Introduction à l'apprentissage

1) Apprentissage supervisé

Données $\mathcal{D} = \{t^{(i)}, x^{(i)}\}_{i=1}^{N}$

vecteurs caractéristiques $x^{(i)}$

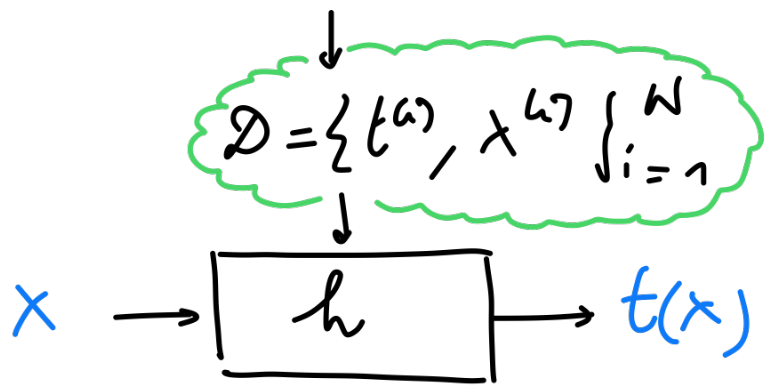
$\rightarrow x^{(i)} = [x_1^{(i)}, \dots, x_D^{(i)}] \in \mathbb{R}^D$

valeurs cibles $t^{(i)} \in \mathbb{R}$

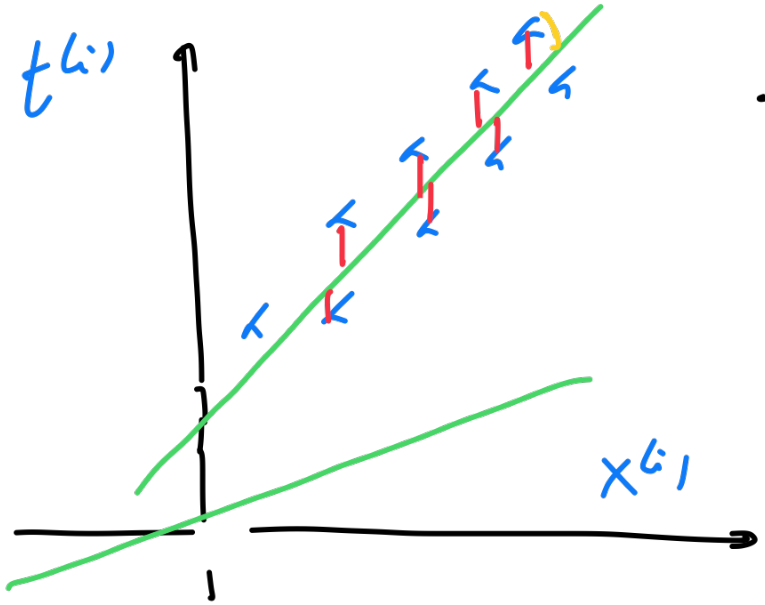
$\{t^{(i)}, x^{(i)}\}$ paire d'entraînement

$\mathcal{X}$: espace d'entrée

$\mathcal{Y}$: espace de sortie



Objectif : $h(x^{(i)}) \approx t^{(i)}$



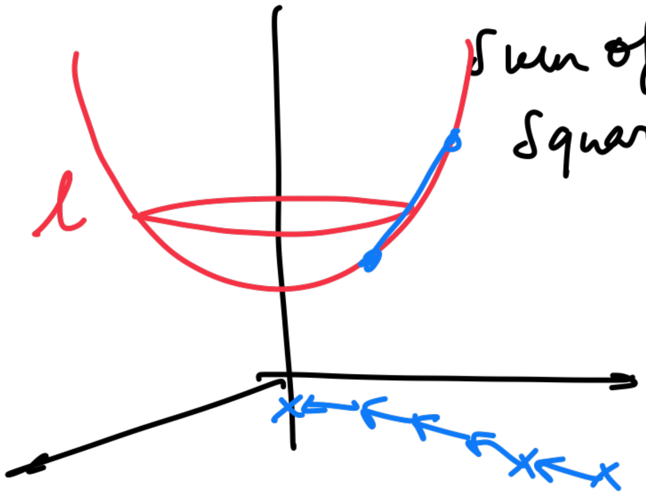
$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_D x_D$$

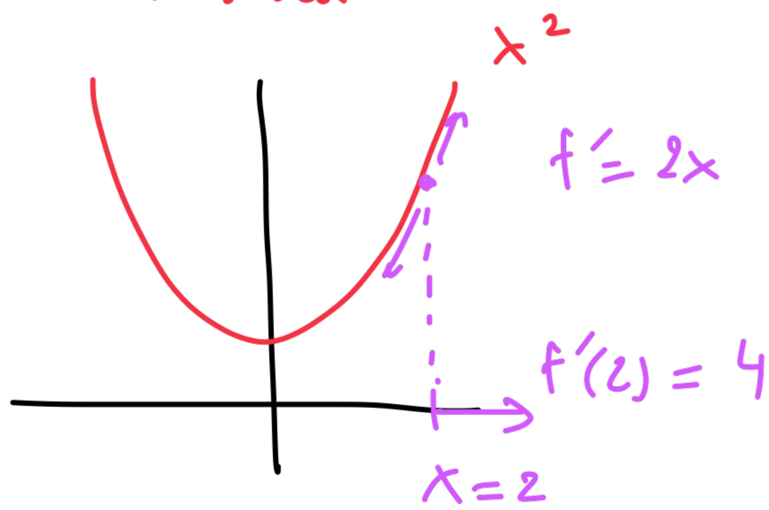
β_j : coefficients de régression

Fonction de coût

$$\ell(\beta) = \frac{1}{N}\sum_{i=1}^{N}\Big(\underbrace{t^{(i)} - \big(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\big)}_{\text{résidu}}\Big)^2$$

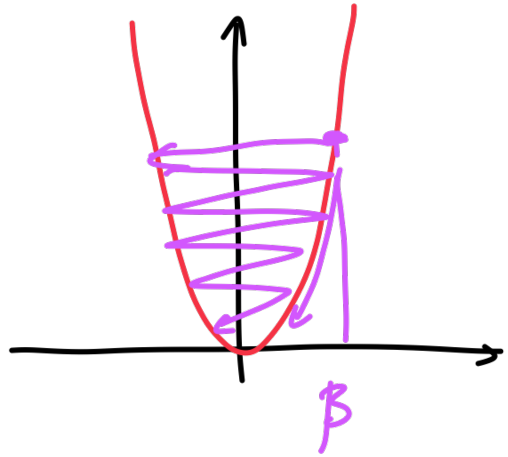
Residual Sum of Squares / OLS





Approche #1 → Calcul de dérivé et résolution $f'(x) = 0$

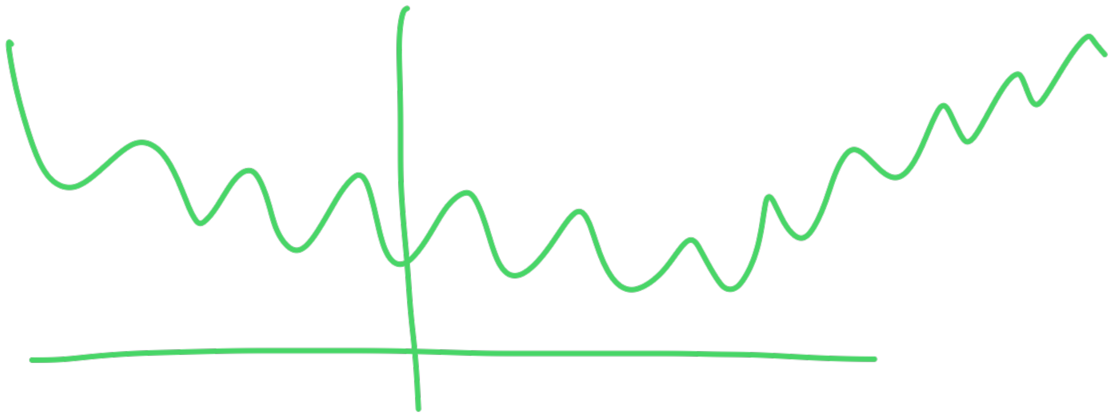
Approche #2 → Descente de gradient

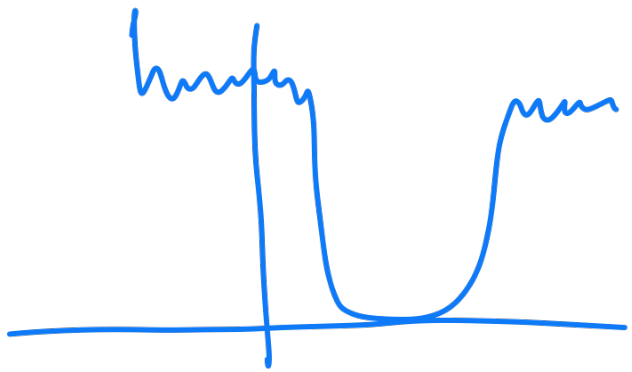


$$\beta^{(0)} = \beta_{init}$$

While iter < maxIter

$$\beta \leftarrow \beta - \eta \text{ grad } \ell(\beta)$$





$$f(\beta) = \beta^2 + 6\beta + \log(\beta) \rightarrow f'(\beta) = 2\beta + 6 + \frac{1}{\beta} = \frac{df}{d\beta}$$

$$f(\beta_0, \beta_1) = \beta_0^2 + \beta_0\beta_1 + \beta_1^2 \longrightarrow \text{grad } f = \left[\frac{\partial f}{\partial \beta_0}, \frac{\partial f}{\partial \beta_1} \right]$$

$$= \left[2\beta_0 + \beta_1, 2\beta_1 + \beta_0 \right]$$

$$\ell(\beta) = \frac{1}{N} \sum_{i=1}^{N} \Big(t^{(i)} - \underbrace{\big(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\big)}_{\langle \tilde{x}^{(i)}, \beta \rangle} \Big)^2$$

$$\frac{\partial \ell}{\partial \beta_0} = \frac{2}{N} \sum_{i=1}^{N} \Big(t^{(i)} - \big(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\big)\Big)\big(-1\big) \leftarrow$$

$$\frac{\partial \ell}{\partial \beta_j} = \frac{2}{N} \sum_{i=1}^{N} \Big(t^{(i)} - \big(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\big)\Big)\big(-x_j^{(i)}\big) \leftarrow$$

$$\tilde{x} = \underbrace{[1, \vec{x}]} \qquad h_\beta(x) = \beta^T \tilde{x} = [\beta_0 \ \beta_1 \cdots \beta_D] \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix}$$

$$\operatorname{grad} \ell(\beta) = \frac{2}{N} \sum_{i=1}^{N} \left(t^{(i)} - \langle \tilde{x}^{(i)}, \beta \rangle \right)\left(-\tilde{x}^{(i)}\right)$$

$$= \left[\frac{\partial \ell}{\partial \beta_0}, \frac{\partial \ell}{\partial \beta_1}, \dots, \frac{\partial \ell}{\partial \beta_D} \right]$$

Alternative #1 → Descente de gradient stochastique

shuffle $(\{t^{(i)}, x^{(i)}\}_{i=1})$

for $\{t^{(i)}, x^{(i)}\} \in D$

$$\beta \leftarrow \beta - \eta \frac{2}{N} \left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)} \right) \right) \left(-\tilde{x}^{(i)}\right)$$

Alternative #2 → MINIBATCH GD

Shuffle $(\{t^{(i)}, x^{(i)}\}_{i=1}^{N})$

for minibatch $\mathcal{B}$ in $\mathcal{D}$

$$\beta \leftarrow \beta + \frac{2\eta}{N} \sum_{i \in \mathcal{B}} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_D x_D^{(i)}) \right) (-\tilde{x}^{(i)})$$

Approche #2: Resolution des Equations normales

$\vec{\beta} = [\beta_0, \beta_1, \ldots, \beta_D]^T \in \mathbb{R}^{D+1}$ $\quad \tilde{x}^{(i)} = [1, x_1^{(i)}, \ldots, x_D^{(i)}] \in \mathbb{R}^{D+1}$

$$\vec{t} = \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix} \in \mathbb{R}^N \qquad \underline{\underline{\tilde{X}}} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_D^{(1)} \\ \vdots & & & \\ 1 & x_1^{(N)} & \cdots & x_D^{(N)} \end{bmatrix} \in \mathbb{R}^{N \times (D+1)}$$

résidus : $\varepsilon^{(i)} = t^{(i)} - \tilde{x}^{(i)T}\beta = t^{(i)} - \langle \tilde{x}^{(i)}, \beta \rangle$

$$\vec{\varepsilon} = \begin{bmatrix} \varepsilon^{(1)} \\ \vdots \\ \varepsilon^{(N)} \end{bmatrix} = \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix} - \begin{bmatrix} — \tilde{x}^{(1)} — \\ \vdots \\ — \tilde{x}^{(N)} — \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix}$$

$$= \vec{t} - \underline{\underline{\tilde{X}}}\,\vec{\beta}$$

$$l(\beta) = \frac{1}{N}\sum_{i=1}^{N} \left(\varepsilon^{(i)}\right)^2 = \frac{1}{N}\vec{\varepsilon}^T\vec{\varepsilon} = \frac{1}{N}\left(\vec{t} - \underline{\underline{\tilde{X}}}\beta\right)^T\left(\vec{t} - \underline{\underline{\tilde{X}}}\beta\right)$$

$$l(\beta) = \frac{1}{N}\left(\vec{t}^T - \beta^T\underline{\underline{\tilde{X}}}^T\right)\left(\vec{t} - \underline{\underline{\tilde{X}}}\beta\right) = \cancel{\vec{t}^T\vec{t}} - \vec{t}^T\underline{\underline{\tilde{X}}}\beta - \beta^T\underline{\underline{\tilde{X}}}^T\vec{t} + \beta^T\underline{\underline{\tilde{X}}}^T\underline{\underline{\tilde{X}}}\beta\,)$$

$$\operatorname{grad} l(\beta) = \left[\frac{\partial l}{\partial \beta_0}, \dots, \frac{\partial l}{\partial \beta_D} \right]$$

Exemple $$x^T \beta = x_0 \beta_0 + x_1 \beta_1 + \dots + \beta_D x_D$$

$$\operatorname{grad}_{\beta} (x^T \beta) = \left[\frac{\partial (x^T\beta)}{\partial \beta_0}, \frac{\partial (x^T\beta)}{\partial \beta_1}, \dots, \frac{\partial (x^T\beta)}{\partial \beta_D} \right] = [x_0, x_1, \dots, x_D]$$

$$\frac{\partial (\vec{t}^T \underline{\underline{\tilde{X}}} \beta)}{\partial \beta} = \underline{\underline{\tilde{X}}}^T \vec{t} \longrightarrow -2 \underline{\underline{\tilde{X}}}^T \vec{t}$$

$$(\beta^T \underline{\underline{\tilde{X}}}^T \vec{t})^T = t^T \underline{\underline{\tilde{X}}} \beta$$

$$\begin{bmatrix}\beta_0 & \beta_1\end{bmatrix}\begin{bmatrix}a_{00} & a_{01}\\ a_{01} & a_{11}\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} = a_{00}\beta_0^2 + a_{11}\beta_1^2 + 2a_{01}\beta_0\beta_1$$

$$\underset{\beta}{\text{grad}}\, f(\beta) = \left[\frac{\partial f}{\partial \beta_0}, \frac{\partial f}{\partial \beta_1}\right] = \left[\underbrace{2a_{00}\beta_0 + 2a_{01}\beta_1}, \underbrace{2a_{11}\beta_1 + 2a_{01}\beta_0}\right]$$

$$= 2\begin{bmatrix}a_{00} & a_{01}\\ a_{01} & a_{11}\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}$$

$$\rightarrow \text{grad}\, \ell(\beta) = 2\,\underline{\underline{\tilde{X}}}^T\underline{\underline{\tilde{X}}}\,\beta$$

$$\text{grad}\, \ell(\beta) = 2\,\underline{\underline{\tilde{X}}}^T\underline{\underline{\tilde{X}}}\,\beta - 2\,\underline{\underline{\tilde{X}}}^T t = 0$$

(Equations normales)

$$(\underline{\underline{\tilde{X}}}^T\underline{\underline{\tilde{X}}})^{-1}$$

$$(\underline{\underline{\tilde{X}}}^T\underline{\underline{\tilde{X}}})^{-1}\,\underline{\underline{\tilde{X}}}^T\underline{\underline{\tilde{X}}}\,\beta = \underline{\underline{\tilde{X}}}^T t$$

$$\beta_{OLS} = (\underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}})^{-1} \underline{\underline{\tilde{X}}}^T t$$