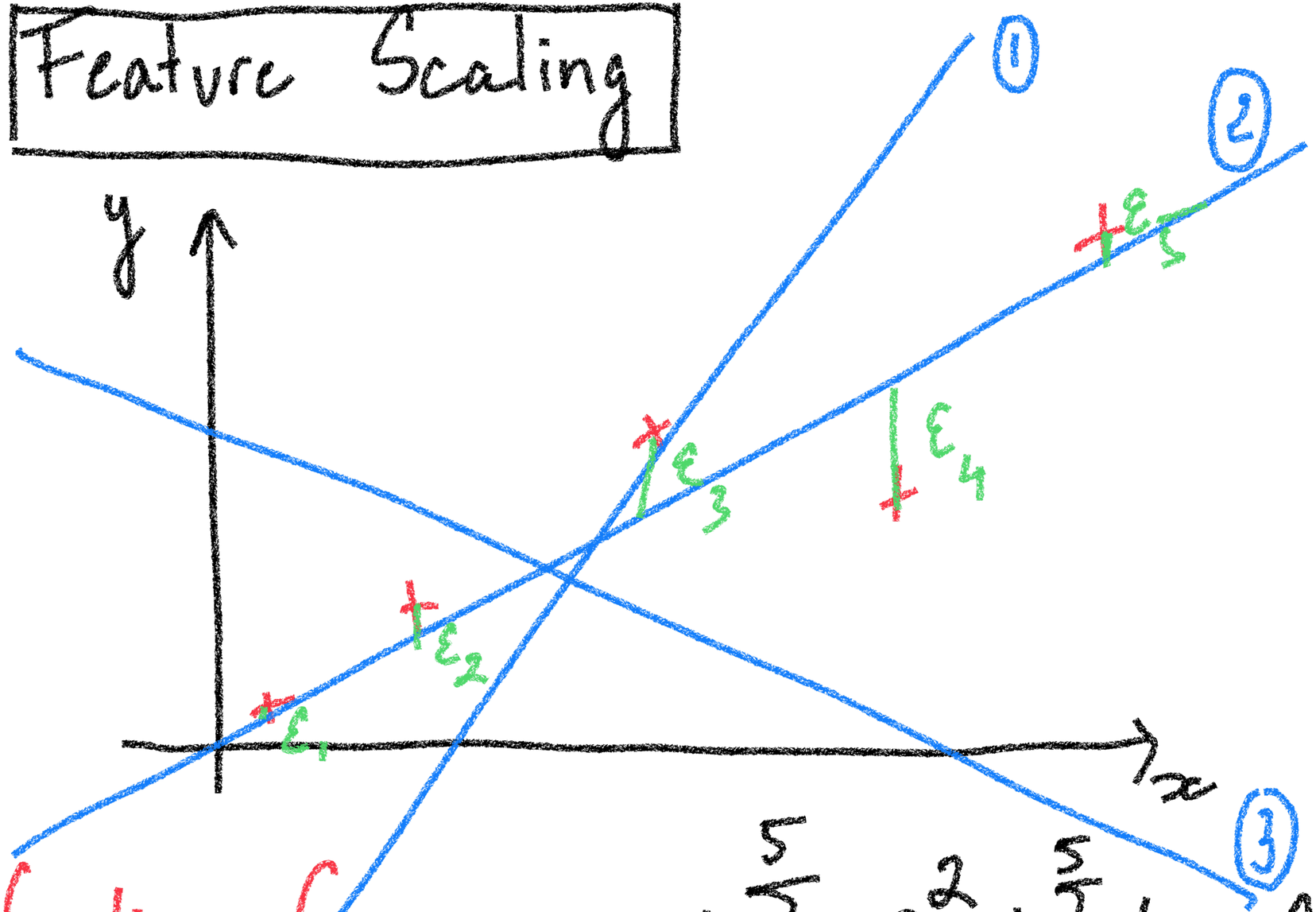
Feature Scaling



x	y
2	4000
3	5000
4,3	4300
⋮	⋮

train dataset

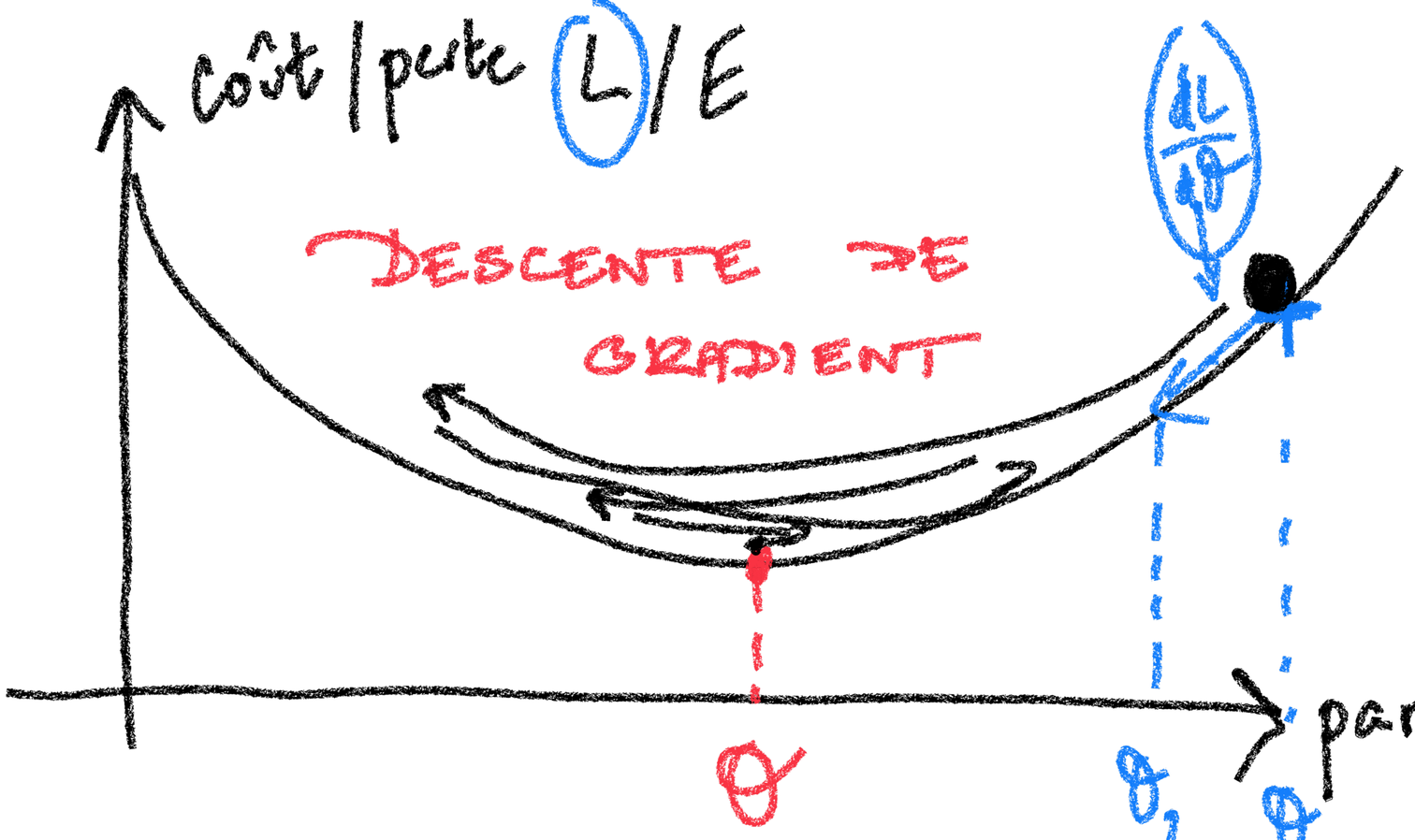
$y = ax + b + \varepsilon$

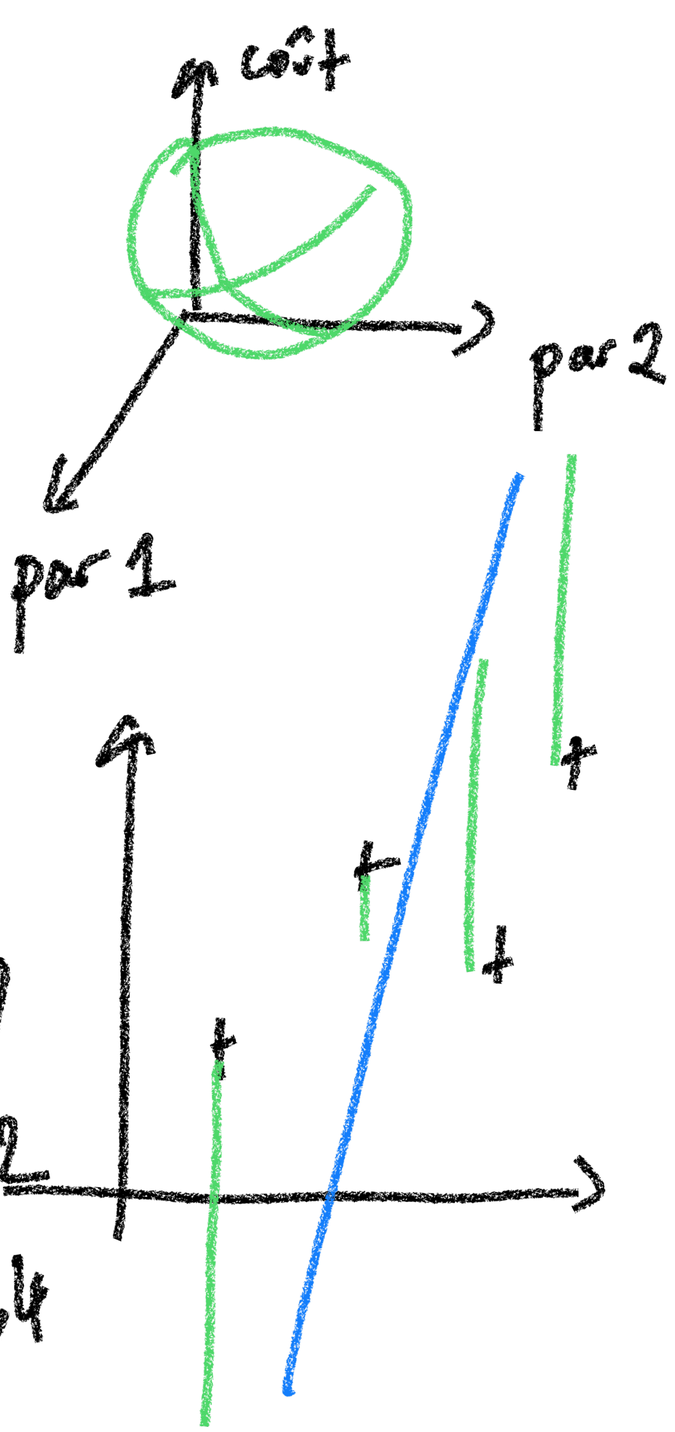
$\varepsilon \sim \mathcal{N}(0, \sigma^2)$

fonction de coût / perte

$$\text{minimiser} \frac{1}{5}\sum_{i=1}^{5} \varepsilon_i^2 \quad \frac{1}{5}\sum_{i=1}^{5}(y_i - \hat{y}_i)^2$$

" MEAN SQUARE ERROR "





$$\theta_2 = \theta_1 - \frac{dL}{d\theta} \cdot lr$$

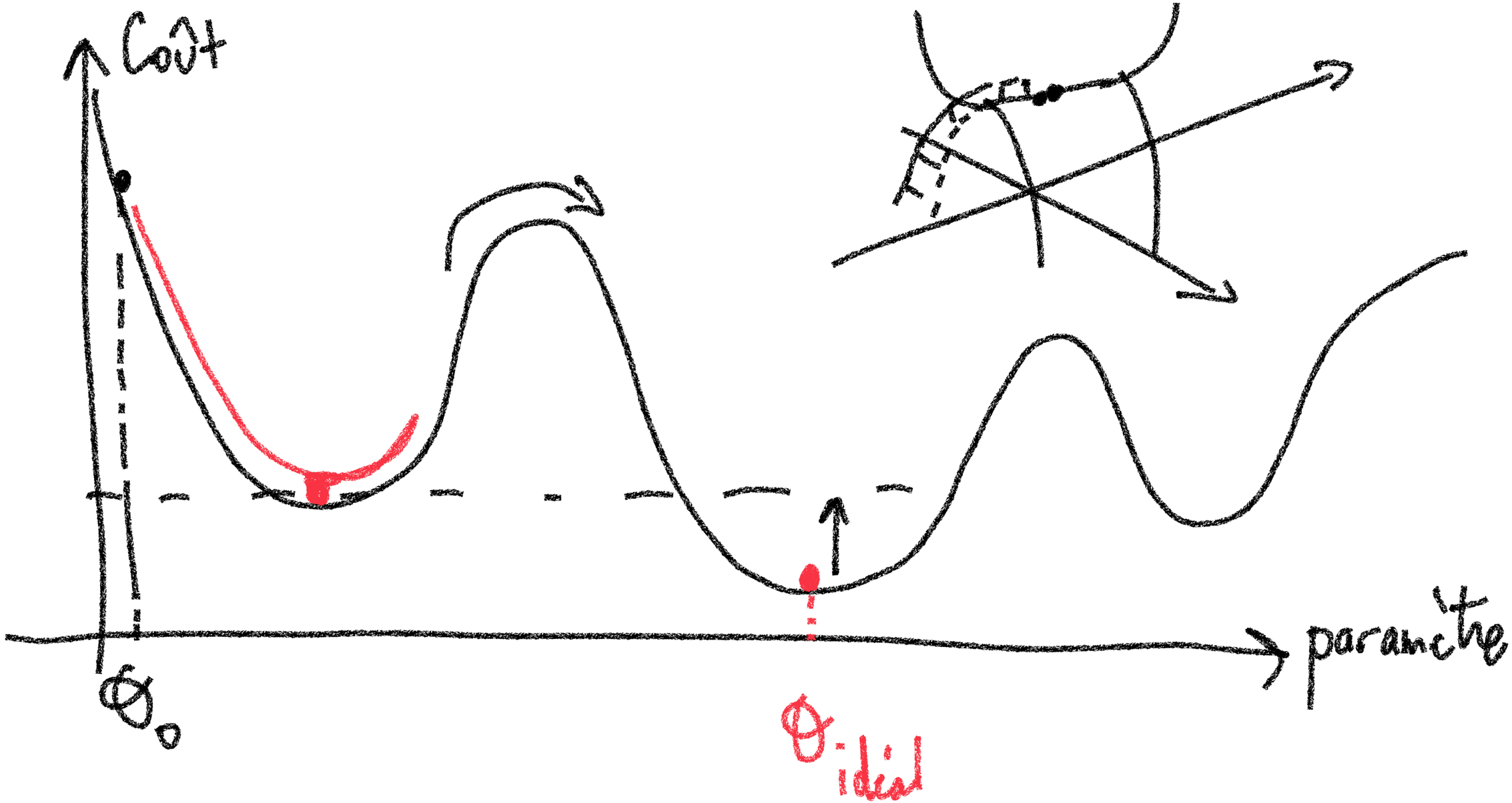
τ learning rate

$\theta_3 = \theta_2 - \dots$

$L = 4,57$

$L = 3,52$

$L = 0,34$


Cout
paramètre
θ_0
$\theta_{idéal}$

x_1	x_2	y
150 000 000	82,7	
320 000 000	78,3	
⋮	⋮	

$$y = a_1 x_1 + a_2 x_2 + a_3$$

$$\begin{cases} a_1 = 2 \\ a_2 = 2 \\ a_3 = 2 \end{cases}$$

On transforme x_1 et x_2 pour qu'ils soient à la même échelle.

$$\begin{cases} a_1' = a_1 - (lr) \dfrac{dL}{da_1} \\ a_2' = a_2 = (lr) \dfrac{dL}{da_2} \end{cases}$$

↖ échelles de grandeur de x_1 et x_2 sont complètement ≠

x_1	x_2	y

df.describe()

	min	max	moyenne
x_1	0	150.10^6	75.10^6
x_2	-5	10	2

Standardisation

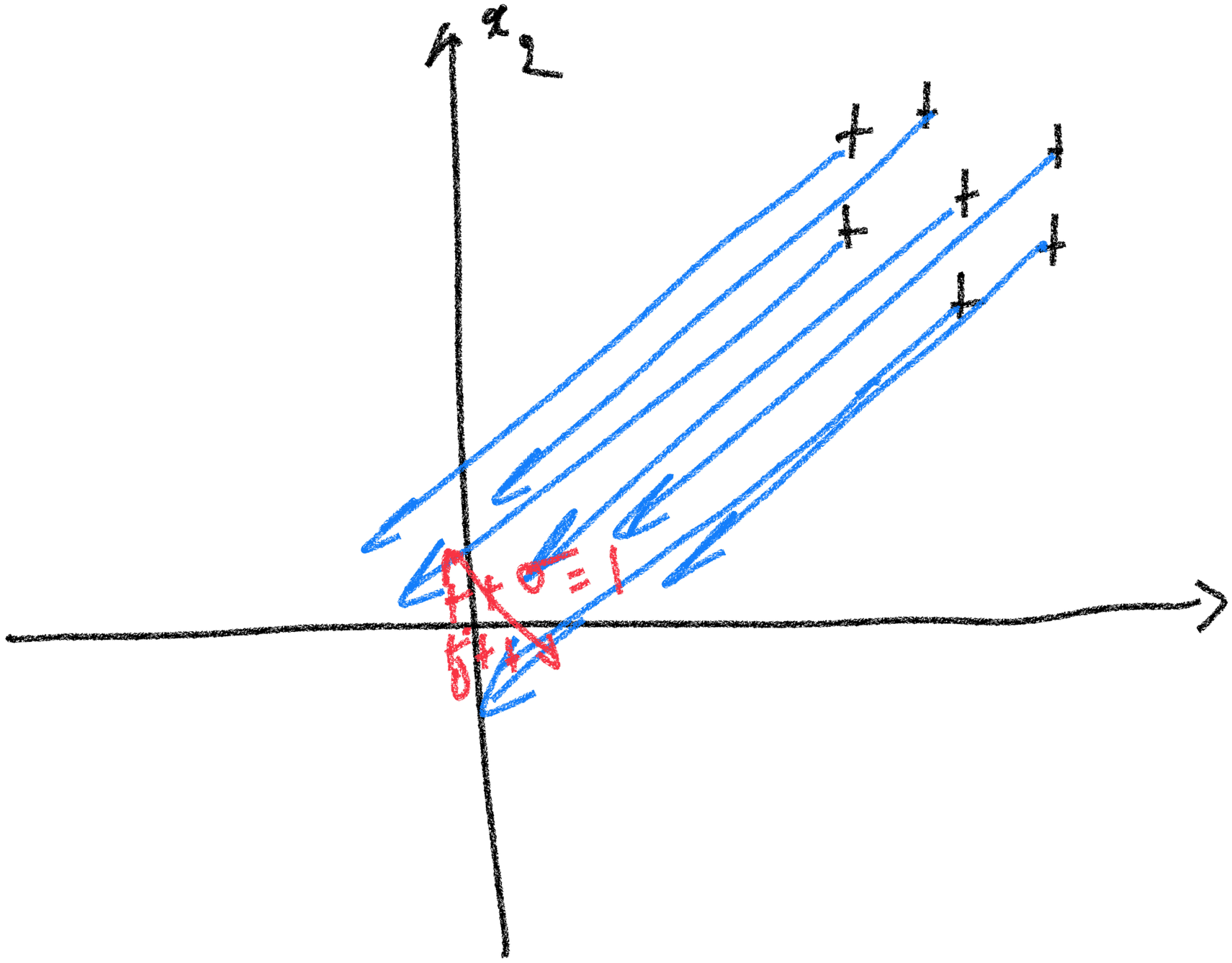
$\mu_2 = 75.10^6$

$\sigma_1 = 5.10^6$

$x_1' \leftarrow \frac{x_1 - \mu_1}{\sigma_1}$

$\mu_1' = 0$

$\sigma_1' = 1$



Standardisation

Feature Normalization

$$x_1' = \frac{x_1 - \min(x_1)}{\max(x_1) - \min(x_1)}$$

x_1	x_1
5	$\frac{5+2}{9} = \frac{7}{9}$
7	1
-2	0

x_1	x_2

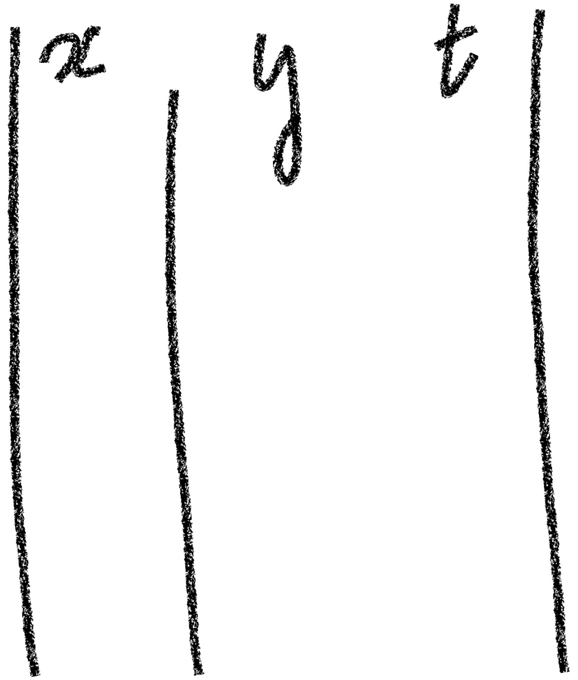
$\min(x_1) = -2$

$\max(x_1) = 7$

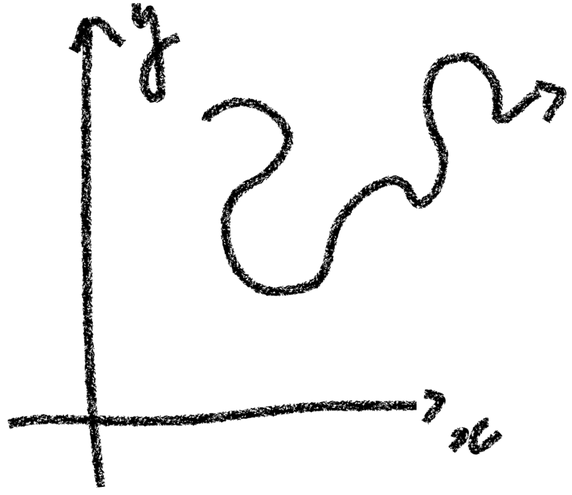
$\max - \min = 9$

x_1, x_2

Feature engineering



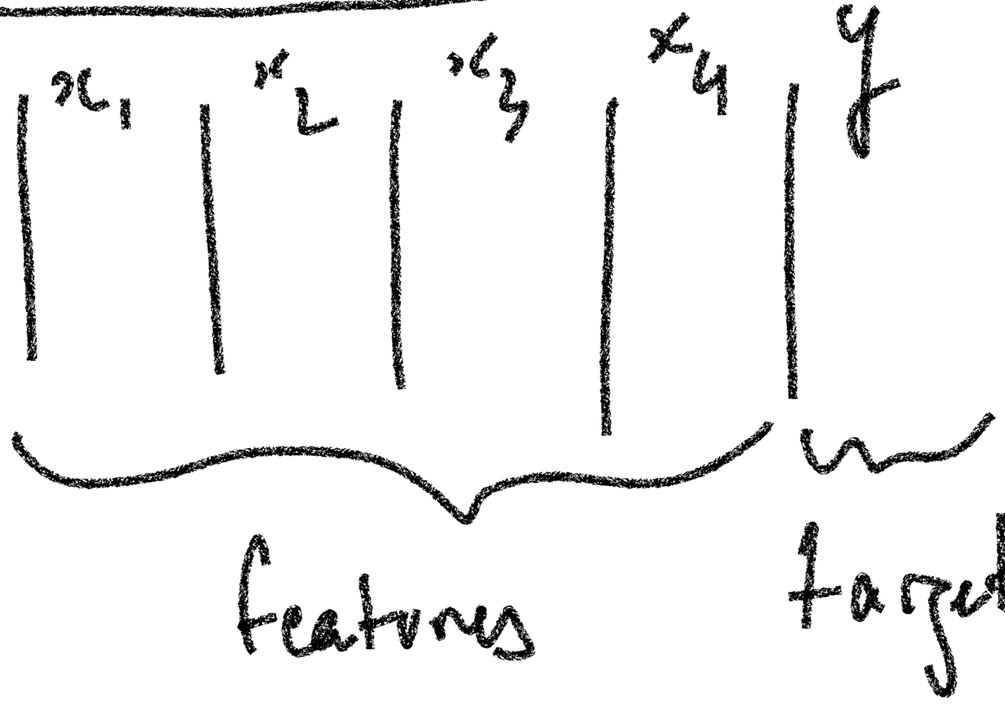
→ conducteur est il ivre ?



$$v_x = x_{t+1} - x_t$$

$$v_y = y_{t+1} - y_t$$

Feature Selection



Si on a trop de corrélation entre 2 variables x_i et x_j, la régression linéaire va moins bien marcher.

Curse of Dimensionality

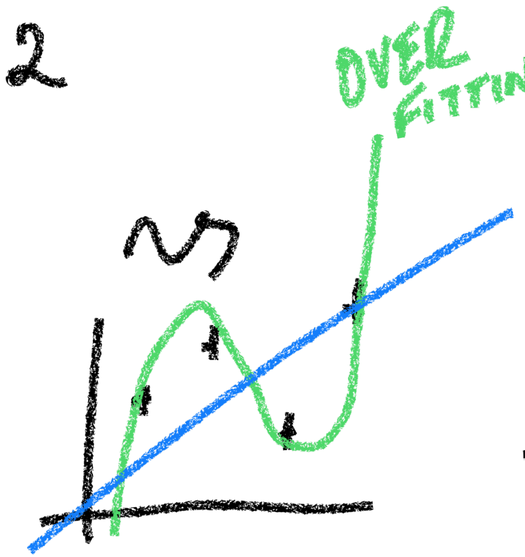
- **Score de corrélation de Pearson** (exc): 0,9 → ?

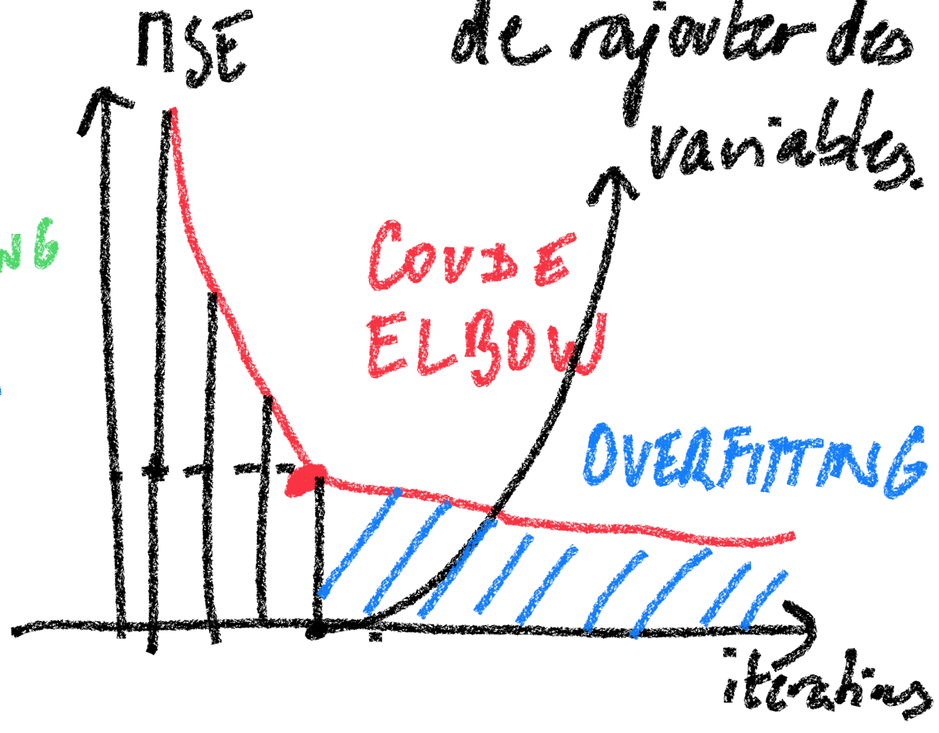
[Matrice de Corrélation] → Vous éliminez une des 2 variables qui ont un score > 0,9

- **Forward Feature Selection**

$y = a_1 x_1 + b$ MSE = 5,2

$y = a_1 x_1 + a_2 x_2 + b$ MSE = ...





On s'arrête de rajouter des variables.

- Feature Scaling : Si vous avez des colonnes qui ont des échelles de grandeur très ≠, vous devez (peut-être) utiliser de la standardisation.
- F. Eng : Si selon vous il manque des variables explicatives importantes, importez les ou construisez les.
- F. Selection : Si vous avez trop de colonnes peu informatives, ou trop similaires, supprimez les.