C10, algorithmes pour l'optimisation sans contrainte

$f \in C(\mathbb{R}^n, \mathbb{R})$.On suppose qu'il existe $\bar{x} \in \mathbb{R}^n$ tel que $f(\bar{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^n$

Objectif : calculer un tel point $\bar{x}$
Nous allons distinguer deux types de méthodes

1. Méthodes de descente
   Il s'agit ici de construire une suite $(x^{(k)})_{k\in\mathbb{N}}$ telle que

   $$f(x^{(k+1)}) \leq f(x^{(k)}) \text{ pour tout } k \in \mathbb{N},$$

   en espérant que $\lim_{k\to+\infty} x^{(k)} = \bar{x}$
2. Si f est de classe C^1, chercher une solution de l'équation $\nabla f(x) = 0$ (puis vérifier que f est bien minimale au point x). Cette équation s'appelle "Equation d'Euler" du problème de minimisation de f

Direction de descente

Definition

$f \in C(\mathbb{R}^n, \mathbb{R})$, $x \in \mathbb{R}^n$, $w \in \mathbb{R}^n$, $w \neq 0$

1. *w est une direction de descente (dd) au point x si il existe $a_0 > 0$ tel que*

$$f(x + aw) \leq f(x) \text{ pour tout } a \in]0, a_0]$$

2. *w est une direction de descente (dds) au point x si il existe $a_0 > 0$ tel que*

$$f(x + aw) < f(x) \text{ pour tout } a \in]0, a_0]$$

Méthode de descente

Definition

$f \in C(\mathbb{R}^n, \mathbb{R})$, *Une méthode de descente consiste à construire une suite de la manière suivante*

Initialisation *Choisir* $x^{(0)} \in \mathbb{R}^n$

Itération *pour* $k \geq 0$

- *Choisir (si c'est possible) une dds au point* $x^{(k)}$*, notée* $w^{(k)}$
- *Prendre* $x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)}$ *avec* α_k *bien choisi de manière à avoir, en particulier,* $f(x^{(k+1)}) < f(x^{(k)})$

Condition nécessaire et condition suffisante pour une dds

Proposition

$f \in C^1(\mathbb{R}^n, \mathbb{R})$, $x \in \mathbb{R}^n$, $w \in \mathbb{R}^n$, $w \neq 0$

1. *(CN) Si w est une dds au point x, alors $w \cdot \nabla f(x) \leq 0$*
2. *(CS) Si $w \cdot \nabla f(x) < 0$ alors w est une dds au point x*

Exemple fondamental :

Si $\nabla f(x) \neq 0$ alors $w = -\nabla f(x)$ est une dds au point x

Démonstration de la proposition

Hypothèse : $f \in C^1(\mathbb{R}^n, \mathbb{R})$, $x \in \mathbb{R}^n$, $w \in \mathbb{R}^n$, $w \neq 0$
$\varphi(t) = f(x + tw)$, de sorte $\varphi \in C^1(\mathbb{R}, \mathbb{R})$, $\varphi'(t) = \nabla f(x + tw) \cdot w$

1. (CN) Si w est une dds au point x,

   il existe $a_0 > 0$ tel que $\varphi(t) = f(x + tw) < f(x)$ pour tout $t \in]0, a_0]$,

   $$\frac{\varphi(t) - \varphi(0)}{t} < 0 \textit{ pour tout } t \in]0, a_0],$$

   et donc, quand $t \to 0$, $\varphi'(t) = \nabla f(x) \cdot w \leq 0$

2. (CS) Si $w \cdot \nabla f(x) < 0$, on a $\varphi'(0) = \nabla f(x) \cdot w < 0$

   Il existe $a_0 > 0$ tel que $\varphi'(t) < 0$ pour tout $t \in]0, a_0]$, et donc (par le théorème des Accroissements Finis) $\varphi(t) < \varphi(0)$ tout $t \in]0, a_0]$

   Ceci prouve que w est une dds au point x

Algorithme du Gradient à Pas Fixe (GPF)

Hypothèse : $f \in C^1(\mathbb{R}^n, \mathbb{R})$. On choisit $\rho > 0$

Initialisation $x^{(0)} \in \mathbb{R}^n$.

Itération pour $k \geq 0$, on choisit $w^{(k)} = -\nabla f(x^{(k)})$

$$x^{(k+1)} = x^{(k)} + \rho w^{(k)}$$

Si $\nabla f(x^{(k)}) \neq 0$, $w^{(k)}$ est une dds au point $x^{(k)}$, mais deux questions :

1. A t'on $f(x^{(k+1)}) \leq f(x^{(k)})$?
2. A t'on $\lim_{k \to +\infty} x^{(k)} = \bar{x}$?

En général, non et non...

Convergence de l'algorithme GPF

Theorem

Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$. On suppose qu'il existe $\alpha > 0$ et $M > 0$ tels que

(mon) $(\nabla f(x) - \nabla f(y)) \cdot (x - y) \geq \alpha |x - y|^2$ *pour tout* $x, y \in \mathbb{R}^n$

(lip) $|\nabla f(x) - \nabla f(y)| \leq M |x - y|$ *pour tout* $x, y \in \mathbb{R}^n$

Alors, si $\rho < \frac{2\alpha}{M^2}$, on a bien

(des) $f(x^{(k+1)}) \leq f(x^{(k)})$ *pour tout* $k \geq 0$

(cve) $\lim_{k \to +\infty} x^{(k)} = \bar{x}$ *et* $\bar{x} = \mathrm{argmin}_{\mathbb{R}^n} f$

petit rappel : l'hypothèse (mon) implique que f est strictement convexe et $f(x) \to +\infty$ quand $|x| \to +\infty$ et donc qu'il existe un unique $\bar{x} \in \mathbb{R}^n$ tel que $f(\bar{x}) \leq f(x)$ pour tout $x \in \mathbb{R}^n$. De plus $\bar{x}$ est caractérisé par $\nabla f(\bar{x}) = 0$

Démonstration du théorème, convergence vers $\bar{x}$

On pose $h(x) = x - \rho \nabla f(x)$

L'algorithme (GPF) est l'algorithme du point fixe pour h. On déjà vu (Cours 6, point fixe de relaxation) que sous les hypothèses (mon)-(lip), si $\rho < \frac{2\alpha}{M^2}$, la fonction h est strictement contractante et $\lim_{k \to +\infty} x^{(k)} = \bar{x}$ avec $\nabla f(\bar{x}) = 0$ (et donc ici $\bar{x} = argmin_{\mathbb{R}^n} f$)

rappel du cours 6, $h(x) = x - \omega g(x)$ avec g vérifiant (mon)-(lip) et $0 < \omega < \frac{2\alpha}{M^2}$, théorème 2.8

Dém. du théorème, (GPF) est une méthode de descente

On suppose que f vérifie (mon)-(lip), $\rho < \frac{2\alpha}{M^2}$, $\nabla f(x^{(k)}) \neq 0$, $x^{(k+1)} = x^{(k)} - \rho \nabla f(x^{(k)})$. On va montrer $f(x^{(k+1)}) < f(x^{(k)})$

Remarque : Les hypothèse (mon)-(lip) impliquent $\alpha \leq M$ car $\alpha |x-y|^2 \leq (\nabla f(x) - \nabla f(y)) \cdot (x-y) \leq M|x-y|^2$
et donc $\rho < \frac{2\alpha}{M^2} < \frac{2}{M}$

$y = x^{(k+1)}$, $x = x^{(k)}$, $y = x - \rho \nabla f(x)$

$$f(y) - f(x) = \int_0^1 \nabla f(x + t(y-x)) \cdot (y-x) dt =$$

$$\int_0^1 (\nabla f(x + t(y-x)) - \nabla f(x)) \cdot (y-x) dt + \nabla f(x) \cdot (y-x)$$

$$\leq \frac{M}{2} |y-x|^2 - \frac{1}{\rho} |y-x|^2 < 0,$$

car $\rho < \frac{2}{M}$

Algorithme du Gradient à Pas Optimal (GP0)

Hypothèse : $f \in C^1(\mathbb{R}^n, \mathbb{R})$

Initialisation $x^{(0)} \in \mathbb{R}^n$

Itération pour $k \geq 0$,

1. on choisit $w^{(k)} = -\nabla f(x^{(k)})$,
2. on choisit (si c'est possible) $\rho_k > 0$ tel que $f(x^{(k)} + \rho_k w^{(k)}) \leq f(x^{(k)} + \rho w^{(k)})$ pour tout $\rho \geq 0$,
3. $x^{(k+1)} = x^{(k)} + \rho_k w^{(k)}$

Si $\nabla f(x^{(k)}) \neq 0$, $w^{(k)}$ est une dds au point $x^{(k)}$, mais trois questions :

1. Existence de ρ_k ?
2. Calcul de ρ_k ?
3. A t'on $\lim_{k \to +\infty} x^{(k)} = \bar{x}$?

Convergence de l'algorithme GP0, questions 1 et 3

Theorem

Soit $f \in C^1(\mathbb{R}^n, \mathbb{R})$ *telle que* $f(x) \to +\infty$ *quand* $\|x\| \to +\infty$ *Alors*

1. *la suite* $(x^{(k)})_{k\in\mathbb{N}}$ *est bien définie par l'alogorithme (GPO) (c'est-à-dire que* ρ_k *existe pour tout* k*),*
2. *la suite* $(x^{(k)})_{k\in\mathbb{N}}$ *est bornée et toute sous suite convergente converge vers un point qui annule* ∇f*,*
3. *si* f *est convexe, toute sous suite convergente converge vers un point appartenant à* $\text{argmin}_{\mathbb{R}^n} f$*,*
4. *si* f *est strictement convexe* $\lim_{k\to+\infty} x^{(k)} = \bar{x}$ *et* $\bar{x} = \text{argmin}_{\mathbb{R}^n} f$

La démonstration sera faite en td, le point difficile est le deuxième item

Calcul de ρ_k dans (GPO), question 2

$x^{(k)}$ est connu,
$w^{(k)} = -\nabla f(x^{(k)}) \neq 0$ (sinon $x^{(k+1)} = x^{(k)}$, l'algorithme s'arrête)
On suppose qu'il existe $\rho_k \geq 0$ tel
$f(x^{(k)} + \rho_k w^k) \leq f(x^{(k)} + \rho w^k)$ pour tout $\rho \geq 0$

On pose $\varphi(\rho) = f(x^{(k)} + \rho w^{(k)})$
Comme $w^{(k)}$ est une dds au point $x^{(k)}$ on a $\rho_k > 0$ et donc
$\varphi'(\rho_k) = 0$, c'est-à-dire

$$\nabla f(x^{(k)} + \rho_k w^{(k)}) \cdot w^{(k)} = 0$$

Ceci permet parfois de calculer ρ_k

Noter aussi $\nabla f(x^{(k+1)}) \cdot w^{(k)} = 0$

Calcul de ρ_k, fonctionnelle quadratique

$f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$ où $A \in \mathcal{M}_n(\mathbb{R})$ est s.d.p. et $b \in \mathbb{R}^n$
$\nabla f(x) = Ax - b$
Pour $w \neq 0$, on cherche ρ tel que

$$\nabla f(x + \rho w) \cdot w = 0,$$

c'est-à-dire $(Ax + \rho Aw - b) \cdot w = 0$ et donc

$$\rho = \frac{(b - Ax) \cdot w}{Aw \cdot w}$$

Ceci permet le calcul de ρ quelquesoit w dds au point x

Le cas d'une fonctionnelle générale est plus compliqué

Algorithme du gradient conjugué (1)

$f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$ où $A \in \mathcal{M}_n(\mathbb{R})$ est s.d.p. et $b \in \mathbb{R}^n$
Rappel : $\nabla f(x) = Ax - b$
idée (1952) :
Construire la suite des itérées $x^{(k)}$ telle que :

1. Pour chaque $k \geq 0$, $x^{(k+1)} = x^{(k)} + \rho_k w^{(k)}$ avec $w^{(k)}$ dds en $x^{(k)}$ et ρ_k optimal dans la direction $w^{(k)}$
2. $Aw^{(k)} \cdot w^{(l)} = 0$ si $l < k$ (orthogonalité des $w^{(k)}$ pour le produit scalaire induit par A)
3. $(b - Ax^{(k)}) \cdot w^{(l)} = 0$ si $l < k$

La propriété 1 est facile à obtenir si $w^{(k)}$ est une dds. Pour les propriétés 2 et 3, la seule égalité facile est

$$(b - Ax^{(k)}) \cdot w^{(k-1)} = -\nabla f(x^{(k)}) \cdot w^{(k-1)} = 0,$$

car ρ_k est optimal dans la direction $w^{(k)}$

Algorithme du gradient conjugué (2)

intérêt (si on peut construire une telle suite) :
La suite s'arrête au plus tard à l'itération n car on alors $(b - Ax^{(n)})$ orthogonal à la famille $\{w^{(0)} \dots w^{(n-1)}\}$ qui forme une base de $\mathbb{R}^n$ (car c'est une famille de n vecteurs orthogonaux 2 à 2 non nuls).
Donc, $Ax^{(n)} = b$

il s'agit donc d'une méthode directe mais on peut même espèrer avoir "presque" la solution en moins de n itérations

Pour avoir les propriètes demandées, il suffit (par miracle) de choisir pour $k > 0$
$w^{(k)} = (b - Ax^{(k)}) + \lambda_k w^{(k-1)}$ avec λ_k choisi pour avoir $Aw^{(k)} \cdot w^{(k-1)} = 0$

(On a alors $Aw^{(k)} \cdot w^{(l)} = 0$ et $(b - Ax^{(k)}) \cdot w^{(l)} = 0$ pour tout $l < k$ et pas seulement pour $l = k - 1$)

Gradient conjugué préconditionné

Intérêt réel de la méthode du gradient conjugué : nul...

Précondionnement, 1980 :
La méthode est extrèmement intéressante si on l'utilise avec un "préconditionnement" consistant à remplacer A par $L^{-1}A(L^t)^{-1}$ et b par $L^{-1}b$, où L est "proche" de la matrice de factorisation de A par la méthode de Choleski

Le bon choix de la matrice de précondionnement dépend du problème considéré

Méthode de Newton pour minimiser f

On cherche par la méthode de Newton un point annulant ∇f.
Rappel : $J_{\nabla f}(x) = H_f(x)$ (on suppose f de classe C^2)

Initialisation $x^{(0)} \in \mathbb{R}^n$
Itération pour $k \geq 0$, $H_f(x^{(k)})(x^{(k+1)} - x^{(k)}) = -\nabla f(x^{(k)})$

Si la matrice $H_f(x^{(k)})$ est s.d.p. (ce qui est presque toujours vrai si f est strictement convexe), cette méthode est une méthode de descente
Elle s'écrit $x^{(k+1)} = x^{(k)} + w^{(k)}$ avec
$w^{(k)} = -(H_f(x^{(k)})^{-1}\nabla f(x^{(k)})$ et, si $\nabla f(x^{(k)}) \neq 0$,

$$w^{(k)} \cdot \nabla f(x^{(k)}) = -(H_f(x^{(k)})^{-1}\nabla f(x^{(k)}) \cdot \nabla f(x^{(k)}) < 0$$

Intérêt de la méthode : convergence quadratique
Inconvénient de la méthode : calculer la Hessienne de f

Méthode de quasi-Newton pour minimiser f

Objectif : éviter de calculer la Hessienne de f
Idée : remplacer dans la méthode de Newton la matrice $H_f(x^{(k)})$ par une matrice $B^{(k)}$, facile à calculer et telle que :

1. $B^{(k)}$ est s.d.p.
2. $B^{(k)}$ approche de mieux en mieux la hessiene de f au cours des itérations

Initialisation $x^{(0)} \in \mathbb{R}^n$
Itération pour $k \geq 0$, $B^{(k)}(x^{(k+1)} - x^{(k)}) = -\nabla f(x^{(k)})$

Question : Comment choisir $B^{(k)}$?

Choix de $B^{(k)}$, méthode (BFGS)

Rappel (idée de Broyden) :
Pour que $B^{(k)}$ se rapproche de la hessienne de f on va demander
$B^{(k)}(x^{(k)} - x^{(k-1)}) = \nabla f(x^{(k)}) - \nabla f(x^{(k-1)})$

Initialisation $B^{(0)}$ s.d.p. (par exemple $B^{(0)} = I$)
Itération pour $k > 0$, on note
$s^{(k)} = x^{(k)} - x^{(k-1)}$, $y^{(k)} = \nabla f(x^{(k)}) - \nabla f(x^{(k-1)})$.
Si $s^{(k)} \neq 0$, on choisit

$$B^{(k)} = B^{(k-1)} + \frac{y^{(k)}(y^{(k)})^t}{(y^{(k)})^t s^{(k)}} - \frac{B^{(k-1)} s^{(k)} (s^{(k)})^t B^{(k-1)}}{(s^{(k)})^t B^{(k-1)} s^{(k)}}$$

On a bien $B^{(k)}$ symétrique et $B^{(k)}(s^{(k)}) = y^{(k)}$

Theorem (1976)

$f \in C^2(\mathbb{R}^n, \mathbb{R})$ strictement convexe et $f(x) \to +\infty$ quand $\|x\| \to +\infty$.
On note $\bar{x} = \text{argmin}_{\mathbb{R}^n} f$. On suppose $H_f(\bar{x})$ est s.d.p.
Alors $\lim_{k \to +\infty} x^{(k)} = \bar{x}$ et la convergence est superlinéaire

Remarque : sous les hypothèses du théorème, $(y^{(k)})^t s^{(k)} > 0$ si $s^{(k)} \neq 0$

Résumé

1. Méthodes de gradient, (GPF) et (GPO). La convergence est en général linéaire
2. Méthode de gradient conjugué pour fonctionnelle quadratique (donc correspond à la résolution d'un système linéaire)
3. Méthode de Newton. La convergence est quadratique
4. Méthode de quasi-Newton. La convergence est superlinéaire (mais, par rapport aux méthodes de gradient, demande, à chaque itération, la résolution d'un système linéaire)