

# CSE 417T: Introduction to Machine Learning

## Lecture 4: Infinite Hypothesis Sets

Henry Chai

09/06/18

# Recall

- Suppose  $\mathcal{H}$  is finite i.e.  $\mathcal{H} = \{h_1, \dots, h_m\}$
- $E_{in}(g)$  = in-sample error of best hypothesis in  $\mathcal{H}$
- $E_{out}(g)$  = out-of-sample error of best hypothesis in  $\mathcal{H}$
- Generalization error of  $g = |E_{in}(g) - E_{out}(g)|$

$$\bullet P\{|E_{in}(g) - E_{out}(g)| > \epsilon\} \leq 2(m)e^{-2\epsilon^2 n}$$

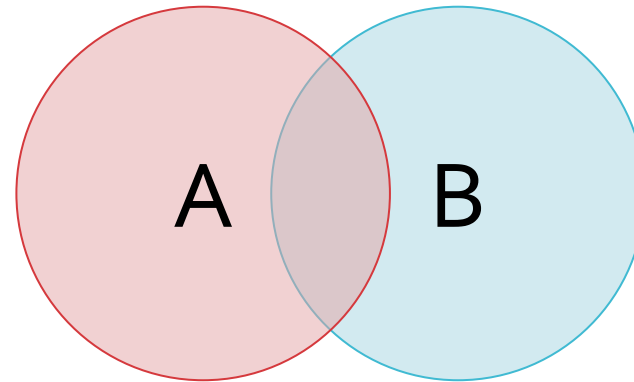


$$\bullet E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2n} \log\left(\frac{2m}{\delta}\right)}$$

with probability at least  $1 - \delta$

The Union  
Bound is bad

$$P\{A \cup B\} = P\{A\} + P\{B\} - P\{A \cap B\}$$



## Recall

- Suppose  $\mathcal{H}$  is finite i.e.  $\mathcal{H} = \{h_1, \dots, h_m\}$

$$P\{|E_{in}(g) - E_{out}(g)| > \epsilon\}$$

$$\leq P\left\{\bigcup_{j=1}^m |E_{in}(h_j) - E_{out}(h_j)| > \epsilon\right\}$$

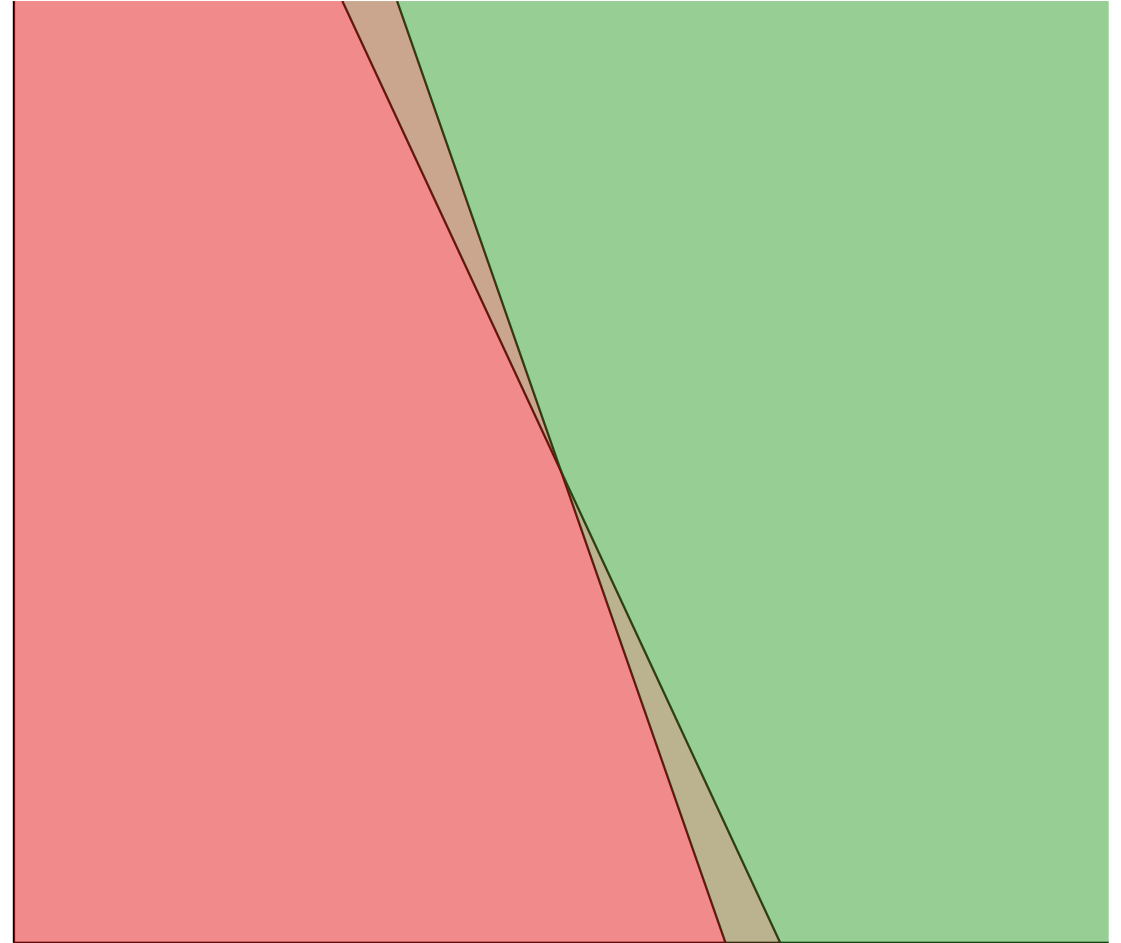
$$\leq \sum_{j=1}^m P\{|E_{in}(h_j) - E_{out}(h_j)| > \epsilon\}$$

$$\leq \sum_{j=1}^m 2e^{-2\epsilon^2 n} = 2(m)e^{-2\epsilon^2 n}$$

# Good News

If two hypotheses  $h_1, h_2 \in \mathcal{H}$  are very similar, then the events  $|E_{in}(h_1) - E_{out}(h_1)| > \epsilon$  and  $|E_{in}(h_2) - E_{out}(h_2)| > \epsilon$  are very likely to overlap

$P\{|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \cap |E_{in}(h_2) - E_{out}(h_2)| > \epsilon\}$  is big

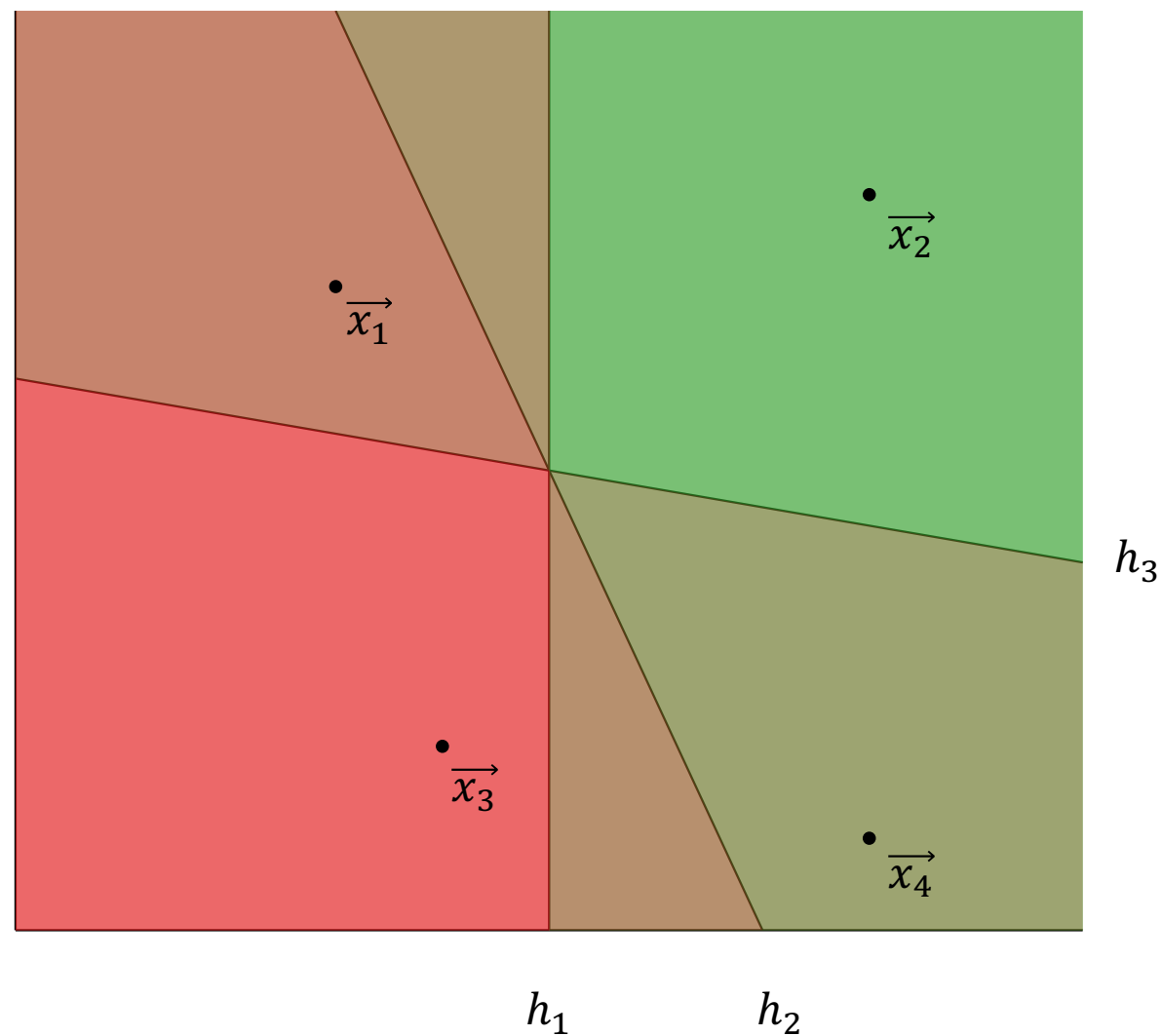


# Dichotomy

- Given some finite sample of points  $(\vec{x}_1, \dots, \vec{x}_n)$  from the input space and single hypothesis  $h \in \mathcal{H}$ , applying  $h$  to each point in  $(\vec{x}_1, \dots, \vec{x}_n)$  results in a **dichotomy**
  - $(h(\vec{x}_1), \dots, h(\vec{x}_n))$  is a vector of  $n + 1$ 's and  $-1$ 's
- Given  $(\vec{x}_1, \dots, \vec{x}_n)$ , each hypothesis in  $\mathcal{H}$  generates a dichotomy but not necessarily a unique dichotomy!
  - The set of dichotomies induced by  $\mathcal{H}$  on  $(\vec{x}_1, \dots, \vec{x}_n)$  is  $\mathcal{H}(\vec{x}_1, \dots, \vec{x}_n) = \{ (h(\vec{x}_1), \dots, h(\vec{x}_n)) \mid h \in \mathcal{H} \}$

# Dichotomy Example

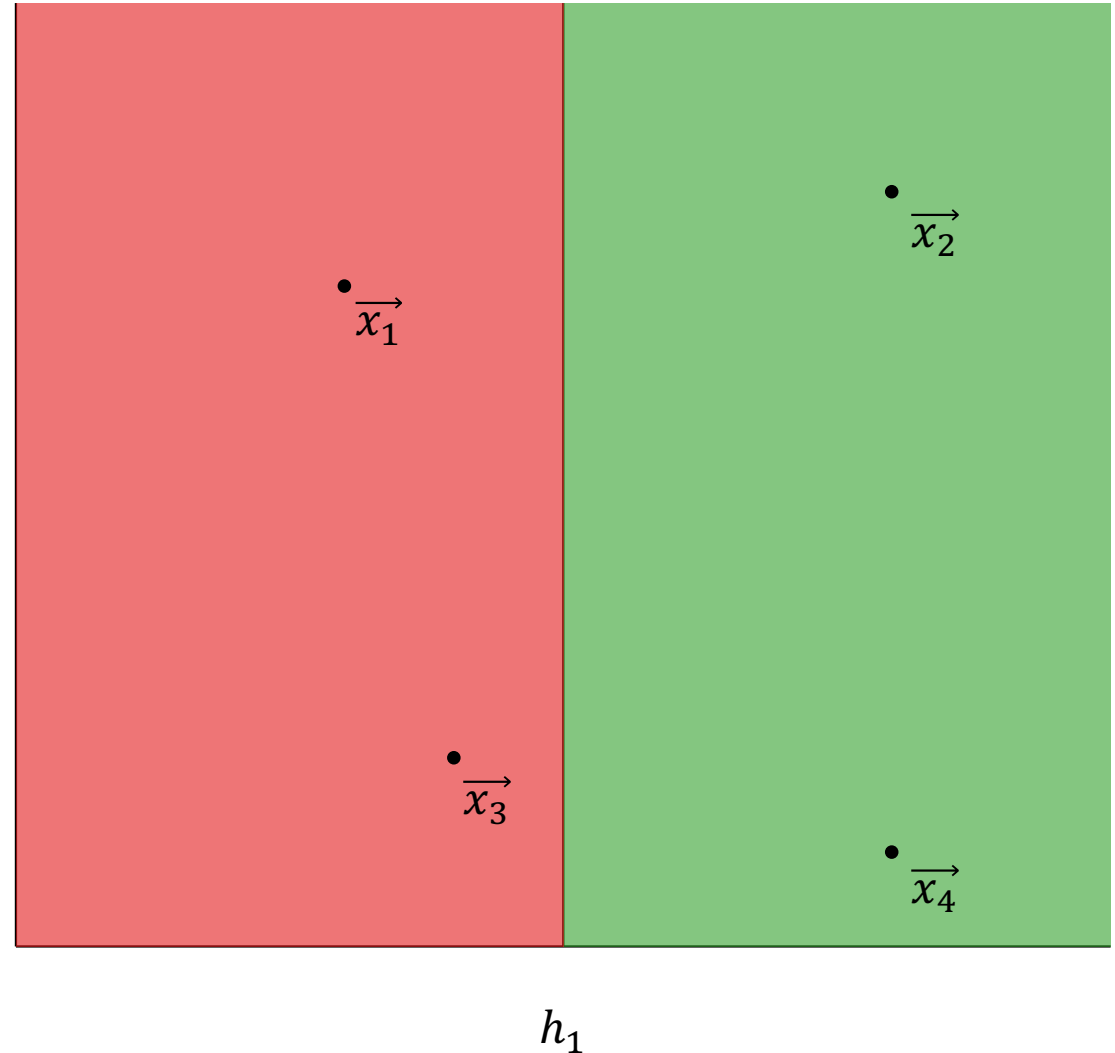
$$\mathcal{H} = \{h_1, h_2, h_3\}$$



# Dichotomy Example

$$\mathcal{H} = \{h_1, h_2, h_3\}$$

$$(h_1(\vec{x}_1), h_1(\vec{x}_2), h_1(\vec{x}_3), h_1(\vec{x}_4)) \\ = (-1, +1, -1, +1)$$

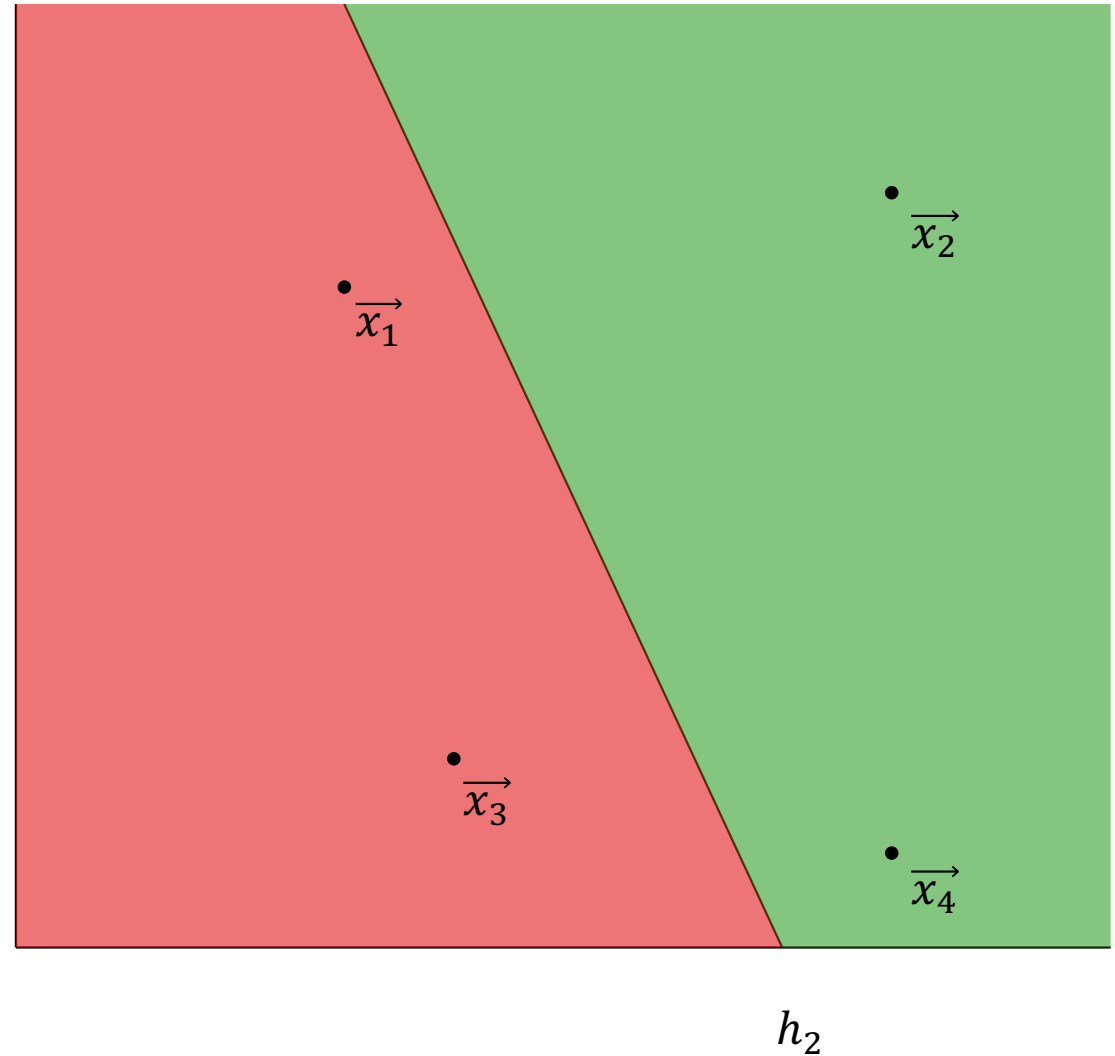




# Dichotomy Example

$$\mathcal{H} = \{h_1, h_2, h_3\}$$

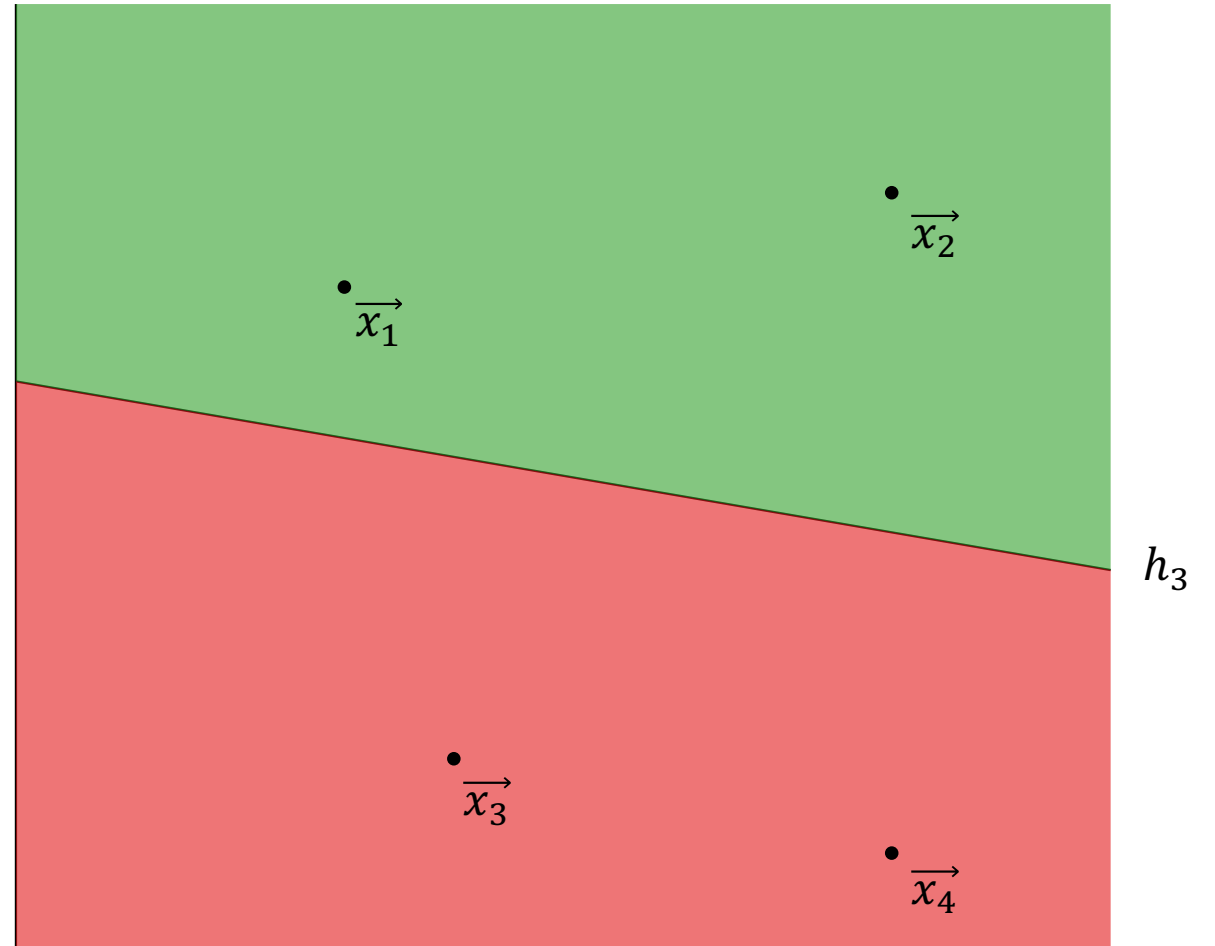
$$(h_2(\vec{x}_1), h_2(\vec{x}_2), h_2(\vec{x}_3), h_2(\vec{x}_4)) \\ = (-1, +1, -1, +1)$$



# Dichotomy Example

$$\mathcal{H} = \{h_1, h_2, h_3\}$$

$$(h_3(\vec{x}_1), h_3(\vec{x}_2), h_3(\vec{x}_3), h_3(\vec{x}_4)) \\ = (+1, +1, -1, -1)$$

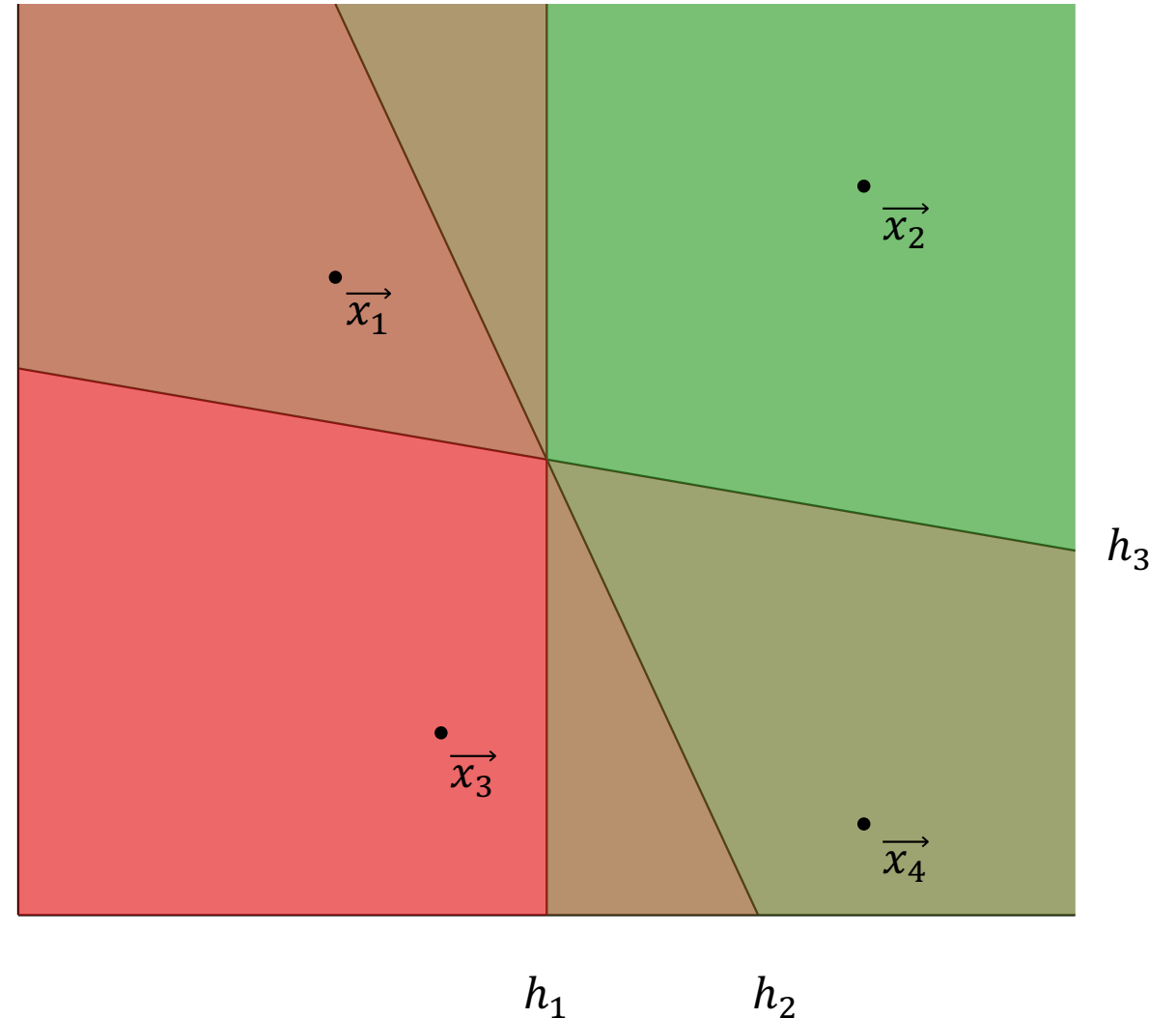


# Dichotomy Example

$$\mathcal{H} = \{h_1, h_2, h_3\}$$

$$\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4) = \{(-1, +1, -1, +1), (+1, +1, -1, -1)\}$$

$$|\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4)| = 2$$

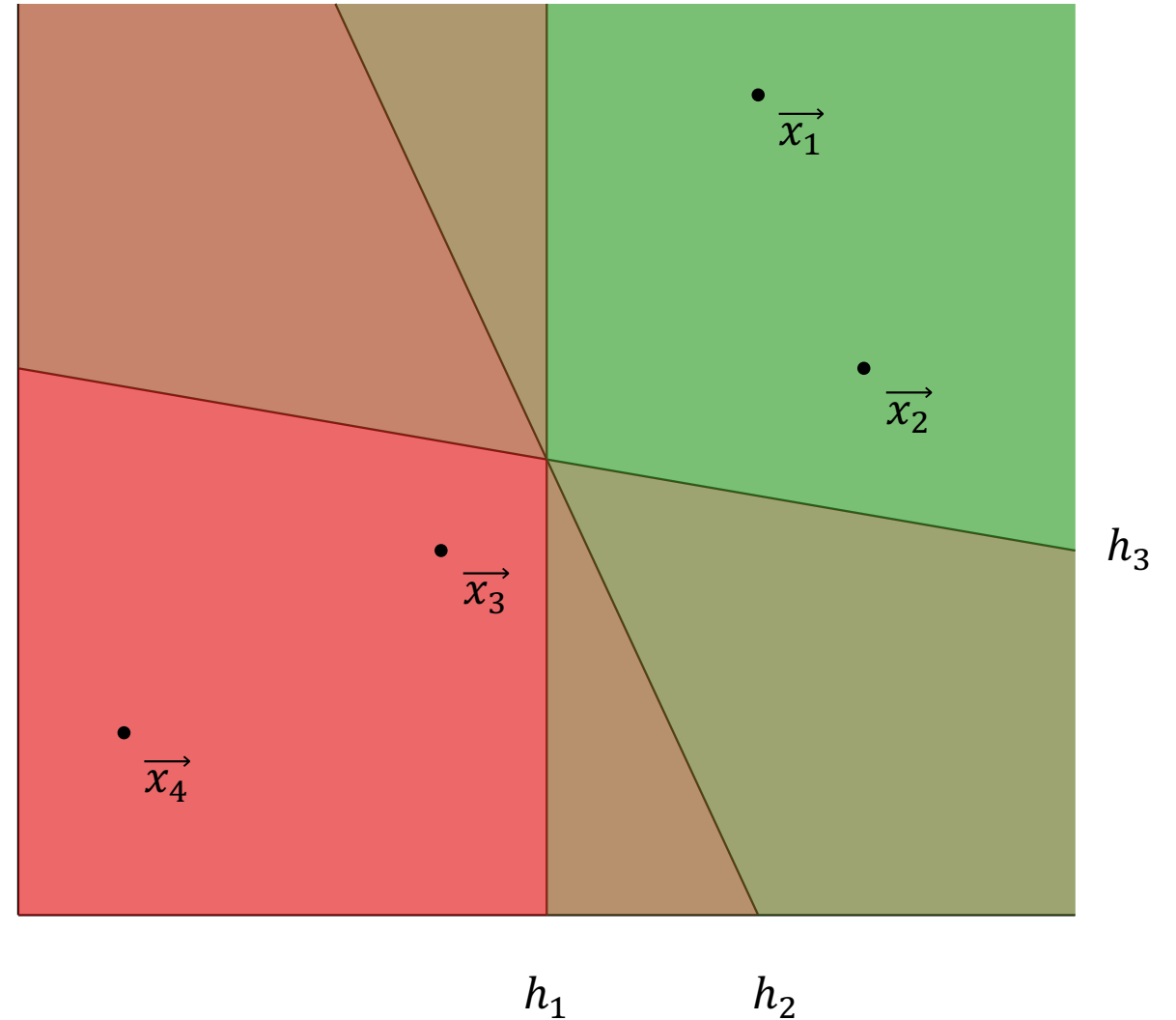


# Dichotomy Example

$$\mathcal{H} = \{h_1, h_2, h_3\}$$

$$\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4) = \{(+1, +1, -1, -1)\}$$

$$|\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4)| = 1$$



# Growth Function

- The growth function of  $\mathcal{H}$  is the largest number of dichotomies  $\mathcal{H}$  can induce across all data sets of size  $n$

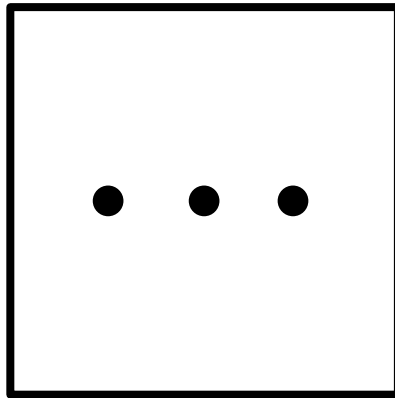
$$m_{\mathcal{H}}(n) = \max_{(\vec{x}_1, \dots, \vec{x}_n) \in \mathcal{X}} |\mathcal{H}(\vec{x}_1, \dots, \vec{x}_n)|$$

# Growth Function (Shattering)

- Observe that  $m_{\mathcal{H}}(n) \leq 2^n \forall \mathcal{H}$  and  $n$
- Given  $\mathcal{H}$ , if  $\exists (\vec{x}_1, \dots, \vec{x}_n) \in \mathcal{X}$  s.t.  $|\mathcal{H}(\vec{x}_1, \dots, \vec{x}_n)| = 2^n$ , then  $\mathcal{H}$  shatters  $(\vec{x}_1, \dots, \vec{x}_n)$
- If  $\exists (\vec{x}_1, \dots, \vec{x}_n) \in \mathcal{X}$  that is shattered by  $\mathcal{H}$ , then  $m_{\mathcal{H}}(n) = 2^n$

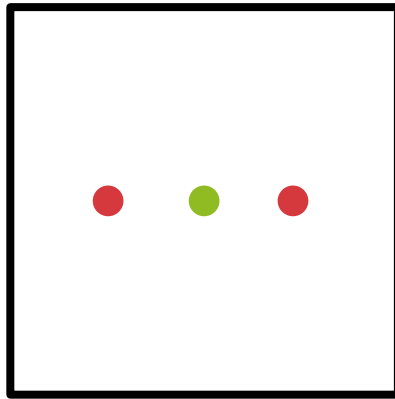
# Growth Function: Example

- $\mathcal{X} = \mathbb{R}^2$  and  $\mathcal{H} =$  Linear classifiers (perceptrons)
- What is  $m_{\mathcal{H}}(3)$ ?



# Growth Function: Example

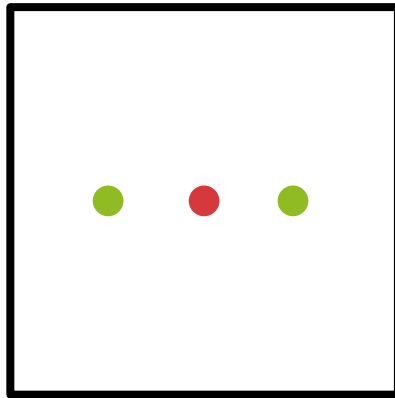
- $\mathcal{X} = \mathbb{R}^2$  and  $\mathcal{H} =$  Linear classifiers (perceptrons)
- What is  $m_{\mathcal{H}}(3)$ ?





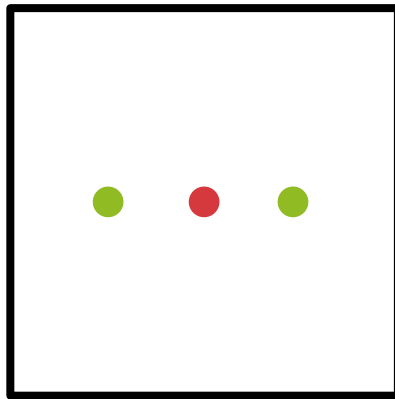
# Growth Function: Example

- $\mathcal{X} = \mathbb{R}^2$  and  $\mathcal{H} =$  Linear classifiers (perceptrons)
- What is  $m_{\mathcal{H}}(3)$ ?

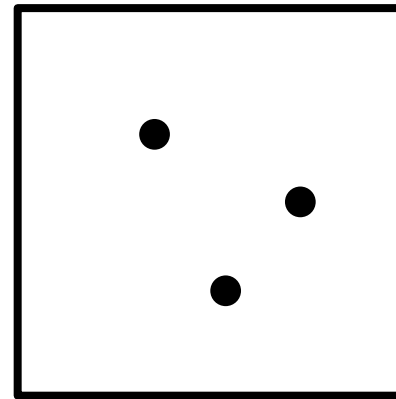


# Growth Function: Example

- $\mathcal{X} = \mathbb{R}^2$  and  $\mathcal{H} =$  Linear classifiers (perceptrons)
- $m_{\mathcal{H}}(3) = 8 = 2^3$



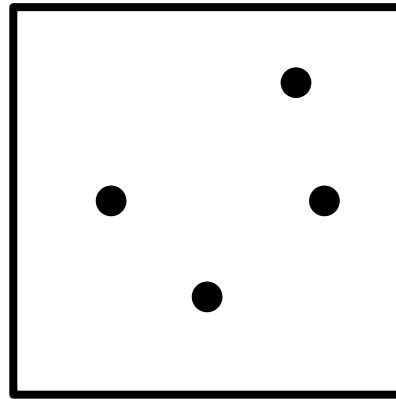
$$|\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3)| = 6$$



$$|\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3)| = 8$$

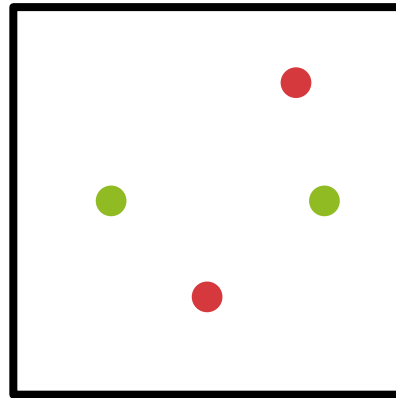
# Growth Function: Example

- $\mathcal{X} = \mathbb{R}^2$  and  $\mathcal{H} =$  Linear classifiers (perceptrons)
- What is  $m_{\mathcal{H}}(4)$ ?



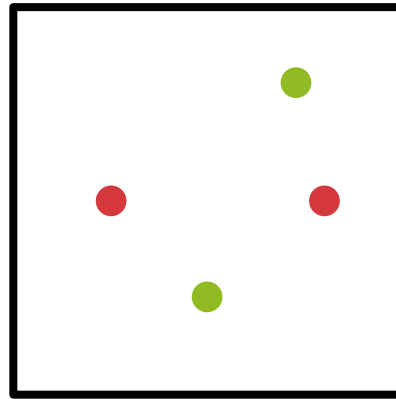
# Growth Function: Example

- $\mathcal{X} = \mathbb{R}^2$  and  $\mathcal{H} =$  Linear classifiers (perceptrons)
- What is  $m_{\mathcal{H}}(4)$ ?



# Growth Function: Example

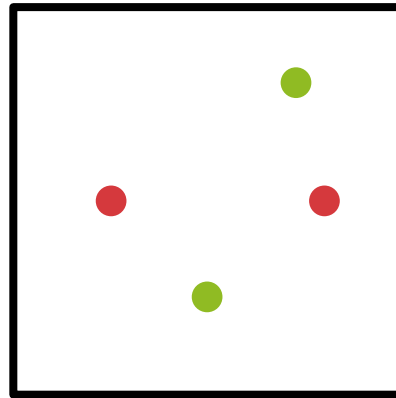
- $\mathcal{X} = \mathbb{R}^2$  and  $\mathcal{H} =$  Linear classifiers (perceptrons)
- What is  $m_{\mathcal{H}}(4)$ ?



$$|\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3)| = 14$$

# Growth Function: Example

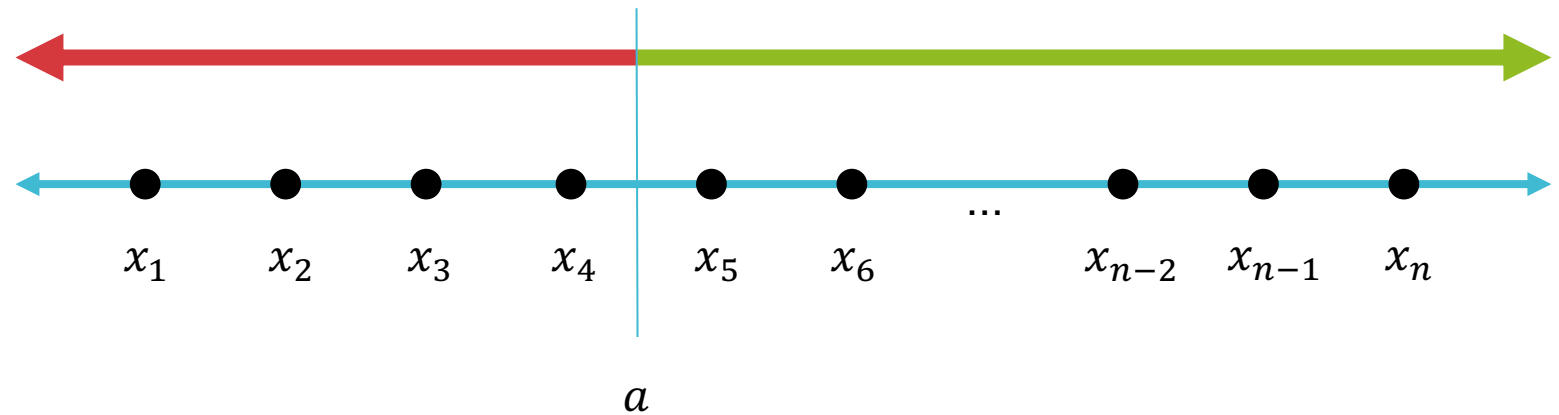
- $\mathcal{X} = \mathbb{R}^2$  and  $\mathcal{H} =$  Linear classifiers (perceptrons)
- $m_{\mathcal{H}}(4) = 14 < 2^4$



$$|\mathcal{H}(\vec{x}_1, \vec{x}_2, \vec{x}_3)| = 14$$

# Growth Function: Example

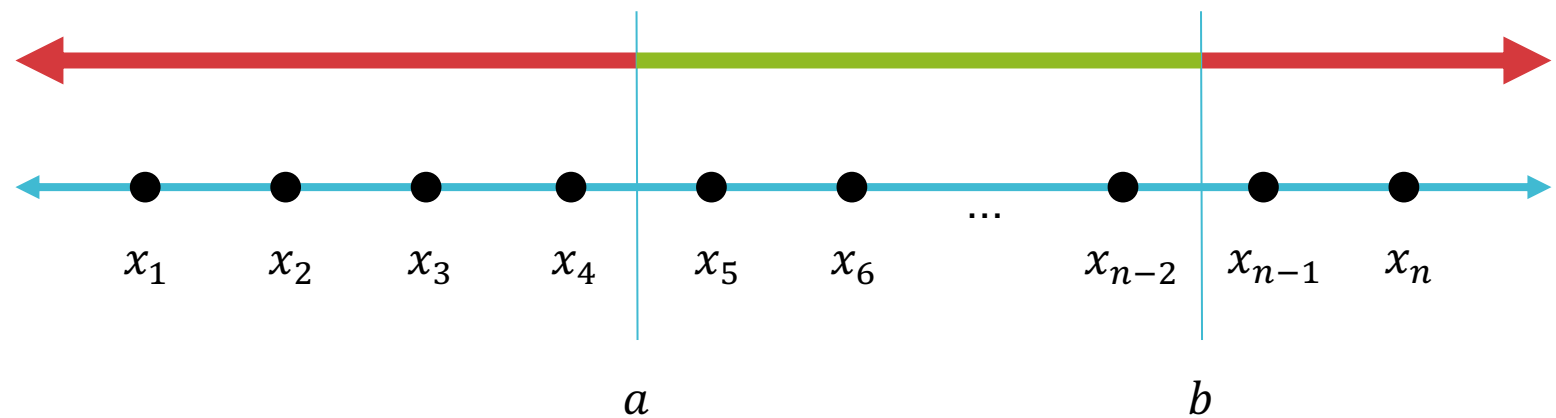
- $\mathcal{X} = \mathbb{R}$  and  $\mathcal{H} =$  Positive rays:  $h(x) = \text{sign}(x - a)$
- $m_{\mathcal{H}}(n) = n + 1$



# Growth Function: Example

- $\mathcal{X} = \mathbb{R}$  and  $\mathcal{H} =$  Positive intervals

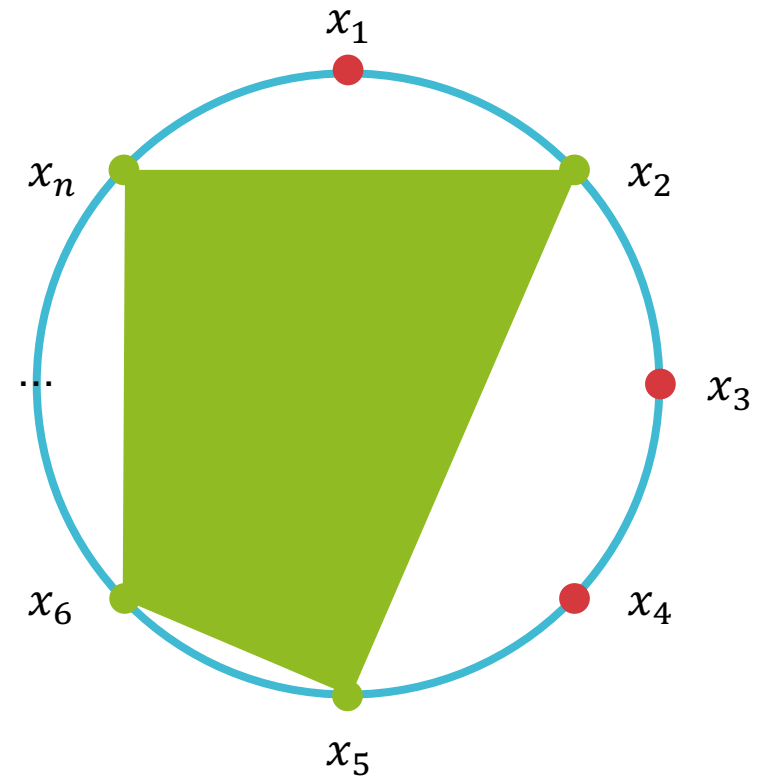
- $m_{\mathcal{H}}(n) = \binom{n+1}{2} + 1 = \frac{n^2}{2} + \frac{n}{2} + 1$





# Growth Function: Example

- $\mathcal{X} = \mathbb{R}^2$  and  $\mathcal{H} = \text{Convex sets}$
- $m_{\mathcal{H}}(n) = 2^n$



# Growth Function (Break Points)

- If  $m_{\mathcal{H}}(k) < 2^k$ , then  $k$  is a break point for  $\mathcal{H}$ 
  - For 2D linear separators,  $k = 4$  is a break point
  - For 1D positive rays,  $k = 2$  is a break point
  - For 1D positive intervals,  $k = 3$  is a break point
  - For 2D convex sets, there are no break points
- Theorem: If there is at least one break point for  $\mathcal{H}$ , then  $m_{\mathcal{H}}(n)$  is polynomial in  $n$

# Recall

- $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2n} \log \left( \frac{2m_{\mathcal{H}}(n)}{\delta} \right)}$

with probability at least  $1 - \delta$

- If  $m_{\mathcal{H}}(n)$  is polynomial in  $n$ , then  $\sqrt{\frac{1}{2n} \log \left( \frac{2m_{\mathcal{H}}(n)}{\delta} \right)} \rightarrow 0$

in the limit as  $n \rightarrow \infty$