Optimisation
Séance 1

Maxime Berar

23 septembre 2024

Rappel : *Gradient Descent*

Pour cet algorithme on choisit $d^{(k)} = -\nabla f(x^{(k)})$ comme direction de descente,

Descente de gradient

$\texttt{choisir}\ x^{(0)} \in \mathrm{dom} f\ \texttt{et poser}\ \ k = 0.$
$\texttt{tant que}\ \|\nabla f(x^{(k)}\| > \varepsilon$
$\quad d^{(k)} = -\nabla f(x^{(k)})$
$\quad \texttt{déterminer}\ \sigma_k > 0$ (rebroussement)
$\quad x^{(k+1)} = x^{(k)} + \sigma_k d^{(k)}$
$\quad k = k + 1$

Direction de plus forte descente dépend de la norme

Steepest Descent dépend de la norme choisie ...

$$d = \arg\min_d \{d^\top \nabla f(x) \mid \|d\| = 1\}$$

- norme euclidienne $\|x\|_2^2 = \sum_{i=1}^n x_i^2$

$$d = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$$

  Steepest Descent avec $||\cdot||_2 \Longrightarrow$ **Gradient Descent**
- norme P-quadratique $\|x\|_P^2 = x^\top P x$

$$d = -\frac{P^{-1}\nabla f(x)}{\sqrt{(\nabla f(x))^\top P^{-1} \nabla f(x)}}$$

Par exemple, dans le cas du problème quadratique, pourquoi ne pas choisir la seconde norme ?

Problème quadratique avec la seconde norme $\|\cdot\|_P^2$

$$f(x) = .5x^\top P x + q^\top x + r, \quad \nabla f(x) = Px + q$$

$$d^{(k)} = -\frac{x^{(k)} + P^{-1}q}{\sqrt{\nabla f(x^{(k)}) P^{-1} \nabla f(x^{(k)})}} = -\frac{x^{(k)} - x^*}{\sqrt{\nabla f(x^{(k)}) P^{-1} \nabla f(x^{(k)})}}$$

Steepest Descent avec $||\cdot||_P$

choisir $x^{(0)} \in \mathrm{dom} f$ et poser $k = 0$.

tant que $\|\nabla f(x^{(k)}\| > \varepsilon$

$d^{(k)} = -x^{(k)} + P^{-1}q$

déterminer $\sigma_k > 0$ (backtracking ! ! !)

$x^{(k+1)} = (1 - \sigma_k)x^{(k)} + \sigma_k x^*$

$k = k + 1$

Si $\sigma_k = 1$, convergence en 1 itération ... pour le problème quadratique

Approximation quadratique d'un problème

Si $\|d\| = 1$ et $\varepsilon > 0$,

$$f(x + \varepsilon d) = f(x) + \varepsilon D_d f(x) + o(\varepsilon)$$

$$f(x + \varepsilon d) = f(x) + \varepsilon D_d f(x) + \frac{\varepsilon^2}{2} d^\top (D_d^2 f(x)) + o(\varepsilon^2)$$

$$f(x + \varepsilon d) = f(x) + \varepsilon d^\top \nabla f(x) + \frac{\varepsilon^2}{2} d^\top H_f(x) d + o(\varepsilon^2)$$

On suppose que f est de classe $\mathcal{C}^2$.

Approximation quadratique d'un problème (2)

Si $H_f(x^{(k)}) \succ 0$, on peut définir une nouvelle fonction quadratique

$$\tilde{f}_k(x) = f(x^{(k)}) + (x - x^{(k)})^\top \nabla f(x^{(k)}) + .5(x - x^{(k)})^\top H_f(x^{(k)})(x - x^{(k)})$$

$$\tilde{f}_k(x) = \underbrace{.5x^\top H_f(x^{(k)})x}_{.5x^\top Px} \underbrace{-x^\top \left(H_f(x^{(k)})x^{(k)} - \nabla f(x^{(k)})\right)}_{+x^\top q}$$

$$+ \underbrace{f(x^{(k)}) - (x^{(k)})^\top \nabla f(x^{(k)}) + .5(x^{(k)})^\top H_f(x^{(k)})x^{(k)}}_{r}$$

pour laquelle, un pas de descente avec la norme induite par $H_f(x^{(k)})$ est

$$d^k = -x^{(k)} + H_f(x^{(k)})^{-1} \left(H_f(x^{(k)})x^{(k)} - \nabla f(x^{(k)})\right)$$

$$d^k = -x^{(k)} + x^{(k)} - H_f(x^{(k)})^{-1} \nabla f(x^{(k)})$$

$$d^k = -H_f(x^{(k)})^{-1} \nabla f(x^{(k)})$$

Succession d'approximation quadratique.

A chaque itération, si $\nabla^2 f(x^{(k)})$ est définie positive, on choisit la direction de descente $d^{(k)}$ qui nous mène au minimum de l'approxiation quadratique

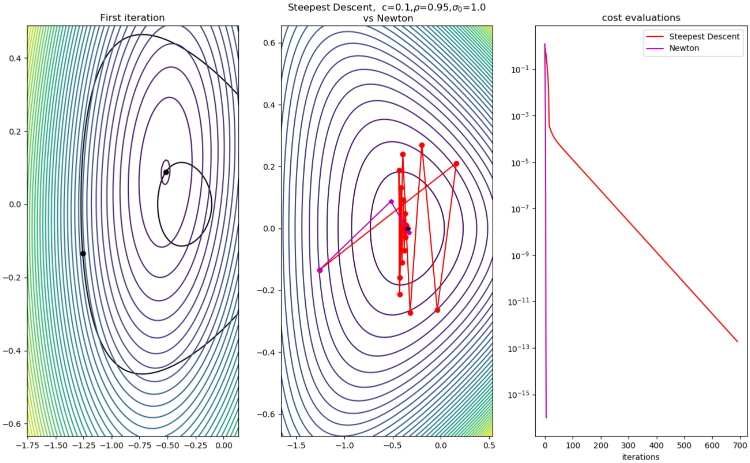
$$d^{(k)} = -[H_f(x^{(k)})]^{-1}\nabla f(x^{(k)})$$

- On obtient bien une direction de descente, car $\nabla f(x^{(k)})^\top d^{(k)} = -\nabla f(x^{(k)})^\top H_f(x^{(k)})\nabla f(x^{(k)}) \leq 0$,

Remarque, le pas est inclus ...

$$d^{(k)} = \underbrace{\sqrt{\nabla f(x^{(k)})^\top [H_f(x^{(k)})]^{-1}\nabla f(x^{(k)})}}_{\text{Pas de Newton}} * \underbrace{-\frac{[H_f(x^{(k)})]^{-1}\nabla f(x^{(k)})}{\sqrt{\nabla f(x^{(k)})^\top [H_f(x^{(k)})]^{-1}\nabla f(x^{(k)})}}}_{\text{Direction de Newton}}$$

Exemple



First iteration
Steepest Descent, c=0.1,ρ=0.95,σ0=1.0
vs Newton
cost evaluations
Steepest Descent
Newton
iterations

Méthode de Newton

Théorême

Soit f de classe $\mathcal{C}^2$, $x \in \mathbb{R}^n$ tel que $\nabla f(\bar{x}) = 0$ et $H_f(\bar{x})$ définie positive et H_f une fonction Lipschitz au voisinage de x.
On considère la suite $(x^{(k)})$ définie par $x^{(0)}$ et

$$x^{(k+1)} = x^{(k)} - [H_f(x^{(k)})]^{-1} \nabla f(x^{(k)}).$$

Alors si $x^{(0)}$ est suffisamment proche de $\bar{x}$

(i) la suite $(x^{(k)})$ converge vers $\bar{x}$;

(ii) la méthode de Newton est d'ordre 2 ;

(iii) la suite $(\|\nabla f(x)^{(k)}\|)$ converge vers 0 de façon quadratique.

Remarques

- Si en un point stationnaire, la matrice hessienne est définie positive et si $x^{(0)}$ est proche de $\bar{x}$, on a une convergence très rapide. Dans le cas contraire l'algorithme peut diverger.
- Par contre le coût de cette méthode est grand : à chaque itération on doit construire et garder en mémoire la matrice hessienne et résoudre $H_f(x^{(k)})d = \nabla f(x^{(k)})$, en utilisant l'algorithme de Cholesky cela fait $O(n^2)$ opérations.
- Les méthodes quasi-Newton remplacent la matrice hessienne par une approximation B^k qui vérifie la relation de la sécante

$$B^k(x^{(k)} - x^{(k-1)}) = \nabla f(x^{(k)}) - \nabla f(x^{(k-1)}).$$

Remarques (cont.)

cf Convex Optimization, Boyed et al.

- Méthode de Newton avec rebroussement : possible, on parle alors de *damped Newton method* ou *guarded Newton method* $\Rightarrow$ *souvent nécessaire en pratique*
  pour contrôler la taille du pas ...

- Autre critère d'arrêt lié au pas de *Newton*

$$\frac{\lambda^2(x^k)}{2} = \frac{\nabla f(x^{(k)})^\top H_f(x^{(k)})^{-1} \nabla f(x^{(k)})}{2} < \varepsilon$$

  que l'on place avant la recherche de pas.

Peut-on construire une preuve de convergence indépendante de la direction de descente choisie ?

- Direction de descente : $\left\langle d, \nabla f(x^k) \right\rangle < 0$
- englobe la descente de gradient, la méthode de Newton ou d'autres variantes à construire.
- à l'exception du gradient conjugué.

Convergence des méthodes de recherche linéaire

élément 0 : f est bornée par en dessous
élément 1 : angle entre la direction de descente d_k et $-\nabla f(x^{(k)})$

$$\cos\theta_k = \frac{-\nabla f(x^{(k)})^\top d^{(k)}}{\|\nabla f(x^{(k)})\| \|d^{(k)}\|}$$

élément 2 : gradient est Lipschitz continu, ie pour tout $(x, \tilde{x}) \in \mathcal{D}^2$

$$\|\nabla f(x) - \nabla f(\tilde{x})\| \leq L\|x - \tilde{x}\|$$

élément 3 : conditions de Wolfe

Conditions de Wolfe

$$f(x^{(k)} + \sigma d^{(k)}) \leq f(x^{(k)}) + c_1\sigma(\nabla f(x^{(k)})^\top d^{(k)}) \qquad (W1)$$
$$\nabla f(x^{(k)} + \sigma d^{(k)})^\top d^{(k)} \geq c_2(\nabla f(x^{(k)})^\top d^{(k)}) \qquad (W2)$$

avec $0 < c_1 < c_2 < 1$.

Théorême

Proposition

Soit $f \in \mathcal{C}^1(\mathbb{R}^n)$ et d une direction de descente de f en x, on suppose que f est minorée sur $\{x + \sigma d | \sigma \geq 0\}$. Alors, si $0 < c_1 < c_2 < 1$, il existe des intervalles dans $\mathbb{R}^+$ qui vérifient les conditions de Wolfe faibles et fortes.

Théorême (Zoutendijk)

Soit f de classe $\mathcal{C}^1$ sur l'ouvert $\mathcal{D} \subset \mathbb{R}^n$ et $x^{(0)} \in \mathcal{D}$ tel que l'ensemble de niveau inférieur $L_f(f(x^{(0)})) = \{x \in \mathcal{D} | f(x) \leq f(x^{(0)})\}$ est fermé, de plus on suppose que f est minoré sur $L_f(f(x^{(0)}))$ et que ∇f est Lipschitz sur $\mathcal{D}$.
Considérons la suite $(x^{(k)})_k$, définie pour tout $k \geq 0$ par $x_{(k+1)} = x^{(k)} + \sigma_k d^{(k)}$, où $d^{(k)}$ est une direction de descente et σ_k vérifie les conditions de Wolfe , alors

$$\sum_{k=0}^{+\infty} \cos^2 \theta_k \|\nabla f(x^{(k)})\|_2^2 < +\infty, \quad \text{où} \quad \cos\theta_k = -\frac{\nabla f(x^{(k)})^\top d^{(k)}}{\|\nabla f(x^{(k)})\| \|d^{(k)}\|}$$

Corollaire

Si f et la méthode de descente $(x^{(k)})_k$ vérifient les hypothèses du théorème précédent et s'il existe $\delta > 0$ tel que pour k suffisamment grand on a $\cos\theta_k \geq \delta$, alors

$$\lim_{k \to +\infty} \|\nabla f(x^{(k)})\| = 0$$

Démonstration

Conditions W2

$$\begin{aligned}
&\nabla f(x^{(k)} + \sigma_k d^{(k)})^\top d^{(k)} \geq c_2(\nabla f(x^{(k)})^\top d^{(k)})\\
&\nabla f(x^{(k+1)})^\top d^{(k)} \geq c_2(\nabla f(x^{(k)})^\top d^{(k)})\\
&\left(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\right)^\top d^{(k)} \geq (c_2 - 1)(\nabla f(x^{(k)})^\top d^{(k)})
\end{aligned}$$

Lipschitz conditions

$$\begin{aligned}
&\|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\| \leq L\|x^{(k+1)} - x^{(k)}\|\\
&\|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\| \leq \sigma_k L\|d^{(k)}\|\\
&\|\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\|\|d^{(k)}\| \leq \sigma_k L\|d^{(k)}\|^2\\
&\left(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\right)^\top d^{(k)} \leq \sigma_k L\|d^{(k)}\|^2
\end{aligned}$$

Combinaison des 2 inégalités

$$\begin{aligned}
&\sigma_k L\|d^{(k)}\|^2 \geq \left(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\right)^\top d^{(k)} \geq (c_2 - 1)(\nabla f(x^{(k)})^\top d^{(k)})\\
&\sigma_k \geq \frac{c_2 - 1}{L}\frac{(\nabla f(x^{(k)})^\top d^{(k)})}{\|d^{(k)}\|^2}
\end{aligned}$$

Démonstration (cont.)

Condition de Wolfe

$$f(x^{(k+1)}) \leq f(x^{(k)}) + c_1 \sigma (\nabla f(x^{(k)})^\top d^{(k)}) \qquad (W1)$$

et résultat précédent

$$\sigma_k \geq \frac{c_2 - 1}{L} \frac{(\nabla f(x^{(k)})^\top d^{(k)})}{\|d^{(k)}\|^2}$$

donnent

$$f(x^{(k+1)}) \leq f(x^{(k)}) - c_1 \frac{1 - c_2}{L} \frac{(\nabla f(x^{(k)})^\top d^{(k)})^2}{\|d^{(k)}\|^2}$$

$$f(x^{(k+1)}) \leq f(x^{(k)}) - c \cos^2 \theta_k \|\nabla f(x^{(k)})\| 2$$

Démonstration (cont.)

$$f(x^{(k+1)}) - f(x^{(k)}) \leq -c\cos^2\theta_k \|\nabla f(x^{(k)})\|^2$$

Somme sur l'ensemble des indices

$$f(x^{(k+1)}) \leq f(x^{(0)}) - c\sum_{j=0}^{k} \cos^2\theta_k \|\nabla f(x^{(k)})\|^2$$

Comme on fait l'hypothèse que f est bornée par en dessous

$$\sum_{j=0}^{+\infty} \cos^2\theta_k \|\nabla f(x^{(k)})\|^2 < \infty$$

Conséquences

La condition $\sum_{j=0}^{+\infty} \cos^2 \theta_k \|\nabla f(x^{(k)})\|^2 < \infty$ (Zoutendijk) implique que

$$\cos^2 \theta_k \|\nabla f(x^{(k)})\|^2 \longrightarrow 0.$$

Si une méthode de descente s'assure que la direction d^k est choisie telle que l'angle θ_k est borné loin de 90°, alors il existe une constante positive δ telle que

$$\cos \theta_k \geq \delta > 0, \quad \text{pour tout } k$$

On a alors immédiatement

$$\lim_{k \to \infty} \|\nabla f(x^{(k)})\| = 0$$

Comment ça s'applique ?

Hypothèse $\nabla f(x^{(k)}) \neq 0$

Condition sur la direction de descente

$$\cos\theta_k \geq \delta > 0, \quad \text{pour tout } k$$

Gradient

$$\cos\theta_k = 1, \quad \text{pour tout } k$$

Steepest

$$d^{(k)} = \arg\min_d \left\{ \nabla f(x^{(k)})^\top d, \|d\|_? = 1 \right\}, \quad \cos\theta_k = -\frac{\nabla f(x^{(k)})^\top d^{(k)}}{\|\nabla f(x^{(k)})\|} \geq \delta$$

Newton

$$d^{(k)} = -\left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)})$$

ok si Hessienne définie positive pour tout k

Dichotomie pour les règles de Wolfe

Comment trouver efficacement un pas $\sigma > 0$, satisfaisant les 2 conditions de Wolfe ?

$$f(x+\sigma d) \leq f(x) + c_1 \sigma d^\top \nabla f(x) \qquad (W1)$$
$$\nabla f(x+\sigma d)^\top d \geq c_2 (\nabla f(x)^\top d) \qquad (W2)$$

Méthode de dichotomie

On construit deux suites (σ_b) et (σ_h) tel que $\sigma_b < \sigma_h$ et tel que les (σ_b) satisfont le critère W_1, et les (σ_h) ne le satisfont pas.

$$\left(\sigma_b^{k+1}, \sigma_h^{k+1}\right) = \begin{cases} \left(\frac{\sigma_b^k + \sigma_h^{(k)}}{2}, \sigma_h^{(k)}\right), & \text{si } \frac{\sigma_b^k + \sigma_h^{(k)}}{2} \text{ satisfait (W1)} \\ \left(\sigma_b^{(k)}, \frac{\sigma_b^k + \sigma_h^{(k)}}{2}\right), & \text{sinon.} \end{cases}$$

Convergence de la dichotomie

$$\nabla f(x+\sigma d)^\top d \geq c_2(\nabla f(x)^\top d) \qquad (W2)$$

Conditions

$$f(x+\sigma_b d) \leq f(x) + c_1\sigma_b \langle d, \nabla f(x)\rangle$$
$$f(x+\sigma_h d) \geq f(x) + c_1\sigma_h \langle d, \nabla f(x)\rangle$$

Pour tout couple de la suite

$$f(x+\sigma_h d) - f(x+\sigma_b^d) \geq c_1(\sigma_h - \sigma_b) \langle d, \nabla f(x)\rangle$$
$$\frac{f(x+\sigma_h d) - f(x+\sigma_b d)}{(\sigma_h - \sigma_b)} \geq c_1 \langle d, \nabla f(x)\rangle$$

A la limite,

$$D_d f(x+\sigma^* d) = \langle \nabla f(x+\sigma^* d), d\rangle \geq c_1 \langle d, \nabla f(x)\rangle \geq c_2 \langle d, \nabla f(x)\rangle$$