Reducing $M$ to finite number
Where did the $M$ come from?

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

- The Bad events $B_m$:
  
  \[ \left| E_{tr}(h_m) - E(h_m) \right| > \epsilon \]  
  with probability $\leq 2e^{-2\epsilon^2 N}$

- The union bound:

  \[
  P[B_1 \text{ or } B_2 \text{ or } \cdots \text{ or } B_M] 
  \leq P[B_1] + P[B_2] + \cdots + P[B_M] 
  \leq 2M e^{-2\epsilon^2 N}
  \]

  consider worst case: no overlaps

---

**Diagram:**

- No overlap: bound is tight
- Large overlap
Can we improve on $M$?
Can we improve on $M$?
Can we improve on $M$?
Can we improve on $M$?

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 $M$ with?

Instead of the whole input space
What can we replace $M$ with?

Instead of the whole input space
Let’s consider a finite set of input points
What can we replace $M$ 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: $h : \{x_1, x_2, \cdots, x_N\} \rightarrow \{-1, +1\}$

Number of hypotheses $|H|$ can be infinite
Number of dichotomies $|H(x_1, x_2, \cdots, x_N)|$:

$\Rightarrow$ Candidate for replacing $M$
Dichotomies: mini-hypotheses

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

- A hypothesis: \( h : \mathcal{X} \rightarrow \{-1, +1\} \)
- A dichotomy: \( h : \{x_1, x_2, \cdots, x_N\} \rightarrow \{-1, +1\} \)
- Number of hypotheses \( |\mathcal{H}| \) can be infinite
- Number of dichotomies \( |\mathcal{H}(x_1, x_2, \cdots, x_N)| \): at most \( 2^N \)
  \( \Rightarrow \) Candidate for replacing \( M \)
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 classifiers

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

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

- What's $m_{\mathcal{H}}(4)$?
- (At least) missing two dichotomies:

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

$h(x) = -1$

$x_1 \ x_2 \ x_3 \ \ldots$

$a$  \hfill $h(x) = +1$

$x_N$

$\mathcal{H}$ 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

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

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

\[ \mathcal{H} \] 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 $h : \mathbb{R}^2 \to \{-1, +1\}$

  $h(x) = +1$ is convex

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

- $\mathcal{H}$ is set of $h : \mathbb{R}^2 \to \{-1, +1\}$
  
  $h(x) = +1$ is convex

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

- $\mathcal{H}$ is set of $h : \mathbb{R}^2 \to \{-1, +1\}$
  
  $h(x) = +1$ 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 $h$
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

\[ \mathbb{P}[|E_{in} - E_{out}| > \epsilon] \leq 2M e^{-2\epsilon^2 N} \]
What’s next?

• Remember the inequality

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

• What happens if we replace \( M \) by \( m_H(N) \)?
  \( m_H(N) \) polynomial \( \Rightarrow \) Good!
What’s next?

- Remember the inequality

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

- What happens if we replace \( M \) by \( m_{\mathcal{H}}(N) \)?

  \( m_{\mathcal{H}}(N) \) polynomial \( \Rightarrow \) Good!

- How to show \( m_{\mathcal{H}}(N) \) is polynomial?
When will $m_H(N)$ be polynomial
Break point of $\mathcal{H}$

- If no data set of size $k$ can be shattered by $\mathcal{H}$, then $k$ is a break point for $\mathcal{H}$
  \[ m_{\mathcal{H}}(k) < 2^k \]

- VC dimension of $\mathcal{H}$: $k - 1$ (the most points $\mathcal{H}$ can shatter)
Break point of $\mathcal{H}$

- If no data set of size $k$ can be shattered by $\mathcal{H}$, then $k$ is a break point for $\mathcal{H}$
  
  $$m_{\mathcal{H}}(k) < 2^k$$

- VC dimension of $\mathcal{H}$: $k - 1$ (the most points $\mathcal{H}$ can shatter)

- For 2-D perceptron: $k = 4$, VC dimension = 3
Break point - examples

- Positive rays: $m_\mathcal{H}(N) = N + 1$
  
  Break point $k = 2$, $d_{VC} = 1$
Break point - examples

- **Positive rays:** $m_{\mathcal{H}}(N) = N + 1$
  
  Break point $k = 2$, $d_{VC} = 1$

- **Positive intervals:** $m_{\mathcal{H}}(N) = \frac{1}{2} N^2 + \frac{1}{2} N + 1$
  
  Break point $k = 3$, $d_{VC} = 2$
Break point - examples

- Positive rays: \( m_\mathcal{H}(N) = N + 1 \)
  Break point \( k = 2, \ d_{VC} = 1 \)
- Positive intervals: \( m_\mathcal{H}(N) = \frac{1}{2} N^2 + \frac{1}{2} N + 1 \)
  Break point \( k = 3, \ d_{VC} = 2 \)
- Convex set: \( m_\mathcal{H}(N) = 2^N \)
  Break point \( k = \infty, \ d_{VC} = \infty \)
We will show

No break point $\Rightarrow m_{\mathcal{H}}(N) = 2^N$

Any break point $\Rightarrow m_{\mathcal{H}}(N)$ is polynomial in $N$
Puzzle

- Break point is $k = 2$
Puzzle

- Break point is $k = 2$
Puzzle

- Break point is $k = 2$
Puzzle

- Break point is $k = 2$
Puzzle

- Break point is $k = 2$
Break point is $k = 2$
**Puzzle**

- Break point is \( k = 2 \)

<table>
<thead>
<tr>
<th></th>
<th>( x_i )</th>
<th>( x_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( x_2 )</td>
<td>( x_3 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The graph on the left shows an incorrect pairing, indicated by the red X.
- The table on the right correctly assigns the values for \( x_1, x_2, \) and \( x_3 \).
Puzzle

- Break point is $k = 2$
Puzzle

- Break point is $k = 2$
Puzzle

- Break point is $k = 2$
Puzzle

- Break point is $k = 2$
Puzzle

- Break point is $k = 2$
Break point is $k = 2$
Bounding $m_{\mathcal{H}}(N)$

- Key quantity:
  
  $B(N, k)$: Maximum number of dichotomies on $N$ points, with break point $k$
Bounding $m_{\mathcal{H}}(N)$

- Key quantity:
  \[ B(N, k) \]: Maximum number of dichotomies on $N$ points, with break point $k$
- If the hypothesis space has break point $k$, then
  \[ m_{\mathcal{H}}(N) \leq B(N, k) \]
Recursive bound on $B(N, k)$

- For any “valid” set of dichotomies, reorganize rows by
  - $S_1$: pattern of $x_1, \cdots, x_{N-1}$ only appears once
  - $S_2^+, S_2^-$: pattern of $x_1, \cdots, x_{N-1}$ appears twice

<table>
<thead>
<tr>
<th># of rows</th>
<th>$S_1$</th>
<th>$S_2^+$</th>
<th>$S_2$</th>
<th>$S_2^-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td></td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>+1</td>
</tr>
</tbody>
</table>

$B(N, k) = \alpha + 2\beta$
Recursive bound on $B(N, k)$

- Focus on $x_1, x_2, \ldots, x_{N-1}$ columns:
  
  $\alpha + \beta \leq B(N - 1, k)$

<table>
<thead>
<tr>
<th></th>
<th># of rows</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_N - 1$</th>
<th>$x_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>$\alpha$</td>
<td>+1</td>
<td>+1</td>
<td>$\ldots$</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−1</td>
<td>+1</td>
<td>$\ldots$</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+1</td>
<td>−1</td>
<td>$\ldots$</td>
<td>−1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−1</td>
<td>+1</td>
<td>$\ldots$</td>
<td>−1</td>
</tr>
</tbody>
</table>

| $S_2^+$ | $\beta$ | +1    | −1    | $\ldots$ | +1    | +1 |
|         |           | −1    | −1    | $\ldots$ | +1    | +1 |
|         |           | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
|         |           | +1    | −1    | $\ldots$ | +1    | +1 |
|         |           | −1    | −1    | $\ldots$ | −1    | +1 |

| $S_2^-$ | $\beta$ | +1    | −1    | $\ldots$ | +1    | −1 |
|         |           | −1    | −1    | $\ldots$ | +1    | −1 |
|         |           | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
|         |           | +1    | −1    | $\ldots$ | −1    | −1 |
|         |           | −1    | −1    | $\ldots$ | −1    | −1 |
Recursive bound on $B(N, k)$

- Now focus on the $S_2 = S^+_2 \cup S^- + 2$ rows
  $\beta \leq B(N - 1, k - 1)$

<table>
<thead>
<tr>
<th></th>
<th># of rows</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>...</th>
<th>$x_{N-1}$</th>
<th>$x_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>$\alpha$</td>
<td>+1</td>
<td>+1</td>
<td>...</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1</td>
<td>+1</td>
<td>...</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>:</td>
<td>:</td>
<td></td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+1</td>
<td>-1</td>
<td>...</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1</td>
<td>+1</td>
<td>...</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>$S^+_2$</td>
<td>$\beta$</td>
<td>+1</td>
<td>-1</td>
<td>...</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1</td>
<td>-1</td>
<td>...</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>:</td>
<td>:</td>
<td></td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+1</td>
<td>-1</td>
<td>...</td>
<td>+1</td>
<td>+1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1</td>
<td>-1</td>
<td>...</td>
<td>-1</td>
<td>+1</td>
</tr>
<tr>
<td>$S^-_2$</td>
<td>$\beta$</td>
<td>+1</td>
<td>-1</td>
<td>...</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1</td>
<td>-1</td>
<td>...</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>:</td>
<td>:</td>
<td></td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+1</td>
<td>-1</td>
<td>...</td>
<td>+1</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1</td>
<td>-1</td>
<td>...</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>
Recursive bound on $B(N, k)$

$$B(N, k) = \alpha + \beta + \beta$$

$$\leq B(N - 1, k) + B(N - 1, k - 1)$$

What’s the upper bound for $B(N, k)$?
Recursive bound on $B(N, k)$

$$B(N, k) = \alpha + \beta + \beta \leq B(N - 1, k) + B(N - 1, k - 1)$$
Recursive bound on $B(N, k)$

\[ B(N, k) = \alpha + \beta + \beta \leq B(N - 1, k) + B(N - 1, k - 1) \]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>......</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Recursive bound on $B(N, k)$

$$B(N, k) = \alpha + \beta + \beta$$

$$\leq B(N - 1, k) + B(N - 1, k - 1)$$
Recursive bound on $B(N, k)$

\[
B(N, k) = \alpha + \beta + \beta \\
\leq B(N - 1, k) + B(N - 1, k - 1)
\]

<table>
<thead>
<tr>
<th>N</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>......</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>......</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>..</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Recursive bound on $B(N, k)$

\[
B(N, k) = \alpha + \beta + \beta \\
\leq B(N - 1, k) + B(N - 1, k - 1)
\]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>......</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>......</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>......</td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Recursive bound on $B(N, k)$

\[ B(N, k) = \alpha + \beta + \beta \leq B(N - 1, k) + B(N - 1, k - 1) \]
Analytic solution for $B(N, k)$ bound

$B(N, k)$ is upper bounded by $C(N, k)$:

\[ C(N, 1) = 1, \quad N = 1, 2, \cdots \]
\[ C(1, k) = 2, \quad k = 2, 3, \cdots \]
\[ C(N, k) = C(N - 1, k) + C(N - 1, k - 1) \]

- Theorem: $C(N, k) = \sum_{i=0}^{k-1} \binom{N}{i}$
Analytic solution for $B(N, k)$ bound

$B(N, k)$ is upper bounded by $C(N, k)$:

\[
C(N, 1) = 1, \quad N = 1, 2, \cdots \\
C(1, k) = 2, \quad k = 2, 3, \cdots \\
C(N, k) = C(N - 1, k) + C(N - 1, k - 1)
\]

- Theorem: $C(N, k) = \sum_{i=0}^{k-1} \binom{N}{i}$
- Boundary conditions: (easy to check)
Analytic solution for $B(N, k)$ bound

$B(N, k)$ is upper bounded by $C(N, k)$:

$$C(N, 1) = 1, \ N = 1, 2, \cdots$$

$$C(1, k) = 2, \ k = 2, 3, \cdots$$

$$C(N, k) = C(N - 1, k) + C(N - 1, k - 1)$$

- **Theorem:** $C(N, k) = \sum_{i=0}^{k-1} \binom{N}{i}$
- **Boundary conditions:** (easy to check)
- **Induction:**

\[
\sum_{i=0}^{k-1} \binom{N}{i} = \sum_{i=0}^{k-1} \binom{N - 1}{i} + \sum_{i=0}^{k-2} \binom{N - 1}{i}
\]

- select $< k$ from $N$ items
- $N$-th item not chosen
- $N$-th item chosen
It is polynomial!

For a given $\mathcal{H}$, the break point $k$ is fixed:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Polynomial with degree $k - 1$
It is polynomial!

- For a given $\mathcal{H}$, the break point $k$ is fixed:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Polynomial with degree $k$

- $\mathcal{H}$ is positive rays: (break point $k = 2$)

$$m_{\mathcal{H}}(N) = N + 1$$
It is polynomial!

- For a given $\mathcal{H}$, the break point $k$ is fixed:

\[ m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i} \]

Polynomial with degree $k$

- $\mathcal{H}$ is 2D perceptrons: (break point $k = 4$)

\[ m_{\mathcal{H}}(N) =? \]
For a given $\mathcal{H}$, the break point $k$ is fixed:

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Polynomial with degree $k$

$\mathcal{H}$ is 2D perceptrons: (break point $k = 4$)

$$m_{\mathcal{H}}(N) \leq \frac{1}{6}N^3 + \frac{5}{6}N + 1$$
Replace $M$ by $m_\mathcal{H}(N)$

- Original bound:
\[
P[\exists h \in \mathcal{H} \text{ s.t. } |E_{tr}(h) - E(h)| > \epsilon] \leq 2M e^{-2\epsilon^2N}
\]

- Replace $M$ by $m_\mathcal{H}(N)$
\[
P[\exists h \in \mathcal{H} \text{ s.t. } |E_{tr}(h) - E(h)| > \epsilon] \leq 2 \cdot 2m_\mathcal{H}(2N) \cdot e^{-\frac{1}{8} \epsilon^2 N}
\]

Vapnik-Chervonenkis (VC) bound
VC Dimension
The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\text{VC}}(\mathcal{H})$, is the largest value of $N$ for which $m_\mathcal{H}(N) = 2^N$

“the most points $\mathcal{H}$ can shatter”
Definition

- The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{VC}(\mathcal{H})$, is the largest value of $N$ for which $m_\mathcal{H}(N) = 2^N$
  
  "the most points $\mathcal{H}$ can shatter"

- $N \leq d_{VC}(\mathcal{H}) \Rightarrow \mathcal{H}$ can shatter $N$ points
Definition

- The VC dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{VC}(\mathcal{H})$, is the largest value of $N$ for which $m_{\mathcal{H}}(N) = 2^N$

  "the most points $\mathcal{H}$ can shatter"

- $N \leq d_{VC}(\mathcal{H}) \Rightarrow \mathcal{H}$ can shatter $N$ points
- $k > d_{VC}(\mathcal{H}) \Rightarrow \mathcal{H}$ cannot be shattered
- The smallest **break point** is 1 above VC-dimension
The growth function

- In terms of a break point $k$:
  \[ m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i} \]

- In terms of the VC dimension $d_{\text{VC}}$:
  \[ m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{\text{VC}}} \binom{N}{i} \]
VC dimension of linear classifiers

- For $d = 2$, $d_{VC} = 3$
VC dimension of linear classifiers

- For $d = 2$, $d_{VC} = 3$
- What if $d > 2$?
VC dimension of linear classifiers

- For $d = 2$, $d_{VC} = 3$
- What if $d > 2$?

In general,

$$d_{VC} = d + 1$$
For $d = 2$, $d_{VC} = 3$

What if $d > 2$?

In general,

$$d_{VC} = d + 1$$

We will prove $d_{VC} \geq d + 1$ and $d_{VC} \leq d + 1$
VC dimension of linear classifiers

To prove $d_{VC} \geq d + 1$
VC dimension of linear classifiers

- To prove $d_{VC} \geq d + 1$
- A set of $N = d + 1$ points in $\mathbb{R}^d$ shattered by the linear hyperplane

\[
X = \begin{bmatrix}
- x_1^T \\
- x_2^T \\
- x_3^T \\
\vdots \\
- x_{d+1}^T
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 & \ldots & 0 \\
1 & 1 & 0 & \ldots & 0 \\
1 & 0 & 1 & 0 \\
\vdots & \vdots & \vdots & 0 \\
1 & 0 & \ldots & 0 & 1
\end{bmatrix}
\]
VC dimension of linear classifiers

- To prove $d_{VC} \geq d + 1$
- A set of $N = d + 1$ points in $\mathbb{R}^d$ shattered by the linear hyperplane

$$X = \begin{bmatrix}
-x_1^T \\
-x_2^T \\
-x_3^T \\
\vdots \\
-x_{d+1}^T
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 & \ldots & 0 \\
1 & 1 & 0 & \ldots & 0 \\
1 & 0 & 1 & & 0 \\
\vdots & & \ddots & \ddots & \ddots \\
1 & 0 & \ldots & 0 & 1
\end{bmatrix}$$

- $X$ is invertible!
Can we shatter the dataset?

For any \( y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix} \), can we find \( w \) satisfying

\[
\text{sign}(Xw) = y
\]
Can we shatter the dataset?

For any \( y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix} \), can we find \( w \) satisfying

\[
\text{sign}(Xw) = y
\]

Easy! Just set \( w = X^{-1}y \)
Can we shatter the dataset?

For any \( y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix} \), can we find \( w \) satisfying

\[
\text{sign}(Xw) = y
\]

Easy! Just set \( w = X^{-1}y \)

So, \( d_{VC} \geq d + 1 \)
VC dimension of linear classifiers

To show $d_{VC} \leq d + 1$, we need to show

We cannot shatter any set of $d + 2$ points
VC dimension of linear classifiers

• To show $d_{VC} \leq d + 1$, we need to show

  We cannot shatter any set of $d + 2$ points

• For any $d + 2$ points

  \[ x_1, x_2, \ldots, x_{d+1}, x_{d+2} \]

• More points than dimensions $\Rightarrow$ linear dependent

  \[ x_j = \sum_{i \neq j} a_i x_i \]

  where not all $a_i$’s are zeros
VC dimension of linear classifiers

\[ x_j = \sum_{i \neq j} a_i x_i \]

Now we construct a dichotomy that cannot be generated:

\[ y_i = \begin{cases} 
\text{sign}(a_i) & \text{if } i \neq j \\
-1 & \text{if } i = j 
\end{cases} \]
VC dimension of linear classifiers

\[ x_j = \sum_{i \neq j} a_i x_i \]

- Now we construct a dichotomy that cannot be generated:
  \[ y_i = \begin{cases} 
  \text{sign}(a_i) & \text{if } i \neq j \\
  -1 & \text{if } i = j 
  \end{cases} \]

- For all \( i \neq j \), assume the labels are correct: \( \text{sign}(a_i) = \text{sign}(w^T x_i) \)
  \[ \Rightarrow a_i w^T x_i > 0 \]
VC dimension of linear classifiers

\[ x_j = \sum_{i \neq j} a_i x_i \]

- Now we construct a dichotomy that cannot be generated:

\[ y_i = \begin{cases} 
\text{sign}(a_i) & \text{if } i \neq j \\
-1 & \text{if } i = j 
\end{cases} \]

- For all \( i \neq j \), assume the labels are correct: \( \text{sign}(a_i) = \text{sign}(w^T x_i) \)
  \[ \Rightarrow a_i w^T x_i > 0 \]

- For \( j \)-th data,

\[ w^T x_j = \sum_{i \neq j} a_i w^T x_i > 0 \]

- Therefore, \( y_j = \text{sign}(w^T x_j) = +1 \) (cannot be \(-1\))
We proved for $d$-dimensional linear hyperplane

\[ d_{VC} \leq d + 1 \text{ and } d_{VC} \geq d + 1 \implies d_{VC} = d + 1 \]
Putting it together

- We proved for $d$-dimensional linear hyperplane

\[ d_{VC} \leq d + 1 \quad \text{and} \quad d_{VC} \geq d + 1 \Rightarrow \quad d_{VC} = d + 1 \]

- Number of parameters $w_0, \cdots, w_d$

  $d + 1$ parameters!
Putting it together

- We proved for \( d \)-dimensional linear hyperplane

\[
d_{VC} \leq d + 1 \text{ and } d_{VC} \geq d + 1 \Rightarrow d_{VC} = d + 1
\]

- Number of parameters \( w_0, \cdots, w_d \)

\( d + 1 \) parameters!

- Parameters create degrees of freedom
Examples

- Positive rays: 1 parameters, $d_{\text{VC}} = 1$

\[ h(x) = -1 \quad a \quad h(x) = +1 \]
**Examples**

- **Positive rays:** 1 parameters, $d_{\text{VC}} = 1$
  
  $h(x) = -1$  \hspace{1cm} \hspace{1cm} $h(x) = +1$
  
  \hspace{1cm} \hspace{1cm} $a$

- **Positive intervals:** 2 parameters, $d_{\text{VC}} = 2$
  
  $h(x) = -1$  \hspace{1cm} $h(x) = +1$  \hspace{1cm} $h(x) = -1$
Examples

- Positive rays: 1 parameters, $d_{VC} = 1$

- Positive intervals: 2 parameters, $d_{VC} = 2$

- Not always true • • •

  $d_{VC}$ measures the effective number of parameters
Number of data points needed

\[ \Pr[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 4m_H(2N)e^{-\frac{1}{8}\epsilon^2N^2} \]

- If we want certain \( \epsilon \) and \( \delta \), how does \( N \) depend on \( d_{VC} \)?
Number of data points needed

\[ P[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 4m_H(2N)e^{-\frac{1}{8}\epsilon^2N} \]

- If we want certain \( \epsilon \) and \( \delta \), how does \( N \) depend on \( d_{\text{VC}} \)?
- Need \( N^d e^{-N} = \text{small value} \)
Number of data points needed

\[ P[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2N} \]

- If we want certain \( \epsilon \) and \( \delta \), how does \( N \) depend on \( d_{\text{VC}} \)?
- Need \( N^d e^{-N} = \text{small value} \)

\( N \) is almost linear with \( d_{\text{VC}} \)
Regularization
The polynomial model

- $\mathcal{H}_Q$: polynomials of order $Q$

$$\mathcal{H}_Q = \{ \sum_{q=0}^{Q} w_q L_q(x) \}$$

- Linear regression in the $\mathcal{Z}$ space with

$$z = [1, L_1(x), \cdots, L_Q(x)]$$
Unconstrained solution

- Input \((x_1, y_1), \cdots, (x_N, y_N) \rightarrow (z_1, y_1), \cdots, (z_N, y_N)\)
- Linear regression:

\[
\text{Minimize } E_{tr}(w) = \frac{1}{N} \sum_{n=1}^{N} (w^T z_n - y_n)^2
\]

\[
\text{Minimize } \frac{1}{N} (Zw - y)^T (Zw - y)
\]

- Solution \(w_{tr} = (Z^T Z)^{-1} Z^T y\)
Constraining the weights

- Hard constraint: $\mathcal{H}_2$ is constrained version of $\mathcal{H}_{10}$
  (with $w_q = 0$ for $q > 2$)
Constraining the weights

- **Hard constraint:** $\mathcal{H}_2$ is constrained version of $\mathcal{H}_{10}$
  
  (with $w_q = 0$ for $q > 2$)

- **Soft-order constraint:** $\sum_{q=0}^{Q} w_q^2 \leq C$

The problem given soft-order constraint:

$$\text{Minimize } 1_N (Zw - y)^T (Zw - y) \text{ s.t. } w^T w \leq C$$
Constraining the weights

- Hard constraint: $\mathcal{H}_2$ is constrained version of $\mathcal{H}_{10}$
  (with $w_q = 0$ for $q > 2$)
- Soft-order constraint: $\sum_{q=0}^{Q} w_q^2 \leq C$
- The problem given soft-order constraint:

  $\min_{w} \frac{1}{N} (Zw - y)^T (Zw - y)$ s.t. $w^T w \leq C$

  smaller hypothesis space

- Solution $w_{\text{reg}}$ instead of $w_{\text{tr}}$
Equivalent to the unconstrained version

- Constrained version:

\[
\min_w E_{tr}(w) = \frac{1}{N}(Zw - y)^T(Zw - y) \quad \text{s.t.} \quad w^T w \leq C
\]

- Optimal when

\[
\nabla E_{tr}(w_{\text{reg}}) \propto -w_{\text{reg}}
\]

Why? If \(-\nabla E_{tr}(w)\) and \(w\) are not parallel, can decrease \(E_{tr}(w)\) without violating the constraint
Equivalent to the unconstrained version

Constrained version:

\[
\min_w E_{\text{tr}}(w) = \frac{1}{N} (Zw - y)^T(Zw - y) \text{ s.t. } w^T w \leq C
\]

Optimal when

\[
\nabla E_{\text{tr}}(w_{\text{reg}}) \propto -w_{\text{reg}}
\]

Assume \( \nabla E_{\text{tr}}(w_{\text{reg}}) = -2 \frac{1}{N} w_{\text{reg}} \)

\[ \Rightarrow \nabla E_{\text{tr}}(w_{\text{reg}}) + 2 \frac{1}{N} w_{\text{reg}} = 0 \]
Equivalent to the unconstrained version

- **Constrained version:**
  \[
  \min_w E_{\text{tr}}(w) = \frac{1}{N} (Zw - y)^T (Zw - y) \quad \text{s.t.} \quad w^T w \leq C
  \]

- **Optimal when**
  \[
  \nabla E_{\text{tr}}(w_{\text{reg}}) \propto -w_{\text{reg}}
  \]

- **Assume** \( \nabla E_{\text{tr}}(w_{\text{reg}}) = -2 \frac{1}{N} w_{\text{reg}} \)
  \[
  \Rightarrow \nabla E_{\text{tr}}(w_{\text{reg}}) + 2 \frac{1}{N} w_{\text{reg}} = 0
  \]

- **\( w_{\text{reg}} \) is also the solution of unconstrained problem**
  \[
  \min_w E_{\text{tr}}(w) + \frac{\lambda}{N} w^T w
  \]

(Ridge regression!)
Equivalent to the unconstrained version

- Constrained version:
  \[
  \min_w E_{\text{tr}}(w) = \frac{1}{N}(Zw - y)^T(Zw - y) \quad \text{s.t.} \quad w^T w \leq C
  \]

- Optimal when
  \[
  \nabla E_{\text{tr}}(w_{\text{reg}}) \propto -w_{\text{reg}}
  \]

- Assume \( \nabla E_{\text{tr}}(w_{\text{reg}}) = -2\frac{\lambda}{N}w_{\text{reg}} \)
  \[
  \Rightarrow \nabla E_{\text{tr}}(w_{\text{reg}}) + 2\frac{\lambda}{N}w_{\text{reg}} = 0
  \]

- \( w_{\text{reg}} \) is also the solution of unconstrained problem
  \[
  \min_w E_{\text{tr}}(w) + \frac{\lambda}{N}w^T w
  \]

(Ridge regression!)

\( C \uparrow \lambda \downarrow \)
Ridge regression solution

\[
\min_w E_{\text{reg}}(w) = \frac{1}{N} \left( (Zw - y)^T (Zw - y) + \lambda w^T w \right)
\]

\[\nabla E_{\text{reg}}(w) = 0 \implies Z^T Z(w - y) + \lambda w = 0
\]
Ridge regression solution

\[
\min_w E_{\text{reg}}(w) = \frac{1}{N} \left( (Zw - y)^T (Zw - y) + \lambda w^T w \right)
\]

\[
\nabla E_{\text{reg}}(w) = 0 \Rightarrow Z^T Z (w - y) + \lambda w = 0
\]

So, \( w_{\text{reg}} = (Z^T Z + \lambda I)^{-1} Z^T y \) (with regularization)
as opposed to \( w_{\text{tr}} = (Z^T Z)^{-1} Z^T y \) (without regularization)
The result

\[ \min_w E_{\text{tr}}(w) + \frac{\lambda}{N} w^T w \]

\( \lambda = 0 \)  \hspace{1cm} \( \lambda = 0.0001 \)  \hspace{1cm} \( \lambda = 0.01 \)  \hspace{1cm} \( \lambda = 1 \)

overfitting  \hspace{1cm} \rightarrow  \hspace{1cm} \rightarrow  \hspace{1cm} \rightarrow  \hspace{1cm} \rightarrow  \hspace{1cm} \text{underfitting}
Equivalent to “weight decay”

Consider the general case

\[
\min_w E_{tr}(w) + \frac{\lambda}{N} w^T w
\]
Equivalent to “weight decay”

- Consider the general case

\[
\min_w E_{tr}(w) + \frac{\lambda}{N} w^T w
\]

- Gradient descent:

\[
w_{t+1} = w_t - \eta \left( \nabla E_{tr}(w_t) + 2\frac{\lambda}{N} w_t \right)
\]

\[
= w_t \left( 1 - 2\eta \frac{\lambda}{N} \right) - \eta \nabla E_{tr}(w_t)
\]

\[\text{weight decay}\]
Variations of weight decay

- Emphasis of certain weights:

\[
\sum_{q=0}^{Q} \gamma_q w_q^2
\]

- Example 1: \( \gamma_q = 2^q \Rightarrow \text{low-order fit} \)
- Example 2: \( \gamma_q = 2^{-q} \Rightarrow \text{high-order fit} \)
Variations of weight decay

- Emphasis of certain weights:

\[
\sum_{q=0}^{Q} \gamma_q w_q^2
\]

- Example 1: \(\gamma_q = 2^q\) \(\Rightarrow\) low-order fit
- Example 2: \(\gamma_q = 2^{-q}\) \(\Rightarrow\) high-order fit

- General Tikhonov regularizer:

\[
w^T H w
\]

with a positive semi-definite \(H\)
General form of regularizer

- Calling the regularizer $\Omega = \Omega(h)$, we minimize

$$E_{\text{reg}}(h) = E_{\text{tr}}(h) + \frac{\lambda}{N} \Omega(h)$$

- In general, $\Omega(h)$ can be any measurement for the “size” of $h$
L1-regularizer: $\Omega(w) = \|w\|_1 = \sum_q |w_q|$

- Usually leads to a **sparse solution**
  - (only few $w_q$ will be nonzero)
Conclusions

- VC dimension
- Regularization

Questions?