Master MIMSE - Année 2

Optimisation Quadratique

Optimisation quadratique sans contraintes

Optimisation quadratique

- Fonction quadratique = polynôme de degré 2,
- On veut

$$\begin{aligned} \mathbf{min} \;\; & \boldsymbol{f(x)} \\ \boldsymbol{s.c.} \;\; & \boldsymbol{g_k(x) \geq 0, \; \forall k} \\ & \boldsymbol{x \in X \subseteq \Re^n.} \end{aligned}$$

- **Intérêt ?**
  - Modélisation de certains problèmes est déjà de degré 2 (par ex. en optimisation stochastique)
  - Contient la Programmation Linéaire
  - Mais est bien plus riche...
  - ...et en tout cas plus difficile...
  - Contient la PLNE ($\boldsymbol{x \in \{0,1\} \leftrightarrow x^2 - x = 0}$)
  - Problèmes d'identification de paramètres : moindres carrés
  - En finance : matrice de covariance des risques SDP, problèmes d'optimisation associés quadratiques
  - Méthodes utilisables pour des fonctions même non quadratiques.

Fonctions quadratiques à une variable

- On veut optimiser une fonction $f : \Re \to \Re$ sur un espace $X \subseteq \Re$.
- X est un intervalle ou une collection d'intervalles.
- On suppose que f est quadratique : c'est un polymôme de degré au plus 2.
- f est indéfiniment dérivable X, mais de toute façon $f^{(3)} = 0$.
- L'optimum est soit sur une borne d'un des intervalles,
- soit en un point $\hat{x}$ de l'intérieur de X vérifiant :

$$f'(\hat{x}) = 0.$$

- De plus comme f est deux fois dérivable,
  - $f''(\hat{x}) > 0$ implique un minimum local,
  - $f''(\hat{x}) < 0$ implique un maximum local.

Application aux fonctions quadratiques

- Pour $f(x) = ax^2 + bx + c$,
- $f'(x) = 2ax + b$
- et $f''(x) = 2a$.

- $a > 0$ entraîne
  f admet un minimum en $\hat{x} = -\frac{b}{2a}$ **si c'est dans** X,
- ou au point de X le plus proche de $\hat{x}$.

- $a < 0$ entraîne
  f admet un minimum en un des points extrèmes de X.

Fonctions à plusieurs variables

- Conditions analogues au cas à une variable, mais qui font appel aux dérivées partielles et aux gradient.
- Optimums toujours éventuellement présents à la frontière de $\boldsymbol{X}$.
- Soit $\boldsymbol{x} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$.
- Si $\boldsymbol{f}$ quadratique, $\frac{\partial f}{\partial x_i}(\boldsymbol{x})$ existe pour tout $\boldsymbol{x}$ et tout $\boldsymbol{i}$.
- $\boldsymbol{\nabla f(x)} = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_2})(\boldsymbol{x})$ gradient de $\boldsymbol{f}$ en $\boldsymbol{x}$.
- Si $\boldsymbol{f}$ a un extremum local en $\hat{\boldsymbol{x}}$ dans l'intérieur de $\boldsymbol{X}$,
- Alors $\frac{\partial f}{\partial x_i}(\hat{\boldsymbol{x}}) = \boldsymbol{0}\ \forall \boldsymbol{i}$
- Ou encore $\boldsymbol{\nabla f(\hat{x}) = 0.}$
- $\boldsymbol{f}$ est deux fois différentiable,
- Soit $\boldsymbol{\nabla^2 f}$ sa matrice hessienne (matrice des $\frac{\partial^2 f}{\partial x_i \partial x_j}$).
- Si $\boldsymbol{\nabla^2 f(\hat{x}) \succ 0}$ (toutes ses valeurs propres > 0) alors minimum local,
- Si $\boldsymbol{-\nabla^2 f(\hat{x}) \succ 0}$ alors maximum local.
- En général, ni l'un ni l'autre : **point-selle !**

Fonctions à plusieurs variables - Convexité

- On dit qu'un espace S est convexe si
  - pour toute paire de points x, y de S
  - et tout $0 \leq \alpha \leq 1$,
  - le point $\alpha x + (1 - \alpha)y \in S$.
- Le segment $[x, y] \subset S$.
- Une fonction f définie sur S convexe est convexe si pour toute paire de points x, y de S et tout $0 \leq \alpha \leq 1$,

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

- De plus,

$$f(x) - f(y) \geq \nabla f(x)^T (x - y).$$

- La courbe de f est en-dessous de ses cordes et au-dessus de ses tangentes/plans tangents.

- Si f est convexe sur X convexe et si $\hat{x}$ est un minimum local de f alors $\hat{x}$ est un minimum global de f.
- Si f est concave sur X convexe et si $\hat{x}$ est un maximum local de f alors $\hat{x}$ est un maximum global de f.
- Si $\forall x \ \nabla^2 f(x) \succ 0$ alors f est convexe.

- f convexe ssi $(-f)$ est concave.

Matrices définies positives et semi-définies positives

- On considère une matrice M carrée et symétrique, de taille $n \times n$.
- M est semidéfinie positive, noté $M \succeq 0$
  ssi $\forall u \in \Re^n$
  $$u^T M u \geq 0.$$
- Soient $Sp(M) = \{\lambda_1; \ldots; \lambda_n\}$ l'ensemble des valeurs propres de M.
- M symétrique entraîne $Sp(M) \subset \Re$.
- M SDP ssi $Sp(M) \subset \Re^+$.
- Toutes les valeurs propres doivent être ≥ 0.
- (Th. de Gram)
  $$M \succeq 0 \Leftrightarrow \exists U \text{ t.q } M = U^T U.$$
- $M \succ 0$ ("définie positive") ssi
- $\forall u \in \Re^n,\ u^T M u > 0$
- ou encore $Sp(M) \subset \Re^{+*}$.
- alors $Tr(M) > 0$ et $det(\hat{M}) > 0$
- (Lemme de Schur)
  $$\left(\begin{array}{c|c} k & b^T \\ \hline b & M \end{array}\right) \succ 0 \text{ ssi } k > 0 \text{ et } M - \frac{1}{k} b b^T \succ 0.$$

Cas d'une fonction quadratique sur $\Re^n$ -1-

- On considère le cas d'une fonction quadratique

$$f(x) = \sum a_i x_i^2 + \sum\sum a_{ij} x_i x_j + \sum b_i x_i + c.$$

- On veut minimiser f sur $\Re^n$.
- On peut écrire f sous la forme

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

avec $A_{ii} = 2a_i$ et $A_{ij} = A_{ji} = a_{ij}$.

- $\nabla f = Ax + b.$
- $\nabla^2 f = A.$

Si $A \succ 0$

- alors f strictement convexe,
- f admet un minimum unique au point $x^* = -A^{-1}b$ annulant le gradient.
- $f(x^*) = \frac{1}{2} b^T A^{-1} b + c.$

Cas d'une fonction quadratique sur $\Re^n$ -2-

Si $A \not\succeq 0$

- alors f pas convexe,
- il existe une valeur propre $\lambda < 0$
- si u est le vecteur propre associé,
- $f(tu) \longrightarrow -\infty$
- f n'admet pas de minimum !!

Si $A \succeq 0$ et $detA = 0$

- Il faut regarder si $kerA$ est un s.e.v. de $Vect\{b\}^T$!
- Si oui, il existe un minimum,
- Sinon il existe u tel que $f(tu) \longrightarrow -\infty$.

Autour du calcul exact...

- On a l'écriture exacte de l'optimum, le cours est fini alors... ?
- Il faut voir si $\nabla^2 f$ est ou non défini positif,
- Résoudre un système de n équations à n inconnues, ou
- Inverser une matrice $n \times n$.
- Matrice des co-déterminants : hyperlourd
- Pivot : OK, mais
- le mauvais conditionnement de $\nabla^2 f$ peut causer des erreurs de calcul qui se répercutent
- Si $(\lambda_{max} - \lambda_{min})/\lambda_{min}$ est grand.
- Difficile si problème de grande taille.
- On peut chercher à s'approcher de l'optimum dans calculer A^{-1}.

Algorithmes de descente

- Pour trouver un minimum local de f quelconque, on a des algorithmes de descente.
- Principe :
  1. Point initial x_0.
  2. Trouver une direction de descente d : une direction telle que $d^T \nabla f(x_n) < 0$.
  3. Eventuellement calculer un pas de descente ρ.
  4. $x_{n+1} = x_n + \rho d$.
  5. Si $\nabla f(x_{n+1}) \neq 0$ réitérer.
- ρ peut être fixe, mais ρ petit : convergence longue, ρ grand : risque de diverger.
- ρ décroissant selon un schéma donné.
- ρ obtenu par recherche linéaire “le meilleur ρ le long de cette direction”.

Recherche linéaire

- But : Minimiser selon ρ la fonction $f(x_n + \rho d)$.
- **Approche 1 :** exprimer $\phi(\rho) = f(x_n + \rho d)$ et optimiser ϕ. On doit tout recalculer quand on change x_n, risques d'erreur, difficultés à programmer.
- **Approche 2 :** il existe des méthodes pour optimiser $f(x_n + \rho d)$ sans avoir à l'exprimer explicitement, par réduction d'intervalles.
- On suppose ϕ **unimodale** : elle décroît jusqu'à ρ^* puis elle recroît.
- Méthode de Fibonacci :

$$s_i = a_i + (b_i - a_i) A_{n-i} / A_{n+2-i}$$

$$\bar{s_i} = a_i + (b_i - a_i) A_{n+1-i} / A_{n+2-i}$$

  avec les A_i les éléments de la suite de Fibonacci 0,1,1,2,3,5,...
- On évalue aux points s_i et $\bar{s_i}$ et on réduit l'intervalle.
- Méthode de la section dorée : on approche les points précédents avec la formule :

$$s_i = a_i + (b_i - a_i) \frac{3 - \sqrt{5}}{2}$$

$$\bar{s_i} = a_i + (b_i - a_i) \frac{\sqrt{5} - 1}{2}$$

- **Approche 3 :** Théorème fondamental :

$$\nabla f(x_n + \rho^* d)^T d = 0.$$

Algorithme de plus forte pente

- On montre que la direction de descente qui fait décroître le plus vite $f(x_n + \rho d)$ est

$$d = -\nabla f(x_n).$$

- Algo très basique.
- En pratique, convergence pas toujours très bonne.
- Si les x_n convergent vers x^* et que f est C^1 sur un voisinage de x^* alors x^* est un minimum local pour f.
- Donc pour f quadratique convexe, le minimum global.

Gradient accéléré

- En pratique, convergence du gradient pas toujours très bonne : présence de “crète”, phénomène de “zigzag” : voir exemple graphique.
- Algo du gradient accéléré :
  - Point courant $\boldsymbol{x_n}$. On pose $\boldsymbol{y_0 = x_n}$
  - on fait plusieurs itérations du gradient : $\boldsymbol{y_1 \ldots y_p}$,
  - On pose $\boldsymbol{d_n = y_p - y_0}$ la “vraie” direction de recherche.
  - $\boldsymbol{x_{n+1} = x_n + \rho d_n}$ etc.

Méthode de Newton-Raphson

- On peut aussi calculer de façon approchée une solution de $\nabla f(x) = 0$ par la méthode de Newton-Raphson :

$$x_1,\ x_{n+1} = x_n - [\nabla^2 f(x_n)]^{-1} \nabla f(x_n),$$

- Vient du développement de Taylor à l'ordre 1 de ∇f.
- f non quadratique : on approche (localement) par une fonction quadratique.
- On peut introduire un pas de recherche ρ.
- f quadratique : x_2 est toujours optimal !
- Attention ! Il faut pouvoir calculer l'inverse de $\nabla^2 f$ (conditionnement de $\nabla^2 f$, erreurs de calcul...)

Algorithmes de quasi-Newton

- Difficultés calculatoires : on approche la méthode (Quasi-Newton).
- Ex. Davidon-Fletcher-Powell (59)

1. $\boldsymbol{H_1 = I}$, $\boldsymbol{x_1}$ point initial.
2. Itération $\boldsymbol{k}$:
   (a) $\boldsymbol{d_k = -H_k g_k}$ (on note $\boldsymbol{g_k = \nabla f(x_k)}$)
   (b) Trouver $\boldsymbol{\rho^*}$
   (c) $\boldsymbol{\sigma_k = \rho^* d_k}$
   (d) $\boldsymbol{x_{k+1} = x_k + \sigma_k}$
   (e) $\boldsymbol{\gamma_k = g_{k+1} - g_k}$
   (f) $\boldsymbol{H_{k+1} = H_k + A_k + B_k}$ avec $\boldsymbol{A_k = \sigma_k \sigma_k^T / \sigma_k^T \gamma_k}$ et $\boldsymbol{B_k = H_k \gamma_k \gamma_k^T H_k / \gamma_k^T H_k \gamma_k}$
   (g) Cas d'arrêt ? Sinon réitérer.

Remarques sur DFP

- **Stabilité**

  On dit que la méthode est *stable* si la valeur de f diminue à chaque itération.

- On montre que DFP est stable si $H_k \succ 0$ pour tout k,
- Par récurrence, on montre que c'est OK.
- **Terminaison quadratique** On dit que la méthode a la *propriété de terminaison quadratique* si elle atteint l'optimum exact en un nombre fini d'opérations.
- On montre c'est vrai si
  - arithmétique exacte,
  - recherche linéaire exacte.
- En fait $H_{n+1} = (\nabla^2 f)^{-1}$.
- En fait $\sum A_k = (\nabla^2 f)^{-1}$
- A_k et B_k de rang 1 : "méthode de rang 1".

DFP complémentaire

- BFGS : Broyden-Fletcher-Goldfarb-Shanno (70)
  1. $\boldsymbol{H_1 = I}$, $\boldsymbol{x_1}$ point initial.
  2. Itération $\boldsymbol{k}$: On fait

$$\boldsymbol{H_{k+1} = (I - \frac{\sigma_k \gamma_k^T}{\gamma_k^T \sigma_k}) H_k (I - \frac{\gamma_k \sigma_k^T}{\gamma_k^T \sigma_k}) + \frac{\sigma_k \sigma_k^T}{\gamma_k^T \sigma_k}.}$$

- Méthode de rang 2
- Mêmes propriétés que DFP
- En plus, la norme de l'erreur d'une certaine matrice est réduite à chaque itération
- Evite une tendance de $\boldsymbol{H_k}$ à devenir singulière dans DFP.

Gradients conjugués

- Directions conjuguées par rapport à $\boldsymbol{H}$: $\boldsymbol{p^T H q} = 0$.
- Dans un algo de descente, si on utilise des directions conjuguées, les points successifs minimisent aussi la fonction dans les directions précédemment utilisées.
- Algo de Fletcher-Reeves
  1. $\boldsymbol{d_1} = -\boldsymbol{\nabla f}(\boldsymbol{x_1})$
  2. $\boldsymbol{d_{k+1}} = -\boldsymbol{\nabla f}(\boldsymbol{x_{k+1}}) + \sum_{r \leq k} ||\boldsymbol{\nabla f}(\boldsymbol{x_{k+1}})||^2 / ||\boldsymbol{\nabla f}(\boldsymbol{x_k})||^2 \boldsymbol{d_r}$.
- Mêmes propriétés que DFP
- En fait : les trois méthodes donnent les mêmes points,
- BFGS plus stable que DFP,
- DFP formule plus simple que BFGS,
- gradients conjugués : on n'utilise que des vecteurs, pas des matrices !