Class 11-12: PAC Learning - VC - Dimension
Online Class

• Attend synchronous class (if possible), participate, ask questions

On Zoom:
• Use the largest display you have. Your phone is not a good choice
• Join Zoom with video ON and audio Muted
• View options: Fit to Window
• To participate:
  • Use Raise Hand in Reactions
  • Unmute and talk
  • Chat to Everyone or Host.
Learning a Binary Classifier
PAC Learning - Valiant 1984

- An unknown probability distribution $\mathcal{D}$ on a domain $\mathcal{U}$
- An unknown correct classification – a partition $c$ of $\mathcal{U}$ to $\text{In}$ and $\text{Out}$ sets
- Input:
  - Concept class $\mathcal{C}$ – a collection of possible classification rules (partitions of $\mathcal{U}$).
  - A training set $\{(x_i, c(x_i)) \mid i = 1, \ldots, m\}$, where $x_1, \ldots, x_m$ are sampled from $\mathcal{D}$.
- Goal: With probability $1 - \delta$ the algorithm generates a good classifier.
  A classifier is good if the probability that it errs on an item generated from $\mathcal{D}$ is $\leq \text{opt}(\mathcal{C}) + \epsilon$, where $\text{opt}(\mathcal{C})$ is the error probability of the best classifier in $\mathcal{C}$. 
Formal Definition

- We have a unit cost function $Oracle(c, D)$ that produces a pair $(x, c(x))$, where $x$ is distributed according to $D$, and $c(x)$ is the value of the concept $c$ at $x$. Successive calls are independent.

- A concept class $\mathcal{C}$ over input set $X$ is PAC learnable if there is an algorithm $L$ with the following properties: For every concept $c \in \mathcal{C}$, every distribution $D$ on $X$, and every $0 \leq \epsilon, \delta \leq 1/2$,
  
  - Given a function $Oracle(c, D)$, $\epsilon$ and $\delta$, with probability $1 - \delta$ the algorithm output an hypothesis $h \in \mathcal{C}$ such that $Pr_D(h(x) \neq c(x)) \leq \epsilon$.
  - The concept class $\mathcal{C}$ is efficiently PAC learnable if the algorithm runs in time polynomial in the size of the problem, $1/\epsilon$ and $1/\delta$.

So far we showed that the concept class "intervals on the line" and "Boolean conjunctions" are efficiently PAC learnable.
Uniform Convergence for Learning Binary Classification

- Given a concept class $\mathcal{C}$, and a training set sampled from $\mathcal{D}$, $\{(x_i, c(x_i)) \mid i = 1, \ldots, m\}$.
- For any $h \in \mathcal{C}$, let $\Delta(c, h)$ be the set of items on which the two classifiers differ: $\Delta(c, h) = \{x \in U \mid h(x) \neq c(x)\}$
- For the realizable case we need a training set (sample) that with probability $1 - \delta$ intersects every set in
  $$\{\Delta(c, h) \mid Pr(\Delta(c, h)) \geq \epsilon\} \quad (\epsilon\text{-net})$$
- For the unrealizable case we need a training set that with probability $1 - \delta$ estimates, within additive error $\epsilon$, every set in
  $$\Delta(c, h) = \{x \in U \mid h(x) \neq c(x)\} \quad (\epsilon\text{-sample}).$$
- Under what conditions can a finite sample achieve these requirements?
  - What sample size is needed?
Uniform Convergence Sets

Given a collection $R$ of sets in a universe $X$, under what conditions a finite sample $N$ from an arbitrary distribution $\mathcal{D}$ over $X$, satisfies with probability $1 - \delta$,

1. $\forall r \in R, \ Pr_{\mathcal{D}}(r) \geq \epsilon \Rightarrow r \cap N \neq \emptyset$ (\(\epsilon\)-net)

2. for any $r \in R$, $\left| Pr_{\mathcal{D}}(r) - \frac{|N \cap r|}{|N|} \right| \leq \epsilon$ (\(\epsilon\)-sample)

- Under what conditions on $R$ can a finite sample achieve these requirements?
- What sample size is needed?
Two fundamental questions:

- What concept classes are PAC-learnable with a given number of training (random) examples?
- What concept class are efficiently learnable (in polynomial time)?

A complete (and beautiful) characterization for the first question, not very satisfying answer for the second one.

Some Examples:

- **Efficiently PAC learnable:** Interval in $R$, rectangular in $R^2$, disjunction of up to $n$ variables, 3-CNF formula, ...
- **PAC learnable, but not in polynomial time (unless $P = NP$):** DNF formula, finite automata, ...
- **Not PAC learnable:** Convex body in $R^2$, $\{\sin(hx) \mid 0 \leq h \leq \pi\}$, ...
$(X, R)$ is called a "range space":

- $X =$ finite or infinite set (the set of objects to learn)
- $R$ is a family of subsets of $X$, $R \subseteq 2^X$.
  - In learning, $R = \{ \Delta(c, h) \mid h \in C \}$, where $C$ is the concept class, and $c$ is the correct classification.
- For a finite set $S \subseteq X$, $s = |S|$, define the projection of $R$ on $S$,
  \[ \Pi_R(S) = \{ r \cap S \mid r \in R \}. \]

- If $|\Pi_R(S)| = 2^s$ we say that $R$ shatters $S$.
- The VC-dimension of $(X, R)$ is the maximum size of $S$ that is shattered by $R$. If there is no maximum, the VC-dimension is $\infty$. 

Vapnik–Chervonenkis (VC) Dimension 1968/1971

8 / 47
The VC-Dimension of a Collection of Intervals

$C = \text{collections of intervals in } [A,B] – \text{can shatter 2 point but not 3. No interval includes only the two red points}$

The VC-dimension of $C$ is 2
Collection of Half Spaces in the Plane

$C$ – all half space partitions in the plane. Any 3 points can be shattered:

- Cannot partition the red from the blue points
- The VC-dimension of half spaces on the plane is 3
- The VC-dimension of half spaces in $d$-dimension space is $d+1$
Axis-parallel rectangles on the plane

4 points that define a convex hull can be shattered.

No five points can be shattered since one of the points must be in the convex hull of the other four.
Convex Bodies in the Plane

• $C$ – all convex bodies on the plane

Any subset of the point can be included in a convex body. The VC-dimension of $C$ is $\infty$
A Few Examples

- $C =$ set of intervals on the line. Any two points can be shattered, no three points can be shattered.
- $C =$ set of linear half spaces in the plane. Any three points can be shattered but no set of 4 points. If the 4 points define a convex hull let one diagonal be 0 and the other diagonal be 1. If one point is in the convex hull of the other three, let the interior point be 1 and the remaining 3 points be 0.
- $C =$ set of axis-parallel rectangles on the plane. 4 points that define a convex hull can be shattered. No five points can be shattered since one of the points must be in the convex hull of the other four.
- $C =$ all convex sets in $\mathbb{R}^2$. Let $S$ be a set of $n$ points on a boundary of a cycle. Any subset $Y \subset S$ defines a convex set that doesn’t include $S \setminus Y$. 
The Main Result

Let $C$ be a concept class with VC-dimension $d$ then

1. $C$ is PAC learnable in the realizable case with

$$m = O\left(\frac{d}{\epsilon} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon} \ln \frac{1}{\delta}\right) \quad (\epsilon\text{-net})$$

samples.

2. $C$ is PAC learnable in the unrealizable case with

$$m = O\left(\frac{d}{\epsilon^2} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta}\right) \quad (\epsilon\text{-sample})$$

samples.

The sample size is not a function of the number of concepts, or the size of the domain!
Sauer’s Lemma

For a finite set $S \subseteq X$, $s = |S|$, define the projection of $R$ on $S$,

$$\Pi_R(S) = \{ r \cap S \mid r \in R \}.$$ 

**Theorem**

Let $(X, R)$ be a range space with VC-dimension $d$, for any $S \subseteq X$, such that $|S| = n$,

$$|\Pi_R(S)| \leq \sum_{i=0}^{d} \binom{n}{i}.$$ 

For $n = d$, $|\Pi_R(S)| \leq 2^d$, and for $n > d \geq 2$, $|\Pi_R(S)| \leq n^d$.

The number of distinct concepts on $n$ elements grows polynomially in the VC-dimension!
Proof

• By induction on $d$, and for a fixed $d$, by induction on $n$.
• True for $d = 0$ or $n = 0$, since $\Pi_R(S) = \{\emptyset\}$.
• Assume that the claim holds for $d' \leq d - 1$ and any $n$, and for $d$ and all $|S'| \leq n - 1$.
• Fix $x \in S$ and let $S' = S - \{x\}$.

\[
|\Pi_R(S)| = |\{r \cap S \mid r \in R\}|
\]
\[
|\Pi_R(S')| = |\{r \cap S' \mid r \in R\}|
\]
\[
|\Pi_{R(x)}(S')| = |\{r \cap S' \mid r \in R \text{ and } x \not\in r \text{ and } r \cup \{x\} \in R\}|
\]

• For $r_1 \cap S \neq r_2 \cap S$ we have $r_1 \cap S' = r_2 \cap S'$ iff $r_1 = r_2 \cup \{x\}$, or $r_2 = r_1 \cup \{x\}$. Thus,

\[
|\Pi_R(S)| = |\Pi_R(S')| + |\Pi_{R(x)}(S')|
\]
Fix $x \in S$ and let $S' = S - \{x\}$.

\[
\begin{align*}
|\Pi_R(S)| &= |\{r \cap S \mid r \in R\}| \\
|\Pi_R(S')| &= |\{r \cap S' \mid r \in R\}| \\
|\Pi_{R(x)}(S')| &= |\{r \cap S' \mid r \in R \text{ and } x \not\in r \text{ and } r \cup \{x\} \in R\}| 
\end{align*}
\]

- The VC-dimension of $(S, \Pi_R(S))$ is no more than the VC-dimension of $(X, R)$, which is $d$.
- The VC-dimension of the range space $(S', \Pi_R(S'))$ is no more than the VC-dimension of $(S, \Pi_R(S))$ and $|S'| = n - 1$, thus by the induction hypothesis $|\Pi_R(S')| \leq \sum_{i=0}^{d} \binom{n-1}{i}$.
- For each $r \in \Pi_{R(x)}(S')$ the range set $\Pi_S(R)$ has two sets: $r$ and $r \cup \{x\}$. If $B$ is shattered by $(S', \Pi_{R(x)}(S'))$ then $B \cup \{x\}$ is shattered by $(X, R)$, thus $(S', \Pi_{R(x)}(S'))$ has VC-dimension bounded by $d - 1$, and $|\Pi_{R(x)}(S')| \leq \sum_{i=0}^{d-1} \binom{n-1}{i}$. 
\[ |\Pi_R(S)| = |\Pi_R(S')| + |\Pi_R(x)(S')| \]

\[
|\Pi_R(S)| \leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\
= 1 + \sum_{i=1}^{d} \left( \binom{n-1}{i} + \binom{n-1}{i-1} \right) \\
= \sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \frac{n^i}{i!} \leq n^d
\]

[We use \( \binom{n-1}{i-1} + \binom{n-1}{i} = \frac{(n-1)!}{(i-1)!(n-i-1)!} \left( \frac{1}{n-i} + \frac{1}{i} \right) = \binom{n}{i} \)]

The number of distinct concepts on \( n \) elements grows polynomially in the VC-dimension!
**Definition**

Let $(X, R)$ be a range space, with a probability distribution $D$ on $X$. A set $N \subseteq X$ is an $\epsilon$-net for $X$ with respect to $D$ if

$$\forall r \in R, \Pr_D(r) \geq \epsilon \Rightarrow r \cap N \neq \emptyset.$$ 

**Theorem**

Let $(X, R)$ be a range space with VC-dimension bounded by $d$. With probability $1 - \delta$, a random sample of size

$$m \geq \frac{8d}{\epsilon} \ln \frac{16d}{\epsilon} + \frac{4}{\epsilon} \ln \frac{4}{\delta}$$

is an $\epsilon$-net for $(X, R)$. 
How to Sample an \( \epsilon \)-net?

- Let \((X, R)\) be a range space with VC-dimension \(d\). Let \(M\) be \(m\) independent samples from \(X\).
- Let \(E_1 = \{\exists r \in R \mid \Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0\}\). We want to show that \(\Pr(E_1) \leq \delta\).
- Choose a second sample \(T\) of \(m\) independent samples.
- Let \(E_2 = \{\exists r \in R \mid \Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2\}\)

**Lemma**

\[ \Pr(E_2) \leq \Pr(E_1) \leq 2\Pr(E_2) \]
Lemma

\[ \Pr(E_2) \leq \Pr(E_1) \leq 2\Pr(E_2) \]

\[ E_1 = \{ \exists r \in R \mid \Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0 \} \]

\[ E_2 = \{ \exists r \in R \mid \Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2 \} \]

\[ \frac{\Pr(E_2)}{\Pr(E_1)} = \Pr(E_2 \mid E_1) \geq \Pr(\| T \cap r \| \geq \epsilon m/2) \geq 1/2 \]

Since \( |T \cap r| \) has a Binomial distribution \( B(m, \epsilon) \),
\[ \Pr(\| T \cap r \| < \epsilon m/2) \leq e^{-\epsilon m/8} < 1/2 \text{ for } m \geq 8/\epsilon. \]
\[ E_2 = \{ \exists r \in R \mid Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2 \} \]

\[ E_2' = \{ \exists r \in R \mid |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2 \} \]

**Lemma**

\[ Pr(E_1) \leq 2Pr(E_2) \leq 2Pr(E_2') \leq 2(2m)^d 2^{-\epsilon m/2}. \]

Choose an arbitrary set \( Z \) of size \( 2m \) and divide it randomly to \( M \) and \( T \). For a fixed \( r \in R \) and \( k = \epsilon m/2 \), let

\[ E_r = \{ |r \cap M| = 0 \text{ and } |r \cap T| \geq k \} = \{ |M \cap r| = 0 \text{ and } |r \cap (M \cup T)| \geq k \} \]

\[
Pr(E_r) = Pr(|M \cap r| = 0 \mid |r \cap (M \cup T)| \geq k) Pr(|r \cap (M \cup T)| \geq k) \\
\leq Pr(|M \cap r| = 0 \mid |r \cap (M \cup T)| \geq k) \leq \frac{(2m-k)}{m} \frac{(2m)}{m} \\
= \frac{m(m-1) \ldots (m-k+1)}{2m(2m-1) \ldots (2m-k+1)} \leq 2^{-\epsilon m/2}
\]
Since $|\Pi_R(Z)| \leq (2m)^d$, 

$$Pr(E'_2) \leq (2m)^d 2^{-\epsilon m/2}.$$ 

$$Pr(E_1) \leq 2Pr(E'_2) \leq 2(2m)^d 2^{-\epsilon m/2}.$$ 

**Theorem**

Let $(X, R)$ be a range space with VC-dimension bounded by $d$. With probability $1 - \delta$, a random sample of size 

$$m \geq \frac{8d}{\epsilon} \ln \frac{16d}{\epsilon} + \frac{4}{\epsilon} \ln \frac{4}{\delta}$$

is an $\epsilon$-net for $(X, R)$.

We need to show that $(2m)^d 2^{-\epsilon m/2} \leq \delta$ for 

$$m \geq \frac{8d}{\epsilon} \ln \frac{16d}{\epsilon} + \frac{4}{\epsilon} \ln \frac{1}{\delta}.$$
We show that \((2m)^d 2^{-\epsilon m/2} \leq \delta\). for \(m \geq \frac{8d}{\epsilon} \ln \frac{16d}{\epsilon} + \frac{4}{\epsilon} \ln \frac{1}{\delta}\).

Equivalently, we require

\[
\epsilon m/2 \geq \ln(1/\delta) + d \ln(2m).
\]

Clearly \(\epsilon m/4 \geq \ln(1/\delta)\), since \(m > \frac{4}{\epsilon} \ln \frac{1}{\delta}\).

We need to show that \(\epsilon m/4 \geq d \ln(2m)\).
Lemma

If \( y \geq x \ln x > e \), then \( \frac{2y}{\ln y} \geq x \).

Proof.

For \( y = x \ln x \) we have \( \ln y = \ln x + \ln \ln x \leq 2 \ln x \). Thus

\[
\frac{2y}{\ln y} \geq \frac{2x \ln x}{2 \ln x} = x.
\]

Differentiating \( f(y) = \frac{\ln y}{2y} \) we find that \( f(y) \) is monotonically decreasing when \( y \geq x \ln x \geq e \), and hence \( \frac{2y}{\ln y} \) is monotonically increasing on the same interval, proving the lemma.

Let \( y = 2m \geq \frac{16d}{\epsilon} \ln \frac{16d}{\epsilon} \) and \( x = \frac{16d}{\epsilon} \), we have

\[
\frac{4m}{\ln(2m)} \geq \frac{16d}{\epsilon},
\]

so

\[
\frac{\epsilon m}{4} \geq d \ln(2m)
\]

as required.
**Theorem**

A random sample of a range space with VC dimension $d$ that with probability at least $1 - \delta$ is an $\epsilon$-net must have size $\Omega\left(\frac{d}{\epsilon}\right)$.

Consider a range space $(X, R)$, with $X = \{x_1, \ldots, x_d\}$, and $R = 2^X$.

Define a probability distribution $D$:

\[
\begin{align*}
\Pr(x_1) &= 1 - 4\epsilon \\
\Pr(x_2) &= \Pr(x_3) = \cdots = \Pr(x_d) = \frac{4\epsilon}{d - 1}
\end{align*}
\]

Let $X' = \{x_2, \ldots, x_d\}$. 
Let $X' = \{x_2, \ldots, x_d\}$.

$$Pr(x_2) = Pr(x_3) = \cdots = Pr(x_d) = \frac{4\epsilon}{d-1}$$

Let $S$ be a sample of $m = \frac{(d-1)}{16\epsilon}$ examples from the distribution $D$.

Let $B$ be the event $|S \cap X'| \leq (d - 1)/2$, then $Pr(B) \geq 1/2$.

With probability $\geq 1/2$, the sample does not hit a set of probability

$$\frac{d-1}{2} \cdot \frac{4\epsilon}{d-1} = 2\epsilon$$

**Corollary**

A range space has a finite $\epsilon$-net iff its VC-dimension is finite.
Let \( X \) be a set of items, \( D \) a distribution on \( X \), and \( C \) a set of concepts on \( X \).

\[
\Delta(c, c') = \{ c \setminus c' \cup c' \setminus c \mid c' \in C \}
\]

We take \( m \) samples and choose a concept \( c' \), while the correct concept is \( c \).

If \( Pr_D(\{ x \in X \mid c'(x) \neq c(x) \}) > \epsilon \) then, \( Pr(\Delta(c, c')) \geq \epsilon \), and no sample was chosen in \( \Delta(c, c') \).

How many samples are needed so that with probability \( 1 - \delta \) all sets \( \Delta(c, c') \), \( c' \in C \), with \( Pr(\Delta(c, c')) \geq \epsilon \), are hit by the sample?
Theorem

The VC-dimension of \((X, \{\Delta(c, c') \mid c' \in C\})\) is the same as \((X, C)\).

Proof.

We show that
\[
\{c' \cap S \mid c' \in C\} \rightarrow \{((c' \setminus c) \cup (c \setminus c')) \cap S \mid c' \in C\}
\]
is a bijection.

Assume that \(c_1 \cap S \neq c_2 \cap S\), then w.o.l.g. \(x \in (c_1 \setminus c_2) \cap S\).

\(x \notin c\) iff \(x \in ((c_1 \setminus c) \cup (c \setminus c_1)) \cap S\) and
\(x \notin ((c_2 \setminus c) \cup (c \setminus c_2)) \cap S\).

\(x \in c\) iff \(x \notin ((c_1 \setminus c) \cup (c \setminus c_1)) \cap S\) and \(x \in ((c_2 \setminus c) \cup (c \setminus c_2)) \cap S\).

Thus, \(c_1 \cap S \neq c_2 \cap S\) iff
\(((c_1 \setminus c) \cup (c \setminus c_1)) \cap S \neq ((c_2 \setminus c) \cup (c \setminus c_2)) \cap S\). The projection on \(S\) in both range spaces has equal size.
PAC Learning

Theorem

In the realizable case, a concept class \( C \) is PAC-learnable iff the VC-dimension of the range space defined by \( C \) is finite.

Theorem

Let \( C \) be a concept class that defines a range space with VC dimension \( d \). For any \( 0 < \delta, \epsilon \leq 1/2 \), there is an

\[
m = O \left( \frac{d}{\epsilon} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon} \ln \frac{1}{\delta} \right)
\]

such that \( C \) is PAC learnable with \( m \) samples.
Unrealizable (Agnostic) Learning

- We are given a training set \( \{(x_1, c(x_1)), \ldots, (x_m, c(x_m))\} \), and a concept class \( \mathcal{C} \).
- No hypothesis in the concept class \( \mathcal{C} \) is consistent with all the training set (\( c \not\in \mathcal{C} \)).
- Relaxed goal: Let \( c \) be the correct concept. Find \( c' \in \mathcal{C} \) such that

\[
\Pr_{\mathcal{D}}(c'(x) \neq c(x)) \leq \inf_{h \in \mathcal{C}} \Pr_{\mathcal{D}}(h(x) \neq c(x)) + \epsilon.
\]

- An \( \epsilon/2 \)-sample of the range space \( (X, \Delta(c, c')) \) gives enough information to identify an hypothesis that is within \( \epsilon \) of the best hypothesis in the concept class.
When does the sample identify the correct rule?
The unrealizable (agnostic) case

- The unrealizable case - \(c\) may not be in \(C\).
- For any \(h \in C\), let \(\Delta(c, h)\) be the set of items on which the two classifiers differ: 
  \(\Delta(c, h) = \{x \in U \mid h(x) \neq c(x)\}\)
- For the training set \(\{(x_i, c(x_i)) \mid i = 1, \ldots, m\}\), let
  \[
  \tilde{Pr}(\Delta(c, h)) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{h(x_i) \neq c(x_i)}
  \]
- Algorithm: choose \(h^* = \arg\min_{h \in C} \tilde{Pr}(\Delta(c, h))\).
- If for every set \(\Delta(c, h)\),
  \[
  |Pr(\Delta(c, h)) - \tilde{Pr}(\Delta(c, h))| \leq \epsilon,
  \]
  then
  \[
  Pr(\Delta(c, h^*)) \leq opt(C) + 2\epsilon.
  \]

where \(opt(C)\) is the error probability of the best classifier in \(C\).
If for every set $\Delta(c, h)$,

$$|Pr(\Delta(c, h)) - \tilde{Pr}(\Delta(c, h))| \leq \epsilon,$$

then

$$Pr(\Delta(c, h^*)) \leq opt(C) + 2\epsilon.$$

where $opt(C)$ is the error probability of the best classifier in $C$. Let $\tilde{h}$ be the best classifier in $C$. Since the algorithm chose $h^*$,

$$\tilde{Pr}(\Delta(c, h^*)) \leq \tilde{Pr}(\Delta(c, \tilde{h})).$$

Thus,

$$Pr(\Delta(c, h^*)) - opt(C) \leq \tilde{Pr}(\Delta(c, h^*)) - opt(C) + \epsilon$$

$$\leq \tilde{Pr}(\Delta(c, \tilde{h})) - opt(C) + \epsilon \leq 2\epsilon$$
**Definition**

An $\varepsilon$-sample for a range space $(X, R)$, with respect to a probability distribution $\mathcal{D}$ defined on $X$, is a subset $N \subseteq X$ such that, for any $r \in R$,

$$\left| \Pr_{\mathcal{D}}(r) - \frac{|N \cap r|}{|N|} \right| \leq \varepsilon .$$

**Theorem**

Let $(X, \mathcal{R})$ be a range space with VC dimension $d$ and let $\mathcal{D}$ be a probability distribution on $X$. For any $0 < \varepsilon, \delta < 1/2$, there is an

$$m = O \left( \frac{d}{\varepsilon^2} \ln \frac{d}{\varepsilon} + \frac{1}{\varepsilon^2} \ln \frac{1}{\delta} \right)$$

such that a random sample from $\mathcal{D}$ of size greater than or equal to $m$ is an $\varepsilon$-sample for $X$ with with probability at least $1 - \delta$. 
Proof of the $\varepsilon$-sample Bound:

Let $N$ be a set of $m$ independent samples from $X$ according to $D$. Let

$$E_1 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \right\}.$$ 

We want to show that $\Pr(E_1) \leq \delta$.

Choose another set $T$ of $m$ independent samples from $X$ according to $D$. Let

$$E_2 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \land \left| \Pr(r) - \frac{|T \cap r|}{m} \right| \leq \varepsilon/2 \right\}$$

Lemma

$$\Pr(E_2) \leq \Pr(E_1) \leq 2 \Pr(E_2).$$
Lemma

\( \Pr(E_2) \leq \Pr(E_1) \leq 2 \Pr(E_2). \)

\[
E_1 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \right\}
\]

\[
E_2 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \land \left| \frac{|T \cap r|}{m} - \Pr(r) \right| \leq \varepsilon/2 \right\}
\]

For \( m \geq \frac{24}{\varepsilon^2} \),

\[
\frac{\Pr(E_2)}{\Pr(E_1)} = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_1)} = \Pr(E_2|E_1) \geq \Pr\left( \left| \frac{|T \cap r|}{m} - \Pr(r) \right| \leq \varepsilon/2 \right)
\]

\[
\geq 1 - 2e^{-\varepsilon^2 m/12} \geq 1/2
\]

[In bounding \( \Pr(E_2|E_1) \) we use the fact that the probability that \( \exists r \in R \) is not smaller than the probability that the event holds for a fixed \( r \).]
Instead of bounding the probability of

\[
E_2 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \land \left| \frac{|T \cap r|}{m} - \Pr(r) \right| \leq \varepsilon/2 \right\}
\]

we bound the probability of

\[
E'_2 = \{ \exists r \in R \mid ||r \cap N| - |r \cap T|| \geq \frac{\varepsilon}{2m} \}.
\]

By the triangle inequality (\(|A| + |B| \geq |A + B|\)):

\[
||r \cap N| - |r \cap T|| + ||r \cap T| - m \Pr_D(r)| \geq ||r \cap N| - m \Pr_D(r)|.
\]

or

\[
||r \cap N| - |r \cap T|| \geq ||r \cap N| - m \Pr_D(r)| - ||r \cap T| - m \Pr_D(r)| \geq \frac{\varepsilon}{2m}.
\]

Since \(N\) and \(T\) are random samples, we can first choose a random sample \(Z\) of \(2m\) elements, and partition it randomly into two sets of size \(m\) each. The event \(E'_2\) is in the probability space of random partitions of \(Z\).
Lemma

\[ Pr(E_1) \leq 2 \Pr(E_2) \leq 2 \Pr(E'_2) \leq 2(2m)^d e^{-\epsilon^2 m/8}. \]

• Since \( N \) and \( T \) are random samples, we can first choose a random sample of \( 2m \) elements \( Z = z_1, \ldots, z_{2m} \) and then partition it randomly into two sets of size \( m \) each.
• Since \( Z \) is a random sample, any partition that is independent of the actual values of the elements generates two random samples.
• We will use the following partition: for each pair of sampled items \( z_{2i-1} \) and \( z_{2i} \), \( i = 1, \ldots, m \), with probability \( 1/2 \) (independent of other choices) we place \( z_{2i-1} \) in \( T \) and \( z_{2i} \) in \( N \), otherwise we place \( z_{2i-1} \) in \( N \) and \( z_{2i} \) in \( T \).
For \( r \in R \), let \( E_r \) be the event
\[
E_r = \left\{ \| r \cap N \| - \| r \cap T \| \geq \frac{\varepsilon}{2}m \right\}.
\]

We have \( E'_2 = \{ \exists r \in R \mid \| r \cap N \| - \| r \cap T \| \geq \frac{\varepsilon}{2}m \} = \bigcup_{r \in R} E_r \).

- If \( z_{2i-1}, z_{2i} \in r \) or \( z_{2i-1}, z_{2i} \notin r \) they don’t contribute to the value of \( \| r \cap N \| - \| r \cap T \| \).
- If just one of the pair \( z_{2i-1} \) and \( z_{2i} \) is in \( r \) then their contribution is \(+1\) or \(-1\) with equal probabilities.
- There are no more than \( m \) pairs that contribute \(+1\) or \(-1\) with equal probabilities. Applying the Chernoff bound we have
\[
Pr(E_r) \leq e^{-(\varepsilon m/2)^2/2m} \leq e^{-\varepsilon^2 m/8}.
\]

- Since the projection of \( X \) on \( T \cup N \) has no more than \((2m)^d\) distinct sets we have the bound.
To complete the proof we show that for

\[ m \geq \frac{32d}{\epsilon^2} \ln \frac{64d}{\epsilon^2} + \frac{16}{\epsilon^2} \ln \frac{1}{\delta} \]

we have

\[ (2m)^d e^{-\epsilon^2 m/8} \leq \delta. \]

Equivalently, we require

\[ \epsilon^2 m/8 \geq \ln(1/\delta) + d \ln(2m). \]

Clearly \( \epsilon^2 m/16 \geq \ln(1/\delta) \), since \( m > \frac{16}{\epsilon^2} \ln \frac{1}{\delta} \).

To show that \( \epsilon^2 m/16 \geq d \ln(2m) \) we use:
Lemma

If \( y \geq x \ln x > e \), then \( \frac{2y}{\ln y} \geq x \).

Proof.

For \( y = x \ln x \) we have \( \ln y = \ln x + \ln \ln x \leq 2 \ln x \). Thus

\[
\frac{2y}{\ln y} \geq \frac{2x \ln x}{2 \ln x} = x.
\]

Differentiating \( f(y) = \frac{\ln y}{2y} \) we find that \( f(y) \) is monotonically decreasing when \( y \geq x \ln x \geq e \), and hence \( \frac{2y}{\ln y} \) is monotonically increasing on the same interval, proving the lemma.

Let \( y = 2m \geq \frac{64d}{e^2} \ln \frac{64d}{e^2} \) and \( x = \frac{64d}{e^2} \), we have \( \frac{4m}{\ln(2m)} \geq \frac{64d}{e^2} \), so

\[
\frac{\epsilon^2 m}{16} \geq d \ln(2m)
\]
as required.
Application: Unrealizable (Agnostic) Learning

- We are given a training set \( \{(x_1, c(x_1)), \ldots, (x_m, c(x_m))\} \), and a concept class \( C \).
- No hypothesis in the concept class \( C \) is consistent with all the training set (\( c \not\in C \)).
- Relaxed goal: Let \( c \) be the correct concept. Find \( c' \in C \) such that

\[
\Pr_D(c'(x) \neq c(x)) \leq \inf_{h \in C} \Pr_D(h(x) \neq c(x)) + \epsilon.
\]

- An \( \epsilon/2 \)-sample of the range space \( (X, \Delta(c, c')) \) gives enough information to identify an hypothesis that is within \( \epsilon \) of the best hypothesis in the concept class.
Definition

A set of functions $\mathcal{F}$ has the *uniform convergence* property with respect to a domain $Z$ if there is a function $m_\mathcal{F}(\epsilon, \delta)$ such that

- for any $\epsilon, \delta > 0$, $m(\epsilon, \delta) < \infty$
- for any distribution $D$ on $Z$, and a sample $z_1, \ldots, z_m$ of size $m = m_\mathcal{F}(\epsilon, \delta)$,

$$Pr(\sup_{f \in \mathcal{F}} |\frac{1}{m} \sum_{i=1}^{m} f(z_i) - E_D[f]| \leq \epsilon) \geq 1 - \delta.$$ 

Let $f_E(z) = 1_{z \in E}$ then $E[f_E(z)] = Pr(E)$. 
Uniform Convergence

Definition

A range space \((X, \mathcal{R})\) has the uniform convergence property if for every \(\epsilon, \delta > 0\) there is a sample size \(m = m(\epsilon, \delta)\) such that for every distribution \(\mathcal{D}\) over \(X\), if \(S\) is a random sample from \(\mathcal{D}\) of size \(m\) then, with probability at least \(1 - \delta\), \(S\) is an \(\epsilon\)-sample for \(X\) with respect to \(\mathcal{D}\).

Theorem

The following three conditions are equivalent:

1. A concept class \(\mathcal{C}\) over a domain \(X\) is agnostic PAC learnable.
2. The range space \((X, \mathcal{C})\) has the uniform convergence property.
3. The range space \((X, \mathcal{C})\) has a finite VC dimension.
Definition

A set of functions $\mathcal{F}$ has the *uniform convergence* property with respect to a domain $Z$ if there is a function $m_{\mathcal{F}}(\epsilon, \delta)$ such that

- for any $\epsilon, \delta > 0$, $m(\epsilon, \delta) < \infty$
- for any distribution $D$ on $Z$, and a sample $z_1, \ldots, z_m$ of size $m = m_{\mathcal{F}}(\epsilon, \delta)$,

\[
Pr(\sup_{f \in \mathcal{F}} |\frac{1}{m} \sum_{i=1}^{m} f(z_i) - E_D[f]| \leq \epsilon) \geq 1 - \delta.
\]

Let $f_E(z) = 1_{z \in E}$ then $E[f_E(z)] = Pr(E)$. 

Uniform Convergence and Learning

Definition

A set of functions $\mathcal{F}$ has the *uniform convergence* property with respect to a domain $Z$ if there is a function $m_{\mathcal{F}}(\epsilon, \delta)$ such that

- for any $\epsilon, \delta > 0$, $m(\epsilon, \delta) < \infty$
- for any distribution $D$ on $Z$, and a sample $z_1, \ldots, z_m$ of size $m = m_{\mathcal{F}}(\epsilon, \delta)$,

$$
Pr\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_{i=1}^{m} f(z_i) - E_D[f] \right| \leq \epsilon \right) \geq 1 - \delta.
$$

- Let $\mathcal{F}_H = \{f_h \mid h \in H\}$, where $f_h$ is the loss function for hypothesis $h$.
- $\mathcal{F}_H$ has the uniform convergence property $\Rightarrow$ an ERM (Empirical Risk Minimization) algorithm "learns" $H$.
- The *sample complexity* of learning $H$ is bounded by $m_{\mathcal{F}_H}(\epsilon, \delta)$.
Some Background

- Let $f_x(z) = \mathbf{1}_{z \leq x}$ (indicator function of the event $\{-\infty, x\}$)
- $F_m(x) = \frac{1}{m} \sum_{i=1}^{m} f_x(z_i)$ (empirical distributed function)
- Strong Law of Large Numbers: for a given $x$,
  $$F_m(x) \xrightarrow{a.s} F(x) = Pr(z \leq x).$$
- Glivenko-Cantelli Theorem:
  $$\sup_{x \in \mathbb{R}} |F_m(x) - F(x)| \xrightarrow{a.s} 0.$$
- Dvoretzky-Keifer-Wolfowitz Inequality
  $$Pr(\sup_{x \in \mathbb{R}} |F_m(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}.$$
- VC-dimension characterizes the uniform convergence property for arbitrary sets of events.