M1 MApI3 - UE OPTIMISATION
Support de cours

Luca Amodei

15 décembre 2017

1 Introduction

Qu'est-ce qu'un problème d'optimisation ?
Soient $X \subset \mathbb{R}^n$ et $f : \mathbb{R}^n \longrightarrow \mathbb{R}$.
Un problème d 'optimisation (P_X) est défini par

$$\begin{cases} \min f(x) \\ \text{ sous la contrainte } \ x \in X \end{cases}$$

. L'ensemble X est appelé domaine ou ensemble des contraintes. Un point $x \in X$ est dit admissible.
. La fonction f est appelée fonction coût (ou objectif, ou critère).
. Chercher une solution du problème revient à chercher un point de minimum local ou global de f dans l'ensemble des contraintes X.

On notera le problème (P_X) sous forme synthétique

$$\begin{cases} \min f(x) \\ x \in X \end{cases} \tag{1}$$

Remarque : On pourrait noter $\inf f(x)$ (borne inférieure de l'ensemble $f(X)$) à la place de $\min f(x)$. De fait, c'est bien un point $x \in X$ qui réalise la valeur de la borne inférieure qui est recherché. C'est donc la notation $\min f(x)$ (minimum de l'ensemble $f(X)$) que l'on utilisera.

Définition 1.1. *Un point $x^* \in X$ est un minimum local de f sur X si et seulement si il existe un voisinage V_{x^*} de x^* tel que $\forall x \in V_{x^*} \cap X$, on a $f(x) \geq f(x^*)$.*

Définition 1.2. *Un point $x^* \in X$ est un minimum global de f sur X si et seulement si $\forall x \in X$, on a $f(x) \geq f(x^*)$.*

Les minima sont dits stricts si les inégalités dans les définitions précédentes sont strictes pour $x \neq x^*$. Cela signifie que le minimum x^* est unique, soit dans le voisinage $V_{x^*} \cap X$ (dans le cas local), soit dans l'ensemble X (dans le cas global).

Qu'en est-il pour une problème de recherche de valeur maximum ?
Il suffit de considérer le problème (Q_X)

$$\begin{cases} \min \ -f(x) \\ x \in X \end{cases}$$

Un point x^* solution du problème précédent (Q_X) vérifie $-f(x^*) \leq -f(x)$, $\forall x \in X$ et est donc solution du problème

$$\begin{cases} \max f(x) \\ x \in X \end{cases}$$

Grandes familles des problèmes d'optimisation

. Optimisation numérique : $X \subset \mathbb{R}^n$
. Optimisation discrète : X est un ensemble fin ou dénombrable
. Commande optimale, problèmes inverses : X est un ensemble de fonctions (dimension de X infinie)
. Optimisation stochastique : X est un ensemble aléatoire
. Optimisation multicritère : on veut optimiser plusieurs fonctions objectifs en même temps (notion d'optimum de Pareto)

Ce cours d'optimisation s'intéresse uniquement à l'optimisation numérique.

Exemples de problèmes d'optimisation

1. Problème de gain optimal.

   On suppose qu'un produit P est fabriqué dans une usine à partir de n matières premières. La quantité de produits P fabriqués est donnée par une fonction $f(x_1, \ldots, x_n)$ (dite fonction de production) qui dépend des quantités $x_1 \geq 0, \ldots, x_n \geq 0$ des n matières premières utilisées. On note $q > 0$ le prix de vente unitaire du produit P, et $p_1 > 0, \ldots, p_n > 0$, les prix unitaires correspondant aux n matières premières,

   La fonction de gain g s'écrit alors

   $$g(x_1, \ldots, x_n) = qf(x_1, \ldots, x_n) - p_1x_1 - p_2x_2 - \cdots - p_nx_n.$$

   On cherche donc la solution du problème

   $$\begin{cases} \max g(x) \\ x \in X \end{cases}$$

   où X est l'ensemble $\mathbb{R}^n_+$.

2. Problème du transport optimal.

   On veut transporter des marchandises de n dépôts à m points de vente. On connaît les quantités stockées s_i dans chaque dépôt i, les demandes d_j dans chaque point de vente j et le coût de transport unitaire c_{ij} pour transporter du dépôt i au point de vente j. On suppose que $\sum_{i=1}^n s_i \geq \sum_{j=1}^m d_j$ (l'offre est supérieure à la demande).

On cherche les quantités x_{ij} transportées entre les dépôts i et les points de vente j qui minimisent le coût total de transport

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}x_{ij}$$

sous les contraintes :

$x_{ij} \geq 0, \forall i = 1, \ldots, n, j = 1, \ldots, m,$
$\sum_{i=1}^{n} x_{ij} = d_j, \forall j = 1, \ldots, m,$ (il faut satisfaire toutes les demandes)
$\sum_{j=1}^{m} x_{ij} \leq s_i, \forall i = 1, \ldots, n,$ (il ne faut pas dépasser les quantités stockées).

La fonction coût est ici linéaire et les contraintes (de type égalités et inégalités) sont également linéaires. Il s'agit de ce qu'on appelle un problème de programmation linéaire. Ces problèmes seront étudiés en détail dans le cours "Fondamentaux de la recherche opérationnelle" du Master 2 MApI3.

Algorithmique de l'optimisation

Un algorithme associé au problème (P_X) consiste à générer à partir d'un point initial $x_0 \in \mathbb{R}^n$ (ou X), une suite (x_k) où $x_k \in \mathbb{R}^n$ (ou X) qui converge vers x^* solution locale ou globale du problème (P_X).

On dit que la convergence est locale si elle a lieu pour des points initiaux x_0 dans un voisinage de x^*. Sinon elle est dite globale.

Vitesse de convergence

Soit la suite (x_k) générée par un algorithme telle que $\lim_k x_k = x^*$. Les définitions suivantes caractérisent les vitesses de convergence de la suite (x_k).

Définition 1.3.

— *La convergence est linéaire s'il existe* $\tau \in]0, 1[$ *tel que*

$$\lim_k \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \tau$$

— *La convergence est superlinéaire si*

$$\lim_k \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0$$

— *La convergence est d'ordre* p *s'il existe* $\tau \geq 0$ *tel que*

$$\lim_k \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} = \tau$$

Pour $p = 2$, *on dit que la convergence est quadratique.*

Définition 1.4. *Soit $f : \mathbb{R}^n \to \mathbb{R}$ de classe C^1. On dit qu'un algorithme générant la suite (x_k) est globalement convergent si $\forall x_0 \in \mathbb{R}^n$ on a $\lim_k \nabla f(x_k) = 0$.*

Cette propriété est justifiée par le fait qu'une solution x^* du problème (P_X) sans contraintes ($X = \mathbb{R}^n$) est un point critique c.-à-d. $\nabla f(x^*) = 0$ (voir plus loin théorème 2.4). Cette condition est évidemment plus faible que la convergence de la suite (x_k) vers une solution x^*.

2 Résultats d'existence et d'unicité

Dans ce chapitre on présente les résultats généraux concernant l'existence et l'unicité de la solution du problème de minimisation (P_X)

$$\begin{cases} \min f(x) \\ x \in X \end{cases} \tag{2}$$

On y énonce des conditions nécessaires et suffisantes pour qu'un point $x^* \in X$ soit un minimum local ou global du problème (P_X).

Nous commençons par un résultat classique d'existence.

Théorème 2.1. *(théorème de Weierstrass) Si X est un ensemble non vide compact ($\Leftrightarrow$ fermé et borné) et $f : X \to \mathbb{R}$ est continue, alors (P_X) admet au moins une solution.*

Dans le cas où X est uniquement fermé, il faut ajouter une propriété supplémentaire pour garantir l'existence d'une solution.

Définition 2.1. *Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est dite coercive sur un ensemble X non borné de $\mathbb{R}^n$ si $f(x) \to +\infty$ lorsque $\|x\| \to +\infty$ avec $x \in X$*
(avec des quantificateurs, cela se traduit par $\forall A > 0$, $\exists \rho > 0$, $(x \in X$ et $\|x\| \geq \rho) \Rightarrow f(x) \geq A)$.

Théorème 2.2. *Si X est un ensemble non vide fermé et $f : X \to \mathbb{R}$ est continue et coercive, alors (P_X) admet au moins une solution.*

Preuve : La démonstration se fait en considérant un point $x_0 \in X$ (X est non vide) et l'ensemble $X_0 = \{x \in X \mid f(x) \leq f(x_0)\}$. Il est facile de voir que X_0 est un ensemble non vide et compact (il est fermé car f est continue, et borné du fait que f est coercive). Le problème (P_{X_0}) admet donc une solution $x^* \in X_0$. Celle-ci est aussi une solution du problème (P_X) car $\forall x \in X \backslash X_0$ on a $f(x^*) \leq f(x_0) < f(x)$.

Unicité de l'optimum

L'unicité de la solution repose en général sur des arguments de convexité.

Théorème 2.3. *Soit (P_X) le problème de minimisation avec X convexe et $f : X \to \mathbb{R}$ convexe. Alors,*
- *tout minimum local de (P_X) est un minimum global de (P_X),*
- *si f est strictement convexe, il y a au plus un minimum de (P_X).*

Rappels sur la convexité

Définition 2.2.
- *On dit qu'un ensemble $X \subset \mathbb{R}^n$ est convexe si et seulement si*

$$\forall x, y \in X,\ \forall \alpha \in [0,1],\ \alpha x + (1-\alpha)y \in X.$$

- *Soit X un ensemble convexe. La fonction $f : X \to \mathbb{R}$ est convexe si et seulement si*

$$\forall x, y \in X,\ \forall \alpha \in [0,1],\ f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y).$$

- *On dit que f est strictement convexe si l'inégalité précédente est stricte pour $x \neq y$ et $\alpha \in]0,1[$.*

2.1 Conditions nécessaires et suffisantes d'optimalité dans le cas sans contraintes

On rappelle les formules de Taylor à l'ordre un et deux qui permettent d'établir les conditions nécessaires et suffisantes d'optimalité.

On suppose que la fonction $f : \mathbb{R}^n \to \mathbb{R}$ est de classe C^1. La formule de Taylor de f au point $x \in \mathbb{R}^n$ à l'ordre un s'écrit :

$$f(x+h) = f(x) + \nabla f(x)^T h + o(\|h\|),$$

où l'égalité est vérifiée pour tout $h \in \mathbb{R}^n$ dans un voisinage de 0. Le reste $o(\|h\|)$ est défini par la fonction $o(r)$ (petit o de r) qui vérifie $\lim_{r\to 0} \frac{o(r)}{r} = 0$. Le reste $o(\|h\|)$ peut donc s'écrire $o(\|h\|) = \|h\|\, \epsilon(\|h\|)$ où la fonction $\epsilon(r)$ vérifie $\lim_{r\to 0} \epsilon(r) = 0$.

On note $\nabla f(x)$ le gradient de f au point x. Il s'agit du vecteur colonne ayant pour coordonnées les dérivées partielles $\frac{\partial f}{\partial x_i}(x)$, $i = 1, \ldots, n$. C'est donc la transposée de la matrice jacobienne $Jf(x)$: $\nabla f(x) = Jf(x)^T$. Le produit vecteur

ligne $\times$ vecteur colonne $\nabla f(x)^T h$ est donc le produit scalaire $\langle \nabla f(x), h \rangle$ (les vecteurs $\nabla f(x)$ et h sont des vecteurs colonnes).

Si l'on suppose que la fonction f est de classe C^2 on obtient la formule de Taylor à l'ordre deux :

$$f(x+h) = f(x) + \nabla f(x)^T h + \frac{1}{2} h^T \nabla^2 f(x) h + o(\|h\|^2),$$

où l'égalité est vérifiée pour tout $h \in \mathbb{R}^n$ dans un voisinage de 0. On note ici $\nabla^2 f(x)$ la matrice hessienne ($n \times n$, symétrique) ayant pour coefficients ij (ligne i, colonne j, $i = 1, \ldots, n$, $j = 1, \ldots, n$) les dérivées secondes $\frac{\partial^2 f}{\partial x_i \, \partial x_j}(x)$ de f au point x.

Dans le cas du problème sans contraintes (c.-à-d. $X = \mathbb{R}^n$) on a les deux conditions nécessaires d'optimalité suivantes.

Condition d'optimalité d'ordre un

Théorème 2.4. *Si f est de classe C^1 et $x^* \in \mathbb{R}^n$ un minimum local du problème $(P_{\mathbb{R}^n})$, alors $\nabla f(x^*) = 0$.*

Un point qui annule le gradient de la fonction f est dit un point critique (ou singulier ou stationnaire) de f. Ce théorème affirme donc que si x^* est un minimum local de f, alors x^* est un point critique.

Il est facile de voir que dans le cas particulier où f est convexe (et différentiable), alors la condition $\nabla f(x^*) = 0$ est aussi suffisante pour que x^* soit un minimum local (et donc global puisque f est convexe). On utilise pour ça l'inégalité $f(y) \geq f(x) + \nabla f(x)^T (y - x)$, $\forall x, y \in \mathbb{R}^n$, qui est satisfaite lorsque f est convexe.

Rappels et compléments sur les fonctions convexes

Proposition 2.1.

- *Si $f : \mathbb{R}^n \to \mathbb{R}$ est différentiable, on a les équivalences suivantes*
  1. *f est convexe,*
  2. *$f(y) \geq f(x) + \nabla f(x)^T (y - x)$, $\forall x, y \in \mathbb{R}^n$,*
  3. *$(\nabla f(y) - \nabla f(x))^T (y - x) \geq 0$, $\forall x, y \in \mathbb{R}^n$.*
- *On a équivalence entre la convexité stricte de f et les inégalités 2. et 3. précédentes strictes, pour $x \neq y$.*

— *Si $f : \mathbb{R}^n \to \mathbb{R}$ est deux fois différentiable, on a les équivalences suivantes*

1. *f est convexe,*
2. *$\forall x \in \mathbb{R}^n$, la matrice hessienne $\nabla^2 f(x)$ est semi-définie positive.*

Remarques :

La proposition précédente reste vraie si la fonction f est définie sur un ouvert convexe $X \subset \mathbb{R}^n$.

Pour la stricte convexité, dans le cas où la fonction est deux fois différentiable, on a uniquement l'implication ($\nabla^2 f(x)$ définie positive $\forall x \in X$) $\Rightarrow$ f strictement convexe. La réciproque n'est pas vraie. Exemple : la fonction $f(x) = x^4$ est strictement convexe mais $f''(0) = 0$.

Conditions d'optimalité d'ordre deux

Théorème 2.5. *Si f est de classe C^2 et $x^* \in \mathbb{R}^n$ un minimum local du problème $(P_{\mathbb{R}^n})$, alors x^* est un point critique et de plus $\nabla^2 f(x^*)$ est semi-définie positive.*

Rappel : On dit que la matrice hessienne $\nabla^2 f(x^*)$ est semi-définie positive si et seulement si

$$h^T \nabla^2 f(x^*) h \geq 0, \, \forall h \in \mathbb{R}^n.$$

On montre que cette propriété équivaut à dire que toutes les valeurs propres de la matrice symétrique $\nabla^2 f(x^*)$ sont des nombres réels positifs ou nuls.

Remarque : La notion de semi-définie positivé d'une matrice n'est considérée que si la matrice est au préalable symétrique.

La condition énoncée dans le théorème précédent devient une condition suffisante si la matrice hessienne $\nabla^2 f(x^*)$ est définie positive.

Théorème 2.6. *Si f est de classe C^2 et x^* un point critique tel que $\nabla^2 f(x^*)$ est définie positive, alors x^* est un minimum local du problème $(P_{\mathbb{R}^n})$.*

Rappel : On dit que la matrice hessienne $\nabla^2 f(x^*)$ est définie positive si et seulement si

$$h^T \nabla^2 f(x^*) h > 0, \, \forall h \in \mathbb{R}^n, \, h \neq 0.$$

On montre que cette propriété équivaut à dire que toutes les valeurs propres de la matrice symétrique $\nabla^2 f(x^*)$ sont des nombres réels strictement positifs.

Cas particulier des fonctions elliptiques

Définition 2.3. *(fonction elliptique)*

On dit que la fonction de classe C^1 $f : \mathbb{R}^n \to \mathbb{R}$ *est elliptique s'il existe* $\alpha > 0$ *(dite constante d'ellipticité) tel que*

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \alpha \|x - y\|^2, \, \forall x, y \in \mathbb{R}^n.$$

Dans le cas où la fonction f est de classe C^2, la propriété d'ellipticité est équivalente à : il existe $\alpha > 0$ tel que

$$y^T \nabla^2 f(x) y \geq \alpha \|y\|^2, \, \forall x, y \in \mathbb{R}^n.$$

Proposition 2.2. *Une fonction elliptique* f *est strictement convexe et coercive.*

Pour les fonctions elliptiques, on a le résultat fondamental suivant.

Proposition 2.3. *Si* $X \subset \mathbb{R}^n$ *est un ensemble convexe fermé et* f *une fonction elliptique, alors le problème* (P_X) *admet une solution unique.*

Un exemple classique de fonction elliptique est la fonction quadratique

$$f(x) = \frac{1}{2} x^T A x - b^T x,$$

définie par une matrice $A \in \mathbb{R}^{n \times n}$ définie positive et un vecteur $b \in \mathbb{R}^n$. On a en effet $\nabla f(x) - \nabla f(y) = A(x - y)$ et

$$(x - y)^T A (x - y) \geq \lambda_1 \|x - y\|^2, \, \forall x, y \in \mathbb{R}^n,$$

où $\lambda_1 > 0$ est la plus petite valeur propre de A.

2.2 Conditions nécessaires et suffisantes d'optimalité dans le cas général

Dans ce paragraphe on énonce des résultats généraux dans le cas où l'ensemble des contraintes X est quelconque.

Pour caractériser un point optimal $x^* \in X$ du problème (P_X), on introduit la notion de direction admissible et plus généralement de direction tangente en x^*.

Définition 2.4. $d \in \mathbb{R}^n$ *est une direction admissible en* $x \in X$ *s'il existe* $\eta > 0$ *tel que* $x + sd \in X$, $\forall s \in [0, \eta]$.

On dit aussi que la direction d est rentrante dans X en x. On note $T_x^a(X)$ l'ensemble des directions admissibles en x (relativement à X). On voit facilement que l'ensemble $T_x^a(X)$ est un cône (c.-à-d. si $d \in T_x^a(X)$, alors $\alpha d \in T_x^a(X)$ pour tout $\alpha > 0$).

La notion de direction tangente généralise la notion de direction admissible en considérant la limite de directions admissibles.

Définition 2.5. *$d \in \mathbb{R}^n$ est une direction tangente en $x \in X$ s'il existe une suite (d_k), $d_k \in \mathbb{R}^n$, vérifiant $\lim_k d_k = d$, et une suite $(\eta_k), \eta_k > 0$, vérifiant $\lim_k \eta_k = 0$, telles que $x + \eta_k d_k \in X$, $\forall k$.*

On voit facilement que cette définition est équivalente à : il existe une suite (x_k), $x_k \in X$, vérifiant $\lim_k x_k = x$, et une suite $(\eta_k), \eta_k > 0$, vérifiant $\lim_k \eta_k = 0$, telles que $\lim_k \frac{x_k - x}{\eta_k} = d$.

On note $T_x(X)$ l'ensemble des directions tangentes en x relativement X.

Proposition 2.4. *$T_x(X)$ vérifie les propriétés suivantes :*

1. *$T_x(X)$ est un cône fermé et $T_x^a(X) \subset T_x(X)$,*
2. *$T_x(X) \subset \overline{\mathbb{R}_+(X - x)}$,*
3. *Si $x \notin \overline{X}$, alors $T_x(X) = \emptyset$,*
4. *Si $x \in \overset{\circ}{X}$, alors $T_x(X) = \mathbb{R}^n$.*

La propriété 4. montre que si X est un ouvert, alors $T_x(X) = \mathbb{R}^n$ pour tout $x \in X$.

Enonçons le résultat général donnant les conditions nécessaires d'optimalité du premier ordre pour le problème (P_X).

Théorème 2.7. *(condition nécessaire de Peano-Kantorovitch) Si f est différentiable et x^* est un minimum local du problème (P_X), alors*

$$\nabla f(x^*)^T d \geq 0, \, \forall d \in T_{x^*}(X).$$

Remarque importante : Il faut bien comprendre la différence fondamentale entre la condition nécessaire d'optimalité dans le cas où $X = \mathbb{R}^n$ (un minimum local x^* est un point critique - théorème 2.4) et le cas général donné par le théorème 2.7. La condition donnée par ce théorème n'implique pas en général qu'un minimum local x^* est un point critique. À noter toutefois le cas particulier où X est un ouvert, où on a alors $T_{x^*}(X) = \mathbb{R}^n$, et donc par le théorème 2.7, on obtient $\nabla f(x^*) = 0$ (x^* est bien dans ce cas un point critique).

Cas particulier où X est un ensemble convexe

Proposition 2.5. *Si X est un ensemble convexe, alors l'inclusion 2. de la proposition 2.4 est une égalité :*

$$T_x(X) = \overline{\mathbb{R}_+(X - x)}.$$

Preuve : Il s'agit de montrer l'inclusion $\overline{\mathbb{R}_+(X-x)} \subset T_x(X)$. Pour cela, il suffit de montrer l'inclusion $\mathbb{R}_+(X - x) \subset T_x^a(X)$. Soient $y \in X$ et $\beta > 0$. Pour tout $\alpha \in [0, \frac{1}{\beta}]$, on a $x + \alpha\beta(y - x) = (1 - \alpha\beta)x + \alpha\beta y$ et $\alpha\beta \in [0, 1]$. On a donc une combinaison convexe de $x \in X$ et de $y \in X$. On en déduit $x + \alpha\beta(y - x) \in X$ ce qui montre que $\beta(y - x) \in T_x^a(X)$ par définition de $T_x^a(X)$. Cette égalité montre alors que $T_x(X)$ est aussi un ensemble convexe.

L'égalité $T_{x^*}(X) = \overline{\mathbb{R}_+(X - x^*)}$ nous permet de reformuler la condition nécessaire d'optimalité du premier ordre. Cette condition devient aussi suffisante si la fonction f est convexe.

Théorème 2.8. *Si f est différentiable, X est convexe et x^* est un minimum local du problème (P_X), alors*

$$\nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in X.$$

Si de plus la fonction f est convexe, alors la condition précédente est suffisante pour que $x^* \in X$ soit un minimum local (et donc global puisque f est convexe) du problème (P_X).

Projection sur un ensemble convexe fermé

La projection d'un point $\tilde{x} \in \mathbb{R}^n$ sur un ensemble convexe fermé X est définie comme solution du problème

$$\begin{cases} \min \frac{1}{2}\|x - \tilde{x}\|^2 \\ x \in X \end{cases} \tag{3}$$

Notons d'abord que ce problème admet bien une solution unique. En effet, la fonction $f(x) = \frac{1}{2}\|x - \tilde{x}\|^2$ est coercive sur X. Il existe donc au moins une solution de ce problème. La solution est unique du fait que la matrice hessienne $\nabla^2 f(x)$ est définie positive : $\nabla^2 f(x) = I_n$. Le théorème précédent permet donc de caractériser la solution unique x^* (appelée projection de $\tilde{x}$ sur X). Sachant que $\nabla f(x^*) = x^* - \tilde{x}$, x^* est donc caractérisé par

$$(x^* - \tilde{x})^T(x - x^*) \geq 0, \forall x \in X.$$

On note $x^* = P_X(\tilde{x})$.

Dans le cas particulier où X est un sous-espace vectoriel, on vérifie que $(P_X(\tilde{x}) - \tilde{x})$ est orthogonal au sous-espace vectoriel X. On montre que l'application P_X est linéaire. Cette propriété est fausse dans le cas général.

Dans le cas général, on a cependant l'inégalité

$$\|P_X(\tilde{x}) - P_X(\tilde{y})\| \leq \|\tilde{x} - \tilde{y}\|, \ \forall \tilde{x}, \tilde{y} \in \mathbb{R}^n$$

(l'application P_X est contractante).

3 Conditions nécessaires et suffisantes dans le cas de contraintes d'égalité

Dans les cas usuels, l'ensemble des contraintes X est défini à partir d'égalités ou d'inégalités de fonctions. Considérons les fonctions $h : \mathbb{R}^n \to \mathbb{R}^p$, et $g : \mathbb{R}^n \to \mathbb{R}^q$, données par $h(x) = (h_1(x), \ldots, h_p(x))$ et $g(x) = (g_1(x), \ldots, g_q(x))$, $\forall x \in \mathbb{R}^n$. On suppose ces fonctions différentiables. On considère des ensembles de contraintes X définis par

$$X = \{x \in \mathbb{R}^n \,|\, h_i(x) = 0, \, i = 1, \ldots, p, \ g_j(x) \leq 0, \, j = 1, \ldots, q\}.$$

La fontion h définit ce qu'on appelle des contraintes de type égalité et la fonction g des contraintes de type inégalité.

Dans ce paragraphe, nous considérons des sous-ensembles de contraintes X définis uniquement par des contraintes de type égalité

$$X = \{x \in \mathbb{R}^n \,|\, h_i(x) = 0, \, i = 1, \ldots, p\}. \tag{4}$$

Nous allons voir que dans ce cas le cône tangent $T_x(X)$ est défini très simplement à partir de la matrice jacobienne $Jh(x)$.

Il est facile de montrer l'inclusion $T_x(X) \subset \mathrm{Ker}(Jh(x))$. L'inclusion réciproque nécessite des hypothèses supplémentaires.

Définition 3.1. *On dit que $x \in X$ est régulier si les vecteurs $\nabla h_i(x)$, $i = 1, \ldots, p$, sont linéairement indépendants.*

Définition 3.2. *On dit que la contrainte $h(x) = 0$ est qualifiée en $x \in X$ si $T_x(X) = \mathrm{Ker}(Jh(x))$.*

On a le résultat suivant.

Proposition 3.1. *Si $x \in X$ est régulier, alors la contrainte $h(x) = 0$ est qualifiée.*

On voit en particulier que dans ce cas le cône tangent $T_x(X)$ est un sous-espace vectoriel de $\mathbb{R}^n$.

Conditions nécessaires du premier ordre

Le théorème du multiplicateur de Lagrange donne les conditions nécessaires d'optimalité du premier ordre pour un minimum local $x^* \in X$ du problème (P_X) avec X défini par des contraintes d'égalité (4).

Théorème 3.1. *(théorème du multiplicateur de Lagrange)*
Soient $f : \mathbb{R}^n \to \mathbb{R}$ et $h : \mathbb{R}^n \to \mathbb{R}^p$ différentiables. On suppose que la contrainte $h(x) = 0$ est qualifiée en $x^* \in X$. Si x^* est un minimum local du problème (P_X), alors il existe $\lambda_1^*, \ldots, \lambda_p^* \in \mathbb{R}$, tels que

$$\nabla f(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla h_i(x^*) = 0. \tag{5}$$

Le vecteur $\lambda^* = (\lambda_1^*, \ldots, \lambda_p^*)$ est appelé multiplicateur de Lagrange. Si le point x^* est régulier, il est déterminé de manière unique à partir de x^*.

Preuve : Les conditions d'optimalité donnent $\nabla f(x^*)^T v \geq 0$, $\forall v \in T_{x^*}(X)$. Or $T_{x^*}(X) = \mathrm{Ker}(Jh(x^*))$, et donc $T_{x^*}(X)$ est un sous-espace vectoriel de $\mathbb{R}^n$. On en déduit que $\nabla f(x^*)^T v = 0$, $\forall v \in T_{x^*}(X)$. Le vecteur $\nabla f(x^*)$ est donc orthogonal à $\mathrm{Ker}(Jh(x^*))$. Sachant que $\mathrm{Ker}(Jh(x^*))^\perp = \mathrm{Im}(Jh(x^*)^T)$, on déduit donc l'existence de $\lambda^* \in \mathbb{R}^p$ tel que $\nabla f(x^*) = -Jh(x^*)^T \lambda^*$. On obtient l'égalité (5) en observant que les colonnes de la matrice $Jh(x^*)^T$ sont les vecteurs $\nabla h_i(x^*)$.

Fonction de Lagrange associée au problème (P_X)

Il est commode d'introduire la fonction de Lagrange associée au problème (P_X). On définit la fonction $L : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ appelée fonction de Lagrange

$$L(x, \lambda) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x).$$

Les conditions nécessaires du premier ordre données par le théorème de Lagrange s'écrivent

$$\nabla_x L(x^*, \lambda^*) = 0,$$

où ∇_x indique le gradient par rapport à la variable x. De fait, il faut aussi spécifier que $x^* \in X$, c.-à-d. x^* vérifie les contraintes $h_i(x^*) = 0$, $i = 1, \ldots, p$.

Ces conditions sont données simplement par l'égalité $\nabla_\lambda L(x^*, \lambda^*) = 0$, où ∇_λ indique le gradient par rapport à la variable λ.

Les conditions nécessaires d'optimalité s'écrivent donc synthétiquement grâce à la fonction de Lagrange

$$\nabla L(x^*, \lambda^*) = 0,$$

où le gradient est pris par rapport à la variable x et la variable λ. On remarque que le système que l'on obtient est un système de $n+p$ variables (x et λ) et $n+p$ équations (n équations pour $\nabla_x L(x^*, \lambda^*) = 0$, p équations pour $\nabla_\lambda L(x^*, \lambda^*) = 0$).

Sensibilité de la valeur optimale de f par rapport aux contraintes

Le multiplicateur de Lagrange permet en outre d'évaluer la sensibilité de la valeur optimale de f par rapport à une variation du niveau de la contrainte.

Considérons des problèmes « perturbés »(P_{X_c}) définis par les contraintes $h(x) = c$, $c \in \mathbb{R}^p$. Supposons que pour chaque c proche de zéro il existe un minimum local unique du problème (P_{X_c}). Pour chaque c on note $x(c)$ cette solution. Ainsi $x^* = x(0)$ correspond à un minimum local pour $c = 0$.

Considérons la valeur de la fonction f à l'optimum, autrement dit la fonction $c \mapsto f(x(c))$. La proposition suivante énonce une propriété fondamentale du multiplicateur de Lagrange λ^* associé au minimum local x^* relativement au gradient de la fonction $f(x(c))$ au point $c = 0$.

Proposition 3.2. *Supposons les fonctions $f : \mathbb{R}^n \to \mathbb{R}$ et $h : \mathbb{R}^n \to \mathbb{R}^p$ de classe C^2. On considère la famille de problèmes (P_{X_c})*

$$\begin{cases} \min f(x) \\ \quad \textit{sous la contrainte} \;\; h(x) = c \end{cases}$$

Supposons que pour $c = 0$, il existe un minimum local régulier x^* et que, avec le multiplicateur de Lagrange associé λ^*, il vérifie les conditions suffisantes d'optimalité stricte du second ordre (théorème 3.3 plus loin). Alors, pour tout c dans un voisinage de 0, il existe un minimum local $x(c)$ solution du problème (P_{X_c}), continument différentiable et tel que $x^* = x(0)$. De plus, on a

$$\nabla_c f(x(0)) = -\lambda^*,$$

où ∇_c indique le gradient de la fonction $c \mapsto f(x(c))$ définie dans un voisinage de 0.

Conditions du second ordre

Le résultat suivant donne les conditions nécessaires du second ordre.

Théorème 3.2.

Soient $f : \mathbb{R}^n \to \mathbb{R}$ et $h : \mathbb{R}^n \to \mathbb{R}^p$ 2 fois différentiables. On suppose que la contrainte $h(x) = 0$ est qualifiée en $x^* \in X$. Si x^* est un minimum local du problème (P_X), alors il existe $\lambda_1^*, \ldots, \lambda_p^* \in \mathbb{R}$, vérifiant les conditions d'optimalité du premier ordre (5) et telles que

$$v^T \nabla^2_{xx} L(x^*, \lambda^*) v \geq 0, \, \forall v \in \mathrm{Ker}(Jh(x^*)) \tag{6}$$

où $\nabla^2_{xx} L(x^*, \lambda^*) = \nabla^2 f(x^*) + \sum_{i=1}^p \lambda_i^* \nabla^2 h_i(x^*)$.

Ces conditions deviennent aussi suffisantes si l'inégalité (6) est stricte.

Théorème 3.3.

Soient $f : \mathbb{R}^n \to \mathbb{R}$ et $h : \mathbb{R}^n \to \mathbb{R}^p$ 2 fois différentiables. On suppose que la contrainte $h(x) = 0$ est qualifiée en $x^* \in X$. S'il existe $\lambda_1^*, \ldots, \lambda_p^* \in \mathbb{R}$, vérifiant l'égalité (5) (condition nécéssaire d'optimalité du premier ordre) et

$$v^T \nabla^2_{xx} L(x^*, \lambda^*) v > 0, \, \forall v \in \mathrm{Ker}(Jh(x^*)), \, v \neq 0, \tag{7}$$

alors x^* est un minimum local strict du problème (P_X).

4 Optimisation sans contraintes - Algorithmes fondamentaux

Soit $x^* \in \mathbb{R}^n$ un minimum local du problème

$$\begin{cases} \min f(x) \\ x \in \mathbb{R}^n \end{cases} \tag{8}$$

Un algorithme de recherche de la solution x^* définit une suite (x_k) telle que $\lim\limits_k x_k = x^*$.

4.1 Méthodes de descente

On veut assurer la décroissance de la valeur de f : $f(x_{k+1}) < f(x_k)$ pour tout k.

On définit la suite (x_k) par la récurrence

$$x_{k+1} = x_k + \rho_k d_k,$$

où $d_k \in \mathbb{R}^n$ et $\rho_k > 0$. Le scalaire $\rho_k > 0$ s'appelle pas de descente dans la direction d_k (à partir du point x_k).

Définition 4.1. *$d_k \in \mathbb{R}^n$ est une direction de descente en x_k si et seulement si $\nabla f(x_k)^T d_k < 0$.*

Par la formule de Taylor à l'ordre un

$$f(x_k + \rho d_k) = f(x_k) + \rho \nabla f(x_k)^T d_k + o(\rho),$$

on constate que si d_k est une direction de descente alors, pour ρ suffisamment petit, on a $f(x_k + \rho d_k) < f(x_k)$.

Si d_k est une direction de descente, se pose alors la question du choix du pas de descente $\rho_k > 0$ approprié de façon à avoir la décroissance stricte $f(x_k + \rho_k d_k) < f(x_k)$. Il faut également ne pas considérer des valeurs de ρ_k trop petites sans quoi l'algorithme ne progresse plus vers la solution du problème x^*.

L'idée naturelle pour le choix d'une direction de descente consiste à prendre $d_k = -\nabla f(x_k)$. En effet $-\nabla f(x_k)$ est bien une direction de descente puisque $\nabla f(x_k)^T(-\nabla f(x_k)) = -\|\nabla f(x_k)\|^2 < 0$ (on suppose bien entendu que $\nabla f(x_k) \neq 0$ sans quoi l'algorithme se termine puisqu'on a alors que x_k est un point critique).

La direction de descente $-\nabla f(x_k)$ est d'ailleurs optimale parmi toutes les directions de descente d telles que $\|d\| = \|\nabla f(x_k)\|$. On a en effet $-d^T \nabla f(x_k) \leq \|\nabla f(x_k)\|^2$ par l'inégalité de Cauchy-Schwarz et donc $-\|\nabla f(x_k)\|^2 \leq d^T \nabla f(x_k)$. La direction $-\nabla f(x_k)$ est appelée direction de plus forte pente de f au point x_k.

Forme générale d'un algorithme de descente de gradient

En sortie de l'algorithme on a une approximation de x^* solution de $\nabla f(x) = 0$.
x_0 donné, $k = 0$.
Tant que le critère d'arrêt n'est pas vérifié
    Prendre $d_k = -\nabla f(x_k)$
    Recherche linéaire : trouver $\rho_k > 0$ tel que $f(x_k + \rho_k d_k) < f(x_k)$
    $x_{k+1} = x_k + \rho_k \, d_k$
    $k = k + 1$
Fin tant que

Ces algorithmes sont généralement faciles à mettre en œuvre mais les conditions pour assurer la convergence sont souvent fortes et la convergence lente.

Choix du pas de descente

— **Méthode du gradient à pas constant**
On prend $\rho_k = \rho > 0$ constant à chaque pas k. On a ainsi la récurrence

$$x_{k+1} = x_k - \rho \nabla f(x_k).$$

— **Méthode de plus profonde descente** (ou méthode à pas optimal, en anglais "steepest descent ")
On détermine $\rho_k > 0$ solution du problème

$$\min_{\rho>0} f(x_k - \rho \nabla f(x_k)) \tag{9}$$

Cette méthode est plus efficace mais la convergence est parfois très lente (phénomène de « zigzag »des itérés successifs dans une vallée allongée). On a en effet la propriété suivante : si ρ_k est solution du problème (9), alors $\nabla f(x_{k+1})^T \nabla f(x_k) = 0$.
En effet la fonction $\phi(\rho) = f(x_k - \rho \nabla f(x_k))$ admet pour dérivée $\phi'(\rho) = -\nabla f(x_k)^T \nabla f(x_k - \rho \nabla f(x_k))$. Si ρ_k est solution du problème (9), alors $\phi'(\rho_k) = 0$. On obtient donc le résultat puisque $x_{k+1} = x_k - \rho_k \nabla f(x_k)$.
Il va de soi que ces deux choix de pas de descente (pas constant, pas optimal) peuvent être aussi bien utilisés pour toute méthode de descente (pas nécessairement la méthode de gradient). La propriété d'orthogonalité de deux gradients successifs n'est cependant vérifiée que dans le cas de la méthode du gradient à pas optimal (plus profonde descente).

— **Méthodes inexactes**
Déterminons des conditions moins restrictives que celles données par la méthode de la plus profonde descente, garantissant toutefois une décroissance suffisante de f. Considérons le cas général d'une direction de descente d à partir d'un point x.

— Inégalité d'Armijo.
Soit $\epsilon_1 \in]0, 1[$. Le pas de descente ρ vérifie l'inégalité d'Armijo si

$$f(x + \rho d) \leq f(x) + \epsilon_1 \rho \nabla f(x)^T d \tag{10}$$

Cette inégalité assure une décroissance suffisante de la fonction f (faire un dessin).
On peut montrer facilement qu'il existe un intervalle $[0, \eta], \eta > 0$, tel que tout $\rho \in [0, \eta]$ satisfait l'inégalité d'Armijo.
Cette condition cependant n'empêche pas le pas de descente ρ de devenir trop petit ce qui a pour effet d'entraîner une « fausse convergence »de l'algorithme (convergence vers un point non critique). Pour

éviter un pas de descente trop petit, on considère la condition supplémentaire dite de courbure.

— Inégalité de courbure.
Soit $\epsilon_2 \in]\epsilon_1, 1[$. Le pas de descente ρ vérifie l'inégalité de courbure si

$$\nabla f(x + \rho d)^T d \geq \epsilon_2 \nabla f(x)^T d \tag{11}$$

Cette seconde condition n'est pas vérifiée pour $\rho = 0$ et donc n'est pas vérifiée pour $\rho > 0$ suffisamment petit (faire un dessin).

Les deux conditions réunies (condition d'Armijo et de courbure) sont appelées conditions de Wolfe. En pratique les valeurs numériques des constantes ϵ_1, ϵ_2, sont $\epsilon_1 = 10^{-4}$ et $\epsilon_2 = 0.9$.

Algorithme pour la recherche d'un pas de descente ρ vérifiant les deux conditions de Wolfe

$\rho_0 > 0$ donné et $\rho_- = 0, \rho_+ = +\infty$.
Tant que ρ_k ne vérifie pas les deux conditions de Wolfe
    Si ρ_k ne vérifie pas l'inégalité d'Armijo (ρ_k est trop grand)
    $\rho_+ = \rho_k$ et $\rho_{k+1} = \frac{\rho_- + \rho_+}{2}$
    Si ρ_k ne vérifie pas l'inégalité de courbure (ρ_k est trop petit)
    $\rho_- = \rho_k$
        Si $\rho_+ < +\infty$, $\rho_{k+1} = \frac{\rho_- + \rho_+}{2}$
        Si $\rho_+ = +\infty$, $\rho_{k+1} = 2\rho_k$
    $k = k + 1$
Fin tant que
$\rho = \rho_k$

Il existe une autre technique simple pour obtenir une valeur de pas de descente ρ qui vérifie l'inégalité d'Armijo et qui ne soit pas trop petite. C'est la technique de rebroussement (en anglais "backtracking "). L'idée consiste de partir d'une valeur suffisamment grande de ρ et de diminuer la valeur de ρ (par un produit $\alpha\rho$ avec $\alpha \in]0, 1[$, α fixé) jusqu'à ce que ρ vérifie l'inégalité d'Armijo.

Algorithme de rebroussement (« backtracking »)

$\alpha \in]0, 1[$ donné et $\rho_0 > 0$ suffisamment grand (pe. $\rho_0 = 1$).
Tant que $f(x + \rho_k d) > f(x) + \epsilon_1 \rho_k \nabla f(x)^T d$
    $\rho_{k+1} = \alpha \rho_k$
    $k = k + 1$
Fin tant que

$\rho = \rho_k$

On a le résultat de validité suivant pour les conditions de Wolfe.

Proposition 4.1. *Soit d une direction de descente de f en x. Si la fonction de mérite $\phi(\rho) = f(x+\rho d)$ dérivable est bornée inférieurement, alors il existe $\rho > 0$ vérifiant les conditions de Wolfe.*

Le théorème suivant donne un résultat de convergence dans le cas où le pas de descente vérife les conditions de Wolfe.

Théorème 4.1. *(théorème de Zoutendijk)*
Soit f différentiable de gradient Lipschitz et bornée inférieurement. Soit l'algorithme de descente $x_{k+1} = x_k + \rho_k d_k$, où pour tout k, d_k est une direction de descente de f en x_k et $\rho_k > 0$ un pas de descente vérifiant les conditions de Wolfe. Alors la série $\sum_k \cos(\theta_k)^2 \|\nabla f(x_k)\|^2$, où $\cos(\theta_k) = \frac{-\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \|d_k\|}$, converge.

Ce théorème implique que s'il existe $c > 0$ tel que $|\cos(\theta_k)| \geq c$, $\forall k$, alors la série $\sum_k \|\nabla f(x_k)\|^2$ est convergente et donc $\lim_k \nabla f(x_k) = 0$. L'algorithme est alors globalement convergent (voir définition 1.4).

Résultats de convergence pour la méthode du gradient

Lemme 4.1. *Si f est différentiable de gradient Lipschitz, avec $L > 0$ constante de Lipschitz, alors*

$$f(y) - f(x) - \nabla f(x)^T (y - x) \leq \frac{L}{2} \|y - x\|^2, \, \forall x, y \in \mathbb{R}^n.$$

Pour la méthode du gradient, sachant que $d_k = -\nabla f(x_k)$, ce lemme implique que si f est de gradient Lipschitz, avec $L > 0$ constante de Lipschitz, alors

$$f(x_{k+1}) \leq f(x_k) - \rho_k \|\nabla f(x_k)\|^2 + \frac{L}{2} \rho_k^2 \|\nabla f(x_k)\|^2 = f(x_k) + \rho_k (\frac{L}{2} \rho_k - 1) \|\nabla f(x_k)\|^2.$$

L'étude de la fonction $\rho \mapsto \varphi(\rho) = \rho(\frac{L}{2}\rho - 1)$ montre que $f(x_{k+1}) < f(x_k)$ si et seulement si $0 < \rho_k < \frac{2}{L}$. Plus précisemment, on a le résultat suivant.

Proposition 4.2. *Soit f différentiable de gradient Lipschitz, avec $L > 0$ constante de Lipschitz, et bornée inférieurement. Alors l'algorithme du gradient $x_{k+1} = x_k - \rho_k \nabla f(x_k)$, à pas fixe $\rho_k = \rho$, avec $\rho < \frac{2}{L}$, ou à pas optimal, est globalement convergent.*

Le théorème de Zoutendijk montre que, sous les mêmes hypothèses, l'algorithme du gradient avec un pas de descente satisfaisant les conditions de Wolfe est globalement convergent. En effet, on a dans ce cas $\cos(\theta_k) = \frac{-\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \, \|d_k\|} = 1$, $\forall k$.

Dans le cas où la fonction f est de plus elliptique, on a le résultat suivant.

Proposition 4.3. *Soit f différentiable, elliptique, de gradient Lipschitz, avec $L > 0$ constante de Lipschitz, et bornée inférieurement. Alors l'algorithme du gradient $x_{k+1} = x_k - \rho_k \nabla f(x_k)$, avec $\rho_k \in [a, b] \subset]0, \frac{2}{L}[$, $\forall k$, converge vers la solution unique x^* du problème $(P_{\mathbb{R}^n})$.*

4.2 Méthode de Newton

La méthode de Newton est une méthode de descente. La direction de descente est obtenue à partir de l'approximation d'ordre deux de la fonction f au point courant x_k.

À partir du développement de Taylor de f à l'ordre deux au point x_k on approxime $f(x)$ par

$$q(x) = f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k).$$

Le point x_{k+1} est obtenu comme solution du problème

$$\min_{x \in \mathbb{R}^n} q(x)$$

Les conditions d'optimalité donnent

$$\nabla q(x_{k+1}) = \nabla^2 f(x_k)(x_{k+1} - x_k) + \nabla f(x_k) = 0.$$

Si l'on suppose que la matrice $\nabla^2 f(x_k)$ est inversible, on obtient alors la récurrence

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k). \tag{12}$$

Le vecteur $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$ est appelé à jouer le rôle de direction de descente. C'est le cas en particulier si $\nabla^2 f(x_k)$ est définie positive. On a en effet $\nabla f(x_k)^T d_k = -\nabla f(x_k)^T \nabla^2 f(x_k)^{-1} \nabla f(x_k) < 0$ si $\nabla f(x_k) \neq 0$.

On a le résultat suivant de convergence. La convergence quadratique rend la méthode de Newton particulièrement efficace.

Théorème 4.2. *Soit f de classe C^3 et x^* un minimum local du problème $(P_{\mathbb{R}^n})$ tel que $\nabla^2 f(x^*)$ est définie positive. Alors, pour x_0 suffisamment proche de x^*, la suite (x_k) obtenue par l'algorithme de Newton (12) converge quadratiquement vers x^*.*

Lorsque les matrices $\nabla^2 f(x_k)$, $\forall k$, sont définies positives on peut considérer la méthode de Newton classique comme une méthode de descente avec un pas de descente égal à 1. On peut cependant aussi envisager dans ce cas des pas de descente $\rho_k > 0$ distincts de 1 pour améliorer la décroissante de f.

Une des limitations de la méthode de Newton vient en particulier du calcul de la matrice hessienne $\nabla^2 f(x_k)$ qui dans certains cas peut s'avérer rédhibitoire. Les méthodes de quasi-Newton constituent des alternatives intéressantes d'un point de vue pratique puisqu'elles évitent le calcul exact de la matrice hessienne $\nabla^2 f(x_k)$ tout en exhibant de bonnes propriétés de convergence (généralement superlinéaire).

4.3 Méthode de Gauss-Newton

Voir exercice 5. TD n°3.

5 Méthode des directions conjuguées

5.1 Cas quadratique - Méthode du gradient conjugué

Soit $A \in \mathbb{R}^{n\times n}$ une matrice définie positive et $q(x)$ la fonction quadratique définie par

$$q(x) = \frac{1}{2}x^T Ax - b^T x,$$

où $b \in \mathbb{R}^n$. On sait que la solution unique x^* du problème

$$\begin{cases} \min\ q(x) \\ x \in \mathbb{R}^n \end{cases}$$

est donnée par la solution (unique) du système

$$Ax = b,$$

obtenu en posant $\nabla q(x) = 0$.

À partir d'un point $x_0 \in \mathbb{R}^n$, la récurrence définie par la méthode du gradient à pas optimal appliquée au problème de minimisation est donnée par

$$x_{k+1} = x_k - \alpha_k g_k,$$

où $g_k = \nabla q(x_k) = Ax_k - b$ et $\alpha_k = \frac{g_k^T g_k}{g_k^T A g_k}$ (voir TD).

La méthode du gradient conjugué consiste à utiliser non pas $-g_k$ comme direction de descente mais un vecteur d_k combinaison linéaire du gradient $-g_k$

et de la direction de descente précédente d_{k-1}. Plus précisemment on prend d_k de la forme $d_k = -g_k + \beta_{k-1} d_{k-1}$. Le coefficient β_{k-1} est choisi de sorte que $d_k^T A d_{k-1} = 0$ (orthogonalité des vecteurs d_k et d_{k-1} au sens du produit scalaire défini par la matrice définie positive A). On obtient facilement

$$\beta_{k-1} = \frac{g_k^T A d_{k-1}}{d_{k-1}^T A d_{k-1}}. \tag{13}$$

On peut ainsi écrire une version préliminaire de l'**algorithme du gradient conjugué [GC1]** :

À partir de x_0, et $d_0 = -g_0 = -(Ax_0 - b)$, faire
$x_{k+1} = x_k + \alpha_k d_k$ avec $\alpha_k = -\frac{g_k^T d_k}{d_k^T A d_k}$ et $g_k = \nabla q(x_k) = Ax_k - b$
$d_{k+1} = -g_{k+1} + \beta_k d_k$ avec $g_{k+1} = \nabla q(x_{k+1}) = Ax_{k+1} - b$ et $\beta_k = \frac{g_{k+1}^T A d_k}{d_k^T A d_k}$
$k = k + 1$

Le coefficient α_k est le pas de descente optimal dans la direction de descente d_k et le coefficient β_k implique l'A-orthogonalité (orthogonalité au sens du produit scalaire défini par la matrice définie positive A) $d_{k+1}^T A d_k = 0$. On remarque aussi que si $g_k = 0$ alors x_k est la solution du problème et l'algorithme se termine.

On va montrer en fait que la suite des directions d_k générées par l'algorithme précédent sont A-orthogonales deux à deux, c.-à-d. $d_i^T A d_j = 0$, $\forall i \neq j$.

Avant de montrer cette propriété, nous allons donner deux résultats fondamentaux qui découlent de l'A-orthogonalité des vecteurs d_k.

Une première observation triviale est l'indépendance d'une suite de vecteurs A-orthogonaux deux à deux.

Proposition 5.1. *Soit $\{d_0, d_1, \ldots, d_k\}$ un ensemble de vecteurs non nuls, A-orthogonaux deux à deux, c.-à-d. $d_i^T A d_j = 0$, $\forall i \neq j$, et $i, j \leq k$. Alors ces vecteurs sont linéairement indépendants.*

La démonstration est immédiate en considérant une combinaison linéaire de ces vecteurs.

Théorème 5.1. *Soit $\{d_0, d_1, \ldots, d_{n-1}\}$ un ensemble de n vecteurs non nuls, A-orthogonaux deux à deux. Alors, pour tout $x_0 \in \mathbb{R}^n$, la suite (x_k) définie par la récurrence*

$$x_{k+1} = x_k + \alpha_k d_k,$$

où $\alpha_k = -\frac{g_k^T d_k}{d_k^T A d_k}$ et $g_k = \nabla q(x_k) = Ax_k - b$, converge vers x^* solution du système $Ax = b$ après n itérations, c.-à-d. $x_n = x^*$.

Preuve : Les n vecteurs d_j sont indépendants et constituent donc une base de $\mathbb{R}^n$. Il existe donc des coefficients γ_j uniques tels que

$$x^* - x_0 = \gamma_0 d_0 + \cdots + \gamma_{n-1} d_{n-1}.$$

En prenant le produit A-scalaire des deux membres de cette égalité par le vecteur d_k et en utilisant l'A-orthogonalité des vecteurs d_j, on déduit que

$$\gamma_k = \frac{d_k^T A(x^* - x_0)}{d_k^T A d_k^T}.$$

Nous allons montrer que $\gamma_k = \alpha_k$ pour tout $k = 1, \ldots, n-1$. De la relation de récurrence définissant la suite (x_k), on déduit $x_k = x_0 + \alpha_0 d_0 + \cdots + \alpha_{k-1} d_{k-1}$. Utilisant l'$A$-orthogonalité des vecteurs d_j, on déduit que $d_k^T A(x_k - x_0) = 0$. On a donc $d_k^T A(x^* - x_0) = d_k^T A(x^* - x_k + x_k - x_0) = d_k^T A(x^* - x_k)$. Sachant que $g_k = Ax_k - b = A(x_k - x^*)$, on déduit que $\gamma_k = \frac{d_k^T A(x^* - x_k)}{d_k^T A d_k^T} = -\frac{g_k^T d_k}{d_k^T A d_k^T} = \alpha_k$. Le résultat est démontré.

Le second résultat fondamental est le suivant.

Théorème 5.2. *Soit $\{d_0, d_1, \ldots, d_{n-1}\}$ un ensemble de n vecteurs non nuls, A-orthogonaux deux à deux. Alors, pour tout $x_0 \in \mathbb{R}^n$, la suite (x_k) définie par la récurrence*

$$x_k = x_{k-1} + \alpha_{k-1} d_{k-1},$$

où $\alpha_{k-1} = -\frac{g_{k-1}^T d_{k-1}}{d_{k-1}^T A d_{k-1}}$ vérifie la propriété suivante : $\forall k, x_k$ minimise la fonction quadratique $q(x)$ non seulement sur la droite $x_{k-1} + \alpha d_{k-1}$ (par définition du coefficient α_{k-1}) mais aussi sur tout l'espace affine $x_0 + \mathcal{B}_k$, où $\mathcal{B}_k = s.e.v.\{d_0, \ldots, d_{k-1}\}$ est le sous-espace vectoriel généré par les vecteurs $d_0, \ldots, d_{k-1}$.

Preuve : La récurrence qui définit x_k montre que $x_k = x_0 + \alpha_0 d_0 + \cdots + \alpha_{k-1} d_{k-1}$ et donc $x_k \in x_0 + \mathcal{B}_k$. Montrons que $g_k = \nabla q(x_k)$ est orthogonal à tous vecteur de $\mathcal{B}_k$, autrement dit $g_k^T d_j = 0, \forall j = 0, \ldots, k-1$. Cette propriété caractérise en effet le minimum unique de $q(x)$ sur le sous-espace affine $x_0 + \mathcal{B}_k$ (justifier cette affirmation). Montrons la propriété par récurrence sur k. Pour $k = 1$ la propriété est satisfaite puisque α_0 est précisemment la valeur de $\alpha \in \mathbb{R}$ qui minimise la fonction $\alpha \mapsto q(x_0 + \alpha d_0)$. Supposons maintenant que $g_k \perp \mathcal{B}_k$ (hypothèse de récurrence). Montrons que $g_{k+1} \perp \mathcal{B}_{k+1}$. De la relation $x_{k+1} = x_k + \alpha_k d_k$ on déduit $Ax_{k+1} - b = Ax_k - b + \alpha_k A d_k$, c.-à-d. $g_{k+1} = g_k + \alpha_k A d_k$. Par définition de α_k, on a $d_k^T g_{k+1} = 0$. Pour $j < k$, l'égalité précédente implique $d_j^T g_{k+1} = d_j^T g_k + \alpha_k d_j^T A d_k$. Le premier terme du membre de droite est nul par hypothèse de récurrence et le second aussi dû à l'A-orthogonalité des directions d_j. On a donc $d_j^T g_{k+1} = 0$. Le résultat est démontré.

Ces deux théorèmes supposent que les directions A-orthogonales d_j sont données à priori. Nous allons montrer que l'algorithme [GC1] définit de fait des directions d_j A-orthogonales. Ce résultat est obtenu en montrant certaines propriétés des vecteurs g_k et d_k. Nous donnons également d'autres expressions équivalentes des coefficients α_k et β_k plus utilisées dans la pratique.

Théorème 5.3. *Si la suite (x_k) générée par l'algorithme du gradient conjugué [GC1] ne se termine pas en x_k (c.-à-d. si $g_j \neq 0$, $\forall j = 0, \ldots, k$,), alors*

1. $s.e.v.\{g_0, g_1, \ldots, g_k\} = s.e.v.\{g_0, Ag_0, \ldots, A^k g_0\}$
2. $s.e.v.\{d_0, d_1, \ldots, d_k\} = s.e.v.\{g_0, Ag_0, \ldots, A^k g_0\}$
3. $d_k^T A d_j = 0$, $\forall i \leq k-1$
4. $\alpha_k = \frac{g_k^T g_k}{d_k^T A d_k}$
5. $\beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$

Preuve : On va montrer les trois propriétés 1., 2. et 3. en même temps par récurrence sur k. Pour $k = 0$ elles sont de toute évidence satisfaites. Supposons qu'elles soient vraies pour k et montrons qu'elles sont vraies pour $k+1$. On a

$$g_{k+1} = g_k + \alpha_k A d_k$$

(voir preuve du théorème précédent). Par hypothèse de récurrence, on a que g_k et Ad_k appartiennent au $\text{s.e.v.}\{g_0, Ag_0, \ldots, A^{k+1} g_0\}$. Donc $g_{k+1} \in \text{s.e.v.}\{g_0, Ag_0, \ldots, A^{k+1} g_0\}$. De plus $g_{k+1} \notin \text{s.e.v.}\{g_0, Ag_0, \ldots, A^k g_0\} = \text{s.e.v.}\{d_0, d_1, \ldots, d_k\}$ car sinon on aurait $g_{k+1} = 0$ du fait que g_{k+1} est orthogonal à $\text{s.e.v.}\{d_0, d_1, \ldots, d_k\}$ (le théorème précédent et l'hypothèse de récurrence impliquent cette propriété). On a donc

$$\text{s.e.v.}\{g_0, g_1, \ldots, g_{k+1}\} = \text{s.e.v.}\{g_0, Ag_0, \ldots, A^{k+1} g_0\},$$

ce qui montre la propriété 1. pour $k+1$. À partir de l'égalité

$$d_{k+1} = -g_{k+1} + \beta_k d_k,$$

et en utilisant l'hypothèse de récurrence sur la propriété 2. ainsi que la propriété 1. démontrée à l'ordre $k+1$, on déduit que la propriété 2. est vraie à l'ordre $k+1$.

Pour montrer la propriété 3. à l'ordre $k+1$ on utilise l'égalité

$$d_{k+1}^T A d_j = -g_{k+1}^T A d_j + \beta_k d_k^T A d_j,$$

déduite de $d_{k+1} = -g_{k+1} + \beta_k d_k$. Si $j = k$, le membre de droite est nul de par la définition de β_k. Considérons $j < k$. Le second terme du membre de droite est

nul par hypothèse récurrence. Par ailleurs, on a $Ad_j \in \text{s.e.v.}\{d_0, d_1, \ldots, d_{j+1}\}$ et donc $Ad_j \in \text{s.e.v.}\{d_0, d_1, \ldots, d_k\}$. Le premier terme du membre de droite est donc aussi nul puisqu'on a $g_{k+1} \perp \text{s.e.v.}\{d_0, d_1, \ldots, d_k\}$ par le théorème précédent. La propriété 3. est donc vérifiée à l'ordre $k+1$.

Pour montrer 4. on utilise l'égalité

$$-g_k^T d_k = g_k^T g_k - \beta_{k-1} g_k^T d_{k-1}.$$

Le second terme du membre de droite est nul par le théorème précédent.

Enfin, pour montrer la propriété 5., on remarque que $g_{k+1}^T g_k = 0$ car $g_k \in \text{s.e.v.}\{d_0, d_1, \ldots, d_k\}$ et g_{k+1} est orthogonal à $\text{s.e.v.}\{d_0, d_1, \ldots, d_k\}$ par le théorème précédent. De l'égalité

$$Ad_k = \frac{1}{\alpha_k}(g_{k+1} - g_k),$$

on déduit que

$$g_{k+1}^T Ad_k = \frac{1}{\alpha_k} g_{k+1}^T g_{k+1}.$$

En utilisant l'expression de α_k donnée par 4., on déduit le résultat.

Remarques :

- Il est clair que si la propriété 3. est vérifiée jusqu'au rang k, on a alors l'orthogonalité $d_i^T Ad_j = 0$, $\forall i \neq j$, $i, j \leq k$.
- Comme cela a déjà été évoqué, l'arrêt de l'algorithme [GC1] survient lorsque $g_k = \nabla q(x_k) = Ax_k - b = 0$. x_k est la solution (unique) du système $Ax = b$ (ou de manière équivalente la solution du problème de minimisation de la fonction quadratique $q(x)$). Si par contre $g_k \neq 0$, l'égalité $d_k = g_k - \beta_{k-1} d_{k-1}$ et l'orthogonalité des vecteurs g_k et d_{k-1} montre que $d_k \neq 0$. L'algorithme se poursuit à l'étape suivante $k+1$.
- L'égalité $g_k^T d_k = -g_k^T g_k^T$ démontrée pour établir la propriété 4. précédente montre que la direction d_k est bien une direction de descente.

On peut maintenant reformuler l'algorithme [GC1] en utilisant les nouvelles expressions des coefficients α_k, β_k et un test d'arrêt sur la valeur de la norme du gradient g_k.

Algorithme du gradient conjugué [GC2]

Soit $\epsilon > 0$ paramètre du test d'arrêt.

Soit x_0 donné, $d_0 = -g_0 = -(Ax_0 - b)$, et $k = 0$.

`tant que` $\|g_k\| > \epsilon$ `faire`

$$\begin{aligned}
&\alpha_k = \frac{g_k^T g_k}{d_k^T A d_k} \\
&x_{k+1} = x_k + \alpha_k d_k \\
&g_{k+1} = g_k + \alpha_k A d_k \\
&\beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k} \\
&d_{k+1} = -g_{k+1} + \beta_k d_k \\
&k = k + 1
\end{aligned}$$

`fin tant que`

On montre que la vitesse de convergence de la suite (x_k) est donnée par

$$\|x_k - x^*\|_A \leq 2 \left(\frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}} \right)^k \|x_0 - x^*\|_A,$$

où λ_n et λ_1 sont respectivement la plus grande et la plus petite valeur propre de la matrice A. On peut comparer cette inégalité avec celle obtenue pour la méthode du gradient à pas optimal (voir TD) :

$$\|x_k - x^*\|_A \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \right)^k \|x_0 - x^*\|_A.$$

5.2 Le cas non quadratique

On peut étendre l'algorithme du gradient conjugué au cas où la fonction objectif f n'est pas quadratique. En général on ne plus alors assurer la convergence en n étapes au maximum comme dans le cas quadratique.

De fait, l'algorithme du gradient conjugué met en jeux à chaque itération k le gradient de la fonction $g_k = \nabla f(x_k)$ (égal à $g_k = Ax_k - b$ dans le cas quadratique) et la matrice hessienne $\nabla^2 f(x_k)$ (constante et égale à A dans le cas quadratique).

Si l'on reprend l'algorithme [GC1] en remplaçant la matrice A par $\nabla^2 f(x_k)$, on obtient l'algorithme suivant [E1GC]

À partir de x_0, et $d_0 = -g_0 = -\nabla f(x_0)$, faire

$x_{k+1} = x_k + \alpha_k d_k$ avec $\alpha_k = -\frac{g_k^T d_k}{d_k^T \nabla^2 f(x_k) d_k}$ et $g_k = \nabla f(x_k)$

$d_{k+1} = -g_{k+1} + \beta_k d_k$ avec $g_{k+1} = \nabla f(x_{k+1})$ et $\beta_k = \frac{g_{k+1}^T \nabla^2 f(x_k) d_k}{d_k^T \nabla^2 f(x_k) d_k}$

$k = k + 1$

Le calcul de la matrice hessienne $\nabla^2 f(x_k)$ à chaque itération rend cependant cet algorithme peu intéressant dans la pratique. On voit en effet que la complexité de cet algorithme est comparable à celle de l'algorithme de Newton pour lequel la vitesse de convergence est en principe bien meilleure (convergence quadratique).

La matrice hessienne intervient dans le calcul des coefficients α_k et β_k. Dans le cas quadratique on a donné une expression équivalente de β_k : on a montré au Théorème 5.3 que $\beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$. Cette expression permet d'utiliser uniquement les gradients de la fonction en x_k et x_{k+1}. D'autre part, on sait que la valeur de α_k, dans le cas quadratique, est obtenue en prenant la valeur optimale du pas de descente dans la direction de descente d_k : α_k est solution du problème

$$\begin{cases} \min f(x_k + \alpha d_k) \\ \alpha > 0 \end{cases}$$

En utilisant ces deux propriétés, on obtient l'algorithme dit de Fletcher-Reeves (calqué sur l'algorithme du gradient conjugué dans le cas quadratique).

Algorithme de Fletcher-Reeves [E2GC]

Soit $\epsilon > 0$ paramètre du test d'arrêt.

Soit x_0 donné, $d_0 = -g_0 = -\nabla f(x_0)$, et $k = 0$.

tant que $\|g_k\| > \epsilon$ faire

$\alpha_k > 0$ minimum de la fonction $\alpha \mapsto f(x_k + \alpha d_k)$
$x_{k+1} = x_k + \alpha_k d_k$
$g_{k+1} = \nabla f(x_{k+1})$
$\beta_k = \frac{g_{k+1}^T g_{k+1}}{g_k^T g_k}$
$d_{k+1} = -g_{k+1} + \beta_k d_k$
$k = k + 1$

fin tant que

La variante dite de Polak-Ribière consiste à remplacer l'expression de β_k dans l'algorithme de Fletcher-Reeves par $\beta_k = \frac{g_{k+1}^T (g_{k+1} - g_k)}{g_k^T g_k}$, expression identique dans le cas où la fonction f est quadratique (on sait en effet que dans ce cas-là $g_{k+1}^T g_k = 0$, voir démonstation du théorème 5.3).

Dans la pratique, le pas optimal α_k peut être obtenu par une méthode de recherche linéaire approchée (conditions de Wolfe).

6 Méthodes de quasi-Newton

On a vu que la méthode de Newton utilise la direction de descente

$$d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k). \tag{14}$$

Il s'agit en effet d'une direction de descente dans le cas où la matrice hessienne $\nabla^2 f(x_k)$ est définie positive.

Dans la pratique le calcul de la matrice hessienne s'avère souvent complexe et coûteux. En s'inspirant de l'expression (14), les méthodes de quasi-Newton utilisent des directions de descente d_k de la forme

$$d_k = -H_k^{-1} \nabla f(x_k), \tag{15}$$

où la matrice $H_k \in \mathbb{R}^{n \times n}$ est définie positive et représente une approximation de la matrice hessienne $\nabla^2 f(x_k)$. Il existe plusieurs familles de méthodes de quasi-Newton suivant les types d'approximations de la matrice hessienne utilisés.

Comme pour toute méthode de descente, on peut améliorer la décroissance de la fonction en utilisant un pas de descente $\rho_k > 0$ le long de la direction de descente d_k. Pour un pas de descente $\rho_k > 0$ vérifiant les conditions de Wolfe, le théorème général 4.1 (théorème de Zoutendijk) affirme qu'on a convergence globale si $\cos(\theta_k) = \frac{-\nabla f(x_k)^T d_k}{\|\nabla f(x_k)\| \, \|d_k\|}$ est uniformément minoré. On a ici $\cos(\theta_k) = \frac{\nabla f(x_k)^T H_k^{-1} \nabla f(x_k)}{\|\nabla f(x_k)\| \, \|H_k^{-1} \nabla f(x_k)\|}$. Un calcul simple montre que

$$\cos(\theta_k) = \frac{1}{\|H_k\|_2 \, \|H_k^{-1}\|_2},$$

où la norme matricielle $\| \, \|_2$ indique la norme 2 subordonnée à la norme euclidienne classique.

On a donc convergence globale s'il existe $\alpha > 0$ tel que $\mathrm{cond}_2(H_k) \leq \alpha, \ \forall k$, où $\mathrm{cond}_2(H_k) = \|H_k\|_2 \, \|H_k^{-1}\|_2$ est ce qu'on appelle le conditionnement pour la norme 2 de la matrice H_k. En effet, on a alors

$$\cos(\theta_k) \geq \frac{1}{\alpha}, \ \forall k.$$

Comment obtenir des familles de matrices H_k définies positives et approximant la matrice hessienne $\nabla^2 f(x_k)$? On utilise pour ça une approche de type

incrémental où la matrice H_{k+1} est construite à partir de la matrice H_k et des vecteurs gradient $\nabla f(x_k)$ et $\nabla f(x_{k+1})$.

Le développement de Taylor à l'ordre un de la fonction $\nabla f(x)$ au point x_{k+1} donne

$$\nabla f(x_k) = \nabla f(x_{k+1}) + \nabla^2 f(x_{k+1})(x_k - x_{k+1}) + o(\|x_k - x_{k+1}\|).$$

Au premier ordre, on a donc

$$\nabla^2 f(x_{k+1})(x_{k+1} - x_k) \approx \nabla f(x_{k+1}) - \nabla f(x_k).$$

Par analogie avec cette égalité approchée, on impose à la matrice H_{k+1} de vérifier l'équation (dite équation de la sécante ou équation de quasi-Newton)

$$H_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k). \tag{16}$$

Bien entendu, cette équation n'est pas suffisante pour déterminer la matrice H_{k+1} à partir des points x_k, x_{k+1}, et des gradients $\nabla f(x_{k+1})$, $\nabla f(x_k)$.

Dans l'approche de type incrémental utilisée, la matrice H_{k+1} est définie à partir de la matrice H_k comme solution du problème d'optimisation suivant : rechercher la matrice H_{k+1} la plus proche de la matrice H_k et vérifiant l'équation de la sécante (16).

Afin d'obtenir des calculs plus simples, on utilise dans l'espace des matrices la norme de Frobenius, notée $\| \ \|_F$, et définie par

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}.$$

Cette norme est obtenue à partir du produit scalaire de Frobenius

$$\langle A, B \rangle_F = \sum_{i,j} a_{ij} b_{ij},$$

suivant le procédé général. On voit qu'il s'agit d'une généralisation aux matrices de la norme vectorielle euclidienne. Cette norme matricielle n'est cependant pas subordonnée à la norme vectorielle euclidienne (ni d'ailleurs à aucune norme vectorielle).

Pour simplifier les notations, on pose $s_k = x_{k+1} - x_k$ et $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. On a le résultat suivant.

Proposition 6.1. *(Broyden - 1965)*
La solution unique, notée H_{k+1}, du problème de minimisation

$$\min_{H\in\mathbb{R}^{n\times n}} \|H - H_k\|_F^2$$

avec la contrainte $Hs_k = y_k$ (équation de la sécante) est donnée par

$$H_{k+1} = H_k + \frac{1}{s_k^T s_k}(y_k - H_k s_k)s_k^T.$$

Preuve : Voir TD.

Il faut noter dans ce résultat que la matrice H_{k+1} diffère de la matrice H_k par la matrice de rang un $\frac{1}{s_k^T s_k}(y_k - H_k s_k)s_k^T$ (le vecteur colonne $(y_k - H_k s_k)$ multiplié par le vecteur ligne s_k^T donne une matrice de rang un - en termes savants, il s'agit du produit tensoriel des deux vecteurs). La matrice $(y_k - H_k s_k)s_k^T$ n'étant pas symétrique, cette proposition ne permet pas d'assurer que si la matrice H_k est symétrique, alors la matrice H_{k+1} l'est aussi. Il sera donc difficile de pouvoir assurer que H_{k+1} est définie positive si H_k l'est aussi.

Pour obtenir une solution symétrique, il suffit de modifier la formulation précédente en ajoutant la contrainte de symétrie $H^T = H$.

Proposition 6.2. *La solution unique, notée H_{k+1}, du problème de minimisation*

$$\min_{H\in\mathbb{R}^{n\times n}} \|H - H_k\|_F^2$$

avec les contraintes $Hs_k = y_k$ (équation de la sécante) et $H^T = H$ (symétrie) est donnée par

$$H_{k+1} = H_k + \frac{1}{s_k^T s_k}\left((y_k - H_k s_k)s_k^T + s_k(y_k - H_k s_k)^T\right) - \frac{(y_k - H_k s_k)^T s_k}{(s_k^T s_k)^2} s_k s_k^T.$$

En considérant une norme de Frobenius pondérée, on obtient une formule de mise à jour plus adaptée, qui permet en particulier de déduire simplement la propriété que H_{k+1} est définie positive si H_k est définie positive.

Pour cela, on observe d'abord que l'égalité de la sécante $H_{k+1}s_k = y_k$ implique que si H_{k+1} est définie positive, alors nécessairement $s_k^T y_k = s_k^T H_{k+1} s_k > 0$. On suppose donc à priori $s_k^T y_k > 0$. Il est facile de voir que cette hypothèse permet de définir une matrice définie positive W telle que $Wy_k = W^{-1}s_k$ (W n'est évidemment pas unique). On considère alors la norme de Frobenius « pondérée »par la matrice W définie par :

$$\|M\|_W = \|WMW\|_F, \forall M \in \mathbb{R}^{n\times n}.$$

Proposition 6.3. *(formule de mise à jour DFP - Davidon, Fletcher, Powell)*

On suppose $s_k^T y_k > 0$ et W une matrice définie positive telle que $W y_k = W^{-1} s_k$. La solution unique, notée H_{k+1}, du problème de minimisation

$$\min_{H \in \mathbb{R}^{n \times n}} \|H - H_k\|_W^2$$

avec les contraintes $H s_k = y_k$ (équation de la sécante) et $H^T = H$ (symétrie) est donnée par

$$H_{k+1} = (I_n - \rho_k y_k s_k^T) H_k (I_n - \rho_k s_k y_k^T) + \rho_k \, y_k y_k^T,$$

avec $\rho_k = \frac{1}{s_k^T y_k}$.

On observe que la solution H_{k+1} ne dépend pas de la matrice définie positive W vérifiant $W y_k = W^{-1} s_k$.

La preuve de la proposition 6.3 est obtenue en appliquant le résultat de la proposition 6.2 au problème

$$\min_{H' \in \mathbb{R}^{n \times n}} \|H' - H'_k\|_F^2$$

avec les contraintes $H'^T = H'$ (symétrie) et $H' s'_k = y'_k$ (équation de la sécante), où $s'_k = W^{-1} s_k = W y_k = y'_k$ et $H'_k = W H_k W$. La solution unique H'_{k+1} de ce problème donne alors $H_{k+1} = W^{-1} H'_{k+1} W^{-1}$.

La direction de descente d_k étant définie par $d_k = -H_k^{-1} \nabla f(x_k)$, on a tout intérêt a définir des formules de mise à jour directement sur la matrice inverse $B_k = H_k^{-1}$. Le produit matrice $\times$ vecteur $-B_k \nabla f(x_k)$ donne immédiatement la direction de descente d_k sans avoir à résoudre le système $H_k x = -\nabla f(x_k)$ qui a un coût numérique supérieur.

La proposition suivante donne la formule de mise à jour BFGS (Broyden, Fletcher, Goldfarb, Shanno) qui, du fait de sa complexité intéressante, est dans la pratique plus utilisée que la formule DFP. Pour la matrice $B_{k+1} = H_{k+1}^{-1}$ l'équation de la tangente devient $B_{k+1} y_k = s_k$.

Proposition 6.4. *(formule de mise à jour BFGS - Broyden, Fletcher, Goldfarb, Shanno)*

On suppose $s_k^T y_k > 0$ et W une matrice définie positive telle que $W s_k = W^{-1} y_k$. La solution unique, notée B_{k+1}, du problème de minimisation

$$\min_{B \in \mathbb{R}^{n \times n}} \|B - B_k\|_W^2$$

avec les contraintes $By_k = s_k$ *(équation de la sécante) et* $B^T = B$ *(symétrie) est donnée par*

$$B_{k+1} = (I_n - \rho_k s_k y_k^T) B_k (I_n - \rho_k y_k s_k^T) + \rho_k\, s_k s_k^T,$$

avec $\rho_k = \frac{1}{s_k^T y_k}$.

Il est intéressant de comparer la formule de mise à jour BFGS avec la formule DFP. On passe d'une formule à l'autre par échange de H_k et B_k et de y_k et s_k.

Pour compléter cette présentation on peut mentionner les formules dites duales des formules DFP et BFGS obtenues en utilisant la formule de mise à jour de Sherman-Morrison-Woodbury (SMW) qui donne l'inverse d'une matrice. Plus exactement, connaissant l'inverse d'un matrice A (supposée inversible), la formule de SMW donne l'inverse de la matrice $(A - BD^{-1}C)$ sous l'hypothèse que les matrices D et $(D - CA^{-1}B)$ sont inversibles. On montre en effet que, sous ces hypothèses, la matrice $(A - BD^{-1}C)$ est inversible et

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}.$$

Pour assurer la consistance de cette formule, il faut considérer bien sûr que les matrices inversibles A et D sont carrées, $A \in \mathbb{R}^{n\times n}$, $D \in \mathbb{R}^{m\times m}$, avec $m, n \in \mathbb{N}^*$ quelconques et $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{m\times n}$.

La formule de SMW est intéressante dans le cas où $m << n$. Il faut la voir comme une formule de mise à jour de la matrice inverse A^{-1} lorsque la matrice A est « perturbée »par une matrice de rang m petit devant n (la matrice $BD^{-1}C$ est de rang $\leq m$).

La formule duale de DFP donne la formule de mise à jour de l'inverse B_k :

$$B_{k+1} = B_k - \frac{1}{y_k^T B_k y_k} B_k y_k y_k^T B_k + \frac{1}{s_k^T y_k} s_k s_k^T.$$

La formule duale de BFGS donne la formule de mise à jour de l'inverse H_k :

$$H_{k+1} = H_k - \frac{1}{s_k^T H_k s_k} H_k s_k s_k^T H_k + \frac{1}{s_k^T y_k} y_k y_k^T.$$

Comme pour les formules DFP et BFGS on passe d'une formule duale à l'autre par échange de H_k et B_k et de y_k et s_k.

La définie positivité de la suite des matrices B_k obtenues par la formule BFGS est précisée par le résultat suivant.

Proposition 6.5. *Si* B_k *est définie positive et* $s_k^T y_k > 0$, *alors* B_{k+1} *est définie positive.*

Enfin, si le long de la direction de descente $d_k = -B_k \nabla f(x_k)$ on prend un pas de descente $\rho_k > 0$ vérifiant les conditions de Wolfe (plus exactement l'inégalité de courbure, voir (11)), on a

$$(\nabla f(x_{k+1}) - \nabla f(x_k))^T (x_{k+1} - x_k) = \rho_k (\nabla f(x_{k+1}) - \nabla f(x_k))^T d_k \geq \rho_k (\epsilon_2 - 1) \nabla f(x_k))^T d_k > 0.$$

La condition $y_k^T s_k > 0$ est alors vérifiée. Si B_k est définie positive, alors B_{k+1} l'est aussi.

La formule de mise à jour BFGS associée avec un pas de descente $\rho_k > 0$ satisfaisant l'inégalité de courbure (11) permet donc de définir pour tout k les directions de descente $d_k = -B_k \nabla f(x_k)$. L'algorithme est initialisé avec une matrice B_0 définie positive. En l'absence de toute information, on peut prendre $B_0 = I_n$, c.-à-d. une première direction de descente donnée par le gradient.

7 Conditions nécessaires et suffisantes d'optimalité dans le cas de contraintes d'égalité et d'inégalité

7.1 Contraintes d'inégalité

Nous commençons pas considérer les conditions nécessaires d'optimalité du premier ordre pour le problème (P_X) dans le cas où les contraintes définies par l'ensemble X sont de type inégalités

$$X = \{x \in \mathbb{R}^n \mid g_j(x) \leq 0,\ j = 1, \ldots, q\}.$$

Les fonctions f et g sont supposées de classe C^1.

Définition 7.1. *Soit $x \in X$. On dit que g_j est active en x si $g_j(x) = 0$. g_j est dite inactive en x si $g_j(x) < 0$.*

Qu'en est-il du cône tangent $T_x(X)$ dans le cas des contraintes d'inégalité ?

Définition 7.2. *Soit $x \in X$. On appelle cône linéarisant en x l'ensemble $L(x)$ défini par*

$$L(x) = \{v \in \mathbb{R}^n \mid \nabla g_j(x)^T v \leq 0,\ \textit{pour tout } j = 1, \ldots, q,\ \textit{tel que } g_j \textit{ est active en } x\}.$$

On montre l'inclusion $T_x(X) \subset L(x)$, pour tout $x \in X$.

Définition 7.3. *On dit que les contraintes $g(x) \leq 0$, sont qualifiées en $x \in X$ si $T_x(X) = L(x)$.*

Pour avoir l'égalité $T_x(X) = L(x)$, on doit faire des hypothèses supplémentaires sur $x \in X$.

Proposition 7.1. *(conditions suffisantes de qualification)*
Soit $x \in X$. Si les vecteurs $\nabla g_j(x)$, pour tous les indices $j = 1, \ldots, q$, tels que g_j est active en x, sont indépendants, alors les contraintes $g(x) \leq 0$, sont qualifiées en x.

Pour établir les conditions nécessaires d'optimalité d'ordre un exprimées à l'aide d'un multiplicateur de Lagrange (voir théorème 3.1 du multiplicateur de Lagrange pour le cas de contraintes d'égalité) il est nécessaire d'utiliser un résultat de géométrie des ensembles convexes.

Lemme 7.1. *(lemme de Farkas)*
Soit $K = \{By \,|\, y \in \mathbb{R}^m, y \geq 0\}$ où $B \in \mathbb{R}^{n \times m}$ et $y \geq 0 \Leftrightarrow y_j \geq 0, \forall j = 1, \ldots, m$. Alors, pour tout $g \in \mathbb{R}^n$, on a une et une seule des deux alternatives suivantes :

1. *$g \in K$,*
2. *il existe $d \in \mathbb{R}^n$, $d \neq 0$, tel que $g^T d < 0$ et $B^T d \geq 0$.*

Remarques :
- L'inégalité $B^T d \geq 0$ est équivalente à $z^T d \geq 0$, $\forall z \in K$.
- L'alternative 2 exprime le fait que l'on peut séparer par un hyperplan (d'équation $d^T x = 0$) l'ensemble convexe fermé K et le vecteur g.

Théorème 7.1. *Soit $x^* \in X$ minimum local du problème (P_X) vérifiant les conditions suffisantes de qualification des contraintes $g(x) \leq 0$ (proposition 7.1). Alors il existe $\mu^* = (\mu_1^*, \ldots, \mu_q^*)^T \in \mathbb{R}^q$, $\mu_j^* \geq 0$, $\forall j$, $j = 1, \ldots, q$, tel que*

$$\nabla f(x^*) + \sum_{j=1}^{q} \mu_j^* \nabla g_j(x^*) = 0, \tag{17}$$

et

$$\mu_j^* g_j(x^*) = 0, \, \forall j = 1, \ldots, q$$

(conditions de complémentarité).

Preuve : Soit $K = \{- \sum_{j | g_j \text{ active en } x^*} \mu_j \nabla g_j(x^*), \forall \mu_j \geq 0\}$. Montrons que $\nabla f(x^*) \in K$. Raisonnons par l'absurde. Si $\nabla f(x^*) \notin K$, par le lemme de Farkas et la définition de K, il existe alors un vecteur d tel que $\nabla f(x^*)^T d < 0$ et $-\nabla g_j(x^*)^T d \geq 0$, pour tout indice j tel que g_j est active en x^*. Par

définition du cône linéaire $L(x^*)$, on a donc $d \in L(x^*)$. Sachant que $L(x^*) = T_{x^*}(X)$ (qualification des contraintes), on aboutit ainsi à une contradiction avec la condition nécessaire d'optimalité donnée par le théorème de Peano-Kantorovich (théorème 2.7) $\nabla f(x^*)^T d \geq, \forall d \in T_{x^*}(X)$.

Enfin, pour obtenir l'égalité (17), il suffit d'ajouter les termes $\mu_j^* \nabla g_j(x^*)$ avec $\mu_j^* = 0$, pour tous les indices j pour lesquels g_j est inactive en x^*. Ceci entraîne les conditions de complémentarité $\mu_j^* g_j(x^*) = 0$, $j = 1, \ldots, q$ (soit $g_j(x^*) = 0$, soit $\mu_j^* = 0$).

7.2 Contraintes d'égalité et d'inégalité

Nous considérons dans ce paragraphe le problème de minimisation (P_X) dans le cas général où les contraintes sont du type égalités et inégalités :

$$X = \{x \in \mathbb{R}^n \,|\, h_i(x) = 0,\, i = 1, \ldots, p,\; g_j(x) \leq 0,\, j = 1, \ldots, q\}.$$

Les fonctions f, g et h sont supposées de classe C^1.

La définition de cône linéarisant s'étend au cas où on a des contraintes d'égalité et d'inégalité.

Définition 7.4. *Soit $x \in X$. On appelle cône linéarisant en x l'ensemble $L(x)$ défini par*

$$\begin{aligned} L(x) = \{v \in \mathbb{R}^n \,|\quad & \nabla h_i(x)^T v = 0, \forall i,\, i = 1, \ldots, p,\ \textit{et} \\ & \nabla g_j(x)^T v \leq 0, \forall j, j = 1, \ldots, q,\ \textit{tel que } g_j \textit{ est active en } x\}. \end{aligned} \tag{18}$$

On a, de la même façon que précédemment, l'inclusion $T_x(X) \subset L(x)$, pour tout $x \in X$. La proposition suivante donne une condition suffisante pour avoir l'égalité $T_x(X) = L(x)$ (on dit qu'on a alors qualification des contraintes en x).

Proposition 7.2. *(condition suffisante de qualification)*
Soit $x \in X$. Si les vecteurs $\nabla f_i(x)$, $\forall i$, $i = 1, \ldots, p$, et $\nabla g_j(x)$, pour tous les indices $j = 1, \ldots, q$, tels que g_j est active en x, sont indépendants, alors les contraintes d'égalité et d'inégalité $h(x) = 0, g(x) \leq 0$, sont qualifiées en x.

Pour obtenir les conditions nécessaires d'optimalité, on fait appel à une version étendue du lemme de Farkas vu précédemment.

Lemme 7.2. *(lemme de Farkas étendu)*

Soit $K = \{By + Cw \,|\, y \in \mathbb{R}^m, y \geq 0, w \in \mathbb{R}^p\}$ *défini par* $B \in \mathbb{R}^{n\times m}$, *et* $C \in \mathbb{R}^{n\times p}$. *Alors, pour tout* $g \in \mathbb{R}^n$, *on a une et une seule des deux alternatives suivantes :*

1. $g \in K$,
2. *il existe* $d \in \mathbb{R}^n$, $d \neq 0$, *tel que* $g^T d < 0$, *et* $B^T d \geq 0$, $C^T d = 0$.

Remarques :

— Les deux propriétés $B^T d \geq 0$, $C^T d = 0$ sont équivalentes à $z^T d \geq 0$, $\forall z \in K$.

— Comme précedemment, l'alternative 2 exprime le fait que l'on peut séparer par un hyperplan (d'équation $d^T x = 0$) l'ensemble convexe fermé K et le vecteur g.

On peut énoncer le théorème de Karush-Kuhn-Tucker (KKT) donnant les conditions nécessaires d'optimalité du premier ordre.

Théorème 7.2. *(théorème KKT)*

Soit $x^* \in X$ *minimum local du problème* (P_X) *vérifiant la condition suffisante de qualification des contraintes* $h(x) = 0$, $g(x) \leq 0$ *(proposition 7.2). Alors il existe* $\lambda^* = (\lambda_1^*, \ldots, \lambda_p^*)^T \in \mathbb{R}^p$, *et* $\mu^* = (\mu_1^*, \ldots, \mu_q^*)^T \in \mathbb{R}^q$, $\mu_j^* \geq 0$, $\forall j$, $j = 1, \ldots, q$, *tel que*

$$\nabla f(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{q} \mu_j^* \nabla g_j(x^*) = 0, \tag{19}$$

et

$$\mu_j^* g_j(x^*) = 0, \; \forall j = 1, \ldots, q$$

(conditions de complémentarité).

Preuve : La démonstration se fait de la même façon que pour le théorème 7.1 en montrant que $\nabla f(x^*)$ appartient à l'ensemble convexe fermé K

$$K = \{-\sum_{i=1}^{p} \lambda_i \nabla h_i(x^*) - \sum_{j|g_j \text{ active en } x^*} \mu_j \nabla g_j(x^*), \; \forall \lambda_i, \; \forall \mu_j \geq 0\}.$$

Remarques :

1. Il est évident que les conditions nécessaires d'optimalité données par les égalités (17) et (19) peuvent s'exprimer respectivement à l'aide des fonctions de Lagrange

$$L(x, \mu) = f(x) + \sum_{j=1}^{q} \mu_j g_j(x),$$

et

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x) + \sum_{j=1}^{q} \mu_j g_j(x).$$

Les égalités (17) et (19) s'écrivent respectivement $\nabla_x L(x^*, \mu^*) = 0$, et $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$.

2. Il est facile de voir que les conditions de complémentarité $\mu_j^* g_j(x^*) = 0, \forall j = 1, \ldots, q$, peuvent s'écrire de manière concise $\mu^{*T} g(x^*) = 0$, du fait que $g(x^*) \leq 0$ et $\mu^* \geq 0$.

Dans le cas où les fonctions f et g sont convexes et h affine, alors les conditions nécessaires d'optimalité données par le théorème KKT sont aussi suffisantes.

Théorème 7.3. *(CNS dans le cas convexe)*
On suppose que les fonctions différentiables f, g sont convexes, que h est affine, et que $x^* \in X$ vérifie la condition suffisante de qualification des contraintes $h(x) = 0$, $g(x) \leq 0$. Alors x^* est solution du problème (P_X) si et seulement si l'égalité (19) est vérifiée (avec $\lambda^* \in \mathbb{R}^p$, $\mu^* \in \mathbb{R}^q$, $\mu^* \geq 0$) ainsi que la condition de complémentarité $\mu^{*T} g(x^*) = 0$.

Preuve :
La preuve dans un sens est donnée par le théorème KKT (théorème 7.2). Il suffit donc de montrer que si (x^*, λ^*, μ^*) vérifie les conditions données par le théorème KKT, alors x^* est solution du problème (P_X).

Il est facile de voir que la fonction $x \mapsto L(x, \lambda^*, \mu^*)$ est convexe. On remarque ici que l'hypothèse h est affine permet de garantir la convexité de l'application $x \mapsto \sum_{i=1}^{p} \lambda_i^* h_i(x)$, ce qui n'aurait pas été possible en supposant uniquement h convexe, car les coefficients λ_i^* ne sont pas nécessairement positifs. L'égalité (19) permet de conclure que

$$L(x^*, \lambda^*, \mu^*) \leq L(x, \lambda^*, \mu^*), \forall x \in \mathbb{R}^n,$$

car on a l'équivalence $\nabla_x L(x^*, \lambda^*, \mu^*) = 0 \Leftrightarrow x^*$ est le minimum de la fonction convexe $x \mapsto L(x, \lambda^*, \mu^*)$. La condition de complémentarité et le fait que $x^* \in X$, implique $L(x^*, \lambda^*, \mu^*) = f(x^*)$. Par ailleurs, pour tout $x \in X$, on a $\sum_{i=1}^{p} \lambda_i^* h_i(x) + \sum_{j=1}^{q} \mu_j^* g_j(x) \leq 0$, car $h(x) = 0$, $g(x) \leq 0$ et $\mu^* \geq 0$. On a donc $L(x, \lambda^*, \mu^*) \leq f(x)$, $\forall x \in X$. Le résultat est démontré.

7.3 Conditions nécessaires et suffisantes du second ordre

De la même façon que dans le cas du problème avec contrainte d'égalité (voir théorème 3.2), les conditions nécessaires d'optimalité du second ordre expriment

une propriété de la matrice hessienne $\nabla^2_{xx} L(x^*, \lambda^*, \mu^*)$. Les fonctions f, g et h sont supposées de classe C^2.

Soit $x^* \in X$ un minimum local du problème (P_X) et $\lambda^* \in \mathbb{R}^p$, $\mu^* \in \mathbb{R}^q$, $\mu^* \geq 0$, des multiplicateurs vérifiant l'équation (19).

À partir du cône linéarisant $L(x^*)$ (voir définition 7.4), on définit le sous-ensemble $C(x^*, \lambda^*, \mu^*)$ appelé cône critique :

$$C(x^*, \lambda^*, \mu^*) = \{v \in \mathbb{R}^n \,|\, v \in L(x^*) \text{ et } \nabla g_j(x)^T v = 0 \text{ pour tout indice } j \text{ tel que } g_j \text{ est active en } x^* \text{ et } \mu_j^* > 0\}.$$

Il faut noter que cet ensemble dépend non seulement de $x^* \in X$ mais également des multiplicateurs des Lagrange λ^* et μ^* vérifiant l'égalité (19).

Théorème 7.4. *(condition nécessaire d'optimalité du second ordre)*
Soit $x^* \in X$ un minimum local du problème (P_X) vérifiant la condition suffisante de qualification des contraintes (proposition 7.2). Soient $\lambda^* \in \mathbb{R}^p$, $\mu^* \in \mathbb{R}^q$, $\mu^* \geq 0$, des multiplicateurs de Lagrange vérifiant les conditions nécessaires d'optimalité du premier ordre (théorème 7.2) associées à x^*. On a alors

$$v^T \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) v \geq 0, \, \forall v \in C(x^*, \lambda^*, \mu^*).$$

Cette condition nécessaire devient une condition suffisante pour que $x^* \in X$ soit un minimum local, si l'inégalité précédente est remplacée par une inégalité stricte pour tout $v \in C(x^*, \lambda^*, \mu^*)$, $v \neq 0$.

Théorème 7.5. *(condition suffisante d'optimalité)*
Soit $x^* \in X$ vérifiant la condition suffisante de qualification des contraintes (proposition 7.2) et les conditions nécessaires d'optimalité du premier ordre avec $\lambda^* \in \mathbb{R}^p$, $\mu^* \in \mathbb{R}^q$, $\mu^* \geq 0$, les multiplicateurs de Lagrange. Si

$$v^T \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) v > 0, \, \forall v \in C(x^*, \lambda^*, \mu^*), \, v \neq 0,$$

alors x^* est un minimum local du problème (P_X).

8 Dualité pour le problème d'optimisation avec contraintes

L'approche par dualité du problème d'optimisation avec contraintes apporte des informations intéressantes pour l'étude du problème et permet en outre de définir de nouvelles méthodes numériques pour le calcul de la solution.

Le théorème de Lagrange (théorème 3.1) montre qu'en un point optimum x^* on peut associer un multiplicateur de Lagrange $\lambda^* \in \mathbb{R}^p$ tel que le gradient de la fonction de Lagrange $\nabla_x L(x^*, \lambda^*)$ s'annule. La fonction de Lagrange L $(x, \lambda) \mapsto L(x, \lambda) = f(x) + \lambda^T h(x)$ associe la fonction objectif et la contrainte et permet de caractériser une solution x^* du problème (P_X) en tant que point singulier de la fonction $x \mapsto L(x, \lambda^*)$, lorsque la variable λ (appelée aussi variable duale) est fixée et égale à la valeur particulière λ^*.

Dans le cas de contraintes générales (égalités et inégalités), le théorème KKT (théorème 7.2) généralise le théorème de Lagrange. En un point optimum x^* on peut associer deux multiplicateurs de Lagrange $\lambda^* \in \mathbb{R}^p$ et $\mu^* \in \mathbb{R}^q_+$, tels que le gradient de la fonction de Lagrange $\nabla_x L(x^*, \lambda^*, \mu^*)$ s'annule.

Problème primal

On peut formuler le problème (P_X) à partir uniquement de la fonction de Lagrange $L : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^q_+ \to \mathbb{R}$, $L(x, \lambda, \mu) = f(x) + \lambda^T h(x) + \mu^T g(x)$.

Pour tout $x \in \mathbb{R}^n$, on définit la fonction primale $\overline{f}(x)$:

$$\overline{f}(x) = \sup_{(\lambda,\mu)\in\mathbb{R}^p\times\mathbb{R}^q_+} L(x, \lambda, \mu).$$

On voit facilement que

$$\overline{f}(x) = \begin{cases} f(x) \text{ si } x \in X \\ +\infty \text{ si } x \neq X \end{cases}$$

On voit ainsi que

$$X = \{x \in \mathbb{R}^n \mid \overline{f}(x) < +\infty\}.$$

L'ensemble des contraintes X est aussi appelé domaine admissible primal.

Le problème primal (P_X)

$$\begin{cases} \inf f(x) \\ x \in X \end{cases}$$

peut donc se formuler

$$\begin{cases} \inf \overline{f}(x) \\ x \in \mathbb{R}^n \end{cases}$$

autrement dit

$$\inf_{x\in\mathbb{R}^n} \left(\sup_{(\lambda,\mu)\in\mathbb{R}^p\times\mathbb{R}^q_+} L(x, \lambda, \mu) \right) \tag{20}$$

(problème d'inf-sup).

Le problème dual consiste à échanger ce problème d'inf-sup en un problème de sup-inf.

Problème dual

Pour tout $(\lambda, \mu) \in \mathbb{R}^p \times \mathbb{R}^q_+$, on définit la fonction duale f^* :

$$f^*(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu),$$

et X^* le domaine admissible dual

$$X^* = \{(\lambda, \mu) \in \mathbb{R}^p \times \mathbb{R}^q_+ \mid f^*(\lambda, \mu) > -\infty\}.$$

Le problème dual est défini par

$$\begin{cases} \sup f^*(\lambda, \mu) \\ (\lambda, \mu) \in X^* \end{cases}$$

On montre facilement que X^* est un ensemble convexe et que la fonction duale f^* est concave.

Écartant le cas sans intérêt où $X^* = \varnothing$, on a

$$\sup_{(\lambda,\mu) \in X^*} f^*(\lambda, \mu) = \sup_{(\lambda,\mu) \in \mathbb{R}^p \times \mathbb{R}^q_+} f^*(\lambda, \mu),$$

et donc le problème dual peut se formuler

$$\sup_{(\lambda,\mu) \in \mathbb{R}^p \times \mathbb{R}^q_+} \left(\inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu) \right) \tag{21}$$

(problème de sup-inf).

Quelle relation existe-t'il entre le problème primal (20) et le problème dual (21) et dans quel cas y-a-t'il équivalence (dans un sens à préciser) des deux problèmes ?

Voici quelques résultats généraux sur la dualité. Pour simplifier les notations on formule les résultats suivants en considérant une fonction de couplage générique φ entre la variable primale $x \in X$ et la variable duale $y \in Y$, $(x, y) \mapsto \varphi(x, y)$.

Le problème inf-sup

$$\inf_{x \in X} \sup_{y \in Y} \varphi(x, y)$$

est appelé problème primal.

On dit que $\overline{x} \in X$ est une solution du problème primal si

$$\sup_{y\in Y} \varphi(\overline{x}, y) = \inf_{x\in X} \sup_{y\in Y} \varphi(x, y),$$

autrement dit $\overline{x} \in X$ réalise l'inf de la fonction primale définie par $x \mapsto \sup_{y\in Y} \varphi(x, y)$.

De même, le problème sup-inf

$$\sup_{y\in Y} \inf_{x\in X} \varphi(x, y)$$

est appelé problème dual.

On dit que $\overline{y} \in Y$ est une solution du problème dual si

$$\inf_{x\in X} \varphi(x, \overline{y}) = \sup_{y\in Y} \inf_{x\in X} \varphi(x, y),$$

autrement dit $\overline{y} \in Y$ réalise le sup de la fonction duale définie par $y \mapsto \inf_{x\in X} \varphi(x, y)$.

L'inégalité suivante entre le problème dual et le problème primal est toujours vérifiée.

Proposition 8.1. *(inégalité de dualité faible)*

On a

$$\sup_{y\in Y} \inf_{x\in X} \varphi(x, y) \leq \inf_{x\in X} \sup_{y\in Y} \varphi(x, y)$$

Preuve : Pour tout $(x', y') \in X \times Y$, on a $\inf_{x\in X} \varphi(x, y') \leq \varphi(x', y')$ et donc $\sup_{y\in Y} \inf_{x\in X} \varphi(x, y) \leq \sup_{y\in Y} \varphi(x', y)$. Le membre de gauche de cette inégalité étant indépendant de $x' \in X$, on déduit le résultat.

La différence $\inf_{x\in X} \sup_{y\in Y} \varphi(x, y) - \sup_{y\in Y} \inf_{x\in X} \varphi(x, y)$ est appelé saut de dualité.

La notion de point-selle joue un rôle clé dans la dualité.

Définition 8.1. *(point-selle)*

Le couple $(\overline{x}, \overline{y}) \in X \times Y$ est un point-selle de φ sur $X \times Y$ si et seulement si

$$\varphi(\overline{x}, y) \leq \varphi(\overline{x}, \overline{y}) \leq \varphi(x, \overline{y}),\ \forall (x, y) \in X \times Y.$$

Proposition 8.2. *(caractérisation d'un point-selle)*

Le couple $(\overline{x}, \overline{y}) \in X \times Y$ est un point-selle de φ sur $X \times Y$ si et seulement si $\overline{x}$ est solution du problème primal, $\overline{y}$ est solution du problème dual, et

$$\sup_{y \in Y} \inf_{x \in X} \varphi(x, y) = \varphi(\overline{x}, \overline{y}) = \inf_{x \in X} \sup_{y \in Y} \varphi(x, y).$$

Preuve : Pour tout $(\overline{x}, \overline{y}) \in X \times Y$, on a

$$\inf_{x \in X} \varphi(x, \overline{y}) \leq \sup_{y \in Y} \inf_{x \in X} \varphi(x, y) \leq \inf_{x \in X} \sup_{y \in Y} \varphi(x, y) \leq \sup_{y \in Y} \varphi(\overline{x}, y), \qquad (22)$$

la seconde inégalité provenant de la dualité faible, les deux autres des définitions de borne sup et borne inf.

Si $(\overline{x}, \overline{y})$ est un point-selle, les termes extrêmes de l'équation (22) sont donc égaux puisque

$$\varphi(\overline{x}, \overline{y}) = \inf_{x \in X} \varphi(x, \overline{y}) = \sup_{y \in Y} \varphi(\overline{x}, y).$$

On a alors égalité partout dans (22). En particulier, l'égalité

$$\inf_{x \in X} \varphi(x, \overline{y}) = \sup_{y \in Y} \inf_{x \in X} \varphi(x, y),$$

montre que $\overline{y}$ est solution du problème dual. L'égalité

$$\inf_{x \in X} \sup_{y \in Y} \varphi(x, y) = \sup_{y \in Y} \varphi(\overline{x}, y),$$

montre que $\overline{x}$ est solution du problème primal. L'égalité du milieu montre que le saut de dualité est égal à zéro.

Réciproquement : on constate qu'on a égalité partout dans (22). On a donc

$$\varphi(\overline{x}, \overline{y}) = \inf_{x \in X} \varphi(x, \overline{y})$$

et

$$\sup_{y \in Y} \varphi(\overline{x}, y) = \varphi(\overline{x}, \overline{y}).$$

Le couple $(\overline{x}, \overline{y})$ est un point-selle.

Cette proposition montre que la propriété pour $\overline{x} \in X$ d'être associé à un point $\overline{y} \in Y$ tel que le couple $(\overline{x}, \overline{y})$ est un point-selle de φ sur $X \times Y$ est plus forte que la propriété d'être simplement solution du problème primal.

Revenons à la dualité donnée par la fonction de Lagrange L définie par $L(x, \lambda, \mu) = f(x) + \lambda^T h(x) + \mu^T g(x)$, avec $x \in \mathbb{R}^n$ variable primale et $(\lambda, \mu) \in \mathbb{R}^p \times \mathbb{R}^q_+$ variable duale.

Supposons que $\overline{x}$ soit solution du problème (P_X). Dans quelles conditions peut-on dire qu'un point dual $(\overline{\lambda}, \overline{\mu}) \in \mathbb{R}^p \times \mathbb{R}^q_+$ est tel que $(\overline{x}, \overline{\lambda}, \overline{\mu})$ est un point-selle de la fonction de Lagrange L ?

Proposition 8.3. *Si $\overline{x} \in X$ est solution du problème (P_X) et s'il existe $(\overline{\lambda}, \overline{\mu}) \in \mathbb{R}^p \times \mathbb{R}^q_+$ tel que*

$$f(\overline{x}) = \min_{x \in \mathbb{R}^n} L(x, \overline{\lambda}, \overline{\mu}),$$

alors $(\overline{x}, \overline{\lambda}, \overline{\mu})$ est un point-selle de la fonction de Lagrange L.

Preuve :

On a $f(\overline{x}) \leq L(\overline{x}, \overline{\lambda}, \overline{\mu}) = f(\overline{x}) + \overline{\mu}^T g(\overline{x})$. Or $\overline{\mu}^T g(\overline{x}) \leq 0$ car $\overline{\mu} \geq 0$ et $g(\overline{x}) \leq 0$. On en déduit $\overline{\mu}^T g(\overline{x}) = 0$ et $f(\overline{x}) = L(\overline{x}, \overline{\lambda}, \overline{\mu})$. Par ailleurs, $L(\overline{x}, \lambda, \mu) = f(\overline{x}) + \mu^T g(\overline{x})$, et donc $L(\overline{x}, \lambda, \mu) \leq f(\overline{x}) = L(\overline{x}, \overline{\lambda}, \overline{\mu})$ pour tout $(\lambda, \mu) \in \mathbb{R}^p \times \mathbb{R}^q_+$. Le résultat est démontré.

Dans le cas convexe, les hypothèses de la proposition précédente sont satisfaites.

Proposition 8.4. *On suppose que les fonctions différentiables f, g sont convexes, que h est affine, et que $\overline{x} \in X$ vérifie la condition suffisante de qualification des contraintes $h(x) = 0$, $g(x) \leq 0$. Alors $\overline{x}$ est un minimum global (P_X) si et seulement si $(\overline{x}, \overline{\lambda}, \overline{\mu})$ est un point-selle de la fonction de Lagrange L, où $(\overline{\lambda}, \overline{\mu}) \in \mathbb{R}^p \times \mathbb{R}^q_+$, sont des multiplicateurs de Lagrange vérifiant les conditions d'optimalité données par le théorème KKT (théorème 7.2).*

Preuve : On utilise la convexité de la fonction $x \mapsto L(x, \overline{\lambda}, \overline{\mu})$ qui implique l'équivalence

$$L(\overline{x}, \overline{\lambda}, \overline{\mu}) = \min_{x \in \mathbb{R}^n} L(x, \overline{\lambda}, \overline{\mu}) \Leftrightarrow \nabla_x L(\overline{x}, \overline{\lambda}, \overline{\mu}) = 0.$$

Pour conclure, on utilise la proposition précédente 8.3 ainsi que le théorème 7.3 (CNS dans le cas convexe).

9 Méthodes numériques pour l'optimisation avec contraintes

9.1 Méthode du gradient projeté

On suppose que X est un ensemble convexe fermé.

La méthode du gradient projeté consiste simplement à modifier la méthode du gradient (utilisée dans le cas du problème sans contraintes) de façon à garantir qu'à chaque étape k, $x_k \in X$.

L'itération est alors définie par

$$x_{k+1} = P_X(x_k - \rho_k \nabla f(x_k)), \tag{23}$$

où P_X est l'opérateur de projection sur l'ensemble convexe fermé X et $\rho_k > 0$ le pas de descente (qui peut être fixé ou dépendre de k).

Cette méthode est bien adaptée aux cas où l'opération de projection sur X est simple à calculer. C'est le cas en particulier lorsque X est défini par des contraintes de bornes : $X = \{x \in \mathbb{R}^n \mid a \leq x \leq b\}$, avec $a, b \in \mathbb{R}^n$.

Propriétés de la méthode du gradient projeté

Proposition 9.1. *Soit $d_k = x_{k+1} - x_k$, où x_{k+1} est défini par l'itération (23) du gradient projeté. On a les propriétés suivantes :*

1. *Si $d_k \neq 0$, alors d_k est un direction de descente, c.-à-d. $\nabla f(x_k)^T d_k < 0$.*
2. *Si $d_k = 0$, alors $\nabla f(x_k)^T (y - x_k) \geq 0$, $\forall y \in X$ (x_k vérifie les conditions nécessaires et suffisantes d'optimalité dans le cas convexe).*
3. *$x_k + \alpha d_k \in X$, $\forall \alpha \in [0, 1]$.*

Preuve :

1. Le point projeté $P_X(x_k - \rho_k \nabla f(x_k))$ vérifie l'inégalité

   $$(x_k - \rho_k \nabla f(x_k) - P_X(x_k - \rho_k \nabla f(x_k)))^T (y - P_X(x_k - \rho_k \nabla f(x_k))) \leq 0, \ \forall y \in X.$$

   Par définition de d_k, on a donc

   $$(d_k + \rho_k \nabla f(x_k))^T (y - x_k - d_k) \geq 0, \ \forall y \in X. \tag{24}$$

   En prenant $y = x_k$, on a donc $\nabla f(x_k)^T d_k \leq -\frac{1}{\rho_k} d_k^T d_k < 0$.
2. Se déduit de (24) avec $d_k = 0$.
3. Sachant que $x_k, x_{k+1} \in X$, le résultat est évident.

Résultats de convergence

Proposition 9.2. *Soit X un convexe fermé et f une fonction de classe C^1, de gradient Lipschitz (constante de Lipschitz $L > 0$) et bornée inférieurement. Alors la suite (x_k) définie par la méthode du gradient projeté $x_{k+1} = P_X(x_k - \rho_k \nabla f(x_k))$, avec $\rho_k \in [a, b] \subset]0, \frac{2}{L}[$, $\forall k$, vérifie*

$$\lim_{k \to +\infty} (x_{k+1} - x_k) = 0.$$

Tout point adhérent $x^* \in X$ de la suite (x_k) vérifie la condition nécessaire d'optimalité (théorème 2.8)

$$\nabla f(x^*)^T (y - x^*), \ \forall y \in X.$$

Preuve : On a

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2,$$

et donc $f(x_{k+1}) \leq f(x_k) - \frac{1}{\rho_k}\|d_k\|^2 + \frac{L}{2}\|d_k\|^2$, en utilisant la définition de d_k et la proposition précédente (voir preuve de la propriété 1.). On a donc

$$(\frac{1}{\rho_k} - \frac{L}{2})\|d_k\|^2 \leq f(x_k) - f(x_{k+1}).$$

Si $\rho_k \in [a, b] \subset]0, \frac{2}{L}[$, alors $\frac{1}{\rho_k} - \frac{L}{2} \geq 0$. La série de terme général $f(x_k) - f(x_{k+1})$ est donc positive. Cette série est convergente puisque ses sommes partielles $S_l = \sum_{k=0}^{l}(f(x_k) - f(x_{k+1})) = f(x_0) - f(x_{l+1})$ sont majorées du fait que fonction f est minorée. La série positive de terme général $\|d_k\|^2$ est donc convergente et $\lim_{k\to+\infty}(x_{k+1} - x_k) = 0$.

Pour tout point d'adhérence x^* de la suite (x_k), on peut extraire une sous-suite (que l'on note également (x_k)) qui converge vers x^*. Du fait que $\rho_k \in [a, b], \forall k$, la suite (ρ_k) converge également vers une valeur $\rho^* \in [a, b]$ (quitte à considérer une sous-suite de (ρ_k) et donc à nouveau une sous-suite de (x_k)). On a également $\lim_{k\to+\infty} x_{k+1} = x^*$ du fait que $\lim_{k\to+\infty}(x_{k+1} - x_k) = 0$. Par passage à la limite de l'égalité

$$x_{k+1} = P_X(x_k - \rho_k \nabla f(x_k)),$$

on a donc

$$x^* = P_X(x^* - \rho^* \nabla f(x^*)).$$

Par définition de l'opérateur de projection, on a donc

$$(x^* - (x^* - \rho^* \nabla f(x^*)))^T(y - x^*) \geq 0, \ \forall y \in X,$$

et donc $\nabla f(x^*)^T(y - x^*) \geq 0, \ \forall y \in X$.

Si la fonction f est de plus elliptique, on a le résultat suivant.

Proposition 9.3. *Soit X un convexe fermé et f une fonction de classe C^1, elliptique (constante d'ellipticité $\alpha > 0$), de gradient Lipschitz (constante de Lipschitz $L > 0$) et bornée inférieurement. Alors la suite (x_k) définie par la méthode du gradient projeté $x_{k+1} = P_X(x_k - \rho_k \nabla f(x_k))$, avec $\rho_k \in [a, b] \subset]0, \frac{2\alpha}{L}[$, $\forall k$, converge vers la solution unique x^* du problème (P_X).*

Preuve : On utilise l'égalité $x^* = P_X(x^* - \rho_k \nabla f(x^*))$, vérifiée par la solution unique x^* du problème (P_X) (voir preuve précédente). Utilisant la propriété que

P_X est contractante et les propriétés de la fonction f, on obtient facilement l'inégalité

$$\|x_{k+1} - x^*\|^2 \leq (1 + L\rho_k^2 - 2\rho_k\alpha)\|x_k - x^*\|^2.$$

Si on prend $\rho_k \in [a, b] \subset]0, \frac{2\alpha}{L}[$, $\forall k$, on a alors

$$\|x_{k+1} - x^*\| \leq \kappa\|x_k - x^*\|,$$

avec $\kappa \in]0, 1[$. La suite (x_k) converge donc vers x^*.

9.2 Méthode SQP

La méthode SQP (en anglais Sequential Quadratic Programming) peut se définir comme une méthode de Newton appliquée à la résolution du système d'optimalité dans le cas général où l'ensemble de contraintes X n'est pas l'espace $\mathbb{R}^n$ entier. On se limite ici au cas où les contraintes sont du type égalité $h(x) = 0$, mais cette méthode est aussi applicable au cas des contraintes d'inégalité $g(x) \leq 0$ et plus généralement mixte (égalité et inégalité).

On considère donc le problème (P_X) avec $X = \{x \in \mathbb{R}^n \,|\, h(x) = 0\}$, où $h : \mathbb{R}^n \to \mathbb{R}^p$.

Les conditions nécessaires d'optimalité du premier ordre pour $x^* \in X$ minimum local de (P_X) sont données par

$$\begin{cases} \nabla_x L(x^*, \lambda^*) &= 0 \\ \nabla_\lambda L(x^*, \lambda^*) &= 0 \end{cases}$$

où $L(x, \lambda) = f(x) + \lambda^T h(x)$ est la fonction de Lagrange associée au problème (P_X) et $\lambda^* \in \mathbb{R}^p$ un multiplicateur de Lagrange associé à x^*.

Le couple (x^*, λ^*) apparaît ainsi comme une solution du système

$$\begin{cases} \nabla f(x) + Jh(x)^T\lambda &= 0 \\ h(x) &= 0 \end{cases} \tag{25}$$

de $n + p$ équations à $n + p$ inconnues (les couples (x, λ)).

Il s'agit donc de résoudre le système (25). On peut pour cela faire appel à la méthode de Newton qui offre des propriétés de convergence intéressantes (convergence quadratique).

Rappel et remarque :

1. L'algorithme de Newton appliquée à la résolution d'un système général $F(z) = 0$, de m équations à m inconnues ($z \in \mathbb{R}^m$), est basé sur la linéarisation de la fonction F autour d'un point courant z_k

$$F(z_{k+1}) \approx F(z_k) + JF(z_k)(z_{k+1} - z_k).$$

L'égalité $F(z_k)+JF(z_k)(z_{k+1}-z_k)=0$ définit ainsi l'itération de Newton

$$z_{k+1} = z_k - (JF(z_k))^{-1}F(z_k)$$

(on suppose ici que la matrice jacobienne $JF(z_k)$ est inversible).

2. Il faut remarquer que si la fonction $F(z)$ est de la forme $\nabla\phi(z)$ avec $\phi : \mathbb{R}^m \to \mathbb{R}$, on retrouve l'itération de Newton associée au problème d'optimisation (voir équation (12)) $\min_{z\in\mathbb{R}^m}\phi(z)$ (il faut noter que $J(\nabla\phi)(z) = \nabla^2\phi(z)$).

L'itération de Newton appliquée au système (25) donnant (x_{k+1}, λ_{k+1}) à partir de (x_k, λ_k) conduit donc à considérer le système

$$H_k \begin{pmatrix} x_{k+1} - x_k \\ \lambda_{k+1} - \lambda_k \end{pmatrix} = - \begin{pmatrix} \nabla_x L(x_k, \lambda_k) \\ h(x_k) \end{pmatrix}, \tag{26}$$

où H_k est la matrice jacobienne du système (25) au point (x_k, λ_k). La matrice H_k est donnée par

$$H_k = \begin{pmatrix} \nabla^2_{xx} L(x_k, \lambda_k) & Jh(x_k)^T \\ Jh(x_k) & 0 \end{pmatrix}.$$

On est amené à se poser la question de l'inversibilité de la matrice H_k du système.

Un résultat classique d'algèbre linéaire (vu en TD) montre que si $A \in \mathbb{R}^{n\times n}$ symétrique est définie positive sur le sous-espace vectoriel $\ker(B)$, où $B \in \mathbb{R}^{p\times n}$ est de rang égal à p, alors la matrice $M = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}$ est inversible.

Ces propriétés sont vérifiées pour $A = \nabla^2_{xx} L(x^*, \lambda^*)$ et $B = Jh(x^*)$ lorsque (x^*, λ^*) vérifie les conditions suffisantes d'optimalité du second ordre données par le théorème 3.3. Pour (x_k, λ_k) suffisamment proche de (x^*, λ^*), on peut supposer que ces propriétés sont également vérifiées.

En posant $d_k = x_{k+1} - x_k$, le système (26) devient

$$\begin{cases} \nabla^2_{xx} L(x_k, \lambda_k) d_k + Jh(x_k)^T(\lambda_{k+1} - \lambda_k) &= -\nabla_x L(x_k, \lambda_k) \\ Jh(x_k) d_k &= -h(x_k) \end{cases} \tag{27}$$

En éliminant de la première ligne le terme $-Jh(x_k)^T\lambda_k$ qui apparaît à gauche et à droite de l'égalité, on obtient le système

$$\begin{cases} \nabla^2_{xx} L(x_k, \lambda_k) d_k + Jh(x_k)^T\lambda_{k+1} &= -\nabla f(x_k) \\ Jh(x_k) d_k &= -h(x_k) \end{cases} \tag{28}$$

Il est remarquable d'observer que ce système exprime les conditions nécessaires d'optimalité du problème

$$\min_{d \in X_k} q_k(d)$$

où q_k est la fonction quadratique

$$q_k(d) = \frac{1}{2} d^T \nabla^2_{xx} L(x_k, \lambda_k) d + \nabla f(x_k)^T d,$$

et $X_k = \{d \in \mathbb{R}^n \mid Jh(x_k)d + h(x_k) = 0\}$. La variable λ_{k+1} du système (28) joue le rôle de multiplicateur de Lagrange associé à la solution du problème $\min_{d \in X_k} q_k(d)$.

La méthode SQP consiste donc à résoudre une suite de problèmes quadratiques $\min_{d \in X_k} q_k(d)$ avec des contraintes linéaires.

Algorithme SQP

(x_0, λ_0) donné

tant que le critère d'arrêt n'est pas satisfait

- calcul de la solution (d_k, λ_{k+1}) du système (28),
- $x_{k+1} = x_k + d_k$,
- $k = k + 1$

fin tant que

Convergence de l'algorithme

Proposition 9.4. *Soit $x^* \in X$ est un minimum local du problème (P_X). Sous les hypothèses que*

- *x^* est régulier (voir définition 3.1),*
- *le couple (x^*, λ^*) satisfait les conditions d'optimalité du premier ordre et les conditions suffisantes d'optimalité du second ordre données par le théorème 3.3,*
- *les fonction f et h sont de classe C^2 et de dérivées secondes lipschitziennes,*

alors si le couple (x_0, λ_0) est suffisamment proche de (x^*, λ^*), la suite $((x_k, \lambda_k))$ générée par l'algorithme SQP converge quadratiquement vers (x^*, λ^*).

9.3 Méthode d'Uzawa

La méthode d'Uzawa est basée sur la résolution du problème dual.

On se place dans le cas où la fonction f est convexe et l'espace des contraintes X de type inégalités $g(x) \leq 0$ avec g convexe (les fonctions f et g sont de classe C^1).

D'après la proposition 8.4 on sait qu'un minimum global x^* est associé à un multiplicateur de Lagrange $\mu^* \geq 0$ tel que le couple (x^*, μ^*) est un point-selle de la fonction de Lagrange $L(x, \mu) = f(x) + \mu^T g(x)$. Ceci légitime le fait d'envisager une approche duale du problème.

On peut également considérer le cas de contraintes d'égalités $h(x) = 0$ avec h affine de façon à conserver la convexité de X et la convexité de la fonction de Lagrange par rappport à la variable x.

On rappelle le problème dual :

$$\sup_{\mu \in \mathbb{R}^q_+} \inf_{x \in \mathbb{R}^n} L(x, \mu).$$

Le problème dual consiste donc à chercher le $\sup$ (en fait le $\max$) de la fonction duale $f^*(\mu) = \inf_{x \in \mathbb{R}^n} L(x, \mu)$.

L'algorithme d'Uzawa est basé la structure sup-inf du problème dual. Il se décompose en deux étapes :

1. à partir d'une valeur courante $\mu_k \in \mathbb{R}^q_+$, calcul de $x_k \in \mathbb{R}^n$ solution de $\inf_{x \in \mathbb{R}^n} L(x, \mu_k)$,
2. réalisation d'une itération de l'algorithme du gradient projeté pour la fonction duale $f^*(\mu)$ à partir du point μ_k de façon à faire croître la valeur de celle-ci ($f^*(\mu_{k+1}) > f^*(\mu_k)$: on cherche de fait le $\max$ de la fonction duale).

L'étape 2. de l'algorithme nécessite le calcul du gradient de la fonction duale $\nabla f^*(\mu)$ au point μ_k. Sachant que par définition $f^*(\mu) = \inf_{x \in \mathbb{R}^n} L(x, \mu)$, on a le résultat suivant.

Proposition 9.5. *Pour tout $\mu \in \mathbb{R}^q_+$, on a*

$$\nabla f^*(\mu) = g(x(\mu)),$$

où $x(\mu)$ est la solution (qu'on suppose unique) de l'équation $\nabla_x L(x, \mu) = 0$.

Preuve :

On dérive par rapport à la variable μ l'égalité $f^*(\mu) = L(x(\mu), \mu)$. On a $\nabla f^*(\mu) = \nabla_\mu L(x(\mu), \mu) + Jx(\mu)^T \nabla_x L(x(\mu), \mu)$. Sachant que $\nabla_x L(x(\mu), \mu) = 0$, et $\nabla_\mu L(x(\mu), \mu) = g(x(\mu))$, on obtient le résultat.

Pour l'étape 2. de l'algorithme, on utilise donc l'égalité $\nabla f^*(\mu_k) = g(x_k)$.

Algorithme d'Uzawa

Soit $\rho > 0$ fixé, $k = 0$ et $\mu_0 \geq 0$ donné

tant que le critère d'arrêt n'est pas satisfait

- calcul de x_k solution du problème $\inf_{x \in \mathbb{R}^n} L(x, \mu_k)$,
- $\mu_{k+1} = P_{\mathbb{R}^q_+}(\mu_k + \rho_k g(x_k))$ avec $\rho_k > 0$
- $k = k + 1$

fin tant que

Supposons que le couple (x^*, μ^*) avec $\mu^* \geq 0$, soit un point-selle de la fonction de Lagrange. On a alors

$$L(x^*, \mu) \leq L(x^*, \mu^*) \leq L(x, \mu^*), \, \forall (x, \mu) \in \mathbb{R}^n \times \mathbb{R}^q_+.$$

L'inégalité de gauche montre que $(\mu - \mu^*)^T g(x^*) \leq 0$, $\forall \mu \in \mathbb{R}^q_+$, qu'on peut reformuler

$$(\mu - \mu^*)^T ((\mu^* + \rho_k g(x^*)) - \mu^*) \leq 0, \, \forall \mu \in \mathbb{R}^q_+,$$

avec $\rho_k > 0$. Cette inégalité montre que

$$\mu^* = P_{\mathbb{R}^q_+}(\mu^* + \rho_k g(x^*)), \tag{29}$$

par définition et unicité de la projection sur un ensemble convexe (ici $\mathbb{R}^q_+$). On voit ainsi que μ^* est un point fixe de l'itération de l'étape 2. de l'algorithme d'Uzawa.

Résultat de convergence

Proposition 9.6. *On suppose que la fonction f est elliptique (constante d'ellipticité $\alpha > 0$), que la fonction g est convexe et Lipschitz (constante de Lipschitz $M > 0$) et que la solution $x^* \in X$ du problème (P_X) vérifie la*

condition suffisante de qualification des contraintes $g(x) \leq 0$. *Alors la suite* (x_k) *définie par l'algorithme d'Uzawa avec* $\rho_k \in [a, b] \subset]0, \frac{2\alpha}{M^2}[$, $\forall k$, *converge vers* x^*.

Preuve :

La condition d'optimalité de x^* est donnée par

$$\nabla f(x^*) + Jg(x^*)^T \mu^* = 0, \tag{30}$$

avec $\mu^* \geq 0$. Par ailleurs, la convexité de g implique

$$g(x) \geq g(x^*) + Jg(x^*)(x - x^*),\ \forall x \in \mathbb{R}^n$$

(inégalité sur toutes les composantes des vecteurs). Sachant que $\mu^* \geq 0$, on a donc $(x - x^*)^T Jg(x^*)^T \mu^* \leq (g(x) - g(x^*))^T \mu^*$. Utilisant l'égalité (30) et l'inégalité précédente, on déduit

$$\nabla f(x^*)^T (x - x^*) + (g(x) - g(x^*))^T \mu^* \geq 0,\ \forall x \in \mathbb{R}^n. \tag{31}$$

Le point x_k solution du problème $\inf_{x \in \mathbb{R}^n} L(x, \mu_k)$, vérifie également l'égalité (30)

$$\nabla f(x_k) + Jg(x_k)^T \mu_k = 0,$$

de laquelle on déduit de la même façon

$$\nabla f(x_k)^T (x - x_k) + (g(x) - g(x_k))^T \mu_k \geq 0,\ \forall x \in \mathbb{R}^n. \tag{32}$$

En prenant $x = x_k$ dans l'inéquation (31) et $x = x^*$ dans l'inéquation (32) et en combinant les deux expressions on obtient

$$(\nabla f(x_k) - \nabla f(x^*))^T (x_k - x^*) \leq -(g(x_k) - g(x^*))^T (\mu_k - \mu^*).$$

L'ellipticité de la fonction f donne alors

$$\alpha \|x_k - x^*\|^2 \leq -(g(x_k) - g(x^*))^T (\mu_k - \mu^*). \tag{33}$$

Sachant que le couple (x^*, μ^*) est un point-selle (voir proposition 8.4), on a l'égalité (29). Grâce à la propriété de contraction de la projection $P_{\mathbb{R}^q_+}$ on a

$$\|\mu_{k+1} - \mu^*\|^2 \leq \|(\mu_k - \mu^*) + \rho_k (g(x_k) - g(x^*))\|^2.$$

En développant le second membre de cette inégalité et en utilisant l'inégalité (33) ainsi que l'inégalité de Lipschitz vérifiée par la fonction g, on obtient

$$\|\mu_{k+1} - \mu^*\|^2 \leq \|\mu_k - \mu^*\|^2 - 2\rho_k \alpha \|x_k - x^*\|^2 + \rho_k^2 M^2 \|x_k - x^*\|^2.$$

On a donc

$$\rho_k (2\alpha - \rho_k M^2) \|x_k - x^*\|^2 \leq \|\mu_k - \mu^*\|^2 - \|\mu_{k+1} - \mu^*\|^2.$$

Sachant que $\rho_k (2\alpha - \rho_k M^2) \geq a(2\alpha - bM^2) > 0$, on déduit que la suite $(\|\mu_k - \mu^*\|)$ est décroissante et donc convergente (puisque minorée par 0). L'inégalité $a(2\alpha - bM^2)\|x_k - x^*\|^2 \leq \|\mu_k - \mu^*\|^2 - \|\mu_{k+1} - \mu^*\|^2$, montre que (x_k) converge vers x^*.

9.4 Méthodes de pénalisation