SHIRAZ UNIVERSITY
Computer Science and Engineering Department
Machine Learning Lab

Learning Theory

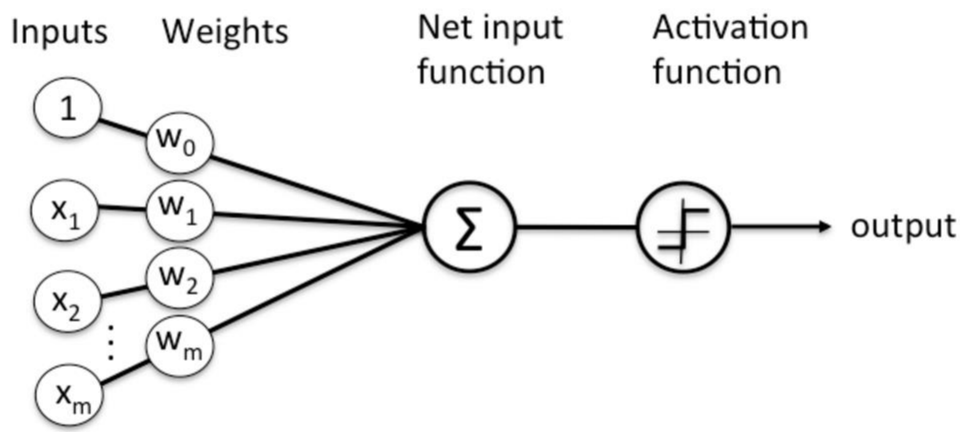
Sattar Hashemi

Based on:

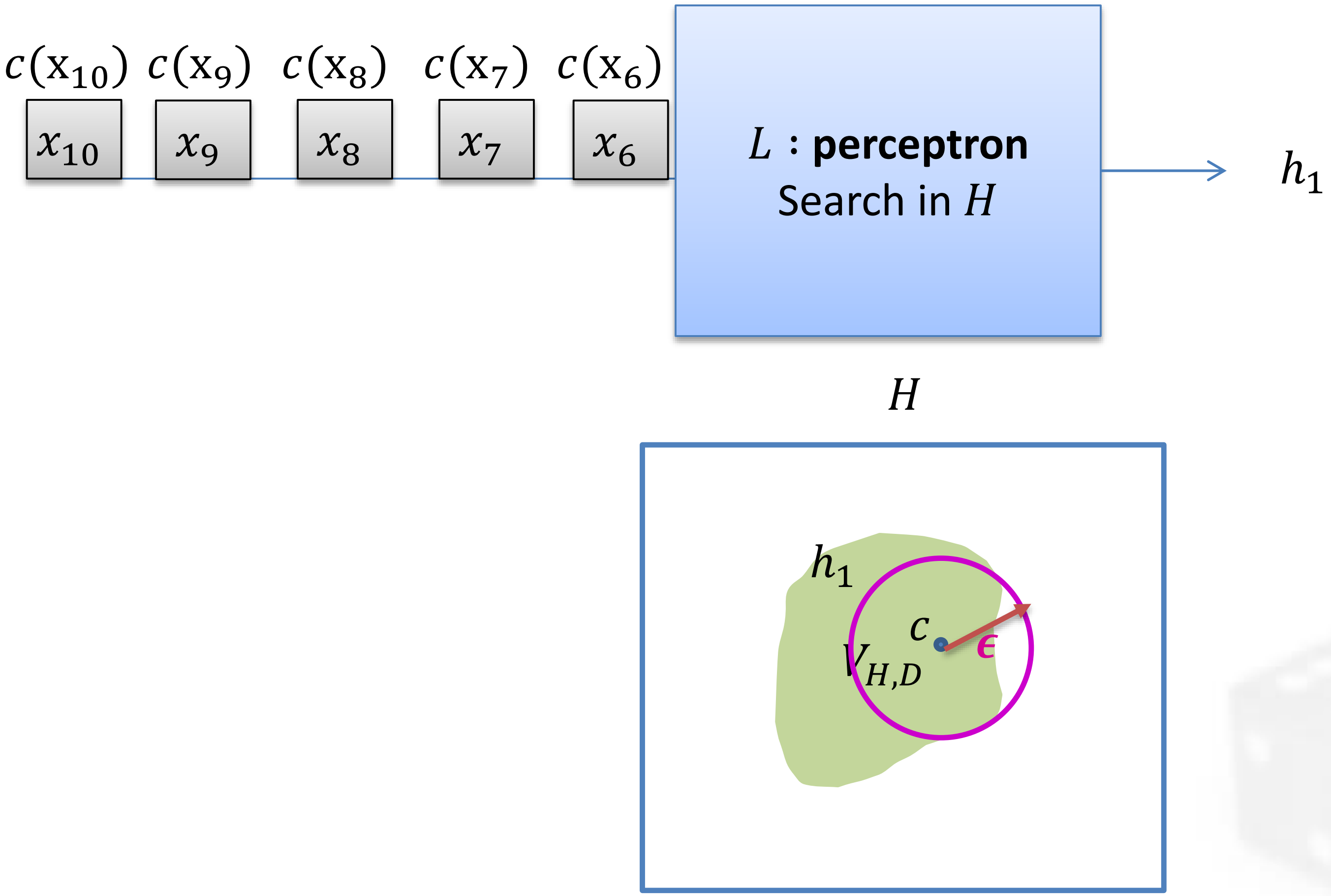
Mitchell, Tom M. "Machine learning, Chapter 7
NG, Andrew. Machine learning Lecture Notes, Learning Theory
Rudin, Walter. "Principles of Mathematical Analysis." (1976).
Vapnik, Vladimir. "Statistical learning theory" (1998).

Example: c as a limit point

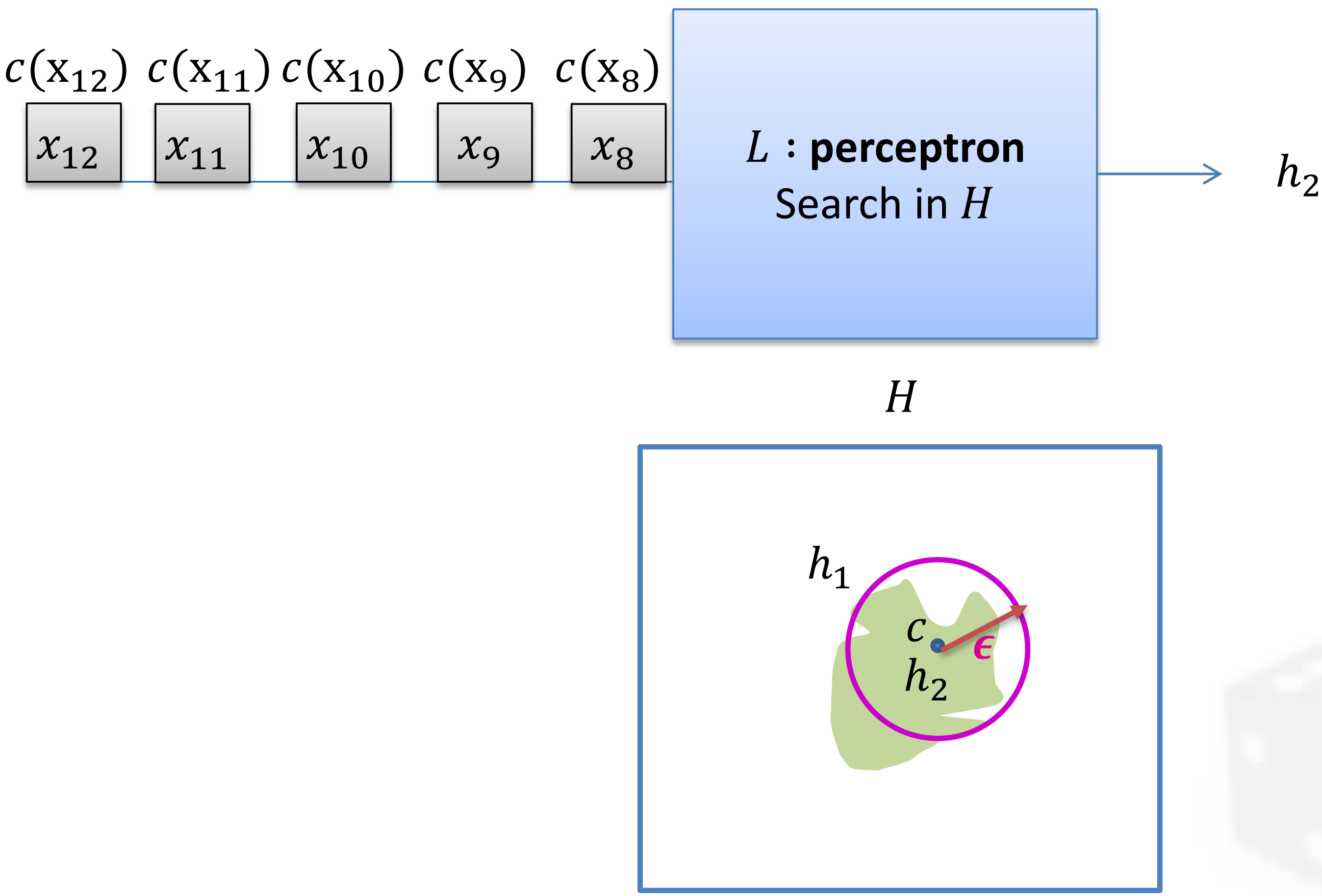
- Assume **Linear separable discrete data**
- Hypothesis representor:
  - **Rosenblatt perceptron**
- So our $|H| \in R$ and $c \in H$
- What happened in Space H ?



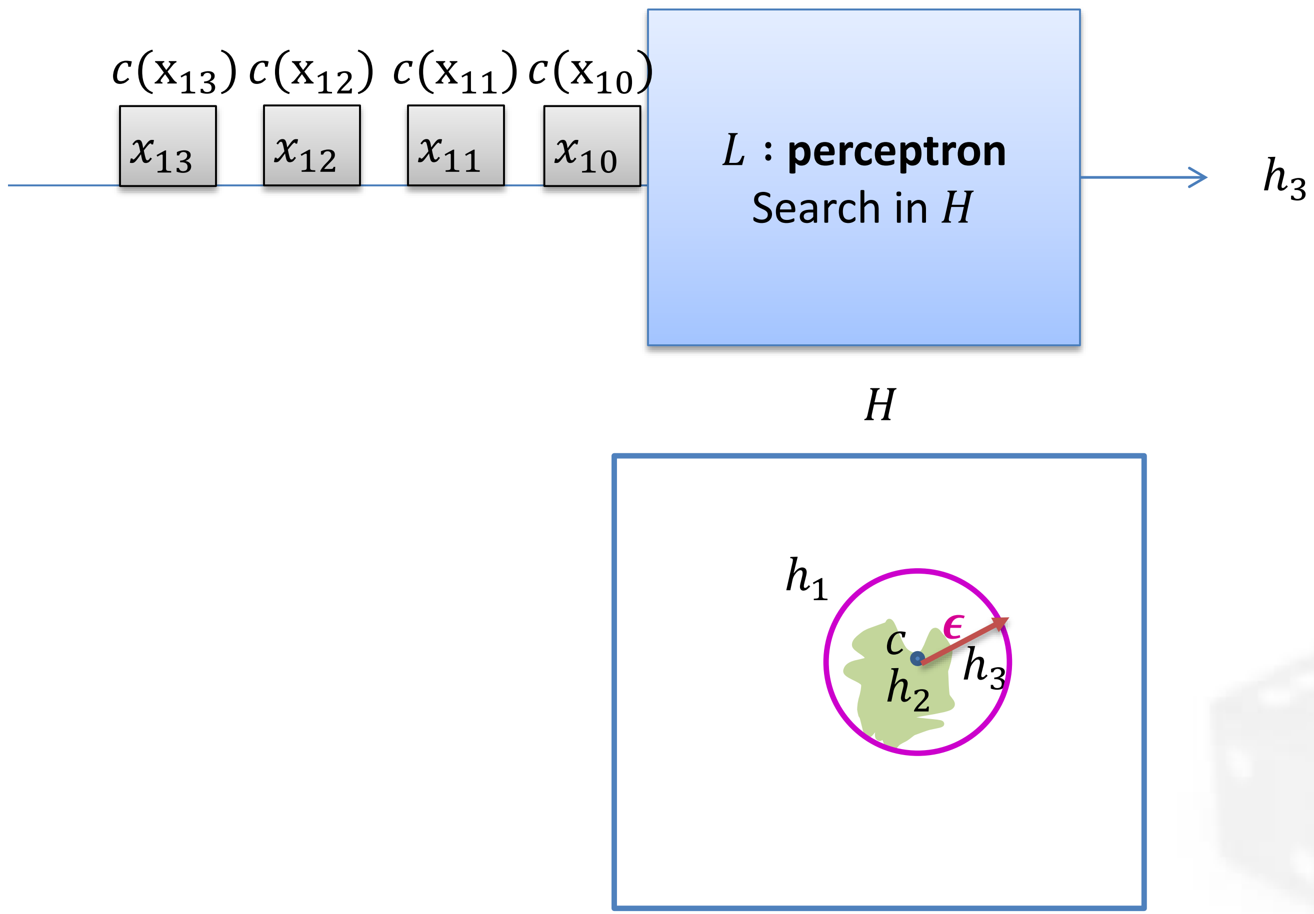
Example: c as a limit point



Example: c as a limit point



Example: c as a limit point



Uniform convergence

Definition* we say that a sequence of function $\{f_n\}, n = 1,2,3, \ldots$ converges uniformly on E to a function of f if for every $\epsilon > 0$ there is an integer N such that $n \geq N$ implies

$$\forall x \in E \quad |f_n(x) - f(x)| \leq \epsilon$$

- In our case :
  - $E = X$
  - $f(x) = c$

*Rudin, Walter. "Principles of Mathematical Analysis"

Loss and Risk function

True error: $error_{\mathcal{D}}(h) = P_{\mathcal{D}}\big(\boldsymbol{I}\big(h(x), c(x)\big) = 1\big)$
Train Error: $P_{D}\big(\boldsymbol{I}\big(h(x), c(x)\big) = 1\big)$
In classification (that whole lecture is about it), very simple **loss function** is

0-1 loss function: $\boldsymbol{I}\big(h(x), c(x)\big)$

Risk function is average of loss function over true distribution

$$E_{\sim P_{\mathcal{D}}}\Big(\boldsymbol{I}\big(h(x), c(x)\big)\Big) = P_{\mathcal{D}}\big(\boldsymbol{I}\big(h(x), c(x)\big) = 1\big)$$

What is the statistics of **Risk**? **Empirical Risk**

Empirical Risk function is average of loss function over training example

$$E_{\sim P_{D}}\Big(\boldsymbol{I}\big(h(x), c(x)\big)\Big) = P_{D}\big(\boldsymbol{I}\big(h(x), c(x)\big) = 1\big)$$

Sample Complexity for Finite H

Empirical Risk Minimization

$R(h) = E_{\sim P_{\mathcal{D}}}\left(\boldsymbol{I}\big(h(x), c(x)\big)\right)$ R is the risk function

$\hat{R}(h) = E_{\sim P_D}\left(\boldsymbol{I}\big(h(x), c(x)\big)\right)$ $\hat{R}$ is the empirical risk function

Main objective of ERM (Empirical Risk Minimization)

$$\boldsymbol{\hat{h} = \mathbf{argmin}_{h \in H}\, \hat{R}(h)}$$

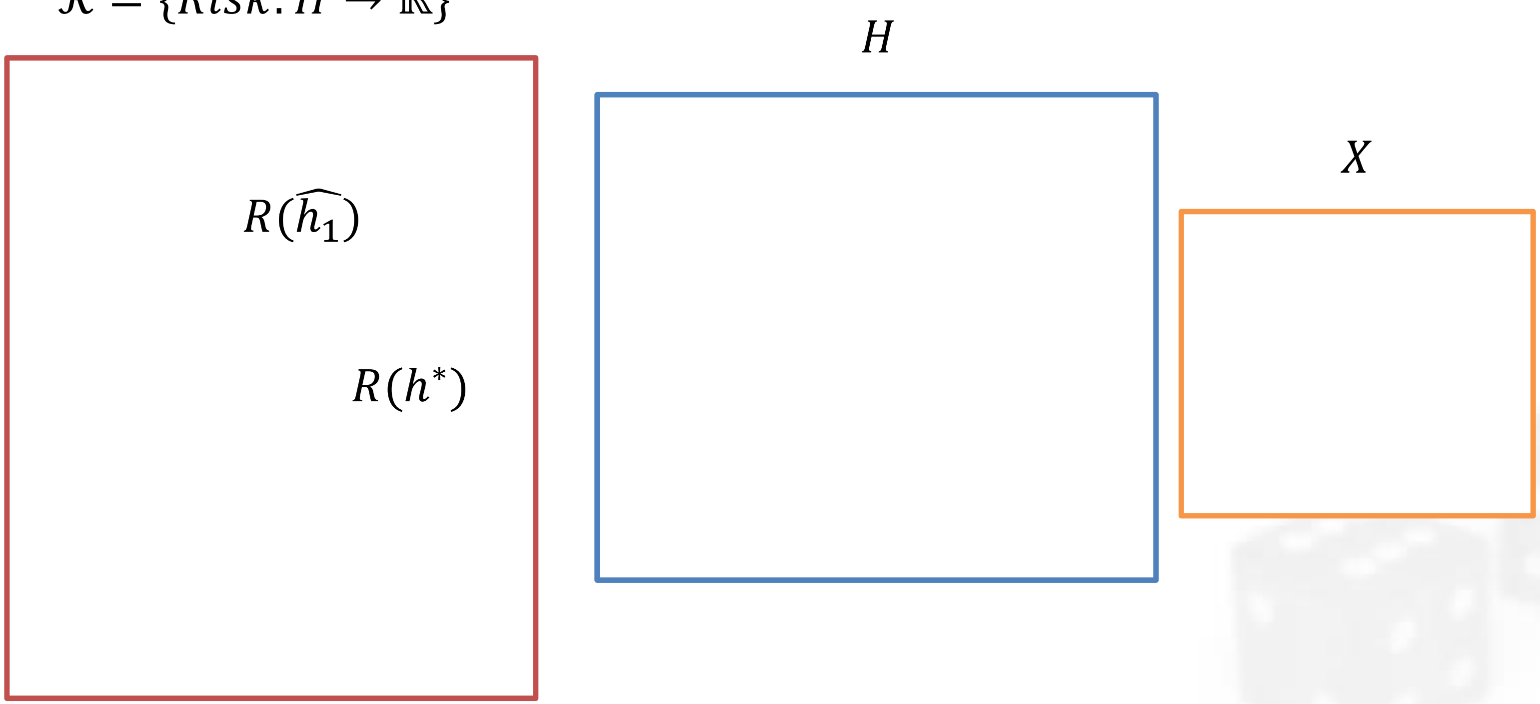
Best hypothesis in H:

$$\boldsymbol{h^* = \mathbf{argmin}_{h \in H}\, R(h)}$$

Goal of learning: find $\hat{h}$ whose $R(\hat{h})$ will be close to $\boldsymbol{R(h^*)}$.

Space of Risk Function*

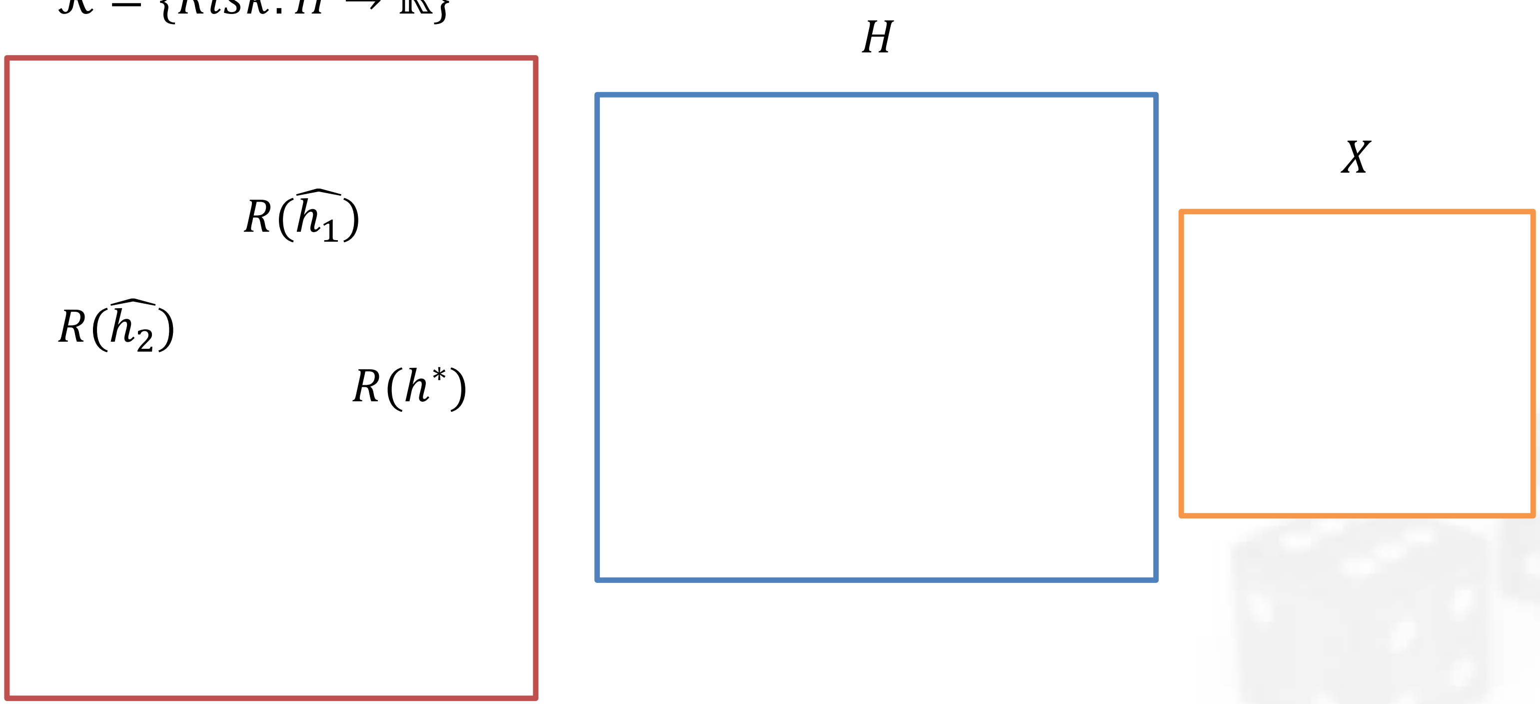
$$\mathcal{R} = \{Risk: H \to \mathbb{R}\}$$



*Vapnik, Vladimir. "Statistical learning theory.

Space of Risk Function*

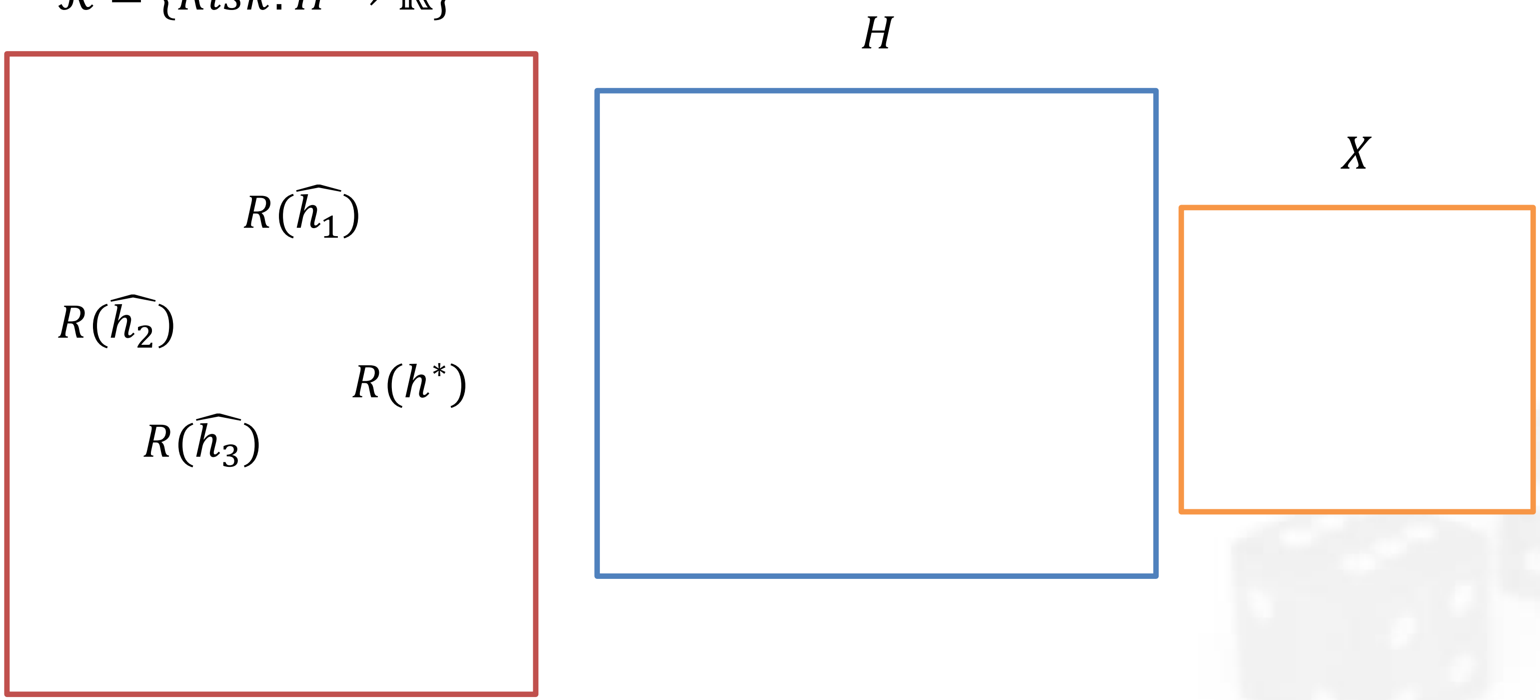
$\mathcal{R} = \{Risk: H \to \mathbb{R}\}$



*Vapnik, Vladimir. "Statistical learning theory.

Space of Risk Function*

$\mathcal{R} = \{Risk: H \rightarrow \mathbb{R}\}$



*Vapnik, Vladimir. "Statistical learning theory.

Hoeffding Bound

- Hoefding Inequality: Let $Z_1, Z_2, \ldots Z_m$ be m independent and identically distributed random variables drawn from the $Bernoulli(\phi)$ distribution. Let $\hat{\phi} = \frac{1}{m}\sum_{i=1}^{m} Z_i$ and $\epsilon > 0$ then

$$P\left(\left|\phi - \hat{\phi}\right| > \epsilon\right) \leq 2e^{-2\epsilon^2 m}$$

Uniform Convergence result

- $\hat{R}(h) \sim Bernolli$
- So based on Hoeffding bound

$$P\left(\left|R(h_i) - \hat{R}(h_i)\right| > \epsilon\right) \leq 2e^{-2\epsilon^2 m}$$

- It is true only for one h not all $h \in H$. We want to talk about convergence in space of Risk function so based on **union bound lemma**

$$P\left(\exists h \in H, \left(\left|R(h_i) - \hat{R}(h_i)\right| > \epsilon\right)\right) \leq 2|H|e^{-2\epsilon^2 m}$$

so

$$P\left(\forall h \in H, \left(\left|R(h_i) - \hat{R}(h_i)\right| \leq \epsilon\right)\right) \geq 1 - 2|H|e^{-2\epsilon^2 m}$$

- This is the **Uniform Convergence** result

Sample Complexity and PAC Learnability

Sample Complexity for Finite H

$$P\left(\forall h \in H, \left(\left|R(h_i) - \hat{R}(h_i)\right| \leq \epsilon\right)\right) \geq 1 - 2|H|e^{-2\epsilon^2 m}$$

$$P\left(\forall h \in H, \left(\left|R(h_i) - \hat{R}(h_i)\right| \leq \epsilon\right)\right) \geq 1 - \delta$$

$$\boldsymbol{m \geq \frac{1}{2\epsilon^2} \ln \frac{2|H|}{\delta}}$$

If $c \notin H \rightarrow m \in O\left(P\left(\frac{1}{\epsilon}\right)^2, n, \log\left(\frac{1}{\delta}\right)\right)$

So **if $\boldsymbol{m \in O(P(n))}$ then $\boldsymbol{c}$ is PAC learnable**.

Infinite H

Are there any measure of complexity that we can use instead of |H|?

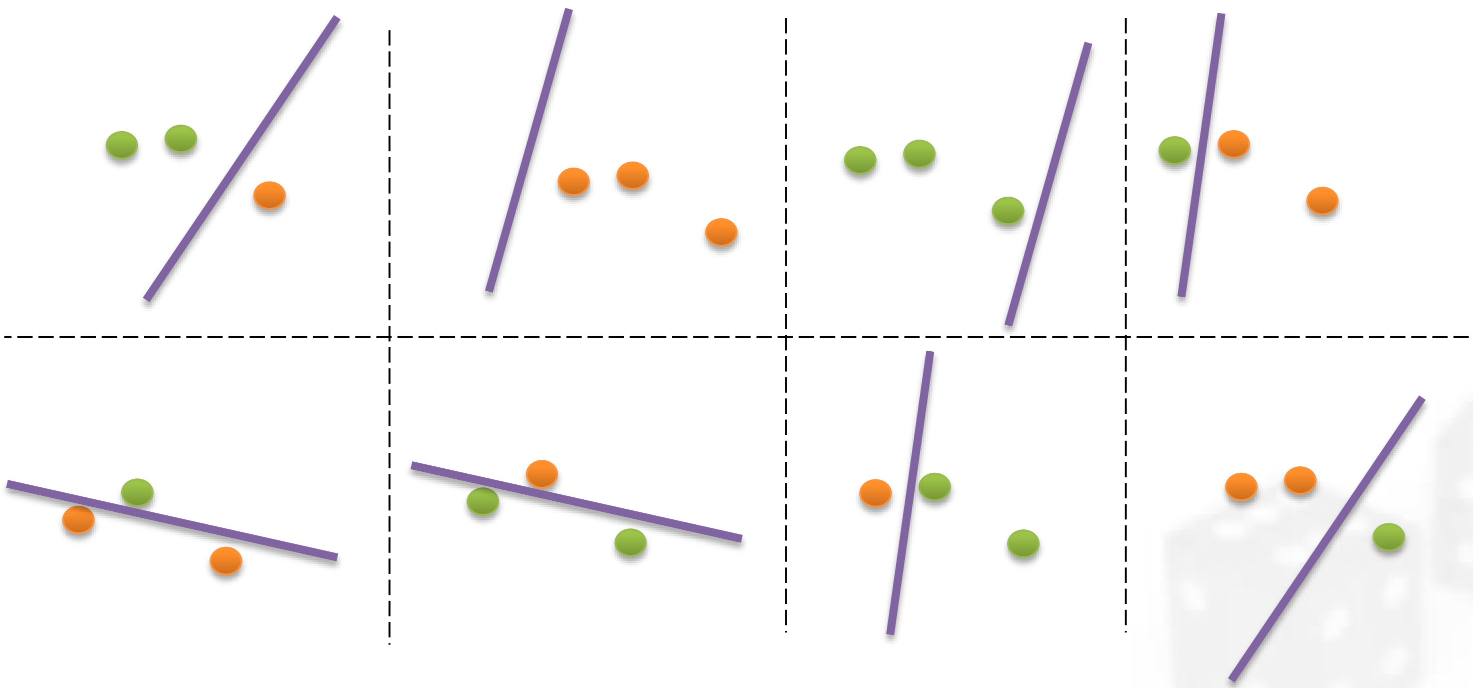
Infinite H

Are there any measure of complexity that we can use instead of |H|?

Answer: The largest subset of X for which H can guarantee zero training error (regardless of the target function c)

Shattering a Set of Instances

- Definition: a **dichotomy** of a set S is a partition of S into two disjoint subset
- Definition: a set of instances S is **shattered** by hypothesis space $\boldsymbol{H}$ if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy



The Vapnik-Chervonenkis Dimension

- **Definition**: the **Vapnik-Chervonenkis Dimension $\boldsymbol{VC(H)}$**, of hypothesis space H define over instance space $X$is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of Xcan be shattered by, then $VC(H) = \infty$

VC dimension: example

- Consider $X = R$ want to learn $c: X \rightarrow \{+, -\}$
- What is VC dimension of
- Open intervals:
- H1: if $x > a$ then $y = +$ else $y = -$
- H2: if $a < x < b$ then $y = +$ else $y = -$

VC dimension: example

- Consider $X = R$ want to learn $c: X \to \{+, -\}$
- What is VC dimension of
- Open intervals:
- H1: if $x > a$ then $y = +$ else $y = -$
  - Answer: VC=1
- H2: if $a < x < b$ then $y = +$ else $y = -$
  - Answer: VC=2

VC dimension: example

- What is VC dimension of lines in a plane?
  H3: $\{((w_0 + w_1x_1 + w_2x_2) > 0 \ \rightarrow y = +\}$
$$VC(H_3) = 3$$
- What about hyper plane in n dimension input space?
$$VC(H_4) = n + 1$$