Identification de paramètres et optimisation
Cours de Master 2 STIM
2015-2016

Sébastien Adam

7 janvier 2016

Plan du Cours

Contexte

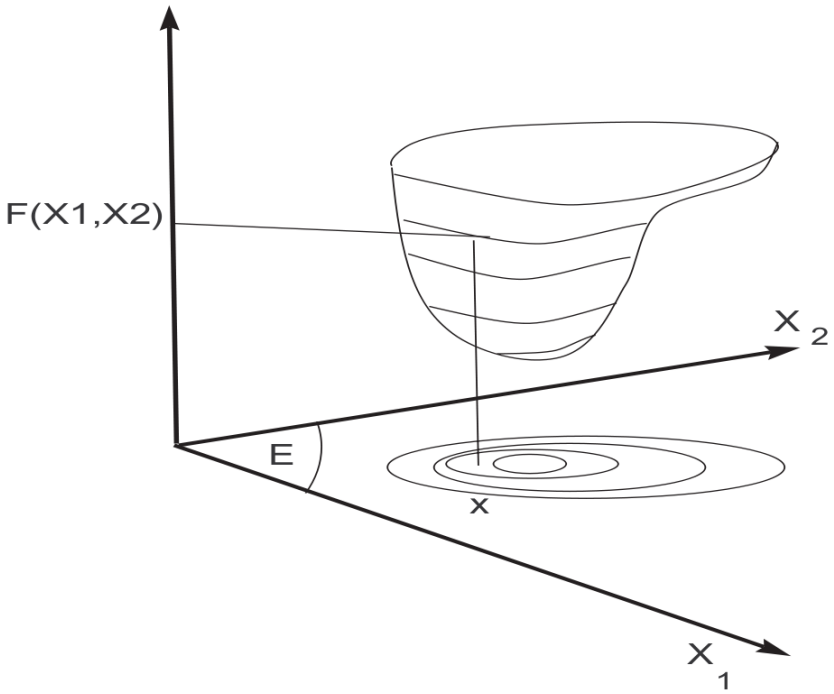
Les moindres carrés, c'est bien ...

- ... sous réserve de vérifier deux hypothèses :
  1. Critère d'erreur quadratique : $J_{MC}(\theta) = \sum_{i=1}^{N} (y_{Si} - y_{mi})^2$
  2. Modèle linéaire par rapport aux paramètres : $Y_M(\theta) = X\theta$
- On sait calculer analytiquement l'optimal :
  $\frac{\partial J_{MC}(\theta)}{\partial \theta} = 0 \Rightarrow X^T X\theta = X^T Y_S \Rightarrow \theta_{MC} = \left(X^T X\right)^{-1} X^T Y_S$
- On peut adapter une estimation avec une nouvelle mesure.

Mais ...

- L'hypothèse de linéarité est forte, même si des "tricks" existent.
- Modèle non linéaire $\rightarrow$ soucis pour annuler la dérivée
- L'inversion de $X^T X$ peut être très coûteuse (si K est grand - RDN, CRF), voire impossible
- Alternative : les méthodes itératives ou méthodes de descente

Principe général des méthodes de descente



Méthodes itératives

- On part d'un point θ_0
- On cherche à « descendre » : $J(\theta_{n+1}) < J(\theta_n)$
- On calcule $\theta_{n+1} = \theta_n + \alpha_n d_n$
  - d_n est un vecteur donnant la direction de l'itération
  - α_n est la longueur du pas
- On continue jusqu'à ... l'arrêt

Nombreux choix à effectuer → nombreuses variantes

- Déterminer la valeur θ_0
- Déterminer la direction d_n
- Déterminer la longueur du pas α_n
- Choisir un critère d'arrêt

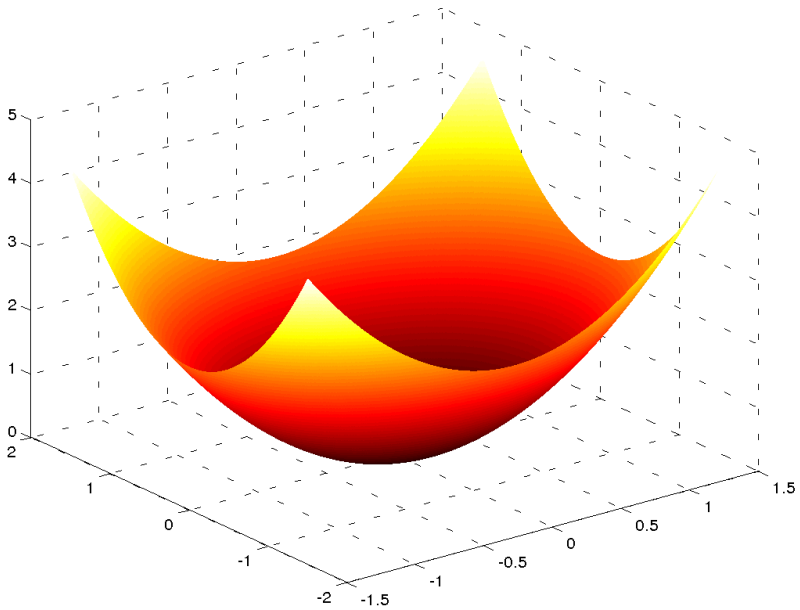
Plan du Cours

Quelques rappels utiles

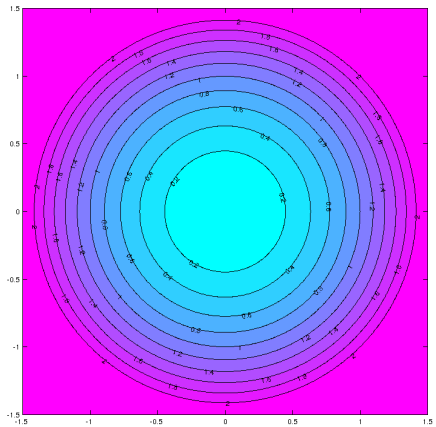
Fonction vectorielle

- On considère des fonctions $J(\theta)$ de $\mathbb{R}^K$ dans $\mathbb{R}$. Ex : $J(\theta) = \theta_1^2 + \theta_2^2$
- On va illustrer les descentes à l'aide de deux représentations :

Valeurs de critères



Lignes d'iso-coût
$\{\theta \in \mathbb{R}^K | J(\theta) = cte\}$

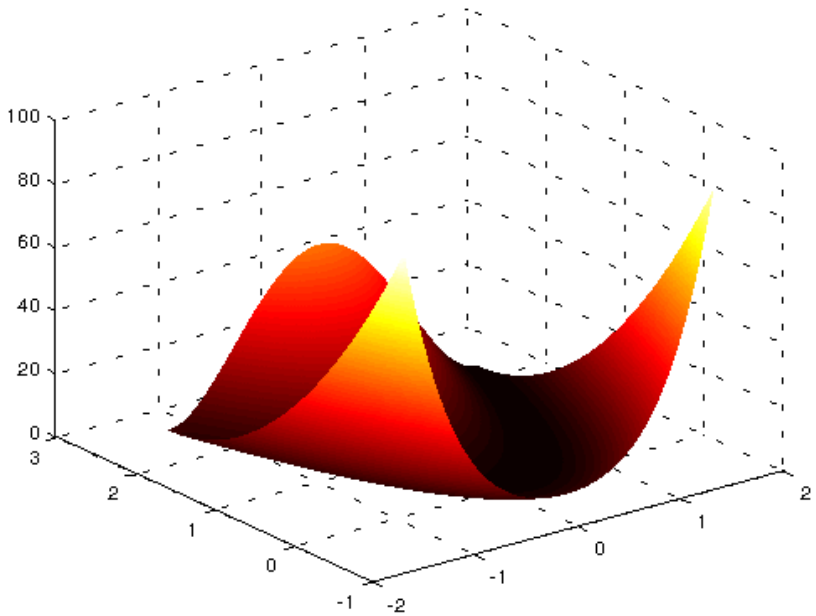


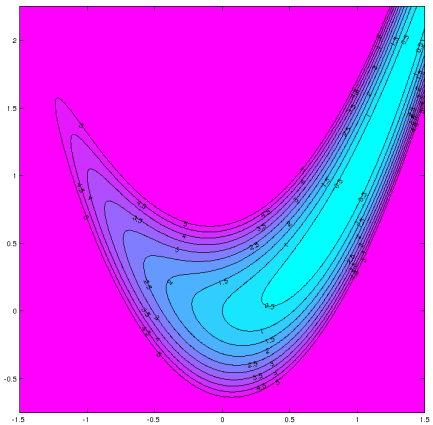
Quelques rappels utiles

Fonction vectorielle

- Ca peut être un peu plus compliqué

$$J(\theta) = (1 - \theta_1)^2 + 10(\theta_2 - \theta_1^2)^2$$





Quelques rappels utiles

Variations de fonctions vectorielles

- On veut descendre : il faut analyser les variations de la fonction $J(\theta)$ au voisinage d'un point θ_0
- Cette analyse doit être faite au sens d'une direction. On parle de dérivée directionnelle, généralisant la dérivée classique
- On appelle dérivée directionnelle de J au point θ_0 et dans la direction $d \in R^K$ la limite : $D_{\theta_0} J(\theta_0, d) = \lim_{\epsilon \to 0} \frac{J(\theta_0 + \epsilon d) - J(\theta_0)}{\epsilon}$
- $D_{\theta_0} J(\theta_0, d)$ "quantifie" la variation du critère lors d'un déplacement d'un vecteur d dans l'espace des paramètres (approximée linéaire). $\frac{D_{\theta_0} J(\theta_0, d)}{||d||}$ est la pente de la fonction en θ_0.
- NB : $\varphi(\epsilon) = J(\theta_0 + \epsilon d) \to D_{\theta_0} J(\theta_0, d) = lim_{\epsilon \to 0} \frac{\varphi(\epsilon) - \varphi(0)}{\epsilon} = \varphi'(0)$
- On peut donc calculer $D_{\theta_0} J(\theta_0, d)$ en calculant $\varphi'(0)$
- Exemples : $J(\theta) = a^T \theta$, $J(\theta) = \|\theta\|^2$

Quelques rappels utiles

Gradient d'une fonction vectorielle

- Une fonction $J(\theta)$ est dite différentiable si ses K dérivées partielles $\frac{\partial J}{\partial \theta_k}$ existent et sont continues. On note alors J est C^1
- Le vecteur composé des différentes dérivées partielles est appelé gradient de J en θ_0 et noté $\nabla J(\theta_0) = \left[\frac{\partial J}{\partial \theta_1}(\theta_0), ..., \frac{\partial J}{\partial \theta_K}(\theta_0)\right]^T$
- $\nabla J(\theta_0)$ est un déplacement dans l'espace des paramètres
- $\nabla\left(\alpha_1 J_1(\theta) + \alpha_2 J_2(\theta)\right) = \alpha_1 \nabla J_1(\theta) + \alpha_2 \nabla J_2(\theta)$
- $\nabla\left(J_1 o J_2(\theta)\right) = J_1'(J_2(\theta)).\nabla J_2(\theta)$
- On peut montrer que : $D_{\theta_0} J(\theta_0, d) = \nabla J(\theta_0)^T d$
- On peut donc calculer un gradient en exprimant sa dérivée directionnelle sous la forme $\nabla J(\theta_0)^T d$ ou $d^T \nabla J(\theta_0)$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla J(\theta_0) = [2\theta_1 + 2\theta_2, 2\theta_1 - 4\theta_2]^T$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla J(\theta_0) = [2\theta_1 + 2\theta_2, 2\theta_1 - 4\theta_2]^T$
- $J(\theta) = a^T \theta$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla J(\theta_0) = [2\theta_1 + 2\theta_2, 2\theta_1 - 4\theta_2]^T$
- $J(\theta) = a^T \theta$
  - $\nabla J(\theta_0) = a^T$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla J(\theta_0) = [2\theta_1 + 2\theta_2, 2\theta_1 - 4\theta_2]^T$
- $J(\theta) = a^T \theta$
  - $\nabla J(\theta_0) = a^T$
- $J(\theta) = \|\theta\|^2$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla J(\theta_0) = [2\theta_1 + 2\theta_2, 2\theta_1 - 4\theta_2]^T$
- $J(\theta) = a^T \theta$
  - $\nabla J(\theta_0) = a^T$
- $J(\theta) = \|\theta\|^2$
  - $\nabla J(\theta_0) = 2\theta^T$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla J(\theta_0) = [2\theta_1 + 2\theta_2, 2\theta_1 - 4\theta_2]^T$
- $J(\theta) = a^T \theta$
  - $\nabla J(\theta_0) = a^T$
- $J(\theta) = \|\theta\|^2$
  - $\nabla J(\theta_0) = 2\theta^T$
- $J(\theta) = sin(a^T \theta)$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla J(\theta_0) = [2\theta_1 + 2\theta_2, 2\theta_1 - 4\theta_2]^T$
- $J(\theta) = a^T \theta$
  - $\nabla J(\theta_0) = a^T$
- $J(\theta) = \|\theta\|^2$
  - $\nabla J(\theta_0) = 2\theta^T$
- $J(\theta) = sin(a^T \theta)$
  - $\nabla J(\theta_0) = a^T cos(a^T \theta)$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

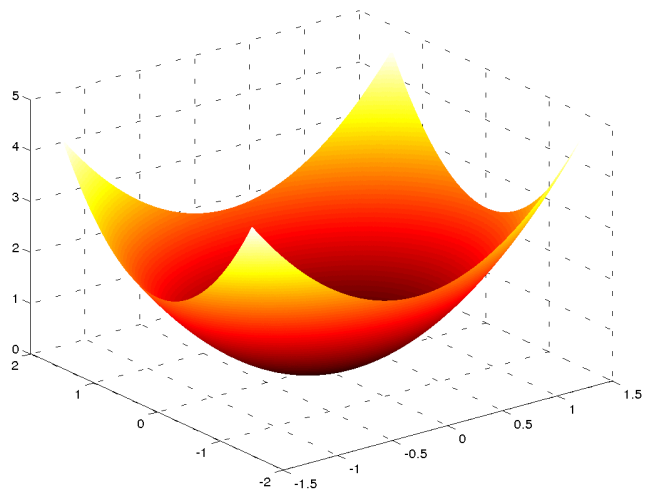
- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla J(\theta_0) = [2\theta_1 + 2\theta_2, 2\theta_1 - 4\theta_2]^T$
- $J(\theta) = a^T \theta$
  - $\nabla J(\theta_0) = a^T$
- $J(\theta) = \|\theta\|^2$
  - $\nabla J(\theta_0) = 2\theta^T$
- $J(\theta) = sin(a^T \theta)$
  - $\nabla J(\theta_0) = a^T cos(a^T \theta)$
- $J(\theta) = \|A\theta - b\|^2$

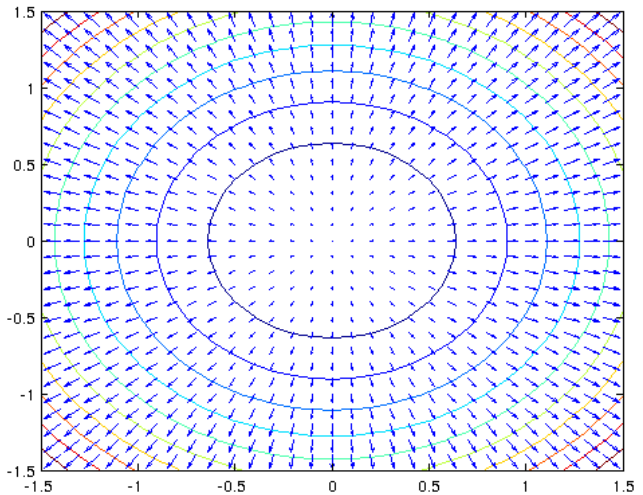
Quelques rappels utiles

Propriétés importante

- La direction donnée par le gradient est celle de plus forte pente en tout point.
- $\forall d$ de norme $\|\nabla J(\theta_0)\|$, Cauchy-Schwarz nous dit que $D_{\theta_0} J(\theta_0, d) = \nabla J(\theta_0)^T d \leq \|\nabla J(\theta_0)\| \, \|d\| \leq \|\nabla J(\theta_0)\|^2$

Illustration

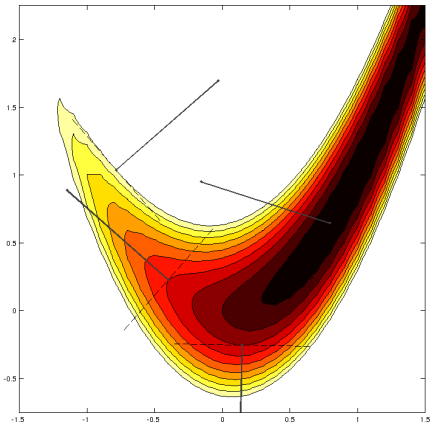




Quelques rappels utiles

Autre propriété intéressante

- Le gradient de J en θ_0 est orthogonal à l'ensemble de niveau $N_J(\theta_0)$: $\nabla J(\theta_0)^T v = 0$ pour tous les vecteurs v tangents à J en θ_0



Taylor-Young au premier ordre

- Taylor-Young au premier ordre approximant linéairement $J(\theta)$ en θ_0 devient $J(\theta) = J(\theta_0) + (\theta - \theta_0)^T \nabla J(\theta_0) + \|\theta - \theta_0\| \epsilon(\theta)$ avec : $\lim_{\theta \to \theta_0} \epsilon(\theta) = 0$

Quelques rappels utiles

Second ordre : Hessien d'une fonction vectorielle

- Si les dérivées partielles de J admettent à leur tour des dérivées partielles, on dit que J possède des dérivées partielles d'ordre 2.
- La dérivée partielle suivant la direction j de la dérivée partielle $\frac{\partial J}{\partial \theta_i}(a)$ est notée : $\frac{\partial^2 J}{\partial \theta_i \partial \theta_j}(a)$
- Si toutes les dérivées partielles sont continues, on dit que J est deux fois continuement différentiable. On note alors J est C^2
- La matrice de dimenson (K, K)

$$\nabla^2 J(a) = \frac{\partial^2 J}{\partial \theta \partial \theta^T}(a) = \begin{pmatrix} \frac{\partial^2 J}{\partial \theta_1^2}(a) & \frac{\partial^2 J}{\partial \theta_1 \partial \theta_2}(a) & \dots & \frac{\partial^2 J}{\partial \theta_1 \partial \theta_K}(a) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 J}{\partial \theta_K \partial \theta_1}(a) & \frac{\partial^2 J}{\partial \theta_K \partial \theta_2}(a) & \dots & \frac{\partial^2 J}{\partial \theta_K^2}(a) \end{pmatrix}$$

est appelée Hessien de J en a. Elle est symétrique par définition.

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla^2 J(\theta_0) = [2, -4]^T$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla^2 J(\theta_0) = [2, -4]^T$
- $J(\theta) = a^T \theta$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla^2 J(\theta_0) = [2, -4]^T$
- $J(\theta) = a^T \theta$
  - $\nabla^2 J(\theta_0) = 0$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla^2 J(\theta_0) = [2, -4]^T$
- $J(\theta) = a^T \theta$
  - $\nabla^2 J(\theta_0) = 0$
- $J(\theta) = \|\theta\|^2$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla^2 J(\theta_0) = [2, -4]^T$
- $J(\theta) = a^T \theta$
  - $\nabla^2 J(\theta_0) = 0$
- $J(\theta) = \|\theta\|^2$
  - $\nabla^2 J(\theta_0) = 2I$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla^2 J(\theta_0) = [2, -4]^T$
- $J(\theta) = a^T \theta$
  - $\nabla^2 J(\theta_0) = 0$
- $J(\theta) = \|\theta\|^2$
  - $\nabla^2 J(\theta_0) = 2I$
- $J(\theta) = \|A\theta - b\|^2$

Quelques rappels utiles

Rappels de dérivation

- $\forall x \in \mathbb{R}^k, a \in \mathbb{R}^k, \frac{\partial a^T x}{\partial x} = \frac{\partial x^T a}{\partial x} = a$
- $\forall x \in \mathbb{R}^k, A \in \mathbb{M}_{k,k}, \frac{\partial x^T A x}{\partial x} = (A + A^T)x$

Exemples

- $K = 2 : J(\theta) = \theta_1^2 - 2\theta_2^2 + 2\theta_1\theta_2 = \theta^T A \theta$
  - $\nabla^2 J(\theta_0) = [2, -4]^T$
- $J(\theta) = a^T \theta$
  - $\nabla^2 J(\theta_0) = 0$
- $J(\theta) = \|\theta\|^2$
  - $\nabla^2 J(\theta_0) = 2I$
- $J(\theta) = \|A\theta - b\|^2$
  - $\nabla^2 J(\theta_0) = 2A^T A$

Quelques rappels utiles

Hessien d'une fonction vectorielle

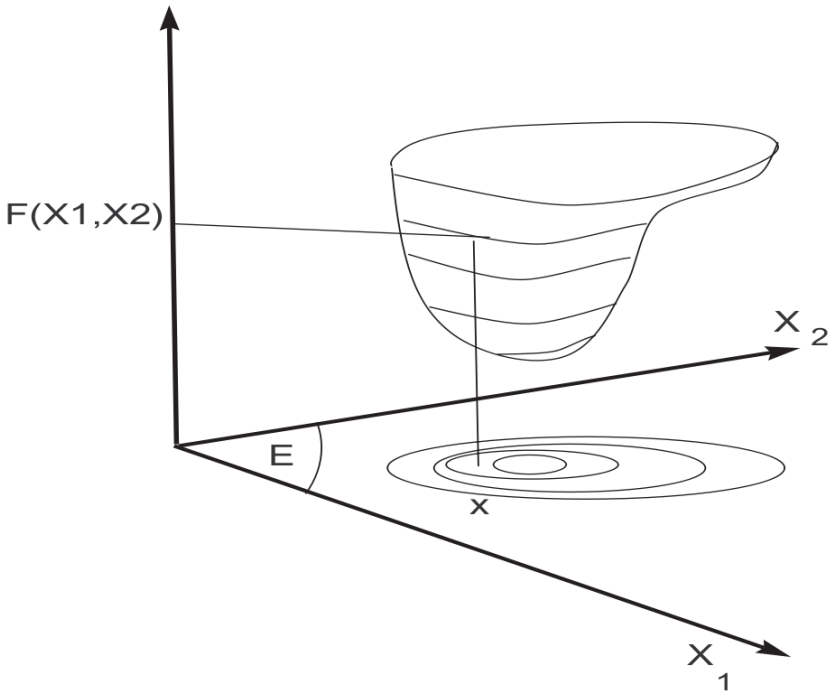
- Soit $d \in \mathbb{R}^K$. La grandeur $\frac{d^T \nabla^2 J(a) d}{\|d\|}$ est appelée courbure de la fonction J en a dans la direction d
- Soit un point θ_0 tel que $\nabla J(\theta_0) = 0$ et $\nabla^2 J(\theta_0)$ définie positive (c'est à dire $D^T \nabla^2 J(\theta_0) D > 0$). Alors θ_0 est un minimiseur local strict de J.
- Attention : ce n'est pas forcément un minimum global
- On a la formule de Taylor-Young au second ordre

$$J(\theta) = J(\theta_0) + (\theta - \theta_0)^T \nabla J(\theta_0) + \frac{1}{2}(\theta - \theta_0)^T \nabla^2 J(\theta_0)(\theta - \theta_0) + \|\theta - \theta_0\|^2 \epsilon(\theta)$$

$$\text{avec : } \lim_{\theta \to a} \epsilon(\theta) = 0$$

Plan du Cours

Méthode du gradient



Méthodes itératives

- On part d'un point θ_0
- On cherche à « descendre » : $J(\theta_{n+1}) < J(\theta_n)$
- On calcule $\theta_{n+1} = \theta_n + \alpha_n d_n$
  - d_n est un vecteur donnant la direction de l'itération
  - α_n est la longueur du pas
- On continue jusqu'à ... l'arrêt

Principes

- Les méthodes de descente de gradient considèrent l'anti-gradient comme direction de descente : $d_n = -\nabla J(\theta_n)$
- On pose donc : $\theta_{n+1} = \theta_n - \alpha_n \nabla J(\theta_n)$

Méthode du gradient

Principes

- Taylor-Young au premier ordre : $J(\theta + h) = J(\theta) + h^T \nabla J(\theta)$
- D'où : $J(\theta_{n+1}) = J(\theta_n - \alpha_n \nabla J(\theta_n)) = J(\theta_n) - \alpha_n \nabla J(\theta_n)^T \nabla J(\theta_n)$
- Donc $J(\theta_{n+1}) = J(\theta_n) - \alpha_n \|\nabla J(\theta_n)\|^2 \leq J(\theta_n)$
- C'est donc bien une direction de descente du critère, au voisinage de θ_n
- On peut montrer que c'est la direction de plus grande descente :

Preuve :
Soit $h \in \mathbb{R}^K$ une direction quelconque de norme $\|\nabla J(\theta)\|$.
On a $J(\theta + h) - J(\theta) = h^T \nabla J(\theta)$
Cauchy-Schwarz nous donne $J(\theta + h) - J(\theta) \leq \left\|h^T\right\| \|\nabla J(\theta)\|$
Donc : $J(\theta + h) - J(\theta) \leq \|\nabla J(\theta)\|^2$

Méthode du gradient

Algorithme général des méthodes de descente de gradient

1. n=0
2. Déterminer θ_0
3. Tant que critère d'arrêt non satisfait
   1. $d_n = -\nabla J(\theta_n)$ (ou, en version normalisée, $d_n = -\frac{\nabla J(\theta_n)}{\|\nabla J(\theta_n)\|^2}$)
   2. Déterminer α_n
   3. $\theta_{n+1} = \theta_n - \alpha_n d_n$
   4. $n = n + 1$
4. Retourner θ_n

Paramétrage de la méthode

- Une valeur initiale θ_0
- Une valeur pour α_n
- Le critère d'arrêt

Méthode du gradient

Méthode à pas fixe

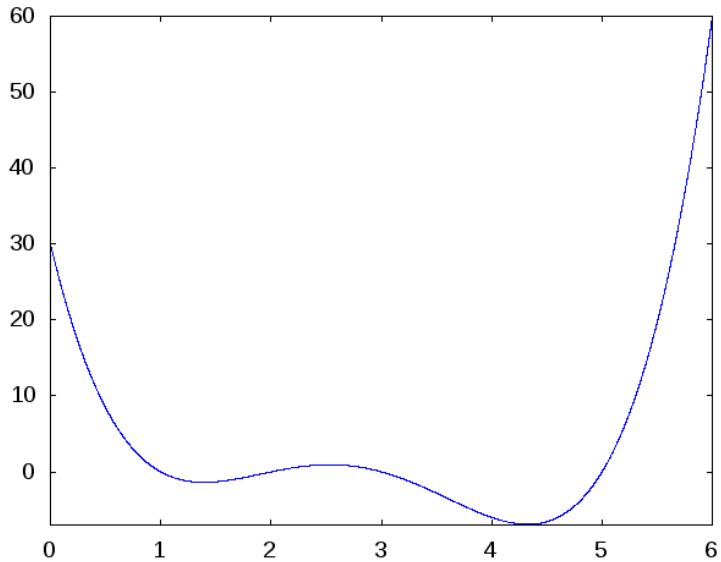
- On fixe $\alpha_n = \alpha$
- $\theta_{n+1} = \theta_n - \alpha \nabla J(\theta_n)$

Impact du choix de α

- $J(\theta) = (\theta-1)(\theta-2)(\theta-3)(\theta-5)$
- $\frac{dJ}{d\theta} = 4\theta^3 - 33\theta^2 + 82\theta - 61$
- Testons quelques valeurs :

θ_0	5	5	5	5	5
α_n	0.001	0.01	0.05	0.17	0.1

$\Rightarrow$ Importance cruciale de θ_0 et α_n



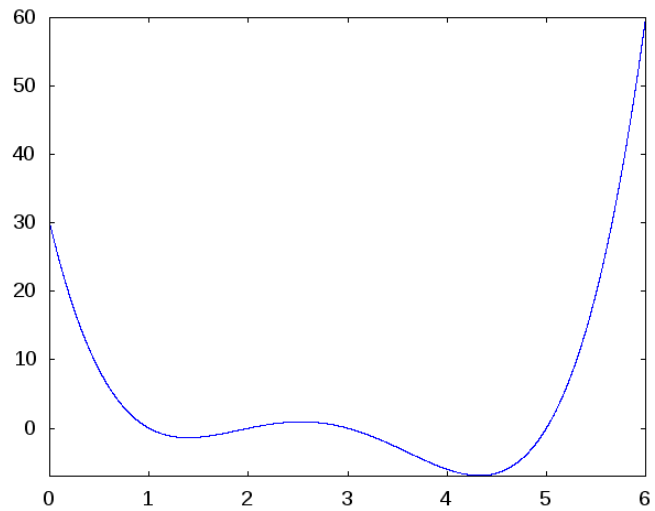
Méthode du gradient

Méthode à pas adaptatif

- On fait évoluer α_n par une régle heuristique
- $\theta_{n+1} = \theta_n - \alpha_n \nabla J(\theta_n)$

Choix de α_n : régle heuristique

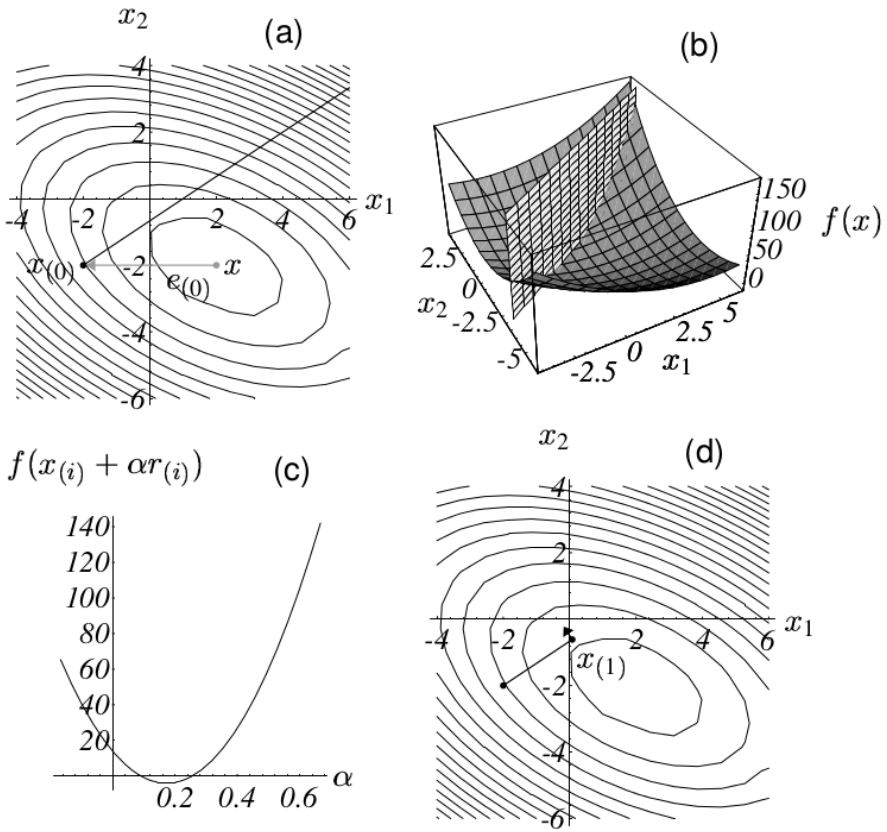
- si $J(\theta_{n+1}) < J(\theta_n)$ alors
  - augmenter α_n
- si $J(\theta_{n+1}) = J(\theta_n)$ alors
  - augmenter α_n
- si $J(\theta_{n+1}) > J(\theta_n)$ alors
  - diminuer α_n et abandonner θ_{n+1}



Méthode du gradient

Méthode à pas optimal

- On cherche à chaque itération la valeur optimale de α_n
- On considère une fonction $\varphi : \alpha_n \to J(\theta - \alpha_n \nabla J(\theta))$



Méthode du gradient

Méthode à pas optimal

- On cherche à chaque itération la valeur optimale de α_n
- On considère une fonction $\varphi : \alpha_n \to J(\theta - \alpha_n \nabla J(\theta))$
- On cherche alors à trouver le α_n qui minimise $\varphi(\alpha_n)$, c'est à dire qui annule la dérivée $\varphi'(\alpha_n) = -\nabla J(\theta_n)^T \nabla J(\theta_n - \alpha_n \nabla J(\theta_n))$

Exemple

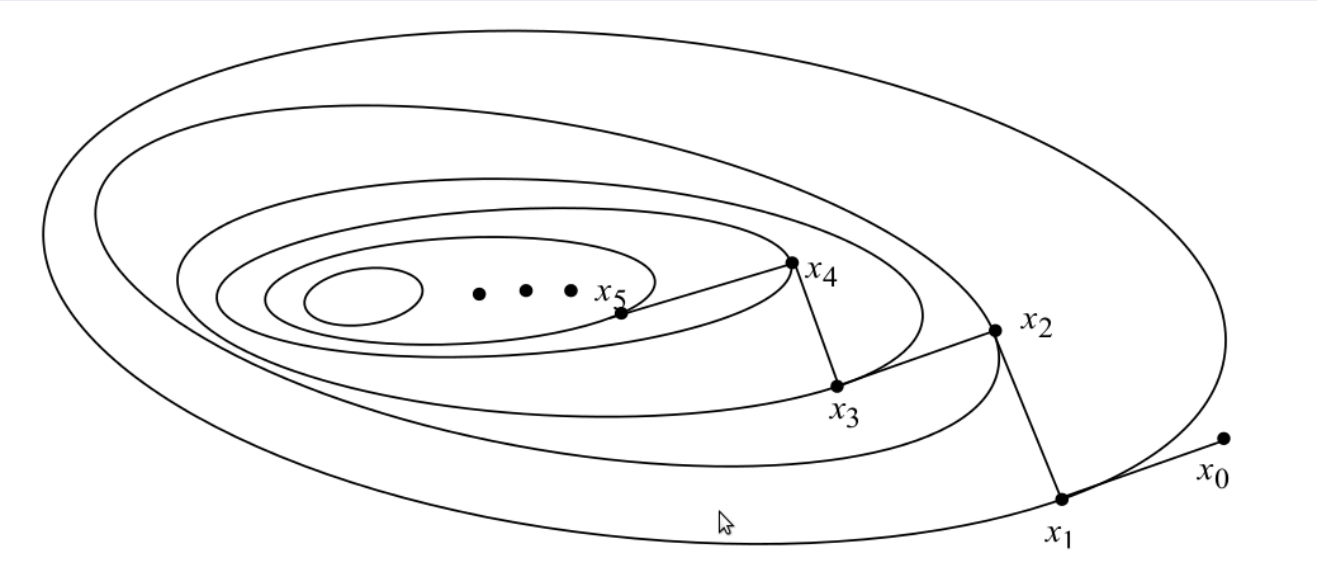
- Considérons $J(\theta) = \frac{\theta_1^2}{2} + \frac{7\theta_2^2}{2}$
- Exprimer $\theta_{n+1} = f(\theta_n)$ en utilisant la méthode du pas optimal

Remarque

- $\varphi'(\alpha_n) = 0 \Rightarrow \nabla J(\theta_n)^T \nabla J(\theta_n - \alpha_n \nabla J(\theta_n)) = 0$
- On a donc $\nabla J(\theta_n)^T \nabla J(\theta_{n+1}) = 0$. Les directions de descentes sont orthogonales

Méthode du gradient

Illustration



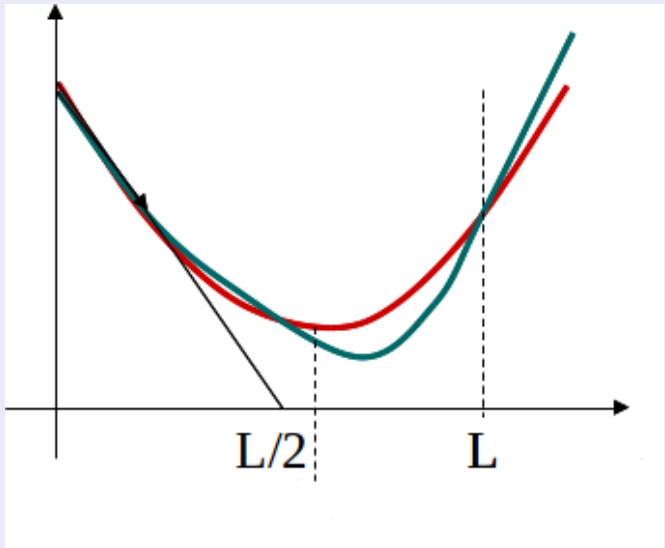
Problèmes

- $\varphi'(\alpha_n) = 0$ rarement calculable analytiquement et doit être effectuée à chaque pas du gradient

Méthode du gradient

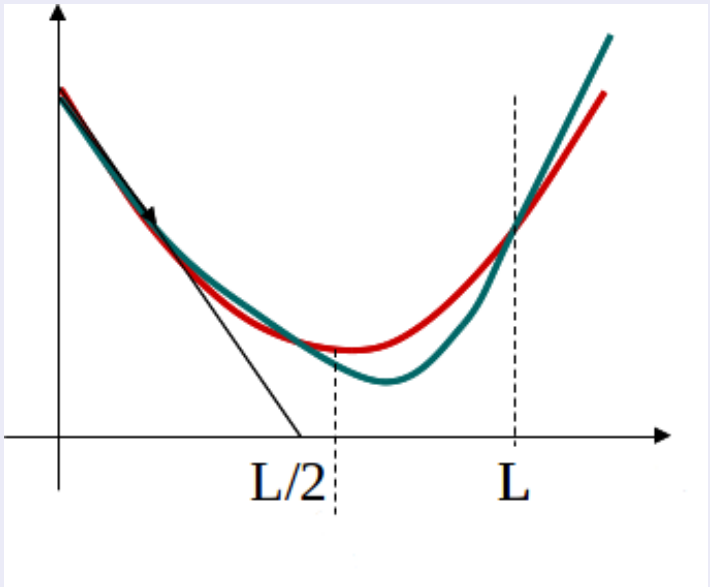
Recherche approchée

- On a $\varphi(\alpha_n) = J(\theta_n - \alpha_n \nabla J(\theta_n))$
- On a $\varphi'(0) = -\nabla J(\theta_n)^T \nabla J(\theta_n) < 0 \rightarrow \varphi$ est décroissante au voisinage de 0
- On approxime $\varphi(\alpha_n)$ par une parabole $\varphi_p(\alpha_n) = a\alpha_n^2 + b\alpha_n + c$ sur $[0, L]$



Méthode du gradient

Recherche approchée



Rappel équation tangente d'une fonction f en a :
$y = f(a) + f'(a)(x - a)$ i.e
$y = c + b\alpha_n$

- $\varphi(0) = \varphi_p(0) \rightarrow c = J(\theta_n)$
- $\varphi'(0) = \varphi'_p(0) \rightarrow b = -\nabla J(\theta_n)^T \nabla J(\theta_n)$
- $\varphi(L) = \varphi_p(L) \rightarrow a = (J(L) - bL - c)/L^2$
- La tangente en 0 coupe l'axe des abcisse en $L/2$: $L = -2c/b$
- $\alpha_n = -b/2a$

Méthode du gradient

Autres alternatives pour déterminer α_n

- Dichotomie
- Méthode du nombre d'or ....
- Toute autre méthode de recherche unidimensionnelle

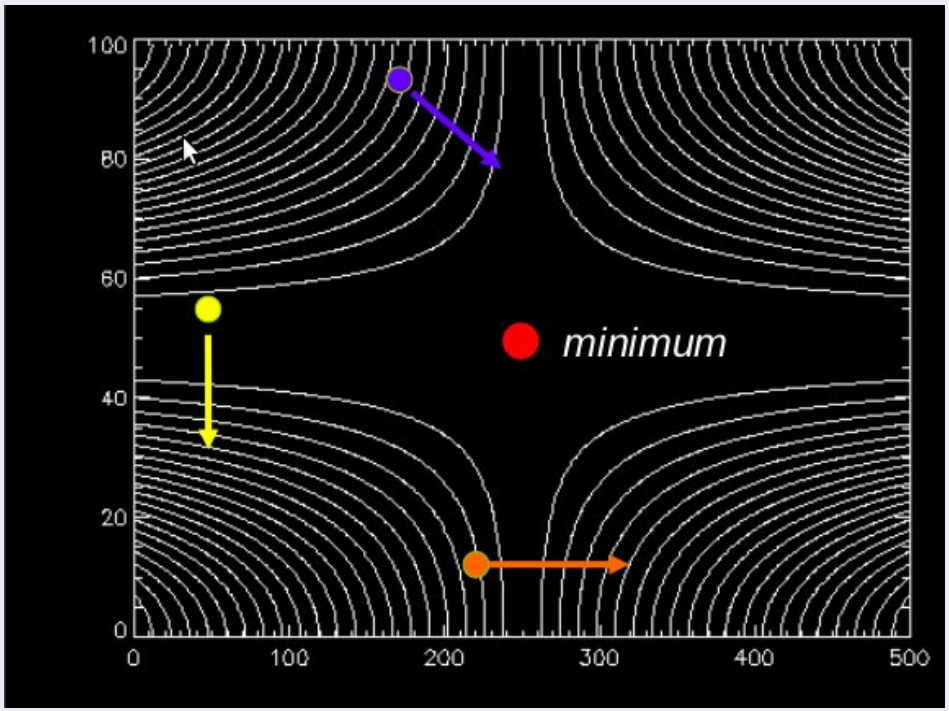
Propriétés de l'algorithme du gradient

- Algorithme simple à mettre en œuvre
- Grand domaine de convergence (domaine où la convergence est assurée)
- Se bloque dans les minimums locaux
- Ralentit au voisinage de la solution

Méthode du gradient

Limites de la méthode du gradient

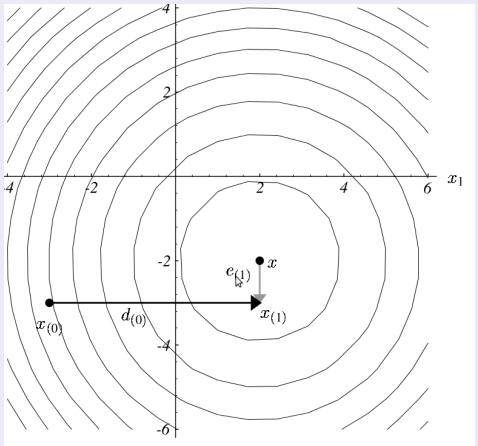
- Constat : la direction de l'anti-gradient n'est pas toujours la direction optimale
- Exemple :



Méthode du gradient

Idée

- Ce qu'on voudrait idéalement : des directions orthogonales qui en choisissant bien le α converge en n itérations.



- Mais il faudrait connaître la cible...

Une alternative

- Méthodes de gradient conjugué

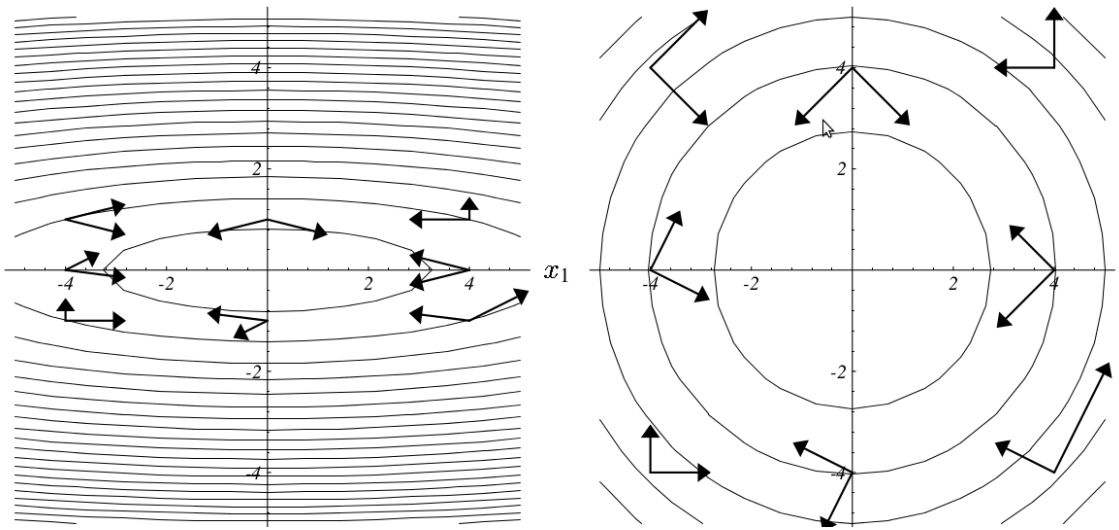
Méthode du gradient conjugué

Origine de la méthode

Initialement conçue pour minimiser en au plus n itérations une fonction quadratique telle que $J(\theta) = \frac{1}{2}\theta^T A\theta + b^T\theta + c$, avec A définie positive.

Directions conjuguées

- La méthode utilise des directions A-orthogonales, aussi appelées A-conjuguées, telles que $x^T Ay = 0$



Méthode du gradient conjugué

Principe de la méthode

- Le principe reste le même : on construit une suite de valeurs $\theta_{n+1} = \theta_n + \alpha_n d_n$ avec une valeur de α_n qui minimise $J(\theta_n + \alpha_n d_n)$
- L'idée fondamentale : on prend des directions d_n mutuellement A-conjuguées
- Pour $K > 2$, il en existe une infinité. On cherche d_n telle qu'elle soit une combinaison linéaire de d_{n-1} avec le gradient en θ_n : $d_n = -\nabla J(\theta_n) + \beta_n d_{n-1}$
- On choisit β_n tel que d_{n-1} et d_n soient A-conjuguées : $d_{n-1}^T A d_n = 0$
- $\beta_n = \frac{d_{n-1}^T A \nabla J(\theta_n)}{d_{n-1}^T A d_{n-1}}$
- d_0 est toujours la direction de l'anti grandient

Méthode du gradient conjugué

Algorithme de la méthode

1. Déterminer θ_0
2. $d_0 = -\nabla J(\theta_0)$
3. Tant que critère d'arrêt non satisfait
   1. $\alpha_n = -\frac{\nabla J(\theta_n) d_n}{d_n^T A d_n}$
   2. $\theta_{n+1} = \theta_n + \alpha_n d_n$
   3. $\beta_n = \frac{\nabla J(\theta_{n+1}) A d_n}{d_n^T A d_n}$
   4. $d_{n+1} = -\nabla J(\theta_{n+1}) + \beta_n d_n$
   5. $n = n + 1$

L'algorithme du gradient conjugué converge en n pas au maximum pour des formes QDP.

Méthode du gradient conjugué

Généralisation à des fonctions non quadratiques

Fletcher et Reeves ont étendu le GC aux fonctions non quadratiques.

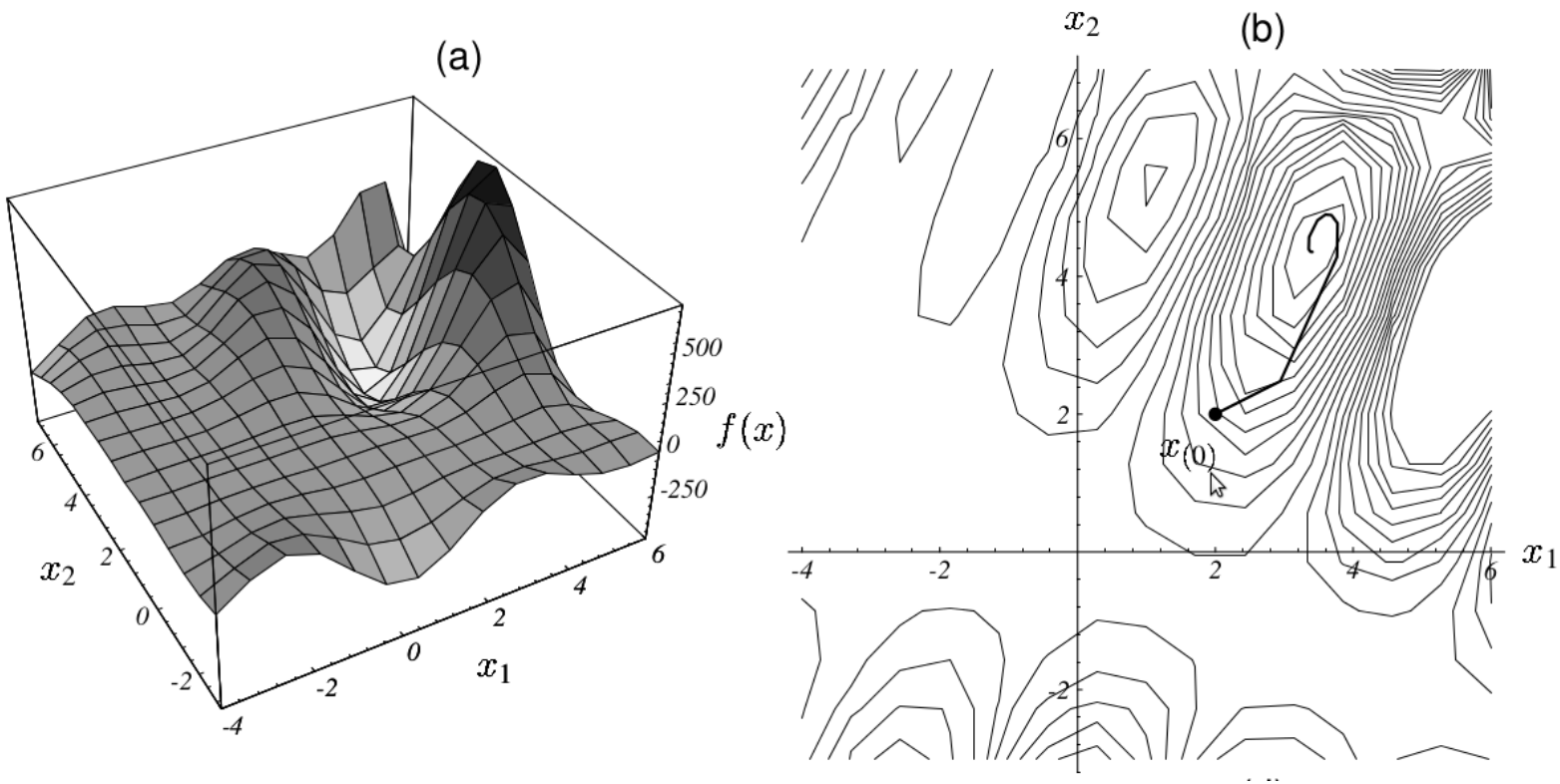
1. Déterminer θ_0
2. $d_0 = -\nabla J(\theta_0)$
3. Tant que critère d'arrêt non satisfait
   1. $\alpha_n \leftarrow$ recherche linéaire du pas optimal suivant d_n
   2. $\theta_{n+1} = \theta_n + \alpha_n d_n$
   3. $\beta_n = \frac{||\nabla J(\theta_{n+1})||^2}{||\nabla J(\theta_n||^2}$
   4. $d_{n+1} = -\nabla J(\theta_{n+1}) + \beta_n d_n$
   5. $n = n + 1$
4. Retourner θ_n

Il existe une variante réputée plus performante (Polak et Ribière) :

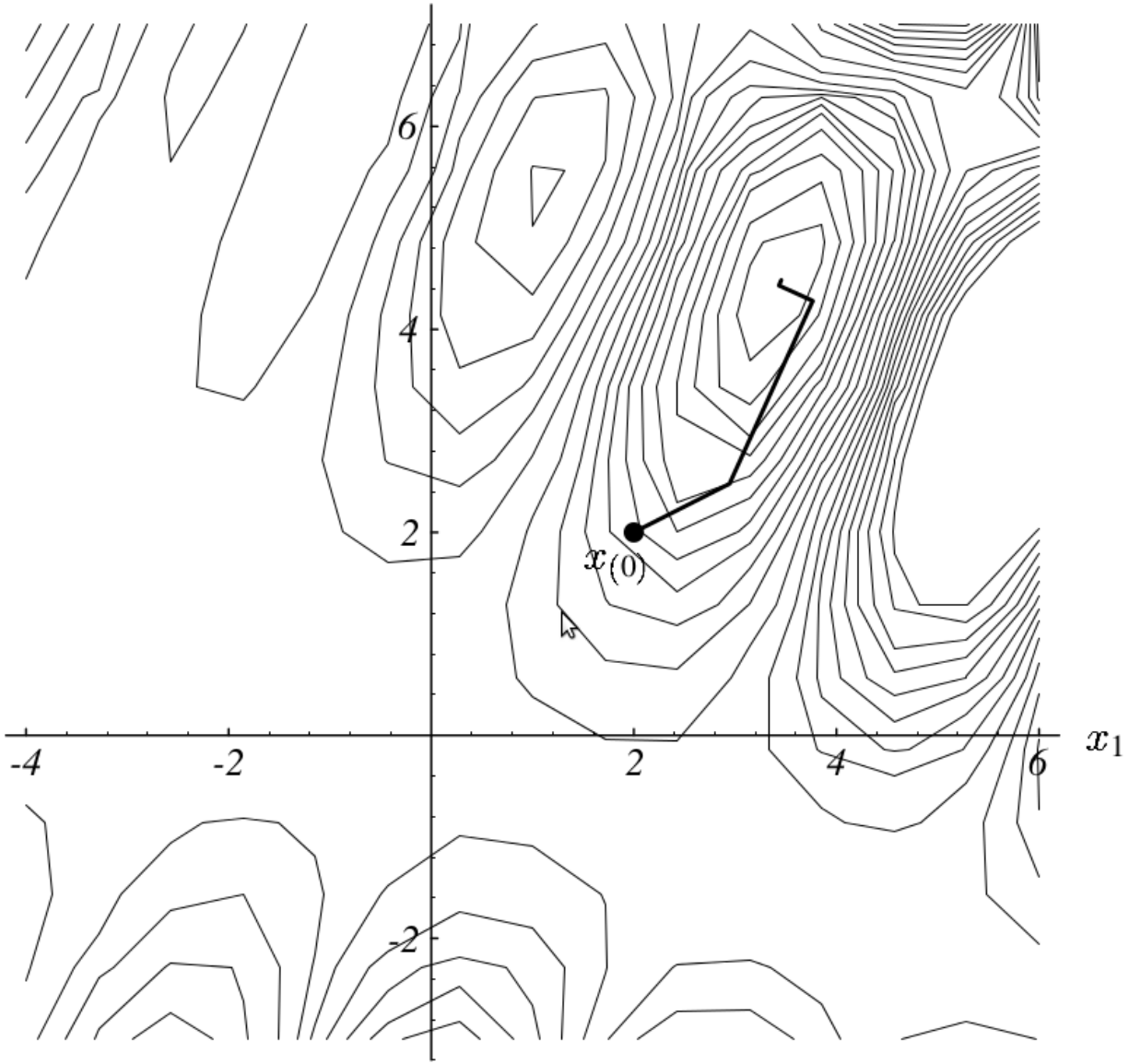
$\beta_n = \frac{(\nabla J(\theta_{n+1}) - \nabla J(\theta_n))^T \nabla J(\theta_{n+1})}{||\nabla J(\theta_n)||^2}$

Il faut parfois redémarrer l'algorithme en repartant du gradient

Illustration de Fletcher-Reeves sur une fonction compliquée



Impact d'un restart



Plan du Cours

Méthode de Newton

Principes

- Approximation quadratique du critère (en supposant que le critère est dans C^2).

Équations

- On pose $\theta_{n+1} = \theta_n + \delta_n$
- Taylor-Young au second ordre nous donne
  $J(\theta_{n+1}) = J(\theta_n + \delta_n) = J(\theta_n) + \nabla J(\theta_n)^T \delta_n + \frac{1}{2}\delta_n^T H(\theta_n)\delta_n$
- Avec

$$H(\theta_n) = \frac{d^2 J}{d\theta d\theta^T}(\theta_n) = \begin{pmatrix} \frac{d^2 J}{d\theta_1^2}(\theta_n) & \frac{d^2 J}{d\theta_1 d\theta_2}(\theta_n) & \cdots & \frac{d^2 J}{d\theta_1 d\theta_K}(\theta_n) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{d^2 J}{d\theta_K d\theta_1}(\theta_n) & \frac{d^2 J}{d\theta_K d\theta_2}(\theta_n) & \cdots & \frac{d^2 J}{d\theta_K^2}(\theta_n) \end{pmatrix}$$

Méthode de Newton

Résolution

$$J(\theta_{n+1}) = J(\theta_n + \delta_n) = J(\theta_n) + \nabla J(\theta_n)^T \delta_n + \frac{1}{2}\delta_n^T H(\theta_n)\delta_n$$

- Si la matrice $H(\theta)$ est définie positive, le minimum de cette quadratique en δ_n existe
- Il sera atteint pour $\frac{dJ}{d\delta_n} = 0$
- $\nabla J(\theta_n) + H(\theta)\delta_n = 0$
- Si $H(\theta_n)$ est inversible : $\delta_n = -H(\theta_n)^{-1} \nabla J(\theta_n)$
- Soit finalement :

$$\theta_{n+1} = \theta_n - H(\theta_n)^{-1} \nabla J(\theta_n)$$

- Remarque : en pratique, pas d'inversion : $H(\theta_n)\,\delta_n = -\nabla J(\theta_n)$

Méthode de Newton

Remarques

- Comparaison avec le gradient :
  - Calculs beaucoup plus lourds à chaque itération
  - Domaine de convergence réduit (Hessien inversible)
  - Direction et pas connus simultanément
  - Convergence beaucoup plus rapide
- Si le modèle est linéaire par rapport aux paramètres et le critère quadratique : moindres carrés

Variantes de Newton

Newton avec line search

- Idée : utiliser la direction de Newton quand elle est définie, et la plus forte pente sinon
- En pratique : δ_n est la solution de $(H(\theta_n) + D)\,\delta_n = -\nabla J(\theta_n)$ où D est une matrice diagonale qui rend la parenthèse définie positive
- Le pas de descente est alors optimisé en mono-dimensionnel

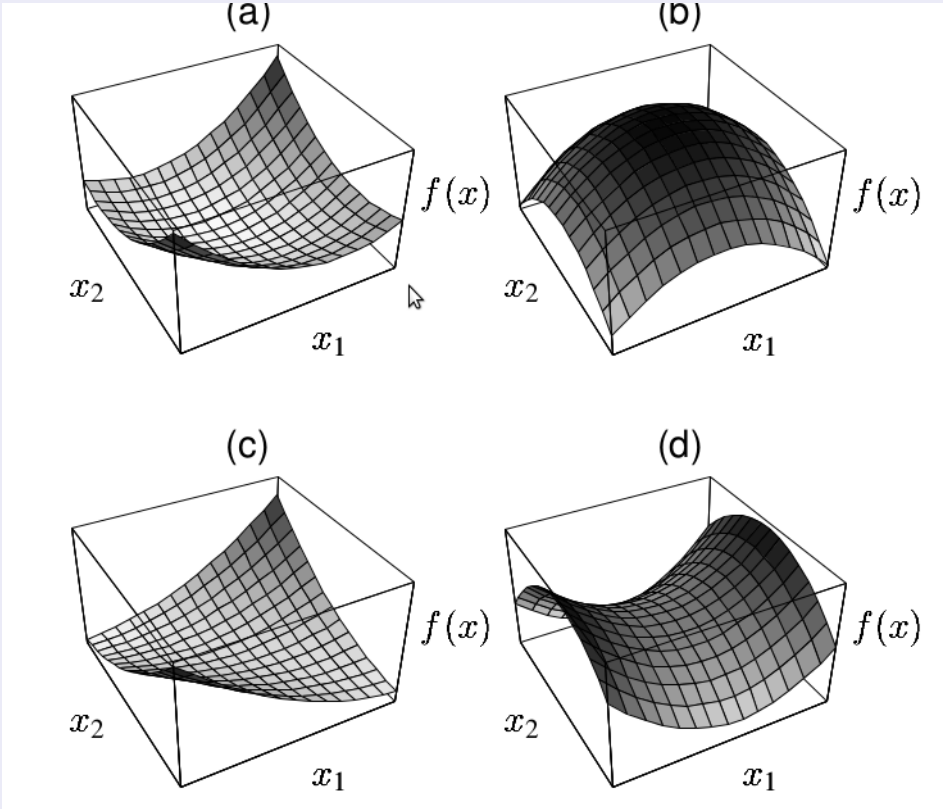
Méthodes quasi-Newton

- Idée : Estimer $(H(\theta_n))^{-1}$ par B_n
- Davidon-Fletcher-Powell : $B_n = B_{n-1} - \frac{B_{n-1} Y_n Y_n^T B_{n-1}}{Y_n^T B_{n-1} Y_n} + \frac{d_{n-1} d_{n-1}^T}{d_{n-1}^T Y_n}$
- Boyden-Fletcher-Goldfarb-Shanno (BFGS) :
  $B_n = B_{n-1} + \left(1 + \frac{Y_n^T B_{n-1} Y_n}{Y_n d_{n-1}}\right) \frac{d_{n-1} d_{n-1}^T}{Y_n^T d_{n-1}} - \frac{B_{n-1} Y_n d_{n-1} + d_{n-1} Y_n^T B_{n-1}}{Y_n^T D_{n-1}}$
- Avec : $Y_n = \nabla J(\theta_n) - \nabla J(\theta_{n-1})$

Comparaison des méthodes de gradient sur des formes QDP

Définitions

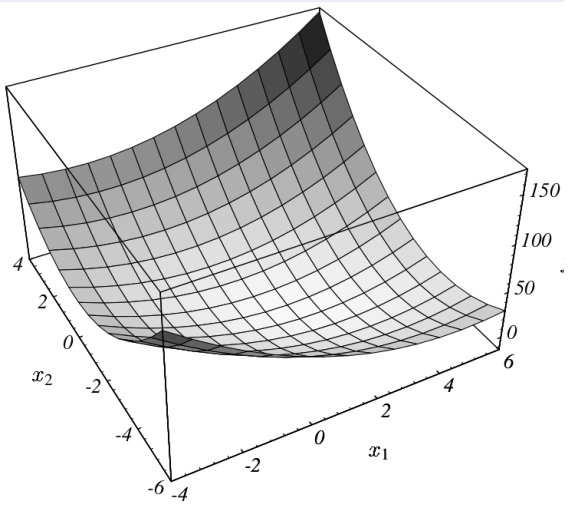
- Une fonction quadratique est définie par $J(\theta) = \frac{1}{2}\theta^T A\theta - b^T\theta + c$
  - A est une matrice, b et θ sont des vecteurs, c est un scalaire
  - Variations suivant A

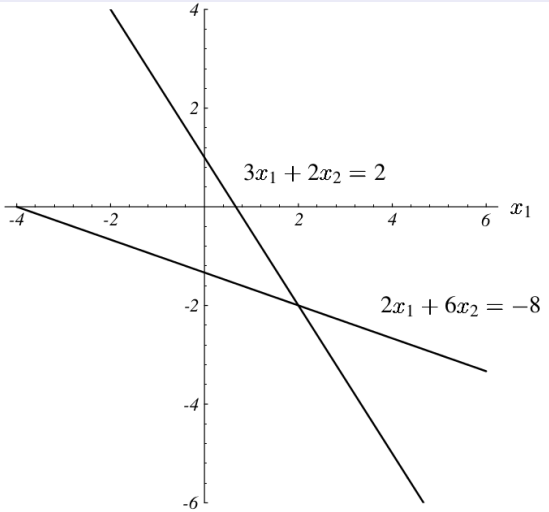


Comparaison des méthodes de gradient sur des formes QDP

Définitions

- On s'intéresse au cas où A est symétrique (i.e. $A^T = A$) et définie positive (i.e $\theta^t A\theta > 0 \ \forall \theta \neq 0$)
- Dans ce cas $\nabla J(\theta) = A\theta - b$ et minimiser J est équivalent à résoudre un système d'équations linéaires (de solution $\theta = A^{-1}b$)
- Exemple : $A = \begin{pmatrix} 3 & 2 \\ 2 & 6 \end{pmatrix}$, $b = \begin{pmatrix} 2 \\ -8 \end{pmatrix}$, $c = 0$





Comparaison des méthodes de gradient sur des formes QDP

Application du gradient

- Le gradient en θ_n vaut $\nabla J(\theta_n) = A\theta_n - b$
- Le pas optimal vaut $\alpha_n = \frac{(A\theta_n - b)^T (A\theta_n - b)}{(A\theta_n - b)^T A (A\theta_n - b)}$, le Hessien vaut A
- Comparons pas fixe, pas adaptatif, pas optimal et Newton