Foundations of Machine Learning
Learning with
Infinite Hypothesis Sets

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu
Motivation

- With an infinite hypothesis set $H$, the error bounds of the previous lecture are not informative.
- Is efficient learning from a finite sample possible when $H$ is infinite?
- Our example of axis-aligned rectangles shows that it is possible.
- Can we reduce the infinite case to a finite set? Project over finite samples?
- Are there useful measures of complexity for infinite hypothesis sets?
This lecture

- Rademacher complexity
- Growth Function
- VC dimension
- Lower bound
Empirical Rademacher Complexity

Definition:

- $G$ family of functions mapping from set $Z$ to $[a, b]$.
- sample $S = (z_1, \ldots, z_m)$.
- $\sigma_i$s (Rademacher variables): independent uniform random variables taking values in $\{-1, +1\}$.

$$\hat{\mathcal{R}}_S(G) = \mathbb{E} \left[ \sup_{g \in G} \frac{1}{m} \left[ \begin{array}{c} \sigma_1 \\ \vdots \\ \sigma_m \\ g(z_1) \\ \vdots \\ g(z_m) \end{array} \right] \right] = \mathbb{E} \left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right].$$

correlation with random noise
Rademacher Complexity

Definitions: let $G$ be a family of functions mapping from $Z$ to $[a, b]$.

- Empirical Rademacher complexity of $G$:
  $$\hat{\mathcal{R}}_S(G) = \mathbb{E}_\sigma \left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right],$$
  where $\sigma_i$s are independent uniform random variables taking values in $\{-1, +1\}$ and $S = (z_1, \ldots, z_m)$.

- Rademacher complexity of $G$:
  $$\mathcal{R}_m(G) = \mathbb{E}_{S \sim D^m} [\hat{\mathcal{R}}_S(G)].$$
Rademacher Complexity Bound

(Koltchinskii and Panchenko, 2002)

**Theorem:** Let $G$ be a family of functions mapping from $Z$ to $[0, 1]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for all $g \in G$:

$$
\mathbb{E}[g(z)] \leq \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(G) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.
$$

$$
\mathbb{E}[g(z)] \leq \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_S(G) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.
$$

**Proof:** Apply McDiarmid’s inequality to

$$
\Phi(S) = \sup_{g \in G} \mathbb{E}[g] - \hat{\mathbb{E}}_S[g].
$$
• **Changing one point of $S$ changes $\Phi(S)$ by at most $\frac{1}{m}$.**

$$\Phi(S') - \Phi(S) = \sup_{g \in G} \{E[g] - \hat{E}_{S'}[g]\} - \sup_{g \in G} \{E[g] - \hat{E}_S[g]\}$$

$$\leq \sup_{g \in G} \{\{E[g] - \hat{E}_{S'}[g]\} - \{E[g] - \hat{E}_S[g]\}\}$$

$$= \sup_{g \in G} \{\hat{E}_S[g] - \hat{E}_{S'}[g]\} = \sup_{g \in G} \frac{1}{m} (g(z_m) - g'(z'_m)) \leq \frac{1}{m}.$$

• **Thus, by McDiarmid’s inequality, with probability at least $1 - \frac{\delta}{2}$**

$$\Phi(S) \leq E[S] \Phi(S) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

• **We are left with bounding the expectation.**
• Series of observations:

\[
E[\Phi(S)] = E_S \left[ \sup_{g \in G} E[g] - \hat{E}_S(g) \right]
\]

\[
= E_S \left[ \sup_{g \in G} E[\hat{E}_{S'}(g) - \hat{E}_S(g)] \right]
\]

(sub-add. of sup) \leq E_{S,S'} \left[ \sup_{g \in G} \hat{E}_{S'}(g) - \hat{E}_S(g) \right]

\[
= E_{S,S'} \left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} (g(z'_i) - g(z_i)) \right]
\]

(swap \( z_i \) and \( z'_i \)) = E_{\sigma,S,S'} \left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (g(z'_i) - g(z_i)) \right]

(sub-additiv. of sup) \leq E_{\sigma,S'} \left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z'_i) \right] + E_{\sigma,S} \left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} -\sigma_i g(z_i) \right]

\[
= 2 E_{\sigma,S} \left[ \sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right] = 2 \mathcal{R}_m(G').
\]
• Now, changing one point of $S$ makes $\hat{R}_S(G)$ vary by at most $\frac{1}{m}$. Thus, again by McDiarmid’s inequality, with probability at least $1 - \frac{\delta}{2}$,

$$R_m(G) \leq \hat{R}_S(G) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$  

• Thus, by the union bound, with probability at least $1 - \delta$,

$$\Phi(S) \leq 2\hat{R}_S(G) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$
Loss Functions - Hypothesis Set

**Proposition:** Let $H$ be a family of functions taking values in $\{-1, +1\}$, $G$ the family of zero-one loss functions of $H$: $G = \{(x, y) \mapsto 1_{h(x) \neq y} : h \in H\}$. Then,

$$\mathcal{R}_m(G) = \frac{1}{2} \mathcal{R}_m(H).$$

**Proof:**

$$\mathcal{R}_m(G) = \mathbb{E}_{S, \sigma} \left[ \sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i 1_{h(x_i) \neq y_i} \right]$$

$$= \mathbb{E}_{S, \sigma} \left[ \sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \frac{1}{2} (1 - y_i h(x_i)) \right]$$

$$= \frac{1}{2} \mathbb{E}_{S, \sigma} \left[ \sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} -\sigma_i y_i h(x_i) \right]$$

$$= \frac{1}{2} \mathbb{E}_{S, \sigma} \left[ \sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i) \right].$$
Corollary: Let $H$ be a family of functions taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \hat{R}(h) + \mathcal{R}_m(H) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$ 

$$R(h) \leq \hat{R}(h) + \hat{\mathcal{R}}_S(H) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$
Remarks

First bound distribution-dependent, second data-dependent bound, which makes them attractive.

But, how do we compute the empirical Rademacher complexity?

Computing $E_{\sigma} \left[ \sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i) \right]$ requires solving ERM problems, typically computationally hard.

Relation with combinatorial measures easier to compute?
This lecture

- Rademacher complexity
- Growth Function
- VC dimension
- Lower bound
Growth Function

- **Definition:** the growth function $\Pi_H : \mathbb{N} \rightarrow \mathbb{N}$ for a hypothesis set $H$ is defined by

$$\forall m \in \mathbb{N}, \quad \Pi_H(m) = \max_{\{x_1, \ldots, x_m\} \subseteq X} \left| \{(h(x_1), \ldots, h(x_m)) : h \in H\} \right|.$$ 

- Thus, $\Pi_H(m)$ is the maximum number of ways $m$ points can be classified using $H$. 


Massart’s Lemma

**Theorem:** Let \( A \subseteq \mathbb{R}^m \) be a finite set, with \( R = \max_{x \in A} \|x\|_2 \), then, the following holds:

\[
E_{\sigma} \left[ \frac{1}{m} \sup_{x \in A} \sum_{i=1}^{m} \sigma_i x_i \right] \leq \frac{R \sqrt{2 \log |A|}}{m}.
\]

**Proof:**

\[
\exp \left( t E_{\sigma} \left[ \sup_{x \in A} \sum_{i=1}^{m} \sigma_i x_i \right] \right) \leq E_{\sigma} \left( \exp \left( t \sup_{x \in A} \sum_{i=1}^{m} \sigma_i x_i \right) \right) \quad \text{(Jensen’s ineq.)}
\]

\[
= E_{\sigma} \left( \sup_{x \in A} \exp \left( t \sum_{i=1}^{m} \sigma_i x_i \right) \right)
\]

\[
\leq \sum_{x \in A} E_{\sigma} \left( \exp \left( t \sum_{i=1}^{m} \sigma_i x_i \right) \right) = \sum_{x \in A} \prod_{i=1}^{m} E_{\sigma} \left( \exp \left[ t \sigma_i x_i \right] \right)
\]

\((\text{Hoeffding’s ineq.}) \leq \sum_{x \in A} \left( \exp \left[ \frac{\sum_{i=1}^{m} t^2 (2|x_i|)^2}{8} \right] \right) \leq |A| e^{\frac{t^2 R^2}{2}}.
\)
• Taking the log yields:

\[
E_{\sigma} \left[ \sup_{x \in A} \sum_{i=1}^{m} \sigma_i x_i \right] \leq \frac{\log |A|}{t} + \frac{tR^2}{2}.
\]

• Minimizing the bound by choosing \( t = \frac{\sqrt{2 \log |A|}}{R} \) gives

\[
E_{\sigma} \left[ \sup_{x \in A} \sum_{i=1}^{m} \sigma_i x_i \right] \leq R \sqrt{2 \log |A|}.
\]
Growth Function Bound on Rad. Complexity

**Corollary:** Let $G$ be a family of functions taking values in $\{-1, +1\}$, then the following holds:

$$\mathcal{K}_m(G) \leq \sqrt{\frac{2 \log \Pi_G(m)}{m}}.$$ 

**Proof:**

$$\hat{\mathcal{K}}_S(G) = \mathbb{E}_\sigma \left[ \sup_{g \in G} \frac{1}{m} \left[ \begin{array}{c} \sigma_1 \\ \vdots \\ \sigma_m \end{array} \right] \cdot \left[ \begin{array}{c} g(z_1) \\ \vdots \\ g(z_m) \end{array} \right] \right]$$

$$\leq \frac{\sqrt{m} \sqrt{2 \log |\{(g(z_1), \ldots, g(z_m)) : g \in G\}|}}{m} \quad \text{(Massart’s Lemma)}$$

$$\leq \frac{\sqrt{m} \sqrt{2 \log \Pi_G(m)}}{m} = \sqrt{\frac{2 \log \Pi_G(m)}{m}}.$$
Corollary: Let $H$ be a family of functions taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

$$R(h) \leq \hat{R}(h) + \sqrt{2 \log \Pi_H(m)} + \sqrt{\log \frac{1}{2m}}.$$

But, how do we compute the growth function? Relationship with the VC-dimension (Vapnik-Chervonenkis dimension).
This lecture

- Rademacher complexity
- Growth Function
- VC dimension
- Lower bound
VC Dimension


- **Definition**: the **VC-dimension** of a hypothesis set $H$ is defined by

  $$\text{VCdim}(H) = \max\{m : \Pi_H(m) = 2^m\}.$$ 

- Thus, the VC-dimension is the size of the largest set that can be fully shattered by $H$.

- Purely combinatorial notion.
Examples

- In the following, we determine the VC dimension for several hypothesis sets.

- To give a lower bound $d$ for $\text{VCdim}(H)$, it suffices to show that a set $S$ of cardinality $d$ can be shattered by $H$.

- To give an upper bound, we need to prove that no set $S$ of cardinality $d+1$ can be shattered by $H$, which is typically more difficult.
Intervals of The Real Line

Observations:

• Any set of two points can be shattered by four intervals.

- -  + -  - +

• No set of three points can be shattered since the following dichotomy “+ - +” is not realizable (by definition of intervals):

+ - +

• Thus, \( VCdim(\text{intervals in } \mathbb{R}) = 2 \).
Hyperplanes

Observations:

• Any three non-collinear points can be shattered:

• Unrealizable dichotomies for four points:

• Thus, $\text{VCdim}(\text{hyperplanes in } \mathbb{R}^d) = d + 1$. 
Observations:

• The following four points can be shattered:

```
+ - + - + - + - + -
+ + + + + + + + +
```

• No set of five points can be shattered: label negatively the point that is not near the sides.

```
+ + + - + + + + +
```

• Thus, $\text{VCdim}(\text{axis-aligned rectangles}) = 4$. 
Convex Polygons in the Plane

Observations:

• $2d + 1$ points on a circle can be shattered by a $d$-gon:

\[ |\text{positive points}| < |\text{negative points}| \]

• It can be shown that choosing the points on the circle maximizes the number of possible dichotomies. Thus, $\text{VCdim}(\text{convex } d\text{-gons}) = 2d + 1$.

Also, $\text{VCdim}(\text{convex polygons}) = +\infty$. 
Observations:

- Any finite set of points on the real line can be shattered by \( \{t \mapsto \sin(\omega t) : \omega \in \mathbb{R} \} \).
- Thus, \( \text{VCdim}(\text{sine functions}) = +\infty \).
Sauer’s Lemma

(Vapnik & Chervonenkis, 1968-1971; Sauer, 1972)

**Theorem:** let $H$ be a hypothesis set with $\text{VCdim}(H) = d$, then, for all $m \in \mathbb{N}$,

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$ 

**Proof:** the proof is by induction on $m+d$. The statement clearly holds for $m = 1$ and $d = 0$ or $d = 1$. Assume that it holds for $(m - 1, d - 1)$ and $(m - 1, d)$.

- Fix a set $S = \{x_1, \ldots, x_m\}$ with $\Pi_H(m)$ dichotomies and let $G = H|_S$ be the set of concepts $H$ induces by restriction to $S$. 

Consider the following families over $S' = \{x_1, \ldots, x_{m-1}\}$:

$$G_1 = G|_{S'} \quad G_2 = \{g' \subseteq S' : (g' \in G) \land (g' \cup \{x_m\} \in G)\}.$$ 

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$\cdots$</th>
<th>$x_{m-1}$</th>
<th>$x_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
</tbody>
</table>

Observe that $|G_1| + |G_2| = |G|$. 
• Since $\text{VCdim}(G_1) \leq d$, by the induction hypothesis,

$$|G_1| \leq \Pi_{G_1}(m - 1) \leq \sum_{i=0}^{d} \binom{m - 1}{i}.$$ 

• By definition of $G_2$, if a set $Z \subseteq S'$ is shattered by $G_2$, then the set $Z \cup \{x_m\}$ is shattered by $G$. Thus,

$$\text{VCdim}(G_2) \leq \text{VCdim}(G) - 1 = d - 1$$

and by the induction hypothesis,

$$|G_2| \leq \Pi_{G_2}(m - 1) \leq \sum_{i=0}^{d-1} \binom{m - 1}{i}.$$ 

• Thus, $|G| \leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i}$

$$= \sum_{i=0}^{d} \binom{m-1}{i} + \binom{m-1}{i-1} = \sum_{i=0}^{d} \binom{m}{i}.$$
Sauer’s Lemma - Consequence

- **Corollary**: let $H$ be a hypothesis set with $\text{VCdim}(H) = d$ then, for all $m \geq d$,

$$\Pi_H(m) \leq \left( \frac{em}{d} \right)^d = O(m^d).$$

- **Proof**:

$$\sum_{i=0}^d \binom{m}{i} \leq \sum_{i=0}^d \binom{m}{i} \left( \frac{m}{d} \right)^{d-i}$$

$$\leq \sum_{i=0}^m \binom{m}{i} \left( \frac{m}{d} \right)^{d-i}$$

$$= \left( \frac{m}{d} \right)^d \sum_{i=0}^m \binom{m}{i} \left( \frac{d}{m} \right)^i$$

$$= \left( \frac{m}{d} \right)^d \left( 1 + \frac{d}{m} \right)^m \leq \left( \frac{m}{d} \right)^d e^d.$$
Remarks

Remarkable property of growth function:

- either $\text{VCdim}(H) = d < +\infty$ and $\Pi_H(m) = O(m^d)$
- or $\text{VCdim}(H) = +\infty$ and $\Pi_H(m) = 2^m$. 
Corollary: Let $H$ be a family of functions taking values in $\{-1, +1\}$ with VC dimension $d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$, 

$$R(h) \leq \hat{R}(h) + \sqrt{2d \log \frac{em}{d}} + \sqrt{\log \frac{1}{2m}}.$$ 

Proof: Corollary combined with Sauer’s lemma.

Note: The general form of the result is 

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\log\frac{m/d}{m/d}}\right).$$
Comparison - Standard VC Bound

(Vapnik & Chervonenkis, 1971; Vapnik, 1982)

**Theorem:** Let $H$ be a family of functions taking values in $\{-1, +1\}$ with VC dimension $d$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for any $h \in H$,

\[
R(h) \leq \widehat{R}(h) + \sqrt{8d \log \frac{2em}{d} + 8 \log \frac{4}{\delta}}.
\]

**Proof:** Derived from growth function bound

\[
\Pr \left[ \left| R(h) - \widehat{R}(h) \right| > \epsilon \right] \leq 4\Pi_H(2m) \exp \left( -\frac{m\epsilon^2}{8} \right).
\]
This lecture

- Rademacher complexity
- Growth Function
- VC dimension
- Lower bound
Theorem: let $H$ be a hypothesis set with VC dimension $d > 1$. Then, for any learning algorithm $L$,

$$\exists D, \exists f \in H, \Pr_{S \sim D^m} \left[ R_D(h_S, f) > \frac{d - 1}{32m} \right] \geq 1/100.$$ 

Proof: choose $D$ such that $L$ can do no better than tossing a coin for some points.

- Let $X = \{x_0, x_1, \ldots, x_{d-1}\}$ be a set fully shattered. For any $\epsilon > 0$, define $D$ with support $X$ by

$$\Pr_D[x_0] = 1 - 8\epsilon \quad \text{and} \quad \forall i \in [1, d - 1], \Pr_D[x_i] = \frac{8\epsilon}{d - 1}.$$
• We can assume without loss of generality that $L$ makes no error on $x_0$.

• For a sample $S$, let $\overline{S}$ denote the set of its elements falling in $X_1 = \{x_1, \ldots, x_{d-1}\}$ and let $S$ be the set of samples of size $m$ with at most $(d - 1)/2$ points in $X_1$.

• Fix a sample $S \in S$. Using $|X - \overline{S}| \geq (d - 1)/2$,

$$
E_{f \sim U}[R_D(h_S, f)] = \sum_f \sum_{x \in X} 1_{h(x) \neq f(x)} \Pr[x] \Pr[f] \\
\geq \sum_f \sum_{x \notin \overline{S}} 1_{h(x) \neq f(x)} \Pr[x] \Pr[f] \\
= \sum_{x \notin \overline{S}} \left( \sum_f 1_{h(x) \neq f(x)} \Pr[f] \right) \Pr[x] \\
= \frac{1}{2} \sum_{x \notin \overline{S}} \Pr[x] \geq \frac{1}{2} \frac{d - 1}{2} \frac{8\epsilon}{d - 1} = 2\epsilon.
$$
• Since the inequality holds for all \( S \in S \), it also holds in expectation: \( \mathbb{E}_{S, f \sim U}[R_D(h_S, f)] \geq 2\epsilon \). This implies that there exists a labeling \( f_0 \) such that \( \mathbb{E}_S[R_D(h_S, f_0)] \geq 2\epsilon \).

• Since \( \Pr_D[X - \{x_0\}] \leq 8\epsilon \), we also have \( R_D(h_S, f_0) \leq 8\epsilon \). Thus,

\[
2\epsilon \leq \mathbb{E}_S[R_D(h_S, f_0)] \leq 8\epsilon \Pr_{S \in S}[R_D(h_S, f_0) \geq \epsilon] + (1 - \Pr_{S \in S}[R_D(h_S, f_0) \geq \epsilon])\epsilon.
\]

• Collecting terms in \( \Pr_{S \in S}[R_D(h_S, f_0) \geq \epsilon] \), we obtain:

\[
\Pr_{S \in S}[R_D(h_S, f_0) \geq \epsilon] \geq \frac{1}{7\epsilon}(2\epsilon - \epsilon) = \frac{1}{7}.
\]

• Thus, the probability over all samples \( S \) (not necessarily in \( S \)) can be lower bounded as

\[
\Pr_S[R_D(h_S, f_0) \geq \epsilon] \geq \Pr_{S \in S}[R_D(h_S, f_0) \geq \epsilon] \Pr[S] \geq \frac{1}{7} \Pr[S].
\]
• This leads us to seeking a lower bound for $\Pr[S]$. The probability that more than $(d - 1)/2$ points be drawn in a sample of size $m$ verifies the Chernoff bound for any $\gamma > 0$:

$$1 - \Pr[S] = \Pr[S_m \geq 8\epsilon m(1 + \gamma)] \leq e^{-8\epsilon m \frac{\gamma^2}{3}}.$$ 

• Thus, for $\epsilon = (d - 1)/(32m)$ and $\gamma = 1$,

$$\Pr[S_m \geq \frac{d-1}{2}] \leq e^{-(d-1)/12} \leq e^{-1/12} \leq 1 - 7\delta,$$

for $\delta \leq .01$. Thus, $\Pr[S] \geq 7\delta$ and

$$\Pr_{S} [R_D(h_S, f_0) \geq \epsilon] \geq \delta.$$
Agnostic PAC Model

**Definition**: concept class $C$ is **PAC-learnable** if there exists a learning algorithm $L$ such that:

- for all $c \in C$, $\epsilon > 0$, $\delta > 0$, and all distributions $D$,
  
  $$\Pr_{S \sim D} \left[ R(h_S) - \inf_{h \in H} R(h) \leq \epsilon \right] \geq 1 - \delta,$$

- for samples $S$ of size $m = poly(1/\epsilon, 1/\delta)$ for a fixed polynomial.
VCDim Lower Bound - Non-Realizable Case

(Anthony and Bartlett, 1999)

**Theorem:** let $H$ be a hypothesis set with VC dimension $d > 1$. Then, for any learning algorithm $L$,

$$\exists D \text{ over } X \times \{0, 1\}, \quad \Pr_{S \sim D^m} \left[ R_D(h_S) - \inf_{h \in H} R_D(h) > \sqrt{\frac{d}{320m}} \right] \geq 1/64.$$

Equivalently, for any learning algorithm, the sample complexity verifies

$$m \geq \frac{d}{320\epsilon^2}.$$
References


References


