Optimisation Numérique

Version 1.0

Version temporaire

François-Xavier Vialard

Mars 2014

Contents

Chapter 1

Introduction

Les problèmes d'optimisation sont omniprésents dans l'industrie, la finance et l'ingéniérie. La nature elle-même fait tendre certains systèmes physiques vers des minimums d'énergie. Une partie essentielle du travail est de modéliser la situation que l'on veut décrire comme un minimum ou maximum d'un certain critère que l'on appelle aussi fonction objectif. Une fois la fonction objectif bien définie et donc les variables sur lesquelles optimiser, une méthode d'optimisation est déterminée en fonction des caractéristiques du problème. En effet, une collection de méthodes est disponible pour différents types de problèmes et la performance d'une méthode peut varier sensiblement selon les cas auxquels on l'applique. Lorsque le résultat numérique est obtenu, on peut vérifier que le résultat est cohérent en vérifiant certaines conditions (seulement nécessaires) d'optimalité. En retour, une étude numérique d'un problème peut apporter des informations importantes pour la modélisation elle-même avec le calcul des sensibilités de la solution obtenue par rapport aux variables, aux paramètres du modèle ou aux contraintes éventuellement présentes.

Voici un exemple de problème de minimisation linéaire généralement en grande dimension:

> **Optimisation d'un plan de production et de distribution:** Une entreprise possède n usines avec un coût de production unitaire $(p_i)_{i\in[1,n]}$ et une limite de production $(M_i)_{i\in[1,n]}$. L'entreprise doit fournir ses m clients dont la demande (réel positif ou nul) est $(D_j)_{j\in[1,m]}$. Le coût d'approvisionnement unitaire des clients est supposé proportionnel à la distance entre le client et l'usine: on se donne donc une matrice $(c_{i,j})_{i,j\in[1,n]\times[1,m]}$ représentant le coût d'approvisionnement du client j par l'usine i. Le problème est d'optimiser le plan de production et de distribution afin de minimiser le coût et de fournir les clients en respectant les contraintes de production maximale de chaque usine. Le problème s'écrit donc avec $x_{i,j}$ la quantité de biens produits par l'entreprise i envoyés au client j,
>
> $$\begin{cases} \text{argmin}_{x\in\mathbb{R}_+^{nm}} \sum_{i,j}(c_{i,j}+p_i)x_{i,j}\,, \\ \text{tel que quel que soit } i\in[1,n]\ \sum_{j=1}^{m} x_{i,j} \le M_i\,, \\ \text{et quel que soit } j\in[1,m]\ \sum_{i=1}^{n} x_{i,j} = D_j\,. \end{cases} \tag{1.1}$$

Ce type de problème est appelé programmation linéaire et la méthode de minimisation la plus connue pour ce type de problèmes est la méthode du simplexe (années 1940). Plus récemment, la méthode des points intérieurs (années 1980) a été développée et est plus performante que la méthode du simplexe dans certains cas, notamment en grande dimension. Dans ce cas, le problème d'optimisation est continu: chaque variable sur laquelle optimiser appartient à un continuum $\mathbb{R}_+$. Pourtant, de nombreux problèmes font intervenir des variables discrètes, ces problèmes d'optimisation dans le cas discret

sont en général plus difficile que les problèmes d'optimisation continue lorsque ces derniers font intervenir des fonctions lisses (fonction objectif et contraintes). En effet, on peut avoir une information locale par la connaissance de la dérivée, ce qui n'est pas possible dans le cas discret.

Ces problèmes d'optimisation sont parfois contraints comme dans l'exemple (1.1). Dans cette exemple, la fonction objectif et les contraintes sont linéaires. Tout autre problème rentre dans la catégorie programmation non-linéaire.

On ne s'intéressera dans ce cours qu'à l'optimisation continue non linéaire. De plus, on ne s'intéressera qu'aux méthodes locales qui déterminent un minimum local, i.e. une solution qui est un minimiseur dans un voisininage de cette solution.

Les méthodes que l'on va étudier sont les méthodes de descente qui sont basées sur la technique suivante: Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ au moins C^1, on dispose donc autour d'un point x_0 d'une information locale

$$f(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + o(\|x - x_0\|) \,. \tag{1.2}$$

Les algorithmes de descente partent d'un point $x \in \mathbb{R}^n$ et itèrent le processus suivant:

- choisir une direction de descente d pour diminuer la valeur de la fonction objectif, par exemple $-\nabla f(x_0)$,
- choisir un pas de descente α et mettre à jour x par $x + \alpha d$.

On appliquera ces méthodes sur différents problèmes jouets tel que l'optimisation de portefeuilles:

> **Optimisation de portefeuilles:** Un investisseur doit répartir ces investissements entre n actions S_i qui ont un rendement r_i qui est aléatoire mais dont on connait l'espérance m_i et la matrice de covariance, *i.e.* $Q := E[(r_i - m_i)(r_j - m_j)]$. On cherche donc à optimiser un équilibre entre le rendement du portefeuille et le risque associé (la variance du portefeuille), ce qui conduit à l'optimisation sous contraintes de
>
> $$f(x) = \frac{1}{2}\langle x, Qx \rangle - \langle x, m \rangle \text{ tel que } \begin{cases} \sum_{i=1}^n x_i = 1 \\ x_i \geq 0 \, \forall i \in [1, \ldots, n] \,. \end{cases}$$
>
> Ce modèle a été proposé par Markowitz. On peut préférer l'optimisation suivante qui fixe un minimum de rendement pour le portefeuille et optimise sur la variance:
>
> $$f(x) = \frac{1}{2}\langle x, Qx \rangle \text{ tel que } \begin{cases} \sum_{i=1}^n x_i = 1, \\ \sum_{i=1}^n x_i r_i \geq r \,, \\ x_i \geq 0 \, \forall i \in [1, \ldots, n] \,. \end{cases}$$

Ce type de problème est connu sous le nom de programmation quadratique et des méthodes spécifiques ont été développées pour sa résolution. On s'intéressera dans ce cours à des méthodes plus générales qui s'appliquent pour des fonctions régulières quelconques.

Les formulations variationnelles sont souvent utilisées en traitement d'image. On donne dans ce qui suit un exemple de régularisation H^1:

> **Problème de zooming:** Une image $I_{n,m}$ est une matrice de taille $n \times m$. Il est naturel de modéliser cette image comme la représentation discrétisée d'une fonction continue $I : [0, a] \times [0, b] \mapsto \mathbb{R}$. On considère que $I_{n,m}$ est donné par la projection $p_{n,m}$ dans L^2 de $I : [0, a] \times [0, b] \mapsto \mathbb{R}$ sur le sous espace des fonctions constantes par morceaux sur une grille cartésienne. Etant donné l'image $I_{n,m}$, comment zoomer sur l'image, c'est à dire représenter l'image par une matrice de taille $N \times M$ où $M > m$ et $N > n$?

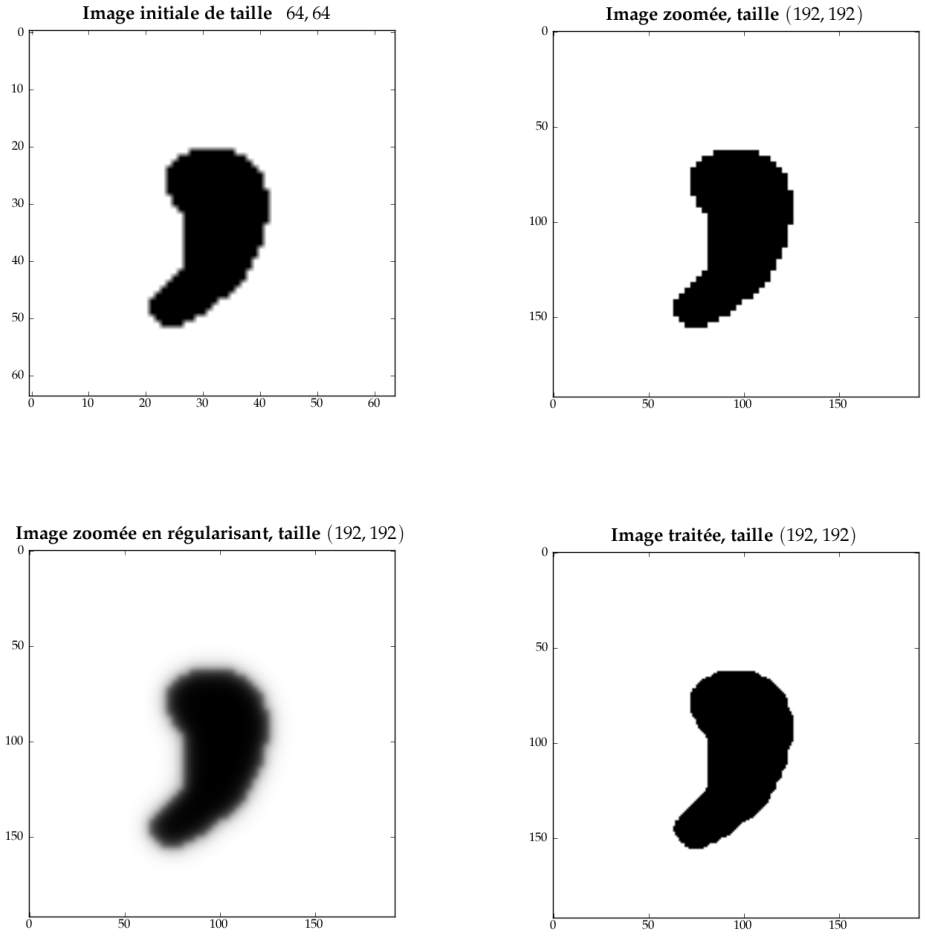
Le problème variationnel qu'on peut introduire dans une forme géneérale est

$$\mathcal{E}(I_{N,M}) = \text{Régularisation}(I_{N,M}) + \text{Similarité}(p_{n,m}(I_{N,M}), I_{n,m})\,. \quad (1.3)$$

Afin de pouvoir utiliser les résultats de cours, on implémentera une méthode numérique pour

$$\text{Régularisation}(I_{N,M}) := \sum_{i=1}^{N-1}\sum_{j=1}^{M-1} (I_{N,M}(i+1,j) - I_{N,M}(i,j))^2 + (I_{N,M}(i,j+1) - I_{N,M}(i,j))^2 \quad (1.4)$$

et le terme de similarité est le carré de la norme L^2. Un exemple d'utilisation de ce modèle est donné par les figures 1 et 1. L'image initiale est binaire et un seuillage est effectué sur l'image régularisée. C'est un exemple jouet car il n'est pas nécessaire de d'utiliser une régularisation H^1 pour régulariser l'image zoomée.



La fonctionnelle $\mathcal{E}$ est une fonction quadratique des valeurs de l'image sur la grille. C'est donc encore un problème de programmation quadratique. La régularisation utilisant la norme H^1 introduit un flou sur les bords des images binaires. Dans l'exemple précédent, on a simplement utilisé un seuillage pour palier ce problème. Une approche variationnelle permettant résoudre ce problème utilise la norme

$\|u\|_\varepsilon = \sum_{i,j} \sqrt{\|\nabla f(x)\|^2 + \varepsilon^2}$ pour une valeur de ε suffisamment faible. Ce cas n'appartient plus au domaine de programmation quadratique.

Plus généralement, certains problèmes sont par nature hautement non linéaires et non convexes. Par exemple, la recherche d'un chemin optimal (dans un sens précisé ci-dessous) dans un ouvert de $\mathbb{R}^n$ pour un produit scalaire g qui dépend du point où l'on se trouve. Le calcul du carré de la longueur d'un chemin donne

$$\int_0^1 g(x)(\dot{x}, \dot{x})\, dt\,, \tag{1.5}$$

où $x : [0,1] \mapsto \mathbb{R}^n$ est un chemin vérifiant $x(0) = a$. On cherche à se rapprocher d'un point b à moindre coût défini par

$$\begin{cases} \mathcal{J}(x \in C^1([0,1], \mathbb{R}^n)) = \int_0^1 g(x)(\dot{x}, \dot{x})\, dt + \|x(1) - b\|^2\,, \\ x(0) = a\,. \end{cases} \tag{1.6}$$

On peut montrer que le problème se ramène à l'optimisation d'une fonction $f : \mathbb{R}^n \mapsto \mathbb{R}$ sur un paramètre p_0 qui permet de reconstruire la trajectoire optimale par une équation différentielle ordinaire:

$$\begin{cases} \dot{x} = a(x,p) \\ \dot{p} = b(x,p)\,. \end{cases} \tag{1.7}$$

et f est donnée par $f(p_0) = g^{-1}(a)(p_0, p_0) + \|x(1) - b\|^2$. On a donc que $x(1)$ est le résultat de l'intégration du système (1.7) pour les conditions initiales $x(0) = a$ et $p(0) = p_0$. Le terme $\|x(1) - b\|^2$ est donc une fonction de p_0. Ce type de problème ne sera pas étudié dans ce cours, mais on développe des méthodes qui peuvent s'appliquer à la fonction f lorsqu'on sait calculer le gradient de f (ce qui peut se faire par la méthode de l'adjoint).

Ce cours s'appuie sur différentes sources et notamment [Bon97] et [NW06, Ber99] (en anglais).

Chapter 2

Prérequis

2.1 Minimisation et condition du premier ordre

2.1.1 Approche théorique

L'optimisation d'un point de vue général peut se présenter comme:

Problème 1. *Étant donné un ensemble X (ensemble de contraintes) et une fonction $f : X \mapsto \mathbb{R}$ (fonction de coût), chercher $x^\star \in X$ tel que*

$$f(x^\star) \leq f(x) \tag{2.1}$$

quel que soit $x \in X$.

La formulation de ce problème est évidemment très vaste et on se concentrera dans ce cours au cas ou le domaine d'optimisation peut-être décrit de manière continue. Par exemple, on rencontrera dans la suite $X = \mathbb{R}^n$, $X = \Omega \subset \mathbb{R}^n$ un ouvert convexe ou $X = S^n \subset \mathbb{R}^{n+1}$ la sphère unité de dimension n.

Définition 1. *Si un tel $x^\star$ existe, on l'appelle alors minimum global de f sur X. Ce minimum est dit strict si*

$$x \neq x^\star \Rightarrow f(x^\star) < f(x)\,. \tag{2.2}$$

A priori, rien ne garantit l'existence de ce minimum. Sur l'intervalle ouvert $]0, 2[$, la fonction $x \mapsto x$ n'a pas de minimum. Pour garantir l'existence d'un minimum, on utilise souvent les hypothèses de continuité et de compacité d'un ensemble de sous-niveau défini par $\{x \in \Omega \mid f(x) \leq f(x_0)\}$:

Proposition 1. *Soit $f : \Omega \subset \mathbb{R}^n \mapsto \mathbb{R}$ une fonction continue sur Ω ouvert. S'il existe $x_0 \in \Omega$ tel que $S_0 = \{x \in \Omega \mid f(x) \leq f(x_0)\}$ est compact alors f admet un minimum global.*

Souvent, certaines hypothèses sont faites sur f pour obtenir ce type de propriété comme par exemple la coercivité: f est dite coercive si

$$\lim_{\|x\| \to +\infty} \|f(x)\| = +\infty \tag{2.3}$$

On ne développera pas de méthodes permettant de garantir la convergence vers un minimum global mais seulement local (en général).

Définition 2. *Soit $\Omega \subset \mathbb{R}^n$ un ouvert et $f : \Omega \mapsto \mathbb{R}$ une fonction. Un vecteur $x_0 \in \Omega$ est dit minimum local de la fonction f s'il existe un réel $\varepsilon > 0$ tel que quel que soit $x \in \Omega$,*

$$\|x - x_0\| \leq \varepsilon \Rightarrow f(x_0) \leq f(x)\,. \tag{2.4}$$

La notion de minimum local ne requiert qu'une notion de voisinage pour être définie donc cette notion s'étend immédiatement aux espaces munis d'une topologie, comme par exemple la sphere S^n. On rappelle la définition de la differentiabilité

Définition 3. *Une fonction $f : \Omega \subset \mathbb{R}^n \mapsto \mathbb{R}^p$ est différentiable au point x_0 si il existe une application linéaire notée $f'(x_0) \in L(\mathbb{R}^n, \mathbb{R}^p)$ et $\varepsilon : \mathbb{R}^n \mapsto \mathbb{R}^p$ avec $\lim_{h\to 0} \varepsilon(h) = 0$ telle que*

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \|h\|\varepsilon(h)\,. \tag{2.5}$$

Si $f'(x_0)$ existe alors cette application linéaire est unique, *i.e.* la seule à vérifier l'équation (2.5). Si une fonction est différentiable en x_0, alors elle est continue en x_0.

Définition 4. *Soit une fonction différentiable $f : \mathbb{R}^n \mapsto \mathbb{R}^p$. Un point $x \in \mathbb{R}^n$ est dit point critique de f si $rg(df(x)) < p$. Dans ce cas, la valeur $f(x)$ est dite valeur critique.*

Pour l'optimisation, on s'intéresse le plus souvent au cas où $p = 1$:

Remarque 1. *Pour une fonction $f : \mathbb{R}^n \mapsto \mathbb{R}$, le point $x \in \mathbb{R}^n$ est un point critique si $f'(x) = 0$.*

Toujours dans le cas où $p = 1$, on peut écrire la dérivée de $f' \in L(\mathbb{R}^n, \mathbb{R}) = \mathbb{R}^{n*}$ comme un vecteur dans $\mathbb{R}^n$ que l'on appelle le gradient et qui est donné au point x_0 par: $\nabla f(x_0) = \begin{bmatrix} \partial_{e_1} f(x_0) \\ \vdots \\ \partial_{e_n} f(x_0) \end{bmatrix}$ avec $(e_1, \ldots, e_n)$ la base canonique de $\mathbb{R}^n$. On a alors,

$$f'(x_0)(v) = \sum_{i=1}^{n} \partial_{e_i} f(x_0) v_i = \langle \nabla f(x_0), v \rangle\,, \tag{2.6}$$

avec $\langle \cdot, \cdot \rangle$ le produit scalaire usuel défini par $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$ pour $x, y \in \mathbb{R}^n$. Attention, le gradient est un vecteur que l'on a présenté en colonne car on l'assimile à un vecteur de $\mathbb{R}^n$ et non une application linéaire dans $L(\mathbb{R}^n, \mathbb{R})$ qui s'écrit

$$L = \begin{bmatrix} \partial_{e_1} f(x_0) & \ldots & \partial_{e_n} f(x_0) \end{bmatrix}\,,$$

comme une matrice de taille $(1, n)$ (1 ligne et n colonnes). On a donc $\nabla f(x_0) = L^t$ où L^t désigne la transposée de L.

Proposition 2 (Condition nécessaire d'optimalité). *Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ une fonction et x_0 un minimum local de f. Alors, si f est différentiable en x_0, on a $\nabla f(x_0) = 0$ ou de manière équivalente, $\partial_i f(x_0) = 0$ quel que soit $i \in [\![1, n]\!]$.*

Preuve: Soit $(e_i)_{i\in[\![1,n]\!]}$ la base canonique de $\mathbb{R}^n$. Par définition du minimum local on a $f(x_0 + \varepsilon e_i) - f(x_0) \geq 0$ quel que soit $\varepsilon > 0$, donc on obtient $\frac{f(x_0+\varepsilon e_i)-f(x_0)}{\varepsilon} \geq 0$ pour $\varepsilon \geq 0$ et $\frac{f(x_0+\varepsilon e_i)-f(x_0)}{\varepsilon} \leq 0$ pour $\varepsilon \leq 0$. En passant à la limite, on obtient $\partial_i f(x) \geq 0$ et $\partial_i f(x) \leq 0$, ce qui prouve le résultat. □

Une conséquence de ce théorème en dimension 1 pour une fonction $f : \mathbb{R} \mapsto \mathbb{R}$ est le théorème des accroissements finis:

Proposition 3 (Accroissements finis). *Soit $f : \mathbb{R} \mapsto \mathbb{R}$ une fonction continue sur $[a, b]$ et dérivable sur $]a, b[$ alors il existe $c \in]a, b[$ tel que*

$$f(b) = f(a) + f'(c)(b - a)\,. \tag{2.7}$$

Preuve: La fonction $G(x) := (f(b)-f(x))-\frac{b-x}{b-a}(f(b)-f(a))$ satisfait $G(b) = G(a) = 0$. Si $G \equiv 0$ alors le resultat est trivialement vérifié. Sinon, quitte à considérer $-G$, on peut supposer qu'il existe $x_0 \in]a, b[$ tel que $G(x_0) < 0$. Comme G est continue sur $[a, b]$ compact, G admet un minimum local en un point c nécessairement différent de a et b pour lequel l'égalité 2.7 est verifiée en utilisant le théorème 2. □

2.1.2 Approximation numérique des dérivées

On propose dans cette section une estimation numérique des dérivées lorsque qu'on dispose d'un programme qui renvoie la valeur (approchée) de la fonction. On appelle **oracle** un tel programme, qui peut retourner aussi la valeur de la dérivée et la dérivée seconde. Si f est différentiable, on a l'approximation

$$\frac{f(x+\varepsilon)-f(x)}{\varepsilon} \simeq f'(x)\,. \tag{2.8}$$

Soit un oracle qui retourne une approximation $\hat{f}$ de la valeur de la fonction avec p chiffres caractéristiques, alors on a

$$\hat{f}(x) = f(x) + \eta\|f\|_\infty\,, \tag{2.9}$$

avec η de l'ordre de 10^{-p}. La formule d'approximation (2.10) devient donc

$$\left\| \frac{\hat{f}(x+\varepsilon)-\hat{f}(x)}{\varepsilon} - \frac{f(x+\varepsilon)-f(x)}{\varepsilon} \right\| \leq \frac{2\|f\|_\infty\eta}{\varepsilon}\,. \tag{2.10}$$

De plus, si f est C^2, on verra (voir la proposition 4 ci-après) que

$$\|\frac{f(x+\varepsilon)-f(x)}{\varepsilon} - f'(x)\| \leq \frac{\|f''\|_\infty}{2}\|\varepsilon\| \tag{2.11}$$

On obtient alors

$$\left\| \frac{\hat{f}(x+\varepsilon)-\hat{f}(x)}{\varepsilon} - f'(x) \right\| \leq \frac{\|f''\|_\infty}{2}\varepsilon + \frac{2\|f\|_\infty\eta}{\varepsilon}\,. \tag{2.12}$$

Si on minimise le second membre, on obtient

$$\varepsilon = 2\sqrt{\frac{\|f\|_\infty}{\|f''\|_\infty}\eta}\,. \tag{2.13}$$

En conclusion, si le système est bien conditionné *i.e.* $\frac{\|f\|_\infty}{\|f''\|_\infty}$ d'ordre 1 alors un bon choix est $\varepsilon = \sqrt{\eta}$.

Si la fonction f est plus régulière, il est préférable d'utiliser une méthode de différences finies centrée qui fait gagner en précision: en effet, le fait de centrer la formule d'approximation permet de se débarasser des termes symétriques:

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)(h,h) + O(\|h\|^3)\,. \tag{2.14}$$

On a alors

$$\frac{f(x+h)-f(x-h)}{2\|h\|} = f'(x) + O(\|h\|^2)\,, \tag{2.15}$$

plus précisément $O(\|h\|^2) = \frac{1}{6}\|f^{(3)}\|_\infty\|h\|^2$ lorsque f est C^3. Dans ce cas, un bon choix de ε peut être donné par le minimum en ε du terme de droite de

$$\left\| \frac{\hat{f}(x+\varepsilon)-\hat{f}(x-\varepsilon)}{2\varepsilon} - f'(x) \right\| \leq \frac{\|f^{(3)}\|_\infty}{6}\varepsilon^2 + \frac{2\|f\|_\infty\eta}{\varepsilon}\,. \tag{2.16}$$

Le minimum est atteint en $\varepsilon^3 = \frac{6\|f\|_\infty}{\|f^{(3)}\|_\infty}\eta$. On obtient donc, sous l'hypothèse que $\|f\|_\infty$ et $\|f^{(3)}\|_\infty$ sont du même ordre de grandeur, qu'une bonne valeur de ε est de l'ordre de $\eta^{1/3}$ et l'erreur commise sur la dérivée est de l'ordre de $\eta^{2/3}$.

Remarque 2. *Attention, si on choisit la différence centrée, on doit faire appel n fois de plus à l'oracle que pour la différence finie non centrée. Cela peut être coûteux lorsque n est grand et lorsque l'évaluation de la fonction est coûteuse.*

2.2 Conditions du second ordre et convexité

2.2.1 Application du développement de Taylor au second ordre

En vue d'obtenir des résultats de convergence pour les méthodes développées dans la suite, on demande souvent plus de régularité sur les fonctions pour disposer d'estimations telles que:

Proposition 4 (Estimation de second-ordre). *Soit $\Omega \subset \mathbb{R}^n$ un ouvert et $f : \Omega \mapsto \mathbb{R}$ une fonction C^1 et vérifiant,*

$$\|\langle \nabla f(x), v\rangle - \langle \nabla f(y), v\rangle\| \leq L\|x-y\|\|v\|\,, \forall v \in \mathbb{R}^n \tag{2.17}$$

pour $x, y \in C$ un convexe de Ω et L un réel strictement positif (on dit que ∇f est $L-$Lipschitz sur C pour la norme induite sur les formes linéaires). Alors, on a

$$f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2\,, \tag{2.18}$$

pour $x, y \in C$.

Preuve: On introduit la fonction $G : \mathbb{R} \mapsto \mathbb{R}$ définie par $G(t) = f(x+t(y-x)) - f(x)$. Par hypothèse sur f, G est C^1 donc on peut écrire

$$G(1) - G(0) = \int_0^1 \langle \nabla f(x + t(y-x)), y-x\rangle\, dt$$
$$f(y) - f(x) - \langle \nabla f(x), y-x\rangle = \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y-x\rangle\, dt$$

Comme ∇f est Lipschitz de constante L sur $[x, y]$, on majore le terme de droite pour obtenir

$$f(y) - f(x) - \langle \nabla f(x), y-x\rangle \leq \int_0^1 Lt\|y-x\|^2\, dt = \frac{L}{2}\|y-x\|^2\,, \tag{2.19}$$

ce qui donne le résultat annoncé. □

On remarque que ce résultat est local, puisqu'on a besoin des hypothèses faites sur f uniquement le long du chemin $x + t(y-x)$. D'autre part, l'hypothèse que ∇f est Lipschitz est localement vérifiée si f est C^2. On obtient donc le corollaire suivant:

Corollaire 1. *Soit f est $C^2(\mathbb{R}^n, \mathbb{R})$ alors l'estimation de la proposition 4 est valable sur tout compact K t.q. $[x, y] \subset K \subset \mathbb{R}^n$ avec $L = \sup_{y\in K} |\frac{\partial^2 f}{\partial_{x_i}\partial_{x_j}}(y)|$.*

On rappelle que si $x \in \Omega \mapsto f'(x) \in L(\mathbb{R}^n, \mathbb{R}^p)$ est C^1 alors f est dite C^2 et on a

$$f''(x_0) \in L(\mathbb{R}^n, L(\mathbb{R}^n, \mathbb{R}^p)) \simeq L_2(\mathbb{R}^n, \mathbb{R}^p)\,, \tag{2.20}$$

où $L_2(\mathbb{R}^n, \mathbb{R}^p)$ désigne l'espace des applications bilinéaires de $\mathbb{R}^n$ dans $\mathbb{R}^p$. C'est à dire que $f''(x_0)$ est à valeurs dans $\mathbb{R}^p$, prend comme arguments deux vecteurs de $\mathbb{R}^n$ et est linéaire en chacun de ces deux arguments. On rappelle le théorème de Schwarz qui dit que $f''(x_0)$ est symétrique si les dérivées partielles $\partial_{i,j} f$ sont continues en x_0. Ce théorème peut s'expliquer par le fait que le quotient

$$\frac{f(x_0+h_1+h_2) - f(x_0+h_1) - f(x_0+h_2) + f(x_0)}{\|h_1\|\|h_2\|}$$

est symétrique en h_1, h_2 et ce rapport tend vers $f''(x_0)(e_1, e_2)$ quand $h_1 = \varepsilon_1 e_1$ et $h_2 = \varepsilon_2 e_2$ et $\varepsilon_1, \varepsilon_2$ tendent vers 0. Le double passage à la limite, est rendu licite (utiliser par exemple le théorème des accroissements finis) sous l'hypothèse de continuité des

dérivées partielles. On peut donc présenter la notation matricielle de la dérivée seconde de $f : \mathbb{R}^n \mapsto \mathbb{R}$ appelée Hessienne de f par

$$\nabla^2 f(x_0) = \begin{bmatrix} \partial_{1,1} f(x_0) & \dots & \partial_{1,n} f(x_0) \\ \vdots & & \vdots \\ \partial_{n,1} f(x_0) & \dots & \partial_{n,n} f(x_0) \end{bmatrix} \tag{2.21}$$

et on a

$$f''(x_0)(h_1, h_2) = \langle h_1, \nabla^2 f(x_0) h_2 \rangle (= h_1^t \nabla^2 f(x_0) h_2) \,. \tag{2.22}$$

On présente une condition nécessaire pour l'obtention d'un minimum mais qui n'est pas suffisante:

Théorème 1. *Si $f : \Omega \mapsto \mathbb{R}$ admet un minimum local en x_0 (Ω ouvert) et est deux fois différentiable en x_0 alors*

$$f''(x_0)(h, h) \geq 0$$

pour tout $h \in \mathbb{R}^n$.

Preuve: La condition nécessaire du premier ordre s'applique dans ce cas et donc $f'(x_0) = 0$. En appliquant la définition de la différentiabilité au second ordre en x_0, on obtient pour $\eta > 0$ et $h \in \mathbb{R}^n$

$$f(x_0 + \eta h) - f(x_0) = \eta^2 \frac{1}{2} f''(x_0)(h, h) + \eta^2 \varepsilon(\eta h) \|h\|^2 \geq 0 \,, \tag{2.23}$$

la dernière inégalité étant vérifiée dans un voisinage de x_0. Enfin, en divisant par η^2, on obtient en passant à la limite:

$$\frac{1}{2} f''(x_0)(h, h) \geq 0 \,. \tag{2.24}$$

□

Remarque 3. *Cette condition n'est pas suffisante pour assurer que x_0 est un minimum local. La fonction définie sur $\mathbb{R}$, $f(x) = x^3$ en est un contre exemple.*

D'autre part, le théorème de Schwarz ne s'applique pas sous les hypothèses du théorème et $f''(x_0)$ peut ne pas être symétrique. Dans la suite, les fonctions que l'on minimisera seront suffisamment régulières et ce cas ne se présentera pas.

On peut quand même proposer une condition suffisante sur la Hessienne pour assurer qu'un point critique est un minimum local:

Théorème 2. *Si x_0 est un point critique de $f : \mathbb{R}^n \mapsto \mathbb{R}$ deux fois différentiable au point x_0 et s'il existe $\alpha > 0$ tel que*

$$\nabla^2 f(x_0)(h, h) \geq \alpha \|h\|^2 \tag{2.25}$$

pour $h \in \mathbb{R}^n$ alors x_0 est un minimum local strict pour f.

Preuve: Le développement limité de f à l'ordre 2 donne, pour $h \in \mathbb{R}^n$

$$f(x_0 + h) - f(x_0) = \frac{1}{2} f''(x_0)(h, h) + \varepsilon(h) \|h\|^2 \,, \tag{2.26}$$

Par hypothèse, il existe $r > 0$ tel que $\|h\| < r$ implique $\|\varepsilon(h)\| < \frac{1}{4}\alpha$. On en déduit

$$f(x_0 + h) - f(x_0) > \frac{1}{2}\alpha \|h\|^2 - \frac{1}{4}\alpha \|h\|^2 = \frac{1}{4}\alpha \|h\|^2 > 0 \,.$$

ce qui donne le résultat □

2.2.2 Utilisation de la convexité

Lorsqu'on a plus d'hypothèses sur la fonction (comme la convexité), on peut obtenir des résultats globaux. On va supposer que Ω est un ouvert convexe de $\mathbb{R}^n$. On rappelle qu'un ensemble S d'un espace affine est dit convexe si quels que soient $a, b \in S$ et $t \in]0,1[$, $ta + (1-t)b \in S$.

Définition 5. *Une fonction $f : \Omega \mapsto \mathbb{R}$ est dite convexe sur Ω si quels que soient $a, b \in S$, on a $f(ta + (1-t)b) \leq tf(a) + (1-t)f(b)$ et strictement convexe si cette inégalité est stricte pour $t \in]0,1[$.*

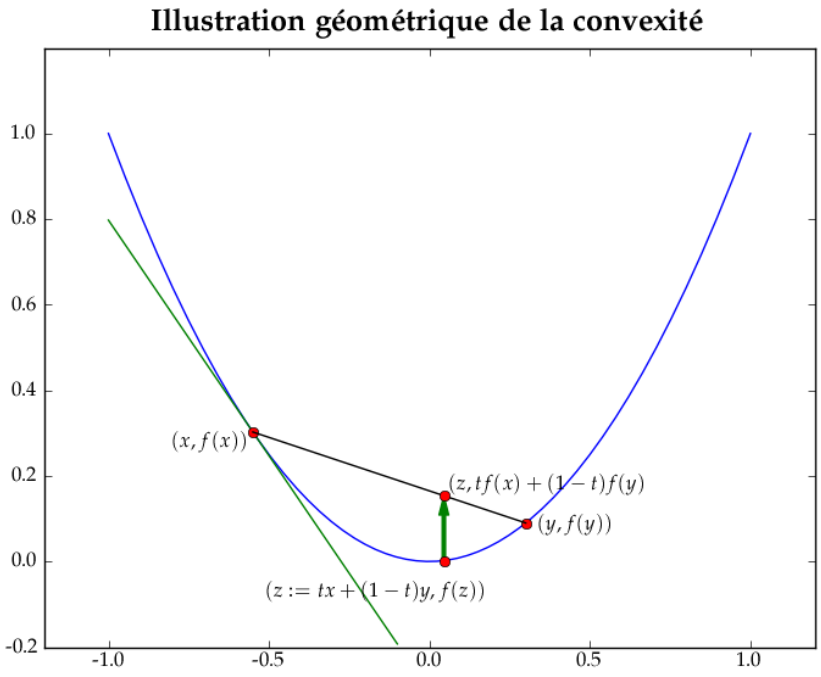


Figure 2.1: La fonction $x \mapsto x^2$ est strictement convexe.

Si de plus f est différentiable alors la propriété de convexité est équivalente à: quels que soient $x, y \in \Omega$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle . \tag{2.27}$$

Cette propriété (le graphe est au dessus de sa tangente) ainsi que la définition sont illustrées dans la figure 2.1. Cette propriété dit qu'une connaissance locale pour une fonction convexe donne de l'information globale sur la fonction:

Proposition 5. *Soit f est une fonction différentiable et convexe sur Ω (convexe). Si $\nabla f(x_0) = 0$ alors x_0 est un minimum global pour f.*

Exercice 1. *Donner une preuve de la proposition en utilisant l'équation* (2.27).

Lorsque la fonction f est C^2 et convexe, la Hessienne de f, $\nabla^2 f$ est positive en tout point. Réciproquement, une fonction f de classe C^2 est convexe si $\nabla^2 f$ est positive en tout point. Beaucoup de propriétés peuvent être prouvées pour des fonctions convexes. En voici quelques unes à démontrer en exercice:

Exercice 2.
- *Soit $f : \mathbb{R} \mapsto \mathbb{R}$ une fonction convexe croissante et $g : \Omega :\mapsto \mathbb{R}$ une fonction convexe. Montrer que $f \circ g$ est convexe.*

- *Montrer qu'un supremum de fonctions convexes est convexe.*
- *Montrer que l'épigraphe d'une fonction f convexe, ensemble défini par*

$$Epi(f) := \{(x, y) \mid y \geq f(x)\} ,$$

est convexe.

On a vu que l'hypothèse de convexité peut apporter des propriétés importantes pour un problème de minimisation convexe (Ω est convexe et f convexe) mais:

Remarque 4. *L'hypothèse de convexité ne garantit pas l'existence d'une solution à un problème de minimisation ni son unicité.*

En revanche, on montre facilement:

Proposition 6. *L'ensemble des minimiseurs est convexe et la stricte convexité assure l'unicité d'un éventuel minimiseur.*

Attention, là encore, l'existence d'un minimiseur n'étant pas assurée, cet ensemble peut être vide.

Exercice 3. *Que dire de la minimisation de la fonction $f : \mathbb{R}^n \mapsto \mathbb{R}$ définie par $f(x) = \langle a, x\rangle$ pour $a \in \mathbb{R}^n$ non nul, sous la contrainte $\sum_{i=1}^n x_i^4 = 1$?*

Définition 6. *Soit $\Omega \subset \mathbb{R}^n$ un domaine convexe. On dit qu'une fonction f de classe $C^2(\Omega, \mathbb{R})$ est elliptique s'il existe un réel $c > 0$ tel que quel que soit $x \in \Omega$,*

$$\nabla^2 f(x) \geq c\, Id\,. \tag{2.28}$$

Une fonction elliptique est en particulier strictement convexe. La réciproque n'est pas vraie comme le montre la fonction $f : x \mapsto x^4$.

Exercice 4. *Montrer qu'une fonction elliptique $f : \mathbb{R}^n \mapsto \mathbb{R}$ est coercive et que par conséquent, si $\Omega \subset \mathbb{R}^n$ est un convexe fermé, le minimum de f sur Ω existe et est unique.*

2.3 Optimisation sous contraintes

2.3.1 Position du problème

On s'intéresse au problème suivant de minimisation sous contraintes (MSC)

$$\textbf{(MSC)}\ \ \operatorname{argmin}_{x\in\mathbb{R}^n} f(x) \text{ tel que } x \text{ satisfait les conditions } \begin{cases} c_i(x) = 0\,,\ i \in \mathcal{E} \\ c_j(x) \geq 0\,,\ j \in \mathcal{I} \end{cases} \tag{2.29}$$

sous l'hypothèse suivante:

$$\text{les fonctions } c_i \text{ pour } i \in \mathcal{E} \cup \mathcal{I} \text{ et la fonction } f \text{ sont } C^1(\mathbb{R}^n, \mathbb{R})\,. \tag{2.30}$$

Evidemment, on peut reformuler le problème d'optimisation (2.29) par $\operatorname{argmin}_{x\in\Omega} f(x)$ où Ω est défini par

$$\Omega = \{x \in \mathbb{R}^n \mid c_i(x) = 0\,,\ i \in \mathcal{E} \text{ et } c_j(x) \geq 0\,,\ j \in \mathcal{I}\} \tag{2.31}$$

Le point important est que le minimum recherché peut se trouver au bord du domaine et les conditions d'optimalité du premier ordre ne sont plus les mêmes que lorsque le minimum est dans l'intérieur du domaine. Pour obtenir les conditions d'optimalité du premier ordre, connues sous le nom de théorème de Karush-Kuhn-Tucker (KKT), il est nécessaire de supposer certaines conditions locales (*i.e.* réalisées au point optimal x considéré) telles que l'indépendance linéaire des contraintes pour les contraintes qui sont actives en x ou d'autres types de conditions. On définit donc:

Définition 7. *L'ensemble des contraintes actives au point $x \in \Omega$ noté $\mathcal{A}(x)$ est défini par*

$$\mathcal{A}(x) = \mathcal{E} \cup \{i \in \mathcal{I} \mid c_i(x) = 0\} \,. \tag{2.32}$$

Une condition (dite de qualification) qui permet d'obtenir le théorème de KKT est la suivante

Définition 8. *Un point $x \in \Omega$ est dit qualifié pour le problème (2.29) si les vecteurs $\{\nabla c_i(x) \mid i \in \mathcal{A}(x)\}$ sont linéairement indépendants.*

L'intérêt de cette condition permet de décrire au premier ordre le voisinage dans Ω d'un point $x \in \Omega$.

Définition 9. *Le cône des directions séquentiellement admissibles au point $x \in \Omega$ est noté $S(x)$ et est défini par*

$$S(x) = \left\{ d \in \mathbb{R}^n \mid \exists\, (x_k)_{k\in\mathbb{N}} \in \Omega^{\mathbb{N}} \text{ de limite } x \text{ et } (t_k)_{k\in\mathbb{N}} \in \mathbb{R}_+^{*\mathbb{N}} \text{ t.q. } \lim_{k\to+\infty} \frac{x_k - x}{t_k} = d \right\} . \tag{2.33}$$

Si une direction d est dans ce cône, on dira que d est une direction (séquentiellement) admissible en x.

Il est facile de voir que $S(x)$ est stable par multiplication par un scalaire strictement positif (propriété définissant un cône). De plus, on a $0 \in S(x)$. On admet la proposition suivante:

Proposition 7. *Si l'hypothèse de qualification (8) est vérifiée au point $x \in \Omega$, alors on a*

$$S(x) = D(x) \tag{2.34}$$

avec $D(x)$ le cône des directions admissibles au premier ordre défini par

$$D(x) = \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \langle d, \nabla c_i(x)\rangle = 0 \text{ pour } i \in \mathcal{E}\,, \\ \langle d, \nabla c_j(x)\rangle \geq 0 \text{ pour } j \in \mathcal{A}(x) \setminus \mathcal{E}\,. \end{array} \right\} . \tag{2.35}$$

Une possible stratégie de preuve repose sur le théorème des fonctions implicites qui utilise directement la condition d'indépendance.

La première condition nécessaire est

Proposition 8. ***Condition nécessaire au premier ordre:*** *Si x est un minimum local pour f alors $\langle \nabla f(x), d\rangle \geq 0$ pour toute direction admissible $d \in S(x)$.*

On remarque qu'on a pas besoin de l'hypothèse de qualification au point x.

Preuve: On considère une suite de points $x_k \in \Omega$ telle que la direction limite $d \neq 0$ est admissible. On remarque qu'il suffit de montrer la condition pour $\|d\| = 1$. On a alors

$$f(x_k) - f(x) = \|x_k - x\|\langle \nabla f(x), d\rangle + o(\|x_k - x\|) \geq 0 \tag{2.36}$$

ce qui implique, en divisant par $\|x_k - x\|$ (c'est possible pour k assez grand) et en passant à la limite,

$$\langle \nabla f(x), d\rangle \geq 0\,.$$

□

2.3.2 Conditions au premier ordre et second ordre

On écrit alors les conditions au premier ordre de Karush-Kuhn-Tucker (KKT):

Théorème 3. *Si $x \in \Omega$ est une solution du problème de minimisation (2.29) et qu'il est qualifié, alors il existe un vecteur $\lambda \in \mathbb{R}^p$ où p est le nombre de contraintes tel que, si on introduit $L(x,\lambda) := f(x) - \sum_{i=1}^{p} \lambda_i c_i(x)$ on a*

$$\nabla_x L(x,\lambda) = 0\,, \tag{2.37}$$
$$c_i(x) = 0 \qquad \text{for } i \in \mathcal{E}\,, \tag{2.38}$$
$$c_j(x) \geq 0 \qquad \text{for } j \in \mathcal{I}\,, \tag{2.39}$$
$$\lambda_j \geq 0 \qquad \text{for } j \in \mathcal{I}\,, \tag{2.40}$$
$$\lambda_i c_i(x) = 0 \qquad \text{for } i \in \mathcal{E} \cup \mathcal{I}\,. \tag{2.41}$$

Remarque 5. *Dans le cas où il n'y a pas de contraintes d'inégalités ($\mathcal{I} = \emptyset$), le résultat est une conséquence d'une propriété bien connue d'algèbre linéaire. En effet, par un développement limité au premier ordre pour le minimiseur x, on obtient: $\langle \nabla f(x), d\rangle = 0$ pour $d \in \cap_{i\in\mathcal{E}} (\nabla c_i(x))^{\perp}$, ce qui implique que*

$$\nabla f(x) \in \mathbf{Vect}(\nabla c_i(x))\,. \tag{2.42}$$

On donne maintenant une ébauche de preuve, suffisament précise pour que l'étudiant puisse la compléter.

Preuve: Dans la suite, on indice les contraintes actives $i = 1, \ldots, k$. Une preuve variationnelle du théorème utilise l'application suivante:

$$\Phi : \lambda \in \mathbb{R}^k \mapsto \frac{1}{2}\|\nabla f(x) - \sum_{i=1}^{k} \lambda_i \nabla c_i(x)\|^2 \tag{2.43}$$

que l'on cherche à minimiser sur l'ensemble convexe défini par

$$E := \left\{\lambda \in \mathbb{R}^k \;\middle|\; \lambda_j \geq 0 \text{ si } k \in \mathcal{A}(x)\right\}\,.$$

L'application Φ est elliptique sur $\mathbb{R}^k$ car sa Hessienne est constante égale à $\langle \nabla c_i(x), \nabla c_j(x)\rangle$ (on rappelle que x est fixé) qui est définie positive par l'hypothèse de qualification. Il existe donc (voir l'exercice 4) un unique minimiseur λ^* de (2.43) sur Ω. On note

$$r := -\nabla f(x) + \sum_{i=1}^{k} \lambda_i^* \nabla c_i(x)\,.$$

Montrons que r est une direction admissible: Notons $\nabla C(x)$ la matrice des gradients des contraintes au point $x \in \Omega$, $\nabla C(x) := \left[\nabla c_1(x) \quad \ldots \quad \nabla c_p(x)\right]$. On a alors, en notant $(e_j)_{j\in[1,k]}$ la base canonique de $\mathbb{R}^k$,

- si $j \in \mathcal{E}$ alors $\langle r, \nabla c_j(x)\rangle = 0$ *i.e.* $\langle r, \nabla C e_j\rangle = 0$,
- si $\lambda_j^* > 0$ pour $j \in \mathcal{A}(x) \setminus \mathcal{E}$ alors $r \in \nabla c_j(x)^{\perp}$ (pourquoi?),
- si $\lambda_j^* = 0$ alors par la condition nécessaire du premier ordre (proposition 8) pour (2.43), on a nécessairement $\langle r, \nabla C(e_j)\rangle \geq 0$.

Ce qui implique les conditions d'admissibilité pour r. On a aussi $[\nabla C]^t r \geq 0$ et $\langle [\nabla C]^t r, \lambda^*\rangle = 0$ et par conséquent $\langle \nabla f(x), r\rangle = -\|r\|^2$.

D'autre part, la condition du premier ordre 8 assure $\langle \nabla f(x), r\rangle \geq 0$ puisque r est une direction admissible (on utilise ici l'hypothèse de qualification). On obtient donc $r = 0$. □

Remarque 6. *L'hypothèse de qualification implique directement que le multiplicateur de Lagrange $(\lambda_i)_{i\in[1,p]}$ donné par le théorème 3 est unique. C'est aussi compris dans la preuve par l'unicité du minimum qui reste vraie lorsqu'on relache les hypothèses sur λ_i.*

En général, cette condition nécessaire n'est évidemment pas suffisante, mais peut le devenir si on dispose d'hypothèses supplémentaires, en particulier des hypothèses de convexité:

Théorème 4. *Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ une fonction convexe de classe C^1. Si les conditions d'égalité sont affines et les conditions d'inégalités sont définies par $c_i(x) \geq 0$ avec c_i concave, alors tout point x^* vérifiant les conditions de KKT pour un λ^* est un minimiseur global.*

Remarque 7. *On rappelle qu'on a utilisé la matrice $\nabla C(x) := \begin{bmatrix}\nabla c_1(x) & \dots & \nabla c_p(x)\end{bmatrix}$ dans la preuve du théorème précédent. On notera aussi $C(x) := \begin{bmatrix} c_1(x) & \dots & c_p(x)\end{bmatrix}$.*

Preuve: On se restreint au sous-espace affine E défini par les contraintes d'égalités affines $\mathcal{E}$ et dans la suite $x \in E$. Le Lagrangien du système $L(x, \lambda^*) = f(x) - \langle \lambda, C(x)\rangle$ est convexe sur E par les hypothèses sur f et c_i et la positivité de λ^*. On a alors

$$f(x) \geq L(x, \lambda^*) \geq L(x^*, \lambda^*) + \langle \nabla_x L(x^*, \lambda^*), x - x^*\rangle = f(x^*)\,. \tag{2.44}$$

□

Comme dans le cas non contraint, on peut dériver des conditions nécessaires au second ordre. On commence par définir le cône pour lequel les conditions du premier ordre ne sont pas suffisantes pour décider si x est un minimum local.

Théorème 5. *Soit (x, λ) un couple de points vérifiant les conditions de KKT pour une fonction f et des fonctions c_i pour $i = 1, \dots, p$ de classe C^2. Le cône $C(x, \lambda)$ des directions critiques est défini par*

$$d \in C(x,\lambda) \equiv \begin{cases} \langle \nabla c_i(x), d\rangle = 0 \text{ pour } i \in \mathcal{E}\,, \\ \langle \nabla c_i(x), d\rangle = 0 \text{ pour } i \in \mathcal{A}(x) \cap \mathcal{I} \text{ et } \lambda_i > 0\,. \\ \langle \nabla c_i(x), d\rangle \geq 0 \text{ pour } i \in \mathcal{A}(x) \cap \mathcal{I} \text{ et } \lambda_i = 0\,. \end{cases} \tag{2.45}$$

Si, de plus, x est un minimum local de f pour lequel les conditions de qualification sont satisfaites, alors

$$\langle d, \nabla^2_{xx} L(x,\lambda) d\rangle \geq 0 \text{ quel que soit } d \in C(x,\lambda)\,. \tag{2.46}$$

Remarque 8. *On vérifie facilement que $C(x, \lambda) \subset D(x)$.*

On ne donne pas la preuve de ce résultat qui peut être trouvé dans [NW06]. L'hypothèse de qualification est importante dans ce théorème. En revanche, elle n'est pas nécessaire dans la condition suffisante énoncée ci-dessous, dont on donne une preuve pour permettre de bien comprendre comment intervient la définition du cône $C(x, \lambda)$.

Théorème 6. *Soit (x, λ) un couple de points vérifiant les conditions de KKT pour une fonction f et des fonctions c_i pour $i = 1, \dots, p$ de classe C^2. Si on suppose que*

$$\langle d, \nabla^2_{xx} L(x,\lambda) d\rangle > 0 \text{ quel que soit } d \in C(x,\lambda) \setminus \{0\}\,. \tag{2.47}$$

alors x est un minimum local strict de f.

Preuve: Supposons que x ne soit pas un minimum local strict de f, alors il existe une suite de points $x_k \neq x$ de limite x vérifiant les contraintes tels que $f(x_k) \leq f(x)$. En utilisant un développement de Taylor au premier ordre, on obtient:

$$\langle \nabla f(x), d\rangle \leq 0\,. \tag{2.48}$$

D'autre part, on a

$$L(x_k, \lambda) = f(x_k) - \langle \lambda, C(x_k) \rangle \leq f(x_k) \leq f(x) = L(x, \lambda)\,. \tag{2.49}$$

On en déduit avec $d_k = \frac{x_k - x}{\|x_k - x\|}$

$$L(x_k, \lambda) - L(x, \lambda) = \|x_k - x\| \langle \nabla_x L, d_k \rangle + \|x_k - x\|^2 \frac{1}{2} \langle d, \nabla^2_{xx} L(x, \lambda) d \rangle + o(\|x_k - x\|^2) \leq 0\,, \tag{2.50}$$

Par compacité de la sphère unité, on peut choisir une suite (x_k) telle que d_k converge dans S_{n-1} vers une limite d qui vérifie nécessairement

$$\begin{aligned} &\langle \nabla c_i(x), d \rangle = 0 \text{ pour } i \in \mathcal{E}\,, \\ &\langle \nabla c_i(x), d \rangle \geq 0 \text{ pour } i \in \mathcal{A}(x) \cap \mathcal{I} \end{aligned}$$

ce qui implique, comme $\lambda_i \geq 0$ pour $i \in \mathcal{A}(x) \cap \mathcal{I}$ et $\lambda_i = 0$ pour $i \in \mathcal{I} \setminus \mathcal{A}(x)$.

$$\sum_{i=1}^{p} \lambda_i \langle \nabla c_i(x), d \rangle \geq 0\,. \tag{2.51}$$

D'autre part, en utilisant la condition du premier ordre de KKT et (2.48), on obtient

$$\langle \nabla f(x), d \rangle = \sum_{i=1}^{p} \lambda_i \langle \nabla c_i(x), d \rangle \leq 0\,. \tag{2.52}$$

et donc

$$\sum_{i=1}^{p} \lambda_i \langle \nabla c_i(x), d \rangle = 0\,.$$

Cela implique que si $\lambda_i > 0$ alors $\langle \nabla c_i(x), d \rangle = 0$ et donc $d \in C(x, \lambda)$. L'équation (2.50) se réécrit

$$L(x_k, \lambda) - L(x, \lambda) = \|x_k - x\|^2 \frac{1}{2} \langle d_k, \nabla^2_{xx} L(x, \lambda) d_k \rangle + o(\|x_k - x\|^2) \leq 0\,. \tag{2.53}$$

ce qui implique $\langle d, \nabla^2_{xx} L(x, \lambda) d \rangle \leq 0$ et contredit les hypothèses du théorème. □

Remarque 9. *Pour assurer qu'un point vérifie les conditions de KKT, on a vu qu'une condition de qualification est exigée. Il existe en fait différentes conditions de qualification et elles peuvent être moins exigeantes que l'indépendance linéaire demandée dans ce cours. Par exemple,*

- *Les contraintes actives sont affines dans un voisinage de x.*
- *Ω est convexe, les contraintes d'égalité sont affines, les contraintes d'inégalités sont concaves et C^1 et il existe un point admissible $x_0 \in \Omega$ tel que $c_i(x_0) > 0$ pour toute contrainte d'inégalité non linéaire telle que $c_i(x) = 0$.*
- ***Mangasarian-Fromovitz**: Les fonctions c_i sont de classe C^1. Il existe un vecteur $w \in \mathbb{R}^n$ tel que*

  $$\begin{aligned} &\langle \nabla c_i(x), w \rangle = 1 \text{ pour } i \in \mathcal{A}(x) \cap \mathcal{I}\,, \\ &\langle \nabla c_i(x), w \rangle = 0 \text{ pour } i \in \mathcal{E}\,. \end{aligned}$$

  et les gradients $(\nabla c_i(x))_{i \in \mathcal{E}}$ sont linéairement indépendants.

On pourra retenir qu'il est possible de se passer de la condition de qualification 8 pour obtenir certains des résultats précédents (mais pas tous).

2.3.3 Multiplicateurs de Lagrange

En optimisation sous contraintes, les multiplicateurs de Lagrange jouent un rôle important dans l'énoncé des résultats. En pratique, les multiplicateurs de Lagrange donnent une indication sur la force qu'une contrainte exerce contre la minimisation de la fonction objectif. Par exemple, si la contrainte est inactive, le multiplicateur de Lagrange associé est nul. Réciproquement, si $\lambda_i = 0$, un perturbation de la contrainte va avoir un effet négligeable au premier ordre. La quantité significative est le gradient de f

$$\nabla f(x) = \sum_{i=1}^{p} \lambda_i \nabla c_i(x)\,, \tag{2.54}$$

qui fait intervenir le produit $\lambda_i \nabla c_i(x)$. On peut donc remarquer que multiplier c_i par un scalaire positif β ne change pas ce produit mais multiplie simplement λ_i par $\frac{1}{\beta}$.

2.3.4 De l'optimisation sous contraintes à l'optimisation sans contraintes

Le reste de ce cours se concentre sur les méthodes d'optimisation sans contraintes. On ne développera pas d'algorithme spécifique dans le cas contraint mais il est important de noter qu'on peut parfois réduire un problème contraint à un problème non contraint:

Cas affine: Sous des contraintes d'égalités affines données par $Ax = b$ pour $A \in L(\mathbb{R}^n, \mathbb{R}^p)$ et b dans l'image de A, le problème (MSC) est équivalent au problème de minimisation non contraint sur $\mathbb{R}^k$ où k est la dimension du noyau de A et $x_0 \in \mathbb{R}^n$ vérifiant $Ax_0 = b$ de $\tilde{f}$ où f est définie par: soit $v_1, \ldots, v_k$ une base de $Ker(A)$ alors on peut minimiser $\tilde{f}(s_1, \ldots, s_k) = f(x_0 + \sum_{i=1}^{k} s_i v_i)$.

Cas de la sphère: par exemple, la projection stéréographique permet de traiter un problème contraint du type

$$\begin{cases} \min_{x \in \mathbb{R}^3} f(x) \\ \sum_{i=1}^{3} x_i^2 = 1\,. \end{cases} \tag{2.55}$$

Approche par pénalisation: Une autre approche est de pénaliser le non respect des contraintes: en transformant le problème de minimisation initial (2.29) en un problème non contraint mais différent du problème initial:

$$\text{argmin}_{x \in \mathbb{R}^n} f(x) + \sum_{i=1}^{p} \alpha_i g(c_i(x)) := f_\alpha\,, \tag{2.56}$$

où g est une fonction régulière, par exemple $g(x) = \|x\|^2$ pour des contraintes d'égalité. Les scalaires α_i sont strictement positifs. On espère alors trouver une solution proche d'une solution du problème initial en minimisant successivement (2.56) pour des valeurs des α_i de plus en plus grandes.

Définition 10. *Soit Ω un fermé de $\mathbb{R}^n$. On dit que g est une fonction de pénalité pour Ω si g vérifie $g(x) = 0$ si $x \in \Omega$ et $g(x) > 0$ si $x \notin \Omega$.*

Proposition 9. *Soientt $f : \mathbb{R}^n \mapsto \mathbb{R}$ une fonction à minimiser pour $x \in \Omega$ un fermé de $\mathbb{R}^n$, g une fonction de pénalité pour Ω et $t_k \to +\infty$ une suite croissante de réels. On suppose que f atteint son minimum sur Ω en un point x^* et qu'il existe $t > 0$ tel que l'ensemble fermé $S := \{x \in \mathbb{R}^n \mid f(x) + tg(x) \leq f(x^*)\}$ est borné. En définissant*

$$x_k := \text{argmin}_{x \in \mathbb{R}^n} f(x) + t_k g(x)\,,$$

on a $\lim_{k \to \infty} f(x_k) = f(x^*)$ et $\lim_{k \to \infty} g(x_k) = 0$.

Preuve: On note $f_k = f + t_k g$, on a alors évidemment $f_{k+1} \geq f_k$ et donc $f_{k+1}(x_{k+1}) \geq f_k(x_k)$. On a donc une suite croissante de valeurs $f_k(x_k)$ majorée par $f(x^*)$. De plus, la suite x_k est bornée car $x_k \in S$. Pour toute suite extraite convergente de x_k, on a que $\lim_{k\to\infty} g(x_k) = 0$ (car $\lim_{k\to\infty} f_k(x_k) = +\infty$ si $g(\lim_{k\to\infty} x_k) > 0$). □

Ce résultat ne dit pas comment choisir la suite t_k par exemple. Ce type de questions doivent être résolues en pratique. Un point faible de la méthode précédente est que le Hessien de la fonction f_α à minimiser est

$$H(x) = \nabla^2 f(x) + \sum_{i=1}^{p} \alpha_i \left(c_i(x)\nabla^2 c_i(x) + \nabla c_i(x)\, {}^t\nabla c_i(x) \right) .$$

Lorsque les valeurs des α_i sont grandes, on voit que

$$H(x) \simeq \sum_{i=1}^{p} \alpha_i \nabla c_i(x)\, {}^t\nabla c_i(x) ,$$

puisqu'on peut supposer que $c_i(x)$ est de l'ordre de $\frac{1}{\alpha_i}$ si x est un minimum de f_α, ce qui implique que le H est le plus souvent dégénéré et l'utilisation de méthodes de type Newton semble difficile.

Méthode du Lagrangien augmenté:

Une amélioration de la méthode précédente dans le cas de contraintes d'égalité (mais généralisable au cas de contraintes d'inégalités) est la définition d'un Lagrangien augmenté $L : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}_+^* \mapsto \mathbb{R}$

$$L(x, \lambda, \mu) = f(x) - \sum_{i=1}^{p} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i=1}^{p} |c_i(x)|^2 . \tag{2.57}$$

On remarque dans ce cas que L et f coïncident sur Ω. On considère donc une suite croissante de réels μ_k et $\lambda^0 = 0 \in \mathbb{R}^p$ et on définit par récurrence la suite

$$x_k := \mathrm{argmin}_{x\in\mathbb{R}^n} L(x, \lambda^k, \mu_k) .$$

Au point x_k, on a donc:

$$\nabla_x L(x_k, \lambda^k, \mu_k) = 0 = \nabla f(x_k) - \sum_{i=1}^{p} (\lambda_i^k - c_i(x)) \nabla c_i(x) . \tag{2.58}$$

Dans ce cas, l'idée est d'itérer des minimisations sur $\mathbb{R}^n$ en mettant à jour le multiplicateur de Lagrange λ et en augmentant μ afin d'obtenir un point vérifiant les conditions du théorème de KKT. Pour cela, on constate que si $x_k \to x^*$ une solution (qualifiée) du problème d'optimisation, λ^k converge vers les multiplicateurs de Lagrange associés au point x^*. Cette remarque suggère un algorithme de type point fixe pour la mise à jour du multiplicateur de Lagrange λ:

$$\lambda_i^{k+1} = \lambda_i^k - c_i(x) . \tag{2.59}$$

L'avantage de cette méthode est de pouvoir éviter le choix de valeurs trop élevées pour μ qui conduit au problème de mauvais conditionnement du Hessien. Sous certaines hypothèses, on peut obtenir des résultats de convergence locale. Pour aller plus loin, on pourra consulter [NW06].

En conclusion, les algorithmes de minimisation sans contraintes sont la base de la minimisation sous contraintes et sont étudiés dans la suite de ce cours.

Chapter 3

Les méthodes de descentes

3.1 Algorithme général

Les méthodes de descente pour des fonctions $f : \Omega \subset \mathbb{R}^n \mapsto \mathbb{R}$ (Ω ouvert de $\mathbb{R}^n$) sont des algorithmes du type suivant:

Initialisation de $x \leftarrow x_0 \in \Omega$.
Tant que (critère d'arrêt) **faire**
  Choix d'une direction *(de descente)* d,
  Choix d'un pas de descente $\varepsilon > 0$ tel que $x + \varepsilon d \in \Omega$,
  $x \leftarrow x + \varepsilon d$.
Fin
Retourner x

Algorithme 1: Méthode de descente générique

Cet algorithme 1 est appelé méthode de descente car on fait en sorte de choisir une direction de descente et un pas tels que $f(x)$ décroît à chaque itération. Un exemple de critère d'arrêt est $\|\nabla f(x_k)\| > \eta$ ce qui est suggéré par le Théorème 7. On s'intéressera particulièrement au taux de convergence de la méthode qui est une mesure d'efficacité de la méthode.

Définition 11.
- *On dit que la convergence de la suite $(x_k)_{k\in\mathbb{N}}$ vers sa limite x^* est linéaire si*
$$\|x^* - x_{k+1}\| \leq \alpha \|x^* - x_k\|, \tag{3.1}$$
pour $\alpha > 0$.
- *Le taux de convergence asymptotique est défini par $r := \limsup_{k\to\infty} \frac{\|x^* - x_{k+1}\|}{\|x^* - x_k\|}$.*
- *La convergence est dite superlinéaire si $r = 0$.*
- *S'il existe $\beta > 0$ et $\gamma > 0$ tels que $\limsup_{k\to\infty} \frac{\|x^* - x_{k+1}\|}{\|x^* - x_k\|^\beta} = \gamma$, on dit que l'ordre de convergence de la suite est β.*

Dans la suite, on discute le choix de la direction et le choix du pas.

3.2 Choix de la direction de descente

Evidemment, cette question a un véritable intérêt quand la dimension de l'espace d'optimisation est au moins 2. On préfère restreindre le terme direction de descente aux cas suivants:

Définition 12. *Une direction de descente pour f au point x est un vecteur $d \in \mathbb{R}^n$ tel que*

$$\forall\, T > 0\ \exists\, \rho \in]0, T[\text{ tel que } f(x + \rho d) < f(x)\,. \tag{3.2}$$

Une direction de descente au sens strict pour une fonction $f \in D^1(\mathbb{R}^n, \mathbb{R})$ est un vecteur $d \in \mathbb{R}^n$ tel que

$$\langle \nabla f(x), d \rangle < 0\,. \tag{3.3}$$

3.2.1 Descente de gradient simple

Une idée intuitive pour choisir la direction de descente est simplement de la définir comme "la" direction dans laquelle la fonction f décroît le plus. Il est évident qu'on peut déterminer un sous-espace vectoriel $E_- := \{v \mid df(v) < 0\}$ sur lequel f décroît au premier ordre. Mais il n'est a priori pas évident de déterminer "la" direction optimale de descente sans une structure additionnelle sur $\mathbb{R}^n$ comme une norme ou un produit scalaire. En effet, plus intuitivement on voudrait minimiser $df(v)$ sur E_-, ce qui n'a bien sûr pas de sens puisque si $E_- \neq \{0\}$ alors $\inf_{v \in E_-} df(v) = -\infty$. Il faut donc se donner une façon de mesurer les vecteurs de E_-. Si on dispose d'une norme sur $\mathbb{R}^n$, alors l'ensemble des directions "optimales" est clairement défini:

$$D_{\|\dot\|}(x) := \operatorname{argmin} \{df_x(v) \mid v \in \mathbb{R}^n \text{ tel que } \|v\| = 1\}\ . \tag{3.4}$$

Dans la suite, on notera parfois D au lieu de $D_{\|\dot\|}(x)$ et $df(v)$ au lieu de $df_x(v)$.

Exercice 5. *Prouver que D est non vide. Prouver que, si la norme vérifie la propriété*

$$\forall (x, y) \in \mathbb{R}^n\ \|x + y\| = \|x\| + \|y\| \Rightarrow x, y \text{ colinéaires} \tag{3.5}$$

alors D est un singleton.

Proposition 10. *Soit $\mathbb{R}^n$ muni du produit scalaire usuel, pour f dérivable en x et $d \in \mathbb{R}^n$, si $\langle d, \nabla f(x) \rangle < 0$ alors d est une direction de descente.*

Exercice 6.

- *Donner une preuve de la proposition 10.*
- *Montrer que, si $\nabla f(x) \neq 0$, alors $D(x) = \{-\nabla f(x)\}$ (en fait, ceci est valable pour un produit scalaire quelconque).*
- *Si f est une fonction convexe sur $[x, y]$ et $f(x) > f(y)$ alors $d = y - x$ est une direction de descente.*

Pour la norme L^1, le résultat de l'exercice 5 ne s'applique pas car cette norme ne vérifie pas la propriété demandée. En fait l'ensemble des directions de descente n'est pas réduit à un unique vecteur dans tous les cas. On peut montrer:

Proposition 11. *Pour $\|x\|_1 = \sum_{i=1}^n |x_i|$, l'ensemble des directions optimales de descente au point x pour f est donné par*

$$\operatorname{Conv} \left\{ -\operatorname{sgn}(\partial_i f(x)) e_i \,\middle|\, |\partial_i f(x)| = \sup_{j=1\dots n} |\partial_j f(x)| \right\}\ , \tag{3.6}$$

où e_i est la base usuelle de $\mathbb{R}^n$.

Preuve: On a, par minoration directe, en notant $\|\nabla f(x)\|_\infty = \sup_{j=1\dots n} |\partial_j f(x)|$

$$\frac{df(v)}{\|v\|_1} = \frac{\sum_{i=1}^n \partial_i f(x) v_i}{\|v\|_1} \geq -\|\nabla f(x)\|_\infty \tag{3.7}$$

Cette borne inférieure est atteinte pour chaque vecteur de l'ensemble

$$\text{Ext} = \left\{ e_i \,\middle|\, |\partial_i f(x)| = \|\nabla f(x)\|_\infty \right\}$$

donc on a Ext $\subset D$. On vérifie alors aisément que $E = \text{Conv}(Ext) \subset D$.
Soit maintenant $v \in \mathbb{R}^n$, on écrit $v = v_E + u$ où v_E est la projection orthogonale de v sur E. On a $df(v) \geq -\|\nabla f(x)\|_\infty |v_E|_1 - \alpha |u|_1$ avec $\alpha := \sup_{e_j \notin \text{Ext}} |df(e_j)|$. En particulier $\alpha < \|\nabla f(x)\|_\infty$, donc dès que $|u|_1 > 0$ on a $df(v) > -\|\nabla f(x)\|_\infty (|v_E|_1 + |u|_1) = -\|\nabla f(x)\|_\infty |v|_1$.
Ce qui donne l'inclusion $D \subset E$ et le résultat. □

illustration graphique de la norme

On se propose maintenant d'étudier la première méthode à notre disposition: choix du gradient comme direction de decente et choix du pas constant. Voici un premier théorème de convergence pour une méthode de gradient à pas constant qui impose que le pas soit suffisamment petit pour obtenir le résultat.

Théorème 7 (Convergence du gradient à pas fixe). *Soit $\Omega \subset \mathbb{R}^n$ un ouvert et $f : \Omega \mapsto \mathbb{R}$ une fonction vérifiant*

- *∇f est $L-$Lispchitz,*
- *f est bornée inférieurement,*
- *il existe $x_0 \in \Omega$ tel que $S_0 = \{x \in \Omega |\, f(x) \leq f(x_0)\}$ est fermé dans $\mathbb{R}^n$*

alors si $\varepsilon < \frac{2}{L}$ la suite $(x_k)_{k\in\mathbb{N}}$ définie par la donnée initiale x_0 et l'équation $x_{k+1} = x_k - \varepsilon \nabla f(x_k) \in \Omega$ vérifie

- *la suite $(f(x_k))_{k\in\mathbb{N}}$ est décroissante de limite finie,*
- *la suite $(\nabla f(x_k))_{k\in\mathbb{N}}$ converge vers 0.*

Preuve: Si pour un certain $k_0 \geq 0$, $\nabla f(x_{k_0}) = 0$ alors la suite x_k est stationnaire en x_{k_0} et le résultat annoncé est donc vérifié. Sinon, $-\nabla f(x_k)$ est toujours une direction de descente en x_k. La proposition 4 donne

$$f(x_{k+1}) \leq f(x_k) - \varepsilon \langle \nabla f(x_k), \nabla f(x_k) \rangle + \frac{\varepsilon^2 L}{2} \|\nabla f(x_k)\|^2 , \tag{3.8}$$

L'hypothèse $\varepsilon < \frac{2}{L}$ donne

$$f(x_{k+1}) - f(x_k) \leq -\varepsilon \|\nabla f(x_k)\|^2 (1 - \frac{\varepsilon L}{2}) \leq 0 . \tag{3.9}$$

Ceci est vrai à condition que $x_{k+1} \in \Omega$. Pour le montrer, on utilise un argument de connexité sur $I = \{t \in [0, \frac{2}{L}[\,|\, \forall s \leq t\,, x_k - s\nabla f(x_k) \in \Omega\}$: I est ouvert non vide (car Ω ouvert) et fermé (car contenu dans S_0 par l'inégalité précédente) dans $[0, \frac{2}{L}]$ donc $I = [0, \frac{2}{L}]$.
La suite $(f(x_k))_{k\in\mathbb{N}}$ est décroissante (strictement car le gradient en x_k est non nul) et converge donc dans $\overline{\mathbb{R}}$. Comme f est bornée inférieurement, cette limite est finie. Donc le terme de gauche de l'équation (3.9) converge vers 0, ce qui donne que la suite $(\nabla f(x_k))_{k\in\mathbb{N}}$ converge vers 0. □

Remarque 10.
- *Attention ce résultat ne donne en aucun cas l'existence d'un minimum local dans $\mathbb{R}^n$. En effet, la fonction $f : x \mapsto \frac{1}{1+x^2}$ satisfait les conditions énoncées et dans ce cas l'algorithme donne une convergence de x_k vers $-\infty$ ou $+\infty$ selon le signe de x_0.*
- *Si on rajoute une hypothèse de compacité pour $\{x \,|\, f(x) \leq f(x_0)\}$ alors l'existence d'un minimum global est évidemment garantie mais pas la convergence de la suite $(x_k)_{k\in\mathbb{N}}$ vers ce minimum.*
- *Si l'algorithme est initialisé avec un point critique de la fonction alors la suite $(x_k)_{k\in\mathbb{N}}$ est constante.*
- *Sous les hypothèses du théorème et en supposant de plus la convergence de la suite x_k, sa limite n'est pas nécessairement un minimum local de la fonction mais seulement un point critique. L'algorithme donne le minimum global sur $[0, -\infty[$ de la fonction $f : x \mapsto x^3$ mais ce n'est qu'un point critique de la même fonction définie sur $\mathbb{R}$ ou toute modification de la cette fonction sur $]-\infty, 1[$ par exemple, qui, de plus, satisferait les conditions du Théorème 7.*

Exercice 7. *Soit $f : \mathbb{R}_+ \mapsto \mathbb{R}$ définie par $f(x) = e^{-x}$. Montrer que f vérifie les hypothèses du Théorème 7 et déterminer la limite d'une suite $(x_k)_{k\in\mathbb{N}}$ définie comme dans le Théorème 7.*

En rajoutant des hypothèses telles que la coercivité de la fonction f, alors on peut caractériser la limite m de la suite $(f(x_k))_{k\in\mathbb{N}}$.

Corollaire 2. *Soit $\Omega \subset \mathbb{R}^n$ un ouvert et $f : \Omega \mapsto \mathbb{R}$, on note $S_k := \{x \in \Omega \,|\, f(x) \leq f(x_k)\}$. Sous les hypothèses du Théorème 7 et en supposant qu'il existe $k_0 \in \mathbb{N}$ tel que S_{k_0} est compact, alors la suite $(f(x_k))_{k\in\mathbb{N}}$ converge vers une valeur critique de la fonction f.*

Preuve: Par compacité de S_{k_0}, il existe une valeur d'adhérence de la suite x_k notée a qui vérifie par continuité de f et ∇f, $f(a) = m$ et $\nabla f(a) = 0$. D'où le résultat. □

Remarque 11. *Rappelons que les hypothèses de la proposition 4 (sur lesquelles repose le Théorème 7) sont formulées de manière globale mais le résultat précédent reste vrai si l'hypothèse ∇f Lipschitz est vérifiée uniquement sur S_{k_0} pour un certain $k_0 \in \mathbb{N}$.*

Les hypothèses utilisées dans le corollaire 2 sont trop faibles pour garantir la convergence de la suite x_k vers un point critique de la fonction f. Un exemple de situation de non-convergence est donné dans l'exercice ci-dessous.

Exercice 8. *Soit $\mathcal{C}$ la courbe dans le plan complexe, définie en coordonnés polaires par $\rho = a + e^{\theta}$ et le potentiel définie sur cette courbe par $V(\theta) = e^{\theta}$ où a est un réel strictement positif. Construire une fonction $f : \mathbb{C} \mapsto \mathbb{R}$ qui étend V et dont le gradient est porté par la tangente à $\mathcal{C}$. Montrer que la suite x_k du Théorème 7 initialisée par $x_0 \in \mathcal{C}$ admet pour valeurs d'adhérence le cercle de rayon a.*

En rajoutant une hypothèse de second ordre sur f, on peut obtenir la convergence vers un point critique (mais pas un minimum local):

Corollaire 3. *Sous les hypothèses du corollaire 2 avec de plus f C^2 et en supposant que tout point critique de f est non dégénéré i.e.*

$$\nabla f(x) = 0 \Rightarrow \mathbf{Jac}(\nabla^2 f)(x) \neq 0\,, \tag{3.10}$$

alors la suite $(x_k)_{k\in\mathbb{N}}$ converge vers un point critique de la fonction f.

Preuve: L'hypothèse donnée par (3.10) implique que les points critiques sont isolés dans S_{k_0}. En effet, soit x un point critique de f et y un point dans un voisinage de x_0 (choisi par la suite) on a

$$\nabla f(y) = \nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x + t(y-x))(y-x)\, dt$$
$$\|\nabla f(y)\| > \frac{1}{2}\|\nabla^2 f(x)(y-x)\|$$

Cette inégalité est vérifiée sur une boule ouverte centrée en x, $B(x, \varepsilon)$ en utilisant la continuité de la dérivée seconde de f en x. Ceci montre donc que $\nabla f(y)$ est non nul pour $y \in B(x, \varepsilon)$ dès que $y \neq x$. Les points d'adhérence de la suite x_k sont donc isolés. De plus, le théorème 7 donne la convergence de $\|x_{k+1} - x_k\|$ vers 0, ce qui implique la convergence de la suite x_k (le montrer en exercice). □

Remarque 12. *Attention, le corollaire 3 n'assure pas la convergence de l'algorithme vers un minimum local de la fonction. Le théorème suivant rajoute une hypothèse de convexité locale et permet souvent d'obtenir des minimums locaux.*

Théorème 8. *Sous les notations du théorème 7, on considère $f : \Omega \mapsto \mathbb{R}$ une fonction de classe D^2 bornée inférieurement sur Ω telle que*

- *il existe une constante $L > 0$ telle que $0 \leq \nabla^2 f \leq L\, Id$ sur l'ensemble $S_0 := \{x \in \Omega \,|\, f(x) \leq f(x_0)\}$ fermé dans $\mathbb{R}^n$,*
- *$\varepsilon < \frac{2}{L}$*

alors on a l'alternative suivante:

- *soit la suite $(x_k)_{k\in\mathbb{N}}$ définie au Théorème 7 converge vers un minimum local de f (sauf si $x_k \equiv x_0$ et dans ce cas x_0 est seulement un point critique de f),*
- *soit $\lim_{k\to+\infty} \|x_k\| = +\infty$.*

Preuve: Si on ne se trouve pas dans le second cas, cela implique qu'il existe une sous-suite convergente de x_k vers une limite notée x_∞. On peut supposer que $f(x_\infty) < f(x_0)$ sinon le résultat est immédiat car dans ce cas, $x_k \equiv x_0$. Par continuité, il existe une boule $B(x_\infty, \varepsilon) \subset S_0$. De plus, x_∞ est un point fixe de l'application $F : x \mapsto x - \varepsilon \nabla f(x)$ pour laquelle on a $\|F'(x)\| = \|1 - \varepsilon\nabla^2 f(x)\| \leq 1$. Cela implique que

$$\|F(x_\infty) - F(y)\| \leq \|x_\infty - y\| \tag{3.11}$$

i.e. $F\left(B(x_\infty, s)\right) \subset B(x_\infty, s)$ pour $s \leq \alpha$. On a montré que $d(x_\infty, x_k)$ est une suite décroissante dont on sait que la limite est 0, ce qui est le résultat annoncé.
Le point x_∞ est un point critique pour f et sur un voisinage ouvert de x_∞ la fonction f est convexe. Par convexité, x_∞ est un minimum pour f sur ce voisinage. C'est donc un minimum local pour f. □

On s'intéresse maintenant à la vitesse de convergence de l'algorithme dans un cas où la convergence est assurée par le théorème suivant:

Théorème 9. *Soit $\Omega \subset \mathbb{R}^n$ un ouvert et $f : \Omega \mapsto \mathbb{R}$ une fonction de classe D^2 bornée inférieurement telle que*

- *il existe deux constantes $l, L > 0$ telle que $l\, Id \leq \nabla^2 f \leq L\, Id$ sur l'ensemble $S_0 := \{x \in \Omega \,|\, f(x) \leq f(x_0)\}$ fermé dans $\mathbb{R}^n$,*
- *$\varepsilon < \frac{2}{L}$*

alors

- *la suite $(x_k)_{k\in\mathbb{N}}$ définie au Théorème 7 converge vers un minimum local strict de f noté x^*,*
- $\|x_{k+1} - x^*\| \leq m(\varepsilon)\|x_k - x^*\|$ *avec* $m(\varepsilon) = \max(|1 - l\varepsilon|, |1 - L\varepsilon|)$.

Preuve: Soit $F(x) = x - \varepsilon\nabla f(x)$, l'hypothèse donne donc $\sup_{x\in S_0} \|F'(x)\| \leq \max(|1 - l\varepsilon|, |1 - L\varepsilon|)$ (attention à bien voir pourquoi: on peut par exemple diagonaliser en base orthonormée) et comme $[x_k, x_{k+1}] \subset S_0$ on obtient

$$\|x_k - x_{k+1}\| = \|F(x_{k-1}) - F(x_k)\| \leq m(\varepsilon)\|x_{k-1} - x_k\|\,. \tag{3.12}$$

Cette inégalité assure la convergence de la suite $(x_k)_{k\in\mathbb{N}}$. Par le Théorème 7, la limite x^* est un point critique de f. En utilisant le lemme 4, on montre que ce point critique est un minimum local strict de f. Il existe alors une boule ouverte centrée en x_∞ contenue dans S_0, on peut donc appliquer l'inégalité des accroissements finis pour k assez grand:

$$\|x^* - x_{k+1}\| \leq m(\varepsilon)\|x^* - x_k\|\,, \tag{3.13}$$

ce qui prouve le second point. □

Corollaire 4. *Pour $\varepsilon = 2/(l+L)$, on obtient $\|x^* - x_0\| \leq \frac{1}{1+m(\varepsilon)}\|x_1 - x_0\| \leq \frac{\|\nabla f(x_0)\|}{c}$*

Exercice 9. *Donner une preuve du corollaire.*

Si on a uniquement à disposition une estimation de $\nabla^2 f(x^*)$, on peut quand même évaluer la vitesse convergence de la suite $(x_k)_{k\in\mathbb{N}}$ (si celle-ci converge vers x^* évidemment):

Proposition 12. *Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ telle que la suite $(x_k)_{k\in\mathbb{N}}$ soit bien définie et converge vers x^* un minimum local de f. On suppose que f est de classe D^2 au point x^*, $\nabla^2 f(x^*)$ est définie positive et $\|Id - \varepsilon\nabla^2 f(x^*)\| < 1$. Alors le taux de convergence asymptotique r vérifie $r \leq \|Id - \varepsilon\nabla^2 f(x^*)\|$.*

Exercice 10. *Donner une preuve de la proposition précédente.*

Le problème de la méthode de gradient est qu'en pratique la convergence peut être très lente si le problème est mal conditionné. Par exemple, la minimisation de

$$f(x, y) = x^2 + 10y^2 \tag{3.14}$$

exhibe une lente convergence comme indiqué dans la figure 3.1.

On peut étudier ce problème dans un cas simple:

$$f(x) = \sum_{i=1}^{n} \frac{1}{2}\lambda_i x_i^2 - b_i x_i\,, \tag{3.15}$$

où λ_i sont des réels strictement positifs. Le cas général d'une matrice définie positive A peut se réduire à ce cas. La descente de gradient à pas constant itère l'application linéaire $x \to Id - \varepsilon A$. On cherche donc à maximiser le taux de contraction de cette application donné par $\max(|1-\varepsilon\lambda|, |1-\varepsilon\Lambda|)$ avec $\lambda = \min_{i=1,\ldots,n} \lambda_i$ et $\Lambda = \max_{i=1,\ldots,n} \lambda_i$. On doit donc minimiser le taux de contraction en fonction du pas ε et on obtient $\varepsilon_{opt} = \frac{2}{\lambda+\Lambda}$, et on obtient la constante de contraction $\frac{\Lambda-\lambda}{\Lambda+\lambda}$.

3.2.2 Changement de produit scalaire

On s'intéresse maintenant à l'effet d'un changement de produit scalaire sur $D(x)$.

Proposition 13. *Soit q une forme quadratique définie positive sur $\mathbb{R}^n$ et Q sa matrice associée $Q := (q(e_i, e_j))$ dans la base canonique. Alors,*

$$\nabla_Q f(x) = Q^{-1}(\nabla f(x)) \tag{3.16}$$

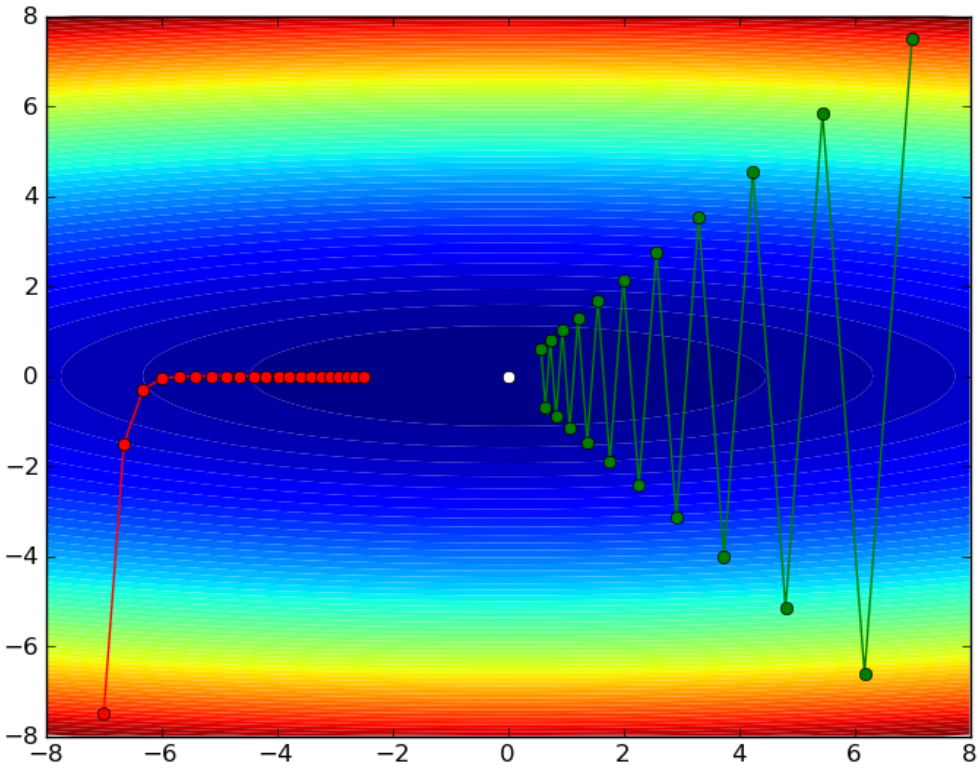


Figure 3.1: Un effet de zig-zag et une convergence lente.

Preuve: Par définition, $\langle \nabla f(x), u \rangle = q(\nabla_Q f(x), u) = \langle Q(\nabla_Q f(x)), u \rangle$. On en déduit que quel que soit $u \in \mathbb{R}^n$, $\langle \nabla f(x) - Q(\nabla_Q f(x)), u \rangle = 0$ et donc $\nabla f(x) = Q(\nabla_Q f(x))$, d'où le résultat. □

Si on choisit la direction de descente définie par $-\nabla f(x)$, alors il faut retenir que changer le produit scalaire change le gradient et en pratique le résultat de la descente de gradient et la vitesse de convergence peuvent être bien différents.

```
illustration numerique
```

Définition 13. *Un algorithme du type 1 est dit robuste si la direction d calculée à chaque itération est une direction de descente.*

3.2.3 Changement d'échelle et changement de variable

Un changement de produit scalaire peut s'interpréter comme un changement de variable global. En effet, soit M l'application linéaire définie par $M(e_i) = \lambda_i e_i$ dans la base canonique de $\mathbb{R}^n$ alors M avec $\lambda_i > 0$. On peut s'intéresser à la minimisation de la fonction $f \circ M$ pour laquelle le gradient s'écrit $M(\nabla f(x))$. On voit donc que le changement de produit scalaire est équivalent à un changement de variable pour le calcul de gradient, *i.e.*

$$\nabla(f \circ M)(x) = \nabla_{M^{-1}} f(M(x)) \,. \tag{3.17}$$

Attention, dans l'équation précédente, le gradient de $f \circ M$ est évalué au point x tandis que le terme de droite est évalué en $M(x)$.

On peut donc penser que le problème de minimisation initial sera plus aisé à résoudre numériquement par un changement de variable adéquat. On a plus généralement:

Proposition 14. *Si ϕ est un difféomorphisme de $\Omega \subset \mathbb{R}^n$ ouvert et $f : \Omega \mapsto \mathbb{R}^n$,*

$$\nabla f \circ \phi = d\phi^*(\nabla f) \tag{3.18}$$

où $d\phi^*(\nabla f)(x) = T\phi^*_{\phi(x)}(\nabla f(\phi(x)))$.

3.2.4 Méthode de Newton

La méthode de Newton correspond à une méthode de descente pour laquelle le choix de la direction est donnée par le gradient pour la métrique définie par la Hessienne:

$$d = -\nabla_{\nabla^2 f} f(x) = -\left(\nabla^2 f\right)^{-1} \nabla f(x)\,. \tag{3.19}$$

Ce choix, permis lorsque la Hessienne est inversible, définit un algorithme robuste (voir définition 13) lorsque $\nabla^2 f$ est une matrice symétrique définie positive. Ce choix de pas utilise l'information de second ordre donnée par la Hessienne: en utilisant un développement limité de f au second ordre, on obtient

$$f(x+v) \simeq f(x) + \langle \nabla f(x), v\rangle + \frac{1}{2}\langle v, \nabla^2 f(x) v\rangle\,, \tag{3.20}$$

Une idée naturelle est de choisir comme direction de descente celle qui minimise cette approximation. En différenciant par rapport à v, on obtient:

$$\nabla f(x) + \nabla^2 f(x) v = 0\,, \tag{3.21}$$

et donc

$$v = -\left(\nabla^2 f\right)^{-1} \nabla f(x)\,.$$

En fait, cette quantité contient non seulement la direction de descente optimale mais aussi le pas optimal à choisir pour minimiser le développement limité au second ordre.

3.3 Choix du pas

Le choix du pas optimal (aussi appelé line search) fait référence aux méthodes d'optimisation de fonctions d'une seule variable réelle puisqu'il s'agit de choisir ε tel que $f(x+\varepsilon d) < f(x)$, la direction de descente d étant préalablement fixée. Ce problème de choix du pas recouvre deux problèmes que l'on peut distinguer de manière artificielle:

- optimisation unidimensionnelle ou choix du pas optimal,
- choix d'un pas approché satisfaisant.

En effet, dans le cas de l'optimisation multidimensionnelle, il peut être inutilement coûteux de chercher le minimum de la fonction le long de la direction de descente, sachant que cette direction n'est pas nécessairement celle sur laquelle se trouve le minimum global de la fonction à minimiser. Par exemple, lorsque le coût d'évaluation de la fonction est important, le pas choisi sera souvent non optimal. En fait, on fait face à un compromis entre réduction de la valeur de la fonction restreinte à la droite de recherche et temps de calcul alloué à cette estimation. Si le pas est mal choisi, l'algorithme peut alterner entre deux valeurs comme le montre l'exemple de la figure 3.2 et ne pas converger vers un minimum local.

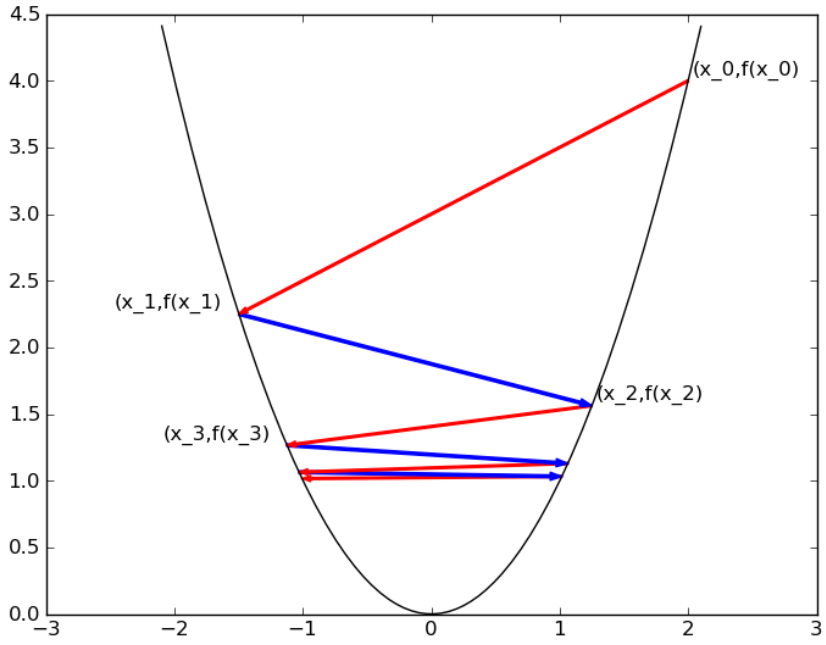


Figure 3.2: Dans cet exemple, $x_{k+1} = x_k + (-1)^{i+1} * (2.0 + 3.0/(2^{i+1})$ et $x_0 = 2$ Le pas fait décroître la fonction objectif $f(x) = x^2$ et la suite ne converge pas vers le minimum 0.

3.3.1 Méthodes de recherche du pas optimal

Si on dispose seulement de l'évaluation numérique de la fonction, alors on peut utiliser une méthode de dichotomie, dont le principe repose sur une méthode de réduction de l'intervalle qui contient le minimum local recherché. On appelle cet intervalle, intervalle de confiance. On se restreint au cas où la fonction est définie sur $\mathbb{R}_+$ car c'est bien le cas qui nous concerne lorsqu'on a déterminé une direction de descente. L'idée essentielle de ce type d'algorithme est de

1. trouver un triplet de points qui contient un minimum (local) du type $a < b < c$ tel que $f(b) \leq \min(f(a), f(b))$
2. mettre à jour le triplet pour réduire l'intervalle $[a, c]$ de façon à faire converger la procédure.

Initialisation de l'intervalle$[a = 0, b]$ pour $b \in \mathbb{R}_+^*$.
Tant que $(f(a) \leq f(b))$ **faire**
  $b \leftarrow \frac{b}{2}$
Fin
Initialisation de $c \leftarrow 2b$
Tant que $(f(b) > f(c))$ **faire**
  $a \leftarrow b$,
  $b \leftarrow c$,
  $c \leftarrow 2c$
Fin
Tant que $(c - a >$ tolérance$)$ **faire**
  Application d'une méthode de réduction du triplet a, b, c.
Fin
Retourner b

Algorithme 2: Méthode de dichotomie

Remarque 13.
- *La première partie de l'algorithme se termine si on a choisi une direction de descente.*
- *La seconde partie se termine si on dispose de plus d'une condition de coercivité sur f. Si ce n'est pas le cas, on peut borner la recherche de b et faire terminer l'algorithme.*
- *La méthode de réduction du triplet se fait par dichotomie ou par interpolation.*

Plusieurs méthodes de réduction d'intervalle sont utilisées en pratique. En général, ces méthodes s'appliquent au cas des fonctions unimodales:

Définition 14. *Une fonction $f : [0, T] \mapsto \mathbb{R}$ est dite unimodale sur $[0, T]$ s'il existe un minimum strict $t^* \in]0, T[$ tel que f est strictement décroissante sur $[0, t^*]$ et strictement croissante sur $[t^*, T]$.*

Remarque 14. *Toute restriction d'une fonction unimodale à un intervalle non vide qui contient le minimum est aussi unimodale.*

La méthode de la section dorée est la suivante:

$[\alpha := \frac{1}{2}(\sqrt{5}-1)]$
Initialisation d'un intervalle $[a, b]$ contenant un minimum local.
$c \leftarrow \alpha a + (1-\alpha)b$ et $v_c = f(c)$
$d \leftarrow a + b - c$ et $v_c = f(d)$
Tant que $(b - a >$ tolérance$)$ **faire**
  Évaluer $v_c \leftarrow f(c)$ ou $v_d \leftarrow f(d)$ si nécessaire
  Si $(v_c < v_d)$ **Alors**
    $b \leftarrow d$
    $d \leftarrow c$ et $v_d \leftarrow v_c$
    $r \leftarrow c$
    $c \leftarrow a + b - d$
  Sinon
    $a \leftarrow c$
    $c \leftarrow d$ et $v_c \leftarrow v_d$
    $r \leftarrow d$
    $c \leftarrow a + b - c$
  Fin Si
Fin
Retourner r

Algorithme 3: Méthode de la section dorée

A chaque itération, la longueur de l'intervalle $[a, b]$ est multipliée par l'inverse du nombre d'or $\alpha := \frac{1}{2}(\sqrt{5}-1) \simeq 0.618$. En particulier, la vitesse de convergence de cette méthode est linéaire. Pourquoi choisit-on le paramètre α défini par le ratio précédent? A priori, on pourrait fixer c et d librement, mais on cherche à

- minimiser le nombre d'évaluations de la fonction,
- réduire l'intervalle $[a, b]$ avec un taux constant.

La seconde condition implique que le choix de c, d se réduit au choix de α et la première condition implique que α est solution de $\alpha^2 = 1 - \alpha$, d'où la valeur de α. Lorsque le nombre d'évaluations de la fonction est fixé préalablement, alors la méthode optimale de localisation du minimum local est une variante de la méthode de la section dorée: la méthode de Fibonacci. Ces méthodes de section permettent de localiser le minimimum des fonctions unimodales:

Théorème 10. *Soit f unimodale sur l'intervalle $[a_0, b_0]$, l'algorithme de la section dorée 3 retourne une estimation c du minimum t^* tel que $|t^* - c| <$ tolérance.*

Preuve: Il suffit de vérifier qu'à chaque étape, f est unimodale sur l'intervalle courant $[a, b]$. Si $f(c) < f(d)$ alors l'intervalle courant devient $[a, d]$, or f n'est pas décroissante sur $[c, d]$ et est décroissante sur $[a, x^*]$ (avec $x^* = \text{argmin}_{[a,b]} f$). On en déduit $x^* \in [a, d]$. Le raisonnement est le même pour le cas $f(c) \geq f(d)$. On en déduit que $x^* \in [a, b]$ à chaque étape et donc que f est unimodale sur $[a, b]$ d'après la remarque 14. Comme $|b_n - a_n| < \alpha^n |b_0 - a_0|$ (où l'indice n représente le nombre d'itérations déjà effectuées), l'algorithme se termine en un nombre fini d'itérations et le résultat vérifie donc $|c - x^*| <$ tolérance. □

En pratique, on utilise cet algorithme sur des fonctions qui ne sont pas régulières ou bien des fonctions pour lesquelles l'évaluation de la fonction et/ou du gradient est coûteuse. Si la fonction n'est pas unimodale, on ne peut pas appliquer le théorème précédent et l'algorithme ne converge pas nécessairement vers un minimum local.

Une autre approche pour la réduction de l'intervalle consiste à interpoler la fonction entre les valeurs connues par un polynôme pour lequel on peut résoudre explicitement

le problème de minimisation. Ces méthodes vont être efficaces si la fonction à minimiser est suffisamment régulière. Prenons par exemple le cas d'une interpolation quadratique, on rappelle que $P \in \mathbb{R}_2[X] \mapsto (P(a), P(b), P(c)) \in \mathbb{R}^3$ est une application linéaire inversible donc il existe un unique polynôme de degré 2 interpolant la fonction f aux points a, b, c.

Exercice 11. *Montrer que le minimum du polynôme interpolant est atteint en*

$$c^* = \frac{1}{2}(a+c) + \frac{1}{2}\frac{(f(a)-f(c))(c-b)(b-a)}{(c-b)f(a)+(b-a)f(c)+(a-c)f(b)} \tag{3.22}$$

On obtient l'algorithme suivant:

Initialisation de $a, b, c \in \mathbb{R}$ tels que $f(c) < f(a) < f(b)$
Tant que ($|c - c^*| >$ tolérance) **faire**
  Calcul de c^*
  Mettre à jour le triplet (a, b, c) en incluant c^*.
Fin
Retourner c^*

Algorithme 4: Méthode quadratique

On admet le théorème suivant:

Théorème 11. *Si $f : \mathbb{R} \mapsto \mathbb{R}$ est C^4 et un unique minimum x^* tel que $f''(x^*) > 0$ alors l'algorithme quadratique converge avec une vitesse superlinéaire:*

$$\limsup_{n\to\infty} \frac{\|c_{k+1} - x^*\|}{\|c_k - x^*\|^{1.32}} < \infty\,. \tag{3.23}$$

Preuve: Ce théorème est admis. □

Si on a aussi accès à la dérivée de f à chaque évaluation de la fonction f alors on peut envisager une interpolation par spline cubique. On peut ainsi obtenir une meilleure approximation de f quand f est suffisamment régulière. On s'attend alors à obtenir de meilleur taux de convergence pour ces fonctions. Plus précisément, l'interpolation cubique est définie par l'inverse de l'application linéaire suivante $P \in \mathbb{R}_3[X] \mapsto (P(a), P'(a), P(b), P'(b)) \in \mathbb{R}^4$ qui est un isomorphisme dès que $a \neq b$. Lorsqu'on utilise un algorithme de mise à jour d'un intervalle $[a, b]$ en minimisant l'interpolation cubique, l'ordre de convergence de l'algorithme sous des hypothèses adéquates est 2.

Exercice 12. *Ecrire un algorithme de même type que l'algorithme 3 utilisant l'interpolation cubique (sans expliciter le calcul du minimum).*

3.3.2 Recherche inexacte

Plutôt que de faire une recherche exacte du minimum sur la fonction unidimensionnelle $t \mapsto f(x + td)$, on peut se contenter d'une diminution suffisante de la valeur de la fonction f. On a déjà mentionné en introduction que cette précision est inutilement coûteuse lorsqu'on cherche à résoudre le problème global. Un autre argument est que pour des méthodes de type Newton ou quasi-Newton, le taux de convergence ne dépend pas en général de l'optimisation unidimensionnelle. La **règle d'Armijo** requiert une diminution suffisante de la fonction en imposant de choisir t tel que

$$f(x + td) \leq f(x) + \rho t \langle \nabla f(x), d \rangle\,, \tag{3.24}$$

pour un réel $\rho \in]0, \frac{1}{2}[$. Si $\langle \nabla f(x), d\rangle < 0$ alors cette condition est vérifiée pour t suffisamment petit. Par exemple, il suffit d'initialiser $\beta \in]0,1[$ et de retenir le plus petit entier n tel que $\beta^n t$ satisfait l'inégalité (3.24). La règle d'Armijo peut conduire à accepter des pas trop petits et elle n'évite donc pas de se retrouver dans une situation où l'algorithme converge vers une valeur qui n'est pas un minimum.
La **règle de Goldstein** permet d'éviter cette situation en imposant:

$$\begin{cases} f(x+td) \leq f(x) + \rho t \langle \nabla f(x), d\rangle \\ f(x+td) \geq f(x) + (1-\rho) t \langle \nabla f(x), d\rangle \,, \end{cases} \tag{3.25}$$

pour $\rho < \frac{1}{2}$. La condition $\rho < \frac{1}{2}$ est nécessaire pour ne pas exclure de la recherche le minimum dans certains cas, comme le montre l'exercice suivant:

Exercice 13. *Soit f une fonction polynômiale de degré 2 telle que $f'(0) < 0$ et $f''(0) > 0$, montrer que le minimum est atteint en x^* pour lequel:*

$$f(x^*) = f(0) + \frac{1}{2} x^* f'(0) \,. \tag{3.26}$$

Lorsqu'on a accès à la dérivée de f, on peut aussi utiliser la **règle de Wolfe**, qui impose pour $0 < \alpha < \beta < 1$

$$\begin{cases} f(x+td) \leq f(x) + \alpha t \langle \nabla f(x), d\rangle \\ \langle \nabla f(x+td), d\rangle \geq \beta \langle \nabla f(x), d\rangle \,. \end{cases} \tag{3.27}$$

Attention, il faut se rappeler que $\langle \nabla f(x), d\rangle < 0$ pour les inégalités précédentes.

Théorème 12. *Soit α, β deux réels tels que $0 < \alpha < \beta < 1$ et f une fonction de classe $C^1(\mathbb{R})$. On suppose que $f'(0) < 0$ et qu'il existe $t_0 > 0$ tel que*

$$f(t_0) \geq f(0) + \alpha t_0 f'(0) \,, \tag{3.28}$$

*alors il existe un intervalle non vide $I \subset]0, t_0[$ tel que tout $t \in I$ satisfait les inégalités (5) définissant la **règle de Wolfe**.*

Preuve: Soit $t_1 := \inf \{t \in]0, t_0[\mid f(t) \geq f(0) + \alpha t f'(0)\}$. Comme $f'(0) < 0$, on a $t_1 > 0$ et par continuité, $f(t_1) = f(0) + \alpha t_1 f'(0)$. Par le théorème des accroissements finis, on a l'existence de $t_2 \in]0, t_1[$ tel que $f'(t_2) = \alpha f'(0) > \beta f'(0)$. Par continuité de f', il existe $\varepsilon > 0$ tel que tout $t \in]t_2 - \varepsilon, t_2 + \varepsilon[$ satisfait la seconde inégalité de la règle de Wolfe. La première est satisfaite dès que $]t_2 - \varepsilon, t_2 + \varepsilon[\subset]0, t_1[$, donc pour ε suffisamment petit. □

A priori, on serait plutôt incité à utiliser β proche de 0 pour obtenir une approximation plus exacte du minimum, mais cela requiert aussi plus d'évaluations de la fonction. En pratique, des valeurs telles que $\alpha = 0.1$ et $\beta = 0.4$ donnent un compromis satisfaisant mais ce choix est très dépendant de la méthode utilisée pour le choix de la direction.

On peut donc proposer l'algorithme basé sur une méthode de dichotomie. On considère une fonction f à minimiser pour laquelle d est une direction de descente au point x.

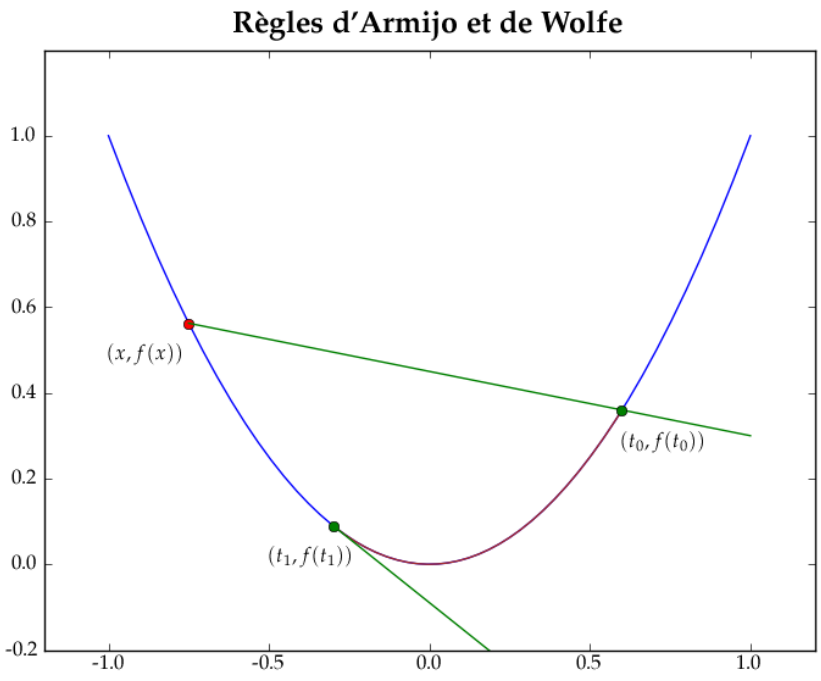


Figure 3.3: Le point x représente le point courant auquel les deux méthodes sont appliquées. Le point t_0 est choisi tel que l'intervalle $]x, t_0[$ soit le plus grand intervalle dont l'adhérence contient x qui satisfait la condition d'Armijo pour $\alpha = 0.1$. L'intervalle $]t_1, t_0[$ satisfait la règle de Wolfe avec $\beta = 0.4$

Initialisation de $\varepsilon = 1$, $a = 0$, $b = \infty$.
Tant que (Règle de Wolfe non satisfaite) **faire**
  Si $(f(x + \varepsilon d) > f(x) + \alpha\varepsilon\langle\nabla f(x), d\rangle)$ **Alors**
    $b = \varepsilon$
    $\varepsilon \leftarrow \frac{a+b}{2}$
  Sinon
    Si $(\langle\nabla f(x + \varepsilon d), d\rangle < \beta\langle\nabla f(x), d\rangle)$ **Alors**
      $a = \varepsilon$
      $\varepsilon \leftarrow \frac{a+b}{2}$ si $b < \infty$ et $\varepsilon \leftarrow 2a$ sinon.
    Sinon
      Retourner ε
    Fin Si
  Fin Si
Fin

Algorithme 5: Exemple d'algorithme pour satisfaire la règle de Wolfe

On peut montrer la proposition suivante pour cet algorithme:

Proposition 15. *Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ une fonction C^1 pour laquelle d vérifie $\langle\nabla f(x), d\rangle < 0$. On a l'alternative suivante: soit l'algorithme termine, soit la fonction n'est pas bornée inférieurement.*

Preuve:

Si b reste constant égal à ∞, alors la suite des itérés de ε est 2^k pour lesquels $f(x+2^k d) \leq f(x) + 2^k \alpha \langle \nabla f(x), d\rangle$ de limite $-\infty$.

Sinon, b est initialisé à une valeur finie à une certaine itération. On voit alors facilement que la suite $[a,b]$ est une suite de segments emboîtés telle que la longueur du segment est divisée de moitié à chaque itération (dès que b a une valeur finie). En particulier, la suite des ε converge vers une valeur finie telle que

$$\langle \nabla f(x+\varepsilon d), d\rangle \leq \beta \langle \nabla f(x), d\rangle \,. \tag{3.29}$$

De plus, on a $f(x+ad) \leq f(x) + a\alpha\langle \nabla f(x), d\rangle$ et d'autre part $f(x+bd) \geq f(x) + b\alpha\langle \nabla f(x), d\rangle$ (attention à bien le comprendre), ce qui implique

$$\begin{aligned} f(x+bd) - f(x+ad) &\geq (b-a)\alpha\langle \nabla f(x), d\rangle \\ \frac{f(x+bd)-f(x+ad)}{b-a} &\geq \alpha\langle \nabla f(x), d\rangle \,. \end{aligned}$$

En passant à la limite sur a et b, on a

$$\langle \nabla f(x+\varepsilon d), d\rangle \geq \alpha \langle \nabla f(x), d\rangle \,. \tag{3.30}$$

On obtient donc une contradiction avec l'inégalité (3.29).

□

On peut proposer un algorithme en général plus efficace que le précédent en utilisant une interpolation quadratique pour l'estimation de ε dans la première condition, par exemple. L'implémentation d'une méthode efficace de linesearch est délicate en pratique et il est recommandé d'utiliser les implémentations disponibles.

3.4 Convergence des méthodes de descente

Jusqu'à présent, on a étudié les propriétés de convergence de la méthode de descente à pas constant. On se propose maintenant d'étudier une méthode de descente incluant une recherche du pas optimal. On utilisera la définition suivante:

Notation 1. *On notera θ l'angle (non orienté) formé par le vecteur d et $-\nabla f(x)$. Il est caractérisé par*

$$\cos\theta = \frac{-\langle \nabla f(x), d\rangle}{\|\nabla f(x)\|\|d\|} \,. \tag{3.31}$$

On a le théorème de Zoutendijk suivant:

Théorème 13. *Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ une fonction C^1 telle que ∇f soit Lipschitz de constante $L > 0$. On suppose que f est bornée inférieurement. Si on considère l'algorithme 1 pour lequel le choix du pas se fait par la règle de Wolfe (on considère $0 < \alpha < \beta$) alors*

$$\sum_{i=0}^{\infty} \cos^2\theta_i \|\nabla f(x_i)\|^2 < \infty \,. \tag{3.32}$$

Remarque 15. *Comme f est bornée inférieurement, l'algorithme 5 termine toujours à condition qu'on de trouver une direction de descente pour laquelle $\langle \nabla f(x), d\rangle < 0$ ce qui est possible si $\nabla f(x) \neq 0$.*
Dans le cas où $\nabla f(x) = 0$, alors la condition nécessaire d'optimalité du premier ordre est satisfaite et on peut faire terminer l'algorithme. Dans ce cas, il serait même préférable d'utiliser une condition de second ordre pour décider de terminer l'algorithme de descente.

Preuve: On note $x_{i+1} = x_i + \varepsilon_i d_i$ et $\nabla f(x_i) = g_i$. On a par la condition de Wolfe:

$$\langle g_{i+1}, d_i \rangle \geq \beta \langle g_i, d_i \rangle \,. \tag{3.33}$$

En soustrayant à cette égalité $\langle g_i, d_i \rangle$, on obtient

$$\langle g_{i+1} - g_i, d_i \rangle \geq (\beta - 1) \langle g_i, d_i \rangle \,, \tag{3.34}$$

(qui est strictement positif si d_i est une direction de descente telle que $\langle g_i, d_i \rangle < 0$). Par l'hypothèse ∇f Lipschitz, on obtient:

$$\langle g_{i+1} - g_i, d_i \rangle \leq \|d_i\|^2 L \varepsilon_i \,, \tag{3.35}$$

et donc on a une minoration de ε_i:

$$\varepsilon_i \geq \frac{1-\beta}{L\|d_i\|^2} (-\langle g_i, d_i \rangle) \,. \tag{3.36}$$

La première partie de la condition de Wolfe qui assure une décroissance "minimale" implique

$$f(x_{i+1}) - f(x_i) \leq \varepsilon_i \alpha \langle g_i, d_i \rangle \leq -\alpha |\langle g_i, d_i \rangle|^2 \frac{1-\beta}{L\|d_i\|^2} \,. \tag{3.37}$$

On obtient donc en sommant sur i:

$$\sum_{i=0}^{k} \alpha |\langle g_i, d_i \rangle|^2 \frac{1-\beta}{L\|d_i\|^2} \leq f(x_0) - f(x_k) \,. \tag{3.38}$$

Ceci se réécrit, pour $k = \infty$

$$\sum_{i=0}^{\infty} \frac{\alpha(1-\beta)}{L} \frac{|\langle g_i, d_i \rangle|^2}{\|d_i\|^2} \leq f(x_0) - \inf f < \infty \,, \tag{3.39}$$

car f est bornée inférieurement, ce qui termine la preuve. □

On voit par exemple que si on quel que soit i, $\cos(\theta_i) \geq \delta > 0$, alors la suite $\|\nabla f(x_i)\|^2$ tend vers 0. En pratique, d'autres cas d'applications de la formule (3.32) peuvent aussi être utiles. On retrouve donc le résultat de la convergence du gradient à pas constant en utilisant seulement la seconde partie de la preuve, et on a aussi la convergence d'une méthode de gradient à pas exact (i.e. minimiseur de la fonction $\varepsilon \mapsto f(x_k + \varepsilon d_k)$). On s'intéresse maintenant à l'ordre de convergence de la méthode de gradient à pas optimal. On a vu que l'ordre de convergence du gradient à pas constant est linéaire dans le cas favorable d'une fonction elliptique. On admet le résultat suivant qui dit qu'un pas optimal ne fait pas mieux en général:

Proposition 16. *Soit Q une forme quadratique définie positive sur $\mathbb{R}^n$ et $b \in \mathbb{R}^n$, on considère la minimisation de*

$$\frac{1}{2}\langle x, Qx \rangle + \langle b, x \rangle \,. \tag{3.40}$$

La suite x_k définie par la méthode de descente de gradient simple avec un pas optimal satisfait à:

$$\|x_{k+1} - x^*\|_Q^2 \leq \left(\frac{\Lambda - \lambda}{\Lambda + \lambda} \right)^2 \|x_k - x^*\|_Q^2 \,. \tag{3.41}$$

Même si cette proposition donne seulement une borne sur la convergence de la méthode, c'est cependant un bon indicateur sur la performance de l'algorithme. De plus, on remarque que $\frac{1}{2}\|x - x^*\|_Q^2 = f(x) - f(x^*)$ ce qui donne donc un autre indicateur de convergence. Cela montre que même dans le meilleur des cas (linesearch exact) la vitesse de convergence est mauvaise si la matrice admet un rapport Λ/λ trop grand. En pratique, pour des systèmes en grande dimension, on ne peut pas espérer un bon conditionnement de la matrice Q et on fait donc face à une convergence très lente. Par exemple, pour obtenir un objectif de 0.01 dans le cas où $\Lambda/\lambda = 500$, on s'attend à obtenir la précision demandée après environ 600 itérations.

Chapter 4

Méthode du gradient conjugué

La méthode du gradient conjugué permet de résoudre des systèmes linéaires en grandes dimensions. Elle a été introduite dans les années 1950 par Hestenes et Stiefel. Cette technique a ensuite été adaptée pour l'optimisation non linéaire. Dans ce cas, c'est une méthode qui moins coûteuse que les méthodes de Newton puisqu'elle n'utilise que le calcul du gradient de la fonction.

4.1 Méthode de gradient conjugué dans le cas linéaire

Si on s'intéresse à l'équation linéaire dans $\mathbb{R}^n$

$$Ax + b = 0 \tag{4.1}$$

avec A une matrice symétrique d'efinie positive alors la solution de ce probl'eme est la solution du problème de minimisation suivant:

$$\operatorname{argmin} \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle \,. \tag{4.2}$$

En effet la différentielle seconde de $\langle x, Ax\rangle$ pour A une application linéaire quelconque est $A^t + A$. Donc, si A est symétrique définie positive alors la fonction (4.4) est strictement convexe. On en déduit que le minimiseur de (4.4) noté x^* est unique et vérifie que le gradient en ce point est nul:

$$Ax^* + b = 0 \tag{4.3}$$

Etudions d'abord le cas simple où A est diagonale à coefficients positifs a_i dans la base canonique de $\mathbb{R}^n$. On peut toujours s'y ramener si on dispose d'une base $(p_i)_{i\in 1\ldots n}$ telle que $\langle p_i, Ap_j\rangle = 0$ pour $i \neq j$. On appelle une telle base orthogonale pour A, une famille de vecteurs **conjugués**. Dans ce cas le problème de minimisation est séparable dans cette base et il suffit de résoudre pour chaque i

$$\operatorname{argmin} \frac{1}{2}a_i x_i^2 + b_i x_i \,. \tag{4.4}$$

En particulier, si on utilise une optimisation unidimensionnelle successive sur chaque direction conjguée, l'optimum est alors atteint en n itérations au plus. Pour générer une base orthogonale pour le produit scalaire A, il suffit de diagonaliser la matrice A ou d'utiliser la méthode d'orthogonalisation de Gram-Schmidt. La méthode du gradient conjugué peut fournir une bonne approximation de la solution en peu d'itérations bien que sa précision soit faible en général. Lorsque $n \simeq 10^6$, ce type de méthodes itératives sont les seules disponibles.

Initialisation de $x \leftarrow x_0 \in \mathbb{R}^n$, $r_0 \leftarrow Ax_0 + b$,
$p_0 \leftarrow -r_0$, $k \leftarrow 0$
Tant que $(r_k \neq 0)$ **faire**
  $x_{k+1} \leftarrow x_k - \frac{\langle p_k, r_k \rangle}{\langle Ap_k, p_k \rangle} p_k$
  $r_{k+1} \leftarrow Ax_{k+1} + b$
  $p_{k+1} \leftarrow -r_{k+1} + \frac{\langle r_{k+1}, Ap_k \rangle}{\langle Ap_k, p_k \rangle} p_k$
  $k \leftarrow k + 1$
Fin
Retourner x_k

Algorithme 6: Méthode du gradient conjugué linéaire

Remarque 16. *On remarque que la condition d'arrêt est équivalente à* $\langle p_k, r_k \rangle = 0$*, pourquoi?*

4.2 Interprétation de la méthode du gradient conjugué

On propose une lecture de la méthode du gradient conjugué qui peut en faciliter sa compréhension. Voici deux points à retenir pour retrouver l'algorithme:

1. On remarque tout d'abord que la première étape est une simple descente de gradient à pas optimal. Ensuite, une fois qu'une direction de descente p_k est choisie, le pas optimal est utilisé.

   On rappelle que l'optimum d'une fonction quadratique (et strictement convexe) du type $t \mapsto f(x_k + tp_k)$ est atteint au point t_k tel que $\langle \nabla f(x_k + t_k p_k), p_k \rangle = 0$, c'est à dire tel que $t_k = -\frac{\langle Ax_k + b, p_k \rangle}{\langle Ap_k, p_k \rangle}$.

2. La nouvelle direction de descente candidate (avant conjugaison) à l'étape k est $-r_{k+1}$ avec le gradient noté r_{k+1}. La seule condition requise par l'algorithme est l'orthogonalisation des deux vecteurs r_{k+1} et p_k par rapport au produit scalaire A. C'est à dire, en notant p_{k+1} la nouvelle direction de descente $p_{k+1} = -r_{k+1} + \beta p_k$ telle que $\langle p_{k+1}, Ap_k \rangle = 0$, ce qui donne la valeur de $\beta = \frac{\langle r_{k+1}, Ap_k \rangle}{\langle p_k, Ap_k \rangle}$.

La propriété remarquable de cet algorithme est que la condition de conjugaison sur d_{k+1} et d_k suffit à assurer la conjugaison de d_{k+1} avec les autres directions de descente d_i pour $i < k$. C'est un résultat qui est donné par le théorème 14. Voici trois points à retenir pour retrouver la preuve de ce théorème:

1. Une première observation est que ce procédé génère uniquement des directions d_i dans le sous-espace vectoriel $F = \mathbf{Vect}\{A^j r_0 \,|\, j \leq i\}$. Si la conjugaison des directions est conservée, leur nombre est donc majoré par $\dim F$ et l'algorithme termine en au plus n itérations avec n la dimension de l'espace.

2. La seconde observation est que si on a trouvé l'optimum sur un sous espace F alors si on choisit le pas optimal pour une direction de descente d orthogonale à ce sous espace F pour le produit scalaire A, l'optimum obtenu est l'optimum global de la fonction sur $F + \mathbb{R}d$. Pour le voir, il suffit de décomposer sur une base orthogonale comme dans l'introduction.

3. La dernière observation est que cette propriété a une conséquence importante: $\langle r_{k+1}, d_i \rangle = 0$ pour $i \geq k$ en d'autres termes $r_{k+1} \in \mathbf{Vect}\{d_i \,|\, j \leq k\}^{\perp}$. Ceci implique la conjugaison de d_i et r_{k+1} pour $i < k$: en utilisant la première observation, on a $Ad_i \in \mathbf{Vect}\{d_i \,|\, j \leq k\}$ si $i < k$.

4.3 Étude théorique

On prouve ici rigoureusement les remarques précédentes.

Théorème 14. *L'algorithme 6 converge en, au plus, n itérations et on a les propriétés suivantes*

- $\mathbf{Vect}\{A^i(r_0) \mid i = 0 \dots k\} = \mathbf{Vect}\{r_i \mid i = 0 \dots k\} = \mathbf{Vect}\{p_i \mid i = 0 \dots k\}$
- $\langle p_{k+1}, Ap_i \rangle = 0$ *pour* $i \leq k$,
- $\langle r_{k+1}, p_i \rangle = 0$ *pour* $i \leq k$,

tant que l'algorithme n'a pas terminé.

Exercice 14. *On montre facilement, en utilisant le théorème, que* $\langle r_{k+1}, r_i \rangle = 0$ *pour* $i \leq k$, *ce qui est une propriété importante.*

Preuve: On montre par récurrence chaque propriété en commençant par la première: L'égalité $\mathbf{Vect}\{r_i \mid i = 0 \dots k\} = \mathbf{Vect}\{p_i \mid i = 0 \dots k\}$ est immédiate en utilisant $p_{k+1} = -r_{k+1} + \alpha_k p_k$. En utilisant l'égalité au rang k et la relation

$$r_{k+1} = r_k + \beta_k A p_k \tag{4.5}$$

on obtient

$$\mathbf{Vect}\{r_i \mid i = 0 \dots k+1\} \subset \mathbf{Vect}\{A^i(r_0) \mid i = 0 \dots k+1\}\,.$$

L'inclusion réciproque s'obtient en écrivant: $A^{k+1}(r_0) = A(\sum_{i=0}^{k} \mu_i p_i)$ et

$$Ap_k = \frac{r_{k+1} - r_k}{\beta_k} \tag{4.6}$$

donc

$$\mathbf{Vect}\{r_i \mid i = 0 \dots k+1\} = \mathbf{Vect}\{A^i(r_0) \mid i = 0 \dots k+1\}\,.$$

Ceci est licite si $\beta_k \neq 0$ par la remarque 16.

On a par construction $\langle p_{k+1}, Ap_k \rangle = 0$, de même $\langle r_{k+1}, p_k \rangle = 0$ et notamment pour $k = 0$. Supposons les propriétés vraies au rang k et montrons les au rang $k+1$ si l'algorithme ne s'est pas terminé. On a directement la dernière propriété:

$$\langle r_{k+1}, p_i \rangle = \langle r_k + \beta_k A p_k, p_i \rangle = 0\,, \tag{4.7}$$

pour $i \leq k-1$. On a aussi par construction de p_{k+1}

$$\langle p_{k+1}, Ap_i \rangle = \langle -r_{k+1} + \alpha_k p_k, Ap_i \rangle\,. \tag{4.8}$$

Or $A(p_i) \in \mathbf{Vect}\{p_j \mid j = 0 \dots k\}$ ceci pour $i \leq k-1$, on obtient donc, en utilisant l'égalité (4.7)

$$\langle p_{k+1}, Ap_i \rangle = 0\,. \tag{4.9}$$

On en déduit le résultat annoncé: puisque la dimension de l'espace est n, l'algorithme doit se terminer en au plus n itérations. □

Proposition 17. *L'itéré* x_k *de l'algorithme est le minimiseur de la fonction* (4.4) *sur le sous-espace affine* $x_0 + \mathbf{Vect}\{p_i \mid i = 0 \dots k-1\}$.

Preuve: Le minimiseur est caractérisé par $Ax + b \in \mathbf{Vect}\{p_i \mid i = 0 \dots k-1\}^{\perp}$ par stricte convexité de la fonction à minimiser. C'est justement la dernière propriété du théorème 14. □

Remarque 17. *En fait, on observe que le nombre d'itérations pour converger est majoré par la dimension de l'espace* $\mathbf{Vect}\{A^i(r_0) \mid i = 0 \dots k\}$ *qui peut-être bien inférieure à la dimension de l'espace. Par exemple, si* A *a seulement* p *valeurs propres alors le nombre d'itérations est majoré par* p *car il existe un polynôme de degré* p *annulant* A*.*

Il est possible d'estimer la vitesse de convergence de l'algorithme en fonction des valeurs propres de A. Ces estimations font souvent intervenir des quotients du type $\frac{\lambda_{n-k}-\lambda_1}{\lambda_{n-k}+\lambda_1}$ avec $\lambda_1 \leq \ldots \leq \lambda_n$ les valeurs propres de A. Plus cette quantité est petite, meilleure est la convergence de x_k vers x^*.

On peut réécrire l'algorithme du gradient conjugué de la manière suivante qui est la forme standard de l'algorithme:

Initialisation de $x \leftarrow x_0 \in \mathbb{R}^n$, $r_0 \leftarrow Ax_0 + b$,
$p_0 \leftarrow -r_0$, $k \leftarrow 0$
Tant que $(r_k \neq 0)$ **faire**
  $\alpha_k \leftarrow \frac{\|r_k\|^2}{\|p_k\|_A^2}$
  $x_{k+1} \leftarrow x_k + \alpha_k p_k$
  $r_{k+1} \leftarrow r_k + \alpha_k A p_k$
  $p_{k+1} \leftarrow -r_{k+1} + \frac{\|r_{k+1}\|^2}{\|r_k\|^2} p_k$
  $k \leftarrow k + 1$
Fin
Retourner x_k

Algorithme 7: Méthode du gradient conjugué linéaire: reformulation

Exercice 15. *Implémenter l'algorithme 7 pour observer le nombre d'itérations obtenues en pratique. Générer par exemple une matrice symétrique définie positive par* $A = NN^t$ *où* N *est une matrice aléatoire dont les coefficients sont i.i.de loi normale standard (en Matlab ou Python (module Scipy), utiliser la fonction* `randn()`*). Que remarque-t-on?*

4.4 Méthode du gradient conjugué dans le cas non-linéaire

La méthode du gradient conjugué repose essentiellement sur la structure de la fonction à minimiser (4.4) mais il est naturel de chercher une adaptation dans des cas plus généraux. Cela a été fait dans les années 60 par Fletcher et Reeves en adaptant l'algorithme 7.

Initialisation de $x \leftarrow x_0 \in \mathbb{R}^n$, $f_0 \leftarrow f(x_0)$,
$g_0 \leftarrow \nabla f(x_0)$, $p_0 \leftarrow -g_0$, $k \leftarrow 0$
Tant que ($|g_k| > \text{tol}$) **faire**
  Estimer α_k par recherche unidimensionnelle
  $x_{k+1} = x_k + \alpha_k p_k$
  $g_{k+1} \leftarrow \nabla f(x_{k+1})$
  $p_{k+1} \leftarrow -g_{k+1} + \frac{\|g_{k+1}\|^2}{\|g_k\|^2} p_k$
  $k \leftarrow k + 1$
Fin
Retourner x_k

Algorithme 8: Méthode du gradient conjugué non-linéaire: Fletcher-Reeves

Il faut faire attention dans l'estimation de α_k car ce paramètre influe sur le fait que p_k soit une direction de descente pour f. Si la recherche est exacte et donc que α_k est un point critique de la fonction unidimensionnelle alors $\langle g_k, p_{k-1} \rangle = 0$ ce qui implique $\langle g_k, p_k \rangle = -\|g_k\|^2 + \frac{\|g_k\|^2}{\|g_{k-1}\|^2} \langle g_k, p_{k-1} \rangle < 0$. Si la recherche du pas n'est pas exacte alors des conditions additionnelles de type Wolfe peuvent être ajoutées pour que p_k soit une direction de descente. Un autre algorithme utilisant la formulation de l'algorithme 6 proposé par Polak-Ribière cherche à conserver la condition d'orthogonalité par rapport au Hessien de f, *i.e.* $p_{k+1} = -g_{k+1} + \frac{\langle g_{k+1}, Ap_k \rangle}{\langle Ap_k, p_k \rangle} p_k$. Le Hessien n'est pas explicite mais Ap_k peut s'approximer par la formule suivante:

$$\alpha_k A p_k = \int_0^1 \nabla^2 f(x_k + s\alpha_k p_k) \alpha_k p_k \, ds = g_{k+1} - g_k \tag{4.10}$$

On peut donc remplacer

$$\frac{\langle g_{k+1}, Ap_k \rangle}{\langle Ap_k, p_k \rangle} \leftarrow \frac{\langle g_{k+1} - g_k, g_{k+1} \rangle}{\langle g_{k+1} - g_k, p_k \rangle} \,, \tag{4.11}$$

et donc $p_{k+1} \leftarrow -g_{k+1} + \frac{\langle g_{k+1} - g_k, g_{k+1} \rangle}{\langle g_{k+1} - g_k, p_k \rangle} p_k$.

Il a été démontré que l'algorithme 8 est globalement convergent et un exemple montre que l'algorithme de Polak-Ribière peut produire une suite pour laquelle aucune valeur d'adhérence n'est un point critique de la fonction. Cependant, l'algorithme de Polak-Ribière donne en pratique de meilleurs résultats que celui de Fletcher-Reeves, qui n'est donc plus utilisé. En fait, dans certains cas, l'algorithme 8 peut être plus lente qu'une descente de gradient simple. D'autre part, on peut se dire que si l'algorithme est initialisé dans un voisinage d'un point critique tel que la Hessienne est définie positive la convergence va être bonne (pour une fonction lisse). Si l'algorithme est redémarré toutes les n itérations alors la convergence des suite extraites $(x_{kn})_{k \in \mathbb{N}}$ est quadratique.

Chapter 5

Méthodes Newtoniennes

5.1 Introduction

La méthode de Newton utilise de l'information de second ordre pour minimiser une fonction f. Plus précisément, cette méthode demande la résolution d'un système linéaire qui fqit intervenir le Hessien de la fonction, ce qui est donc plus exigeant en calcul que les méthodes de gradient simple. Généralement, dans le cadre d'application de la méthode de Newton, on dispose du calcul de $f(x_0), \nabla f(x_0), \nabla^2 f(x_0)$.

Comment utiliser cette information pour obtenir un pas et une direction de descente? Le développement de Taylor à l'ordre 2 de f, une fonction $C^2(\mathbb{R}^n, \mathbb{R})$ s'écrit

$$f(x+s) = f(x) + \langle \nabla f(x), s \rangle + \frac{1}{2}\langle s, \nabla^2 f(x) s \rangle + o(\|s\|^2) \tag{5.1}$$

Minimiser cette approximation de second ordre donne l'optimum comme solution du système:

$$\nabla f(x) + \nabla^2 f(x) s = 0\,, \tag{5.2}$$

qui admet une unique solution s si $\nabla^2 f(x)$ est inversible. Cette solution sécrit donc

$$s = -[\nabla^2 f(x)]^{-1}(\nabla f(x)) \tag{5.3}$$

Si en plus $\nabla^2 f(x)$ est définie positive, s est une direction de descente admissible puisque dans ce cas:

$$\langle \nabla f(x), -[\nabla^2 f(x)]^{-1} \nabla f(x) \rangle < 0.$$

On peut donc proposer l'algorithme suivant:

Initialisation de $x \leftarrow x_0 \in \Omega$, $H \leftarrow \nabla^2 f(x_0)$.
[η est le seuil de tolérance sur le gradient]
Tant que ($\|\nabla f(x)\| \geq \eta$) **faire**
  $d \leftarrow -H^{-1}\nabla f(x)$,
  $\varepsilon := \operatorname{argmin}\{f(x+\eta d) \mid \eta > 0\}$ (pas optimal)
  $\varepsilon := 1$ (pas fixe)
  $x \leftarrow x + \varepsilon d$
  $H \leftarrow \nabla^2 f(x)$
Fin
Retourner x

Algorithme 9: Méthode de Newton à pas optimal/fixe

Remarque 18. *On remarque que l'équation* (5.2) *revient à la recherche d'une solution de l'équation* $\nabla f(y) = 0$ *en utilisant un développement au premier ordre de* ∇f.

5.2 La méthode de Newton pour la résolution numérique d'équations

La remarque 18 fait le lien entre le problème de minimisation et la résolution de l'équation $\nabla f(x_0) = 0$ (on rappelle que c'est seulement une condition nécessaire pour que x_0 soit un minimum). La méthode de Newton pour résoudre des équations du type $F(x) = 0$ s'applique donc au cas particulier où $F = \nabla f$.

La méthode de Newton pour la résolution d'équations utilise une information à l'ordre 1 sur F. En supposant par exemple que F est $C^1(\mathbb{R}^n, \mathbb{R}^n)$, on utilise le développement de Taylor:

$$F(x+s) = F(x) + DF(x)(s) + o(\|s\|)\,. \tag{5.4}$$

L'idée est de résoudre l'équation "approximante" $F(x) + DF(x)(s) = 0$ qui a une unique solution si $DF(x)$ est inversible donnée par $s = -DF(x)^{-1}(F(x))$.

Remarque 19. *Évidemment, on retrouve bien la formule* (5.3) *lorsqu'on remplace* F *par* ∇f *et* DF *par* $\nabla^2 f$.

On peut donc proposer l'algorithme suivant:

Initialisation de $x \leftarrow x_0 \in \mathbb{R}^n$
[η *est le seuil de tolérance sur la valeur de la fonction]*
Tant que $(\|F(x)\| \geq \eta)$ **faire**
  $x \leftarrow x - DF(x)^{-1}(F(x))$
Fin
Retourner x

Algorithme 10: Méthode de Newton pour la résolution d'équations

Théorème 15. Convergence locale quadratique de la méthode de Newton: *Soit* $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ *une application* $C^1(\mathbb{R}^n, \mathbb{R}^n)$ *telle que* dF *soit localement Lipschitz. On suppose qu'il existe* $x^* \in \mathbb{R}^n$ *tel que* $F(x^*) = 0$ *et* $dF(x^*)$ *inversible. Alors, il existe* $\varepsilon > 0$ *tel que l'algorithme 10 initialisé dans la boule* $B(x^*, \varepsilon)$ *converge et on a dans ce cas*

$$\|x_{k+1} - x^*\| \leq C\|x_k - x^*\|^2 \tag{5.5}$$

pour une certaine constante C.

Preuve: On suppose que la constante de Lipschitz de dF est M dans un voisinage de x^*. Comme $dF(x^*)$ est inversible et dF est Lipschitz (donc continue), il existe $\varepsilon_1 > 0$ et $L > 0$ tels que $dF(y)$ est inversible si $y \in B(x^*, \varepsilon_1)$ et $\|dF(y)^{-1}\| \leq L$. En particulier, tant que $x \in B(x^*, \varepsilon_1)$, la mise à jour de x est bien définie. Il reste à montrer que x_k (suite générée par l'algorithme 10) converge bien vers x^*. On peut donc écrire

$$\|F(x^*) - F(x) - dF(x)(x^* - x)\| \leq \frac{M}{2}\|x^* - x\|^2 \tag{5.6}$$

On obtient alors, comme $F(x^*) = 0$ et en multipliant par $dF(x)^{-1}$:

$$\| - dF(x)^{-1}(F(x)) - (x^* - x)\| \leq \frac{ML}{2}\|x^* - x_k\|^2$$

$$\|x_{k+1} - x^*\| \leq \frac{ML}{2}\|x^* - x_k\|^2\,.$$

Soit $\varepsilon = \frac{1}{2}\min(\frac{2}{ML}, 1)$, si $\|x^* - x_k\| < \varepsilon$, la suite x_k est bien définie et reste dans la boule $B(x^*, \varepsilon)$. On a de plus l'inégalité (5.5) pour $C = \frac{ML}{2}$ et donc

$$\|x_{k+1} - x^*\| \leq \frac{1}{2}\|x^* - x_k\|\,, \tag{5.7}$$

ce qui prouve la convergence de la suite x_k vers x^*. □

Ce théorème prouve la convergence de l'algorithme de Newton pour la résolution d'équations uniquement dans un voisinage de la solution. Une estimation de ce voisinage est même donnée dans la preuve (mais est fonction de quantité qui implique la connaissance de F sur un voisinage).

Discussion d'exemples

5.3 Méthode de Newton pour la minimisation

En ce qui concerne l'algorithme de Newton pour la minimisation, on constate immédiatement:

- Le théorème 15 donne la convergence quadratique de l'algorithme de Newton à pas fixe dans un voisinage d'un minimum local vérifiant la condition suffisante du second-ordre (voir l'équation (2.25)) $\nabla^2 f$ définie positive.
- Notons encore une fois que l'algorithme 9 n'est pas bien défini si la Hessienne n'est pas inversible et même si c'est le cas, cela ne suffit pas à garantir que la direction choisie est une direction de descente. L'algorithme peut donc diverger, ce qui est le cas en pratique.
- La proposition précédente montre que l'algorithme peut converger vers un point selle de f pour lequel la Hessienne est inversible mais qui n'est pas un minimum local.

On discute maintenant de la convergence de la méthode de Newton à pas optimal. La convergence n'est pas garantie par le théorème 15 puisque maintenant le pas est optimal. On peut prouver le théorème suivant:

Proposition 18 (Convergence de la méthode de Newton à pas optimal)**.** *Soit $f \in C^2(\Omega, \mathbb{R})$ et $x_0 \in \Omega$ tels que*

$$0 < c\,Id \leq \nabla^2 f(x) \leq K\,Id \tag{5.8}$$

pour $x \in S_0 := \{x \in \Omega \mid f(x) \leq f(x_0)\}$ que l'on suppose fermé dans $\mathbb{R}^n$ alors on a

- *la suite des valeurs $f(x_k)$ est strictement décroissante ou la suite x_k est stationnaire.*
- *si $\inf f > -\infty$ alors la suite $\nabla f(x_k)$ converge vers 0, S_0 contient au moins un minimiseur local de f et $\lim_{k\to\infty} d(x_k, LM) = 0$ où LM est l'ensemble des minimiseurs locaux de f.*
- *Si S_0 est convexe alors x_k converge vers l'unique minimiseur de f sur Ω.*

Preuve: Si le pas ou la direction de descente est nul, cela implique que le point courant est un minimum local (car $\nabla^2 f$ est définie positive) et la suite $(x_k)_{k\in\mathbb{N}}$ est stationnaire. Sinon, la direction d est une direction de descente et le choix du pas est optimal. En particulier, la suite x_k est strictement décroissante et est contenue dans

S_0. Lorsque f est bornée inférieurement, $f(x_k)$ est convergente vers une limite finie. Par la Proposition 4, on obtient:

$$f(x_{k+1}) \leq \inf_{\rho} f(x_k) + \rho\langle \nabla f(x_k), d_k\rangle + \frac{\rho^2}{2} K\|d_k\|^2$$
$$f(x_{k+1}) - f(x_k) \leq -\frac{\langle \nabla f(x_k), d_k\rangle^2}{2K\|d_k\|^2} \leq -\frac{c^2}{K}\|d_k\|^2 \leq 0\,,$$

On en déduit que $\lim_{k\mapsto\infty} d_k = 0$ et $\lim_{k\mapsto\infty} \nabla f(x_k) = 0$. En utilisant le corollaire 4, on en déduit l'existence d'un minimum local x^* tel que

$$\|x_k - x^*\| \leq \frac{1}{c}\|\nabla f(x_k)\| \tag{5.9}$$

On obtient donc le second point et le dernier point est la conséquence de la stricte convexité de f sur S_0 convexe. □

La vitesse de convergence quadratique est l'avantage le plus important de la méthode de Newton (à pas fixe) et est conservé pour le pas optimal:

Théorème 16. *Si l'algorithme converge vers un point x^* pour lequel $\nabla^2 f(x^*)$ est définie positive et $\nabla^2 f$ est $M-$Lipschtizienne au voisinage de x^* alors si la suite ne stationne pas en x^* pour k assez grand, on a*

- *le pas optimal calculé à chaque étape tend vers un,*
- *la convergence est au moins quadratique et* $\limsup_{k\to\infty} \frac{\|x^*-x_{k+1}\|}{\|x^*-x_k\|^2} \leq \frac{M}{c}$.

Preuve: Posons $\phi(t) = f(x_k + td_k)$, on a $\phi'(t) = \langle \nabla f(x_k + td_k), d_k\rangle$ et $\phi''(x_k + td_k) = \langle \nabla^2 f(x_k + td_k)d_k, d_k\rangle$ et en particulier, ϕ'' est Lipschitz de constante $\|d_k\|^3 M$. Puisque t_k est un pas optimal, nécessairement $\phi'(t_k) = 0$. On a donc

$$|(1-t_k)\langle d_k, [\nabla^2 f(x_k)]^{-1} d_k\rangle| = |\phi'(t_k) - \phi'(0) - \phi''(0)t_k| \leq \frac{M}{2} t_k^2 \|d_k\|^3 \tag{5.10}$$

Si d_k est nul pour un certain k, alors la suite est stationnaire à partir de k et le résultat est clair. Sinon, on en déduit pour $\varepsilon > 0$ suffisamment petit et k assez grand:

$$\left|\frac{1}{t}_k - 1\right| \leq \frac{M}{2c-\varepsilon}|t_k|\|d_k\| = \frac{M}{2c-\varepsilon}\|x_{k+1} - x_k\| \to 0 \tag{5.11}$$

ce qui prouve que le pas optimal tend vers 1.

Par continuité, $\nabla f(x^*) = 0$ et en utilisant un développement limité de $\nabla f(x)$ on obtient avec $e_k = x^* - x_k$

$$\|\nabla f(x^*) - \nabla f(x_k) - \nabla^2 f(x_k)e_k\| \leq \frac{M}{2}\|x^* - x_k\|^2$$
$$\|e_k + d_k\| \leq \frac{M}{2c-\varepsilon}\|x^* - x_k\|^2\,,$$

en multipliant la première ligne par $[\nabla^2 f(x_k)]^{-1}$. On obtient alors

$$\|d_k\| \leq \|e_k\|(1 + \frac{M}{2c-\varepsilon}\|e_k\|)\,.$$

On a de plus $e_{k+1} = e_k - t_k d_k$, on a donc

$$\|e_{k+1}\| \leq |t_k - 1|\|d_k\| + \frac{M}{2c-\varepsilon}\|x^* - x_k\|^2\,. \tag{5.12}$$

En utilisant l'inégalité (5.10), on a

$$|t_k - 1|\|d_k\| \leq \frac{M}{2c-\varepsilon} t_k^2 \|d_k\|^2 \leq \frac{M}{2c-\varepsilon}(1+o(1))\|e_k\|^2\,,$$

donc on obtient

$$\limsup_{k\to\infty} \frac{|t_k-1|\|d_k\|}{\|e_k\|^2} \leq \frac{M}{2c}$$

On en déduit le résultat

$$\limsup_{k\to\infty} \frac{\|e_{k+1}\|}{\|e_k\|^2} \leq \frac{M}{c}\,. \tag{5.13}$$

□

Remarque 20. *Attention, l'hypothèse $\nabla^2 f(x^*)$ définie positive est importante pour obtenir la vitesse de convergence quadratique comme le montre l'exemple suivant: $f(x,y) = x^4 + 6y^2$ est strictement convexe mais la hessienne au minimum 0 n'est pas inversible et l'algorithme initialisé avec $x_0 = (1,1)$ converge seulement linéairement.*

L'étude de la méthode de Newton à pas optimal montre que ce pas tend vers 1, il est donc naturel de s'attendre à ce qu'une méthode de Newton à pas fixé à 1 donne le même type de résultat. Finalement, ce qu'il faut retenir c'est qu'il y a un voisinage attractif pour un minimum local x_0 quel que soit le pas utilisé et la convergence est quadratique.

Une propriété importante de la méthode est son invariance par changement d'échelle. C'est un avantage de cette méthode par rapport à un choix de direction de descente comme le gradient usuel. En particulier, il n'est pas nécessaire dans une certaine mesure de redimensionner le problème considéré.

Théorème 17. *La méthode de Newton est invariante par transformation affine qui est la composition par une application linéaire inversible M et une translation z: $m(x) = Mx + z$. L'algorithme 9 pour f produit des suites de points x_k telles que $m(x_k)$ soit une suite générée par l'algorithme 9 pour $f \circ m$. Si on note $Newton_f(x)$ l'itéré de la première étape de la méthode de Newton, on a*

$$Newton_f(m(x)) = m(Newton_{f\circ m}(x)) \tag{5.14}$$

Preuve: On a $\nabla(f\circ m) = M^t(\nabla f)\circ m$ et $M^t[(\nabla^2 f)\circ m]M$ donc la direction de descente s'écrit

$$d_{f\circ m}(x) = \left[M^t[(\nabla^2 f)\circ m]M\right]^{-1} M^t(\nabla f)\circ m = M^{-1} d_f(m(x))\,, \tag{5.15}$$

Ensuite, on écrit

$$\begin{aligned}\inf_{t\in\mathbb{R}_+} (f\circ m)(x + tM^{-1}d_f(m(x))) &= \inf_{t\in\mathbb{R}_+} f(M(x+tM^{-1}d_f(m(x))) + z) \\ &= \inf_{t\in\mathbb{R}_+} f(m(x) + td_f(m(x)))\,. \end{aligned} \tag{5.16}$$

En notant $\text{Newton}_f(x) := x + t_{argmin} d$ où $d = -[\nabla^2 f(x)]^{-1}\nabla f(x)$ et t_{argmin} réalise le minimum de $f(x+td)$, on a donc bien:

$$\text{Newton}_f(m(x)) = m(\text{Newton}_{f\circ m}(x))\,,$$

ce qui est le résultat attendu. □

Remarque 21. *Cette preuve montre que le résultat est aussi valable pour la méthode de Newton à pas fixe.*

5.4 Pour aller plus loin

L'algorithme de Newton exige que la Hessienne de f soit définie positive pour que la direction choisie soit une direction de descente (par exemple f est elliptique). Si on souhaite minimiser une fonction quelconque, on doit proposer une alternative pratique lorsque la Hessienne ne vérifie pas cette condition. À l'aide de la partie 3.4, on peut proposer facilement un algorithme qui converge, celui-ci est du à Goldstein et Price.

Initialisation de $x \leftarrow x_0 \in \Omega$, $H \leftarrow \nabla^2 f(x_0)$ et $\nu > 0$
[η est le seuil de tolérance sur le gradient]
Tant que ($\|\nabla f(x)\| \geq \eta$) **faire**
  Si (H définie positive) **Alors**
    Si ($\frac{|\langle H^{-1}\nabla f(x), \nabla f(x)\rangle|}{\|\nabla f(x)\|\|H^{-1}(\nabla f(x))\|} \geq \nu$) **Alors**
      $d \leftarrow -H^{-1}\nabla f(x)$
    Fin Si
  Sinon
    $d \leftarrow -\nabla f(x)$
  Fin Si
  $\varepsilon :=$ choix du pas par la règle de Wolfe
  $x \leftarrow x + \varepsilon d$
  $H \leftarrow \nabla^2 f(x)$
Fin
Retourner x

Algorithme 11: Un algorithme de Newton modifié

Le théorème de Zoutendijk est valable pour cet algorithme et on obtient donc la terminaison de l'algorithme puisque la suite des gradients $\nabla f(x_k)$ converge vers 0. Évidemment, l'avantage de cet algorithme est de bénéficier de la convergence quadratique pour les cas favorables.

Une variante naturelle de l'algorithme de Newton est d'introduire une boucle pour satisfaire la condition de positivité de $\nabla^2 f(x) + \lambda Id$ pour $\lambda > 0$. Évidemment si λ est grand, la direction de descente obtenue est proche du gradient $\nabla f(x)$.

Exercice 16. *Écrire un algorithme qui utilise l'idée précédente et qui termine comme pour l'algorithme 11.*

Chapter 6

Introduction aux méthodes de quasi-Newton

6.1 Motivation

L'idée des méthodes de quasi-Newton est de remplacer le calcul de la Hessienne et du système linéaire associé par une approximation (moins coûteuse à calculer) de l'inverse de la Hessienne. Bien que le calcul de la Hessienne ne pose pas de problème dans beaucoup de cas, certains problèmes (en grande dimension notamment) nécessitent l'utilisation de ces techniques. Comme c'est une technique dérivée de la méthode de Newton, on peut étudier le développement de cette idée pour la résolution d'équations du type $f(x) = 0$ en dimension 1.

La méthode de Newton propose d'utiliser l'approximation linéaire pour résoudre l'équation précédente ce qui donne la suite:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}\,. \tag{6.1}$$

On souhaite donc éviter de calculer $f'(x_k)$ et utiliser les calculs précédents pour en obtenir une approximation. Limitons nous par exemple à l'information de $f(x_k)$ et $f(x_{k-1})$. On peut approximer $f'(x_k)$ grossièrement (a priori) par $\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}$. On obtient la méthode de la sécante:

$$x_{k+1} = x_k - f(x_k)\frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}\,. \tag{6.2}$$

On peut montrer le théorème suivant:

Proposition 19. *Soit $f : \mathbb{R} \mapsto \mathbb{R}$ une fonction C^1 qui admet une racine $x^* \in \mathbb{R}$ i.e. $f(x^*) = 0$ vérifiant $f'(x^*) \neq 0$ et C^2 au voisinage de x^*. Alors la méthode de la sécante converge si initialisée dans un voisinage de x^* et l'ordre de convergence est $\frac{1+\sqrt{5}}{2}$.*

Preuve: On admet cette proposition. Pour indication, une méthode de preuve est de faire un développement limité au voisinage de x^* et de trouver une relation du type

$$|e_{k+1}| \leq M|e_k||e_{k-1}|\,, \tag{6.3}$$

pour une certaine constante $M > 0$. On peut se convaincre du résultat en remplacant e_{k+1} par e_k^α et e_{k-1} par $e_k^{1/\alpha}$, puisqu'alors α est racine de $\alpha = 1 + \frac{1}{\alpha}$. Attention, tout cela doit être détaillé. □

Le commentaire essentiel est que l'ordre de convergence est intéressant pour une méthode qui ne nécessite pas de calcul de dérivée. C'est a priori encore plus intéressant dans le cas de la méthode de Newton pour la minimisation puisque dans ce cas, on peut essayer d'éviter le calcul de la Hessienne.

6.2 Méthodes de quasi-Newton

L'idée centrale des méthodes de quasi-Newton est de mettre à jour une approximation du Hessien de f noté B_k à chaque itération de la méthode de descente. Le modèle de second-ordre est donc:

$$f(y) \simeq f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2}\langle y - x_k, B_k(y - x_k) \rangle \,, \tag{6.4}$$

où B_k est symétrique définie positive et mise à jour à chaque itération afin de vérifier la condition de quasi-Newton:

$$B_k(x_{k-1} - x_k) = \nabla f(x_{k-1}) - \nabla f(x_k) \,. \tag{6.5}$$

Cette condition est l'extension en dimension supérieure de la méthode de la sécante sauf que la matrice B_k peut (doit) garder une certaine mémoire des m ises à jour précédentes. On remarque que es deux conditions précédentes (B_k symétrique définie positive et l'équation 6.5 dite équation de la sécante) ne peuvent être satisfaites simultanément que si

$$\langle x_{k-1} - x_k, \nabla f(x_{k-1}) - \nabla f(x_k) \rangle = \langle x_{k-1} - x_k, B_k(x_{k-1} - x_k) \rangle > 0 \,.$$

Exercice 17. *Pourquoi la condition précédente est-elle nécessaire et suffisante? (On rappelle qu'une condition pour qu'une matrice symétrique soit définie positive est que ces mineurs principaux soient strictement positifs).*

Ces conditions sont remplies si, par exemple, la règle de Wolfe est vérifiée (voir la preuve de la proposition 20). Pour sélectionner une telle matrice parmi l'ensemble des matrices vérifiant ces conditions, la première méthode (historiquement) propose de résoudre le problème d'optimisation suivant:

$$\text{argmin}_B \, \|B - B_{k-1}\| \text{ telle que } B(x_k - x_{k-1}) = \nabla f(x_k) - \nabla f(x_{k-1}) \,, \tag{6.6}$$

et B symétrique définie positive. La norme choisie correspond au produit scalaire usuel sur les matrices $\langle A, B \rangle = \text{tr}(A^t B)$ pondéré par le Hessien moyen

$$\bar{H} := \int_0^1 \nabla^2 f(x_{k-1} + s\varepsilon_{k-1} d_{k-1}) \, ds$$

qui vérifie

$$\bar{H}(x_k - x_{k-1}) = \nabla f(x_k) - \nabla f(x_{k-1}) \,, \tag{6.7}$$

comme suit:

$$\|A\|_{\bar{H}}^2 = \|\bar{H}^{-1/2} A \bar{H}^{-1/2}\|^2 \,.$$

L'idée pour introduire une telle pondération est de sélectionner une norme qui est invariante par transformation linéaire. Le point important est que le résultat de la minimisation ne dépend pas de de la matrice $\bar{H}$ à condition que l'équation (6.7) soit vérifiée et que la matrice soit inversible. Concrètement, cela signifie que le résultat ne dépend que de $u_k := x_k - x_{k-1}$ et $v_k := \nabla f(x_k) - \nabla f(x_{k-1})$. On peut obtenir la formule suivante:

$$B_k = (I - \varepsilon_k v_k u_k^t) B_{k-1} (I - \varepsilon_k u_k v_k^t) + \varepsilon_k v_k v_k^t \,, \tag{6.8}$$

avec $\varepsilon_k = \frac{1}{v_k^t u_k}$. Ensuite, l'inverse qui est utilisé dans la formule de calcul de la direction de descente peut être calculé de manière explicite. On omet la formule car cette méthode n'est plus utilisée en pratique. En effet, la méthode de référence a été proposée par Broyden, Fletcher, Goldfarb and Shanno (BFGS) qui suit la même approche mais sur l'inverse de B_k noté H_k. La formule de BFGS est la suivante

BFGS: $$H_k = (I - \varepsilon_k u_k v_k^t) H_{k-1} (I - \varepsilon_k v_k u_k^t) + \varepsilon_k u_k u_k^t \,. \tag{6.9}$$

Cette formule est obtenue par minimisation de

$$\|H - H_{k-1}\|_{\bar{H}}^2 = \|\bar{H}^{1/2}(H - H_{k-1})\bar{H}^{1/2}\|^2 \,,$$

sous les contraintes H symétrique et $Hv_k = u_k$.

Exercice 18. *Résoudre le problème de minimisation explicitement pour obtenir la formule* (6.9).

On obtient donc l'algorithme suivant:

Initialisation de $x \leftarrow x_0 \in \Omega$, $H \leftarrow A \in \mathrm{SDP}_n(\mathbb{R})$.
$k \leftarrow 0$
[η est le seuil de tolérance sur le gradient]
Tant que ($\|\nabla f(x_k)\| \geq \eta$) **faire**
  $d \leftarrow -H\nabla f(x_k)$,
  $\varepsilon_k \leftarrow \mathrm{argmin}\{f(x_k + \varepsilon d_k) \mid \varepsilon > 0\}$ en utilisant la règle de Wolfe et en testant le pas $\varepsilon = 1$)
  $x_{k+1} \leftarrow x_k + \varepsilon_k d_k$,
  Mettre à jour H par la formule (6.9),
  $k \leftarrow k + 1$.
Fin
Retourner x

Algorithme 12: méthode de BFGS

Proposition 20. *L'algorithme 12 vérifie quelque soit $k \geq 0$, H_k est définie positive et la formule de mise à jour est invariante par transformation linéaire inversible.*

Preuve: On vérifie d'abord la condition de positivité $\langle x_{k-1} - x_k, \nabla f(x_{k-1}) - \nabla f(x_k)\rangle > 0$: La condition de Wolfe sécrit pour $c \in]0,1[$

$$\langle \nabla f(x_{k-1} + \varepsilon d_{k-1}), d_{k-1}\rangle > c\langle \nabla f(x_{k-1}), d_{k-1}\rangle \,. \tag{6.10}$$

On obtient donc

$$\langle d_k, \nabla f(x_k - \nabla f(x_{k-1}))\rangle \geq (c-1)\langle \nabla f(x_{k-1}), d_{k-1}\rangle > 0$$

car $\langle \nabla f(x_{k-1}), d_{k-1}\rangle < 0$, ce qui donne à une constante positive (le pas optimal) près la condition de positivité recherchée. Ceci est donc vérifié tant que l'algorithme n'a pas terminé.

Pour démontrer que H_k reste symétrique définie positive, on utilise la formule (6.9). Pour un vecteur quelconque $w \in \mathbb{R}^n$, on a:

$$\langle w, H_k w\rangle = \langle r, H_{k-1} r\rangle + \varepsilon_k \langle u_k, w\rangle^2 \,, \tag{6.11}$$

où r est défini par $r = w - \varepsilon_k v_k \langle u_k, w\rangle$. Si (par hypothèse de récurrence) H_{k-1} est définie positive, le premier terme est strictement positif sauf si $r = 0$ et le second terme est strictement positif sauf si $\langle u_k, w\rangle = 0$. Donc, $\langle w, H_k v\rangle > 0$ sauf si $w = \varepsilon_k v_k \langle u_k, w\rangle = 0$. La matrice H_k est donc bien symétrique définie positive.

On vérifie maintenant l'invariance par changement de repère. Soit M une matrice inversible, alors l'expression correspondante à $\bar{H}$ dans le nouveau repère est $\bar{H}^* =$

$M^t \bar{H} M$. On obtient alors $u_k^* = M^{-1}u_k$ et $v_k = M^t v_k$. Si on suppose qu'initialement, on a $A^* = M^{-1}A[M^t]^{-1}$, alors on a quelque soit $k \geq 0$, $H_k^* = M^{-1}H_k[M^t]^{-1}$ car

$$\begin{aligned}(I - \varepsilon_k u_k^* v_k^{t*}) H_{k-1}^* (I - \varepsilon_k v_k^* u_k^{t*}) = \\ M^{-1}(M - \varepsilon_k u_k v_k^t M) M^{-1} H_{k-1} [M^t]^{-1} (M^t - \varepsilon_k M^t v_k u_k)[M^t]^{-1}\end{aligned} \tag{6.12}$$

ce qui signifie $H_k^* = M^{-1}H_k[M^t]^{-1}$. □

La proposition suivante est très intéressante sur un plan théorique puisqu'elle relie l'algorithme de BFGS et l'algorithme du gradient conjugué.

Proposition 21. *Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ une fonction quadratique du type $f(x) = \frac{1}{2}\langle x, Bx\rangle + \langle b, x\rangle$ avec $b \in \mathbb{R}^n$ et $B \in SDP_n(\mathbb{R})$. Si on remplace la règle de Wolfe par le calcul d'un pas optimal, la suite des directions de descente générées par l'algorithme de BFGS sont conjuguées par rapport à B et quelque soit $1 \leq i \leq k-1$, on a $H_k y_i = s_i$.*

Preuve: On prouve le résultat par récurrence. On suppose donc le résultat vrai au rang k. Notons tout d'abord que $\langle g_{k+1}, u_k\rangle = 0$ car le pas est optimal. On peut voir $\langle g_{k+1}, u_i\rangle = 0$ pour $i \leq k$ car

$$\langle g_{k+1}, u_i\rangle = \langle g_{i+1}, u_i\rangle + \langle g_{i+2} - g_{i+1}, u_i\rangle + \ldots + \langle g_{k+1} - g_k, u_i\rangle . \tag{6.13}$$

Or $g_{j+1} - g_j = \lambda_j B u_j$ pour un réel λ_j et par hypothèse de récurrence, $\langle Bu_j, u_i\rangle = 0$ pour $k+1 \geq j > i$, donc tous les termes de droite de l'équation (6.13) sont nuls.

La formule de mise à jour peut donc être utilisée pour voir:

$$H_k(v_i) = u_i$$

car c'est vrai par construction pour $i = k$ et pour $i < k$ l'hypothèse de récurrence permet d'obtenir $H_k v_i = H_{k-1}v_i + R(v_i)$ avec $R(v_i) = \varepsilon_k \langle v_i, u_k\rangle (u_k - H_{k-1}v_k) + \varepsilon_k u_k \langle v_i, (u_k - H_{k-1}v_k)\rangle$. On voit facilement que $R(v_i) = 0$ car $\langle v_i, u_k\rangle = \langle Gu_i, u_k\rangle = \langle Gu_i, u_k\rangle$ et $\langle v_i, H_{k-1}v_k\rangle = \langle H_{k-1}v_i, v_k\rangle = \langle u_i, v_k\rangle = 0$ par la première remarque.

Il reste à montrer $\langle u_{k+1}, Gu_i\rangle = 0$ pour $i \geq k$. On remarque que $Gu_i = \beta_i(g_i - g_{i-1}) = \beta_i v_i$ donc $\langle H_k g_{k+1}, v_i\rangle = \langle g_{k+1}, H_k v_i\rangle = \langle g_{k+1}, u_i\rangle = 0$ par la première remarque. □

Pour la matrice $A = Id$ dans le cas de la proposition précédente, l'algorithme donne les mêmes itérés que l'algorithme du gradient conjugué. La méthode du gradient conjugué non-linéaire peut toujours être utile pour des problèmes en grande dimension où le stockage de la matrice Hessienne est difficile; une méthode de quasi-Newton à mémoire limitée a aussi été développée pour répondre à ces problèmes.

On peut montrer que la méthode de BFGS converge sous les hypothèses suivantes

Théorème 18. *Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ une fonction de classe C^2 telle qu'il existe $x_0 \in \mathbb{R}^n$ tel que l'ensemble de niveau $F_{x_0} = \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$ est convexe et il existe des constantes strictement positives m, M satisfaisant*

$$m\|v\|^2 \leq \nabla^2 f(x) \leq M\|v\|^2 , \tag{6.14}$$

quelque soit $v \in \mathbb{R}^n$ et $x \in \mathbb{R}^n$. Alors, l'algorithme 12 converge (pour une tolérance nulle) vers l'unique minimiseur de f.

Preuve: Ce théorème est admis. □

Le principal intérêt de la méthode est d'obtenir une convergence super-linéaire (point commun avec la méthode de la sécante).

Théorème 19. *Soit $f : \mathbb{R}^n \mapsto \mathbb{R}$ une fonction de classe C^2 telle qu'il existe $x^* \in \mathbb{R}^n$ un minimum local satisfaisant $\nabla^2 f(x)$ définie positive et dont le hessien est Lipschitz au voisinage de x^*. Si la suite générée par l'algorithme 12 converge vers x^* alors l'ordre de convergence de la suite x_k est super-linéaire.*

Preuve: Ce théorème est admis. □

Il faut retenir que numériquement la formule de mise à jour de BFGS est reconnue comme étant la plus efficace parmi un grand nombre de formules proposées. Cette formule de mise à jour est stable numériquement et tend à corriger des erreurs d'approximations du Hessien au fur et à mesure des itérations.

Exercice 19. *Tenter d'établir une comparaison (sur des tests numériques par exemple) la plus juste possible entre la méthode de Newton et la méthode de quasi-Newton.*

Bibliography

[Ber99] D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.

[Bon97] Gilbert J.-C. Lemaréchal C. Sagastizábal C. Bonnans, J.-F. *Optimisation Numerique: Aspects theoriques et pratiques.* Springer Mathématiques et Applications, 1997.

[NW06] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer Series in Operations Research, 2006.