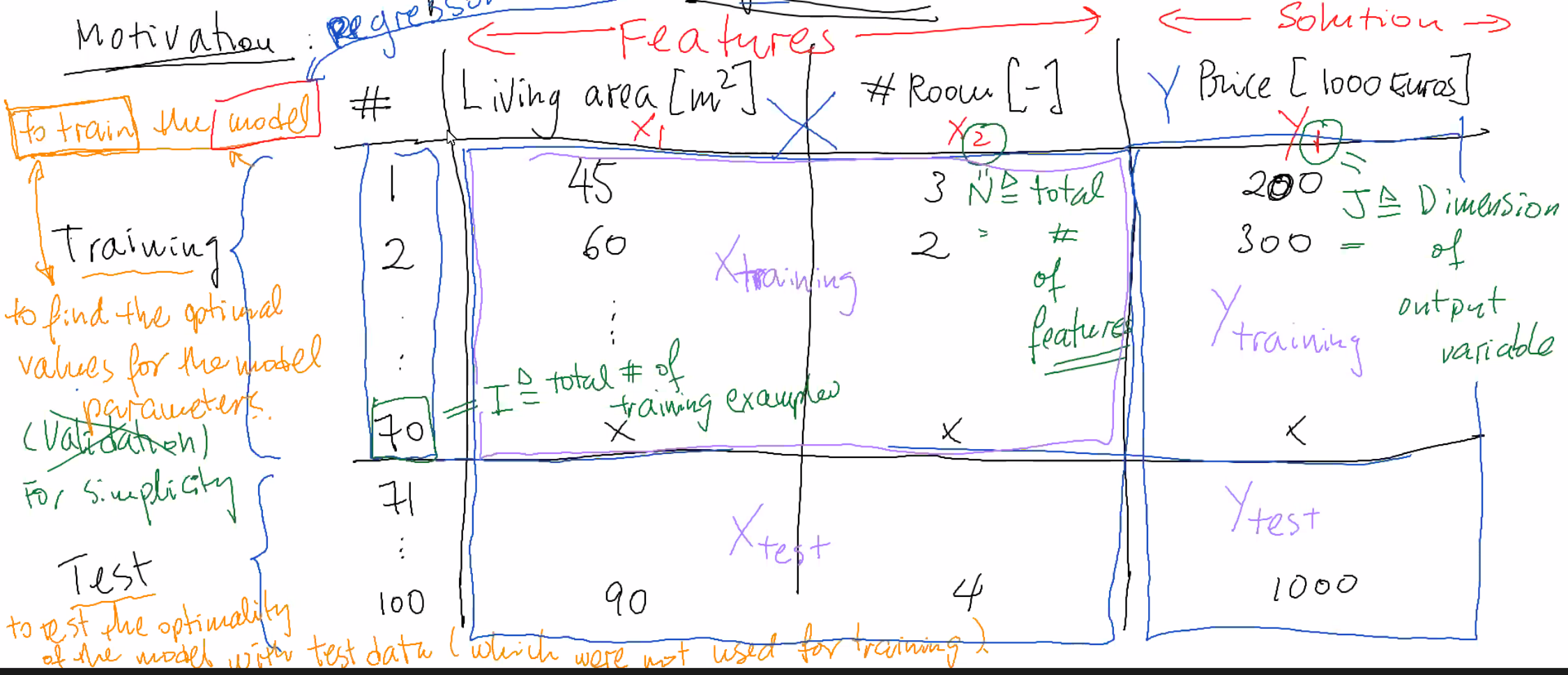
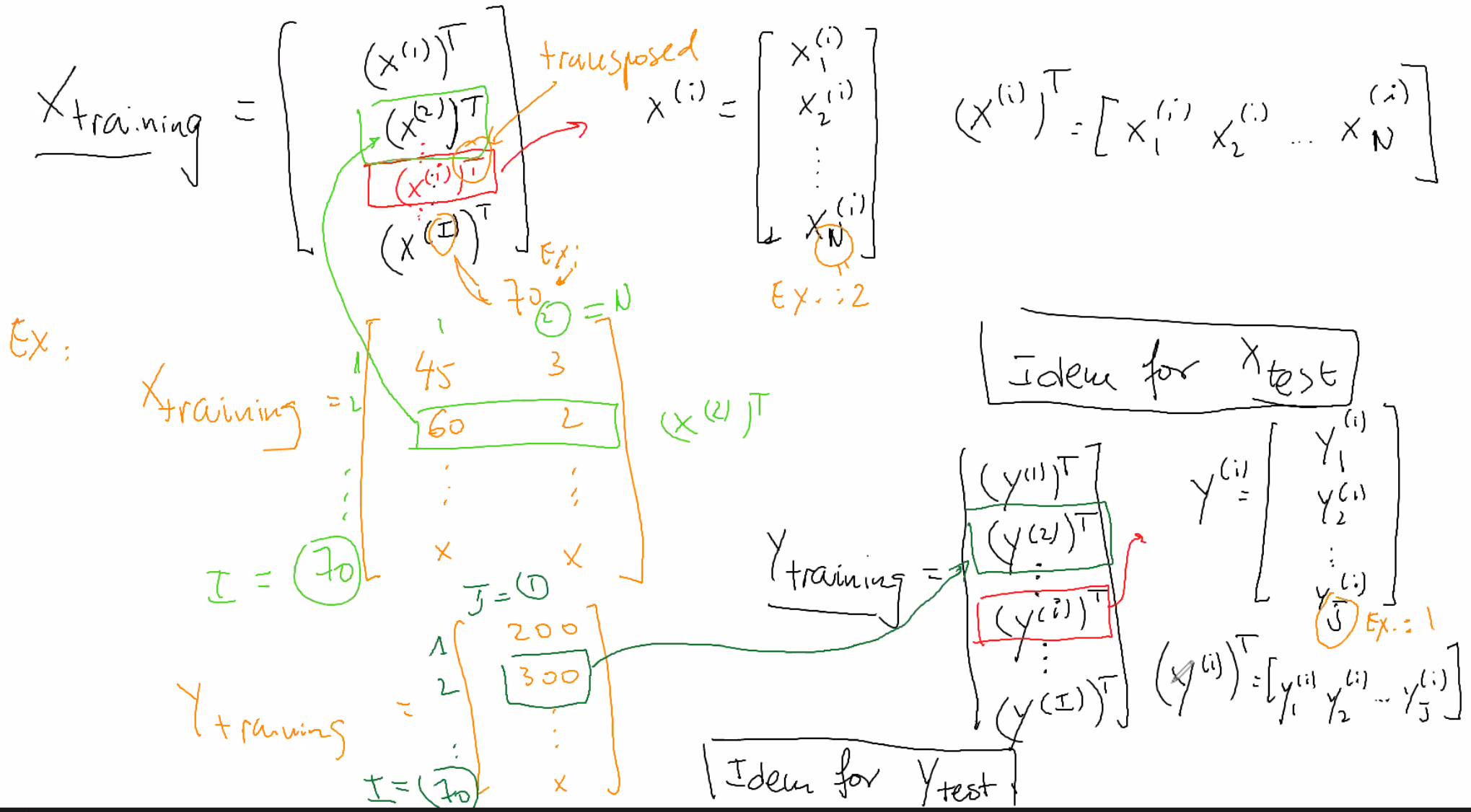
# Lesson 5 : (Non-regularized) Regression

## Motivation : Regressor





$X_{\text{train}} = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \vdots \end{bmatrix}$ transposed $x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \end{bmatrix}$ $(x^{(i)})^T$ $\begin{bmatrix} (i) & (i) & (i) \end{bmatrix}$

Regression pb = problem

Find the hyperplane that optimally represents the relation between some given data and their solutions.

similar — mathematically

Hypothesis $f^{on}$ = function.

$y \sim \hat{y}^{(i)} = h_\theta(x^{(i)})$ ← an example of features

model parameters

Actual output variable — Estimated output variable

$X \in \mathbb{R}^{N \times 1}$

$X^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_N^{(i)} \end{bmatrix}$, $Y^{(i)} = \begin{bmatrix} y_1^{(i)} \\ \vdots \\ y_J^{(i)} \end{bmatrix}$, $Y$

Variable ≠ parameters

- model parameters $\theta$
- Hyper parameters

$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$ ← Intercept or Bias

For linear regression : $\hat{y}^{(i)} = h_\theta(x^{(i)}) = \theta_0 1 + \theta_1 x_1^{(i)} + \cdots + \theta_N x_N^{(i)}$

$\begin{bmatrix} 1 \\ x_1^{(i)} \\ x_2^{(i)} \end{bmatrix}$ (intercept term)

Zoom Réunion

## Aside comment :

→ Linear hypothesis for : $\hat{y}^{(i)} = h_\theta(x^{(i)}) = \theta^T x^{(i)}$

(logistic regression / Sigmoid) hypothesis for : $\hat{y}^{(i)} = h_\theta(x^{(i)}) = \frac{1}{1+e^{-\theta^T x^{(i)}}}$

## Regression pb (mathematically defined) :

Ex :

| Index | Value |
|---|---|
| 0 | 3 |
| 1 | 2 |
| 2 | 1 |

$z = 2$

$\min z = 1$

$\operatorname{argmin} z = 2$

How ?

$$\min_\theta \; ? \; \underbrace{E}_{\text{Error}}(X, Y, \theta) \Rightarrow E_{min}$$

$$\theta^* = \operatorname*{argmin}_\theta E(X, Y, \theta)$$

Def : Error for :

$E \triangleq$ Sum of squared Errors (SSE) $= \frac{1}{2} \sum_{i=1}^{I} (\hat{y}^{(i)} - y^{(i)})^2 = \frac{1}{2} \sum_{i=1}^{I} (h_\theta(x^{(i)}) - y^{(i)})^2$

in this course

if $I = 1$ → # MSE : Mean Square Error

# Approaches to solve the regression pb. ( $\min_{\theta} E(x, y, \theta)$ )

→ 1) Batch Gradient Descent (BGD)

2) Stochastic Gradient Descent (SGD)

3) closed-form solution (CFS).

$$\nabla_\theta E = \begin{bmatrix} \frac{\partial E}{\partial \theta_0} \\ \frac{\partial E}{\partial \theta_n} \\ \frac{\partial E}{\partial \theta_N} \end{bmatrix}$$

$$\theta^{(itera)} = \begin{bmatrix} \theta_0^{(itera)} \\ \theta_n^{(itera)} \\ \theta_N^{(itera)} \end{bmatrix}$$

## 1) Batch Gradient Descent (BGD)

"Pack" "Group"

"Partial Derivatives"

### Pseudo-code (vectorial version)

Initialize $\theta$ randomly → $\theta^{(0)}$

For itera = 1, 2, ... Itera-max:

$$\theta^{(itera)} \leftarrow \theta^{(itera-1)} - \alpha \nabla_\theta E$$

iteration
epoch
episode

Gradient of E w.r.t. $\theta$

learning rate (with small values (0.001 or 0.0001))

### Pseudo-code (scalar version)

Initialize $\theta$ randomly → $\theta^{(0)}$

For itera = 1, 2, ... Itera-max

For n = 0, 1, ... N:

$$\theta_n^{(itera)} \leftarrow \theta_n^{(itera-1)} - \alpha \frac{\partial E}{\partial \theta_n^{(itera-1)}}$$

index for intercept

features

Zoom Réunion

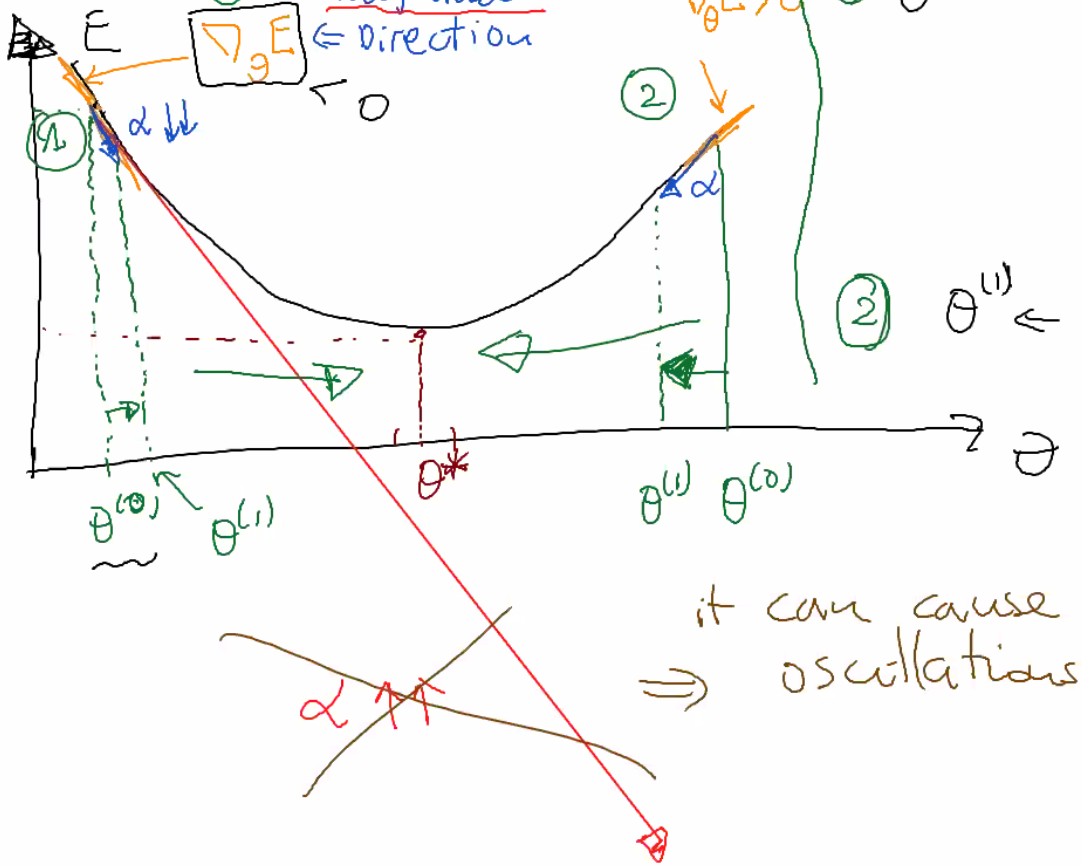
$\theta^{(itera)} \leftarrow \theta^{(itera-1)} - \alpha \nabla_\theta E$

$\alpha$ ← magnitude

$\nabla_\theta E$ ← Direction

Example:

$\theta$ : 1 Dimens.on




$\nabla_\theta E > 0$

① $\theta^{(1)} \leftarrow \theta^{(0)} - \alpha \nabla_\theta E$

$\alpha > 0$, $\nabla_\theta E > 0$, $< 0$

② $\theta^{(1)} \leftarrow \theta^{(0)} - \alpha \nabla_\theta E$

$> 0$, $> 0$, $> 0$, $< 0$

it can cause

$\Rightarrow$ oscillations or instability

Intelligence artificielle (OA)
vous pouvez remettre le tableau d'avant 2sec svp

Répondre

Taper ici pour rechercher

FRA 14:56 02/11/2020

$$\frac{\partial E}{\partial \theta_n} = \frac{\partial}{\partial \theta_n}\left[\frac{1}{2}\sum_{i=1}^{I}\left(\hat{y}^{(i)} - y^{(i)}\right)^2\right] = \frac{\partial}{\partial \theta_n}\left[\frac{1}{2}\sum_{i=1}^{I}\left(\underbrace{h_\theta(x^{(i)})}_{\theta^T \bar{x}^{(i)}} - y^{(i)}\right)^2\right]$$

If $J=1$

$$= \frac{\partial}{\partial \theta_n}\left[\frac{1}{2}\sum_{i=1}^{I}\left(\theta^T \bar{x}^{(i)} - y^{(i)}\right)^2\right]$$

In general,

$$E = \frac{1}{2}\sum_{i=1}^{I}\left\|\hat{y}^{(i)} - y^{(i)}\right\|_2^2 = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\hat{y}_j^{(i)} - y_j^{(i)}\right)^2$$

$$= \frac{1}{2}\sum_{i=1}^{I}\frac{\partial}{\partial \theta_n}\left[\left(\theta^T \bar{x}^{(i)} - y^{(i)}\right)^2\right]$$

$$= \frac{1}{2}\,2\sum_{i=1}^{I}\left(\theta^T \bar{x}^{(i)} - y^{(i)}\right)\frac{\partial\left(\theta^T \bar{x}^{(i)}\right)}{\partial \theta_n}$$

$$\frac{\partial}{\partial \theta_n}\left(\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \cdots + \theta_n x_n^{(i)} + \cdots \theta_N x_N^{(i)}\right)$$

$0 \quad 0 \quad x_n^{(i)} \quad 0$

$$= \sum_{i=1}^{I}\left(h_\theta(x^{(i)}) - y^{(i)}\right) x_n^{(i)}$$

Alternatives: [ mini Batch, stochastic choice ]

Batch !

- Pseudo-code (scalar version) :

$$\theta_n^{(\text{itera})} \leftarrow \theta_n^{(\text{itera}-1)} - \alpha \sum_{i=1}^{I}\left(h_\theta(x^{(i)}) - y^{(i)}\right) x_n^{(i)}$$

# 2) Stochastic Gradient Descent (SGD)

## pseudo-code (scalar version)

Initialize $\theta$ randomly → $\theta^{(0)}$ ← iteration

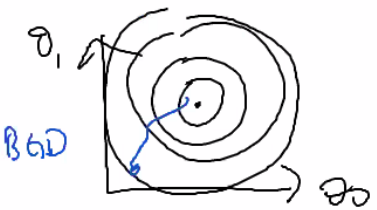
For itera = 1, 2, ..., Itera_max :

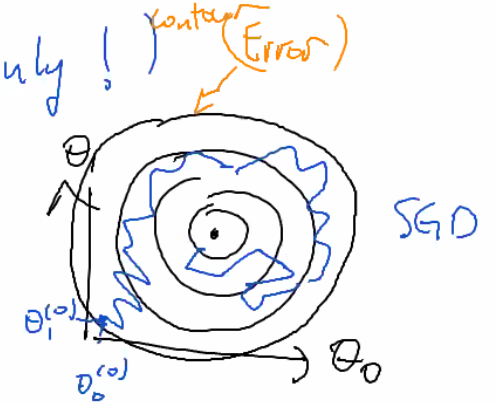
For $n = 0, 1, \dots, N$ :

$i \sim \mathcal{U}[1, \text{Itera\_max}]$

choose randomly following the uniform distribution between 1 and Itera_max (integer values only !)

$$\theta_n^{(\text{itera})} \leftarrow \theta_n^{(\text{itera}-1)} - \alpha \left( h_\theta(x^{(i)}) - y^{(i)} \right) x_n^{(i)}$$




Note : Precision ↓ , Efficiency ↑

Purely mathematic

## 3) Closed-form solution (CFS)

Note.: Precision ↑, Efficiency ↑ (if $N \nearrow$)

$$\boxed{\theta^* = (X^T X)^{-1} X^T y}$$

$X = X_{training} =$

$$\begin{array}{c} \\ 1 \\ \vdots \\ I \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 \cdots & N \end{array} \\ \begin{bmatrix} (\vec{x}^{(1)})^T \\ \vdots \\ (\vec{x}^{(I)})^T \end{bmatrix} \end{array} = \begin{bmatrix} 1 & (x^{(1)})^T \\ \vdots & \vdots \\ 1 & (x^{(I)})^T \end{bmatrix}$$

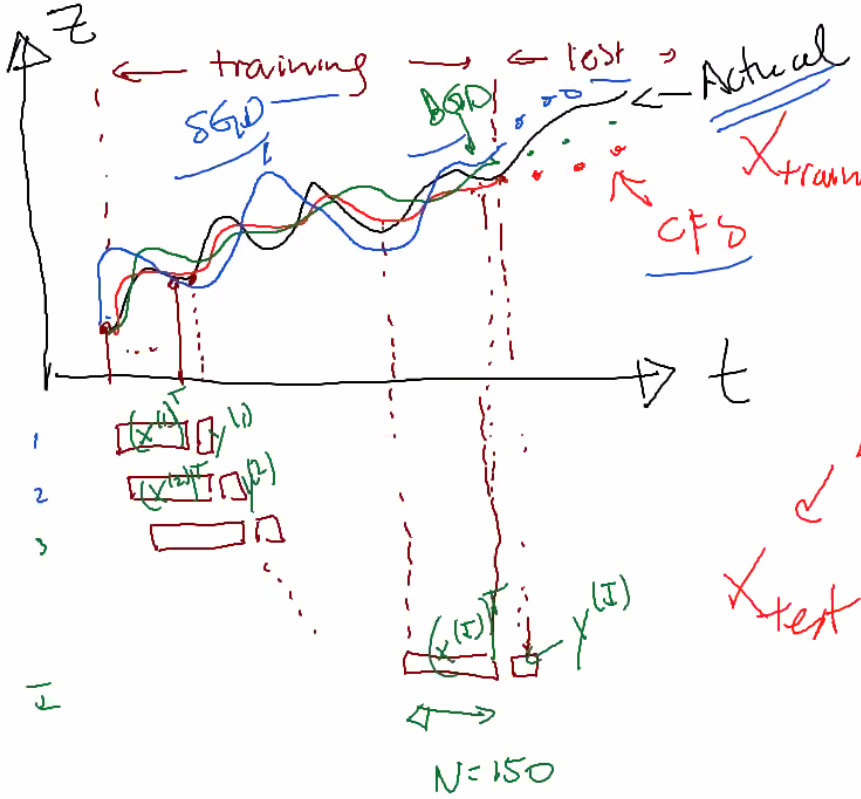
$1 = x_0^{(1)}$

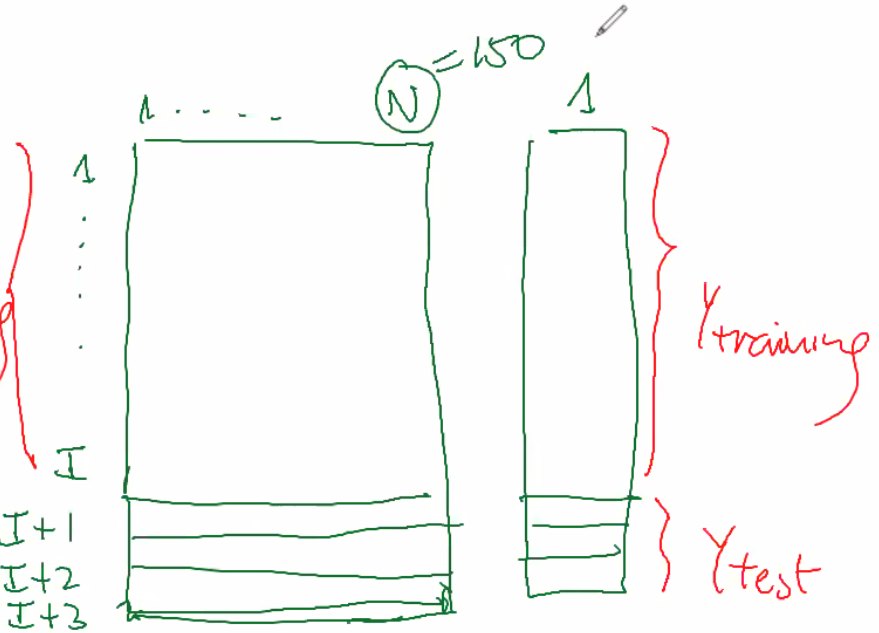
$1 = x_0^{(I)}$

$y = y_{training} =$

$$\begin{array}{c} \\ 1 \\ \vdots \\ I \end{array} \begin{array}{c} \begin{array}{ccc} 1 & \cdots & J \end{array} \\ \begin{bmatrix} (y^{(1)})^T \\ \vdots \\ (y^{(I)})^T \end{bmatrix} \end{array}$$

# Example (Homework)




1) BGD $\Leftrightarrow$ $\theta^*_{BGD} \Rightarrow \hat{Y}_{BGD}$ → training, test

2) SGD $\Leftrightarrow$ $\theta^*_{SGD} \Rightarrow \hat{Y}_{SGD}$ → training, test

3) CFS $\Leftrightarrow$ $\theta^*_{CFS} \Rightarrow \hat{Y}_{CFS}$ → training, test