Introduction to Machine Learning (PART2)

Marc Sebban

Laboratoire Hubert Curien, UMR CNRS 5516
University of Jean Monnet Saint-Étienne (France)

Outline

General Introduction - Machine Learning Settings

Supervised Learning

- True risk – Empirical risk – Loss functions
- Curse of Dimensionality – Overfitting/Underfitting
- Regularization – Norms – Sparsity

Statistical Learning Theory

- **Generalization bounds**
  - **Uniform convergence**
  - **Uniform stability**
  - **Algorithmic robustness**
- **Bias/Variance trade-off**

Model Selection/Cross-Validation

Learning from Imbalanced Data/Data Preparation

k-Nearest Neighbor Classifier

Ensemble Methods – Theory of Boosting (Adaboost)

Statistical Learning Theory

Difference between h and h^*

Reminder

- **Classifier learned from** S: h is the hypothesis learned by your machine learning algorithm from the training data

$$h = \arg\min_{h_i \in \mathcal{H}} \hat{\mathcal{R}}(h_i) = \arg\min_{h_i \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

Difference between h and h^*

Reminder

- **Classifier learned from** S: h is the hypothesis learned by your machine learning algorithm from the training data

$$h = \arg\min_{h_i \in \mathcal{H}} \hat{\mathcal{R}}(h_i) = \arg\min_{h_i \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

- **Optimal classfier**: h^* is the best hypothesis at test time

$$h^* = \arg\min_{h_i \in \mathcal{H}} \mathcal{R}(h_i) = \arg\min_{h_i \in \mathcal{H}} \mathbb{E}_{z \sim \mathcal{D}_{\mathcal{Z}}} \ell(h, z)$$

Difference between h and h^*

Reminder

- **Classifier learned from** S: h is the hypothesis learned by your machine learning algorithm from the training data

$$h = \arg\min_{h_i \in \mathcal{H}} \hat{\mathcal{R}}(h_i) = \arg\min_{h_i \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

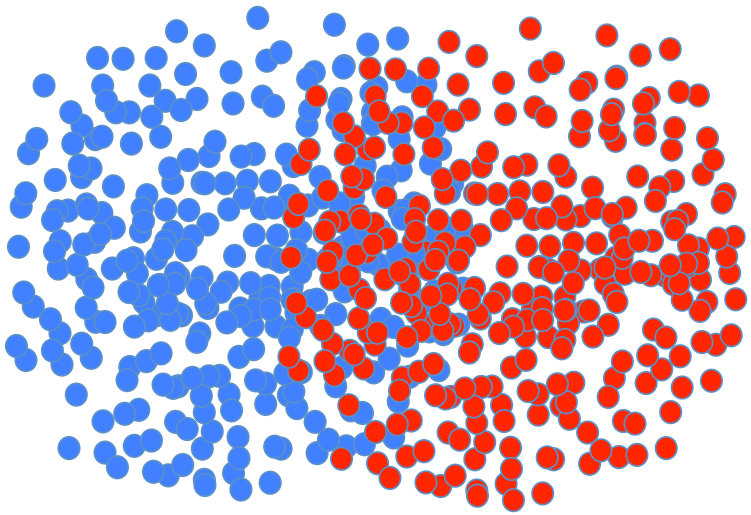
- **Optimal classfier**: h^* is the best hypothesis at test time

$$h^* = \arg\min_{h_i \in \mathcal{H}} \mathcal{R}(h_i) = \arg\min_{h_i \in \mathcal{H}} \mathbb{E}_{z \sim \mathcal{D}_{\mathcal{Z}}} \ell(h, z)$$

Bad news: most of the time $h \neq h^*$...

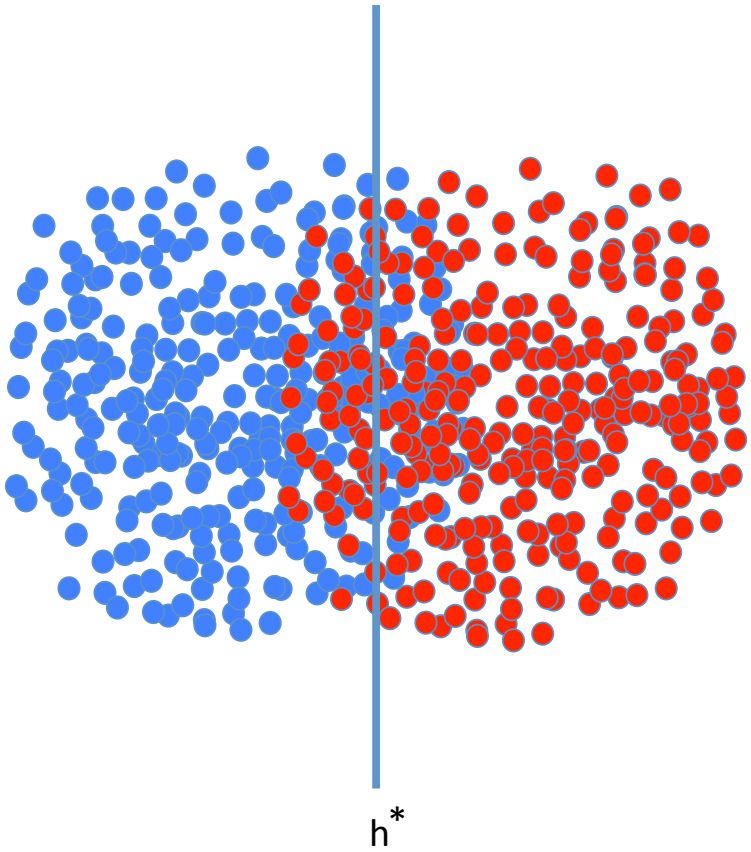
h versus h^*: A toy example

Given the following distribution $\mathcal{D}_{\mathcal{Z}}$ and $\mathcal{H}$: the set of lines $\in \mathbb{R}^2$.



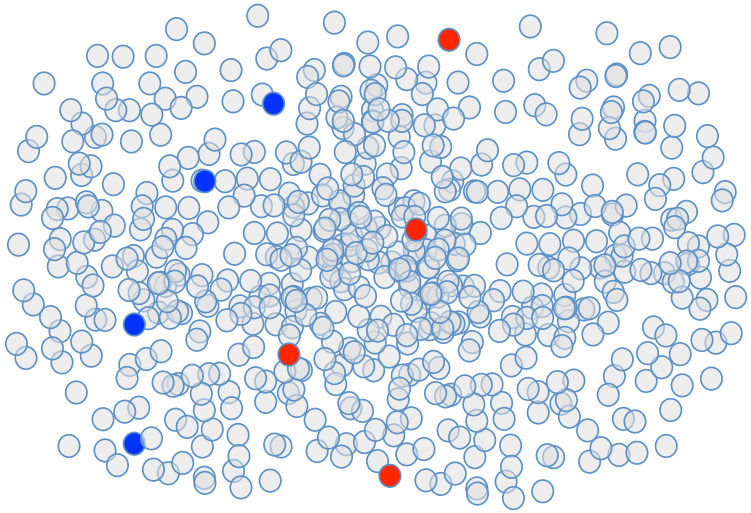
h versus h^*: A toy example

We know that: $h^* = \arg\min_{h_i \in \mathcal{H}} \mathcal{R}(h_i)$ with $\mathcal{R}(h) = \mathbb{E}_{z \sim \mathcal{D}_{\mathcal{Z}}} \ell(h, z)$.



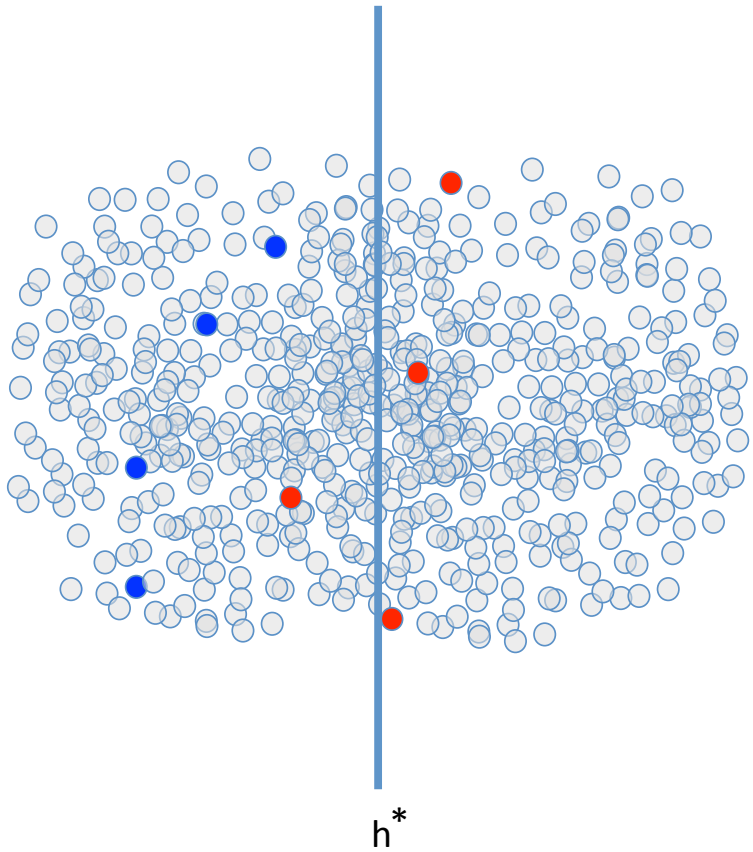
h versus h^*: A toy example

Let us define a sample S drawn from $\mathcal{D}_{\mathcal{Z}}$.



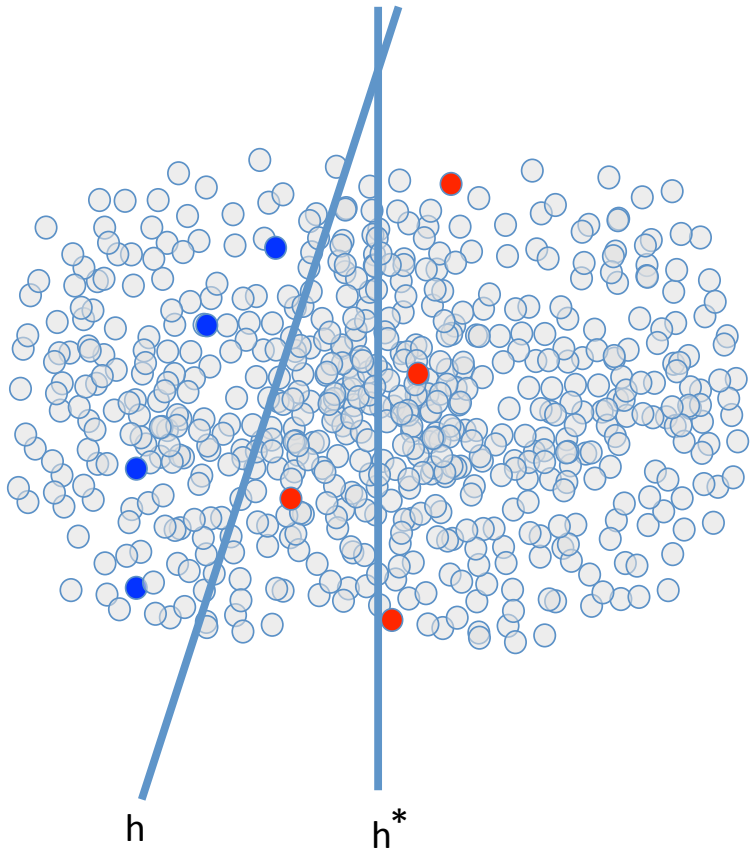
h versus h^*

$h^* = \arg\min_{h_i \in \mathcal{H}} \mathcal{R}(h_i)$ makes errors on S...



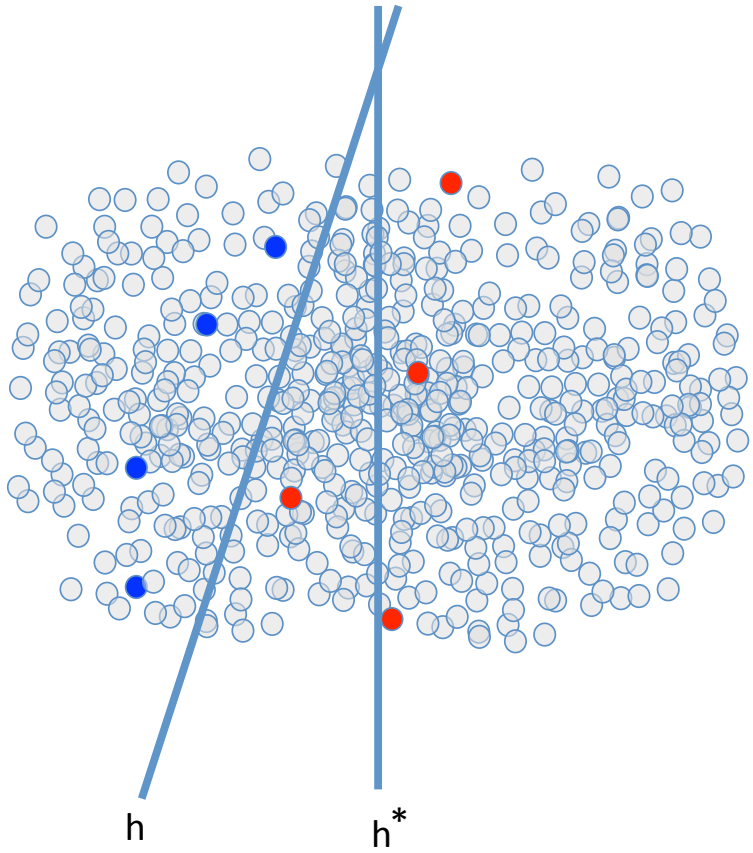
h versus h^*: A toy example

...while $h = \arg\min_{h_i \in \mathcal{H}} \hat{\mathcal{R}}(h_i)$ does not...



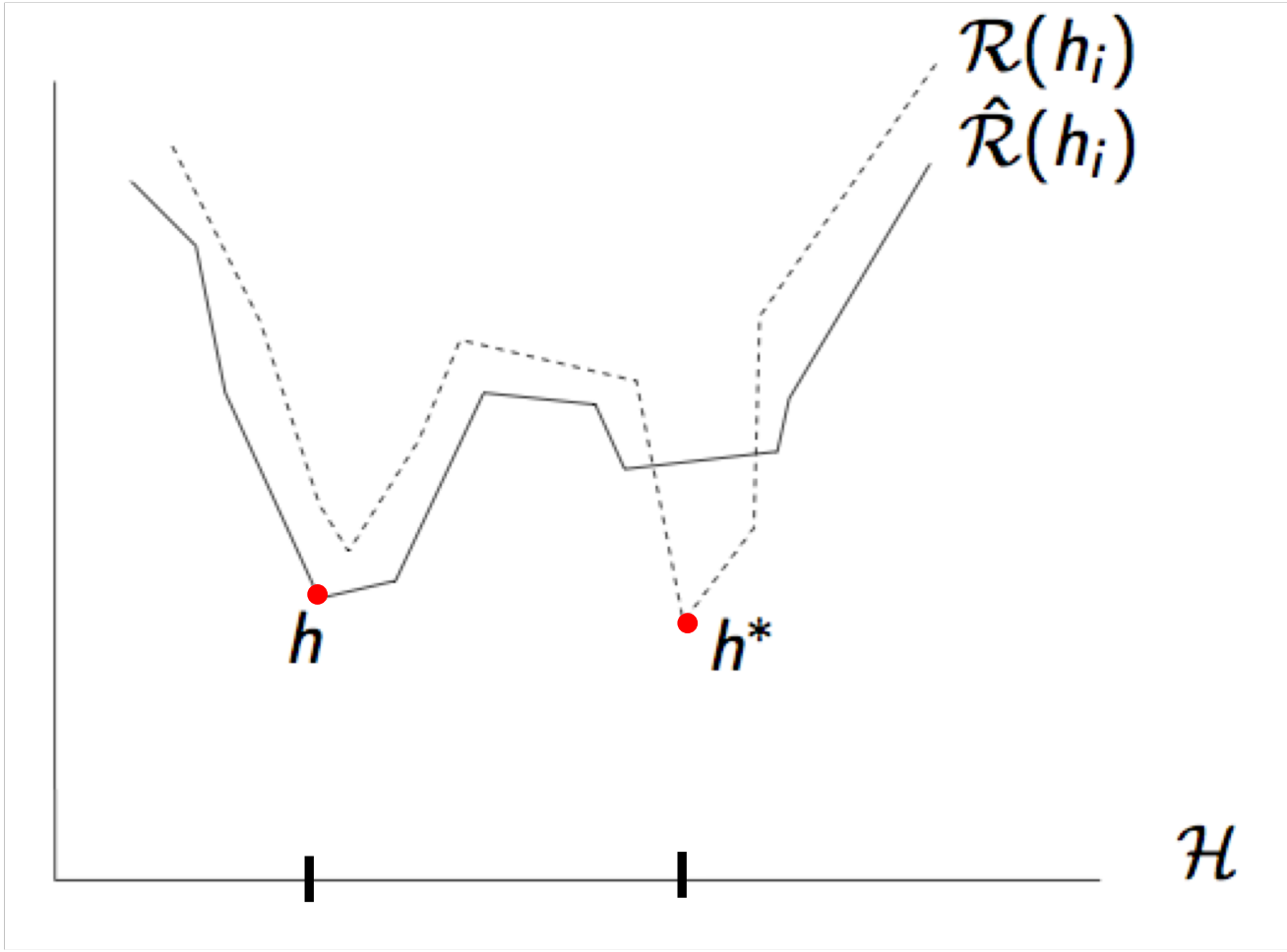
h versus h^*: A toy example

Thus, $\forall S, \forall h \in \mathcal{H}, \hat{\mathcal{R}}(h) \leq \hat{\mathcal{R}}(h^*)$ while $\mathcal{R}(h^*) \leq \mathcal{R}(h)$!



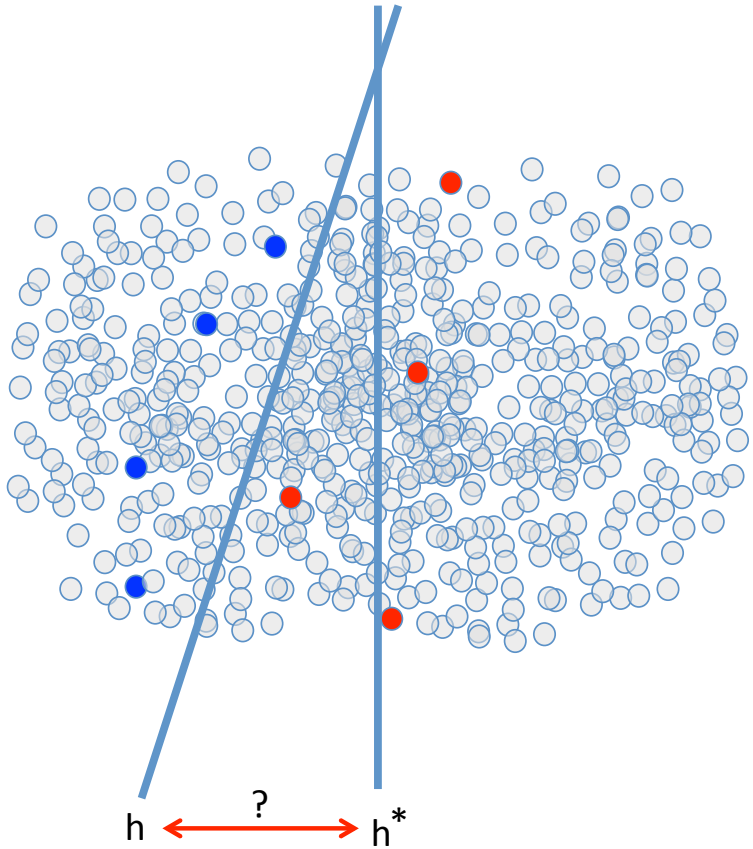
h versus h^*: A toy example

Thus, $\forall S, \forall h \in \mathcal{H}, \hat{\mathcal{R}}(h) \leq \hat{\mathcal{R}}(h^*)$ while $\mathcal{R}(h^*) \leq \mathcal{R}(h)$!



h versus h^*: A toy example

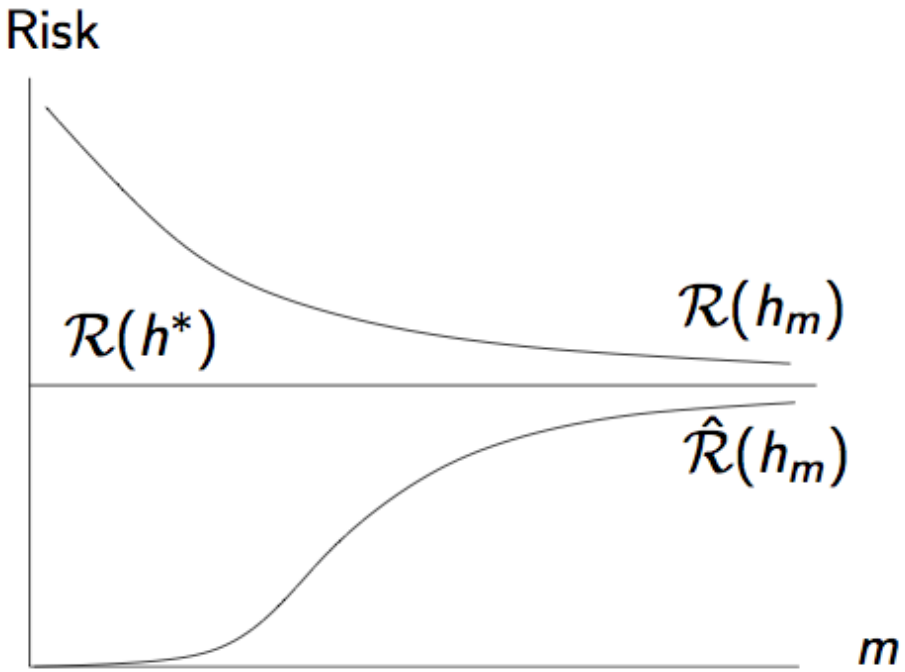
Theoretical question: Under what conditions h converges towards h^*?

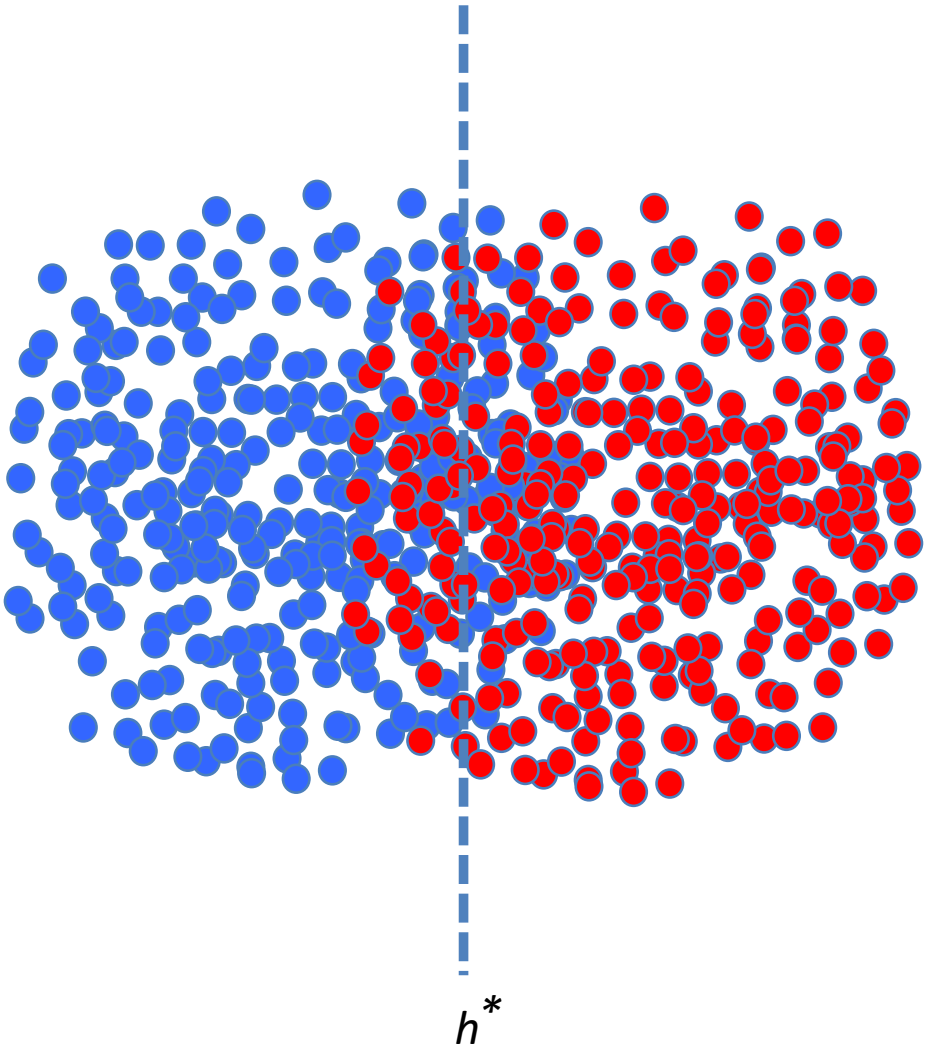


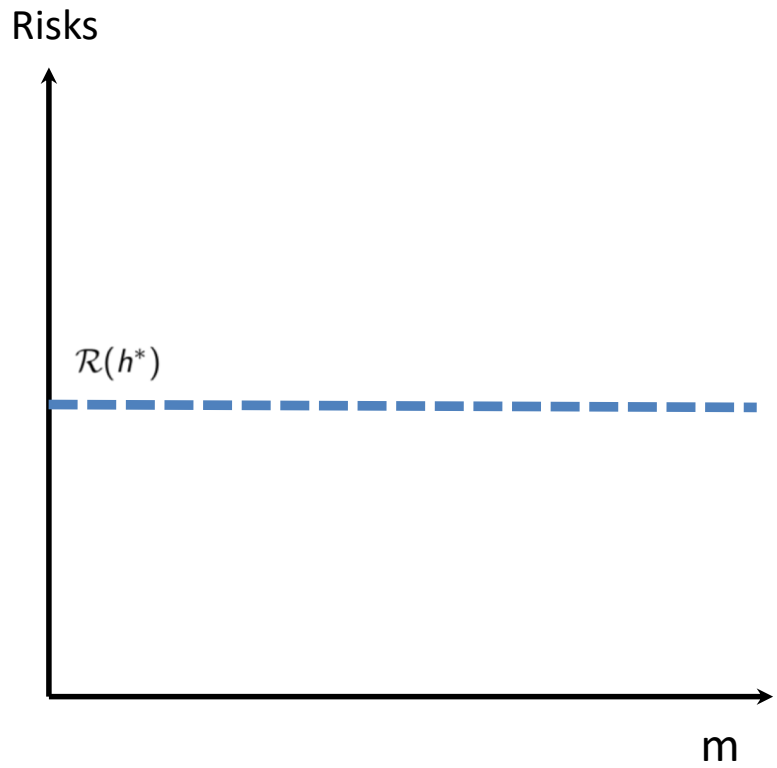
Link between $\hat{\mathcal{R}}(h)$, $\mathcal{R}(h)$ and $\mathcal{R}(h^*)$

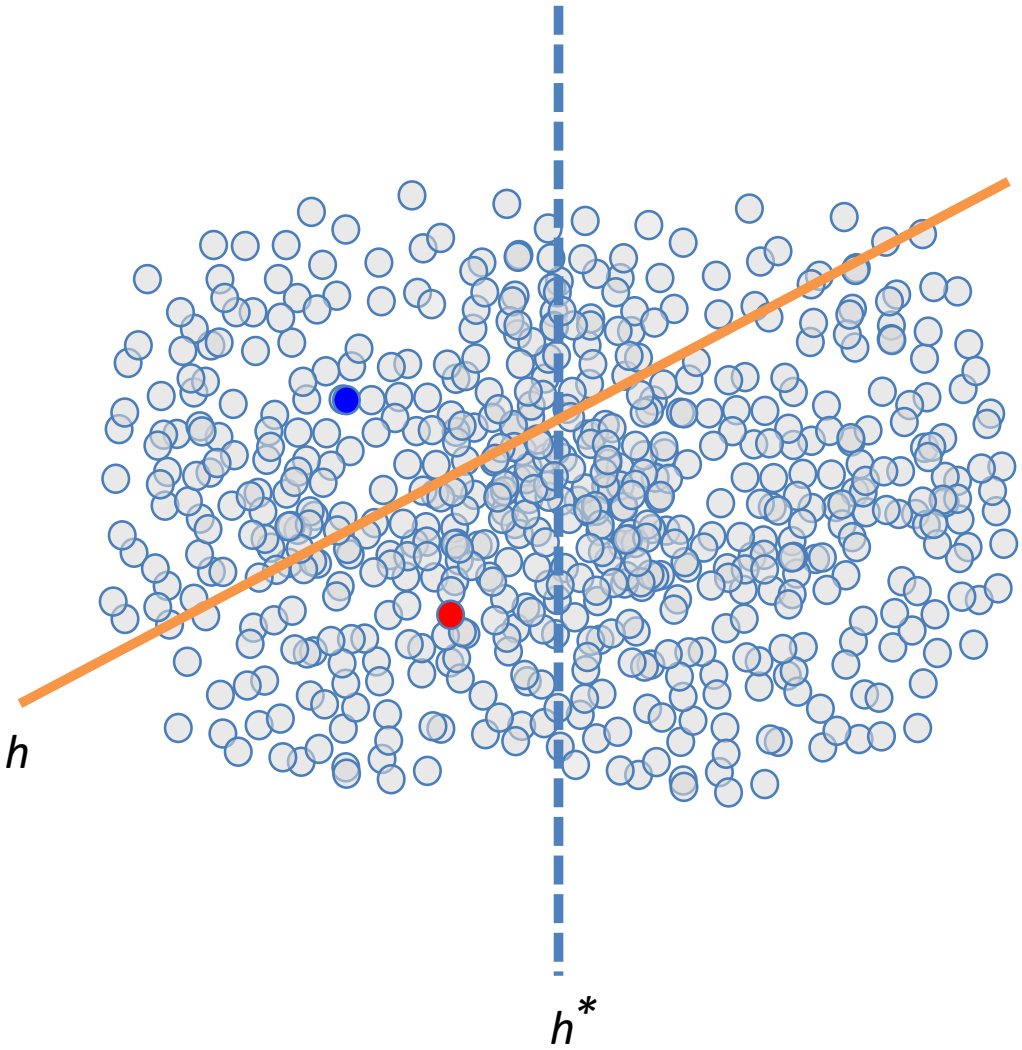
Intuition

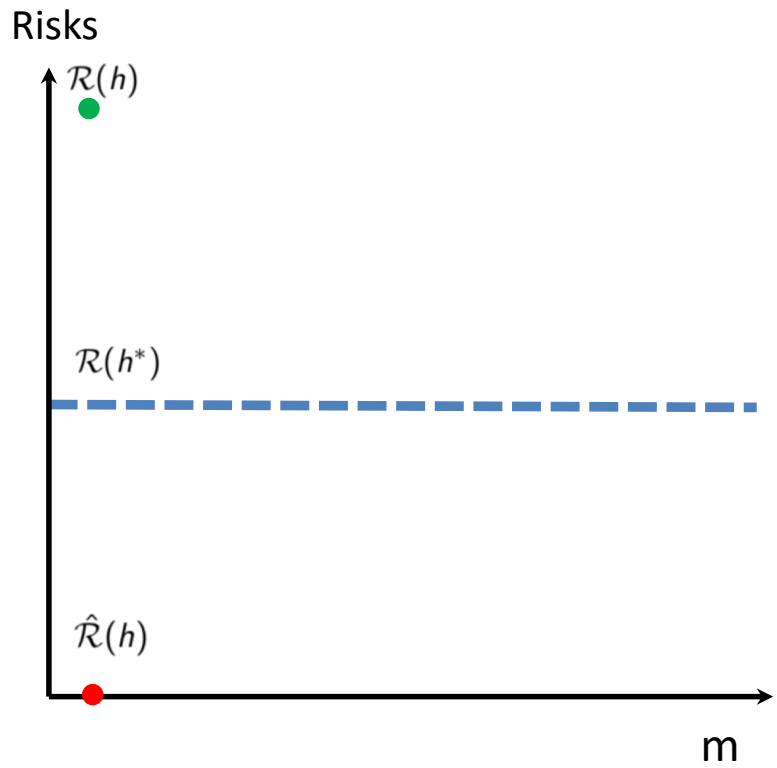
The law of large numbers prompts us to increase the size of the learning set and to search for **the minimal size** m that allows $\hat{\mathcal{R}}(h)$, with a large probability, to be as close to $\mathcal{R}(h)$ and $\mathcal{R}(h^*)$ as possible.

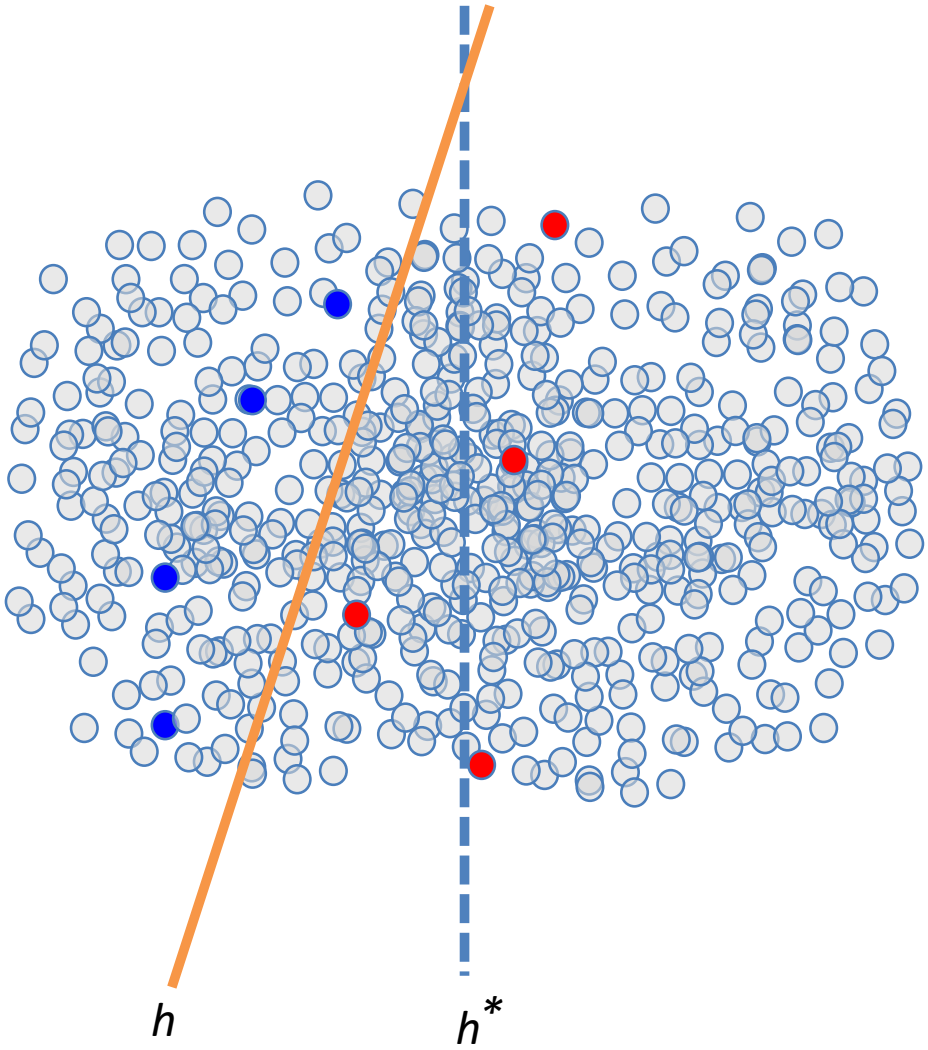


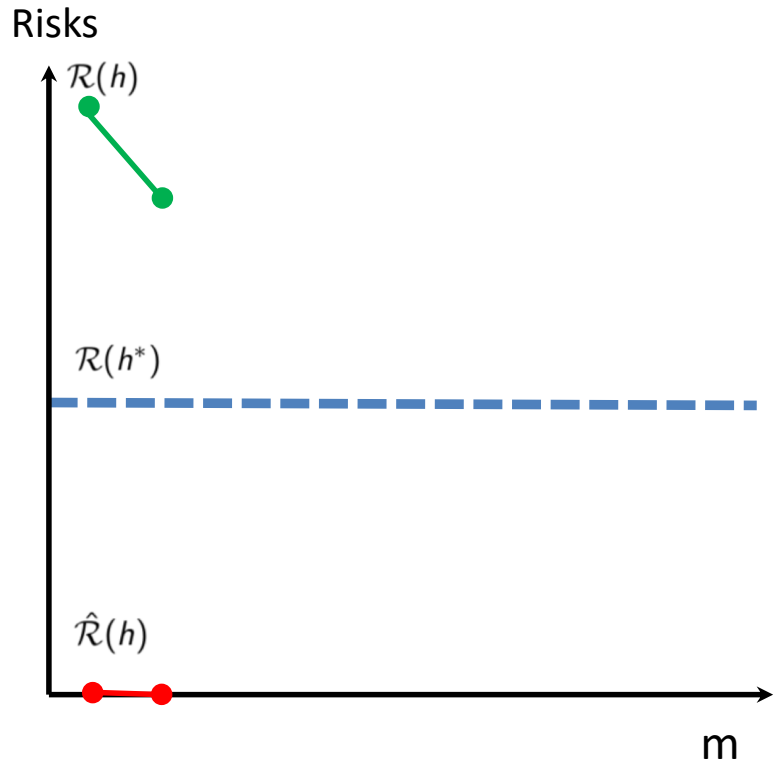

h^*

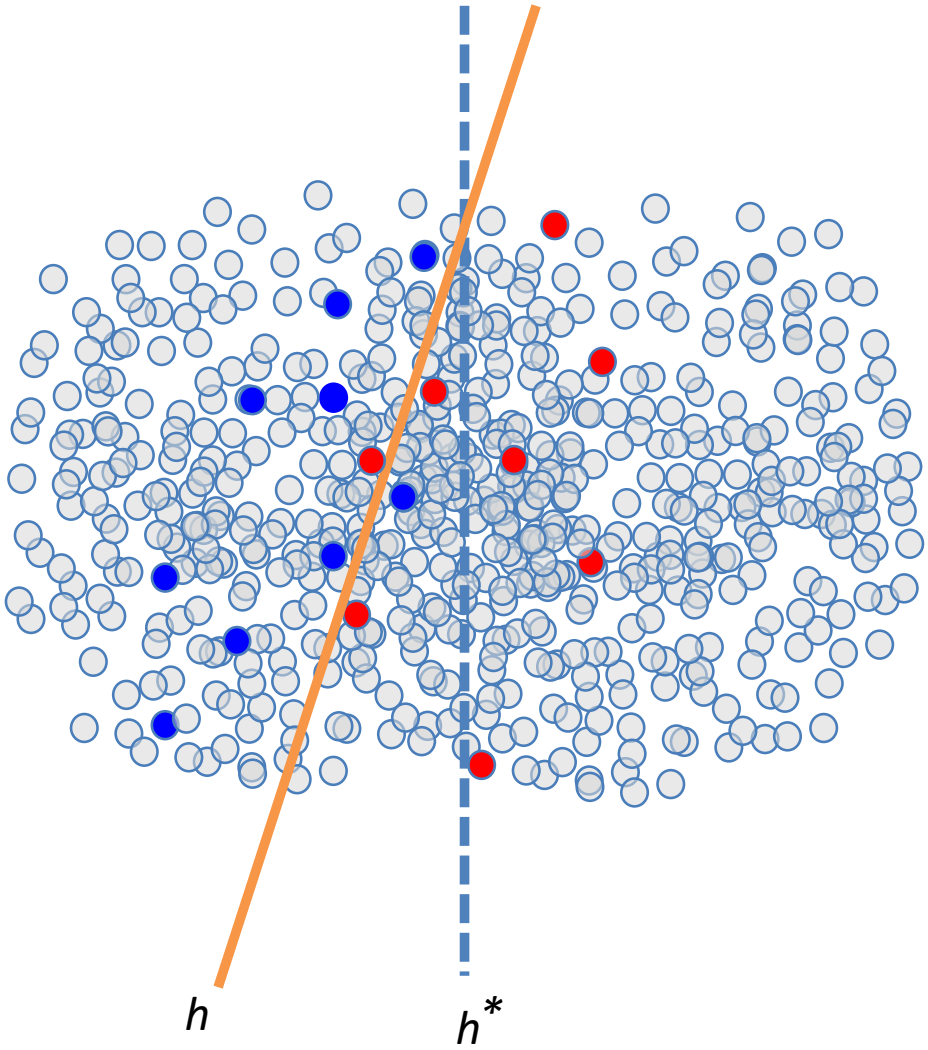

Risks
$\mathcal{R}(h^*)$
m

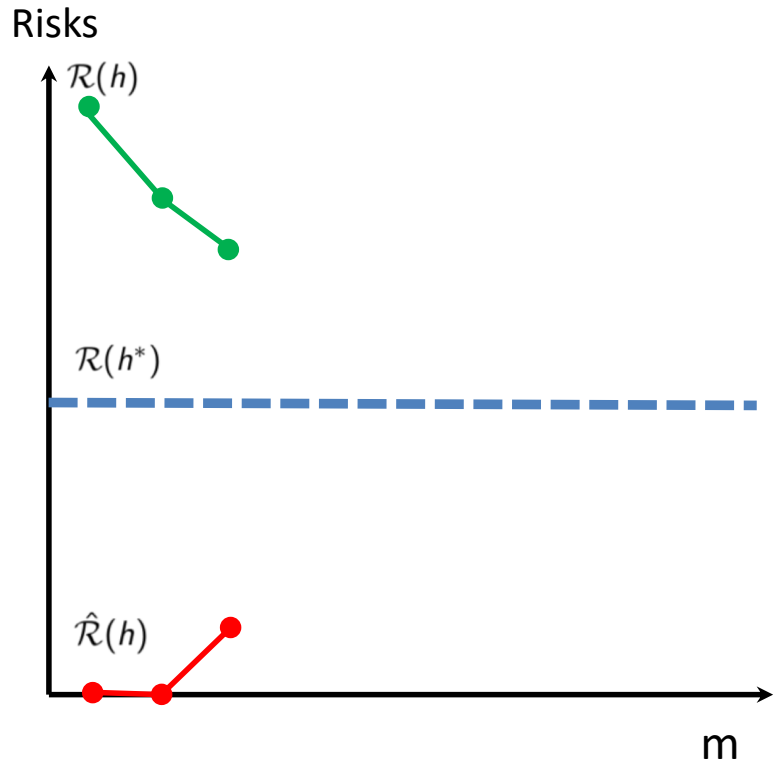

h
h*

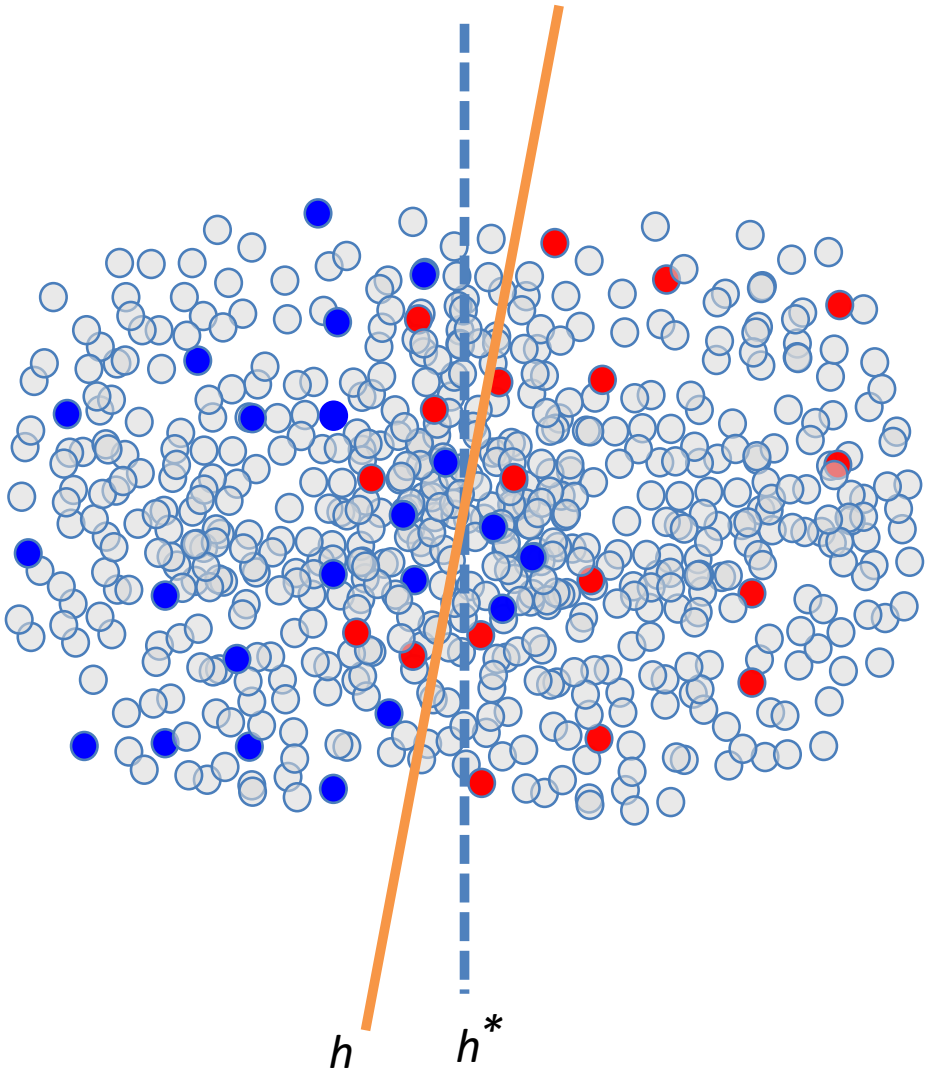

Risks
$\mathcal{R}(h)$
$\mathcal{R}(h^*)$
$\hat{\mathcal{R}}(h)$
m

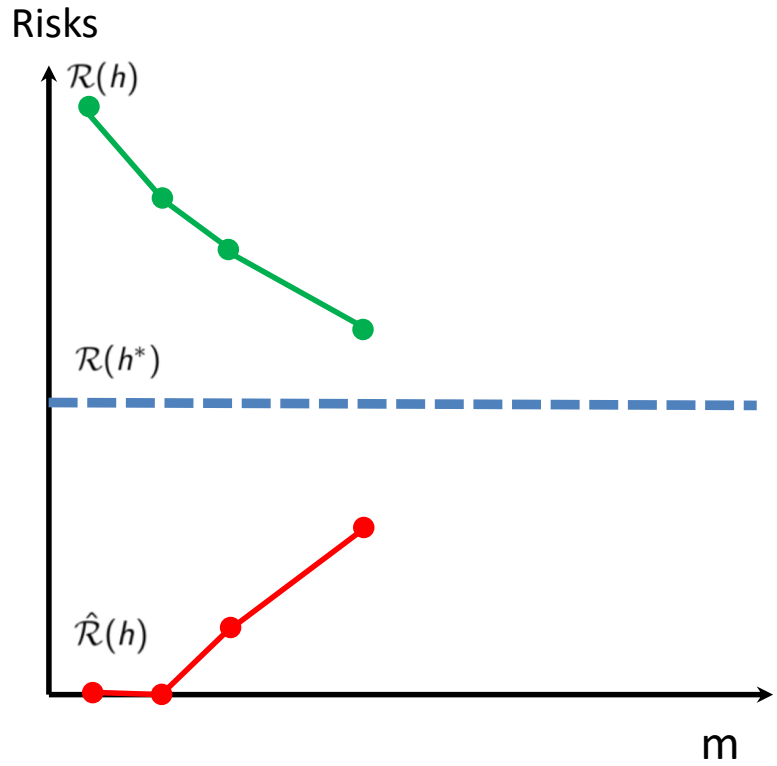

h
h*

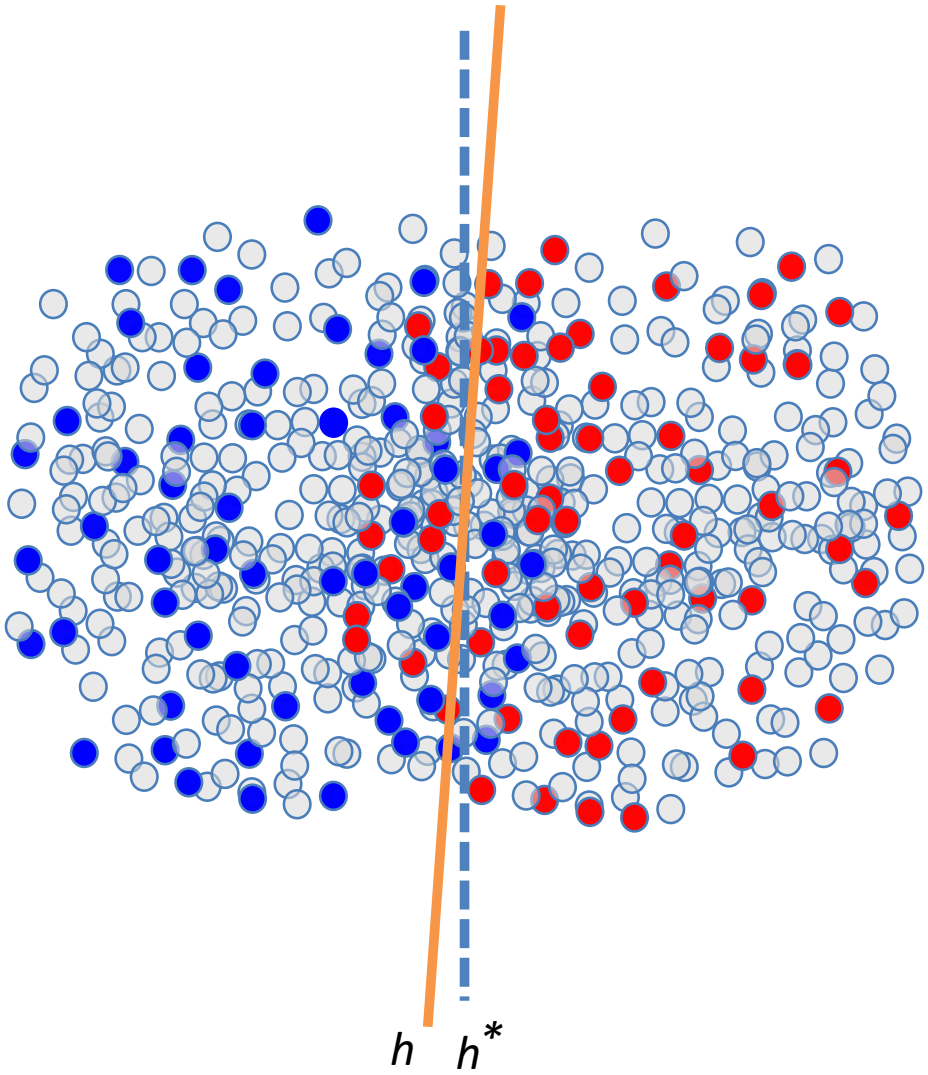

Risks
$\mathcal{R}(h)$
$\mathcal{R}(h^*)$
$\hat{\mathcal{R}}(h)$
m

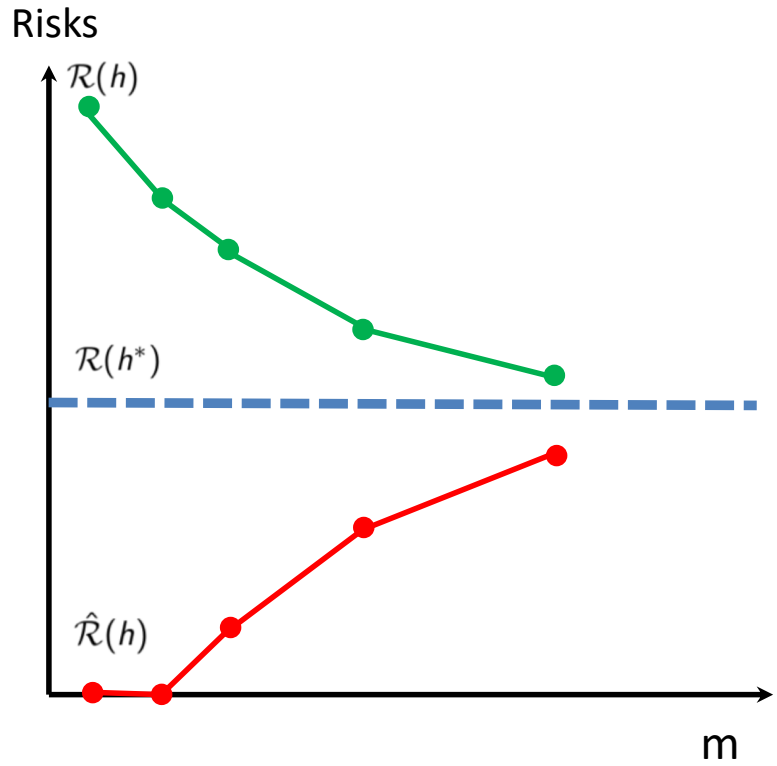

h
h*

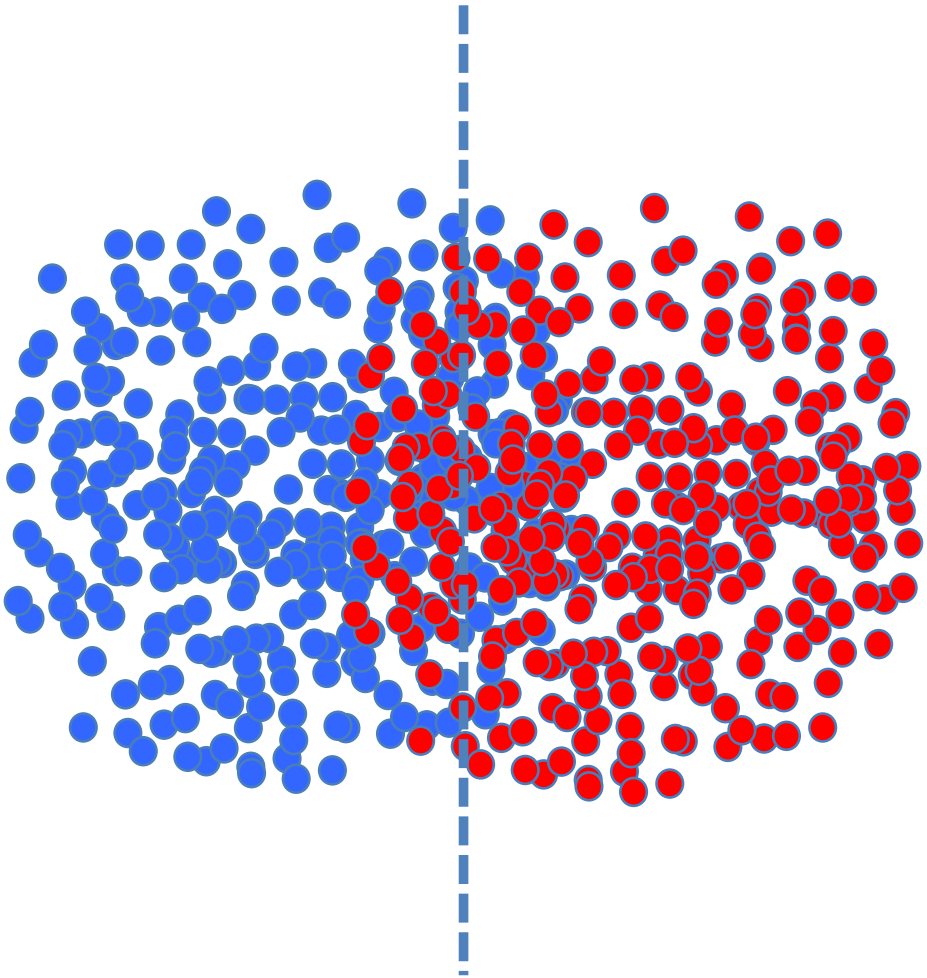

Risks
$\mathcal{R}(h)$
$\mathcal{R}(h^*)$
$\hat{\mathcal{R}}(h)$
m

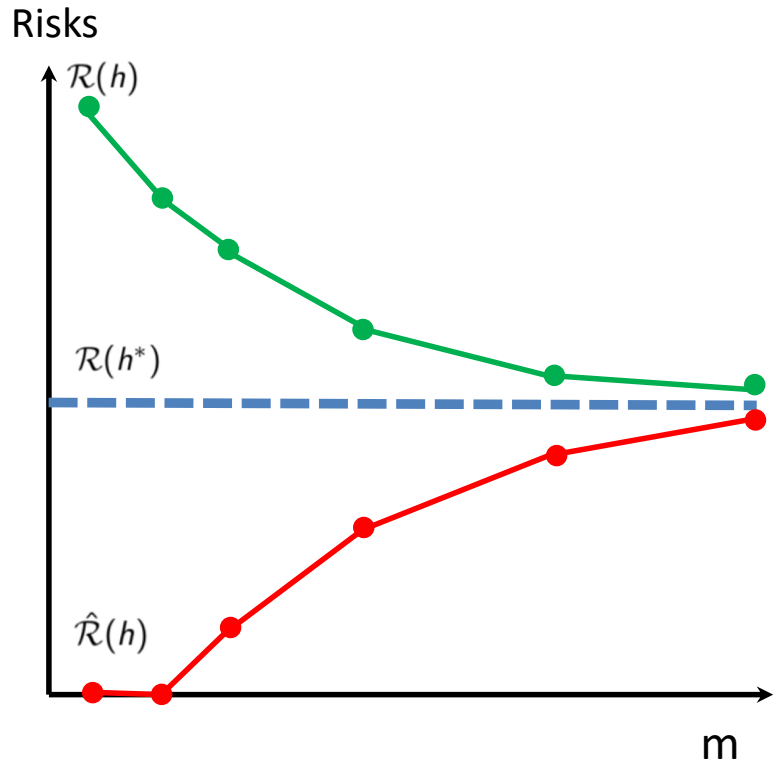

h
h*

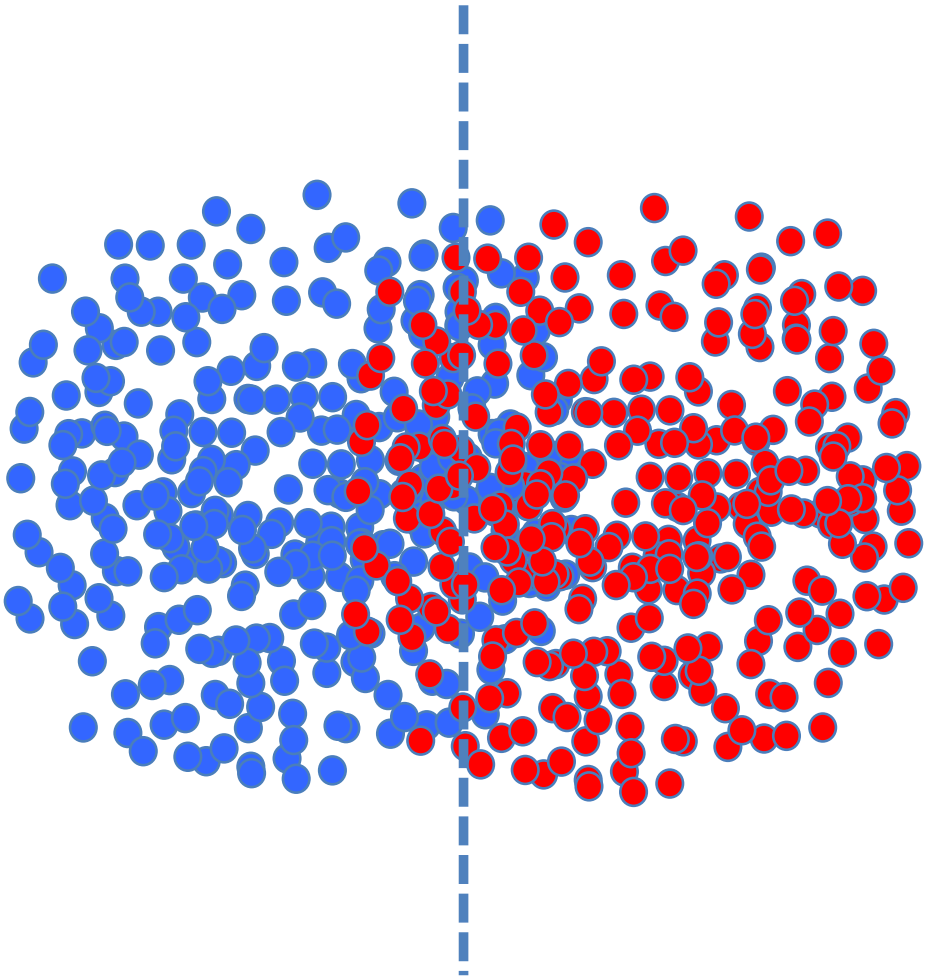

Risks
$\mathcal{R}(h)$
$\mathcal{R}(h^*)$
$\hat{\mathcal{R}}(h)$
m

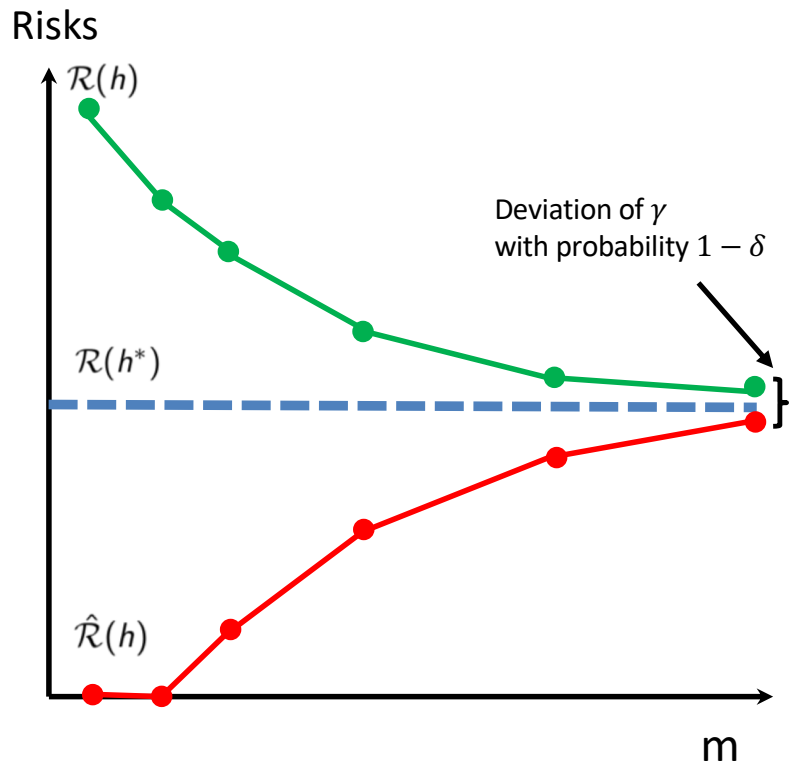

h
h*


Risks
$\mathcal{R}(h)$
$\mathcal{R}(h^*)$
$\hat{\mathcal{R}}(h)$
m


$h^* \approx h$


Risks
$\mathcal{R}(h)$
$\mathcal{R}(h^*)$
$\hat{\mathcal{R}}(h)$
m


$h^* \approx h$


Risks
$\mathcal{R}(h)$
Deviation of γ
with probability $1-\delta$
$\mathcal{R}(h^*)$
$\hat{\mathcal{R}}(h)$
m

Statistical Learning Theory: Empirical Risk Minimization

Condition for trusting h obtained by the minimization of the empirical risk $\hat{\mathcal{R}}(h)$

PAC (Probably Approximately Correct) Condition [Valiant 84]

$$h = \arg\min_{h_i \in \mathcal{H}} \hat{\mathcal{R}}(h_i) \text{ and } h^* = \arg\min_{h_i \in \mathcal{H}} \mathcal{R}(h_i)$$

$$\forall h \in \mathcal{H}, \forall \mathcal{D}_{\mathcal{Z}}, \forall \gamma \geq 0, \forall \delta \leq 1,$$

$$P(|\mathcal{R}(h) - \mathcal{R}(h^*)| \leq \gamma) \geq 1 - \delta$$

Since only $\hat{\mathcal{R}}(h)$ is observable, we need:

1. to relate $\hat{\mathcal{R}}(h)$ to $\mathcal{R}(h)$...
2. and $\mathcal{R}(h)$ to $\mathcal{R}(h^*)$

Statistical Learning Theory

$$\forall h \in \mathcal{H}, \forall \mathcal{D}_{\mathcal{Z}}, \forall \gamma \geq 0, \forall \delta \leq 1,$$

$$P(|\mathcal{R}(h) - \mathcal{R}(h^*)| \leq \gamma) \geq 1 - \delta$$

Question

Under what conditions the previous inequality holds?

Statistical Learning Theory

$$\forall h \in \mathcal{H}, \forall \mathcal{D}_{\mathcal{Z}}, \forall \gamma \geq 0, \forall \delta \leq 1,$$
$$P(|\mathcal{R}(h) - \mathcal{R}(h^*)| \leq \gamma) \geq 1 - \delta$$

Question

Under what conditions the previous inequality holds?

One has to differentiate two different situations:

- When $|\mathcal{H}|$ is finite with two scenarios:
  1. $f \in \mathcal{H}$.
  2. $f \notin \mathcal{H}$.
- When $|\mathcal{H}|$ is infinite (e.g. the set of lines in $\mathbb{R}^2$).

When $|\mathcal{H}|$ is finite and $f \in \mathcal{H}$

Definition (Consistent Learner)

- Input $S = \{z_i = (\mathbf{x_i}, y_i)\}_{i=1}^{m}$.
- Output $h \in \mathcal{H}$ consistent with the sample (i.e $\hat{\mathcal{R}}(h) = 0$) if one exists.

What are the conditions to not select an hypothesis h such that $\hat{\mathcal{R}}(h) = 0$ and $\mathcal{R}(h) \geq \gamma$, therefore which would not be h^* with $\hat{\mathcal{R}}(h^*) = 0$?

When $|\mathcal{H}|$ is finite and $f \in \mathcal{H}$

Definition (Consistent Learner)

- Input $S = \{z_i = (\mathbf{x_i}, y_i)\}_{i=1}^{m}$.
- Output $h \in \mathcal{H}$ consistent with the sample (i.e $\hat{\mathcal{R}}(h) = 0$) if one exists.

What are the conditions to not select an hypothesis h such that $\hat{\mathcal{R}}(h) = 0$ and $\mathcal{R}(h) \geq \gamma$, therefore which would not be h^* with $\hat{\mathcal{R}}(h^*) = 0$?

Theorem

$$m \geq \frac{1}{\gamma}[ln(|\mathcal{H}|) + ln(\frac{1}{\delta})]$$

labeled examples are sufficient so that, with probability $1 - \delta$, *all* $h \in \mathcal{H}$ *with* $\mathcal{R}(h) \geq \gamma$ *have* $\hat{\mathcal{R}}(h) > 0$ *(then,* h *won't be induced because it is not consistent).*

Sample complexity

Sketch of proof

Assume that there exist k **bad** hypotheses $h_1, h_2, \ldots, h_k$ such that $\mathcal{R}(h_i) \geq \gamma$

Sample complexity

Sketch of proof

Assume that there exist k **bad** hypotheses $h_1, h_2, \ldots, h_k$ such that $\mathcal{R}(h_i) \geq \gamma$

1. Take one of the h_i. The probability that h_i is consistent with one training example is $\leq 1 - \gamma$. The probability for h_i to be consistent with first m training samples is $\leq (1 - \gamma)^m$

Sample complexity

Sketch of proof

Assume that there exist k **bad** hypotheses $h_1, h_2, \ldots, h_k$ such that $\mathcal{R}(h_i) \geq \gamma$

1. Take one of the h_i. The probability that h_i is consistent with one training example is $\leq 1 - \gamma$. The probability for h_i to be consistent with first m training samples is $\leq (1 - \gamma)^m$
2. Now, the probability that **at least** one of the h_i is consistent with first m training examples is $\leq k(1 - \gamma)^m \leq |\mathcal{H}|(1 - \gamma)^m$

Sample complexity

Sketch of proof

Assume that there exist k **bad** hypotheses $h_1, h_2, \ldots, h_k$ such that $\mathcal{R}(h_i) \geq \gamma$

1. Take one of the h_i. The probability that h_i is consistent with one training example is $\leq 1 - \gamma$. The probability for h_i to be consistent with first m training samples is $\leq (1 - \gamma)^m$
2. Now, the probability that **at least** one of the h_i is consistent with first m training examples is $\leq k(1 - \gamma)^m \leq |\mathcal{H}|(1 - \gamma)^m$
3. To control the magnitude of this probability, let set $|\mathcal{H}|(1 - \gamma)^m \leq \delta$

Sample complexity

Sketch of proof

Assume that there exist k **bad** hypotheses $h_1, h_2, \ldots, h_k$ such that $\mathcal{R}(h_i) \geq \gamma$

1. Take one of the h_i. The probability that h_i is consistent with one training example is $\leq 1 - \gamma$. The probability for h_i to be consistent with first m training samples is $\leq (1 - \gamma)^m$
2. Now, the probability that **at least** one of the h_i is consistent with first m training examples is $\leq k(1 - \gamma)^m \leq |\mathcal{H}|(1 - \gamma)^m$
3. To control the magnitude of this probability, let set $|\mathcal{H}|(1 - \gamma)^m \leq \delta$
4. Using in (2) the fact that $1 - x \leq e^{-x}$, we get $m \geq \frac{1}{\gamma}[ln(|\mathcal{H}|) + ln(\frac{1}{\delta})]$

Two views on the same result

Finite hypothesis space, realizable case

Sample complexity

$$m \geq \frac{1}{\gamma}[ln(|\mathcal{H}|) + ln(\frac{1}{\delta})]$$

labeled examples are sufficient so that, with probability $\geq 1 - \delta$, all $h \in \mathcal{H}$ with $\mathcal{R}(h) \geq \gamma$ have $\hat{\mathcal{R}}(h) > 0$.

Contrapositive ($A \rightarrow B$ implies $\neg B \rightarrow \neg A$)

With probability at least $1 - \delta$, for all $h \in \mathcal{H}$ such that $\hat{\mathcal{R}}(h) = 0$

$$\mathcal{R}(h) \leq \frac{1}{m}(ln(|\mathcal{H}|) + ln(\frac{1}{\delta}))$$

What if $|\mathcal{H}|$ is finite and we don't know if $f \notin \mathcal{H}$?

In practise, we don't know if $f \in \mathcal{H}$.

Can we (i) say with probability $1-\delta$ that $\forall h \in \mathcal{H}$, $|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \leq \gamma$ and (ii) relate $\hat{\mathcal{R}}(h)$ to $\mathcal{R}(h^*)$?

- Called "uniform convergence".
- Motivates optimizing over S, even if we cannot find a perfect function.

When $|\mathcal{H}|$ is finite

Lemma 1: Union Bound

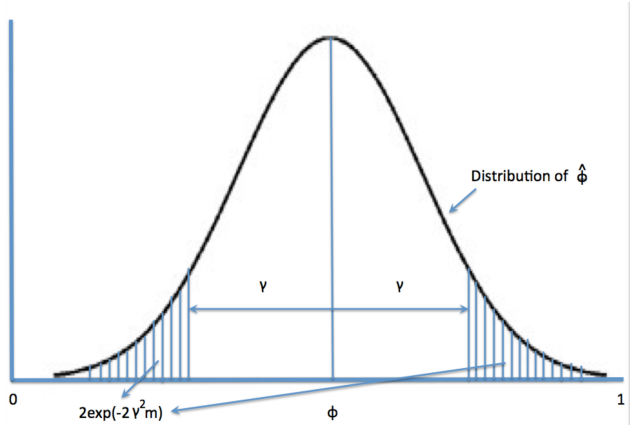
Let $A_1, A_2, ..., A_k$ be k events (not necessary independent). Then

$$P(A_1 \cup A_2, ..., A_k) \leq P(A_1) + P(A_2) + ... + P(A_k)$$

Lemma 2: Hoeffding Inequality

Let $Z_1, Z_2, ..., Z_m$ be m Bernouilli random variables with mean ϕ (i.e. $P(Z_i = 1) = \phi$). Let $\hat{\phi} = \frac{1}{m} \sum_{i=1}^{m} Z_i$ and let any $\gamma > 0$ be fixed. Then

$$P(|\hat{\phi} - \phi| > \gamma) \leq 2exp(-2\gamma^2 m)$$



When $|\mathcal{H}|$ is finite

Hoeffding inequality (Lemma 2) can be applied on $\mathcal{R}(h)$ and $\hat{\mathcal{R}}(h)$ with $\ell(h, z_i)$ a Bernouilli random variable with mean $\mathcal{R}(h)$.

Theorem

For a given $h \in \mathcal{H}, \forall \gamma \geq 0, \forall m > 0, \forall \mathcal{D}_{\mathcal{Z}}$

$$P(|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \geq \gamma) \leq 2e^{-2\gamma^2 m}$$

However, we need a bound that holds uniformly over the whole space of hypotheses. Therefore, we are interested in:

$$P(\sup_{h \in \mathcal{H}} |\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \geq \gamma)$$

When $|\mathcal{H}|$ is finite

Uniform convergence

Let A_j be the event $|\mathcal{R}(h_j) - \hat{\mathcal{R}}(h_j)| \geq \gamma$. By Lemma 2, we get

$$P(A_j) \leq 2e^{-2\gamma^2 m}$$

$$\begin{aligned} P(\sup_{h \in \mathcal{H}} |\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \geq \gamma) &= P(A_1 \cup \ldots \cup A_{|\mathcal{H}|}) \\ &\leq \sum_{i=1}^{|\mathcal{H}|} P(A_i) \text{ (Lemma 1)} \\ &= \sum_{i=1}^{|\mathcal{H}|} 2e^{-2\gamma^2 m} \\ &= 2|\mathcal{H}|e^{-2\gamma^2 m} \end{aligned}$$

Bound on m

From $\forall h \in \mathcal{H}, P(|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \geq \gamma) \leq 2|\mathcal{H}|e^{-2\gamma^2 m}$, $\forall$ probability δ:

$$\begin{aligned}
\text{If} \quad & 2|\mathcal{H}|e^{-2\gamma^2 m} && = \delta \\
\Leftrightarrow \quad & e^{2\gamma^2 m} && = \frac{2|\mathcal{H}|}{\delta} \\
\Leftrightarrow \quad & 2\gamma^2 m && = ln\frac{2|\mathcal{H}|}{\delta} \\
\Leftrightarrow \quad & m && = \frac{1}{2\gamma^2} ln\frac{2|\mathcal{H}|}{\delta}
\end{aligned}$$

So, if $m \geq \frac{1}{2\gamma^2} ln\frac{2|\mathcal{H}|}{\delta}$ then with probability $1 - \delta$, $\forall \gamma$, we have

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \gamma$$

Bound on m

From $\forall h \in \mathcal{H}, P(|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \geq \gamma) \leq 2|\mathcal{H}|e^{-2\gamma^2 m}$, $\forall$ probability δ:

$$\begin{aligned}
\text{If} \quad & 2|\mathcal{H}|e^{-2\gamma^2 m} && = \delta \\
\Leftrightarrow \quad & e^{2\gamma^2 m} && = \frac{2|\mathcal{H}|}{\delta} \\
\Leftrightarrow \quad & 2\gamma^2 m && = ln\frac{2|\mathcal{H}|}{\delta} \\
\Leftrightarrow \quad & m && = \frac{1}{2\gamma^2} ln\frac{2|\mathcal{H}|}{\delta}
\end{aligned}$$

So, if $m \geq \frac{1}{2\gamma^2} ln\frac{2|\mathcal{H}|}{\delta}$ then with probability $1 - \delta$, $\forall \gamma$, we have

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \gamma$$

- There is nothing more practical than a good theory !! [J.C. Maxwell]
- Usually, m is fixed. Given m and δ, can we say something on the deviation γ?

Bound on γ given m and a fixed probability δ

$$\begin{aligned} \text{If} \quad & 2|\mathcal{H}|e^{-2\gamma^2 m} && = \delta \\ \Leftrightarrow \quad & e^{2\gamma^2 m} && = \frac{2|\mathcal{H}|}{\delta} \\ \Leftrightarrow \quad & \gamma && = \sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}} \end{aligned}$$

So, with a probability at least $1-\delta$, we have $\forall h \in \mathcal{H}$:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}}$$

or said differently $P\left(|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \leq \sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}}\right) \geq 1-\delta$

Bound on γ given m and a fixed probability δ

$$\begin{aligned} \text{If} \quad & 2|\mathcal{H}|e^{-2\gamma^2 m} && = \delta \\ \Leftrightarrow \quad & e^{2\gamma^2 m} && = \frac{2|\mathcal{H}|}{\delta} \\ \Leftrightarrow \quad & \gamma && = \sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}} \end{aligned}$$

So, with a probability at least $1-\delta$, we have $\forall h \in \mathcal{H}$:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}}$$

or said differently $P\left(|\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \leq \sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}}\right) \geq 1-\delta$

Reminder: PAC condition

$$\forall \mathcal{D}_{\mathcal{Z}}, \forall \gamma \geq 0, \forall \delta \leq 1, P(|\mathcal{R}(h) - \mathcal{R}(h^*)| \leq \gamma) \geq 1-\delta$$

When $|\mathcal{H}|$ is finite

We know that with a probability $\geq 1-\delta, \forall h \in \mathcal{H}$

$$\gamma = \sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}} \quad \text{and} \quad |\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \leq \gamma$$

so

$$\mathcal{R}(h) - \gamma < \hat{\mathcal{R}}(h) < \mathcal{R}(h) + \gamma$$

Therefore,

$$\begin{aligned}
\mathcal{R}(h) &\leq \hat{\mathcal{R}}(h) + \gamma \\
&\leq \hat{\mathcal{R}}(h^*) + \gamma \text{ (because } h = \arg\min_{h_i \in \mathcal{H}} \hat{\mathcal{R}}(h_i)) \\
&< (\mathcal{R}(h^*) + \gamma) + \gamma \\
&= \mathcal{R}(h^*) + 2\gamma \\
&= \mathcal{R}(h^*) + 2\sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}}
\end{aligned}$$

When $|\mathcal{H}|$ is finite

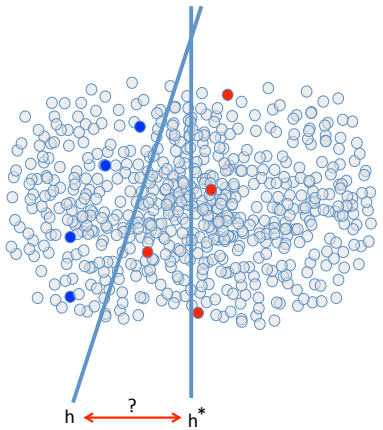
Condition of validity of the ERM principle

$$\forall \mathcal{D}_{\mathcal{Z}}, \forall \gamma \geq 0, \forall \delta \leq 1, P(|\mathcal{R}(h) - \mathcal{R}(h^*)| \leq \gamma) \geq 1 - \delta$$

Theorem

Let m, δ be fixed. With probability at least $1 - \delta$,

$$\mathcal{R}(h) \leq \mathcal{R}(h^*) + 2\sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}}$$

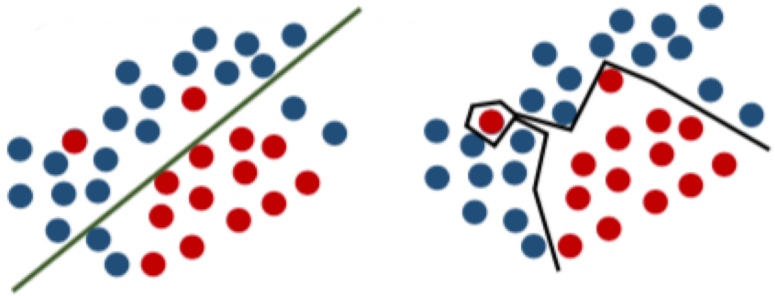


Theorem

Let m, δ be fixed. With probability at least $1-\delta$,

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}}$$

$|\mathcal{H}|$ plays the role of a complexity measure which is related to the risk of overfitting



$|\mathcal{H}|$ small

$|\mathcal{H}|$ large

$\mathcal{H}$: {

$\mathcal{H}$: {

When $|\mathcal{H}|$ is finite

Theorem

Let m, δ be fixed. With probability at least $1-\delta$,

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}}$$

$$\mathcal{R}(h) \leq \mathcal{R}(h^*) + 2\sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}}$$

Infinite Case

What about if $|\mathcal{H}| = \infty$?

We need a complexity measure as a surrogate of $|\mathcal{H}|$.

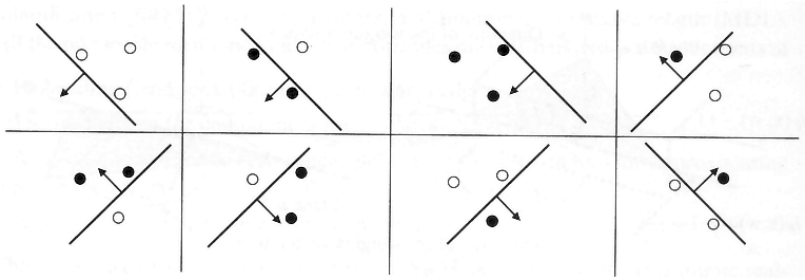
Scenario when $|\mathcal{H}|$ is infinite

When $|\mathcal{H}|$ is infinite

The VC dimension $d_{\mathcal{H}}$ (for Vapnik-Chervonenkis dimension) is a measure of the **capacity** (or **complexity**) of the class of hypotheses $\mathcal{H}$.

Definition

- The VC dimension $d_{\mathcal{H}}$ of a class of hypotheses $\mathcal{H}$ is defined as the cardinality of the largest set of points that a hypothesis $h \in \mathcal{H}$ can **shatter**.
- A set of points is **shattered** if for all assignments of labels to those points, there exists a hypothesis $h \in \mathcal{H}$ that makes no error. Said differently, S is shattered by $\mathcal{H}$ if $\mathcal{H}$ realizes **all possible dichotomies of** S (*i.e.* 2^m).



When $|\mathcal{H}|$ is infinite - Uniform convergence analysis

From the VC dimension $d_{\mathcal{H}}$, we can define the following upper bound:

Theorem

Let $\mathcal{H}$ be a class of hypotheses, $\forall h \in \mathcal{H}, \forall \delta \geq 0, \forall m > 0$, the following bound holds:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{d_{\mathcal{H}}(ln\frac{2m}{d_{\mathcal{H}}} + 1) + ln\frac{4}{\delta}}{m}}$$

Using the same math as before, we also get:

Theorem

Let $\mathcal{H}$ be a class of hypotheses, $\forall h \in \mathcal{H}, \forall \delta \geq 0, \forall m > 0$, the following bound also holds:

$$\mathcal{R}(h) \leq \mathcal{R}(h^*) + 2\sqrt{\frac{d_{\mathcal{H}}(ln\frac{2m}{d_{\mathcal{H}}} + 1) + ln\frac{4}{\delta}}{m}}$$

When $|\mathcal{H}|$ is infinite - Uniform convergence analysis

Corollary

To guarantee that $\mathcal{R}(h) \leq \mathcal{R}(h^*) + 2\gamma$ it suffices that m is on the order of the VC-dim, i.e.:

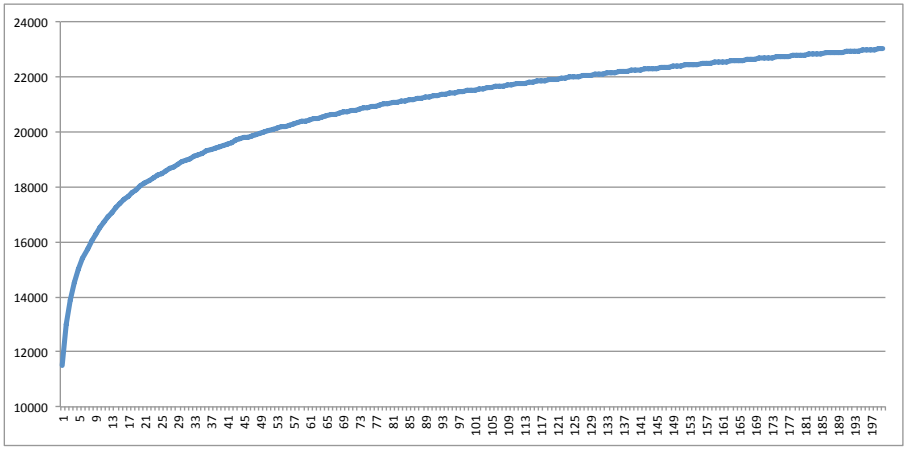
$$m = \mathcal{O}_{\delta,\gamma}(d_{\mathcal{H}}),$$

where we treat γ and δ as constant.

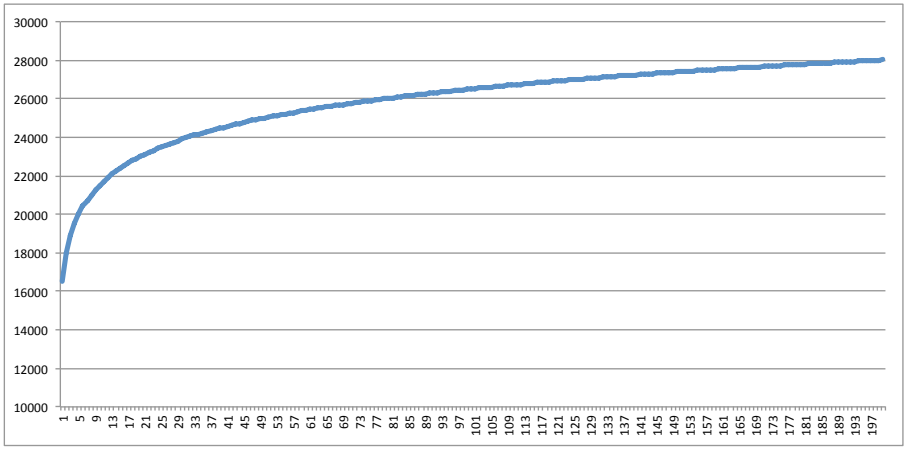
Intuition behind the Corollary

The number of training examples you need is roughly linear in the VC-dimension of the hypothesis class. For most reasonable hypothesis classes, it turns out that the VC-dimension is very similar to the number of parameters of your model. (e.g. Linear classifier in d dimensions $\rightarrow$ $d_{\mathcal{H}} = d + 1$).

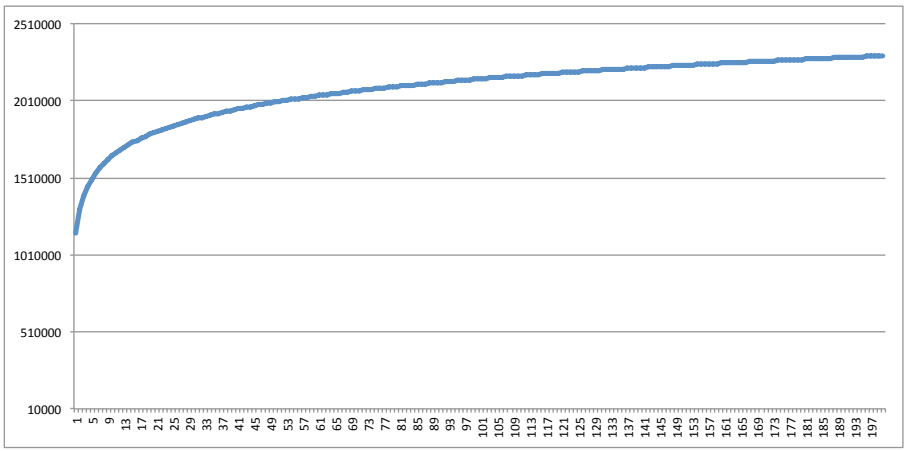
Note that if γ and δ are (too) small... we get pessimistic bounds.



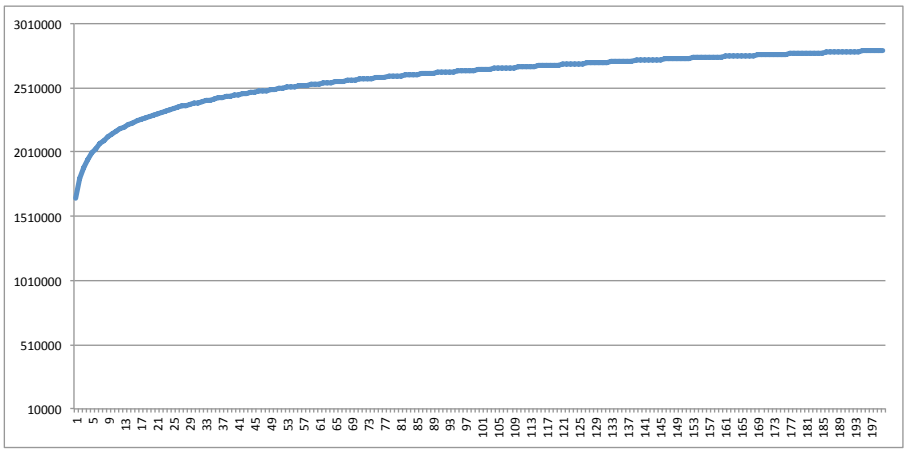
δ=0.01 γ=0.01



δ=0.001 γ=0.01



δ=0.01 γ=0.001



δ=0.001 γ=0.001

Uniform Convergence

We have proven the following two bounds:

- Finite case:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{1}{2m} ln \frac{2|\mathcal{H}|}{\delta}}$$

- Infinite case:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{d_{\mathcal{H}}(ln \frac{2m}{d_{\mathcal{H}}} + 1) + ln \frac{4}{\delta}}{m}}$$

Both cases can be formalized as follows:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{\text{“Complexity Measure”}}{m}}$$

Bias-Variance trade-off

Both cases can be formalized as follows:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{\text{“Complexity Measure”}}{m}}$$

TRUE RISK

Behavior at test time on unknow samples

BIAS

Capacity to learn the training samples

VARIANCE

Penalty term

Bias-Variance trade-off

Both cases can be formalized as follows:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{\text{“Complexity Measure”}}{m}}$$

TRUE RISK

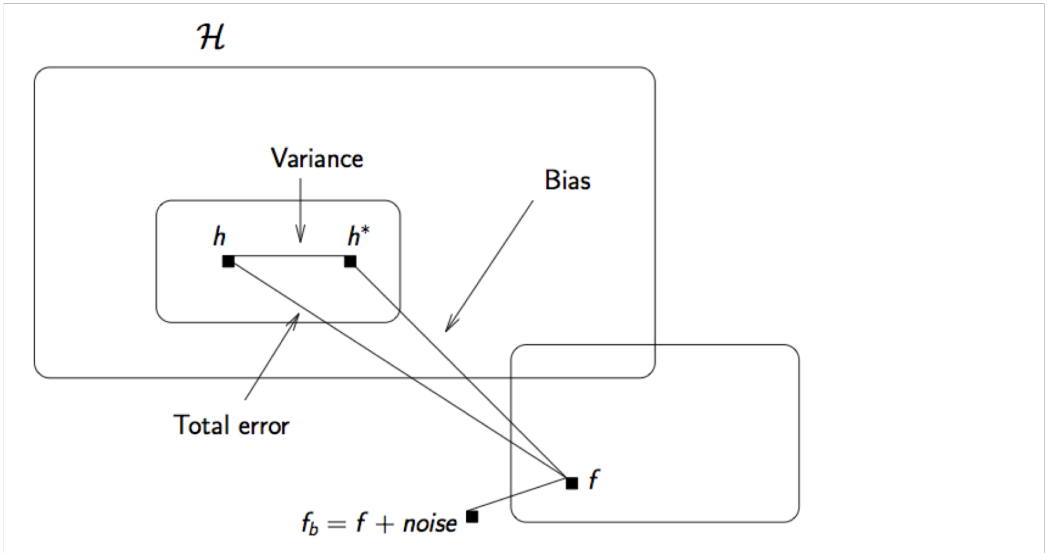
Behavior at test time on unknow samples

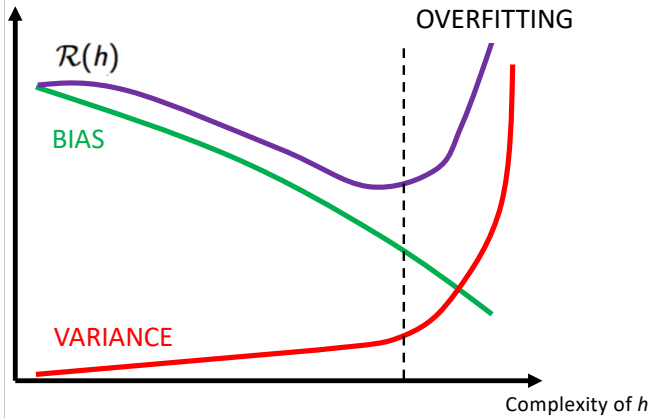
BIAS

Capacity to learn the training samples

VARIANCE

Penalty term





Some remarks about the VC theory

1. It gives **pessimistic bounds** which have to hold $\forall h \in \mathcal{H}$.
2. The only property that matters is the **"size" of the hypothesis space** and not on **how the algorithm searches the space**.
3. Therefore, the VC theory is meaningful when the learning algorithm performs minimization of $\hat{\mathcal{R}}(h)$ in the **full hypothesis space**.
4. It is **useless for local algorithms**, like the k-NN which has an infinite $d_{\mathcal{H}}$.
5. It does not take into account the **training examples**.

Rademacher Complexity

Rademacher Complexity

There exist other measures of the complexity of $\mathcal{H}$ such as the **Rademacher complexity** (Bartlett & Mendelson, 2002).

Rademacher Complexity

Let $S = \{z_i = (\mathbf{x_i}, y_i)\}_{i=1}^m$ be the training set. The empirical Rademacher complexity of $\mathcal{H}$ is

$$Rad_m(\mathcal{H}, S) = \mathbb{E}\left(sup_{h \in \mathcal{H}} | \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(z_i) | \right)$$

where $\sigma_1, \sigma_2, ..., \sigma_m$ are m iid Rademacher random variables with $p(\sigma_i = 1) = p(\sigma_i = -1) = \frac{1}{2}$.

Intuitively, $Rad_m(\mathcal{H}, S)$ is large (i.e. σ_i and $h(z_i) = 1$ often agree) if we can find a classifier $h \in \mathcal{H}$ that "looks like" random noise, that is, is highly correlated with $\sigma_1, ..., \sigma_m$.

Generalization bound

Let $\mathcal{H}$ be a class of hypotheses, with probability $1-\delta$, $\forall m > 0$:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + 2Rad_m(\mathcal{H}, S) + \sqrt{\frac{4}{m} ln(\frac{2}{\delta})}$$

Other theoretical frameworks

However, the previous remarks about the VC-dim still hold for the Rademacher complexity-based bounds.
Two analytical frameworks take into account the algorithm L to derive generalization bounds: **Uniform stability** and **Algorithmic robustness**.
The goal is to bound:

$$P(|\mathcal{R}(L, h_S) - \hat{\mathcal{R}}(L, h_S)| \geq \gamma)$$

which differs from what we studied so far:

$$P(sup_{h \in \mathcal{H}} |\mathcal{R}(h) - \hat{\mathcal{R}}(h)| \geq \gamma)$$

Uniform stability

Uniform stability

Variance versus Stability

- Statistical learning theory prompts us to **reduce the variance without altering the bias**.
- Having a **low variance** is equivalent to having **high stability**.
- How to **relate the generalization error $\mathcal{R}_h$ to the stability** of an algorithm L which induces h?

Intuitively, an algorithm L is said **stable** if it is robust to small changes in the training sample, i.e., the variation in its output h is small.

Uniform stability

Given a training set S of size m, we build $\forall i = 1, ..., m$:

- $S^{\setminus i} = \{z_1, ..., z_{i-1}, z_{i+1}, ..., z_m\}$ by removing the i-th element of S.
- $S^i = \{z_1, ..., z_{i-1}, z_i', z_{i+1}, ..., z_m\}$ by replacing the i-th element by z_i' drawn i.i.d. from $\mathcal{D}_{\mathcal{Z}}$.

Definition (Uniform stability [Bousquet and Elisseef 2002])

An algorithm L has uniform stability $\frac{\beta}{m}$ with respect to a loss function ℓ if the following holds:

$$\forall S, \forall i \in \{1, ..., m\}, sup_z |\ell(h_S, z) - \ell(h_{S^{\setminus i}}, z)| \leq \frac{\beta}{m},$$

where β is a positive constant, h_S and $h_{S^{\setminus i}}$ are the hypothesis learned by L from S and $S^{\setminus i}$ respectively. By the triangle inequality, we get:

$$\forall S, \forall i \in \{1, ..., m\}, sup_z |\ell(h_S, z) - \ell(h_{S^i}, z)| \leq 2\frac{\beta}{m}$$

Uniform stability

Generalization bound using uniform stability

Let S be a training sample of size m and $\delta > 0$. For any algorithm L with uniform stability $\frac{\beta}{m}$ with respect to a loss function ℓ bounded by M, with probability $1 - \delta$, we have:

$$\mathcal{R}_{h_S} \leq \hat{\mathcal{R}}_{h_S} + \frac{2\beta}{m} + (4\beta + M)\sqrt{\frac{ln\frac{1}{\delta}}{2m}}$$

Theorem [Kearns and Ron 1999]

An algorithm L having an hypothesis space of finite VC-dimension is stable in the sense that its stability is bounded by its VC-dimension.

Corollary

Using the stability as a complexity measure does not give worse bounds than using the VC-dimension.

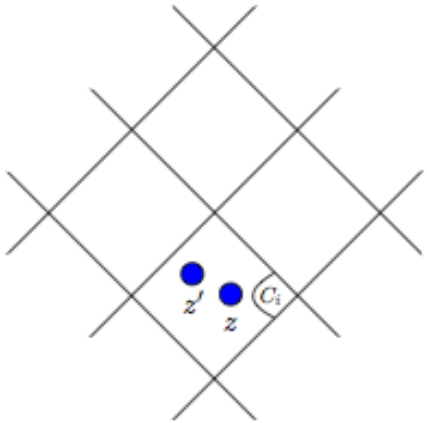
Algorithmic robustness

Algorithmic robustness

Definition (Algorithmic robustness)

Algorithm L is $(K, \epsilon(\cdot))$-robust, for $K \in \mathbb{N}$ and $\epsilon(\cdot) : \mathcal{Z}^n \to \mathbb{R}$, if $\mathcal{Z}$ can be partitioned into K disjoint sets, denoted by $\{C_i\}_{i=1}^K$, such that the following holds:

$$\forall z \in S, \forall z' \in \mathcal{Z}, \forall i \in [K] : \text{if } z, z' \in C_i, \text{then } |\ell(h_\mathcal{T}, z) - \ell(h_\mathcal{T}, z')| \leq \epsilon(S),$$



Algorithmic robustness

Theorem (Robustness bound)

If an algorithm L is $(K, \epsilon(\cdot))$-robust, then with probability $1-\delta$, we have:

$$\mathcal{R}_{h_S} \leq \hat{\mathcal{R}}_{h_S} + \epsilon(S) + \sqrt{\frac{2K \ln 2 + 2 \ln(1/\delta)}{m}},$$

where h_T is the hypothesis learned by $\mathcal{A}$ from $\mathcal{T}$.

Generalization Bounds and Bias-Variance trade-off

Uniform Convergence
$d_{\mathcal{H}}$: VC-dimension

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + \sqrt{\frac{d_{\mathcal{H}}(ln\frac{2m}{d_{\mathcal{H}}} + 1) + ln\frac{4}{\delta}}{m}}$$

Rademacher complexity
Rad_m: Rademacher complexity

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + 2Rad_m(\mathcal{H}, S) + \sqrt{\frac{4}{m} ln(\frac{2}{\delta})}$$

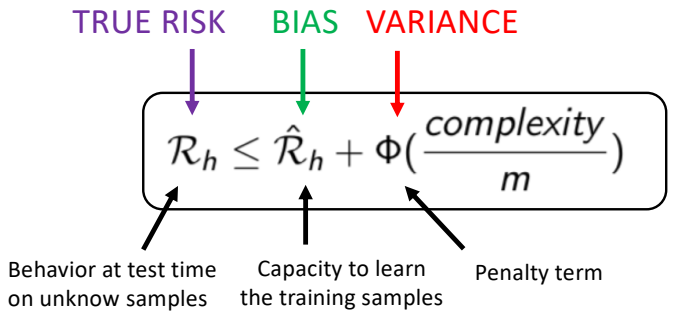
Uniform stability
β: stability constant

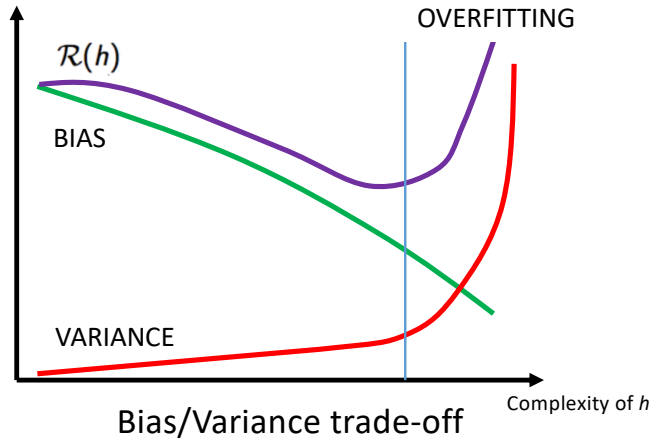
$$\mathcal{R}_{h_S} \leq \hat{\mathcal{R}}_{h_S} + \frac{2\beta}{m} + (4\beta + M)\sqrt{\frac{ln\frac{1}{\delta}}{2m}}$$

Algorithmic robustness

$$\mathcal{R}_{h_S} \leq \hat{\mathcal{R}}_{h_S} + \epsilon(S) + \sqrt{\frac{2K \ln 2 + 2\ln(1/\delta)}{m}}$$

All these bounds share a common shape:





Bias/Variance trade-off

48

Estimation of the generalization risk $\mathcal{R}_h$

Estimation of the generalization risk $\mathcal{R}_h$

$$\mathcal{R}_h \leq \hat{\mathcal{R}}_h + \Phi(\frac{complexity}{m})$$

How to assess $\mathcal{R}_h$ in practice?

Generalization bounds cannot be used in practice because they are often too pessimistic.

Use the empirical $\hat{\mathcal{R}}_h$ as an estimate of $\mathcal{R}_h$ is not a good idea because the training error is likely to be lower than the actual generalization error.

Cross Validation

Cross-Validation is a technique for assessing how a model learned by a ML algorithm from the training set S will generalize at test time.

Estimation of the generalization risk $\mathcal{R}_h$

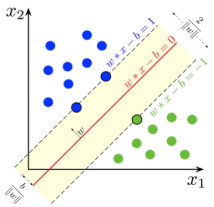
Hold-out k cross-validation algorithm

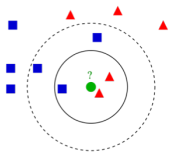
Input: A learning algorithm L and a learning set S
Output: An estimate $\hat{\mathcal{R}}'(h)$
Split S randomly in k subsets $S_1, ..., S_k$;
for *i=1 to k* **do**
  Run L on $S - S_i$ and induce classifier h_i;
end
Deduce the estimate $\hat{\mathcal{R}}'(h)$ of the true risk s.t. $\hat{\mathcal{R}}'(h) = \frac{1}{k} \sum_{i=1}^{k} \hat{\mathcal{R}}'(h_i)$ where $\hat{\mathcal{R}}'(h_i)$ is error of h_i on S_i;

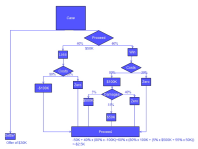
Model Selection and Hyperparameter tuning

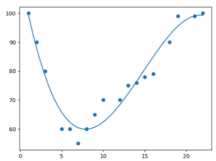
Model Selection and Hyperparameter tuning

Popular Supervised Learning Methods



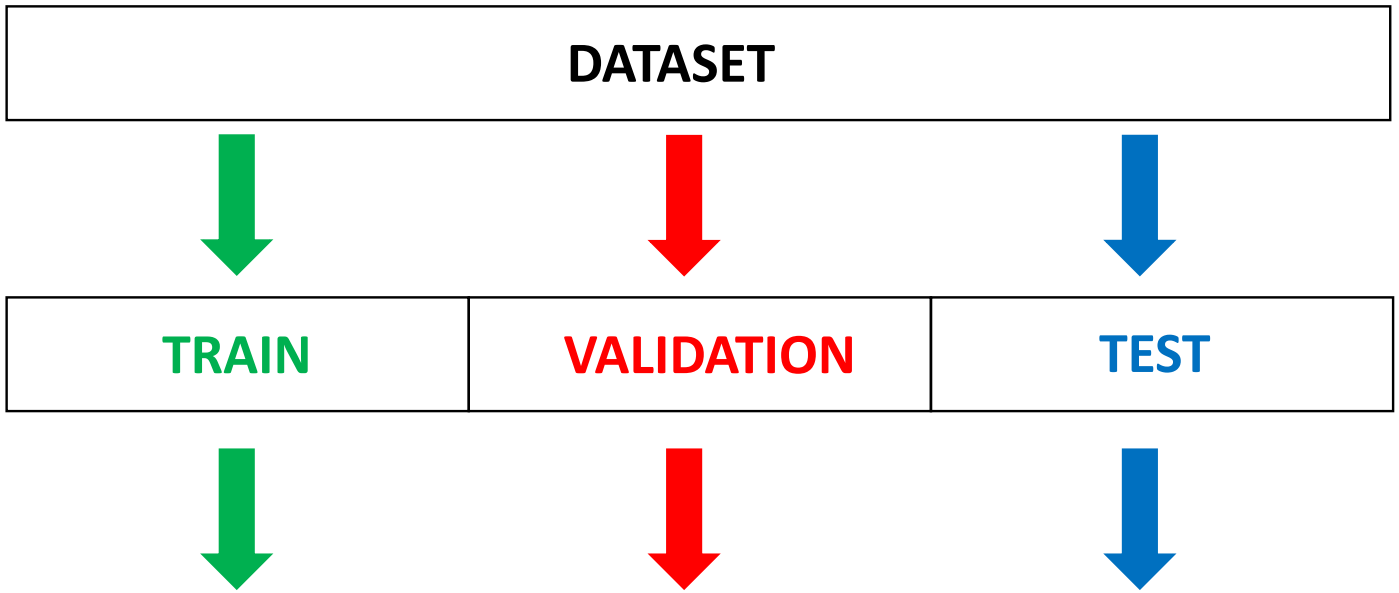






ML method	Learned model ($\boldsymbol{\theta}$ parameters)	Hyperparameters ($\boldsymbol{\lambda}, \ldots$)
Deep Learning	Weights of the network	# of hidden layers, # of epochs, batch size, learning rate, ...
Support Vector Machine	Coefficients of the hyperplane	kernel, regularization parameter C
k-Nearest Neighbors	None (« lazy » algorithm)	k (# of neighbors), metric
Decision tree	Decision rules	split measure, depth, min # of samples in a leaf, etc.
Regression, SVR	Coefficients of the polynomial	degree of the polynomial, kernel

Train, Validation and Test Sets



Train multiple models
Learn the parameters θ with:
- Different ML algorithms
- Different hyperparameters

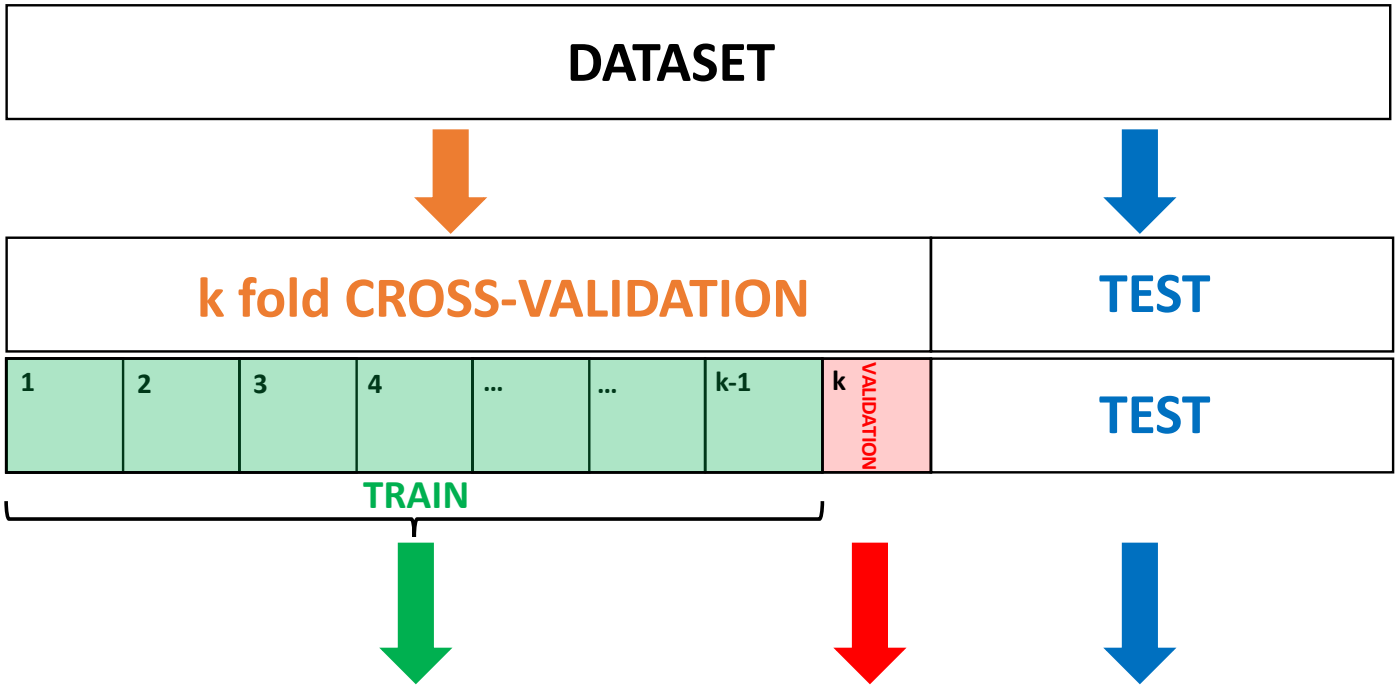
Validate the models
- Tune hyperparameters λ, ...
- Select best model

Evaluate the model
- Compute the error on the TEST

41

Model Selection and Hyperparameter tuning

k fold – Cross Validation



Train multiple models
Learn the parameters θ with:
- Different ML algorithms
- Different hyperparameters

Validate the models
- Tune hyperparameters λ, ...
- Select best model

Evaluate the model
- Compute the error on the TEST

42

How to address the Bias-Variance trade-off by analyzing the training and validation/test errors?

TRUE RISK BIAS VARIANCE

$$\mathcal{R}_h \leq \hat{\mathcal{R}}_h + \Phi\left(\frac{complexity}{m}\right)$$

Consider a cat classification task. An "ideal" classifier (such as a human) might achieve nearly perfect performance in this task (i.e. the Bayesian error $\epsilon_B \approx 0$).

Scenario 1

Suppose your algorithm performs as follows:

- Training error = 1%
- Validation error = 11%

How to address the Bias-Variance trade-off by analyzing the training and validation/test errors?

TRUE RISK BIAS VARIANCE

$$\mathcal{R}_h \leq \hat{\mathcal{R}}_h + \Phi\left(\frac{complexity}{m}\right)$$

Consider a cat classification task. An "ideal" classifier (such as a human) might achieve nearly perfect performance in this task (i.e. the Bayesian error $\epsilon_B \approx 0$).

Scenario 1

Suppose your algorithm performs as follows:

- Training error $= 1\%$
- Validation error $= 11\%$

What problem does it have? We estimate the **bias** as 1%, and the **variance** as 10% $(= 11\% - 1\%)$. Thus, it has **high variance** and small **bias**. The classifier has very low training error, but it is failing to generalize to the validation set. This is **overfitting**. This can be fixed by training on a massive training set.

How to address the Bias-Variance trade-off by analyzing the training and validation/test errors?

TRUE RISK BIAS VARIANCE

$$\mathcal{R}_h \leq \hat{\mathcal{R}}_h + \Phi\left(\frac{complexity}{m}\right)$$

Scenario 2

Now consider this:

- Training error = 15%
- Validation error = 16%

How to address the Bias-Variance trade-off by analyzing the training and validation/test errors?

TRUE RISK BIAS VARIANCE

$$\mathcal{R}_h \leq \hat{\mathcal{R}}_h + \Phi(\frac{complexity}{m})$$

Scenario 2

Now consider this:

- Training error = 15%
- Validation error = 16%

We estimate the **bias** as 15%, and **variance** as 1%. This classifier is fitting the training set poorly, but its error on the validation set is barely higher than the training error. This classifier therefore has **high bias**, but **low variance**. This is **underfitting**. This can be fixed by increasing the expressiveness of h.

Harder problem

Suppose now that you are building a speech recognition system, and find that 14% of the audio clips have so much background noise or are so unintelligible that even a human cannot recognize what was said. In this case, even the most "optimal" speech recognition system might have error around 14% (i.e. $\epsilon_B = 0.14$ - note that this error can be estimated from a set of experts).

Suppose that on this speech recognition problem, your classifier h achieves:

- Training error $= 15\%$
- Validation error $= 30\%$

We can conclude that:

1. There is not much room for improvement in terms of bias or in terms of training set performance. The bias can be rewritten as follows:

$$\textit{Bias} = \text{ Unavoidable bias } + \text{ Avoidable bias } = \epsilon_B + 1\%$$

2. There is ample room for improvement in the errors due to variance. This can be done by learning on a massive training set (all variance is "avoidable").

Multiple Choice Questions **MCQ2**