– Optimisation –
Résumé de cours

D. Bresch-Pietri

Mines Paris – PSL

3 avril 2023

Conditions d'optimalité : cas général

Existence

Théorème 1 (Weierstrass)

Si f est une fonction réelle continue sur un compact $K \subset \mathbb{R}^n$ alors le problème de recherche de minimum global

$$\min_{x \in K} f(x)$$

possède une solution $x^* \in K$.

Théorème 2

Soit $f : \mathbb{R}^n \to \mathbb{R}$ continue et telle que $\lim_{\|x\| \to \infty} f(x) = +\infty$. Alors, pour tout F fermé non-vide de $\mathbb{R}^n$, il existe une solution au problème $\min_{x \in F} f(x)$.

Conditions d'optimalité : cas général

Condition nécessaire sur un ouvert

Théorème 3

Soit Ω un ouvert de $\mathbb{R}^n$. Une condition nécessaire pour que x^* soit un optimum local de $\Omega \ni x \mapsto f(x) \in \mathbb{R}$ fonction deux fois différentiable est

$$\left\{\nabla f(x^*) = 0, \nabla^2 f(x^*) \geq 0\right\}$$

Condition suffisante sur un ouvert

Théorème 4

Une condition suffisante pour que x^* soit un optimum local de $\Omega \ni x \mapsto f(x) \in \mathbb{R}$ fonction deux fois différentiable sur Ω ouvert de $\mathbb{R}^n$ est

$$\left\{\nabla f(x^*) = 0, \nabla^2 f(x^*) > 0\right\}$$

Analyse convexe

Définition 8

On dit que l'application $f : E \to \mathbb{R}$ (E convexe de $\mathbb{R}^n$) est convexe si

$$\forall (x,y) \in E \times E \quad \forall \lambda \in [0,1] \quad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

Théorème 5

Soient E convexe de $\mathbb{R}^n$ et $f : E \to \mathbb{R}$ continue. Les deux propositions suivantes sont équivalentes :

1. f convexe
2. $\mathrm{Epi}(f) = \{(x,y) \mid x \in E\,,\ y \geq f(x)\}$ est convexe

Par ailleurs, dans le cas où $E^\circ \neq \emptyset$, ceci est équivalent à

3. $\forall x \in E^\circ \quad \exists\, \alpha_x \in \mathbb{R}^n \quad \forall y \in E \quad f(y) \geq f(x) + \alpha_x^T (y - x)$

Analyse convexe

Définition 9

Soit $f : E \to \mathbb{R}$ convexe. Un vecteur $v \in \mathbb{R}^n$ est appelé sous-gradient de f au point $x_0 \in \mathbb{R}^n$ si

$$\forall x \in E \quad f(x) \geq f(x_0) + v^T(x - x_0) \tag{1}$$

L'ensemble de tous les sous-gradients en x_0 est appelé sous-différentiel de f en x_0 et noté $\partial f(x_0)$

$$\partial f(x_0) = \left\{ v \in \mathbb{R}^n \mid \forall x \in E \quad f(x) \geq f(x_0) + v^T(x - x_0) \right\} \tag{2}$$

Analyse convexe

Théorème 9

Soit f une application différentiable de Ω dans $\mathbb{R}$. Les propositions suivantes sont équivalentes

1. f est convexe
2. $\forall (x,y) \in \Omega^2,\ f(y) \geq f(x) + (\nabla f(x))^T (y-x)$
3. ∇f est monotone : $\forall (x,y) \in \Omega^2 \quad (\nabla f(x) - \nabla f(y))^T (x-y) \geq 0$

Et, si f deux fois différentiable, ces propriétés sont équivalentes à

4. $\forall x \in \Omega,\ \nabla^2 f(x) \geq 0$

Analyse convexe

Définition 10

Soit $E \subset \mathbb{R}^n$ convexe. On dit que $f : E \to \mathbb{R}$ est **fortement convexe** (ou α-convexe) s'il existe $\alpha > 0$ tel que $f - \frac{\alpha}{2} \|.\|^2$ est convexe.

Théorème 11

Soit f une application différentiable de Ω dans $\mathbb{R}$, et $\alpha > 0$. Les propositions suivantes sont équivalentes

- f est α-convexe sur Ω
- $\forall (x, y) \in \Omega^2 \quad f(y) \geq f(x) + (\nabla f(x))^T (y - x) + \frac{\alpha}{2} \|x - y\|^2$
- $\forall (x, y) \in \Omega^2 \quad \left((\nabla f(x))^T - (\nabla f(y))^T\right)(x - y) \geq \alpha \|x - y\|^2$

Et si f est deux fois différentiable, ces propriétés sont équivalentes à

- $\forall x \in \Omega \quad \nabla^2 f(x) \geq \alpha I$

Conditions d'optimalité : cas convexe

Théorème 6

Soient $f : \mathbb{R}^n \to \mathbb{R}$ convexe. Alors x^* est un minimiseur global de f si et seulement si $0 \in \partial f(x^*)$. De plus, tout minimiseur local est global.

NB : si f différentiable :

$$x^* \text{ est un minimiseur global de } f \text{ si et seulement si } \nabla f(x^*) = 0.$$

Théorème 12

Soit $f : \mathbb{R}^n \to \mathbb{R}$ fortement convexe. Alors f admet un unique minimum (global) sur tout fermé convexe de $\mathbb{R}^n$.

Méthodes diff. sans contraintes : gradient

$$x^{k+1} = x^k + l^k p^k$$

Algorithme 1 (Gradient à pas optimal)

À partir de $x^0 \in \mathbb{R}^n$ quelconque, itérer

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

où $l^k \in \operatorname{argmin}_{l \in \mathbb{R}} f(x^k - l \nabla f(x^k))$.

Théorème 13

Si f est α-convexe, différentiable et de gradient ∇f Lipschitzien sur tout borné, alors l'algorithme du gradient à pas optimal converge vers l'unique solution x^* du problème d'optimisation $\min_{x \in \mathbb{R}^n} f(x)$.

Méthodes diff. sans contraintes : gradient

Définition 13

On appelle **condition d'Armijo** (de paramètre c_1) sur les itérations $(x^k, p^k, l^k)_{k\in\mathbb{N}}$ l'inéquation

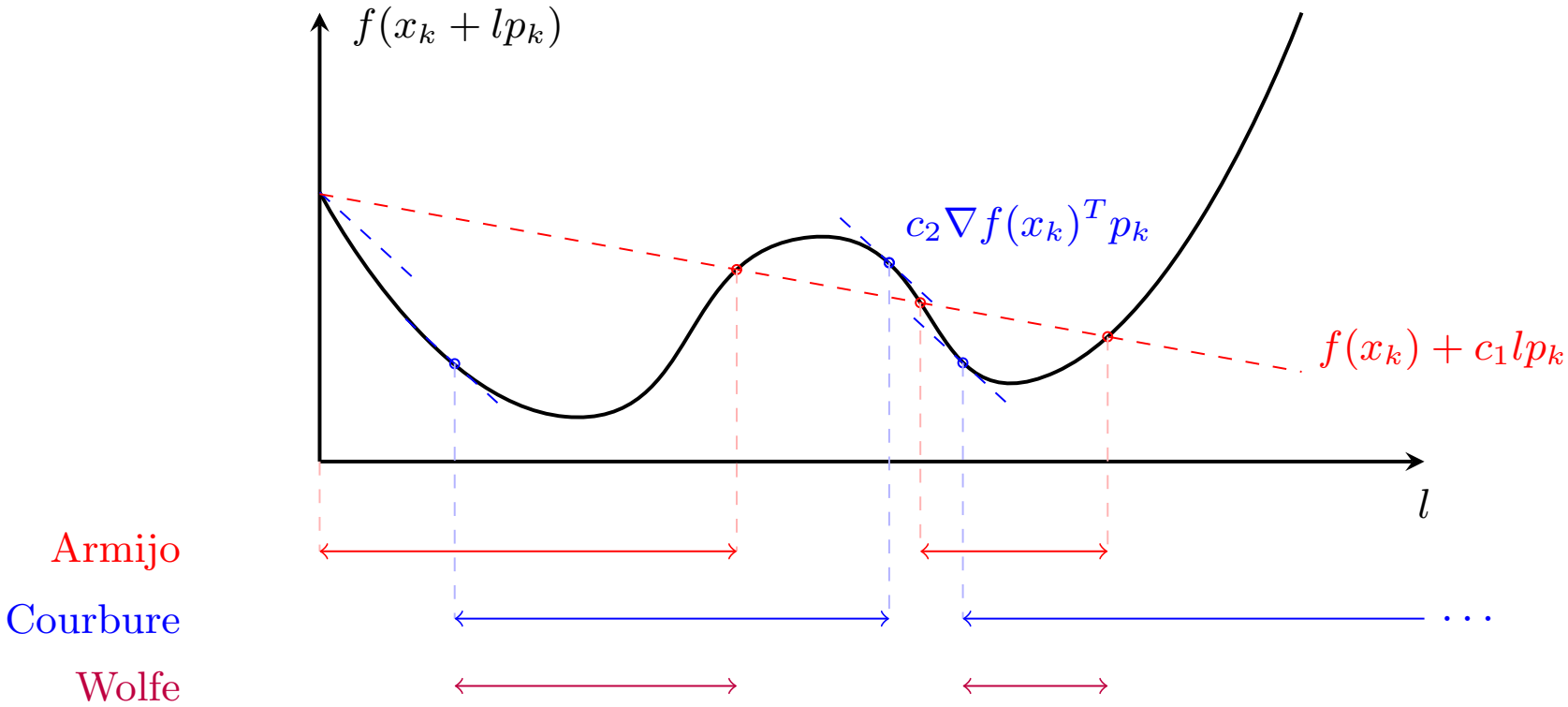
$$f(x^k + l^k p^k) \leq f(x^k) + c_1 l^k \nabla f(x^k)^T p^k \tag{3}$$

On appelle **condition de courbure** (de paramètre c_2) sur les itérations $(x^k, p^k, l^k)_{k\in\mathbb{N}}$ l'inéquation

$$\nabla f(x^k + l^k p^k)^T p^k \geq c_2 \nabla f(x^k)^T p^k \tag{4}$$

On appelle **conditions de Wolfe** ces deux conditions avec $0 < c_1 < c_2 < 1$.

Méthodes diff. sans contraintes : gradient



Théorème 16

Soit f différentiable, bornée inférieurement et telle que ∇f Lispchitzien. Alors on a la convergence

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

Méthodes diff. sans contraintes : gradient

Algorithme 2 (Gradient stochastique)

A partir de $x^0 \in \mathbb{R}^n$, $\theta \in]0,2[$ et une loi de probabilité discrète $(p_i)_{i=1,\dots,n}$ ($p_i > 0$ et $\sum_{i=1}^n p_i = 1$), itérer

- choisir avec une probabilité p_i l'indice i ($i \in \{1,\dots,n\}$)
- $$x_j^{k+1} = \begin{cases} x_j^k - \dfrac{\theta}{L_i}\dfrac{\partial f}{\partial x_i}(x) \text{ si } j = i \\ x_j^k \text{ sinon} \end{cases}$$

où $\left|\frac{\partial^2 f}{\partial x_i^2}(x)\right| \leq L_i$.

Méthodes diff. sans contraintes : second ordre

Algorithme 3 (Newton)

À partir de $x^0 \in \mathbb{R}^n$ quelconque, itérer

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

Théorème 18

Soit $\mathbb{R}^n \ni x \mapsto f(x) \in \mathbb{R}$ deux fois différentiable, possédant un unique minimum global x^* tel que $\nabla^2 f(x^*)$ est définie positive (on note $\lambda > 0$ sa plus petite valeur propre) et tel que $\mathbb{R}^n \ni x \mapsto \nabla^2 f(x) \in \mathcal{M}_n(\mathbb{R})$ est localement Lipschitz au voisinage de x^* (on note C sa constante Lipschitz). L'algorithme de Newton **converge quadratiquement** vers x^* si on l'initialise en un point x^0 tel que $\left\|x^0 - x^*\right\| \leq \frac{2\lambda}{3C}$.

Méthodes diff. sans contraintes : second ordre

Algorithme 4 (Algorithme de BFGS)

À partir de $x^0 \in \mathbb{R}^n$ quelconque et de $R^0 = I(n)$ (d'autres choix de matrice définie positive sont possibles), itérer

$$
\begin{aligned}
p^k &= -R^k \nabla f(x^k) \\
x^{k+1} &= x^k + l^k p^k, \quad l^k \text{ satisfaisant les conditions de Wolfe} \\
s^k &= x^{k+1} - x^k \\
y^k &= \nabla f(x^{k+1}) - \nabla f(x^k) \\
R^{k+1} &= \left(I - \gamma^k s^k (y^k)^T\right) R^k \left(I - \gamma^k y^k (s^k)^T\right) + \gamma^k s^k (s^k)^T
\end{aligned}
$$

CV **superlinéaire** sous certaines hypothèses de Lipschitziannité.

Méthodes diff. sans contraintes : gradient conjugué

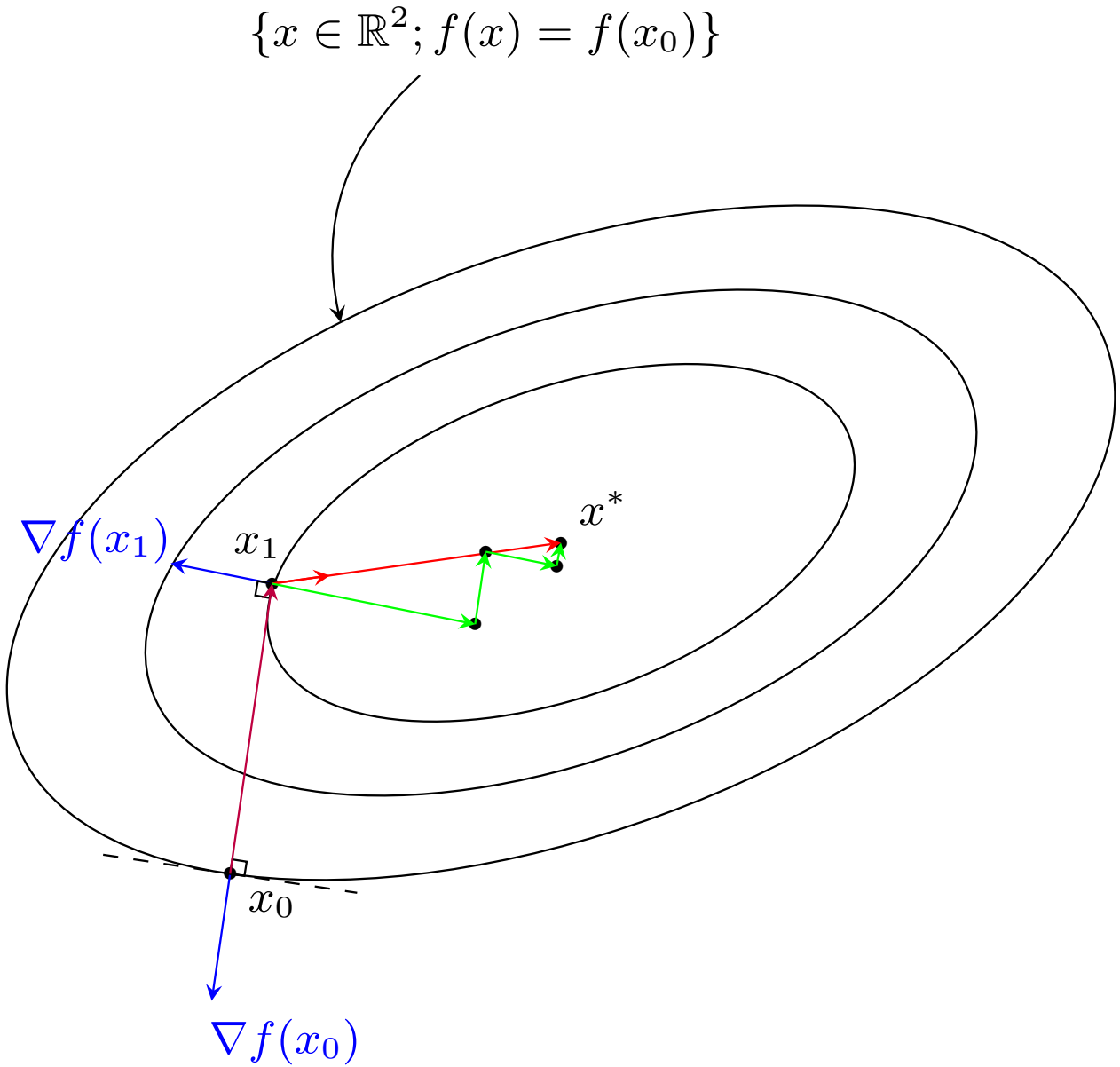
Méthode pour les fonctions quadratiques **de grande taille**.
Directions A-conjuguées : $p_i^T A p_j = 0$

Algorithme 5 (Algorithme du gradient conjugué)

À partir de $x^0 \in \mathbb{R}^n$ quelconque calculer $r^0 = Ax^0 - b$ et $p^0 = -r^0$. Itérer

$$
\begin{aligned}
l^k &= \frac{(r^k)^T r^k}{(p^k)^T A p^k} \\
x^{k+1} &= x^k + l^k p^k \\
r^{k+1} &= r^k + l^k A p^k \\
\beta^{k+1} &= \frac{\left\| r^{k+1} \right\|^2}{\left\| r^k \right\|^2} \\
p^{k+1} &= -r^{k+1} + \beta^{k+1} p^k
\end{aligned}
$$

Méthodes diff. sans contraintes : gradient conjugué



Méthodes diff. sans contraintes : gradient conjugué

Théorème 21

Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive, $K(A) = \frac{\lambda_1}{\lambda_n}$ le rapport entre la plus petite et la plus grande des valeurs propres de A (aussi appelé nombre de conditionnement). Les itérations de l'algorithme 5 du gradient conjugué appliqué à $\mathbb{R}^n \ni x \mapsto \phi(x) = \frac{1}{2}x^T Ax - b^T x \in \mathbb{R}$ avec $b \in \mathbb{R}^n$, **convergent en** n **étapes** et vérifient l'inégalité

$$\left\|x^k - x^*\right\|_A \leq 2\left(\frac{\sqrt{K(A)}-1}{\sqrt{K(A)}+1}\right)^k \left\|x^0 - x^*\right\|_A$$

Extension aux fonctions non-linéaires : algorithme de Fletcher-Reeves et Polak-Ribière. **Convergence superlinéaire**.

Méthodes diff. sans contraintes

Méthode	*Convergence*	*Avantages/Inconvénients*
Gradient	linéaire	
pas optimal		sous-pb à résoudre
Wolfe		
stochastique		plus lent mais utile pour gradt numérique et gde taille
Newton	quadratique	pb de complexité
Quasi-newton	superlinéaire	requiert stockage hessien
Grad. conjugué	superlinéaire et exacte	pour pb de grande taille

Optimisation sous contraintes : égalités

Contraintes égalités

$$\min_{c(x,u)=0} f(x,u)$$

Lagrangien :

$$\mathcal{L}(x,u,\lambda) = f(x,u) + \lambda^T c(x,u) \in \mathbb{R}$$

Théorème 23

Il existe λ^* tel que $(x^*,u^*,\lambda^*) \in \mathbb{R}^{2n+m}$ est un point stationnaire de $\mathcal{L}$ ssi (x^*,u^*) est un point stationnaire de f sous la contrainte c.

Optimisation sous contraintes : égalités

Multiplicateurs de Lagrange λ :

Théorème 24

Si $(x^*, u^*, \lambda^*) \in \mathbb{R}^{2n+m}$ est un point stationnaire de $\mathcal{L}$, alors

$$(\nabla f)^T (x^*, u^*) = -(\lambda^*)^T (\nabla c)^T (x^*, u^*)$$

Théorème 25 (coût marginal)

Si $(x^*, u^*, \lambda^*) \in \mathbb{R}^{2n+m}$ est un point stationnaire de $\mathcal{L}$, alors

$$\frac{\partial f^*}{\partial c} = (\lambda^*)^T$$

Optimisation sous contraintes : inégalités

$$min_{c(x)\leq 0} f(x) \tag{5}$$

Théorème 26 (Conditions KKT)

Considérons un point $x^* \in \mathbb{R}^n$. Notons la famille des indices des **contraintes actives en** x^* par $I = \{i \in \{1,..,m \text{ tel que } c_i(x^*) = 0\}\}$. Supposons que la famille $(\nabla c_i(x^*))_{i\in I}$ est une famille libre (on dit que **les contraintes sont qualifiées**).
Alors, si x^* est une solution du problème (5), $\nabla f(x^*)$ appartient au cône convexe engendré par $(-\nabla c_i(x^*))_{i\in I}$.

$$\exists \lambda_i \geq 0, i = 1,...,m \text{ tels que } \nabla f(x^*) = -\sum_{i=1}^{m} \lambda_i \nabla c_i(x^*)$$
$$\text{et } \lambda_i c_i(x^*) = 0\,, i = 1,..,n$$

Optimisation sous contraintes : inégalités

Théorème 26

Soit le problème d'optimisation $\min_{c(x)\leq 0} f(x)$ où les fonctions f **et** c **sont différentiables et convexes**. On suppose qu'**il existe** $x \in \mathbb{R}^n$ **tel que** $c_i(x) < 0$ **pour** c_i **non affine**, alors **les conditions KKT sont nécessaires et suffisantes** pour que x^* soit un minimum global.

Base de l'algorithme des contraintes actives pour QP et SQP (non-linéaire).

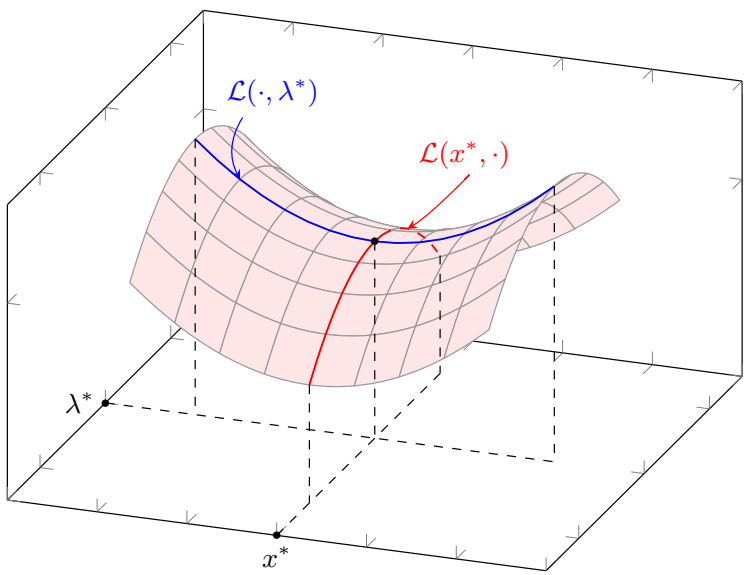
Optimisation sous contraintes : dualité

$$\mathcal{L}(x,\lambda) = f(x) + \lambda^T c(x) \in \mathbb{R}$$

Définition 20

On dit que (x^*,λ^*) est un point selle de $\mathcal{L}$ si x^* est un minimum pour $X \ni x \mapsto \mathcal{L}(x,\lambda^*)$ et λ^* est un maximum pour $L \ni \lambda \mapsto \mathcal{L}(x^*,\lambda)$ ou encore si

$$\sup_{\lambda \in L} \mathcal{L}(x^*,\lambda) = \mathcal{L}(x^*,\lambda^*) = \inf_{x \in X} \mathcal{L}(x,\lambda^*) \tag{6}$$



Optimisation sous contraintes : dualité

Théorème 27 (Théorème du point selle)

Si (x^*, λ^*) est un point selle de $\mathcal{L}$ sur $X \times L$ alors

$$\sup_{\lambda \in L} \inf_{x \in X} \mathcal{L}(x, \lambda) = \mathcal{L}(x^*, \lambda^*) = \inf_{x \in X} \sup_{\lambda \in L} \mathcal{L}(x, \lambda)$$

Prob. dual **Prob. primal**

Théorème 28 (Optimalité du point selle)

Si (x^*, λ^*) est un point selle de $\mathcal{L}$ sur $X \times (\mathbb{R}^+)^m$, alors x^* est solution du problème (5).

Si f, c convexes et x^* solution de (5) tq contraintes qualifiées alors (x^*, λ^*) point selle.

Optimisation sous contraintes : dualité

Algorithme 8 (Algorithme d'Uzawa)

À partir de $x^0 \in \mathbb{R}^n$, $\lambda^0 \in \mathbb{R}^m$, $\alpha \in \mathbb{R}^+$ quelconques, on note $\mathbb{R}^m \ni \lambda \mapsto P(\lambda) \in (\mathbb{R}^+)^m$ la projection sur $(\mathbb{R}^+)^m$, itérer

$$\text{résoudre } \min_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda^k), \text{ on note } x^{k+1} \text{ la solution}$$

$$\lambda^{k+1} = P\left(\lambda^k + \alpha c(x^{k+1})\right)$$

Méthodes diff. sous contraintes

On remplace la recherche de minima sous contrainte par celle de :

- **points stationnaires du Lagrangien** (contraintes égalités)
  conditions KKT : le signe des multiplicateurs asociés aux contraintes actives permet de conclure
  $\rightarrow$ *Algorithme des contraintes actives*

- **points selles du Lagrangien** $X \times (\mathbb{R}^+)^m$
  $\rightarrow$ *Algorithme d'Uzawa*

Ces deux algorithmes s'appuient sur la résolution de sous-problèmes sans contraintes.

Eléments avancés d'analyse convexe

Fenchel

Définition 21

Soit $f : \mathbb{R}^n \to \mathbb{R}$. On appelle transformée de Fenchel de f la fonction f^* définie par

$$f^*(\varphi) = \sup_{x \in \mathbb{R}^n} \left(\varphi^T x - f(x) \right)$$

Pour toute fonction $f : \mathbb{R}^n \to \mathbb{R}$ continue, la biconjuguée $f^{\star\star}$ est définie par

$$f^{**}(x) = \sup_{\varphi \in \mathbb{R}^n} \left(x^T \varphi - f^{\star}(\varphi) \right)$$

Eléments avancés d'analyse convexe

Théorème 32 (Théorème de Moreau-Fenchel)

Soit $f : \mathbb{R}^n \to \mathbb{R}$ continue. $f^\star$ est convexe et la biconjuguée de f satisfait $f^{\star\star} = f$ ssi f est convexe.

Théorème 33 (Régularisation des fonctions fortement convexes)

Soit $f : \mathbb{R}^n \to \mathbb{R}$ fortement convexe. Alors $f^\star$ est de classe $\mathcal{C}^1$ sur $\mathbb{R}^n$ avec

$$\forall \varphi \in \mathbb{R}^n \quad \nabla f^\star(\varphi) = argmax_{x \in \mathbb{R}^n} \left(\varphi^T x - f(x)\right)$$

Eléments avancés d'analyse convexe

Opérateur proximal

Définition 22

Soient $f : \mathbb{R}^n \to \mathbb{R}$ et $\mu > 0$. L'opérateur proximal de f (pour le paramètre μ), noté $\mathrm{Prox}_{\mu f}$, est défini, pour tout $x \in \mathbb{R}^n$, par

$$\mathrm{Prox}_{\mu f}(x) = \mathrm{argmin}_{s \in \mathbb{R}^n} \left(f(s) + \frac{1}{2\mu} \|s - x\|^2 \right) \tag{7}$$

L'enveloppe de Moreau de f (pour le paramètre μ), notée $M_{\mu f}$, est définie, pour tout $x \in \mathbb{R}^n$, par

$$M_{\mu f}(x) = \min_{s \in \mathbb{R}^n} \left(f(s) + \frac{1}{2\mu} \|s - x\|^2 \right)$$

Eléments avancés d'analyse convexe

Théorème 34 (Régularisation de Moreau-Yosida)

Soient f convexe et $\mu > 0$. L'enveloppe de Moreau de f est de classe $\mathcal{C}^1$ sur $\mathbb{R}^n$ avec

$$\forall x \in \mathbb{R}^n \quad \nabla M_{\mu f}(x) = \frac{1}{\mu}\left(x - \mathrm{Prox}_{\mu f}(x)\right)$$

Théorème 35

Soit f convexe et soit $\mu > 0$. Les trois propriétés suivantes sont équivalentes :

1. $x^* \in \mathbb{R}^n$ est un minimiseur (global) de f
2. $x^* = \mathrm{Prox}_{\mu f}(x^*)$
3. x^* est un minimiseur (global) de $M_{\mu f}$

Méthodes non-diff sans contraintes : sous-gradient

Algorithme 13

A partir de $x_0 \in \mathbb{R}^n$ quelconque, itérer

$$x^{k+1} = x^k - lg^k\,, \quad g_k \in \partial f(x^k)$$

Théorème 38

Si f est convexe de minimiseur $x^* \in \mathbb{R}^n$, et que son sous-gradient est borné au moins localement par $G > 0$, alors l'Algorithme 9 de sous-gradient à pas fixe garantit, pour tout $\varepsilon > 0$,

$$\exists k \in \mathbb{N} \quad |f(x^k) - f(x^*)| \leq \frac{lG^2}{2}(1+\varepsilon)$$

Méthodes non-diff sans contraintes : sous-gradient

1. $\lim_{k\to\infty} l^k = 0$ et $\sum_{k=0}^{\infty} l^k = +\infty$
2. $\sum_{k=0}^{\infty} l^k = +\infty$ et $\sum_{k=0}^{\infty} (l^k)^2 < +\infty$

Algorithme 14

A partir de $x_0 \in \mathbb{R}^n$ quelconque et $\hat{f} = f(x_0)$, itérer

$$x^{k+1} = x^k - l^k g^k\,, \quad g_k \in \partial f(x^k)$$
$$\hat{f} = \min\left\{\hat{f}, f(x^{k+1})\right\}$$

Théorème 39

Si f est convexe, que son sous-gradient est borné au moins localement par $G > 0$ et que la suite (l^k) satisfait la condition C1 ou C2 ci-dessus, alors l'Algorithme 14 converge vers un minimum de f.

Méthodes non-diff sans contraintes : min. proximale

Algorithme 15

A partir de $x^0 \in \mathbb{R}^n$ quelconque, itérer

$$x^{k+1} = \mathrm{Prox}_{l^k f}(x^k)$$

avec (l_k) choisie hors ligne, satisfaisant par exemple les propriétés C1 ou C2.

On lui préfère la variante suivante.

Méthodes non-diff sans contraintes : gradient proximal

$$\min_{x\in\mathbb{R}^n}[f(x)=g(x)+h(x)], \quad \text{avec } g \text{ différentiable}$$

Algorithme 16

A partir de $x^0 \in \mathbb{R}^n$ quelconque, itérer

$$x^{k+1} = \text{Prox}_{lh}(x^k - l\nabla g(x_k))$$

Théorème 40

Soit $g : \mathbb{R}^n \to \mathbb{R}$ de gradient L-Lipschitzien. Pour $l \leq 1/L$, l'Algorithme 16 de gradient proximal assure que, pour x^* minimiseur de f,

$$f(x^k) - f(x^*) \leq \frac{1}{2lk}\|x_0 - x^*\|^2$$

Méthodes non-diff sans contraintes : faisceaux

Etant donné un faisceau d'informations

$$\{(x_i, f(x_i), g_i) \mid g_i \in \partial f(x_i)\,,\ i = 1, \ldots, k\}$$

obtenu après k itérations, on construit une approximation linéaire par morceaux de la fonction f

$$\forall y \in \mathbb{R}^n \quad \varphi_k(y) = \max_{i=1,\ldots,k} \left\{ f(x_i) + g_i^T (y - x_i) \right\} \tag{8}$$

et on résout $\min_{y \in \mathbb{R}^n} \varphi_k(y)$ qui se reformule comme un problème de programmation linéaire (LP).

Méthodes non-diff sans contraintes

Méthode	*Convergence*	*Avantages/Inconvénients*
Sous-gradient		oscillatoire
pas fixe	-	pas de CV
pas variable	sous-linéaire	
Prox/Gradient proximal	sous-linéaire	méthode de descente calcul de Prox
Faisceaux	?	pas de descente stockage faisceaux LP simple

Quelques perspectives rapides...

On a traité de l'optimisation **continue** de **dimension finie**. Certains éléments mathématiques s'étendent

- aux fonctions à valeurs dans $\mathbb{R} \cup \{\pm\infty\}$
- aux espaces de Hilbert

Et il y a d'autre pans :

- **optimisation combinatoire** (variables de décision entières, ordonnancement)
- **optimisation de trajectoires** (Euler-Lagrange, contrôle optimal)
- optimisation robuste ou stochastique, analyse de sensibilité

Et, souvent, la partie la plus compliquée en pratique est de formaliser le problème que l'on veut résoudre, cad à formuler f et c...