



Bienvenue!

**ÉCOLE D'ÉTÉ FRANCOPHONE
EN APPRENTISSAGE PROFOND**

21-25 août 2017



IVADO

HEC Montréal
Polytechnique Montréal
Université de Montréal



MILA

Optimisation pour les réseaux profonds

École d'été du MILA

Nicolas Le Roux

Google Brain team

27/3/17

1 Optimisation simple

2 Trois difficultés

3 Trouver un bon optimiseur

4 Commentaires additionnels

Optimisation ?

- Un modèle est une fonction paramétrique
- On définit à la main la fonction d'erreur
- On cherche les paramètres qui minimisent l'erreur moyenne
- Entraîner un modèle = trouver les meilleurs paramètres = optimisation

Descente de gradient du premier ordre

- Expression à minimiser: $f(\theta)$

Descente de gradient du premier ordre

- Expression à minimiser: $f(\theta)$
- Exemple: $f(\theta) = \frac{1}{N} \sum_i (x_i \theta - y_i)^2$

Descente de gradient du premier ordre

- Expression à minimiser: $f(\theta)$
- Exemple: $f(\theta) = \frac{1}{N} \sum_i (x_i \theta - y_i)^2$
- Initialisation: θ_0

Descente de gradient du premier ordre

- Expression à minimiser: $f(\theta)$
- Exemple: $f(\theta) = \frac{1}{N} \sum_i (x_i \theta - y_i)^2$
- Initialisation: θ_0
- Itérer:
 - ▶ Calcul du gradient: $g(\theta_t) = f'(\theta_t)$

Descente de gradient du premier ordre

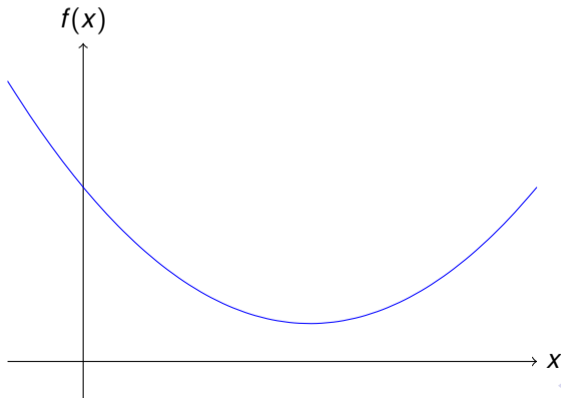
- Expression à minimiser: $f(\theta)$
- Exemple: $f(\theta) = \frac{1}{N} \sum_i (x_i \theta - y_i)^2$
- Initialisation: θ_0
- Itérer:
 - ▶ Calcul du gradient: $g(\theta_t) = f'(\theta_t)$
 - ▶ Mise à jour: $\theta_{t+1} = \theta_t - \alpha g(\theta_t)$

Descente de gradient du premier ordre

- Expression à minimiser: $f(\theta)$
- Exemple: $f(\theta) = \frac{1}{N} \sum_i (x_i \theta - y_i)^2$
- Initialisation: θ_0
- Itérer:
 - ▶ Calcul du gradient: $g(\theta_t) = f'(\theta_t)$
 - ▶ Mise à jour: $\theta_{t+1} = \theta_t - \alpha g(\theta_t)$
- Cet algorithme converge pour un “bon” α [Cau47].

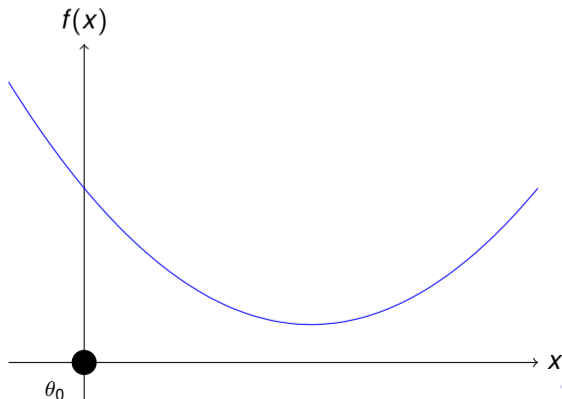
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



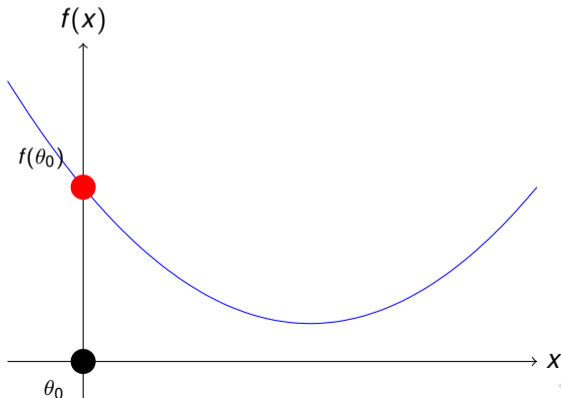
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



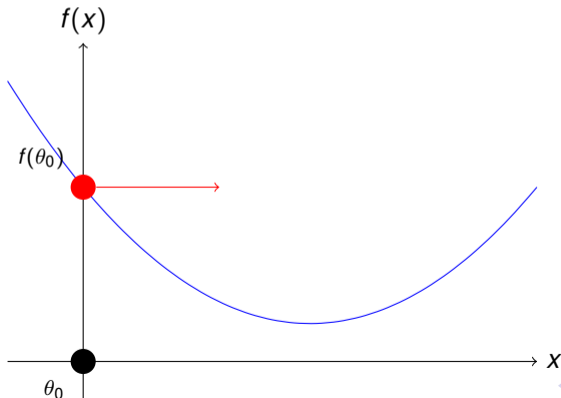
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



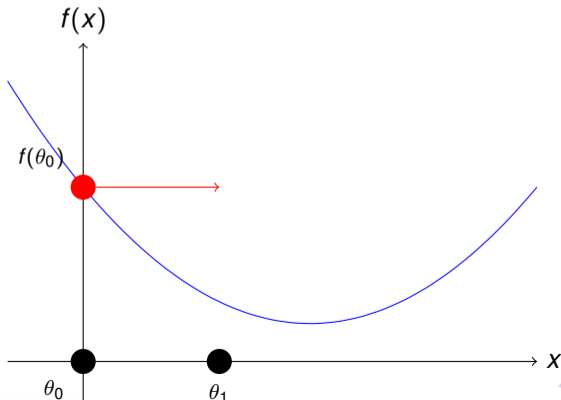
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



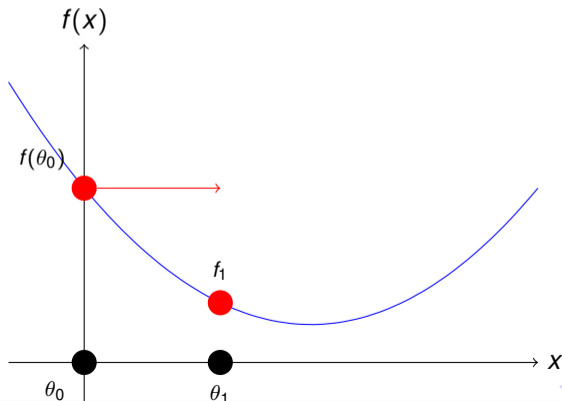
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



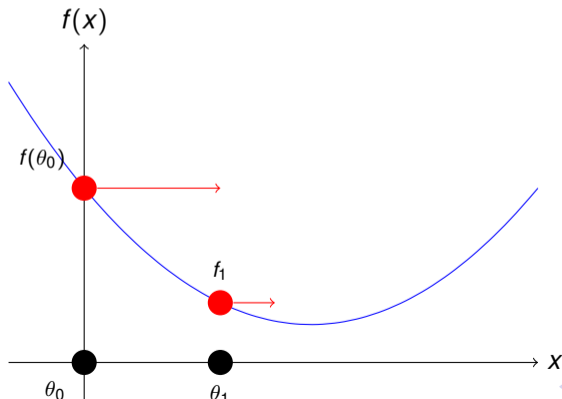
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



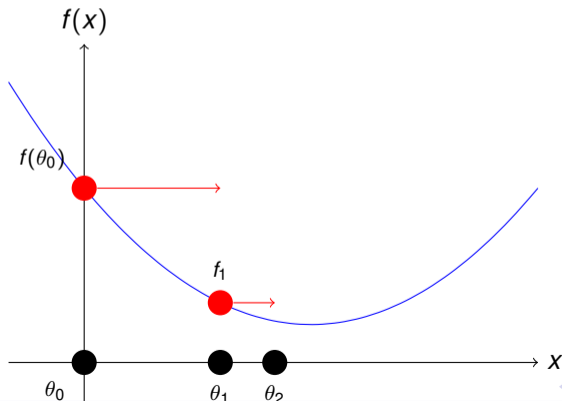
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



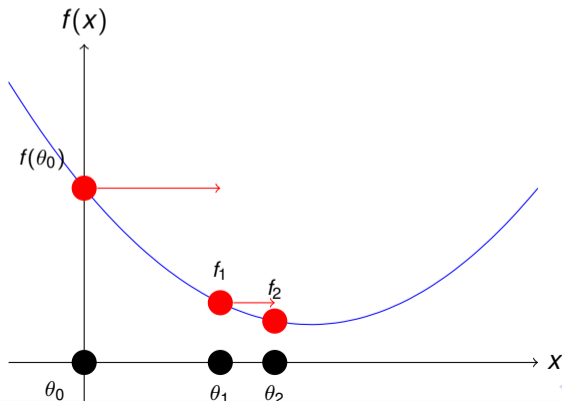
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



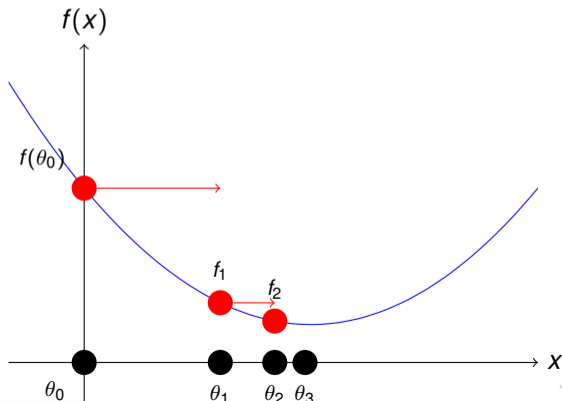
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



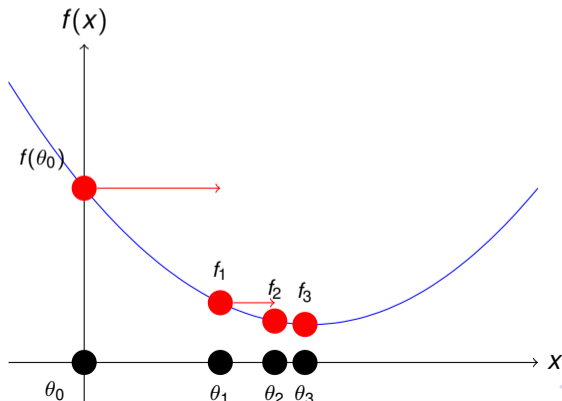
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



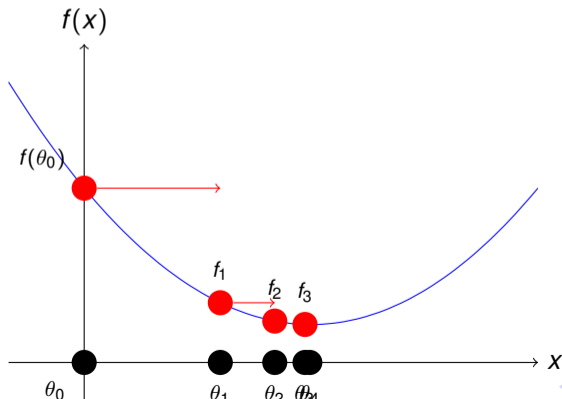
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



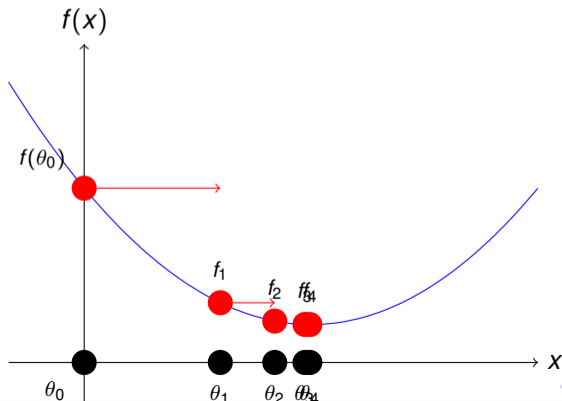
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



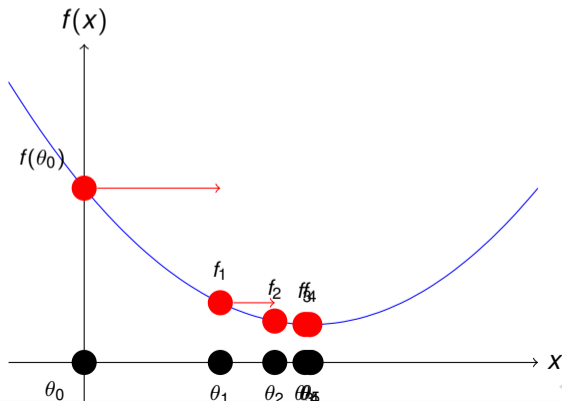
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



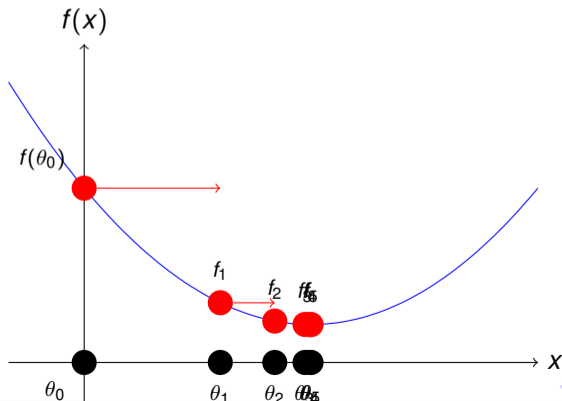
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



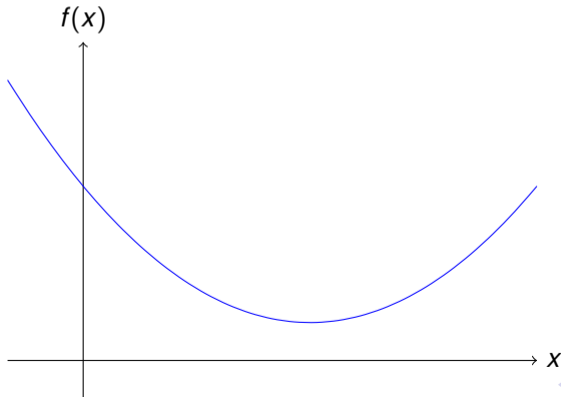
1D - Bon pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



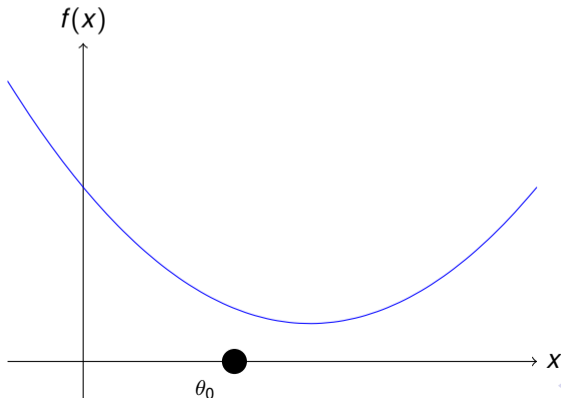
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



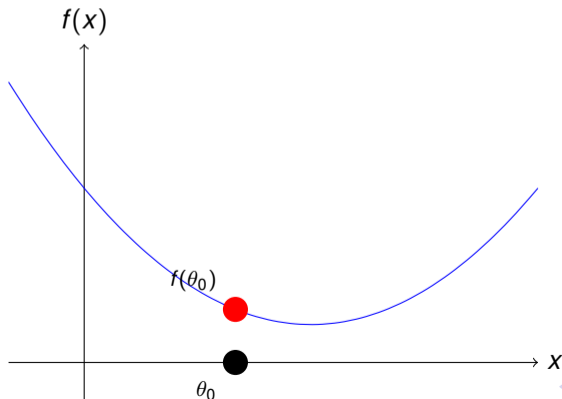
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



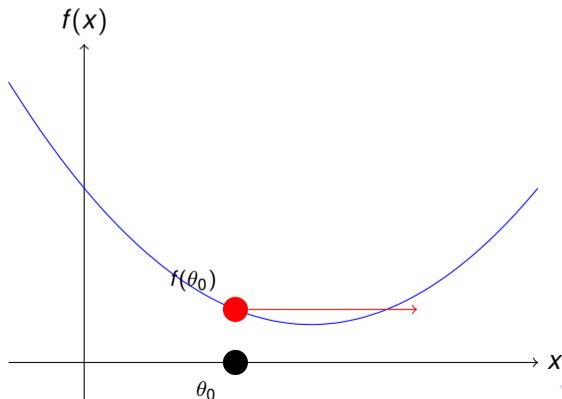
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



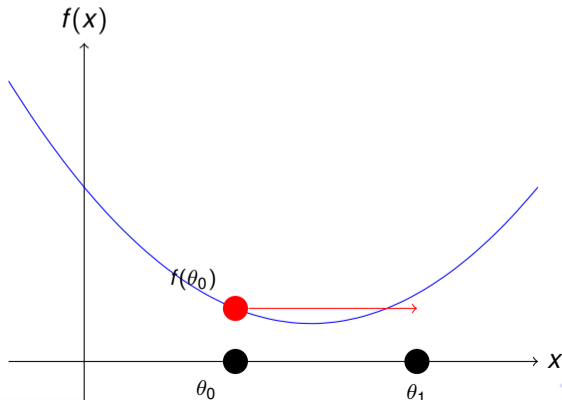
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



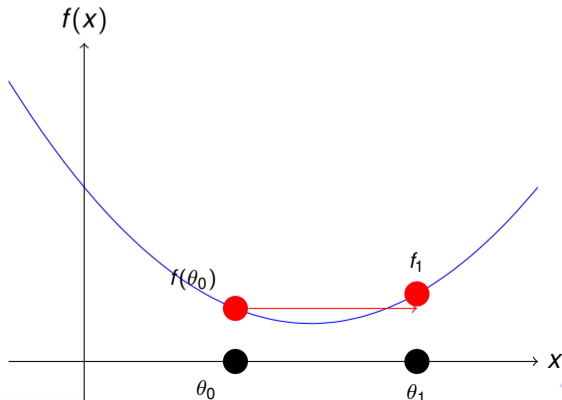
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



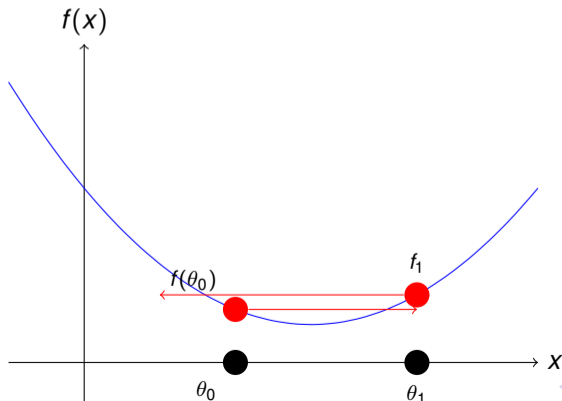
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



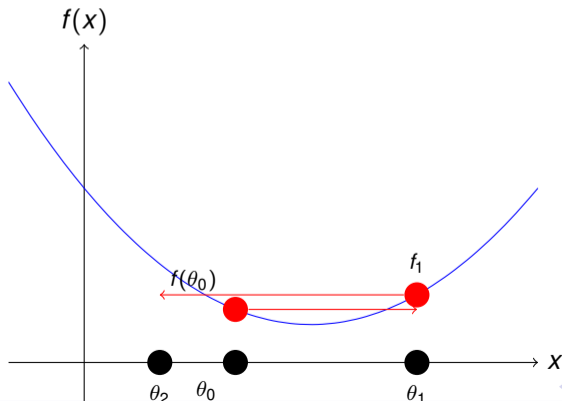
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



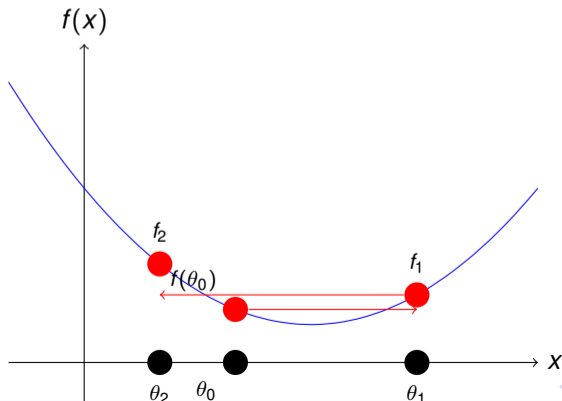
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



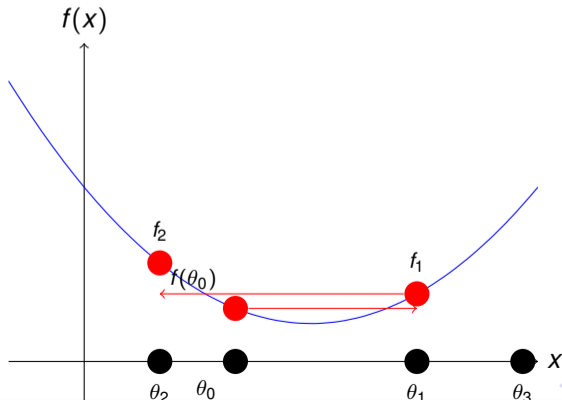
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



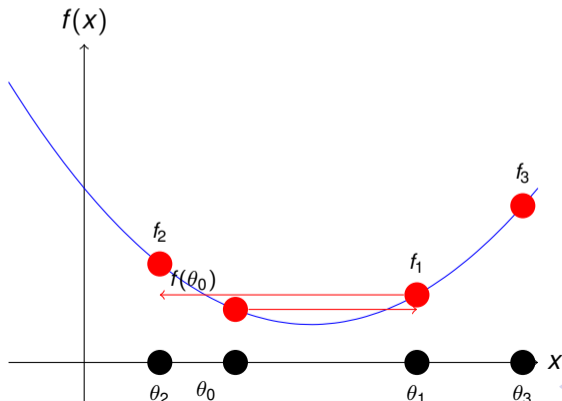
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



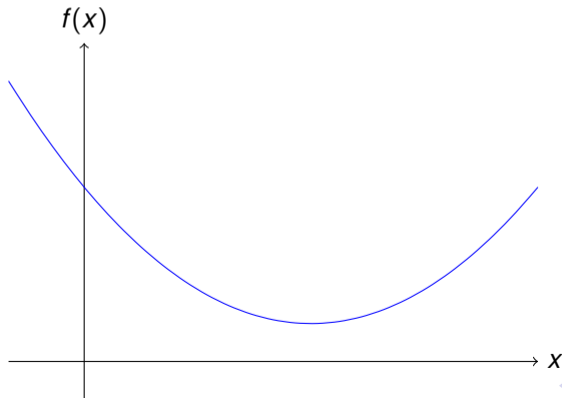
1D - Grand pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



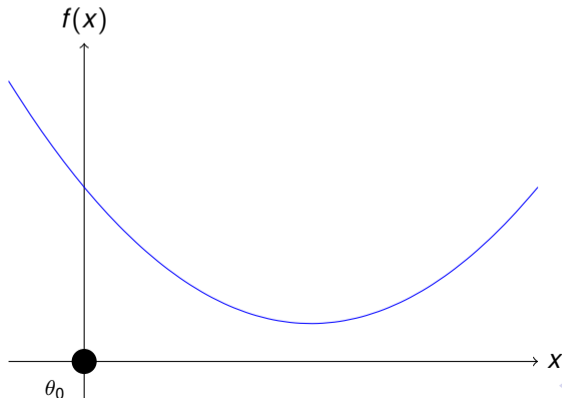
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



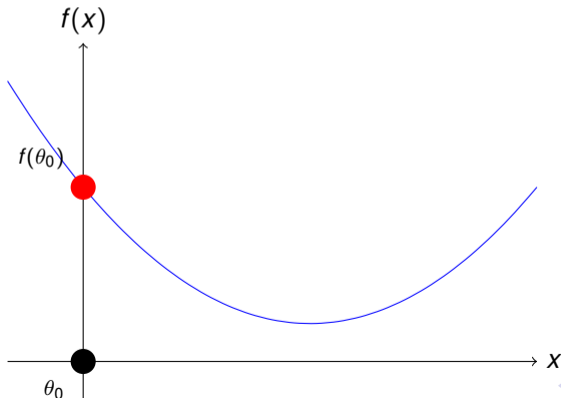
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



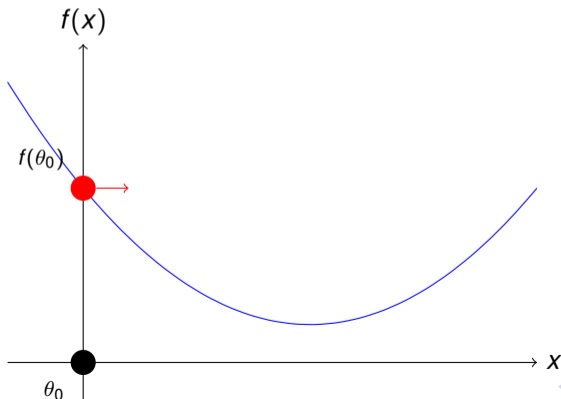
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



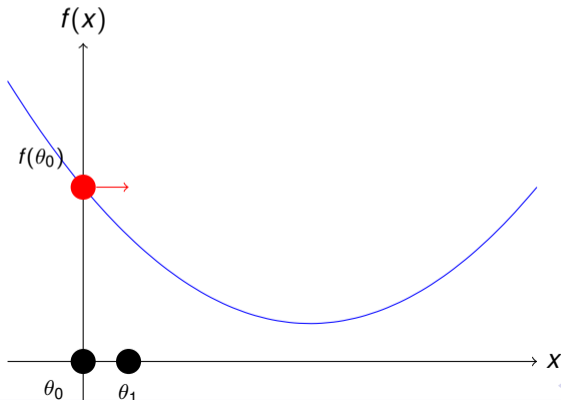
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



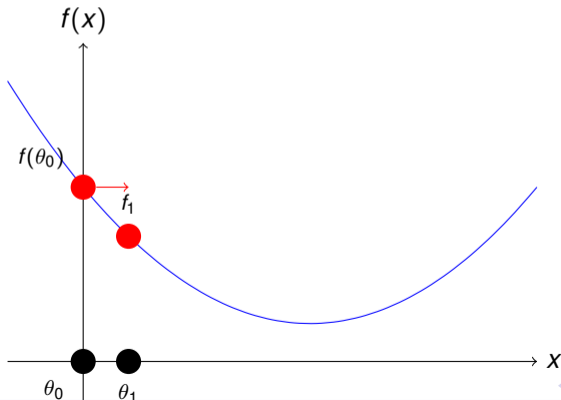
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



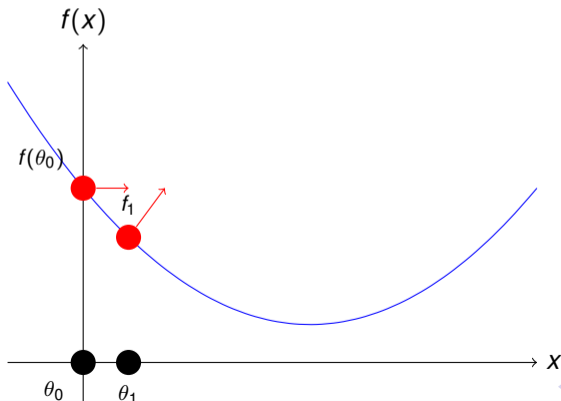
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



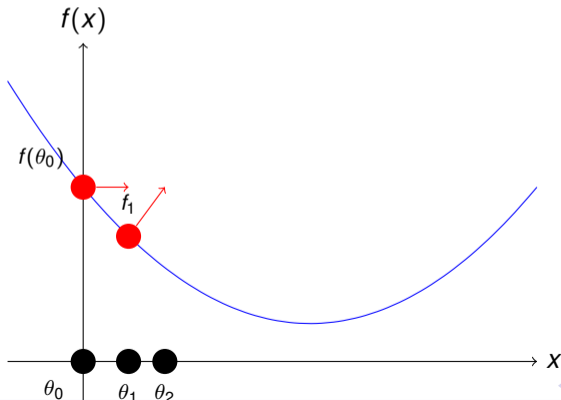
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



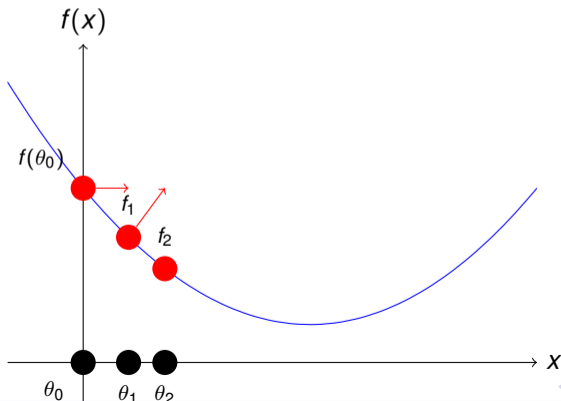
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



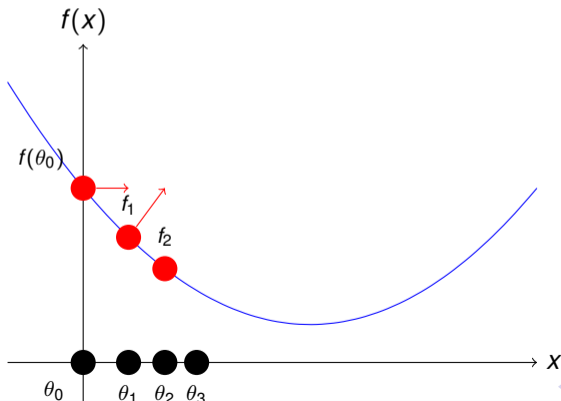
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



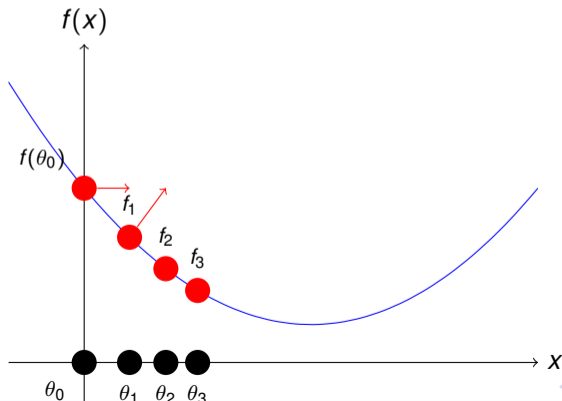
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



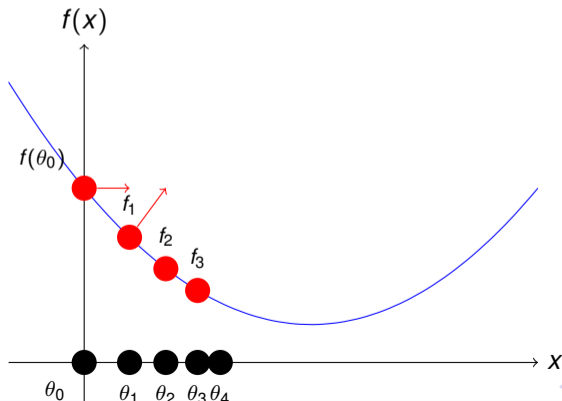
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



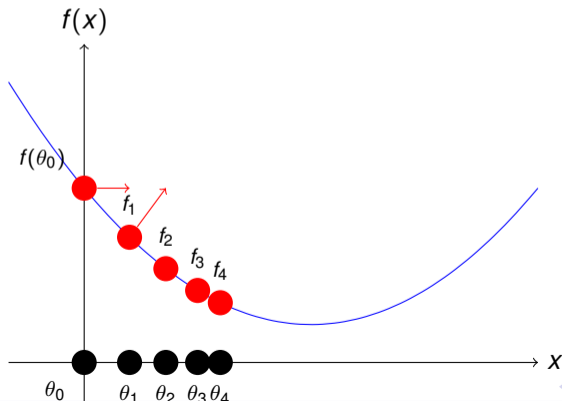
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



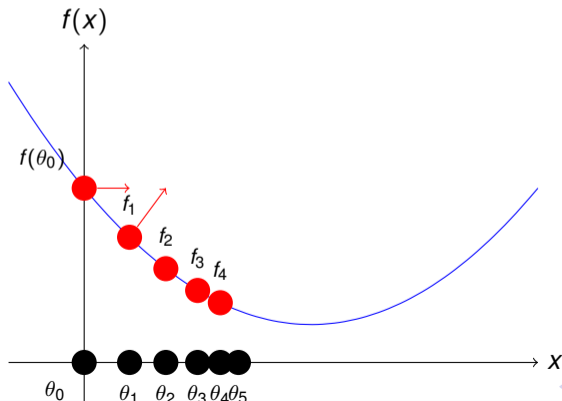
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



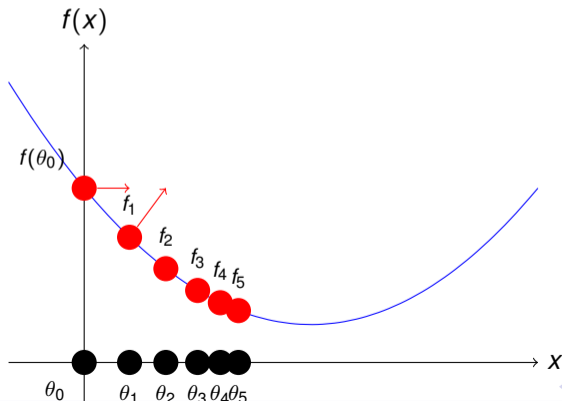
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



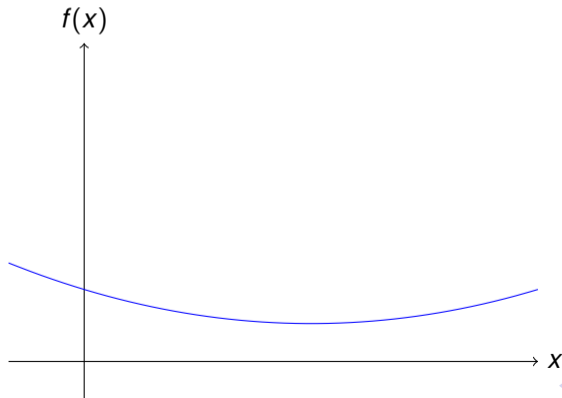
1D - Petit pas de gradient

$$\theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



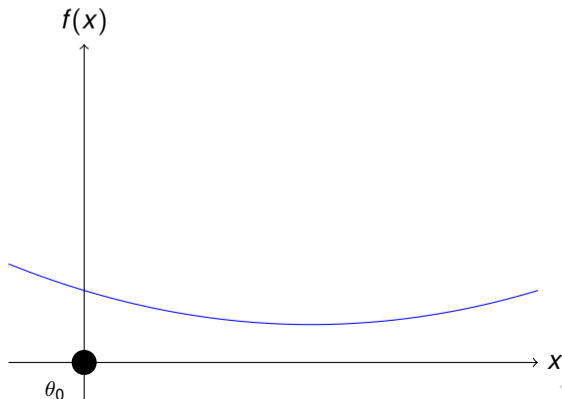
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



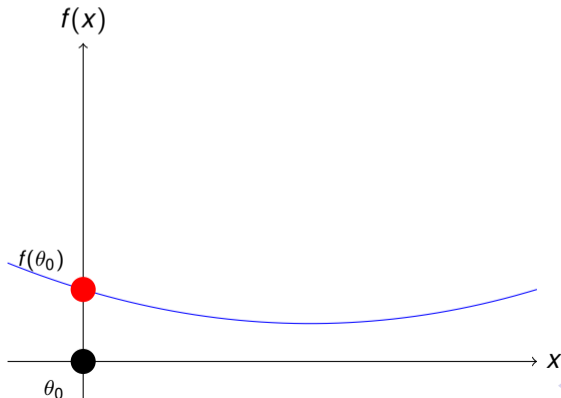
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



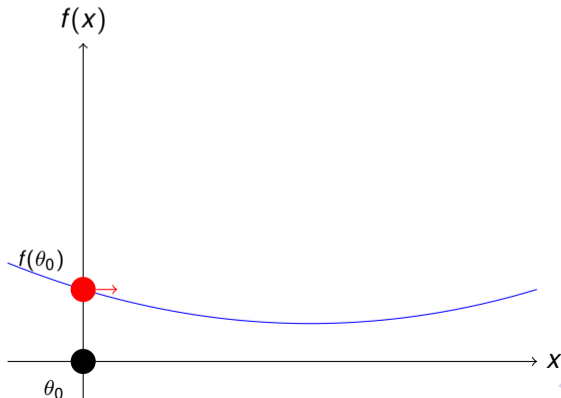
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



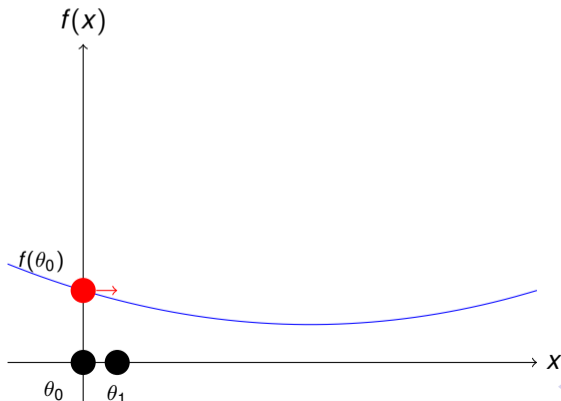
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



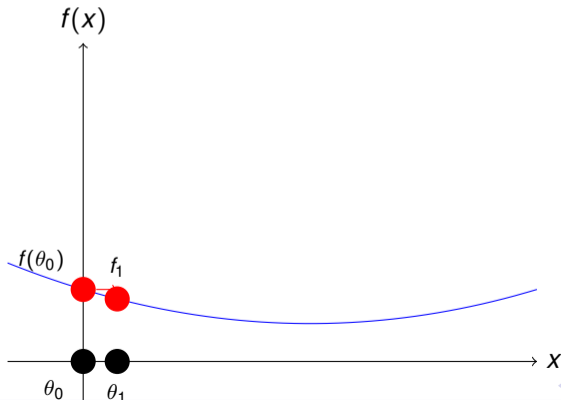
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



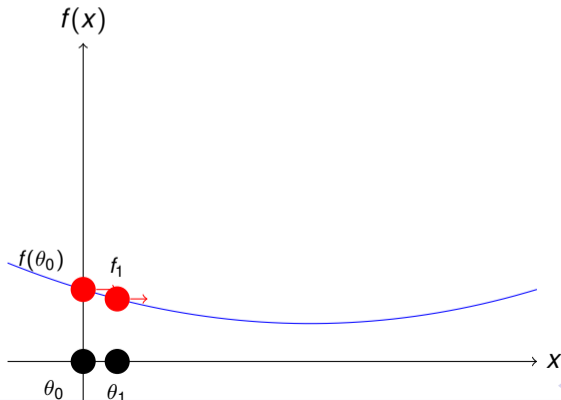
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



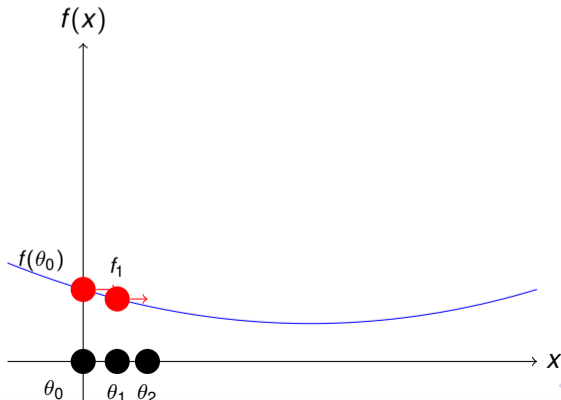
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



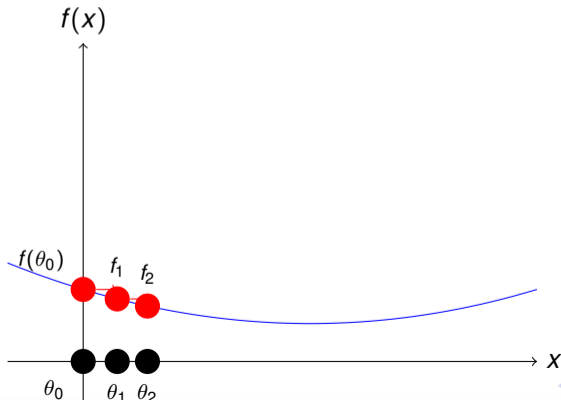
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



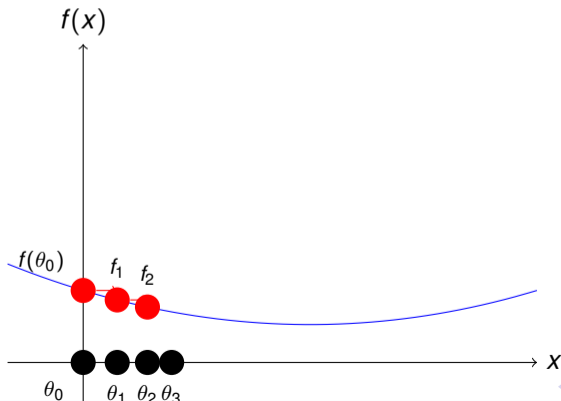
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



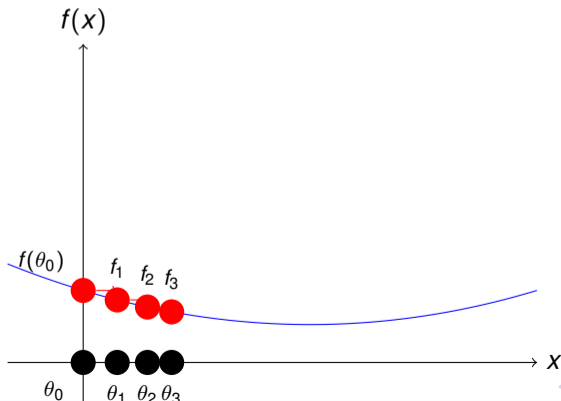
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



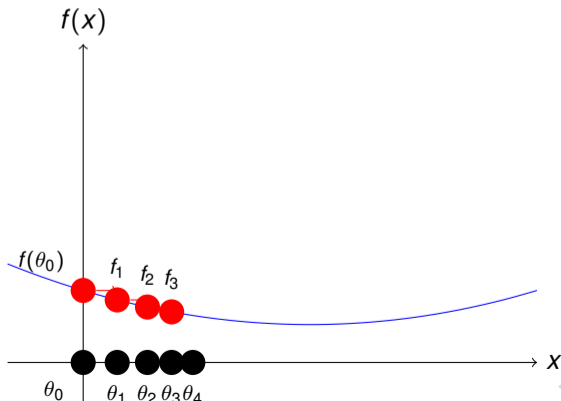
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



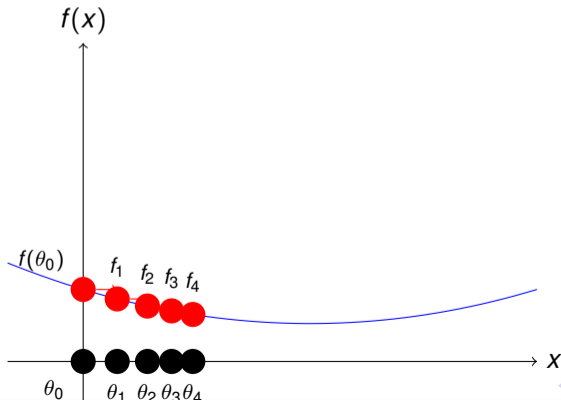
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



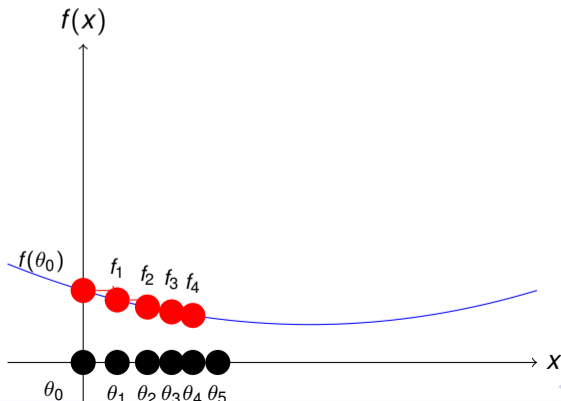
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



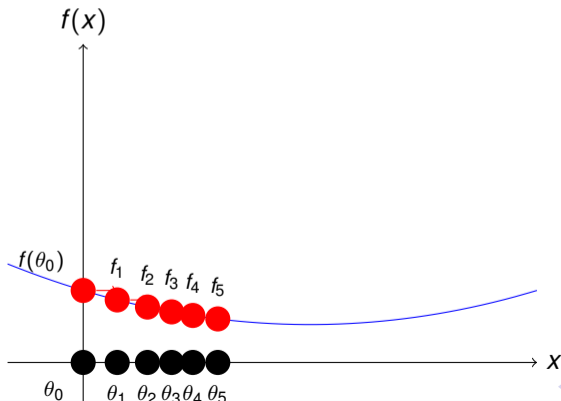
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



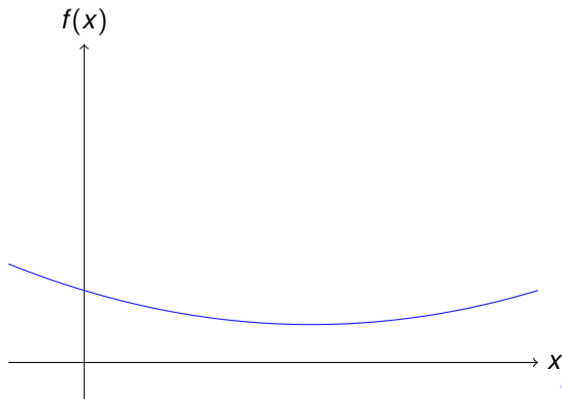
1D - “Bon” pas de gradient

$$\theta_{t+1} = \theta_t - \frac{3}{2} \frac{df(x)}{dx}$$



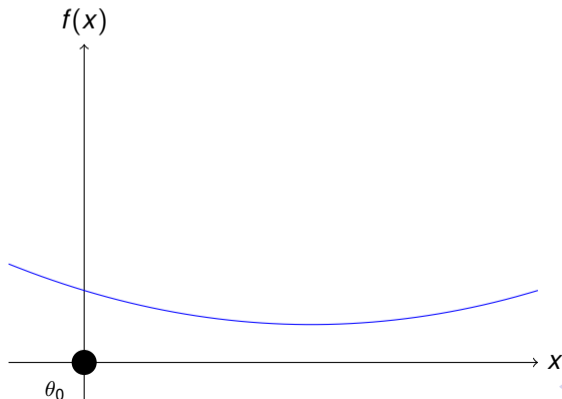
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



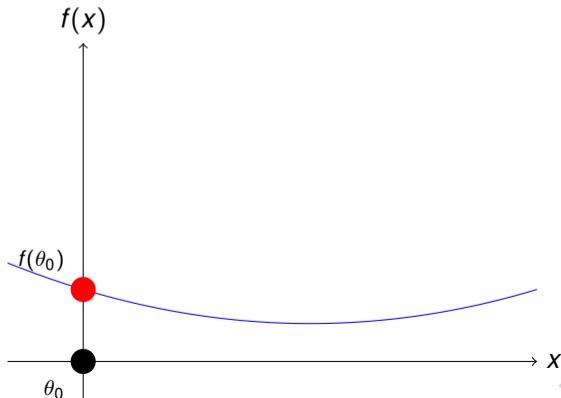
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



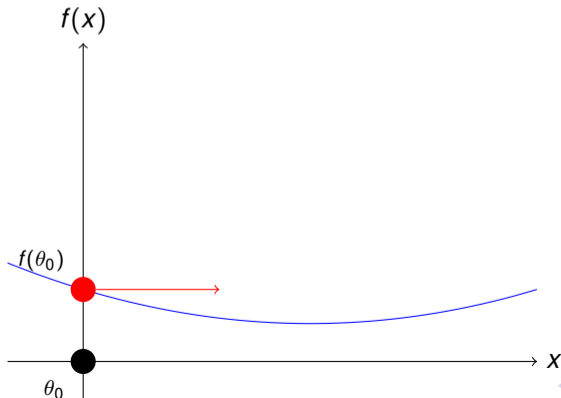
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



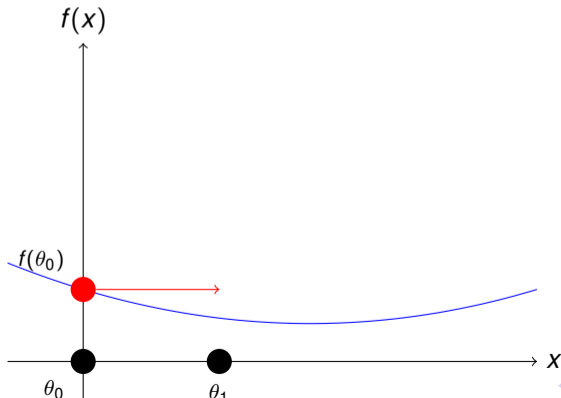
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



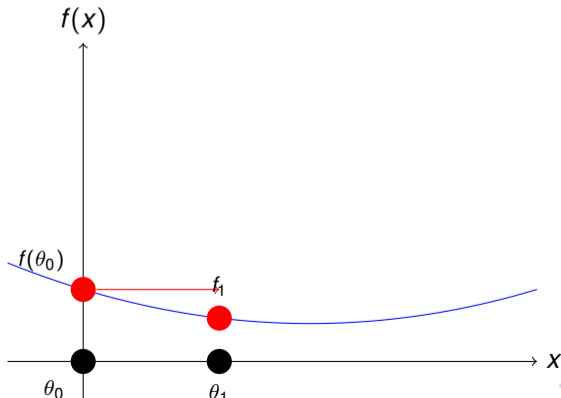
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



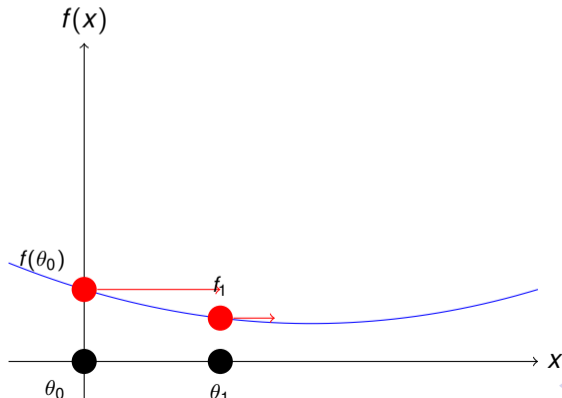
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



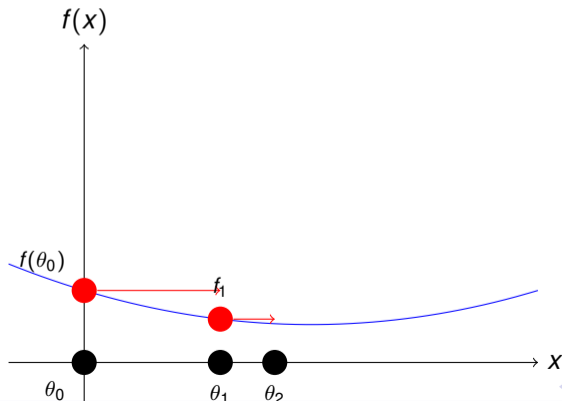
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



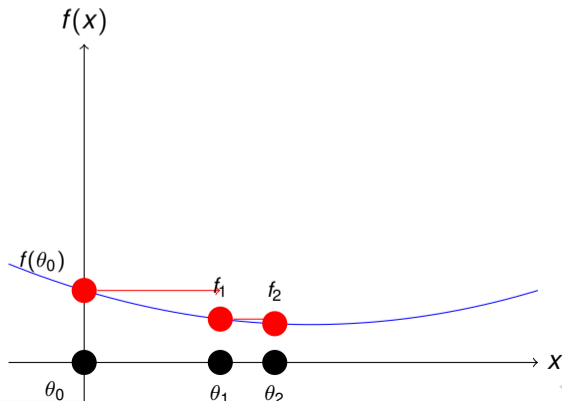
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



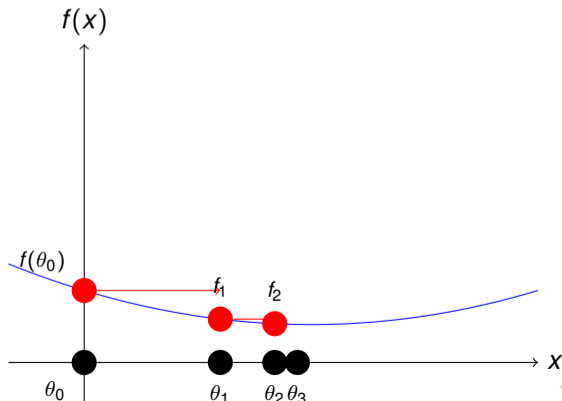
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



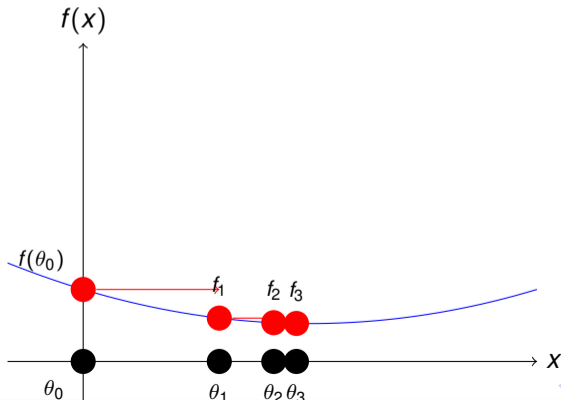
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



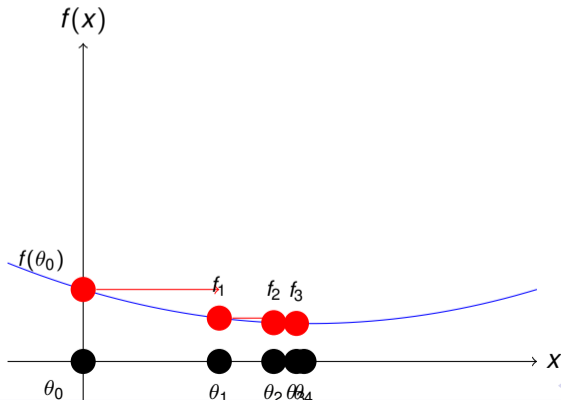
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



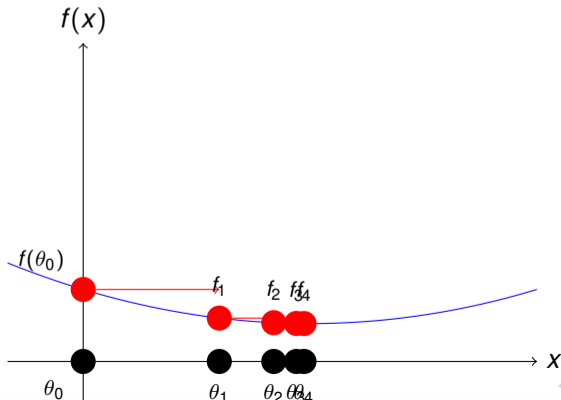
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



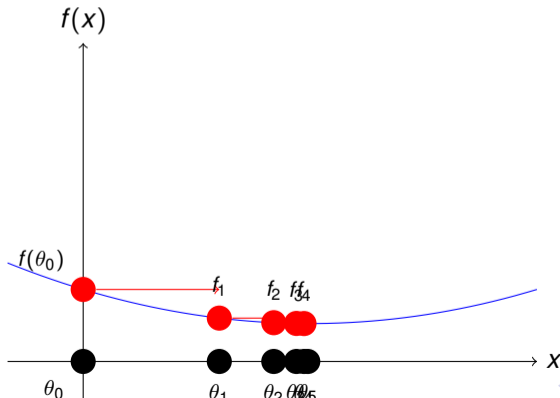
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



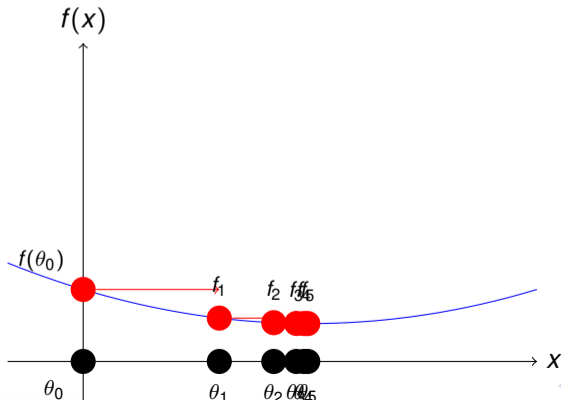
1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



1D - “Grand” pas de gradient

$$\theta_{t+1} = \theta_t - 6 \frac{df(x)}{dx}$$



Fonction convexe en 1D

- Le pas de gradient influe sur la convergence
- Il dépend de la fonction à optimiser
- S'il y a convergence, c'est forcément au minimum

1 Optimisation simple

2 **Trois difficultés**

3 Trouver un bon optimiseur

4 Commentaires additionnels

Difficulté 1: la haute dimension

- En pratique, θ n'est pas un nombre
- C'est un vecteur qui peut contenir des millions de paramètres

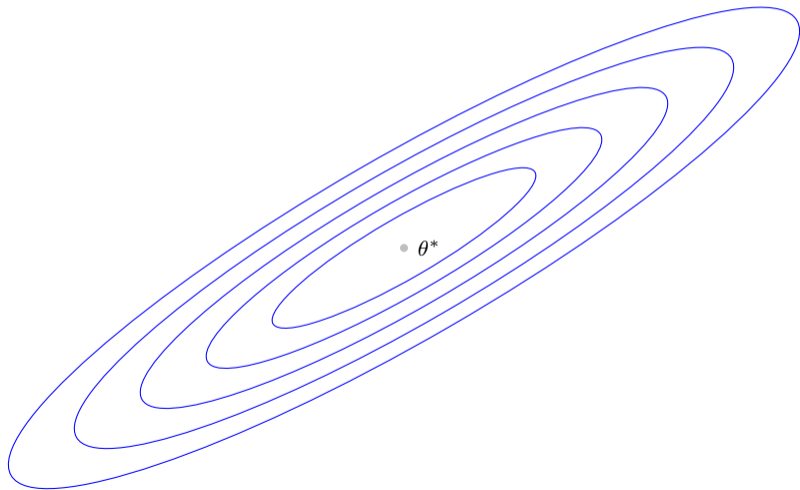
Difficulté 1: la haute dimension

- En pratique, θ n'est pas un nombre
- C'est un vecteur qui peut contenir des millions de paramètres
- La descente de gradient marche-t-elle toujours?

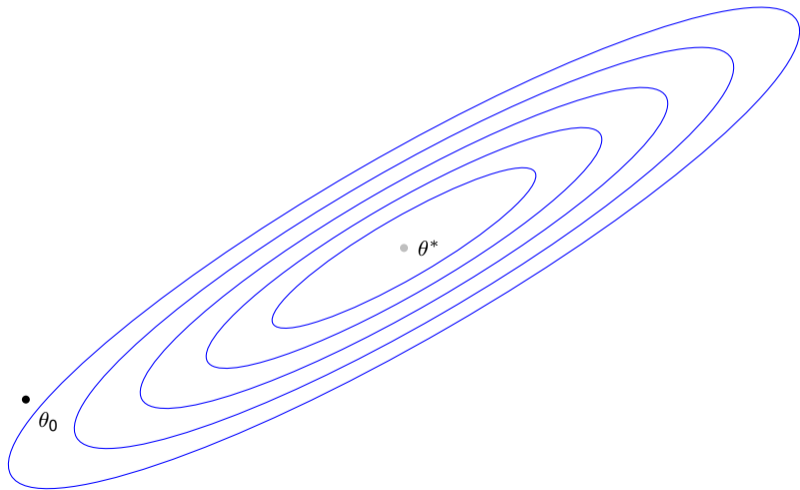
Difficulté 1: la haute dimension

- En pratique, θ n'est pas un nombre
- C'est un vecteur qui peut contenir des millions de paramètres
- La descente de gradient marche-t-elle toujours?
- Oui mais...

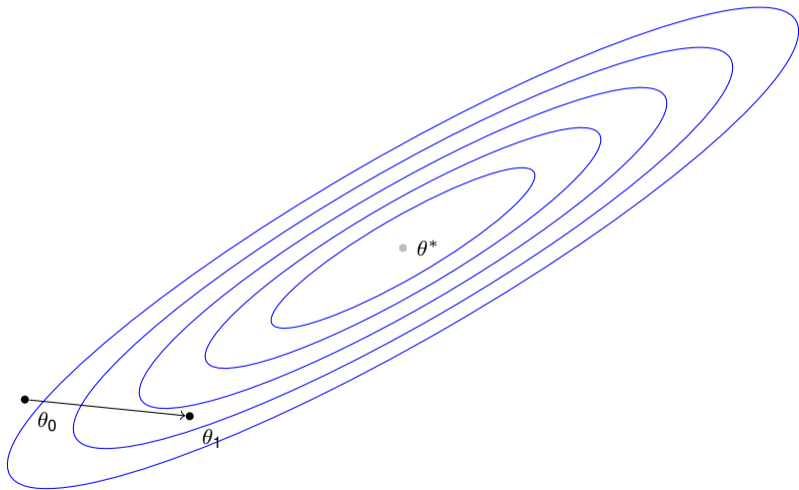
“Haute” dimension - Bon pas de gradient



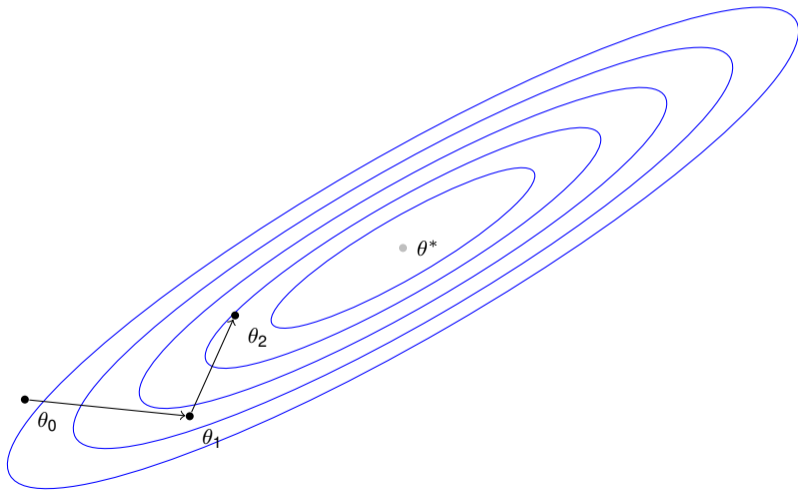
“Haute” dimension - Bon pas de gradient



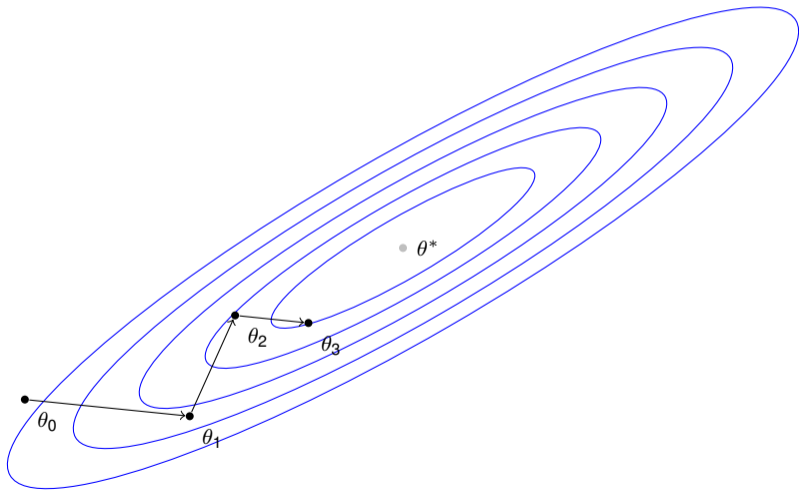
“Haute” dimension - Bon pas de gradient



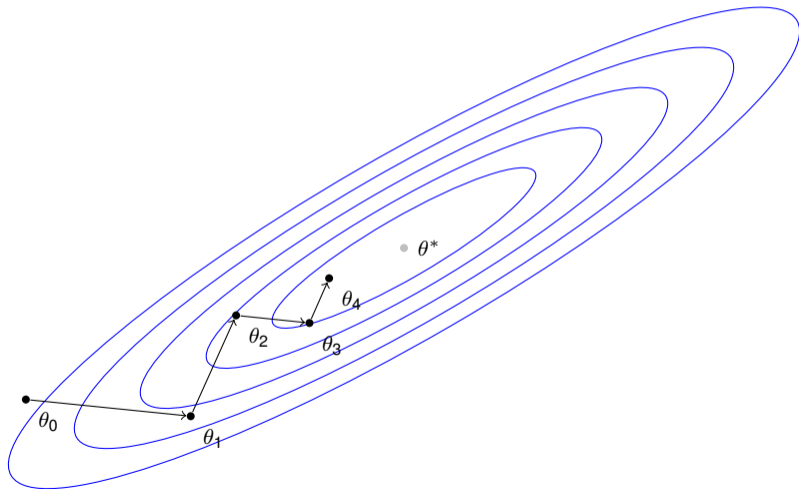
“Haute” dimension - Bon pas de gradient



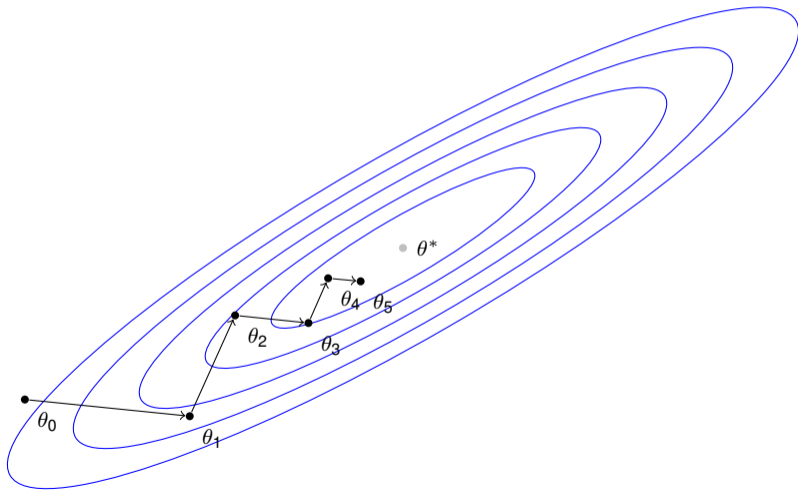
“Haute” dimension - Bon pas de gradient



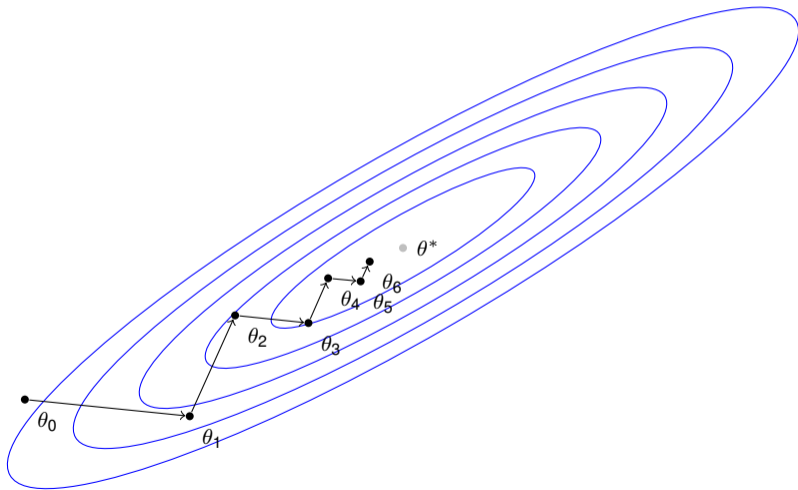
“Haute” dimension - Bon pas de gradient



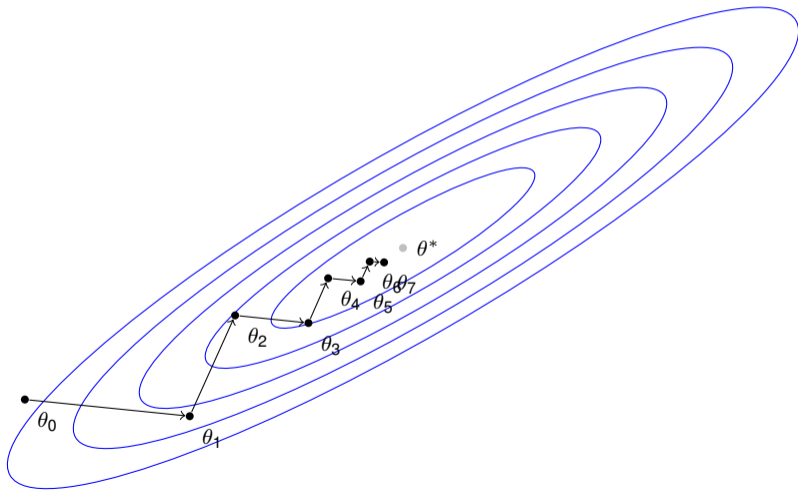
“Haute” dimension - Bon pas de gradient



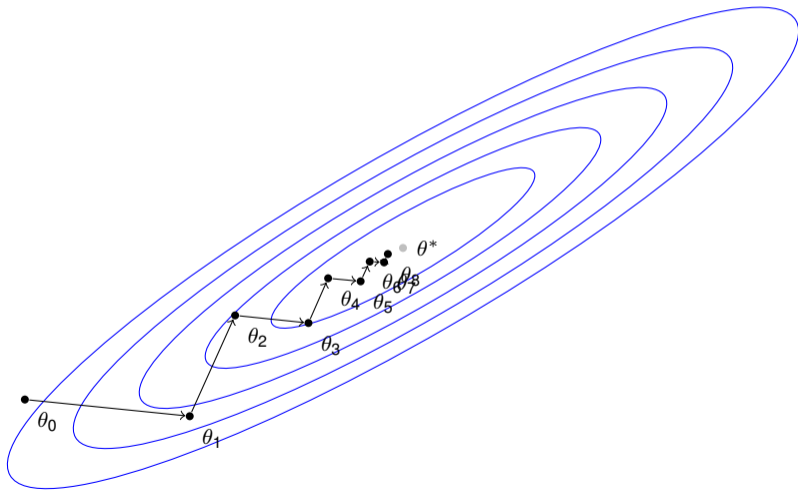
“Haute” dimension - Bon pas de gradient



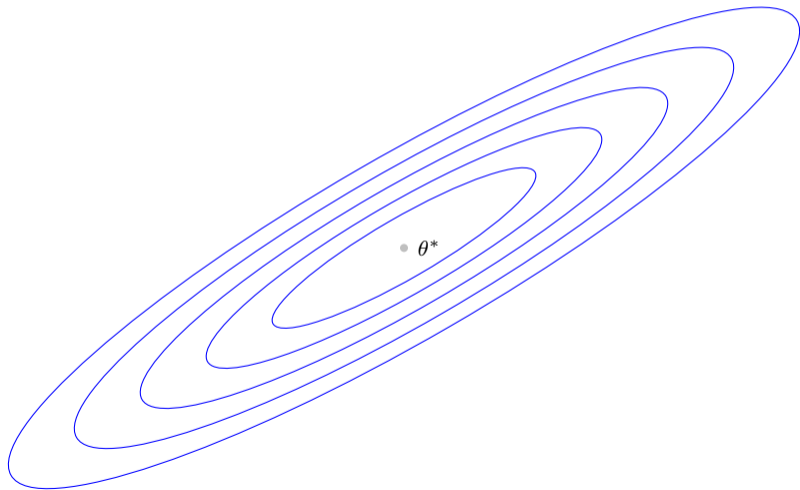
“Haute” dimension - Bon pas de gradient



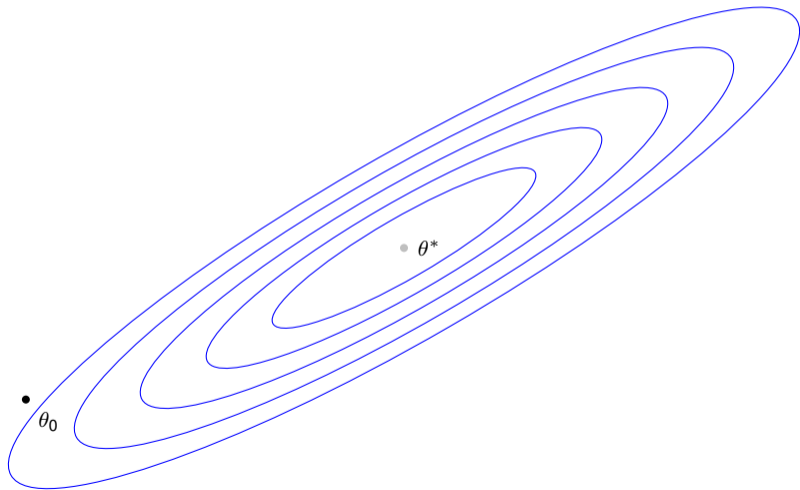
“Haute” dimension - Bon pas de gradient



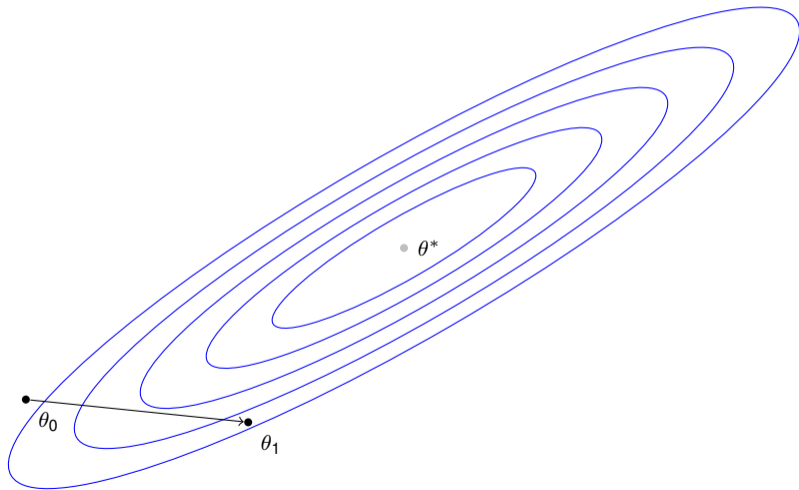
“Haute” dimension - Grand pas de gradient



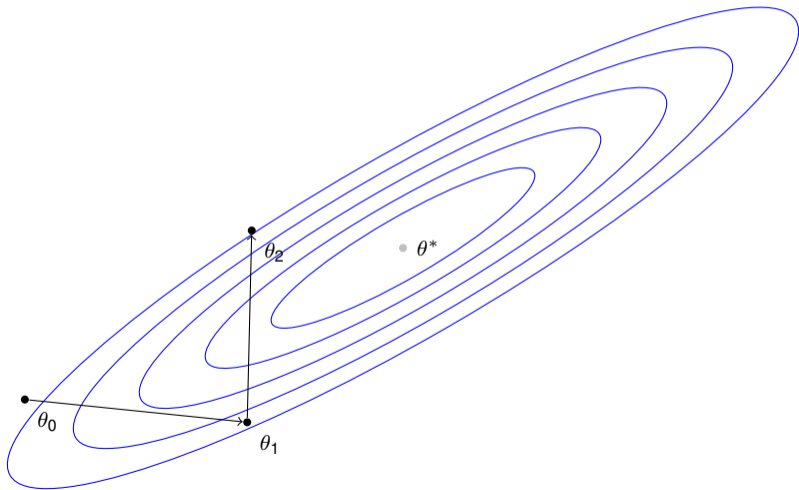
“Haute” dimension - Grand pas de gradient



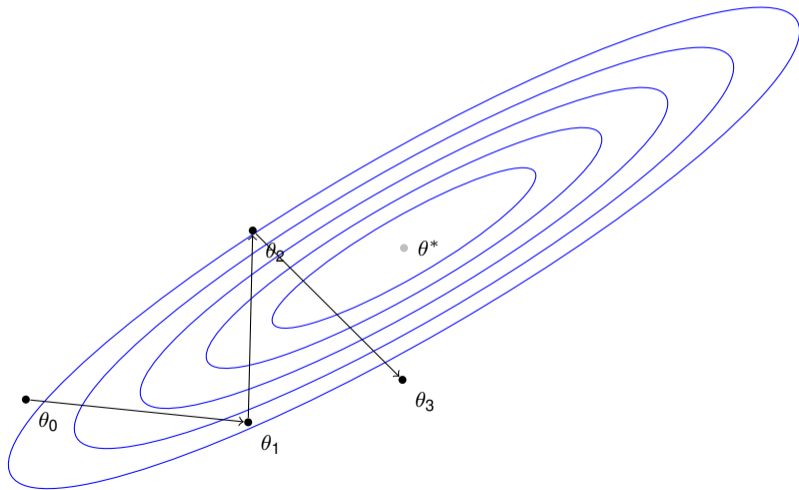
“Haute” dimension - Grand pas de gradient



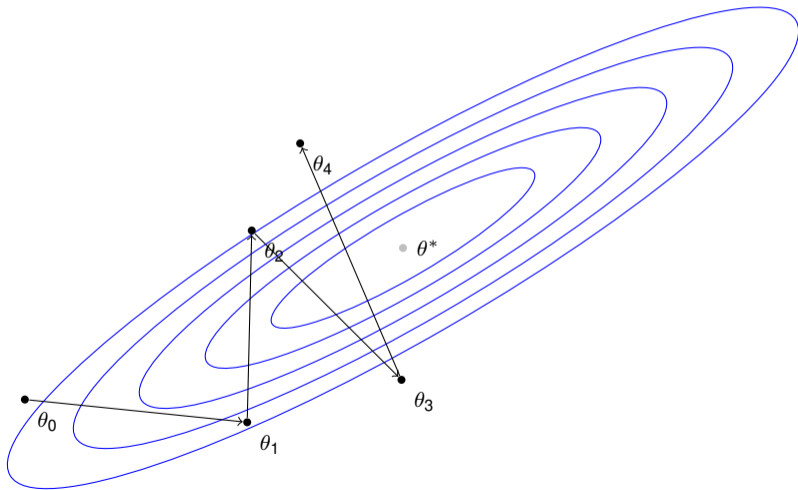
“Haute” dimension - Grand pas de gradient



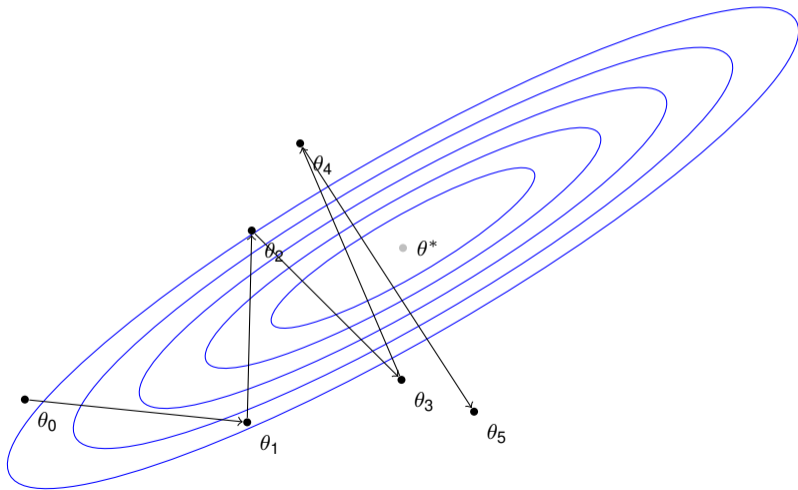
“Haute” dimension - Grand pas de gradient



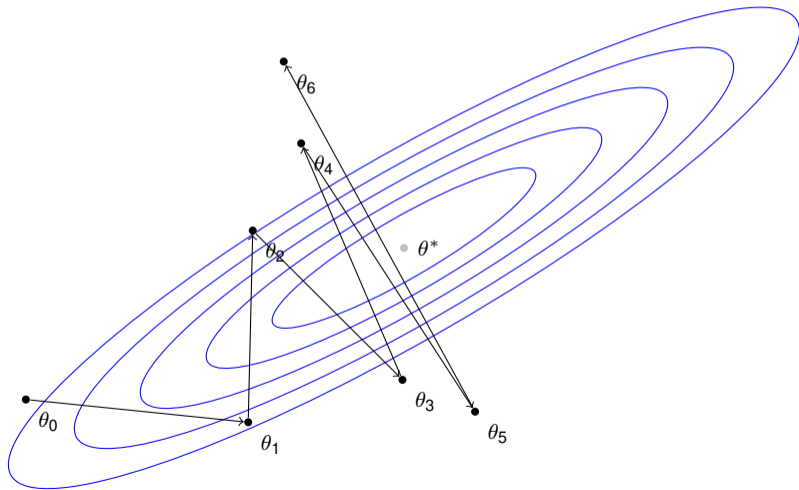
“Haute” dimension - Grand pas de gradient



“Haute” dimension - Grand pas de gradient



“Haute” dimension - Grand pas de gradient



Le conditionnement

- α doit être petit pour les directions “courbées”
- Progrès très faible dans les directions “plates”

Le conditionnement

- α doit être petit pour les directions “courbées”
- Progrès très faible dans les directions “plates”
- Le rapport de courbure s’appelle **conditionnement**

Le conditionnement

- α doit être petit pour les directions “courbées”
- Progrès très faible dans les directions “plates”
- Le rapport de courbure s’appelle **conditionnement**
- Certaines méthodes cherchent à *reconditionner*

Un reconditionnement simple

- Exemple $x \in \mathbb{R}^d$
- Modèle linéaire: $f(\theta, x) = \sum_j \theta_{(j)} x_{(j)}$
- La courbure est proportionnelle à l'amplitude de x_i

Un reconditionnement simple

- Exemple $x \in \mathbb{R}^d$
- Modèle linéaire: $f(\theta, x) = \sum_j \theta_{(j)} x_{(j)}$
- La courbure est proportionnelle à l'amplitude de x_i
- Solution: normaliser les données.

Difficulté 2: Stochasticité

- Calculer le gradient exact coûte cher
- Peut-on s'en sortir avec gradient approché ?

Difficulté 2: Stochasticité

- Calculer le gradient exact coûte cher
- Peut-on s'en sortir avec gradient approché ?
 - ▶ OUI ! [RM51]

Difficulté 2: Stochasticité

- Calculer le gradient exact coûte cher
- Peut-on s'en sortir avec gradient approché ?
 - ▶ OUI ! [RM51]
- On a souvent $f(\theta) = \frac{1}{N} \sum_i f_i(\theta)$
- On ne va calculer le gradient que pour une seule f_i

Propriétés du gradient stochastique

- L'algorithme "nécessite" un pas décroissant.
- En théorie, $1/t$. En pratique, souvent moins vite.

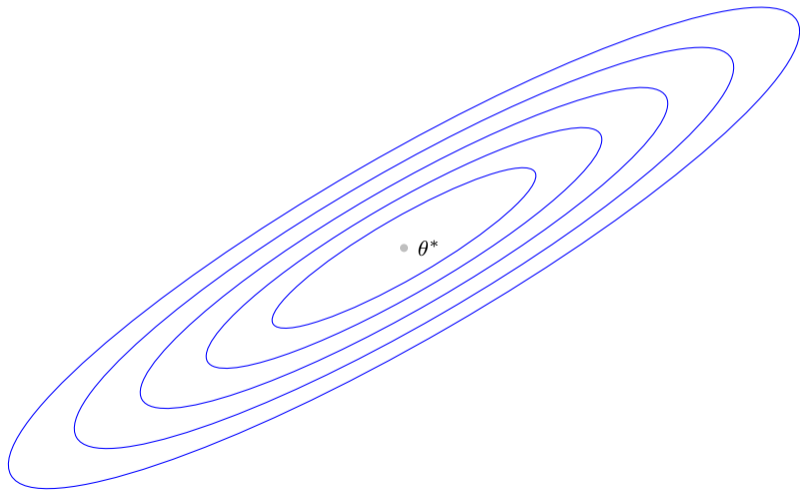
Propriétés du gradient stochastique

- L'algorithme "nécessite" un pas décroissant.
- En théorie, $1/t$. En pratique, souvent moins vite.
- La convergence est plus lente que le gradient batch.
- Cela n'est vrai que pour l'erreur d'entraînement.

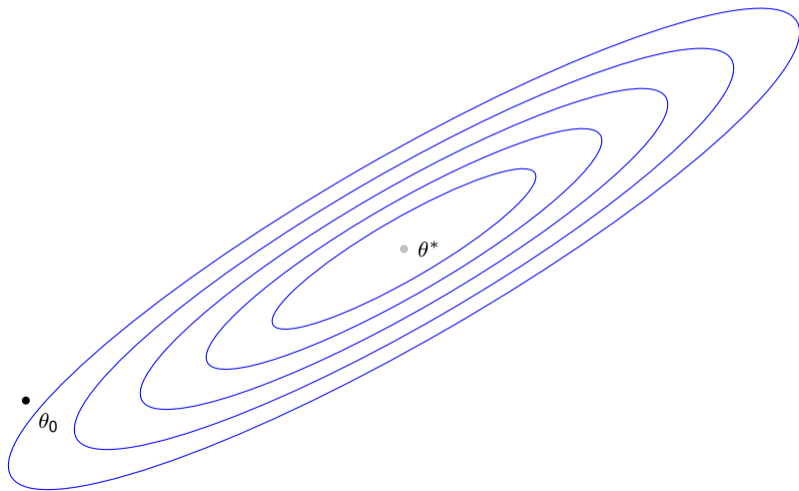
Propriétés du gradient stochastique

- L'algorithme "nécessite" un pas décroissant.
- En théorie, $1/t$. En pratique, souvent moins vite.
- La convergence est plus lente que le gradient batch.
- Cela n'est vrai que pour l'erreur d'entraînement.
- Même convergence sur l'erreur de test [BB08].
 - ▶ Convergence en nombre de *mises à jour*.

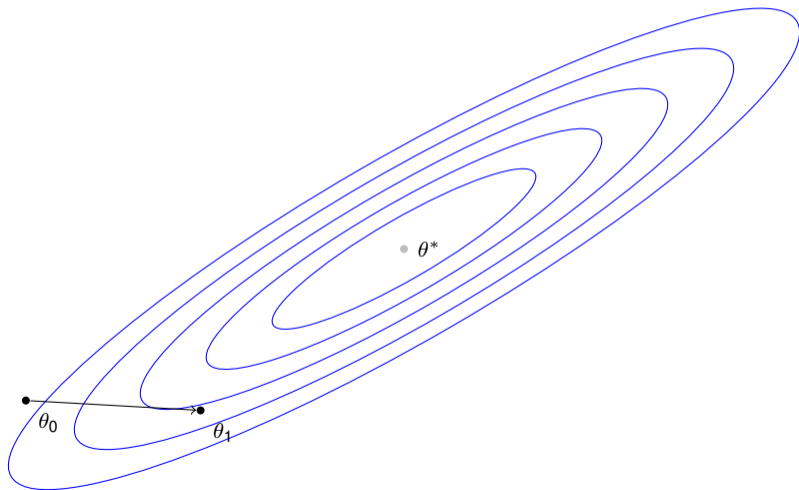
Gradient stochastique - Pas constant



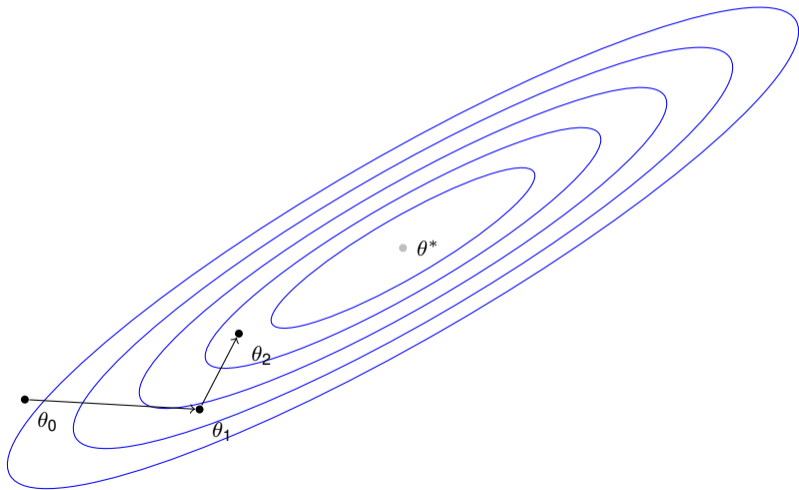
Gradient stochastique - Pas constant



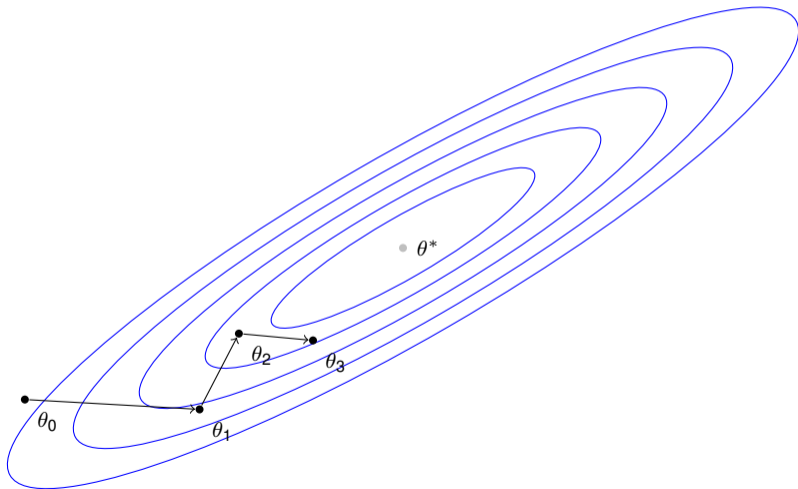
Gradient stochastique - Pas constant



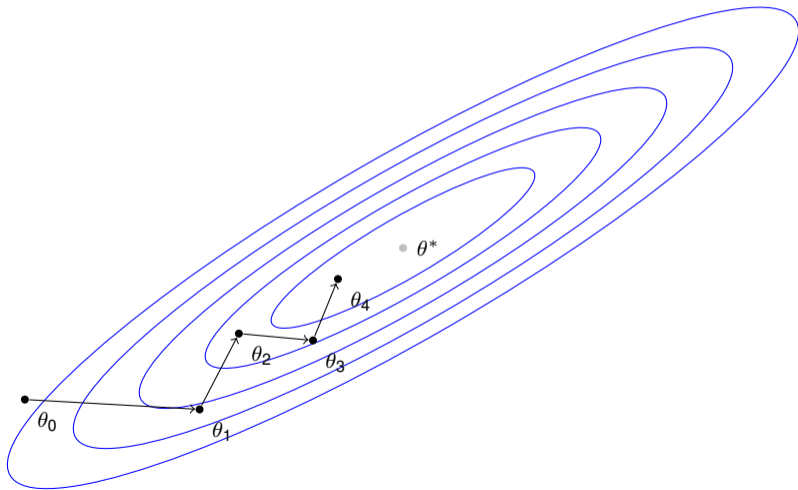
Gradient stochastique - Pas constant



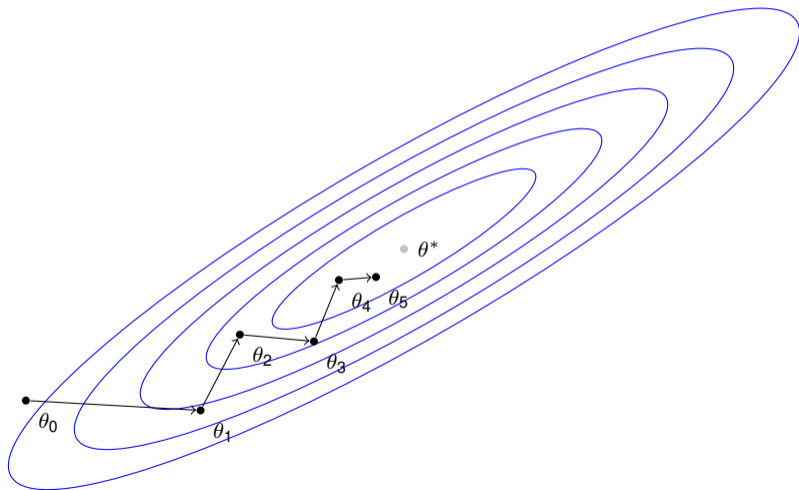
Gradient stochastique - Pas constant



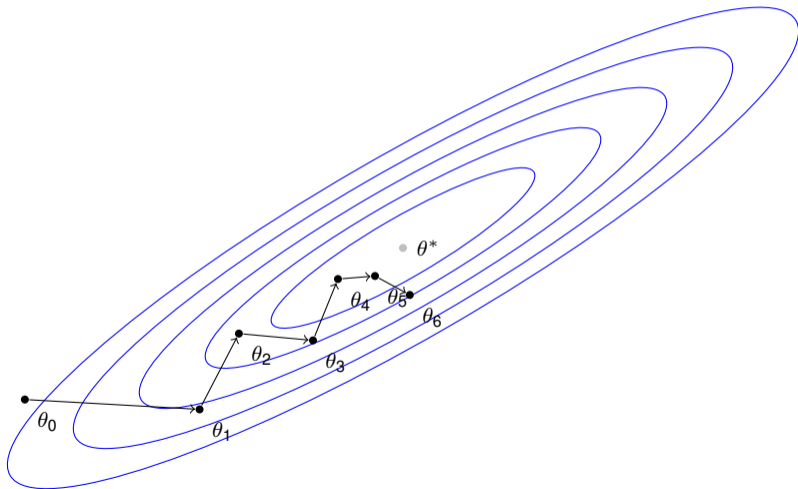
Gradient stochastique - Pas constant



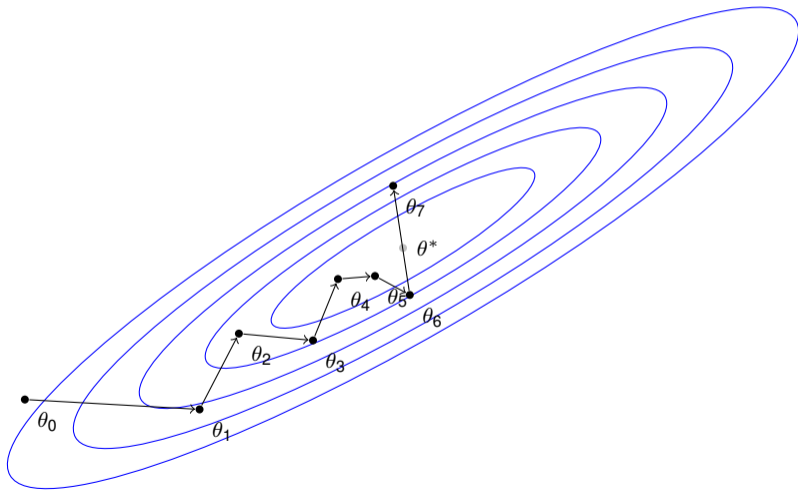
Gradient stochastique - Pas constant



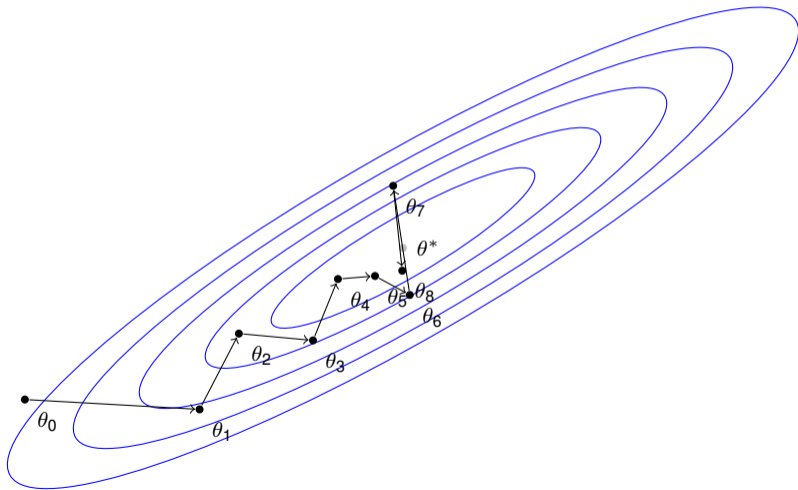
Gradient stochastique - Pas constant



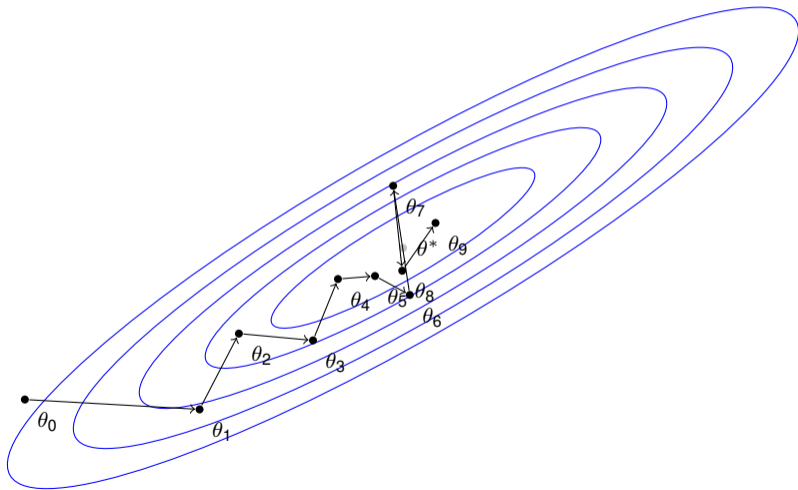
Gradient stochastique - Pas constant



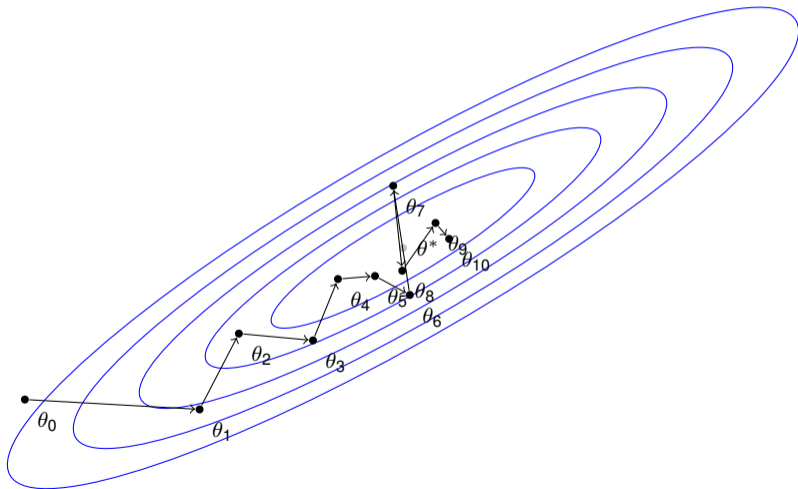
Gradient stochastique - Pas constant



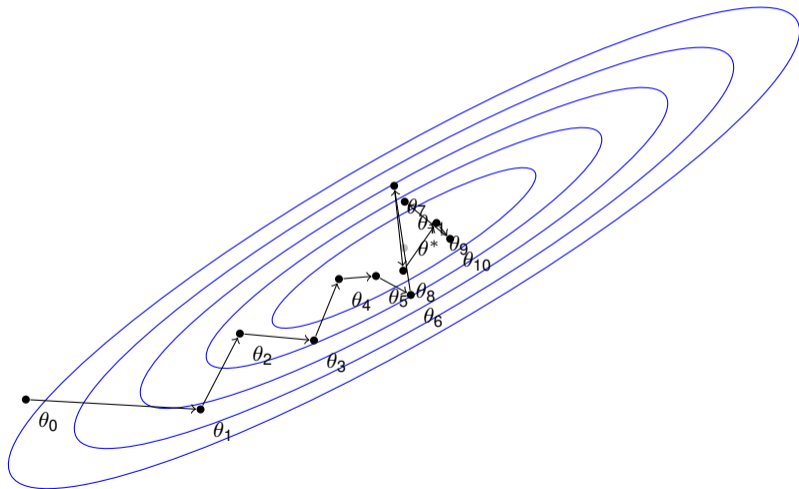
Gradient stochastique - Pas constant



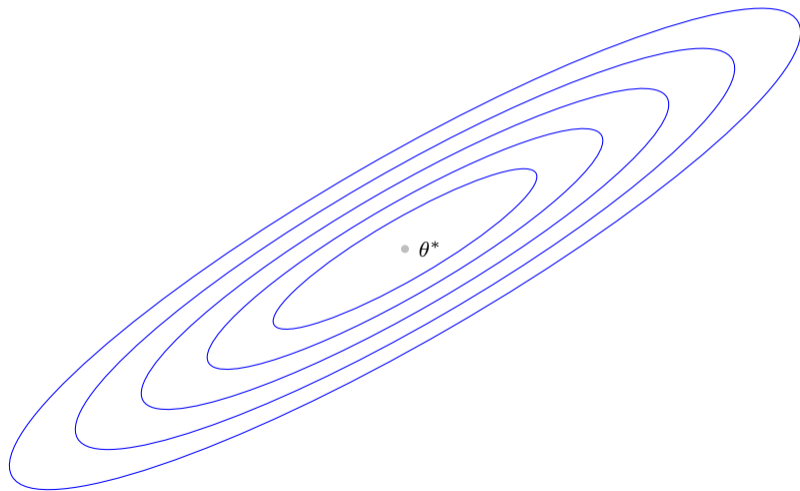
Gradient stochastique - Pas constant



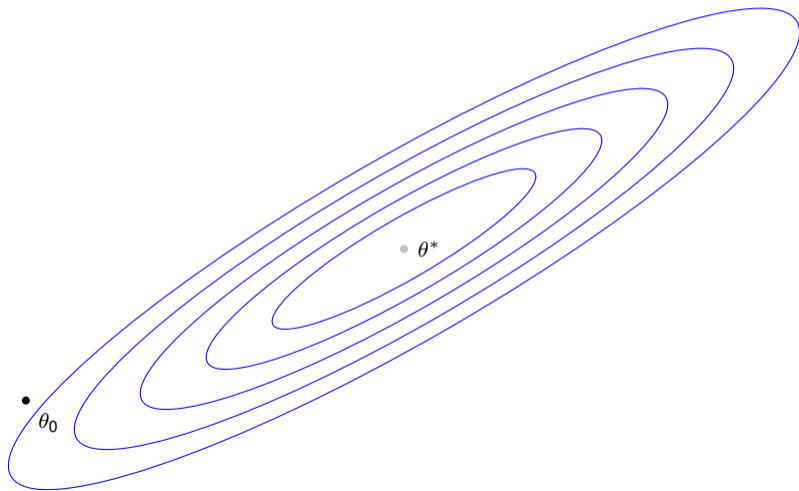
Gradient stochastique - Pas constant



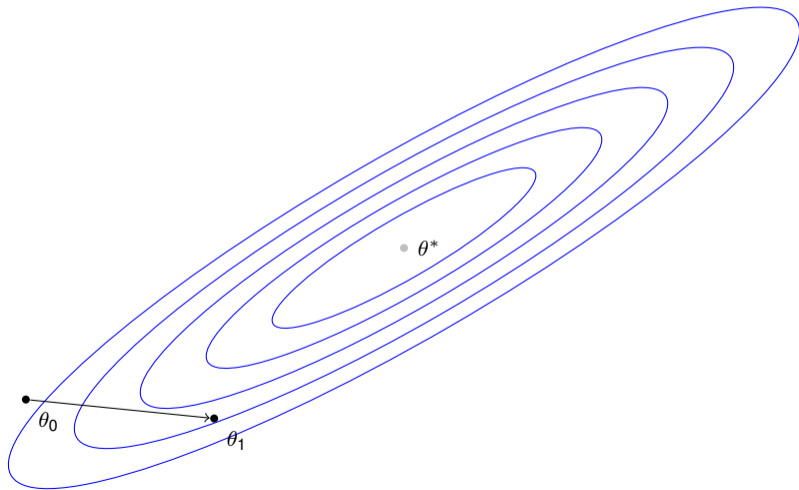
Gradient stochastique - Pas décroissant



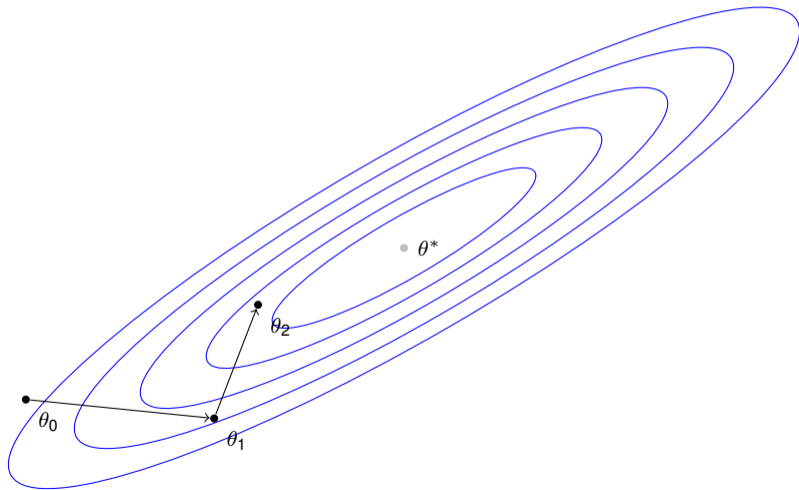
Gradient stochastique - Pas décroissant



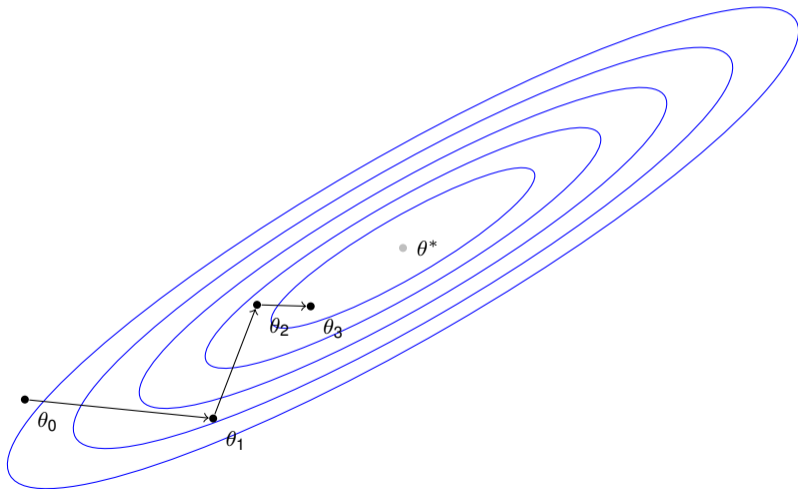
Gradient stochastique - Pas décroissant



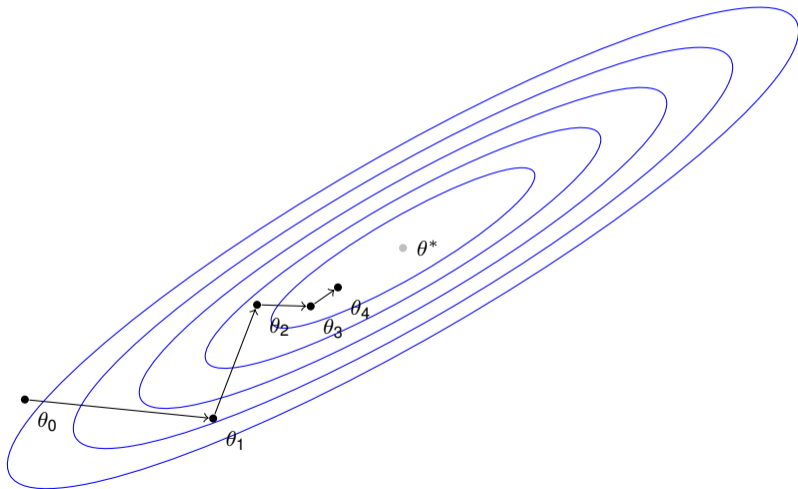
Gradient stochastique - Pas décroissant



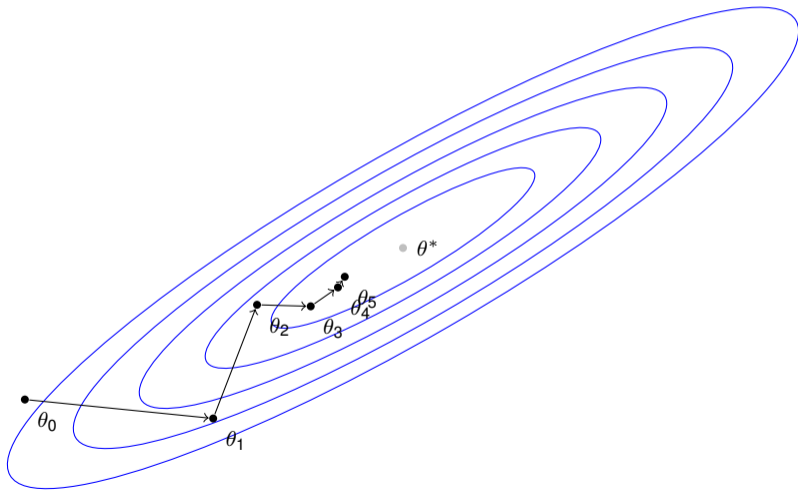
Gradient stochastique - Pas décroissant



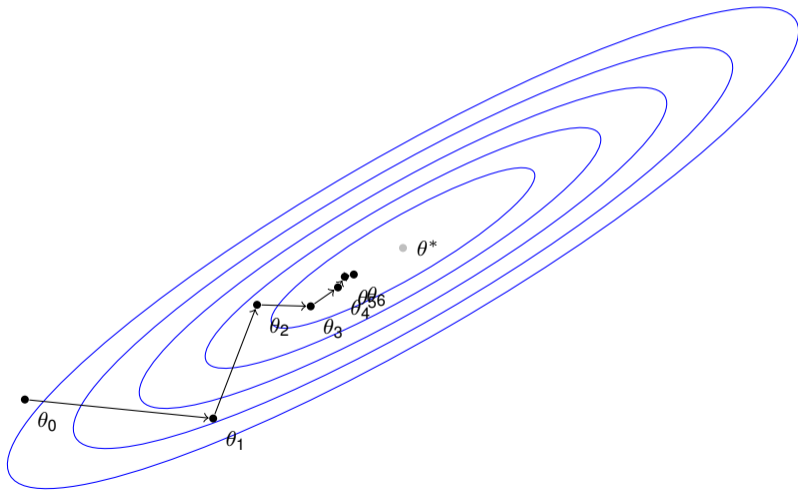
Gradient stochastique - Pas décroissant



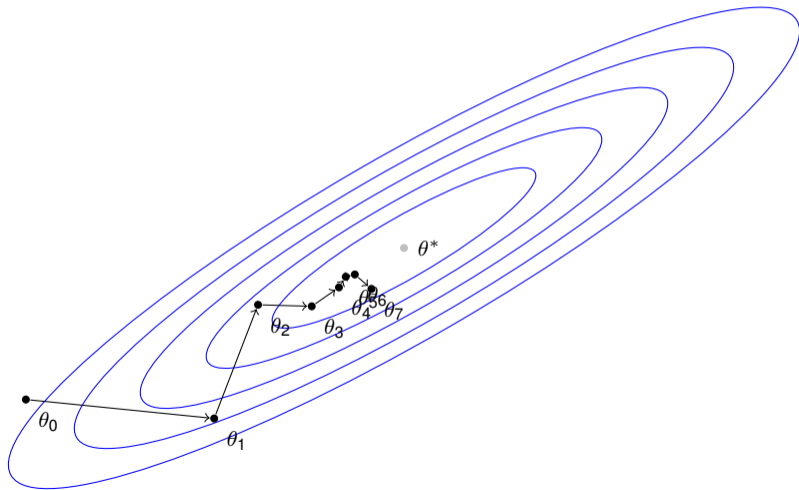
Gradient stochastique - Pas décroissant



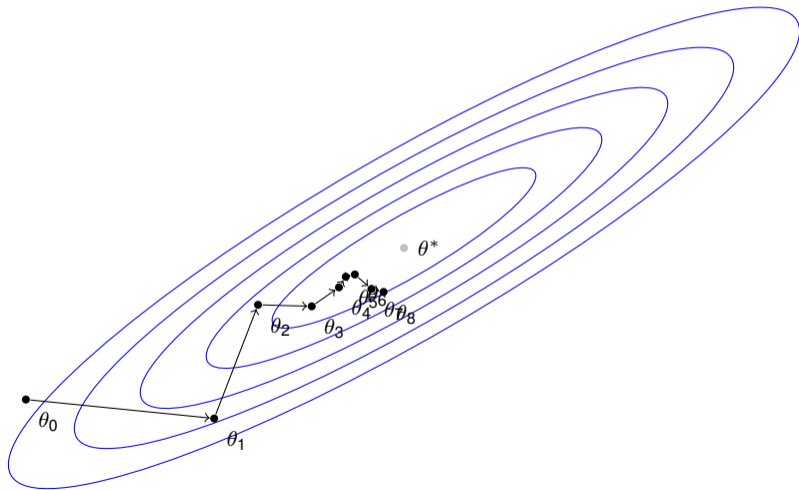
Gradient stochastique - Pas décroissant



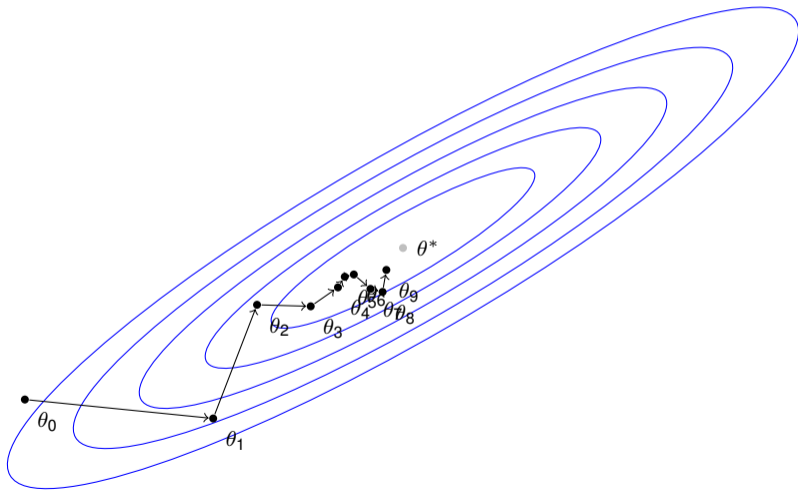
Gradient stochastique - Pas décroissant



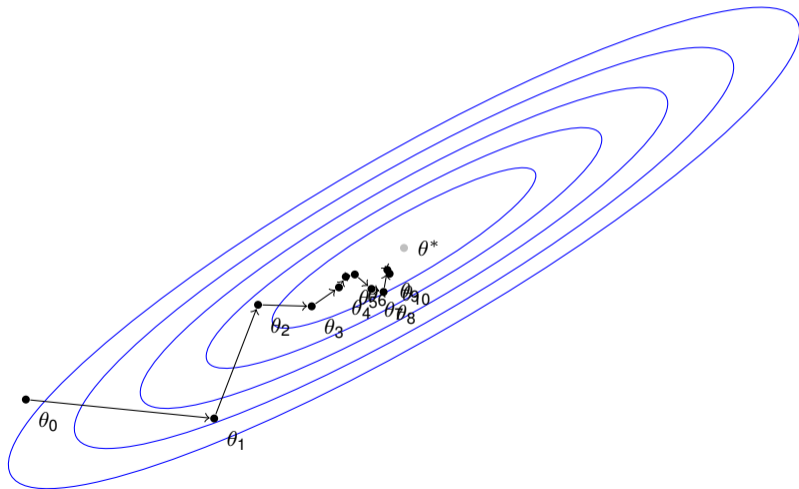
Gradient stochastique - Pas décroissant



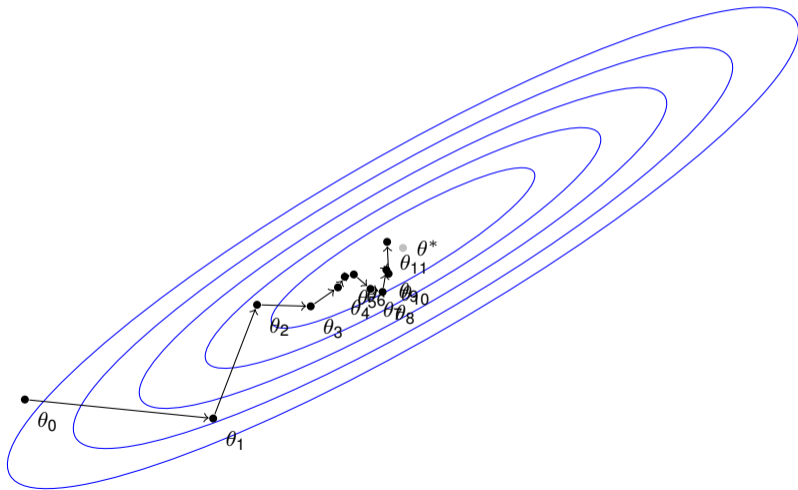
Gradient stochastique - Pas décroissant



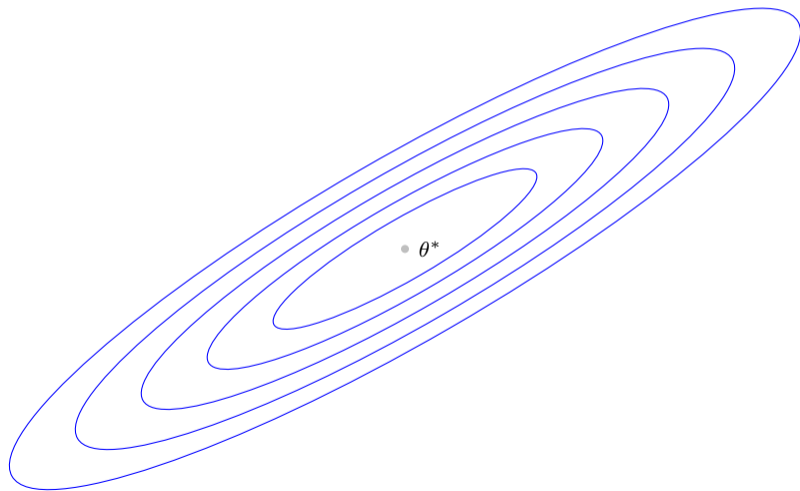
Gradient stochastique - Pas décroissant



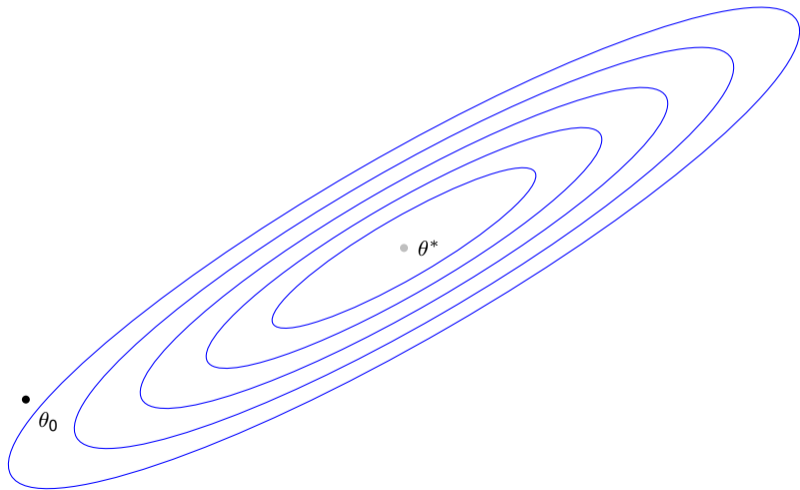
Gradient stochastique - Pas décroissant



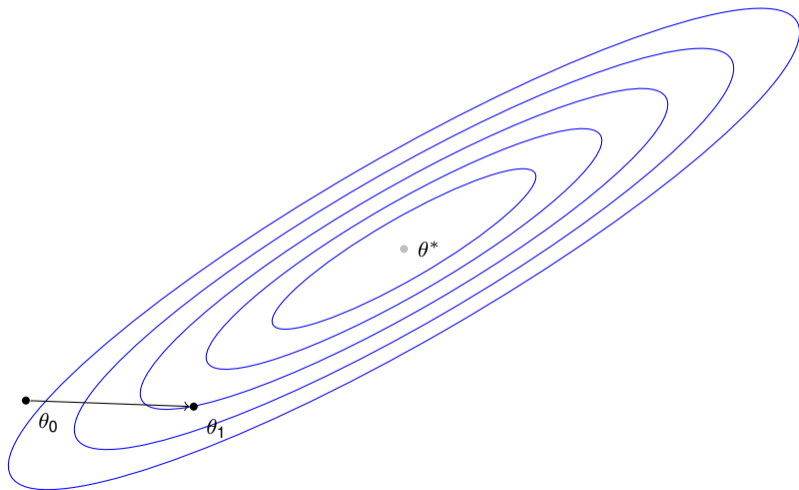
Gradient stochastique - Pas très décroissant



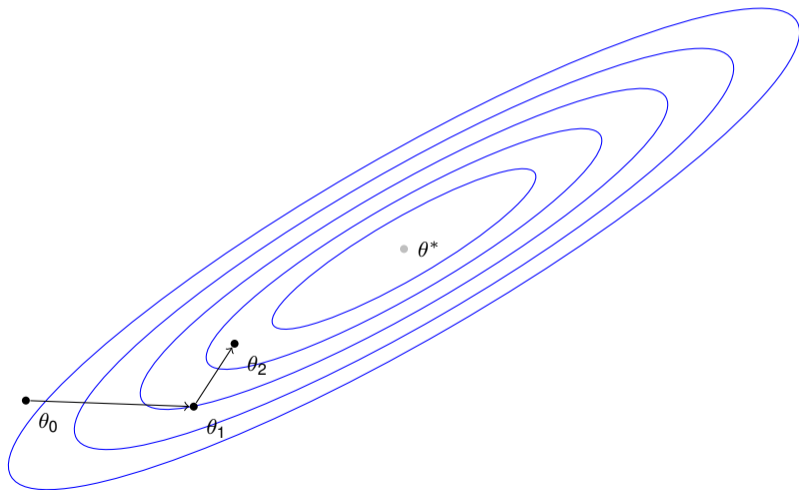
Gradient stochastique - Pas très décroissant



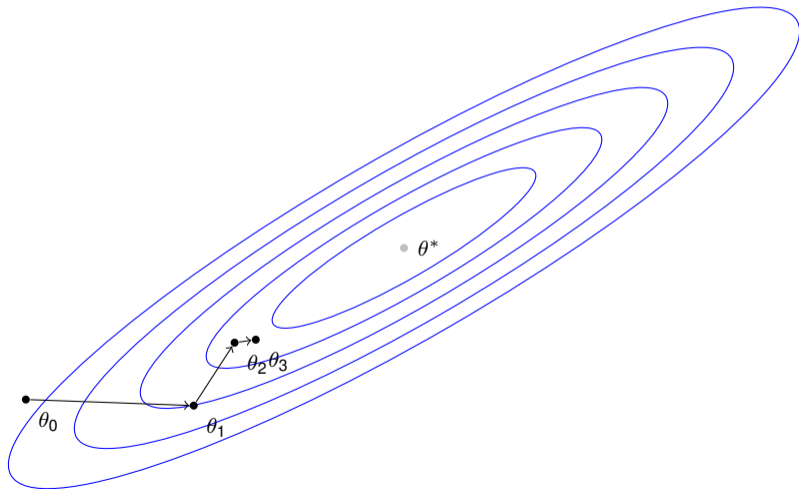
Gradient stochastique - Pas très décroissant



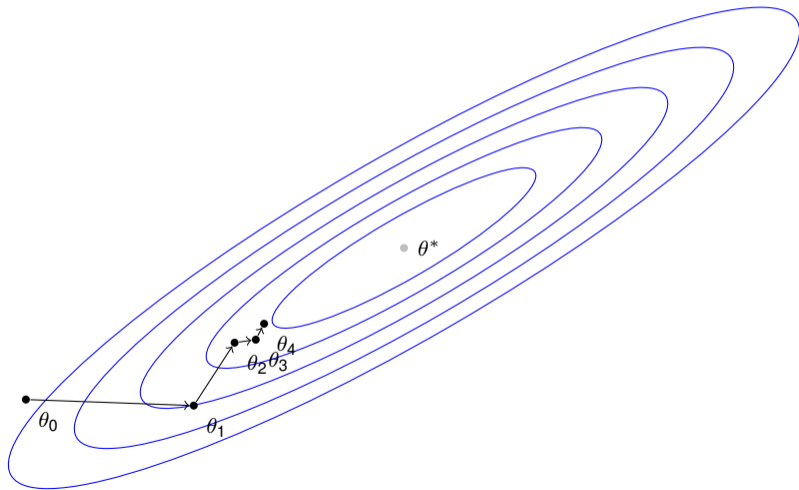
Gradient stochastique - Pas très décroissant



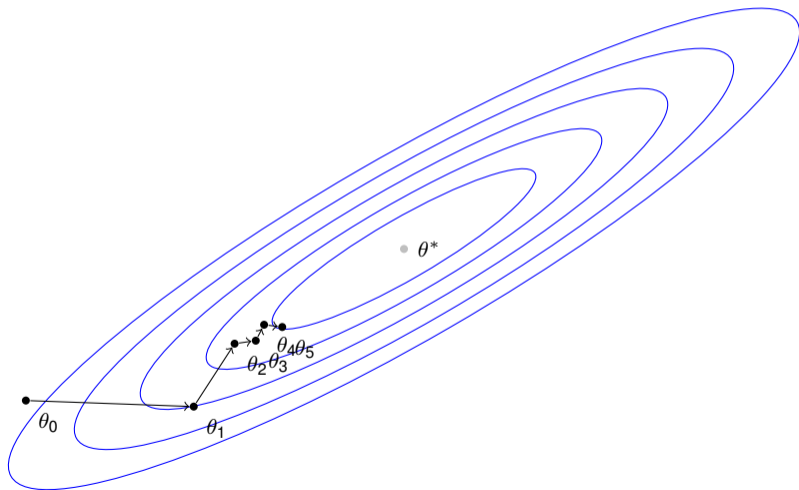
Gradient stochastique - Pas très décroissant



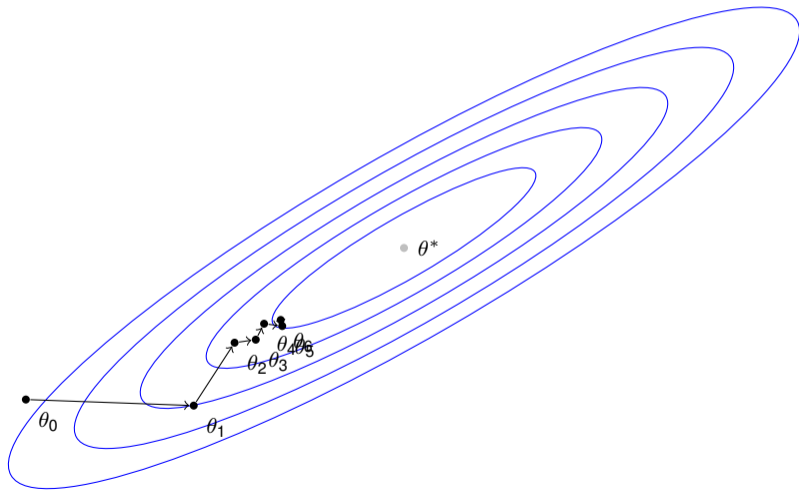
Gradient stochastique - Pas très décroissant



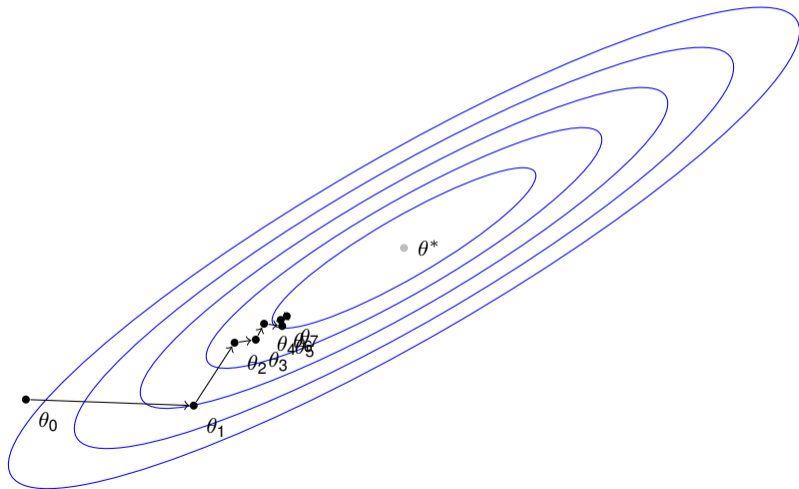
Gradient stochastique - Pas très décroissant



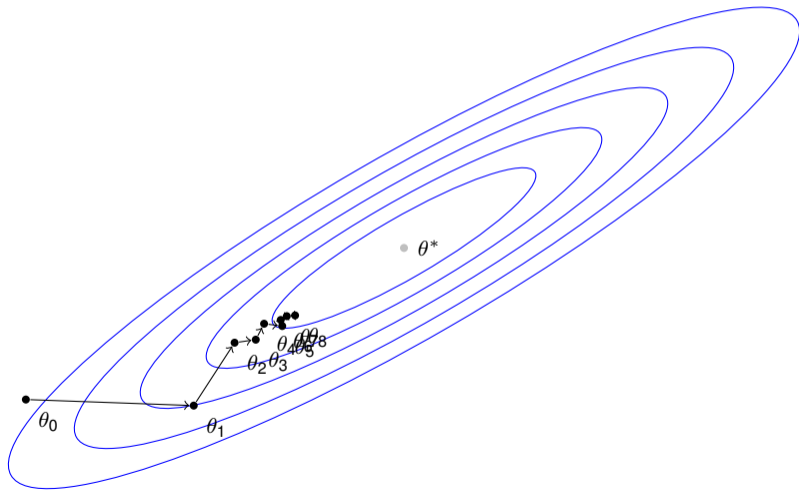
Gradient stochastique - Pas très décroissant



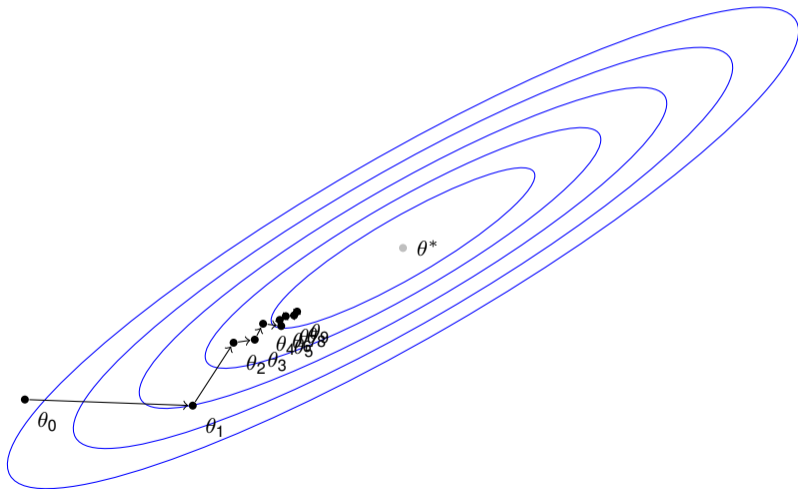
Gradient stochastique - Pas très décroissant



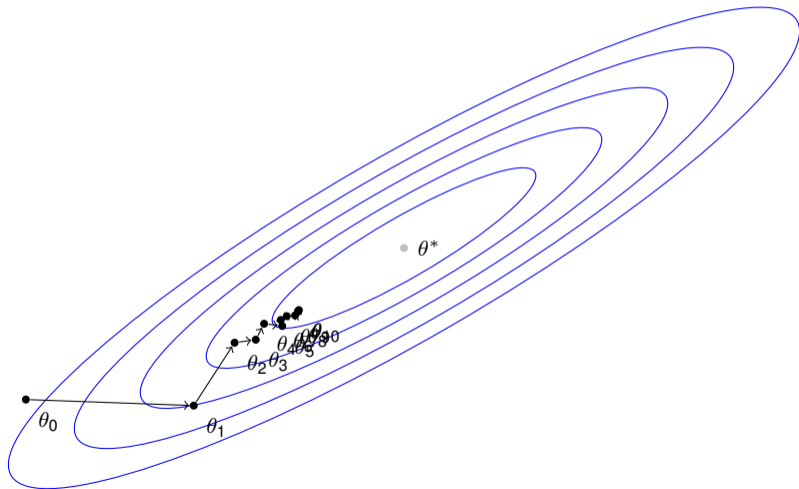
Gradient stochastique - Pas très décroissant



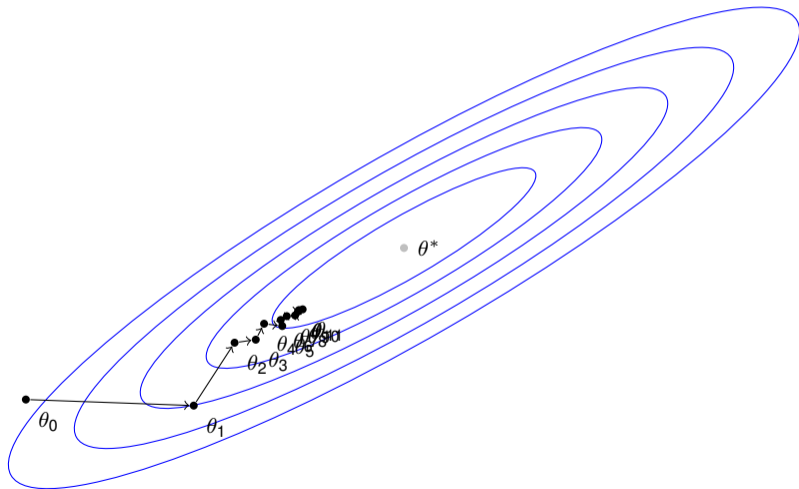
Gradient stochastique - Pas très décroissant



Gradient stochastique - Pas très décroissant



Gradient stochastique - Pas très décroissant

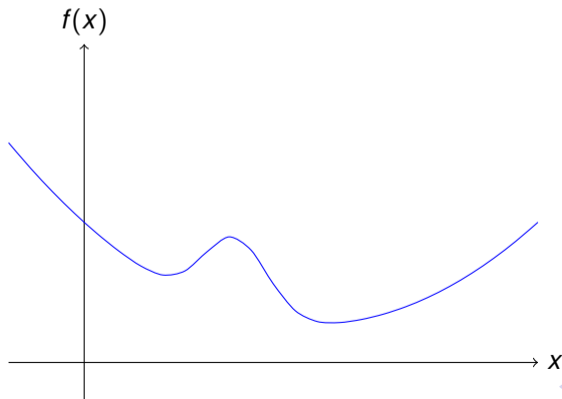


Difficulté 3: non-convexité

- Preuves de convergence valides dans le cas convexe
- Les réseaux profonds sont non-convexes
- Théorie encore balbutiante mais des résultats empiriques

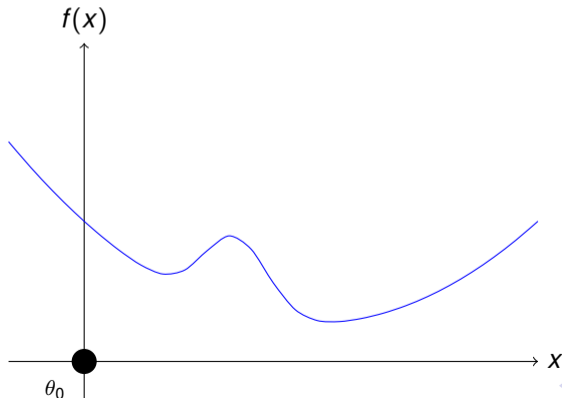
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



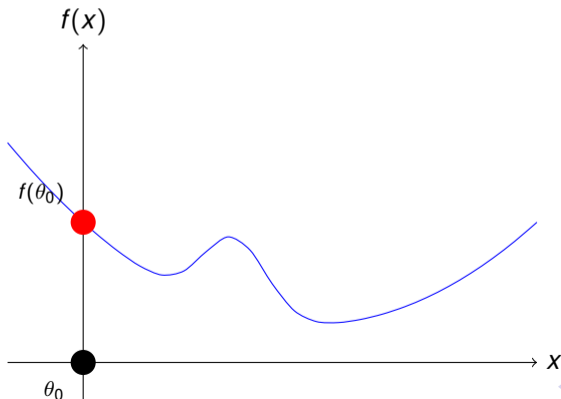
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



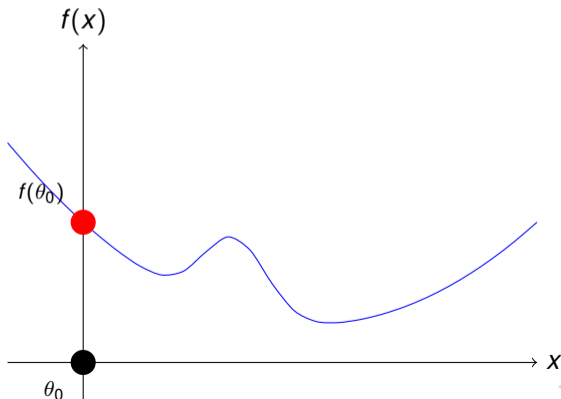
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



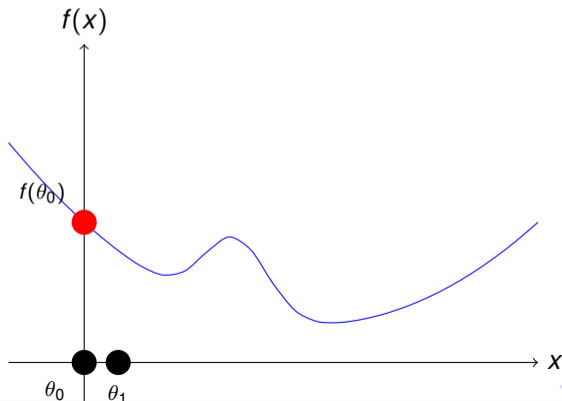
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



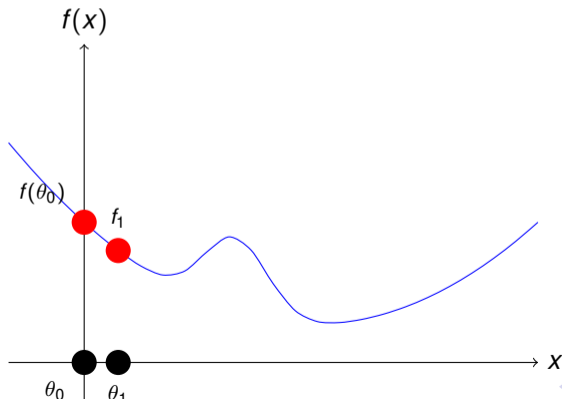
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



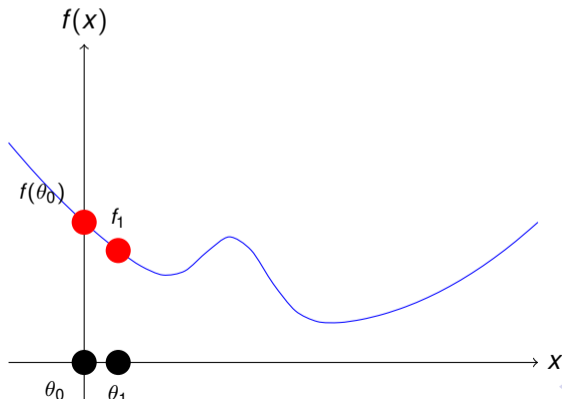
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



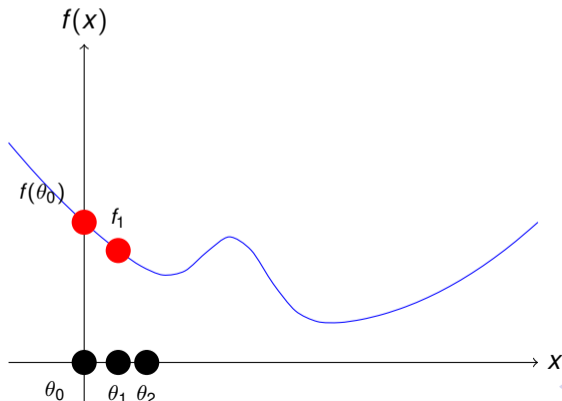
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



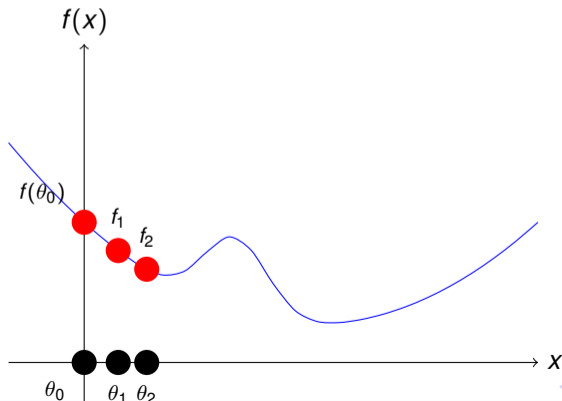
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



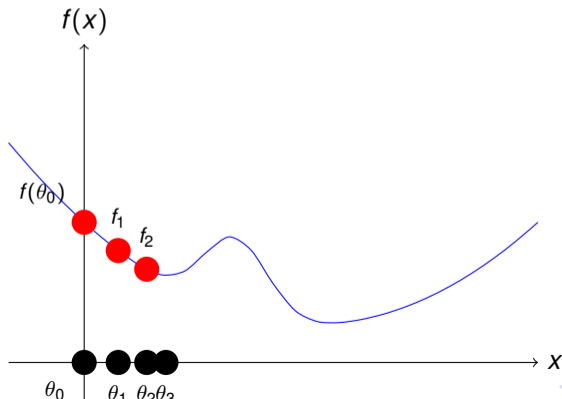
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



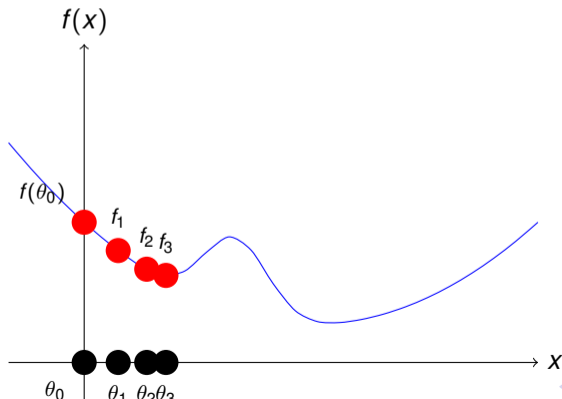
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



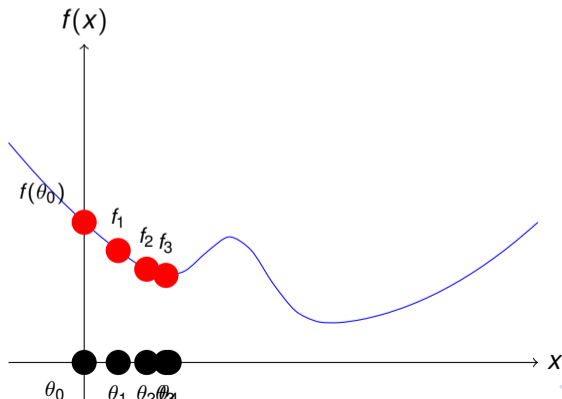
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



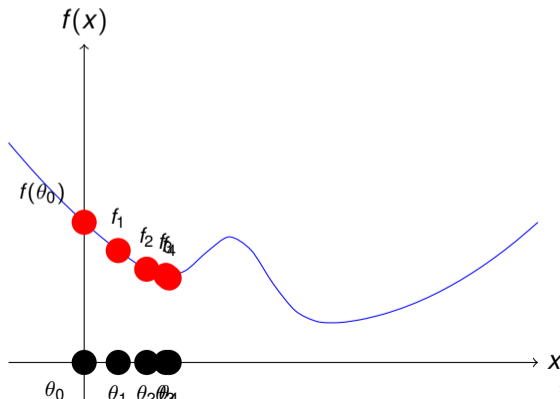
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



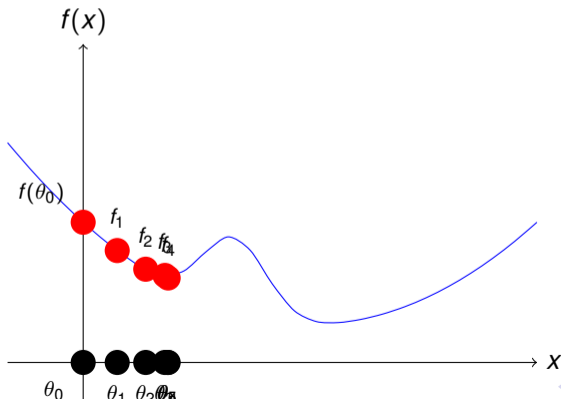
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



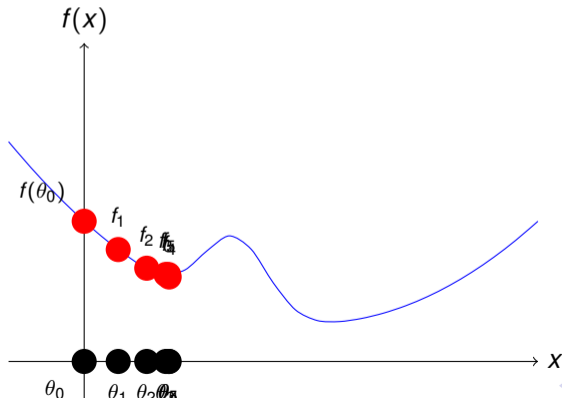
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



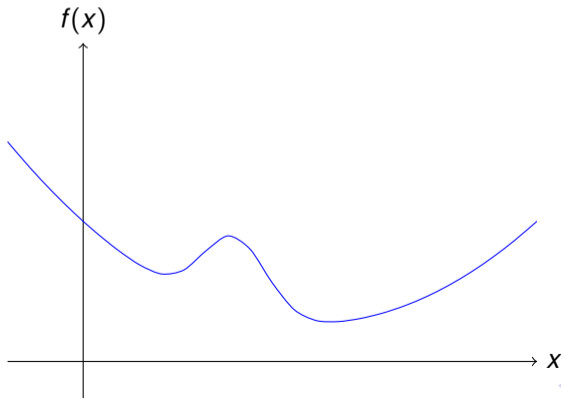
Fonction non-convexe - Petit pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



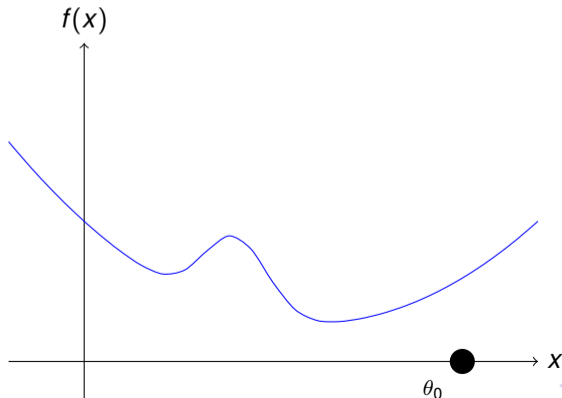
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



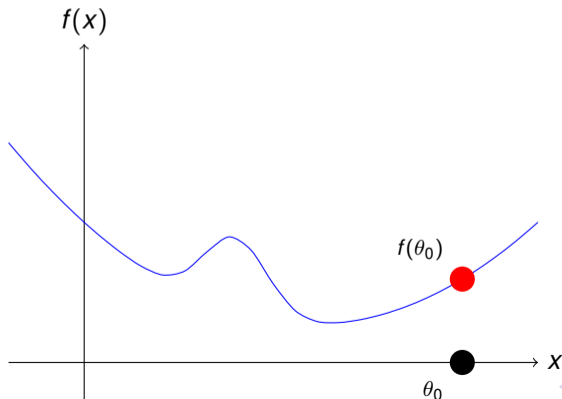
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



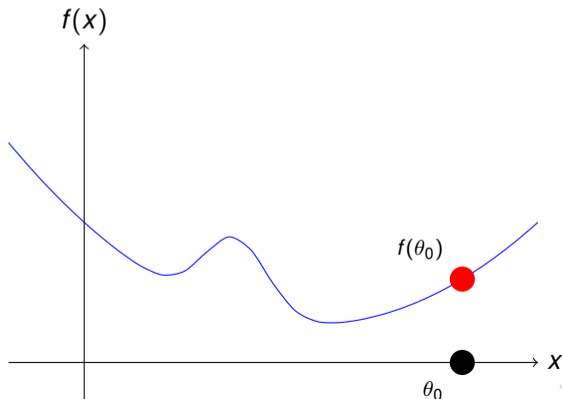
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



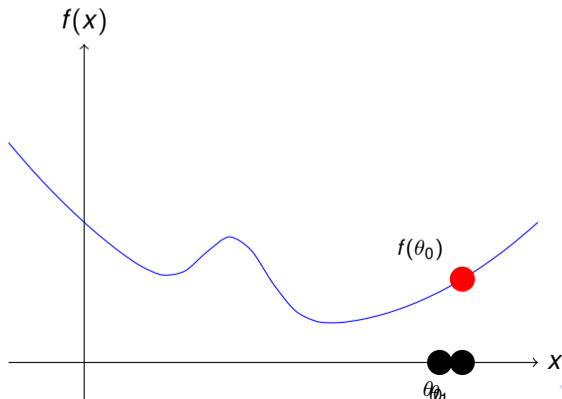
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



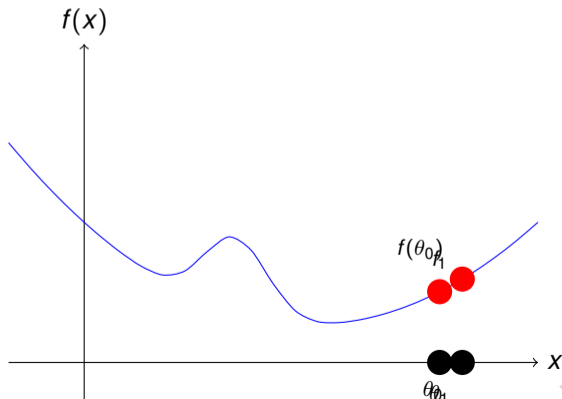
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



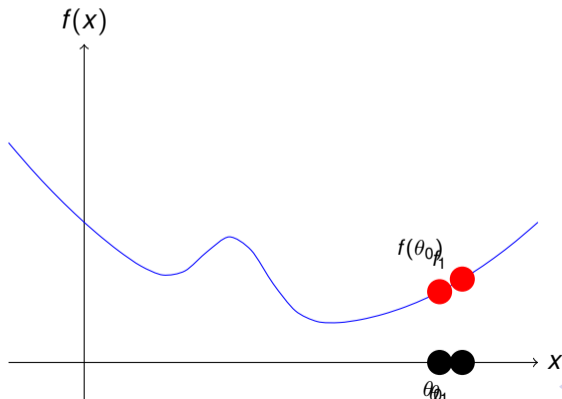
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



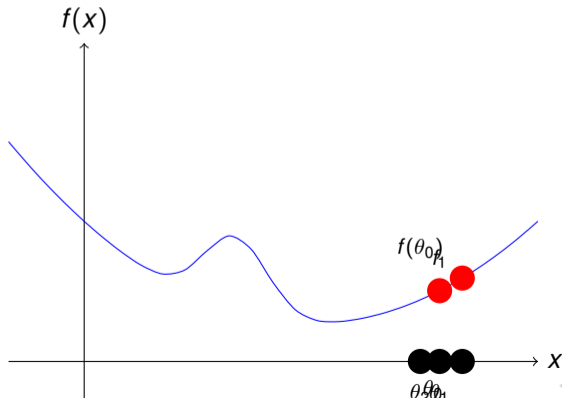
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



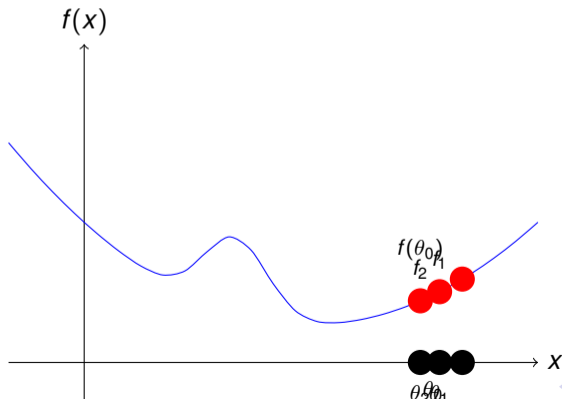
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



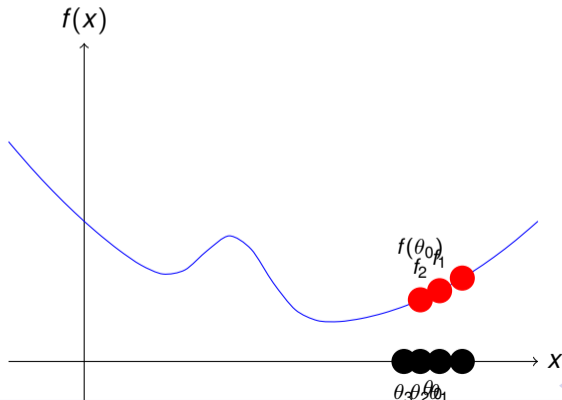
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



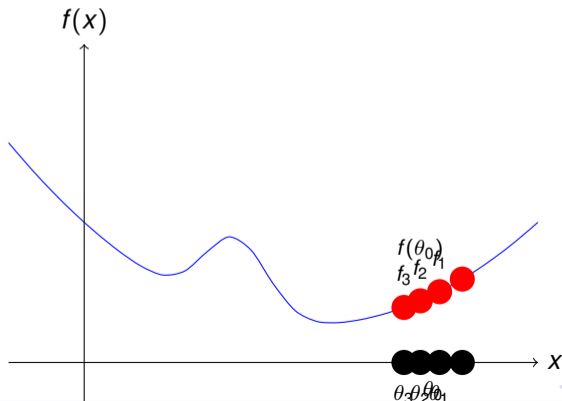
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



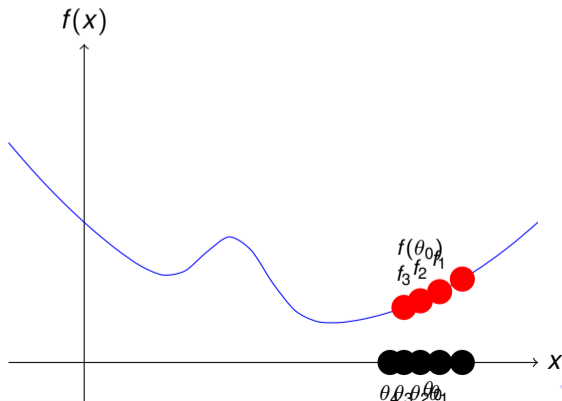
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



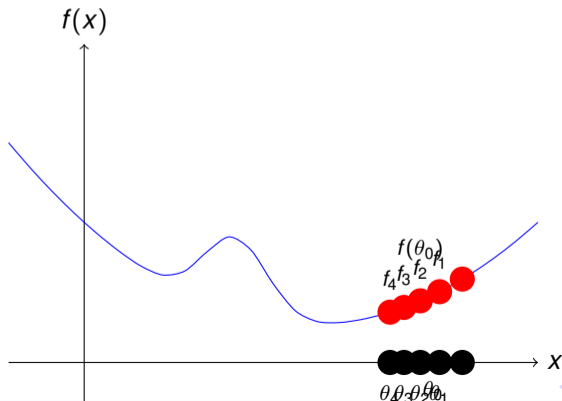
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



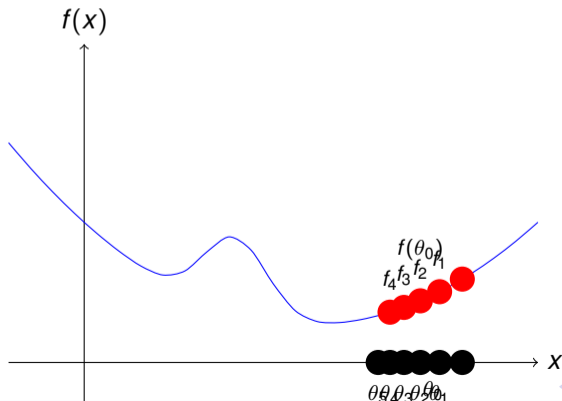
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



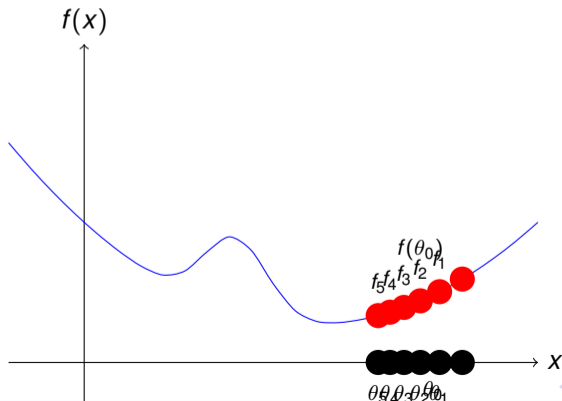
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



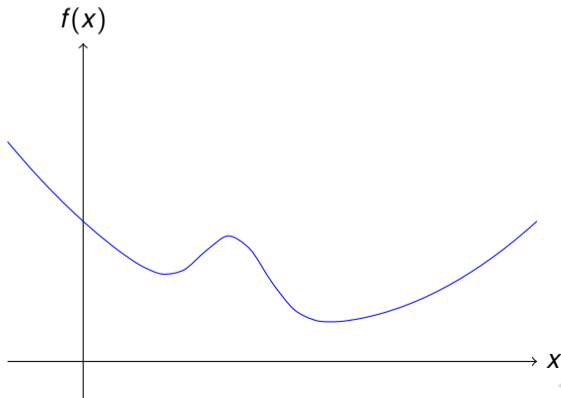
Fonction non-convexe - Petit pas (2)

$$\theta_0 = 5 \quad \theta_{t+1} = \theta_t - \frac{1}{2} \frac{df(x)}{dx}$$



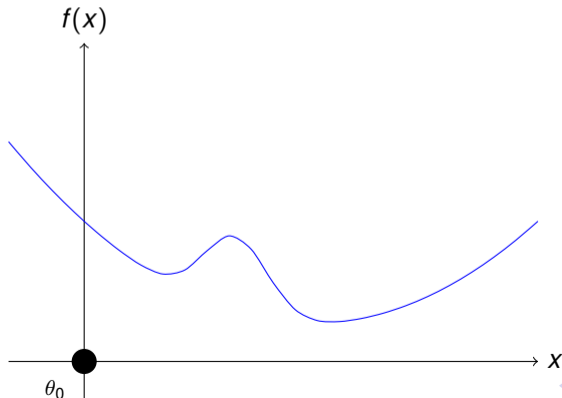
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



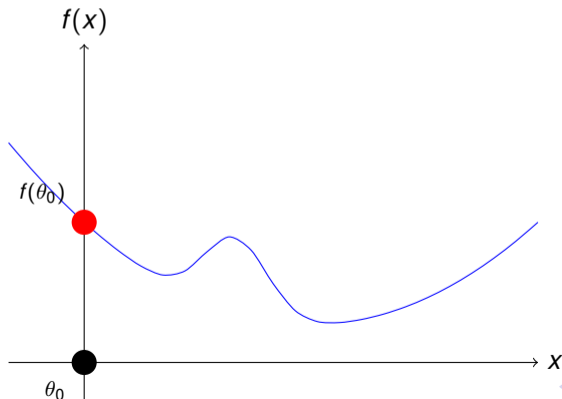
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



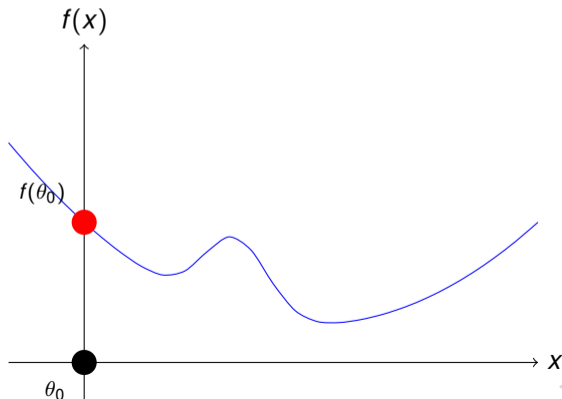
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



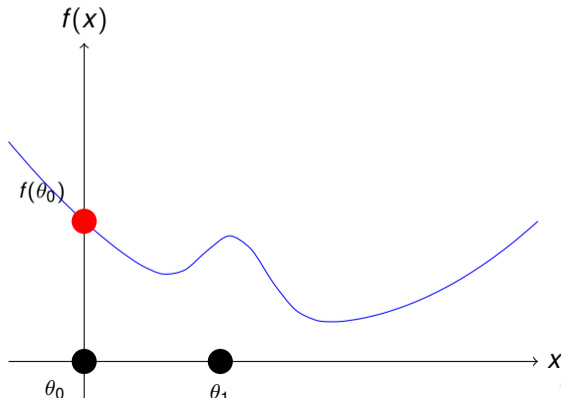
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



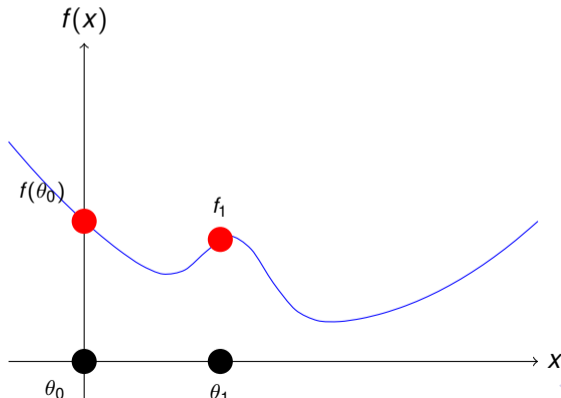
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



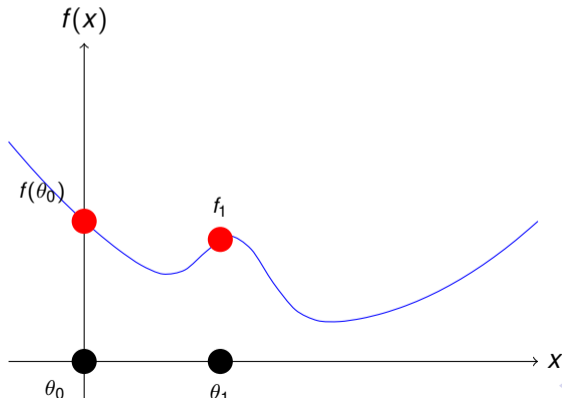
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



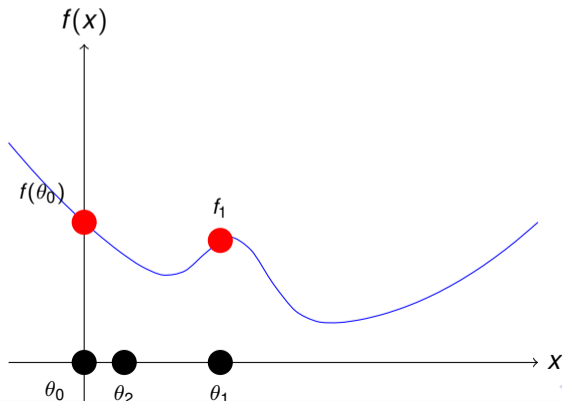
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



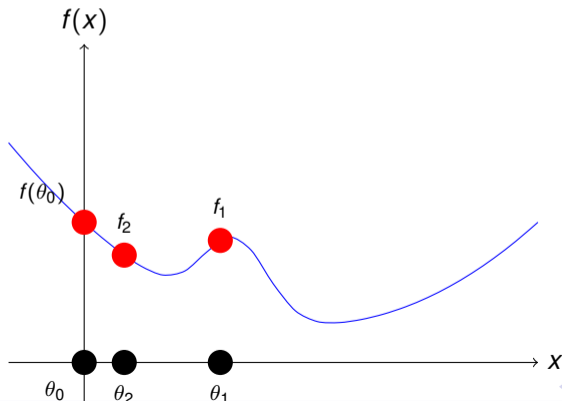
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



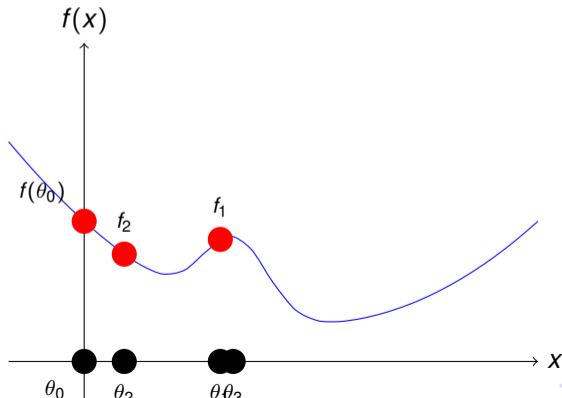
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



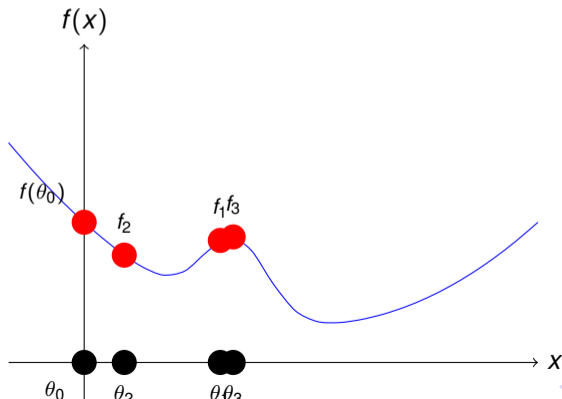
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



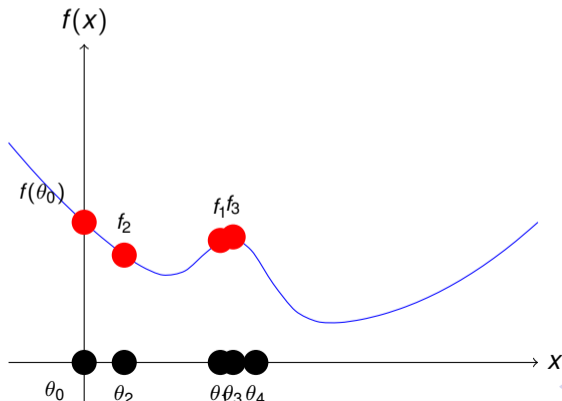
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



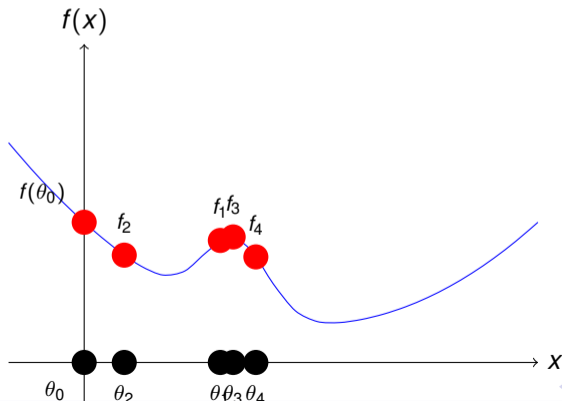
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



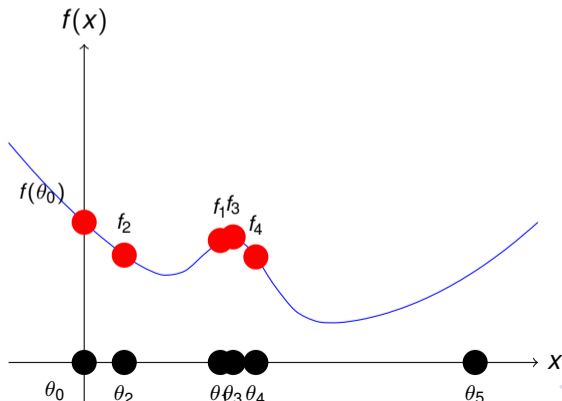
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



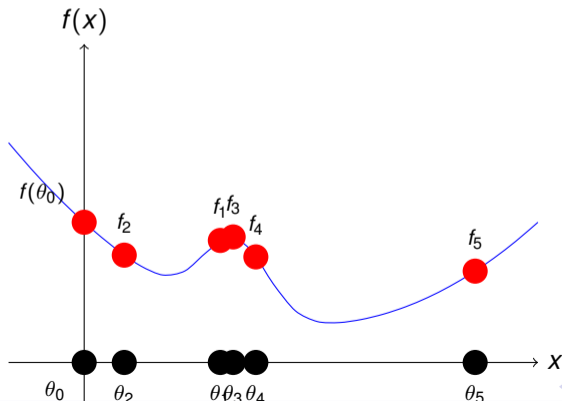
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



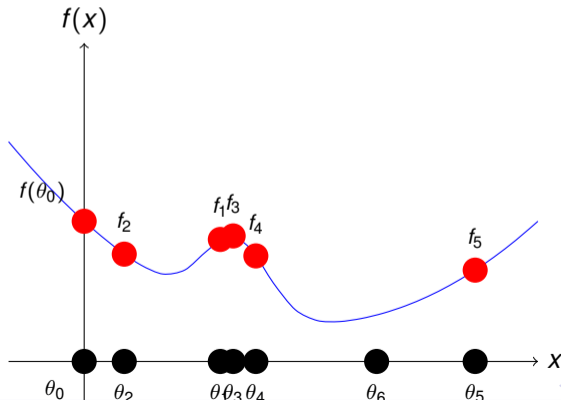
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



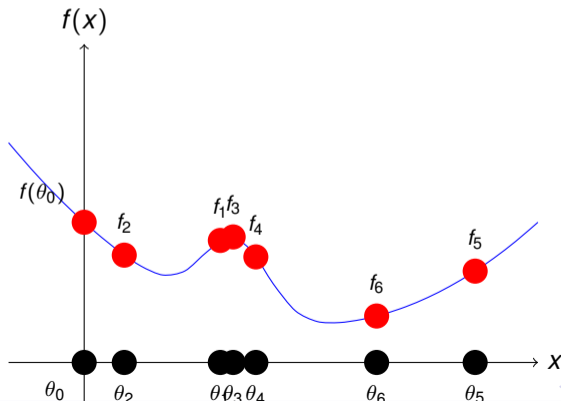
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



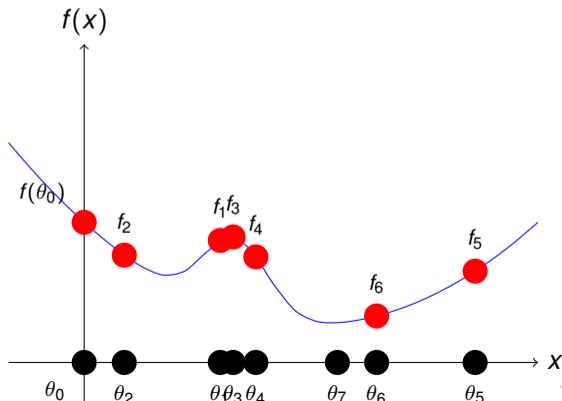
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



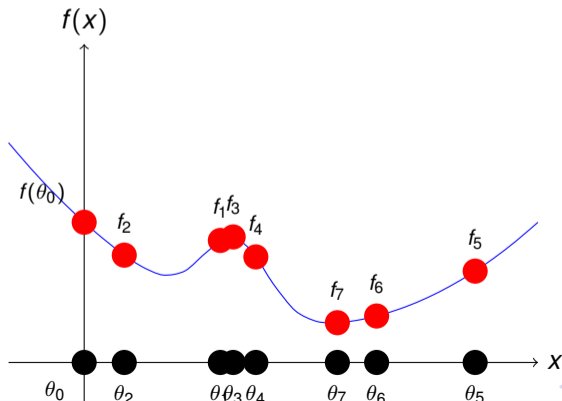
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



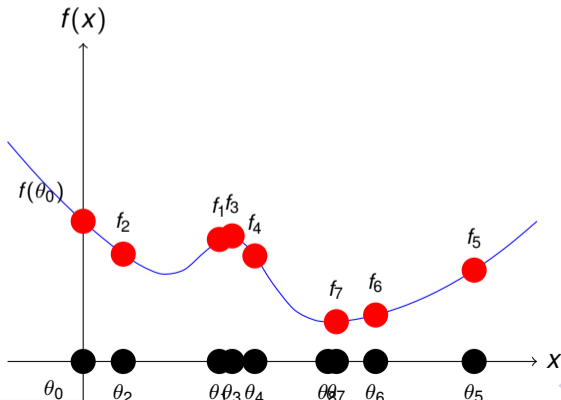
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



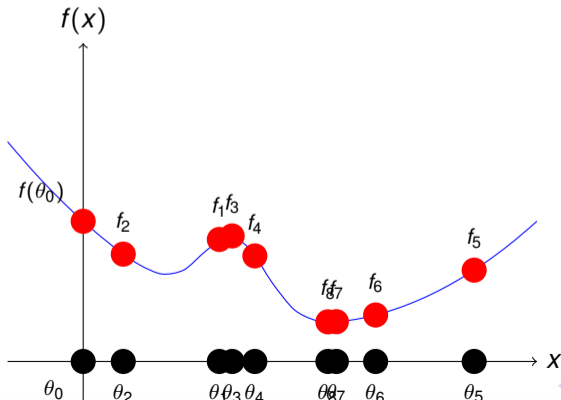
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



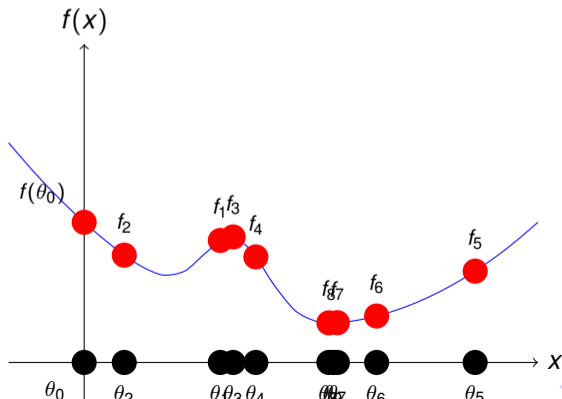
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



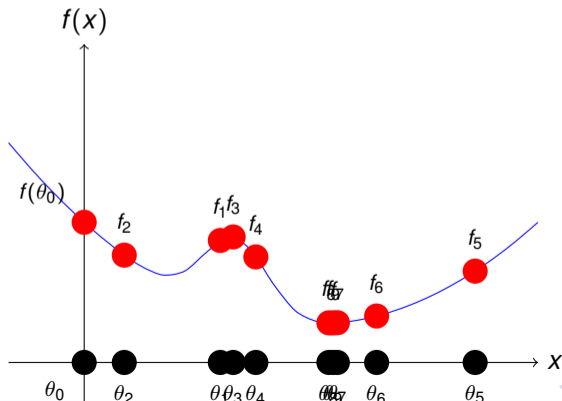
Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



Fonction non-convexe - Grand pas

$$\theta_0 = 0 \quad \theta_{t+1} = \theta_t - 2 \frac{df(x)}{dx}$$



Non-convexité - À retenir

- La solution dépend de l'initialisation [Ben+07]
- La solution dépend de l'optimiseur [Wil+17]

Non-convexité - À retenir

- La solution dépend de l'initialisation [Ben+07]
- La solution dépend de l'optimiseur [Wil+17]
- Hormis ces deux points, intuition en 1D complètement fausse

Non-convexité en haute dimension

- Plateaux
- Points-selle [Dau+14; AHV17]
- Peu de mauvais minima locaux, beaucoup se ressemblent [Cho+15]

- 1 Optimisation simple
- 2 Trois difficultés
- 3 Trouver un bon optimiseur**
- 4 Commentaires additionnels

Qualités d'un bon optimiseur

- Rapide (en temps de calcul)
- Facile à régler
- Gère bien le conditionnement
- Gère les plateaux et points-selle

Qualités d'un bon optimiseur

- Rapide (en temps de calcul)
- Facile à régler
- Gère bien le conditionnement
- Gère les plateaux et points-selle
- Aucun algorithme ne fait tout ça aujourd'hui

Idée 1: Utiliser plusieurs pas de gradient

- Idée: identifier la courbure dans chaque direction et utiliser un pas de gradient différent pour chacune

Idée 1: Utiliser plusieurs pas de gradient

- Idée: identifier la courbure dans chaque direction et utiliser un pas de gradient différent pour chacune
- Plusieurs méthodes mais Gauss-Newton est la plus populaire [LeC+98]

Idée 1: Utiliser plusieurs pas de gradient

- Idée: identifier la courbure dans chaque direction et utiliser un pas de gradient différent pour chacune
- Plusieurs méthodes mais Gauss-Newton est la plus populaire [LeC+98]
- De nombreuses variantes existent [DHS11; KB14]

Idée 2: Blanchir les données

- Blanchir les données aide pour un modèle linéaire

Idée 2: Blanchir les données

- Blanchir les données aide pour un modèle linéaire
- Les activations des couches cachées ne sont plus blanchies
- Les blanchir à leur tour aide le conditionnement

Idée 2: Blanchir les données

- Blanchir les données aide pour un modèle linéaire
- Les activations des couches cachées ne sont plus blanchies
- Les blanchir à leur tour aide le conditionnement
- Cette méthode s'appelle Batch-Norm [IS15]

Idée 3: Utiliser la mémoire

- Résoudre le mauvais conditionnement
 - ▶ On veut réduire α dans les directions courbées
 - ▶ On veut augmenter α dans les directions plates

Idée 3: Utiliser la mémoire

- Résoudre le mauvais conditionnement
 - ▶ On veut réduire α dans les directions courbées
 - ▶ On veut augmenter α dans les directions plates
- Limiter le bruit de la stochasticité
 - ▶ On veut aller vite quand tous les gradients sont d'accord
 - ▶ On veut aller lentement quand les gradients sont en désaccord

Idée 3: Utiliser la mémoire

- Résoudre le mauvais conditionnement
 - ▶ On veut réduire α dans les directions courbées
 - ▶ On veut augmenter α dans les directions plates
- Limiter le bruit de la stochasticité
 - ▶ On veut aller vite quand tous les gradients sont d'accord
 - ▶ On veut aller lentement quand les gradients sont en désaccord
- Ces deux propriétés demandent de la mémoire.

Momentum [Nes83]

- Momentum garde une mémoire des gradients passés
- Ajout d'un terme de vitesse v_t

Momentum [Nes83]

- Momentum garde une mémoire des gradients passés
- Ajout d'un terme de vitesse v_t
 - ▶ Initialisation: $v_0 = 0, \theta_0$.

Momentum [Nes83]

- Momentum garde une mémoire des gradients passés
- Ajout d'un terme de vitesse v_t
 - ▶ Initialisation: $v_0 = 0, \theta_0$.
 - ▶ Calcul du gradient: $g(\theta_t) = f'(\theta_t)$

Momentum [Nes83]

- Momentum garde une mémoire des gradients passés
- Ajout d'un terme de vitesse v_t
 - ▶ Initialisation: $v_0 = 0, \theta_0$.
 - ▶ Calcul du gradient: $g(\theta_t) = f'(\theta_t)$
 - ▶ Mise à jour de la vitesse: $v_t = \beta v_{t-1} + g(\theta_t)$

Momentum [Nes83]

- Momentum garde une mémoire des gradients passés
- Ajout d'un terme de vitesse v_t
 - ▶ Initialisation: $v_0 = 0, \theta_0$.
 - ▶ Calcul du gradient: $g(\theta_t) = f'(\theta_t)$
 - ▶ Mise à jour de la vitesse: $v_t = \beta v_{t-1} + g(\theta_t)$
 - ▶ Mise à jour: $\theta_{t+1} = \theta_t - \alpha v_t$

Momentum [Nes83]

- Momentum garde une mémoire des gradients passés
- Ajout d'un terme de vitesse v_t
 - ▶ Initialisation: $v_0 = 0, \theta_0$.
 - ▶ Calcul du gradient: $g(\theta_t) = f'(\theta_t)$
 - ▶ Mise à jour de la vitesse: $v_t = \beta v_{t-1} + g(\theta_t)$
 - ▶ Mise à jour: $\theta_{t+1} = \theta_t - \alpha v_t$
- Ajout d'un deuxième hyperparamètre.

- 1 Optimisation simple
- 2 Trois difficultés
- 3 Trouver un bon optimiseur
- 4 Commentaires additionnels**

Généralisation

- But de l'apprentissage est de bien généraliser
- Aucune différence dans le monde convexe
- Optimiseur a un impact sur la généralisation dans les réseaux profonds
- Raison encore peu claire

Robustesse: un point essentiel

- Pas les mêmes qualités nécessaires pour un entraînement unique et pour un entraînement régulier
- La difficulté de réglage est un point critique qui n'apparaît pas dans les courbes
- Attention aux modifications du pas de gradient pendant l'optimisation [TH12]
- L'erreur qui vous importe est l'erreur de généralisation, pas d'entraînement

Merci !

nlr@google.com

Références I



Sanjeev Arora, Moritz Hardt, and Nisheeth Vishnoi. "Off the convex path". In: (2017). URL: <http://www.offconvex.org>.



Yoshua Bengio et al. "Greedy layer-wise training of deep networks". In: *Advances in neural information processing systems*. 2007, pp. 153–160.



Olivier Bousquet and Léon Bottou. "The tradeoffs of large scale learning". In: *Advances in neural information processing systems*. 2008, pp. 161–168.



Augustin Louis Cauchy. "Méthode générale pour la résolution des systèmes d'équations simultanées". In: *Comptes rendus des séances de l'Académie des sciences de Paris 25* (1847), pp. 536–538.



Anna Choromanska et al. "The loss surfaces of multilayer networks". In: *Artificial Intelligence and Statistics*. 2015, pp. 192–204.



Yann N Dauphin et al. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization". In: *Advances in neural information processing systems*. 2014, pp. 2933–2941.



John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization". In: *Journal of Machine Learning Research* 12.Jul (2011), pp. 2121–2159.

Références II



Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International Conference on Machine Learning*. 2015, pp. 448–456.



Diederik Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).



Yann A LeCun et al. "Efficient backprop". In: *Neural networks: Tricks of the trade*. Springer, 1998.



Yurii Nesterov. "A method of solving a convex programming problem with convergence rate $O(1/k^2)$ ". In: *Soviet Mathematics Doklady*. Vol. 27. 2. 1983, pp. 372–376.



Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.



Tijmen Tieleman and Geoffrey Hinton. "Lecture 6: Neural networks for Machine Learning". In: *COURSERA* (2012). URL:

http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.



Ashia C Wilson et al. "The Marginal Value of Adaptive Gradient Methods in Machine Learning". In: *arXiv preprint arXiv:1705.08292* (2017).