Where did the $|\mathcal{H}|$ come from?

- The Bad events $\mathcal{B}_m$:
  - $|E_{\text{tr}}(h_m) - E(h_m)| > \epsilon$ with probability $\leq 2e^{-2\epsilon^2 N}$
Where did the $|\mathcal{H}|$ come from?

- The Bad events $\mathcal{B}_m$:
  - $|E_{tr}(h_m) - E(h_m)| > \epsilon$ with probability $\leq 2e^{-2\epsilon^2N}$

- The union bound:
  \[
  \mathbb{P}[\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \ldots \text{ or } \mathcal{B}_M] \leq \mathbb{P}[\mathcal{B}_1] + \mathbb{P}[\mathcal{B}_2] + \ldots + \mathbb{P}[\mathcal{B}_M] \leq 2 |\mathcal{H}| e^{-2\epsilon^2N}
  \]

  - consider worst case: no overlaps

![Diagram showing the Bad events $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ with varying overlaps.](image)

  - No overlap: bound is tight
  - Large overlap
Theory of Generalization
A simple solution

• For each particular \( h \),
  \[
P[| E_{tr}(h) - E(h) | > \epsilon] \leq 2e^{-2\epsilon^2N}
\]

• If we have a hypothesis set \( \mathcal{H} \), we want to derive the bound for \( P[\sup_{h \in \mathcal{H}} | E_{tr}(h) - E(h) | > \epsilon] \)
  \[
P[| E_{tr}(h_1) - E(h_1) | > \epsilon] \text{ or ... or } P[| E_{tr}(h|\mathcal{H}|) - E(h|\mathcal{H}|) | > \epsilon]
\]
  \[
  \leq \sum_{m=1}^{\mathcal{H}} P[| E_{tr}(h_m) - E(h_m) |] \leq 2|\mathcal{H}| e^{-2\epsilon^2N}
\]

• Because of union bound inequality \( P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i) \)
Uniform convergence

- When our learning algorithm $\mathcal{A}$ picks the hypothesis $g$:
  - $P[\exists h \in \mathcal{H} \mid E_{tr}(h) - E(h) > \epsilon] \leq 2 |\mathcal{H}| e^{-2\epsilon^2N}$
- Subtract both sides from 1
  - $P[\neg \exists h \in \mathcal{H} \mid E_{tr}(h) - E(h) > \epsilon] = P[\forall h \in \mathcal{H} \mid E_{tr}(h) - E(h) \leq \epsilon]$
    - $\geq 1 - 2 |\mathcal{H}| e^{-2\epsilon^2N}$
What uniform convergence tell us?

\[ P[ \neg \exists h \in \mathcal{H} \mid E_{tr}(h) - E(h) > \epsilon] = P[ \forall h \in \mathcal{H} \mid E_{tr}(h) - E(h) \leq \epsilon] \]

\[ \geq 1 - 2 |\mathcal{H}| e^{-2\epsilon^2 N} \]

- Given \( \epsilon \) and some \( \delta > 0 \), how large must \( N \) be before we can guarantee that with probability at least \( 1 - \delta \), training error will be within \( \epsilon \) of generalization error?

- Set \( \delta = 2 |\mathcal{H}| e^{-2\epsilon^2 N} \), solve \( N \)

\[ N \geq \frac{1}{2\epsilon^2} \log \frac{2 |\mathcal{H}|}{\delta} \]

- The training set size \( N \) that a certain method or algorithm requires in order to achieve a certain level of performance is also called the algorithm’s sample complexity.
What uniform convergence tell us?

\[ P[\neg \exists h \in \mathcal{H} \mid E_{tr}(h) - E(h) > \epsilon] = P[\forall h \in \mathcal{H} \mid E_{tr}(h) - E(h) \leq \epsilon] \]

• 
\[ \geq 1 - 2|\mathcal{H}|e^{-2\epsilon^2N} \]

• Given \( N \) and some \( \delta \), we have

  • \( |E_{tr}(h) - E(h)| \leq \sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}} \)

  • i.e \( |E_{tr}(h) - E(h)| \leq \gamma \) for all \( h \in \mathcal{H} \)
What uniform convergence tell us?

\[ P[\neg \exists h \in \mathcal{H} \mid E_{tr}(h) - E(h) > \epsilon] = P[\forall h \in \mathcal{H} \mid E_{tr}(h) - E(h) \leq \epsilon] \]

\[ \geq 1 - 2|\mathcal{H}|e^{-2\epsilon^2N} \]

• Given \( N \) and some \( \delta \), we have

\[ |E_{tr}(h) - E(h)| \leq \sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}} \]

• i.e \( |E_{tr}(h) - E(h)| \leq \gamma \) for all \( h \in \mathcal{H} \)

• What about the best hypothesis in training data?
What uniform convergence tell us?

- Given $N$ and some $\delta$, we have

  \[ |E_{tr}(h) - E(h)| \leq \sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}} \]

- i.e. $|E_{tr}(h) - E(h)| \leq \gamma$ for all $h \in \mathcal{H}$

- What about the best hypothesis in training data? $\hat{h} = \arg\min_{h \in \mathcal{H}} E_{tr}(h)$

- Define the best hypothesis as $h^* = \arg\min_{h \in \mathcal{H}} E(h)$

- We have $E(\hat{h}) \leq E_{tr}(\hat{h}) + \gamma \leq E_{tr}(h^*) + \gamma \leq E(h^*) + 2\gamma$
What uniform convergence tell us?

- What about the best hypothesis in training data? \( \hat{h} = \arg\min_{h \in \mathcal{H}} E_{tr}(h) \)

- Define the best hypothesis as \( h^* = \arg\min_{h \in \mathcal{H}} E(h) \)
  - We have \( E(\hat{h}) \leq E_{tr}(\hat{h}) + \gamma \leq E_{tr}(h^*) + \gamma \leq E(h^*) + 2\gamma \)
  - So we have
    \[
    E(\hat{h}) \leq \left( \min_{h \in \mathcal{H}} E(h) \right) + 2\sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}}
    \]
    - Connection with bias/variance tradeoff
What uniform convergence tell us?

• What about the best hypothesis in training data? \( \hat{h} = \arg \min_{h \in \mathcal{H}} E_{tr}(h) \)

• Define the best hypothesis as \( h^* = \arg \min_{h \in \mathcal{H}} E(h) \)
  - We have \( E(\hat{h}) \leq E_{tr}(\hat{h}) + \gamma \leq E_{tr}(h^*) + \gamma \leq E(h^*) + 2\gamma \)
  - So we have
    \[
    E(\hat{h}) \leq (\min_{h \in \mathcal{H}} E(h)) + 2\sqrt{\frac{1}{2N} \log \frac{2|\mathcal{H}|}{\delta}}
    \]
    - Connection with bias/variance tradeoff
  - Further, given \( \epsilon \) and some \( \delta > 0 \), is suffices that
    \[
    N \geq \frac{1}{2\epsilon^2} \log \frac{2|\mathcal{H}|}{\delta} = O\left(\frac{1}{\epsilon^2} \log \frac{|\mathcal{H}|}{\delta}\right)
    \]
Can we improve on $|\mathcal{H}|$?
Can we improve on $|\mathcal{H}|$?
Can we improve on $|\mathcal{H}|$?
Can we improve on $|\mathcal{H}|$?

- The event that $|E_{tr}(h_1) - E(h_1)| > \epsilon$ and $|E_{tr}(h_2) - E(h_2)| > \epsilon$ are largely overlapped.
What can we replace $\mathcal{H}$ with?

- Instead of the whole input space
What can we replace $\mathcal{H}$ with?

- Instead of the whole input space
- Let's consider a finite set of input points
What can we replace \( \mathcal{H} \) with?

- Instead of the whole input space
- Let’s consider a finite set of input points
- How many patterns of colors can you get?
Dichotomies: mini-hypotheses

- A hypothesis: $h : \mathcal{X} \rightarrow \{-1, +1\}$
- A dichotomy: $h : \{x_1, x_2, \ldots, x_N\} \rightarrow \{-1, +1\}$
Dichotomies: mini-hypotheses

- A hypothesis: \( h : \mathcal{X} \rightarrow \{-1, +1\} \)
- A dichotomy: \( h : \{x_1, x_2, \ldots, x_N\} \rightarrow \{-1, +1\} \)
- Number of hypotheses \(|\mathcal{H}|\) can be infinite
- Number of dichotomies \(|\mathcal{H}(x_1, x_2, \ldots, x_N)|\) at most \(2^N\)
Dichotomies: mini-hypotheses

• A hypothesis: $h : \mathcal{X} \rightarrow \{-1, +1\}$

• A dichotomy: $h : \{x_1, x_2, \ldots, x_N\} \rightarrow \{-1, +1\}$

• Number of hypotheses $|\mathcal{H}|$ can be infinite

• Number of dichotomies $|\mathcal{H}(x_1, x_2, \ldots, x_N)|$ at most $2^N$

  • $\Rightarrow$ Candidate for replacing $|\mathcal{H}|$

• Why?
Theory of Generalization

Symmetrization lemma

• Imagine we have the ghost dataset $S'$ with also size $N$:

\[ P[\sup_{h \in \mathcal{H}} | E_{tr}(h) - E(h) | > \epsilon] \leq 2P[\sup_{h \in \mathcal{H}} | E_{tr}(h) - E'_{tr}(h) | > \frac{\epsilon}{2}] \]
Theory of Generalization

Growth function

- Imagine we have the ghost dataset $S'$ with also size $N$:
  
  \[ P[\sup_{h \in \mathcal{H}} |E_{tr}(h) - E(h)| > \epsilon] \leq 2P[\sup_{h \in \mathcal{H}} |E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}] \]

- By union bound:
  
  \[ P[\sup_{h \in \mathcal{H}_{S \cup S'}} |E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}] \leq |\mathcal{H}_{S \cup S'}| P[|E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}] \]
Theory of Generalization

Growth function

- Imagine we have the ghost dataset $S'$ with also size $N$:
  
  \[ P[\sup_{h \in \mathcal{H}} |E_{tr}(h) - E(h)| > \epsilon] \leq 2P[\sup_{h \in \mathcal{H}} |E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}] \]

- By union bound:
  
  \[ P[\sup_{h \in \mathcal{H}_{S \cup S'}} |E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}] \leq |\mathcal{H}_{S \cup S'}| P[|E_{tr}(h) - E'_{tr}(h)| > \frac{\epsilon}{2}] \]

- How to bound $|\mathcal{H}_{S \cup S'}|$
Theory of Generalization

Deduce the dimension

• Why do we need to consider every possible hypothesis?

• $P[\sup_{h \in \mathcal{H}} |E_{tr}(h) - E(h)| > \epsilon]$

• If we omit one hypothesis, we might miss the biggest gap

• However, are the events of each hypothesis having a big generalization gap are likely to be independent?

• No
The growth function

• The growth function counts the most dichotomies on any N points:
  
  \[ m_{\mathcal{H}}(N) = \max_{x_1, \ldots, x_N \in \mathcal{X}} |\mathcal{H}(x_1, \ldots, x_N)| \]
The growth function

- The growth function counts the most dichotomies on any $N$ points:

  $$m_{\mathcal{H}}(N) = \max_{x_1, \ldots, x_N \in \mathcal{X}} |\mathcal{H}(x_1, \ldots, x_N)|$$

- The growth function satisfies:

  $$m_{\mathcal{H}}(N) \leq 2^N$$
Growth function for linear classifiers

- Compute $m_{\mathcal{H}}(3)$ in 2-D space

- What’s $|\mathcal{H}(x_1, x_2, x_3)|$?
Growth function for linear classifiers

- Compute $m_{\mathcal{H}}(3)$ in 2-D space when $\mathcal{H}$ is perceptron (linear hyperplanes)

$$m_{\mathcal{H}}(3) = 8$$
Growth function for linear classifiers

- Compute $m_{\mathcal{H}}(3)$ in 2-D space when $\mathcal{H}$ is perceptron (linear hyperplanes)
Growth function for linear classifiers

- Compute $m_{\mathcal{H}}(3)$ in 2-D space when $\mathcal{H}$ is perceptron (linear hyperplanes)

- Doesn’t matter because we only counts the most dichotomies
Growth function for linear classifier

• What’s $m_\mathcal{H}(4)$?
Growth function for linear classifier

- What’s $m_{\mathcal{H}}(4)$?
- (At least) missing two dichotomies:
Growth function for linear classifier

• What’s $m_{\mathcal{H}}(4)$?

• (At least) **missing** two dichotomies:

- $m_{\mathcal{H}}(4) = 14 < 2^4$
Example I: positive rays

\[ h(x) = -1 \quad \text{for} \quad x_1, x_2, x_3, \ldots \]

\[ h(x) = +1 \quad \text{for} \quad a, x_N \]

\[ \mathcal{H} \text{ is set of } h : \mathbb{R} \rightarrow \{-1, +1\} \]

\[ h(x) = \text{sign}(x - a) \]

\[ m_{\mathcal{H}}(N) = N + 1 \]
Example II: positive intervals

\[ h(x) = -1 \]
\[ h(x) = +1 \]
\[ h(x) = -1 \]

\[ x_1 \quad x_2 \quad x_3 \quad \ldots \quad x_N \]

\( \mathcal{H} \) is set of \( h : \mathbb{R} \rightarrow \{-1, +1\} \)

Place interval ends in two of \( N + 1 \) spots

\[ m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1 \]
Example III: convex sets

• $\mathcal{H}$ is set of $h : \mathbb{R}^2 \rightarrow \{-1, +1\}$

  • $h(x) = +1$ is convex

• How many dichotomies can we generate?
Example III: convex sets

- $\mathcal{H}$ is set of $h : \mathbb{R}^2 \rightarrow \{-1, +1\}$
  - $h(x) = +1$ is convex
- How many dichotomies can we generate?
Example III: convex sets

- $\mathcal{H}$ is set of $h : \mathbb{R}^2 \rightarrow \{-1, +1\}$
  - $h(x) = +1$ is convex
- How many dichotomies can we generate?
Example III: convex sets

- $\mathcal{H}$ is set of $h : \mathbb{R}^2 \rightarrow \{-1, +1\}$
  - $h(x) = +1$ is convex
- $m_\mathcal{H}(N) = 2^N$ for any $N \Rightarrow$ We say the $N$ points are “shattered” by $h$
Shattered

- Given a set $S = \{x^{(i)}, \ldots, x^{(d)}\}$ (no relation to the training set) of points $x^{(i)} \in \mathcal{X}$, we say that $\mathcal{H}$ shatters $S$ if $\mathcal{H}$ can realize any labeling on $S$. I.e., if for any set of labels $\{y^{(i)}, \ldots, y^{(d)}\}$, there exist some $h \in \mathcal{H}$ so that $h(x^{(i)}) = y^{(i)}$ for all $i = 1, \ldots, d$
The 3 growth functions

- $\mathcal{H}$ is positive rays:
  - $m_\mathcal{H}(N) = N + 1$

- $\mathcal{H}$ is positive intervals:
  - $m_\mathcal{H}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$

- $\mathcal{H}$ is convex sets:
  - $m_\mathcal{H}(N) = 2^N$
What’s next?

• Remember the inequality

\[ \mathbb{P}[|E_{\text{in}} - E_{\text{out}}| > \epsilon] \leq 2|\mathcal{H}| e^{-2\epsilon^2 N} \]

• What happens if we replace \(|\mathcal{H}|\) by \(m_{\mathcal{H}}(N)\)

• \(m_{\mathcal{H}}(N)\) polynomial \(\Rightarrow\) Good!
What’s next?

• Remember the inequality
  
  \[ \mathbb{P}[|E_{\text{tr}} - E| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2N} \]

• What happens if we replace $|\mathcal{H}|$ by $m_{\mathcal{H}}(N)$
  
  • $m_{\mathcal{H}}(N)$ polynomial $\Rightarrow$ Good!
  
  • Why?

• How to show $m_{\mathcal{H}}(N)$ is polynomial?