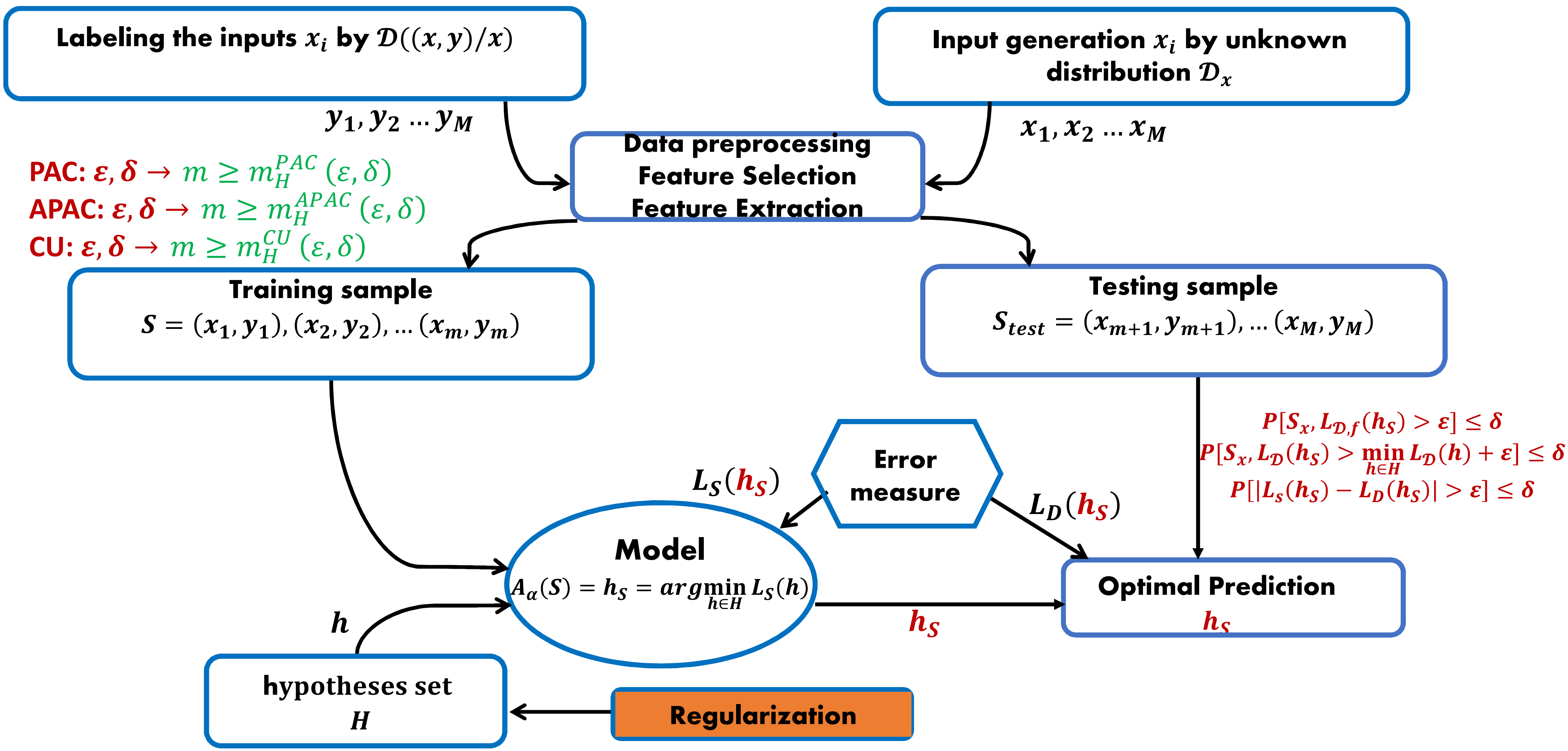
Part 3: Overfitting Underfitting

1. Validation/Cross-Validation
2. Regularization:
   1. Minimisation of regularized cost
   2. Algorithm stability
   3. Tikhonov Regularizer
   4. Stability-Adaption tradeoff

Supervised Learning Passive Offline Algorithm (SLPOA)



Recall

Definition: APAC learning model

H follows agnostic PAC learning, if there exist $m_{\mathcal{H}}:\ (0,1)^2\ \rightarrow \mathbb{N}$ and A_α.

Having the following property: $\forall \varepsilon, \delta \in (0,1), \forall\, \mathcal{D}$ on $X \times Y$.

Then, if we run A_α on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ generated $(i.i.d.)$ such that S is selected with a probability at least $(1-\delta)$, A_α will generate the hypothesis h_S such that:

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in H} L_{\mathcal{D}}(h) + \varepsilon.$$

In other words:

$$P_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}(h_S) > \min_{h \in H} L_{\mathcal{D}}(h) + \varepsilon\right] \leq \delta\ \ for\ all\ m \geq m_H(\varepsilon, \delta)$$

Recall

Definition: Uniform Convergence

We say that H has the uniform convergence property with respect to (Z, l), if there exist:

- a function $m_H^{CU}(\varepsilon, \delta)$: $[0,1]^2 \longrightarrow \mathbb{N}$, such that: $\forall(\varepsilon, \delta) \in [0,1]^2 and\ \forall\ \mathcal{D}$ over Z.
- S is a sample of size $m \geq m_H^{CU}(\varepsilon, \delta)$, whose points are drawn $(i.i.d.)$ by $\mathcal{D}$, such that with probability of at least $(1-\delta)$, S is ε-representative:

$$P[|L_s(h) - L_D(h)| \leq \varepsilon] \geq 1 - \delta \Leftrightarrow P[|L_s(h) - L_D(h)| > \varepsilon] \leq \delta$$

Recall

- If $|H| \approx \infty \boldsymbol{|L_s(h_S) - L_D(h_S)| \leq g}(d_{CV}) \boldsymbol{= \varepsilon \in V(0)}$
$$d_{CV} < \infty \; or \; N_C < \infty \Leftrightarrow CU \;\Leftrightarrow APAC \Leftrightarrow PAC$$
- If $|H| < \infty$
  - $If\ \ target\ function\ exist\ then\ PAC$
  - $If\ Target\ function\ is\ stochastic\ then\ CU\ \Longrightarrow APAC$

Example: Regression

$$x_i = \left(x_i^1, x_i^2\right)$$

$$f(x_i) = w_0 x_i^1 + w_1 x_i^2 + w_2 x_i^1 x_i^2 + w_3 \left(x_i^1\right)^2 + w_4 \left(x_i^2\right)^2 = y_i$$

If $L_D(h) \gg L_S(h)$ then we reduce the degree of polynomial that is to reduce the size of $w = (w_0, w_1, w_2, w_3, w_4)$

Motivation

If $L_D(h) \gg L_S(h)$ we have the Overfitting Problem .

To remedy this problem, we should penalize the model parameters.

Objective:

- How to penalize the learning.

Tool:

- Regularization.

1. Minimization of regularized cost

$$A_\alpha(S) = h_S \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \{L_S(w)\}$$

If we have the Overfitting Problem , we penalize the model parameters as following .

$$\begin{cases} \boldsymbol{Min} & \boldsymbol{L_S(w)} \\ \boldsymbol{s.t} & \boldsymbol{\|w\| < C} \\ & \boldsymbol{w \in \mathbb{R}^d} \end{cases}$$

- $\Longrightarrow L(w, \lambda) = L_S(w) + \lambda(\|w\| - C) = L_S(w) + \mathrm{R(w)}$

$$\Rightarrow A_\alpha(S) = h_S \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \{L_S(w) + R(w)\}$$

- $d(w, 0) = \|w\|$
- $\|w\| = \|w\|_1 = \sum_{i=1}^{d} |w_i|$ with $w \in l_1$
- $\|w\| = \|w\|_2^2 = \sum_{i=1}^{d} {w_i}^2$ with $\mathrm{w} \in l_2$
- $L_S \in L_2$

1. Minimization of regularized cost

Definition : RLM algorithm

RLM is a learning algorithm used to minimize the sum of the empirical error and the regularization function $R: \mathbb{R}^d \longrightarrow \mathbb{R}$.

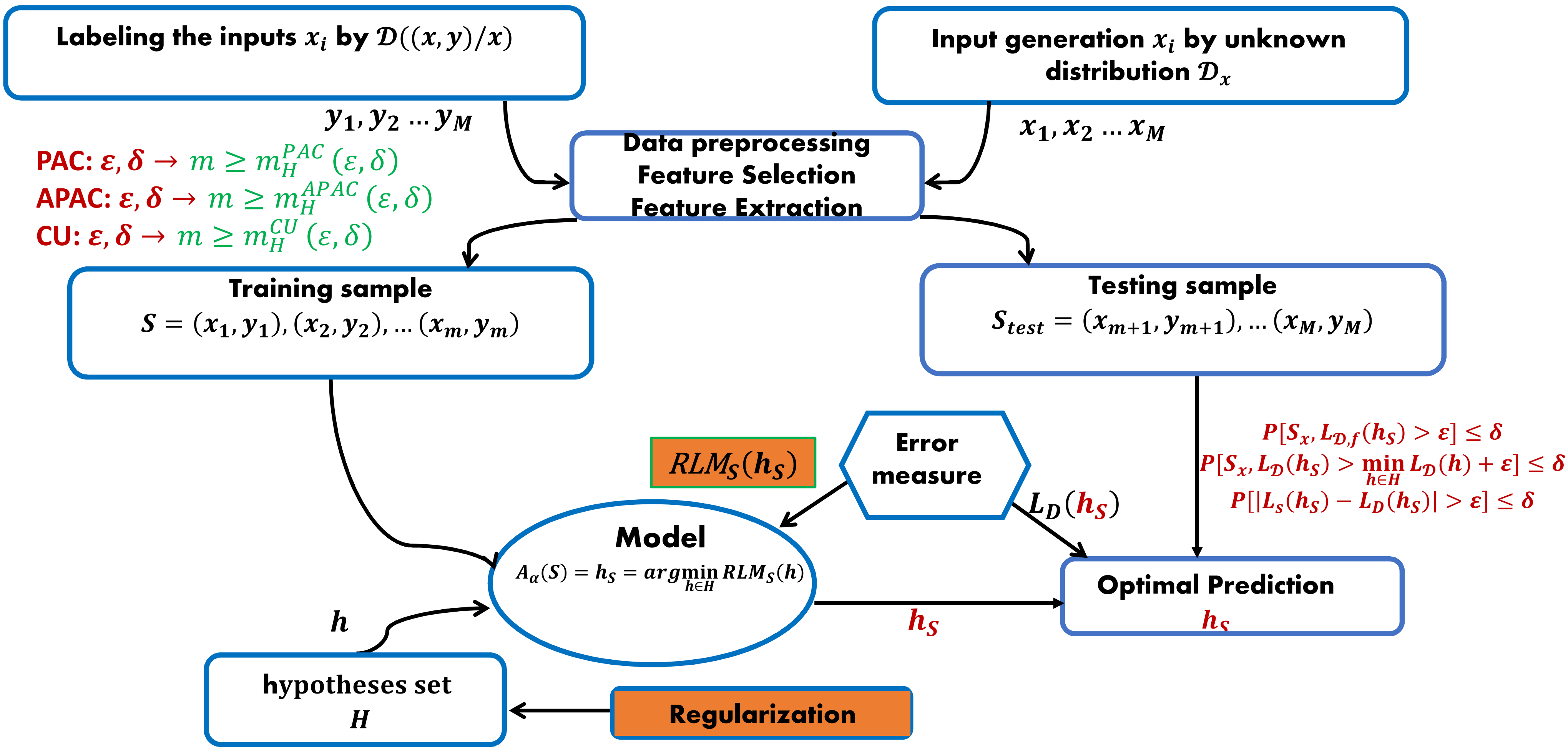
It generates the following hypothesis:

$$A_\alpha(S) = h_S \in \underset{w}{\operatorname{argmin}}\{L_S(w) + R(w)\}$$

Notice :

- The RLM algorithm shares the same similarities with SRM and MDL, such that the complexity of the hypotheses is measured by a regularization function $R(w)$.
- Similarly to MDL, there exist many types of regularization functions that depend on the type of problem to deal with.

Supervised Learning Passive Offline Algorithm (SLPOA)



1. Minimization of regularized cost

Definition : Tikhonov regularizer

Tikhonov regularization function has the following from :

$$R(w) = \lambda\|w\|^2$$

With :

$\lambda > 0$ is a scalar. And $\|.\|$ is the l_2 norm:

$$\|w\| = \sqrt{\sum_{i=1}^{d} {w_i}^2}$$

The learning rule becomes :

$$A_\alpha(S) = h_S = \underset{w}{\operatorname{argmin}}(L_S(w) + \lambda\|w\|^2)$$

1. Minimization of regularized cost – Ridge Regression

Consider a problem of linear regression and the training data $(x_1, y_1), \ldots, (x_m, y_m)$.

We want to minimize: L_2 norm

$$L_S(w) = \frac{1}{m}\sum_{i=1}^{m}(w^T x_i - y_i)^2 = \frac{1}{m}(Xw - y)^T(Xw - y)$$

In that case the solution is: $w_{lin} = (X^T X)^{-1} X^T y$

Hard constraint:

$$w_i = 0 \text{ pour } i > 0.$$

Soft constraint:

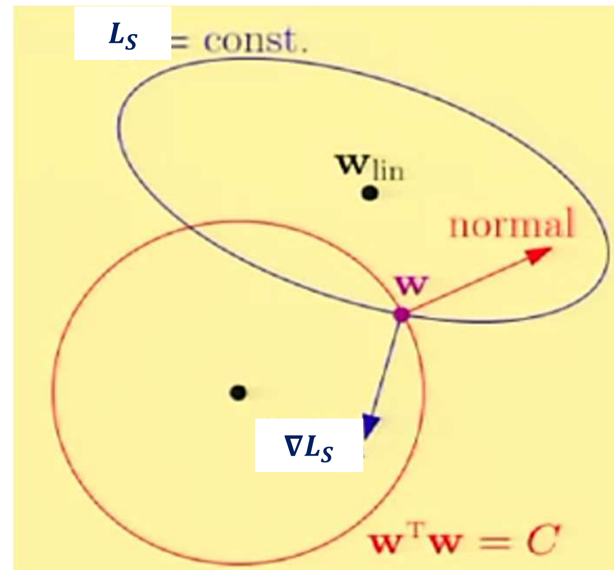
$$\sum_{i=0}^{d} w_i^2 \leq C$$

Instead of eliminating weights, we are going to minimize their values.

1. Minimization of regularized cost – Ridge Regression

We can use the multiples of Lagrange to solve that problem:

$$\begin{cases} Min \quad L_S(w) = \dfrac{1}{m}(Xw - y)^T(Xw - y) \\ w^T w \leq C \end{cases}$$



the solution is w_{reg} instead of w_{lin}.

So:

$$\nabla L_S(w_{reg}) \propto -w_{reg}$$
$$= -2\lambda w_{reg}$$
$$\nabla L_S(w_{reg}) + 2\lambda w_{reg} = 0$$

The minimisation problem becomes:

$$Min \quad L_S(w) + \lambda w^T w$$

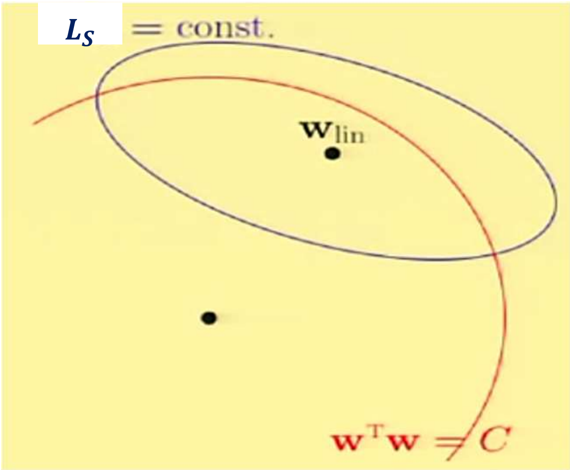
$$Min\, L_S(w) + \lambda w^T w = \frac{1}{m}(Xw - y)^T(Xw - y) + \lambda w^T w$$

1. Minimization of regularized cost – Ridge Regression

$$\nabla\left(\frac{1}{m}(Xw-y)^T(Xw-y)+\lambda w^T w\right)=0$$

$$X^T(Xw-y)+\lambda w=0$$

$$w_{reg}=(X^TX+\lambda I)^{-1}X^Ty$$



Hence, the unconditional minimization problem.

If C ↑ so λ ↓ : non-severe regularization.

If C ↓ so λ ↑ : severe regularization.

Definition : Ridge regression

Ridge regression is a combination between linear regression (having the squared cost) and Tikhonov regularization. The learning rule becomes:

$$A_\alpha(S)=\underset{w\in\mathbb{R}^d}{\operatorname{argmin}}\left\{\lambda\|w\|_2^2+\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(\langle w,x_i\rangle-y_i)^2\right\}$$

2. Model's Stability

Definition :

A Model is said to be stable if a small changing in its inputs implies a small changing in its outputs.

Notice:

Let A_α be a Learning Model, let $S = \{z_1, \dots, z_m\}$ be a training set of size m, let $A_\alpha(S)$ be the output of the Model A_α.

Consider another example z'. Let $S^{(i)}$ be the training set obtained by replacing the example z_i in S by z' :

$$S^{(i)} = \{z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_m\}$$

Replacing z_i by z' defines the small changing in its inputs, this means we train $A_\alpha(S^{(i)})$ instead of $A_\alpha(S)$.

2. Model's Stability

Theorem :

Soit la distribution D, soit $S = \{z_1, \dots, z_m\}$ une séquence $i.i.d$ d'exemples, soit z' un autre exemple $i.i.d$. Soit $U(m)$ une distribution uniforme sur $[m]$. Donc $\forall A_\alpha$:

$$E_{S \sim D^m}\left[L_D\left(A_\alpha(S)\right) - L_S\left(A_\alpha(S)\right)\right] = E_{(S,z') \sim D^{m+1}, i \sim U(m)}\left[\, l\left(A_\alpha\left(S^{(i)}\right), z_i\right) - l(A_\alpha(S), z_i)\right]$$

Notice :

- If the différence $l\left(A_\alpha\left(S^{(i)}\right), z_i\right) - l(A_\alpha(S), z_i)$ is large, we say the Model overfits the data.
- If the différence $l\left(A_\alpha\left(S^{(i)}\right), z_i\right) - l(A_\alpha(S), z_i)$ is small, we say the Model is stable.

2. Model's Stability

Definition : Model's stability

Let $\varepsilon: \mathbb{N} \longrightarrow \mathbb{R}$ be a monotonic decreasing function, we say that the Model A_α is stable (On-Average-Replace-One-Stable) with two rates $\varepsilon(m)$, if $\forall D$:

$$E_{(S,z')\sim D^{m+1}, i\sim U(m)}\left[\, l\left(A_\alpha\left(S^{(i)}\right), z_i\right) - l(A_\alpha(S), z_i)\right] \leq \varepsilon(m)$$

Notice :

- With this definition, the Model A_α doesn't suffer from overfitting if and only if it is sable.
- To have a good Model, it should not overfit the data, moreover its empirical error should be small :

$$L_D(A_\alpha(\mathrm{S})) \approx L_S(A_\alpha(\mathrm{S})) \text{ and } L_S(A_\alpha(\mathrm{S})) \approx 0$$

3. Tikhonov Regularizer – Lipschitz Cost Function

Definition : ρ-Lipschitz function

Let $C \subset \mathbb{R}^d$, we say that the function $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is ρ-Lipschitz on C, if $\forall w_1, w_2 \in C$, we have :

$$\|f(w_1) - f(w_2)\| \leq \rho \|w_1 - w_2\|$$

Corollary :

Let's suppose that the cost function is convex and ρ-Lipschitz. So, RLM algorithm, having the Tikhonov regularizer $\lambda\|w\|^2$, is stable with the rate $\varepsilon(m) = \frac{2\rho^2}{\lambda m}$. So :

$$\boldsymbol{E_{S \sim D^m}\left[L_D\left(A_\alpha(S)\right) - L_S\left(A_\alpha(S)\right)\right] \leq \frac{2\rho^2}{\lambda m}}$$

3. Tikhonov Regularizer – Smooth Cost Function

Definition : β-Smooth function

We say that the function $f : \mathbb{R}^d \to \mathbb{R}$ is β- Smooth if its gradient is β-Lipschitz.

That means $\forall v, w$:

$$\|\nabla f(v) - \nabla f(w)\| \leq \beta \|v - w\|$$

- This implies that, $\forall v, w$ we have :

$$f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2$$

- If f is β-Smooth and convex, we have $\forall v, w$:

$$f(w) + \langle \nabla f(w), v - w \rangle \leq f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2$$

- If $\forall v$ we have $f(v) \geq 0$, and if f is β-Smooth, we say that f is a self-bounded function :

$$\|\nabla f(w)\|^2 \leq 2\beta f(w)$$

3. Tikhonov Regularizer – Smooth Cost Function

Corollary :

Let's suppose that the cost function is convex, β-Smooth and non-negative. So, RLM algorithm, having the Tikhonov regularizer $\lambda\|w\|^2$ such that $\lambda \geq \frac{2\beta}{m}$, is stable with rate :

$$\varepsilon(m) = \frac{48\beta}{\lambda m} E\left[L_S\left(A_\alpha(S)\right)\right]$$

So :

$$E_{S\sim D^m}\left[L_D\left(A_\alpha(S)\right) - L_S\left(A_\alpha(S)\right)\right] \leq \frac{48\beta}{\lambda m} E\left[L_S\left(A_\alpha(S)\right)\right]$$

Notice :

- For the two types of the cost function,

when $\lambda \longrightarrow \infty$, $E_{S\sim D^m}\left[L_D\left(A_\alpha(S)\right) - L_S\left(A_\alpha(S)\right)\right] \longrightarrow 0$

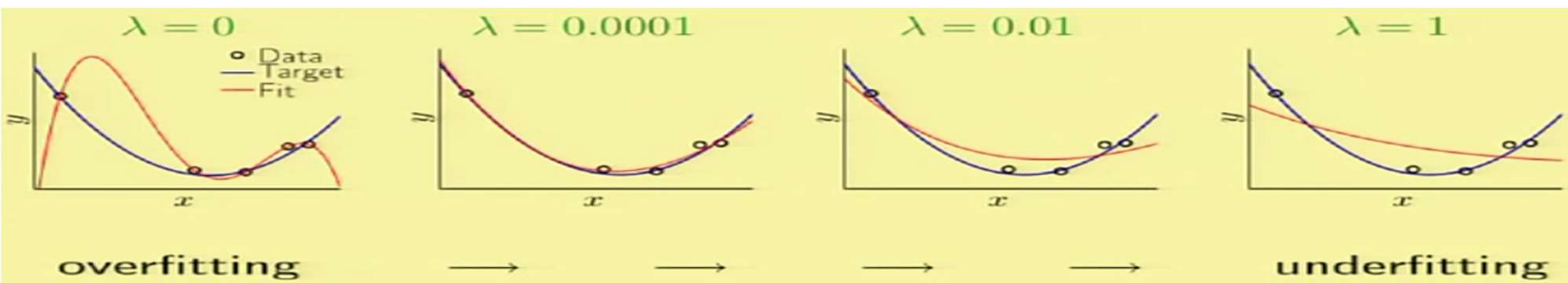
4. Stability-Adaption tradeoff

We can write the estimation of the generalization error as:

$$E_S[L_D(A_\alpha(S))] = E_S\left(L_S(A_\alpha(S))\right) + E_S[L_D(A_\alpha(S)) - L_S(A_\alpha(S))]$$

- The first term is the empirical error, it implies **the adaption** of the Model $A_\alpha$to training data $\boldsymbol{S}$.
- The second term is the difference between the general error and the empirical error, it implies the **stability** of the Model $A_\alpha$to small changings of inputs.

- If λ increases $\longrightarrow$ $L_S(A_\alpha(S))$ increases $\longrightarrow$ adaption decreases $\longrightarrow$ underfitting.
- If λ decreases $\longrightarrow L_D(A_\alpha(S)) - L_S(A_\alpha(S))$ increases $\longrightarrow$ stability decreases $\longrightarrow$ overfitting.



4. Stability-Adaption tradeoff - Lipschitz Cost Function

Corollary :

Let's suppose that the cost function is convex and ρ-Lipschitz. So, RLM algorithm, having the Tikhonov regularizer $\lambda\|w\|^2$, such that $\forall w^*$:

$$E_S\left[L_D\left(A_\alpha(S)\right)\right] \leq L_D(w^*) + \lambda\|w^*\|^2 + \frac{2\rho^2}{\lambda m}$$

Corollary : APAC learning for convex-Lipschitz bounded problems.

Let (H, Z, l) be a learning problem convex, Lipschitz and bounded having the parameters ρ and B.

For any training set of size m, let $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$.

So, RLM algorithm having the Tikhonov regularizer $\lambda\|w\|^2$ meets :

$$E_S\left[L_D\left(A_\alpha(S)\right)\right] \leq \min_{w \in H} L_D(w) + \rho B\sqrt{\frac{8}{m}}$$

In particular, $\forall \varepsilon > 0$, if $m \geq \frac{8\rho^2 B^2}{\varepsilon^2}$, so for any distribution D :

$$E_S\left[L_D\left(A_\alpha(S)\right)\right] \leq \min_{w \in H} L_D(w) + \varepsilon$$

4. Stability-Adaption tradeoff - Smooth Cost Function

Corollary :

Let's suppose that the cost function is convex and β-Lipschitz and non-negative. So, RLM algorithm, having the Tikhonov regularizer $\lambda\|w\|^2$ and for any $\lambda \geq \frac{2\beta}{m}$, such that $\forall w^*$:

$$E_S\left[L_D\left(A_\alpha(S)\right)\right] \leq (1 + \frac{48\beta}{\lambda m})(L_D(w^*) + \lambda\|w^*\|^2)$$

Corollary : APAC learning for convex-smooth bounded problems.

Let (H, Z, l) be a learning problem convex-Smooth and bounded having the parameters β and B. Suppose that $l(0, z) \leq 1$ for any $z \in Z$.

$\forall \varepsilon \in [0,1]$, let $m \geq \frac{150\beta B^2}{\varepsilon^2}$ and $\lambda = \varepsilon/3B^2$, so for any distribution D :

$$E_S\left[L_D\left(A_\alpha(S)\right)\right] \leq \min_{w \in H} L_D(w) + \varepsilon$$