Optimisation

Fabrice Rossi

Plan

Table des matières

1 Introduction

Régression linéaire multiple

- Expliquer y par une combinaison linéaire des $(x_j)_{1\leq j\leq p}$

$$\hat{y} = \sum_{j=1}^{n} a_j x_j = \langle \mathbf{a}, \mathbf{x} \rangle$$

- Comment choisir les a_j ?
- Minimisation de l'erreur quadratique sur un ensemble d'exemples $(y_i, \mathbf{x}_i)_{1\leq i\leq n}$

$$\begin{aligned} \mathbf{a}^* &= \arg\min_{\mathbf{a}} \sum_{i=1}^{n} \left(y_i - \langle \mathbf{a}, \mathbf{x}_i \rangle\right)^2 \\ &= \arg\min_{\mathbf{a}} \|\mathbf{y} - \mathbf{a}^T X\|_2^2, \end{aligned}$$

avec $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$

Régression contrainte

- Pas de contrôle de $\mathbf{a}$:
  - explosion possible des coefficients
  - problème mal conditionné
  - trop de coefficients non nuls
- Régression *ridge* :

– contrainte sur $\|\mathbf{a}\|_2 = \sqrt{\sum_{j=1}^{p} a_j^2}$

$$\begin{array}{rl} \text{minimiser} & \|\mathbf{y} - \mathbf{a}^T X\|_2^2 \\ \text{avec} & \|\mathbf{a}\|_2 \leq \lambda \end{array}$$

– Régression parcimonieuse (*lasso*) :
  – contrainte sur $\|\mathbf{a}\|_1 = \sum_{j=1}^{p} |a_j|$

$$\begin{array}{rl} \text{minimiser} & \|\mathbf{y} - \mathbf{a}^T X\|_2^2 \\ \text{avec} & \|\mathbf{a}\|_1 \leq \lambda \end{array}$$

Modèle de Markowitz en finance

– Objectif : investissement dans un portefeuille d'actifs avec minimisation du risque sous contrainte de rendement
– Modèle :
  – r_i rendement de l'actif i
  – évolution stochastique : $\mathbf{r}$ est de moyenne $\boldsymbol{\mu}$ et de covariance Σ
  – proportion des actifs dans le portefeuille, $\mathbf{x}$ ($\sum_i x_i = 1 = \mathbf{1}^T\mathbf{x}$)
  – espérance de rendement : $\boldsymbol{\mu}^T\mathbf{x}$
  – variance du rendement : $\mathbf{x}^T\Sigma\mathbf{x}$
– Problème

$$\begin{array}{rl} \text{minimiser} & \mathbf{x}^T\Sigma\mathbf{x} \\ \text{avec} & \mathbf{1}^T\mathbf{x} = 1 \\ & \boldsymbol{\mu}^T\mathbf{x} \geq r_{\min} \end{array}$$

Forme générale

– un problème d'optimisation ($\mathcal{P}$) est défini par

$$\begin{array}{rl} \text{minimiser sur } \mathbb{R}^n & J(\mathbf{x}) \\ \text{avec} & g_i(\mathbf{x}) \leq 0\,, 1 \leq i \leq p \\ & h_j(\mathbf{x}) = 0\,, 1 \leq j \leq q \end{array}$$

– vocabulaire :
  – J (à valeur dans $\mathbb{R} \cup \{\infty\}$) est la **fonction de coût**, la **fonction objectif** ou encore le **critère**
  – les g_i sont les **contraintes d'inégalité**
  – les h_j sont les **contraintes d'égalité**
  – l'**ensemble des contraintes** est

$$\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^n | g_i(\mathbf{x}) \leq 0\,, 1 \leq i \leq p \text{ et } h_j(\mathbf{x}) = 0\,, 1 \leq j \leq q\}$$

  ensemble des points admissibles ou **réalisables**

Notes

Énoncé ici dans $\mathbb{R}^n$, mais applicable plus généralement.

Optima locaux et globaux

– optimum local : meilleure valeur *localement* au sens de la métrique de l'espace et de l'ensemble des contraintes
– formellement :
  – $\mathcal{C} \subset X$ espace métrique et J de $\mathcal{C}$ dans $\mathbb{R}$
  – $x^* \in \mathcal{C}$ réalise un *minimum local* de J sur $\mathcal{C}$ ssi $\exists$ une boule ouverte B centrée en x^* telle que $\forall x \in B \cap \mathcal{C}$, $J(x) \geq J(x^*)$

  - inégalité *stricte* pour $x \neq x^*$ → minimum local *strict*
- $x^* \in \mathcal{C}$ réalise un *minimum global* de J sur $\mathcal{C}$ ssi $\forall x \in \mathcal{C}$, $J(x) \geq J(x^*)$
- Propriété : minimum de $J \Leftrightarrow$ maximum de $-J$

Notes

Attention aux effets de $\mathcal{C}$: montrer ce qui se passe avec x sur $\mathbb{R}^+$, par exemple.

État de l'art

- problème de moindres carrés
  - J est quadratique et il n'y a pas de contraintes
  - résolution facile (inverse ou pseudo-inverse de matrice)
- programmation linéaire :
  - $J(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ (ou affine)
  - contraintes linéaires (ou affines)
  - résolution relativement facile
- programmation convexe :
  - les fonctions J et g_i sont convexes et les h_i affines
  - résolution possible
- cas général :
  - résolution complète (c.-à-d. trouver un minimum **global**) très difficile (coût exponentiel en la taille du problème)
  - résolution locale envisageable

2 Outils mathématiques

Outils fondamentaux

- deux outils mathématiques fondamentaux pour l'analyse des problèmes d'optimisation
- différentiabilité :
  - approximation linéaire locale
  - tendances locales, par exemple plus grande pente
  - *caractérisation des optima*
- convexité :
  - pour les ensembles : on peut se promener sur un segment
  - pour les fonctions : quand on se promène sur un segment, la fonction reste ≪ sous le segment ≫
  - *existence d'optima*
- lien important entre les deux propriétés : une fonction convexe est minorée *globalement* par son approximation linéaire *locale*

Rappels

- espace vectoriel normé :
  - espace vectoriel
  - norme (définie positive, sous-additive et homogène)
- espace de Banach : espace vectoriel normé complet
- espace de Hilbert :
  - espace vectoriel
  - produit scalaire (bilinéaire, symétrique, défini positif)
  - complet

Notes

- norme :
  - $\|x\| \geq 0$ et $\|x\| = 0$ si et seulement si $x = 0$
  - $\|\alpha x\| = |\alpha|\|x\|$
  - $\|x + y\| \leq \|x\| + \|y\|$
- complet : les suites de Cauchy ont une limite
- produit scalaire :
  - $\langle x, x\rangle \geq 0$ et $\langle x, x\rangle = 0$ si et seulement si $x = 0$
  - $\langle x, y\rangle = \langle y, x\rangle$
  - $\langle ax + by, z\rangle = a\langle x, z\rangle + b\langle y, z\rangle$

2.1 Différentiabilité

Différentiabilité

- J est définie sur ouvert U d'un Banach X et à valeurs dans $\mathbb{R}$
- J est **différentiable** (au sens de **Fréchet**) en $x \in U$ s'il existe une forme linéaire *continue* DJ_x telle que
$$\lim_{\|h\|_X \to 0} \frac{J(x+h) - J(x) - DJ_x(h)}{\|h\|_X} = 0,$$
c'est-à-dire $J(x + h) = J(x) + DJ_x(h) + o(\|h\|_X)$
- si X est un espace de Hilbert, on définit le **gradient** de J en x comme l'élément de X, noté $\nabla J(x)$, tel que
$$DJ_x(h) = \langle \nabla J(x), h\rangle_X$$
Théorème de représentation de Riesz ($X' = X$)

Notes

Rappels :

- forme linéaire continue : il existe M tel que pour tout h, $|DJ_x(h)| \leq M\|h\|_X$
- théorème de représentation de Riesz : identification du dual topologique d'un espace de Hilbert avec lui même. Conséquence : toute forme linéaire continue définie sur un Hilbert s'exprime sous la forme d'un produit scalaire.

Différentiabilité

- si $J(x) = \langle a, x\rangle_X$, $\nabla J(x) = a$
- dans $\mathbb{R}^n$, si $J(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$, alors $\nabla J(\mathbf{x}) = (A + A^T)\mathbf{x}$
- cas particulier, si $J(\mathbf{x}) = \|\mathbf{x}\|^2$, alors $\nabla J(\mathbf{x}) = 2\mathbf{x}$
- dans $\mathbb{R}$, $\nabla J(x) = J'(x)$
- dans $\mathbb{R}^n$,
$$\nabla J(\mathbf{x}) = \left(\frac{\partial J}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial J}{\partial x_n}(\mathbf{x})\right)^T$$

Notes

Principe des solutions : étudier $J(x + h) - J(x)$. Par exemple

$$\begin{aligned} J(x+h) - J(x) &= (x+h)^T A(x+h) - x^T Ax \\ &= x^T Ah + h^T Ax + h^T Ah \\ &= (x^T A + x^t A^T)h + h^T Ah \\ &= \left\langle (A + A^T)x, h\right\rangle_{\mathbb{R}^n} + \langle h, Ah\rangle_{\mathbb{R}^n} \end{aligned}$$

Or, le théorème de Cauchy-Schwartz donne

$$| \langle h, Ah \rangle_{\mathbb{R}^n} | \leq \|h\|_{\mathbb{R}^n} \|Ah\|_{\mathbb{R}^n},$$

et par continuité de A,

$$| \langle h, Ah \rangle_{\mathbb{R}^n} | \leq \|A\| \|h\|^2_{\mathbb{R}^n} = o(\|h\|_{\mathbb{R}^n}).$$

Et donc finalement, $\nabla J(\mathbf{x}) = (A + A^T)\mathbf{x}$.

Différentielle au sens de Gâteaux

- différentielle au sens de Gâteaux : dérivée directionnelle linéaire continue
- J est définie sur un ouvert U d'un Banach X et à valeurs dans $\mathbb{R}$
- J est **Gâteaux-différentiable** en $x \in U$ ssi la dérivée directionnelle

$$J'(x, h) = \lim_{t \to 0^+} \frac{J(x + th) - J(x)}{t},$$

existe pour tout $h \neq 0$ et si $h \mapsto J'(x, h)$ est une forme linéaire continue

Propriétés

- linéaire n'est pas automatique : $J(u, v) = \frac{u^2 v^2}{(u^2+v^2)^{\frac{3}{2}}}$ en $(0, 0)$ avec $J(u, v) = 0$
- la notion de gradient s'applique aussi dans le cas Gâteaux-différentiable : $J'(x, h) = \langle \nabla J(x), h \rangle_X$ (pour X un Hilbert)
- Fréchet différentiable implique Gâteaux différentiable
- réciproque fausse :
  - $J(x, y) = \frac{x^6}{(y-x^2)^2+x^8}$ avec $J(0, 0) = 0$
  - J n'est pas continue en $(0, 0)$, mais est Gâteaux différentiable
  - principe de l'exemple : approche linéaire ok, mais approche « curviligne » discontinue

Notes

Premier exemple :

$$\begin{aligned} \frac{J(ta, tb) - J(0, 0)}{t} &= \frac{t^4 a^2 b^2}{t(t^2 a^2 + t^2 b^2)^{\frac{3}{2}}} \\ &= \frac{a^2 b^2}{(a^2 + b^2)^{\frac{3}{2}}} \end{aligned}$$

Deuxième exemple :

$$\begin{aligned} \frac{J(ta, tb) - J(0, 0)}{t} &= \frac{t^6 a^6}{t((tb - t^2 a^2)^2 + t^8 a^8)} \\ &= \frac{t^5 a^6}{t^2 (b - ta^2)^2 + t^8 a^8} \\ &= \frac{t^3 a^6}{(b - ta^2)^2 + t^5 a^8} \end{aligned}$$

et donc $\lim_{t \to 0^+} \frac{J(ta,tb)-J(0,0)}{t} = 0$. Mais

$$J(x, x^2) = \frac{1}{x^2},$$

donc $J(x, y)$ n'a pas de limite finie en $(0, 0)$.

Deuxième ordre

- différentielle : approximation linéaire locale
- différentielle à l'ordre deux : approximation *quadratique* locale
- J est définie sur un ouvert U d'un Banach X et à valeurs dans $\mathbb{R}$
- J est différentiable à l'ordre deux en $x \in U$ si elle est différentiable dans un voisinage de x et si $u \mapsto DJ_u$ est différentiable en x
- Remarque :
  - DJ_u est à valeurs dans X' le dual (topologique) de X
  - différentielle d'une fonction à valeurs dans un Banach V : il suffit de considérer $\|J(x+h) - J(x) - DJ_x(h)\|_V$ (qui doit être $o(\|h\|_X)$)
  - Gâteaux ou Fréchet

Deuxième ordre

- plus simplement, J est Fréchet différentiable à l'ordre deux en $x \in U$ ssi il existe une application linéaire continue DJ_x et une application bilinéaire symétrique continue D^2J_x telles que
$$J(x+h) = J(x) + DJ_x(h) + \frac{1}{2}D^2J_x(h,h) + o(\|h\|_X^2)$$
- dans un Hilbert
  - $DJ_x(h) = \langle \nabla J(x), h \rangle_X$
  - $D^2J_x(h,h) = \left\langle \nabla^2 J(x)(h), h \right\rangle_X$ pour un endomorphisme symétrique continu $\nabla^2 J(x)$ (appelé le **Hessien** par abus de langage)

Deuxième ordre

- dans $\mathbb{R}^n$, la matrice Hessienne est donnée par
$$\nabla^2 J = \begin{pmatrix} \frac{\partial^2 J}{\partial x_1^2} & \frac{\partial^2 J}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 J}{\partial x_1 \partial x_n} \\ \frac{\partial^2 J}{\partial x_2 \partial x_1} & \frac{\partial^2 J}{\partial x_2^2} & \cdots & \frac{\partial^2 J}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 J}{\partial x_n \partial x_1} & \frac{\partial^2 J}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 J}{\partial x_n^2} \end{pmatrix}$$
- dans $\mathbb{R}^n$, si $J(\mathbf{x}) = \mathbf{x}^T A \mathbf{x}$, alors $\nabla^2 J(\mathbf{x}) = (A + A^T)$

Notes

Deuxième exemple, on avait

$$J(x+h) - J(x) = \left\langle (A + A^T)x, h \right\rangle_{\mathbb{R}^n} + \langle h, Ah \rangle_{\mathbb{R}^n} .$$

On peut écrire

$$\begin{aligned} \langle h, Ah \rangle_{\mathbb{R}^n} &= \frac{1}{2} \left(\langle h, Ah \rangle_{\mathbb{R}^n} + \langle Ah, h \rangle_{\mathbb{R}^n} \right) \\ &= \frac{1}{2} \left(h^T A^T h + h^T A h \right) \\ &= \frac{1}{2} \left\langle (A^T + A)h, h \right\rangle_{\mathbb{R}^n} . \end{aligned}$$

2.2 Convexité

Convexité

- convexe : contient les segments entre ses points

- formellement, $\mathcal{C} \subset X$ (un espace vectoriel) est **convexe** ssi

$$\forall x, y \in \mathcal{C},\ \forall t \in [0,1],\ (1-t)x + ty \in \mathcal{C}$$

- une fonction J de $\mathcal{C} \subset X$ dans $\mathbb{R} \cup \{\infty\}$ est **convexe** ssi :
  - $\mathcal{C}$ est convexe et,
  - $\forall x, y \in \mathcal{C},\ \forall t \in [0,1]$,

$$J\left((1-t)x + ty\right) \leq (1-t)J(x) + tJ(y)$$

  - convexité **stricte** si quand $x \neq y$ et $\forall t \in]0,1[$,

$$J\left((1-t)x + ty\right) < (1-t)J(x) + tJ(y)$$

- J est (strictement) concave si $-J$ est (strictement) convexe

Exemples

- remarque : J est convexe ssi toutes les fonctions $g(\lambda) = J(x + \lambda h)$ sont convexes
- les fonctions affines $J(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b$ sont convexes
- sur $\mathbb{R}^n$:
  - toute norme est convexe (inégalité triangulaire + homogénéité)
  - $J(\mathbf{x}) = \max_i x_i$ est convexe
  - $|x|$ est convexe
  - que dire de $J(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ avec A symétrique ?

Notes

Démonstration de la remarque :

- supposons J convexe. Soit x et h fixés. Soit λ_1 et λ_2 tels que $x + \lambda_1 h \in \mathcal{C}$ et $x + \lambda_2 h \in \mathcal{C}$. Alors $\forall t \in [0,1]$, $\lambda = (1-t)\lambda_1 + t\lambda_2$ est tel que $x + \lambda h \in \mathcal{C}$. En effet,

$$\begin{aligned} x + \lambda h &= x + ((1-t)\lambda_1 + t\lambda_2)h \\ &= (1-t)x + tx + ((1-t)\lambda_1 + t\lambda_2)h \\ &= (1-t)(x + \lambda_1 h) + t(x + \lambda_2 h) \end{aligned}$$

  Ce qui permet de conclure grâce à la convexité de $\mathcal{C}$. La même construction montre immédiatement par convexité de J que

$$g((1-t)\lambda_1 + t\lambda_2) \leq (1-t)g(\lambda_1) + tg(\lambda_2),$$

  ce qui donne la convexité de g
- dans l'autre sens, on se donne donc x et y dans $\mathcal{C}$, et on considère $(1-t)x+ty$, soit $x+t(y-x)$ et on considère le g correspondant à $h = y-x$. On a $g(0) = J(x)$ et $g(1) = J(y)$. Par convexité de g,

$$g((1-t)\lambda_1 + t\lambda_2) \leq (1-t)g(\lambda_1) + tg(\lambda_2),$$

  ce qui donne pour $\lambda_1 = 0$ et $\lambda_2 = 1$

$$J((1-t)x + ty) = g(t) \leq (1-t)J(x) + tJ(y).$$

  On en déduit la convexité de J.

Analyse de $J(x) = max_i x_i$: comme $\lambda \in [0,1]$,

$$\begin{aligned} (1-\lambda)x_i &\leq (1-\lambda)\max_j x_j \\ \lambda y_i &\leq \lambda \max_j y_j, \end{aligned}$$

et donc

$$(1-\lambda)x_i + \lambda y_i \leq (1-\lambda)J(x) + \lambda J(y),$$

ce qui assure la convexité de J.

Propriétés

- domaine de définition d'une fonction convexe

$$\mathrm{dom}(J) = \{x \in \mathcal{C} | J(x) < \infty\}$$

- J est dite **propre** quand $\mathrm{dom}(J) \neq \emptyset$
- si $\mathcal{C} \subset \mathbb{R}^n$, une fonction convexe propre est continue sur l'intérieur de son domaine
  - dans le cas Banach, il faut supposer que J est majorée et non identiquement égale à $-\infty$ sur un ouvert (non vide) inclus dans $\mathrm{dom}(J)$ pour obtenir la continuité

Notes

Démonstration :

- dans un Banach, on suppose J majorée sur un voisinage de u_0 : J est alors continue en u_0. Considérons en effet $B = B(u_0, \rho)$ une boule ouverte incluse dans $\mathrm{dom}(J)$ et telle que $J < M$ sur B. Soit $u \in B(u_0, \rho/2)$. On considère $u_1 = u_0 + t(u - u_0)$ avec $t = \frac{\rho}{2\|u-u_0\|}$. On a donc $u_1 \in B$. Par convexité, on a

$$\begin{aligned} J(u) &\leq \frac{1}{t} J(u_1) + \frac{t-1}{t} J(u_0) \\ J(u) - J(u_0) &\leq \frac{1}{t} \left(J(u_1) - J(u_0)\right) \\ J(u) - J(u_0) &\leq \frac{M - J(u_0)}{t} = \frac{2\|u - u_0\|(M - J(u_0))}{\rho} \end{aligned}$$

  De la même façon, on considère $u_2 = u_0 - t(u - u_0)$. De nouveau, $u_2 \in B$. Comme $u_0 = \frac{1}{1+t}(u_2 + tu)$, par convexité, on a

$$\begin{aligned} J(u_0) &\leq \frac{1}{1+t} J(u_2) + \frac{t}{1+t} J(u) \\ J(u_0) + tJ(u_0) &\leq J(u_2) + tJ(u) \\ J(u_0) - J(u) &\leq \frac{1}{t} \left(J(u_2) - J(u_0)\right) \\ J(u_0) - J(u) &\leq \frac{M - J(u_0)}{t} = \frac{2\|u - u_0\|(M - J(u_0))}{\rho}. \end{aligned}$$

  On en déduit donc que $|J(u) - J(u_0)| \leq M'\|u - u_0\|$ pour une valeur M' indépendante de u. Donc $\lim_{u \to u_0} J(u) = J(u_0)$.
- soit maintenant $u \neq u_0$ dans l'intérieur de $\mathrm{dom}(J)$. Par définition, il existe $\epsilon > 0$ tel que $w = u + \epsilon(u - u_0) = u_0 + (1 + \epsilon)(u - u_0)$ soit dans l'intérieur de $\mathrm{dom}(J)$. Pour tout v dans l'intérieur de $\mathrm{dom}(J)$, on peut définir $v' = (1 + \frac{1}{\epsilon})(v - w) + w$, ce qui conduit à

$$\begin{aligned} v' - u_0 &= \left(1 + \frac{1}{\epsilon}\right)(v - u - \epsilon(u - u_0)) + (1 + \epsilon)(u - u_0) \\ v' - u_0 &= \left(1 + \frac{1}{\epsilon}\right)(v - u). \end{aligned}$$

  On considère alors $B(u, \mu)$ avec $\mu = \frac{\epsilon}{1+\epsilon}\rho$. Pour tout $v \in B(u, \mu)$, on a $v' \in B(u_0, \rho)$. Comme $v = \frac{\epsilon}{1+\epsilon} v' + \left(1 - \frac{\epsilon}{1+\epsilon}\right) w$ on a par convexité de J

$$\begin{aligned} J(v) &\leq \frac{\epsilon}{1+\epsilon} J(v') + \left(1 - \frac{\epsilon}{1+\epsilon}\right) J(w), \\ J(v) &\leq \frac{\epsilon}{1+\epsilon} M + \left(1 - \frac{\epsilon}{1+\epsilon}\right) J(w), \end{aligned}$$

  ce qui montre que J est majorée sur $B(u, \mu)$. On applique alors le résultat précédent pour conclure à la continuité de J en u.

– en dimension n, on utilise l'existence de $n+1$ points affinement indépendants pour montrer que J est majorée sur l'intérieur de l'enveloppe convexe des points considérés.

Caractérisation par les différentielles

– soit J de $\mathcal{C} \subset H$ (Hilbert) dans $\mathbb{R}$ et Gâteaux différentiable sur $\mathcal{C}$ un convexe, J est convexe ssi

$$\forall x, y \in \mathcal{C},\ J(y) \geq J(x) + \langle \nabla J(x), y - x \rangle_H$$

– en d'autres termes : l'approximation linéaire *locale* est un minorant *global*

– preuve :

$\Rightarrow$ passage à la limite sur t positif de

$$\frac{J\left(x + t(y-x)\right) - J(x)}{t} \leq J(x) - J(y)$$

$\Leftarrow$ combinaison de deux applications de la minoration

$$\begin{array}{rcl} J(x) & \geq & J(x + t(y-x)) - t \left\langle \nabla J(x + t(y-x)), y - x \right\rangle_H \\ J(y) & \geq & J(x + t(y-x)) + (1-t) \left\langle \nabla J(x + t(y-x)), y - x \right\rangle_H \end{array}$$

Notes

$\Rightarrow$ on considère donc $(1-t)x + ty$, ce qui par convexité conduit à

$$J(x + t(y-x)) \leq (1-t)J(x) + tJ(y)$$

puis à l'équation du transparent car $t > 0$. Par passage à la limite, on a

$$J'(x, y-x) = \langle \nabla J(x), y - x \rangle_H \leq J(x) - J(y)$$

$\Leftarrow$ on applique donc deux fois la minoration, pour les couples $\{x+t(y-x), x\}$ et $\{x+t(y-x), y\}$, ce qui donne les deux inégalités du transparent. Comme $1 > t > 0$, on obtient

$$\begin{array}{rcl} (1-t)J(x) & \geq & (1-t)J(x + t(y-x)) - (1-t)t \left\langle \nabla J(x + t(y-x)), y - x \right\rangle_H \\ tJ(y) & \geq & tJ(x + t(y-x)) + t(1-t) \left\langle \nabla J(x + t(y-x)), y - x \right\rangle_H . \end{array}$$

La convexité s'en déduit par sommation.

Caractérisation par les différentielles

– autre caractérisation, J est convexe ssi ∇J est un opérateur monotone

$$\forall x, y \in \mathcal{C},\ \langle \nabla J(y) - \nabla J(x), y - x \rangle_H \geq 0$$

– remarque : la convexité stricte est impliquée par une version stricte des conditions (minoration globale et monotonie)

– ordre 2 : si J est C^2, elle est convexe ssi $\nabla^2 J(x)$ est positive en tout $x \in \mathcal{C}$

Exemples

- que dire de $J(x) = \frac{1}{x^2}$?
- maximum « assoupli » : $J(\mathbf{x}) = \log\left(\sum_{i=1}^{n} \exp x_i\right)$
- $J(A) = \log\det A$ est concave sur l'ensemble des matrices symétriques définies positives

Notes

- le domaine de définition de $J(x) = \frac{1}{x^2}$ n'est pas convexe...
- pour le soft max, on calcule le Hessien, puis on montre qu'il est défini positif en utilisant Cauchy-Schwarz
- pour le déterminant, on se ramène à l'étude d'une fonction sur $\mathbb{R}$ en considérant les matrices $Z + tV$. On utilise alors le fait qu'une matrice symétrique définie positive possède une racine carrée, ce qui permet d'introduire les valeurs propres d'une matrice bien choisie et de calculer explicitement la dérivée seconde de la fonction de t.

3 Résultats théoriques d'optimisation sans contrainte

3.1 Existence et unicité d'un minimum

Existence d'un minimum

- on s'intéresse à $(\mathcal{P})$ $\min_{\mathbf{x}\in\mathbb{R}^n} J(x)$: pas de contrainte
- remarque : $\inf_{\mathbf{x}\in\mathbb{R}^n} J(x)$ est toujours bien défini, le problème est de *réaliser* un minimum
- J est à valeurs dans $\mathbb{R} \cup \{+\infty\}$
- J est coercitive (*coercive*) si $\lim_{\|x\|\to\infty} J(x) = +\infty$
- Résultat : si J est propre (non identiquement infinie), continue et coercitive alors $(\mathcal{P})$ a *au moins* une solution

Notes

L'idée est de montrer que $d = \inf_{\mathbf{x}\in\mathbb{R}^n} J(x) < \infty$ est atteint :

- J coercive implique qu'une suite $J(x_p)$ convergeant vers d est portée par des x_p bornées. Soit en effet une telle suite. Il existe un P tel que $p \geq P$ implique $J(x_p) < d + 1$. Comme J est coercive, il existe un M tel que $\|x\| \geq M$ implique $J(x) > d + 1$. Donc $p \geq P$ implique $\|x_p\| < M$. Donc les $\|x_p\|$ sont bornées par $\max(M, \|x\|_1, \ldots, \|x\|_P)$.
- la continuité garantit que d est atteint en toutes les valeurs d'adhérence de la suite des x_p. Or comme $(x_p)_{p\geq 1}$ est un fermé borné de $\mathbb{R}^n$, il est compact et admet donc au moins une valeur d'adhérence.

Unicité

- En général, on a des minima, pas un minimum
- Cas particulier : si J est **strictement** convexe, alors $(\mathcal{P})$ a *au plus* une solution
- Preuve :
  - soit d un minimum atteint en x et y
  - si $x \neq y$, par convexité stricte, on a

$$J\left(\frac{1}{2}x + \frac{1}{2}y\right) < \frac{1}{2}J(x) + \frac{1}{2}J(y) = d,$$

    ce qui est impossible, puisque d est le minimum de J
- Remarque :
  - un minimum *local* d'une fonction convexe est *global*

– l'ensemble des points qui réalisent le minimum d'une fonction convexe est convexe

Notes

– première remarque : soit un minimum local x^*, par convexité pour tout y

$$J(tx^* + (1-t)y) \leq tJ(x^*) + (1-t)J(y)$$

Quand t est suffisamment proche de 1, $J(tx^* + (1-t)y) \geq J(x^*)$ par hypothèse, ce qui conduit à $(1-t)J(x^*) \leq (1-t)J(y)$ d'où la conclusion.

– deuxième remarque : si x et y réalisent le minimum d, on a

$$J(tx + (1-t)y) \leq tJ(x) + (1-t)J(y) = d,$$

ce qui montre que tout point de la forme $tx + (1-t)y$ réalise le minimum, d'où la convexité de l'ensemble de ces points.

Existence et unicité

– Si J est strictement convexe et coercitive alors $(\mathcal{P})$ admet une solution unique

– Cas intéressant, les fonctions **elliptiques** (α-convexes ou fortement convexes) :

– J est elliptique ssi existe une constante d'ellipticité $\alpha > 0$ telle que $\forall x, y, t \in [0,1]$,

$$J\left((1-t)x + ty\right) \leq (1-t)J(x) + tJ(y) - \frac{\alpha}{2}t(1-t)\|x-y\|^2$$

– elliptique $\Rightarrow$ strictement convexe et coercitive

Notes

– on montre d'abord que J convexe implique $J(x) \geq \langle h, x\rangle + \delta$ pour un certain couple (h, δ). On considère l'épigraphe de J, c'est-à-dire

$$Epi(J) = \{(x,t) | J(x) \leq t\}.$$

Cet ensemble est clairement convexe et fermé (par convexité de J et par continuité de J). On se donne un élément x_0 du domaine de J (supposé non vide) et $\lambda_0 < J(x_0)$. Alors $(x_0, \lambda_0) \notin Epi(J)$. On sait qu'un point peut être séparé strictement d'un convexe fermé dans un espace de Hilbert. Il existe donc v, α et β tels que $\langle v, x\rangle + \alpha t > \beta > \langle v, x_0\rangle + \alpha\lambda_0$ pour tout $(x,t) \in Epi(J)$. En particulier, on a $\langle v, x\rangle + \alpha J(x) > \beta$. Il est clair que $\alpha > 0$, puisque que l'épigraphe n'est pas borné en t pour x fixé (et que le cas $x = x_0$ permet d'exclure $\alpha = 0$). Donc $J(x) > \left\langle \frac{1}{\alpha}v, x\right\rangle + \frac{\beta}{\alpha}$.

– on montre ensuite que le caractère elliptique implique la coercitivité (et même $J(x) \geq \gamma\|x\|^2 - \lambda$ pour un certain couple $\gamma > 0$, λ). Comme pour l'étape précédente, on fixe x_0 du domaine de J. Par ellipticité et le résultat précédent, on a pour tout x

$$\begin{aligned} \frac{1}{2}J(x) + \frac{1}{2}J(x_0) - \frac{\alpha}{8}\|x - x_0\|^2 &\geq J\left(\frac{x+x_0}{2}\right) \\ &\geq \langle h, x/2\rangle + \langle h, x_0/2\rangle + \delta \end{aligned}$$

et donc

$$\begin{aligned} J(x) &\geq \frac{\alpha}{4}\|x - x_0\|^2 + \langle h, x\rangle + \langle h, x_0\rangle + \delta - J(x_0) \\ &\geq \frac{\alpha}{4}\|x\|^2 + \left\langle h - \frac{\alpha}{2}x_0, x\right\rangle + \mu, \end{aligned}$$

avec $\mu = \frac{\alpha}{4}\|x_0\|^2 + \langle h, x_0\rangle + \delta - J(x_0)$. D'après l'inégalité de Cauchy-Swartz, on a donc

$$J(x) \geq \frac{\alpha}{4}\|x\|^2 - (\|h\| + \frac{\alpha}{2}\|x_0\|)\|x\| + \mu,$$

ce qui permet de conclure à la coercitivité. De plus, il existe clairement une constante λ telle que

$$J(x) \geq \frac{\alpha}{8}\|x\|^2 - \lambda.$$

Fonctions elliptiques

- Caractérisations à l'ordre 1 :
  - si J est différentiable au sens de Gâteaux, J est α-convexe ssi :
    - $\forall x, y,\ J(y) \geq J(x) + \langle \nabla J(x), y - x\rangle + \frac{\alpha}{2}\|x - y\|^2$
    - $\forall x, y,\ \langle \nabla J(y) - \nabla J(x), y - x\rangle \geq \alpha\|x - y\|^2$
- Caractérisation à l'ordre 2 :
  - si J est C^2, J est α-convexe ssi :

$$\forall x, h,\ \left\langle \nabla J^2(x)h, h\right\rangle \geq \alpha\|h\|^2$$

- Preuves dans le même esprit que pour la caractérisation des fonctions convexes
- Exemple : $J(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$ avec A symétrique

Notes

- première inégalité à l'ordre 1 :
  $\Rightarrow$ passage à la limite sur t (même principe que pour convexité simple)
  $\Leftarrow$ combinaison d'applications de l'inégalité aux couples $\{x + t(y - x), x\}$ et $\{x + t(y - x), y\}$
- deuxième inégalité à l'ordre 1 :
  $\Rightarrow$ simple somme de l'inégalité 1 appliquée en (x, y) et en (y, x)
  $\Leftarrow$ on considère la fonction $\phi(t) = J(x + t(y - x))$. On a

$$\phi(1) = \phi(0) + \int_0^1 \phi'(t)dt.$$

Or, $\phi'(t) = \langle \nabla J(x + t(y - x))), y - x\rangle$, donc

$$\int_0^1 \phi'(t)dt - \langle \nabla J(x), y - x\rangle = \int_0^1 \langle \nabla J(x + t(y - x))) - \nabla J(x), y - x\rangle\, dt,$$

et donc, d'après la majoration

$$\int_0^1 \phi'(t)dt - \langle \nabla J(x), y - x\rangle \geq \int_0^1 \alpha t\|x - y\|^2 dt = \frac{\alpha}{2}\|x - y\|^2,$$

soit finalement

$$J(y) - J(x) \geq \langle \nabla J(x), y - x\rangle + \frac{\alpha}{2}\|x - y\|^2.$$

On conclut en appliquant utilisant la preuve de l'inégalité à l'ordre 2.

3.2 Conditions d'optimalité

Condition d'optimalité du 1er ordre

- si J Gâteaux différentiable en x^* qui réalise un minimum local de J, alors $\nabla J(x^*) = 0$
- preuve :
  - pour t assez petit $J(x^*) \leq J(x^* + th)$ et donc $J'(x, h) \geq 0$
  - par linéarité, $\forall h$, $\langle \nabla J(x^*), h \rangle = 0$ et donc $\nabla J(x^*) = 0$
  - la linéarité est indispensable, cf $J(u, v) = \frac{u^2 v^2}{(u^2+v^2)^{\frac{3}{2}}}$
- ce n'est pas une condition suffisante, par ex. $J(x) = x^3$ en zéro
- vocabulaire :
  - $\nabla J(x^*) = 0$ est l'**équation d'Euler**
  - les x^* tels que $\nabla J(x^*) = 0$ sont les **points critiques** ou **stationnaires**

Notes

$J'(x, h) \geq 0$ est évident, de même que $J'(x, -h) \geq 0$. Par linéarité, on a aussi $\langle \nabla J(x^*), -h \rangle \leq 0$, ce qui donne bien $\langle \nabla J(x^*), h \rangle = 0$ puis la conclusion.

Cas convexe

- si J est Gâteaux différentiable et convexe alors x^* réalise un minimum global si et seulement si $\nabla J(x^*) = 0$
- preuve : une fonction convexe est minorée globalement par son approximation linéaire locale
- extension :
  - J convexe n'est pas toujours différentiable, par ex. $J(x) = |x|$
  - un **sous-gradient** de J en x est un vecteur v tel que

$$\forall y,\ J(y) - J(x) \geq \langle v, y - x \rangle$$

  - la **sous-différentielle** de J en x est l'ensemble des sous-gradients, notée $\partial J(x)$ (dans $\mathbb{R}^n$, c'est un ensemble non vide, compact et convexe)
  - exemple $\partial |x|(0) = [-1, 1]$
  - CNS d'optimalité $0 \in \partial J(x^*)$

Notes

Pour tout y, on a

$$J(y) \geq J(x^*) + \langle \nabla J(x^*), y - x^* \rangle_H ,$$

donc si $\nabla J(x^*) = 0$, $J(x^*)$ est un minimum global.

Pour le cas non différentiable, la difficulté réside dans la preuve du caractère nécessaire de la condition $0 \in \partial J(x^*)$, son caractère suffisant étant aussi immédiat que dans le cas différentiable.

Condition d'optimalité du 2ème ordre

- si J est C^2 et si x^* réalise un minimum de J alors :
  - $\nabla J(x^*) = 0$
  - $\left\langle \nabla^2 J(x^*) h, h \right\rangle \geq 0$ pour tout h
- preuve par passage à la limite du développement limité à l'ordre 2
- ce n'est pas une condition suffisante, par ex. $J(x) = x^3$
- si J est C^2 et si
  - $\nabla J(x^*) = 0$
  - il existe $\alpha > 0$ tel que $\left\langle \nabla^2 J(x^*) h, h \right\rangle \geq \alpha \|h\|^2$ ($\nabla^2 J(x)$ est défini positif)
  - alors x^* réalise un minimum local (strict) de J

- preuve identique
- **ellipticité locale**
- ce n'est pas une condition nécessaire, par ex. $J(x) = x^4$

Notes

- CN : on a
$$J(x+h) = J(x) + \langle \nabla J(x), h \rangle + \frac{1}{2} \left\langle \nabla^2 J(x)h, h \right\rangle + o(\|h\|^2).$$
En x^* un minimum, on a $\nabla J(x^*) = 0$ et donc
$$J(x^* + th) = J(x^*) + \frac{1}{2} t^2 \left\langle \nabla^2 J(x^*)h, h \right\rangle + o(t^2\|h\|^2),$$
soit donc pour t suffisamment petit,
$$t^2 \left\langle \nabla^2 J(x^*)h, h \right\rangle + o(t^2\|h\|^2) \geq 0,$$
ce qui conduit à $\left\langle \nabla^2 J(x^*)h, h \right\rangle \geq 0$.
- CS : les hypothèses conduisent à
$$J(x^* + h) - J(x^*) \geq \frac{\alpha}{2}\|h\|^2 + o(\|h\|^2),$$
et donc $J(x^* + h) - J(x^*) \geq 0$ pour h suffisamment petit.

4 Algorithmes

Principes
- algorithmes numériques :
  - pas de résultat exact
  - notion de convergence
  - coût par itération
- hypothèses algorithmiques :
  - fonctions plus ou moins régulières
  - compromis coût d'une itération et nombre d'itérations
- hypothèses mathématiques :
  - pour garantir la convergence
  - convexité plus ou moins forte

4.1 Fonctions définies sur $\mathbb{R}$

Section dorée
- minimisation de f de $\mathbb{R}$ dans $\mathbb{R}$
- outil de base : permet de construire des algorithmes pour J définie sur $\mathbb{R}^n$
- méthode de la section dorée :
  - trois points initiaux tels que $f(x_1) > f(x_2)$ et $f(x_3) > f(x_2)$
  - on cherche le minimum dans $]x_1, x_3[$
  - on évalue f en x_4 choisit dans le plus grand de deux intervalles déterminés par x_2 par exemple $[x_1, x_2]$
  - on fonction de $f(x_4)$:
    - on passe dans $[x_4, x_3]$ si $f(x_4) > f(x_2)$
    - sinon on passe dans $[x_1, x_2]$
  - réduction optimale de la longueur de l'intervalle de recherche si le grand sous intervalle est $\frac{1+\sqrt{5}}{2}$ plus long que le petit (nombre d'or)

Méthode de Newton

- méthode de recherche d'un zéro d'une fonction f :
  - approximation linéaire de f

$$f(x+h) = f(x) + hf'(x) + o(h)$$

  - au premier ordre, $f(x+h) = 0$ conduit à $h = -\frac{f(x)}{f'(x)}$
  - algorithme : $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$
  - convergence quadratique : $|x_{n+1} - x^*| \leq \mu |x_{n+1} - x^*|^2$
- application à la minimisation de f :
  - on cherche à résoudre l'équation d'Euler $f'(x) = 0$
  - algorithme

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

4.2 Fonctions définies sur $\mathbb{R}^n$

Algorithme de relaxation

- méthode naïve : optimisations successives par rapport à chaque variable
- algorithme :
  1. point initial $\mathbf{x}^0$ (avec $J(\mathbf{x}^0) < \infty$)
  2. pour $k \geq 1$ croissant
     (a) pour i allant de 1 à n :

$$x_i^{k+1} = \arg\min_{x \in \mathbb{R}} J(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x, x_{i+1}^k, \ldots, x_n^k)$$

     (b) tester la convergence $\|\mathbf{x}^{k+1} - \mathbf{x}^k\| < \epsilon$
- remarque : calcul du minimum dans $\mathbb{R}$ par un algorithme adapté
- *Attention* : très mauvais algorithme en pratique (mais analysable théoriquement...)

Convergence

- on suppose :
  - J *coercive*,
  - J C^1 et α*-convexe*,
  - ∇J M*-Lipschitzienne* :

$$\forall \mathbf{x}, \mathbf{y} \ \|\nabla J(\mathbf{x}) - \nabla J(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\|$$

- alors l'algorithme de relaxation converge vers le minimum de J

Notes

Pour simplifier la preuve, on définit $\mathbf{x}^{k,i} = \left(x_1^{k+1}, \ldots, x_i^{k+1}, x_{i+1}^k, \ldots, x_n^k\right)$ pour $1 \leq i \leq n$ (et donc $\mathbf{x}^{k,n} = \mathbf{x}^{k+1}$) Par extension $\mathbf{x}^{k,0} = \mathbf{x}^k$.

- par construction $J(\mathbf{x}^{k+1}) \leq J(\mathbf{x}^k)$, donc comme J est coercive, $(\mathbf{x}^k)_k$ est bornée ;
- les $(\mathbf{x}^k)_k$ sont donc un fermé borné de $\mathbb{R}^n$. Par continuité de J (impliquée par la convexité), $J(\mathbf{x}^k)_k$ est donc bornée et converge ;
- on montre ensuite que $\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \to 0$. Considérons la fonction

$$x \mapsto g(x) = J(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x, x_{i+1}^k, \ldots, x_n^k).$$

  Par définition x_i^{k+1} minimise g. Comme J est C^1, g l'est aussi et on a donc $g'(x_i^{k+1}) = 0$, soit

$$\frac{\partial J}{\partial x_i}(\mathbf{x}^{k,i}) = 0.$$

Or, $\mathbf{x}^{k,i} - \mathbf{x}^{k,i-1} = (0, \ldots, 0, x_i^{k+1} - x_i^k, 0, \ldots, 0)^T$ et donc $\left\langle \nabla J(\mathbf{x}^{k,i}), \mathbf{x}^{k,i} - \mathbf{x}^{k,i-1} \right\rangle = 0$. Comme J est α-convexe et C^1, on a

$$\begin{aligned} J(\mathbf{x}^{k,i-1}) &\geq J(\mathbf{x}^{k,i}) + \left\langle \nabla J(\mathbf{x}^{k,i}), \mathbf{x}^{k,i} - \mathbf{x}^{k,i-1} \right\rangle + \frac{\alpha}{2}\|\mathbf{x}^{k,i} - \mathbf{x}^{k,i-1}\|^2 \\ &\geq J(\mathbf{x}^{k,i}) + \frac{\alpha}{2}(x_i^{k+1} - x_i^k)^2 \end{aligned}$$

En sommant les inégalités pour i allant de 1 à n, on obtient

$$\begin{aligned} J(\mathbf{x}^{k,0}) &\geq J(\mathbf{x}^{k,n}) + \frac{\alpha}{2}\sum_{i=1}^{n}(x_i^{k+1} - x_i^k)^2 \\ J(\mathbf{x}^k) &\geq J(\mathbf{x}^{k+1}) + \frac{\alpha}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2. \end{aligned}$$

La convergence de $J(\mathbf{x}^k)$ implique donc celle de $\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2$ vers 0.

– Soit $\mathbf{x}^*$ le point qui réalise le minimum de J. Par α-convexité, on a

$$\left\langle \nabla J(\mathbf{x}^k) - \nabla J(\mathbf{x}^*), \mathbf{x}^k - \mathbf{x}^* \right\rangle \geq \alpha\|\mathbf{x}^k - \mathbf{x}^*\|^2,$$

et comme $\nabla J(\mathbf{x}^*) = 0$, on obtient grâce à l'inégalité de Cauchy-Schwarz

$$\alpha\|\mathbf{x}^k - \mathbf{x}^*\| \leq \|\nabla J(\mathbf{x}^k)\|$$

– Par optimalité de x_i^{k+1},

$$\|\nabla J(\mathbf{x}^k)\|^2 = \sum_{i=1}^{n}\left(\frac{\partial J}{\partial x_i}(\mathbf{x}^k) - \frac{\partial J}{\partial x_i}(\mathbf{x}^{k,i})\right)^2.$$

Or, comme ∇J est M-Lipschitzienne, on a donc

$$\left(\frac{\partial J}{\partial x_i}(\mathbf{x}^k) - \frac{\partial J}{\partial x_i}(\mathbf{x}^{k,i})\right)^2 \leq \|\nabla J(\mathbf{x}^k) - \nabla J(\mathbf{x}^{k,i})\|^2 \leq M^2(x_i^{k+1} - x_i^k)^2$$

et donc

$$\|\nabla J(\mathbf{x}^k)\|^2 \leq M^2\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2.$$

On en conclut donc $\|\nabla J(\mathbf{x}^k)\|^2 \to 0$ puis la convergence de l'algorithme.

Remarque : on a

$$\|\mathbf{x}^k - \mathbf{x}^*\| \leq \frac{M}{\alpha}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|,$$

ce qui justifie le critère d'arrêt. De plus, $\frac{M}{\alpha}$ est le conditionnement de la matrice H.

Méthodes de descentes

– principe général :
  1. point initial $\mathbf{x}_0$
  2. pour $k \geq 1$ croissant
     (a) choisir une **direction de descente** $\mathbf{d}_k \neq \mathbf{0}$
     (b) choisir un **pas de descente** $\rho_k > 0$
     (c) poser $\mathbf{x}_{k+1} = \mathbf{x}_k + \rho_k\mathbf{d}_k$
     (d) tester la convergence (par ex. $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < \epsilon$)
– il faut qu'on puisse descendre :
  – on doit pouvoir trouver ρ_k tel que $J(\mathbf{x}_{k+1}) < J(\mathbf{x}_k)$
  – si J est convexe, $J(\mathbf{x}_{k+1}) \geq J(\mathbf{x}_k) + \rho_k \langle v, \mathbf{d}_k \rangle$ pour tout sous-gradient v : on doit donc avoir $\langle v, \mathbf{d}_k \rangle < 0$ pour au moins un sous-gradient v
– solution classique :
  – algorithme du **gradient**
  – $\mathbf{d}_k = -\nabla J(\mathbf{x}_k)$

Convergence

- L'algorithme du gradient converge sous des conditions assez fortes :
  - J est C^1, coercitive et strictement convexe
  - ∇J est M-Lipschitzienne
  - si $\rho_k \in [\beta_1, \beta_2]$ avec $0 < \beta_1 < \beta_2 < \frac{2}{M}$
- On peut obtenir une preuve plus simple en supposant que J est α-convexe C^1 et de gradient M-Lipschitzien :
  - on suppose alors que $\rho_k \in]0, \frac{2\alpha}{M^2}[$
  - on obtient une vitesse de convergence linéaire

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \sqrt{1 - 2\alpha\rho_k + M^2\rho_k^2}\|\mathbf{x}_k - \mathbf{x}^*\|,$$

  optimale de valeur $\sqrt{1 - \frac{\alpha^2}{M^2}}$ (pour $\rho_k = \frac{\alpha}{M^2}$)

Notes

Preuve du premier cas :

- comme J est coercitive, C^1 et strictement convexe, il existe un unique minimum atteint en $\mathbf{x}^*$ et unique solution de l'équation d'Euler.
- On sait qu'on a (cf au dessus) :

$$J(\mathbf{x}_{k+1}) = J(\mathbf{x}_k) + \int_0^1 \langle \nabla J(\mathbf{x}_k + t(\mathbf{x}_{k+1} - \mathbf{x}_k)), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle \, dt.$$

  Or, par Cauchy Schwarz et comme ∇J est M-Lipschitzienne,

$$\begin{aligned} |\langle \nabla J(\mathbf{x}_k + t(\mathbf{x}_{k+1} - \mathbf{x}_k)) - \nabla J(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle| &\leq \|J(\mathbf{x}_k + t(\mathbf{x}_{k+1} - \mathbf{x}_k)) - \nabla J(\mathbf{x}_k)\|\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \\ &\leq tM\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2. \end{aligned}$$

  On a donc

$$J(\mathbf{x}_{k+1}) - J(\mathbf{x}_k) \leq \langle \nabla J(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{M}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2.$$

  Par définition, $\langle \nabla J(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle = -\rho_k\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$ et donc d'après les hypothèses sur ρ_k,

$$J(\mathbf{x}_{k+1}) - J(\mathbf{x}_k) \leq -\left(\frac{1}{\beta_2} - \frac{M}{2}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2,$$

  donc avec $\kappa = -\left(\frac{1}{\beta_2} - \frac{M}{2}\right) < 0$,

$$J(\mathbf{x}_{k+1}) - J(\mathbf{x}_k) \leq \kappa\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$$

- la minoration obtenue au dessus montre que la suite des $J(\mathbf{x}_k)$ décroît strictement. Comme J est minorée, la suite converge.
- les $\mathbf{x}_k$ sont bornés par coercitivité.
- la minoration permet aussi de conclure directement à la convergence de $\mathbf{x}_{k+1} - \mathbf{x}_k$ vers 0 en écrivant :

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \leq -\kappa^{-1}\left(J(\mathbf{x}_k) - J(\mathbf{x}_{k+1})\right).$$

  Or, par définition, $\nabla J(\mathbf{x}_k) = \frac{\mathbf{x}_{k+1} - \mathbf{x}_k}{\rho_k}$. La borne inférieure sur ρ_k permet de conclure que $\nabla J(\mathbf{x}_k)$ tend vers 0.
- L'unique valeur d'adhérence des $\mathbf{x}_k$ est alors $\mathbf{x}^*$ ce qui permet de conclure à la convergence de $\mathbf{x}_k$ vers $\mathbf{x}^*$.

Preuve avec des hypothèses plus fortes :

- comme J est elliptique, elle possède un unique minimum $\mathbf{x}^*$. On note $\mathbf{u}_k = \mathbf{x}_k - \mathbf{x}^*$. On a

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \rho_k(\nabla J(\mathbf{x}_k) - \nabla J(\mathbf{x}^*)),$$

  et donc

$$\|\mathbf{u}_{k+1}\|^2 = \|\mathbf{u}_k\|^2 + \rho_k^2\|\nabla J(\mathbf{x}_k) - \nabla J(\mathbf{x}^*)\|^2 - 2\rho_k\langle \mathbf{x}_k - \mathbf{x}^*, \nabla J(\mathbf{x}_k) - \nabla J(\mathbf{x}^*)\rangle$$

– Comme ∇J est M-Lipschitzienne, on a

$$\|\nabla J(\mathbf{x}_k) - \nabla J(\mathbf{x}^*)\| \leq M\|\mathbf{u}_k\|.$$

Comme J est α-convexe, on a

$$\langle \mathbf{x}_k - \mathbf{x}^*, \nabla J(\mathbf{x}_k) - \nabla J(\mathbf{x}^*)\rangle \geq \alpha\|\mathbf{u}_k\|^2.$$

Donc

$$\|\mathbf{u}_{k+1}\|^2 \leq (1 - 2\alpha\rho_k + M^2\rho_k^2)\|\mathbf{u}_k\|^2$$

– alors $\rho_k \in]0, \frac{2\alpha}{M^2}[$ implique la convergence :
  – la convergence est linéaire $\|\mathbf{u}_{k+1}\| \leq \sqrt{1 - 2\alpha\rho_k + M^2\rho_k^2}\|\mathbf{u}_k\|$
  – vitesse optimale de $\sqrt{1 - \frac{\alpha^2}{M^2}}$ (pour $\rho_k = \frac{\alpha}{M^2}$)
  – liée au conditionnement du Hessien

Choix du pas

– pas constant
– pas variable :
  – adaptation du pas au problème
  – pas optimal :
    – $\rho_k = \arg\min_{\rho>0} J(\mathbf{x}_k - \rho\nabla J(\mathbf{x}_k))$
    – avantage : meilleure réduction possible par itération
    – inconvénient : coût de la recherche
  – pas approximativement optimal :
    – rechercher un ρ_k qui réduit « assez » J
    – comme $J(\mathbf{x}_k - \rho\nabla J(\mathbf{x}_k)) = J(\mathbf{x}_k) - \rho\|\nabla J(\mathbf{x}_k)\|^2 + o(\rho\|\nabla J(\mathbf{x}_k)\|)$, on peut demander une réduction d'au moins $\alpha\rho\|\nabla J(\mathbf{x}_k)\|^2$ (avec $\alpha < \frac{1}{2}$)

Notes

La preuve de la convergence de l'algorithme à pas optimal, sous les hypothèses classiques (J est α-convexe C^1 et de gradient M-Lipschitzien) est très proche de celle utilisée pour l'algorithme de relaxation :

– par construction $J(\mathbf{x}_{k+1}) \leq J(\mathbf{x}_k)$, donc dès que J est minorée, $J(\mathbf{x}_k)$ converge (c'est le cas ici, puisque que J est α-convexe et possède donc un minimum unique)
– comme J est coercive (car α-convexe), $(\mathbf{x}_k)$ est bornée
– comme ρ_k réalise le minimum de $f(\rho) = J(\mathbf{x}_k - \rho\nabla J(\mathbf{x}_k))$, $f'(\rho_k) = 0$, donc

$$\langle \nabla J(\mathbf{x}_{k+1}), \nabla J(\mathbf{x}_k)\rangle = 0$$

– comme J est α-convexe et que $\langle \nabla J(\mathbf{x}_{k+1}), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle = 0$ on a

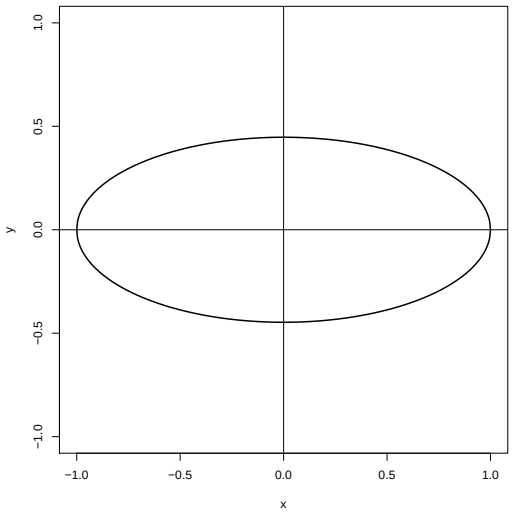
$$J(\mathbf{x}_k) - J(\mathbf{x}_{k+1}) \geq \frac{\alpha}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$$

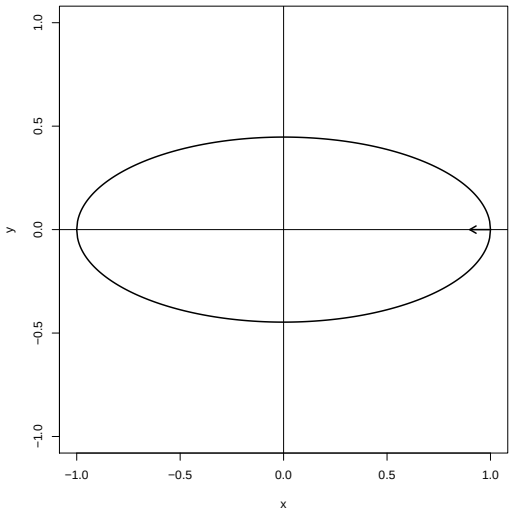
donc $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \to 0$
– comme ∇J est M-Lipschitzienne, $\|\nabla J(\mathbf{x}_k)\| \leq M\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$ et donc $\nabla J(\mathbf{x}_k) \to \mathbf{0}$
– de ce fait, toute valeur d'adhérence de $(\mathbf{x}_k)$ est le minimum $\mathbf{x}^*$, ce qui assure la convergence de l'algorithme
– on a de plus $\alpha\|\mathbf{x}_k - \mathbf{x}^*\| \leq \|\nabla J(\mathbf{x}_k)\|$ (par α-convexité et Cauchy Schwarz), ce qui donne une majoration de l'erreur de la forme $\frac{M}{\alpha}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|$, comme dans le cas de l'algorithme de relaxation

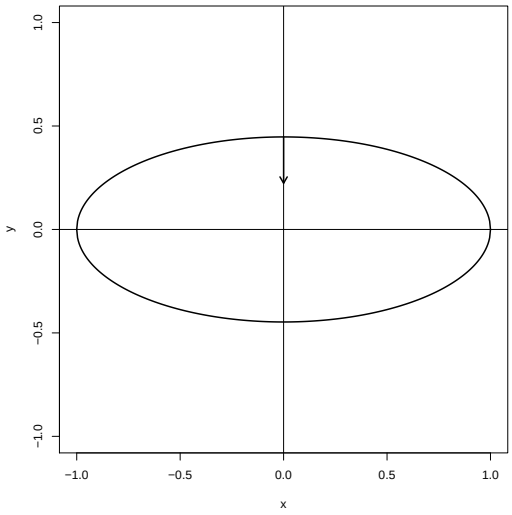
---

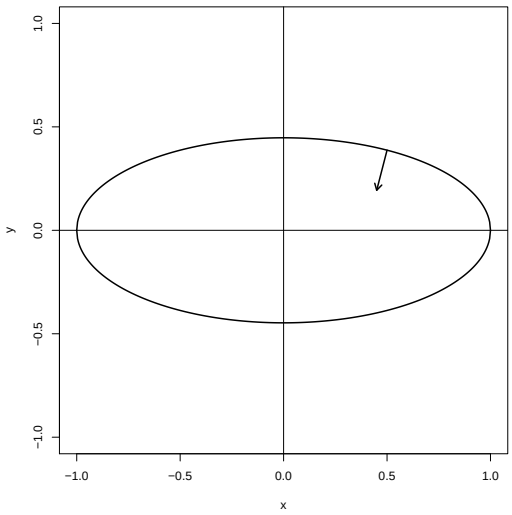
Gradient conjugué

- Analyse de l'algorithme du gradient sur le cas particulier des formes quadratiques $J(\mathbf{x}) = \frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle - \langle \mathbf{b}, \mathbf{x}\rangle$ pour A définie positive
- si A est mal conditionnée, l'algorithme du gradient est mauvais









Méthode du gradient conjugué

- idée fondamentale : utiliser des directions de descentes *conjuguées* c.-à-d. orthogonales au sens de A
- autrement dit : ne pas perturber l'optimisation précédente
- principe général :
  - algorithme de descente à pas optimal
  - directions de descente obtenues par récurrence à partir des gradients aux points précédents
- applicable à des fonctions quelconques (pas seulement quadratiques)

Algorithme

1. point initial $\mathbf{x}_0$, gradient initial $\mathbf{g}_0 = A\mathbf{x}_0 - b$, direction de descente initiale $\mathbf{w}_0 = \mathbf{g}_0$ (on suppose que $\mathbf{g}_0 \neq \mathbf{0}$)
2. pour $k \geq 1$ croissant
   (a) $\rho_{k-1} = \frac{\langle \mathbf{g}_{k-1}, \mathbf{w}_{k-1} \rangle}{\langle A\mathbf{w}_{k-1}, \mathbf{w}_{k-1} \rangle}$
   (b) $\mathbf{x}_k = \mathbf{x}_{k-1} - \rho_{k-1}\mathbf{w}_{k-1}$
   (c) $\mathbf{g}_k = A\mathbf{x}_k - b$
   (d) $\mathbf{g}_k = \mathbf{0}$ fin de l'algorithme
   (e) sinon :
      i. $\alpha_k = -\frac{\langle \mathbf{g}_k, A\mathbf{w}_{k-1} \rangle}{\langle A\mathbf{w}_{k-1}, \mathbf{w}_{k-1} \rangle}$
      ii. $\mathbf{w}_k = \mathbf{g}_k + \alpha_k \mathbf{w}_{k-1}$

- ρ_k est un pas optimal
- α_k est tel que $\langle \mathbf{w}_k, A\mathbf{w}_{k-1} \rangle = 0$

Propriétés

- pour tout $l < k$:
  - $\langle \nabla J(\mathbf{x}_k), \nabla J(\mathbf{x}_l) \rangle = 0$
  - $\langle \nabla J(\mathbf{x}_k), \mathbf{w}_l \rangle = 0$
  - $\langle \mathbf{w}_k, A\mathbf{w}_l \rangle = 0$
- l'algorithme trouve le minimum de $J(\mathbf{x}) = \frac{1}{2}\langle \mathbf{x}, A\mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle$ en au plus n itérations (n est l'ordre de A)
- erreurs numériques : critère d'arrêt sur $\|\nabla J(\mathbf{x}_k)\|$

Cas non linéaire

1. point initial $\mathbf{x}_0$, gradient initial $\mathbf{g}_0 = \nabla J(\mathbf{x}_0)$, direction de descente initiale $\mathbf{w}_0 = \mathbf{g}_0$ (on suppose que $\mathbf{g}_0 \neq \mathbf{0}$)
2. pour $k \geq 1$ croissant
   (a) $\rho_{k-1} = \arg\min_{\rho>0} J(\mathbf{x}_{k-1} - \rho \mathbf{w}_{k-1})$
   (b) $\mathbf{x}_k = \mathbf{x}_{k-1} - \rho_{k-1}\mathbf{w}_{k-1}$
   (c) $\mathbf{g}_k = \nabla J(\mathbf{x}_k)$
   (d) $\mathbf{g}_k = \mathbf{0}$ fin de l'algorithme
   (e) sinon :
      i. choisir un α_k
      ii. $\mathbf{w}_k = \mathbf{g}_k + \alpha_k \mathbf{w}_{k-1}$

– Diverses formules pour le calcul de α_k, par exemple $\alpha_k = \frac{\|\mathbf{g}_k\|^2}{\|\mathbf{g}_{k-1}\|^2}$ (Fletcher-Reeves)
– « Redémarrage » : $\mathbf{w}_k = \mathbf{g}_k$ de temps en temps

Méthode de Newton

– on cherche un zéro de ∇J, or

$$\nabla J(\mathbf{x} + \mathbf{h}) = \nabla J(\mathbf{x}) + \nabla^2 J(\mathbf{x})\mathbf{h} + o(\|\mathbf{h}\|)$$

donc au premier ordre

$$\mathbf{h} = -\nabla^2 J(\mathbf{x})^{-1} \nabla J(\mathbf{x})$$

– correspond aussi au minimum de l'approximation au deuxième ordre

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

– si $\nabla^2 J(\mathbf{x})$ est définie positive, $\mathbf{h}$ est une direction de descente

$$\left\langle \nabla J(\mathbf{x}), -\nabla^2 J(\mathbf{x})^{-1} \nabla J(\mathbf{x}) \right\rangle < 0$$

Notes

On considère en effet la fonction

$$U(\mathbf{h}) = J(\mathbf{x}) + \mathbf{h}^T \nabla J(\mathbf{x}) + \frac{1}{2}\mathbf{h}^T \nabla^2 J(\mathbf{x})\mathbf{h}.$$

On a alors

$$\nabla U(\mathbf{h}) = \nabla J(\mathbf{x}) + \nabla^2 J(\mathbf{x})\mathbf{h},$$

et donc $\nabla U(\mathbf{h}) = 0$ donne

$$\mathbf{h} = -\nabla^2 J(\mathbf{x})^{-1} \nabla J(\mathbf{x})$$

Algorithmes

– Newton pur :
   1. point initial $\mathbf{x}_0$
   2. pour $k \geq 1$ croissant
      (a) calculer $\mathbf{d}_k = -\nabla^2 J(\mathbf{x_k})^{-1} \nabla J(\mathbf{x_k})$
      (b) tester la convergence par $\left\langle \nabla J(\mathbf{x_k}), \nabla^2 J(\mathbf{x_k})^{-1} \nabla J(\mathbf{x_k}) \right\rangle < \epsilon$
      (c) poser $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$
– variante dite « gardée » :
   – mise à jour par $\mathbf{x}_{k+1} = \mathbf{x}_k + \rho_k \mathbf{d}_k$
   – recherche de $\rho_k \leq 1$ par réduction suffisante
– on peut montrer la convergence de l'algorithme gardé sous les hypothèses classiques (α-convexité) et si le Hessien est Lipschitzien

Résumé

- algorithme du gradient :
  - coût en $O(n)$
  - grand nombre d'itérations
- gradient conjugué :
  - coût en $O(n)$
  - nombre d'itérations beaucoup plus faible que le gradient simple
- méthode Newton :
  - coût en $O(n^3)$
  - nombre d'itérations faible (convergence quadratique près du minimum)
- solutions intermédiaires :
  - méthodes de quasi-Newton (par ex. BFGS)
  - approximation de $\nabla^2 J(\mathbf{x})^{-1}$
  - coût en $O(n^2)$ ou même $O(n)$
  - convergence rapide