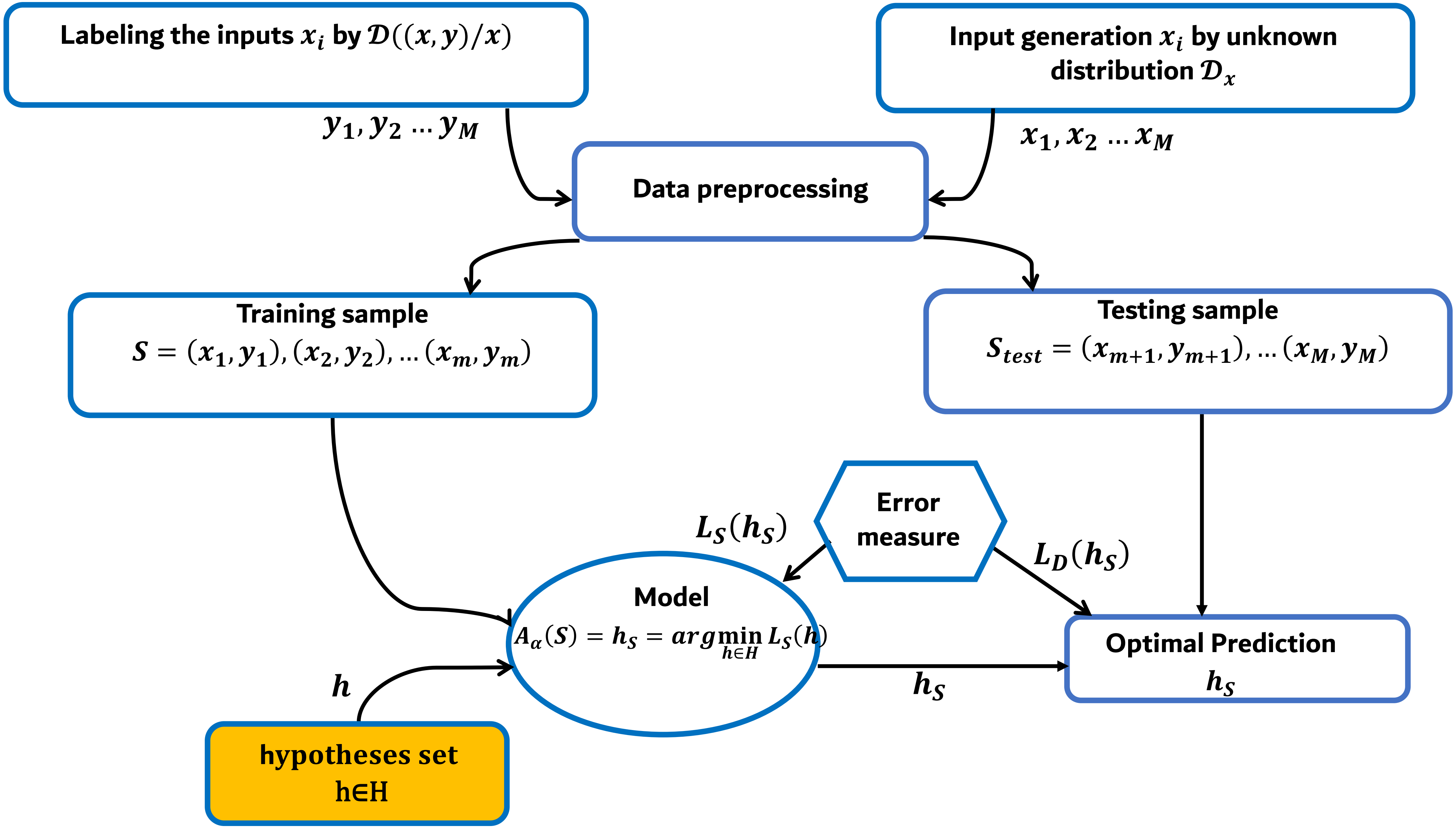
Part 1: Machine learning theory

1. Learning framework
2. Uniform convergence
3. Learnability of infinite size hypotheses set
    1. No-Free-Lunch theorem
    2. Infinite hypothesis class: Exemple
    3. VC dimension
    4. Covering number
4. Tradeoff Bias/Variance
5. Non-Uniform learning

Supervised Learning Passive Offline Algorithm (SLPOA)



Recall (classification)

Definition: shuttering

Let $\boldsymbol{H}$ be a set of functions from X to {0,1} and $A \subseteq X$ a finite set.

We say that $\boldsymbol{H}$ shutters A if the restriction of H over A is of finite cardinality:

$$|\boldsymbol{H}_{\boldsymbol{A}}| = \boldsymbol{2}^{|\boldsymbol{A}|}$$

Such that:

$$\boldsymbol{H}_{\boldsymbol{A}} = \{\boldsymbol{h}(\boldsymbol{a}_{\boldsymbol{1}}), \dots, \boldsymbol{h}\left(\boldsymbol{a}_{|\boldsymbol{A}|}\right) : \boldsymbol{h} \in \boldsymbol{H}\}$$

Recall (classification)

Definition: VC Dimension

The VC dimension is a property of H which measures the maximum size of a set A to be shuttered by H:

$$\boldsymbol{d_{VC}(H)} = \begin{cases} \mathbf{max}\{|\boldsymbol{A}|, \boldsymbol{A\ is\ shuttered\ by\ H}\} \\ +\infty\ \boldsymbol{there\ is\ no\ maximum\ for\ A} \end{cases}$$

Definition: Growth function

Let H be a class of hypothesis, the **growth function** of H is $\Pi_H: \mathbb{N} \longrightarrow \mathbb{N}$, such that:

$$\boldsymbol{\Pi_H(m)} = \max_{\substack{A \subset X \\ |A| = m}} |\boldsymbol{H_A}|$$

$\boldsymbol{H_A}$ **is the restriction of** $\boldsymbol{H}$ **on** $\boldsymbol{A}$**.**

Recall (classification)

Theorem:

Let H be a class of hypotheses such that $d_{VC}(H) = +\infty$. So H is not PAC.

Notice:

- $\forall H$ and $\forall m$, $\Pi_H(m) \leq 2^m$
- If H shutters the class of size m, So: $\Pi_H(m) = 2^m$
- **If $\boldsymbol{d_{VC}(H)} = \max\{|A|, A \text{ is shuttered by } H\} < \boldsymbol{m}$, So: $\boldsymbol{\Pi_H(m) < 2^m}$**

Lemma:

$\forall H$ and $\forall A \subseteq X$: $|H_A| \leq |\{B \subseteq A : H \text{ shutters } B\}|$

Lemma: Sauer

Let H be a class of hypotheses such that: $d_{VC}(H)(\approx) \leq d < +\infty$

Then: $\forall m$, $\Pi_H(m)(\approx) \leq \sum_{i=0}^{d} C_m^i \Rightarrow \log \Pi_H(m) \leq \log \sum_{i=0}^{d} C_m^i$

In particular, if $m > d + 1$, so: $\Pi_H(m)(\approx) \leq \left(\frac{me^1}{d}\right)^d \Rightarrow \log \Pi_H(m) \leq \log \left(\frac{me^1}{d}\right)^d$

$$\Rightarrow 4 + \sqrt{\log(\Pi_H(m))} \leq 4 + \sqrt{\log \left(\frac{me^1}{d}\right)^d}$$

Recall(classification)

Theorem: Generalization bound of VC

Let H be a class of hypotheses and Π_H is its growth function. So, for any D and for any $\delta \in [0,1]$:

$$\boldsymbol{P_{S\sim D^m}(|L_D(h) - L_S(h)| \leq \varepsilon) \geq 1 - \delta}$$

Such that: $\boldsymbol{\varepsilon = \frac{4+\sqrt{log\left(\Pi_H(2m)\right)}}{\delta\sqrt{2m}}}$

Lemma: Sauer

Let H be a class of hypotheses such that: $d_{VC}(H)(\approx) \leq d < +\infty$

Then: $\forall m$, $\Pi_H(m)(\approx) \leq \sum_{i=0}^{d} C_m^i \Rightarrow \log \Pi_H(m) \leq \log \sum_{i=0}^{d} C_m^i$

In particular, if $m > d + 1$, so: $\Pi_H(m)(\approx) \leq \left(\frac{me^1}{d}\right)^d \Rightarrow \log \Pi_H(m) \leq \log \left(\frac{me^1}{d}\right)^d$

$$\Rightarrow 4 + \sqrt{\log\left(\Pi_H(m)\right)} \leq 4 + \sqrt{\log\left(\frac{me^1}{d}\right)^d}$$

Recall(classification)

Theorem:

Let H be a class of hypotheses in $X \times \{0,1\}$.

Let l be the classification loss function.

We have equivalence between:

1. H follows a uniform convergence.
2. H is agnostic PAC learnable by ERM.
3. H is agnostic PAC learnable.
4. H is PAC learnable.
5. H is PAC learnable by ERM.
6. $d_{VC}(H)$ is finite.

Notice:

The VC dimension is a tool characterizing the PAC learning.

Covering number

1. **Background**
2. **Covering numbers in a general metric space**
3. **Covering numbers in Euclidean space**
4. **Uniform convergence in a Real-valued Function class $\boldsymbol{H}$**

1. Background

Definition: Metric space

(M, d) is called a metric space that consists of a set M together with a metric $d: M \times M \to [0, \infty)$ that satisfies the following for all $x, y, z \in M$:

- $d(x, y) = 0 \quad \Longrightarrow \quad x = y.$
- $d(x, y) = d(y, x).$
- $d(x, z) \leq d(x, y) + d(y, z).$

Definition: Open d-ball

An open d-ball centered at $x \in M$ is defined as:

$$B_{d,\varepsilon}(x) = \{y \in M \mid d(x, y) < \varepsilon\}$$

2. Covering numbers in a general metric space

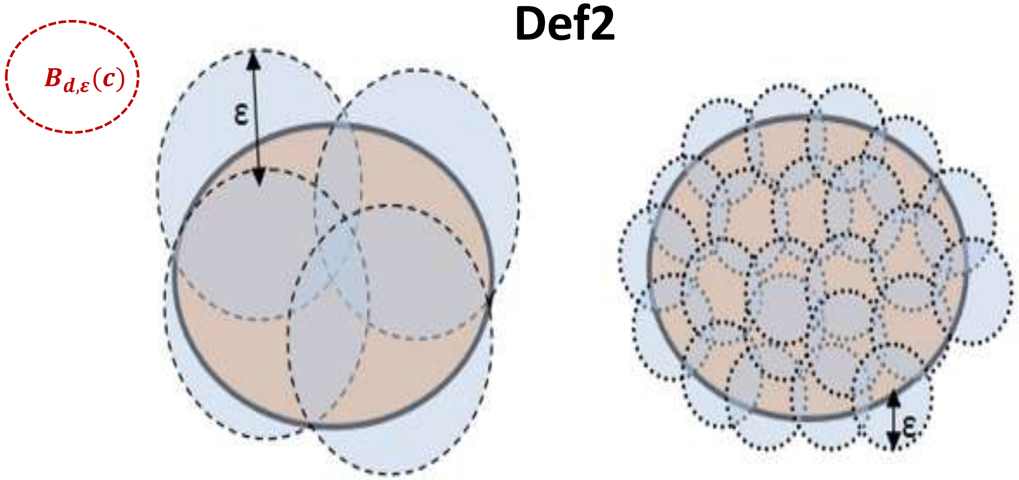
Definition: ε-cover

Let (M, d) be a metric space. Let $W \subseteq M$ and let $\varepsilon > 0$

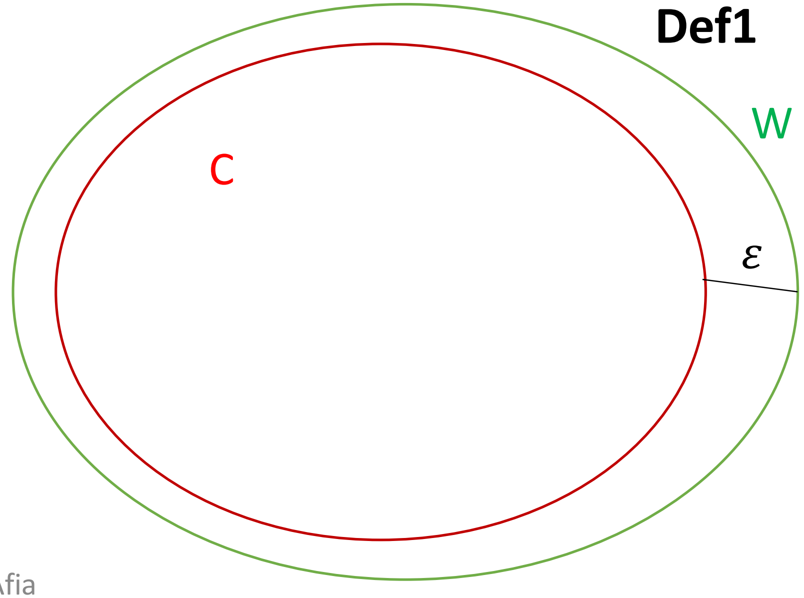
- **Def1:** A set $C \subseteq W$ is said to be ε-cover of W with respect to d if $(\forall w \in W)(\exists c \in C)$ such that: $d(w, c) < \varepsilon$

- **Def2:** $C \subseteq W$ is an ε-cover of W with respect to d if the union of (open) d-balls of radius ε centered at points in C contains W: $\bigcup_{c \in C} B_{d,\varepsilon}(c) \supseteq W$

Def2



Def1



2. Covering numbers in a general metric space

Definition: ε-covering number

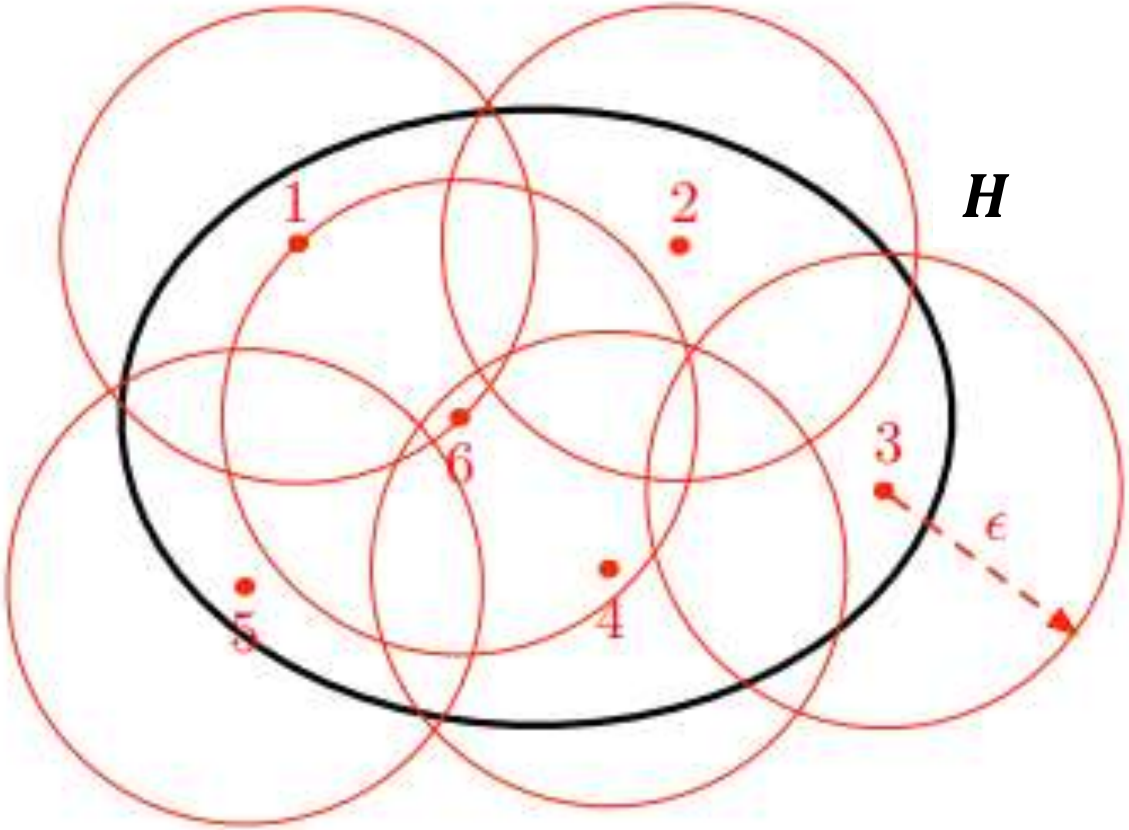
The ε-covering number $\mathcal{N}(\boldsymbol{\varepsilon}, \boldsymbol{W}, \boldsymbol{d})$ of W with respect to d is defined as the cardinality of the smallest ε-cover of W if W has a finite ε-cover with respect to d. Otherwise, if W does not have a finite ε-cover with respect to d, ε-covering number is equal to infinity.

$$\boldsymbol{\mathcal{N}(\varepsilon, W, d)} = \begin{cases} \boldsymbol{min\{|C|, C\ is\ an\ \varepsilon - cover\ of\ W\ with\ respect\ to\ d\}} \\ \infty \qquad\qquad \boldsymbol{if\ W\ does\ not\ have\ a\ finite\ \varepsilon - cover} \end{cases}$$

2. Covering numbers in a general metric space

Example:

For instance, for the H shown in the figure the set of points {1, 2, 3, 4, 5, 6} is a covering. However, the covering number is 5 as point 6 can be removed from the set C and the resulting points are still a covering. C= {1, 2, 3, 4, 5}



3. Covering numbers in Euclidean space

Consider now $M = \mathbb{R}^n$. We can define a number of different metrics on $\mathbb{R}^n$, including in particular the following:

$$d_1(x, y) = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|$$

$$d_2(x, y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$d_\infty(x, y) = \max_i |x_i - y_i|$$

3. Covering numbers in Euclidean space

Accordingly, for any $W \subseteq \mathbb{R}^n$, we can define the corresponding covering numbers $\mathcal{N}(\varepsilon, W, d)$ for $p = 1,2,\infty$.

It is easy to see that:

$$d_1(x,y) \leq d_2(x,y) \leq d_\infty(x,y)$$

Therefore, the corresponding covering numbers satisfy the relation:

$$\mathcal{N}(\varepsilon, W, d_1) \leq \mathcal{N}(\varepsilon, W, d_2) \leq \mathcal{N}(\varepsilon, W, d_\infty)$$

4.Uniform covering numbers for a real-valued function class H

Definition: uniform covering number

Let H be a class of real-valued functions on X:

$$H = \{h |\, h: X \longrightarrow \mathbb{R}\}$$

And let $S = \{x_1, \ldots, x_m\} \subset X$. Then the $H(S) = H_S \subseteq \mathbb{R}$.

For any $\varepsilon > 0$ and $m \in N$, the uniform d_p covering numbers of H for $p = 1,2,\infty$ are defined as:

$$\boldsymbol{\mathcal{N}_p(\varepsilon, H, m)} = \begin{cases} \boldsymbol{\max_{S \subset X:\, |S| = m} \mathcal{N}(\varepsilon, H_S, d_p) \;\; if\; \mathcal{N}(\varepsilon, H_S, d_p)\; is\; finite\; for\; all\; S \subset X} \\ \infty \qquad\qquad \boldsymbol{otherwise} \end{cases}$$

Notice: The number of "uniform" refers to the maximum over all S $\subset X$. It has no relationship with uniform convergence.

4.Uniform covering numbers for a real-valued function class H

Background:

- $S = \{(x, y)\} \subseteq X \times Y$: **is data set with** $Y \subseteq \mathbb{R}$
- $\mathbb{R}^X$: **real space Function**
  - $\mathbb{R}^X = \{h \in \mathbb{R}^X \,|\, h: X \longrightarrow \mathbb{R}\} \Longrightarrow \forall x \in X \;\; h(x) \in \mathbb{R}$

- $S_h = \{(x, h(x))\} \subseteq X \times \hat{Y}$: **is data set with label** $h(x)$
- $\hat{Y} \subseteq \mathbb{R}$ then $\hat{Y}^X \subseteq \mathbb{R}^X$
  - $\hat{Y}^X = H = \{h \in \hat{Y}^X \,|\, h: X \longrightarrow \hat{Y}\} \subseteq \hat{Y}^X \Longrightarrow \;\forall x \in X \;\; h(x) \in \hat{Y}$
  - $l: Y \times \hat{Y} \longrightarrow \mathbb{R}^+$such that $l\left(y, \hat{y} \in \hat{Y}\right)$ is loss function betwen y and $\hat{y}$
  - $l_h: X \times Y \longrightarrow \mathbb{R}^+$such that $l\left(y, h(x) \in \hat{Y}\right)$ is loss function betwen y and $h(x)$

4.Uniform covering numbers for a real-valued function class H

Let's assume that H takes values in some set $\hat{Y} \subseteq \mathbb{R}$, so that $H \subseteq \hat{Y}^X$.

We will require the **loss function l** to be bounded. we will assume $\exists B > 0$ such that:

$(\forall y \in Y)(\forall \hat{y} \in \hat{Y}) \qquad 0 \leq l(y, \hat{y}) \leq B$ and $l: Y \times \hat{Y} \longrightarrow [0, B]$

Definition: The loss function class

We will find it useful to define for any function class $H \subseteq \hat{Y}^X$ and loss $l: Y \times \hat{Y} \longrightarrow [0, B]$ the loss function class $l_H \subseteq [0, B]^{X \times Y}$ given by:

$$l_H = \left\{ l_h: X \times Y \longrightarrow [0, B] \mid l_h(x, y) = l\big(y, h(x) \in \hat{Y}\big)\ for\ some\ h \in H \right\}$$

Uniform convergence in a Real-valued Function class H

Theorem: generalization bound

Let the sets $Y, \hat{Y} \subseteq \mathbb{R}$. Let $H \subseteq \hat{Y}^X$, and let $l: Y \times \hat{Y} \longrightarrow [0, B]$.

Let D be any distribution on $X \times Y$.

For any $\varepsilon > 0$:

$$\underset{S\sim D^m}{\mathrm{P}}\left(\sup_{h\in H}|L_D(h) - L_S(h)| \geq \varepsilon\right) \leq \delta = 4\,\mathcal{N}_1\left(\frac{\varepsilon}{8}, l_H, 2m\right) e^{-m\varepsilon^2/32B^2}$$

Uniform convergence in a Real-valued Function class H

Lemma: L-Lipschitz loss

Let $Y, \hat{Y} \subseteq \mathbb{R}$.

Let $H \subseteq \hat{Y}^X$, and let $l: Y \times \hat{Y} \longrightarrow [0, B]$.

l is Lipschitz in its second argument with Lipschitz constant $L > 0$, if and only if:

$$|l(y, \hat{y}_1) - l(y, \hat{y}_2)| \leq L|\hat{y}_1 - \hat{y}_2| \quad \forall y \in Y, \hat{y}_1, \hat{y}_2 \in \hat{Y}$$

Then for any $m \in N$

$$\mathcal{N}_1(\varepsilon, l_F, m) \leq \mathcal{N}_1(\frac{\varepsilon}{L}, H, m)$$

Uniform convergence in a Real-valued Function class H

Corollary: generalization bound

Let $Y, \hat{Y} \subseteq \mathbb{R}$.

Let $H \subseteq \hat{Y}^X$, and let $l: Y \times \hat{Y} \longrightarrow [0, B]$ such that l is Lipchitz in its second argument with Lipschitz constant $L > 0$.

Let D be any distribution on $X \times Y$.

For any $\varepsilon > 0$:

$$\underset{S \sim D^m}{\mathrm{P}} \left(\sup_{h \in H} |L_D(h) - L_S(h)| \geq \varepsilon \right) \leq \delta = 4\, \mathcal{N}_1 \left(\frac{\varepsilon}{8L}, H, 2m \right) e^{-m\varepsilon^2/32B^2}$$