M1 Apprentissage / 2A

Cours 1: Régression

Alexis Lechervy

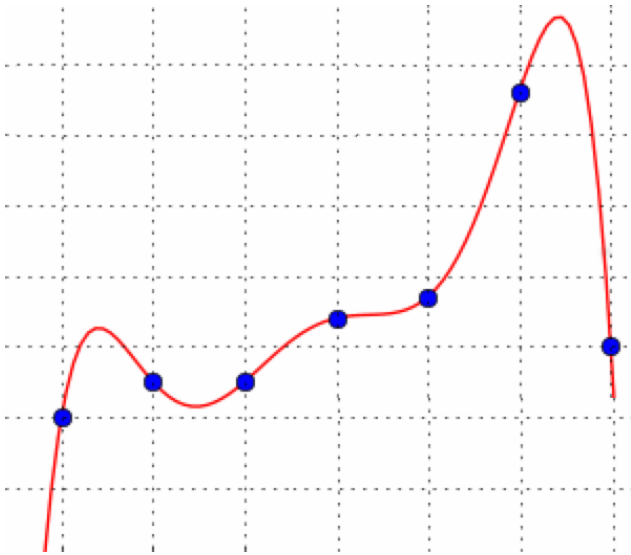
Sommaire

La régression

Principes

- La régression est un apprentissage supervisé. On connait la "vrai valeur" associé à des exemples d'apprentissage.
- L'objectif de la régression est de pouvoir prédire une valeur réel pour un exemple donnée.
- On recherche d'une fonction f_a de paramètre a vérifiant $y = f_a(x)$.

Exemple



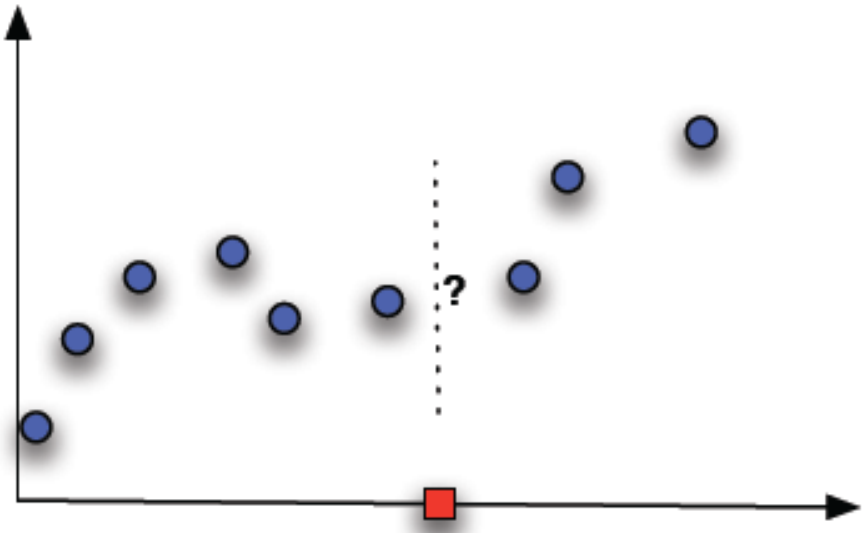
Exemple d'utilisation de la régression

Exemples d'utilisation

- Pression $\rightarrow$ Température d'ébullition.
- Vitesse + taux d'humidité de la route $\rightarrow$ Distance de freinage.
- Taux d'alcool dans le sang $\rightarrow$ temps de réaction.
- ...

But

On souhaite prédire la valeur cible y à partir d'une nouvelle entrée x.

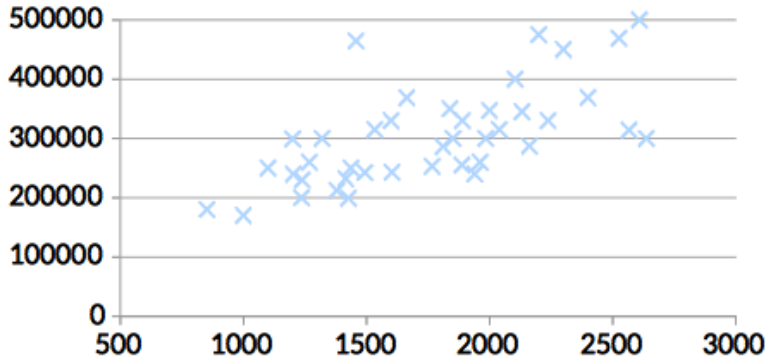


Sommaire

Exemple introductif

Objectif

On cherche à prédire le prix de maisons en millier de dollars à Portland en fonction de leurs surfaces en pied^2. On dispose pour cela de 47 exemples.



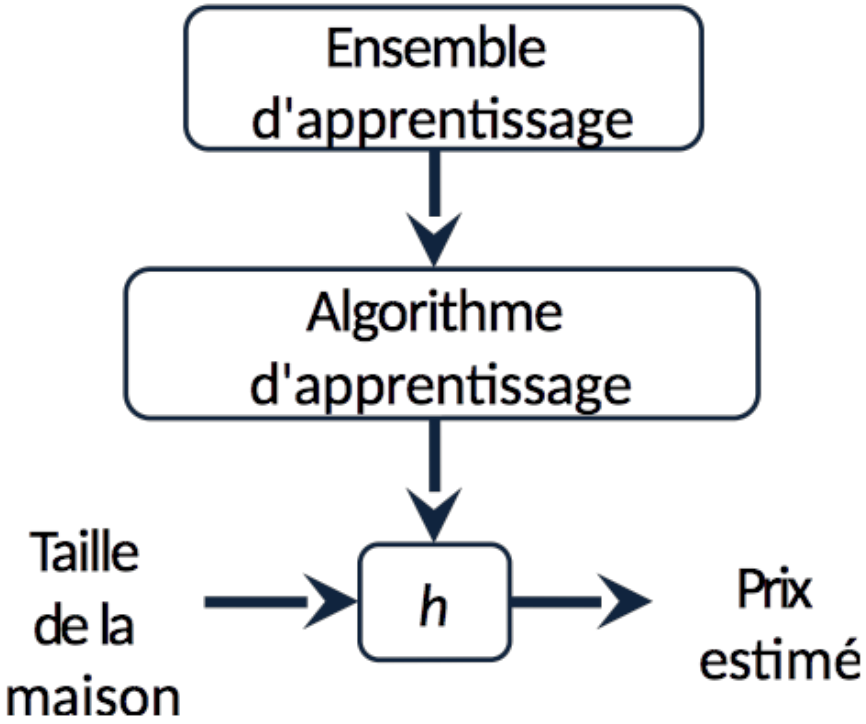
Données

Taille en pied^2	Prix en 1000$
2104	460
1416	232
1534	315
852	178
...	...

Notations

- m nombre d'exemples d'apprentissage.
- x variable d'entrée, descripteur.
- y variable de sortie/cible.

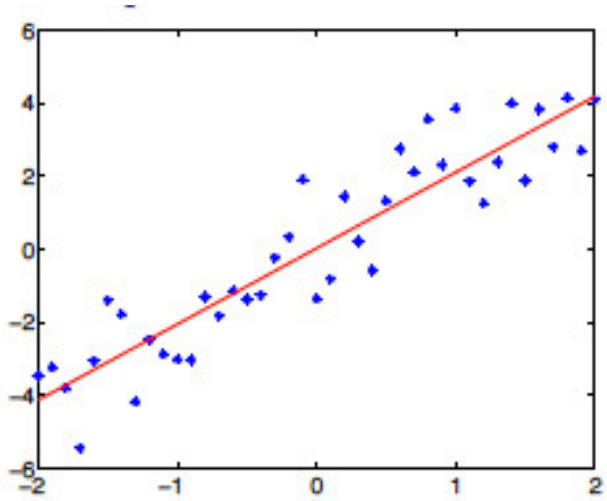
Objectif



Comment représenter l'hypothèse h ?

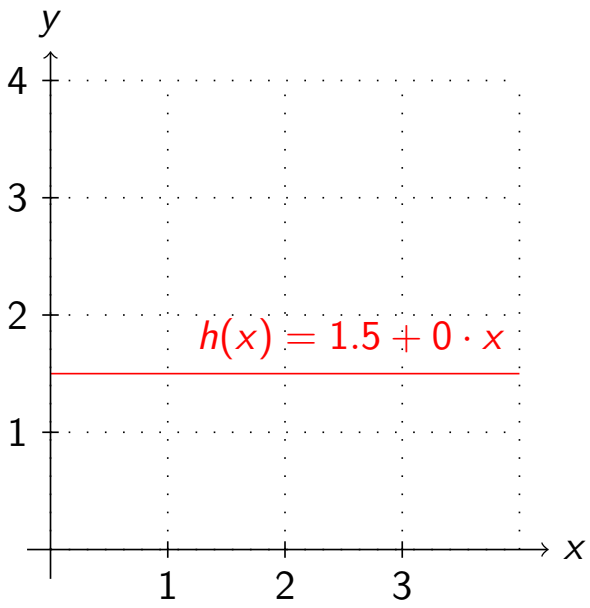
Dans le cas de la régression linéaire à une seule variable, on a :

$$h_\theta(x) = \theta_0 + \theta_1 x.$$

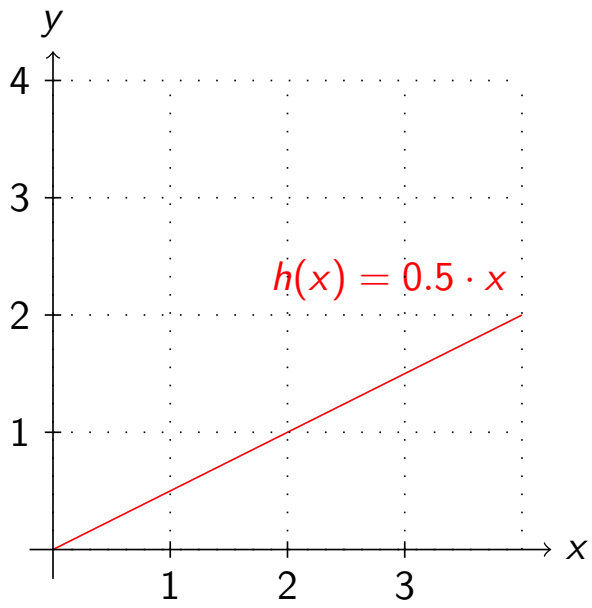


Comment choisir les θ ?

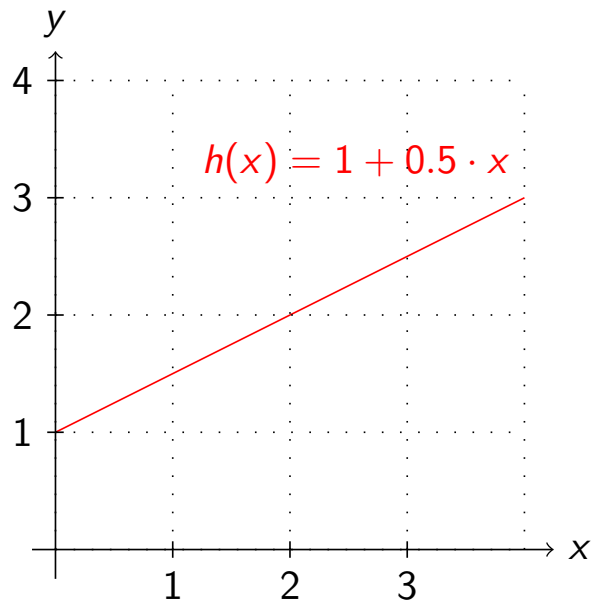
Exemples de valeurs de θ



$\theta_0 = 1.5, \theta_1 = 0$



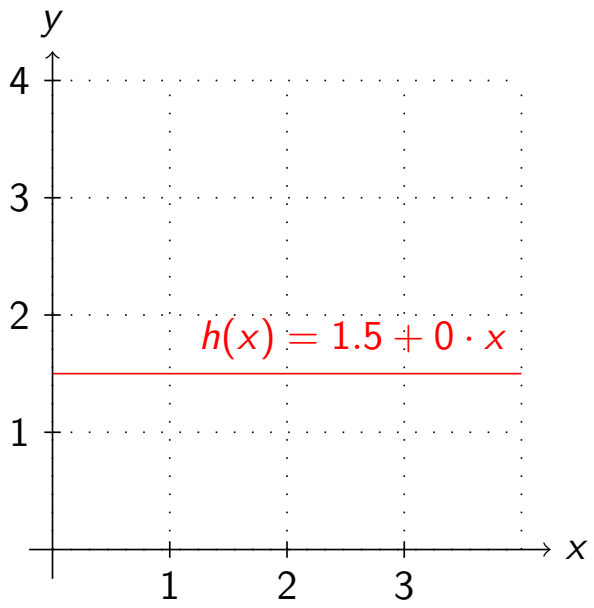
$\theta_0 = 0, \theta_1 = 0.5$



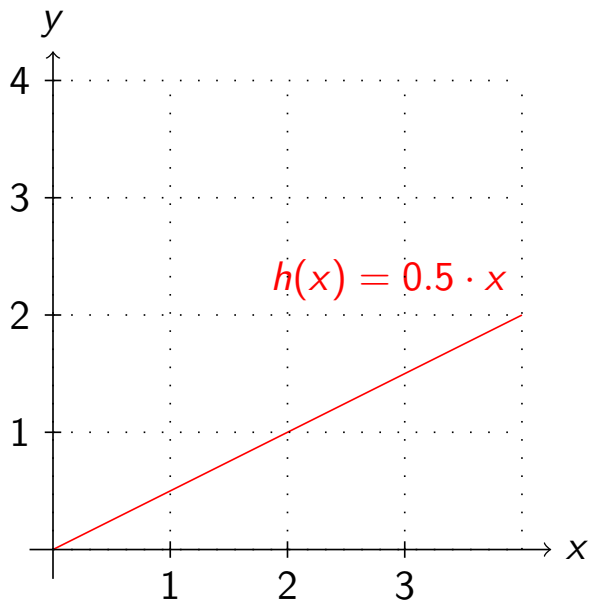
$\theta_0 = 1, \theta_1 = 0.5$

Comment choisir les θ ?

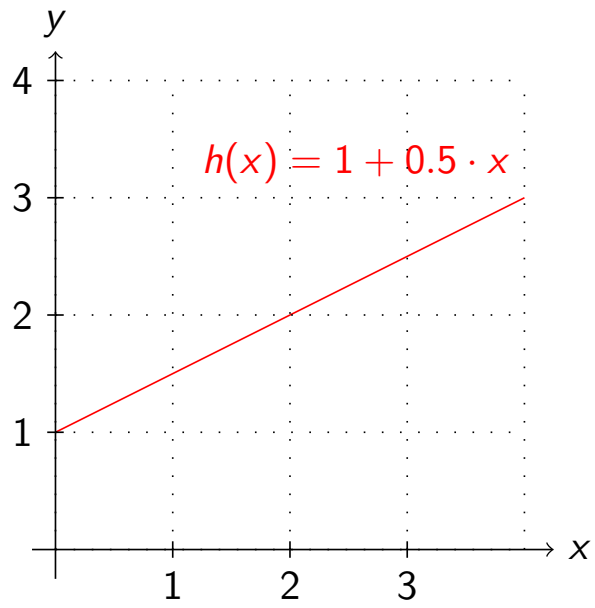
Exemples de valeurs de θ



$\theta_0 = 1.5, \theta_1 = 0$



$\theta_0 = 0, \theta_1 = 0.5$



$\theta_0 = 1, \theta_1 = 0.5$

Idée !

Choisir θ_0, θ_1 tel que $h_\theta(x)$ soit le plus proche possible de y pour les couples d'apprentissage (x, y).

Erreur de prédiction

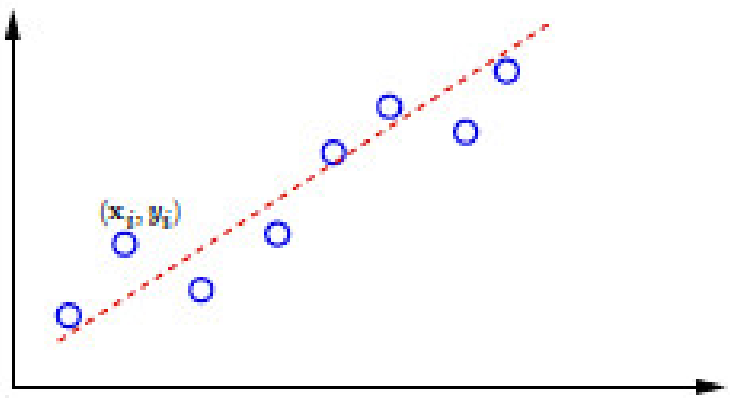
Le critère de moindre carré

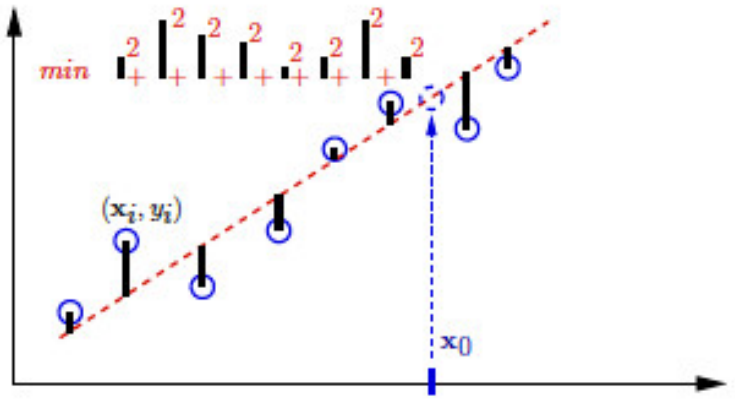
- On peut mesurer l'erreur de prédiction en terme de moyenne des distances au carrés :

$$J(y, \tilde{y}) = \|y - \tilde{y}\|^2.$$

- On cherche donc les θ minimisant la fonction de coût suivantes :

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - h_\theta(x_i))^2.$$





Résumé

Version générale

Hypothèse

$$h_\theta(x) = \theta_0 + \theta_1 x.$$

Paramètres

$$\theta_0, \theta_1.$$

Fonction de coût

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - h_\theta(x_i))^2.$$

Objectif

$$\arg \min_{\theta_0, \theta_1} J(\theta_0, \theta_1).$$

Résumé

Version générale

Hypothèse

$$h_\theta(x) = \theta_0 + \theta_1 x.$$

Paramètres

$$\theta_0, \theta_1.$$

Fonction de coût

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - h_\theta(x_i))^2.$$

Objectif

$$\arg\min_{\theta_0, \theta_1} J(\theta_0, \theta_1).$$

Version simplifiée

Hypothèse

$$h_\theta(x) = \cancel{\theta_0} + \theta_1 x.$$

Paramètres

$$\theta_0 = 0, \theta_1.$$

Fonction de coût

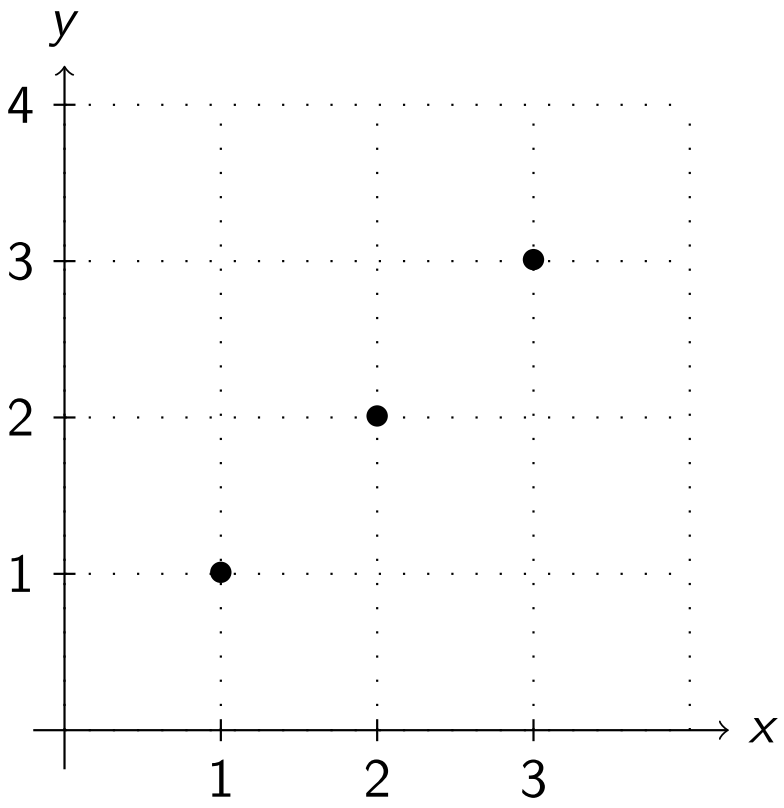
$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - \theta_1 x_i)^2.$$

Objectif

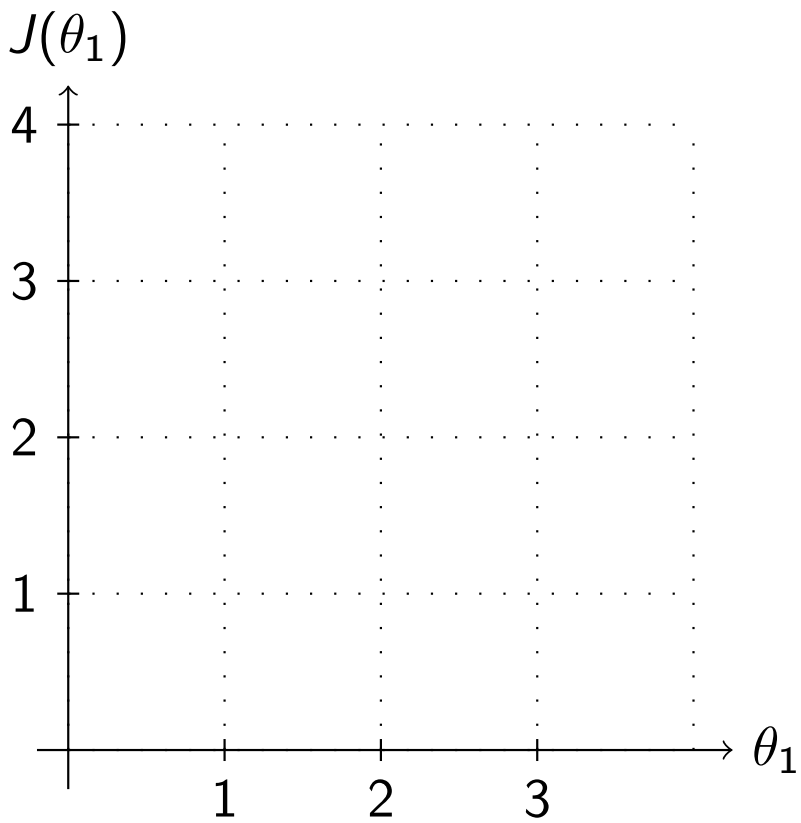
$$\arg\min_{\theta_1} J(\theta_1).$$

Intuition sur la fonction de coût simplifier

$h_\theta(x)$ (pour un θ_1 fixé)

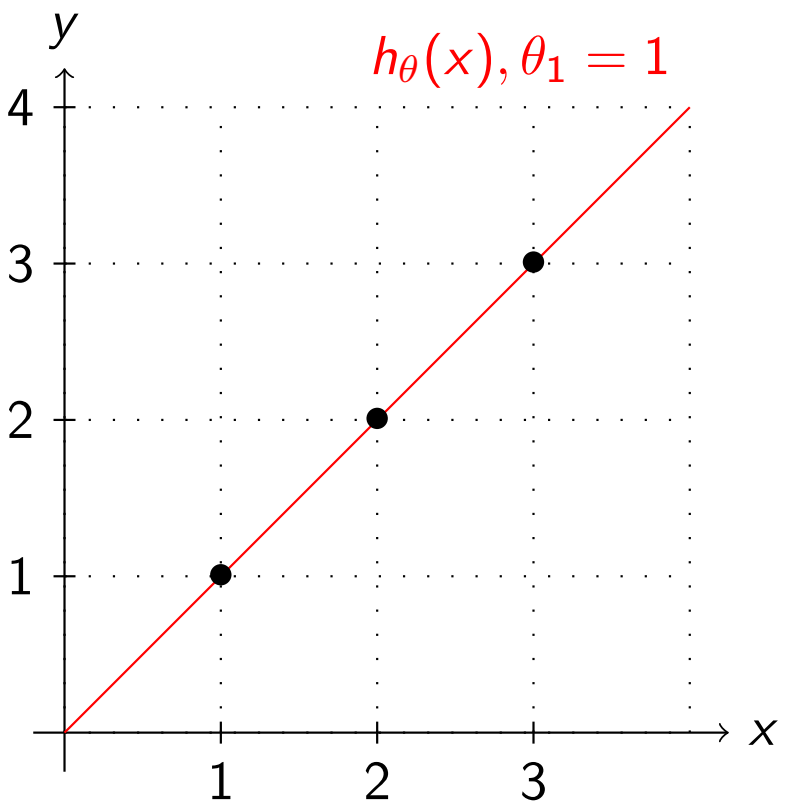


$J(\theta_1)$ (fonction du paramètre θ_1)

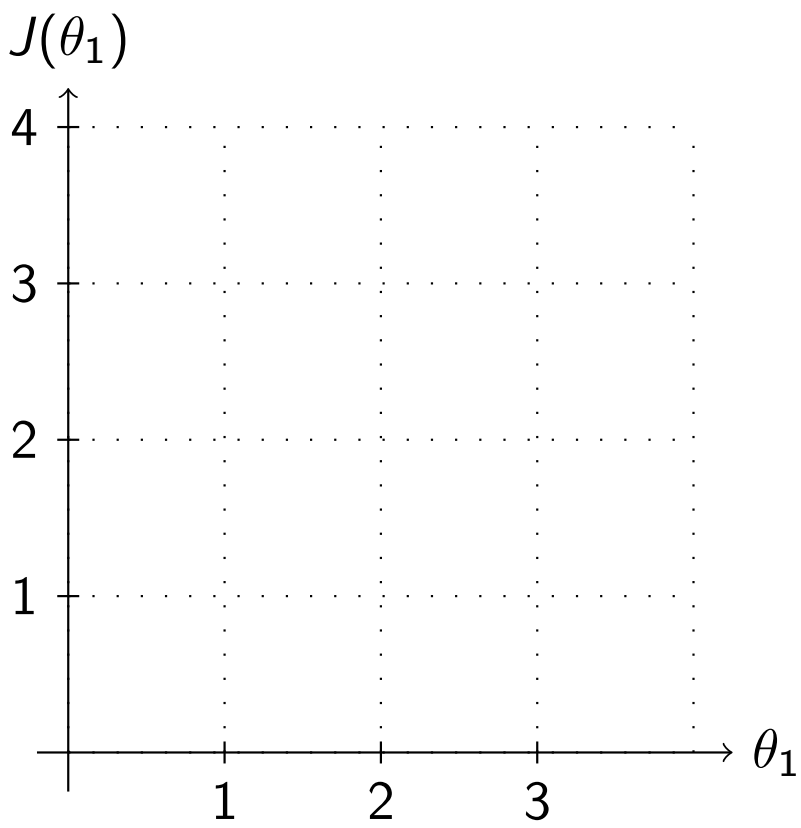


Intuition sur la fonction de coût simplifier

$h_\theta(x)$ (pour un θ_1 fixé)

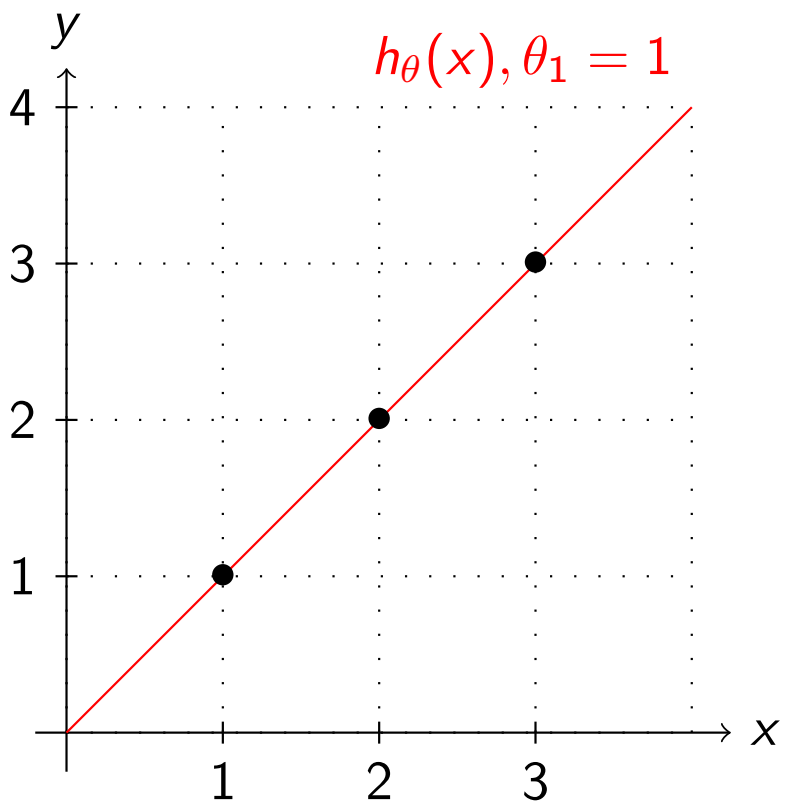


$J(\theta_1)$ (fonction du paramètre θ_1)

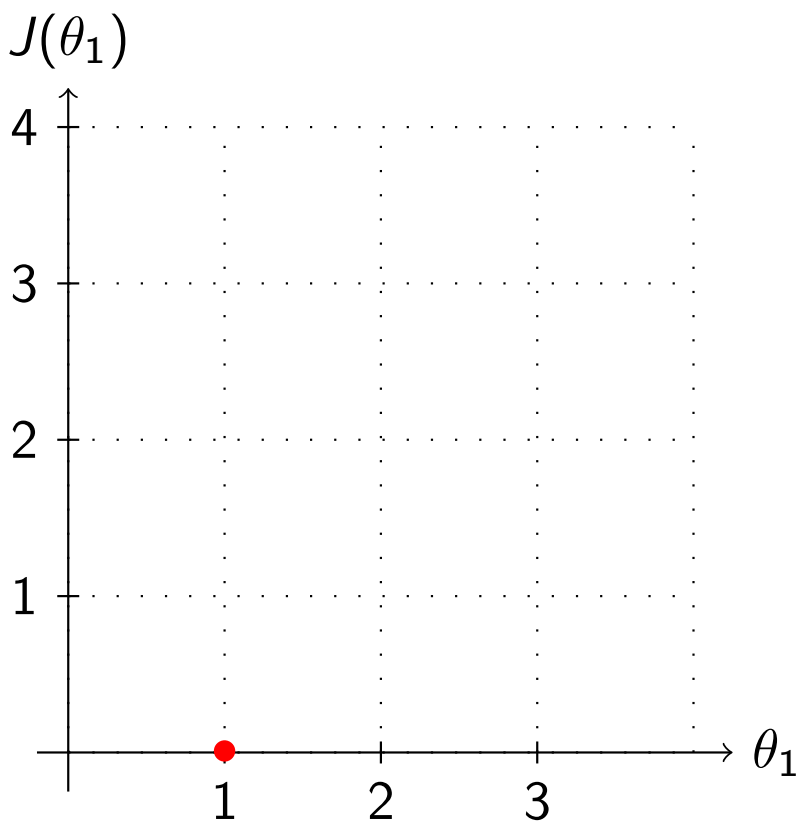


Intuition sur la fonction de coût simplifier

$h_\theta(x)$ (pour un θ_1 fixé)



$J(\theta_1)$ (fonction du paramètre θ_1)

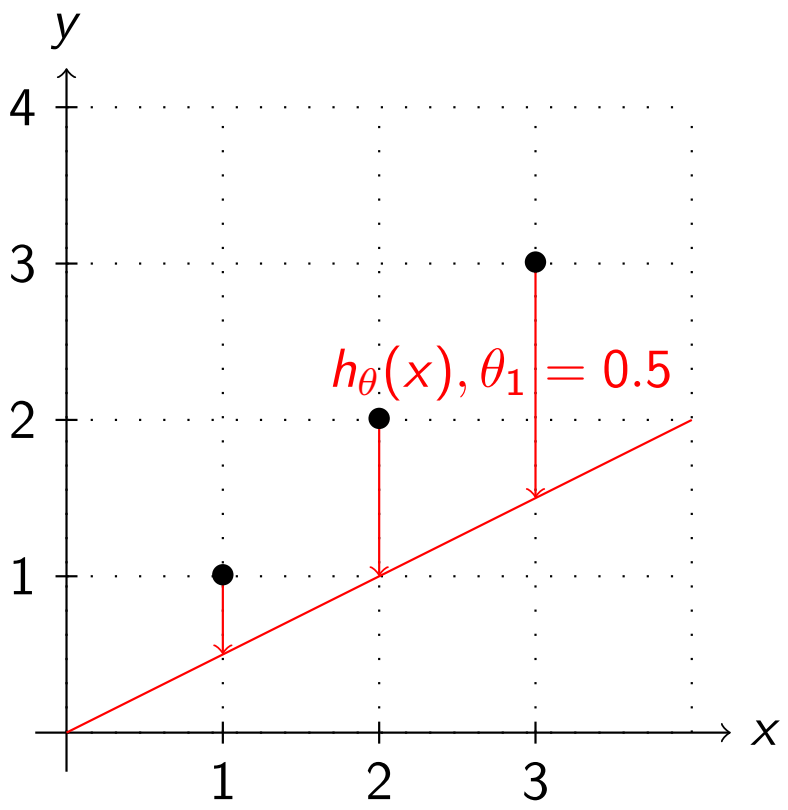


Calcul de la fonction de coût

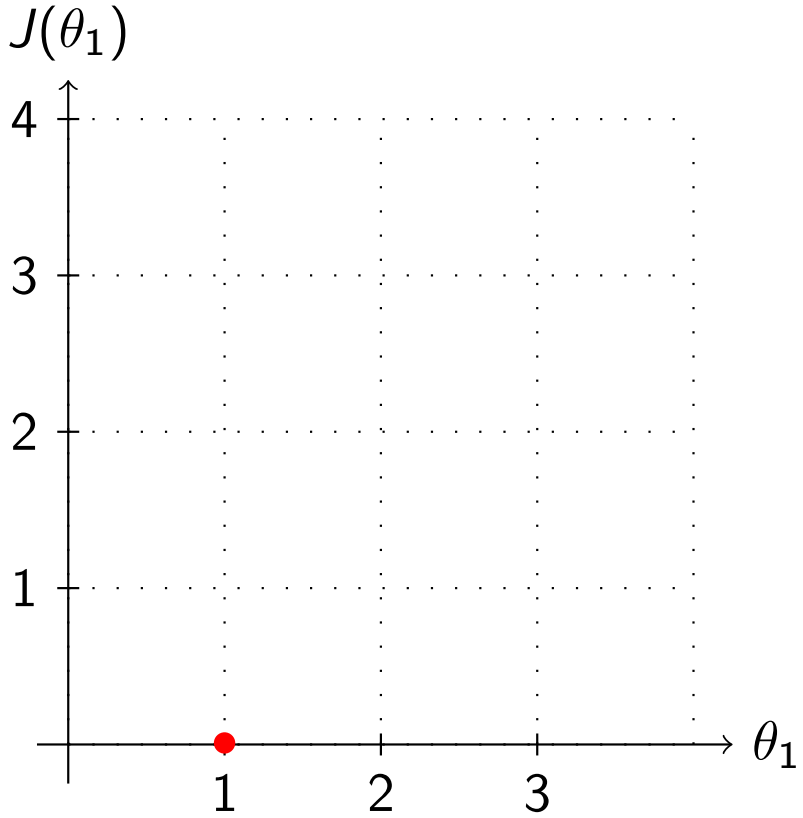
$$J(\theta_1 = 1) = \frac{1}{3}(0^2 + 0^2 + 0^2) = 0.$$

Intuition sur la fonction de coût simplifier

$h_\theta(x)$ (pour un θ_1 fixé)



$J(\theta_1)$ (fonction du paramètre θ_1)

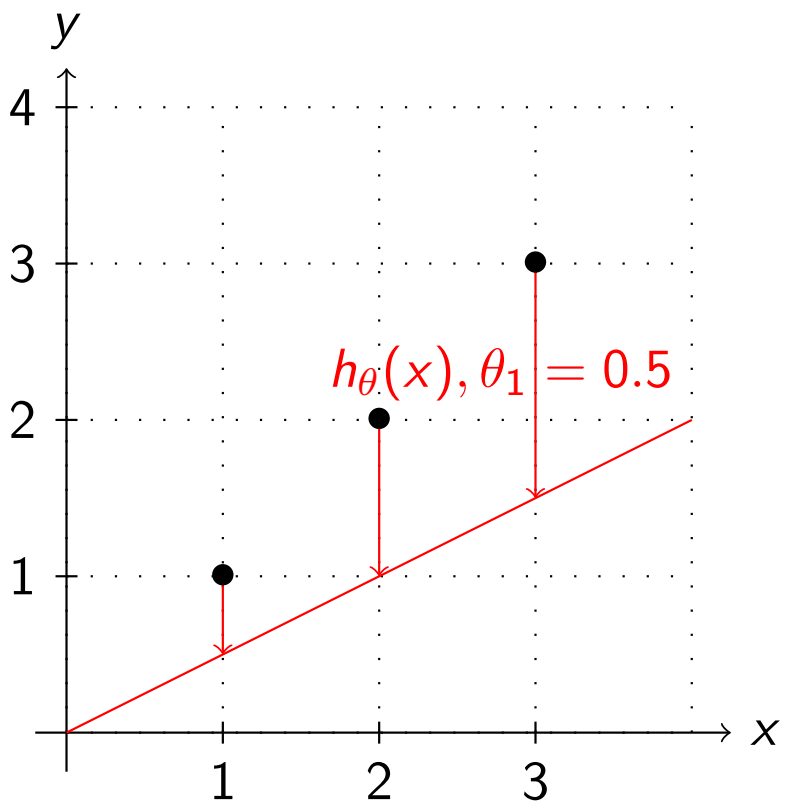


Calcul de la fonction de coût

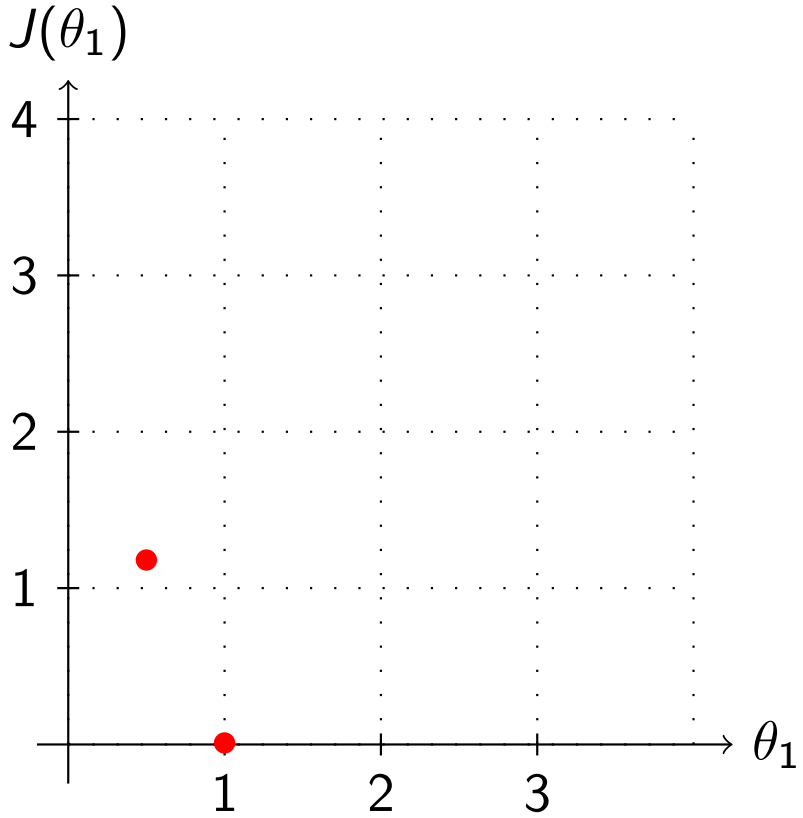
$$J(\theta_1 = 0.5) = \frac{1}{3}([0.5-1]^2 + [1-2]^2 + [1.5-3]^2) = \frac{3.5}{3} \simeq 1.17.$$

Intuition sur la fonction de coût simplifier

$h_\theta(x)$ (pour un θ_1 fixé)



$J(\theta_1)$ (fonction du paramètre θ_1)

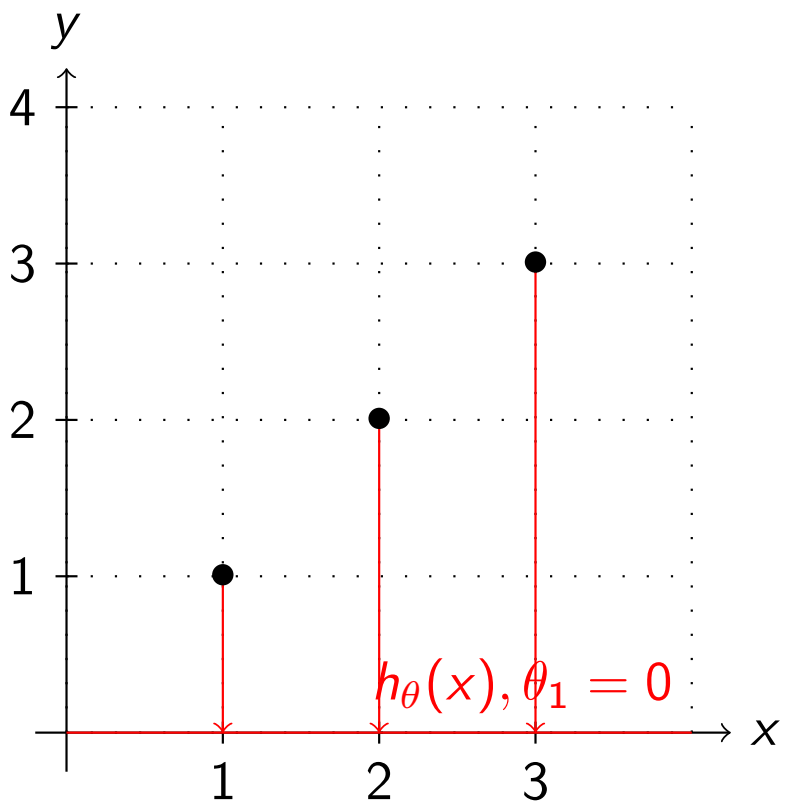


Calcul de la fonction de coût

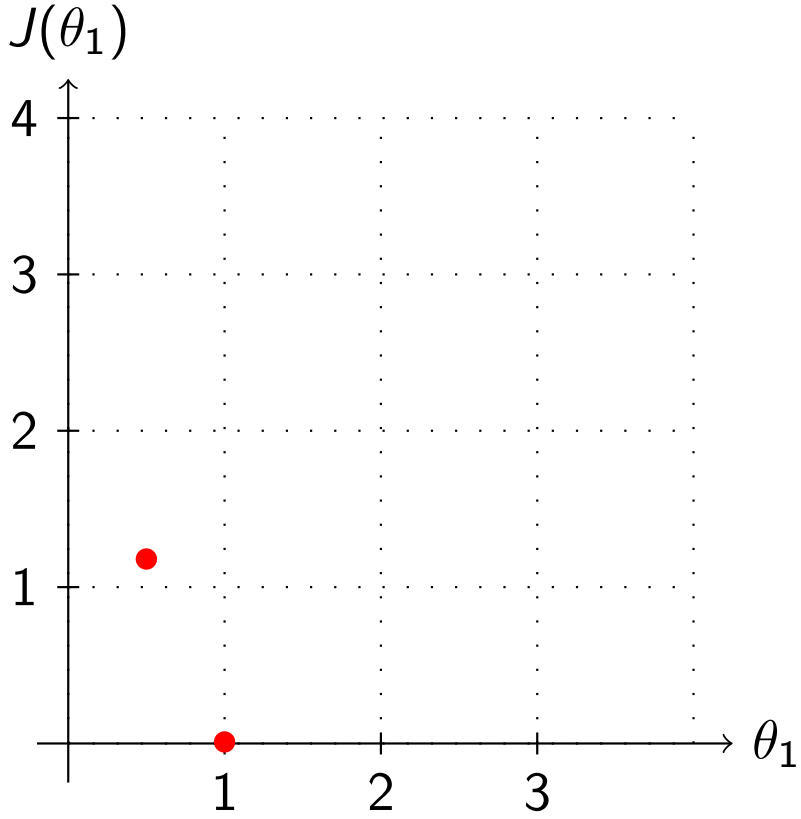
$$J(\theta_1 = 0.5) = \frac{1}{3}([0.5-1]^2 + [1-2]^2 + [1.5-3]^2) = \frac{3.5}{3} \simeq 1.17.$$

Intuition sur la fonction de coût simplifier

$h_\theta(x)$ (pour un θ_1 fixé)



$J(\theta_1)$ (fonction du paramètre θ_1)

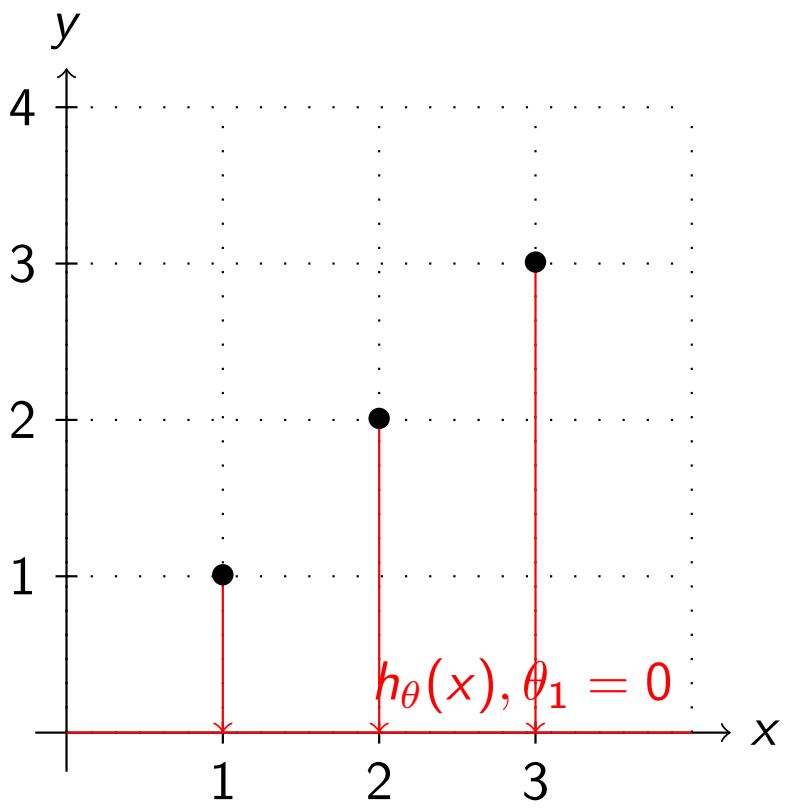


Calcul de la fonction de coût

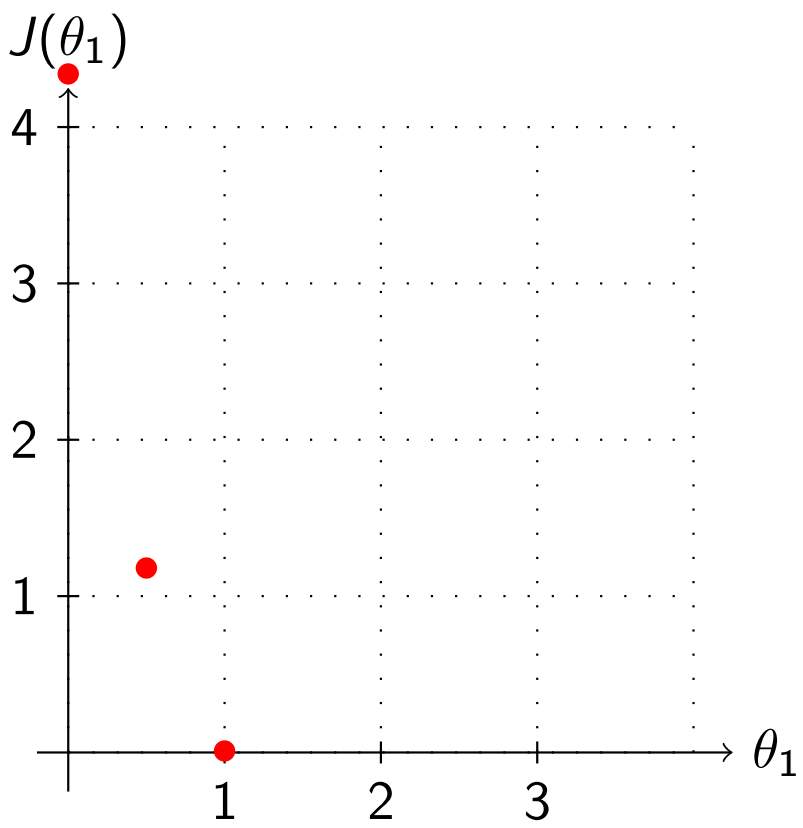
$$J(\theta_1 = 0) = \frac{1}{3}(1^2 + 2^2 + 3^2) = \frac{14}{3} \simeq 4.33.$$

Intuition sur la fonction de coût simplifier

$h_\theta(x)$ (pour un θ_1 fixé)



$J(\theta_1)$ (fonction du paramètre θ_1)

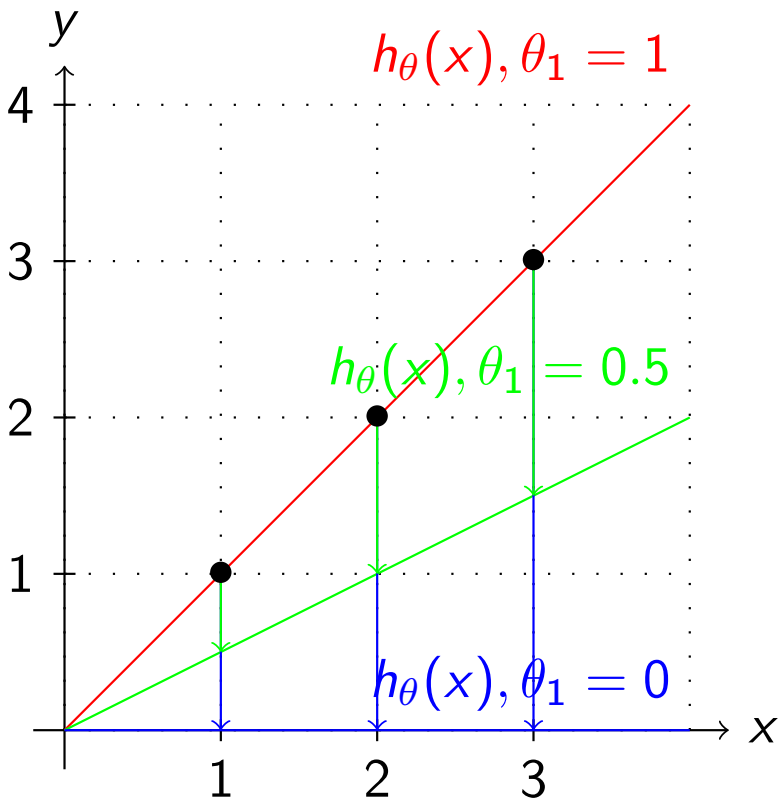


Calcul de la fonction de coût

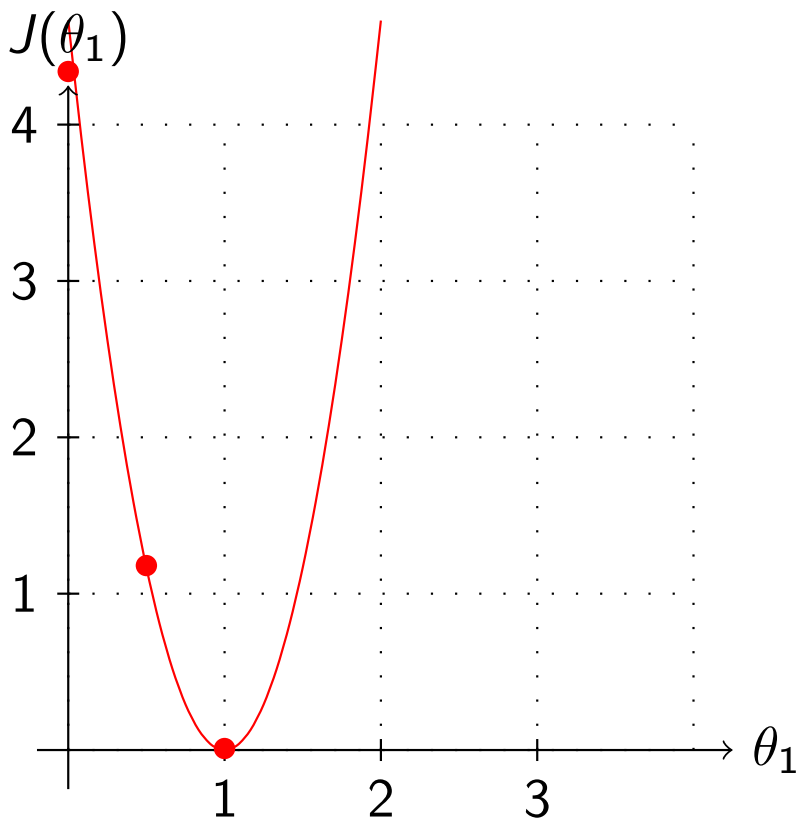
$$J(\theta_1 = 0) = \frac{1}{3}(1^2 + 2^2 + 3^2) = \frac{14}{3} \simeq 4.33.$$

Intuition sur la fonction de coût simplifier

$h_\theta(x)$ (pour un θ_1 fixé)



$J(\theta_1)$ (fonction du paramètre θ_1)



Minimisation de la fonction de coût dans le cas simplifié

Fonction de coût

$$J(\theta_1) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \theta_1 x_i)^2.$$

Objectif

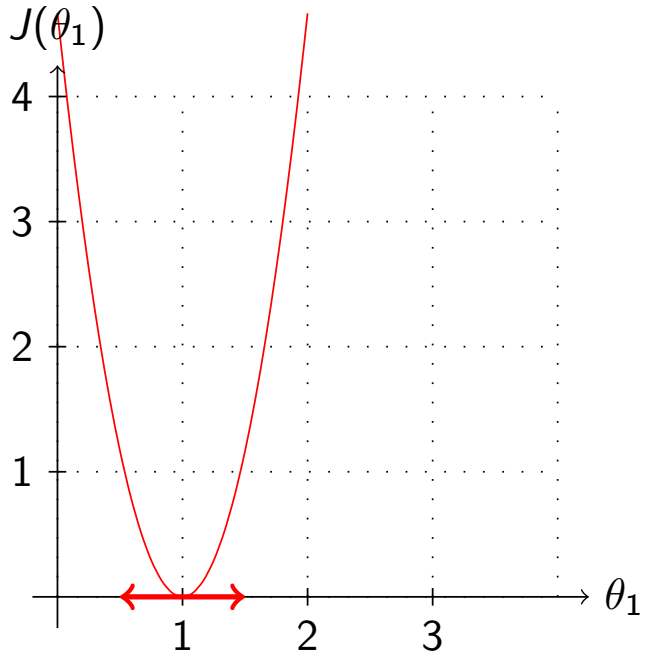
$$\arg\min_{\theta_1} J(\theta_1).$$

Dérivée

$$\begin{aligned}\frac{dJ(\theta_1)}{d\theta_1} &= \frac{2}{m}\sum_{i=1}^{m}\theta_1 x_i^2 - x_i y_i \quad (1)\\ &= 2\left(\theta_1\overline{x^2} - \overline{xy}\right) \quad (2)\end{aligned}$$

Condition nécessaire du minimum

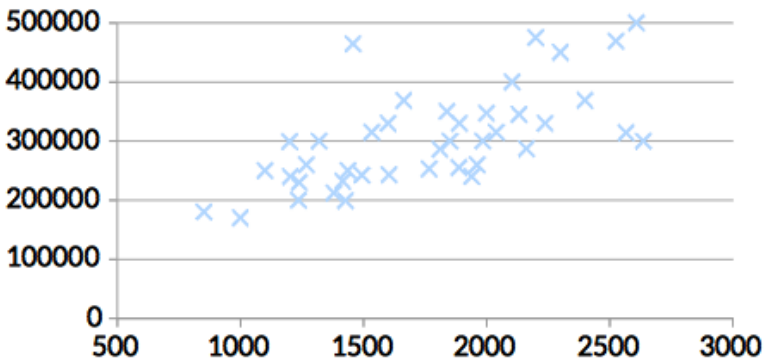
Le minimum de la fonction $J(\theta_1)$ annule la dérivée en θ_1.



$$\Longrightarrow \theta_1 = \frac{\overline{xy}}{\overline{x^2}}$$

Retour à notre exemple introductif

$h_\theta(x)$ pour (θ_0, θ_1) fixé



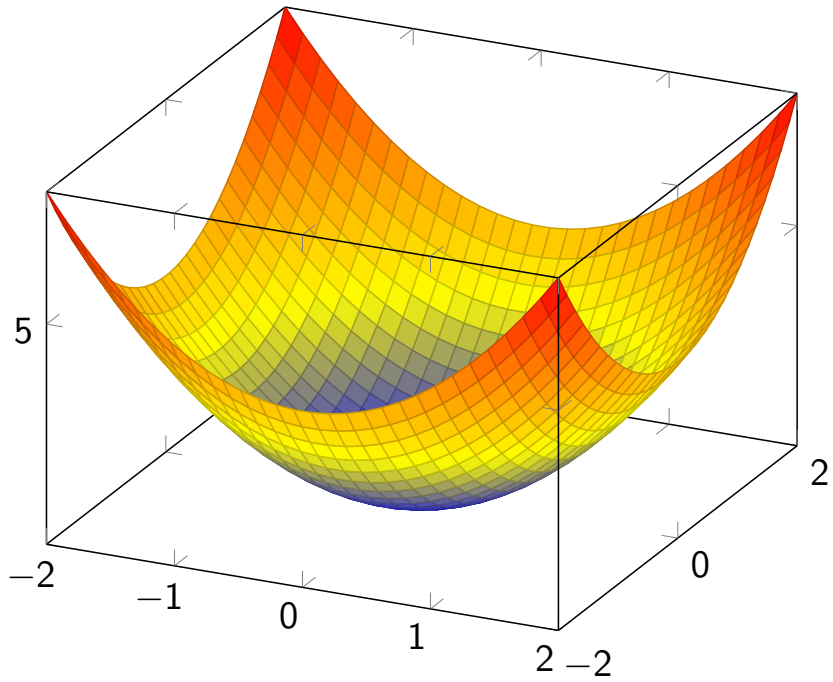
Fonction de coût

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - h_\theta(x_i))^2.$$

Objectif

$$\arg \min_{\theta_0, \theta_1} J(\theta_0, \theta_1).$$

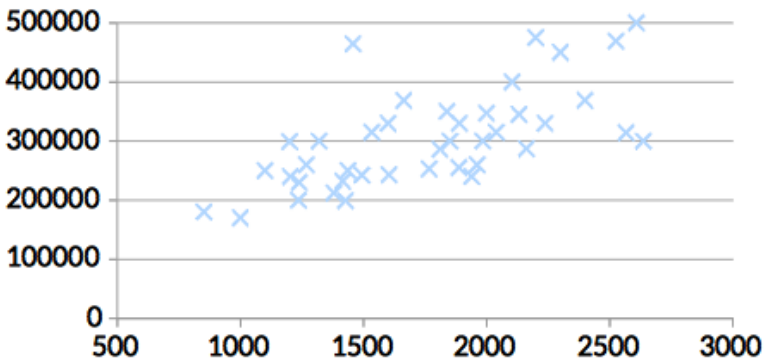
$J(\theta_0, \theta_1)$ fonction de paramètre (θ_0, θ_1)



$\Longrightarrow$ Condition nécessaire pour trouver le minimum : annuler la dérivée de $J(\theta_0, \theta_1)$.

Retour à notre exemple introductif

$h_\theta(x)$ pour (θ_0, θ_1) fixé



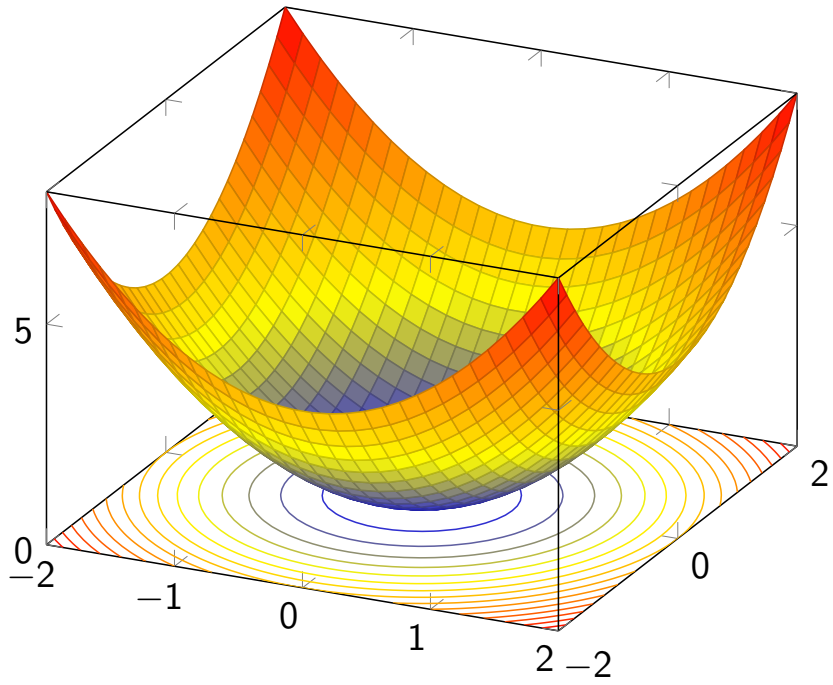
Fonction de coût

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - h_\theta(x_i))^2.$$

Objectif

$$\arg\min_{\theta_0, \theta_1} J(\theta_0, \theta_1).$$

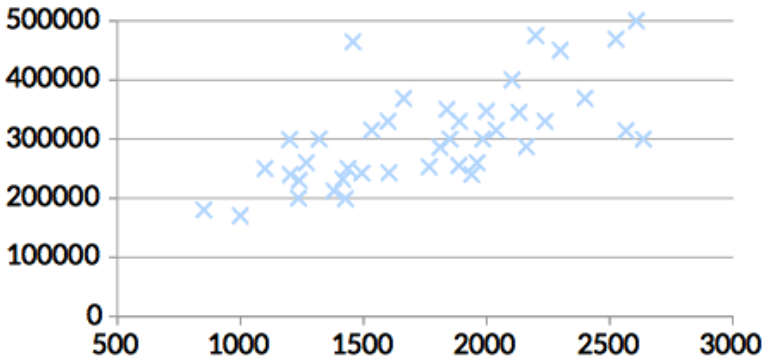
$J(\theta_0, \theta_1)$ fonction de paramètre (θ_0, θ_1)



$\Longrightarrow$ Condition nécessaire pour trouver le minimum : annuler la dérivée de $J(\theta_0, \theta_1)$.

Retour à notre exemple introductif

$h_\theta(x)$ pour (θ_0, θ_1) fixé



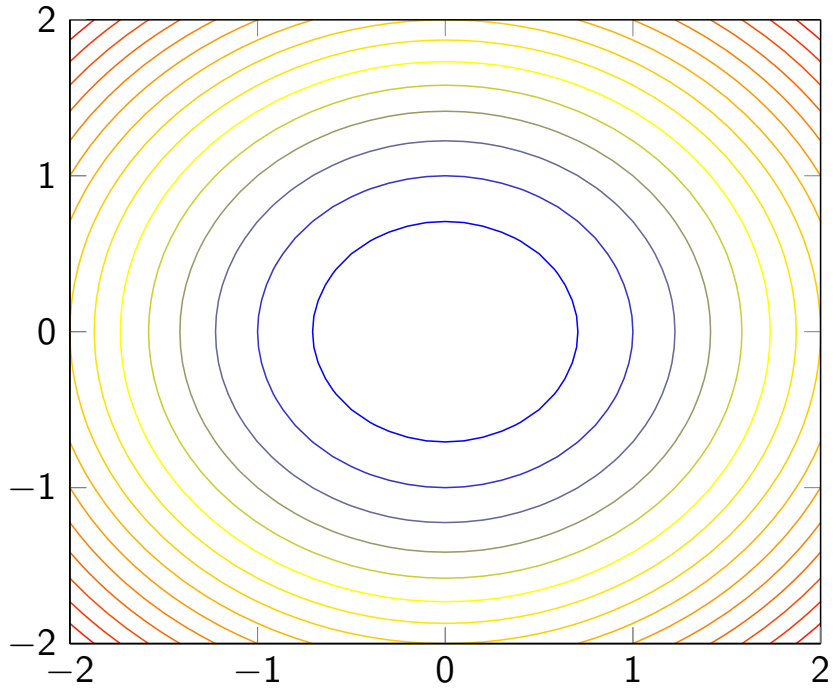
Fonction de coût

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - h_\theta(x_i))^2.$$

Objectif

$$\arg\min_{\theta_0, \theta_1} J(\theta_0, \theta_1).$$

$J(\theta_0, \theta_1)$ fonction de paramètre (θ_0, θ_1)



$\Longrightarrow$ Condition nécessaire pour trouver le minimum : annuler la dérivée de $J(\theta_0, \theta_1)$.

Minimisation par moindre carré dans le cas 1D

Formulation

- Nous nous restreignons, dans un premier temps, au cas à une seule dimension :

$$\theta^\star = \arg\min_\theta J(\theta) = \arg\min_\theta \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \theta_0 - \theta_1 x_i\right)^2 .$$

Condition d'optimalité

Une condition nécessaire pour minimiser J est d'annuler les dérivées de J suivantes θ_0 et θ_1. Cette condition est suffisante si la fonction J est convexe.

Annulation de la dérivée de J suivant θ_0

$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0} = \frac{\partial}{\partial \theta_0} \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \theta_0 - \theta_1 x_i\right)^2, \tag{3}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial \theta_0} \left(y_i - \theta_0 - \theta_1 x_i\right)^2, \tag{4}$$

$$= \frac{1}{m} \sum_{i=1}^{m} 2 \left(y_i - \theta_0 - \theta_1 x_i\right) \cdot (-1) = 0, \tag{5}$$

$$\Longrightarrow \frac{1}{m} \sum_{i=1}^{m} 2 \left(y_i - \theta_0 - \theta_1 x_i\right) = 0. \tag{6}$$

Annulation de la dérivée de J suivant θ_0

$$\frac{\partial J(\theta_0,\theta_1)}{\partial \theta_0} = \frac{\partial}{\partial \theta_0}\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\theta_0-\theta_1 x_i\right)^2, \tag{3}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial \theta_0}\left(y_i-\theta_0-\theta_1 x_i\right)^2, \tag{4}$$

$$= \frac{1}{m}\sum_{i=1}^{m}2\left(y_i-\theta_0-\theta_1 x_i\right)\cdot(-1)=0, \tag{5}$$

$$\Longrightarrow \frac{1}{m}\sum_{i=1}^{m}2\left(y_i-\theta_0-\theta_1 x_i\right)=0. \tag{6}$$

Annulation de la dérivée de J suivant θ_1

$$\frac{\partial J(\theta_0,\theta_1)}{\partial \theta_1} = \frac{-2}{m}\sum_{i=1}^{m}\left(y_i-\theta_0-\theta_1 x_i\right)x_i = 0.$$

Conditions d'optimalités

On obtient donc les conditions d'optimalités suivantes :

$$\frac{1}{m}\sum_{i=1}^{m}(y_i - \theta_0 - \theta_1 x_i) = 0, \qquad \frac{1}{m}\sum_{i=1}^{m}(y_i - \theta_0 - \theta_1 x_i)\, x_i = 0.$$

Remarques

Soit $\epsilon_i = y_i - \theta_0 - \theta_1 x_i$, on a :

- Les erreurs de prédiction ont une moyenne nulle

$$\frac{1}{m}\sum_{i=1}^{m}\epsilon_i = 0.$$

- Les erreurs de prédiction ne sont pas corrélées avec les données

$$\frac{1}{m}\sum_{i=1}^{m}\epsilon_i x_i = 0.$$

Calcul de la solution

Calcul de θ_0

$$\frac{1}{m}\sum_{i=1}^{m}(y_i - \theta_0 - \theta_1 x_i) = 0 \quad \Longrightarrow \quad \theta_0 = \frac{1}{m}\sum_{i=1}^{m}(y_i - \theta_1 x_i) \tag{7}$$

$$\Longrightarrow \quad \theta_0 = \frac{1}{m}\sum_{i=1}^{m} y_i - \theta_1 \frac{1}{m}\sum_{i=1}^{m} x_i \tag{8}$$

$$\Longrightarrow \quad \theta_0 = \bar{y} - \theta_1 \bar{x}. \tag{9}$$

Calcul de la solution

Calcul de θ_0

$$\theta_0 = \bar{y} - \theta_1 \bar{x}. \tag{10}$$

Calcul de θ_1

$$\frac{1}{m}\sum_{i=1}^{m}(y_i - \theta_0 - \theta_1 x_i)\, x_i = 0 \quad \Longrightarrow \quad \frac{1}{m}\sum_{i=1}^{m}(y_i - (\bar{y} - \theta_1 \bar{x}) - \theta_1 x_i)\, x_i = 0 \tag{11}$$

$$\Longrightarrow \quad \frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y} - \theta_1(-\bar{x} + x_i))\, x_i = 0 \tag{12}$$

$$\Longrightarrow \quad \theta_1 = \frac{\frac{1}{m}\sum_{i=1}^{m}(y_i - \bar{y})\, x_i}{\frac{1}{m}\sum_{i=1}^{m}(-\bar{x} + x_i)\, x_i} \tag{13}$$

$$\Longrightarrow \quad \theta_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{cov(x, y)}{var(x)}. \tag{14}$$

Calcul de la solution

Calcul de θ_0

$$\theta_0 = \bar{y} - \theta_1 \bar{x}. \tag{15}$$

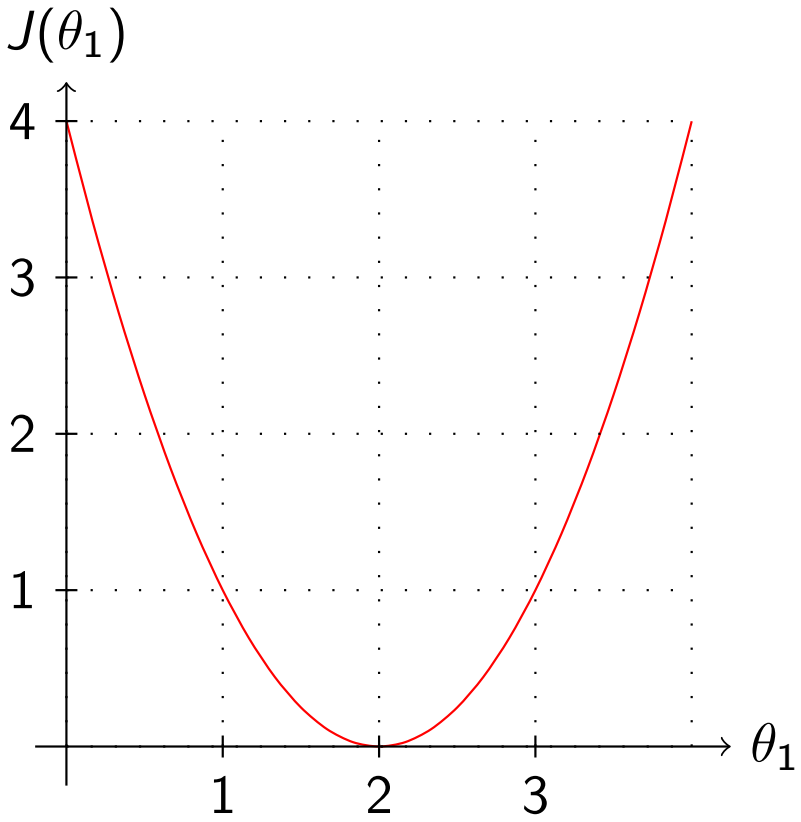
Calcul de θ_1

$$\theta_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}. \tag{16}$$

Méthode de résolution alternative (cas simplifié)

Idée

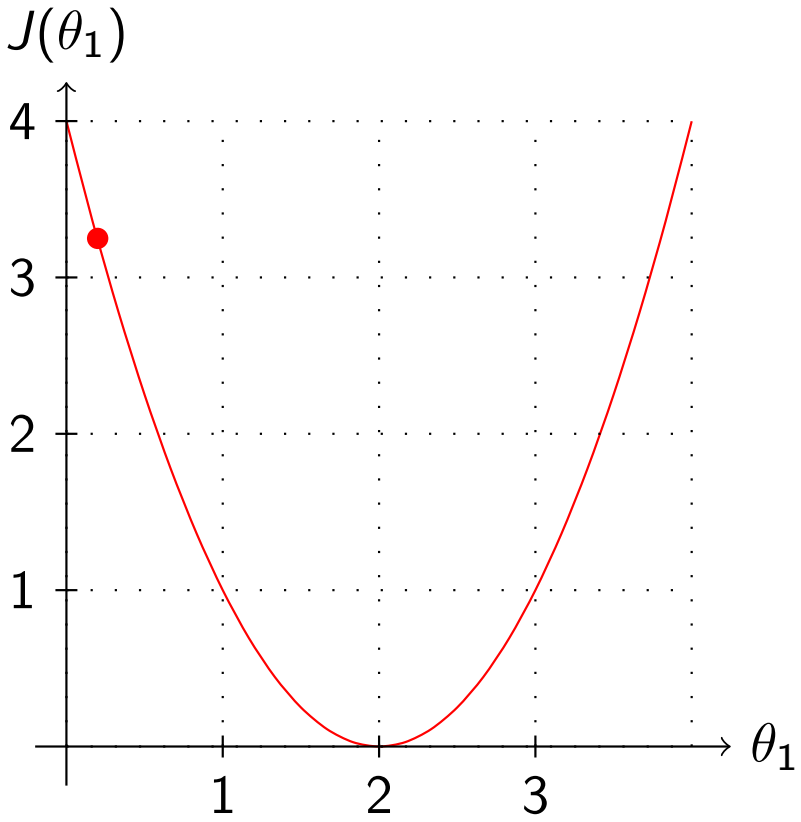
On peut atteindre le minimum en se dirigeant itérativement dans la direction de plus forte pente.



Méthode de résolution alternative (cas simplifié)

Idée

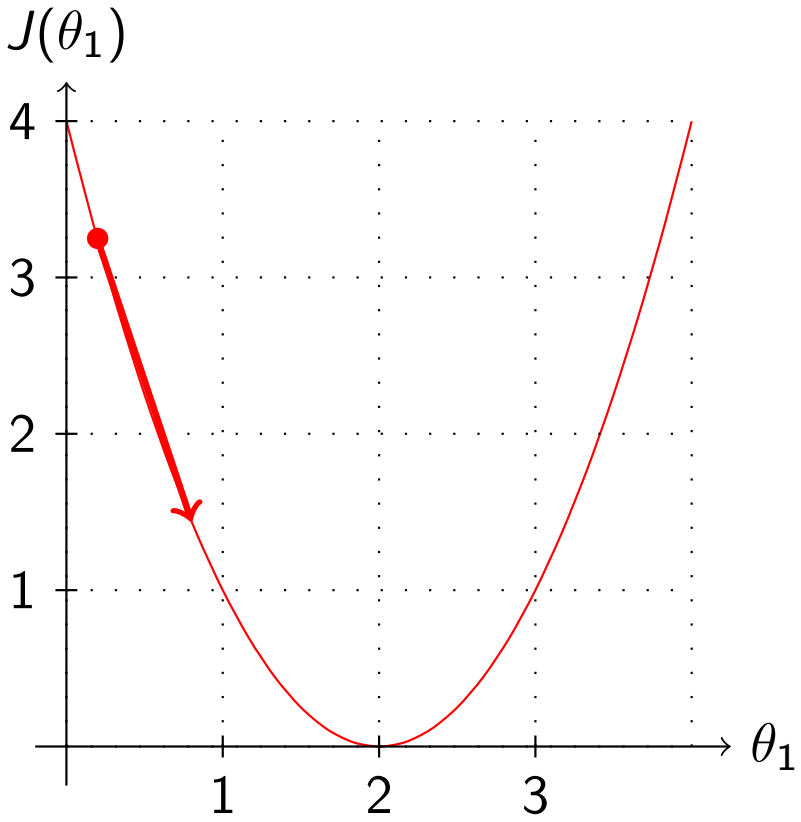
On peut atteindre le minimum en se dirigeant itérativement dans la direction de plus forte pente.



Méthode de résolution alternative (cas simplifié)

Idée

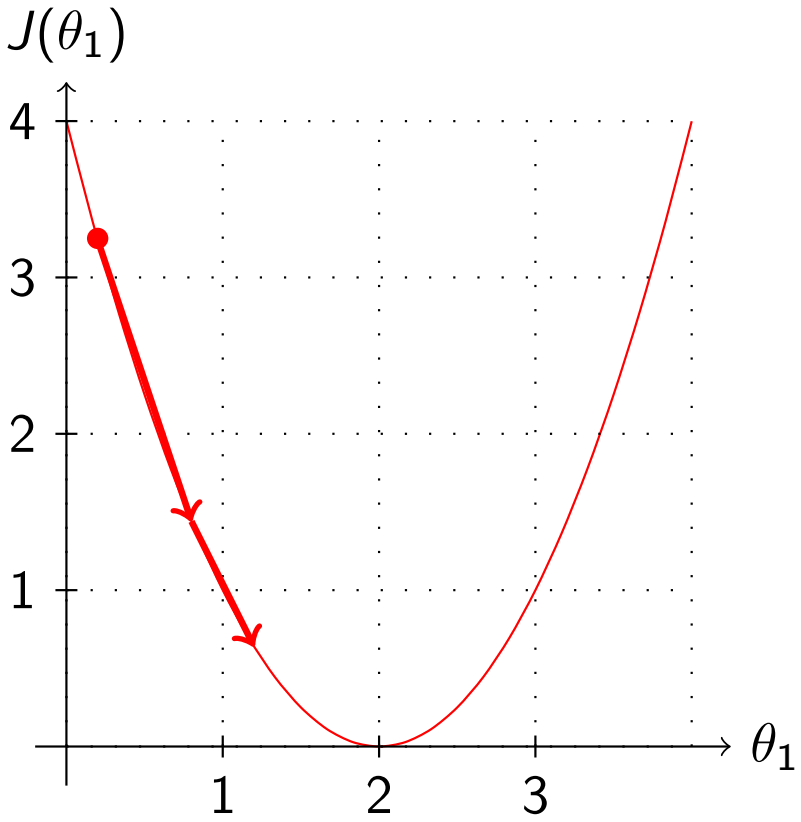
On peut atteindre le minimum en se dirigeant itérativement dans la direction de plus forte pente.



Méthode de résolution alternative (cas simplifié)

Idée

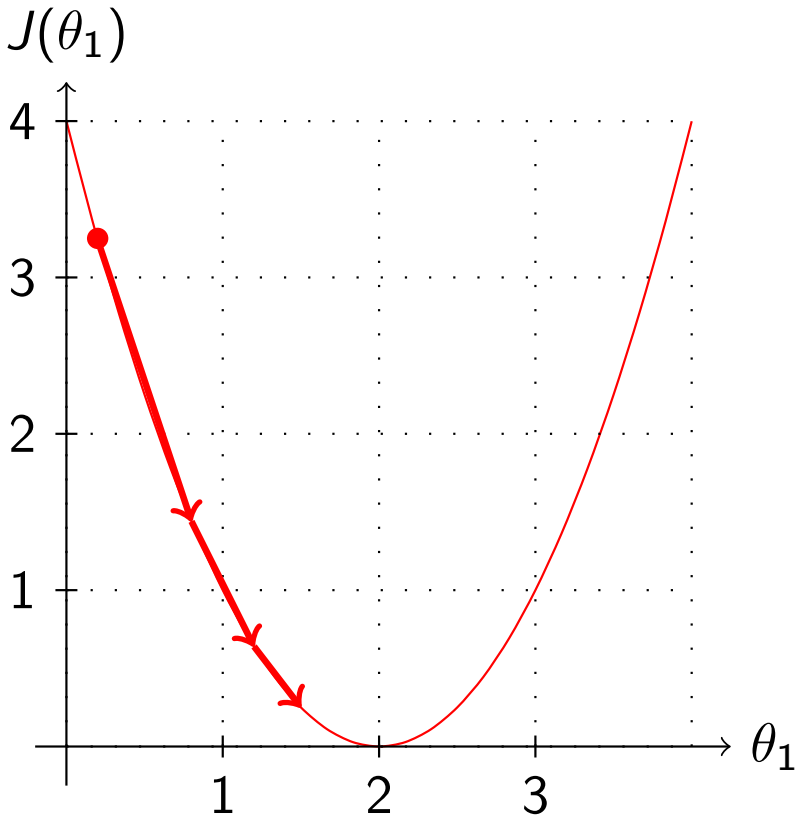
On peut atteindre le minimum en se dirigeant itérativement dans la direction de plus forte pente.



Méthode de résolution alternative (cas simplifié)

Idée

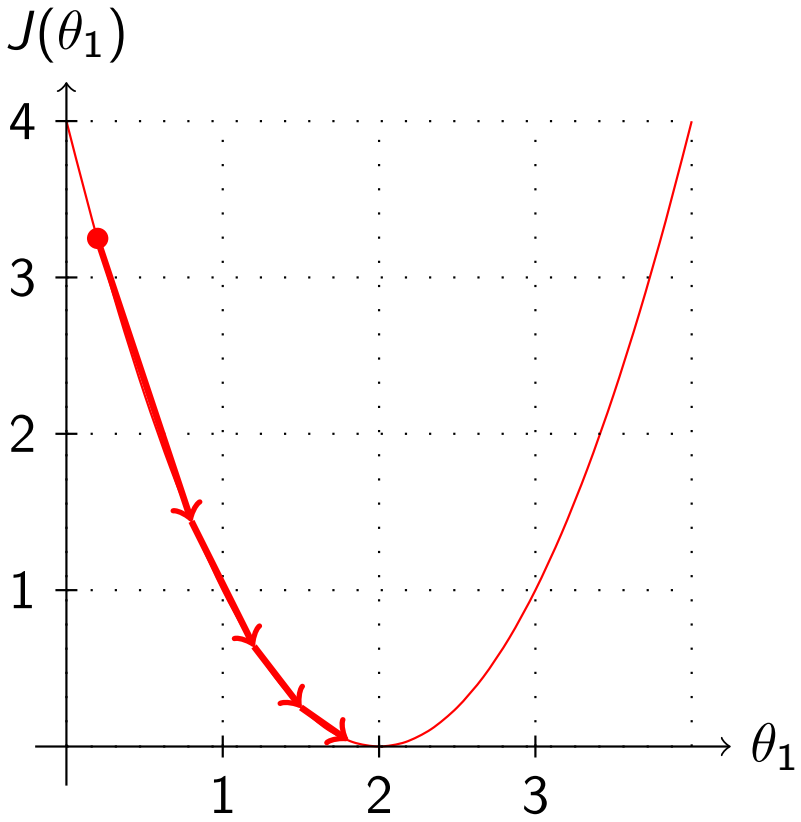
On peut atteindre le minimum en se dirigeant itérativement dans la direction de plus forte pente.



Méthode de résolution alternative (cas simplifié)

Idée

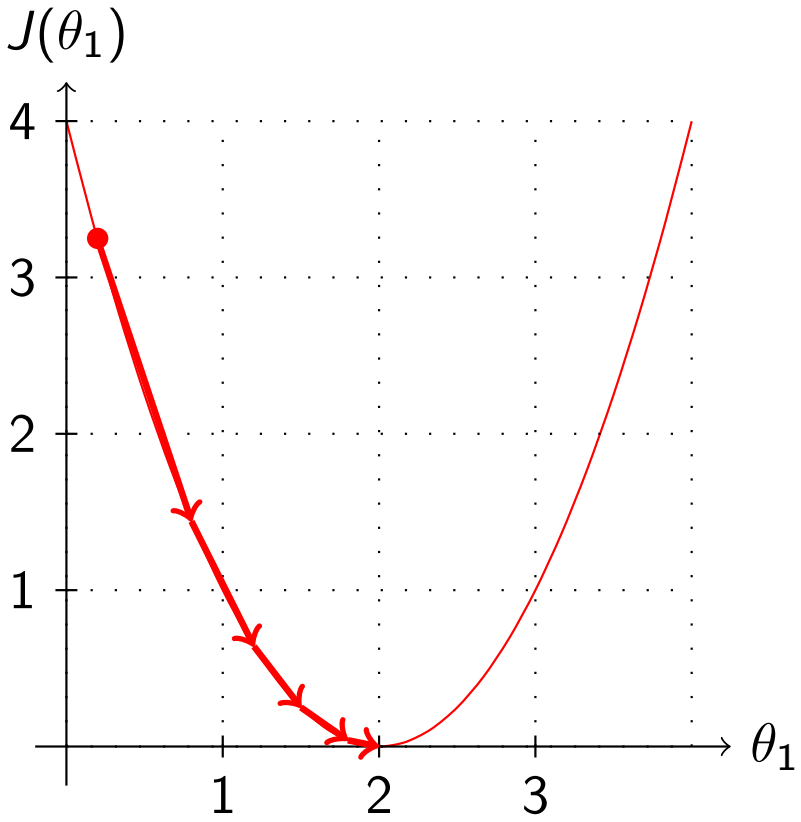
On peut atteindre le minimum en se dirigeant itérativement dans la direction de plus forte pente.



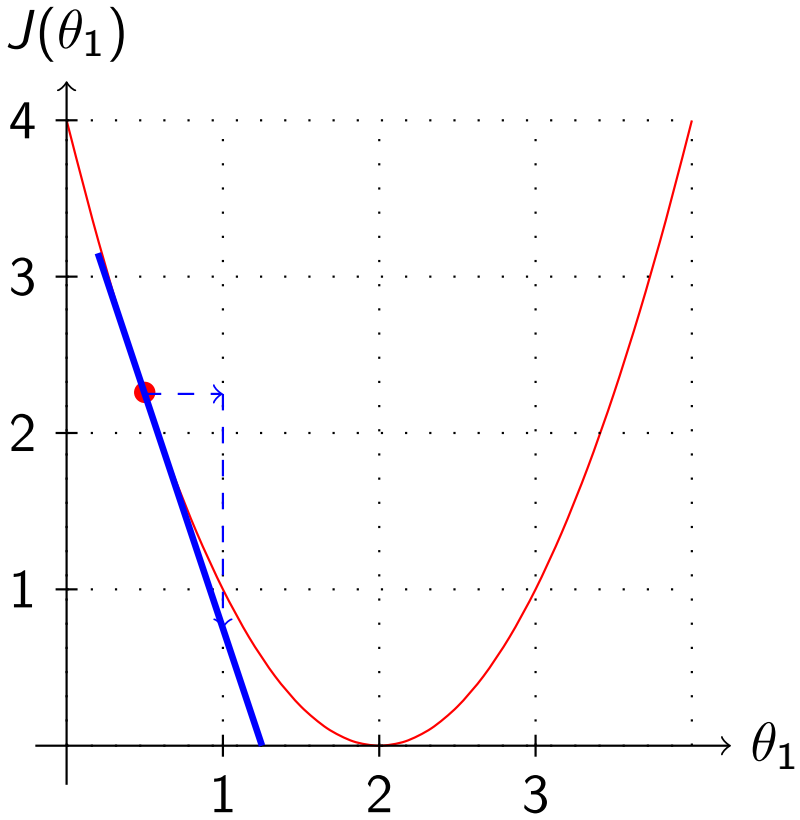
Méthode de résolution alternative (cas simplifié)

Idée

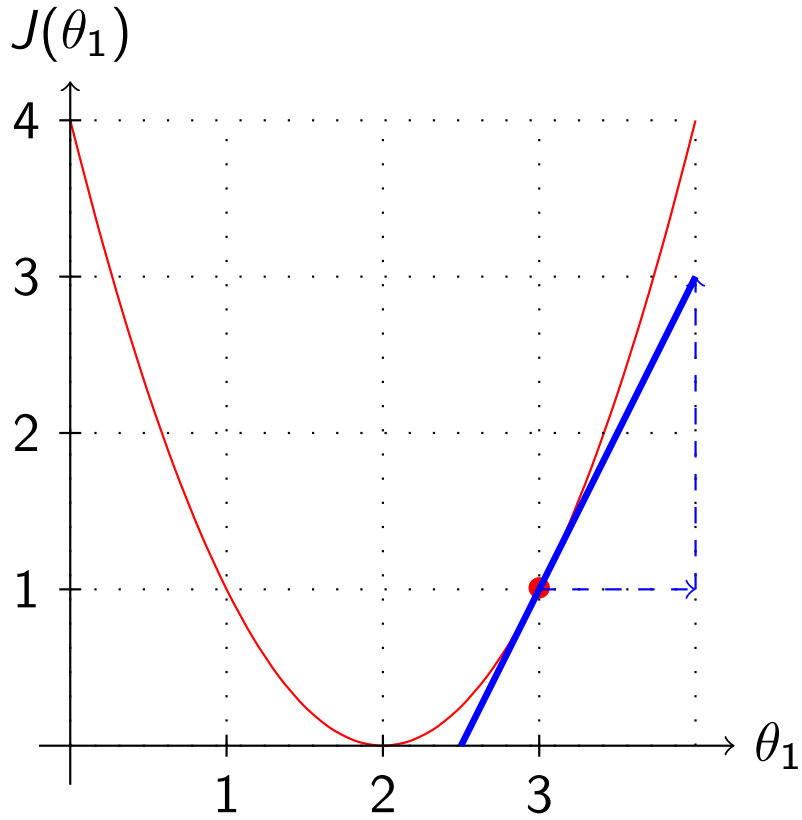
On peut atteindre le minimum en se dirigeant itérativement dans la direction de plus forte pente.



Direction de la pente.

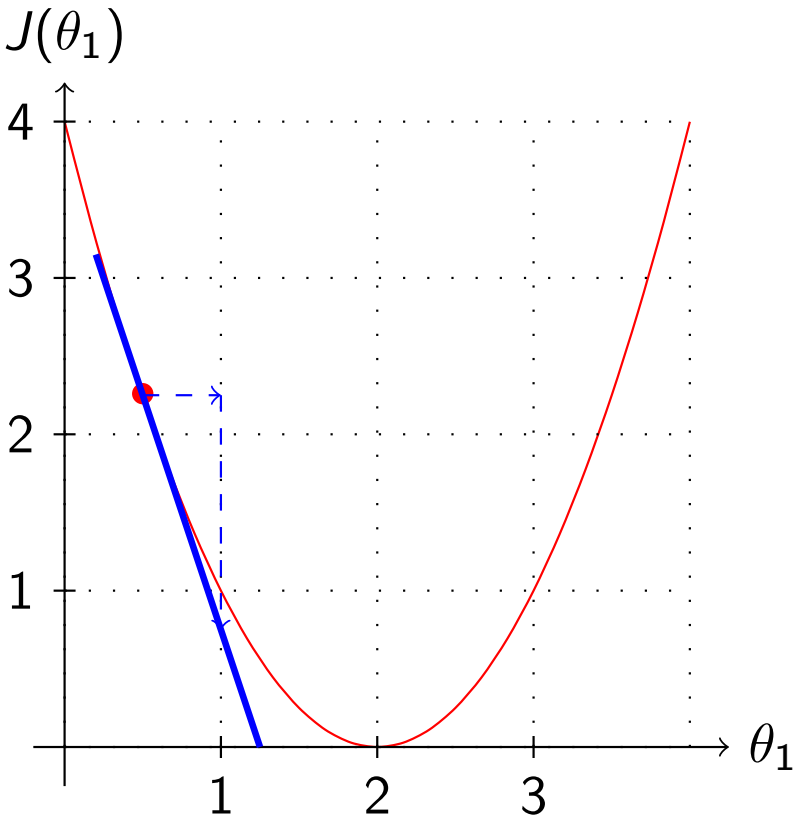


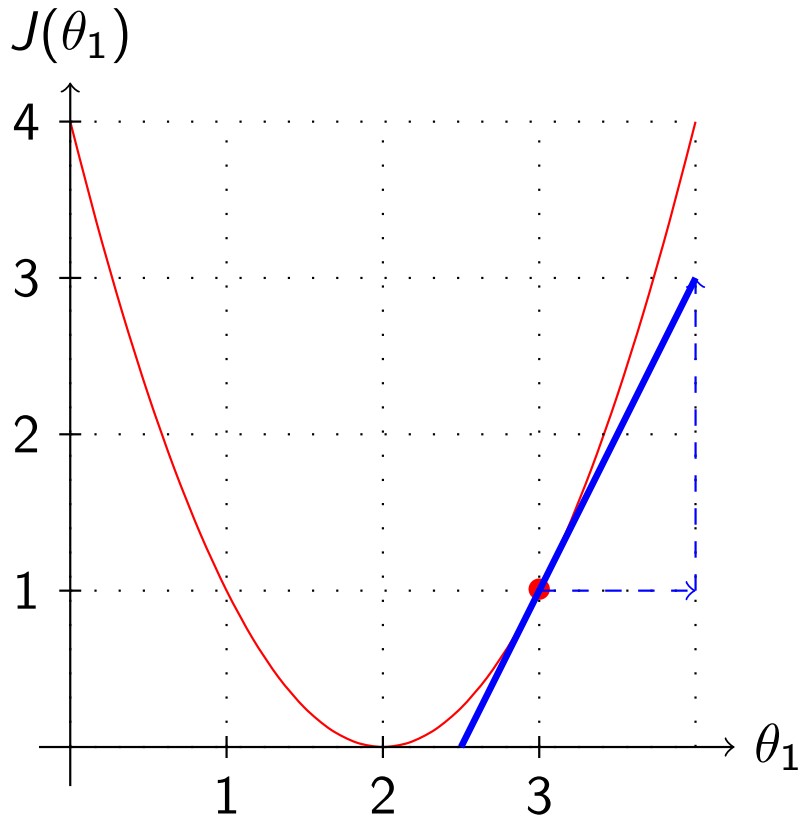
Valeur de la dérivée en 0.5 : -3
Direction de la pente : positive.



Valeur de la dérivée en 3 : 2
Direction de la pente : négative.

Direction de la pente.





Comment trouver la direction de la pente ?

Le signe opposé de la dérivée nous donne la direction de la pente.

Résolution par descente de gradient du cas simplifié

Problème d'optimisation à résoudre

$$\arg\min_{\theta_1} J(\theta_1).$$

Solution par descente de gradient

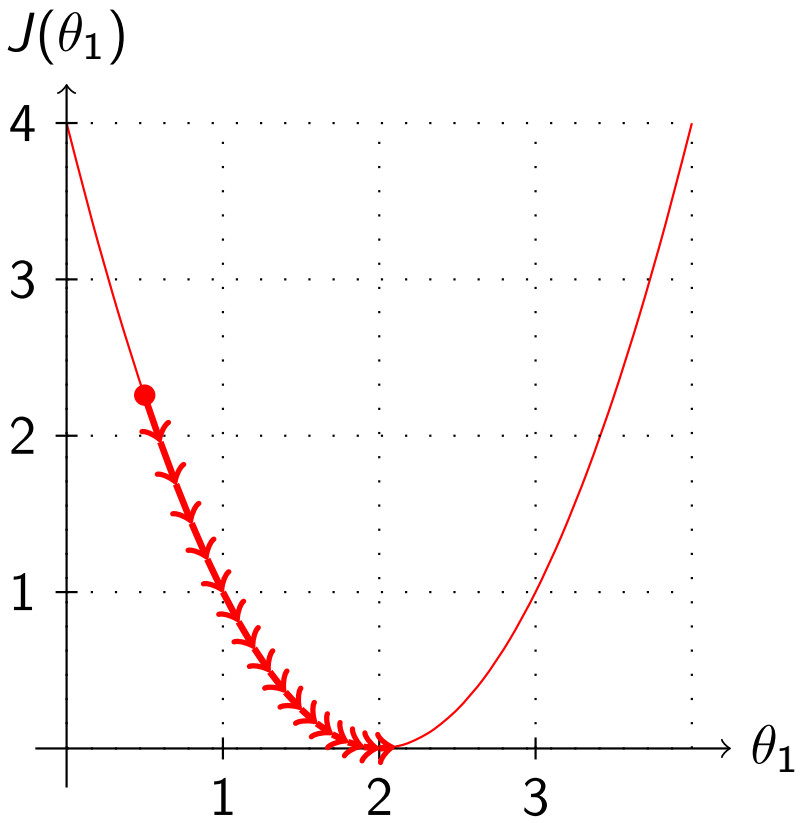
Initialiser avec un θ_1 choisi aléatoirement.
Répéter jusqu'à convergence :

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_i} J(\theta_1).$$

α est une constante correspondant au taux d'apprentissage.
Lorsque l'on s'approche du minimum la dérivée tend vers 0, en conséquence les pas sont de plus en plus petit. Il n'est pas nécessaire d'avoir une valeur de α qui décroit.

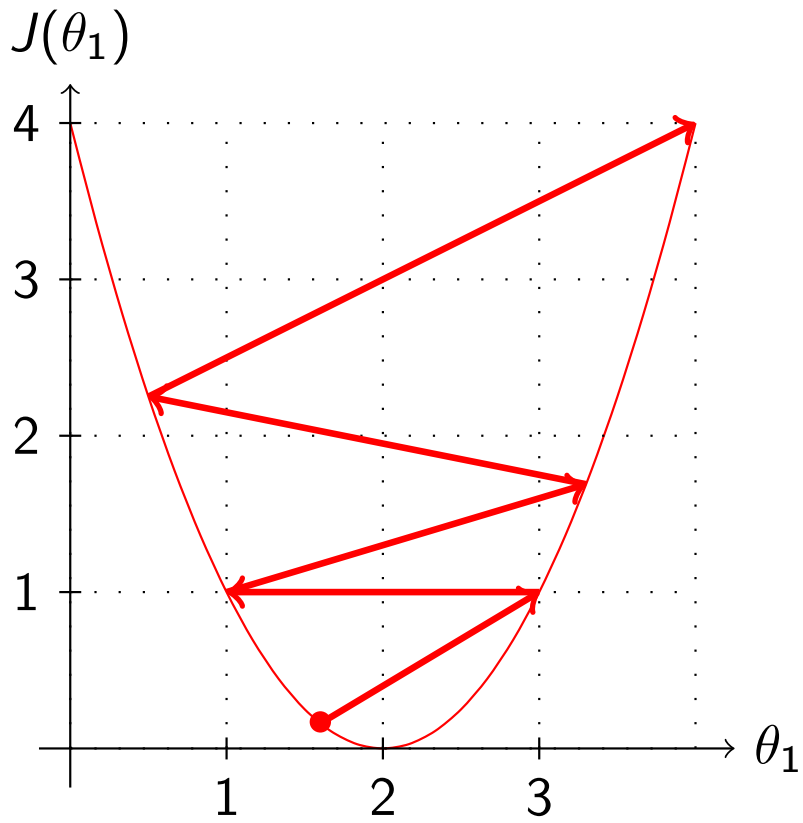
Bien choisir la valeur de α

Valeur de α trop petite



La convergence est très lente.

Valeur de α trop grande



L'algorithme diverge.

Résolution par descente de gradient du cas 1D

Problème d'optimisation à résoudre

$$\arg \min_{\theta_0, \theta_1} J(\theta_0, \theta_1).$$

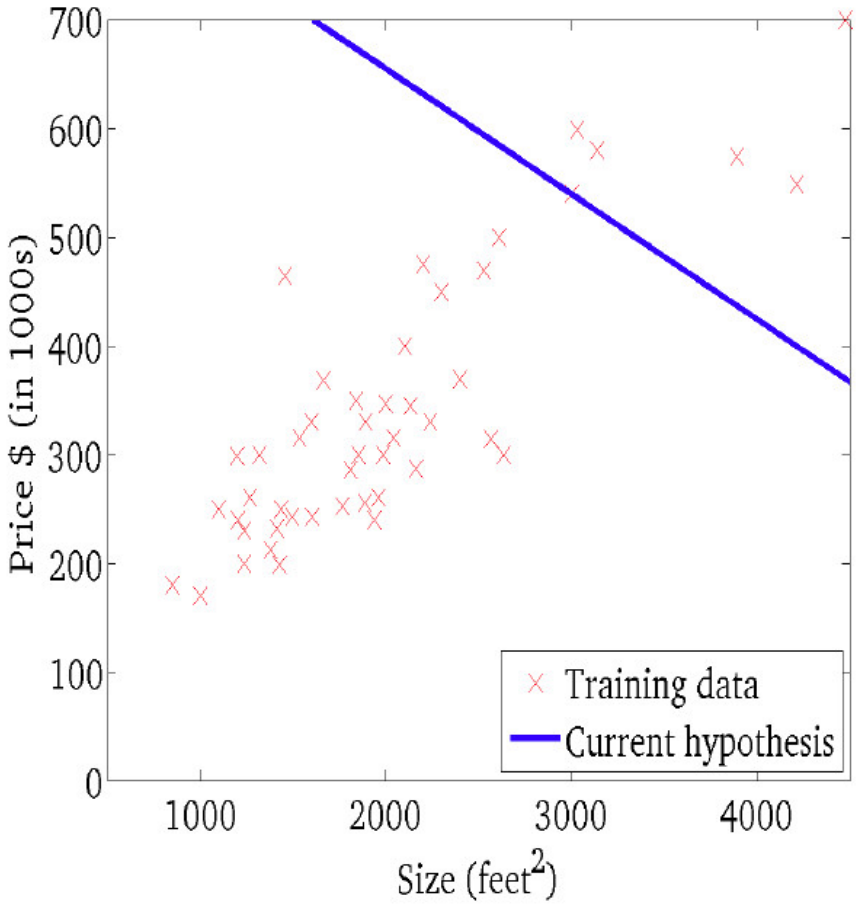
Solution par descente de gradient

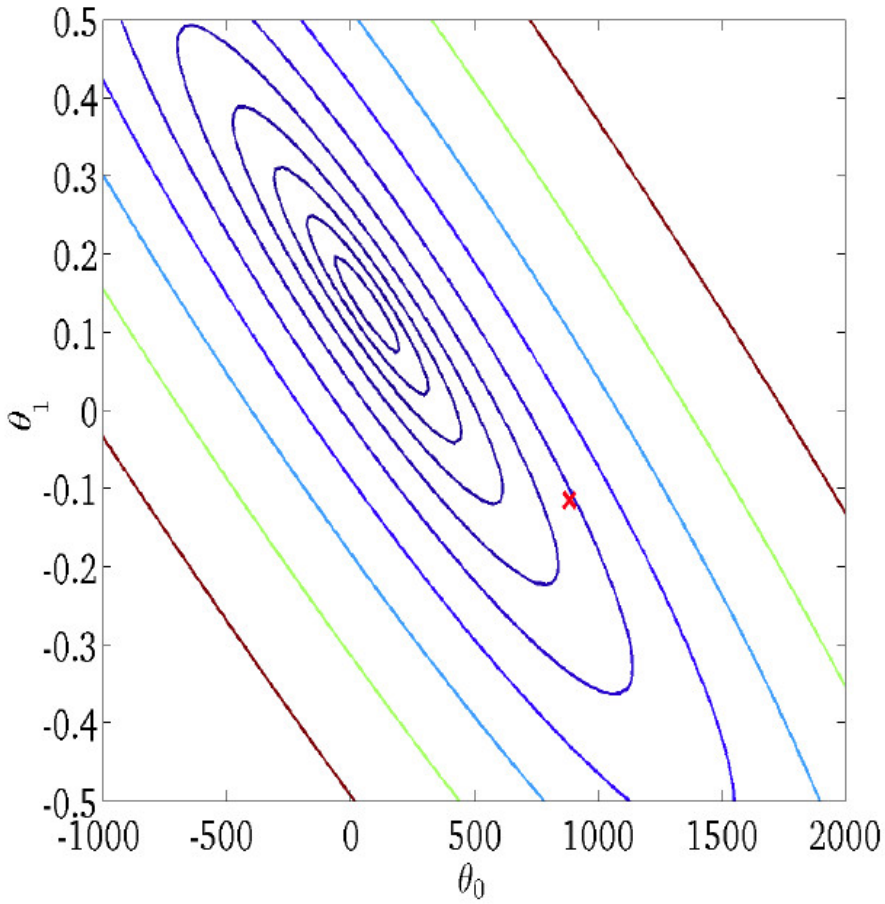
Initialiser avec θ_0, θ_1 choisies aléatoirement.
Répéter jusqu'à convergence :

$$\theta_i := \theta_i - \alpha \frac{\partial}{\partial \theta_i} J(\theta_i) \text{ avec } i \in \{0, 1\}.$$

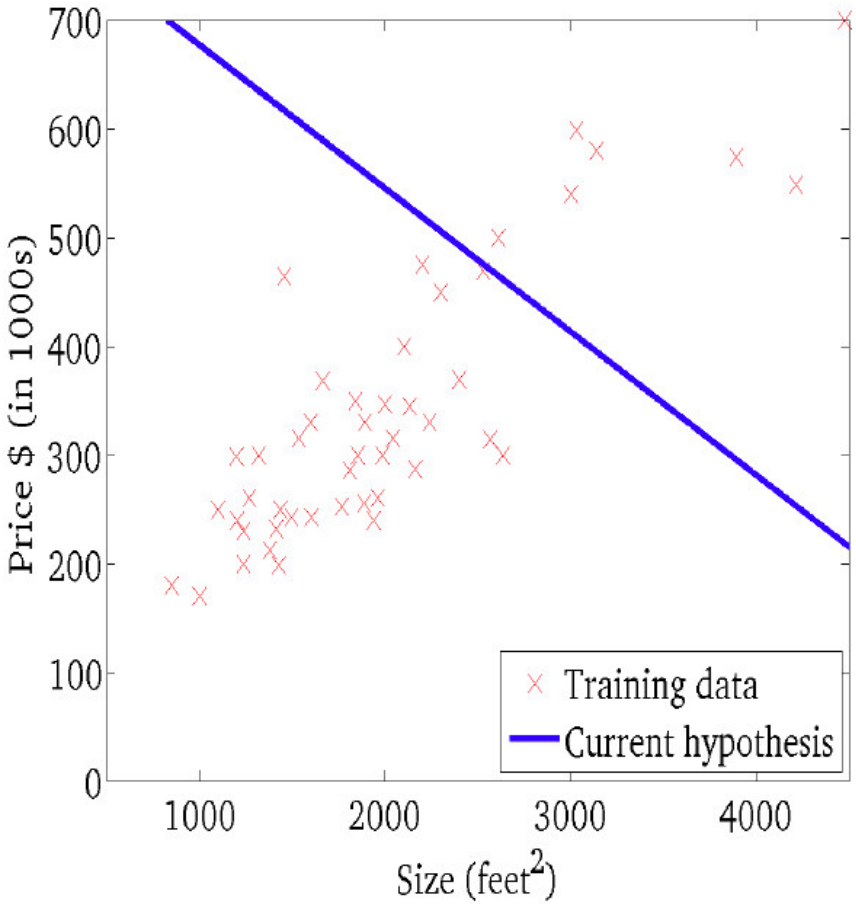
α est une constante correspondant au taux d'apprentissage.
Important : θ_0 et θ_1 sont à mettre à jours simultanément.

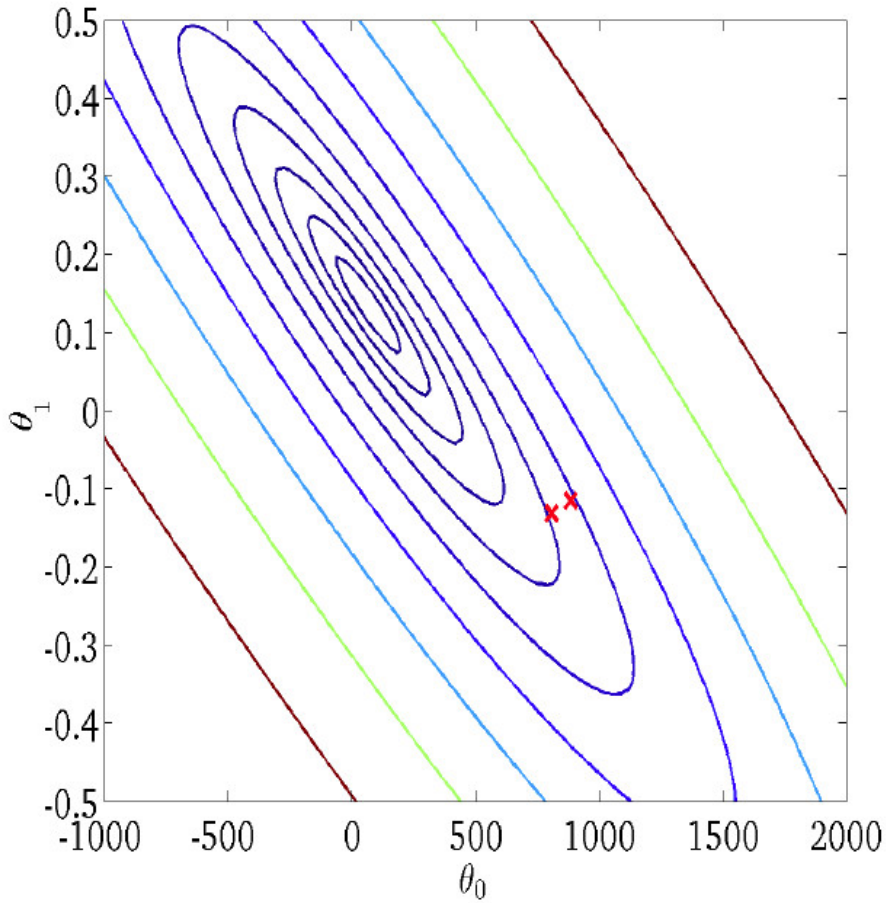
Application de la descente de gradient à notre exemple



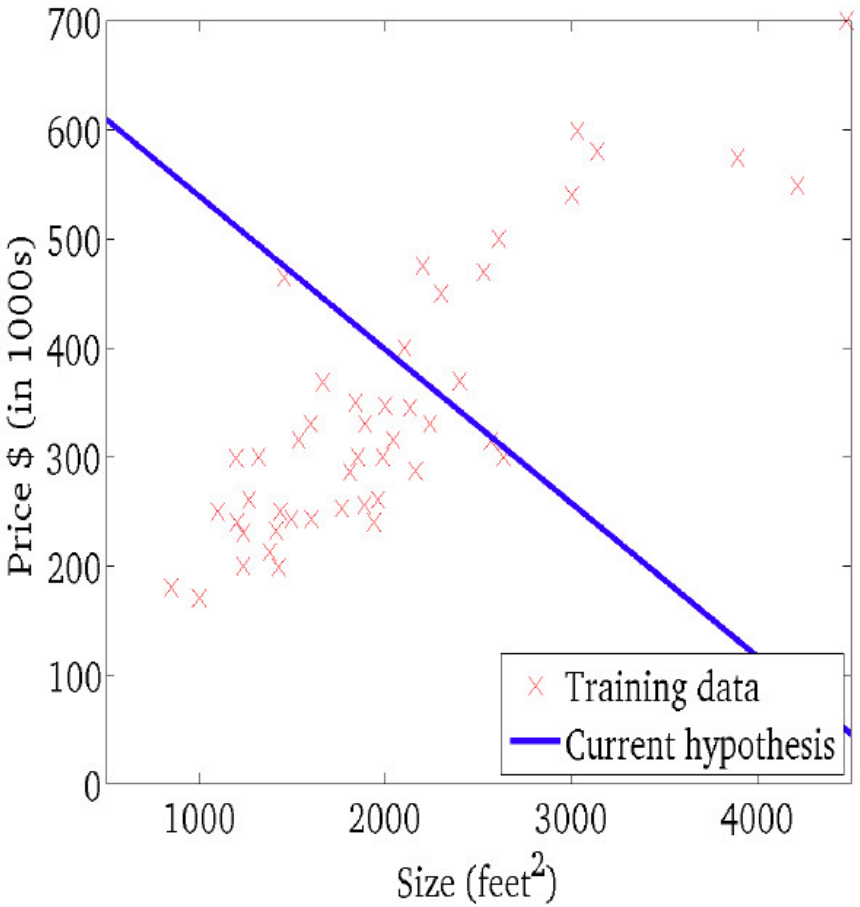


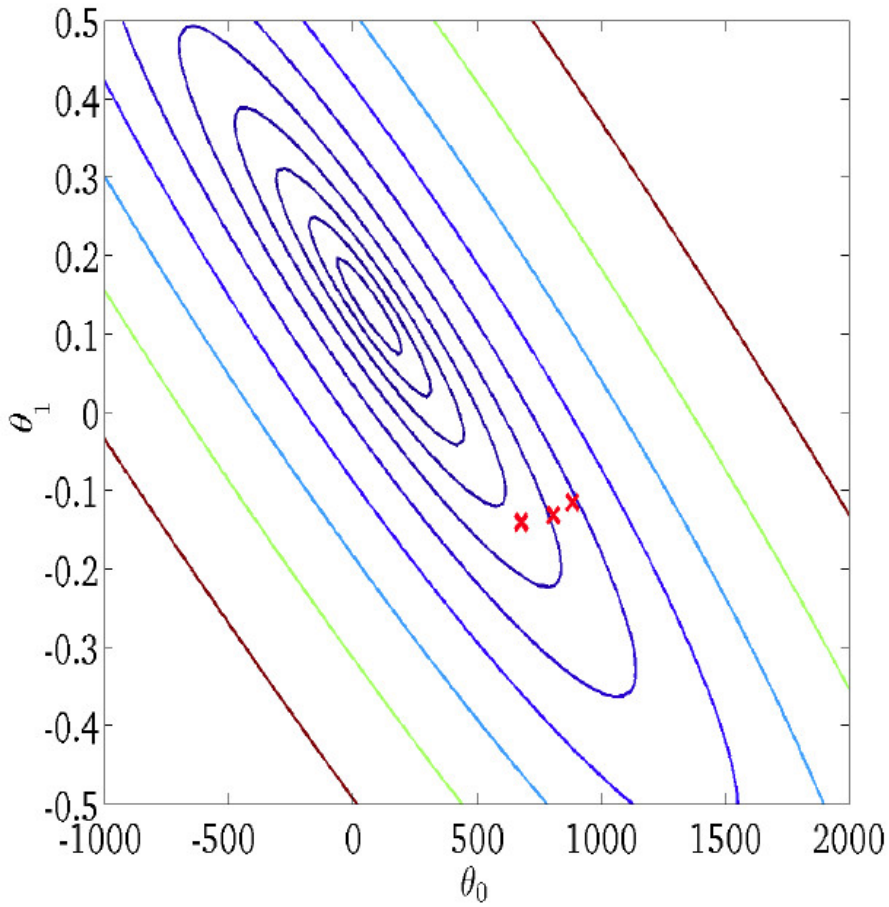
Application de la descente de gradient à notre exemple



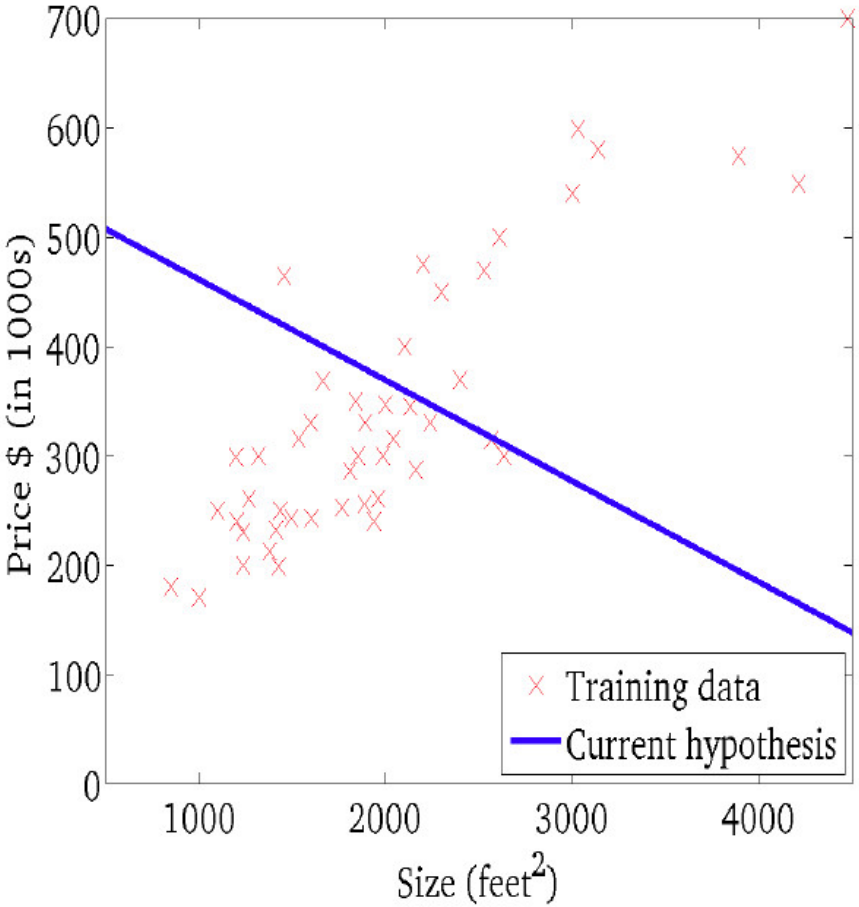


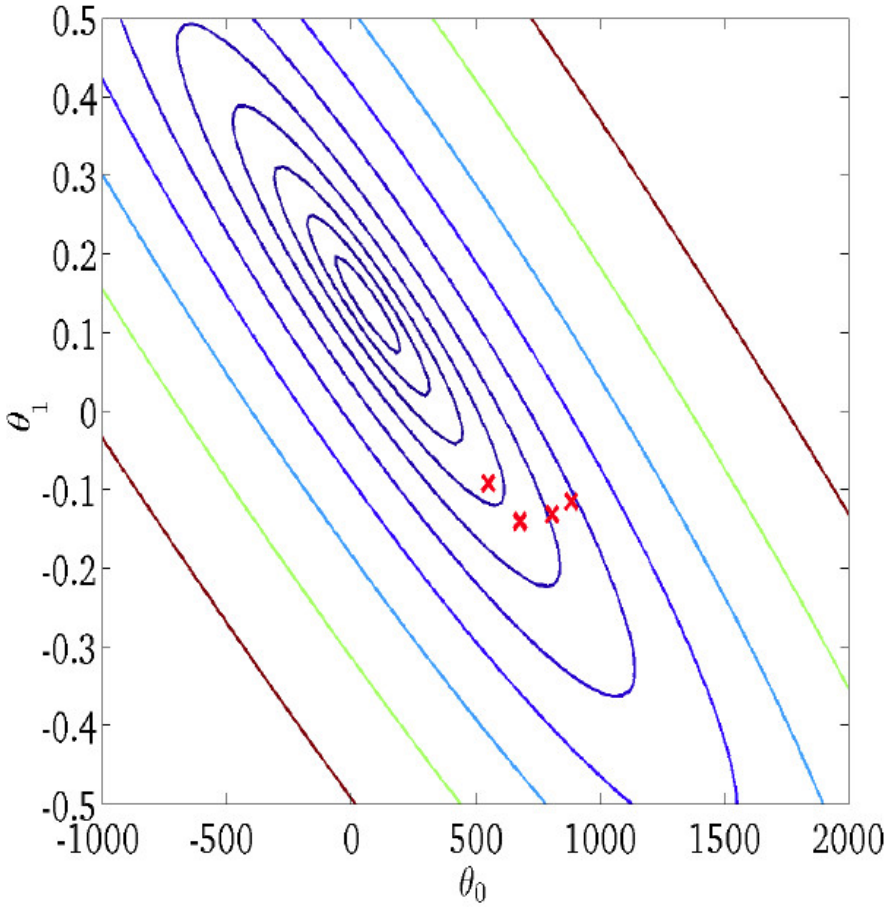
Application de la descente de gradient à notre exemple



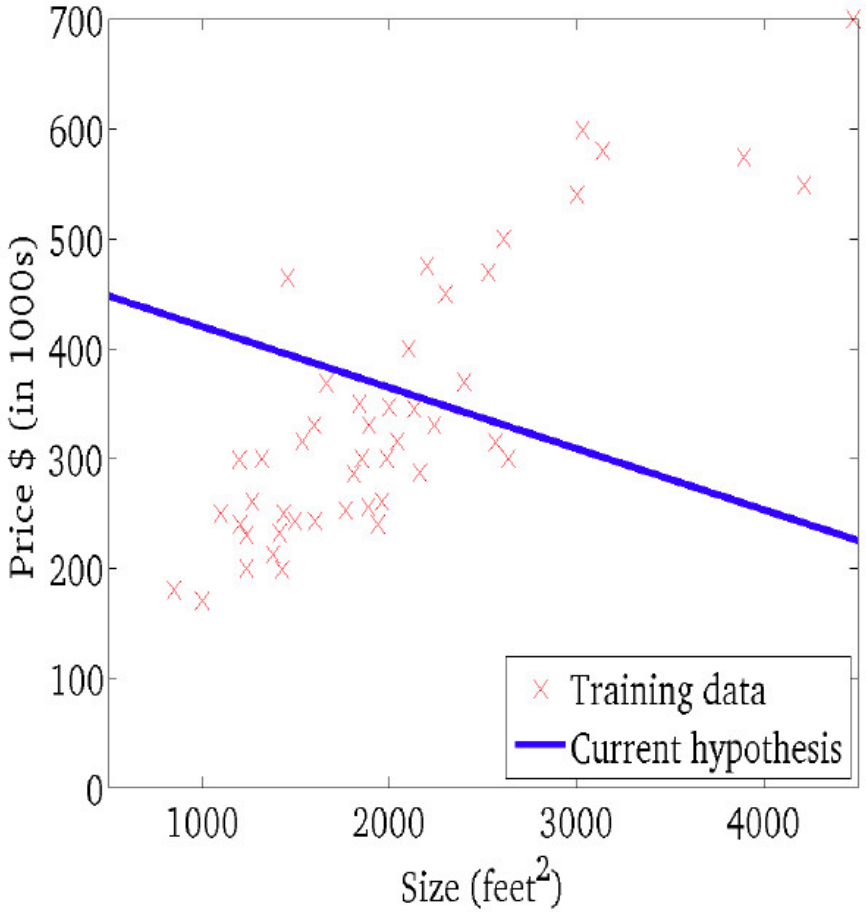


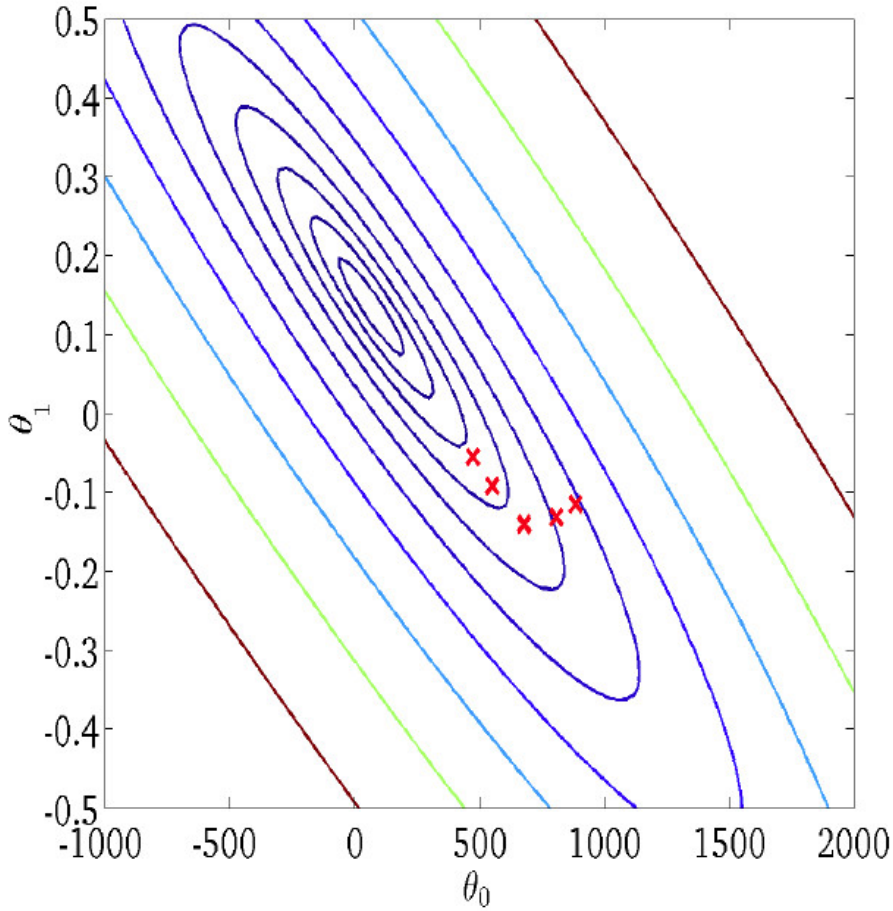
Application de la descente de gradient à notre exemple



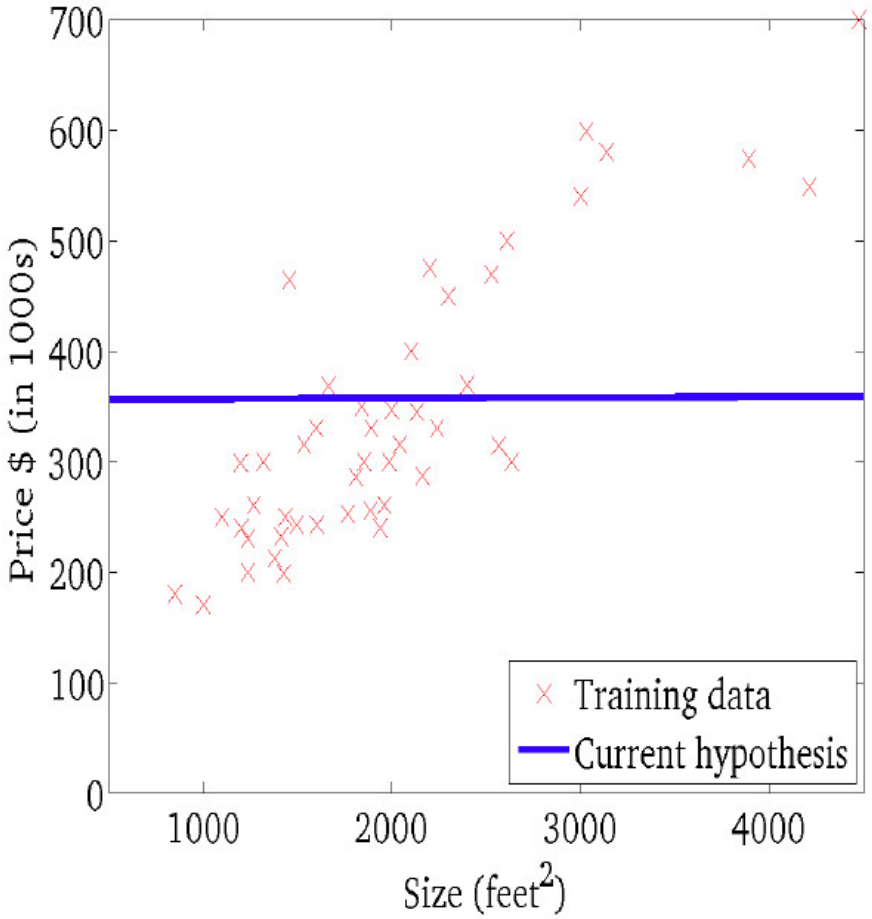


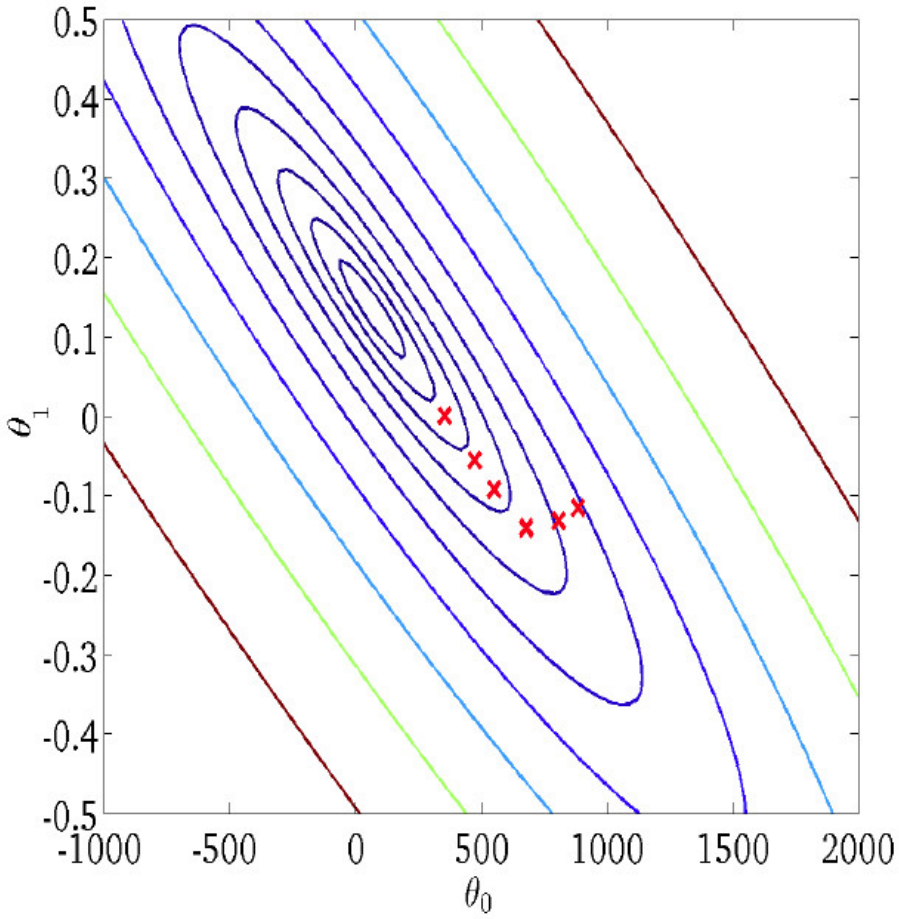
Application de la descente de gradient à notre exemple



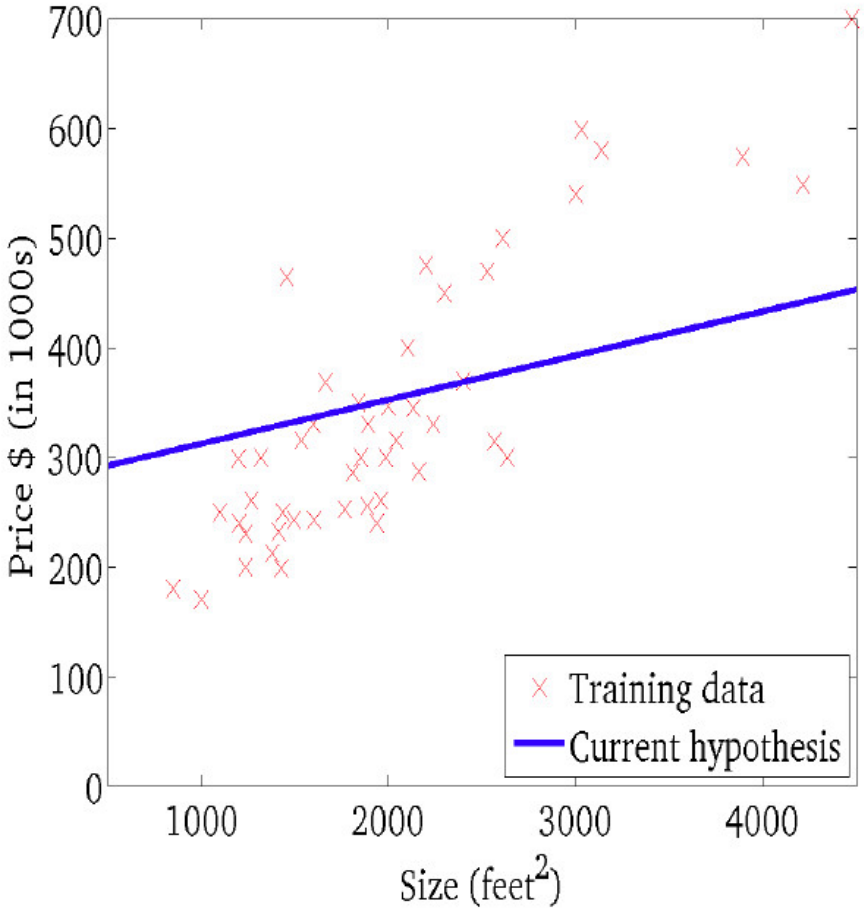


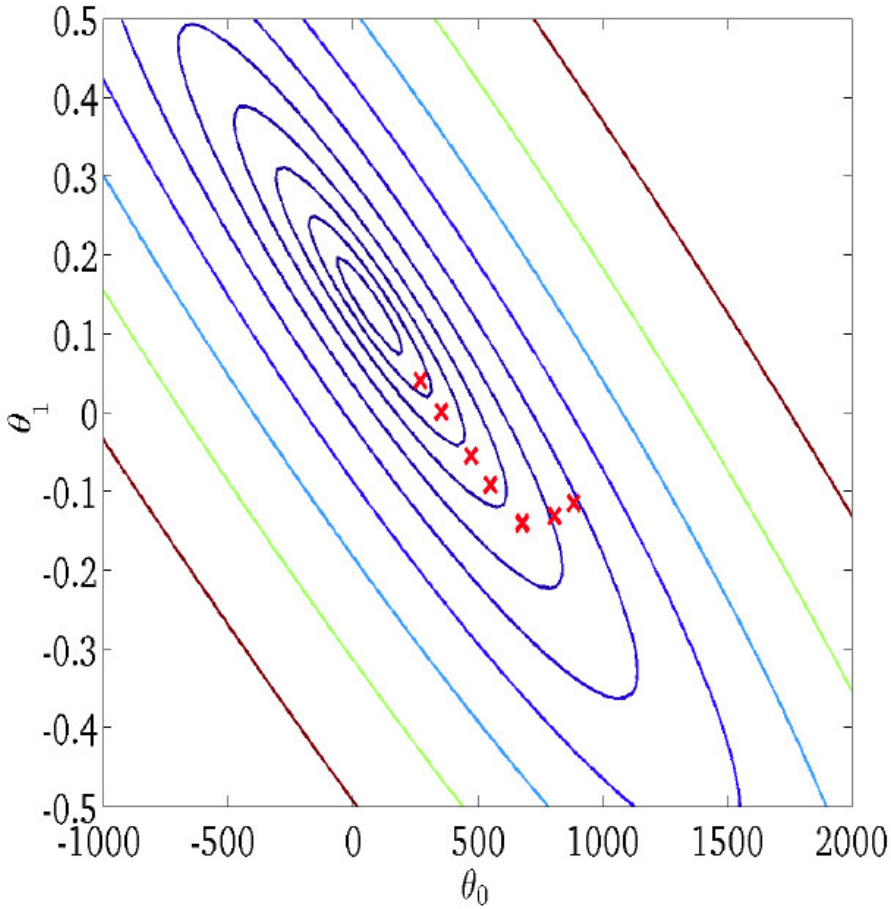
Application de la descente de gradient à notre exemple



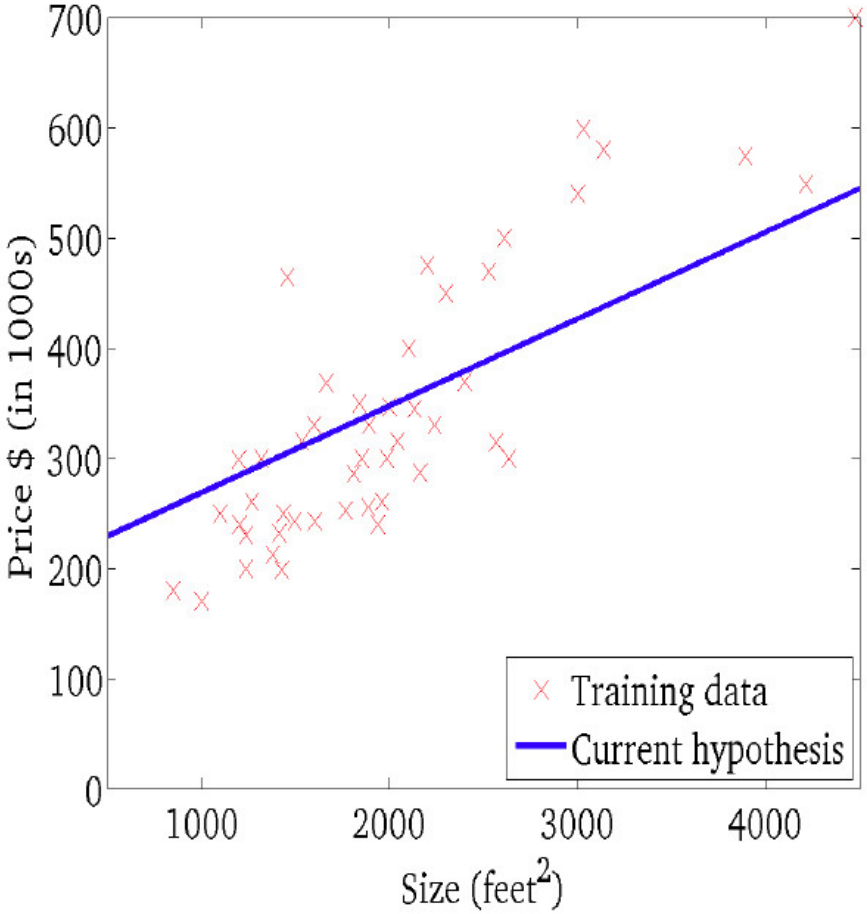


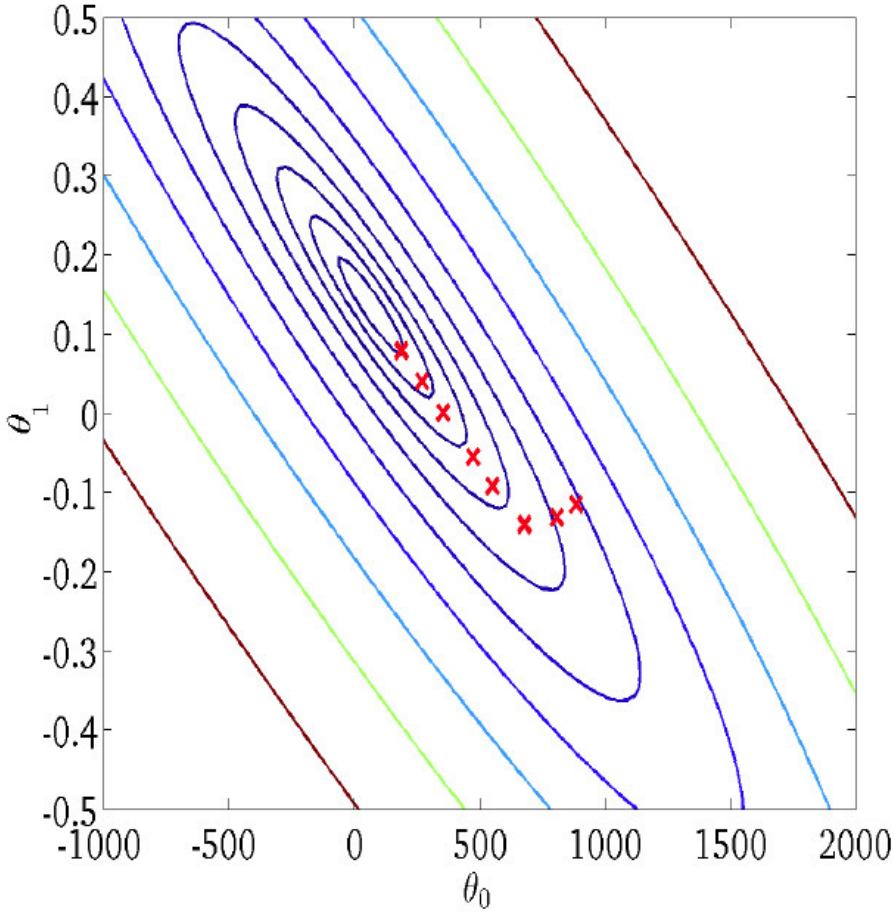
Application de la descente de gradient à notre exemple



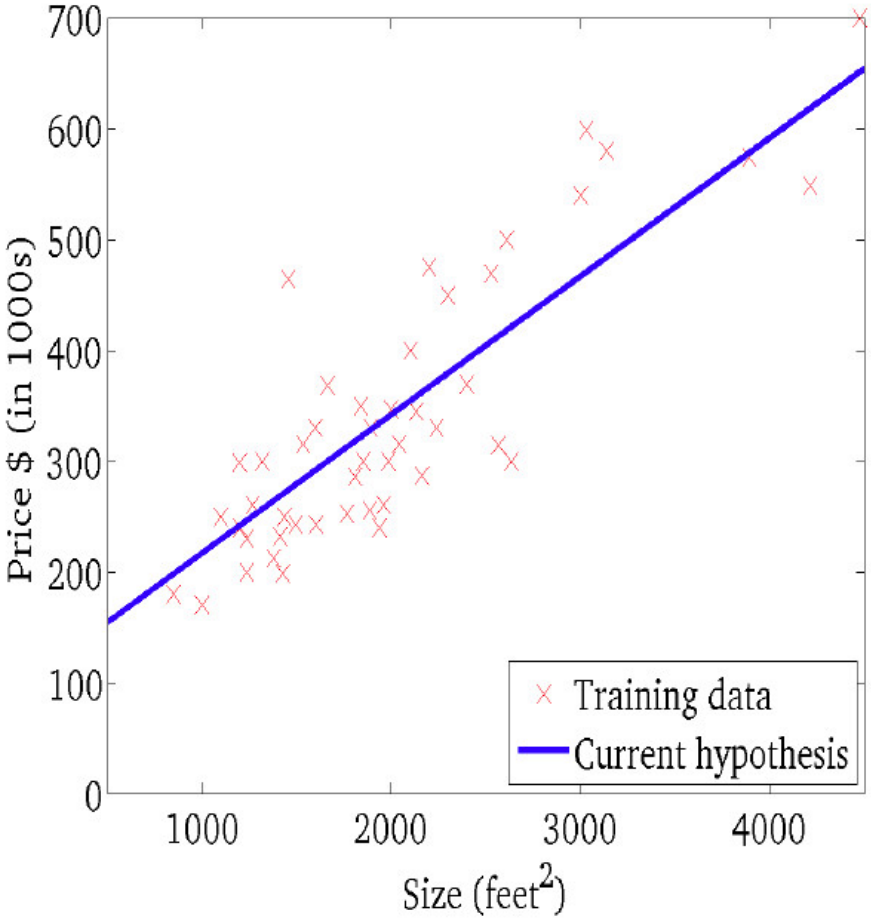


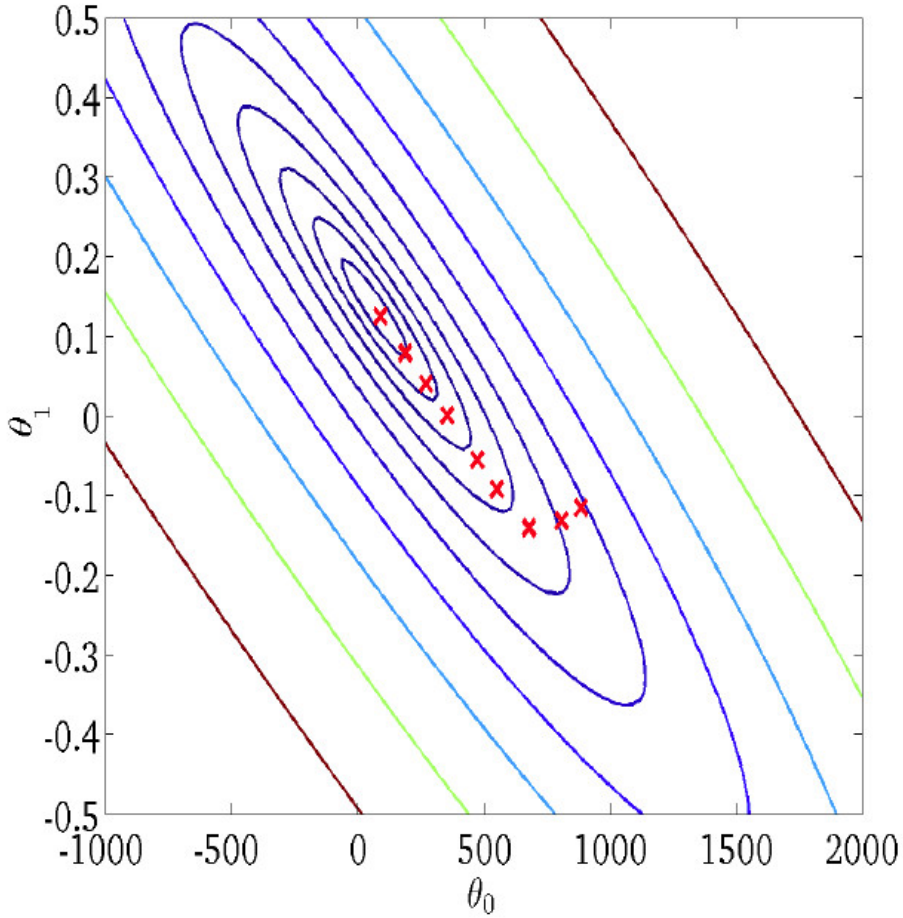
Application de la descente de gradient à notre exemple





Application de la descente de gradient à notre exemple





Sommaire

Retour sur l'exemple introductif

Données

Taille en pied^2	Prix en 1000$
2104	460
1416	232
1534	315
852	178
...	...

Régression linéaire à un seul paramètre

$$h_\theta(x) = \theta_0 + \theta_1 x.$$

Extension du problème à plusieurs paramètres

Données

Taille en pied^2	Nbr de salle de bain	Nb d'étage	Age de la maison	Prix en 1000$
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

Notations

- d nombre de dimensions du descripteur.
- $x^{(i)}$ la $i^{\text{ème}}$ dimension du descripteur du l'exemple x.
- $x_j^{(i)}$ la $i^{\text{ème}}$ dimension du descripteur du $j^{\text{ème}}$ exemple.

Extension du problème à plusieurs paramètres

Données

Taille en pied^2 $x^{(1)}$	Nbr de salle de bain $x^{(2)}$	Nb d'étage $x^{(3)}$	Age de la maison $x^{(4)}$	Prix en 1000\$ y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

Notations

- d nombre de dimensions du descripteur.
- $x^{(i)}$ la $i^{\text{ème}}$ dimension du descripteur du l'exemple x.
- $x_j^{(i)}$ la $i^{\text{ème}}$ dimension du descripteur du $j^{\text{ème}}$ exemple.

Extension du problème à plusieurs paramètres

Données

Taille en pied^2 $x^{(1)}$	Nbr de salle de bain $x^{(2)}$	Nb d'étage $x^{(3)}$	Age de la maison $x^{(4)}$	Prix en 1000\$ y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

Régression linéaire à multi paramètres

$$h_\theta(x) = \theta_0 + \theta_1 x^{(1)} + \cdots + \theta_d x^{(d)} \quad (17)$$

Simplification de la formulation

Notation

On pose $x = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(d)} \end{bmatrix}$ et $\theta = \begin{bmatrix} \theta^1 \\ \vdots \\ \theta^d \end{bmatrix}$

Régression linéaire à multi paramètres

$$h_\theta(x) = \theta_0 + \theta_1 x^{(1)} + \cdots + \theta_d x^{(d)} \quad (18)$$

$$= \theta^\top x + \theta_0. \quad (19)$$

Simplification de la formulation

Notation

On pose $x = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(d)} \end{bmatrix}$ et $\theta = \begin{bmatrix} \theta^1 \\ \vdots \\ \theta^d \end{bmatrix}$

Régression linéaire à multi paramètres

$$h_\theta(x) = \theta_0 \cdot 1 + \theta_1 x^{(1)} + \cdots + \theta_d x^{(d)} \quad (18)$$

$$= \theta^\top x + \theta_0. \quad (19)$$

Simplification de la formulation

Notation

On pose $x = \begin{bmatrix} 1 \\ x^{(1)} \\ \vdots \\ x^{(d)} \end{bmatrix}$ et $\theta = \begin{bmatrix} \theta_0 \\ \theta^1 \\ \vdots \\ \theta^d \end{bmatrix}$

Régression linéaire à multi paramètres

$$h_\theta(x) = \theta_0 \cdot 1 + \theta_1 x^{(1)} + \cdots + \theta_d x^{(d)} \quad (18)$$

$$= \theta^\top x + \cancel{\theta_0}. \quad (19)$$

Erreur de prédiction

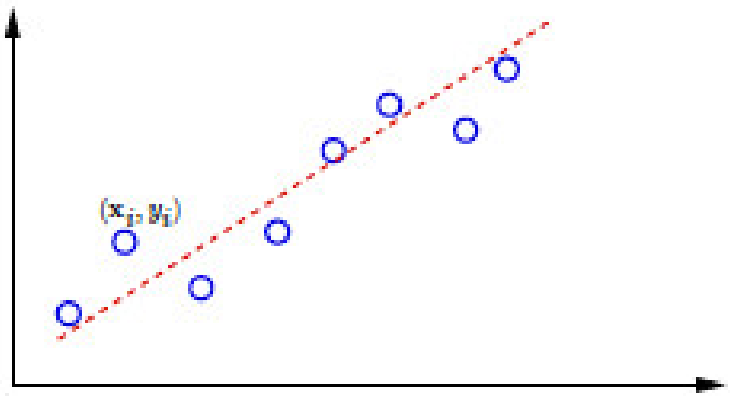
Le critère de moindre carré

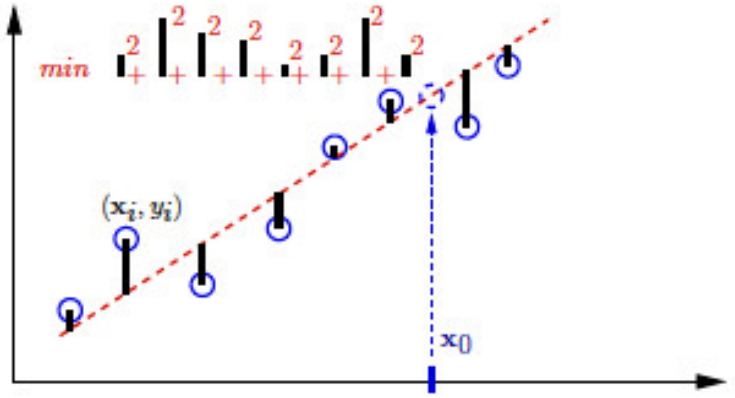
- On peut mesurer l'erreur de prédiction en terme de moyenne des distances au carrés :

$$Loss(y, \tilde{y}) = \|y - \tilde{y}\|^2.$$

- On cherche donc les θ minimisant la fonction de coût suivantes :

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - h_\theta(x_i))^2.$$





Résumé

Contexte choisie

- Classe de fonction utilisée : les fonctions linéaires.

$$h_\theta(x) = \theta^\top x.$$

- Critère de sélection : minimisation du risque empirique.

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - h_\theta(x_i))^2.$$

Problématique : la régression par moindre carré

On cherche les $\theta^\star$ optimal tel que

$$\theta^\star = \arg\min_\theta \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \sum_{j=0}^{d} \theta_j x_i^{(j)} \right)^2, \text{ en considérant } x_i^{(0)} = 1.$$

Minimisation par moindre carrée dans le cas nD

Notations

$$\text{On pose : } X = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(d)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_m^{(1)} & \cdots & x_m^{(d)} \end{bmatrix}, \; y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, \; \theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_d \end{bmatrix}.$$

Formulation

$$\theta^\star = \arg\min_\theta J(\theta) = \arg\min_\theta \frac{1}{m} \|y - \hat{y}\|^2 \tag{20}$$

$$= \arg\min_\theta \frac{1}{m} (y - X\theta)^\top (y - X\theta). \tag{21}$$

Rappels mathématiques

Addition/Multiplication et transposé

- $(AB)^\top = B^\top A^\top$,
- $(A+B)^\top = A^\top + B^\top$.

Dérivée et matrice

- $\frac{\partial a^\top b}{\partial a} = \frac{\partial b^\top a}{\partial a} = b$,
- $\frac{\partial a^\top B a}{\partial a} = 2Ba$.

Condition nécessaire d'optimalité

Annulation de la dérivée de J suivant θ

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\frac{1}{m}(y - X\theta)^\top (y - X\theta), \tag{22}$$

$$= \frac{1}{m}\frac{\partial}{\partial \theta}\left[y^\top y - \theta^\top X^\top y - y^\top X\theta + \theta^\top X^\top X\theta\right], \tag{23}$$

$$= \frac{1}{m}(0 - X^\top y - (y^\top X)^\top + 2X^\top X\theta), \tag{24}$$

$$= -\frac{2}{m}(X^\top y - X^\top X\theta), \tag{25}$$

$$\Longrightarrow X^\top y - X^\top X\theta = 0. \tag{26}$$

Condition nécessaire d'optimalité

Annulation de la dérivée de J suivant θ

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta}\frac{1}{m}\left(y - X\theta\right)^\top \left(y - X\theta\right), && (22)\\
&= \frac{1}{m}\frac{\partial}{\partial \theta}\left[y^\top y - \theta^\top X^\top y - y^\top X\theta + \theta^\top X^\top X\theta\right], && (23)\\
&= \frac{1}{m}(0 - X^\top y - (y^\top X)^\top + 2X^\top X\theta), && (24)\\
&= -\frac{2}{m}(X^\top y - X^\top X\theta), && (25)\\
\Longrightarrow\ & X^\top y - X^\top X\theta = 0. && (26)
\end{aligned}$$

Solution

$$\begin{aligned}
X^\top X\theta &= X^\top y, && (27)\\
\theta^\star &= (X^\top X)^{-1}X^\top y. && (28)
\end{aligned}$$

Méthode de résolution alternative

Analyse de la solution analytique

Le calcul de la pseudo-inverse est en $O(D^3)$. Si le nombre de dimensions des descripteurs est grand, le calcul de la pseudo-inverse peut s'avérer très couteux.

Solution de résolution alternative : la descente de gradient

- Cette solution consiste à atteindre le minimum d'une fonction par des descentes successive selon son gradient (la plus forte pente).
- On change itérativement la valeur de θ selon son gradient de $J(\theta)$:

$$\theta \longleftarrow \theta - \alpha \nabla_\theta J(\theta).$$

Résolution d'un problème de régression par descente de gradient

A chaque itération on met à jour θ

$$\theta \longleftarrow \theta - \alpha \nabla_\theta J(\theta), \tag{29}$$

$$\longleftarrow \theta - \alpha \nabla_\theta \left[(y - X\theta)^\top (y - X\theta) \right], \tag{30}$$

$$\longleftarrow \theta + \frac{2\alpha}{m} X^\top (y - X\theta), \tag{31}$$

$$\longleftarrow \theta + \frac{2\alpha}{m} \sum_{i=1}^{m} (y_i - h_\theta(x_i))\, x_i. \tag{32}$$

Online gradient descent regression

Afin d'accélérer les calcules, on peut utiliser une descente de gradient stochastique. On approxime le gradient de J par le gradient de la fonction de coût pour un exemple :

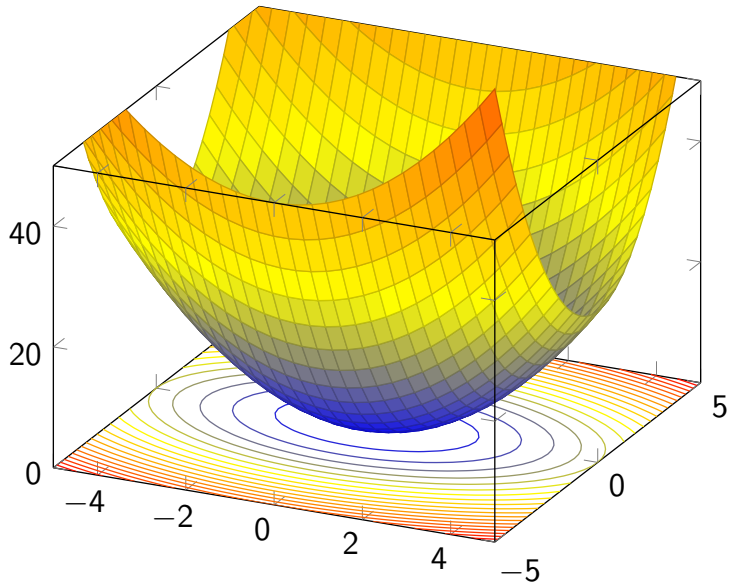
$$\theta \longleftarrow \theta + 2\alpha(y_i - h_\theta(x_i))x_i$$

Amplitude des entrées

Retour à notre exemple introductif

- $x^{(1)}$ = Taille en pied^2 de la maison, valeur entre 0 et 2000.
- $x^{(2)}$ = Nombre de salle de bain, enter 1 et 5.

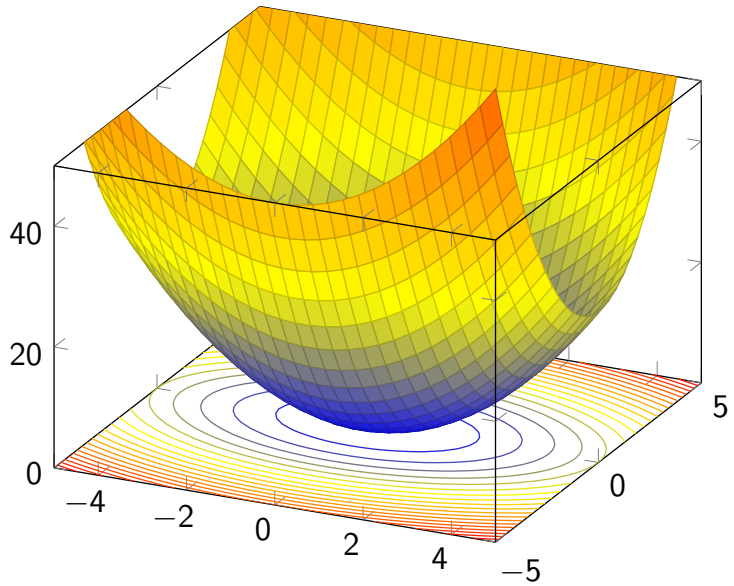
Projection de la fonction de coût sur θ_1 et θ_2 :



Amplitude des entrées

Retour à notre exemple introductif

Projection de la fonction de coût sur θ_1 et θ_2 :



Problème

Problème si chaque dimension du descripteur n'est pas dans la même plage de valeur, la fonction de coût "prend la forme d'un tunnel de half-pipe". La résolution du problème d'optimisation en prenant la plus forte pente est très lente.

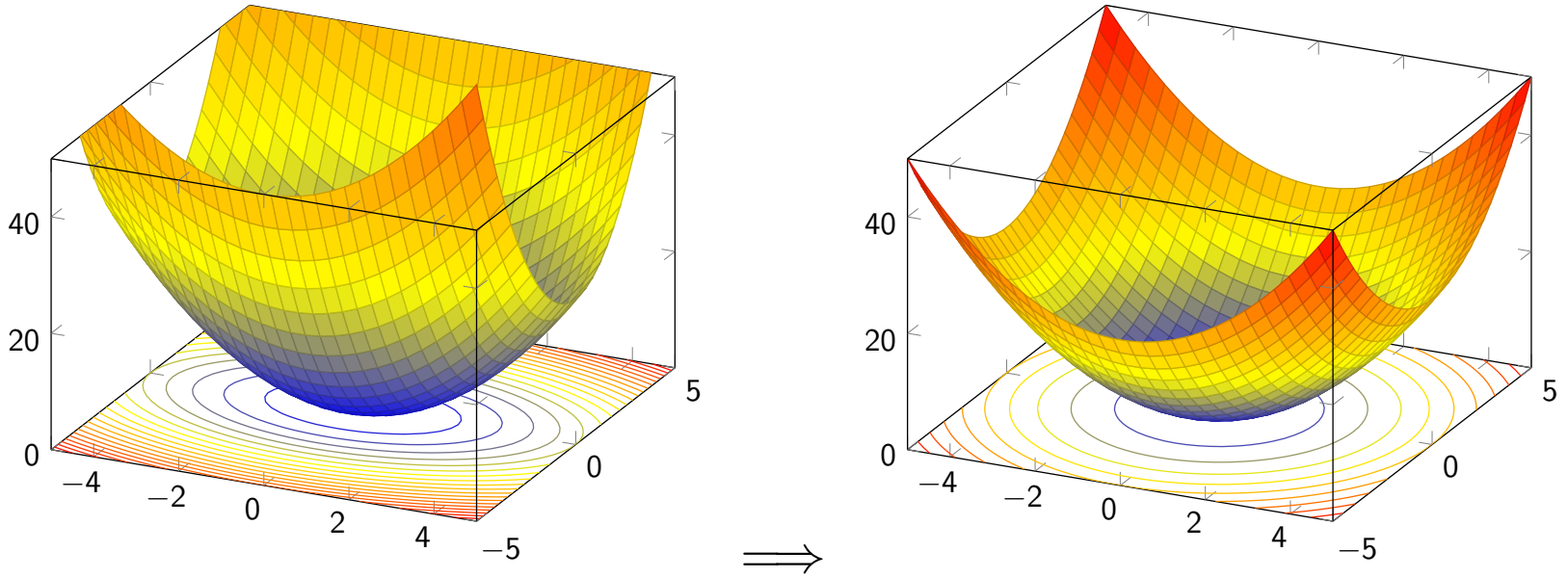
Normalisation des entrées

Solution du problème précédent : Normalisation des entrées

L'amplification des modifications de θ dépend de x_i. Généralement on normalise les données de façon à avoir les entrées entre -1 et 1 et de moyenne nulle :

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sigma}.$$

avec $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ et $\sigma^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$.

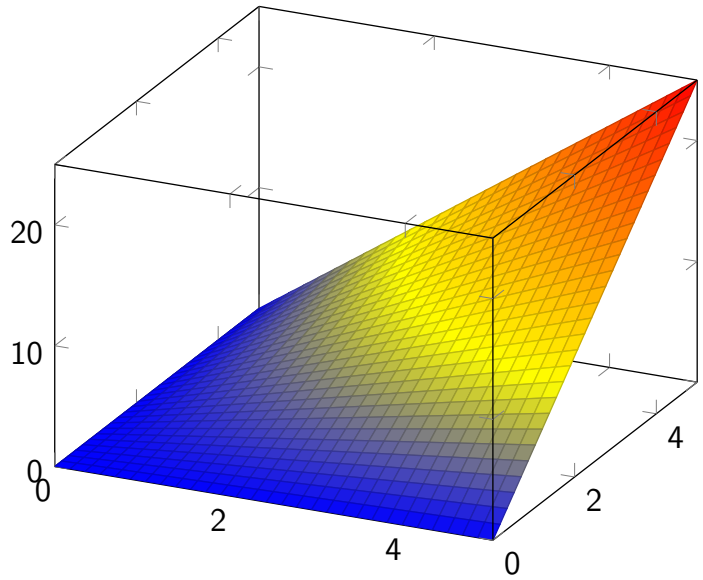


Sommaire

Exemple de régression non linéaire ?

Exemple

On souhaites faire une régression sur le prix de maison x en utilisant leurs largeurs x_ℓ et leurs longueurs x_L. On a une répartition des prix qui ressemble à :



Ce problème n'est pas linéaire, il est de la forme :

$$y = \theta_0 + \theta_1 x_\ell x_L.$$

$\Longrightarrow$ Comment le résoudre ?

Exemple de régression non linéaire ?

Idée

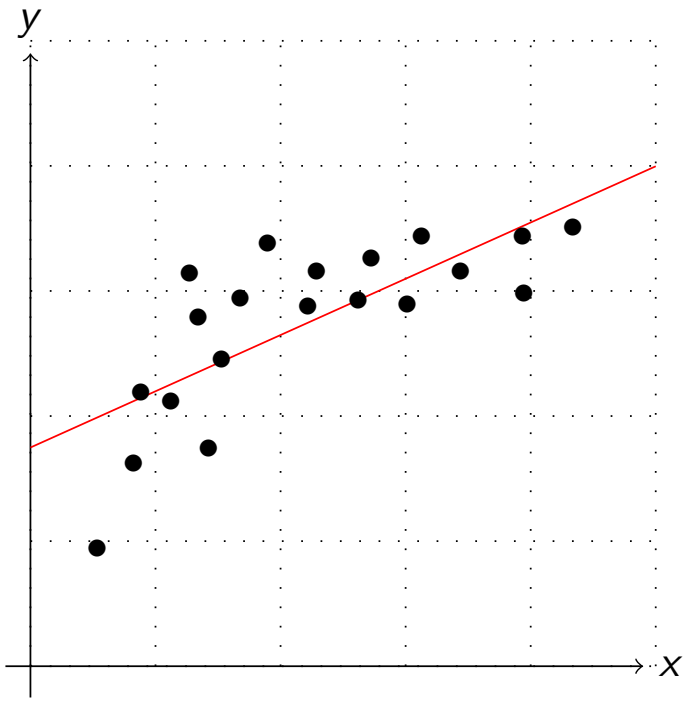
En réécrivant le problème à l'aide de l'air $x_s = x_\ell x_L$, on retrouve un problème linéaire que l'on sais résoudre :

$$y = \theta_0 + \theta_1 x_\ell x_L \rightarrow y = \theta_0 + \theta_1 x_s.$$

Astuce

Se ramener à un problème que l'on sait résoudre.

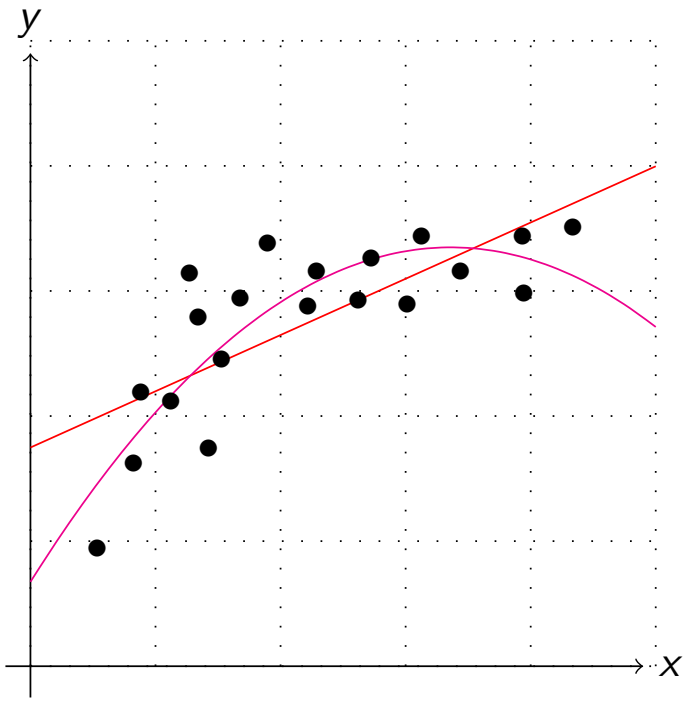
Autres exemples



Régression de type :

- linéaire $y = \theta_0 + \theta_1 x$.

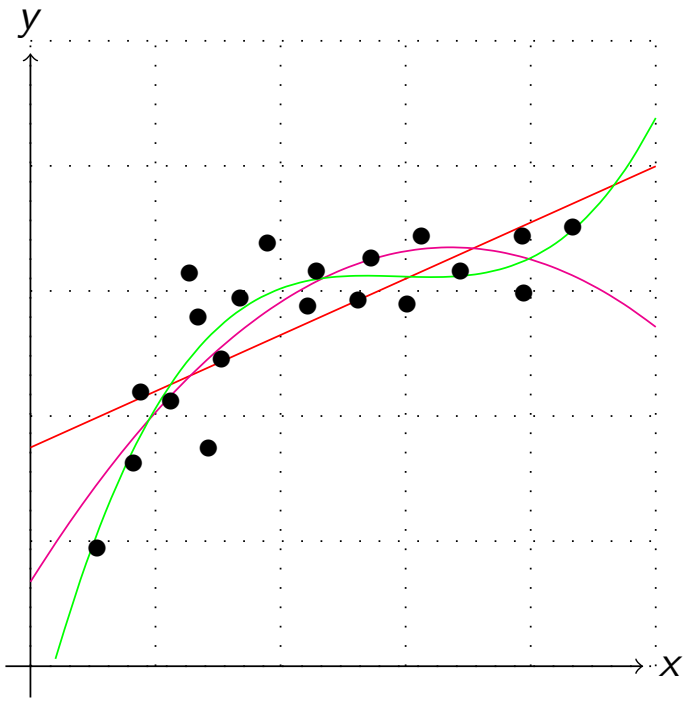
Autres exemples



Régression de type :

- linéaire $y = \theta_0 + \theta_1 x$.
- quadratique $y = \theta_0 + \theta_1 x + \theta_2 x^2$.

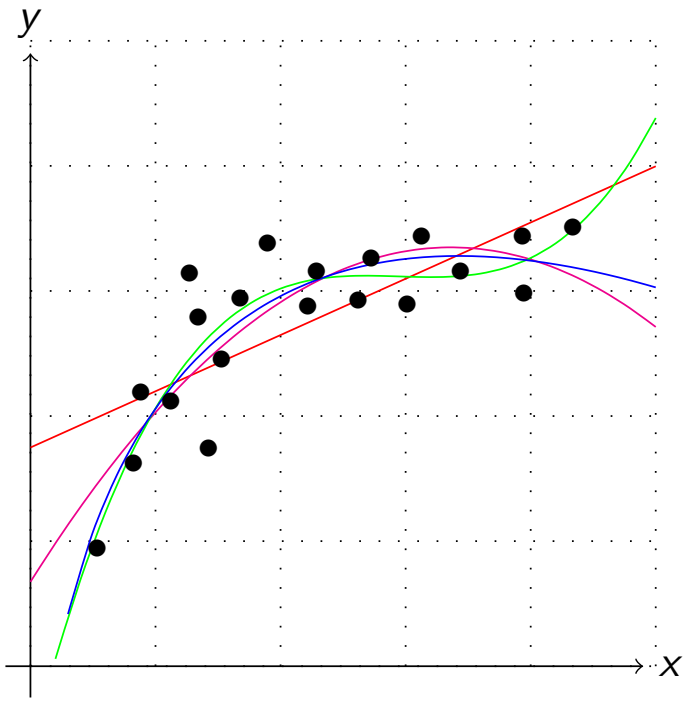
Autres exemples



Régression de type :

- linéaire $y = \theta_0 + \theta_1 x$.
- quadratique $y = \theta_0 + \theta_1 x + \theta_2 x^2$.
- cubique $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$.

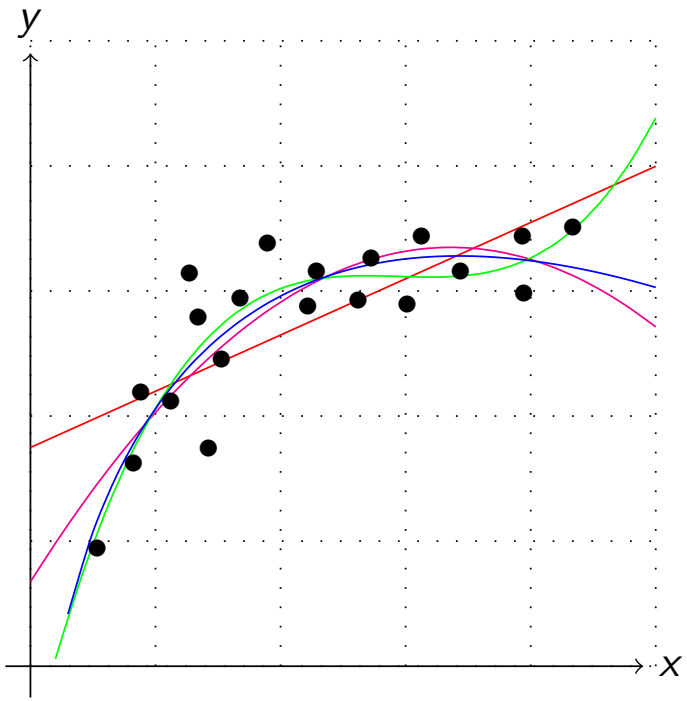
Autres exemples



Régression de type :

- linéaire $y = \theta_0 + \theta_1 x$.
- quadratique $y = \theta_0 + \theta_1 x + \theta_2 x^2$.
- cubique $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$.
- autre $y = \theta_0 + \theta_1 x + \theta_2 \sqrt{x}$.

Autres exemples



Régression de type :

- linéaire $y = \theta_0 + \theta_1 x$.
- quadratique $y = \theta_0 + \theta_1 x + \theta_2 x^2$.
- cubique $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$.
- autre $y = \theta_0 + \theta_1 x + \theta_2 \sqrt{x}$.

Comment faire pour ces régressions ?

Solution

Se ramener au cas linéaire !

Exemples

- linéaire $y = \theta_0 + \theta_1 x$.

Solution

Se ramener au cas linéaire !

Exemples

- linéaire $y = \theta_0 + \theta_1 x$.
- quadratique $y = \theta_0 + \theta_1 x + \theta_2 x^2 \rightarrow y = \theta_0 x^{(0)} + \theta_1 x^{(1)} + \theta_2 x^{(2)}$ avec $x = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}$.

Solution

Se ramener au cas linéaire !

Exemples

- linéaire $y = \theta_0 + \theta_1 x$.
- quadratique $y = \theta_0 + \theta_1 x + \theta_2 x^2 \rightarrow y = \theta_0 x^{(0)} + \theta_1 x^{(1)} + \theta_2 x^{(2)}$ avec $x = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}$.
- cubique $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \rightarrow y = \theta_0 x^{(0)} + \theta_1 x^{(1)} + \theta_2 x^{(2)} + \theta_3 x^{(3)}$ avec $x = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix}$.

Solution

Se ramener au cas linéaire !

Exemples

- linéaire $y = \theta_0 + \theta_1 x$.
- quadratique $y = \theta_0 + \theta_1 x + \theta_2 x^2 \rightarrow y = \theta_0 x^{(0)} + \theta_1 x^{(1)} + \theta_2 x^{(2)}$ avec $x = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}$.
- cubique $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \rightarrow y = \theta_0 x^{(0)} + \theta_1 x^{(1)} + \theta_2 x^{(2)} + \theta_3 x^{(3)}$ avec $x = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \end{bmatrix}$.
- autre $y = \theta_0 + \theta_1 x + \theta_2 \sqrt{x} \rightarrow y = \theta_0 x^{(0)} + \theta_1 x^{(1)} + \theta_2 x^{(2)}$ avec $x = \begin{bmatrix} 1 \\ x \\ \sqrt{x} \end{bmatrix}$.

Régression pour des fonctions non linéaires

Hypothèse sur la famille de fonction

On considère les fonctions de type

$$h_\theta(x) = \theta_0 + \theta_1\varphi_1(x) + \cdots + \theta_d\varphi_d(x).$$

Résolution

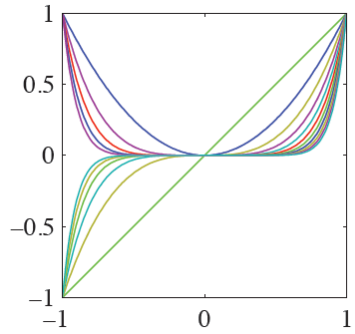
On calcul les valeurs $\varphi_i(x)$ pour tout i, puis on effectue une régression linéaire dans ce nouvelle espace.

$$\{\forall i, x^{(i)} = \varphi_i(x)\} \longrightarrow h_\theta(x) = \theta_0 + \theta_1 x^{(1)} + \cdots + \theta_d x^{(d)}.$$

Exemple de fonction φ

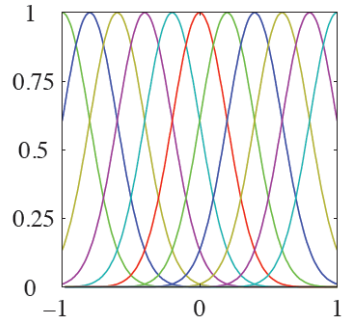
La régression polynomiale

$$\varphi_i = x^i.$$



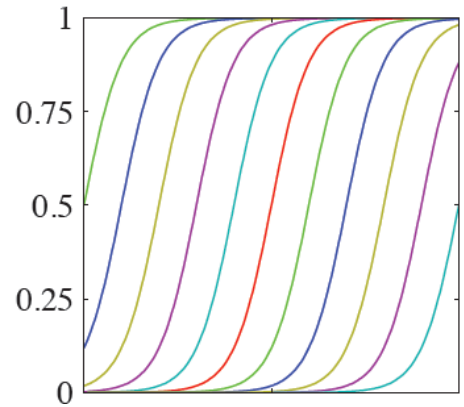
Combinaison d'exponentielle

$$\varphi_i = \exp\left(-\frac{(x-\mu_i)^2}{2\sigma^2}\right)$$



Combinaison de sigmoïde

$$\varphi_i = s\left(\frac{x-\mu_i}{\sigma}\right) \text{ avec } s(a) = \frac{1}{1+\exp(-a)}.$$



Merci de votre attention.