the following are equivalent:

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 finite VC-dimension.

Proof Sketch:

- $1 \rightarrow 2$ was shown earlier (SSBD Corollary 4.4).
- $2 \rightarrow 3$, $3 \rightarrow 4$, $2 \rightarrow 5$, and $5 \rightarrow 4$ are immediate from definitions.
- $4 \rightarrow 6$ and $5 \rightarrow 6$ comes from the no-free-lunch theorem.
- We will show $6 \rightarrow 1$ using Rademacher complexity later.

$\square$

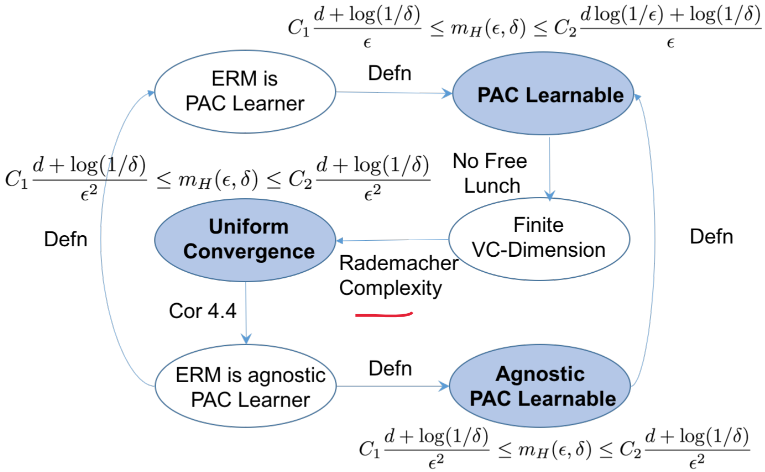


Fundamental Theorem: These are equivalent

It is possible to get more refined bounds in terms of the VC-dimension.

Theorem: (SSBD Theorem 6.8 The fundamental theorem of statistical learning - quantitative version)
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 VC-dim$(\mathcal{H}) = d < \infty$. Then there are absolute constants C_1, C_2 such that:

- $\mathcal{H}$ is PAC learnable with sample complexity
$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}.$$
- $\mathcal{H}$ is agnostic PAC learnable with sample complexity
$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}.$$
- $\mathcal{H}$ has uniform convergence property with sample complexity
$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m^{UC}_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2}.$$



Fundamental Theorem: Quantitative Bounds

Rademacher Complexity

With infinite function classes, using the union bound over all functions will no longer give a finite bound.

We will look at Rademacher complexity, which can be used together with VC-dimension, or other assumptions, to bound the sample complexity of infinite classes.

- To simplify notation, we will compose our hypothesis class with the loss function. Denote

$$\mathcal{F} = \ell \circ \mathcal{H} = \{z \to \ell(h, z) : h \in \mathcal{H}\}.$$

We will use the empirical and expected losses of $f \in \mathcal{F}$

$$L_D(f) = E_{z \sim \mathcal{D}}[f(z)], \qquad L_S(f) = \frac{1}{m}\sum_{i=1}^{m} f(z_i).$$

- Recall that we used the notion of *representativeness* when we studied uniform convergence: we have uniform convergence if the samples are ϵ-representative with high probability for all distributions.
- For this section, it suffices to look at one-sided *representativeness* of S with respect of $\mathcal{F}$ as

$$\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f)).$$

When we discuss uniform convergence
we used 2-sided version

$$\sup_{f \in \mathcal{F}} |L_D(f) - L_S(f)|$$

  - If S has good representativeness (small), functions with small empirical risk will also have small expected risk.
- We do not know $\mathcal{D}$ and would like to estimate or bound the representativeness error from data.

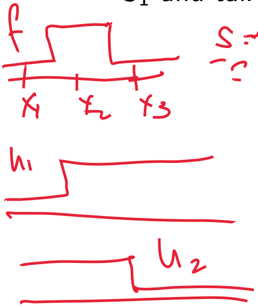
Recall example

$H = \{f, h_1, h_2\}$

- Given S, one possibility of estimating its representativeness is by randomly splitting it into disjoint sets S_1 and S_2 and measuring

$$\frac{1}{m}\sup_{f\in\mathcal{F}}(m_1 L_{S_1}(f) - m_2 L_{S_2}(f))$$

where m_1 and m_2 are the sizes of S_1 and S_2.
- Rademacher complexity averages the estimates over the random splits generated from m coin tosses: heads goes into S_1 and tail into S_2.



$S = \cancel{\{x_1, x_2\}}$

$= \{(x_1,0),(x_2,1)\}$

Let $S_1 = \{(x_1,0)\}$, $S_2 = \{(x_2,1)\}$

Estimate representativeness

$\frac{1}{2}\sup_{h\in H}(L_{S_1}(h) - L_{S_2}(h)) = \frac{1}{2}(1-0) = \frac{1}{2}$

Other splits $S_1 = \{(x_2,1)\}$, $S_2 = \{(x_1,1)\}$: 0

$S_1 = \{(x_1,0),(x_2,1)\}$, $S_2 = \{\}$: $\frac{1}{2}$

$S_1 = \{\}$, $S_2 = \{(x_1,0),(x_2,1)\}$: 0

Rademacher comp : $\frac{1}{4}\left(\frac{1}{2} + 0 + \frac{1}{2} + 0\right) = \frac{1}{4}$

- Let $\mathcal{F} \circ S$ be the set of all possible evaluation of functions $f \in \mathcal{F}$ on S

$$\mathcal{F} \circ S = \{(f(z_1), \ldots, f(z_m)) : f \in \mathcal{F}\}.$$

The Rademacher complexity of $\mathcal{F}$ with respect to S is:

$$R(\mathcal{F} \circ S) = E_{\boldsymbol{\sigma} \sim \{\pm 1\}^m} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right],$$

where σ_i are i.i.d. sampled with $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$. We can also define the Rademacher complexity of a set $A \subset \mathbb{R}^m$ as

$$R(A) = E_{\boldsymbol{\sigma}} \left[\sup_{a \in A} \frac{1}{m} \sum_{i=1}^{m} \sigma_i a_i \right],$$

- Rademacher complexity is a measure of the maximum correlation of the functions with random binary (± 1) sequences.
  - Consider a $\{-1, 1\}$-valued function f with $\frac{1}{m}\sum_{i=1}^{m} \sigma_i f(z_i) = c$. If f agrees entirely with $\boldsymbol{\sigma}$, then $c = 1$.
    - The value c can be related to classification accuracy of f when σ is the label: the accuracy is $(c+1)/2$.
    - Since $\boldsymbol{\sigma}$ is randomly selected, we are effectively asking how well the function class is able to fit (agree with) noise as labels.
  - Consider class $\mathcal{F}$ of $\{-1, 1\}$-valued functions. If the set S of points is shattered by $\mathcal{F}$, then we can always find a function that agrees entirely with any $\boldsymbol{\sigma}$, hence $R(\mathcal{F} \circ S) = 1$.
  - If we only have one function in $\mathcal{F}$, the Rademacher complexity is 0.
    - Rademacher complexity is always greater than or equal to 0, but may be larger than 1 for real-valued, rather than binary functions.

$E[\sigma_i f(z_i)]$
$= E(\sigma_i) f(z_i)$

Exercise 8

In this experiment, we will estimate the Rademacher complexity of linear SVM, Gaussian SVM and decision trees. We will use randomly generated 1000 20-dimensional binary vectors as the input set. The parameter C is set to 1 for both linear and Gaussian SVM, and the parameter *gamma* is set to 1 in Gaussian SVM.

Before running your experiment, predict roughly what the estimated Rademacher complexities of the three classifier classes would be.

Rademacher complexity has various useful properties.

Lemma: (SSBD Lemma 26.6) For any $A \subset \mathbb{R}^m$, scalar $c \in \mathbb{R}$, and vector $\mathbf{b} \in \mathbb{R}^m$ we have

$$R(\{c\mathbf{a} + \mathbf{b} : \mathbf{a} \in A\}) = |c|R(A).$$

Proof:

- Let $A' = \{c\mathbf{a} + \mathbf{b} : \mathbf{a} \in A\}$. Then

$$
\begin{aligned}
R(A') &= E_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{a}\in A} \frac{1}{m}\sum_{i=1}^{m} \sigma_i (ca_i + b_i)\right] \\
&= E_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{a}\in A} \frac{1}{m}\sum_{i=1}^{m} \sigma_i ca_i + \cancel{\frac{1}{m}\sum_{i=1}^{m} \sigma_i b_i}\right] \\
&= |c| E_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{a}\in A} \frac{1}{m}\sum_{i=1}^{m} \sigma_i a_i\right] = |c| R(A).
\end{aligned}
$$

0

$E(\sigma_i b_i)$

$= E[\sigma_i] b_i$

Note that

- The components with b_i disappears on the third line because σ_i is equally likely to be ± 1
- If c is positive, moving c outside the expectation is straightforward.
- If c is negative, we can move the negative sign onto $\boldsymbol{\sigma}$ instead and note that taking expectation with $-\boldsymbol{\sigma}$ gives the same result as with $\boldsymbol{\sigma}$.

Exercise 9:

Let $\mathcal{H} = \{f(z) + g(z) | f \in \mathcal{F}, g \in \mathcal{G}\}$. Express $R(\mathcal{H} \circ S)$ in terms of $R(\mathcal{F} \circ S)$ and $R(\mathcal{G} \circ S)$.

Answer: $R(\mathcal{H} \circ S) = R(\mathcal{F} \circ S) + R(\mathcal{G} \circ S)$

$$R(\mathcal{H} \circ S) = E_\sigma\left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(z_i)\right]$$

$$= E_\sigma\left[\sup_{f \in \mathcal{F}, g \in \mathcal{G}} \frac{1}{m} \sum \sigma_i \left(f(z_i) + g(z_i)\right)\right]$$

$$= E_\sigma\left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum \sigma_i f(z_i) + \sup_{g \in \mathcal{G}} \sum \sigma_i g(z_i)\right]$$

$$= R(\mathcal{F} \circ S) + R(\mathcal{G} \circ S)$$

It turns out that, on average, the Rademacher complexity can be use to upper bound the representativeness value. So small Rademacher complexity implies small representativeness value.

Lemma: (SSBD Lemma 26.2)

$$E_{S\sim\mathcal{D}^m}[\mathrm{Rep}_{\mathcal{D}}(\mathcal{F},S)] \leq 2E_{S\sim\mathcal{D}^m}R(\mathcal{F}\circ S).$$

Proof in the Appendix. It is often easier to bound the Rademacher complexity rather than directly bounding the representativeness value.

$$\mathrm{Rep}_D(F,S) = \sup_{f\in F}\left(L_D(f) - L_S(f)\right)$$

sample another sample

To provide generalization bound, we will use McDiarmid's inequality

Lemma (McDiarmid's inequality): (SSBD Lemma 26.4) Let V be some set and let $f : V^m \to \mathbb{R}$ be a function of m variables such that for some $c > 0$ for all $i \in [m]$ for all $x_1, \ldots, x_m, x_i' \in V$, we have

To get Hoeffding, set $f : X \to [a,b]$

$c = \frac{(b-a)}{m}$

$$|f(x_1, \ldots, x_m) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)| \leq c.$$

Let $X_1, \ldots, X_m$ be m independent random variables taking values in V. Then with probability at least $1 - \delta$ we have

$\epsilon = \frac{(b-a)}{m}\sqrt{\ln\left(\frac{2}{\delta}\right)\frac{m}{2}}$

$$|f(X_1, \ldots, X_m) - E[f(X_1, \ldots, X_m)]| \leq c\sqrt{\ln\left(\frac{2}{\delta}\right) m/2}.$$

$\frac{\delta}{2} = \exp\left(\frac{-2\epsilon^2 m}{(b-a)^2}\right)$

$\delta = 2\exp($

We would like to apply McDiarmid's inequality on the representiveness value

$$\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f)).$$

To do that we need to bounded the constant c when used with the representativeness error.

Lemma: Assume that for all z and $h \in \mathcal{H}$ we have that $|\ell(h, z)| \leq c$. Let $f(S) = \mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S)$. Then

$$|f(z_1, \ldots, z_m) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m)| \leq 2c/m.$$

The proof is provided in the Appendix.

Theorem: (SSBD Theorem 26.5) Assume that for all z and $h \in \mathcal{H}$ we have that $|\ell(h,z)| \le c$. Then with probability at least $1-\delta$, for all $h \in \mathcal{H}$

$$L_{\mathcal{D}}(h) - L_S(h) \le 2E_{S' \sim D^m} R(\ell \circ \mathcal{H} \circ S') + c\sqrt{\frac{2\ln(2/\delta)}{m}}.$$

Proof:

- By the previous Lemma, the representativeness error $\text{Rep}_{\mathcal{D}}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}}(L_{\mathcal{D}}(f) - L_S(f))$ satisfies the bounded difference condition in McDiarmid's inequality with constant
$c' =$ $2c/m$.
- Furthermore we know that the average representativeness error is bounded by twice the average Rademacher complexity. The result follows from combining this with McDiarmid's inequality. □

McDiarmid $L_D(h) - L_S(h)$
$\le E[L_D(h) - L_S(h)] + c'\sqrt{\ln(\frac{2}{\delta})\frac{m}{2}}$

Replace c' with $\frac{2c}{m}$ to get $c\sqrt{\frac{2\ln 2/\delta}{m}}$

Implications:

- The term $c\sqrt{\frac{2\ln(2/\delta)}{m}}$ does not depend on $\mathcal{H}$ other than through the magnitude bound c. Becomes small quickly regardless of what function class is used.
- By bounding the expected Rademacher complexity $E_{S'\sim D^m} R(\ell \circ \mathcal{H} \circ S')$, we can bound the representativeness error.
  - The bound holds uniformly for all $h \in \mathcal{H}$.
  - The bound is distribution dependent.
  - For analysis, we often get a worst case bound on $R(\ell \circ \mathcal{H} \circ S)$ for any S. This allows us to give a distribution independent bound that holds for any distribution.

Massart's lemma allows us to bound the Rademacher complexity of finite function classes.

Lemma (Massart): (SSBD Lemma 26.8) Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ be a finite set of vectors in $\mathbb{R}^m$. Define $\bar{\mathbf{a}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{a}_i$. Then

$$R(A) \le \max_{\mathbf{a}\in A} \|\mathbf{a} - \bar{\mathbf{a}}\|_2 \frac{\sqrt{2\log(N)}}{m}.$$

grow like $\sqrt{m}$

$O\left(\sqrt{\frac{\log N}{m}}\right)$

The proof is in the Appendix.

Using Massart's Lemma and Rademacher complexity, we can now show that finite VC-dimension implies uniform convergence, completing the fundamental theorem.

- Let $(\mathbf{x}_1, y_i), \dots, (\mathbf{x}_m, y_m)$ be the training set. Sauer's lemma tells us that if $\mathrm{VCdim}(\mathcal{H}) = d$, then

$$|\{(h(\mathbf{x}_1), \dots, h(\mathbf{x}_m)) :\in \mathcal{H}\}| \leq \left(\frac{em}{d}\right)^d .$$

- Let $A = \{(\mathbb{1}_{[h(\mathbf{x}_1) \neq y_1]}, \dots, \mathbb{1}_{[h(\mathbf{x}_m) \neq y_m]}) : h \in \mathcal{H}\}$ denote the vectors generated by the function class composed with the 0-1 loss. We also have $|A| \leq \left(\frac{em}{d}\right)^d$.
- To use Massart's lemma, we need to bound $\|\mathbf{a} - \bar{\mathbf{a}}\|_2$ for $\mathbf{a} \in A$.
    - Each component of $\mathbf{a} - \bar{\mathbf{a}}$ has magnitude at most 1, hence

$$\|\mathbf{a} - \bar{\mathbf{a}}\|_2 \leq \sqrt{\sum_{i=1}^{m} \underbrace{(a_i - \bar{a}_i)^2}_{\leq 1}} \leq \sqrt{m}.$$

- Combining Sauer's lemma that shows $|A| \leq \left(\frac{em}{d}\right)^d$ with Massart's lemma, we get

$$R(A) \leq \sqrt{\frac{2d \log(em/d)}{m}}.$$

- Applying SSBD Theorem 26.5, we get

$$L_{\mathcal{D}}(h) - L_S(h) \leq \sqrt{\frac{8d \log(em/d)}{m}} + \sqrt{\frac{2 \ln(2/\delta)}{m}}.$$

- Repeating the argument for the minus 0-1 loss (to get two sided bound), and applying the union bound, we get

δ/2

$$
\begin{aligned}
|L_{\mathcal{D}}(h) - L_S(h)| &\leq \sqrt{\frac{8d\log(em/d)}{m}} + \sqrt{\frac{2\ln(4/\delta)}{m}} \\
&\leq 2\sqrt{\frac{8d\log(em/d) + 2\ln(4/\delta)}{m}},
\end{aligned}
$$

simplify

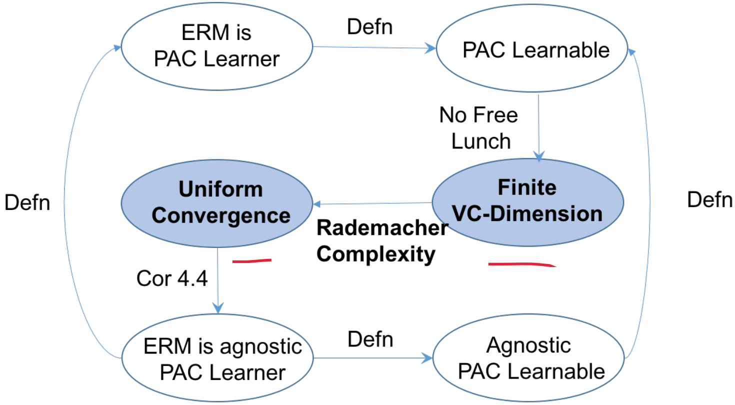
where the second inequality comes from concavity of the square root.

- To ensure that this is less than ϵ, it suffices to have

$$
m \geq \frac{4}{\epsilon^2}(8d\log(m) + 8d\log(e/d) + 2\log(4/\delta)).
$$

- Using SSBD Lemma A.2, is suffices that

$$
m \geq 4\frac{32d}{\epsilon^2}\log\left(\frac{64d}{\epsilon^2}\right) + \frac{8}{\epsilon^2}(8d\log(e/d) + 2\log(4/\delta)).
$$



Fundamental Theorem: These are equivalent

Measures of Complexity

We have seen two measures of complexity of function classes with infinite number of functions.

- VC Dimension
  - Can bound the number of functions on m points.
  - Combinatorial parameter: largest number of points that can be shattered.
- Rademacher Complexity
  - Average over all partitions into two sets, where maximize difference in the two sets using functions in the class.
  - Roughly measures how well the function class can fit random classifications.
  - Defined for a single sample. Can estimate the expected Rademacher complexity using a single sample.

Other commonly used complexity measures

- Covering number
  - How many balls of radius r is required such that each members in the set is within at least one ball?
  - A type of discretization of the space.
- Packing number
  - How many points can we fit into the set such that all points are a distance of at least r from each other?
  - Closely related to covering. If cannot fit any more point, all members of the set must be within distance r of one of the existing points.

Reference

Some material are taken directly from SSBD.

- SSBD Chapters 2, 3, 4, 5, 6, 26, 28

Outline

Rademacher Complexity Proofs

Lemma: (SSBD Lemma 26.2)

$$E_{S\sim\mathcal{D}^m}[\mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S)] \leq 2E_{S\sim\mathcal{D}^m}R(\mathcal{F}\circ S).$$

Proof:

- Let $S' = \{z'_1, \ldots, z'_m\}$ be another i.i.d. sample. Then $L_D(f) = E_{S'}[L_{S'}(f)]$, giving

$$L_{\mathcal{D}}(f) - L_S(f) = E_{S'}[L_{S'}(f)] - L_S(f) = E_{S'}[L_{S'}(f) - L_S(f)].$$

- Taking supremum over $f \in \mathcal{F}$ and using the fact that sup of expectation is smaller than expectation of sup

$$\sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f)) = \sup_{f \in \mathcal{F}} E_{S'}[L_{S'}(f) - L_S(f)]$$
$$\leq E_{S'}\left[\sup_{f \in \mathcal{F}} (L_{S'}(f) - L_S(f))\right].$$

- Taking expectation on both sides

$$E_S[\sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f))] \leq E_{S,S'}\left[\sup_{f \in \mathcal{F}} (L_{S'}(f) - L_S(f))\right]$$
$$= \frac{1}{m} E_{S,S'}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} (f(z_i') - f(z_i))\right]$$

- Let σ_i be a random variable such that $P[\sigma_i = 1] = P[\sigma_i = -1] = 1/2$. As z_i and z_i' are i.i.d. random variables, we have

$$E_{S,S'}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}(f(z_i') - f(z_i))\right] = E_{S,S',\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(f(z_i') - f(z_i))\right]$$

- We also have

$$\begin{aligned} & E_{S,S',\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(f(z_i') - f(z_i))\right] \\ & \leq E_{S,S',\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i') + \sup_{f\in\mathcal{F}}\sum_{i=1}^{m}-\sigma_i f(z_i))\right] \\ & = 2mE_{S\sim\mathcal{D}^m}R(\mathcal{F}\circ S), \end{aligned}$$

where the third line is because the prob of $\boldsymbol{\sigma}$ is the same as the prob of $-\boldsymbol{\sigma}$. □

Lemma: Assume that for all z and $h \in \mathcal{H}$ we have that $|\ell(h,z)| \leq c$. Let $f(S) = \mathrm{Rep}_{\mathcal{D}}(\mathcal{F}, S)$. Then

$$|f(z_1, \ldots, z_m) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m)| \leq 2c/m.$$

Proof:

- Let $S = \{z_1, \ldots, z_m\}$ and $S' = \{z_1, \ldots, z_j', \ldots, z_m\}$ differ in element j.
- Substituting the definition of f

$$|f(S) - f(S')| = \left| \sup_{h \in \mathcal{H}} \left(E_D[\ell(h,z)] - \frac{1}{m} \sum_{z \in S} \ell(h,z)\right) - \sup_{h \in \mathcal{H}} \left(E_D[\ell(h,z)] - \frac{1}{m} \sum_{z \in S'} \ell(h,z)\right) \right|.$$

- Let h^* maximize $f(S)$. Substituting, we get

$$
\begin{aligned}
|f(S) - f(S')| &= \Bigg| E_D[\ell(h^*, z)] - \frac{1}{m}\sum_{z\in S}\ell(h^*, z) \\
&\qquad - \sup_{h\in\mathcal{H}}\Big(E_D[\ell(h, z)] - \frac{1}{m}\sum_{z\in S'}\ell(h, z)\Big)\Bigg| \\
&\leq \Bigg| E_D[\ell(h^*, z)] - \frac{1}{m}\sum_{z\in S}\ell(h^*, z) \\
&\qquad - E_D[\ell(h^*, z)] + \frac{1}{m}\sum_{z\in S'}\ell(h^*, z)\Bigg| \\
&= \Bigg| \frac{1}{m}\sum_{z\in S'}\ell(h^*, z) - \frac{1}{m}\sum_{z\in S}\ell(h^*, z)\Bigg|
\end{aligned}
$$

where the second line is because h^* may not maximize $f(S')$.

- As all the elements except one are the same in S and S', we have

$$\begin{aligned}|f(S)-f(S')| &\leq \left|\frac{1}{m}\sum_{z\in S'}\ell(h^*,z)-\frac{1}{m}\sum_{z\in S}\ell(h^*,z)\right| \\ &= \frac{1}{m}\left|\ell(h^*,z_j')-\ell(h^*,z_j)\right| \\ &\leq \frac{2c}{m}.\end{aligned}$$

$\square$

Proof of Massart's Lemma

Lemma (Massart): (SSBD Lemma 26.8) Let $A = \{\mathbf{a}_1, \dots, \mathbf{a}_N\}$ be a finite set of vectors in $\mathbb{R}^m$. Define $\bar{\mathbf{a}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{a}_i$. Then

$$R(A) \leq \max_{\mathbf{a}\in A} \|\mathbf{a} - \bar{\mathbf{a}}\|_2 \frac{\sqrt{2\log(N)}}{m}.$$

Proof: (Massart's Lemma)

- From SSBD Lemma 26.6, we can work with $\bar{\mathbf{a}} = 0$.
- Let $\lambda > 0$ and $A' = \{\lambda \mathbf{a}_1, \ldots, \lambda \mathbf{a}_N\}$. Then

$$
\begin{aligned}
mR(A') &= E_{\boldsymbol{\sigma}}\left[\max_{\mathbf{a}\in A'}\langle \boldsymbol{\sigma}, \mathbf{a}\rangle\right] = E_{\boldsymbol{\sigma}}\left[\log\left(\max_{\mathbf{a}\in A'} e^{\langle \boldsymbol{\sigma}, \mathbf{a}\rangle}\right)\right] \\
&\leq E_{\boldsymbol{\sigma}}\left[\log\left(\sum_{\mathbf{a}\in A'} e^{\langle \boldsymbol{\sigma}, \mathbf{a}\rangle}\right)\right] \\
&\leq \log\left(E_{\boldsymbol{\sigma}}\left[\sum_{\mathbf{a}\in A'} e^{\langle \boldsymbol{\sigma}, \mathbf{a}\rangle}\right]\right) \quad \text{Jensen's Inequality} \\
&\leq \log\left(\sum_{\mathbf{a}\in A'} \prod_{i=1}^{m} E_{\sigma^i}\left[e^{\sigma_i a_i}\right]\right),
\end{aligned}
$$

where we exploited independence of σ_i in the last step.

- From SSBD Lemma A.6

$$E_{\sigma^i}\left[e^{\sigma_i a_i}\right] = \frac{\exp(a_i) + \exp(-a_i)}{2} \leq \exp(a_i^2/2)$$

giving

$$\begin{aligned} mR(A') &\leq \log\left(\sum_{\mathbf{a}\in A'}\prod_{i=1}^{m}\exp(a_i^2/2)\right) = \log\left(\sum_{\mathbf{a}\in A'}\exp(\|\mathbf{a}\|^2/2)\right) \\ &\leq \log\left(|A'|\max_{\mathbf{a}\in A'}\exp(\|\mathbf{a}\|^2/2)\right) = \log|A'| + \max_{\mathbf{a}\in A'}(\|\mathbf{a}\|^2/2). \end{aligned}$$

- From the previous lemma, $R(A) = \frac{1}{\lambda} R(A')$ giving

$$R(A) \leq \frac{\log |A'| + \max_{\mathbf{a}' \in A'}(\|\mathbf{a}'\|^2/2)}{\lambda m}$$
$$= \frac{\log |A| + \lambda^2 \max_{\mathbf{a} \in A}(\|\mathbf{a}\|^2/2)}{\lambda m}.$$

- Setting $\lambda = \sqrt{2 \log(|A|) / \max_{\mathbf{a} \in A} \|\mathbf{a}\|^2}$ and rearranging gives the result $\square$