Joseph Salmon

Analyse 2: Algorithme pour l'optimisation sans contrainte

Joseph Salmon

Septembre 2014

INSTITUT Mines-Télécom

TELECOM ParisTech

Joseph Salmon

Plan du cours

La descente de gradient : intuition

- Enjeu : minimiser f (dans $\mathbb{R}^d$) en trouvant un nouveau point pour lequel f diminue le plus.
- Approximation du premier ordre :

$$f(x) \approx f(x^0) + \langle \nabla f(x^0), x - x^0 \rangle$$

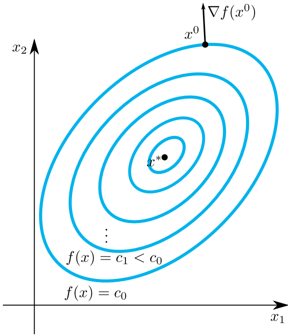
- Solution : il faut "s'aligner" avec la direction opposée au gradient $x - x_0 = -\alpha \nabla f(x^0)$
  $\alpha > 0$ contrôle la "vitesse" avec laquelle on progresse dans la direction. Ce paramètre est appelé le **pas** de la méthode.

La descente de gradient : algorithme

Data: initialisation x^0, nb max. d'itérations T, critère d'arrêt ε, pas α
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T$ **do**
    $x^{t+1} \leftarrow x^t - \alpha \nabla f(x^t)$
    STOP si critère d'arrêt inférieur à ε
end

Critères d'arrêts possibles :

- $\|\nabla f(x^t)\| \leq \varepsilon$
- $f(x^{t+1}) - f(x^t) \leq \varepsilon$
- $\|x^{t+1} - x^t\| \leq \varepsilon$ ou $\frac{\|x^{t+1} - x^t\|}{\|x^t\|} \leq \varepsilon$

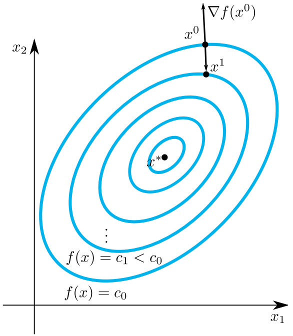


La descente de gradient : algorithme

Data: initialisation x^0, nb max. d'itérations T, critère d'arrêt ε, pas α
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T$ **do**
  $x^{t+1} \leftarrow x^t - \alpha \nabla f(x^t)$
  STOP si critère d'arrêt inférieur à ε
end

Critères d'arrêts possibles :

- $\|\nabla f(x^t)\| \leq \varepsilon$
- $f(x^{t+1}) - f(x^t) \leq \varepsilon$
- $\|x^{t+1} - x^t\| \leq \varepsilon$ ou $\frac{\|x^{t+1} - x^t\|}{\|x^t\|} \leq \varepsilon$

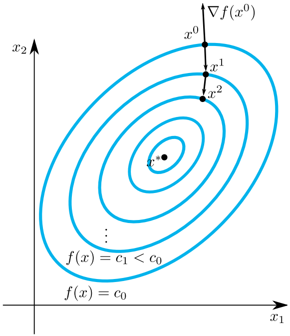


La descente de gradient : algorithme

Data: initialisation x^0, nb max. d'itérations T, critère d'arrêt ε, pas α
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T$ **do**
    $x^{t+1} \leftarrow x^t - \alpha \nabla f(x^t)$
    STOP si critère d'arrêt inférieur à ε
end

Critères d'arrêts possibles :

- $\|\nabla f(x^t)\| \leq \varepsilon$
- $f(x^{t+1}) - f(x^t) \leq \varepsilon$
- $\|x^{t+1} - x^t\| \leq \varepsilon$ ou $\frac{\|x^{t+1} - x^t\|}{\|x^t\|} \leq \varepsilon$

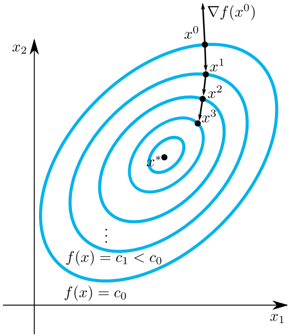


La descente de gradient : algorithme

Data: initialisation x^0, nb max. d'itérations T, critère d'arrêt ε, pas α
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T$ **do**
    $x^{t+1} \leftarrow x^t - \alpha \nabla f(x^t)$
    STOP si critère d'arrêt inférieur à ε
end

Critères d'arrêts possibles :

- $\|\nabla f(x^t)\| \leq \varepsilon$
- $f(x^{t+1}) - f(x^t) \leq \varepsilon$
- $\|x^{t+1} - x^t\| \leq \varepsilon$ ou $\frac{\|x^{t+1} - x^t\|}{\|x^t\|} \leq \varepsilon$



Analyse 2: Algorithme pour l'optimisation sans contrainte

Joseph Salmon

La descente de gradient : algorithme

Data: initialisation x^0, nb max. d'itérations T, critère d'arrêt ε, pas α
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T$ **do**

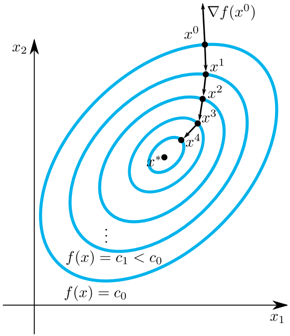
$x^{t+1} \leftarrow x^t - \alpha \nabla f(x^t)$

STOP si critère d'arrêt inférieur à ε

end

Critères d'arrêts possibles :

- $\|\nabla f(x^t)\| \leq \varepsilon$
- $f(x^{t+1}) - f(x^t) \leq \varepsilon$
- $\|x^{t+1} - x^t\| \leq \varepsilon$ ou $\frac{\|x^{t+1} - x^t\|}{\|x^t\|} \leq \varepsilon$

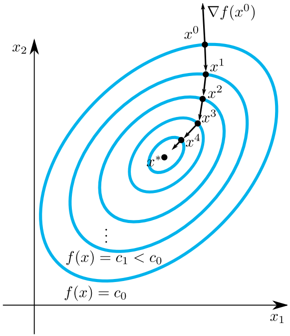


La descente de gradient : algorithme

Data: initialisation x^0, nb max. d'itérations T, critère d'arrêt ε, pas α
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T$ **do**
  $x^{t+1} \leftarrow x^t - \alpha \nabla f(x^t)$
  STOP si critère d'arrêt inférieur à ε
end

Critères d'arrêts possibles :

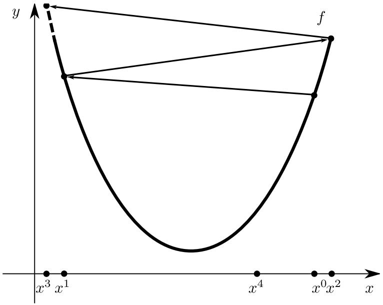
- $\|\nabla f(x^t)\| \leq \varepsilon$
- $f(x^{t+1}) - f(x^t) \leq \varepsilon$
- $\|x^{t+1} - x^t\| \leq \varepsilon$ ou $\frac{\|x^{t+1} - x^t\|}{\|x^t\|} \leq \varepsilon$



Attention au choix du pas (cas 1D)

$$\boxed{x^{t+1} = x^t - \alpha \nabla f(x^t)}$$

α : paramètre crucial pour obtenir la convergence vers un minimum



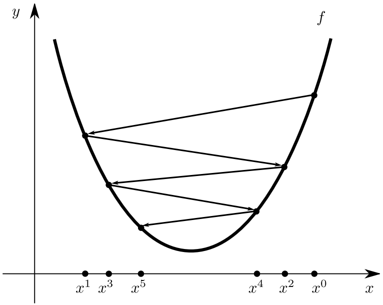
Divergence : pas beaucoup trop grand

Joseph Salmon

Attention au choix du pas (cas 1D)

$$\boxed{x^{t+1} = x^t - \alpha \nabla f(x^t)}$$

α : paramètre crucial pour obtenir la convergence vers un minimum

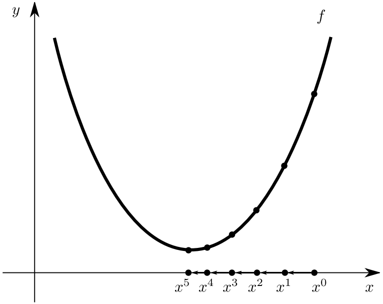


Convergence lente : pas trop grand

Attention au choix du pas (cas 1D)

$$\boxed{x^{t+1} = x^t - \alpha \nabla f(x^t)}$$

α : paramètre crucial pour obtenir la convergence vers un minimum

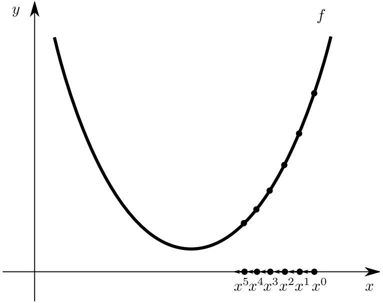


Convergence rapide : bon pas

Attention au choix du pas (cas 1D)

$$\boxed{x^{t+1} = x^t - \alpha \nabla f(x^t)}$$

α : paramètre crucial pour obtenir la convergence vers un minimum

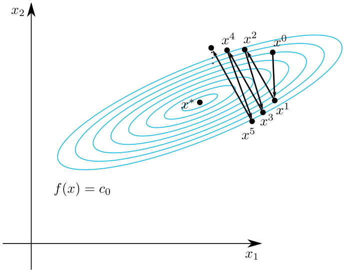


Convergence lente : pas trop petit

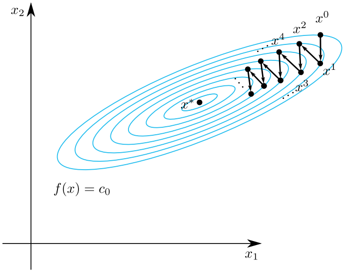
Attention au choix du pas (cas 2D)

$$\boxed{x^{t+1} = x^t - \alpha \nabla f(x^t)}$$

α : paramètre crucial pour obtenir la convergence vers un minimum



Trop grand pas



Trop petit pas

Recherche linéaire I

Parfois, il faut choisir le pas à chaque itération : α^t évolue avec les itérations. On note $d^t = -\nabla f(x^t)$ une direction de descente

Règle de la minimisation

Minimisation sur l'amplitude : il faut résoudre le problème 1D :

$$f(x^t + \alpha^t d^t) = \min_{\alpha \geq 0} f(x^t + \alpha d^t)$$

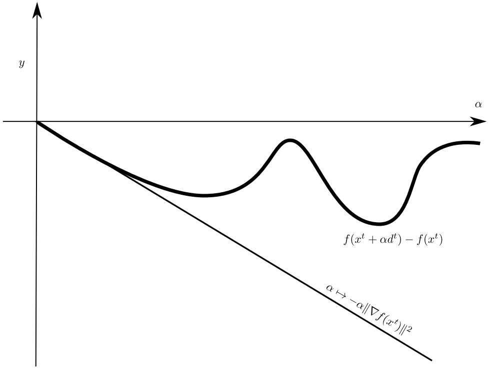
Rem: Pour cela il faut que le problème 1D soit simple à résoudre

Recherche linéaire II

Règle d'Armijo (ou du *backtracking* géométrique)

En fixant $s > 0$, $\sigma \in]0,1[$, et $\beta \in]0,1[$, il s'agit de choisir $\alpha^t = \beta^{m_t} s$: où m_t est le premier entier non nul tel que

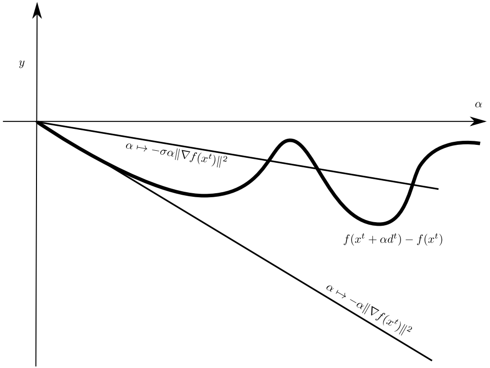
$$f(x^t + \beta^m s d^t) - f(x^t) \leq \sigma \beta^m s \langle \nabla f(x^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(x^t)\|^2$$



Recherche linéaire II

Règle d'Armijo (ou du *backtracking* géométrique)

En fixant $s > 0$, $\sigma \in]0,1[$, et $\beta \in]0,1[$, il s'agit de choisir $\alpha^t = \beta^{m_t} s$: où m_t est le premier entier non nul tel que
$f(x^t + \beta^m s d^t) - f(x^t) \leq \sigma \beta^m s \langle \nabla f(x^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(x^t)\|^2$

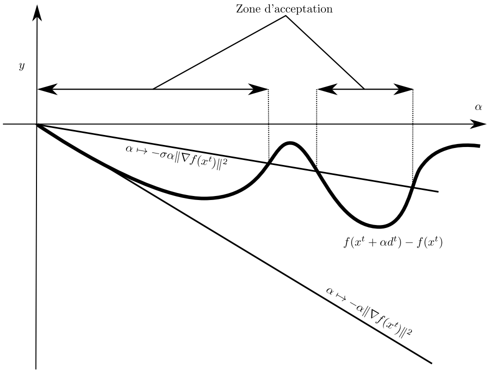


Analyse 2: Algorithme pour l'optimisation sans contrainte

Recherche linéaire II

Règle d'Armijo (ou du *backtracking* géométrique)

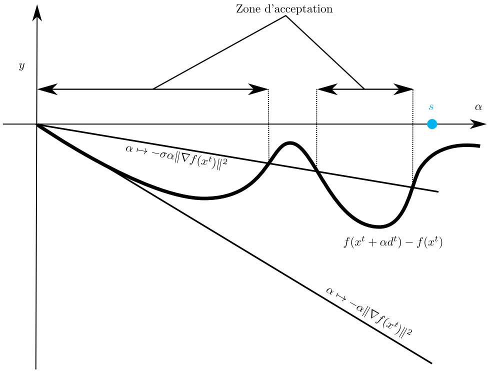
En fixant $s > 0$, $\sigma \in]0,1[$, et $\beta \in]0,1[$, il s'agit de choisir $\alpha^t = \beta^{m_t} s$: où m_t est le premier entier non nul tel que $f(x^t + \beta^m s d^t) - f(x^t) \leq \sigma \beta^m s \langle \nabla f(x^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(x^t)\|^2$



Recherche linéaire II

Règle d'Armijo (ou du *backtracking* géométrique)

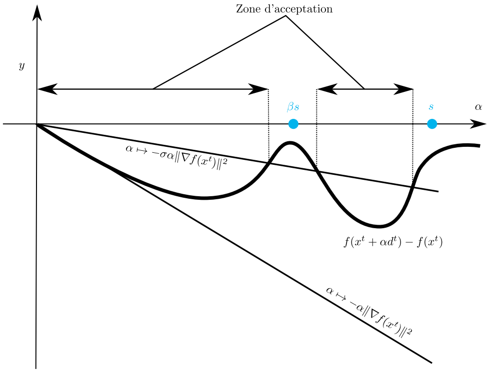
En fixant $s > 0$, $\sigma \in]0,1[$, et $\beta \in]0,1[$, il s'agit de choisir $\alpha^t = \beta^{m_t} s$: où m_t est le premier entier non nul tel que $f(x^t + \beta^m s d^t) - f(x^t) \leq \sigma \beta^m s \langle \nabla f(x^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(x^t)\|^2$



Recherche linéaire II

Règle d'Armijo (ou du *backtracking* géométrique)

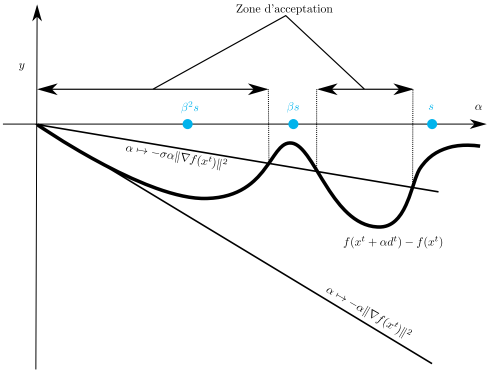
En fixant $s > 0$, $\sigma \in]0,1[$, et $\beta \in]0,1[$, il s'agit de choisir $\alpha^t = \beta^{m_t} s$: où m_t est le premier entier non nul tel que
$f(x^t + \beta^m s d^t) - f(x^t) \leq \sigma \beta^m s \langle \nabla f(x^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(x^t)\|^2$



Recherche linéaire II

Règle d'Armijo (ou du *backtracking* géométrique)

En fixant $s > 0$, $\sigma \in]0,1[$, et $\beta \in]0,1[$, il s'agit de choisir $\alpha^t = \beta^{m_t} s$: où m_t est le premier entier non nul tel que
$$f(x^t + \beta^m s d^t) - f(x^t) \leq \sigma \beta^m s \langle \nabla f(x^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(x^t)\|^2$$

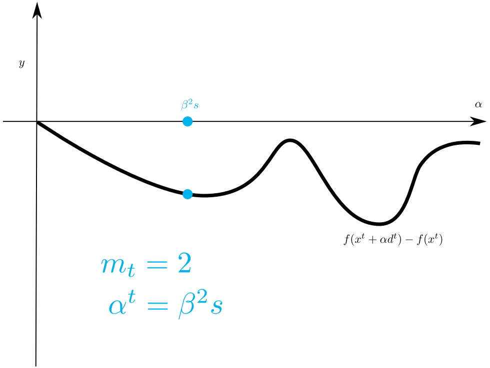


Analyse 2: Algorithme pour l'optimisation sans contrainte

Recherche linéaire II

Règle d'Armijo (ou du *backtracking* géométrique)

En fixant $s > 0$, $\sigma \in]0,1[$, et $\beta \in]0,1[$, il s'agit de choisir $\alpha^t = \beta^{m_t} s$: où m_t est le premier entier non nul tel que
$f(x^t + \beta^m s d^t) - f(x^t) \leq \sigma \beta^m s \langle \nabla f(x^t), d_t \rangle = -\sigma \beta^m s \|\nabla f(x^t)\|^2$



Analyse 2: Algorithme pour l'optimisation sans contrainte

Recherche linéaire III

Règle d'Armijo (ou du *backtracking*)

En pratique on fait souvent les choix, *cf.* Bertsekas (1999) :

- $s = 1$
- $\beta = 1/2$ ou $\beta = 1/10$
- $\sigma \in [10^{-5}, 10^{-1}]$

Détour par la méthode de Newton

Objectif : la méthode de Newton (ou Newton-Raphson) sert à trouver les zéros d'une fonction, *i.e.*, résoudre $f(x) = 0$
L'idée : approximation locale par une fonction affine

$$f(x) \approx f(x^0) + f'(x^0)(x - x^0)$$

La règle de mise à jour est donc :

$$\boxed{x^{t+1} \leftarrow x^t - \frac{f(x^t)}{f'(x^t)}}$$

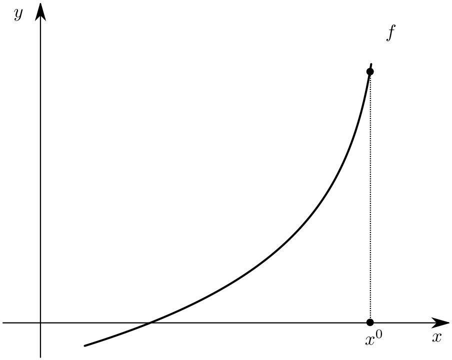
Détour par la méthode de Newton II

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T-1$ **do**

$$x^{t+1} \leftarrow x^t - \frac{f(x^t)}{f'(x^t)}$$

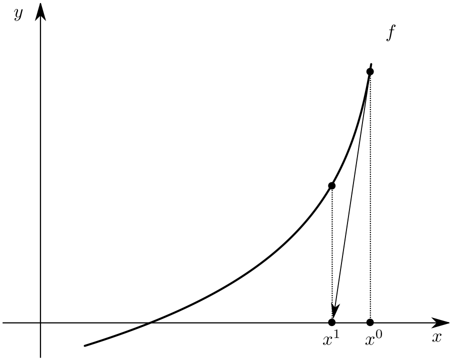
STOP si critère d'arrêt inférieur à ε

end



Détour par la méthode de Newton II

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T-1$ **do**
 $x^{t+1} \leftarrow x^t - \frac{f(x^t)}{f'(x^t)}$
 STOP si critère d'arrêt inférieur à ε
end



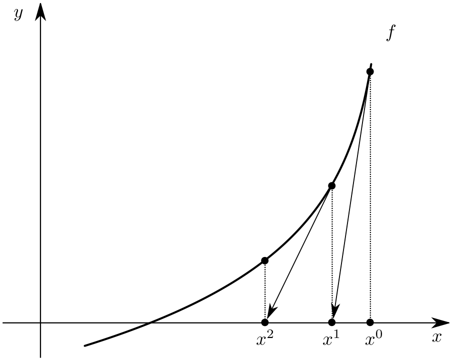
Détour par la méthode de Newton II

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T-1$ **do**

$$x^{t+1} \leftarrow x^t - \frac{f(x^t)}{f'(x^t)}$$

STOP si critère d'arrêt inférieur à ε

end



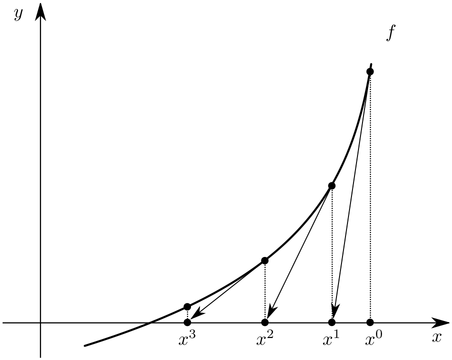
Détour par la méthode de Newton II

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T-1$ **do**

$$x^{t+1} \leftarrow x^t - \frac{f(x^t)}{f'(x^t)}$$

STOP si critère d'arrêt inférieur à ε

end



Détour par la méthode de Newton II

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε

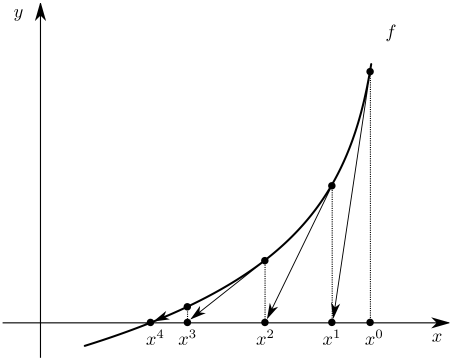
Result: un point x_T "proche" du minimum de la fonction f

for $1 \leq t \leq T-1$ **do**

$x^{t+1} \leftarrow x^t - \frac{f(x^t)}{f'(x^t)}$

STOP si critère d'arrêt inférieur à ε

end



Méthode de Newton pour la minimisation

Localement, en un point x^0 une fonction deux fois différentiable ressemble à :

$$f(x) \approx f(x^*) + \langle \nabla f(x^*), x - x^* \rangle + \frac{1}{2}(x - x^*)^\top \nabla^2 f(x^*)(x - x^*)$$

- Enjeu : minimiser en x l'approximation (quadratique) précédente
- Solution : CNO

$$\nabla f(x^*) + \nabla^2 f(x^*)(x - x^*) = 0$$

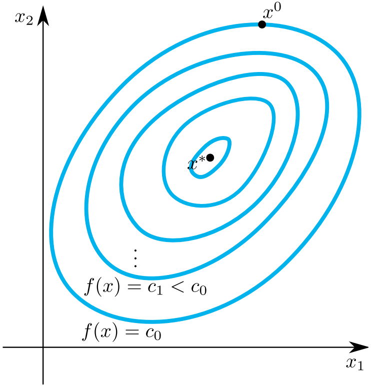
- Nouvelle règle de mise à jour :

$$\boxed{x^{t+1} \leftarrow x^t - (\nabla^2 f(x^t))^{-1} \nabla f(x^t)}$$

Rem: C'est donc la méthode de Newton appliquée à la recherche de zéros d'une approximation du gradient de f

Méthode de Newton : algorithme

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T-1$ **do**
$$x^{t+1} \leftarrow x^t - (\nabla^2 f(x^t))^{-1} \nabla f(x^t)$$
STOP si critère d'arrêt inférieur à ε
end



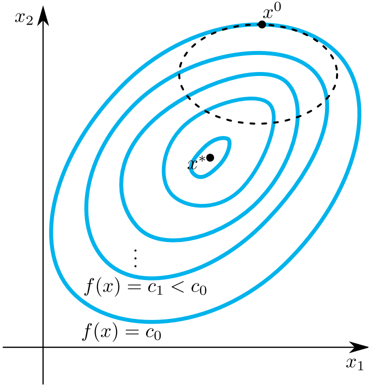
Méthode de Newton : algorithme

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T-1$ **do**

$$x^{t+1} \leftarrow x^t - (\nabla^2 f(x^t))^{-1} \nabla f(x^t)$$

STOP si critère d'arrêt inférieur à ε

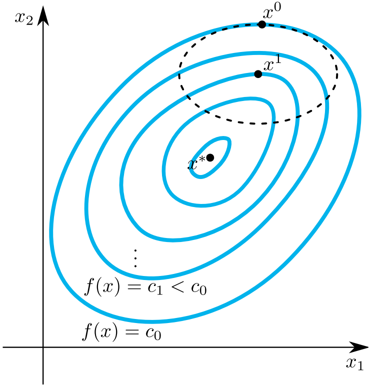
end



Joseph Salmon

Méthode de Newton : algorithme

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T-1$ **do**
    $x^{t+1} \leftarrow x^t - (\nabla^2 f(x^t))^{-1} \nabla f(x^t)$
    STOP si critère d'arrêt inférieur à ε
end



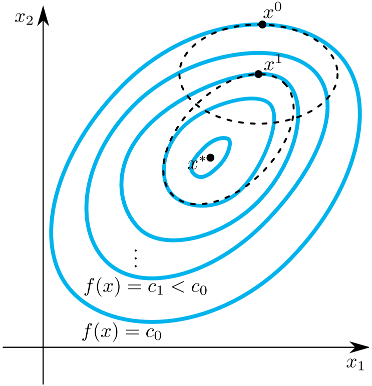
Méthode de Newton : algorithme

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T-1$ **do**

$x^{t+1} \leftarrow x^t - (\nabla^2 f(x^t))^{-1} \nabla f(x^t)$

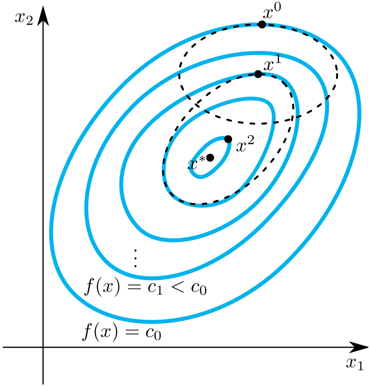
STOP si critère d'arrêt inférieur à ε

end



Méthode de Newton : algorithme

Data: point initial x^0, nombre max. d'itérations T, critère d'arrêt ε
Result: un point x_T "proche" du minimum de la fonction f
for $1 \leq t \leq T-1$ **do**
  $x^{t+1} \leftarrow x^t - (\nabla^2 f(x^t))^{-1} \nabla f(x^t)$
  STOP si critère d'arrêt inférieur à ε
end



Références I

D. P. Bertsekas.
Nonlinear programming.
Athena Scientific, 1999.