Master 1 Spécialité : ingéniérie mathématique

2016-2017

Optimisation déterministe et stochastique

Laurent Guillopé

Laboratoire de mathématiques Jean Leray
Département de mathématiques, UFR Sciences et techniques
Université de Nantes

www.math.sciences.univ-nantes.fr/~guillope/m1-ods/

Version : 11 mai 2017

Table des matières

Prologue

Le cadre général de ces notes est l'étude des optima d'une fonction $J : E \to \mathbb{R}$: meilleure minoration avec la borne $\inf_{x\in E} J(x)$, valeur minimale $\inf_{x\in E} J(x)$ atteinte au point de minimum $\operatorname{argmin}_{x\in E}(J(x))$. Au signe près, les méthodes déployées à cet effet sont évidemment valides pour la recherche de maxima de J : l'ensemble $\operatorname{argmax}_{x\in E}(J(x))$ des points de maxima de J coïncident avec celui $\operatorname{argmin}_{x\in E}(-J(x))$ des points de minima de $K = -J$, les valeurs extrémales étant opposées : $J_* = -K^*$ où $J_* = \min_{x\in E} J(x)$ et $K^* = \max_{x\in E} K(x)$.

Si les problèmes d'optimisation sont ultra fréquents, leur grande variété empêche la validité de méthodes générales. Dans la suite, on aborde certains cas où E est une partie d'un espace de dimension finie (domaine ouvert, adhérence de domaine régulier, surface ou plus généralement hypersurface,...) dont la classification fait émerger différent types de problèmes (minimum local, minimum avec contraintes d'inégalités, minima contraints en égalité) et la fonction J de régularité diverse (de classe $\mathcal{C}^2$, linéaire, quadratique, convexe,...).

Les problèmes et méthodes rencontrés seront à caractère stochastique ou déterministe : stochastique dans l'énoncé même du problème (maxima de vraisemblance, processus de décision markovien) ou bien dans la voie de résolution (recherches à la Monte-Carlo, calcul d'espérances conditionnelles)

Les différents résultats sont accompagnés d'exemples : les outils comme `R` et `sage` permettent des expérimentations et des traitements aisés.

Terminons en citant le potentiel de Lennard-Jones

$$V_{LJ}(x_1,\dots,x_n) = \sum_{1\le k<\ell\le n} \Phi(\|x_k - x_\ell\|), \quad x_1,\dots,x_n \in \mathbb{R}^d$$

avec $\Phi(r) = r^{-12} - r^{-6}$ qui est central dans certaines modélisations chimiques, comme l'est le potentiel de Thompson

$$V_T(x_1,\dots,x_n) = \sum_{1\le k<\ell\le n} \|x_k - x_\ell\|^{-\alpha}, \quad x_1,\dots,x_n \in \mathbb{S}^d.$$

La recherche d'un minimum (pour $\ell > 3$, ℓ souvent de l'ordre de 100) est un véritable défi : le nombre des minima locaux est exponentiel en n, des minima locaux proches de minima globaux leur sont difficilement discernables, des points selle sont présents,... Ces potentiels $V_{LJ}; V_T$ ont donc de nombreuses oscillations, comme la fonction modèle $J(x,y) = (x\sin(2y)+y\sin(2x))^2\cosh(x\sin(x)/10)+(x\cos y - y\sin x))^2\cosh(y\cos(2y)/10)$ dont le graphe et les lignes de niveau sont tracées dans la figure 1

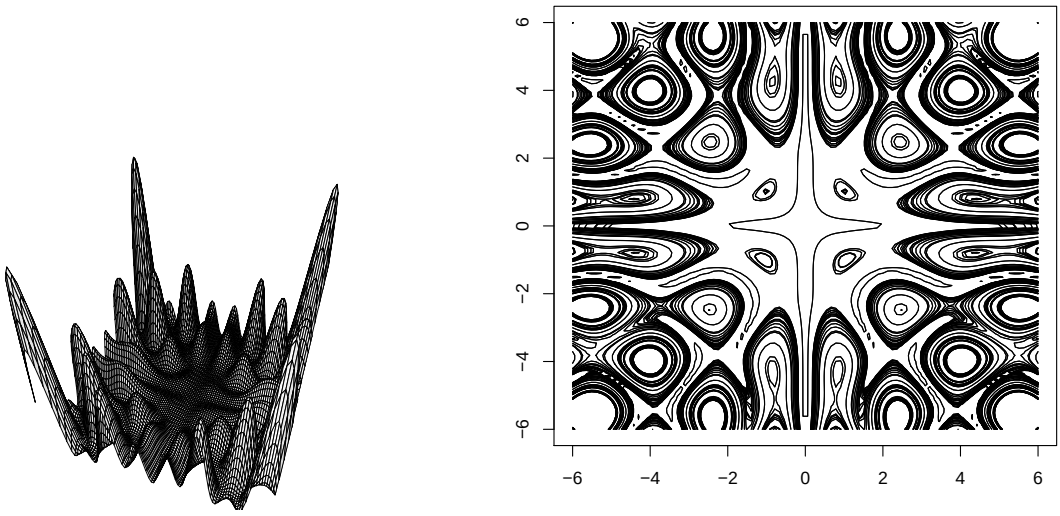


FIGURE 1 . Une fonction oscillante de deux variables : graphe de la fonction et lignes de niveau.

1. Remarques sur l'usage de R et sage

Les différents calculs et figures ont été effectués avec R [1], sauf ceux nécessitant du calcul formel, qui ont été exécutés avec sage [2].

La première partie de la figure A.1 a été tracée par le code R suivant

```
x<-seq(-0.9,0.9,length.out=100)
y<-seq(0.1,4,length.out=100)
K<-function(v,r) {log(v*(5/2)*(1-r**2))+(v+(5/2))/(v*(5/2)*(1-r**2))+log(v*(5/2))+4/v+4/(5/2)}
Kv<-Vectorize(K)
z=outer(x,y,Kv)
lev=NULL
for (rr in seq(-0.9,0.9,length.out=60)) {lev=c(lev,K(0.3,rr))}
for (v in seq(0.1,4,length.out=30)) {lev=c(lev,K(v,0))}
for (r in seq(-0.7,0.7,length.out=30)) {lev=c(lev,K(5/2,r))}
lev=c(lev,K(0,5/2))
contour(x,y,z,xlim=c(-0.9,0.9),ylim=c(1.5,4),level=lev,xlab="rho",ylab="v_2",drawlabels=FALSE)
```

Le calcul de la hessienne (21) a été effectué avec sage suivant

```
var('r,v,w')
L(v,w,r)=log(v*w*(1-r**2))+(v+w)/(v*w*(1-r**2))+log(v*w)+4/v+4/w
gradL=L.gradient()
critical=solve([gradL[0] == 0, gradL[1] == 0, gradL[2] == 0],r,v,w, solution_dict=True)
critical
H=L.hessian()
L.hessian()(8/3,8/3,1/2)
numerical_approx(L(8/3,8/3,1/2))
for i in range(3) :
   numerical_approx(L.hessian()(8/3,8/3,1/2).eigenvalues()[i])
```

R, et ses bibliothèques fournit de multiples procédures d'optimisation : la fonction nlm optimise des fonctions d'une variable sur un intervalle : c'est une méthode de Newton (avec calcul de la dérivée numérique $(J'(x) \simeq (J(x+h) - J(x))/h)$, si la dérivée

1. R : www.r-project.org/. Le paquet sweave permet d'entrelacer de manière naturelle l'exécution de code R et la composition suivant TEX.

2. sage : www.sagemath.org/. Le paquet sagetex permet de combiner exécution de code sage et composition avec LATEX.

exacte n'est pas disponible). La fonction `optim` met à disposition de multiples méthodes d'optimisation : Nelder-Mead, BFGS, CG, L-BFGS-B, SANN, Brent.

2. Extrema du modèle quadratique

Dans ce paragraphe introductif, J, U seront supposées définies sur $\mathbb{R}^d$ tout entier, dérivables autant nécessaire que ce soit. On précisera éventuellement des propriétés locales, valables pour des fonctions définies localement. Cette partie introductive est concentrée sur le modèle quadratique, qui régit l'étude de beaucoup de problèmes de minima : un fait et quatre remarques, qui seront développées ultérieurement.

La formule de Taylor centrée en x à l'ordre 2 énonce, pour J de classe $\mathcal{C}^2$,

$$J(x+h) = J(x) + \langle \nabla J(x), h\rangle + \frac{\operatorname{Hess} J(x)[h]}{2} + o(\|h\|^2), \quad h \to 0.$$

Le modèle quadratique J_x de J centré en x est obtenu en considérant les termes d'ordre au plus 2 dans le développement de Taylor de J centré en x :

$$J_x(h) = J(x) + \langle \nabla J(x), h\rangle + \frac{\operatorname{Hess} J(x)[h]}{2}. \tag{1}$$

Le modèle quadratique $\widetilde{J}_x$ basé en x est défini suivant $\widetilde{J}_x(y) = J_x(y-x)$: si J est quadratique $J(y) = \widetilde{J}_x(y-x)$.

FAIT 0.1: *Si* $\operatorname{Hess} J(x)$ *est inversible,* J_x *a un unique point critique* $h_* = -[\operatorname{Hess} J(x)]^{-1}\nabla J(x)$. *Si* $\operatorname{Hess} J(x)$ *est définie positive,* h_* *est l'unique point de minimum global de* J_x

DÉMONSTRATION. Pour simplifier, notons $C = J(x)$, $v = \nabla J(x)$ et $A = \operatorname{Hess} J(x)$. L'application

$$U : h \in \mathbb{R}^d \mapsto U(h) = C + \langle v, h\rangle + \langle Ah, h\rangle/2 \in \mathbb{R}$$

a comme gradient $\nabla U(h) = v + Ah$. Si A est inversible, la fonction U a donc un seul point critique $h_* = -A^{-1}v$, pour lequel on peut écrire

$$\begin{aligned} U(h_*+k) &= C + \langle v, h_*+k\rangle + A(h_*+k), h_*+k\rangle/2 \\ &= C + \langle v, h_*\rangle + \langle Ah_*, h_*\rangle/2 + \langle v, k\rangle + \langle Ah_*, k\rangle + \langle Ak, k\rangle/2 \\ &= U(h_*) + \langle Ak, k\rangle/2. \end{aligned}$$

Si A est définie positive, alors le dernier terme est positif, non nul si k est non nul : h_* est un minimum global strict. □

Dans ce cas, les lignes de niveau sont des ellipsoïdes, *i. e.* des ellipses en dimension $d = 2$.

△ REMARQUE 0.1: [Discussion sur A symétrique non définie positive] Reprenant les notations précédentes, si A est définie négative, h_* est un maximum global. Si A inversible n'est pas définie (positive ou négative), alors A inversible a des éléments propres $(v_\pm, \lambda_\pm)$ avec $\pm\lambda_\pm > 0$: $U(h_* + tv_\pm) = U(h_*) + t^2\lambda_\pm\|v_\pm\|^2$ et h_* est un minimum (maximum resp.) pour U restreinte à la droite $h_* + \mathbb{R}tv_+$ (resp. $h_* + \mathbb{R}v_-$).

Dans le cas A non inversible, décomposons $v = v_- + v_0 + v_+, k = k_- + k_0 + k_+$ suivant la décomposition

$$\mathbb{R}^d = K_+ \oplus \ker A \oplus K_- \quad \text{avec } K_\pm = \bigoplus_{\pm\lambda>0} \ker(A-\lambda).$$

Alors, si $h_* = -A_\pm^{-1}(v_+ + v_-)$ où $A_\pm$ est l'automorphisme de $K_- \oplus K_+$ obtenu par restriction de A

$$\begin{aligned} U(h_* + k) &= C + \langle v, h_* + k\rangle + \langle A(h_* + k), h_* + k\rangle/2 \\ &= U(h_*) + \langle v_0, k_0\rangle + \langle v_+ + v_-, k_+ + k_-\rangle + \langle Ah_*, k_+ + k_-\rangle \\ &\qquad + \langle Ak_+, k_+\rangle/2 + \langle Ak_-, k_-\rangle/2 \\ &= U(h_*) + \langle v_0, k_0\rangle + \langle Ak_+, k_+\rangle/2 + \langle Ak_-, k_-\rangle/2 \end{aligned}$$

Comme fonction de k, la fonction U est linéaire non nulle (et donc ni majorée, ni minorée) dans la direction du noyau $\ker A$ si v_0 est non nul, constante sinon, alors qu'elle est minorée ou majorée sur les sous-espaces K_+ (resp. K_-) avec extremum l'origine (s'ils sont non triviaux). Si les espaces K_+ et K_- ne sont pas triviaux, on dit que le point critique h_* est un point selle : en dimension d avec K_+ et K_- de dimension $d_\pm > 1$, on a des directions issues de l'origine le long desquelles la fonction J diminue (resp. augmente), le graphe de la fonction (qui est donc une surface dans $\mathbb{R}^3$) a des apparences de selle de cheval (ou col de montagne). ▽

△ Remarque 0.2: [Heuristique de la méthode de Newton] Soit x proche du minimum x_* de J : on va approcher ce point critique $x_* = x_*(J)$ de J par le point critique $x_*(\widetilde{J}_x)$ du modèle quadratique basé en x : $h_* = x_*(\widetilde{J}_x) - x$ est le minimum de l'approximation (1) taylorienne J_x centrée en x. Ainsi, vu que $h_* = -(\mathrm{Hess}\, J(x))^{-1}\nabla J(x)$, on obtient

$$x_*(\widetilde{J}_x) = x - (\mathrm{Hess}\, J(x))^{-1}\nabla J(x).$$

Le point $x_*(\widetilde{J}_x)$ est *heuristiquement* une approximation de $x_*(J)$: il est remarquable que, sous des hypothèses souvent vérifiées, l'itération

$$x_{k+1} = x_k - (\mathrm{Hess}\, J(x_k))^{-1}\nabla J(x_k)$$

converge vers l'extremum $x_*(J)$: c'est l'itération de Newton-Raphson, qui est généralisée en l'itération de Newton-Lagrange pour des problèmes de minimum sous contraintes. Cette itération de Newton vaut pour l'approximation du zéro x_* de $F : U(\subset \mathbb{R}^n) \to \mathbb{R}^n$ sous l'hypothèse (entre autres) d'inversibilité de la différentielle $dF(x_*)$. ▽

△ Remarque 0.3: [Régression linéaire et moindres carrés] La régression linéaire vise à estimer des paramètres $\Theta \in \mathbb{R}^p$ d'une variable aléatoire (à expliquer ou prédire) modélisant la réponse d'un système à une entrée $x \in \mathbb{R}^q$ (variables dites explicatives ou prédictives). Plus précisément on fait l'hypothèse de la modélisation par la variable aléatoire $Y(x, \Theta) = \langle x, \Theta\rangle$ qui, complémentée par un bruit ε, décrit la réponse y (on a supposé pour simplifier $p = q$). Le bruit (vu comme erreur) ε est supposé de moyenne nulle et de variance σ^2, par exemple pour un modèle gaussien $\varepsilon \sim \mathcal{N}(0, \sigma)$. On considère un échantillon $Y(x_1, \Theta), \ldots, Y(x_m, \Theta)$ dont on compare les valeurs à celles observées $y_1, \ldots, y_m$: à travers le choix du paramètre Θ, on cherche à minimiser les normes des résidus $r_\Theta(x_j, y_j) = y_j - Y(x_j, \Theta)$, soit la norme totale $\|R_\Theta(\mathbf{x}, \mathbf{y})\| = (\sum_{j=1}^m \|r_\Theta(x_j, y_j)\|^2)^{1/2}$ du résidu total $R_\Theta(\mathbf{x}, \mathbf{y}) = (r_\Theta(x_j, y_j))_{j=1}^m$. Cette minimisation du résidu (introduite par Gauß) est équivalente à la maximisation de la vraisemblance

$$(\sigma\sqrt{2\pi})^{-m}\mathrm{e}^{-\sum_j \varepsilon_j^2/(\sqrt{2}\sigma)} = \mathrm{e}^{-\|R_\Theta(\mathbf{x},\mathbf{y})\|^2}$$

dans le cas du modèle gaussien $\varepsilon \sim \mathcal{N}(0, \sigma)$. Notant

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} \in \mathbb{R}^m, \quad \mathbf{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \in \mathrm{Hom}(\mathbb{R}^p, \mathbb{R}^m),$$

on obtient

$$\mathbf{X}\Theta = {}^{\mathsf{T}}\big(\langle x_1,\Theta\rangle,\dots,\langle x_m,\Theta\rangle\big) \in \mathbb{R}^m$$

puis

$$\begin{aligned} M_{\mathbf{x},\mathbf{y}}(\Theta) = \|R_\Theta(\mathbf{x},\mathbf{y})\|^2 &= \sum_{j=1}^m r_\Theta(x_j,y_j)^2 = \sum_{j=1}^m (y_j - Y(x_i,\Theta))^2 \\ &= \sum_{j=1}^m (y_j - \langle x_j,\Theta\rangle)^2 = \|\mathbf{y}-\mathbf{X}\Theta\|^2_{\mathbb{R}^m} \\ &= \|\mathbf{y}\|^2 - 2\langle \mathbf{y},\mathbf{X}\Theta\rangle + \langle \mathbf{X}\Theta,\mathbf{X}\Theta\rangle = \|\mathbf{y}\|^2 - 2\langle {}^{\mathsf{T}}\mathbf{X}\mathbf{y},\Theta\rangle + \langle {}^{\mathsf{T}}\mathbf{X}\mathbf{X}\Theta,\Theta\rangle \end{aligned}$$

dont les points critiques Θ_* relativement aux variations en Θ annulent le gradient

$$\nabla_\Theta M_{\mathbf{x},\mathbf{y}}(\Theta) = 2(-{}^{\mathsf{T}}\mathbf{X}\mathbf{y} + {}^{\mathsf{T}}\mathbf{X}\mathbf{X}\Theta),$$

soit $\Theta_* = ({}^{\mathsf{T}}\mathbf{X}\mathbf{X})^{-1}{}^{\mathsf{T}}\mathbf{X}\mathbf{y}$, à supposer que ${}^{\mathsf{T}}\mathbf{X}\mathbf{X}$ soit inversible. Si $\mathbf{X}$ est inversible, alors le minimum est déterminé suivant $\Theta = \mathbf{X}^{-1}\mathbf{y}$, mais en général $\mathbf{X}$ n'est pas inversible (ne serait-ce parce que $\mathbf{X}$ est une matrice non carrée !) alors que ${}^{\mathsf{T}}\mathbf{X}\mathbf{X}$ l'est !

Gauß résolut la détermination des *moindres carrés* (*i. e.* la minimisation de) $\|R_\Theta(\mathbf{x},\mathbf{y})\|^2 = \sum_{j=1}^m \|r_\Theta(x_j,y_j)\|^2$ dans le cas où les résidus $r_\Theta(x_j,y_j)$ sont linéaires dans sa détermination de l'orbite de Ceres à partir des données astronomiques collectées par G. Pazzi, méthode présentée indépendamment par A. Legendre et généralisée sous le nom de Gauß-Newton dans le cas non linéaire[3].

La matrice ${}^{\mathsf{T}}\mathbf{X}\mathbf{X}$ est inversible si et seulement si elle est injective. Un $u \in \ker {}^{\mathsf{T}}\mathbf{X}\mathbf{X} = \ker \mathbf{X}$ vérifie $0 = \langle x_1,u\rangle = \ldots = \langle x_m,u\rangle$, *i. e.* il est orthogonal à $\mathrm{Vect}(u_1,\dots,u_m)$: X est injective si et seulement si la famille $(x_1,\dots,x_m)$ engendre $\mathbb{R}^p$. On fera l'hypothèse que les valeurs prédictives $x_1,\dots,x_m$ engendrent $\mathbb{R}^p$. Pour la régression linéaire simple où on utilise des variables prédictives du type $x_i = (1,t_i) \in \mathbb{R}^2$ de telle manière que $Y(x_i,\Theta) = \alpha + \beta t_i$ où on a noté $\Theta = (\alpha,\beta)$, la condition est que les $t_1,\dots,t_m$ ne soient pas tous égaux.

Plus généralement, soit A une matrice de type (m,n) et $b \in \mathbb{R}^m$. Un problème de moindres carrés consiste en la minimisation de $K_{A,b}(x) = \|Ax-b\|^2$ pour $x \in \mathbb{R}^n$. Vu la somme directe $\mathrm{Im}\, A \oplus_\perp (\mathrm{Im}\, A)^\perp = \mathrm{Im}\, A \oplus_\perp \ker {}^{\mathsf{T}}A$, on peut écrire $b = Av_b + k_b$ avec v_b un vecteur de $\mathbb{R}^n$ et $k_b \in \ker {}^{\mathsf{T}}A$ ce qui induit $\langle Ax,b\rangle = \langle Ax,Av_b\rangle + \langle x,{}^{\mathsf{T}}Ak_b\rangle = \langle Ax,Av_b\rangle$ et donc

$$\begin{aligned} \|Ax-b\|^2 &= \langle Ax,Ax\rangle - 2\langle Ax,b\rangle + \|b\|^2 = \langle Ax,Ax\rangle - 2\langle Ax,Av_b\rangle + \|b\|^2 \\ &= \|Ax-Av_b\|^2 - \|Av_b\|^2 + \|b\|^2 = \|A(x-v_b)\|^2 - \|Av_b\|^2 + \|b\|^2, \end{aligned}$$

dont le lieu des points de minimum est le sous-espace affine $v_b + \ker A$. On a l'identité des noyaux $\ker A$ et $\ker {}^{\mathsf{T}}AA$.

On peut retrouver ces résultats en arguant des arguments de convexité. La fonction $K_{A,b}$ est convexe (comme fonction quadratique avec hessienne $\mathrm{Hess}\, K_{A,b}(x) = {}^{\mathsf{T}}AA \geq 0$) et son gradient $\nabla_x K_{A,b}(x) = 2{}^{\mathsf{T}}AA(x-v_b)$. Ses points critiques sont les seuls vecteurs du sous-espace $v_b + \ker A$ qui sont donc les minima de la fonction convexe $K_{A,b}$. ▽

△ Remarque 0.4: [Pseudo-inverses] On convient de dire que la matrice A a pour inverse généralisé B si $ABA = A$. Dans ces notes sont apparus deux tels inverses généralisés. D'une part, si A est symétrique avec π_A la projection sur l'orthogonal de son noyau

3. Par exemple, pour un résidu $y - \langle x,\theta\rangle$ remplacé par l'expression type $\log(1+\exp(y\langle,x,\theta\rangle)$

et $A_\pm$ la restriction de A à cet orthogonal, on a introduit dans la remarque 0.1, l'opérateur $\widetilde{A} = A_\pm^{-1}\pi_A$, qui coïncide avec A^{-1} si A est inversible et vérifie $A\widetilde{A} = \widetilde{A}A = \pi_A$.

Par ailleurs, dans la remarque 0.3, si $\mathbf{X}$ d'ordre (m,p) est injective, on a introduit l'opérateur $X^\dagger = ({}^\mathsf{T}\mathbf{X}\mathbf{X})^{-1}{}^\mathsf{T}\mathbf{X}$ qui vérifie $\mathbf{X}^\dagger\mathbf{X} = I_{\mathbb{R}^p}$: c'est le pseudo-inverse de Moore-Penrose. ▽

△ Remarque 0.5: Soit $\mathbf{w} : (\Omega, \mathcal{A}, dP) \to \mathbb{R}$ intégrable. Le problème

$$\text{argmin}_{v\in\mathbb{R}}\, \mathbb{E}\left((v-\mathbf{w})^2\right)$$

est résoluble aisément puisque la fonction à minimiser est quadratique

$$\mathbb{E}((v-\mathbf{w})^2) = \mathbb{E}(v^2 - 2v\mathbf{w} + \mathbf{w}^2) = v^2 - 2v\mathbb{E}(\mathbf{w}) + \mathbb{E}(\mathbf{w}^2)$$

de gradient $2(v - \mathbb{E}(\mathbf{w}))$, gradient nul pour $v = \mathbb{E}(\mathbf{w})$. Cet exemple sera repris dans la méthode de gradient stochastique (cf. section **??**) applicable à la minimisation de fonctions J du type $J(v) = \mathbb{E}_\Omega(j(v,\cdot))$ avec $j : (v,\omega) \in C \times \Omega \to j(x,\omega) \in \mathbb{R}$. ▽

Chapitre 1

Programmation stochastique

1. Recherche stochastique

Les recherches d'un minimum par exploration aléatoire du domaine de définition E de J sont souvent simples à mettre en place pour des fonctions aux propriétés variées, qu'elles soient peu nécessairement, dotées d'une multitude de minima locaux ou définies dans des domaines de dimension d élevée. Elles se révèlent cependant souvent convergentes, confirmant des arguments heuristiques ou confirmés par des développements théoriques. La vitesse de convergence est souvent difficile à estimer, la plupart du temps faible (d'autant plus faible que l'approximation x_k est proche du minimum x_*), au contraire de certaines méthodes utilisant la différentiabilité (par ex. la méthode de Newton). La programmation est aisée et rapide, basée uniquement sur la fonction J (pas de calcul de gradient, exact ou approché, ni a fortiori de hessienne) et la simulation de variables aléatoires $X_1, \dots, X_k, \dots$ convergent vers un point de minimum et telles que la suite des valeurs $J \circ X_1, \dots, J \circ X_k, \dots$ converge en décroissant vers la valeur minimum J_*.

À mise en place peu contraignante, algorithme sans grande garantie d'efficacité. Ainsi de l'absence de test d'arrêt universel pour la suite de points (x_n) candidats pour le minimum de J : on peut considérer le nombre N d'itérations, les différences $|J(x_n) - J(x_{n-\ell})|$ pour $\ell = 1, \dots, L$ (adaptée pour les fonctions J de coût/objectif sans bruit) ou les écarts $\|x_n - x_{n-\ell}\|$ pour $\ell = 1, \dots, L$ (cas de convergence vers un extremum unique).

Le choix de mesures de probabilité insufflant la recherche aléatoire (sur le domaine de recherche donné ou après localisation la recherche) s'impose parfois, par exemple si E est un pavé $\prod_{k=1}^{d} [a_k, b_k]$: on prend la distribution uniforme produit, voire un produit de gaussienne lorsque $E = \mathbb{R}^d$. En général, on utilisera une méthode de rejet ou de projection en incluant E dans un pavé.

1.1. Recherche aléatoire à la Monte-Carlo. Cette recherche est appelée parfois *recherche aléatoire en aveugle*. L'énoncé suivant précise les propriétés de convergence d'une simulation $x_0, x_1, \dots$ obtenue par tirages aléatoires indépendants sur $(E, \mathbb{P})$.

THÉORÈME 1.1 ([38, thm. 2.1, p. 40]): *Soit $(E, \mathbb{P})$ un espace de probabilité, $(\Omega = E^{\mathbb{N}}, \mathbb{P}^{\otimes \mathbb{N}})$ l'espace des suites $\omega = (\pi_k(\omega))_{k \geq 0}$ avec π_k la k-ième projection $\Omega \to E$ incarnant des suites de tirages indépendants et J une fonction bornée sur E avec un unique minimum x_* tel que, pour tous $\varepsilon, \eta > 0$ assez petits,*

$$\eta_\varepsilon = \inf_{\|x - x_*\| \geq \varepsilon} J(x) - J(x_*) > 0, \quad \mathbb{P}[J \geq J(x_*) + \eta] \leq \delta_\eta < 1. \tag{2}$$

Soit $X_0 : \Omega \to E$ une variable aléatoire. Pour $k \geq 1$, définissons les variables aléatoires $X_k : \Omega \to E$ par la relation de récurrence

$$X_k(\omega) = \begin{cases} \pi_k(\omega) & \text{si } J(\pi_k(\omega)) < J(X_{k-1}(\omega)), \\ X_{k-1}(\omega) & \text{sinon.} \end{cases}$$

Alors la suite (X_k) converge p. s. vers x_*, avec $J_k = J \circ X_k \to J(x_*)$.

△ REMARQUE 1.1: Si E est un compact de $\mathbb{R}^d$ avec la mesure uniforme de Lebesgue $\pi_E = \lambda_d/\lambda_d(E)$ (ou $f(x)\lambda_d$ avec la densité f continue positive non nulle) et J une fonction continue sur E ayant un point de minimum unique, les hypothèses du théorème sont bien assurées. Cela résulte du fait que le minimum de J est atteint sur $\{\|x-x_*\| \geq \varepsilon\}$ et ne peut être x_* d'une part, que $\{J \leq J(x_*)+\eta\}$ contient une boule ouverte non vide (donc de mesure non nulle) d'autre part. ▽

Algorithme 1.1 (recherche aléatoire à l'aveugle)

```
Choisir x_0 ; k = 0
tant que k < K faire
    Tirer x
    si J(x) < J(x_k) alors
        x_{k+1} = x
    sinon
        x_{k+1} = x_k
    fin si
    k = k + 1
fin tant que
```

DÉMONSTRATION. La suite de variables aléatoires $(J_k = J \circ X_k)_{k\geq 0}$ est décroissante, elle converge donc simplement, soit J_∞ sa limite. Par construction des X_k, on a, pour $j = 0, \ldots, k$, les inégalités. $J_k = J \circ X_k \leq J_j = J \circ X_j \leq J \circ \pi_j$. Notons $J_* = J(x_*)$. Ainsi pour $\eta > 0$, vu l'indépendance des tirages (π_j) et des fonctions $(J \circ \pi_j)$,

$$\begin{aligned}\mathbb{P}_\Omega(J_k - J_* \geq \eta) &\leq \mathbb{P}_\Omega(J \circ \pi_j \geq J_* + \eta, j = 0, \ldots, k) \\ &= \prod_{j=0}^{k} \mathbb{P}_\Omega(J \circ \pi_j \geq J_* + \eta) = \mathbb{P}(J \geq J_* + \eta)^{k+1} \leq \delta_\eta^{k+1}\end{aligned}$$

et par suite

$$\mathbb{P}_\Omega(J_\infty \geq J_* + \eta) \leq \mathbb{P}_\Omega(J_k \geq J_* + \eta) \leq \delta_\eta^{k+1},$$

soit $\mathbb{P}_\Omega(J_\infty \geq J_* + \eta) = 0$ en faisant tendre $k \to \infty$. Ainsi $J_\infty < J_* + \eta$ p. s. pour tout $\eta > 0$: considérant une suite de rationnels $\eta_k \to 0^+$, on conclut que $J_\infty \leq J_*$ p. s., soit $J_\infty = J_*$ p. s.

Par ailleurs, pour $\varepsilon > 0$ et avec le $\eta_\varepsilon > 0$ de l'hypothèse (2),

$$\mathbb{P}_\Omega(\|X_k - x_*\| > \varepsilon) \leq \mathbb{P}_\Omega(J_k \geq J_* + \eta_\varepsilon) \leq \delta_{\eta_\varepsilon}^{k+1} \to_{k\to\infty} 0.$$

et donc $X_k \xrightarrow{\mathbb{P}} x_*$ en probabilité.

Vu que les fonctions J_k sont uniformément bornées par la borne $\|J\|_\infty$ supposée finie, il y a convergence en moyenne quadratique et convergence ponctuelle presque sûrement pour une sous-suite J_{k_j}. Par décroissance de la suite (J_n), c'est en fait toute la suite (J_n) qui converge presque sûrement vers J_*. L'unicité du minimum x_*, alliée aux hypothèses (2), implique $X_k \to x_*$ presque sûrement sur Ω. □

▷ EXEMPLES 1.1:

1.1.1 Soit $\|\ \|$ la norme euclidienne sur $\mathbb{R}^d$ et la fonction J_d définie sur $[1,3]^d$ par $J_d(m) = \|m\|^2/d$ avec $m_d = (1, \ldots, 1)$ comme unique minimum et $J_d = 1$ comme valeur minimale. Le tableau suivant illustre la dégradation de la performance suivant l'augmentation de la dimension : la fonction a été normalisée de telle manière que le minimum soit toujours 1.

d	1	2	3	4	5
min	1.000002	1.001672	1.013262	1.035949	1.146557
d	6	7	8	9	10
min	1.187828	1.259963	1.288603	1.453986	1.532829

TABLE 1. Une recherche stochastique en aveugle pour le minimum de $J_d(m) = \|m\|^2/d$ sur le pavé $[1,3]^d$ avec 10^6 tirages pour $d = 1, \dots, 10$.

1.1.2 Pour la fonction J définie sur sur $[-1,1]^3$ par

$$J(x,y,z) = \left| \text{dét} \begin{pmatrix} 1 & x & x^2 \\ 1 & y & y^2 \\ 1 & z & z^2 \end{pmatrix} \right| = |(x-y)(y-z)(x-z)|$$

avec un minimum (non unique), la valeur du minimum obtenu suivant l'algorithme 1.1 est décrit dans la table 1.1 suivant le nombre de tirages. ◁

n	1	2	3	4	5	6	7	8	9	10
min J	0.076	0.005	0.0016	9.3e-05	5.3e-06	8.7e-06	1.6e-06	3.8e-07	2.2e-08	1.4e-09

TABLE 2. Recherche en aveugle sur $[-1,1]^3$ avec $N = 10^n$ tirages pour le minima de la fonction $J = |(x-y)(y-z)(x-z)|$.

Pour estimer la vitesse de convergence en fonction de la dimension d de $E \subset \mathbb{R}^d$ avec la mesure uniforme $|\quad|_d/|E|_d$, considérons le cube $V_* = \prod_{i=1}^d [x_{*j} - a/2, x_{*j} + a/2]$ centré en x_*, d'arête de longueur a et inclus dans l'intérieur de E et le test d'arrêt $X_k \in V_*$. Vu l'indépendance des tirages, la probabilité ρ_k de ne pas arriver dans V_* avant l'itération d'ordre k est

$$\rho_k = \mathbb{P}(X_j \notin V_*, j = 1, \dots, J) = \prod_{j=1}^{J} \mathbb{P}(X_j \notin V_*) = \mathbb{P}(x \notin V_*)^k = (1 - |V_*|)^k .$$

La probabilité d'arriver dans V_* au cours des k premières itérations est $1 - \rho_k$. En dimension d, on a $\mathbb{P}(V_*) = a^d/|E|_d$, soit

$$k = \frac{\log \rho_k}{\log(1 - a^d/|E|_d)} \sim_{d\to\infty} \frac{\log(1/\rho_k)}{a^d/|E|_d} \to +\infty \text{ si } d \to \infty,$$

ce qui indique une convergence de plus en plus lente quand d devient grand. Pour $\rho = 0.05$, $x_* = 0$, $a = 0.03$ et $U = [-1/2, 1/2]^d$, on a les valeurs

d	2	4	6	8	10
n	$3.33e^3$	$3.69\ e^6$	$4.11e^9$	$4.57e^{12}$	$5.40e^{15}$

1.2. Recherche aléatoire localisée. L'algorithme de *recherche localisée à incréments aléatoires* construit x_{k+1} à partir de x_k par ajout d'un incrément aléatoire d_k (suivant, par exemple, une loi uniforme sur une boule ou une gaussienne centrée) : si $x_k + d_k$ est dans $\mathbf{dom}\, J$ et $J(x_k + d_k) < J(x_k)$, on pose $x_{k+1} = x_k + d_k$, sinon $x_{k+1} = x_k$. Un d_k utile (et qui sera retenu dans l'itération) est un vecteur d_k de direction $d_k/\|d_k\|$ et de taille $\|d_k\|$ qui permette de baisser la valeur de J : $J(x_k + d_k) < J(x_k)$. C'est l'équivalent des vecteurs de descente introduit systématiquement dans les méthodes déterministes de descente.

Une variante considère le projeté [1] $p_{k+1} = \mathbf{pr}_E(x_k + d_k)$ sur le domaine de définition E et on teste $J(p_{k+1}) < J(x_k)$ pour poser $x_{k+1} = p_{k+1}$ en cas de succès et $x_{k+1} = x_k$ sinon.

Algorithme 1.2 (recherche localisée à incréments aléatoires)

```
Choisir x_0
tant que k < K faire
    Tirer d
    si x_k + d ∈ dom J & J(x_k + d) < J(x_k) alors
        x_{k+1} = x_k + d
    sinon
        x_{k+1} = x_k
    fin si
    k = k + 1
fin tant que
```

L'incrément aléatoire d_k suit souvent une distribution uniforme ou une loi normale $\mathcal{N}(0, \Sigma)$ avec Σ adapté à la forme de $\mathbf{dom}\, J$ (plus ou moins effilé suivant les axes de coordonnées).

THÉORÈME 1.2: *Soit $E \subset \mathbb{R}^n$ de mesure de Lebesgue finie non nulle, $J : E \to \mathbb{R}$ continue avec $J_* = \inf_{x\in E} J(x)$. Soit $(X_k)_{k\geq 1}$ une chaîne de Markov à valeurs dans E telle que*

- *X_1 soit de loi uniforme π_E sur E,*
- *X_k étant donné, on effectue un tirage de $Y \in \mathbb{R}^n$ suivant la loi de densité uniforme sur E : si $X_k + Y \in E$ et $J(X_k + Y) < J(X_k)$, alors on pose $X_{k+1} = X_k + Y$, sinon $X_{k+1} = X_k$.*

Alors la suite $J \circ X_k$ converge presque sûrement (et donc aussi en probabilité) vers J_*. En outre, $\mathbb{E}(\tau_\varepsilon) = \pi(E_\varepsilon)^{-1}$ où on a noté $E_\varepsilon = \{x | J(x) < J_* + \varepsilon\}$ et τ_ε le temps d'atteinte $\tau_\varepsilon = \inf\{k; X_k \in E_\varepsilon\}$.

DÉMONSTRATION. Notons $p_\varepsilon = \pi_E(E_\varepsilon)$ et $q_\varepsilon = 1 - p_\varepsilon$. Alors $P(X_1 \in E_\varepsilon) = p_\varepsilon$,

$$P(X_k \in E_\varepsilon, X_{k-1} \in \overline{E}_\varepsilon) = \int_{\overline{E_\varepsilon}} \int_{E_\varepsilon} d\pi_E(x) d\pi_E(y) = p_\varepsilon P(X_{k-1} \in \overline{E_\varepsilon}),$$

et

$$P(X_k \in E_\varepsilon) = P(X_{k-1} \in E_\varepsilon) + p_\varepsilon P(X_{k-1} \in \overline{E_\varepsilon}) = q_\varepsilon P(X_{k-1} \in E_\varepsilon) + p_\varepsilon,$$

soit finalement

$$P(X_k \in E_\varepsilon) = 1 - q_\varepsilon^k$$

Vu que la suite $(A_{\varepsilon k} = \{J \circ X_k < \varepsilon + J_*\})_{k\geq 1}$ est croissante, on a $\lim J \circ X_k \to J_*$ presque sûrement. La convergence en probabilité en résulte.

Concernant le temps d'atteinte τ_ε, l'égalité $P(\tau_\varepsilon = k) = P(X_k \in E_\varepsilon, X_{k-1} \in \overline{E_\varepsilon}) = p_\varepsilon q_\varepsilon^{k-1}$ et donc

$$\mathbb{E}(\tau_\varepsilon) = \sum_{k=1}^{\infty} kP(\tau_\varepsilon = k) = p_\varepsilon \sum_{k=1}^{\infty} kq_\varepsilon^{k-1} = 1/p_\varepsilon. \qquad \square$$

1. Si E un convexe d'un espace de Hilbert (de dimension finie ou pas), le projeté $\mathbf{pr}_E(m)$ est défini comme le point de E réalisant le minimum de la distance $\|m - u\|_2$ pour $u \in E$.

THÉORÈME 1.3: *Soit $E \subset \mathbb{R}^n$ de mesure de Lebesgue finie non nulle et π_E la mesure de probabilité uniforme sur E, $J : E \to \mathbb{R}$ continue avec $J_* = \min_{x\in E} J(x)$. Soit $(X_k)_{k\geq 1}$ une suite de variables aléatoires à valeurs dans E telle que*

- *X_1 soit de loi uniforme π_E sur E,*
- *X_k étant donné, on effectue un tirage de $Y \in \mathbb{R}^n$ suivant la loi de densité $q(y)dy$: si $X_k + Y \in E$ et $J(X_k + Y) < J(X_k)$, alors $X_{k+1} = X_k + Y$, sinon $X_{k+1} = X_k$.*

Il est supposé que $\inf_{\|x\|\leq R} q(x) > 0$ pour tout $R > 0$.

(1) La suite $J \circ X_k$ converge en probabilité vers J_*.

(2) Soit $\varepsilon > 0$, $E_\varepsilon = \{x \in E | J(x) < J_* + \varepsilon\}$, $\overline{E}_\varepsilon = \{x \in E | J(x) \geq J_* + \varepsilon\}$ son complémentaire et $\tau_\varepsilon = \inf\{k \in \mathbb{N} | X_k \in E_\varepsilon\}$ le temps d'atteinte de E_ε. Alors

$$\pi_E(E_\varepsilon)^{-1}\alpha_\varepsilon\beta_\varepsilon^{-2}(1 + \pi_E(E_\varepsilon)\beta_\varepsilon) \leq \frac{\mathbb{E}(\tau_\varepsilon) - \pi_E(E_\varepsilon)}{\pi_E(\overline{E_\varepsilon})} \leq \pi_E(E_\varepsilon)^{-1}\beta_\varepsilon\alpha_\varepsilon^{-2}(1 + \pi_E(E_\varepsilon)\alpha_\varepsilon)$$

où

$$\alpha_\varepsilon = \inf_{\substack{x\in\overline{E_\varepsilon}\\ z\in E_\varepsilon}} q(z - x), \quad \beta_\varepsilon = \sup_{\substack{x\in\overline{E_\varepsilon}\\ z\in E_\varepsilon}} q(z - x)$$

△ REMARQUE 1.2:
Pour la fonction $J : E \to \mathbb{R}$ vérifiant, pour un certain $x_* \in E$, $J(x_*) = -1$ et $J \geq 0$ sur $E \setminus \{x_*\}$, il n'y a pas de convergence : l'hypothèse introduisant le δ_η du théorème est importante. ▽

DÉMONSTRATION. Notons $p_\varepsilon = \pi_E(E_\varepsilon)$ et $q_\varepsilon = 1 - p_\varepsilon$. La convergence en probabilité de $J \circ X_k$ signifie que, pour tout $\varepsilon > 0$

$$P(|J \circ X_k - J_*| \geq \varepsilon) = P(J \circ X_k \geq J_* + \varepsilon) = P(X_k \in \overline{E}_\varepsilon) \to 0$$

Vu que X_1 est uniformément distribué sur E, on a

$$P(\tau_\varepsilon = 1) = P(X_1 \in E_\varepsilon) = \pi_E(E_\varepsilon) = p_\varepsilon.$$

Par ailleurs, avec P_{k-1} la loi induite par X_{k-1},

$$P(\tau_\varepsilon = k) = P(X_k \in E_\varepsilon, X_{k-1} \notin E_\varepsilon) = \int_{\overline{E_\varepsilon}} \left[\int_{E_\varepsilon} q(z - x)d\pi_E(z)\right] dP_{k-1}(x)$$

d'où

$$\alpha_\varepsilon p_\varepsilon P(X_{k-1} \in \overline{E_\varepsilon}) \leq P(\tau_\varepsilon = k) \leq \beta_\varepsilon p_\varepsilon P(X_{k-1} \in \overline{E_\varepsilon}) \tag{3}$$

Par ailleurs,

$$P(X_k \in \overline{E_\varepsilon} | X_{k-1} \in \overline{E_\varepsilon}) = \frac{P(X_k \in \overline{E_\varepsilon}, X_{k-1} \in \overline{E_\varepsilon})}{P(X_{k-1} \in \overline{E_\varepsilon})} = \frac{P(X_{k-1} \in \overline{E_\varepsilon}) - P(\tau_\varepsilon = k)}{P(X_{k-1} \in \overline{E_\varepsilon})}$$

et donc

$$1 - p_\varepsilon\beta_\varepsilon \leq P(X_k \in \overline{E_\varepsilon} | X_{k-1} \in \overline{E_\varepsilon}) \leq 1 - p_\varepsilon\alpha_\varepsilon.$$

puis

$$\begin{aligned} P(X_k \in \overline{E_\varepsilon}) &= P(X_k \in \overline{E_\varepsilon} | X_{k-1} \in \overline{E_\varepsilon})P(X_{k-1} \in \overline{E_\varepsilon}) \\ &\leq (1 - p_\varepsilon\alpha_\varepsilon)P(X_{k-1} \in \overline{E_\varepsilon}) \leq (1 - p_\varepsilon\alpha_\varepsilon)^{k-1}P(X_1 \in \overline{E_\varepsilon}) = (1 - p_\varepsilon\alpha_\varepsilon)^{k-1}q_\varepsilon \\ P(X_k \in \overline{E_\varepsilon}) &\geq (1 - p_\varepsilon\beta_\varepsilon)P(X_{k-1} \in \overline{E_\varepsilon}) \geq (1 - p_\varepsilon\beta_\varepsilon)^{k-1}P(X_1 \in \overline{E_\varepsilon}) = (1 - p_\varepsilon\beta_\varepsilon)^{k-1}q_\varepsilon, \end{aligned}$$

ce qui assure la convergence en probabilité de $J \circ X_k$ vers J_*. Pour l'espérance du temps d'atteinte τ_ε de E_ε, l'inégalité

$$p_\varepsilon \alpha_\varepsilon q_\varepsilon (1 - p_\varepsilon \beta_\varepsilon)^{k-2} \leq P(\tau_\varepsilon = k) \leq p_\varepsilon \beta_\varepsilon q_\varepsilon (1 - p_\varepsilon \alpha_\varepsilon)^{k-2}, \quad k \geq 2,$$

donne la majoration

$$\mathbb{E}(\tau_\varepsilon) = \sum_{k=1}^{\infty} kP(\tau_\varepsilon = k) \leq p_\varepsilon + p_\varepsilon \beta_\varepsilon q_\varepsilon \sum_{k=2}^{\infty} k(1 - p_\varepsilon \alpha_\varepsilon)^{k-2} = p_\varepsilon + p_\varepsilon^{-1} q_\varepsilon \beta_\varepsilon \alpha_\varepsilon^{-2} (1 + p_\varepsilon \alpha_\varepsilon)$$

et une minoration analogue, ce qui conclut la preuve. □

▷ Exemples 1.2:

1.2.1 Pour le premier exemple de la recherche stochastique aveugle (cf. table **??**), on a suivi 10 000 trajectoires pour chaque valeur de d : la table 3 donne le nombre moyen d'itérations nécessaires pour atteindre le minimum à 10^{-13} près.

d	1	2	3	4	5	6	7	8	9	10
$\overline{n}$	7	11	19	32	56	101	186	350	670	1 277
$sd(n)$	4.6	6.7	11.3	20.2	38.6	72.8	142.5	286.9	551.0	1 103.0
$\max(n)$	32	51	92	223	388	705	1 189	2 875	4 741	9 268

Table 3. Nombre moyen d'itérations, avec écart-type et max sur 1 000 recherches suivant une recherche stochastique localisée pour le minimum de la fonction $J_d(m) = \|m\|^2/d$ sur le pavé $[1,3]^d, d = 1, \ldots, 10$, avec atteinte du minimum à 10^{-13} près.

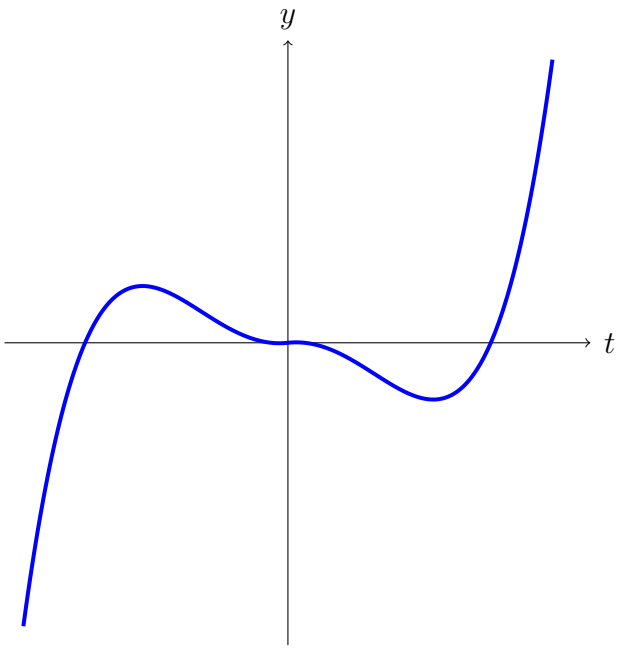


Figure I.1 . Le graphe de $y = t^4 - 16t^2 + 5t$ pour $t \in [-5, 5]$.

1.2.2 Soit J définie sur le pavé $[-8, 8]^2$ suivant $J(t_1, t_2) = K(t_1) + K(t_2)$ où le polynôme de degré 4 $K(t) = t^4 - 16t^2 + 5t$ a 3 extrema locaux (cf. Fig. I.1). La fonction J a quatre minima locaux et $x_* \simeq (-2.9035, -2.9035)$ avec $J(x_*) = -156.6646628$ comme minima global. Avec point de départ $x_0 = (4, 6.4)$, on a effectué $n = 100$

recherches sur $N = 10^6$ itérations : l'algorithme de recherche en aveugle (resp. localisé) a donné comme valeurs minimales moyennes $\overline{J(x_N)} = -156.662296$ (resp. -156.6640332).

1.2.3 Deux fonctions classiques sont prises comme tests de méthodes de recherche de minima

1.2.4 la fonction R_{2p}, dite de Rosenbrock, définie par

$$R_{2p}(x) = \sum_{j=1}^{p} \left[100(x_{2j} - x_{2j-1}^2)^2 + (1 - x_{2j-1})^2\right], \quad x \in \mathbb{R}^{2p},$$

illustre l'exemple d'une fonction avec un minimum à bassin d'attraction très allongé, donc difficilement détectable ;

1.2.5 La fonction de Rastringin définie par

$$J(x, y) = x^2 + y^2 + 30(\sin^2 x + \sin^2 y), \quad (x, y) \in \mathbb{R}^2$$

a de multiples minima locaux. ◁

1.2.6 Soit la fonction $J : \mathbb{R}^6 \to \mathbb{R}$ définie par

$$10^4 \cdot J(t, u, v, w, x, y) = tw(204 + 607x^2)(t + u + v) + uv(187 + 437y^2)(t + 1.57u + w).$$

Sa minimisation avec les contraintes

$$\begin{aligned} &t, u, v, w, x, y \geq 0, \\ &g_1 = 10^5 - (62twx^2(t + u + v) + 58uvy^2(t + 1.57u + w)) \geq 0, \\ &g_2 = tuvwxy - 2070 \geq 0, \end{aligned}$$

modélise le coût d'un transformateur [2, p. 265, Pb 3].

La fonction J n'est pas coercive sur $\mathbb{R}^6_+$, vu que $J(0, 0, v, w, x, y) = 0$. Cependant, les contraintes déterminent un ensemble borné : vu la contrainte g_2, la partie homogène H_5 de degré 5 de la contrainte g_1 restreinte au simplexe $\Sigma_5 = \{t + u + v + w + x + y = 1, t, u, v, w, x, y \geq 0\}$ atteint son minimum m_* en un point intérieur de ce simplexe ; pour un point du domaine contraint on a donc

$$m_* \|(t, u, v, w, x, y)\|_1^5 \leq H_5(t, u, v, w, x, y) \leq 10^5$$

et donc la majoration pour la norme $\|(t, u, v, w, x, y)\|_1 \leq 10/m_*^{1/5}$ sur le domaine contraint.

La recherche de minimum par une méthode de recherche de type Monte-Carlo simple a été comparée avec des recherches localisées avec une gaussienne et une distribution uniforme. La figure I.2 est typique de l'efficacité relative de ces méthodes. C'est la méthode avec gaussienne qui est la plus efficace.

1.3. Recuit simulé. L'algorithme de recuit simulé a sa source dans la conjonction d'une part des mesures de Boltzmann-Gibbs sur un ensemble fini (éventuellement très grand, comme le groupe symétrique $\mathfrak{S}_n$ de cardinal $n!$), d'autre part de la chaîne de Markov introduite par Metropolis-Hastings et ayant comme mesure stationnaire la mesure de probabilité connue seulement par une densité non nécessairement normalisée (comme mesure de probabilité, *i. e.* de masse totale 1). Cet algorithme (comme les algorithmes génétiques ci-dessous) donne une approximation du minimum global du programme : cet avantage est contrebalancé par la relative lenteur de cet algorithme.

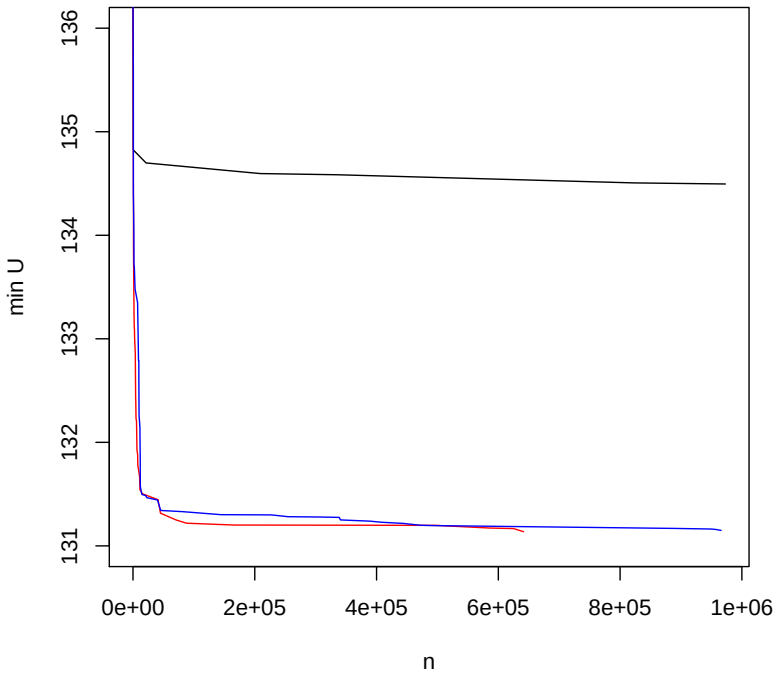


FIGURE I.2 . Obtention d'un minimum approché en fonction du nombre d'itérations : en noir avec une recherche MC, en rouge avec une recherche localisée avec gaussienne, en bleu recherche localisée avec distribution uniforme. On a ici comme valeur minimum $J_* = 131.1371$, avec contraintes $g_1 = 0.1178336$ et $g_2 = 0.00234629$: une application de Lagrange-Newton avec point de départ $(4.804282, 4.211786, 9.839239, 10.23329, 0.9209863, 1.103241)$ converge en 4 itérations vers un point de minimum sur $\{g_1 = 0, g_2 = 0\}$ avec $J_* = 131.0652360127$.

THÉORÈME 1.4 (Mesures de Boltzmann-Gibbs): *Soit M un ensemble fini, $J : M \to \mathbb{R}$ une fonction avec ensemble de minima $M_* = \operatorname{argmin} J$. Alors la mesure de Boltzmann-Gibbs à température $T > 0$*

$$\pi_T^{\mathrm{BG},J}(x) = \frac{\mathrm{e}^{-J(x)/T}}{\sum_{y\in M} \mathrm{e}^{-J(y)/T}}$$

converge vers la mesure uniforme π_0^{BG} de support M_* lorsque $T \to 0^+$.

DÉMONSTRATION. Notons J_* la valeur minimale de J. Alors, pour $x \in M$,

$$\pi_T^{\mathrm{BG},J}(x) = \frac{\mathrm{e}^{-(J(x)-J_*))/T}}{\#M_* + \sum_{y:J(y)>J_*} \mathrm{e}^{-(J(y)-J_*)/T}} \to_{T\to 0^+} \begin{cases} \frac{1}{\#M_*} & \text{si } J(x) = J_*, \\ 0 & \text{sinon.} \end{cases}$$ □

Soit (Ω, μ) un espace de probabilité. Soit π une mesure de densité $\pi(x), x \in \Omega$ absolument continue par rapport à μ, *i. e.* $\pi = \pi(x)\mu$ de masse finie, non nécessairement de masse unité. La méthode de Metropolis-Hastings introduit une chaîne de Markov avec une unique mesure invariante qui est la mesure de probabilité $\pi_1 = \pi / \int_\Omega \pi(x)\mu$: sans connaître le facteur $\int_\Omega \pi(x)\mu$, on arrive à simuler la mesure π_1 !

THÉORÈME 1.5 (Metropolis-Hastings): *Soit (Ω, μ) un espace de probabilités, π une mesure de densité $\pi(x)$ sur Ω, $q(x,y)\mu_y$ avec $x, y \in \Omega$ une probabilité de transition markovienne et A la fonction d'acceptation*

$$A(x,y) = \min\left(\frac{\pi(y)q(x,y)}{\pi(x)q(y,x)}, 1\right),$$

avec $\pi(x), q(x,y) > 0$. Alors le processus $(x_k)_{k\geq 0}$ déterminé par la transition de x_k à x_{k+1} selon

(1) le tirage de y_k suivant la loi $q(x_k, y)\mu_y$

(2) suivi du choix de x_{k+1} suivant la règle

$$x_{k+1} = \begin{cases} y_k & \text{avec probabilité } A(x_k, y_k), \\ x_k & \text{sinon;} \end{cases}$$

induit une chaîne de Markov de probabilité de transition

$$P(x,y) = A(x,y)q(x,y)\mu + \left(1 - \int_\Omega A(x,y)q(x,y)\mu\right)\delta_x$$

dont la mesure stationnaire est la mesure de probabilité de $\pi(x)\mu / \int_\Omega \pi(y)\mu_y$.

Algorithme 1.3 L'algorithme du recuit simulé de recherche d'un minimum

Tirer x_1 suivant la loi g
tant que $k < k_{\max}$ **faire**
  tirer ζ suivant la loi g
  si $J(x_k + \zeta) < J(x_k)$ **alors**
    $x_{k+1} = x_k + \zeta$
  sinon
    tirer π suivant le Bernoulli $\mathcal{B}(\exp((J(x_k) - J(x_k + \zeta))/T_k))$
    si $\pi = 1$ **alors**
      $x_{k+1} = x_k + \zeta$
    sinon
      $x_{k+1} = x_k$
    fin si
  fin si
  $k = k + 1$
fin tant que

L'algorithme de recuit simulé reprend les transitions de Metropolis-Hastings avec une probabilité de transition $q(x,y)$ symétrique, de telle manière que la fonction d'acceptation soit déterminé par le noyau de Boltzmann-Gibbs

$$A_T(x,y) = \exp((J(x) - J(y))/T) \text{ tel que } A_T(x,y) > 1 \text{ si et seulement si } J(y) < J(x)$$

et suivant un schéma de température $(T_k)_{k\geq 1}$ décroissant vers 0 de manière appropriée, comme cela doit être le cas pour refroidir un métal en vue d'obtenir un composite de bonne qualité : ce terme *recuit simulé* rappelle métaphoriquement la prudence à descendre en température pour atteindre un état de qualité, celui de minimum en l'occurrence. On pourra prendre, par exemple, une décroissance logarithmique $T_k = C/\log(1+k)$ ou une décroissance exponentielle[2] $T_k = a^{[k/K]}T_0$: le choix des

2. Si u est réel, $[u]$ désigne sa partie entière.

constantes $C, K > 0, a \in (0,1)$ est crucial pour la convergence et découle de manière empirique. Il a été montré que les schémas de températures tels que $\lim_{k\to\infty} T_k = 0$ et $\sum_{k=1}^{\infty} \exp(-h_*/T_k) = +\infty$ sont convergents.

Heuristiquement, si $T_k = (1+[(\log k)/K])^{-1}$, la température est constante sur chaque pallier $[\mathrm{e}^{KN}, \mathrm{e}^{K(N+1)}) \cap \mathbb{N}$ où le processus de Markov homogène de Metropolis-Hastings converge de manière significative vers $\pi_{T_k}^{\mathrm{BG}}$: sur ce pallier, la distribution de la suite x_k approche la mesure de Boltzmann-Gibbs de température T_k. Vu la convergence $T_k \to 0$, la suite x_k converge en probabilité vers la mesure π_0^{BG} localisée sur l'ensemble des minima de J. La stricte positivité de Q donne parfois des itérés x_k pouvant s'échapper du voisinage d'un extremum local vers l'extremum global recherché.

Les lignes 7-8 de l'algorithme 1.3 peuvent être traduites en code R suivant

```
p = exp(-(J(x[k]+z)-J(x[k]))/T[k])
x[k+1]= x[k] + z*(runif(1)<p)
```

▷ Exemple 1.3: Soit la fonction J définie sur $[0,1]$ par $J(x) = (\cos(50x) + \sin(20x))^2$. Elle a plusieurs maxima locaux. ◁

1.4. Algorithmes génétiques. Instances particulières de méthodes dites d'*évolution*, les algorithmes génétiques [AG] sont inspirés par l'évolution des espèces avec des phases de sélection, croisement et mutation. Ils mettent en place l'évolution d'une population de cardinal N, *i. e.* N trajectoires itératives simultanées interagissant entre elles et laissant espérer une convergence de certains individus de la population vers un point d'optimum de la fonction d'objectifs (fonction de coût pour un minimum, fonction d'adaptivité à maximiser pour continuer la métaphore de l'évolution). C'est une extension des modes de recherche précédents, où on envisage plusieurs trajectoires (indépendantes) soit pour attester de la convergence, soit pour avoir des espérances du nombre d'itérations nécessaires pour obtenir une précision espérée. Les AGs ont été initialement développés en mode discret pour la recherche d'extrema sur des espaces finis de grand cardinal : un exemple typique consiste en l'espace $\mathcal{C}_{\mathcal{A},M} = \mathcal{A}^M$ de mots codés c de longueur M sur un alphabet $\mathcal{A}$: un c est un chromosome, caractérisé par une chaîne d'acides aminées. Nous en donnons ici un exemple en mode intrinsèquement continu : cet exemple n'est pas induit par un AG dérivant de la discrétisation opérée *de facto* lors de la transcription d'un modèle continu dans un ordinateur au caractère intrinsèquement discret.

Un individu est un point du domaine convexe fermé $\Omega \subset \mathbb{R}^n$, une population de N individus est un N-uplet $\mathbf{x} = (x_1, \ldots, x_N) \in \Omega^N$. L'algorithme génétique consiste à l'itération de trois opérations de base, la sélection, le croisement et la mutation :

(1) Pour préparer la sélection dans une population $\mathbf{x}$ de cardinal N, on commence par classer les individus suivant la valeur de leur fonction de coût J (supposée non négative), de la plus petite à la plus grande, soit une population ordonnée $\widetilde{\mathbf{x}}$ avec pour $\widetilde{x}_i$ la probabilité $p_i = J(\widetilde{x}_i)/\sum_j J(x_j)$: on calcule la fonction de répartition $F_{J,\mathbf{x}}$: $F_{J,\mathbf{x}}(i) = \sum_{k=1}^{i} p_k$ On fait N tirages aléatoires pour la loi $\mathcal{U}(0,1)$ uniforme sur $(0,1)$, retenant à chaque tirage ρ l'individu $\widetilde{x}_1$ si $\rho < F_{J,\mathbf{x}}(1)$, $\widetilde{x}_i$ si $F_{J,\mathbf{x}}(i-1) < \rho \leq F_{J,\mathbf{x}}(i)$ sinon. Cette méthode de sélection retient tendanciellement les individus de coût petit.

(2) Pour le croisement modulé par les paramètres $p_c, a \in (0,1)$, on répartit aléatoirement les individus en paires de la population, puis pour chaque paire (x,y) et suivant le succès du tirage de Bernoulli $\mathcal{B}(p_c)$, la paire (x,y) est remplacée par la paire de combinaisons convexes $(ax + (1-a)y, (1-a)x + ay)$.

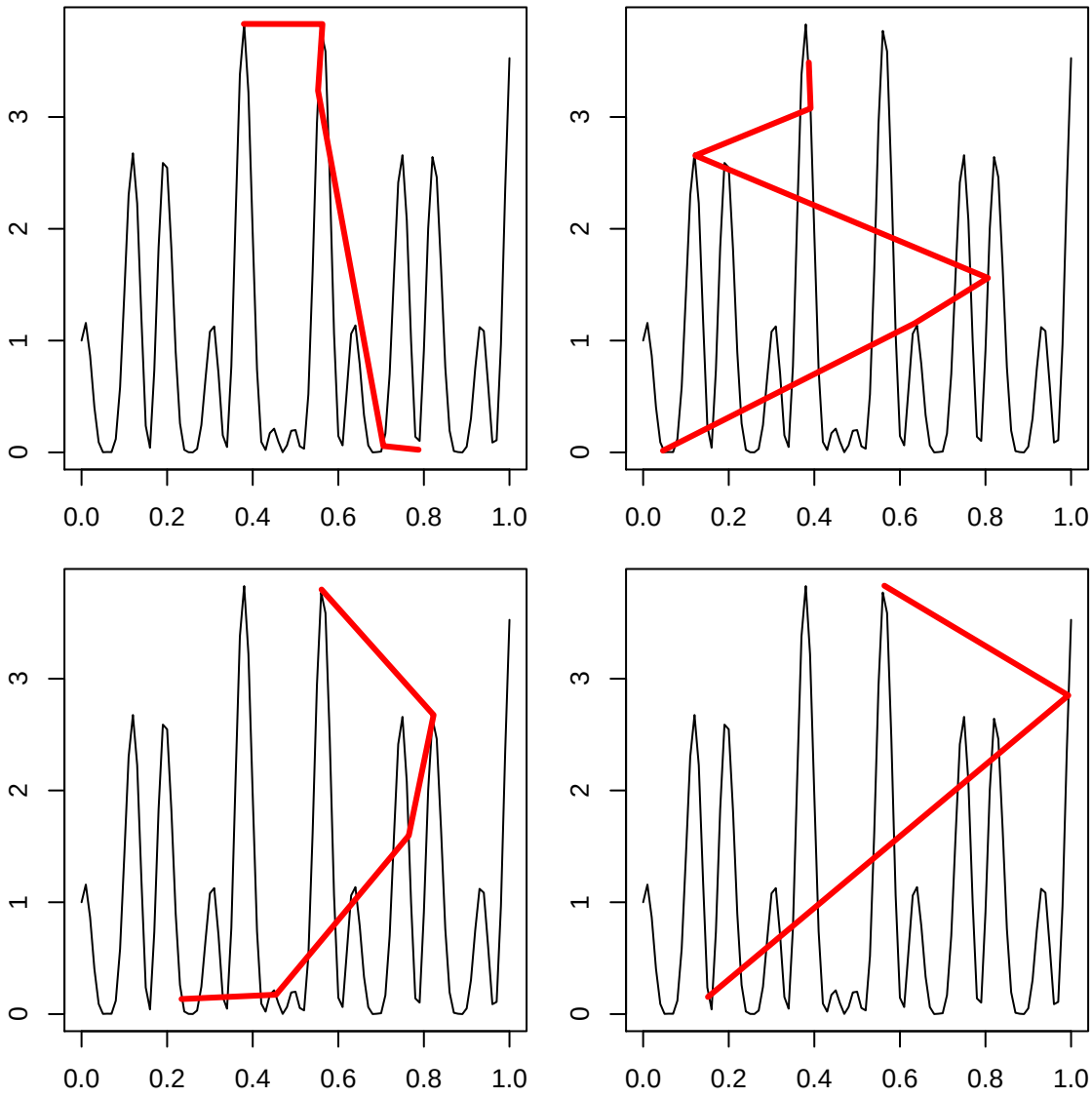


FIGURE I.3 . Des convergences de recuit-simulé pour les maxima de la fonction $J(x) = (\cos(50x) + \sin(20x))^2$ sur $[0,1]$ et suivant quatre schémas de températures $(1/\log(1+t), (1+t)^{-2}, 100/(\log(1+t)), 100/(\log(1+t))))$.

(3) L'opération de mutation modulée par un paramètre $p_m \in (0,1)$ et un poids $\sigma > 0$ consiste à remplacer chaque individu x après tirage aléatoire favorable d'un Bernoulli $\mathcal{B}(p_m)$ en un individu $p_\Omega(x+\sigma\varepsilon)$ où ε est la gaussienne normale $\mathcal{N}(0,\sigma)$ et p_Ω est la projection sur le convexe Ω.

▷ EXEMPLE 1.4: Soit J la fonction sur $\mathbb{R}^2$ définie par $J(m) = \|m\|_2^2$. On considère une population de $N = 10$ individus avec $p_c = 0.6$, $p_m = 0.9$ et $\sigma = 0.1$. Le tableau 1.4 indique des résultats pour la recherche par AG du minimum de J dans $[-1,1]^2$. ◁

2. Optimisation convexe suivant gradient stochastique

Considérons le programme

$$\min_{x\in C} \mathbb{E}_\Omega[u(x,\omega)], \quad \operatorname{argmin}_{x\in C} \mathbb{E}_\Omega[u(x,\omega)], \tag{4}$$

Algorithme 1.4 Algorithme génétique pour $J : \Omega(\subset \mathbb{R}^n) \to \mathbb{R}$

1: Choisir N, p_c, a, p_m, σ
2: Construire une population de N individus (assimilés à des points de Ω)
3: **tant que** $k \leq k_{\max}$ **faire**
4: Ordonner la population et opérer la sélection
5: Croiser des paires d'individus avec les poids p_c, a
6: Muter les individus avec les poids p_c, σ
7: **fin tant que**

0	1	2	5	6	7
.166926	0.1149113	0.04807564	0.02083576	0.01965758	0.01254407
8	9	10	15	19	38
0.006468982	0.003227048	0.0006965387	0.0003170936	1.001663e-05	9.359423e-06

TABLE 4. Suite des valeurs minimales pour un AG appliqué à la fonction $m \mapsto \|m\|^2$ pour m dans le pavé $[-1, 1]^2$.

où $(\Omega, \mathcal{T}, P)$ est un espace probabilisé, C un convexe (éventuellement compact) de l'espace des variables de décision, $u : C \times \Omega \to \mathbb{R}$ une fonction souvent supposée convexe relativement à la variable x. Si l'espérance est calculable aisément, on est ramené au programme déjà considéré pour la fonction U définie par

$$U(x) = \mathbb{E}_\Omega[u(x, \omega)] = \int_\Omega u(x, \omega) dP(\omega),$$

avec par exemple les méthodes de *gradient déterministe* (dit parfois *gradient complet*) consistant en des itérations de descente dans la direction de ∇U. Néanmoins, ce calcul de l'espérance est parfois insurmontable, du fait de la taille des données ou même de l'arrivée progressive des termes constituant l'espérance.

▷ EXEMPLE 1.5: Dans la théorie de l'apprentissage (*machine learning* en anglais), on considère la recherche du point de minimum de fonctions exprimant un coût U_e (ou risque, perte, voire *loss*,...) empirique

$$U_e(x) = \frac{1}{n} \sum_{j=1}^{n} [\ell(x, \omega_j) + U_0(x)] = \frac{1}{n} \sum_{j=1}^{n} \ell(x, \omega_j) + U_0(x), \tag{5}$$

à comparer avec le coût espéré $\mathbb{E}[(\ell(x, \omega) + U_0(x)])$. Comme fonction ℓ de coût [3], on a l'exemple celle des moindres carrés

$$u : (x = (a, b), \omega = (\alpha, \beta)) \in \mathbb{R}^{p \times n} \times \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^p \mapsto \|\beta - a\alpha - b\|_2^2$$

ou celle de régression logistique

$$u : (x, \omega = (\alpha, \beta)) \in \mathbb{R}^p \times \mathbb{R}^p \times \{\pm 1\} \mapsto \log(1 + \mathrm{e}^{-\beta \langle x, \alpha \rangle}),$$

tandis que U_0 est une fonction de régularisation (par exemple la combinaison de Ridge et Lasso $U_0(x) = \lambda \|x\|_2^2 + (1 - \lambda)\|x\|_1$). Le nombre d'observations n est très grand [4] et rend rédhibitoire le calcul du gradient pour cette somme. Considérant l'ensemble des entiers

3. Les fonctions classiques sont du type $(w, (x, y)) \in C_w \times \mathbb{R}^n_{(x,y)} \mapsto \ell(y, \langle x, w \rangle)$, avec la fonction $v \mapsto \ell(y, v)$ convexe.

4. Dans le contexte de données massives, la dimension p décrivant chaque donnée observée est aussi très grande. La dépendance des vitesses de convergence des algorithmes et de leur complexité vis-à-vis des paramètres n, p est d'importance, cf. Bottou [**?**].

$\Omega_n = [\![1, n]\!]$ avec la mesure uniforme, la moyenne empirique (5) est du type $U = \mathbb{E}_{\Omega_n}[u]$ avec $u : (x, i) \in \mathbb{R}^N \times \Omega_n \to u(x, i)$. Le programme de minimisation porte sur des moyennes : l'apparition des propriétés de moyenne statistique est naturelle, et de fait, ces techniques d'approximation statistique se sont révélées fécondes. Par ailleurs, dans ce cas particulier d'une fonction de coût somme de n termes, une étape de l'itération ne dépend pas de n et le gradient $\nabla_x u(x, \omega)$ est un estimateur non biaisé du gradient $\nabla_x \mathbb{E}[u]$. ◁

Le gradient $\nabla_x U(x) = \nabla_x \mathbb{E}[u(x, \omega)]$ de l'espérance est égale à l'espérance $\mathbb{E}[\nabla_x u(x, \omega)]$ du gradient grâce à des hypothèses convenables sur la fonction $u(x, \omega)$. On approche ce gradient $\nabla_x U$ en considérant une approximation à la Monte-Carlo de l'espérance du gradient $\nabla_x u(x, \omega)$

$$\nabla_x \mathbb{E}_\Omega[u(x, \omega)] = \mathbb{E}_\Omega[\nabla_x u(x, \omega)] = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \nabla_x u(x, \omega_k),$$

avec $\omega_1, \omega_k, \ldots,$ représente une suite de tirages aléatoires indépendants identiquement distribués sur Ω. On suppose que la fonction u est différentiable par rapport à la variable x, avec les conditions assurant la dérivation sous l'intégrale, ou plus généralement u avec un sous-gradient $\xi(x, \omega) \in \partial_x u(x, \omega)$ pour presque tout ω, de telle sorte que $\Xi(x) = \mathbb{E}_\Omega[\xi(x, \omega)]$ soit un sous-gradient dans le sous-différentiel $\partial_x U$. Le gradient de la fonction coût du programme

$$\frac{1}{J} \sum_{j=1}^{J} \nabla_x u(x, \omega_j)$$

correspond à l'approximation de Monte-Carlo du gradient complet $\nabla_x \mathbb{E}[u(x, \omega)]$. L'inconvénient majeur de cette méthode est la définition de l'entier K avant la résolution du problème d'optimisation approché : au cas où cette taille soit insuffisante, il faut, après extension de l'échantillon, redémarrer l'optimisation.

Face au programme (4), ou même le programme de la moyenne empirique suivant l'itération

$$x_{k+1} = x_k - a_k \Gamma_k \frac{1}{J} \sum_{j=1}^{J} \nabla_x u(x_j, \omega_j) \tag{6}$$

on est amené donc à introduire l'itération élémentaire (et de complexité dérisoire)

$$x_{k+1} = P_C \left[x_k - a_k \Gamma_k \nabla_x u(x_k, \omega_k) \right], \quad k \geq 1, \tag{7}$$

avec un seul terme aléatoire (drastique simplification, représentant un bruit aléatoire qui l'espère-t-on ne nuit pas à la convergence recherchée) qui contribue à l'approximation de l'espérance du gradient tout en orientant la descente vers le point de minimum : d'une part la descente vers le minimum peut avoir lieu avec une amplitude a_k *constante* (*i. e.* dans un intervalle $[a_-, a_+]$ avec $a_- > 0$), cela sera forcé avec un a_k ne décroissant pas trop rapidement, d'autre part une décroissante assez rapide pour avoir l'approximation de Monte-Carlo. Le choix d'une suite $\mathbf{a} \in \ell^2 \setminus \ell^1$ combine ces deux actions.

△ Remarques 1.3:

(1) Vu que l'algorithme ne doit pas mémoriser les états visités durant les itérations, il semble pouvoir aussi traiter des données arrivant au fil de l'eau.

(2) Le facteur matriciel Γ_k (introduit dans les itérations de descente déterministe pour un meilleur conditionnement du vecteur gradient, facteur quasi-newtonien par ex.) améliore peu les performances de l'algorithme du gradient stochastique.

(3) L'introduction de l'aléa ω transforment les variables x_k en variables aléatoires : la suite $(x_k(\omega))$ est en fait un processus aléatoire, de plus markovien : la variable aléatoire x_{k+1} ne dépend que de l'aléa ω_k et de la variable $x_k(\omega)$ au temps k.

(4) Pour la somme empirique (5), on tirera au sort un entier $j_k \in [\![1, n]\!]$ pour poser

$$x_{k+1} = x_k - a_k \nabla_x \left[\ell(x_k, \omega_{j_k}) + U_0(x_k)\right].$$

(5) La méthode du gradient stochastique n'est pas une méthode de descente, considérée en espérance cette méthode l'est.

(6) S'il y a redondance parmi les données (et c'est le cas dans les masses de données à la base des procédures d'apprentissage), la méthode de gradient complet gaspille des ressources en prenant en compte tous les gradients $\nabla_x \ell(x, i)$. ▽

Sous hypothèse de convexité de la fonction u (convexité forte ou lisse) et pour un choix approprié de la suite $\mathbf{a}$ des amplitudes (souvent dans $\ell^2 \setminus \ell^1$), on montre que ces actions complémentaires assurent une convergence vers les minimum du programme, point de maximum ou valeur minimale, en moyenne quadratique, voire presque sûrement.

Le théorème suivant établit une convergence en moyenne quadratique vers le point de minimum. D'autres résultats indiquent la convergence de $\mathbb{E}[U(x_k)]$ vers $U(x_*)$ avec un reste $\mathcal{O}(1/\sqrt{k})$, amélioré en $\mathcal{O}(1/k)$] avec une hypothèse de forte convexité [**?**].

$$\mathbb{E}[\|x_k - x_*\|^2] = \mathcal{O}(k^{-\beta}), \quad \mathbb{E}[U(x_k) - U(x_*)] = \mathcal{O}(\sqrt{k^{-1/2}})$$

avec β dans certains intervalles de $(0, 1]$, avec des convergences plus ou moins renforcées (voire exponentielle) sous des hypothèses de convexité convenables.

Les itérations successives de type gradient sont accompagnées de réalisations de la variable ω qui permet d'évaluer une espérance comme sait le faire la méthode de Monte-Carlo.

▷ Exemple 1.6: Soit V une variable aléatoire sur Ω, de carré intégrable et dont on souhaite calculer l'espérance $\mathbb{E}[V] = \int_\Omega V(\omega) dP(\omega)$. Une variable aléatoire fluctue autour de son espérance (sa moyenne). Cette espérance correspond au point $U(x_*)$ autour duquel la dispersion de V est minimale, *i. e.* qui minimise la fonction $U : x \in \mathbb{R} \mapsto U(x) = \mathbb{E}_\Omega[(x - V(\omega))^2]$. Développant l'espérance, on obtient

$$U(x) = x^2 - 2x\mathbb{E}_\Omega[V] + \mathbb{E}_\Omega[V^2],$$

et $\operatorname{argmin}_{x\in\mathbb{R}} \mathbb{E}_\Omega[(x - V(\omega))^2] = \mathbb{E}_\Omega[V]$: l'espérance correspond à la valeur autour de laquelle la dispersion de la variable est minimale. Si on introduit la moyenne $M_k(\overline{\omega}) = \frac{1}{k}\sum_{j=1}^k V(\omega_j)$ pour $\overline{\omega} = (\omega_1, \omega_2, \dots) \in \Omega^{\mathbb{N}}$ un échantillon de tirages indépendants identiquement distribués sur Ω, la loi forte des grands nombres énonce que la moyenne M_k converge presque sûrement vers $\mathbb{E}_\Omega[V]$. Par ailleurs, on a

$$\begin{aligned} M_{k+1}(\overline{\omega}) &= \frac{k}{k+1}\left[M_k(\overline{\omega}) + \frac{V(\omega_{k+1})}{k}\right] = M_k(\overline{\omega}) - \frac{M_k(\overline{\omega})}{k+1} + \frac{V(\omega_{k+1})}{k+1} \\ &= M_k(\overline{\omega}) - \frac{1}{k+1}\left[M_k(\overline{\omega}) - V(\omega_{k+1})\right] \\ &= M_k(\overline{\omega}) - \frac{1}{k+1}\nabla_x \left[\frac{[x - V(\omega_{k+1})]^2}{2}\right]_{x=M_k(\overline{\omega})} \end{aligned}$$

On retrouve l'itération (7) avec $a_k = (k+1)^{-1}$ et $u(x, \omega) = |x - V(\omega)|^2/2$: la loi des grands nombres assure donc la convergence de l'itération (7) dans ce cas particulier. L'approximation de l'espérance à la Monte-Carlo peut être présentée comme une itération de

type *méthode gradient stochastique* :

$$U(x) = \mathbb{E}_\Omega[(x - V(\omega)^2]/2 = \mathbb{E}_\Omega[u_V(x,\omega)] \text{ avec } u_V(x,\omega) = (x - V(\omega))^2/2.$$

qui permet d'éviter le calcul de l'espérance $\mathbb{E}_\Omega[(u-V)^2]$ dont on prendrait le gradient : on remplace le calcul du gradient $\nabla_x \mathbb{E}[u_V(x,\omega)]$ par la considération immédiate du gradient $\nabla_x u_X(x,\omega)$ prise en l'aléa ω. On remarquera que si le pas a_k tend vers 0 (au contraire des méthodes de descente de gradient déterministes), il ne le fait pas trop rapidement (la famille (a_k) n'est pas sommable). La convergence du gradient stochastique est (dans cet exemple et sans doute plus généralement) celle de la loi des grands nombres : on s'attend donc à une convergence presque sûre, de même que des estimations sur le type de convergence (corrélat du théorème de la limite centrale). ◁

THÉORÈME 1.6 (Approximation stochastique [19]): *Soit C convexe fermé avec opérateur de projection P_C, $(\Omega, \mathcal{T}, dP)$ un espace probabilisé, $(\Omega, \mathcal{T}, dP)$ un espace probabilisé, $u : (x,\omega) \in C \times \Omega \mapsto u(x,\omega)$ et $U : x \in C \mapsto U(x) = \mathbb{E}[u(x,\omega)]$ différentiable en x, m-convexe. On suppose le gradient $x \in C \mapsto \nabla_x u(x,\omega)$ uniformément borné et L-lipschitzien. Supposons que x_* dans l'intérieur de C soit solution du programme* $\inf_{x\in C} U(x)$ *Alors, il existe $\theta > 0$ tel que l'itération stochastique*

$$x_{k+1}(\omega) = P_C\left[x_k(\omega) - \frac{\theta}{k}\nabla_x u(x_k,\omega_k)\right]$$

converge en moyenne quadratique vers x_* avec

$$x_k(\omega) \overset{L^2}{=} x_* + \mathcal{O}(1/\sqrt{k}), \quad U(x_k(\omega)) \overset{L^1}{=} U(x)_* + \mathcal{O}(1/k),$$

DÉMONSTRATION. Soit $e_k(\omega) = \|x_k(\omega) - x_*\|_2^2$ et $E_k = \mathbb{E}[e_k]$. L'itération x_k dépend de l'aléa ω que nous considérons dans $\Omega^{\mathbb{N}}$ afin d'alléger les notations : $x_k(\omega)$ dépend en fait seulement de la suite finie $(\omega_1, \ldots, \omega_{k-1})$: les espérances $\mathbb{E}_{\Omega^{\mathbb{N}}}[u(x_k,\omega)]$ sont en fait des espérances sur Ω^{k-1}. On s'intéresse donc aux convergences de quantités liées aux processus dépendant de l'espace probabilisé Ω en moyenne (quadratique ou L^1).

Grâce au caractère contractant de la projection P_C et du fait que $x_* \in C$,

$$\begin{aligned} e_{k+1}(\omega) &= \|P_C[x_k(\omega) - \alpha_k \nabla_x U(x_k,\omega_k] - x_*\|_2^2 = \|P_C[x_k(\omega) - \alpha_k \nabla_x u(x_k,\omega_k)] - P_C[x_*]\|_2^2 \\ &\leq \|x_k(\omega) - \alpha_k \nabla_x(x_k,\omega_k) - x_*\|_2^2 \\ &= e_k - 2\alpha_k \langle x_k(\omega) - x_*, \nabla_x u(x_k,\omega_k)\rangle + \alpha_k^2 \|\nabla_x u(x_k,\omega_k\|_2^2 \end{aligned}$$

Par ailleurs,

$$\begin{aligned} \mathbb{E}\left[\langle x_k(\omega) - x_*, \nabla_x u(x_k,\omega_k)\rangle\right] &= \mathbb{E}_{\omega_{[k-1]}}\left[\mathbb{E}_{\omega_k}[\langle x_k - x_*, \nabla_x u(x_k,\omega_k)\rangle]\right] \\ &\leq \mathbb{E}_{\omega_{[k-1]}}\left[\langle x_k - x_*, \mathbb{E}[\nabla_x u(x_k,\omega_k)\rangle]\right] \\ &= \mathbb{E}[\langle x_k - x_*, \nabla U_x(x_k)\rangle] \end{aligned}$$

et donc en prenant l'espérance de l'inégalité précédente et en notant $M^2 = \sup_{x\in C} \mathbb{E}[\|\nabla_x u(x,\omega)\|_2^2]$ (constante finie d'après l'hypothèse)

$$E_{k+1} \leq E_k - 2\alpha_k \mathbb{E}[\langle x_k - x_*, \nabla_x U(x_k)\rangle] + \alpha_k^2 M^2. \tag{8}$$

Par m-convexité, on a

$$\langle x' - x, \nabla U(x') - \nabla U(x)\rangle \geq m\|x - x'\|^2.$$

Alors, avec l'inégalité de Euler-Fermat $\langle x - x_*, \nabla U(x_*)\rangle \geq 0$ au point de minimum x_*, on obtient

$$\mathbb{E}[\langle x_k - x_*, \nabla U(x_k)\rangle] \geq \mathbb{E}[\langle x_k - x_*, \nabla U(x_k) - \nabla U(x_*)\rangle] \geq m\mathbb{E}[\|x_k - x_*\|_2^2] = mE_k$$

et donc

$$E_{k+1} \leq (1-2m\alpha_k)E_k + \alpha_k^2 M^2,$$

soit, sous l'hypothèse $\alpha_k = \theta/k$,

$$E_{k+1} \leq (1-2m\theta/k)E_k + \theta^2 M^2/k^2.$$

Le lemme suivant portant sur une suite déterministe, permettra de conclure.

LEMME 1.1: *Soit $\theta > 1/(2m)$, $C = \theta^2 M^2$ et $\kappa = \max(C(2m\theta-1)^{-1}, a_1)$. Si (E_k) vérifie la relation de récurrence $E_{k+1} \leq (1-2m/k)E_k + C/k^2$, alors $E_k \leq \kappa/k$ pour tout entier $k \geq 1$.*

Preuve du lemme. □

Le lemme assure donc de l'estimation quadratique $\sqrt{E_k} = \|x_k - x_*\|_{L^2(\Omega)} = \mathcal{O}(1/\sqrt{k})$.

On obtient une convergence en moyenne plus rapide pour les valeurs $U(x_k(\omega))$. En effet, le gradient ∇U étant L-lipschitzien et nul au point de minimum x_* supposé intérieur à C, on obtient

$$U(x) \leq U(x_*) + \frac{L}{2}\|x - x_*\|_2^2$$

et par suite

$$\mathbb{E}[U(x_k) - U(x_*)] \leq \frac{LE_k}{2} \leq \frac{L\kappa}{2k}.$$

□

△ REMARQUE 1.4: L'erreur espérée sur la valeur $U(x_*)$ au terme de k itérations est d'ordre $\mathcal{O}(1/k)$ et celle sur l'approximation du point de minimum en $\mathcal{O}(k^{-1/2})$. Ces bornes dépendent de constantes (M, L, θ et surtout $m\ldots$) bien choisies. Par exemple, si $U : x \in [-1,1] \mapsto x^2/10$ donne lieu à l'itération $x_{k+1} = x_k - U'(x_k)/k = (1-1/(5k))x_k$: on a pris $\alpha_k = 1/k$ avec $m = 1$, alors que $m = 0.2$ est la constante de m-convexité. On a, en remarquant $1 - u^{-1} = (1 + (u-1)^{-1})^{-1}$ et usant de l'inégalité de convexité $-\log(1+v) \geq -v$,

$$x_k = \prod_{\ell=1}^{k-1}(1-1/(5\ell)) = \exp\left[-\sum_{\ell=1}^{k-1}\log(1+(5\ell-1)^{-1})\right] \geq \exp\left[\sum_{\ell=1}^{k}(5\ell-1)^{-1}\right]$$
$$\geq \exp\left[-1/4 - \int_1^{k-2}(5t-1)^{-1}dt\right] \geq C'\exp[-\log(5k-11)/5] \geq C'k^{-1/5}$$

pour une certaine constante $C' > 0$, ce qui indique bien la lenteur de la convergence (pour $k = 10^9$, l'erreur est minorée par 0.015. Si on avait pris $a = 1/m = 5$, on aurait atteint la solution $x_* = 1$ en une itération à partir de $x_1 = 1$. ▽

THÉORÈME 1.7 (Approximation stochastique [19]): *Plaçons-nous dans le cadre du théorème précédent, hormis l'hypothèse de m-convexité. Soit un entier N. En choisissant un pas constant $a(N) = C(U,X)/\sqrt{N}$ adapté à N, le point $\widetilde{x}_N$ obtenu comme iso-barycentre des N premiers points de l'itération minimise en moyenne la valeur minimale $U(x_*)$ à $\mathcal{O}(N^{-1/2})$ près,* i. e. *il existe une constante C_1 telle que $\mathbb{E}(U(\overline{x}*) - U(x_N)) \leq C_1/\sqrt{N}$.*

DÉMONSTRATION. Par convexité de la fonction U„

$$U(x) \geq U(x_k) + \langle x - x_k, \nabla_x U(x_k)\rangle,$$

d'où on obtient

$$\mathbb{E}[\langle x_k - x_*, \nabla_x U(x_k)\rangle] \geq \mathbb{E}[U(x_k) - U(x_*)]$$

puis, en reprenant l'inégalité (8),

$$\alpha_\ell \mathbb{E}[U(x_\ell) - U(x_*)] \leq E_\ell - E_{\ell+1} + \alpha_\ell^2 M^2$$

et donc en sommant

$$\sum_{\ell=1}^{k} \alpha_\ell \mathbb{E}[U(x_\ell) - U(x_*)] \leq \sum_{\ell=1}^{k} [E_\ell - E_{\ell+1}] + M^2 \sum_{\ell=1}^{k} \alpha_\ell^2 \leq E_1 + M^2 \sum_{\ell=1}^{k} \alpha_\ell^2$$

puis, avec les coefficients convexes $\lambda_\ell = \alpha_\ell / \sum_{j=1}^{k} \alpha_j$

$$\mathbb{E}\left[\sum_{\ell=1}^{k} \lambda_\ell U(x_\ell) - U(x_*)\right] \leq \frac{E_1 + M^2 \sum_{\ell=1}^{k} \alpha_\ell^2}{\sum_{\ell=1}^{k} \alpha_\ell}$$

Considérons le point $\widetilde{x}_k = \sum_{\ell=1}^{k} \lambda_\ell x_\ell$: la convexité de U donne donc

$$\mathbb{E}\left[U(\widetilde{x}_k) - U(x_*)\right] \leq \frac{E_1 + M^2 \sum_{\ell=1}^{k} \alpha_\ell^2}{\sum_{\ell=1}^{k} \alpha_\ell}$$

Supposons maintenant un nombre donné N d'itérations, avec un pas constant $\alpha = \alpha(N)$ que l'on va déterminer pour optimiser les inégalités précédentes. Le point $\widetilde{x}_N$ est l'isobarycentre $\widetilde{x}_N = N^{-1} \sum_{\ell=1}^{N} x_\ell$ qui vérifie

$$\mathbb{E}\left[U(\widetilde{x}_N) - U(x_*)\right] \leq \frac{E_1}{\alpha N} + M^2 \alpha \leq \frac{D_X^2}{\alpha N} + M^2 \alpha \tag{9}$$

où on a introduit $D_X = \max_{x \in X} \|x - x_1\|_2$ qui vérifie $D_X^2 \geq E_1$. Alors, en prenant $\alpha = D_X/(M\sqrt{N})$ (qui minimise le membre de droite dans (9)), on obtient

$$\mathbb{E}[U(\widetilde{x}_N) - U(x_*] \leq \frac{D_X M}{\sqrt{N}},$$

ce qui était annoncé. □

△ Remarque 1.5: Le résultat est plus faible que celui du théorème. Cependant, les hypothèses sont plus légères. Le résultat vaut aussi pour toute N-itération avec un pas du type $\theta/\sqrt{N}$. Ces qualités justifient à nommer cette méthode de recherche de point de minimum approché comme la méthode d'*approximation stochastique robuste*. ▽

Terminons avec quelques commentaires sur l'efficacité asymptotique et la moyennisation. Le prédicontionnement du gradient (par la multiplication d'une matrice convenablement calibrée) sera aussi utile que dans le cas de méthodes déterministes (par ex. la méthode de Newton-Raphson). Cela mène à la notion d'algorithme de gradient stochastique Newton efficace avec pas décroissant en $a_k = (k+\beta)^{-1}$ et matrice A d'ajustement donnant l'itération $x_{k+1} = x_k - a_k A_k x_k$ tel que

$$\sqrt{k}(x_k - x_*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, H^{-1}\Gamma H^{-1})$$

Un tel algorithme Newton efficace a été introduit par Polyak en ajoutant une moyennisation dans l'itération stochastique :

$$x_{k+1}^M = \frac{1}{k+1} \sum_{j=1}^{k} x_j$$

ou sa forme récursive

$$x_{k+1}^M = x_k^M + \frac{1}{k+1}(x_{k+1} - x_k^M)$$

où Cesàro assure la convergence presque sure de la moyenne (x_k^M) comme corrélat de celle de (x_k).

La mise en place effective de l'algorithme de gradient stochastique pose diverses questions. Tout d'abord quel critère d'arrêt choisir : la norme $\|x_{k+1} - x_k\|$ converge vers 0 comme a_k et ne peut pas être considérée pour un test de convergence, le gradient $\|\nabla u(x_k, \omega_k)\|_2$ ne converge pas nécessairement vers 0, au contraire de l'espérance $\mathbb{E}[\|\nabla u(x_k, \omega_k)\|_2]$, qu'on estimera par

$$\left(\sum_{\ell=1}^{k} a_\ell\right)^{-1} \left(\sum_{\ell=1}^{k} a_\ell \nabla_u(x_\ell, \omega_\ell)\right)^{-1}.$$

L'espérance de la variable aléatoire $\nabla_x u(x_k, \omega_k)$ converge vers $\nabla_x J(x_*)$, étant possiblement à la base de tests de convergence. L'observation de la convergence sera souvent effectuée par des examens huristiques.

La deuxième question porte que le choix des paramètres $\alpha, \beta.\gamma$ de la suite $a_k = \gamma/(k+\beta)^\alpha$. Enfin, la moyennisation est à déclencher après un certain nombre d'itérations à choisir de manière plus ou moins arbitraire.

3. L'algorithme espérance/maximisation

L'algorithme $\mathbb{EM}$ pour *Espérance/Maximisation* est une technique largement utilisée en statistique paramétrique : calcul complexe de maxima de vraisemblance, analyse de modèles à données manquantes, étude de mélanges. Il a été introduit formellement en 1997 par Dempster, Laird, and Rubin [11], mais est présent dès la fin du XIXe siècle sous de multiples avatars.

Après le développement d'un exemple de régression linéaire avec une variable manquante suivant une méthode qui se révèle être de type $\mathbb{EM}$, le cadre général de la méthode $\mathbb{EM}$ est posé, la convergence explicitée pour deux exemples (loi exponentielle et loi normale) est établie, le résultat de monotonie général est explicité, puis cette section se termine par l'étude de quelques exemples particuliers, éclairés parfois par des simulations numériques.

3.1. Données observées, données latentes : un exemple en régression linéaire [32]. Étant donné le modèle linéaire $x = A\Lambda + \varepsilon$, la détermination du paramètre (vectoriel) $\Lambda \in \mathbb{R}^\ell$ en fonction des valeurs explicatives $A \in \mathbb{R}^{n_x} \otimes \mathbb{R}^\ell$ et des données observées $x \in \mathbb{R}^{n_x}$ est bien établie théoriquement : c'est le Λ_* minimisant [5] $\|x - A\Lambda\|$, soit $\Lambda_* = ({}^{\mathsf{T}}AA)^{-1}{}^{\mathsf{T}}Ax$. Cependant, le calcul de l'opérateur $M_A = ({}^{\mathsf{T}}AA)^{-1}{}^{\mathsf{T}}A$ n'est pas toujours simple ou adapté aux statistiques [6] des données observées x. Compléter les variables observées x en $y = {}^{\mathsf{T}}({}^{\mathsf{T}}x, {}^{\mathsf{T}}z)$ permet éventuellement de se ramener à un modèle linéaire augmenté/complété $y = A_c\Lambda + \varepsilon$ où, grâce aux propriétés particulières de A_c, le calcul $\mapsto ({}^{\mathsf{T}}A_cA_c)^{-1}{}^{\mathsf{T}}A_cy$ apparaît aisé en termes des statistiques de y. Cette section introductive est consacrée à la mise en place d'un processus itératif basculant du modèle initial $x \simeq A\Lambda$ vers le modèle augmenté $y \simeq A_c\Lambda$ et vice-versa : ce processus évite le calcul de l'opérateur problématique M_A et détermine par approximations successives le paramètre Λ_*, processus à la convergence rigoureusement établie (comme pour d'autres

5. Autrement dit, le Λ_* maximise la vraisemblance $\exp(-\|x - A\Lambda\|^2/2\sigma^2)/(\sqrt{2\pi}\sigma)^{\dim x}$, l'erreur ε étant supposée gaussienne centrée de variance σ^2.

6. :Les coefficients α, β de la régression linéaire simple en moindres carrés $\min_{\alpha,\beta} \|x - \alpha - \beta t\|_2^2$ a une expression simple en termes des statistiques de x : $\beta_* = (\overline{t * x} - \overline{t}\ \overline{x})/(\overline{t^2} - \overline{t}^2)$ et $\alpha_* = \overline{x} - \beta_*\overline{x}$.

modèles linéaires). Cet exemple exhibe une incarnation simple de l'algorithme $\mathbb{EM}$, processus fort utilisé en statistique bien que sa convergence ne soit pas établie en général, autrement que de manière heuristique.

Soit $\mathcal{Y} = \{1,2\} \times \{1,2,3\}$ et $\mathcal{X} = \mathcal{Y} \setminus \{(2,3)\}$. On considère le modèle de régression linéaire [7]

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad (i,j) \in \mathcal{X},$$

où $\alpha_1 + \alpha_2 = 0$, $\beta_1 + \beta_2 + \beta_3 = 0$ et les variables $(\varepsilon_\kappa)_{\kappa\in\mathcal{Y}}$, sont des variables gaussiennes centrées indépendantes $\mathcal{N}(0,\sigma)$ avec σ constant. Les données indexées sur $\mathcal{X}$ sont reportées dans le tableau 5 où la valeur x_{23} est manquante dans $\mathcal{Y}$: relativement à $\mathcal{Y}$, les données $(x_{ij})_{(i,j)\in\mathcal{X}}$ sont dites *incomplètes*.

x_{11}	x_{12}	x_{13}
10	15	17
x_{21}	x_{22}	x_{23}
22	23	?

TABLE 5. Un jeu incomplet $x \in \mathbb{R}^X$ de données observées indexées par $\mathcal{X}$, avec $x_{2,3}$ manquante.

Les paramètres à déterminer sont $\Lambda = {}^{\mathsf{T}}(\mu, \alpha_1, \beta_1, \beta_2)$ comme maximum de la vraisemblance

$$\begin{aligned}\prod_{\kappa\in\mathcal{X}} \mathcal{N}(0,\sigma)(\varepsilon_\kappa) &= [\sqrt{2\pi}\sigma]^{-\#\mathcal{X}} \prod_{\kappa\in\mathcal{X}} \exp[-\varepsilon_\kappa^2/(2\sigma^2)] \\ &= [\sqrt{2\pi}\sigma]^{-\#\mathcal{X}} \prod_{(i,j)\in\mathcal{X}} \exp\left[-\frac{(x_{ij}-\mu-\alpha_i-\beta_j)^2}{2\sigma^2}\right]\end{aligned}$$

ou minimum des *moindres carrés*

$$\sum_{(i,j)\in\mathcal{X}} (x_{ij}-\mu-\alpha_i-\beta_j)^2 = \|x - A\Lambda\|_2^2$$

avec les conventions

$$x = {}^{\mathsf{T}}(x_{11}, x_{21}, x_{12}, x_{22}, x_{13}), \quad A = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix}.$$

Dans la suite, on identifiera implicitement les espaces de vecteurs colonnes $\mathbb{R}^{\mathcal{X}} \simeq \mathbb{R}^{\#\mathcal{X}}$ et $\mathbb{R}^{\mathcal{Y}} \simeq \mathbb{R}^{\#\mathcal{Y}}$ suivant l'ordre de la définition précédente de x.

7. C'est le modèle linéaire $x = \mu t_m + \alpha_1 t_a + \beta_1 t_{b_1} + \beta_2 t_{b_2} + \varepsilon$ avec les données explicatives (les *régresseurs*) $t_m, t_a, t_{b_1}, t_{b_2}$ et les variables expliquées/observées x listées dans le tableau

(i,j)	t_m	t_a	t_{b_1}	t_{b_2}	x	(i,j)	t_m	t_a	t_{b_1}	t_{b_2}	x
(1,1)	1	1	1	0	10	(2,1)	1	-1	1	0	22
(1,2)	1	1	0	1	15	(2,2)	1	-1	0	1	23
(1,3)	1	1	-1	-1	17						

On a noté ici x (resp. t_*) ce qui est habituellement noté y (resp. x_*) !

Le minimum $\Lambda_* \in \operatorname{argmin}_\Lambda \|x - A\Lambda\|$ est unique (la matrice A est injective), donné par $\Lambda_* = M_A x$ avec $M_A = ({}^\mathsf{T}AA)^{-1}{}^\mathsf{T}A$. Dans le cas présent, ni l'inverse $({}^\mathsf{T}AA)^{-1}$ (qui est bien défini vu que A est injective), ni la composée $({}^\mathsf{T}AA)^{-1}{}^\mathsf{T}A$

$$({}^\mathsf{T}AA)^{-1} = \frac{1}{12}\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & 1 & 1 \\ -1 & 1 & 5 & -1 \\ -1 & 1 & -1 & 5 \end{pmatrix}, \quad ({}^\mathsf{T}AA)^{-1}{}^\mathsf{T}A = \frac{1}{12}\begin{pmatrix} 1 & 3 & 1 & 3 & 4 \\ 3 & -3 & 3 & -3 & 0 \\ 5 & -3 & -1 & -3 & -4 \\ -1 & -3 & 5 & 3 & -4 \end{pmatrix}$$

ne peuvent être formulés simplement [8], en terme de statistiques comme dans la note 6 ou l'expression (10) du lemme 1.2 ci-dessous.

On va remplacer cette régression *complexe* (car sans formulation globale en fonction des statistiques) par un procédé itératif mettant en jeu à la k-ème itération une régression *simple* (le lemme 1.2 ci-dessous indique le caractère computationnel simple de $M_{A_c} y_k$) ,régression déterminant les paramètres Λ_k pour un modèle à données complètes $y_k = {}^\mathsf{T}({}^\mathsf{T}x, z_k)$ et précédée par une évaluation de la *donnée manquante* z_k qui revient en fait à un calcul d'espérance conditionnelle. Cette formulation, un peu cachée dans cet exemple, est fondamentale dans la méthode générale $\mathbb{EM}$ dont l'appellation *Espérance/Maximisation* est ainsi clairement motivée. Pour ce modèle linéaire, le théorème 1.8 établit la convergence de la suite Λ_k vers le paramètre de régression Λ_* du modèle à données manquantes : ce paramètre Λ_* est aussi celui déterminé par une régression du modèle à données complètes $y_\infty = (x, z_\infty)$ obtenu par la complétion des données incomplètes du tableau 5 par la donnée asymptotique $z_\infty = \lim_{k\to+\infty} z_k$.

Pour le modèle sans donnée manquante, *i. e.* avec des données complètes $y \in \mathbb{R}^{\mathcal{Y}}$, les paramètres de la régression $\Lambda_c = \operatorname{argmin}_\Lambda \|y - A_c\Lambda\|$, où

$$A_c = \texttt{rbind}(A, (1,-1,-1,-1)) = \begin{pmatrix} A \\ \widetilde{A} \end{pmatrix}, \quad \widetilde{A} = (1,-1,-1,-1),$$

sont déterminés pareillement par $\Lambda_c = ({}^\mathsf{T}A_cA_c)^{-1}{}^\mathsf{T}A_c y$, qui a une expression plutôt simple (et *classique*) en termes de moyennes statistiques (moments d'ordre 1) :

LEMME 1.2: *Soit* $M_{A_c} = ({}^\mathsf{T}A_cA_c)^{-1}{}^\mathsf{T}A_c$. *Alors*

$$M_{A_c} y = (\overline{y}, \overline{y}_{1\cdot} - \overline{y}, \overline{y}_{\cdot 1} - \overline{y}, \overline{y}_{\cdot 2} - \overline{y}), \quad y \in \mathbb{R}^{\#\mathcal{Y}}, \tag{10}$$

où on a noté $\overline{y} = \sum_{\kappa\in\mathcal{Y}} y_\kappa/6$, $\overline{y}_{i\cdot} = \sum_{j\in\{1,2,3\}} y_{ij}/3$ *et* $\overline{y}_{\cdot j} = \sum_{i\in\{1,2\}} y_{ij}/2$.

DÉMONSTRATION. Le point $M_{A_c} y$ est le point de minimum $\Lambda_c = (\mu, \alpha_1, \beta_1, \beta_2) = \operatorname{argmin}_\Lambda \|y - A_c\Lambda\|_2$: c'est le point critique de $\Lambda \mapsto \|y - A_c\Lambda\|_2^2$, ses quatre dérivées partielles par rapport à $\mu, \alpha_1, \beta_1, \beta_2$ sont donc nulles

$$\begin{aligned} 0 &= \sum_{(i,j)\in\mathcal{Y}} [y_{ij} - \mu - \alpha_i - \beta_j], \\ 0 &= \sum_{j\in\{1,2,3\}} ([y_{1j} - \mu - \alpha_1 - \beta_j] - [y_{2j} - \mu + \alpha_1 - \beta_j]), \\ 0 &= \sum_{i\in\{1,2\}} [y_{i1} - \mu - \alpha_i - \beta_1] - [y_{i3} - \mu - \alpha_i + \beta_1 + \beta_2] \\ 0 &= \sum_{i\in\{1,2\}} [y_{i2} - \mu - \alpha_i - \beta_2] - [y_{i3} - \mu - \alpha_i + \beta_1 + \beta_2] \end{aligned}$$

8. Les calculs ont été effectués avec R.

Le vecteur $\Lambda_c = {}^{\mathsf{T}}(\overline{y}, \overline{y}_{1\cdot} - \overline{y}, \overline{y}_{\cdot 1} - \overline{y}, \overline{y}_{\cdot 2} - \overline{y},)$ est solution de ces 4 équations. Par exemple, pour la dernière,

$$\begin{aligned}
-\frac{1}{2}\partial_{\beta_2}\|y - A_c\Lambda\|^2 &= \sum_{i\in\{1,2\}} [y_{i2} - \mu - \alpha_i - \beta_2] - [y_{i3} - \mu - \alpha_i + \beta_1 + \beta_2] \\
&= \sum_{i\in\{1,2\}} [y_{i2} - y_{i3}] - 2\beta_2 - 2(\beta_1 + \beta_2) \\
&= 2\overline{y}_{\cdot 2} - 2\overline{y}_{\cdot 3} - 2\beta_1 - 4\beta_2 \\
&= 2\overline{y}_{\cdot 2} - (6\overline{y} - 2\overline{y}_{\cdot 1} - 2\overline{y}_{\cdot 2}) - 2\beta_1 - 4\beta_2 \\
&= 2(\overline{y}_{\cdot 1} - \beta_1) + 4(\overline{y}_{\cdot 2} - \beta_2) - 6\overline{y} = 0
\end{aligned}$$

où on a utilisé l'identité $\alpha_1 + \alpha_2 = 0$ dans la première ligne, l'identité générale $3\overline{y} = \overline{y}_{\cdot 1} + \overline{y}_{\cdot 2} + \overline{y}_{\cdot 3}$ pour la pénultième et les identités $\beta_1 = \overline{y}_{\cdot 1} - \overline{y}, \beta_2 = \overline{y}_{\cdot 2} - \overline{y}$ dans la dernière.

Que ce minimum soit déterminé uniquement provient de l'injectivité de la matrice A_c, injectivité qui induit l'inversibilité de ${}^{\mathsf{T}}A_cA_c$ et par suite l'unicité du minimum Λ_c. □

Le calcul exact de Λ_* nécessite celui de l'inverse de la matrice ${}^{\mathsf{T}}AA$: on remplace ce calcul par l'approximation asymptotique $\Lambda_* = \lim_{k\to+\infty}\Lambda_k$ où l'itération du rang k au rang $k+1$ comporte deux étapes basées sur des données complétées $y_k = {}^{\mathsf{T}}({}^{\mathsf{T}}x, z_k)$:

(1) évaluation de $z_k = \mu_k + \alpha_{2,k} + \beta_{3,k} = \mu_k - \alpha_{1,k} - \beta_{1,k} - \beta_{2,k} = \widetilde{A}\Lambda_k$,

(2) détermination du minimum $\Lambda_{k+1} = \operatorname{argmin}_\Lambda \|y_k - A_c\Lambda\|_2$ avec $y_k = \begin{pmatrix} x \\ z_k \end{pmatrix}$.

Le test d'arrêt de l'itération $\Lambda_k \to \Lambda_{k+1}$ est la constatation de la stabilité des estimées de valeurs manquantes ou la nullité approximative de la somme des carrés des résidus $\|x - A\Lambda\|^2$.

Vu $z \sim \mathcal{N}(z_k, \sigma)$ pour la donnée manquante $z = x_{23}$ du tableau 5 et en notant $y = {}^{\mathsf{T}}({}^{\mathsf{T}}x, z)$ le vecteur complet de données, on obtient

$$\|y - A_c\Lambda\|^2 = \left\| \begin{pmatrix} x \\ z \end{pmatrix} - \begin{pmatrix} A \\ \widetilde{A} \end{pmatrix} \Lambda \right\|^2 = \|x - A\Lambda\|_2^2 + (z - \widetilde{A}\Lambda)^2.$$

L'affectation $z_k = \widetilde{A}\Lambda_k$ de la première étape établit l'équivalence de la minimisation de l'espérance conditionnelle

$$\begin{aligned}
E(\|y - A_c\Lambda\|^2 | x, \Lambda_k) &= \|x - A\Lambda\|^2 + E((z - \widetilde{A}\Lambda)^2 | x, \Lambda_k) \\
&= \|x - A\Lambda\|^2 + \int_{\mathbb{R}} (z - \widetilde{A}\Lambda)^2 \mathrm{e}^{-(z-z_k)^2/(2\sigma^2)} \frac{dz}{\sqrt{2\pi}\sigma} \\
&= \|x - A\Lambda\|^2 + (z_k - \widetilde{A}\Lambda)^2 + \sigma^2 \\
&= \left\| \begin{pmatrix} x \\ z_k \end{pmatrix} - \begin{pmatrix} A \\ \widetilde{A} \end{pmatrix} \Lambda \right\|^2 + \sigma^2 = \|y_k - A_c\Lambda\|^2 + \sigma^2
\end{aligned}$$

et de la régression $\operatorname{argmin}_\Lambda \|y_k - A_c\Lambda\|$ de la deuxième étape de l'itération.

LEMME 1.3: *Soit $x \in \mathbb{R}^{\#\mathcal{X}}$. Soit (z_k) une suite vérifiant la relation de récurrence*

$$z_k = \widetilde{A}\Lambda_k, \quad y_k = {}^{\mathsf{T}}({}^{\mathsf{T}}x, z_k), \quad \Lambda_{k+1} = \operatorname{argmin}_\Lambda \|y_k - A_c\Lambda_k\|.$$

Quelle que soit la donnée initiale z_0, la suite (z_k) est convergente vers $z_\infty = 10\overline{x} - 3(\overline{x}_{1\cdot} + \overline{x}_{\cdot 1} + \overline{x}_{\cdot 2})$.

DÉMONSTRATION. Vu $\overline{y_k} = (5\overline{x}+z_k)/6$ d'une part, $\overline{(y_k)}_{1\cdot} = \overline{x}_{1\cdot}$, $\overline{(y_k)}_{\cdot 1} = \overline{x}_{\cdot 1}$ et $\overline{(y_k)}_{\cdot 2} = \overline{x}_{\cdot 2}$ d'autre part, on obtient en reprenant les formules du lemme 1.2

$$\begin{aligned} z_{k+1} = \widetilde{A}\Lambda_{k+1} &= \mu_{k+1} - \alpha_{1,k+1} - \beta_{1,k+1} - \beta_{2,k+1} \\ &= \overline{y_k} - (\overline{y_{k1\cdot}} - \overline{y_k}) - (\overline{y_{k\cdot 1}} - \overline{y_k}) - (\overline{y_{k\cdot 2}} - \overline{y_k}) \\ &= 4\frac{5\overline{x}+z_k}{6} - \overline{x}_{1\cdot} - \overline{x}_{\cdot 1} - \overline{x}_{\cdot 2} = \frac{2z_k}{3} + L(x) \end{aligned}$$

avec $L(x) = 10\overline{x}/3 - \overline{x}_{1\cdot} - \overline{x}_{\cdot 1} - \overline{x}_{\cdot 2}$. Ainsi $z_{k+1} - 3L(x) = 2(z_k - 3L(x))/3$, ce qui donne la convergence géométrique de la suite z_k avec pour limite[9] $3L(x)$. □

Le théorème suivant est le résultat central de cette section : on y voit la répétition du couple *Espérance* (c'est l'égalité $z_k = \widetilde{A}\Lambda_k$) et la *maximisation* (c'est le calcul de l'optimum $\operatorname{argmin}_\Lambda \|{}^{\mathsf{T}}({}^{\mathsf{T}}x, z_k) - A_c\Lambda_k\|$)

THÉORÈME 1.8: *Soit $x \in \mathbb{R}^{\#\mathcal{X}}$, $A \in M_{\#\mathcal{X},\ell}$, $\widetilde{A} \in \mathbb{R}^\ell$, $A_c = {}^{\mathsf{T}}({}^{\mathsf{T}}A, \widetilde{A}) \in M_{\#\mathcal{X},\ell+1}$ et $z_\infty = z_\infty(x)$ la limite d'une suite z_k vérifiant la relation de récurrence*

$$z_k = \widetilde{A}\Lambda_k, \qquad \Lambda_{k+1} = \operatorname{argmin}_\Lambda \|{}^{\mathsf{T}}({}^{\mathsf{T}}x, z_k) - A_c\Lambda_k\|.$$

Les deux points de minimum

$$\Lambda_{c*} = \operatorname{argmin}_\Lambda \|{}^{\mathsf{T}}({}^{\mathsf{T}}x, z_\infty) - A_c\Lambda\|, \quad \Lambda_* = \operatorname{argmin}_\Lambda \|x - A\Lambda\|$$

sont égaux, coïncidant avec la limite de la suite $\Lambda_k = M_{A_c}\begin{pmatrix} x \\ z_k \end{pmatrix}$.

DÉMONSTRATION. Vu que $\Lambda_{c*} = M_{A_c}{}^{\mathsf{T}}({}^{\mathsf{T}}x, z_\infty)$ et $M_{A_c} = ({}^{\mathsf{T}}A_cA_c)^{-1}{}^{\mathsf{T}}A_c$, on a

$$ {}^{\mathsf{T}}A_c \begin{pmatrix} x \\ z_\infty \end{pmatrix} = {}^{\mathsf{T}}A_cA_c\Lambda_{c*} \tag{11}$$

Par ailleurs

$${}^{\mathsf{T}}A_c = ({}^{\mathsf{T}}A{}^{\mathsf{T}}\widetilde{A}), \quad {}^{\mathsf{T}}A_cA_c = ({}^{\mathsf{T}}A{}^{\mathsf{T}}\widetilde{A})\begin{pmatrix} A \\ \widetilde{A} \end{pmatrix} = {}^{\mathsf{T}}AA + {}^{\mathsf{T}}\widetilde{A}\widetilde{A}$$

et

$${}^{\mathsf{T}}\widetilde{A}\widetilde{A}\Lambda_{c*} = {}^{\mathsf{T}}\widetilde{A}(\mu_\infty - \alpha_{1\infty} - \beta_{1\infty} - \beta_{2\infty}) = {}^{\mathsf{T}}\widetilde{A}(\mu_\infty + \alpha_{2\infty} + \beta_{3\infty}) = {}^{\mathsf{T}}\widetilde{A}z_\infty$$

Ainsi l'équation (11) devient

$${}^{\mathsf{T}}Ax + {}^{\mathsf{T}}\widetilde{A}z_\infty = {}^{\mathsf{T}}AA\Lambda_{c*} + {}^{\mathsf{T}}\widetilde{A}\widetilde{A}\Lambda_{c*} = {}^{\mathsf{T}}AA\Lambda_{c*} + {}^{\mathsf{T}}\widetilde{A}z_\infty$$

soit ${}^{\mathsf{T}}Ax = {}^{\mathsf{T}}AA\Lambda_{c*}$ et donc $\Lambda_* = \Lambda_{c*}$ par unicité de Λ_*. □

△ REMARQUE 1.6: La suite $(z_k)_{k\geq 0}$ vérifie

$$z_{k+1} = \widetilde{A}({}^{\mathsf{T}}A_cA_c)^{-1}{}^{\mathsf{T}}A_c\begin{pmatrix} x \\ z_k \end{pmatrix} = Bz_k + c$$

avec $B = \widetilde{A}({}^{\mathsf{T}}AA + {}^{\mathsf{T}}\widetilde{A}\widetilde{A})^{-1}{}^{\mathsf{T}}\widetilde{A}$ et $c = \widetilde{A}({}^{\mathsf{T}}AA + {}^{\mathsf{T}}\widetilde{A}\widetilde{A})^{-1}{}^{\mathsf{T}}\widetilde{A}x$. L'opérateur B a même spectre non nul[10] que l'opérateur $\widetilde{B} = ({}^{\mathsf{T}}AA + {}^{\mathsf{T}}\widetilde{A}\widetilde{A})^{-1}{}^{\mathsf{T}}\widetilde{A}\widetilde{A}$. Le lemme suivant assure que le rayon spectral de $\widetilde{B}$ est strictement inférieur à 1 : si $z_\infty = (1-B)^{-1}c$, l'égalité $z_{k+1} - z_\infty = (1-B)^{-1}(z_k - z_\infty)$ implique la convergence géométrique de la suite (z_k) vers z_∞.

9. Avec les données du tableau 5, $z_\infty = 27$.
10. Si θ est non nul, les opérateurs A et B induisent des isomorphismes $\ker(BA - \theta) \simeq \ker(AB - \theta)$.

	1	2	3	4	5	6
I	F 3.5	B 4.2	A 6.7	D 6.6	C 4.1	E 3.8
II	B 8.9	F 1.9	D ??	A 4.5	E 2.4	C 5.8
III	C 9.6	E 3.7	F −2.7	B 3.7	D 6.0	A 7.0
IV	D 10.5	C 10.2	B 4.6	E 3.7	A 5.1	F 3.8
V	E ??	A 7.2	C 4.0	F −3.3	B 3.5	D 5.0
VI	A 5.9	D 7.6	E −0.7	C 3.0	F 4.0	B 8.6

TABLE 6. Valeurs parcellaires pour la longueur des pousses de brin de blé, avec indication du traitement. Les valeurs pour les parcelles $(II, 3)$ et $(V, 1)$ sont manquantes.

Ainsi, la discussion précédente peut être reprise dans le cas général où les données y sont complétées par des variables z (de dimension quelconque) et les variables explicatives synthétisées A par $\widetilde{A}$. Ce n'est intéressant que si la complétion $A_c = \begin{pmatrix} A \\ \widetilde{A} \end{pmatrix}$ donne un calcul de M_{A_c} aisé, comme c'est le cas de notre exemple (cf. lemme 1.2). ▽

LEMME 1.4: *Soit M, P opérateurs auto-adjoints positifs avec P inversible. Alors $(P + M)^{-1}M$ a un rayon spectral strictement inférieur à 1.*

DÉMONSTRATION. Si N est auto-adjoint positif, on note par $\sqrt{N}$ l'unique opérateur auto-adjoint positif de carré N. Les valeurs propres non nulles de $(P + M)^{-1}M$ et $\sqrt{M}(P + M)^{-1}\sqrt{M}$ coïncident : le spectre de $(P + M)^{-1}M$ est donc positif. Soit (θ, u) un de ses éléments propres. Vu

$$\theta u = (P + M)^{-1}Mu = u - (P + M)^{-1}Pu, \tag{12}$$

$1 - \theta$ est aussi valeur propre de $(P + M)^{-1}P$ dont le spectre est positif, comme il a été montré pour $(P + M)^{-1}M$. On obtient donc $\theta \in [0, 1]$: si $\theta = 1$, l'égalité (12) donne $(P + M)^{-1}Pu = 0$ soit $u = 0$ puisque P a été supposé inversible. Ainsi $\theta \in [0, 1)$ et $\rho_\infty((P + M)^{-1}M) < 1$. □

△ REMARQUE 1.7: Cette introduction d'une suite pour calculer la solution d'une équation du type $Mx = b$ peut paraître inutilement compliquée. En fait, ce procédé est monnaie courante dans les méthodes de Jacobi, Gauß-Seidel : on écrit $M = M_1 - M_2$ avec M_1 inversible, d'inverse M_1^{-1} aisément calculable (par exemple M_1 est diagonale ou triangulaire supérieure) et $M_1^{-1}M_2$ de rayon spectral strictement inférieur à 1 : $x = \lim_{k\to\infty} \sum_{\ell=0}^{k} (M_1^{-1}M_2)^\ell M_1^{-1}b$. D'autre part, il est parfois constaté qu'*il est plus rapide d'aller d'un point A à un point B en passant par le plan complexe*, où les imaginaires restent parfois *cachés*. ▽

△ REMARQUE 1.8: L'exemple 2.3.3 de [**?**] est analogue : il porte sur l'observation des différences de longueur de pousses de brins de blé soumis à divers traitements. Ces pousses sont plantées dans un carré de $36 = 6 \times 6$ parcelles indexées par $(i, j) \in \{1, 2, \ldots, 6\}^2$, chacune recevant un traitement caractérisé par A, B, C, D, E ou F comme indiqué dans le tableau 6 ; de plus chaque ligne L_i contient les 6 traitements dans un ordre aléatoire.

Supposons un instant un tableau carré $\mathbb{T} = (y_{ij})_{i,j=1}^{N}$ d'ordre N analogue au tableau de longueurs 6, sans donnée manquante et avec N traitements différents, chaque traitement étant appliqué sur N cellules exactement.

LEMME 1.5: *Soit* $\mathbb{Y} = (y_{ij}) \in \mathbb{R}^{N^2}$ *et les variables explicatives* $\overline{L_i}, \overline{C_j}, \overline{T_k}$ *définies pour chaque ligne, colonne et traitement resp. comme les moyennes des longueurs sur les* N *cellules correspondantes. Notons* $\mathbb{X} = (1, \overline{L_i}, \overline{C_j}, \overline{T_{k(i,j)}})$ *d'ordre* $(N^2, 4)$ *et* $\Lambda = (\theta_0, \theta_L, \theta_C, \theta_T)$ *d'ordre* $(1, 4)$. *Si* $\mathbb{X}$ *est injective, le vecteur* $\Lambda_* = ({}^T\mathbb{X}\mathbb{X})^{-1}\,{}^T\mathbb{X}\mathbb{Y}$, *solution unique de la régression* $\min_\Lambda \|\mathbb{Y} - \mathbb{X}\Lambda\|_2$, *est donné par*

$$\Lambda_* = (-2\overline{y}, 1, 1, 1)$$

où $\overline{y}$ *est la moyenne des données observées* $\overline{y} = \sum_{i,j=1}^{N} y_{ij}/N^2$.

Ainsi la valeur y_{ij} *du modèle linéaire* $y_{ij} \sim \theta_0 + \theta_L\overline{L}_i + \theta_C\overline{C}_i + \theta_T\overline{T}_{k(i,j)}$ *obtenue par régression pour la cellule* (i, j) *est*

$$y_{ij} = -2\overline{y} + \overline{L_i} + \overline{C_j} + \overline{T_{k(i,j)}}, \quad 1 \le i, j \le N. \tag{13}$$

DÉMONSTRATION. Il s'agit de vérifier les annulations

$$0 = \sum_{ij}[y_{ij} - \theta_0 - \theta_L\overline{L}_i - \theta_c\overline{C}_i - \theta_T\overline{T}_{k(i,j)}], \quad 0 = \sum_{ij}[y_{ij} - \theta_0 - \theta_L\overline{L}_i - \theta_c\overline{C}_i - \theta_T\overline{T}_{k(i,j)}]\overline{L}_i,$$

$$0 = \sum_{ij}[y_{ij} - \theta_0 - \theta_L\overline{L}_i - \theta_c\overline{C}_i - \theta_T\overline{T}_{k(i,j)}]\overline{C}_j, \quad 0 = \sum_{ij}[y_{ij} - \theta_0 - \theta_L\overline{L}_i - \theta_c\overline{C}_i - \theta_T\overline{T}_{k(i,j)}]\overline{T}_{k(i,j)},$$

correspondant à l'annulation des dérivées par rapport à $\theta_0, \theta_L, \theta_C, \theta_T$. Par exemple, la troisième annulation provient des identités

$$\sum_{i,j} y_{ij}\overline{L_i} = N\sum_i \overline{L_i}^2 = \sum_{i,j}\overline{L_i}^2$$

et

$$-\sum_{i,j}\theta_0\overline{L_i} = -\theta_0 n^2\overline{y} = 2\overline{y}^2N^2 = \sum_{i,j}\overline{C_j}\,\overline{L_i} + \sum_{i,j}\overline{T_{k(i,j)}}\,\overline{L_i}. \qquad \square$$

Dans notre cas du tableau 6 avec $N = 6$ et deux données manquantes pour les parcelles $(2, 3)$ et $(5, 1)$, la régression linéaire n'a pas de solution aussi simple. L'algorithme $\mathbb{EM}$ approche la solution Λ_* : après l'initialisation où des valeurs arbitraires sont affectées aux données manquantes $y_{2,3}$ et $y_{5,1}$ (par exemple en prenant des moyennes sur les lignes ou colonnes correspondantes aux parcelles déficientes), il consiste en l'itération des deux étapes successives comme précédemment

(1) appliquer la régression avec données complètes,
(2) calculer les valeurs estimées suivant (13) et affecter ces valeurs estimées aux parcelles à données manquantes. ▽

3.2. L'algorithme $\mathbb{EM}$. Soit $\mathbf{x} = (x_1, \ldots, x_r)$ un échantillon de loi $P_X = P_{\theta(X)} \in \{P_\theta, \theta \in \Lambda \subset \mathbb{R}^d\}$. On cherche à estimer le paramètre $\theta = \theta(P_X)$ à partir d'une réalisation de $\mathbf{x} = (x_1, \ldots x_r)$. De manière (assez) générale, l'estimateur $\widehat{\theta}_r(\mathbf{x}) \in \operatorname{argmax}_\theta p_\theta(\mathbf{x})$ du maximum de vraisemblance converge vers le paramètre $\theta(X)$ de la distribution P_X quand le cardinal r de l'échantillon $\mathbf{x}$ tend vers l'infini (cf. appendice A). La recherche de la valeur $\widehat{\theta}_r(\mathbf{x})$ maximisant la vraisemblance associée à l'échantillon $\mathbf{x}$ s'en trouve fortement motivée.

Soit la loi P_θ est discrète avec $p_\theta(e) = P_\theta(X = e), e \in E$, soit la loi est continue de densité $dp_\theta(x)$ *i. e.* $P_\theta(X \in A) = \int_A dp_\theta(x)dx$ avec $A \subset \mathbb{R}^d$. Dans le premier cas, l'interprétation du maximum de vraisemblance [MV] est heuristiquement bien fondée : plus la probabilité $p_\theta(e)$ est grande, plus la vraisemblance est avérée.

Étant donné l'échantillon $\curvearrowleft = (x_1, \dots, x_r)$ de mesures indépendantes, la vraisemblance est, suivant que la distribution est discrète ou continue,

$$p_\theta(\mathbf{x}) = \prod_{i=1}^{r} p_\theta(x_i) = \prod_{i=1}^{r} \begin{cases} p_\theta(X = x_i), & \text{si } P_X \text{ est discrète,} \\ p_\theta(x_i), & \text{si } P_X \text{ est continue de densité } dp_\theta(x). \end{cases}$$

Si la variable x prend un nombre dénombrable de valeurs $(v_k)_{k\in\mathbb{N}}$ et l'échantillon $\mathbf{x} = (x_1, \dots, x_r)$ atteint N_k fois la valeur v_k pour $k = 0, \dots$ (comme dans les exemples 3.5.2 et 3.5.1 ci-dessous), on a

$$p_\theta(\mathbf{x}) = \prod_{k\geq 0} p_\theta(v_k)^{N_k}$$

où les N_k sont presque tous nuls avec $\sum_{k\geq 0} N_k = r$.

L'estimateur $\widehat{\theta}_r(\mathbf{x})$ du paramètre θ induit par l'échantillon $\curvearrowleft$ est un élément de l'ensemble de maxima

$$\operatorname{argmax}_\theta p_\theta(\mathbf{x}) = \operatorname{argmax}_\theta \ell(\theta, \mathbf{x})$$

où $\ell(\theta, \mathbf{x}) = \log p_\theta(\mathbf{x}) = \sum_{i=1}^{r} \log p_\theta(x_i)$ note la log-vraisemblance de l'échantillon $\mathbf{x}$ relativement à la distribution P_θ.

Souvent, il n'est pas possible de résoudre explicitement l'équation $\nabla_\theta \ell(\theta, \mathbf{x}) = 0$ de criticité du maximum $\widehat{\theta}_{\mathbf{x}}$ pour la vraisemblance $p_\theta(\mathbf{x})$. L'algorithme d'*Espérance/maximisation*, dit $\mathbb{EM}$ [11], vise à pallier cette incapacité par une construction itérative. dans l'espace des paramètres Λ d'un point de maximum pour la vraisemblance $p_\theta(\mathbf{x})$. Il s'introduit naturellement dans diverses situations d'information manquante :

- — données réellement manquantes,
- — paramètres inconnus dans $p_\theta(x)$ comme les paramètres α_j (vérifiant $\alpha_J \geq 0, \sum_{j=1}^{J} = 1$) de mélange pour la distribution $p_\theta(x) = \sum_{j=1}^{J} \alpha_j p_{j\theta}(x)$,
- — calcul de MV se simplifiant considérablement après introduction de données supplémentaires (inobservées éventuellement).

Dans la suite, on note $\mathbf{x}$ les données observées avec fonction de vraisemblance $p_\theta(\mathbf{x})$, $\mathbf{z}$ les données manquantes et $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ les données complètes avec fonction de vraisemblance $p_\theta(\mathbf{y})$. On considère l'espérance conditionnelle le long des données non observées de la log-vraisemblance pour le paramètre θ de la totalité des données

$$Q_{\widetilde{\theta}}(\theta) = \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}(\log p_\theta(\mathbf{x}, \mathbf{z})) = \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}(\ell(\theta, \mathbf{x}, \mathbf{z})) = \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}(\ell(\theta, \mathbf{y})),$$

espérance prise conditionnellement aux données observées $\mathbf{x}$ suivant la mesure de paramètre $\widetilde{\theta}$. La mise en œuvre de la méthode $\mathbb{EM}$ présuppose que la détermination du maximum de la vraisemblance $p_\theta(\mathbf{y})$ des données complètes $\mathbf{y}$ est des plus aisée, débouchant sur des déterminations effectives de points critiques, au contraire de la vraisemblance $p_\theta(\mathbf{x})$ obtenue à partir des seules données observées $\mathbf{x}$.

Étant donnée une valeur du paramètre initiale θ_0, on procède de manière itérative avec au rang k deux étapes

($\mathbb{E}$) le calcul de l'espérance $Q_{\theta_k}(\theta)$,

($\mathbb{M}$) la maximisation $\theta_{k+1} \in \operatorname{argmax}_\theta Q_{\theta_k}(\theta)$.

Le test d'arrêt consiste en général à l'évaluation de la différence $|\ell(\theta_{k+1}, \mathbf{x}) - \ell(\theta_k, \mathbf{x})|$. En fait, le seul résultat établi rigoureusement affirme que la suite $(\theta_k)_{k\leq 0}$ induit une suite de vraisemblances $(p_{\theta_k}(\mathbf{x}))_{k\leq 0}$ croissante (cf. 3.4). Il a été constaté de très nombreux exemples où l'algorithme $\mathbb{EM}$ converge vers le maximum (global sur Λ) de vraisemblance

11. La méthode $\mathbb{EM}$ se trouve dans d'autres contextes statistiques : par ex. en statistique bayésienne, elle est utilisée pour calculer le mode de la distribution *a posteriori*.

$\operatorname{argmax}_\theta p_\theta(\mathbf{x})$, ce qui a contribué à la popularisation de la méthode$\mathbb{EM}$, même si des exemples adhoc exhibent des convergences hors des vraisemblances (globale) maximales : points selle, maxima locaux.

3.3. Deux exemples éclairants. Commençons par un exemple artificiel ayant la vertu de montrer clairement le fonctionnement de l'algorithme $\mathbb{EM}$. On ajoute des variables manquantes $\mathbf{z}$ de lois exponentielles à un échantillon iid $\mathbf{x}$ suivant une loi exponentielle, obtenant ainsi des données $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ iid de loi exponentielle e_θ de moyenne θ, constituées de données observées $\mathbf{x} = (x_1, \dots, x_r)$, complétées par des données manquantes $\mathbf{z} = (z_1 \dots, z_s)$. La vraisemblance est

$$p_\theta(\mathbf{y}) = \prod_{i=1}^{r+s} \left[\theta^{-1}\mathrm{e}^{-y_i\theta^{-1}}\right] = \theta^{-s-r}\mathrm{e}^{-\theta^{-1}\sum_{i=1}^{r+s} y_i}$$

où $y_j = x_j$ si $j = 1, \dots, r$ et $y_{r+k} = z_k$ pour $k = 1, \dots, s$. L'espérance de sa log-vraisemblance est ainsi

$$\begin{aligned}
\mathbb{E}^{\mathbf{z}|\mathbf{x}}_{\widetilde{\theta}}(\ell(\theta, \mathbf{x}, \mathbf{z})) &= -(r+s)\log\theta - \theta^{-1}\mathbb{E}^{\mathbf{z}|\mathbf{x}}_{\widetilde{\theta}}\left(\sum_{j=1}^{r} x_j + \sum_{k=1}^{s} z_k\right) \\
&= -(r+s)\log\theta - \theta^{-1}\left(\sum_{j=1}^{r} x_j + \sum_{k=1}^{s} \mathbb{E}^{\mathbf{z}|\mathbf{x}}_{\widetilde{\theta}}(z_k)\right) \\
&= -(r+s)\log\theta - \theta^{-1}\left(r\overline{\mathbf{x}} + s\widetilde{\theta}\right),
\end{aligned}$$

où on a noté la moyenne empirique $\overline{\mathbf{x}} = \sum_{j=1}^{r} x_j/r$ et utilisé que la loi $e_{\widetilde{\theta}}$ a pour moyenne $\widetilde{\theta}$. Cette espérance est maximale en $\theta_{\max}$ caractérisé par

$$0 = \nabla_\theta \mathbb{E}^{\mathbf{z}|\mathbf{x}}_{\widetilde{\theta}}(\ell(\theta, \mathbf{x}, \mathbf{z})) = -\frac{r+s}{\theta} + \theta^{-2}(r\overline{\mathbf{x}} + s\widetilde{\theta})$$

soit

$$\theta_{\max} = \frac{r\overline{\mathbf{x}} + s\widetilde{\theta}}{r+s}.$$

Ainsi l'itération de $\mathbb{EM}$ est donnée par $\theta_{k+1} = (r\overline{\mathbf{x}} + s\theta_k)/(r+s)$, soit

$$\theta_{k+1} - \overline{\mathbf{x}} = \frac{s}{r+s}(\theta_k - \overline{\mathbf{x}}) = \left(\frac{s}{r+s}\right)^{k+1}(\theta_0 - \overline{\mathbf{x}}) \to 0 \text{ si } k \to \infty.$$

S'il y a convergence de θ_k vers le MV, le paramètre de maximum de la vraisemblance $p_\theta(\mathbf{x})$ est donc la valeur $\widehat{\theta}_r(\mathbf{x}) = \overline{\mathbf{x}}$ et $\overline{\mathbf{x}} \to \mathbb{E}^{e(\theta)}(t) = \theta$ lorsque la taille r de l'échantillon $\mathbf{x}$ tend vers l'infini d'après la loi des grands nombres. On aura remarqué que l'estimateur du maximum de vraisemblance de la vraisemblance $p_\theta(\mathbf{x}) = \theta^{-r}\mathrm{e}^{-\theta^{-1}\sum_{j=1}^{r} x_j}$ (sans variable manquante) est au point $\theta = \overline{\mathbf{x}}$.

Reprenons l'exemple précédent en considérant des variables, tant observées que manquantes, supposées iid et de loi normale $\mathcal{N}(\mu, v)$. Les estimateurs pour la moyenne et la variance s'obtiennent aisément par le MV associé à un échantillon. On suppose donc les données $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ indépendantes et identiquement distribuées suivant la loi normale $\mathcal{N}(\theta)$ avec $\theta = (\mu, v)$, constituées de données $\mathbf{x} = (x_1, \dots, x_r)$ observées et complétées par des données manquantes/latentes $\mathbf{z} = (z_1 \dots, z_s)$. Vu l'expression de la vraisemblance

$$p_\theta(\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}} \exp\left((y_i - \mu)^2/(2v)\right),$$

où $n = r+s$, le calcul de l'espérance conditionnelle $\mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}(\ell(\theta, \mathbf{x}, \mathbf{z}))$ de la log-vraisemblance

$$\begin{aligned}\log p_\theta(\mathbf{y}) &= -\frac{n}{2}\log v - \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2/v \\ &= -\frac{n}{2}\log v - \sum_i y_i^2/(2v) + \mu\sum_i y_i/v - n\mu^2/(2v)\end{aligned}$$

nécessite seulement celui des deux premiers moments des données $\mathbf{y}$

$$\begin{aligned}\mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}\left(\sum_{i=1}^{n} y_i \middle| \mathbf{x}\right) &= \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}\left(\sum_{j=1}^{r} x_j + \sum_{k=1}^{s} z_k \middle| \mathbf{x}\right) \\ &= \sum_{j=1}^{r} x_j + \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}\left(\sum_{k=1}^{s} z_k \middle| \mathbf{x}\right) = r\overline{\mathbf{x}} + s\widetilde{\mu} \\ \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}\left(\sum_{i=1}^{r} y_i^2 \middle| \mathbf{x}\right) &= \sum_{j=1}^{r} x_j^2 + \sum_{k=1}^{s}[\mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}((z_k-\widetilde{\mu})^2) + \widetilde{\mu}^2] = r\overline{\mathbf{x}^2} + s(\widetilde{\mu}^2+\widetilde{v}).\end{aligned}$$

Ainsi

$$\mathbb{E}_{\widetilde{\theta}}(\log p_\theta(\mathbf{y}|\mathbf{x})) = -\frac{n}{2}\log v - (r\overline{\mathbf{x}^2} + s(\widetilde{\mu}^2+\widetilde{v}))/(2v) + \mu(r\overline{\mathbf{x}} + s\widetilde{\mu})/v - n\mu^2/(2v)$$

Le point maximum $\theta_{\max} = (\mu_{\max}, v_{\max})$ est obtenu par l'annulation [12] du gradient $\nabla_\theta \mathbb{E}_{\widetilde{\theta}}(\log p_\theta(\mathbf{y}|\mathbf{x}))$, soit

$$\begin{cases}\nabla_\mu \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}(\log p_\theta(\mathbf{y}|\mathbf{x})) &= (r\overline{\mathbf{x}} + s\widetilde{\mu})/v - 2n\mu/v \\ \nabla_v \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}(\log p_\theta(\mathbf{y}|\mathbf{x})) &= -\frac{n}{2v} + (r\overline{\mathbf{x}^2} + s(\widetilde{\mu}^2+\widetilde{v}))/(2v^2) - \mu(r\overline{\mathbf{x}} + s\widetilde{\mu})/v^2 + n\mu^2/(2v^2)\end{cases}$$

et par suite

$$\begin{cases}\mu_{\max} &= (r/n)\overline{\mathbf{x}} + (s/n)\widetilde{\mu} \\ v_{\max} &= (r/n)\overline{\mathbf{x}^2} + (s/n)(\widetilde{\mu}^2+\widetilde{v}) - [(r/n)\overline{\mathbf{x}} + (s/n)\widetilde{\mu}]^2\end{cases}$$

La récurrence associée au processus itératif $\mathbb{EM}$ prend la forme

$$\begin{cases}\mu_{k+1} &= (r/n)\overline{\mathbf{x}} + (s/n)\mu_k \\ v_{k+1} &= (r/n)\overline{\mathbf{x}^2} + (s/n)(\mu_k^2+v_k) - [(r/n)\overline{\mathbf{x}} + (s/n)\mu_k]^2\end{cases}$$

et peut s'exprimer suivant

$$\begin{cases}\mu_{k+1}-\mu_\infty &= (s/n)(\mu_k-\mu_\infty) \\ v_{k+1}-v_\infty &= (s/n)(v_k-v_\infty) + (s/n)(\mu_k-\mu_\infty)(2(r/n)\overline{\mathbf{x}} + \mu_k + \mu_\infty)\end{cases}$$

où $(\mu_\infty, v_\infty) = (\overline{\mathbf{x}}, \overline{\mathbf{x}^2} - \overline{\mathbf{x}}^2)$ désigne le point fixe du système dynamique précédent, dont on démontre la convergence géométrique [13] quelle que soit la donnée initiale θ_0.

L'algorithme $\mathbb{EM}$ converge vers les expressions des paramètres $\theta = (\mu, v)$ déterminées par le MV pour les variables gaussiennes indépendantes identiquement distribuées suivant $\mathcal{N}(\mu, v)$, dont le calcul explicite est bien connu.

12. La maximisation de la log-vraisemblance d'une loi normale pour un échantillon $\mathbf{Y}$ donne une estimation de la moyenne M et de la variance V comme combinaison linéaire en les $\overline{\mathbf{Y}}$ et $\overline{\mathbf{Y}^2}$: $M = \overline{\mathbf{Y}}, V = \overline{\mathbf{Y}^2} - M^2$. Il suffit de remplacer $\overline{\mathbf{Y}}$ et $\overline{\mathbf{Y}^2}$ par les expressions $\mathbb{E}(n\overline{\mathbf{x}}|\mathbf{y}, \theta))$ et $\mathbb{E}(n\overline{\mathbf{x}^2}|\mathbf{y}, \theta)$.

13. Soit $(u_k), (\varepsilon_k)$ des suites telles que $u_{k+1} = \alpha(u_k + \varepsilon_k)$ avec $0 < |\alpha| < 1$ et $|\varepsilon_k| \leq C|\alpha|^k$. Alors on vérifie par récurrence que $|u_k| \leq |\alpha|^{k-1}|u_1| + (k-1)C\alpha^k$ pour $k \geq 1$, d'où la convergence de u_k vers 0.

3.4. Convergences de $\mathbb{EM}$. Soit $\mathbf{x}$ les données observées et $\mathbf{z}$ les données manquantes (naturellement ou après introduction spécifique). On suppose que les lois $P_\theta, \theta \in \Lambda$ sont à densité $p_\theta(y)$ par rapport à une mesure indépendante de θ : $dP_\theta = p_\theta(y)dy$ par exemple pour la loi modélisant les données complètes $\mathbf{y}$. Ainsi

$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}) p_\theta(\mathbf{z}|\mathbf{x})$$

soit, pour les log-vraisemblances correspondantes (par ex. $\ell(\theta, \mathbf{x}) = \log p_\theta(\mathbf{x}), \ldots$),

$$\ell(\theta, \mathbf{x}) = \ell(\theta, \mathbf{x}, \mathbf{z}) - \log p_\theta(\mathbf{z}|\mathbf{x})$$

où on cherche le maximum de $\ell(\theta, \mathbf{x})$ pour un θ^*, qui maximisera tout autant la vraisemblance $p_\theta(\mathbf{x})$. Les données $\mathbf{z}$ étant manquantes (défaut d'observation ou autre raison), on remplace le membre de droite de l'égalité précédente par son espérance en $\mathbf{z}$ conditionnellement aux données $\mathbf{x}$ observées et pour un paramètre $\widetilde{\theta} \in \Lambda$. Ainsi,

$$\ell(\theta, \mathbf{x}) = \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}[\ell(\theta, \mathbf{x}, \mathbf{z})] - \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}[\log p_\theta(\mathbf{z}|\mathbf{x})] \tag{14}$$

$$= \int_Z \ell(\theta, \mathbf{x}, \mathbf{z}) p_{\widetilde{\theta}}(\mathbf{z}|\mathbf{x}) d\mathbf{z} - \int_Z \log p_\theta(\mathbf{z}|\mathbf{x}) p_{\widetilde{\theta}}(\mathbf{z}|\mathbf{x}) d\mathbf{z} \tag{15}$$

Cette égalité va permettre de minorer $\ell(\theta, \mathbf{x})$ par une fonction qu'on peut maximiser : ce majorant donnera une estimation inférieure du maximum de la log-vraisemblance $\ell(\theta, \mathbf{x})$. La méthode $\mathbb{EM}$ apparaît comme un cas particulier des méthodes de type MM, pour *majoration d'un minorant* : un maximum de J est minoré par le maximum d'une fonction minorante $J \geq J_-$, soit $\max J \geq \max J_-$.

Heuristiquement, on va chercher dans le membre de droite de (14) à majorer le premier terme $Q_{\widetilde{\theta}}(\theta)$ (on compte sur l'optimisation aisée de la vraisemblance pour des modèles à données complètes), alors que le second terme $H_{\widetilde{\theta}}(\theta)$ sera contrôlé convenablement grâce à l'inégalité de Jensen appliquée à la fonction concave log.

Le lemme suivant est corollaire de l'inégalité de Jensen pour une fonction convexe :

LEMME 1.6: *Pour tous $\theta, \widetilde{\theta} \in \Lambda$,*

$$H_{\widetilde{\theta}}(\theta) = -\mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}[\log p_\theta(\mathbf{z}|\mathbf{x}]) \geq -\mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}[\log p_{\widetilde{\theta}}(\mathbf{z}|\mathbf{x})] = H_{\widetilde{\theta}}(\widetilde{\theta}).$$

DÉMONSTRATION. Si φ est concave, l'inégalité de Jensen pour la variable aléatoire X prend la forme

$$\mathbb{E}(\varphi(X)) = \int_\Omega \varphi(X(\omega)) dP(\omega) \leq \varphi\left(\int_\Omega X(\omega) dP(\omega)\right) = \varphi(\mathbb{E}(X))$$

où on a supposé X et $\varphi(X)$ intégrables. Ainsi pour la fonction concave log, on obtient

$$\mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}\left[\log\left(\frac{p_\theta(\mathbf{z}|\mathbf{x})}{p_{\widetilde{\theta}}(\mathbf{z}|\mathbf{x})}\right)\right] \leq \log \mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}\left[\frac{p_\theta(\mathbf{z}|\mathbf{x})}{p_{\widetilde{\theta}}(\mathbf{z}|\mathbf{x})}\right] = \log\left[\int_Z p_\theta(\mathbf{z}|\mathbf{x}) d\mathbf{z}\right] = \log 1 = 0. \qquad \square$$

Ainsi le second terme $H_{\widetilde{\theta}}(\theta) = -\mathbb{E}_{\widetilde{\theta}}^{\mathbf{z}|\mathbf{x}}[\log p_\theta(\mathbf{z}|\mathbf{x}])$ du membre de droite de (14) est minoré par $H_{\widetilde{\theta}}(\widetilde{\theta})$ pour tout θ : étant donné $\widetilde{\theta}$, tout $\theta_{\max}$ maximisant le premier terme $Q_{\widetilde{\theta}}(\theta)$ du membre de droite de (14) induit un accroissement $\ell(\theta_{\max}, \mathbf{x}) \geq \ell(\widetilde{\theta}, \mathbf{x})$ pour la log-vraisemblance déterminée par l'échantillon $\mathbf{x}$

$$\ell(\theta_{\max}, \mathbf{x}) = Q_{\widetilde{\theta}}(\theta_{\max}) + H_{\widetilde{\theta}}(\theta_{\max}) \geq Q_{\widetilde{\theta}}(\widetilde{\theta}) + H_{\widetilde{\theta}}(\widetilde{\theta}) = \ell(\theta_k, \mathbf{x}).$$

L'étape $(\mathbb{M})$ de l'itération d'ordre k de l'algorithme $\mathbb{EM}$ consiste à trouver un θ_{k+1} tel que

$$\theta_{k+1} = \text{argmax}_\theta \left[\int_Z \ell(\theta, \mathbf{x}, \mathbf{z}) p_{\theta_k}(\mathbf{z}|\mathbf{x}) d\mathbf{z} \right] = \text{argmax}_\theta \left[\mathbb{E}_{\theta_k}^{\mathbf{z}|\mathbf{x}}(\ell(\theta, \mathbf{x}, \mathbf{z})) \right].$$

On a prouvé ainsi le théorème

THÉORÈME 1.9: *Soit (θ_k) la suite de paramètres obtenue à partir de l'échantillon observé $\mathbf{x}$ par itération $\mathbb{EM}$ pour la loi $p_\theta(x)$ et des données manquantes $\mathbf{z}$. Alors les suites $(p_{\theta_k}(\mathbf{x}))_{k\geq 0}$ et $(\ell(\theta_k, \mathbf{x}))_{k\geq 0}$ sont croissantes, stationnaires si et seulement si $Q_{\theta_k}(\theta_{k+1}) = Q_{\theta_k}(\theta_k)$.*

Même si la suite de vraisemblance $(p_{\theta_k}(\mathbf{x}))_{k\geq 0}$ est croissante convergente, il n'est pas sûr que la suite (θ_k) converge vers un maximum absolu de la vraisemblance, ni même qu'elle converge. C'est le cas dans les exemples considérés ici : on peut démontrer des résultats de convergence avec des hypothèses de convexité et de continuité. À l'inverse, il y a des exemples où la suite (θ_k) ne converge pas ou, si elle converge, sa limite est un point selle ou un maximum local de la log-vraisemblance $\ell(\theta, \mathbf{x})$.

Outre ces défauts de convergence, la méthode $\mathbb{EM}$ présente des lacunes théoriques, avec une estimation quasi-absente de la vitesse de convergence (cruciale lorsque des millions d'observation multi-factorielles sont examinées) ou des difficultés à réaliser le calcul d'une des deux étapes : espérance (obligeant le recours à des méthodes de Monte-Carlo), maximum de la log-vraisemblance (par ex. dans l'étude de mélanges). Malgré ces déficiences, il apparaît une efficacité remarquable dans de nombreuses situations pour la méthode $\mathbb{EM}$, ce qui explique son ubiquité dans les applications.

3.5. Quelques exemples supplémentaires.

3.5.1. *Veuves et enfants à charge.* Afin d'analyser le nombre d'enfants à charge de veuves, on considère un échantillon $\mathbf{x}$ de $r = 4\,075$ veuves, dont N_n ont n enfants à charge avec $N = (N_0 = 3062, N_1 = 587, 284, 103, 33, 4, 2)$. Que beaucoup de veuves aient peu d'enfants à charge suggère qu'une simple loi de Poisson ne modélise pas finement la situation. On introduit un mélange d'une population A de veuves sans enfant et d'une population B de veuves avec une distribution de Poisson $\mathcal{P}(\mu)$ relativement au nombre d'enfants, $\mathcal{P}(\mu)$ étant de moyenne μ : une veuve appartient à la population A avec probabilité p. Le paramètre cherché est donc $\theta = (p, \mu)$ déterminant le mélange $p\delta_0 + (1-p)\mathcal{P}(\mu)$, mélange donnant pour les données observées comme vraisemblance

$$g_\theta(\mathbf{x}) = [p + (1-p)\mathrm{e}^{-\mu}]^{N_0} \prod_{n\geq 1} \left[\frac{(1-p)\mu^n \mathrm{e}^{-\mu}}{n!} \right]^{N_n}$$

et log-vraisemblance, à des constantes additives près,

$$\ell(\theta, \mathbf{x}) = N_0 \log(p + (1-p)\mathrm{e}^{-\mu}) + \sum_{n\geq 1} N_n \left[\log(1-p) + n \log \mu - \mu \right].$$

L'annulation du gradient $\nabla_\theta \ell(\theta, \mathbf{x})$ ne donne pas lieu à une résolution explicite par des fonctions classiques du point de maximum de la vraisemblance.

On complète la variable X (nombre d'enfants à charge) par la variable Z à valeurs dans $\{0, 1\}$ et exprimant l'appartenance à la population A (si $Z = 1$) ou B pour les veuves sans enfant. On note par N_A, N_B le nombre de veuves sans enfant dans les populations A et B resp. de l'échantillon. Vu que $N_B = N_0 - N_A$, les données totales sont

représentées par $\mathbf{y} = (\mathbf{x}, \mathbf{z})$, avec vraisemblance

$$g_\theta(\mathbf{y}) = p^{N_A}((1-p)\mathrm{e}^{-\mu})^{N_0-N_A} \prod_{n\geq 1} \left[\frac{(1-p)\mu^n \mathrm{e}^{-\mu}}{n!}\right]^{N_n}$$

et log-vraisemblance associée (à des termes constants près)

$$\ell(\theta, \mathbf{y}) = N_A \log p + (N_0 - N_A)(\log(1-p) - \mu) + \sum_{n\geq 1} N_n \left[\log(1-p) + n\log\mu - \mu\right]. \tag{16}$$

L'étape $(\mathbb{E})$ est basée sur le calcul de l'espérance suivant :

LEMME 1.7: *Pour $\theta = (p, \mu)$,*

$$\mathbb{E}_\theta^{\mathbf{z}|\mathbf{x}}(N_A) = \frac{N_0 p}{p + (1-p)\mathrm{e}^{-\mu}} \tag{17}$$

DÉMONSTRATION. On a, pour $n = 0, \ldots, N_0$,

$$\begin{aligned} P_\theta(N_A = n|\mathbf{x}) &= \frac{P_\theta(N_A = n, N_B = N_0 - n)}{P_\theta(N_A + N_B = N_0)} \\ &= C_{p,\mu}\binom{N_0}{n} p^n((1-p)\mathrm{e}^{-\mu})^{N_0-n} \\ &= \binom{N_0}{n} \frac{p^n((1-p)\mathrm{e}^{-\mu})^{N_0-n}}{(p + (1-p)\mathrm{e}^{-\mu})^{N_0}} \end{aligned}$$

avec la constante de normalisation

$$C_{p,\mu} = (p + (1-p)\mathrm{e}^{-\mu})^{-N_0} = P_\theta(N_A + N_B = N_0)^{-1}$$

choisie telle que la somme des probabilités soit 1. La loi de Z conditionnellement à N_0 est donc une loi binomiale $B(N_0; p(p+(1-p)\mathrm{e}^{-\mu})^{-1})$, somme de N_0 Bernoulli indépendantes de paramètre $q_\theta = q_{(p,\mu)} = p(p + (1-p)\mathrm{e}^{-\mu})^{-1}$. Ainsi [14]

$$\mathbb{E}_\theta(Z|\mathbf{x}) = \mathbb{E}_\theta(Z|N_0) = N_0 p(p + (1-p)\mathrm{e}^{-\mu})^{-1}$$

□

14. Il s'agit d'un calcul classique

$$\begin{aligned} \mathbb{E}_\theta(Z|\mathbf{x}) &= \mathbb{E}_\theta(Z|N_0) C_{p,\mu} \sum_{n=0}^{N_0} nP(Z = n, x_B = N_0 - n) \\ &= C_{p,\mu} \sum_{n=0}^{N_0} n\binom{x_0}{n} p^n((1-p)\mathrm{e}^{-\mu})^{N_0-n} \\ &= C_{p,\mu}((1-p)\mathrm{e}^{-\mu})^{N_0} \sum_{n=0}^{N_0} n\binom{N_0}{n} (p(1-p)^{-1}\mathrm{e}^{\mu})^n \\ &= C_{p,\mu}((1-p)\mathrm{e}^{-\mu})^{N_0} N_0 p(1-p)^{-1}\mathrm{e}^{\mu}(1 + p(1-p)^{-1}\mathrm{e}^{\mu})^{N_0-1} \\ &= C_{p,\mu}((1-p)\mathrm{e}^{-\mu})^{N_0} p(1-p)^{-1}\mathrm{e}^{\mu})(1 + p(1-p)^{-1}\mathrm{e}^{\mu})^{N_0-1} \\ &= N_0 p(p + (1-p)\mathrm{e}^{-\mu})^{-1} \end{aligned}$$

où on a utilisé

$$\sum_{n=0}^{N} n\binom{N}{n} u^n = u\frac{d}{du}\left[\sum_{n=0}^{N}\binom{N}{n}u^n\right] = u\frac{d}{du}[(1+u)^N] = Nu(1+u)^{N-1}.$$

Par suite, vu (16) et en notant $\overline{p} = 1 - p$,

$$\begin{aligned} Q_{\theta_k}(\theta) &= \mathbb{E}_{\theta_k}^{Z|\mathbf{x}}(\ell(\theta, \mathbf{x})) \\ &= \mathbb{E}_{\theta_k}^{Z|\mathbf{x}}(Z)\log p + (N_0 - \mathbb{E}_{\theta_k}^{Z|\mathbf{x}}(Z))[\log \overline{p} - \mu] + \sum_{n\geq 1} N_n[\log \overline{p} + n\log\mu - \mu] \end{aligned}$$

avec gradient relativement à $\theta = (p, \mu)$

$$\nabla_\theta Q_{\theta_k}(\theta) = \left(\frac{\mathbb{E}_{\theta_k}^{Z|\mathbf{x}}(Z)}{p} - \frac{\sum_{n\geq 0} N_n - \mathbb{E}_{\theta_k}^{Z|\mathbf{x}}(Z)}{\overline{p}}, \mathbb{E}_{\theta_k}^{Z|\mathbf{x}}(Z) - \sum_{n\geq 0} N_n + \frac{\sum_{n\geq 0} nN_n}{\mu} \right)$$

ayant un seul zéro (le point de maximum !) donné par

$$(p^*, \mu^*) = \left(\frac{\mathbb{E}_{\theta_k}^{Z|\mathbf{x}}(Z)}{\sum_{n\geq 0} N_n}, \frac{\sum_{n\geq 1} nN_n}{\sum_{n\geq 0} N_n - \mathbb{E}_{\theta_k}^{Z|\mathbf{x}}(Z)} \right). \tag{18}$$

L'itération (18), où $\mathbb{E}_{\theta_k}^{Z|\mathbf{x}}(Z)$ est défini par (17), devient

$$\begin{aligned} \theta_{k+1} = (p_{k+1}, \mu_{k+1}) &= \left(\frac{N_0}{\sum_{n\geq 0} N_n} \frac{p_k}{p_k + \overline{p_k}\mathrm{e}^{-\mu_k}}, \frac{\sum_{n\geq 1} nN_n}{\sum_{n\geq 0} N_n - N_0 p_k/(p_k + \overline{p_k}\mathrm{e}^{-\mu_k})} \right) \\ &= \left(\frac{\pi_0 p_k}{p_k + \overline{p_k}\mathrm{e}^{-\mu_k}}, \frac{\overline{N}}{1 - \pi_0 p_k/(p_k + \overline{p_k}\mathrm{e}^{-\mu_k})} \right) \end{aligned}$$

où on a posé $\pi_0 = N_0/r$ et $\overline{N} = \sum_{n\geq 0} nN_n/r$ en rappelant $r = \sum_{n\geq 0} N_n$. Ainsi, en reportant la relation de la première composante dans la seconde

$$(p_{k+1}, \mu_{k+1}) = \left(\frac{\pi_0 p_k}{p_k + (1 - p_k)\mathrm{e}^{-\mu_k}}, \frac{\overline{N}}{1 - p_{k+1}} \right).$$

soit $p_{k+1} = h(p_k)$ où h est définie suivant

$$h(x) = \frac{\pi_0 x}{x + (1 - x)\mathrm{e}^{-\overline{N}/(1-x)}}$$

En partant de $(p_0, \mu_0) = (0.75, 0.4)$, la suite (p_k, μ_k) construite par itération $\mathbb{EM}$ converge vers $p_\infty = 0.6150567, \mu_\infty = 1.0378391$ avec convergence rapide, puis lente. Pour étudier la convergence de la suite p_k, il suffit de considérer l'itération $p_{k+1} = h(p_k)$.

3.5.2. *Jumeaux, filles et garçons*[24]. Dans une étude sur les jumeaux, on cherche à estimer les probabilités p de paires de vrais jumeaux et q de paires avec au moins une fille. Pour ce faire, on observe un échantillon de N paires de jumeaux où on compte f (resp. g, m) paires de jumeaux filles (resp. garçons et mixte fille/garçon).

Le triplet $x = (f, g, m)$ constitue les données observées et $\theta = (p, q)$ est le paramètre à déterminer. Si on connaît exactement quelle paire de jumeaux de même sexe est une paire de vrais jumeaux, il est facile d'estimer p et q : on postule donc comme données complètes les 5-uplets $y = (f_1, f_2, g_1, g_2, m)$ où f_1, g_1 décomptent les paires de vraies jumelles et de vrais jumeaux resp. et où les identités $f_2 = f - f_1, g_2 = g - g_1$ valent. Les données complètes suivent une loi multinomiale

$$\texttt{binom}(N, (pq, (1-p)q^2, p(1-q), (1-p)(1-q)^2, (1-p)2q(1-q)))$$

donnant à l'échantillon $y = (f_1, f_2, g_1, g_2, m)$ la vraisemblance

$$g(y, \theta) = \binom{f+g+m}{f_1, f_2, g_1, g_2, m} (pq)^{f_1}[(1-p)q^2])^{f_2}(p(1-q))^{g_1}[(1-p)(1-q)^2]^{g_2}[(1-p)2q(1-q)]^m$$

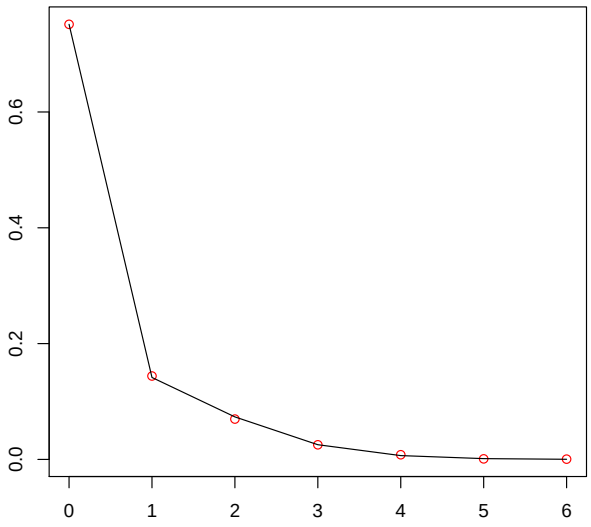


FIGURE I.4 . Comparaison de la distribution $p_\infty\delta_0 + (1-p_\infty)\mathcal{P}(\mu_\infty)$ (`lines((0:6),c(p,rep(0,6))+(1-p)*dpois(0:6,mu))`) et les données (`plot(0:6,data/sumdata,type='p',col="red")`).

vu qu'une paire de vrais jumeaux requiert un seul choix de sexe (d'où les facteurs $(pq)^{f_1}$ et $(p(1-q))^{g_1}$) et des faux jumeaux deux choix de sexe (et donc les facteurs $[(1-p)q^2])^{f_2}$ et $[(1-p)(1-q)^2]^{g_2}$).

À des constantes (additives) près, la log-vraisemblance totale est donnée par

$$\ell(y,\theta) = (f_1+g_1)\log p+(f_2+g_2+m)\log(1-p)+(f_1+2f_2+m)\log q+(g_1+2g_2+m)\log(1-q)$$

Le calcul de l'espérance $Q_{\theta_n}(\theta) = \mathbb{E}_{\theta_n}(\ell(y,\theta))$ se déduit de celle des espérances de f_1, f_2, g_1, g_2, m qui apparaissent linéairement dans la log-vraisemblance $\log g(y,\theta)$

$$\begin{aligned}
f_{1,n} &:= \mathbb{E}_{\theta_n}(f_1|x) = f\frac{p_nq_n}{p_nq_n+(1-p_n)q_n^2},\\
f_{2,n} &:= \mathbb{E}_{\theta_n}(f_2|x) = f\frac{(1-p_n)q_n^2}{p_nq_n+(1-p_n)q_n^2},\\
g_{1,n} &:= \mathbb{E}_{\theta_n}(g_1|x) = g\frac{p_n(1-q_n)}{p_n(1-q_n)+(1-p_n)(1-q_n)^2},\\
g_{2,n} &:= \mathbb{E}_{\theta_n}(g_2|x) = g\frac{(1-p_n)(1-q_n)^2}{p_n(1-q_n)+(1-p_n)(1-q_n)^2},\\
m_n &:= \mathbb{E}_{\theta_n}(m|x) = m,
\end{aligned}$$

en appliquant la règle de Bayes [15], par exemple

$$\mathbb{E}_{\theta_n}(f_1|x) = \mathbb{E}_{\theta_n}(f_1|g) = \frac{\mathbb{E}_{\theta_n}(f_1 1_g)}{P(g,\theta_n)} = \frac{fp_nq_n}{p_nq_n+(1-p_n)q_n^2}$$

et l'additivité

$$\mathbb{E}_{\theta_n}(f_1|x)+\mathbb{E}_{\theta_n}(f_2|x) = \mathbb{E}_{\theta_n}(f|x) = f.$$

15. La règle de Bayes $P(A|B) = P(A,B)/P(B)$ donne pour une variable discrète X la formule $P(A) = \sum_x P(A,X=x) = \sum_x P(A|X=x)P(X=x)$ soit $\mathbb{E}_{\theta_n}(Y|X=x) = \int_\Omega Y(\omega)P(d\omega|X=x) = \int_{X=x} Y(\omega)P(d\omega)/P(X=x) = \mathbb{E}(Y1_{X=x})/P(X=x)$. Le nombre m étant fixé ainsi que le paramètre $\theta_n = (p_n,q_n)$, la variable g_1 est binomiale de paramètre $\alpha_n = p_nq_n/(p_nq_n+p_n(1-q_n)^2)$, d'où son espérance $\mathbb{E}(g_1|m) = m\alpha_n$.

En résulte le calcul de l'espérance $Qtheta_n(\theta) = \mathbb{E}_{\theta_n}(\ell(y,\theta)|x)$

$$\begin{aligned} Q_{\theta_n}(\theta) = &(f_{1,n} + g_{1,n}) \log p + (f_{2,n} + g_{2,n} + m_n) \log(1-p) \\ &+ (f_{1,n} + 2f_{2,n} + m_n) \log q + (g_{1,n} + 2g_{2,n} + m_n) \log(1-q) \end{aligned}$$

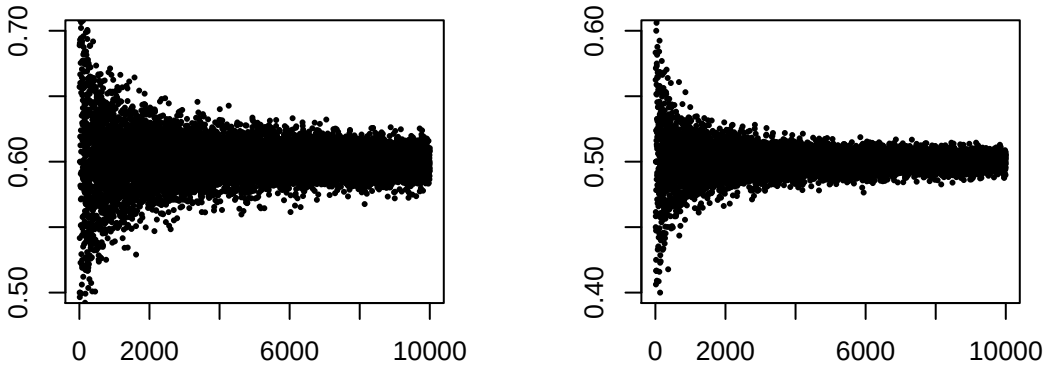
où les variables p, q sont séparées : son point d'optimum [16] est donc

$$p_{n+1} = \frac{f_{1,n} + g_{1,n}}{f+g+m}, \quad q_{n+1} = \frac{f_{1,n} + 2f_{2,n} + m}{f+g+2m+f_{2,n}+g_{2,n}} = \frac{2f - f_{1,n} + m}{2(f+g+m) - f_{1,n} - g_{1,n}}.$$

On a supposé que le nombre de vrais jumeaux et celui de jumeaux de sexe mâle était modélisé par des lois binomiales (cachées). L'itéré p_n prend la forme

$$\mathbb{E}(\#\text{succès}|x,\theta_n)/\mathbb{E}(\#\text{essais}|x,\theta_n)$$

avec le dénominateur fixé égal à $f+g+m$, alors que q_n est de forme analogue avec un dénominateur variable vu que le nombre de choix de sexe dépend du nombre de paires de vrais jumeaux comparé à celui de paires à faux jumeaux.



(a) (b)

FIGURE I.5 . Calcul via $\mathbb{EM}$ de p et q en (a) et (b) resp. pour des échantillons de taille de 5 à 10000 avec $p = 0.6$ et $q = 0.5$.

3.5.3. *Ampoules électriques*[16]. On suppose que la durée de vie d'ampoules électriques suit une distribution exponentielle e_θ de moyenne θ. Pour estimer θ, on teste les durées vie de n ampoules, soit $\mathbf{t} = (t_1, \ldots, t_n)$ avec moyenne $\bar{\mathbf{t}}$. Par ailleurs, on teste un autre lot de m ampoules, sur une période de temps T, au bout de laquelle r ampoules sont mortes.

La donnée manquante est la durée de vie τ pour une ampoule du second lot : cette variable de durée de vie est censurée à droite et à gauche, donnant la variable cachée binaire, suivant que l'ampoule est morte au temps T ou non.

La vraisemblance des données complètes pour les échantillons $\mathbf{t}$ et $\tau = (\tau_1, \ldots, \tau_m)$

$$p(\mathbf{t}, \tau|\theta) = \prod_{i=1}^{n} \left[\theta^{-1} \mathrm{e}^{-t_i/\theta}\right] \prod_{j=1}^{m} \left[\theta^{-1} \mathrm{e}^{-\tau_j/\theta}\right]$$

soit pour la log-vraisemblance

$$\log p(\mathbf{t}, \tau|\theta) = -(n+m) \log \theta - \theta^{-1} \left(n\bar{\mathbf{t}} + \sum_{j=1}^{m} \tau_j \right)$$

16. Pour des constantes α et β, la fonction $p \in (0,1) \mapsto \alpha \log p + \beta \log(1-p)$ atteint son optimum en $p^* = \alpha/(\alpha+\beta)$.

Au temps T, l'espérance de durée de vie totale d'une ampoule (qui n'a pas de mémoire de la période $[0,T]$) encore en fonction est

$$\mathbb{E}_\Lambda(\tau|\text{ampoule en fonction au temps } T) = T + \Lambda,$$

alors que celle d'une ampoule morte est

$$\mathbb{E}_\Lambda(\tau|\text{ampoule morte au temps } T) = \frac{\int_0^T u\Lambda^{-1}\mathrm{e}^{-u/\Lambda}du}{\int_0^T \Lambda^{-1}\mathrm{e}^{-u/\Lambda}du} = \Lambda - T\frac{\mathrm{e}^{-T/\Lambda}}{1-\mathrm{e}^{-T/\Lambda}}.$$

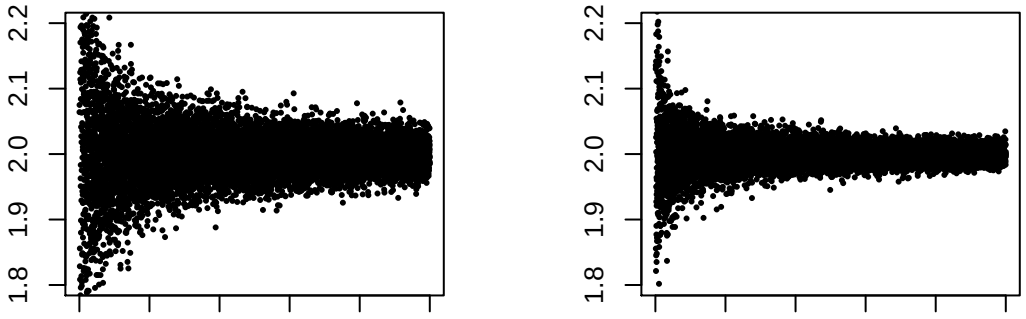
Ainsi, au temps T, vu que $m-r$ ampoules du lot 2 sont encore en fonction et r ampoules mortes, l'espérance

$$\mathbb{E}_{\theta_k}(\log p(\mathbf{t},\tau|\theta)) = -(n+m)\log\theta - \theta^{-1}\left[n\bar{\mathbf{t}} + (m-r)(T+\theta_k) + r\left(\theta_k - T\frac{\mathrm{e}^{-T/\theta_k}}{1-\mathrm{e}^{-T/\theta_k}}\right)\right].$$

La maximisation de $E_{\theta_k}(\log p(\mathbf{t},\tau|\theta))$ en la variable θ donne

$$\theta_{k+1} = \frac{n\bar{\mathbf{t}} + (m-r)(T+\theta_k) + r\left(\theta_k - T\frac{\mathrm{e}^{-T/\theta_k}}{1-\mathrm{e}^{-T/\theta_k}}\right)}{n+m} = \frac{n\bar{\mathbf{t}} + \left(m - \frac{r}{1-\mathrm{e}^{-T/\theta_k}}\right)T + m\theta_k}{n+m}$$

Si on choisit la distribution des durées de vie uniformes $\sim U(0,\theta)$, l'algorithme $\mathbb{EM}$ ne



(a) (b)

FIGURE I.6 . Estimateur θ avec échantillons simulés de $N = 0,\dots 10000$ ampoules : (a) moyenne simple (b) complété par un échantillon observé au temps T et application de $\mathbb{EM}$.

peut être appliqué : les méthodes de vraisemblance maximale doivent être utilisées avec soin quand le domaine de la distribution dépend du paramètre ? ? ? ? ? ?

3.5.4. *Mélanges de tirages de Bernoulli.* Soit un mélange $\alpha_1 B(p_1) + \alpha_2 B(p_2)$ de deux tirages à pile ou face de Bernoulli, où $\alpha_2 = 1-\alpha_1$: le résultat *Face* a comme probabilité p_i dans le tirage $B(p_i)$. Les données observées consistent en M suites $\mathbf{y}_m, m = 1,\dots,M$: la suite de tirages y_m est de longueur ℓ_m, avec f_m résultats *Face* et suivant la loi de Bernoulli $B(\pi_m)$ sans que $\pi_m \in \{p_1,p_2\}$ ne soit connu. Le paramètre à estimer est $\theta = (\alpha_1,p_1,p_2)$

La vraisemblance totale pour un tirage $\mathbf{y}$ de longueur $\ell_\mathbf{y}$, avec $f_\mathbf{y}$ résultats *Face* et suivant a loi $B(p_\zeta)$ ($\zeta \in \{1,2\}$) est :

$$p(\zeta,\mathbf{y}|\theta) = \alpha_\zeta p_\zeta^{f_\mathbf{y}}(1-p_\zeta)^{\ell_\mathbf{y}-f_\mathbf{y}}.$$

Ainsi la log-vraisemblance des résultats des M tirages observés a pour espérance relativement à la loi $dP(\theta_k)$

$$\begin{aligned}\mathbb{E}_M(\theta,\theta_k) &= \sum_{m=1}^{M}\sum_{\zeta\in\{1,2\}} \log p(\zeta,\mathbf{y}_m|\theta)p(\zeta|\mathbf{y}_m,\theta_k)\\ &= \sum_{m=1}^{M}\sum_{\zeta\in\{1,2\}} \log\left[\alpha_\zeta p_\zeta^{f_m}(1-p_\zeta)^{\ell_m-f_m}\right] p(\zeta|\mathbf{y}_m,\theta_k).\end{aligned}$$

En notant, pour $\zeta=1,2$,

$$p_{\zeta,m}(\theta)=p(\zeta|\mathbf{y}_m,\theta)=\frac{p(\zeta,\mathbf{y}_m|\theta)}{p(\mathbf{y}_m|\theta)}=\frac{\alpha_\zeta p_\zeta^{f_m}(1-p_\zeta)^{\ell_m-f_m}}{\sum_{\eta\in\{1,2\}}\alpha_\eta p_\eta^{f_m}(1-p_\eta)^{\ell_m-f_m}}, \tag{19}$$

le maximum $\theta_{k+1}=\operatorname{argmax}_\theta \mathbb{E}(\theta,\theta_k)$ est obtenu en annulant le gradient

$$\begin{aligned}\nabla_{\alpha_1}\mathbb{E}_M(\theta,\theta_k) &= \frac{\sum_{m=1}^M p(1|\mathbf{y}_m,\theta_k)}{\alpha_1}+\frac{\sum_{m=1}^M p(2|\mathbf{y}_m,\theta_k)}{\alpha_1-1}\\ \nabla_{p_\zeta}\mathbb{E}_M(\theta,\theta_k) &= \frac{\sum_{m=1}^M p(\zeta|\mathbf{y}_m,\theta_k)f_m}{p_\zeta}+\frac{\sum_{m=1}^M p(\zeta|\mathbf{y}_m,\theta_k)(\ell_m-f_m)}{p_\zeta-1}\end{aligned}$$

soit

$$\alpha_{1,k+1}=\frac{1}{M}\sum_{m=1}^M p_{1,m}(\theta_k),\quad p_{\zeta,k+1}=\frac{\sum_{m=1}^M p_{\zeta,m}(\theta_k)\ell_m}{\sum_{m=1}^M p_{\zeta,m}(\theta_k)f_m},\quad \zeta=1,2.$$

avec $p_{\zeta,m}(\theta)=p(\zeta|\mathbf{y}_m,\theta)$ donné par (19).

3.5.5. *Mélange de lois normales.* On considère le mélange de deux variables normales $\mathcal{N}(\mu_1,v_1),\mathcal{N}(\mu_2,v_2)$, pondérées resp. par $w\in(0,1)$ et $1-w$. Avec

$$\gamma_{\mu,v}(t)=\frac{\mathrm{e}^{-(t-\mu)^2/(2v)}}{\sqrt{2\pi v}},\quad t\in\mathbb{R},$$

la densité de probabilité du mélange est donc

$$p_\theta(t)=w\gamma_{\mu_1,v_1}(t)+(1-w)\gamma_{\mu_2,v_2}(t),\quad t\in\mathbb{R},$$

avec paramètre $\theta=(\mu_1,\mu_2,v_1,v_2,w)$. La variable cachée z est binaire, valant 1 ou 2 et de loi $B(w)+1$ avec $B(w)$ la loi de Bernoulli $B(w)$: le point x suit la loi $N(\mu_1,v_1)$ ou la loi $N(\mu_2,v_2)$ selon que $z=1$ ou 2 resp. : on pose $w(1)=w$ et $w(2)=1-w$.

La vraisemblance complète associée à l'échantillon $\mathbf{y}=((x_1,z_1),\dots,(x_r,z_r))$ est

$$p_\theta(\mathbf{y})=\prod_{n=1}^r\left[w(z_n)\gamma_{\mu_{z_n},v_{z_n}}(x_n)\right].$$

Par suite, la log-vraisemblance vaut, à des constantes additives près,

$$\ell_\theta(\mathbf{y})=\sum_{n=1}^r\left[\log w(z_n)-(x_n-\mu_{z_n})^2/(2v_{z_n})-\frac{1}{2}\log v_{z_n}\right]$$

Définissons $p_\theta(x)$ suivant

$$p_\theta(x)=P_\theta((x,z)|z=1)=\frac{w\gamma_{\mu_1,v_1}(x)}{w\gamma_{\mu_1,v_1}(x)+(1-w)\gamma_{\mu_2,v_2}(x)},$$

et $\overline{(\mu_1,\mu_2,v_1,v_2,w)} = (\mu_2,\mu_1,v_2,v_1,1-w)$. Alors, avec $p^\ell_\theta(\mathbf{x}) = \sum_{n=1}^r p_\theta(x_n)x_n^\ell$ pour $\ell \in \mathbb{N}$, l'espérance conditionnelle est

$$\begin{aligned}
E_{\theta_k}(\ell_\theta(\mathbf{y})) &= \sum_{n=1}^r [\ell_\theta((x_n,1))P((x_n,z)|\theta_k,z=1) + \ell_\theta(x_n,2))P((x_n,z)|\theta_k,z=2)] \\
&= \sum_{n=1}^r \left[\left[\log w - (x_n-\mu_1)^2/(2v_1) - \frac{1}{2}\log v_1 \right] p_{\theta_k}(x_n) \right. \\
&\qquad \left. + \left[\log(1-w) - (x_n-\mu_2)^2/(2v_2) - \frac{1}{2}\log v_2 \right] p_{\overline{\theta_k}}(x_n) \right] \\
&= \left(\log w - \mu_1^2/(2v_1) - \frac{1}{2}\log v_1 \right) p^0_{\theta_k}(\mathbf{x}) + \mu_1/v_1 p^1_{\theta_k}(\mathbf{x}) - (\mathbf{x})/(2v_1)p^2_{\theta_k} \\
&\quad + \left(\log(1-w) - \mu_2^2/(2v_2) - \frac{1}{2}\log v_2 \right) p^0_{\overline{\theta_k}}(\mathbf{x}) + \mu_2/v_2 p^1_{\overline{\theta_k}}(\mathbf{x}) - (\mathbf{x})/(2v_2)p^2_{\overline{\theta_k}} \\
&= \sum_{j\in\{1,2\}} p^0_{\theta_k(j)}(\mathbf{x}) \left(\log w(j) - \mu_j^2/(2v_j) - \frac{1}{2}\log v_j \right) + p^1_{\theta_k(j)}(\mathbf{x})\mu_j/v_j - p^2_{\theta_k(j)}(\mathbf{x})/(2v_j)
\end{aligned}$$

où on a noté $\theta(j) = (\mu_1,\mu_2,v_1,v_2,w(j))$ pour $j=1,2$. L'espérance $E_{\theta_k}(\ell(\theta,\mathbf{y}))$ atteint son maximum au point θ_{k+1} déterminé par l'annulation des dérivées partielles

$$\begin{aligned}
\nabla_{\mu_j} E_{\theta_k}(\ell(\theta,\mathbf{y})) &= -p^0_{\theta_k(j)}(\mathbf{x})\mu_j/v_j + p^1_{\theta_k(j)}(\mathbf{x})/v_j \\
\nabla_{v_j} E_{\theta_k}(\ell(\theta,\mathbf{y})) &= p^0_{\theta_k(j)}(\mathbf{x}) \left(\mu_j^2/(2v_j^2) - \frac{1}{2v_j} \right) - p^1_{\theta_k(j)}(\mathbf{x})\mu_j/v_j^2 + p^2_{\theta_k(j)}(\mathbf{x})/(2v_j^2) \\
\nabla_w E_{\theta_k}(\ell(\theta,\mathbf{y})) &= \frac{p^0_{\theta_k(1)}}{w} + \frac{p^0_{\theta_k(2)}}{w-1}
\end{aligned}$$

soit

$$\begin{aligned}
\mu_{1,k+1} &= \frac{p^1_{\theta_k}(\mathbf{x})}{p^0_{\theta_k}(\mathbf{x})}, \quad \mu_{2,k+1} = \frac{p^1_{\overline{\theta_k}}(\mathbf{x})}{p^0_{\overline{\theta_k}}(\mathbf{x})} \\
v_{1,k+1} &= \frac{p^2_{\theta_k}(\mathbf{x}) - 2\mu_{1,k+1}p^1_{\theta_k}(\mathbf{x}) + \mu^2_{1,k+1}p^0_{\theta_k}(\mathbf{x})}{p^0_{\theta_k}(\mathbf{x})}, \\
v_{2,k+1} &= \frac{p^2_{\overline{\theta_k}}(\mathbf{x}) - 2\mu_{2,k+1}p^1_{\overline{\theta_k}}(\mathbf{x}) + \mu^2_{2,k+1}p^0_{\overline{\theta_k}}(\mathbf{x})}{p^0_{\overline{\theta_k}}(\mathbf{x})}, \\
w_{k+1} &= \frac{p^0_{\theta_k}(\mathbf{x})}{N}
\end{aligned}$$

où on a utilisé $p^0_{\theta_k(1)}(\mathbf{x}) + p^0_{\theta_k(2)}(\mathbf{x}) = N$.

3.5.6. *Modèles multinomiaux.* On étudie un modèle multinomial [17] tri-valué de paramètres $((1-\theta)/2, \theta/4, (\theta+2)/4)$, pour lequel on a un échantillon de $199 = 38+36+125$

17. Si $p_1+\ldots+p_K = 1$, le modèle multinomial $B(N;p_1,\ldots,p_K)$ a pour support les K-uplets d'entiers dans $\{0,\ldots,N\}^K$ avec comme probabilité

$$P(x_1=N_1,\ldots,x_K=N_K|N_1+\ldots+N_K=N) = \frac{N!}{N_1!\ldots N_K!}\prod_{k=1}^K p_k^{N_k}.$$

Le cas particulier $K=2$ est la loi binomiale $M(N;p,1-p) = B(N;p)$ de De Moivre-Laplace, égale à la loi de la somme de N variables indépendantes de Bernoulli $B(p) = B(1;p)$; la moyenne de Bernoulli

observations se répartissant en trois groupes correspondants au trois valeurs de cardinaux respectifs $38, 36$ et 125.

On considère le modèle quadri-valué $M(199; (1-\theta)/2, \theta/4, \theta/4, 1/2)$ avec les valeurs (latentes) z_1, z_2, z_3, z_4 au dessus des valeurs (observées) x_1, x_2, x_3, x_3 resp. Un échantillon pour ce modèle se répartissant en groupes de cardinaux N_1, N_2, N_3, N_4 donne des groupes de cardinaux $M_1 = N_1, M_2 = N_2, M_3 = N_3 + N_4$ dans le modèle tri-valué.

Le paramètre du maximum de vraisemblance associée à un échantillon observé (à supposer qu'on en ait un !) du modèle quadri-valué est déterminé suivant

(20)
$$\operatorname{argmax}_\theta \left[\left(\frac{1-\theta}{2}\right)^{N_1} \left(\frac{\theta}{4}\right)^{N_2+N_3} \left(\frac{1}{2}\right)^{N_4} \right] = \operatorname{argmax}_\theta \left[(N_1 \log(1-\theta) + (N_2+N_3)\log\theta \right]$$
$$= \frac{N_2+N_3}{N_1+N_2+N_3}$$

La log-vraisemblance $\log p_\theta(\mathbf{z}|\mathbf{x})$ est linéaire en les N_i, ainsi l'espérance $\mathbb{E}_{\widetilde{\theta}}(\log p_\theta(\mathbf{z}|\mathbf{x}))$ exige les calculs des espérances de $N_1, N_2, \ldots$

$$\mathbb{E}_\theta(N_1|\mathbf{x}) = \mathbb{E}_\theta(N_1|M_1 = 38, M_2 = \ldots) = 38$$
$$\mathbb{E}_\theta(N_2|\mathbf{x}) = \mathbb{E}_\theta(N_1|M_1 = 38, M_2 = 34, \ldots) = 34$$
$$\mathbb{E}_\theta(N_3|\mathbf{x}) = 125\frac{\theta}{\theta+2}$$
$$\mathbb{E}_\theta(N_4|\mathbf{x}) = 125\frac{2}{\theta+2}$$

Les évaluations pour les espérances de N_j pour $j = 1, 2$ proviennent des relations $M_j = N_j$ resp. et celles pour N_3 et N_4 résultent du fait que N_3, N_4 sont les cardinaux resp. des valeurs d'un échantillon de $N_3 + N_4 = M_3 = 125$ observations suivant une loi binomiale $B(125; (\theta/4)/(\theta/4+1/2), (1/2)/(\theta/4+1/2)) = B(125; \theta/(\theta+2), 2/(\theta+2))$.

Ainsi, d'après (20), l'estimateur θ_{k+1} déduit de l'optimisation de l'espérance $\mathbb{E}_{\theta_k}(\log p_\theta(\mathbf{z}|\mathbf{x}))$ donne la relation de récurrence

$$\theta_{k+1} = \frac{\mathbb{E}_\theta(N_2|\mathbf{x}) + \mathbb{E}_\theta(N_3|\mathbf{x})}{\mathbb{E}_\theta(N_1|\mathbf{x}) + \mathbb{E}_\theta(N_2|\mathbf{x}) + \mathbb{E}_\theta(N_3|\mathbf{x})} = \frac{34 + 125\frac{\theta_k}{\theta_k+2}}{38 + 34 + 125\frac{\theta_k}{\theta_k+2}} = \frac{159\theta_k + 68}{197\theta_k + 144}.$$

Avec $\theta_0 = 0.5$, on a $\theta_{11} = 0.6268214979\ldots = \theta_\infty$ à 10^{-10} près : θ_∞, point fixe de la transformation $M : \theta_k \mapsto \theta_{k+1}$ [18], est la racine positive du trinôme $197\theta^2 - 15\theta - 68$.

Terminons en remarquant que le point $\theta_{\max}$ pour le maximum de la vraisemblance

$$\left(\frac{1-\theta}{2}\right)^{M_1} \left(\frac{\theta}{4}\right)^{M_2} \left(\frac{\theta+2}{4}\right)^{M_3}$$

étant p, la moyenne de la loi binomiale $B(N;p)$ est Np qu'on retrouve via le calcul

$$\sum_{k=0}^{N} \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k} k = p\frac{\partial}{\partial p}\left(\sum_{k=0}^{N} \frac{N!}{k!(N-k)!} p^k (1-q)^{N-k} \right)_{|p=q} = p\frac{\partial}{\partial p}\left(p+1-q\right)^N)_{|p=q} = pN.$$

18. Avec des a, b, c, d positifs et h définie par $h(x) = (ax+b)/(cx+d)$, l'itération $\theta_{k+1} = h(\theta_k)$ est représentée graphiquement en traçant dans le premier quadrant la branche d'hyperbole $\mathcal{H} : y = h(x), x \geq 0$ et la diagonale qui rencontre $\mathcal{H}$ au point limite $(\theta_\infty, h(\theta_\infty) = \theta_\infty)$: l'itération des points $(\theta_k, h(\theta_k))$ a une abscisse monotone ou oscillant de part et d'autre θ_∞, suivant le signe de $ad - bc$ (qui détermine le type de monotonie de h).

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_m$	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	$\mathbf{x}_{13}$	$\mathbf{x}_{1m_1}$	$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	$\mathbf{x}_{23}$	$\mathbf{x}_{2m_2}$
w_1	1	1	-1	-1	2	2	-2	-2	?	?	?	?
w_2	1	-1	1	-1	?	?	?	?	2	2	-2	-2

TABLE 7. Les données supposées observées de Murray de moyennes nulles et distribuées suivant une loi normale bivaluée

associé à l'échantillon réparti en parties de cardinaux (M_1, M_2, M_3) peut être déterminé directement comme dans (20) en étudiant la dérivée

$$\frac{d}{d\theta}\log\left[(1-\theta)^{38}\theta^{34}(\theta+2)^{125}\right] = \frac{38}{\theta-1}+\frac{34}{\theta}+\frac{125}{\theta+2} = \frac{197\theta^2-15\theta-68}{(\theta-1)\theta(\theta+2)}$$

qui redonne le polynôme exhibé précédemment et dont l'unique zéro positif est le paramètre $\theta \in (0,1)$ cherché !

3.5.7. *Convergence vers un point selle.* Murray [11, p. 27] introduit un exemple où la vraisemblance d'un modèle normal bivalué a deux maxima globaux et un point selle, avec des suites de l'algorithme $\mathbb{EM}$ convergent vers le point selle.

On considère une densité bivaluée $w = {}^{\mathsf{T}}(w_1, w_2)$ avec une distribution normale $w \sim \mathcal{N}(\mu, \Sigma)$ avec $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$ et $\Sigma = \begin{pmatrix} v_1 & \rho\sqrt{v_1v_2} \\ \rho\sqrt{v_1v_2} & v_2 \end{pmatrix}$. La densité vaut $p(w|\mu,\Sigma) = (2\pi)^{-1}(\det\Sigma)^{-1/2}\exp(-\frac{1}{2}\langle w-\mu, \Sigma^{-1}(w-\mu)\rangle)$ pour $w \in \mathbb{R}^2$.

Soit un échantillon de taille n simulant w dont $m_k, k = 1,2$ valeurs avec la seule coordonnée celle d'indice k (l'autre étant donc manquante). On suppose que les données absentes sont totalement aléatoires, de telle sorte que les données observées peuvent être considérées comme un échantillon aléatoire de taille $m = n - m_1 - m_2$ de la loi bivaluée w et une paire indépendantes d'échantillons aléatoires de taille m_1, m_2 de la distribution $w_i \sim \mathcal{N}(\mu_i, \sqrt{v_i})$. Au terme additif $n\log(2\pi)$ près, la log-vraisemblance $\ell(\theta|\mathbf{x})$ des données observées $\mathbf{x} = (x_1, \dots, x_m, x_{11}, \dots, x_{1m_1}, x_{21}, \dots, x_{2m_2})$ vérifie, avec $\theta = (\mu, \mu_1, \mu_2, \Sigma = \begin{pmatrix} v_1 & \sqrt{v_1v_2}\rho \\ \sqrt{v_1v_2}\rho & v_2 \end{pmatrix})$

$$-2\ell(\theta|\mathbf{x}) = m\log\det\Sigma + \sum_{i=1}^{m}\langle x_i-\mu, \Sigma^{-1}(x_i-\mu)\rangle + \sum_{k=1}^{2}\left\{m_k\log v_k + \frac{1}{v_k}\sum_{i_k=1}^{m_k}(x_{ki_k}-\mu_k)^2\right\}$$

Considérons les données observées

On prend μ, μ_1, μ_2 tous nuls. Vu que $\Sigma^{-1} = \begin{pmatrix} v_2 & -\sqrt{v_1v_2}\rho \\ -\sqrt{v_1v_2}\rho & v_1 \end{pmatrix} / \det\Sigma$,

$$\begin{aligned} -2\ell(v_1,v_2,\rho|\mathbf{x}) &= m\log\det\Sigma + \sum_{i=1}^{m}\langle x_i, \Sigma^{-1}x_i\rangle + \sum_{k=1}^{2}\left[m_k\log v_k + \frac{1}{v_k}\sum_{i_k=1}^{m_k}x_{ki_k}^2\right] \\ &= 4\log(v_1v_2(1-\rho^2)) + 4\frac{v_2+v_1}{v_1v_2(1-\rho^2)} + 4\log(v_1v_2) + \frac{16}{v_1} + \frac{16}{v_2} \\ &= 4\left[2\log(v_1v_2) + \log(1-\rho^2) + (v_1^{-1}+v_2^{-1})((1-\rho^2)^{-1}+4)\right] \end{aligned}$$

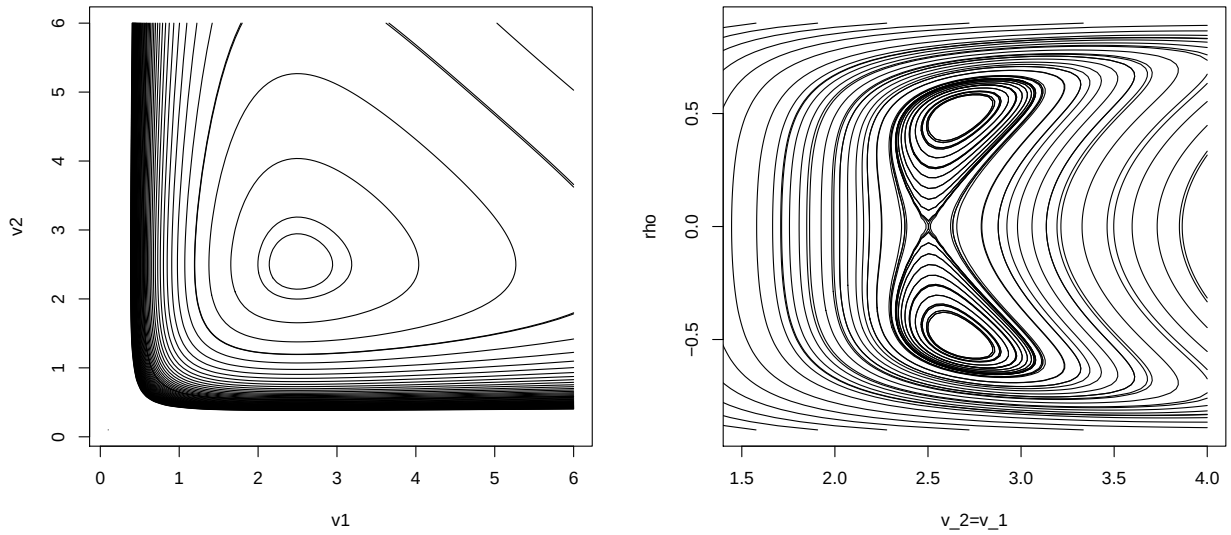


FIGURE I.7 . Les lignes de niveaux sur le plan $\rho = 0$ et sur le plan $v_1 = v2$.

avec ses point critiques déterminés par les 3 équations

$$0 = \frac{2}{v_k} - \frac{1}{v_k^2}((1-\rho^2)^{-1} + 4), \quad k = 1, 2,$$
$$0 = 2\rho\left(\frac{1}{\rho^2 - 1} + \frac{v_1^{-1} + v_2^{-1}}{(1-\rho^2)^2}\right),$$

soit deux points de maxima en $(v_1, v_2, \rho)_\pm = (8/3, 8/3, \pm 1/2)$ avec hessiennne

$$\text{Hess}\,\ell()_{(8/3,8/3,1/2)} = \begin{pmatrix} 9/32 & 0 & -1/4 \\ 0 & 9/32 & -1/4 \\ -1/4 & -1/4 & 16/9 \end{pmatrix} \tag{21}$$

de spectre $(0.2019\ldots, 1.8571\ldots, 0.2812500\ldots)$ et un point selle $(5/2, 5/2, 0)$ de hessienne la matrice diagonale $\text{Diag}(8/25, 8/25, -2/5)$

Pour la loi normale bivaluée $w \sim \mathcal{N}(\mu, \Sigma)$, on a $w_2|w_1 \sim \mathcal{N}(\mu_{2,1}, v_{2,1})$ où

$$\mu_{2,1} = \mu_2 + \rho\sqrt{v_2/v_1}(w_1 - \mu_1), \quad v_{2,1} = v_2(1-\rho^2).$$

Par conséquent,

$$E(w_2|w_1, \theta) = \mu_{2,1}, \quad E(w_2^2|w_1, \theta) = v_{2,1} + \mu_{2,1}^2$$

soit, si les moyennes μ_1, μ_2 sont nulles,

$$E(w_2|w_1, \theta) = \rho\sqrt{v_2/v_1}w_1, \quad E(w_2^2|w_1, \theta) = v_2(1-\rho^2) + \rho^2 v_2/v_1 w_1^2$$

$$
\begin{aligned}
-2Q_{\theta_k}(\theta) &= n \log \det \Sigma + E\left(\sum_{i=1}^{n} \langle y_i, \Sigma^{-1} y_i \rangle \,\middle|\, \mathbf{x}, \theta_k \right) \\
&= n \log \det \Sigma + \sum_{i=1}^{m} E(\langle x_i, \Sigma^{-1} x_i \rangle | \mathbf{x}, \theta_k) + \sum_{i=1}^{m_1} \frac{E(v_2 x_{1i}^2 - 2\rho\sqrt{v_1 v_2} x_{1i} z_{1i} + v_1 z_{1i}^2 | \mathbf{x}, \theta_k)}{v_1 v_2 (1-\rho^2)} \\
&\quad + \sum_{i=1}^{m_2} \frac{E(v_2 z_{2i}^2 - 2\rho\sqrt{v_1 v_2} z_{2i} x_{2i} + v_1 x_{2i}^2 | \mathbf{x}, \theta_k)}{v_1 v_2 (1-\rho^2)} \\
&= n \log \det \Sigma + \sum_{i=1}^{m} \langle x_i, \Sigma^{-1} x_i \rangle + \sum_{i=1}^{m_1} \frac{v_2 x_{1i}^2 - 2\rho\sqrt{v_1 v_2} x_{1i} E(z_{1i} | \mathbf{x}, \theta_k) + v_1 E(z_{1i}^2 | \mathbf{x}, \theta_k)}{v_1 v_2 (1-\rho^2)} \\
&\quad + \sum_{i=1}^{m_2} \frac{v_2 E(z_{2i}^2 | \mathbf{x}, \theta_k) - 2\rho\sqrt{v_1 v_2} x_{2i} E(z_{2i} | \mathbf{x}, \theta_k) + v_1 x_{2i}^2}{v_1 v_2 (1-\rho^2)} \\
&= n \log \det \Sigma + \sum_{i=1}^{m} \langle x_i, \Sigma^{-1} x_i \rangle \\
&\quad + \sum_{i=1}^{m_1} \frac{v_2 x_{1i}^2 - 2\rho\sqrt{v_1 v_2} \rho_k \sqrt{v_{2k}/v_{1k}} x_{1i}^2 + v_1 [v_{2k}(1-\rho_k^2) + \rho_k^2 v_{2k}/v_{1k} x_{1i}^2]}{v_1 v_2 (1-\rho^2)} \\
&\quad + \sum_{i=1}^{m_2} \frac{v_2 [v_{1k}(1-\rho_k^2) + \rho_k^2 v_{1k}/v_{2k} x_{2i}^2] - 2\rho\sqrt{v_1 v_2} \rho_k \sqrt{v_{2k}/v_{1k}} x_{2i}^2 + v_1 x_{2i}^2}{v_1 v_2 (1-\rho^2)} \\
&= n \log \det \Sigma + \sum_{i=1}^{m} \langle x_i, \Sigma^{-1} x_i \rangle \\
&\quad + \sum_{i=1}^{m_1} \frac{v_1^{-1} x_{1i}^2 - 2\rho (v_1 v_2)^{-1/2} \rho_k \sqrt{v_{2k}/v_{1k}} x_{1i}^2 + v_2^{-1} [v_{2k}(1-\rho_k^2) + \rho_k^2 v_{2k}/v_{1k} x_{1i}^2]}{1-\rho^2} \\
&\quad + \sum_{i=1}^{m_2} \frac{v_1^{-1} [v_{1k}(1-\rho_k^2) + \rho_k^2 v_{1k}/v_{2k} x_{2i}^2] - 2\rho\sqrt{v_1 v_2} \rho_k \sqrt{v_{1k}/v_{2k}} x_{2i}^2 + v_2^{-1} x_{2i}^2}{1-\rho^2}
\end{aligned}
$$

soit pour les données du tableau 7

$$
\begin{aligned}
-2Q\theta_k(\theta) &= 12 \log(v_1 v_2 (1-\rho^2)) + 4 \frac{v_1^{-1} + v_2^{-1}}{1-\rho^2} \\
&\quad + 16 \frac{v_1^{-1} - 2\rho (v_1 v_2)^{-1/2} \rho_k \sqrt{v_{2k}/v_{1k}} + v_2^{-1} [v_{2k}(1-\rho_k^2)/4 + \rho_k^2 v_{2k}/v_{1k}]}{1-\rho^2} \\
&\quad + 16 \frac{v_1^{-1} [v_{1k}(1-\rho_k^2)/4 + \rho_k^2 v_{1k}/v_{2k}] - 2\rho (v_1 v_2)^{-1/2} \rho_k \sqrt{v_{1k}/v_{2k}} + v_2^{-1}}{1-\rho^2}.
\end{aligned}
$$

et en particulier pour $\rho_k = 0$

$$
-\frac{1}{2} Q_{v_{1k}, v_{2k}, \rho_k = 0}(\theta) = 3 \log(v_1 v_2 (1-\rho^2)) + \frac{5[v_1^{-1} + v_2^{-1}] + v_2^{-1} v_{2k} + v_1^{-1} v_{1k}}{1-\rho^2}
$$

dont le seul point critique[19] est $\theta_{k+1} = ((5+v_{1k})/3, (5+v_{2k})/3, 0)$. Ainsi la suite $(\theta_k)_{k\geq 0}$ vérifiant

$$\rho_{k+1} = 0, v_{1(k+1)} = \frac{5+v_{1k}}{3}, v_{2(k+1)} = \frac{5+v_{2k}}{3}$$

est une suite $\mathbb{EM}$ convergente vers le point $(5/2, 5/2, 0)$, point selle de la vraisemblance $\exp(\ell(\theta|\mathbf{x}))$. associée à l'ensemble de données observées $\mathbf{x}$.

19. La fonction $Q_{v_{1k},v_{2k},\rho_k=0}(\theta)$ est coercive sur $\mathbb{R}_+^2 \times (-1,1)$. En effet $\log u + u^{-1} \geq 1$ pour $u > 0$, ce qui implique des inégalités du type $\log(abc) \geq (ac)^{-1} + (bc)^{-1} \geq 2$ pour $a, b > 0, c \in (0,1)$ par exemple.

Chapitre 2

Programmation différentiable

Cette section est consacré à des problèmes du type $\inf_{x\in C} U(x)$ où C est de le forme $C = \{x \in \Omega, g(x) = 0, h(x) \geq 0\}$ avec U, g, h différentiables sur l'ouvert Ω de $\mathbb{R}^n$. Les contraintes peuvent être absentes, voire écartées après une paramétrisation. Néanmoins, il est préférable parfois de garder la présentation géométrique des

▷ Exemple 2.1: Le problème de la surface d'une boîte privée d'une face et de volume 1, peut apparaître comme un problème avec contraintes ou sans : les programmes

$$\inf_{\substack{x,y,z>0\\ xyz=1}} [2(yz+zx)+xy)] \qquad \inf_{x,y>0}\left[\frac{2}{x}+\frac{2}{y}+xy\right] \tag{22}$$

sont équivalents, où le second a comme point de minimum $m_* = (\sqrt[3]{2}, \sqrt[3]{2})$ avec hessienne $\operatorname{Hess} U(x,y) = 1/(xy)^3 \begin{pmatrix} 2y^2 & (xy)^2 \\ (xy)^2 & 2x^2 \end{pmatrix}$ définie positive en m_*. Voir l'exemple **??** pour un argument de coercivité. ◁

1. Extrema locaux différentiables

Un problème d'*optimisation sans contrainte* considéré ici est du type

$$\inf_{x\in\Omega} J(x)$$

où Ω est un ouvert [1] d'un espace vectoriel, souvent de dimension finie, et J une fonction continue au minimum.

Définition 2.1: *Soit J définie et différentiable sur un ouvert V de l'espace vectoriel E. Le point $x \in V$ est un* point critique *(dit parfois* stationnaire*) de J si $dJ_x = 0$ ou de manière équivalente $\nabla J(x) = 0$.*

Théorème 2.1 (Euler, Fermat): *Si la fonction différentiable $J : V \to \mathbb{R}$ est minimum (ou maximum) en $x_* \in V$, alors x_* est point critique de J.*

Démonstration. Si x_* est minimum de J, alors, pour $h \in R^N$ et $t > 0$ avec th assez petit

$$\frac{J(x_*+th)-J(x_*)}{t} \geq 0,$$

d'où

$$dJ_{x_*}(h) + \|h\|\varepsilon(th) \geq 0$$

et donc $dJ_{x_*}(h) \geq 0$ en faisant $t \to 0$. S'ensuit aussi $-dJ_{x_*}(h) = dJ_{x_*}(-h) \geq 0$ et donc $dJ_{x_*}(h) = 0$ pour tout vecteur h, soit la nullité de la différentielle dJ_{x_*}. □

▷ Exemples 2.2:

1. L'ouvert Ω peut être défini par une fonction F sous la forme $\Omega = \{F > 0\}$: cette inégalité, « ouverte », n'est pas considérée comme une contrainte, au contraire des inégalités larges des sections suivantes.

2.2.1 La fonction affine $J_C : x \in E \mapsto \langle c, x\rangle$ n'a pas de point critique, excepté si $c = 0$, auquel cas tout point de E est critique. Les optima de fonctions linéaires sur des polyèdres ne seront jamais à l'intérieur de ceux-ci : la programmation linéaire, qui considère ces situations, n'aura pas la tache facile, alors que souvent l'étude des phénomènes linéaires en analyse est plus aisée que le non-linéaire !

2.2.2 La fonction $J_\pm$ définie par $J_\pm(x, y) = x^2 \pm y^2$ a 0 comme point critique : l'origine 0 est un minimum global strict de J_+, alors que J_- n'a ni minimum, ni maximum (local ou global).

2.2.3 La fonction $J(x, y) = \cos x + y^2$ a comme points critiques $(k\pi, 0)$ pour tout $k \in \mathbb{Z}$.

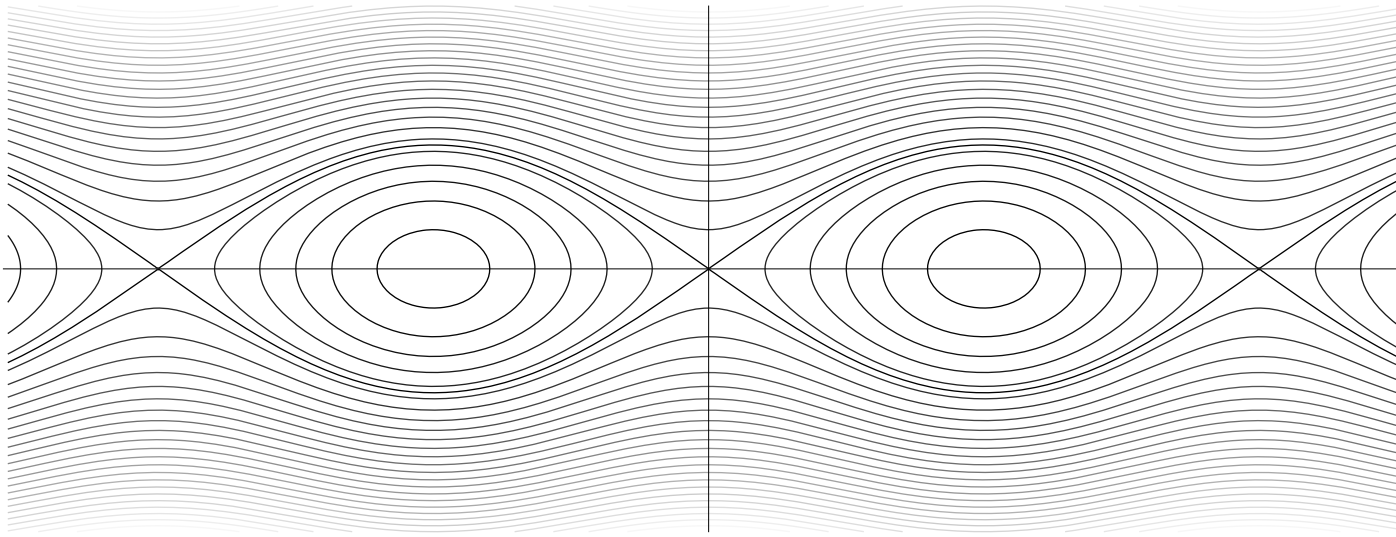


FIGURE II.1 . Les courbes de niveau de la fonction d'objectifs $J(x, y) = \cos x + y^2$.

2.2.4 Soit $E_{a,b} = \mathcal{C}^2([0, 1]) \cap \{y(0) = a, y(1) = b\}$ et $\ell_{a,b}(y) = \int_0^1 \sqrt{1 + y'^2(x)}dx$ la longueur de la courbe plane induite par le graphe de $y \in E_{a,b}$: on cherche à minimiser la longueur $\ell_{a,b}(y)$ des courbes de $(0, a)$ à $(1, b)$ qui sont des graphes $G_y = \{(x, y(x)), x \in [0, 1]\}$. Si $y \in E_{a,b}$ est un minimum de $\ell_{a,b}$, alors pour tout $h \in E_{0,0}$, $y + th$ est dans $E_{a,b}$ et l'annulation au premier ordre de $t \mapsto \ell_{a,b}(y + th)$ donne

$$\begin{aligned} 0 = \frac{d}{dt}\left[\ell_{a,b}(y + th)\right]_{|t=0} &= \frac{d}{dt}\left[\int_0^1 \frac{(y'(x) + th'(x))h'(x)}{\sqrt{1 + (y'(x) + th'(x))^2}}\right]_{|t=0} \\ &= \int_0^1 \frac{y'(x)h'(x)}{\sqrt{1 + y'^2(x)}}dx = -\int_0^1 \frac{d}{dx}\left[\frac{y'(x)}{\sqrt{1 + y'^2(x)}}\right] h(x)dx \\ &= -\int_0^1 \frac{y''(x)h(x)}{(1 + y'^2(x))^{3/2}}dx, \end{aligned}$$

où une intégration par parties a été effectuée à la troisième égalité. Ainsi, par densité de $E_{0,0}$, on a $y'' = 0$ et par suite $y(x) = a + x(b - a), x \in [0, 1]$: la courbe de longueur minimale est nécessairement un segment de droite, correspondant

à l'unique fonction affine $y_{a,b} \in E_{a,b}$. C'est effectivement un minimum : pour $y \in E_{a,b}$ et avec $\| \ \|_2$ la norme euclidienne du plan $\mathbb{R}^2$,

$$1+(b-a)^2 = \langle (1,b-a), \int_0^1 (1,y'(x))dx\rangle \leq \|(1,b-a)\|_2 \left\| \int_0^1 (1,y'(x))dx \right\|_2$$
$$\leq \sqrt{1+(b-a)^2} \int_0^1 \|(1,y'(x))\|_2 dx$$

soit

$$\ell_{a,b}(y_{a,b}) = \sqrt{1+(b-a)^2} \leq \int_0^1 \sqrt{1+y'(x)^2}dx = \ell_{a,b}(y).$$

2.2.5 Soit Ω un ouvert régulier de $\mathbb{R}^2$, représentant une membrane élastique horizontale. La fonction $u : \Omega \to \mathbb{R}$ modélisant les variations de la membrane à l'équilibre sous l'action d'une force verticale d'intensité f est un minimum pour le problème

$$\inf_{v \in H_0^1(\Omega)} \int_\Omega [|\nabla v|^2(m) - f(m)v(m)]dm$$

et vérifie l'équation

$$\Delta u(m) = f(m), \quad m \in \Omega, \qquad u(m) = 0, \quad m \in \partial\Omega$$

Le cadre de ces deux derniers exemples est en dimension infinie : ils ne seront guère abordés dans la suite. ◁

THÉORÈME 2.2: *Soit J critique en x_*.*

(i) Si x_* est un minimum (maximum resp.) local de J, la hessienne $\mathrm{Hess}(J)_{x_*}$ est une forme quadratique positive (négative resp.).

(ii) Si la hessienne $\mathrm{Hess}(J)_{x_*}$ est une forme quadratique définie positive (négative resp.) et x_* point critique de J, alors x_* est un minimum (maximum resp.) local de J.

▷ EXEMPLES 2.3:

2.3.1 Soit la fonction d'objectif $J_\pm$ définie par $J_\pm(x,y) = x^2 \pm y^4$: 0 est minimum global de J_+, mais pas de J_-.

2.3.2 Reprenant la fonction $J(x,y) = \cos x + y^2$ de l'exemple précédent, on a au voisinage de $M_k = (0, k\pi)$

$$J(k\pi+u,y) = y^2 + (-1)^k \cos u = (-1)^k + y^2 - (-1)^k \frac{u^2}{2} + (u^2+y^2)\varepsilon(u^2+y^2)$$

Si k est impair, le point M_k est un minimum local, alors que si k est pair, c'est un point de minimum le long de la verticale $\{(k\pi, y)\}$ et de maximum le long de l'horizontale $\{(k\pi+u, 0)\}$, *i. e.* M_k est un point selle[2].

2.3.3 Soit A une matrice symétrique d'ordre n, b un vecteur de $\mathbb{R}^n$ et c une constante. La fonction J définie par

$$J(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c$$

a comme différentielle $dJ_x(h) = \langle Ax, h\rangle + \langle b, h\rangle$ soit comme gradient $\nabla J_x = Ax + b$ et hessienne $\mathrm{Hess}(J)_x = A$.

Les points critiques x de J sont caractérisés donc par l'équation $Ax + b = 0$: si A est inversible, J a un unique point critique $x_* = -A^{-1}b$.

2. Le point x_* est un point selle de la fonction U si en partant de x_* suivant certaines directions la fonction U croît alors que suivant d'autres elle décroît.

C'est un minimum local si A est définie positive. Il existe une matrice B symétrique telle que $A = B^2$: la matrice A est diagonalisable dans une base orthonormée, *i. e.* il existe une matrice orthogonale P, une matrice diagonale $D(\lambda_1, \dots, \lambda_n)$ avec les λ_i tous positifs non nuls vu que A est définie positive, tels que $A = PD(\lambda_1, \dots, \lambda_n)^{\mathsf{T}}P$; il suffit de prendre $B = PD(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})^{\mathsf{T}}P$. Ainsi

$$\begin{aligned} J(x) - c &= \frac{1}{2}\langle B^2x, x\rangle - \langle B^2x_*, x\rangle = \frac{1}{2}\langle Bx, Bx\rangle - \langle Bx_*, Bx\rangle \\ &= \frac{1}{2}(\langle B(x - x_*), B(x - x_*)\rangle - \frac{1}{2}\langle Bx_*, Bx_*\rangle \\ &= \frac{1}{2}\|B(x - x_*)\|^2 - \frac{1}{2}\|Bx_*\|^2 \end{aligned}$$

Le point x_* est clairement un minimum global strict.

Si A est définie négative, x_* est un maximum local. Écrivant $A = -B^2$, on obtient pareillement

$$J(x) = -\frac{1}{2}\|B(x - x_*)\|^2 + \frac{1}{2}\|Bx_*\|^2 + c. \qquad \triangleleft$$

DÉMONSTRATION. Soit x_* minimum de J. Vu que ∇J_{x_*} est nul, la formule de Taylor à l'ordre 2 s'écrit, pour h donné et t assez petit

$$0 \leq J(x_* + th) - J(x_*) = \frac{\mathrm{Hess}(J)_{x_*}(th)}{2} + \|th\|^2\varepsilon(th) = t^2\left(\frac{\mathrm{Hess}(J)_{x_*}(h)}{2} + \|h\|^2\varepsilon(th)\right)$$

soit, en divisant par $t^2 > 0$,

$$0 \leq \frac{\mathrm{Hess}(J)_{x_*}(h)}{2} + \|h\|^2\varepsilon(th)$$

et par suite $\mathrm{Hess}(J)_{x_*}(h) \geq 0$ pour tout h en faisant tendre $t \to 0$: voilà la première assertion démontrée.

La matrice hessienne étant définie positive, il existe [3] une constante $C = C_{x_*}$ telle que

$$\mathrm{Hess}(J)_{x_*}(h) \geq 4C\|h\|^2, \quad h \in \mathbb{R}^n$$

Ainsi,

$$J(x_* + h) - J(x_*) = \frac{\mathrm{Hess}(J)_{x_*}(h)}{2} + \|h\|^2\varepsilon(h) \geq (2C - |\varepsilon(h)|)\|h\|^2 \geq C\|h\|^2 > 0$$

pour h non nul suffisamment petit de telle sorte que $|\varepsilon(h)| \leq C$: on a bien montré que x_* était un minimum local strict de J. □

Terminons par la définition des types de convergence de la suite $(x_k)_{k\geq 0}$ vers une limite x_* :

DÉFINITION 2.2: *Soit $q_{k,d} = \|x_{k+1} - x_*\|/\|x_k - x_*\|^d$. La convergence de x_k vers x_* est dite*

- linéaire *s'il existe $k_0, \theta \in (0,1)$ tels que $q_{k,1} \leq \theta$ pour $k \geq k_0$ (autrement dit* $\limsup q(k,1) \in (0,1)$*),*
- superlinéaire *si $q_{k,1} \to 0$;*
- quadratique *s'il existe une constante C telle que $q_{k,2} \leq C$*

3. On a vu que qu'une matrice symétrique A était de la forme $A = PD(\lambda_1, \dots, \lambda_n)^{\mathsf{T}}P : < Ax, x >= \langle PD(\lambda_1, \dots, \lambda_n)^{\mathsf{T}}Px, x\rangle = \langle D(\lambda_1, \dots, \lambda_n)^{\mathsf{T}}Px, {}^{\mathsf{T}}Px\rangle \geq \inf_i(\lambda_i)\langle {}^{\mathsf{T}}Px, {}^{\mathsf{T}}Px\rangle = \inf_i(\lambda_i)\langle x, x\rangle$ vu que P est orthogonale. On peut prendre comme constante C le quart de la plus petite valeur propre (qui est positive non nulle vu que le caractère défini positif de la hessienne) de $\mathrm{Hess}(J)_{x_*}$.

— d'ordre d *s'il existe une constante* C *telle que* $q_{k,d} \leq C$

Pour une convergence linéaire, on a $\|x_k - x_*\| \leq \|x_0 - x_*\|\theta^k$ *: vu* $\theta^k = 10^{-|\log_{10}\theta|k}$ *: ainsi si* $C = 10^{-\ell}$*, l'itéré* x_k *a ses* $E[|\log_{10}\theta|k + \ell - 1$ *premières décimales justes : le nombre de décimales justes est linéaire en* k*, avec taux d'accroissement* $|\log_{10}\theta|$ *d'autant plus élevé que* θ *est proche de zéro.*

Pour une convergence quadratique, posant $u_k = C\|x_k - x_*\|$*, on a* $u_k \leq u_{k-1}^2 \leq \ldots \leq u_{k_0}^{2^{k_0}}$*, soit* $\|x_k - x_*\| \leq \theta_0^{2^{k-k_0}}/C$ *avec* $\theta_0 = C\|x_{k_0} - x_*\|$ *: choisissant* k_0 *tel que* $\theta_0 < 1$*, on obtient donc* $\|x_k - x_*\| \leq C_0\theta_0^{2^k} = C_0 10^{-|\log_{10}\theta_0|2^k}$*. Il y a doublement du nombre de décimales exactes à chaque itération.*

△ Remarque 2.1: Pour exhiber la convergence linéaire, il est parfois commode de tracer les points $(k, \ln(\|x_k - x_*\|)$: si la convergence est effectivement linéaire, le graphe laisse apparaître une droite. La convergence quadratique est visible en terme de doublement du nombre de décimales exactes à chaque itération (par ex ; pour une méthode de type Newton). ▽

Différents types de programmes se ramènent à des problèmes sous contraintes.

(1) Si la partie E est paramétrée, par ex. suivant $m \in \mathcal{E} \mapsto p(m) \in E$, le problème sous contrainte $\min_{x\in E} J(x)$ se ramène au problème (sans contrainte) $\inf_{m\in\mathcal{E}} J \circ p(m)$. L'exemple 2 dans **??** ci-dessous peut être considéré comme le problème $\inf_{\theta\in\mathbb{R}t\in\mathbb{R}}[(t - \cos\theta)^2 + (L - t - \sin\theta)^2]$.

(2) On transforme le programme $\min_{g(m)=0} J(m)$ en le programme avec pénalisation P_ε : $\min_{m\in E}[K(m)+\varepsilon^{-1}|g(m)|^2]$. Heuristiquement, le minimum m_*^ε du programme P_ε tend à vérifier la contrainte $g(m) = 0$ du fait du facteur ε^{-1} lorsque $\varepsilon \to 0^+$: un point m ne vérifiant pas la contrainte est *pénalisé* par le facteur ε^{-1} à être un minimum du programme P_ε.

2. Dissections pour des fonctions d'une variable

Si les problèmes d'optimisation univariée apparaissent très fréquemment *per se*, ils sont tout autant des composantes essentielles dans la résolution de programmes plurifactoriels.

Une fonction $J : [a, b] \to \mathbb{R}$ admettant un minimum x_* est dite *unimodale* si elle est strictement croissante (décroissante resp.) à droite de x_* (à gauche resp.) [4]. On suppose que le minimum est dans un intervalle $[a_0, b_0]$. On va montrer trois méthodes itératives qui construisent une suite d'intervalles $[a_k, b_k]$ encadrant le point de minimum x_* et de longueur tendant vers 0 donnant une convergence linéaire vers x_* : $b_k - a_k \leq C\theta^M$, soit par exemple $0 < x_* - a_k \leq C\theta^M$ où M compte le nombre d'opérations non élémentaires (*i. e.* évaluation $J(x)$ ou $J'(x)$) nécessaires pour atteindre la valeur $[a_k, b_k]$ ($M(k) = 2k$ pour les méthodes de trisection ou quadrisection uniformes, $M(k) = k$ pour les autres). Chaque méthode est déterminée par l'étape itérative de passage d'un intervalle $[a, b]$ à un intervalle $[a_+, b_+]$ encadrant le minimum x_*.

2.1. Minimum comme zéro de la dérivée. Si J est de classe $\mathcal{C}^1$ et de dérivée $f = J'$, on peut chercher le minimum comme solution de l'équation $f(x) = 0$ dans l'intervalle $[a, b]$ pour lequel $f(a) < 0, f(b) > 0$: la méthode de dichotomie pour approcher le zéro x_* suit l'algorithme 2.1. Cet algorithme converge linéairement avec l'itéré x_k dans un

4. *Stricto sensu* et en extension des *distributions unimodales* de la statistique où un *mode* désigne une point de maximum de la distribution discrète ou de la fonction de densité, une fonction unimodale est une fonction avec un seul minimum : on utilise ici dans le cas d'une variable une définition un peu plus restrictive.

Algorithme 2.1 La dichotomie pour approcher le zéro de f continue avec $f(a)f(b) < 0$.

```
Choisir a, b tels que f(a)f(b) < 0
tant que b − a > η faire
    m := (a + b)/2
    si f(m)f(b) > 0 alors
        b := m
    sinon
        a := m
    fin si
fin tant que
```

intervalle de longueur $\varepsilon_k = 2^{-k}(b-a)$ et pour chaque itération l'évaluation d'une dérivée de J (qu'il faut expliciter).

Lorsque J n'est pas dérivable (mais toujours supposée unimodale), il faut au moins 3 intervalles successifs pour, à partir des ordres des valeurs de J aux extrémités de ces intervalles, localiser le minimum de J dans un ou deux intervalles jointifs parmi ces trois. Les deux méthodes de multisection suivantes empruntent des stratégies différentes

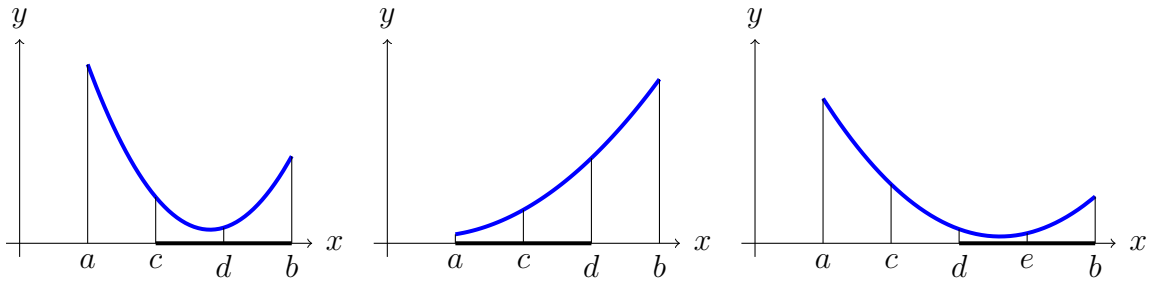


FIGURE II.2 . Trisections et quadrisection d'un intervalle $[a, b]$, avec sous-intervalle (en gras) contenant (nécessairement) le minimum pour l'itération suivante.

de sélection des sous-intervalles, la méthode dite de dissection dorée ayant une vitesse de décroissance des longueurs d'intervalles itérés plus rapide.

2.2. Trisection uniforme. On peut faire une trisection de $[a, b]$ en trois intervalles $[a, b] = [a, c] \cup [c, d] \cup [d, b]$ de même longueur, puis déterminer la valeur minimum $J_{\text{dis}} = J(x_{\text{dis}})$ de J sur l'ensemble fini $\{J(a), J(b), J(c), J(d)\}$: le minimum (continu) de J appartient à l'intervalle $[a_+, b_+]$ encadrant le minimum discret x_{dis} et l'intervalle $[a_+, b_+]$ a une longueur au plus dans un rapport $2/3$ avec $b-a$. À l'itération k, on a un intervalle de longueur au plus $(2/3)^k(b_0 - a_0)$ après avoir effectué $N = 2 + 2k$ évaluations de la fonction : l'erreur en fonction du nombre N d'opérations est donnée par $\varepsilon_N \approx \theta^N$ avec $\theta = (2/3)^{1/2} \approx 0.8165$.

2.3. Quadrisection uniforme. On peut raffiner la méthode précédente par des subdivisions itératives de l'intervalle $[a, b]$ contenant le minimum en quatre intervalles de même longueur : $[a, b] = [a, c] \cup [c, d] \cup [d, e] \cup [e, b]$. Avec les valeurs de J en a, c, d, e, b (celles en a, d, b étant déjà connues de l'itération précédente), on a $x_* \in [d, b]$ si $J(d) > J(e)$, $x_* \in [a, d]$ si $J(c) \leq J(d) \leq J(e)$ et $x_* \in [c, e]$ sinon : on prendra $[a_+, b_+] = [d, b], [a, d]$ et $[c, e]$ resp.. À l'itération k, intervalle $[a_k, b_k]$ sera de longueur $2^{-k}(b_0 - a_0)$, avec $N = 3+2k$ évaluations de la fonction J : le fait d'avoir construit des sous-intervalles

avec les milieux induit la nécessité de calculs de 2 valeurs à chaque itération. L'erreur en fonction du nombre N d'opérations est donc $\varepsilon_N \approx \theta^N$ avec $\theta = (1/2)^{1/2} \approx 0.70711$.

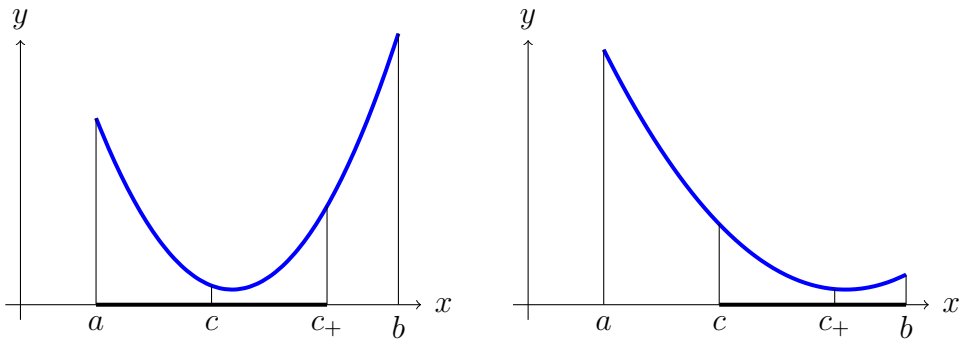


FIGURE II.3 . Dissection dorée, avec $[a_+, b_+] = [a, c_+]$ et $[c, b]$: $\varphi^{-1} = \frac{c_+ - a}{b-a} = \frac{c_+ - c}{b-c} = \frac{c-a}{c_+-a}$.

2.4. Trisection dorée. Cette méthode, raffinant celle de trisection uniforme, construit une suite d'intervalles $([a_k, b_k])$, la donnée de chacun de ces intervalles étant complétée par un $c_k \in [a_k, b_k]$ tel que $J(c_k) \leq \min(J(a_k), J(b_k))$ et convenablement placé relativement aux extrémités a_k, b_k afin d'avoir un taux θ meilleur que celui pour la méthode de trisection uniforme ci-dessus. Supposons donné donc un intervalle $[a, b]$ avec un point c intérieur (*i. e.* un triplet $t = (a < c < b)$) vérifiant $J(c) \leq \min(J(a), J(b))$ et expliquons la construction du triplet $(a_+ < c_+ < b_+)$ successeur. L'intervalle $[a, b]$ contient le minimum de la fonction unimodale J : cette localisation est améliorée par le choix d'un c_+ dans l'intervalle $[c, b]$: on a $J(c) \leq J(c_+)$ ou $J(c_+) \leq J(c)$, soit $x_* \in [a, c_+]$ ou $x_* \in [c, b]$ resp. ; on choisira donc comme triplet $t_+ = (a < c < c_+)$ ou $t_+ = (c < c_+ < b)$ resp.

Algorithme 2.2 L'algorithme de la section dorée

$a_0 = a$; $b_0 = b$
$\varphi = (1 + \sqrt{5})/2$; $k = 1$
tant que $k < K$ & $b_k - a_k > tol$ **faire**
  $c_+ = a_k + \varphi^{-2}(b_k - a_k)$
  $c = a_k + \varphi^{-1}(b_k - a_k)$
  si $J(c) < J(c_+)$ **alors**
    $a_{k+1} = a_k$; $b_{k+1} = c_{+k}$
  sinon
    si $J(c) > J(c_+)$ **alors**
      $a_{k+1} = c$; $b_{k+1} = b_k$
    sinon
      si $J(c) = J(c_+)$ **alors**
        $a_{k+1} = c$; $b_{k+1} = c_+$
      fin si
    fin si
  fin si
  $k = k + 1$
fin tant que

Quelle est une bonne stratégie pour le choix itératif c_k afin d'espérer une décroissance maximale de la longueur de l'intervalle $[a_k, b_k]$ avec son triplet associé $t_k = (a_k < c_k < b_k)$?

D'une part, si α est la longueur relative de $[a, c]$ dans $[a, b]$ (*i. e.* $\alpha = (c-a)/(b-a)$) et α_+ celle de $[c, c_+]$ dans $[a, b]$, l'intervalle déterminant t_+ a sa longueur en proportion $\alpha + \alpha_+$ ou $1 - \alpha$ avec celle de $[a, b]$: on choisira c_+ avec une égalité de ces proportions, soit $\alpha_+ = 1 - 2\alpha$.

D'autre part, le choix de c_+ optimum présuppose que celui de c l'ait été à l'itération précédente : cette invariance d'échelle signifie que la proportion $\alpha_+/(1-\alpha)$ plaçant c_+ dans l'intervalle $[c, b]$ soit la même que la proportion α qui a été utilisée pour placer c dans $[a, b]$, autrement dit $\alpha_+/(1-\alpha) = \alpha$. Avec l'équation précédente $\alpha_+ = 1 - 2\alpha$, on obtient donc pour α l'équation $\alpha^2 - 3\alpha + 1 = 0$ avec sa solution $\alpha_\infty \in [0, 1]$ donnée par $\alpha_\infty = (3 - \sqrt{5})/2 \approx 0.38197$: $1 - \alpha_\infty = \varphi^{-1}$ où φ est le nombre d'or[5]. La suite des intervalles $[a_k, b_k]$ a ses longueurs décroissantes suivant $\varphi^{-k}(b_0 - a_0)$: la convergence vers le minimum est linéaire. À chaque itération, il y a une évaluation de la fonction J : l'erreur en fonction du nombre N d'opérations est donc $\varepsilon_N \simeq \varphi^{-N}$ avec $\varphi^{-1} \approx 0.61803$, convergence meilleure que pour les méthodes de dissection précédentes !

2.5. Autres configurations. Les méthodes de dissection ne donnent que des minima locaux pour des fonctions non unimodales (par ex. $x \in [0, 10] \mapsto 2\cos(x) + \sin(5x)$). La localisation de minima globaux reste un problème à aborder d'autre manière (cette limitation se retrouve dans d'autres méthodes).

En dimension supérieure, la méthode du simplexe mouvant[6], due à Nelder et Mead, permettent des optimisations sans faire appel à des gradients, *i. e.* pour des fonctions supposées continues seulement. La méthode consiste, par des contractions et expansions adéquates, à faire évoluer un simplexe vers le minimum.

3. La méthode de Newton-Raphson

Étant donnée la fonction J définie et dérivable sur un ouvert V, un minimum $x_* \in V$ annule le gradient ∇J. La recherche du minimum peut être initiée donc par la résolution de l'équation $\nabla J(x) = 0$: il faudra prouver que le point critique détecté est un point de minimum, des conditions du second ordre assurant parfois un minimum local (si J est convexe, on verra qu'un tel minimum local est automatiquement un minimum global). Si $F = \nabla J$, $F'(x) = \operatorname{Hess} J(x)$; de manière générale, les zéros de F sont des minima du potentiel Φ avec $\Phi(x) = \|F(x)\|^2/2$ (problème de moindres carrés non linéaires).

La méthode de Newton résout l'équation $F(x) = 0$ où F de classe $\mathcal{C}^1$ est définie sur un ouvert de $E = \mathbb{R}^N$ (avec une preuve tout autant valable pour E Banach) souvent $\mathbb{R}^N$) et à valeurs dans E, avec sa dérivée $F'(x)$ inversible. Considérons le modèle linéaire de F donné par la formule de Taylor et centré en un point x voisin du minimum x_* :

$$F_\ell(y) = F(x) + F'(x)(y - x)$$

soit pour $y = x_*$

$$x_* = x - F'(x)^{-1} F(x) \tag{23}$$

vu $F(x_*) = 0$ et où on a supposé $F'(x)$ inversible.

5. Le nombre $\varphi = 2 - \alpha_\infty = (1 + \sqrt{5})/2$ est appelé *nombre d'or*, ce qui justifie le nom de *section dorée* pour cette méthode de recherche d'optimum. Le nombre d'or est la proportion $\varphi = a/b$ entre les longueurs a et b telle qu'elle le soit aussi entre $a + b$ et a : $a/b = (a + b)/a$ impose φ comme racine positive de l'équation $x^2 = x + 1$.

6. Un simplexe en dimension 1 est un intervalle !

Par ailleurs, pour le programme de minimisation de J, on peut aussi considérer le modèle quadratique J_q de J centré en un point x (voisin de l'extremum x_*) et donné par le développement de Taylor de J à l'ordre 2

$$J_q(y) = J(x) + \langle \nabla J(x), y - x\rangle + \frac{\text{Hess}\, J(x)[y-x]}{2}$$

et dont le point de minimum est $y_* = x - (\text{Hess}\, J(x))^{-1}\nabla J(x)$. L'hypothèse naturelle dans ce cas de minimisation de J est $\text{Hess}\, J_{x_*}$ définie positive : la détermination newtonienne de points critiques de J impose seulement $\text{Hess}\, J_{x_*}$ inversible (la matrice $F'(x)$ n'est pas en général symétrique).

Ces deux arguments amènent à poser le processus itératif de Newton, pour l'approximation du zéro de F ou celle d'un extremum de J avec $F = \nabla J$ suivant

$$x_{k+1} = x_k - F'(x_k)^{-1}F(x_k)$$

qui converge vers x_* lorsque x_0 est assez proche du zéro x_* de F, comme le théorème suivant l'énonce : il y a convergence parfois pour des x_0 plus éloignés. Le test d'arrêt pourra être par exemple un nombre maximum d'itérations, $\|F(x_k)\| \le \varepsilon$ pour une racine de $F = 0$ et $\|\nabla J(x_k)\| \le \varepsilon$ ou $|J(x_k) - J(x_{k+1})| \le \varepsilon$ pour un extremum de J.

Algorithme 2.3 (algorithme de Newton-Raphson pour $F = 0$)

1: Choisir x_0 ; $k = 0$
2: **tant que** ! test **faire**
3: $\quad x_{k+1} = x_k - F'(x_k)^{-1}F(x_k)$
4: $\quad k = k + 1$
5: **fin tant que**

THÉORÈME 2.3: *Soit $E = \mathbb{R}^N$ et $X_* \in E$ un zéro de $F : U(\subset E) \to E$ supposée de classe $\mathcal{C}^2$ avec $F'(X_*)$ inversible. Il existe ρ_0, A, B telle que si $\|X_0 - X_*\| \le \rho_0$, alors la suite $\mathbf{X} = (X_k)_{k\ge 0}$ de point initial X_0 et telle que $X_{k+1} = X_k - F'(X_k)^{-1}F(X_k)$ converge vers X_* avec convergence quadratique,* i. e. *telle que. $\|X_k - X_*\| \le B(A\|X_0 - X_*\|)^{2^k}$.*

DÉMONSTRATION. Soit T la transformation définie par $T(X) = X - F'(X)^{-1}F(X)$ de dérivée $T'(X) = -(F'(X)^{-1})'F(X)$ nulle en X_* : il existe $\rho > 0$ tel que la transformation T est bien définie sur une boule $\overline{B}(X_*, \rho)$ où F' est inversible en raison de l'inversibilité de $F'(X_*)$ et la continuité de F'. Soit $\sqrt{A}$ un majorant de $\max(\|(F'(Z)^{-1})'\|, \|F'\|)$ sur cette même boule. Pour X dans la boule $\overline{B}(X_*, \rho)$, l'inégalité des accroissements finis induit les majorations

$$\begin{aligned}\|T(X) - X_*\| = \|T(X) - T(X_*)\| &\le \sup_{Z\in[X_*,X]} \|T'(Z)\|\|X - X_*\| \\ &\le \sup_{Z\in[X_*,X]} \|(F'(Z)^{-1})'\| \sup_{Z\in[X_*,X]} \|F(Z)\|\|X - X_*\|\end{aligned}$$

et

$$\|F(Z)\| = \|F(Z) - F(X_*)\| \le \sup_{Y\in[X_*,X]} \|F'(Y)\|\|Z - X_*\|, \quad Z \in [X_*, X]$$

d'où résulte

$$\|T(X) - X_*\| \le A\|X - X_*\|^2$$

Ainsi l'application T envoie la boule $\overline{B}(X_*, \rho_0)$ dans elle-même dès que $A\rho_0 \le 1$: elle y est hypercontractante si $A\rho_0 < 1$ vu

$$A\|X_k - X_*\| = A\|T(X_{k-1}) - X_*\| \le (A\|X_{k-1} - X_*\|)^2 \le (A\|X_0 - X_*\|)^{2^k}$$

et la suite $(X_k)_{k\geq 0}$ converge vers X_* dès que son initialisation X_0 appartient à la boule $B(X_*, \rho_0)$ avec $\rho_0 < \min(\rho, 1/A)$. La convergence est quadratique. □

La convergence de la suite de Newton $(T^k(x_0))$ n'est assurée que si x_0 est proche du zéro x_* : le terme initial x_0 sera choisi après une méthode par dichotomie ou par une recherche stochastique, laissant espérer une assez bonne approximation d'un zéro de F, bon point de démarrage d'une suite de Newton convergente.

▷ EXEMPLES 2.4:

2.4.1 L'itération de Héron[7] $x_{n+1} = \frac{1}{2}(A/x_n + x_n)$ approchant $\sqrt{A}$ est la suite de Newton-Raphson associée à la fonction $f(x) = x^2 - A$. Si on effectue le changement de variable $z = (x - \sqrt{A})/(x + \sqrt{A})$ (x dans $\mathbb{R}$ ou $\mathbb{C}$), la suite (x_n) devient la suite (z_n) telle que $z_{n+1} = z_n^2$ qui converge quadratiquement vers 0 si $|z_0| < 1$ (et vers ∞ si $|z_0| > 1$: si z_O est de module 1, la suite $\mathbf{z}$ peut être périodique ou pas) : il y a convergence vers racine carrée $A_{1/2}$ de A (réel ou complexe) si x_0 est dans le demi-plan médiateur de $\sqrt{A}$ et $-\sqrt{A}$ contenant $A_{1/2}$.

2.4.2 Soit F définie sur $\mathbb{R}$ par $F(x) = -x^5 + x^3 + 4x$: cette application a 3 zéros $z_0 = 0$ et $z_\pm = \pm\sqrt{(1+\sqrt{17})/2}$. La suite itérée de Newton $x_{k+1} = x_k - (-5x_k^4 + 3x_k^2 + 4)^{-1}(-x_k^5 + x_k^3 + 4x_k)$ démarrant en $x_0 = 1$ est périodique $x_{2k} = 1$ et $x_{2k+1} = -1$ pour $k \geq 0$.

2.4.3 Avec la condition initiale $x_0 = 1.1$, l'algorithme de Newton-Raphson converge vers le minimum $x_* = 0$ de $\operatorname{ch} x$, mais il y a divergence pour la fonction $\log \operatorname{ch} x$, comme l'indique le tableau 2.4. Les itérations successives sont de la forme $x_{k+1} = x_k - \operatorname{th} x_k$ et $x_{k+1} = x_k - \operatorname{th} x_k/(1 - \operatorname{th}^2 x_k)$ resp.

k	$\log \operatorname{ch} x_k$	$\operatorname{ch} x_k$
0	1.1	1.1
1	-1.128553	0.299501
2	1.234131	0.008645105
3	-1.695166	2.153658e-07
4	5.71536	3.335192e-21
5	-23021.36	0

TABLE 1. Les itérés de Newton pour les fonctions $\log \operatorname{ch} x$ et $\operatorname{ch} x$.

2.4.4 Soit J définie par $J(m) = (x-2)^2[(x-2)^2 + y^2] + (y+1)^2$ avec minimum en $m_* = (2, -1)$. La suite des itérés de Newton-Raphson avec point initial $m_0 = (1, 1)$ exhibe bien le doublement de précision à chaque itéré (cf. tableau 2.4).

2.4.5 La fonction $F(x, y) = (x^3 - 3xy^2 - 1, 3x^2y - y^3$ a comme racines $(-1/2, \pm\sqrt{3}/3), (-1, 0)$. Le point $(-0.6, 0.6)$ donne une itération convergente vers la racine $(-0.5, \sqrt{3}/2)$. Le partitionnement du plan en trois bassins d'attraction correspondant à chaque racine donne des figures fractales. ◁

L'intérêt de la méthode de Newton est sa convergence quadratique si X_0 est assez proche du point fixe (ou extremum de J si $F = \nabla J$) X_*. Néanmoins, pour des

7. Héron d'Alexandrie, 1er siècle apr. J.-C. Le calcul de racine carrée $\sqrt{A}$ intervient dans la formule de Héron donnant l'aire $\mathcal{A}_T$ d'un triangle T en fonction des longueurs a, b, c des côtés : $\mathcal{A}_T = \sqrt{p(p-a)(p-b)(p-c)}$ où p est le demi-périmètre $p = (a+b+c)/2$.

k	x	y	$J(x,y)$
0	1	1	6
1	1	-0.5	1.5
2	1.3913043478260869179	-0.69565217391304345895	0.40920737132871887187
3	1.7459441207973582788	-0.94879809419466143439	0.064891623476885137989
4	1.9862783399912478099	-1.0482080865937803971	0.0025309302111358828943
5	1.9987342021435530182	-1.0001699932311651775	1.63168926685733965e-06
6	1.9999995656676057276	-1.0000016016866475344	2.7540453490388559 6e-12
7	1.9999999999986086685	-1.0000000000001887379	1.9714253277290989944e-24
8	2	-1	0

TABLE 2. Les itérations convergentes de l'algorithme de Newton pour la fonction $J(x,y) = (x-2)^2((x-2)^2+y^2)+(y+1)^2$ avec point de démarrage $m = (1,1)$.

problèmes conséquents, le calcul de l'inverse de la hessienne est très coûteux, voire impossible. Les algorithmes de quasi-Newton (Davidon-Fletcher-Powell [DFP] et Broyden-Fletcher-Goldfarb-Shanno [BFGS], cf. Section 6) calculent donc une approximation de la hessienne à partir des gradients. Ces méthodes dites de quasi-Newton reposent sur une version voisine du théorème 2.3.

THÉORÈME 2.4: *Soit $X_* \in \mathbb{R}^N$ un zéro d'une application $F : U(\subset \mathbb{R}^N) \to \mathbb{R}^N$ où U est un voisinage ouvert de X_*. Supposons F de classe $\mathcal{C}^1$ et $F'(X_*)$ inversible. Il existe $\varepsilon > 0$ tel que si $\|X_0 - X_*\| \leq \varepsilon$ la suite $(X_k)_{k\geq 0}$ vérifiant*

$$X_{k+1} = X_k - M_k^{-1}F(X_k), \quad \|M_k - F'(X_k)\| \leq \varepsilon, \quad k \geq 0$$

est bien définie et converge linéairement vers X_*.

Si $(M_k - F'(X_*))(X_k - X_*) = o(\|X_k - X_*\|)$ (resp. $O(\|X_k - X_*\|^2)$ et F' est de classe $\mathcal{C}^1$), la convergence est superlinéaire (resp. quadratique).

D'autres méthodes existent avec une convergence plus forte que celle de Newton, telle la méthode de Halley utilisée pour approcher des solutions d'équations $f(x) = 0$ en une variable : c'est une méthode itérative définie par

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2f'(x_n)^2 - f(x_n)f''(x_n)} = x_n - \frac{f(x_n)}{f'(x_n)}\left[1 - \frac{f(x_n)}{f'(x_n)} \cdot \frac{f''(x_n)}{2f'(x_n)}\right]^{-1} \tag{24}$$

avec une fonction d'itération F définie par $F(x) = x - \frac{2f(x)f'(x)}{2f'(x)^2 - f(x)f''(x)}$ ayant x_* comme point fixe. On remarque que la méthode de Halley provient de la méthode itérative de Newton-Raphson pour la fonction $g = f/\sqrt{f'}$, avec la condition de croissance stricte pour f au voisinage de son zéro x_* (sinon prendre $g_- = -f/\sqrt{-f'}$).

PROPOSITION 2.1: *Soit f de classe $\mathcal{C}^3$ et x_* une racine simple de f. La méthode itérative de Halley (24) converge si x_0 est suffisamment proche de la racine x_*, cette convergence étant cubique.*

DÉMONSTRATION. On écrit la formule de Lagrange deux fois

$$0 = f(x_*) = f(x_n) + f'(x_n)(x_* - x_n) + \frac{f''(x_n)}{2}(x_* - x_n)^2 + \frac{f'''(\xi_n)}{6}(x_* - x_n)^3$$

$$0 = f(x_*) = f(x_n) + f'(x_n)(x_* - x_n) + \frac{f''(\eta_n)}{2}(x_* - x_n)^2$$

où ξ_n, η_n sont entre x_* et x_n. En multipliant la première égalité par $2f'(x_n)$ et la seconde par $f''(x_n)(x_* - x_n)$, puis en les retranchant, on obtient

$$\begin{aligned}0 =&2f(x_n)f'(x_n) + 2f'(x_n)^2(x_* - x_n) + f'(x_n)f''(x_n)(x_* - x_n)^2 + \frac{f'(x_n)f'''(\xi_n)^3}{3}(x_* - x_n)^3\\ &- f(x_n)f''(x_n)(x_* - x_n) - f'(x_n)f''(x_n)(x_* - x_n)^2 - \frac{f''(x_n)f''(\xi_n)}{2}(x_* - x_n)^3\end{aligned}$$

ce qui, avec les termes $f'(x_n)f''(x_n)(x_* - x_n)^2$ s'annulant, donne après regroupements

$$\begin{aligned}0 = &2f(x_n)f'(x_n) + (2f'(x_n)^2 - f(x_n)f''(x_n))(x_* - x_n)\\ &+ \left[\frac{f'(x_n)f'''(\xi_n)}{3} - \frac{f''(x_n)f''(\eta_n)}{2}\right](x_* - x_n)^3\end{aligned}$$

En mettant le deuxième terme dans le membre de droite, et en divisant par $2f'(x_n)^2 - f(x_n)f''(x_n)$, on obtient

$$x_* - x_n = \frac{-2f(x_n)f'(x_n)}{2f(x_n)^2 - f(x_n)f''(x_n)} - \frac{2f'(x_n)f'''(\xi_n) - 3f''(x_n)f''(\eta_n)}{6(2f'(x_n)^2 - f(x_n)f'''(x_n))}(x_* - x_n)^3$$

soit

$$x_* - x_{n+1} = \frac{2f'(x_n)f'''(\xi_n) - 3f''(x_n)f'(\eta_n)}{12f'(x_n)^2 - 6f(x_n)f''(x_n)}(x_* - x_n)^3 \tag{25}$$

La méthode itérative de Halley pour la fonction f étant celle de Newton pour la fonction $f/\sqrt{f'}$, on sait que la suite (x_n) converge vers le zéro x_* de f. Ainsi le facteur du second membre dans (25) approche $K_* = (2f'''(x_*) - 3f''(x_*))/12f'(x_*)$. Soit K_0 un majorant de la valeur absolue de ce facteur sur un voisinage B_0 de x_* et n_0 assez grand tel que $x_n \in B_*$ si $n \geq n_0$. Ainsi pour $n \geq n_0$, on a la majoration $|x_{n+1} - x_*| \leq K|x_n - x_*|^3$ ou encore $\sqrt{K}|x_{n+1} - x_*| \leq [\sqrt{K}|x_n - x_*|]^3$ et donc

$$|x_{n+1} - x_*| \leq K^{-1/2}\left[\sqrt{K}|x_{n_1} - x_*|\right]^{3^{n+1-n_1}} \leq K^{-1/2}\rho_1^{3^{n+1-n_1}} = (\sqrt{K}\rho_1^{3^{n_1}})^{-1}\rho_1^{3^{n+1}}$$

soit une convergence cubique en ayant choisi n_1 assez grand tel que $\rho_1 = \sup_{n\geq n_1}[\sqrt{K}|x_n - x_*|] < 1$. □

4. Méthodes de descente

Tournant le dos à la résolution de $\nabla J(x) = 0$ traitée dans la section précédente, on présente deux exemples de méthodes de *descente* vers le minimum. Au contraire de la méthode de Newton qui imposait l'usage de la hessienne, ces méthodes requièrent une dérivabilité d'ordre 1 pour leur énoncé (même si l'existence de dérivées d'ordre 2 permet d'établir une convergence linéaire). Ces méthodes itératives sont basées sur le choix à l'étape k d'une direction d_k et d'un scalaire t_k tels que $J(x_k) \geq J(x_k + t_k d_k)$: on descend du point x_k au point x_{k+1} en diminuant (en général strictement !) les valeurs de la fonction J. On choisira un test d'arrêt convenable comme précédemment pour l'algorithme de Newton, par exemple $\|\nabla J(x_k)\| \leq \varepsilon$ ou $J(x_k) - J(x_{k+1}) \leq \varepsilon$. Ainsi, les méthodes de descente construisent itérativement la suite $(x_k)_{k\geq 0}$ suivant l'algorithme 2.4.

Une direction d est dite *de descente* au point x si la fonction $t \in [0, +\infty) \mapsto J(x + td)$ est décroissante à partir de $t = 0$: ainsi sa dérivée à droite en $t = 0$ doit être négative, soit $\langle \nabla J(x), d\rangle \leq 0$.

Algorithme 2.4 : algorithme de descente

Choisir x_0 ; $k = 0$
tant que ! test **faire**
  Choisir une direction de descente d_k
  Choisir le pas $t_k = RDD(J, x_k + \mathbb{R}_+ d_k)$
  Définir $x_{k+1} = x_k + t_k d_k$
  $k = k + 1$
fin tant que

LEMME 2.1: *(i) La direction $-\nabla J(x)$ est une direction de descente.*

(ii) Si la hessienne $\operatorname{Hess} J(x_*)$ *est inversible, la direction* $-(\operatorname{Hess} J(x))^{-1}\nabla J(x)$ *utilisée dans l'algorithme de Newton est une direction de descente au voisinage du minimum* x_*.

DÉMONSTRATION. Pour le (i), on a bien $-\langle \nabla J(x), \nabla J(x)\rangle \leq 0$. Pour le (ii), x_* étant un minimum et la hessienne inversible, la hessienne $\operatorname{Hess} J(x_*)$ est définie positive au voisinage de x_* et par continuité sur un voisinage de x_* : il en est de même de son inverse $(\operatorname{Hess} J(x))^{-1}$ et par suite $-\langle (\operatorname{Hess} J(x))^{-1}\nabla J(x), \nabla J(x)\rangle \leq 0$ au voisinage de x_*. □

▷ EXEMPLES 2.5:

2.5.1 La hessienne de la fonction $U(x, y) = x^4 + y^4 - 4xy$ a pour points critiques $(0, 0), \pm(1, 1)$ et hessienne $\operatorname{Hess} U(x, y) = 4\begin{pmatrix} 3x^2 & -1 \\ -1 & 3y^2 \end{pmatrix}$, définie positive en $(1, 1)$. La Hessienne en $(0.5, 0.5)$ est inversible, ses directions de Newton n'étant pas de descente.

2.5.2 La fonction $U(x, y) = 1000(x^3 - xy^2)^2 + (x^3 + xy)^2 + y^6$ admet $m_* = (0, 0)$ comme minimum local strict. Sa hessienne est inversible le long de l'axe $y = 0$ sauf en l'origine m_*, mais la direction de Newton n'est jamais une direction de descente. ◁

On en déduit l'algorithme 2.5 de descente dans la direction du gradient à pas optimal

Algorithme 2.5 : algorithme de descente dans la direction du gradient à pas optimal

Choisir x_0 ; $k = 0$;
tant que ! test **faire**
  $d_k = -\nabla J(x_k)$
  $t_k = \operatorname{argmin}_{t>0}(J(x_k + t d_k))$
  $x_{k+1} = x_k + t_k d_k$
  $k = k + 1$
fin tant que

Introduisons les notions de fonction *coercive* et de fonction *elliptique* :

DÉFINITION 2.3: *(i) La fonction J définie sur E est dite* coercive *si $J(x) \to +\infty$ lorsque $\|x\| \to +\infty$. De manière générale*[8]*, J définie sur un ouvert Ω est dite coercive si $J(x) \to +\infty$ lorsque $x \to \partial\Omega$ ou $x \to \partial\Omega \cup \{\infty\}$ si Ω non borné.*

(ii) Soit $m > 0$. La fonction J définie sur E et différentiable est dite m-elliptique *si*

$$\langle \nabla J(x) - \nabla J(y), x - y\rangle \geq m\|x - y\|^2, \quad x, y \in E \tag{26}$$

8. Le cas où Ω est l'orthant positif $\mathbb{O}_n = \mathbb{R}^n_{+*}$ est d'intérêt.

△ REMARQUE 2.2: La coercivité de J s'exprime ainsi : pour tout M, il existe un compact $K_M \subset \Omega$ tel que $J(x) \geq M$ pour x hors de K_M. Ainsi une fonction coercive sur Ω continue y est minorée et atteint son minimum : en effet, étant donné $x_0 \in \Omega$, si x_* est un minimum (qui existe d'après le théorème de Weierstrass sur les fonctions continues) de J sur le compact $K_{J(x_0)+1}$, on a $J(x_*)$ minorant $J(x)$ pour x dans $K_{J(x_0)+1}$ et aussi sur son complémentaire où $J(x) \geq J(x_0)+1 \geq J(x_*)$ vu que $x_0 \in K_{J(x_0)+1}$. ▽

▷ EXEMPLES 2.6:

2.6.1 La fonction $m \in \{m \in \mathbb{R}^n, \|m\|_2 < 1\} \to -\log(1-\|m\|_2^2)$ est coercive.

2.6.2 Soit Σ_n le simplexe

$$\Sigma_n = \{(x_0,\dots,x_n), \sum_{k=0}^{n} x_k = 1, x_0 > 0, \dots, x_n > 0\}.$$

La fonction $J_n : x \in \Sigma_n \mapsto -\sum_{k=0}^n \log x_k$ est coercive : pour $\varepsilon > 0$, la partie $K_\varepsilon = \{x \in \Sigma_n, x_0 \geq \varepsilon, \dots, x_n \geq \varepsilon\}$ est compacte et pour $x \notin K_\varepsilon$, $J_n(x) \geq \log(\varepsilon^{-1})$. On montre pareillement que la fonction $L_n : x \in \mathbb{O}_n \mapsto \sum_{k=1}^n x_k - \log\prod_{k=1}^n x_k$ est coercive sur l'orthant $\mathbb{O}_n = \mathbb{R}^n_{+*}$ en considérant par exemple la famille de compacts $K_\varepsilon = \{\varepsilon \leq x_k \leq 1+\varepsilon^{-1}, k = 1,\dots,n\}$ et en remarquant que $L_1(t) \geq 0$ pour $t \in \mathbb{O}_1$ et $L_n(x) \geq \min[\log(\varepsilon^{-1}), 1+\varepsilon^{-1} - \log(1+\varepsilon^{-1})]$ pour $x \in \mathbb{O}_n \setminus K_\varepsilon$.

2.6.3 Le problème de la boîte (22) est coercif, vu les inégalités $2x + 1/xy \geq \sqrt{y}$ et $y + 1/xy \geq \sqrt{x}$. ◁

Le choix du pas t_k est évidement crucial : outre le pas fixe (voire contraint comme dans l'algorithme 2.7) ou le pas optimal (cf. algorithme 2.5), il y a d'autres choix possibles, moins gourmands en calcul, mais garantissant cependant des convergences. Nous noterons le choix d'une de ces méthodes de *recherche sur une demi-droite* par $RDD(J, x_k + \mathbb{R}_+ d_k)$. Par exemple il y a la méthode par *retour en arrière* qui est souvent préférée et donne lieu à des résultats de convergence (globale pour des fonctions convexes) similaires à ceux des méthodes de pas optimal. Nous développons un peu ces propriétés dans la sous-section suivante, avant d'introduire d'autres algorithmes de descente. Cette mé-

Algorithme 2.6 : recherche linéaire sur la demi-droite $x + \mathbb{R}_+ d$

1: Choisir $\alpha \in (0.1, 0.99)$, $\tau \in (0,1)$; $t = 1$
2: **tant que** $J(x+td) > J(x) + \alpha t \langle \nabla J(x), d\rangle$ **faire**
3: $\quad t = \tau t$
4: **fin tant que**

thode de recherche linéaire est appelée ainsi car elle réduit le pas initial $t = 1$ d'un facteur $\tau, \tau^2, \dots$ jusqu'à ce que la condition $J(x+td) \leq J(x) - ct\langle -\nabla J(x), d\rangle$ soit remplie, condition qui, vu le choix de $c \in (0,1)$, est atteinte pour t petit $J(x+td) \approx J(x) - t\langle -\nabla J(x), d\rangle >$ et d est une direction de descente (*i. e.* $\langle -\nabla J(x), d\rangle \geq 0$).

4.1. Recherche linéaire. L'obtention du minimum exact d'une fonction restreinte sur la demi-droite de descente $x + \mathbb{R}_+ d$ est rarement possible : on se limite à trouver un point $x + \alpha d$ de décroissance suffisante pour J, *i. e.* $J(x+\alpha d) < J(x) - \varepsilon$, afin que la méthode itérative de descente converge.

▷ EXEMPLE 2.7: Pour la fonction $J(x) = x^2$, considérons des suites numériques $(d_k)_{k\geq 0}$, $(\alpha_k)_{k\geq 0}$ induisant la suite de descente d'itérés $(x_k)_{k\geq 0}$ vérifiant $x_{k+1} = x_k + \alpha_k d_k$. Si $d_k = (-1)^k$, $\alpha_k = 2 + 3 * 2^{-k-1}$ et $x_0 = 2$, alors $x_k = (-1)^k + (-2)^{-k}$ oscille asymptotiquement entre -1 et 1 : le pas de descente est trop grand et la suite $(x_k)_{k\geq 0}$ évite le point de

minimum $x_* = 0$. Pour $d_k = -1$, $\alpha_k = 2^{-k-1}$ et $x_0 > 1$, la suite $(x_{k+1} = x_0 - 1 + 2^{-k-1})_{k\geq 0}$ converge vers $x_0 - 1$ qui n'est pas le minimum si $x_0 \neq 1$: le pas de descente est trop petit. ◁

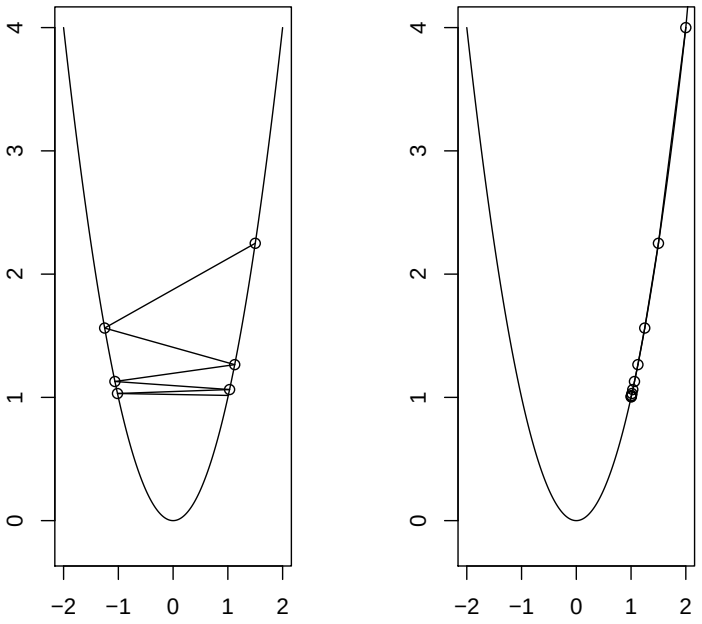


FIGURE II.4 . Des descentes non convergentes vers le point de minimum : à gauche pas $\alpha_k = 2 + 3 * 2^{-k-1}$ (grand) et à droite $\alpha_k = 2^{-k-1}$ (petit).

Dans la direction de descente d à partir de x, il y a des conditions assurant la descente pour un point sur la demi-droite issue de x : celle d'Armijo

$$J(x + \alpha d) \leq J(x) - c_1 \alpha \langle -\nabla J(x), d\rangle$$

et celle de Wolfe

$$\langle -\nabla J(x + \alpha d), d\rangle \leq c_2 \langle -\nabla J(x), d\rangle$$

avec les constantes $0 < c_1 < c_2 < 1$ et α assez petites.

La première condition assure une décroissante suffisante : la constante c_1 est petite, par exemple $c_1 = 10^{-4}$.

LEMME 2.2: *Soit* $\alpha \in (0, 1)$.

Si $\langle d, \nabla U(x)\rangle < 0$, *alors il existe un* t_* *tel que* $U(x + td) < U(x) + \alpha\langle d, \nabla U(x)\rangle$ *pour* $t \in (0, t_*)$.

Si $\alpha < 1/2$, *l'itération*

$$t_0 = 1 \text{ puis } t_{k+1} = \frac{1}{2} \frac{-t_k \langle d, \nabla U(x)\rangle}{U(x + t_k d) - U(x) - t_k \langle d, \nabla U(x)\rangle}$$

tant que $U(x + td) \geq U(x) + \alpha\langle d, \nabla U(x)\rangle$ *retourne un pas vérifiant la règle de Armijo.*

DÉMONSTRATION. Soit $U_{x,d,t}(\alpha) = U(x+td) - U(x) - \alpha\langle d, \nabla U(x)\rangle$, qui vérifie $U_{x,d,t}(0) = 0$ et $\partial_t[U_{\alpha,x,d}(\alpha t)]_{|t=0} = (1 - \alpha)\langle d, \nabla U(x)\rangle < 0$, ainsi pour $t_* > 0$ suffisamment petit, on a $U_{\alpha,x,d}(t) < 0$ pour $t \in (0, t_*)$, ce qui est annoncé.

Pour la deuxième assertion, si $U_{x,d,\alpha}(t) < 0$, ce qui est la condition d'Armijo, on retient le pas $t = 1$ et le point $m + t\nabla U(x)\rangle$. Sinon, on interpole $t \to U(x + td)$ par la

parabole passant par les deux points $(0, U(x))$ et $(t, U(x+td))$ avec pente de la tangente au premier point égale à $\langle d, \nabla U(x)\rangle$:

$$\tau \in \mathbb{R} \mapsto \frac{U_{x,d,1}(t)}{t^2}\tau^2 + \frac{\langle d, \nabla U(x)\rangle}{t}\tau + U(x)$$

et on considère le point de minimum de cette parabole donné par le paramètre

$$t_+ = -\frac{t\langle d, \nabla U(x)\rangle}{2U_{x,d,t}(1)}$$

Si $\langle d, \nabla U(x)\rangle < 0$, cette stratégie de remplacement suivie tant que $U_{x,d,t}(\alpha) \geq 0$, transforme un t en un

$$t_+ = \frac{-t\langle d, \nabla U(x)\rangle}{2U_{x,d,t}(1)} \leq \frac{-t\langle d, \nabla U(x)\rangle}{2[U_{x,d,t}(1) - U_{x,d,t}(\alpha)]} = \frac{t}{2(1-\alpha)}$$

qui appartient à $(0, t/(2(1-\alpha)))$. Pour $\alpha < 1/2$, et donc $1/(2(1-\alpha)) < 1$, on réduit ainsi la taille de l'intervalle $(0, t)$, jusqu'à ce que $U_{x,d,t}(\alpha) < 0$, ce qui est la condition de Armijo, et donc un t tel que $U_{x,d,t}(\alpha) > 0$ avec convergence annoncée vu la première partie du lemme. □

La condition de Wolfe (dite de courbure) est utilisée avec un $c_2 = 0.9$ pour des descentes quasi-Newton, $c_2 = 0.1$ pour des descentes en gradient conjugué non linéaire. Les conditions de Wolfe sont valides pour α assez petit : partant d'un $\alpha_0 > 0$ et fixant $\tau \in (0, 1)$, on itère $\alpha_{k+1} = \tau\alpha_k$ jusqu'à ce que les conditions de Wolfe soient vérifiées pour $x, d, x + \alpha d$. Dans le cas de la descente de Newton-Raphson, le pas $\alpha = 1$ vérifie pour k assez grand les conditions de Wolfe : en général on testera les conditions de Wolfe pour $\alpha_0 = 1$ avant de tester $\tau\alpha, \tau^2\alpha/4, \ldots$.

THÉORÈME 2.5 (Zoutendijk): *Soit $(x_k)_{k\geq 0}$ une suite de descente telle que $x_{k+1} = x_k + \alpha_k d_k$ et vérifiant les conditions de Wolfe. Supposons J continûment différentiable dans un voisinage ouvert N de $J^{-1}((-\infty, J(x_0)])$, avec constante de Lipschitz uniforme pour le gradient ∇J :*

$$\|\nabla J(x) - \nabla J(y)\|_2 \leq L\|x - y\|_2, \quad x, y \in N.$$

Soit $\theta_k \in [-\pi/2, \pi/2]$ l'angle entre la direction de descente d_k et l'opposé du gradient $\nabla J(x_k)$. Alors la série $\sum_{k\geq 0} \cos^2\theta_k \|\nabla J(x_k)\|_2^2$ est convergente.

DÉMONSTRATION. La deuxième condition de Wolfe et la condition de Lipschitz donnent

$$(c_2 - 1)\langle \nabla J(x_k), d_k\rangle \leq \langle \nabla J(x_{k+1}) - \nabla J(x_k), d_k\rangle \leq \alpha_k L\|d_k\|_2^2$$

et par suite

$$\alpha_k \geq \frac{1 - c_2}{L}\frac{\langle -\nabla J(x_k), d_k\rangle}{\|d_k\|_2^2}$$

et, après insertion dans la première condition de Wolfe

$$J(x_{k+1}) \leq J(x_k) - c_1\frac{1 - c_2}{L}\frac{\langle \nabla J(x_k), d_k\rangle^2}{\|d_k\|_2^2}$$

et par suite, avec $c = c_1(1 - c_2)/L$,

$$J(x_{k+1}) \leq J(x_k) - c\cos^2\theta_k\|\nabla J(x_k)\|_2^2$$

et la convergence de la série par sommation télescopique. □

Ainsi si l'angle entre la direction de descente d_k et l'opposé du gradient est restreint dans $[-\pi/2+\delta, \pi/2-\delta]$, alors $\|\nabla J(x_k)\|_2$ tend vers 0, ce qui implique la convergence d'une sous-suite de $(x_k)_{k\geq 0}$ (si la fonction est coercive par exemple) vers un point stationnaire de J, en particulier vers le minimum de J si celui-ci est l'unique point critique. Le théorème de Zoutendijk ne précise en rien le type de convergence de x_k.

Dans le cas d'une fonction J strictement convexe (par ex., s'il existe m, M tels que $0 < m \leq \operatorname{Hess} J(x) \leq M$ pour $x \in E$), il y a convergence linéaire de la méthode de descente à pas maximal et de celle à pas modulé par un retour en arrière [7, 9.3.1].

4.2. Descente à pas fixe. L'algorithme de gradient pour une fonction J m-elliptique décrit dans le théorème suivant est valable dans tout Hilbert E et indique une vitesse de convergence vers le point de minimum.

THÉORÈME 2.6 (Descente dans la direction du gradient (à pas fixe/contrôlé)): *Soit J définie sur l'espace de Hilbert E* [9]*, dérivable et m-elliptique telle qu'il existe une constante $M > 0$ avec*

$$\|\nabla J(v) - \nabla J(w)\| \leq M\|v - w\|, \quad v, w \in E.$$

Alors pour tous a, b tels que $0 < a \leq b < 2m/M^2$ et toute suite $(t_k)_{k\geq 0}$ telle que $t_k \in [a, b]$, la suite vérifiant

$$x_{k+1} = x_k - t_k \nabla J(x_k)$$

converge vers le minimum x_* de J de manière géométrique :

$$\|x_k - x_*\| \leq \beta^k \|x_0 - x_*\|, \quad k \geq 0$$

pour un $\beta \in (0, 1)$ dépendant de a, b, m et M.

Algorithme 2.7 : algorithme à pas contrôlé

```
Choisir x_0 ; k = 0 ;
tant que ! test faire
    d_k = e_{k%%n+1}
    t_k ∈ [a, b]
    x_{k+1} = x_k + t_k d_k
    k = k + 1
fin tant que
```

▷ EXEMPLE 2.8: La recherche du minimum x_* de $\langle x, Ax\rangle/2 - \langle b, x\rangle$, avec A définie positive, est équivalent à la résolution de l'équation linéaire $Ax = b$, soit $x_* = A^{-1}b$. L'itération est donnée par $x_{k+1} = x_k - t_k(Ax_k - b)$ avec

$$x_{k+1} - x_* = (1 - t_k A)(x_k - x_*)$$

Suivant les notations de l'énoncé et si $(\lambda_k(A))$ est le spectre de A, on a $m = \inf_k(\lambda_k(A))$ et $M = \sup_k(\lambda_k(A))$. ◁

9. Dans ce théorème et comme l'indiquent clairement hypothèses et preuve, la norme utilisée est celle déterminée par le produit scalaire définissant la structure hilbertienne de E : en dimension finie, par équivalence des normes, les estimations d'erreur valent pour n'importe quelle norme, alors qu'en dimension infinie on n'a pas de résultat en dehors de E Hilbert au contraire de la méthode de Newton-Raphson valable dans tout Banach.

△ REMARQUE 2.3: Le choix de t_k comme le minimum de $t \in \mathbb{R}^+ \mapsto J(x_k - t\nabla J(x_k))$ correspond à une stratégie de descente la plus raide dans la direction du gradient. Dans ce cas, les segments $[x_{k-1}, x_k]$ et $[x_k, x_{k+1}]$ sont orthogonaux. En effet la dérivée de $t \mapsto J(x_k + t\nabla J(x_k))$ est donnée par

$$\langle \nabla J(x_k + t_k \nabla J(x_k)), \nabla J(x_k)\rangle = \langle \nabla J(x_{k+1}), \nabla J(x_k)\rangle$$

vu que $x_{k+1} = x_k + t_k \nabla J(x_k)$: cette dérivée est nulle en le minimum t_k. Avec ce choix de (t_k), la ligne brisée $\dots x_{k-1}x_k x_{k+1} \dots$ est à angles droits, ce qui laisse penser que d'autres algorithmes que celui du gradient à pas fixe donnent une descente plus rapide vers le minimum. ▽

DÉMONSTRATION. Vu que $\nabla J(x_*) = 0$, on a

$$x_{k+1} - x_* = (x_k - x_*) - t_k(\nabla J(x_k) - \nabla J(x_*)).$$

Par suite, vu $t_k \geq 0$,

$$\begin{aligned}\|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - 2t_k\langle \nabla J(x_k) - \nabla J(x_*), x_k - x_*\rangle + t_k^2 \|\nabla J(x_k) - \nabla J(x_*)\|^2 \\ &\leq (1 - 2mt_k + M^2 t_k^2)\|x_k - x_*\|^2 \leq \beta^2 \|x_k - x_*\|^2\end{aligned}$$

avec $\beta < 1$. En effet, le trinôme

$$T_{m,M}(t) = 1 - 2mt + M^2 t^2 = 1 - (m/M)^2 + (Mt - m/M)^2$$

admet $t_{\min} = m/M^2$ comme point de minimum et vérifie les inégalités

$$T_{m,M}(m/M^2) \leq T_{m,M}(t) < T_{m,M}(0) = T_{m,M}(2m/M^2) = 1$$

pour $t \in [a, b] \subset (0, 2m/M^2)$ et donc $\beta^2 = \sup_{t\in[a,b]} T_{m,M}(t) < 1$. □

Si la régularité est plus faible, et gardant l'hypothèse de convexité, on prouve une convergence plus lente :

PROPOSITION 2.2 ([18, 5.1]): *Soit U une fonction $\mathcal{C}^{1,1}$ convexe, avec gradient M_U-lipschitzien et R_U la distance de l'origine à* $\operatorname{argmin} U$. *Pour $a \in (0, 2/M_U)$, l'itération*

$$x_0 = 0, \quad x_{k+1} = x_k - a\nabla U(x_k), k \geq 0$$

induit la convergence en décroissant de la suite $(U(x_k))$ avec

$$U(x_k) \leq U_* + \frac{M_U R_U^2}{k}.$$

DÉMONSTRATION. Soit $x_* \in \operatorname{argmin} U$ avec $\|x_*\| = R_U$ (par ex. x_* est la projection de l'origine sur le fermé $\operatorname{argmin} U$). On a

$$U(x) + \langle h, \nabla U(x)\rangle \leq U(x + h) \leq U(x) + \langle h, \nabla U(x)\rangle + \frac{M_U}{2}\|h\|^2, \tag{27}$$

avec la première inégalité provenant de la convexité de U et la seconde du caractère lipchitzien du gradient

$$U(x+h) - U(x) - \langle h, \nabla U(x)\rangle = \langle h, \int_0^t [\nabla U(x - th) - \nabla U(x)]dt\rangle = \|h\|^2 M_U \int_0^t t dt = \frac{M_U\|h\|^2}{2}$$

De cette deuxième inégalité résulte

$$U(x_{k+1}) \leq U(x_k) - a\|\nabla U(x_k)\|^2 + \frac{M_U}{2}a^2\|\nabla U(x_k\|^2 = U(x_k) - A\|\nabla U(x_k)\|^2$$

où on a noté $A = a(1 - aM_U/2)$. Par suite

$$\|\nabla U(x_k)\|^2 \leq A^{-1}[U(x_k) - U(x_{k+1})]$$

ce qui assure que cette méthode est de descente et donne la majoration

$$\sum_{k=0}^{K} \|\nabla U(x_k)\|^2 \leq A^{-1}(U(x_0) - U(x_*)) = A^{-1}(U(0) - U(x_*)) \leq \frac{M_U R_U^2}{2A}$$

où on a utilisé à nouveau la seconde inégalité dans (27) avec $x = x_*$ et $h = -x_*$, en remarquant $\nabla U(x_*) = 0$. Ainsi, en utilisant toujours la convexité,

$$\begin{aligned}\|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - 2a\langle x_k - x_*, \nabla U(x_k) + a^2\|\nabla U(x_k)\|^2 \\ &\leq \|x_k - x_*\|^2 - 2a(U(x_k) - U(x_*)) + a^2\|\nabla U(x_k\|^2\end{aligned}$$

et donc, en sommant K inégalités comme ci-dessus

$$\|x_K - x_*\|^2 \leq \|x_0 - x_*\|^2 - 2a\sum_{k=0}^{K-1}[U(x_k) - U(x_*)] + a^2\sum_{k=0}^{K-1}\|\nabla U(x_k\|^2$$

soit

$$\sum_{k=0}^{K-1}[U(x_k) - U(x_*)] \leq \frac{R_U^2}{2a} + aA^{-1}M_U R_U^2/4 = \frac{R_U^2}{a(2 - aM_U)}.$$

Vu que les $U(x_k)$ sont décroissants, on obtient

$$U(x_K) - U(x_*) \leq \frac{R_U^2}{a(2 - aM_U)}K^{-1},$$

où le membre de droite est minimum pour $a \in (0, 2/M_U)$ en $a = 1/M_U$, ce qui achève la preuve. □

4.3. Descente à relaxation. La méthode de relaxation choisit la direction de descente en balayant successivement les directions du repère (orthonormé) canonique $(e_j)_{j=1}^n$:

$$d_1 = e_1, d_2 = e_2, \ldots, d_n = e_n, d_{n+1} = e_1, \ldots, d_{2n} = e_n, d_{2n+1} = d_1, \ldots, d_{3n} = e_n, \ldots.$$

soit $d_k = e_{k\%\%n+1}$ pour $k \leq 1$.

THÉORÈME 2.7 (Relaxation): *Soit J définie sur $E = \mathbb{R}^n$, de classe $\mathcal{C}^2$ et m-elliptique. Soit $(e_i)_{i=1,\ldots,n}$ la base canonique de $\mathbb{R}^n$ et la suite*[10] *$(d_k = e_{k\%\%n+1})$ Alors la suite $(x_k)_{k\geq 0}$ telle que*

$$x_{k+1} = x_k + t_k d_k \text{ avec } t_k = \operatorname{argmin}[t \in \mathbb{R} \mapsto J(x_k + td_k))]$$

converge vers le minimum x_* de J.

△ REMARQUE 2.4: Cette méthode peut-être parallélisée aisément si la fonction J est du type $J(x) = \sum_I V(x_I)$ avec $\{I\}$ une partition de $[\![1, n]\!]$. ▽

Algorithme 2.8 : algorithme de relaxation

1: Choisir x_0 ; $k = 0$;
2: **tant que** ! test **faire**
3: $d_k = e_{k\%\%n+1}$
4: $t_k = \operatorname{argmin}(t \in \mathbb{R} \mapsto J(x_k + td_k)$ (ou min. approché)
5: $x_{k+1} = x_k + t_k d_k$
6: $k = k + 1$
7: **fin tant que**

10. Suivant `R`, le reste de la division euclidienne de k par n est noté $k\%\%n$.

DÉMONSTRATION. Admettons le lemme suivant qui est basé sur des propriétés de convexité examinées dans le chapitre 3.

LEMME 2.3: *(i) Si J est de classe $\mathcal{C}^2$, J est m-elliptique si et seulement si*

$$\text{Hess}\, J(x)(v) \geq m\|v\|^2, \quad x, v \in E. \tag{28}$$

(ii) Si J est elliptique, alors J est strictement convexe et coercive avec

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

DÉMONSTRATION. La première partie provient de la caractérisation de la convexité pour les fonctions 2-fois différentiable dans la proposition 3.2.

Pour la seconde, on applique l'inégalité de convexité avec $m = 0$ à la fonction $x \mapsto J(x) - m\|x$
$^2/2$ (qui est convexe d'après la première partie de ce lemme) : après simplifications on retrouve exactement la formule donnée dans cette seconde partie. □

La fonction J, m-elliptique, est donc coercive et strictement convexe : elle admet donc un unique minimum x_*, qui vérifie de plus $\nabla J(x_*) = 0$. Vu la stricte convexité de J en restriction à la droite $t \in \mathbb{R} \mapsto x_k + td_k \in \mathbb{R}^n$, le réel t_k est défini de manière unique à partir de d_k et x_k et il vérifie

$$0 = \frac{d}{dt}[J(x_k + td_k)]_{|t=t_k} = \langle \nabla J(x_{k+1}), d_k\rangle = \langle \nabla J(x_{k+1}), (x_{k+1} - x_k)/t_k\rangle.$$

Il y a des conditions initiales qui amènent en un nombre fini d'itérations sur le point de minimum : on supposera dans la suite que ce n'est pas le cas, et de manière un peu plus extensive que la suite $(t_k)_{k\geq 0}$ ne contient pas de zéros.

La suite $(J(x_k))_{k\geq 1}$ est décroissante, la suite $(x_k)_{k\geq 1}$ est bornée, on désignera par K un compact contenant la suite $(x_k)_{k\geq 0}$. La formule de Taylor avec reste intégral à l'ordre 2

$$J(x_k) = J(x_{k+1}) + \langle \nabla J(x_{k+1}), x_k - x_{k+1}\rangle + \int_0^1 (1-u)\,\text{Hess}\, J(x_{k+1} + u(x_k - x_{k+1}))[x_{k+1} - x_k]du$$

induit l'inégalité

$$J(x_k) - J(x_{k+1}) \geq \frac{m}{2}\|x_{k+1} - x_k\|^2$$

on en déduit que la suite $\sigma_x^1 = (\|x_{k+1} - x_k\|)$, de même que les suites $\sigma_x^\ell = (\max_{1\leq j\leq \ell} \|x_{k+j} - x_k\|)$, convergent vers 0.

Choisissons l'entier $k_i = n[k/n] + i + 1$ tel que $k_i - k \leq n$ et $d_{k_i} = e_i$. Alors

$$\partial_i J(x_{k_i}) = \langle \nabla J(x_{k_i}), e_i\rangle = \langle \nabla J(x_{k_i}), d_{k_i}\rangle = 0.$$

Alors, d'après l'inégalité d'ellipticité,

$$m\|x_k - x_*\|^2 \leq |\langle \nabla J(x_k) - \nabla J(x_*), x_k - x_*\rangle| \leq \|\nabla J(x_k)\|\|x_k - x_*\|$$

et donc

$$\begin{aligned} m\|x_k - x_*\| \leq \|\nabla J(x_k)\| &= \sqrt{\sum_{i=1}^n \partial_i J(x_k)^2} \\ &\leq \sqrt{\sum_{i=1}^n [\partial_i J(x_k) - \partial_i J(x_{k_i})]^2} \leq n^2 \sup_{\substack{1\leq \ell, m\leq n \\ x\in K}} |\partial^2_{\ell m} J(x)| \sup_{i=1,\dots,n} \|x_k - x_{k_i}\| \end{aligned}$$

Vu la convergence de la suite σ_x^n vers 0 avec (x_k) bornée et l'uniforme continuité locale des dérivées partielles $\partial_{x_i} J$, la convergence $x_k \to x_*$ en résulte. □

4.4. Gradient conjugué. La méthode de type gradient-conjugué[11] est une méthode de descente qui conclut la recherche de minimum de la fonction quadratique $J_{A,b}(x) = \langle Ax, x\rangle/2 - \langle b, x\rangle$ sur $\mathbb{R}^n$ en au plus n itérations : cette prévision théorique est à amender sérieusement pour toute réalisation du programme informatique, où l'accumulation d'erreurs (déjà incluses dans la discrétisation apportée par l'ordinateur) empêche *de facto* cette convergence. Cette méthode, baignée de géométrie, repose sur le choix de directions de descente successives A-conjuguées, *i. e.* orthogonales deux à deux pour la forme bilinéaire symétrique $\langle Ax, y\rangle$

DÉFINITION 2.4: *Soit S matrice symétrique d'ordre n. Deux vecteurs (non nuls) u et v de $\mathbb{R}^n$ sont dits* S-conjugués *si $\langle Su, v\rangle = 0$. Une famille $\mathcal{F} = (u_i)_{i\in I}$ de vecteurs non nuls est dite S-conjuguée si les vecteurs de toute paire de vecteurs distincts de $\mathcal{F}$ sont S-conjugués.*

Dans la suite on considérera la relation de A-conjugaison pour une matrice A symétrique définie positive : la forme bilinéaire symétrique $(u, v) \mapsto \langle Au, v\rangle$ induit un produit scalaire et une norme notés $\langle u, v\rangle_A$ et $\|u\|_A = \sqrt{\langle Au, u\rangle}$ resp. La relation de A-conjugaison coïncide avec celle de A-orthogonalité, *i. e.* d'orthogonalité relativement au produit scalaire $\langle\ ,\ \rangle_A$. Pour un opérateur A symétrique, il existe des bases dont les vecteurs sont deux à deux A-conjugués, par ex. en considérant une base constituée de vecteurs propres orthogonaux.

PROPOSITION 2.3: *Soit A symétrique définie positive, $J_{A,b}$ définie sur $\mathbb{R}^n$ par $J_{A,b}(x) = \langle Ax, x\rangle/2 - \langle b, x\rangle$ et $(d_i)_{0\le i<n}$ une base de directions A-conjuguées.*

Alors la suite $(x_k)_{0\le k<n}$ vérifiant

$$x_{k+1} = x_k + \alpha_{k+1} d_k \text{ avec } \alpha_{k+1} = \operatorname{argmin}_t(J_{A,b}(x_k + t d_k))$$

converge vers l'unique minimum $x_* = A^{-1}b$ de $J_{A,b}$ en au plus n itérations.

De plus, le résidu

$$r_k = \nabla J_{A,b}(x_k) = Ax_k - b = A(x_k - x_*)$$

est orthogonal[12] *pour le produit scalaire $\langle\ ,\ \rangle$ à tous les $d_i, i < k$.*

DÉMONSTRATION. Choisissons x_0 quelconque non nul. Pour $k \ge 0$, par minimalité de $\alpha_{k+1} = \operatorname{argmin}_t[J_{A,b}(x_k + t d_k)]$, on a

$$0 = \langle \nabla J_{A,b}(x_k + \alpha_{k+1} d_k), d_k\rangle = \langle A(x_k + \alpha_{k+1} d_k) - b, d_k\rangle,$$

soit

$$\alpha_{k+1} = -\frac{\langle Ax_k - b, d_k\rangle}{\|d_k\|_A^2} = -\frac{\langle r_k, d_k\rangle}{\|d_k\|_A^2}. \tag{29}$$

Par définition de la suite $(x_k)_{k\ge 0}$, on a, pour $k < n$,

$$x_k = x_{k-1} + \alpha_k d_{k-1} = \ldots = x_0 + \alpha_1 d_0 + \alpha_2 d_1 + \ldots + \alpha_k d_{k-1},$$

11. Ce n'est pas le gradient qui est conjugué, mais les directions de descentes (déterminées par le gradient) : respectant l'usage, on gardera cette appellation au sens déficient.

12. Vu que $r_k = A(x_k - x_*)$, cette propriété est équivalente à la A-orthogonalité des vecteurs d'erreur $x_k - x_*$ avec les $d_0, \ldots, d_{k-1}$.

et donc, relativement à la base $(d_j)_{0\le j<n}$, le vecteur $x_k - x_0$ a ses $n-k$ dernières coordonnées nulles, avec ses k premières coordonnées égales aux coefficient $\alpha_1, \dots, \alpha_k$ définis dans (29). D'autre part, relativement à cette même base A-orthogonale. on a

$$x_* = x_0 + \sum_{k=0}^{n-1} \sigma_k d_k, \quad \sigma_k = \frac{\langle x_* - x_0, d_k\rangle_A}{\|d_k\|_A^2}.$$

Le vecteur $x_k - x_0$, combinaison linéaire des $d_0, \dots, d_{k-1}$, est A-conjugué à d_k, ainsi,

$$\langle x_* - x_0, d_k\rangle_A = \langle x_* - x_0, d_k\rangle_A + \langle x_0 - x_k, d_k\rangle_A = \langle x_* - x_k, d_k\rangle_A$$

En outre, vu que $Ax_* = b$,

$$\langle x_* - x_k, d_k\rangle_A = \langle Ax_* - Ax_k, d_k\rangle = \langle b - Ax_k, d_k\rangle = -\langle r_k, d_k\rangle$$

On en déduit la formule suivante pour la coordonnée α_{k+1} de $x_{k+p} - x_0$ $(1 \le p \le n-k)$ dans la base $(d_j)_{0\le j<n}$

$$\alpha_{k+1} = -\frac{\langle r_k, d_k\rangle}{\|d_k\|_A^2} = \frac{\langle x_* - x_k, d_k\rangle_A}{\|d_k\|_A^2} = \frac{\langle x_* - x_0, d_k\rangle_A}{\|d_k\|_A^2} = \sigma_k$$

soit $\alpha_{k+1} = \sigma_k, k = 0, \dots, n-1$. Ainsi, $x_k = x_*$ si $\alpha_{k+p} = 0, 0 \le p \le n-k-1$: l'itération est stationnaire à partir de x_k dans ce cas.

Enfin, on montre la A-orthogonalité du résidu r_k avec les $d_i, i < k$ par récurrence sur k. Il n'y a rien à démontrer pour $k = 0$. Supposant la propriété vraie au rang $k-1$, on a

$$\langle r_k, d_{k-1}\rangle = \langle \nabla J_{A,b}(x_k), d_{k-1}\rangle = 0$$

vu la construction de x_k, comme minimum de la restriction de $J_{A,b}$ à la droite $x_{k-1} + \mathbb{R}d_{k-1}$. Par ailleurs,

$$r_k = Ax_k - b = A(x_{k-1} + \alpha_k d_{k-1}) - b = r_{k-1} + \alpha_k A d_{k-1}$$

et donc pour $j < k-1$,

$$\langle r_k, d_j\rangle = \langle r_{k-1} + \alpha_k A d_{k-1}, d_j\rangle = \langle r_{k-1}, d_j\rangle + \alpha_k \langle d_{k-1}, d_j\rangle_A$$

où la nullité dans le dernier membre du premier terme provient de l'hypothèse de récurrence et celle du second de la A-conjugaison de la base $(d_j)_{0\le j<n}$. □

L'algorithme du type gradient-conjugué construit la famille de directions $\mathcal{D} = (d_j)$ progressivement, en rajoutant une direction en fin d'itération (orthogonalisée comme dans la méthode de Gram-Schmidt) : à l'itération k, il est construit une nouvelle approximation x_{k+1} du point de minimum, puis calculé le résidu $r_{k+1} = \nabla J(x_{k+1})$ qui est orthogonal (relativement au produit scalaire $\langle\ ,\ \rangle$) à l'espace $\mathrm{Vect}(d_0, \dots, d_k)$. La direction d_{k+1} introduite est sous la forme $d_{k+1} = -r_{k+1} + \beta_{k+1} d_k$ avec β_{k+1} déterminé par l'orthogonalité $d_{k+1} \perp_A \mathrm{Vect}(d_0, \dots, d_k)$, qui est presque assurée : il suffit de vérifier la A-orthogonalité de d_{k+1} et d_k, résultant du choix effectué pour le coefficient $\beta_{k+1} = \frac{\langle r_{k+1}, d_k\rangle_A}{\|d_k\|_A^2}$. Ce choix présuppose que le résidu r_{k+1} est non nul : si il l'est, x_{k+1} est le point de minimum (qui annule le gradient $\nabla J(x_{k+1}) = r_{k+1}$) et il l'est sûrement pour $k+1 \ge n$ puisque on ne peut avoir plus de n directions conjuguées ! L'itération, dite du gradient conjugué quadratique, est décrite dans l'algorithme 2.9.

Nous allons donner une expression équivalente au coefficient β_{k+1}, ce qui donnera un algorithme un peu différent de typeg gradient-conjugué pour des fonctions non quadratiques. Rappelons que le coefficient α_{k+1} est déterminé par l'extrémalité de $t \mapsto J_{A,b}(x_k + td_k)$ en $t = \alpha_{k+1}$

$$0 = \langle \nabla J_{A,b}(x_k + \alpha_{k+1} d_k), d_k\rangle = \langle A(x_k + \alpha_{k+1} d_k) - b, d_k\rangle$$

Algorithme 2.9 : algorithme du gradient conjugué sur fonction quadratique (I)

Choisir x_0 ; $r_0 = Ax_0 - b$; $d_0 = -r_0$; $k = 0$
tant que $\|r_k\| \geq \varepsilon$ **faire**
  $\alpha_{k+1} = -\dfrac{\langle r_k, d_k\rangle}{\|d_k\|_A^2}$
  $x_{k+1} = x_k + \alpha_{k+1} d_k$
  $r_{k+1} = Ax_{k+1} - b$
  $\beta_{k+1} = \dfrac{\langle r_{k+1}, d_k\rangle_A}{\|d_k\|_A^2}$
  $d_{k+1} = -r_{k+1} + \beta_{k+1} d_k$
  $k = k + 1$
fin tant que

soit $\alpha_{k+1} = -\langle r_k, d_k\rangle / \|d_k\|_A^2$, alors que le coefficient β_{k+1} est déterminé par la A-conjugaison de d_{k+1} et d_k

$$0 = \langle d_{k+1}, d_k\rangle_A = \langle -r_{k+1} + \beta_{k+1} d_k, d_k\rangle_A,$$

soit $\beta_{k+1} = \langle r_{k+1}, d_k\rangle_A / \|d_k\|_A^2$. Vu que $d_k = -r_k + \beta_k d_{k-1}$

$$\alpha_{k+1} = \frac{\langle r_k, r_k\rangle}{\|d_k\|_A^2} - \beta_k \frac{\langle r_k, d_{k-1}\rangle}{\|d_k\|_A^2} = \frac{\|r_k\|^2}{\|d_k\|_A^2}$$

et, si $\alpha_{k+1} \neq 0$,

$$\beta_{k+1} = \frac{\langle r_{k+1}, d_k\rangle_A}{\|d_k\|_A^2} = \frac{\langle r_{k+1}, A(x_{k+1} - x_k)\rangle}{\alpha_{k+1}\|d_k\|_A^2} = \frac{\langle r_{k+1}, r_{k+1} - r_k\rangle}{\|r_k\|^2} = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}$$

le produit scalaire $\langle r_{k+1}, r_k\rangle$ étant nul vu que $r_k \in \mathrm{Vect}(d_k, d_{k-1})$ orthogonal à r_{k+1} d'après la dernière assertion de la proposition précédente. En résulte une seconde forme (2.10) de l'algorithme du gradient conjugué pour le modèle quadratique $J_{A,b}$, algorithme qui converge en au plus $n = \dim A$ itérations vers le minimum de la fonctionnelle quadratique. Cet algorithme GC permet d'introduire des algorithmes de type GC pour le

Algorithme 2.10 : algorithme du gradient conjugué sur fonction quadratique (II)

Choisir x_0 ; $r_0 = Ax_0 - b$; $d_0 = -r_0$; $k = 0$
tant que $\|r_k\| \geq \varepsilon$ **faire**
  $\alpha_{k+1} = \dfrac{\|r_k\|^2}{\|d_k\|_A^2}$
  $x_{k+1} = x_k + \alpha_{k+1} d_k$
  $r_{k+1} = Ax_{k+1} - b$
  $\beta_{k+1} = \dfrac{\|r_{k+1}\|^2}{\|r_k\|^2}$
  $d_{k+1} = -r_{k+1} + \beta_{k+1} d_k$
  $k = k + 1$
fin tant que

minimum de fonctions J non quadratiques, dont la convergence, valable sous des hypothèses particulières, ne sera pas abordée ici : on a reporté dans l'algorithme 2.11 les choix de Fletcher-Reeves et de Polak-Ribière.

Algorithme 2.11 : Algorithme du gradient conjugué (J non quadratique)

Choisir x_0 ; r_0 ; $d_0 = -r_0$; $k = 0$
tant que $\|\nabla J(x_k)\| \geq \varepsilon$ **faire**
  $\alpha_{k+1} = RDD(J, x_k + \mathbb{R}_+ d_k)$
  $x_{k+1} = x_k + \alpha_{k+1} d_k$
  $r_{k+1} = \nabla J(x_{k+1})$
  $\beta_{k+1} = \dfrac{\|r_{k+1}\|^2}{\|r_k\|^2}$ (Fletcher-Reeves) ou $\beta_{k+1} = \dfrac{\langle r_{k+1}, r_{k+1} - r_k \rangle}{\|r_k\|^2}$ (Polak-Ribière)
  $d_{k+1} = -r_{k+1} + \beta_{k+1} d_k$
  $k = k + 1$
fin tant que

5. Moindres carrés

La méthode des moindres carrés intervient dans l'estimation des paramètres de modèles : la grandeur $y \in \mathbb{R}$ est mesurée sur un dispositif en réponse à une entrée $\mathbf{x} \in \mathbb{R}^n$ (dite variable explicative, prédicteur ou facteur : pas exemple x une image et y l'image floue observée dans le dispositif) ; cette mesure, que l'on souhaite expliquer (voire prédire), est modélisée par la variable aléatoire $Y(\Lambda, \mathbf{x})$ dépendant du paramètre $\Lambda \in \mathbb{R}^k$ et complétée par une erreur ε : $y - Y(\Lambda, \mathbf{x}) = \varepsilon \sim \mathcal{N}(0, 1)$. Par exemple, un modèle linéaire est modélisé par $Y(\Lambda, x) = \langle \Lambda, x \rangle$ avec $x, \Lambda \in \mathbb{R}^n$. La collection des mesures $(y_j)_{j=1}^m$, en nombre m nettement supérieur à celui des degrés de libertés $n = \dim \Lambda$ et dépendant des variables explicatives non aléatoires $(\mathbf{x}_j \in \mathbb{R}^n)_{j=1}^m$, est supposé être un échantillon issu de la réalisation de m variables $(Y(\Lambda, \mathbf{x}_j) + \varepsilon_j)_{j=1}^m$ indépendantes et identiquement distribuées de loi normale $\mathcal{N}(0, \sigma^2)$. La vraisemblance associée à ces observations $(\mathbf{y}_j = Y(\Lambda, \mathbf{x}_j) + \varepsilon_j)_{j=1}^m$, des réalisations de la variable aléatoire Y est donnée par

$$V(\mathbf{y}; \Lambda, \sigma, \mathbf{x}) = \prod_{j=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_j - Y(\Lambda, \mathbf{x}_j))^2}{2\sigma^2}\right) = \frac{\exp\left(-\frac{\sum_{j=1}^m (y_j - Y(\Lambda, \mathbf{x}_j))^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{m/2}}$$

L'écart-type σ étant fixé, le maximum de la vraisemblance $V(y; \Lambda, \sigma, \mathbf{x})$ correspond au minimum de $\Lambda_*(\mathbf{x}, \mathbf{y})$ pour la norme ℓ_2 du résidu $R_\Lambda(x, y) = (y_j - Y(\Lambda, \mathbf{x}_j)) \in \mathbb{R}^m$ (comme Gauß l'a considéré dans ses observations astronomiques de Cérès),

$$\|R_\Lambda(\mathbf{x}, \mathbf{y})\|_2^2 = \sum_{j=1}^m (y_j - Y(\Lambda, \mathbf{x}_j))^2$$

mais on pourrait prendre une norme ℓ_1 ou ℓ_∞ pour cette régression. Le minimiseur $\Lambda_*(\mathbf{y}, \mathbf{x})$ est un estimateur de Λ, estimateur qui, sous les hypothèses convenables, converge vers Λ quand le nombre m de mesures tend vers $+\infty$.

▷ Exemple 2.9: Supposons $Y(\lambda, x) = \mathrm{e}^{\lambda x}$ avec $\Lambda \in \mathbb{R}$. Alors

$$R_\lambda(\mathbf{y}, \mathbf{x}) = (y_j - Y(\lambda, x_j)) = (y_j - \mathrm{e}^{\lambda x_j}) \in \mathbb{R}^m, \quad \nabla_\Lambda R(\lambda) = (-x_j \mathrm{e}^{\lambda x_j}),$$

$$J_\lambda(y, x) = \frac{1}{2}\|R_\lambda(y, x)\|_2^2 = \sum_{j=1}^m \left|y_j - \mathrm{e}^{\lambda x_j}\right|^2 /2, \quad \nabla_\lambda J(\lambda) = \sum_j (\mathrm{e}^{\lambda x_j} - y_j) x_j \mathrm{e}^{\lambda x_j}$$

$$\operatorname{Hess} J(\lambda) = \sum_j x_j^2 \mathrm{e}^{\lambda x_j}(2\mathrm{e}^{\lambda x_j} - y_j)$$ ◁

Dit *de moindres carrés linéaires*, le modèle le plus simple à analyser, est celui où le résidu $R_\Lambda(\mathbf{x},\mathbf{y})$ est affine dans la variable Λ : $R_\Lambda = A\Lambda - b$ avec $A = A(\mathbf{x},\mathbf{y}) \in \mathrm{Hom}(\mathbb{R}^k,\mathbb{R}^p)$ et $b = b(\mathbf{x},\mathbf{y}) \in \mathbb{R}^p$ où $\mathbf{x} = (x_j) \in (\mathbb{R}^\ell)^m, \mathbf{y} = (y_j) \in \mathbb{R}^m$. Pour le modèle linéaire, on a $R_\Lambda = (y_j - \langle \Lambda, x_j\rangle)_{j=1}^m = \widetilde{\mathbf{Y}} - \widetilde{\mathbf{X}}\Lambda$ avec $\widetilde{\mathbf{Y}} = {}^\mathsf{T}(y_1,\dots,y_m) \in \mathbb{R}^m$ et $\widetilde{\mathbf{X}} = {}^\mathsf{T}(x_1,\dots,x_m) \in \mathrm{Hom}(\mathbb{R}^k,\mathbb{R}^m)$.

Les minima de

$$J(\Lambda) = \|R_\Lambda(\mathbf{x},\mathbf{y})\|^2 = \|A\Lambda - b\|^2 = \langle {}^\mathsf{T}AA\Lambda,\Lambda\rangle - 2\langle {}^\mathsf{T}Ab,\Lambda\rangle + \|b\|^2$$

sont des points critiques de J : $\nabla J(\Lambda) = 2{}^\mathsf{T}A(A\Lambda - b) = 0$. L'application J étant convexe, tout minimum local est global et l'ensemble de ses minima est convexe, réduit à un point si J est strictement convexe : c'est le cas si et seulement si l'application L'application ${}^\mathsf{T}AA$ est bijective, autrement dit si et seulement si A est injective vu que $\ker A = \ker {}^\mathsf{T}AA$. Ainsi, dans le cas de A injective, l'opérateur $A^\dagger = ({}^\mathsf{T}AA)^{-1}{}^\mathsf{T}A$ est appelé *pseudo-inverse* de A, déterminant le minimum

$$\Lambda_* = ({}^\mathsf{T}AA)^{-1}{}^\mathsf{T}Ab = A^\dagger b. \tag{30}$$

△ Remarque 2.5: La minimisation de $\|A\Lambda - b\|_2$, comme substitut de la résolution de l'équation linéaire $A\Lambda = b$, a lieu dans de multiples cas. Citons l'exemple de la recherche des minima de U soumis aux contraintes $h_j = 0, j = 1,\dots,m$: la recherche itérative des multiplicateurs de Lagrange $\Lambda_* = (\lambda_{*j})$ tels que $\nabla U(x_*) = \langle \Lambda_*, \nabla h(x_*)\rangle$ par la méthode de Newton-Lagrange est initiée avec un x_0 minimum approché obtenu par une recherche stochastique et avec un multiplicateur Λ_0 minimisant $\|\nabla U(x_0) - \langle \Lambda, \nabla h(x_0)\rangle\|_2$. En considérant la matrice $A(x)$ dont les vecteurs colonnes sont les $\nabla h_j(x), j = 1,\dots,p$ et $b(x) = \nabla U(x)$, la condition de Lagrange est équivalente à $A(x)^\mathsf{T}\Lambda = b(x)$, qu'on remplace par l'étude de $\mathrm{argmin}_\Lambda \|A(x)^\mathsf{T}\Lambda - b(x)\|_2$. L'itération de Newton-Lagrange (x_k,Λ_k) est souvent convergente si l'itération stochastique, suivie de cette recherche de multiplicateurs de Lagrange minimisants a donné un point suffisamment approché du minimum. ▽

Si R est non linéaire, on a, pour $J = \|R\|^2/2$

$$\begin{aligned}\nabla J(\Lambda) &= \langle R(\Lambda), \nabla R(\Lambda)\rangle = {}^\mathsf{T}(\nabla R(\Lambda))R(\Lambda),\\ \mathrm{Hess}\, J(\Lambda) &= \langle \nabla R(\Lambda), \nabla R(\Lambda)\rangle + \langle \mathrm{Hess}\, R(\Lambda), R(\Lambda)\rangle.\end{aligned}$$

où on a considéré $R(\Lambda)$ comme un vecteur colonne de m lignes et $\nabla R(\Lambda)$ une matrice de type $m \times n$. La méthode de Gauß-Newton pour des moindres carrés non linéaires (exemple de méthode dite de *quasi-Newton*) est d'omettre le terme $\langle \mathrm{Hess}\, R(\Lambda), R\rangle$ (heuristiquement négligeable si le résidu R l'est ou la non linéarité est faible avec une dérivée seconde $\mathrm{Hess}\, R(\lambda)$ petite) de l'itération classique de Newton avec quasi-hessienne ${}^\mathsf{T}(\nabla R(\Lambda))\nabla R(\Lambda)$ dont l'inverse est appliqué au gradient $\nabla J(\Lambda) = {}^\mathsf{T}(\nabla R(\Lambda))R$

$$\Lambda_{k+1} = \Lambda_k - ({}^\mathsf{T}\nabla R_{\Lambda_k}\nabla R_{\Lambda_k})^{-1}{}^\mathsf{T}\nabla R_{\Lambda_k}\, R(\Lambda_k)$$

Théorème 2.8 ([12]): *Soit U un ouvert convexe de $\mathbb{R}^n$, $R : U \to \mathbb{R}^m$ de classe $\mathcal{C}^2$ et $J = \frac{1}{2}\|R\|^2$ On suppose que ∇R est lipchitzien avec $\|\nabla R\|_2 \le \alpha$ et qu'il existe $x_* \in U$, $\lambda_-, \sigma > 0$ tels que $\langle \nabla R(x_*), R(x_*)\rangle = 0$, λ_- est la plus petite valeur propre de $\langle \nabla R(x_*), \nabla R(x_*)\rangle$ et*

$$\|({}^\mathsf{T}\nabla R(x) - {}^\mathsf{T}\nabla R(x_*))R(x_*)\|_2 \le \sigma\|x - x_*\|_2, \quad x \in U. \tag{31}$$

Si $\sigma < \lambda_-$, alors pour tout $c \in (1, \lambda_-/\sigma)$, il existe $\varepsilon > 0$ tel que pour tout $x_0 \in B(x_*, \varepsilon)$ la suite de Gauß-Newton démarrant en x_0 et vérifiant

$$x_{k+1} = x_k - ({}^{\mathsf{T}}\nabla R(x_k)\nabla R(x_k))^{-1}\,{}^{\mathsf{T}}\nabla R(x_k)R(x_k), \quad k \geq 0$$

est bien définie, converge vers x_* et vérifie

$$\|x_{k+1} - x_*\| \leq \frac{c\sigma}{\lambda_-}\|x_k - x_*\|_2 \left[\sigma + \frac{c\alpha\gamma}{2\lambda_-}\|x_k - x_*\|_2\right] \tag{32}$$

et

$$\|x_{k+1} - x_*\| \leq \frac{c\sigma + \lambda_-}{2\lambda_-}\|x_k - x_*\|_2 < \|x_k - x_*\|_2,$$

dernière inégalité qui assure la convergence linéaire de la suite (x_k).

△ Remarque 2.6: La direction de la méthode de Gauß-Newton est une direction de descente

$$\langle -({}^{\mathsf{T}}\nabla R(x)\nabla R(x))^{-1}\,{}^{\mathsf{T}}\nabla R(x)\nabla R(x), \nabla J(x)\rangle = -\langle ({}^{\mathsf{T}}\nabla R(x)\nabla R(x))^{-1}\nabla J(x), \nabla J(x)\rangle \leq 0.$$

Sauf si $R(x_*) = 0$ (en vertu de (31) avec $\sigma = 0$ et (32)), la méthode de Gauß-Newton n'a pas la convergence quadratique de l'algorithme newtonien : il a l'avantage d'œuvrer sans exiger le calcul de hessiennes (les estimer par différences finies est souvent d'un coût prohibitif). ▽

Démonstration. Posons $R_0 = R(x_0)$, $R_* = R(x_*)$ et $G_0 = \nabla R(x_0)$. Il existe $\varepsilon_1 > 0$ tel que ${}^{\mathsf{T}}G_0G_0$ soit non singulière

$$\left\|({}^{\mathsf{T}}G_0G_0)^{-1}\right\| \leq \frac{c}{\lambda_-}, \quad x_0 \in B(x_*, \varepsilon_1).$$

Soit $\varepsilon = \min(\varepsilon_1, (\lambda_- - c\sigma)/(c\alpha\gamma))$. Alors x_1 est bien défini et vérifie

$$\begin{aligned} x_1 - x_* &= x_0 - x_* - ({}^{\mathsf{T}}G_0G_0)^{-1}\,{}^{\mathsf{T}}G_0R_0 \\ &= -({}^{\mathsf{T}}G_0G_0)^{-1}[{}^{\mathsf{T}}G_0R_0 + {}^{\mathsf{T}}G_0G_0(x_* - x_0)] \\ &= -({}^{\mathsf{T}}G_0G_0)^{-1}[{}^{\mathsf{T}}G_0R_* - {}^{\mathsf{T}}G_0(R_* - R_0 - G_0(x_* - x_0)] \end{aligned}$$

La formule de Taylor donne la majoration

$$\|R_* - R_0 - G_0(x_* - x_0)\| \leq \frac{\gamma}{2}\|x_0 - x_*\|^2.$$

Vu ${}^{\mathsf{T}}G_*R(x_*) = 0$, on a $\left\|{}^{\mathsf{T}}G_0R_*\right\| \leq \sigma\|x_* - x_0\|$. Vu $\|G_0\| \leq \alpha$, il en résulte

$$\begin{aligned} \|x_1 - x_*\| &\leq \left\|({}^{\mathsf{T}}G_0G_0)^{-1}\right\| \left[\left\|{}^{\mathsf{T}}G_0R_*\right\| + \|G_0\|\|R_* - R_0 - G_0(x_* - x_0)\|\right] \\ &\leq \frac{c}{\lambda_-}\left[\sigma\|x_0 - x_*\| + \frac{\alpha\gamma}{2}\|x_0 - x_*\|^2\right]. \end{aligned}$$

ce qui prouve l'estimée annoncée pour $k = 0$. On a donc

$$\begin{aligned} \|x_1 - x_*\| &\leq \|x_0 - x_*\|\frac{c}{\lambda_-}\left[\sigma + \frac{\alpha\gamma}{2}\|x_0 - x_*\|\right] \leq \|x_0 - x_*\|\left[\frac{c\sigma}{\lambda_-} + \frac{\lambda_- - c\sigma}{2\lambda_-}\right] \\ &\leq \|x_0 - x_*\|\left[\frac{c\sigma + \lambda_-}{2\lambda_-}\right] < \|x_0 - x_*\|. \end{aligned}$$

On recommence pareillement pour prouver les estimations de la k-ème itération. □

▷ Exemple 2.10: Reprenons l'exemple 2.9 avec le seul paramètre λ et 3 mesures $y = (2, 4, y_3)$ pour les entrées $x = (1, 2, 3)$: la fonction de moindres carrés est donc $J(\lambda) = |2 - \mathrm{e}^{\lambda}|^2 + |4 - \mathrm{e}^{2\lambda}|^2 + |y_3 - \mathrm{e}^{3\lambda}|^2$.

On compare la convergence des méthodes de Newton et Gauß-Newton pour différentes valeurs de y_3 et deux choix pour l'initialisation λ_0. Le test d'arrêt est $\|\lambda_k - \lambda_{k-1}\|_2 < 10^{-16}$. ◁

y_3	λ_0	$\#_N$	$\#_{GN}$	λ_{N*}	λ_{GN*}	$J(\lambda_{N*})$
8	1	7	6	0.693147180559945286	0.693147180559945286	1.578e-30
	0.6	6	5	0.693147180559945286	0.693147180559945397	
3	1	9	9	0.440049858082299961	0.4400498582750344	1.639
	0.5	5	7	0.440049858082300016	0.440049858162159968	
-1	1	11	25	0.0447439841906622018	0.0447439858926492254	6.976
	0	5	23	0.0447439841906622365	0.0447439858763129528	
-4	1	12	*	-0.37192873255882386		16.435
	-0.3	5	*	-0.371928732558823694		
-8	1	13	*	-0.791486337059211342		41.145
	-0.7	5	*	-0.791486337059211342		

Table 3. Comparaison de Newton et Gauß-Newton pour les moindres carrés $J(\lambda) = |2 - \mathrm{e}^{\lambda}|^2 + |4 - \mathrm{e}^{2\lambda}|^2 + |y_3 - \mathrm{e}^{3\lambda}|^2$: $*$ signifie le défaut de convergence ; pour $y_3 = 8$, la valeur exacte du minimum est $\lambda_\infty = \log 2$.

6. La méthode BFGS

La méthode de BFGS (Broyden, Fletcher, Goldfarb, Shanno) est une méthode quasi-Newton pour la recherche d'un minimum d'une fonction J. Le modèle quadratique de J en $x_{k+1} = x_k + \alpha_k d_k$ est $m_{k+1}(u) = J(x_{k+1}) + \langle \nabla J_{x_{k+1}}, u\rangle + \mathrm{Hess}\, J_{x_{k+1}}(u)/2$. Une condition raisonnable est que son gradient coïncide avec celui de J aux points x_k et x_{k+1} : la condition est bien vérifiée en x_{k+1}, celle en x_k s'exprime suivant

$$\nabla J_{x_k} = \nabla m_{k+1}(x_{k+1} - x_k) = \nabla J_{x_{k+1}} + \mathrm{Hess}\, J_{x_{k+1}}(x_{k+1} - x_k)$$

Ainsi, si $H_{k+1} = \mathrm{Hess}\, J_{x_{k+1}}$, $s_k = x_{k+1} - x_k$ et $\delta_k = \nabla J_{x_{k+1}} - \nabla J_{x_k}$, on obtient l'équation dite de la *sécante*

$$H_{k+1} s_k = \delta_k$$

avec la condition nécessaire de positivité $\langle s_k, \delta_k\rangle = \langle s_k, H_{k+1} s_k\rangle > 0$. Si J est quadratique, cette équation de la sécante est toujours vérifiée, quels que soient x, y avec $s = y - x$ et $\delta = \nabla J_y - \nabla J_x$: la méthode quasi-Newton requiert que cette équation de la sécante vaille pour x_{k+1} et x_k seulement. La méthode BFGS construit l'inverse approché B_{k+1} de H_{k+1} par une modification de B_k en résolvant le problème de minimisation

$$\min_{\substack{B = {}^{\mathsf{T}}B \\ B\delta_k = s_k}} \|B - B_k\|_W^2 \tag{33}$$

dans l'espace des matrices M_n d'ordre n muni de la norme

$$\|A\|_W = \|W^{1/2} A W^{1/2}\|_F, \quad \langle A, B\rangle_F = \mathrm{tr}({}^{\mathsf{T}}AB), \quad A, B \in M_n$$

avec W un opérateur tel que l'inverse W_k de la hessienne moyennée sur le segment $[x_k, x_{k+1}]$

$$W_k^{-1} = \int_0^1 \mathrm{Hess}\, J(x_k + (x_{k+1} - x_k)u)\, du$$

pour laquelle $W_k\delta_k = s_k$ comme l'assure la formule de Taylor

$$\delta_k = \nabla J_{x_{k+1}} - \nabla J(x_k) = \int_0^1 \mathrm{Hess}\, J(x_k + t(x_{k+1} - x_k))(x_{k+1} - x_k)dt = W_k^{-1}s_k.$$

Ainsi B_{k+1}, solution symétrique de $B\delta_k = s_k$, est celle qui est la plus proche de B_k pour la norme $\|\ \|_W$. La solution du programme (33) est donnée par le lemme suivant

LEMME 2.4: *Soient A, W définies positives, $x, y \in \mathbb{R}^n$ avec $Wy = x$ et $\langle x, y\rangle > 0$. Le programme*

$$\min_{\substack{B={}^{\mathsf{T}}B \\ Bx=y}} \|B - A\|_W^2$$

a pour solution unique

$$B_* = \left(1 - \frac{P_{y,x}}{\langle x, y\rangle}\right) A \left(1 - \frac{P_{x,y}}{\langle x, y}\right) + \frac{P_{y,y}}{\langle x, y\rangle} \tag{34}$$

où $P_{x,y}$ est l'opérateur de rang 1 tel que $P_{x,y}(v) = \langle y, v\rangle x, v \in \mathbb{R}^n$. L'opérateur B_* est défini positif si A l'est.

DÉMONSTRATION. Relativement au produit scalaire $\langle\ ,\ \rangle_W$ (qui sera aussi utilisé pour les gradients ci-dessous), la forme linéaire $H \in M_n \mapsto \langle M, Hx\rangle$ est représentée par l'opérateur $P_{W^{-1}M,W^{-1}x}$. En effet, vu que W est symétrique,

$$\begin{aligned}\langle P_{W^{-1}M,W^{-1}x}, H\rangle_W &= \langle W^{1/2}P_{W^{-1}M,W^{-1}x}W^{1/2}, W^{1/2}HW^{1/2}\rangle_F \\ &= \mathrm{tr}({}^{\mathsf{T}}(P_{W^{-1/2}M,W^{-1/2}x})W^{1/2}HW^{1/2}) \\ &= \mathrm{tr}(P_{W^{-1/2}x,W^{-1/2}M}W^{1/2}HW^{1/2}) \\ &= \mathrm{tr}(P_{x,M}H) = \mathrm{tr}(P_{Hx,M}) = \langle Hx, M\rangle.\end{aligned}$$

Le gradient relativement à la variable B du lagrangien

$$\mathcal{L}(B, \Lambda, M) = \|B - A\|_W^2 - 2\langle \Lambda, B - {}^{\mathsf{T}}B\rangle_W - 4\langle M, Bx - y\rangle_{\mathbb{R}^n}, \quad B \in M_n, \Lambda \in M_n, M \in \mathbb{R}^n$$

prend donc la forme

$$\frac{1}{2}\nabla\mathcal{L}(B) = B - A - (\Lambda - {}^{\mathsf{T}}\Lambda) - 2P_{W^{-1}M,y}$$

On en déduit au point critique (B_*, Λ_*, M_*)

$$A + \Lambda_* - {}^{\mathsf{T}}\Lambda_* + 2P_{W^{-1}M_*,x} = B_* = {}^{\mathsf{T}}B_* = A + {}^{\mathsf{T}}\Lambda_* - \Lambda_* + 2P_{x,W^{-1}M_*}$$

soit

$$\Lambda_* - {}^{\mathsf{T}}\Lambda_* = P_{x,W^{-1}M_*} - P_{W^{-1}M_*,x}$$

et

$$B_* = A + P_{y,W^{-1}M_*} + P_{W^{-1}M_*,y} \tag{35}$$

puis, vu $B_*x = y$,

$$y = Ax + \langle W^{-1}M_*, x\rangle y + W^{-1}M_*\langle y, x\rangle.$$

Prenant le produit scalaire avec x, on obtient

$$\langle y - Ax, x\rangle = 2\langle W^{-1}M_*, x\rangle\langle x, y\rangle$$

et donc

$$W^{-1}M_* = \frac{y - Ax - \langle W^{-1}M_*, x\rangle y}{\langle x, y\rangle} = \frac{y - Ax - \frac{\langle y - Ax, x\rangle}{2\langle x, y\rangle}y}{\langle x, y\rangle}$$

soit, en n'oubliant pas que A est symétrique

$$P_{y,W^{-1}M_*} = \frac{P_{y,y} - P_{y,x}A}{\langle x,y\rangle} - \frac{\langle y - Ax, x\rangle P_{y,y}}{2\langle x,y\rangle^2} = \frac{-P_{y,x}A}{\langle x,y\rangle} + \frac{P_{y,y}}{2\langle x,y\rangle} + \frac{\langle Ax, x\rangle P_{y,y}}{2\langle x,y\rangle^2}$$
$$P_{W^{-1}M_*,y} = \frac{P_{y,y} - AP_{x,y}}{\langle x,y\rangle} - \frac{\langle y - Ax, x\rangle P_{y,y}}{2\langle x,y\rangle^2} = \frac{-AP_{x,y}}{\langle x,y\rangle} + \frac{P_{y,y}}{2\langle x,y\rangle} + \frac{\langle Ax, x\rangle P_{y,y}}{2\langle x,y\rangle^2}$$

et finalement, en reprenant (35),

$$B_* = \left(1 - \frac{P_{y,x}}{\langle x,y\rangle}\right) A \left(1 - \frac{P_{x,y}}{\langle x,y\rangle}\right) + \frac{P_{y,y}}{\langle x,y\rangle}$$

Posons $Q_{x,y} = 1 - P_{x,y}/\langle x,y\rangle$. Alors

$$\langle z, B_* z\rangle = \langle z, {}^{\mathsf{T}}Q_{x,y}AQ_{x,y}z\rangle + \frac{\langle z,y\rangle^2}{\langle x,y\rangle} = \langle {}^{\mathsf{T}}Q_{x,y}z, AQ_{x,y}z\rangle + \frac{\langle z,y\rangle^2}{\langle x,y\rangle}$$

ce qui assure B_* défini positif vu $\langle x,y\rangle > 0$ et A défini positif. □

△ REMARQUE 2.7: Grâce à la formule de Sherman-Morrison-Woodbury donnant l'inverse d'une perturbation de rang 1 d'une matrice

$$(A + P_{x,y})^{-1} = A^{-1} - P_{A^{-1}x,{}^{\mathsf{T}}A^{-1}y}/(1 + \langle y, A^{-1}x\rangle)$$

on obtient l'inverse

$$B_{k+1}^{-1} = B_k - \frac{P_{By,By}}{\langle y, By\rangle} + \frac{P_{x,x}}{\langle x,y\rangle},$$

formule vérifiée directement en multipliant le membre de gauche par B_{k+1}. ▽

Algorithme 2.12 : L'algorithme BFGS

1: Choisir x_0 et B_0 ; $k = 0$
2: **tant que** ! test **faire**
3: $d_k = -B_k \nabla J_k$
4: $\alpha_k = \operatorname{argmin}(t \mapsto x_k + td_k)$
5: $s_k = x_{k+1} - x_k$; $\delta_k = \nabla J_{x_{k+1}} - \nabla J_{x_k}$
6: $B_{k+1} = \left(1 - \frac{P_{s_k,\delta_k}}{\langle s_k,\delta_k\rangle}\right) B_k \left(1 - \frac{P_{\delta_k,s_k}}{\langle s_k,\delta_k\rangle}\right) + \frac{P_{s_k,s_k}}{\langle s_k,\delta_k\rangle}$
7: $k = k + 1$
8: **fin tant que**

Le résultat de convergence pour la méthode BFGS est établi rigoureusement pour une fonction d'objectif à hessienne positive :

THÉORÈME 2.9: *Soit J de classe C^2 telle que l'ensemble de sous-niveau $C_0 = \{J(x) \leq J(x_0)\}$ soit convexe et la fonction J y vérifie $m\|v\|^2 \leq \operatorname{Hess} J_x(v) \leq M\|v\|^2$ pour $v \in \mathbb{R}^n$. Soit B_0 définie positive. Alors la suite (x_k) de l'algorithme (6) converge vers l'unique minimum x_* de J sur C_0.*

△ REMARQUE 2.8: On prend souvent $B_0 = Id$: la première itération est une descente dans la direction du gradient, alors que la k-ème itération s'apparente plus à une descente à la Newton à mesure que B_k se rapproche de l'inverse de la hessienne $\operatorname{Hess} J$ (cela n'est pas toujours le cas !). Les études théoriques et expérimentales indiquent des propriétés auto-correctives fortes de la méthode BFGS, qui tendent à rapprocher effectivement B_k de l'inverse de la hessienne. ▽

7. Optimisation avec contraintes

7.1. Optimisation avec contraintes d'égalités.

DÉFINITION 2.5: *Soit la fonction d'objectifs* $U : V(\text{ouvert de } \mathbb{R}^n) \to \mathbb{R}$ *et* $g : V \to \mathbb{R}^m$ *une fonction dite de* contrainte. *Il leur est associé le problème de minimisation (avec contrainte d'égalités)*

$$\min_{\substack{x\in V\\ g(x)=0}} U(x) \tag{36}$$

En écrivant la fonction (vectorielle) g *suivant ses coordonnées* $g(x) = (g_1(x), \dots, g_m(x))$, *on dit que* g *réalise* m *contraintes (numériques).*

On conviendra d'écrire encore ∇g : c'est une collection $(\nabla g_1, \dots, \nabla g_m)$ de m vecteurs de $\mathbb{R}^n$, considérée comme une matrice à m colonnes et n lignes[13]. Pour $h \in \mathbb{R}^n$, l'image $dg(h) \in \mathbb{R}^m$ s'interprète comme $\langle \nabla g, h\rangle = (\langle \nabla g_1, h\rangle, \dots, \langle \nabla g_m, h\rangle)$ ou comme le produit matriciel ${}^{\mathsf{T}}\nabla g X_h$ où X_h est le vecteur (colonne) des coordonnées de h.

Soit C_g la partie $C_g = g^{-1}(0) = \{x \in V : g(x) = 0\}$, appelée domaine des points *admissibles* ou lieu des contraintes (la variable x est la variable de *contrôle* ou de *décision*). On supposera dans la suite, et sans le répéter, que la partie C_g des points admissibles n'est pas vide.

Si g est affine de la forme $g(x) = \ell(x) + C$ avec ℓ linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^m$ et $C \in \mathbb{R}^m$, le lieu C_g est une droite (si $n = 2, m = 1$ ou $n = 3, m = 2$), un plan ($n = 3, m = 1$), en général un sous-espace affine de $\mathbf{R}^n$, sous-espace de dimension $n - m$ si l'application $\ell : \mathbb{R}^n \to \mathbb{R}^m$ est surjective (cf. Lemme 2.5 et remarques connexes).

Si g est non linéaire, alors sous des conditions adéquates, la partie C_g est une courbe ($n = 2, m = 1$ ou $n = 3, m = 2$), une surface ($n = 3, m = 1$), une sous-variété en général : les problèmes de recherche de minimum avec contraintes sont donc la recherche de minima pour des fonctions définies sur des parties de $\mathbb{R}^n$ dépourvues de calcul différentiel tel qu'il existe sur les ouverts des espaces $\mathbb{R}^N$. Les conditions habituelles de régularité consistent en la régularité $\mathcal{C}^1$ de g et la surjectivité de la différentielle $dg(x)$ pour $x \in C_g$ qui permet l'application du théorème des fonctions implicites : celui-ci assure l'existence de paramétrage local $\varphi : t \in B_{\mathbb{R}^{n-m}}(0, \rho) \mapsto \varphi(t) \in C_g$ du lieu des contraintes au voisinage de $x_* = \varphi(0)$. La méthode de substitution permet de se ramener, en théorie, à un problème de minimisation sans contrainte $\min_{t\in U}(J \circ \varphi(t))$: on écartera cette méthode, car le paramétrage φ est rarement explicite !

DÉFINITION 2.6: *Soient* U, g *définies sur* $V(\text{ouvert de }\mathbb{R}^n$ *et de classe* $\mathcal{C}^1$ *déterminant le problème d'optimisation avec contrainte d'égalités (36).*

Si $m = 1$, *le lagrangien* $\mathcal{L}_{U,g}$ *associé est la fonction définie sur* $V \times \mathbb{R}$ *par*

$$(x, \lambda) \in V \times \mathbb{R} \mapsto \mathcal{L}_{U,g}(x, \lambda) = U(x) - \lambda g(x).$$

De manière plus générale si $m > 1$, *le lagrangien est défini suivant*

$$(x, \lambda_1, \dots, \lambda_m) \in V \times \mathbb{R}^m \mapsto \mathcal{L}_{U,g}(x, \lambda_1, \dots, \lambda_m) = U(x) - \sum_{k=1}^{m} \lambda_k g_k(x)$$

ou encore

$$(x, \Lambda) \in V \times \mathbb{R}^m \mapsto \mathcal{L}_{U,g}(x, \Lambda) = U(x) - \langle \Lambda, g(x)\rangle_{\mathbb{R}^m}$$

Les coefficients $\lambda, \Lambda, \lambda_1, \dots \lambda_m$ *sont appelés* multiplicateurs de Lagrange

13. C'est un élément du produit tensoriel $\mathbb{R}^m \otimes \mathbb{R}^n$, espace vectoriel de dimension mn.

THÉORÈME 2.10: *Soit x_* un minimum du programme*

$$\min_{\substack{x\in V\\ g(x)=0}} U(x), \tag{37}$$

où $U : V(\text{ouvert de } \mathbb{R}^n) \to \mathbb{R}$ et $g : V \to \mathbb{R}^m$ de classe $\mathcal{C}^1$. Si la différentielle $dg(x_*)$ est surjective, alors il existe un multiplicateur unique Λ_* tel que (x_*, Λ_*) soit un point critique du lagrangien $\mathcal{L}_{U,g}$, i. e.

$$\partial_j U(x_*) = \langle \Lambda_*, \partial_j g(x_*)\rangle, \quad j = 1, \dots, n$$

soit

$$\nabla U(x_*) = \langle \Lambda_*, \nabla g(x_*)\rangle. \tag{38}$$

△ REMARQUES 2.9:

(1) Si g est numérique, la condition de surjectivité signifie simplement la non nullité de dg_{x_*}, *i. e.* que x_* n'est pas un point critique de g. La condition (38) s'écrit simplement $\nabla U(x_*) = \lambda_* \nabla g(x_*)$.

(2) Les $n+m$ variables x_*, Λ_* sont déterminées par les n équations $\nabla_x \mathcal{L}_{U,g}(x_*, \Lambda_*) = 0$ complétées par les m équations de contrainte $g(x_*) = 0$ qui correspondent exactement à l'équation $\nabla_\Lambda \mathcal{L}_{U,g}(x_*, \Lambda_*) = 0$. Par expérience, on commence souvent par résoudre les variables Λ_*, puis les x_*, celles qui nous intéresse comme point de minimum du programme avec contraintes $\min_{g=0} U$.

(3) Les gradients ∇U et ∇g sont orthogonaux aux courbes de niveau de U et g respectivement : ainsi, si x_* est extremum de U sous la contrainte g et les gradients $\nabla U(x_*), \nabla g(x_*)$ non nuls, les courbes de niveau de U et g sont tangentes en x_*, comme l'illustre la figure 2.13.

(4) Les gradients $\nabla g_1(x_*), \dots, \nabla g_m(x_*)$ sont linéairement indépendants si et seulement si la différentielle $dg(x_*)$ est surjective, comme l'exprime le lemme général suivant.

LEMME 2.5: *Soit B matrice d'ordre (m, n) représentant un opérateur de $\mathbb{R}^n$ dans $\mathbb{R}^m$. Notons par $b_1, \dots, b_m$ ses m lignes considérées comme vecteurs de $\mathbb{R}^n$. Alors B est surjectif si et seulement si la famille $(b_1, \dots, b_m)$ est libre.*

DÉMONSTRATION. L'identité

$$\langle u, Bv\rangle_{\mathbb{R}^m} = \langle {}^{\mathsf{T}}Bu, v\rangle_{\mathbb{R}^n}, \quad u \in \mathbb{R}^m, v \in \mathbb{R}^n$$

établit l'égalité $\ker {}^{\mathsf{T}}B = (\operatorname{Im} B)^\perp$, soit l'équivalence ${}^{\mathsf{T}}B$ injective et B surjective ($\operatorname{Im} B = \mathbb{R}^m$ est équivalent à $(\operatorname{Im} B)^\perp = 0$). La famille $(b_1, \dots, b_m)$ est l'image par ${}^{\mathsf{T}}B$ de la base canonique de $\mathbb{R}^m$, libre donc si et seulement si ${}^{\mathsf{T}}B$ injective. □

Si B est surjective, son noyau est l'orthogonal $\operatorname{Vect}(b_1, \dots, b_m)^\perp$, de dimension $n - m$ puisque $\operatorname{Vect}(b_1, \dots, b_m)$ est de dimension m.

(5) Soit E un espace vectoriel, V un sous-espace défini par $V = \operatorname{Vect}(b_1, \dots, b_m)$ et $w \in E$. Supposer que tout u tel que $\langle u, b_i\rangle = 0$ pour $i = 1, \dots, m$ vérifie $\langle u, w\rangle = 0$, c'est exprimer l'appartenance de w à l'orthogonal de l'orthogonal de V, soit $w \in (V^\perp)^\perp$. Vu que $(V^\perp)^\perp = V$, on en déduit que $w \in V$,*i. e.* w est combinaison linéaire des b_i : $w = \alpha_1 b_1 + \dots + \alpha_m b_m$: si la famille $(b_1, \dots, b_m)$ est libre, le m-uplet de coefficients $(\alpha_1, \dots, \alpha_m)$ est unique. ▽

▷ EXEMPLES 2.11:

2.11.1 Soit A matrice symétrique inversible d'ordre n, $\ell \in \mathbb{R}^n$, B une application linéaire $B : \mathbb{R}^n \to \mathbb{R}^m$ représentée par les vecteurs $b_j, j = 1, \dots, m$ ($Bx = (\langle b_j, x\rangle_{\mathbb{R}^n})$) et $k \in \mathbb{R}^m$. On suppose B surjective, *i. e.* les vecteurs b_j linéairement indépendants : l'opérateur ${}^{\mathsf{T}}B$ associe b_j au j-ème vecteur de la base canonique de $\mathbb{R}^m$. Le problème de minimisation

$$\min_{Bx=k} \left(\frac{\langle Ax, x\rangle}{2} + \langle \ell, x\rangle \right)$$

a comme lagrangien associée

$$\mathcal{L}(x, \Lambda) = \frac{\langle Ax, x\rangle}{2} + \langle \ell, x\rangle - {}^{\mathsf{T}}\Lambda(Bx - k)$$

dont les points critiques sont donnés par le système

$$0 = Ax + \ell - \sum_j \lambda_j b_j = Ax + \ell - {}^{\mathsf{T}}B\Lambda, \quad 0 = Bx - k.$$

soit

$$\begin{pmatrix} A & {}^{\mathsf{T}}B \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ -\Lambda \end{pmatrix} = \begin{pmatrix} -\ell \\ k \end{pmatrix}. \tag{39}$$

La première équation se réécrit

$$x = A^{-1}\left({}^{\mathsf{T}}B\Lambda - \ell \right)$$

et la contrainte $Bx = k$ donne $BA^{-1}\ell + k = BA^{-1}{}^{\mathsf{T}}B\Lambda$. L'application $BA^{-1}{}^{\mathsf{T}}B$ est injective : si $BA^{-1}{}^{\mathsf{T}}Bv = 0$, alors $0 = \langle BA^{-1}{}^{\mathsf{T}}Bv, v\rangle = \langle A^{-1}{}^{\mathsf{T}}Bv, {}^{\mathsf{T}}Bv\rangle$, soit vu que A^{-1} est défini positif, ${}^{\mathsf{T}}Bv = 0$ et $v = 0$ puisque ${}^{\mathsf{T}}B$ est injective. Ainsi $BA^{-1}{}^{\mathsf{T}}B$ est un isomorphisme et le multiplicateur Λ_* est donné par

$$\Lambda_* = \left(BA^{-1}{}^{\mathsf{T}}B \right)^{-1} (BA^{-1}\ell + k)$$

et le point critique sous contrainte x_* par

$$x_* = A^{-1}\left({}^{\mathsf{T}}B \left(BA^{-1}{}^{\mathsf{T}}B \right)^{-1} (BA^{-1}\ell + k) - \ell \right).$$

Dans (39), l'inversibilité de la matrice vaut si B est surjective et A inversible, comme le dit la caractérisation générale du lemme suivant.

LEMME 2.6: *Soit A matrice symétrique d'ordre n et B d'ordre (m, n). La matrice $K = \begin{pmatrix} A & {}^{\mathsf{T}}B \\ B & 0 \end{pmatrix}$ est régulière si et seulement si B est surjective et tout $u \in \ker B$ avec $Au \in (\ker B)^{\perp}$ est nul.*

2.11.2 [DÉMONSTRATION./*Il suffit de caractériser la nullité de $\ker K$. Un vecteur (u, v) est dans ce noyau si et seulement si $Bu = 0$ et $Au + {}^{\mathsf{T}}Bv = 0$. Le lemme résulte de l'égalité $\operatorname{Im} {}^{\mathsf{T}}B = (\ker B)^{\perp}$ et de la surjectivité de B équivalente à l'injectivité de ${}^{\mathsf{T}}B$.* □

2.11.3 Considérons la minimisation de la distance $d(M, P)$ où M, P sont deux points du plan, M sur la droite $x + y = L$ et P sur le cercle de rayon 1 et centré à

l'origine. En considérant le carré $d(M,P)^2$ pour éviter une racine carrée, nous obtenons le problème d'optimisation sous contrainte

$$\min_{a^2+b^2=1, x+y=L} ((x-a)^2+(y-b)^2)$$

Le lagrangien est donc

$$\mathcal{L}(a,b,x,y,\lambda,\mu) = (x-a)^2+(y-b)^2-\lambda(a^2+b^2-1)-\mu(x+y-L)$$

dont l'annulation du gradient donne le système de six équations à six inconnues

$$0=\frac{\partial\mathcal{L}}{\partial a}=2(a-x)-2\lambda a, \qquad 0=\frac{\partial\mathcal{L}}{\partial b}=2(b-y)-2\lambda b,$$
$$0=\frac{\partial\mathcal{L}}{\partial x}=2(x-a)-\mu, \qquad 0=\frac{\partial\mathcal{L}}{\partial y}=2(y-b)-\mu,$$
$$0=\frac{\partial\mathcal{L}}{\partial \lambda}=-a^2-b^2+1, \qquad 0=\frac{\partial\mathcal{L}}{\partial \mu}=-x-y+L.$$

résoluble aisément

$$a_*=b_*=\pm 1/\sqrt{2}, \quad x_*=y_*=L/2, \quad \lambda_*=1\mp L/\sqrt{2}, \mu_*=L\mp\sqrt{2}. \qquad \triangleleft$$

Démonstration. Commençons par le cas où la contrainte $g=0$ est affine (comme c'est le cas en *programmation linéaire*) : il existe $B:\mathbb{R}^n\to\mathbb{R}^m$ et $k\in\mathbb{R}^m$ tels que $g(x)=Bx+k$ pour $x\in\mathbb{R}^n$. La différentielle de g est constante, avec $dg(y)=B, y\in\mathbb{R}^n$. Vu que x_* vérifie la contrainte $g(x_*)=Bx_*+k=0$ on peut réécrire le programme sous la forme

$$\min_{\substack{x\in V\\ B(x-x_*)=0}} [U(x)]$$

Soit $v\in\ker B$. La droite $D_{x_*,v}$ passant par x_* et de direction v est paramétrée suivant $t\in\mathbb{R}\mapsto\gamma_{x_*,v}(t)=x_*+tv$: la restriction de U à un voisinage $\gamma_{x_*,v}([-\varepsilon,\varepsilon])$ de x_* sur la droite $D_{x_*,v}$ a un minimum en x_* : $t=0$ est donc point critique de $U\circ\gamma_{x_*,v}$, soit

$$0=\frac{d(U\circ\gamma_{x_*,v})}{dt}(0)=\langle\nabla U(x_*),\gamma'_{x_*,v}(0)\rangle=\langle\nabla U(x_*),v\rangle.$$

Ainsi, le gradient $\nabla U(x_*)$ est orthogonal à tout vecteur de $\ker B=(\operatorname{Im}{}^{\mathsf{T}}B)^\perp$: comme il a été souligné dans la remarque 2.9.5, c'est un vecteur de $(\operatorname{Im}{}^{\mathsf{T}}B^\perp)^\perp=\operatorname{Im}{}^{\mathsf{T}}B$, ainsi $\nabla U(x_*)$ est une combinaison linéaire des $b_1,\dots,b_m$, et ce de manière unique.

La preuve du cas général fait appel au théorème des fonctions implicites qui donne une vision analytico-géométrique du lieu contraint $g=0$ au voisinage de x_*, sous couvert de l'hypothèse de surjectivité de $dg(x_*)$. Ainsi, le sous-espace linéaire $\ker dg(x_*)$ coïncide avec l'ensemble des vecteurs tangents $\gamma'(0)$ des courbes $\gamma: t\in[-\eta,\eta]\mapsto\gamma(t)$ tracées sur le lieu contraint $g=0$ et passant par x_* avec $\gamma(0)=x_*$. Le caractère critique de l'origine $t=0$ des fonctions $U\circ\gamma$ donne la condition $\langle\nabla U(x_*),v\rangle=0$ pour tout $v\in\ker dg(x_*)$, ce qui permet de conclure comme ci-dessus à $\nabla U(x_*)\in\operatorname{Vect}(\nabla g_1(x_*),\dots,\nabla g_m(x_*))$. $\square$

Comme dans le cas sans contrainte, l'étude au deuxième ordre donne, pour un point critique, une autre condition nécessaire satisfaite par un minimum et une condition suffisante impliquant la qualité de minimum :

Théorème 2.11: *Soit le programme*

$$\min_{\substack{x\in V\\ g(x)=0}} U(x), \tag{40}$$

où $U : V(\text{ouvert de } \mathbb{R}^n) \to \mathbb{R}$ *et* $g : V \to \mathbb{R}^m$ *de classe* $\mathcal{C}^1$ *et son lagrangien associé* $\mathcal{L} = \mathcal{L}_{U,g}$ *défini sur* $V \times \mathbb{R}^m$. *Il est supposé lagrangien* $\mathcal{L}_{U,g}$ *critique en* (x_*, Λ_*).

(i) Si x_* *est un minimum local de* U *sous la contrainte* $g(x) = 0$, *la hessienne* $\mathrm{Hess}_x(\mathcal{L}_{U,g})_{x_*}$ *restreinte à l'espace* $\ker dg_{x_*}$ *est une forme quadratique positive.*

(ii) Si la hessienne $\mathrm{Hess}_x(\mathcal{L}_{U,g})_{x_*}$ *restreinte au sous-espace* $\ker dg_{x_*}$ *est une forme quadratique définie positive, alors* x_* *est un minimum local strict de* U *sous la contrainte* $g = 0$.

Démonstration. Soit $\mathcal{L}$ le lagrangien $\mathcal{L}(x, \Lambda) = U(x) - \langle \Lambda_*, g(x) \rangle$, dont la dérivée $d_x \mathcal{L}_{x_*, \Lambda_*}$ (par rapport à la variable x) est nulle au point (x_*, Λ_*). En omettant la variable Λ fixée à sa valeur en Λ_*, la formule de Taylor à l'ordre 2 pour $\mathcal{L}$ s'écrit

$$\mathcal{L}(x_* + h) = \mathcal{L}(x_*) + \frac{\mathrm{Hess}(\mathcal{L})_{x_*}(h)}{2} + \|h\|^2 \varepsilon(h)$$

pour h au voisinage de 0 dans $\mathbb{R}^n$. Si $x_* + h$ est dans le niveau $S_{g,x_*} = \{x \in U | g(x) = 0\}$, *i. e.* $g(x_* + h) = 0$, cette formule devient

$$U(x_* + h) = U(x_*) + \frac{\mathrm{Hess}(\mathcal{L})_{x_*}(h)}{2} + \|h\|^2 \varepsilon(h)$$

Commençons par le cas où g est linéaire comme dans la démonstration de la proposition précédente. Si x_* est un minimum et en remarquant que $x_* + th \in S_{g,x_*}$ si $h \in \ker dg_{x_*}$, on a pour t petit

$$0 \leq U(x_* + th) - U(x_*) = t^2 \left[\frac{\mathrm{Hess}(\mathcal{L})_{x_*}(h)}{2} + \|h\|^2 \varepsilon(th) \right]$$

soit, en faisant $t \to 0$, $\mathrm{Hess}(\mathcal{L})_{x_*}(h) \geq 0$: la forme quadratique $\mathrm{Hess}(\mathcal{L})_{x_*}$ en restriction au sous-espace $\ker dg_{x_*}$ est positive.

D'autre part, si la hessienne $\mathrm{Hess}(\mathcal{L})_{x_*}$ induit sur $\ker dg_{x_*}$ une forme quadratique définie positive, il existe $m > 0$ tel que

$$\mathrm{Hess}(\mathcal{L})_{x_*}(h) \geq 2m\|h\|^2, \quad h \in \ker dg_x$$

alors, pour h non nul dans $\ker dg_x$

$$U(x_* + h) - U(x_*) = \|h^2\| \left[\frac{\mathrm{Hess}(\mathcal{L})_{x_*}(h/\|h\|)}{2} + \varepsilon(h) \right] \geq \|h^2\| [m + \varepsilon(h)] > 0$$

la dernière égalité valant pour h assez petit. Ainsi x_* est un minimum local de U sous la contrainte $g = 0$.

Le cas g non linéaire se traite comme précédemment en faisant appel au théorème des fonctions implicites. □

△ Remarques 2.10:

(1) La formule de Taylor

$$U(x_* + h) = U(x_*) + \langle \nabla_{x_*} U, h \rangle + \frac{\mathrm{Hess}_x(U)_{x_*}(h)}{2} + \|h\|^2 \varepsilon(h)$$

a le terme $\langle \nabla U_{x_*}, h \rangle$ d'ordre 1 en général non nul, ce qui ne permet pas de faire l'analyse précédente.

(2) Si g est affine, alors la hessienne $\mathrm{Hess}_x(\langle \Lambda, g(x) \rangle)$ est nulle, d'où $\mathrm{Hess}_x(\mathcal{L}_{U,g}) = \mathrm{Hess}_x(U)$, ce qui n'est pas général pour des g non-linéaires. ▽

Une caractérisation de la positivité de la forme quadratique $\mathrm{Hess}_{x_*}(\mathcal{L})$ restreinte à $\ker dg_{x_*}$ en termes de signes de mineurs principaux dominants, analogue à la proposition **??**, est fournie par le théorème suivant (prouvé [14] dans l'appendice A).

THÉORÈME 2.12: *Soit A matrice symétrique d'ordre n et q_A la forme quadratique associée ($q_A(v) = \langle Av, v\rangle$ pour $v \in \mathbb{R}^n$), B matrice d'ordre (m, n) avec $m \leq n$, de rang m et dont le mineur principal dominant d'ordre maximum $B_{mm} = (b_{ij})_{1\leq i,j\leq m}$ est inversible, $q_{A,B}$ la forme quadratique sur $K_B = \ker B$ obtenue par restriction de q_A sur K_B.*

La forme $q_{A,B}$ est définie positive (négative resp.) si et seulement si les déterminants des mineurs principaux dominants de la matrice

$$\begin{pmatrix} 0_m & B \\ {}^{\mathsf{T}}B & A \end{pmatrix} \tag{41}$$

d'ordre $2m + r$ avec $r = 1, \ldots, n - m$ sont non nuls du signe de $(-1)^m$ (du signe de $(-1)^{m+r}$ resp.).

△ REMARQUES 2.11:

(1) La matrice (41) est la matrice hessienne $\mathrm{Hess}_{\Lambda,x}\,\mathcal{L}_q$ du lagrangien

$$\mathcal{L}_q(x, \Lambda) = \frac{\langle Ax, x\rangle}{2} + \langle \Lambda, Bx - k\rangle, \quad x \in \mathbb{R}^n, \Lambda \in \mathbb{R}^m$$

associé au modèle quadratique du programme avec contrainte d'égalité

$$\inf_{Bx=k} \left[\frac{\langle Ax, x\rangle}{2}\right]. \tag{q}$$

Les mineurs principaux dominants sont des déterminants de hessienne partielle $\mathrm{Hess}_{\Lambda,x_1,\ldots,x_{2m+r}}\,\mathcal{L}_q$ pour $r = 1, \ldots, n - m$.

(2) Dans le cas général, la matrice analogue à la matrice (41) est

$$\begin{pmatrix} 0_m & dg(x_*) \\ {}^{\mathsf{T}}dg(x_*) & \mathrm{Hess}_x\,\mathcal{L}(x_*, \Lambda_*) \end{pmatrix} = \begin{pmatrix} 0_m & {}^{\mathsf{T}}\nabla g(x_*) \\ \nabla g(x_*) & \mathrm{Hess}_x\,\mathcal{L}(x_*, \Lambda_*) \end{pmatrix}$$

dite matrice bordante la hessienne $\mathrm{Hess}_x\,\mathcal{L}$: la matrice $\mathrm{Hess}_x\,\mathcal{L}$ est bordée par la matrice dg. La matrice

$$\begin{pmatrix} \mathrm{Hess}_x\,\mathcal{L}(x_*, \Lambda_*) & {}^{\mathsf{T}}dg(x_*) \\ dg(x_*) & 0_m \end{pmatrix}$$

est souvent utilisée de manière équivalente : on passe de l'une à l'autre en échangeant des lignes et des colonnes, ce qui ne modifie pas leurs déterminants. Les déterminants principaux de la première d'ordre $2m + r$ avec $1 \leq r \leq n - m$ s'obtiennent à partir de la seconde en omettant des lignes et colonnes correspondant aux variables de décision x_i d'indice $m + r + 1, \ldots, n$.

(3) Ainsi, les déterminants de hessienne bordée intervenant dans les conditions du deuxième ordre d'un problème d'optimisation sous contrainte d'égalité sont des déterminants de matrice hessienne pour les lagrangiens relativement à des variables $x_1, \ldots, x_r$ avec $r \geq 2m + 1$ et $|d_{(x_1,\ldots,x_m)}g| \neq 0$ et les variables Lagrangiennes $\Lambda = (\lambda_1, \ldots, \lambda_m)$. ▽

Rassemblant les résultats des théorèmes 2.11 (ii) et 2.12, on obtient la condition suffisante au deuxième ordre assurant un minimum local strict en termes de mineurs principaux de matrices bordantes (*i. e.* des matrices extraites de lagrangiens).

14. Pour une démonstration détaillée et complète, la seule référence connue à l'auteur de ces notes est l'article : G. DEBREU, *Definite and semidefinite quadratic forms, Econometrica*, 20#2 (1952) 295-300.

PROPOSITION 2.4: *Soit U ouvert de $\mathbb{R}^n$, $J : x = (x_1, \dots, x_n) \in U \mapsto J(x) \in \mathbb{R}$ et $g = (g_1, \dots, g_m) : U \to \mathbb{R}^m$ régulière imposant m contraintes numériques $g_1 = 0, \dots, g_m = 0$. Soit (x_*, Λ_*) un point critique du lagrangien $\mathcal{L} = \mathcal{L}_{J,g} : (x, \Lambda) \in U \times \mathbb{R}^m \mapsto J(x) - \langle \Lambda, g(x) \rangle \in \mathbb{R}$. Il existe une permutation $y_1, \dots, y_n$ des variables $x_1, \dots, x_n$ telle que la sous-matrice $(\partial_{y_i} g_j)_{1 \le i,j \le m}$ soit inversible en x_*. Pour ces variables y, la restriction de $\mathrm{Hess}_y \mathcal{L}(x_*, \Lambda_*)$ au sous-espace $\ker dg(x_*)$ des directions de contrainte est définie positive (resp. négative) si et seulement si les $n - m$ déterminants*

$$d_{2m+k} = \det \mathrm{Hess}_{y_1,\dots,y_{m+k},\lambda_1,\dots,\lambda_m} \mathcal{L}_{U,g}(x_*, \Lambda_*), \quad k = 1, \dots, n - m \tag{42}$$

sont non nuls et du signe de $(-1)^m$ (resp. du signe de $(-1)^{m+k}$). Dans ces conditions, le point x_* est un minimum local strict de la fonction J au voisinage de x_* sur le lieu des contraintes $\{g = 0\}$.

▷ EXEMPLE 2.12: Soit m, n des entiers avec $m < n$, $J_\pm$ la fonction sur $\mathbb{R}^n$ définie par $\pm J_\pm(x) = x_{m+1}^2 + \dots + x_n^2, x \in \mathbb{R}^n$ et g la contrainte définie par $g(x) = (x_1, \dots, x_m) = x(m) \in \mathbb{R}^m$. La hessienne partielle du lagrangien $J_m(x) - \langle \Lambda, x(m) \rangle$ vis-à-vis des variables ordonnées $(x_1, \lambda_1, \dots, x_m, \lambda_m, x_{m+1}, \dots, x_{m+k})$ une forme diagonale par blocs

$$\mathrm{Hess}_{x_1,\lambda_1,\dots x_m,\lambda_m,x_{m+1},\dots,x_{m+k}} \mathcal{L}_{J_\pm,g} = \begin{pmatrix} D_m(K) & 0_{2m,k} \\ 0_{2m,k} & \pm I_k \end{pmatrix}, \quad k = 1, \dots, n - m \tag{43}$$

où $K = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}$ est la matrice d'ordre 2 de la forme quadratique UV, $D_m(K)$ la matrice diagonale par blocs K d'ordre $2m$, I_ℓ la matrice identité d'ordre ℓ et $0_{k,\ell}$ la matrice nulle à k lignes et ℓ colonnes. La fonction J_+ (resp. J_-) est l' exemple le plus simple d'un minimum (resp. maximum) local strict pour un programme non linéaire avec contrainte d'égalité : on vérifie bien que les déterminants $d^\pm_{2m+k}$ du type (42) et (43)

$$d^\pm_{2m+k} = \begin{vmatrix} 0_m & 0_{m,k} & -I_m \\ 0_{k,m} & \pm I_k & 0_{k,m} \\ -I_m & 0_{m,k} & 0_m \end{vmatrix} = \begin{vmatrix} D_m(K) & 0_{2m,k} \\ 0_{k,2m} & \pm I_k \end{vmatrix} = (-1)^m (\pm 1)^k, \quad k = 1, \dots, n - m$$

sont du signe de $(-1)^m$ pour J_+ et du signe de $(-1)^{m+k}$ pour J_-. ◁

Décrivons en basse dimension les différents cas, synthétisés dans le tableau 4.

— Le cas $n = 2, m = 1$ correspond à l'optimisation d'une fonction restreinte à une courbe du plan : pour un minimum, on a l'unique condition suffisante

$$d_3 = \det \mathrm{Hess}_{x,y,\lambda}(\mathcal{L}) = \begin{vmatrix} \partial_{xx}\mathcal{L} & \partial_{xy}\mathcal{L} & \partial_x g \\ \partial_{yx}\mathcal{L} & \partial_{yy}\mathcal{L} & \partial_y g \\ \partial_x g & \partial_y g & 0 \end{vmatrix} < 0$$

Il aura été supposé $\partial_x g \neq 0$.

— Le cas $n = 3, m = 1$ correspond à l'optimisation d'une fonction restreinte à une surface de l'espace à trois dimensions : les conditions $d_4 < 0, d_3 < 0$ sur les déterminants

$$d_4 = \det \mathrm{Hess}_{x,y,z,\lambda}(\mathcal{L}) = \begin{vmatrix} \partial_{xx}\mathcal{L} & \partial_{xy}\mathcal{L} & \partial_{xz}\mathcal{L} & \partial_x g \\ \partial_{yx}\mathcal{L} & \partial_{yy}\mathcal{L} & \partial_{yz}\mathcal{L} & \partial_y g \\ \partial_{xz}\mathcal{L} & \partial_{yz}\mathcal{L} & \partial_{zz}\mathcal{L} & \partial_z g \\ \partial_x g & \partial_y g & \partial_z g & 0 \end{vmatrix}$$

$$d_3 = \det \mathrm{Hess}_{x,y,\lambda}(\mathcal{L}) = \begin{vmatrix} \partial_{xx}\mathcal{L} & \partial_{xy}\mathcal{L} & \partial_x g \\ \partial_{yx}\mathcal{L} & \partial_{yy}\mathcal{L} & \partial_y g \\ \partial_x g & \partial_y g & 0 \end{vmatrix}$$

sont suffisantes pour un minimum local strict. On aura supposé $\partial_x g$ non nul.

— Le cas $n = 3, m = 2$ correspond à l'optimisation d'une fonction sur la courbe $\{g = 0, h = 0\}$ de l'espace à trois dimensions : la condition de minimum est unique, soit la positivité du déterminant d_5 d'ordre 5

$$d_5 = \det \mathrm{Hess}_{x,y,z,\lambda,\mu}(\mathcal{L}) = \begin{vmatrix} \partial_{xx}\mathcal{L} & \partial_{xy}\mathcal{L} & \partial_{xz}\mathcal{L} & \partial_x g & \partial_x h \\ \partial_{yx}\mathcal{L} & \partial_{yy}\mathcal{L} & \partial_{yz}\mathcal{L} & \partial_y g & \partial_y h \\ \partial_{xz}\mathcal{L} & \partial_{yz}\mathcal{L} & \partial_{zz}\mathcal{L} & \partial_z g & \partial_z h \\ \partial_x g & \partial_y g & \partial_z g & 0 & 0 \\ \partial_x h & \partial_y h & \partial_z h & 0 & 0 \end{vmatrix}.$$

On aura supposé $\begin{vmatrix} \partial_x g & \partial_y g \\ \partial_x h & \partial_y h \end{vmatrix}$ non nul : g et h désignent les deux contraintes (numériques) de ce problème d'optimisation, avec lagrangien $\mathcal{L}(x, y, z, \lambda, \mu) = U(x, y, z) - \lambda g(x, y, z) - \mu h(x, y, z)$.

Notons les cas non couverts par le tableau où les énoncés généraux indiquent des points critiques soit dégénérés (un déterminant nul), soit de type selle (en dimension 2 avec une hessienne non dégénérée, le déterminant de la hessienne est négatif). Ces cas ne seront pas développés plus avant ici.

n	m	signe déterminant	optimum
1	0	$d_1 > 0$	minimum
1	0	$d_1 < 0$	maximum
2	0	$d_2 > 0, d_1 > 0$	minimum
2	0	$d_2 > 0, d_1 < 0$	maximum
2	1	$d_3 < 0$	minimum
2	1	$d_3 > 0$	maximum
3	0	$d_3 > 0, d_2 > 0, d_1 > 0$	minimum
3	0	$d_3 < 0, d_2 > 0, d_1 < 0$	maximum
3	1	$d_4 < 0, d_3 < 0$	minimum
3	1	$d_4 < 0, d_3 > 0$	maximum
3	2	$d_5 > 0$	minimum
3	2	$d_5 < 0$	maximum

TABLE 4. Des conditions suffisantes d'extremum en $n \leq 3$ variables de choix et sous m contraintes : l'indice r dans d_r fait référence à l'ordre du déterminant mentionné.

▷ EXEMPLES 2.13:

2.13.1 Pour la distance entre un point sur un cercle et un autre sur une droite ne rencontrant pas ce cercle (cf. exemple 2.11.4), la hessienne du lagrangien $\mathcal{L}(a, x, b, y, \lambda, \mu) = (x - a)^2 + (y - b)^2 - \lambda(a^2 + b^2 - 1) - \mu(x + y - L)$ est

$$\mathrm{Hess}_{a,x,b,y,\lambda,\mu}\,\mathcal{L} = \begin{pmatrix} 2-2\lambda & -2 & 0 & 0 & -2a & 0 \\ -2 & 2 & 0 & 0 & 0 & -1 \\ 0 & 0 & 2-2\lambda & -2 & -2b & 0 \\ 0 & 0 & -2 & 2 & 0 & -1 \\ -2a & 0 & -2b & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 \end{pmatrix}$$

de déterminant $d_6 = 16[(a-b)^2 - 2\lambda(a^2+b^2)]$, avec

$$d_5 = |\mathrm{Hess}_{a,x,b,\lambda,\mu}\,\mathcal{L}| = \begin{vmatrix} 2-2\lambda & -2 & 0 & -2a & 0 \\ -2 & 2 & 0 & 0 & -1 \\ 0 & 0 & 2-2\lambda & -2b & 0 \\ -2a & 0 & -2b & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \end{vmatrix} = 8(a^2+b^2)(1-\lambda).$$

La condition de régularité du système de variables pour les contraintes g_1, g_2 est bien observée : la matrice carrée $\begin{pmatrix} \partial_a g_1 & \partial_x g_1 \\ \partial_a g_2 & \partial_x g_2 \end{pmatrix}$ est non singulière. Cela n'aurait pas été le cas si l'on avait choisi de calculer la hessienne suivant l'ordre des variables a, b, x, y, λ, μ.

Ainsi, pour $(M_+, P) = ((1/\sqrt{2}, 1/\sqrt{2}), (L/2, L/2))$, on a $\lambda = 1 - L/\sqrt{2} < 0$, d'où $d_5 > 0$ et $d_4 = L/\sqrt{2} > 0$, ce qui indique un minimum local, alors qu'à l'autre point critique $(M_-, P) = (-M_+, P)$ $\lambda = 1 + L/\sqrt{2} > 0$, d'où $d_5 < 0$ ce qui exclut le caractère défini pour la forme. La condition $d_{a,x}g$ inversible est bien vérifié aux points critiques (M, P, Λ). Remarquons cependant que si $|d_{a,x}| = 2a$ est nul, alors $|d_{b,x}| = 2b$ ne l'est pas vu que $a^2 + b^2 = 1$: le choix de « premières variables » vérifiant les conditions de surjectivité pour dg est toujours possible !

2.13.2 Soit la fonction d'objectif U définie par $U(x, y, z) = x^3+y^3+z^3$ et la contrainte $x^{-1} + y^{-1} + z^{-1} = 1$. Le lagrangien $\mathcal{L}$ associé à ce problème d'optimisation a comme dérivée partielle $\partial_x\mathcal{L}(x, y, z, \lambda) = 3x^2 + \lambda x^{-2}$: les points critiques de $\mathcal{L}$ sont solution de

$$x^4 = y^4 = z^4 = -\lambda/3, \quad x^{-1} + y^{-1} + z^{-1} = 1$$

et la matrice hessienne bordée est

$$\mathrm{Hess}_{(x,y,z),\lambda}\,\mathcal{L} = \begin{pmatrix} 6x - 2\lambda x^{-3} & 0 & 0 & x^{-2} \\ 0 & 6y - 2\lambda y^{-3} & 0 & y^{-2} \\ 0 & 0 & 6z - 2\lambda z^{-3} & z^{-2} \\ x^{-2} & y^{-2} & z^{-2} & 0 \end{pmatrix}$$

Les points critiques de $\mathcal{L}$ sont donc $M_* = (3, 3, 3), \lambda_* = -243$; $N_{3*} = (1, 1, -1), \mu_* = -3$ et les deux autres solutions N_{1*}, N_{2*} obtenues par permutation de x, y, z. Dans le premier cas et second cas, les matrices hessiennes bordées sont

$$\mathrm{Hess}(\mathcal{L})_{(M_*,\lambda_*)} = \begin{pmatrix} 36 & 0 & 0 & \frac{1}{9} \\ 0 & 36 & 0 & \frac{1}{9} \\ 0 & 0 & 36 & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & 0 \end{pmatrix} \quad \mathrm{Hess}(\mathcal{L})_{(N_{*3},\mu_*)} = \begin{pmatrix} 12 & 0 & 0 & 1 \\ 0 & 12 & 0 & 1 \\ 0 & 0 & -12 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

avec mineurs principaux dominants de contrainte en M_*

$$d_4 = -48, \quad d_3 = \begin{vmatrix} 36 & 0 & \frac{1}{9} \\ 0 & 36 & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & 0 \end{vmatrix} = -\frac{8}{9},$$

et en N_{3*}

$$d_4 = \left|\mathrm{Hess}(\mathcal{L})_{(N_{*3},\mu_*)}\right| = 144, \quad d_3 = \begin{vmatrix} 12 & 0 & 1 \\ 0 & -12 & 1 \\ 1 & 1 & 0 \end{vmatrix} = 0.$$

Dans le premier cas, le point critique est un minimum local, alors que dans le second on a un point selle.

2.13.3 En deux variables, *i. e.* pour $U(x,y)=x^3+y^3$ sous la contrainte $x^{-1}+y^{-1}=1$, on a des résultats analogues : un seul point critique en $(2,2)$. Le tracé des courbes de niveau de U et du lieu contraint $\mathcal{G}=\{g=0\}$ (l'hyperbole $x+y=xy$ privée de l'origine, courbe à trois composantes connexes) donne un autre éclairage des résultats (cf. Fig. II.5). On a pour la hessienne bordée en $(M_*,\lambda_*)=((2,2),48)$

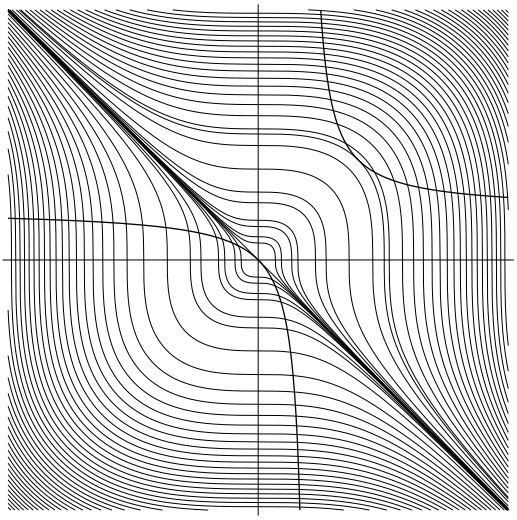


FIGURE II.5 . Les courbes de niveaux $x^3+y^3=h$ et la courbe $x^{-1}+y^{-1}=1$.

$$d_3=\left|\det \operatorname{Hess}_{(M_*,\lambda_*)}\mathcal{L}\right|=-\frac{195}{2},$$

ce qui permet d'affirmer que $(2,2)$ est un minimum local strict : c'est un minimum pour x^3+y^3 global sur la composante de $\mathcal{G}$ contenant $(2,2)$, mais pas sur les autres composantes (où la fonction varie entre 0 et $\pm\infty$). ◁

Le minimum x_* du programme est, avec son multiplicateur de Lagrange Λ_* un point critique du lagrangien $\mathcal{L}$, *i. e.* un zéro de $\nabla_{x,\Lambda}\mathcal{L}=(\nabla_x U-\langle\Lambda,\nabla g\rangle,g)$. La différentielle de ce gradient est

$$K=\begin{pmatrix}\operatorname{Hess}_x\mathcal{L} & -{}^{\mathsf{T}}\nabla g\\ -\nabla g & 0\end{pmatrix}.$$

△ REMARQUE 2.12: La régularité de A et la surjectivité de B ne sont pas suffisantes pour la régularité de $K=\begin{pmatrix}A & {}^{\mathsf{T}}B\\ B & 0\end{pmatrix}$, comme l'exemple $A=\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$ et $B=(-1,1)$ le montrent. À l'inverse, si B est surjective, alors K est régulière dès que la restriction de A à $\ker B$ est définie : dans ce cas, la forme quadratique associée à K a au moins m valeurs propres positives et m négatives. ▽

La conjonction du point de minimum x_* et de son multiplicateur Λ_* donne un zéro du gradient $\nabla_{x,\Lambda}\mathcal{L}$: reprenant le théorème général 2.3 d'approximation du zéro X_* d'une fonction F, nous avons le cas particulier dit de Newton-Lagrange, où la condition $F'(X_*)$ inversible est simplement l'inversibilité de la hessienne $\operatorname{Hess}\mathcal{L}(x_*,\Lambda_*)$:

THÉORÈME 2.13 (Newton-Lagrange): *Soient U et g de classe $\mathcal{C}^3$ dans le voisinage du point stationnaire x_* avec multiplicateur de Lagrange Λ_* relativement au lagrangien*

$\mathcal{L}(x,\Lambda) = U(x) - \langle \Lambda, g(x)\rangle$. *Supposons* $\mathrm{Hess}_{x,\Lambda}\,\mathcal{L}(x_*,\Lambda_*)$ *inversible. Il existe alors un voisinage* V *de* (x_*,Λ_*) *tel que l'itération de Newton* $(x_k,\Lambda_k)_{k\geq 0}$ *avec* $(x_0,\Lambda_0)\in V$ *et pour* $k\geq 0$

$$\begin{pmatrix} x_{k+1}\\ \Lambda_{k+1}\end{pmatrix} = \begin{pmatrix} x_k\\ \Lambda_k\end{pmatrix} - \begin{pmatrix} \mathrm{Hess}_x\,\mathcal{L}(x_k,\Lambda_k) & -{}^{\mathsf{T}}\nabla_x g(x_k)\\ -\nabla_x g(x_k) & 0\end{pmatrix}^{-1}\begin{pmatrix} \nabla_x U(x_k) - \langle \Lambda_k, \nabla_x g(x_k)\rangle\\ -g(x_k)\end{pmatrix},$$

est définie et convergente quadratiquement vers (x_*,Λ_*).

△ Remarques 2.13:

(1) L'itération de Newton provient d'un problème quadratique induit par le développement de Taylor du Lagrangien et de la contrainte

$$\mathcal{L}(x_*,\Lambda_*) = \mathcal{L}(x,\Lambda_*) + \nabla_x\mathcal{L}(x,\Lambda_*)(d) + \frac{\mathrm{Hess}_x\,\mathcal{L}(x,\Lambda_*)(d)}{2} + \dots$$
$$g(x_*) = g(x) + \langle \nabla g(x), d\rangle + \dots$$

où on a noté $d = x - x_*$. Le problème avec variable de décision d

$$\inf_{g(x)+\langle\nabla g(x),d\rangle=0}\left[\nabla_x\mathcal{L}(x,\Lambda_*)(d) + \frac{\mathrm{Hess}_x\,\mathcal{L}(x,\Lambda_*)(d)}{2}\right]$$

a comme équation de Lagrange

$$\begin{pmatrix} A & {}^{\mathsf{T}}B\\ B & 0\end{pmatrix}\begin{pmatrix} x\\ -\Lambda\end{pmatrix} = \begin{pmatrix} -\ell\\ B\end{pmatrix}.$$

(2) L'itéré x_k n'est pas nécessairement dans le lieu des contraintes $g(x)=0$.

(3) Le x_0 initialisant l'itération de Newton-Lagrange est soit pris (en toute confiance) au hasard, soit obtenu comme approximation de x_* par une recherche stochastique appropriée. Pour le Λ_0 initial, on ne peut utiliser d'équation du type $\nabla J(x_0) - \langle \Lambda_0, \nabla g(x_0)\rangle$ qui n'est valable en général que pour x_*. Comme approximation, on cherchera Λ_0 comme le point de minimum du programme (quadratique) de minimisation $\min_\Lambda \|\nabla J(x_0) - \langle \Lambda, \nabla g(x_0)\rangle\|^2$.

(4) La résolution de programmes avec contraintes d'inégalité se ramène à une famille de programmes avec contraintes d'égalités, suivant les faces de l'ensemble admissible ou par pénalisations de la fonction U incorporant les contraintes d'inégalités la remplaçant par $U_\varepsilon = U - \varepsilon^{-1}\sum_j \log(g_j)$. D'autres méthodes de traitement de ces derniers existent : on se limite ici à la seule méthode de Newton. ▽

▷ Exemples 2.14:

2.14.1 Dans l'exemple 1.2. ? ? ? initié par une recherche stochastique, il a été considéré que la solution était dans le domaine des contraintes d'équation $g_1 = 0 \wedge g_2 = 0$ et l'itération de Newton-Lagrange a été lancée : cela a amélioré le point de minimum obtenu par recherche stochastique. ◁

7.2. Optimisation avec contraintes d'inégalités. Cette section est dédiée à l'étude du problème d'optimisation

$$\min_{\substack{x\in V\\ h_j(x)\geq 0, j=1,\dots,p}} U(x)$$

où V est un ouvert de $\mathbb{R}^n$.

On supposera que le domaine des réalisables $\{h_j(x)\geq 0\}$ est non vide (ce qui parfois n'est pas si facile à montrer) et que le point x_* de minimum est dans ce domaine.

Algorithme 2.13 Newton-Lagrange pour programme avec contrainte d'égalité

Choisir x_0, Λ_0 ; $k = 0$
Calculer $g(x_0), \nabla g(x_0),\ \nabla U(x_0)$ et $\nabla_x \mathcal{L}(x_0, \Lambda_0) = \nabla_x U(x_0) - \langle \Lambda_0, \nabla_x g(x_0) \rangle$
tant que $\|\nabla_x \mathcal{L}(x_k, \Lambda_k)\| + \|g(x_k)\| \geq \varepsilon$ & $k \leq K$ **faire**
Calculer $\mathrm{Hess}_x\, \mathcal{L}(x_k, \Lambda_k)$
Calculer $\begin{pmatrix} x_{k+1} \\ \Lambda_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ \Lambda_k \end{pmatrix} - \begin{pmatrix} \mathrm{Hess}_x\, \mathcal{L}(x_k, \Lambda_k) & -{}^{\mathsf{T}}\nabla g(x_k) \\ \nabla g(x_k) & 0 \end{pmatrix}^{-1} \begin{pmatrix} \nabla U(x_k) - \langle \Lambda_k, \nabla_x g(x_k) \rangle \\ g(x_k) \end{pmatrix}$
Calculer $g(x_{k+1}), \nabla g(x_{k+1}),\ \nabla U(x_{k+1})$ et $\nabla_x \mathcal{L}(x_{k+1}, \Lambda_{k+1})$
$k = k + 1$
fin tant que

DÉFINITION 2.7: *Soient pour $j = 1, \ldots, p$ des fonctions $h_j : V \to \mathbb{R}$. La contrainte $h_j(x) \geq 0$ est dite* saturée *en $x_* \in V$, et le point x_** actif *pour la contrainte h_j, si $h_j(x_*) = 0$; on note S_{x_*} l'ensemble des indices $j \in \{1, \ldots, p\}$ tels que la contrainte $h_j(x) \geq 0$ soit saturée en x_*. Ainsi, si $j \notin S_{x_*}$, $h_j(x_*) > 0$.*

Les contraintes $h_j(x) \geq 0, j = 1, \ldots, p$ sont régulières *en x_* si la matrice jacobienne de $h_{S_{x_*}} = (h_j)_{j \in S_{x_*}}$ est de rang $\#S_{x_*}$* i. e. *si les gradients $\nabla h_j(x_*), j \in S_{x_*}$ sont linéairement indépendants.*

THÉORÈME 2.14 (Karush-Kuhn-Tucker[20, 22]): *Soit $U : V (\text{ouvert de } \mathbb{R}^n) \to \mathbb{R}$ et $h_j : V \to \mathbb{R}^m$ de classe $\mathcal{C}^1$. Soit x_* un point régulier pour les contraintes $h_j(x) \geq 0, j = 1, \ldots, p$,* i. e. *les gradients $(\nabla h_j(x_*))$ linéairement indépendants. Si x_* est un minimum de U sur V avec contraintes d'inégalités*

$$\min_{\substack{x \in V \\ h_j(x) \geq 0, j = 1, \ldots, p}} U(x) \tag{44}$$

il existe alors un unique $\Lambda_* = (\lambda_{*1}, \ldots, \lambda_{*p})$ tels que
— *le lagrangien $\mathcal{L}(x, \lambda_1, \ldots, \lambda_p) = U(x) - \sum_{j=1}^p \lambda_j h_j(x)$ est critique en les variables x au point (x_*, Λ_*), soit*

$$\nabla_x \mathcal{L}(x_*, \Lambda_*) = \nabla_x U(x_*) - \sum_{j=1}^p \lambda_{*j} \nabla h_j(x_*) = 0, \tag{45}$$

— *$\lambda_{*j} = 0$ si x_* n'est pas actif pour la contrainte h_j ;*
— *$\lambda_{*j} \geq 0$ si x_* est actif pour la contrainte h_j.*

△ REMARQUES 2.14:

(1) Les conditions nécessaires du théorème précédent peuvent se réécrire

$$\nabla_x U(x_*) = \sum_{j \in S_{x_*}} \lambda_{*j} \nabla h_j(x_*),$$
$$\lambda_j \geq 0, \quad j \in S_{x_*}, \qquad \lambda_j = 0, \quad j \notin S_{x_*}.$$

(2) En introduisant des variables s_j, écrivons le lagrangien associé au problème d'optimisation avec seules contraintes d'égalités, équivalent au problème (44)

$$\min_{h_j(x) - s_j^2 = 0, j = 1, \ldots, p} U(x)$$

avec gradient pour le lagrangien $\mathcal{L}(x,s,\Lambda) = U(x) - \sum_{j=1}^p \lambda_j(h_j(x) - s_j^2)$

$$\nabla_{x,(s_j),(\lambda_j)}\mathcal{L} = \left(\partial_{x_i}U - \sum_{k=1}^p \lambda_k \partial_{x_i} h_k, (2\lambda_j s_j)_j, (h_j - s_j^2)_j\right).$$

Les conditions d'annulation du gradient $\nabla_{x,(s_j),(\lambda_j)}U$ du théorème 2.10 redonne les conditions d'annulation du théorème 2.14 (mais pas les conditions de positivité sur les λ_j).

(3) Pour un problème de maximisation avec contraintes d'inégalités,

$$\max_{h_j(x)\geq 0, j=1,\dots,p} U(x)$$

les conditions (nécessaires) KKT prennent la forme de l'existence de multiplicateurs de Lagrange $\Lambda_* = (\lambda_{*1},\dots,\lambda_{*p})$ tels que $\nabla_x\mathcal{L}(x_*,\Lambda_*) = 0$ si $\mathcal{L} = U - \sum_{j=1}^p \lambda_j h_j$, avec $\lambda_j = 0$ si x_* est inactif pour la contrainte $h_j \geq 0$ et $\lambda_j \leq 0$ sinon. En effet, c'est un problème de minimisation pour $-U$ et $\mathcal{L}_{-U,\Lambda} = -\mathcal{L}_{U,-\Lambda}$.

(4) Cette condition est nécessaire, mais nullement suffisante : la fonction $U(x,y) = y - x^2$ en $M_* = (0,0)$ avec la contrainte $h(x,y) = y \geq 0$ vérifie les conditions KKT, sans que M_* ne soit ni un minimum (M_* est un maximum sur le bord $y = 0$), ni un maximum (M_* est un minimum sur la direction transverse $\{(0,y), y \geq 0\}$), cf. figure II.6. ▽

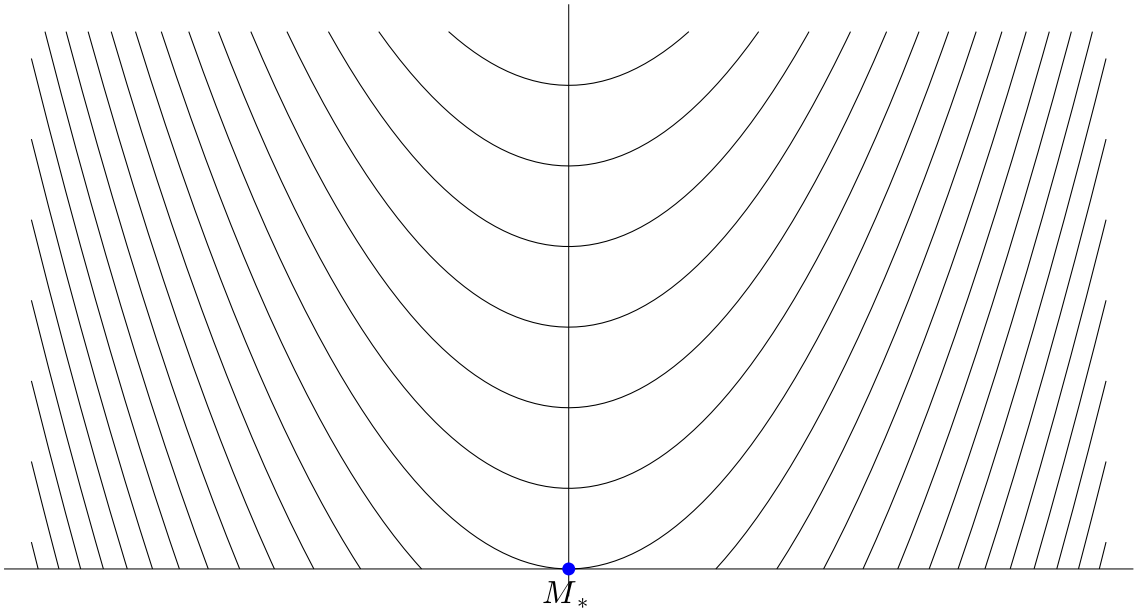


FIGURE II.6 . Les lignes de niveau de $U(x,y) = y - x^2$ sur le domaine $y \geq 0$ et le point M_* vérifiant KKT.

▷ EXEMPLE 2.15: Optimisons $U(x,y) = x + y$ sous les contraintes $xy \geq 2$ et $y \leq -2x + 5$. Le lagrangien

$$\mathcal{L}(x,y,\lambda,\mu) = x + y - \lambda(xy - 2) - \mu(5 - 2x - y)$$

a pour gradient

$$\nabla_{(x,y)}\mathcal{L} = (1 - \lambda y + 2\mu, 1 - \lambda x + \mu) \tag{46}$$

Soit (x, y, λ, μ) annulant ce gradient. Examinons les différents cas :

- en un point intérieur, *i. e.* non actif pour aucune des contraintes, on a $\lambda = 0 = \mu$, ce qui est incompatible avec l'annulation du gradient (46) (une fonction linéaire non nulle n'a pas de point critique) ;
- en un point actif pour la première contrainte seule (*i. e.* sur l'hyperbole $xy = 2$), on a $\mu = 0$ et la condition de nullité (46) donne les solutions $M_\pm = (\pm\sqrt{2}, \pm\sqrt{2})$ avec multiplicateur $\lambda_\pm = \pm 1/\sqrt{2}$: (M_+, λ_+) satisfait la condition KKT pour un minimum, (M_-, λ_-) pour un maximum. ;
- en un point actif pour la seconde contrainte seule (*i. e.* sur la droite $x + 2y = 5$), on a $\lambda = 0$ et la condition de nullité (46) est impossible ;
- les points $P = (2, 1)$ et $Q = (1/2, 4)$ sont les seuls points actifs simultanément pour les deux contraintes : les conditions d'annulation lagrangienne (46) donnent $(\lambda_P, \mu_P) = (1/3, -1/3)$ en P et $(\lambda_Q, \mu_Q) = (-1/3, -7/6)$ en Q : Q (et les multiplicateurs (λ_P, μ_P)) vérifie la condition nécessaire KKT pour un maximum, alors que P ne vérifie ni la condition pour un minimum, ni pour un maximum.

On a $U(P) = 3, U(Q) = 9/2, U(M_\pm) = \pm 2\sqrt{2}$: on en déduit Q maximum et M_+ minimum de U sur la composante bornée de $xy \geq 2, 2x + y \leq 5$, alors que U est majorée par $U(M_-)$ avec infimum $-\infty$ sur l'autre composante. ◁

Démonstration. Si x_* n'est pas actif pour la contrainte h_ℓ, *i. e.* $h_\ell(x_*) > 0$, alors, vu que $h_\ell(y) > 0$ pour y au voisinage de x_*, x_* est un minimum local de U du problème où on a omis la contrainte h_ℓ : il suffit d'établir le théorème pour un problème où x_* est actif pour toutes les contraintes, à charge de rajouter les multiplicateurs de Lagrange λ_ℓ nuls correspondants au contraintes non saturées en x_*.

Comme dans la preuve du théorème 2.10, supposons dans un premier temps les contraintes linéaires : $h_j(x) = h_j(x_*) + \langle b_j, x - x_* \rangle$ avec $b_j = \nabla h_j(x_*)$. Si w vérifie $\langle w, \nabla h_j(x_*) \rangle \geq 0$ pour $j \in S_{x_*}$, alors, au voisinage de x_*, la demi-droite $t \geq 0 \mapsto x_* + tw$ est dans le domaine contraint et la dérivée à droite en $t = 0$

$$\frac{d}{dt}[U(x_* + tw)]_{t=0^+} = \langle \nabla U_{x_*}, w \rangle$$

est positive, soit $\langle \nabla U_{x_*}, w \rangle \geq 0$. Ainsi d'après le Lemme de Farkas-Minkowski, ∇U_{x_*} est combinaison linéaire à coefficients positifs des $b_j = \nabla h_j(x_*)$ où $j \in S_{x_*}$.

Le cas de contraintes non linéaires en inégalités pour des points réguliers est traité avec le théorème des fonctions implicites. □

Lemme 2.7 (Lemme de Farkas-Minkowski): *Soit $b_1, \ldots, b_p, v$ vecteurs de $\mathbb{R}^n$. Si lorsque tous les $\langle w, b_i \rangle$ sont positifs ou nuls il en est de même pour $\langle w, v \rangle$, alors v une combinaison linéaire des b_i à coefficients positifs.*

△ Remarque 2.15: Si on définit le dual conique positif K^+ de la partie K suivant $K^+ = \{w | \langle w, k \rangle \geq 0, k \in K\}$, le lemme de Farkas-Minkowski énonce l'égalité $K = (K^+)^+$ si K est le cône $K = \{\lambda_1 b_1 + \ldots + \lambda_m b_m | \lambda_j \geq 0, j = 1, \ldots, m\}$. ▽

Démonstration. Montrons tout d'abord ce lemme dans le cas où la famille $\mathbf{b} = (b_k)$ est une base de l'espace E : $v = \sum_j \lambda_j b_j$. Soit $\widetilde{\mathbf{b}} = (\widetilde{b}_\ell)$ la base duale de $\mathbf{b}$, *i. e.* l'unique famille $\widetilde{\mathbf{b}}$ telle que $\langle \widetilde{b}_k, b_\ell \rangle = \delta_{k\ell}$. On a tous les $\langle b_k, \widetilde{b}_\ell \rangle \geq 0$ et donc par suite aussi $\langle v, \widetilde{b}_j \rangle = \lambda_j \geq 0$.

Dans le cas général, nous allons utiliser le théorème de Hahn-Banach géométrique qui énonce l'existence d'une forme affine ℓ séparant strictement un point v et un cône C

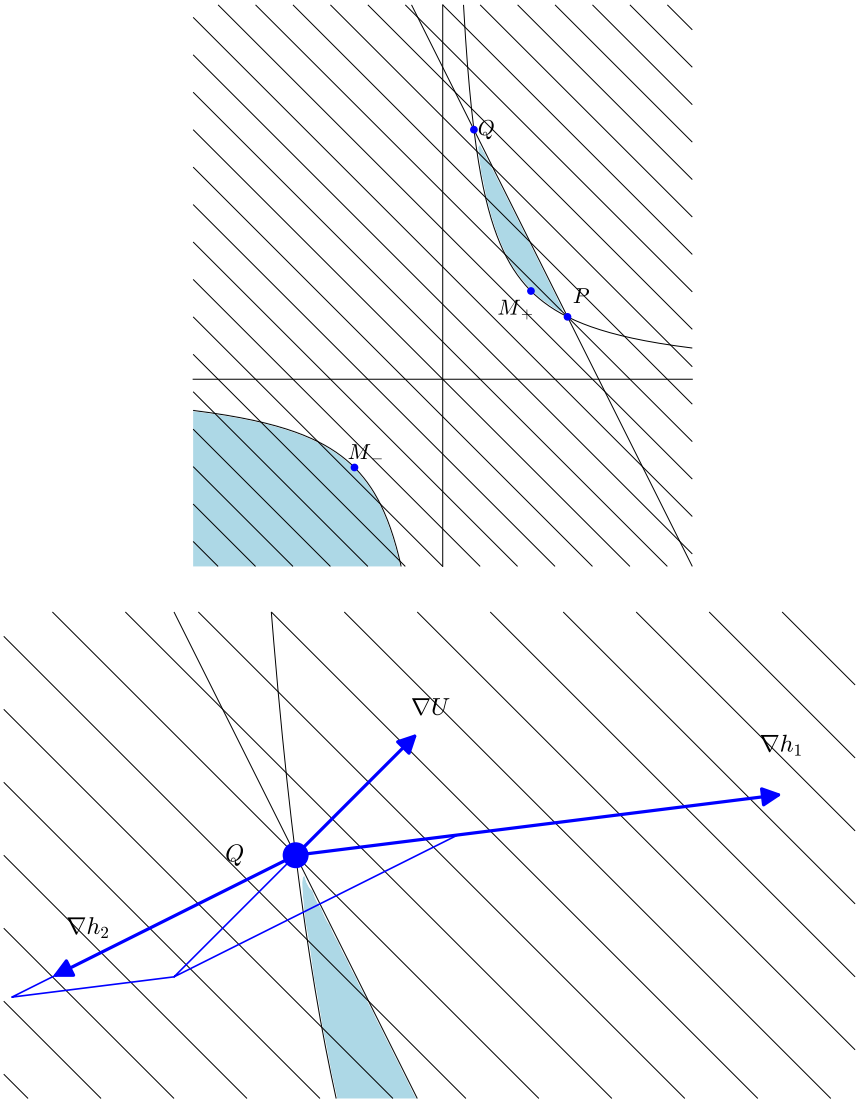


FIGURE II.7 . Le domaine admissible déterminé par les contraintes $xy \geq 2$ et $2x+y \leq 5$ et les gradients en Q : $6\nabla U = -2\nabla h_1 - 7\nabla h_2$, condition KKT de signe sur les multiplicateurs de Lagrange $(\lambda = -2, \mu = -7)$ associée à un maximum.

convexe fermé ne contenant pas ce point, en complément du Lemme de la fermeture du cône démontré *infra* :

LEMME 2.8 (Lemme de la fermeture du cône): *Soit* $\mathbf{v} = (v_1, \ldots, v_m)$ *une famille de* m *vecteurs de l'espace vectoriel* E*. Alors la partie* $C_{\mathbf{v}} = \{\sum_{i=1}^{m} \alpha_i v_i, \alpha_i \geq 0\}$ *est un cône convexe fermé.*

Soit donc $C_{\mathbf{v}}$ qui est un cône convexe fermé d'après le lemme précédent. S'il ne contient pas b, alors Hahn-Banach assure de l'existence de ℓ tel que $\ell(b) > \alpha > \ell(v)$ pour tout $v \in C_{\mathbf{v}}$. Vu que $0 \in C_{\mathbf{v}}$, on a $\alpha > 0$ et donc aussi $\ell(b) > 0$. Soit λ représentant ℓ : on a $\ell(v) = \langle \lambda, v \rangle$. Tous les $\langle \lambda, v_i \rangle$ sont négatifs : en effet, d'une part $v + \rho v_i$ est dans le cône $C_{\mathbf{v}}$ si $\rho \geq 0$ et $v \in C_{\mathbf{v}}$, d'autre part $\langle \lambda, v \rangle = \ell(v)$ est bornée sur $C_{\mathbf{v}}$ par α. En résulte, d'après l'hypothèse, que $\langle \lambda, b \rangle = \ell(b) \leq 0$, ce qui n'est pas. Ainsi, nous avons montré l'appartenance de b au cône $C_{\mathbf{v}}$. □

Preuve du Lemme de la fermeture du cône. La partie C est un cône convexe : reste à montrer qu'elle est fermée.

Commençons par le montrer lorsque les $v_1, \dots, v_n$ sont linéairement indépendants, auquel cas on peut se placer dans le sous-espace $\widetilde{E}$ engendré (librement) par ces vecteurs. Les coordonnées (ξ_j) de $x \in C$ relativement à la base $(v_1, \dots, v_n)$ sont données par le produit $\langle x, e_j \rangle$ avec la base duale (e_j) de (v_j) : une suite de vecteurs dans $\widetilde{E}$ converge si et seulement si les suites de coordonnées convergent aussi. Ainsi, un point adhérent à C a des coordonnées positives relativement à la base (v_j) et est donc dans C.

Supposons les (v_j) linéairement dépendants. Il suffit de montrer tout d'abord que tout z dans C est combinaison linéaire à coefficients positifs de vecteurs $v_K = (v_k)_{k\in K}$ linéairement indépendants avec $K \subset \{1, \dots, m\}$, *i. e.* vecteur d'un certain C_K (qui est partie du cône C) : ainsi le cône C, union $C = \cup_K C_K$ de cônes fermés pour un nombre fini de parties K de $\{1, \dots, n\}$, est fermé.

Soit donc $v = \sum_{i\in I} \alpha_i v_i \in C$ et $\sum_{i\in I_+} \beta_i v_i - \sum_{i\in I_-} \beta_i v_i$ une combinaison linéaire du vecteur nul avec les $\beta_i > 0$ et I_-, I_+ deux parties de $\{1, \dots, m\}$, disjointes et d'union non vide. Alors

$$v = \sum_{i\in I_+} (\alpha_i - \rho\beta_i) v_i + \sum_{i\in I_-} (\alpha_i + \rho\beta_i) v_i + \sum_{i\notin I_-\cup I_+} \alpha_i v_i$$

et on peut choisir un ρ tel que tous les coefficients de cette combinaison linéaire sont positifs et au moins un est nul : si I_+ est vide on prend $\rho = \max_{i\in I_-} \alpha_i/\beta_i$ sinon on prend $\rho = \min_{i\in I_+} \alpha_i/\beta_i$. Ainsi le vecteur v apparaît comme combinaison linéaire d'au plus $\#I - 1$ vecteurs parmi les $\{v_i, i \in I\}$: réitérant cette construction, on aboutit nécessairement à une expression de v comme combinaison linéaire d'une famille $(v_k)_{k\in K}$ libre. □

Il est courant d'avoir des contraintes de positivité sur les variables de choix : il est utile d'avoir un énoncé spécifique pour cette situation [15]

PROPOSITION 2.5: *Soit x_* une solution du programme*

$$\min_{\substack{x\in V \\ h_j(x)\geq 0, j=1,\dots,p \\ x_k \geq 0, k=1,\dots q}} U(x_1, \dots, x_n).$$

On suppose que x_* est actif pour toutes les contraintes, régulier relativement à l'ensemble de ces contraintes. Soit $\mathcal{L}$ le lagrangien

$$\mathcal{L}(x_1, \dots, x_n, \lambda_1, \dots, \lambda_p) = U(x_1, \dots, x_n) - \lambda_1 h_1(x) - \dots - \lambda_p h_p(x)$$

Alors il existe $\Lambda_* = (\lambda_{*1}, \dots, \lambda_{*p})$ tels que

$$\partial_{x_k}\mathcal{L}(x_*, \Lambda_*) \geq 0 \text{ si } x_k^* = 0 \text{ et } k \in [1, q], \quad \partial_{x_k}\mathcal{L}(x_*, \Lambda_*) = 0 \text{ sinon,}$$
$$\lambda_j^* \geq 0 \text{ si } h_j(x_*) = 0, \quad \lambda_j^* = 0 \text{ si } h_j(x_*) > 0, \qquad j = 1, \dots, p.$$

DÉMONSTRATION. Le théorème précédent s'applique au lagrangien

$$\mathcal{M}(x, \lambda_1 \dots, \lambda_p, \mu_1, \dots, \mu_q) = U(x) - \sum_{j=1}^{p} \lambda_j h_j(x) - \sum_{k=1}^{q} \mu_k x_k.$$

Le lagrangien $\mathcal{M}$ est critique : en particulier, si k est l'indice d'une contrainte sur une coordonnée, $\partial_{x_k}\mathcal{M} = \partial_{x_k}\mathcal{L} - \mu_k$ est nul en (x_*, Λ_*, M_*). De plus $\mu_k \geq 0$: en résulte

15. On adaptera l'énoncé lorsque les contraintes de positivité portent sur des variables d'indice dans la partie $K \subset \{1, \dots n\}$.

$\partial_{x_k}\mathcal{L}(x_*, \Lambda_*) \geq 0$. Les autres conditions proviennent sans changement du théorème précédent vu que $\partial_{x_k}\mathcal{M} = \partial_{x_k}\mathcal{L}$ si $k \geq q+1$. □

▷ EXEMPLE 2.16: Soit la fonction $U(x,y) = -xy$ à minimiser sur le domaine

$$x \geq 0, y \geq 0, x + y \leq 6.$$

Le lagrangien est $\mathcal{L}(x, y, \lambda) = -xy - \lambda(6 - x - y)$. On a les conditions suivantes :

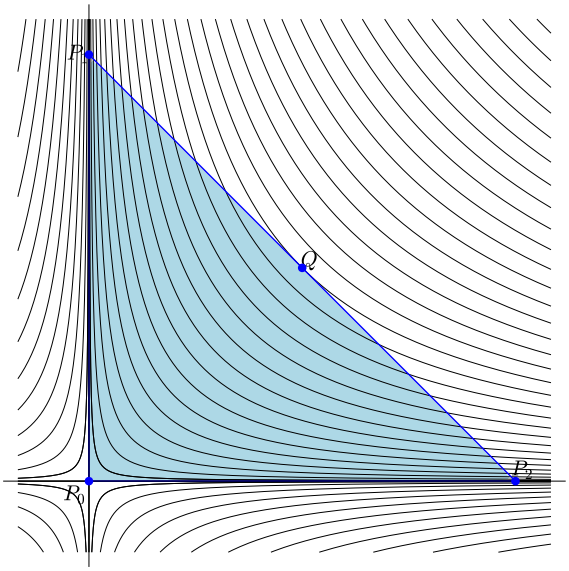


FIGURE II.8 . Les lignes de niveau de $U(x,y) = -xy$ sur le domaine $x \geq 0, y \geq 0, x + y \leq 6$.

— à l'intérieur du triangle, on a l'équation au point critique $\nabla U = 0$ qui revient au système $(-y, -x) = 0$ soit $x = y = 0$ qui correspond au sommet $P_0 = (0,0)$ non intérieur au triangle ;
— sur le côté $\{x + y = 6, x > 0, y > 0\}$, $\lambda \geq 0$, $0 = \partial_y\mathcal{L} = -x + \lambda$ et $0 = \partial_x\mathcal{L} = -y + \lambda$, soit $\lambda = 3$ et $x = y = 3$; le point $Q = (3,3)$ avec le multiplicateur $\lambda = 3 \geq 0$ vérifie les conditions KKT pour un minimum ;
— sur le côté $\{x = 0, y \in (0,6)\}$, $\lambda = 0$ (car la contrainte $6 - x - y \geq 0$ est inactive) et $0 \leq \partial_x\mathcal{L} = -y + \lambda$, soit $y = 0$ qui est hors de (l'intérieur de) ce côté ;
— sur $\{y = 0, x \in (0,6)\}$, $\lambda = 0$ (car la contrainte $6 - x - y \geq 0$ est inactive) et $0 \leq \partial_y\mathcal{L} = -x + \lambda$, soit $x = 0$ qui est hors de (l'intérieur de) ce côté ;

Hors les sommets $P_0 = (0,0), P_1 = (0,6), P_2 = (6,0)$ du triangle, seul le point Q est un point de minimum possible ; vu que $U(P_i) = 0, i = 0,1,2$ et $U(Q) = -9$, c'est Q qui est le minimum. Par ailleurs, on vérifie que les conditions KKT pour un minimum ne sont pas satisfaites aux sommets du triangle. ◁

7.3. Optimisation avec contraintes d'égalités et d'inégalités. Reprenant ce qui précède, on a le théorème général pour un problème d'optimisation avec contraintes en (in)égalités :

THÉORÈME 2.15: *Soient $U : V(\text{ouvert de } \mathbb{R}^n) \to \mathbb{R}$ et $h_j, g_i : V \to \mathbb{R}$ de classe $\mathcal{C}^1$. Soit x_* un point de minimum du programme*

$$\min_{\substack{x\in V\\ g_i(x)=0, i=1,\dots,m\\ h_j(x)\geq 0, j=1,\dots,p\\ x_k\geq 0, k=1,\dots,q}} U(x).$$

Désignons par J_{x_*} (resp. K_{x_*}) l'ensemble des indices de contraintes $h_j \geq 0$ (resp. $x_k \geq 0$) saturées en x_* et $\mathcal{L}_{x_*}$ le lagrangien défini par

$$\mathcal{L}_{x_*}(x, \Lambda, M) = U(x) - \sum_{i=1}^{m} \lambda_i g_i(x) - \sum_{j\in J_{x_*}} \mu_j h_j(x).$$

On suppose le minimum x_* régulier pour les contraintes actives, i. e. *la différentielle de $(x \mapsto ((g_i)_{i=1}^m, (h_j)_{j\in J_{x_*}}, (x_k)_{k\in K_{x_*}})$ surjective en x_* (ce qui est équivalent à l'indépendance linéaire des gradients $(\nabla g_i(x_*))_{i=1}^m, (\nabla h_j(x_*))_{j\in J_{x_*}}, (\nabla x_k(x_*))_{k\in K_{x_*}}$).*

Il existe alors des multiplicateurs $\Lambda_* = (\lambda_{*i})_{i=1,\dots,m}, M_* = (\mu_{*j})_{j\in J_{x_*}}$ tels que

$$\mu_{*j} \geq 0, j \in J_{x_*}, \quad \partial_{x_k}\mathcal{L}_{x_*}(x_*, \Lambda_*, M_*) \geq 0, k \in K_{x_*}.$$

et le point (x_*, Λ_*, M_*) soit critique pour le lagrangien $\mathcal{L}_{x_*}$ i. e.

$$\nabla U(x_*) = \langle \Lambda_*, \nabla g(x_*)\rangle + \langle M_*, \nabla h(x_*)\rangle.$$

De manière analogue à ce qui a été développé dans la section précédente, cet énoncé est équivalent à l'énoncé (apparemment moins général) sans aucune condition de positivité sur des coordonnées : c'est cet énoncé que nous allons démontré suivant une stratégie de pénalisation (qui est valable aussi pour des programmes avec contraintes d'égalité seulement).

Dans un voisinage $B_{x_*}(\subset \mathbb{R}^n)$ d'un point de minimum $x_* nC$ pour U, on remplace le potentiel U et son ensemble de points admissibles C, par un potentiel U_ε défini sur B_{x_*} et qui devient de plus en plus grand en dehors de l'ensemble des contraintes $C \cap B_{x_*}$: x_* est approché par une suite de minima x_ε pas nécessairement dans C (peu importe !) qui convergent (à sous-suite près) vers x_*, point qui vérifie une relation lagrangienne avec des multiplicateur de Lagrange adaptés.

DÉMONSTRATION. Soit $x_* \in V \cap C$ minimum de U et $r > 0$ suffisamment petit tel que $h_j(x) > 0$ sur $\overline{B(x_*, r)}$ pour tous les indices non actifs en x_*, *i. e.* $j \neq S_{x_*}$. On se ramène tout d'abord à un problème sans contrainte en introduisant la fonction pénalisée U_ε définie par

$$U_\varepsilon(x) = U(x) + \|x - x_*\|^2 + \varepsilon^{-1}P(x)$$

avec la pénalité

$$P(x) = \sum_{i=1}^{m} g_i(x)^2 + \sum_{j\in S_{x_*}} h_j^-(x)^2 \geq 0$$

où on a noté $h_j^- = \min(0, h_j)$. Si $x \in V \cap C$, les deux derniers termes sont nuls, sinon on a un surcoût par rapport à U de l'ordre de ε^{-1} ; en outre, le terme $\|x - x_*\|^2$ transforme x_* en un minimum strict sur la boule $\overline{B(x_*, r)}$. Enfin, la fonction U_ε est de classe $\mathcal{C}^1$, mais pas plus à cause des termes h_j^-.

Du fait de la compacité de la boule fermée, il existe un x_ε minimum de U_ε sur $\overline{B(x_*, r)}$. Ainsi, si $M_* = \sup_{x\in\overline{B(x_*,r)}} |U(x)|$, on a

$$P(x_\varepsilon) = \varepsilon(U_\varepsilon(x_\varepsilon) - U(x_\varepsilon) - \|x_\varepsilon - x_*\|^2) \leq \varepsilon(U_\varepsilon(x_*) - U(x_\varepsilon)) \leq 2M_*\varepsilon$$

et

$$U(x_\varepsilon) + \|x_\varepsilon - x_*\|^2 \leq U_\varepsilon(x_\varepsilon) \leq U_\varepsilon(x_*) = U(x_*) = \inf_{x\in V\cap C} U(x).$$

Par compacité de la boule $\overline{B}(x_*, r)$, il existe une suite $\varepsilon_k \to 0^+$ telle que x_{ε_k} converge[16] vers $\underline{x}$ avec

$$\sum_{i=1}^m g_i(\underline{x})^2 + \sum_{j\in S_{x_*}} h_j^-(\underline{x})^2 \leq 0, \quad U(\underline{x}) + \|\underline{x} - x_*\|^2 \leq U(x_*)$$

La première inégalité implique $\underline{x}$ dans le lieu des contraintes C. De la deuxième inégalité résulte $\underline{x} = x_*$, x_ε dans la boule ouverte $B(x_*, r)$ pour ε assez petit, alors que l'annulation du gradient[17] de la fonction U_ε en x_ε donne

$$\nabla U(x_\varepsilon) + 2(x_\varepsilon - x_*) + \varepsilon^{-1}\left[\sum_{i=1}^m 2g_i(x_\varepsilon)\nabla g_i(x_\varepsilon) + \sum_{j\in S_{x_*}} 2h_j^-(x_\varepsilon)\nabla h_j(x_\varepsilon)\right] = 0 \tag{47}$$

Notons $\lambda_i^\varepsilon = -2\varepsilon^{-1}g_i(x_\varepsilon), \mu_i^\varepsilon = -2\varepsilon^{-1}h_j^-(x_\varepsilon)$ (qui est ≥ 0). Si ces suites ne sont pas bornées, supposons par exemple que $|\lambda_1^\varepsilon| = \max_{i,j}(|\lambda_i^\varepsilon|, |\mu_i^\varepsilon|)$ (dans une sous-suite (ε_k) appropriée) et divisons (47) par λ_1^ε : on obtient à la limite, en passant à des sous-suites convergentes au besoin et en notant $\underline{\lambda}_i = \lim_{\varepsilon\to 0}(\lambda_i^\varepsilon/\lambda_1^\varepsilon)$ et $\underline{\mu}_j$ pareillement,

$$0 = \nabla g_1(x_*) + \sum_{i=2}^m \underline{\lambda}_i \nabla g_i(x_*) + \sum_{j\in S_{x_*}} \underline{\mu}_j \nabla h_j(x_*),$$

ce qui contredit l'hypothèse d'indépendance linéaire des $(\nabla g_i(x_*))_{i=1}^m, (\nabla h_j(x_*))_{j\in S_{x_*}}$. On en déduit la convergence de $(\lambda_i^\varepsilon)_{i=1}^m, (\mu_j^\varepsilon)_{j\in S_{x_*}}$ à sous-suite près vers $(\lambda_{*i})_{i=1}^m, (\mu_{*j})_{j\in S_{x_*}}$ avec les $(\mu_{*j})_{j\in S_{x_*}} \geq 0$ et la relation

$$\nabla U(x_*) = \sum_{i=1}^m \lambda_{*i}\nabla g_i(x_*) + \sum_{j\in S_{x_*}} \mu_{*j}\nabla h_j(x_*)$$

attendue et qui conclut la preuve. □

Comme précédemment et en général, les conditions KKT identifient des points candidats à être un minimum ou un maximum : l'examen des valeurs prises par la fonction d'objectif U permet d'établir les maxima et minima globaux.

▷ Exemple 2.17: Soit le programme

$$\text{Optimiser } U(x,y) = x^3 + y \text{ sous les contraintes } x^2 + 2y^2 - 1 = 0 \text{ et } x \geq 0. \tag{48}$$

On considère le lagrangien

$$\mathcal{L}((x,y),\lambda) = x^3 + y - \lambda(1 - x^2 - 2y^2)$$

avec dérivées

$$\partial_x\mathcal{L} = 3x^2 + 2\lambda x, \quad \partial_y\mathcal{L} = 1 + 4\lambda y.$$

16. De manière un peu abusive, ce pour alléger le cours de la preuve, on sera amené à écrire $\varepsilon \to 0$ pour signifier la convergence suivant une suite $\varepsilon_n \to 0^+$ (ou une suite extraite), le contexte permettant aisément de surseoir à cet abus d'expression.

17. La fonction $\varphi : u \in \mathbb{R} \mapsto (u^-)^2 \in \mathbb{R}$ est dérivable sur $\mathbb{R}$, de dérivée $\varphi'(u) = 2u^-$: c'est vrai sur $\mathbb{R}^*$ (où la fonction est de classe $\mathcal{C}^\infty$) et aussi à l'origine puisque $\varphi(u) = \mathcal{O}(u^2)$ au voisinage de $u = 0$. La dérivabilité de $(h^-)^2 = \varphi \circ h$, avec gradient $2h^-\nabla h$ en résulte.

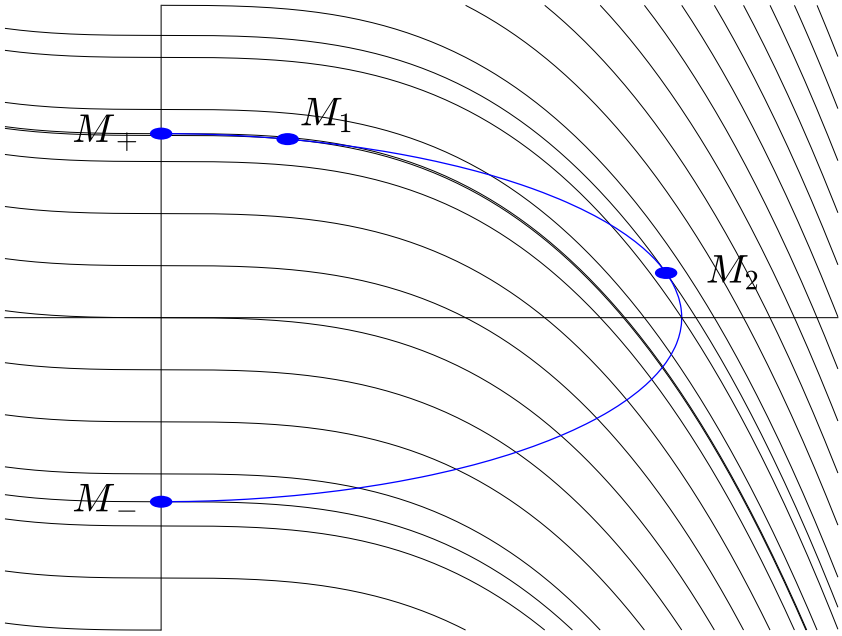


FIGURE II.9 . Les courbes de niveau de la fonction d'objectifs $U(x,y) = x^3 + y$ et le domaine des réalisables $\{x^2 + 2y^2 = 1, x \geq 0\}$.

À l'intérieur $(x > 0)$ de la demi-ellipse $x^2 + 2y^2 = 1$, on a le système

$$3x^2 + 2\lambda x = 0, \quad 1 + 4\lambda y = 0, \quad x^2 + 2y^2 = 1$$

et donc, vu que $x = 0$ est exclus, $x = -2\lambda/3, y = -1/(4\lambda), (2\lambda/3)^2 + 2(4\lambda)^{-2} = 1$, soit deux points $M_1 \simeq (0.243, 0.6859)$ avec $\lambda_1 = -0.3645$ et $M_2 \simeq (0.97, 0.1718)$ avec $\lambda_2 = -1.455$ dans le demi-plan $x > 0$ avec un déterminant de hessienne bordée

$$\begin{vmatrix} 6x + 2\lambda & 0 & 2x \\ 0 & 4\lambda & 4y \\ 2x & 4y & 0 \end{vmatrix} = -16[6xy^2 + \lambda(x^2 + 2y^2)]$$

valant -5.143058 en M_1 (indiquant un minimum local) et 20.52992 en M_2 (indiquant un maximum local). Aux extrémités $M_\pm = (0, \pm 1/\sqrt{2})$ de la demi-ellipse, on a $\partial_x \mathcal{L} = 3x^2 = 0$, qui est compatible avec un minimum ou maximum.

Vu les valeurs $U(M_1) \simeq 0.7000$, $U(M_2) \simeq 1.084$ et $U(M_\pm) \simeq \pm 0.707$, on en déduit le maximum (global) en M_2 et le minimum en M_- pour le programme (48). ◁

Chapitre 3

Programmation convexe

À l'instar du caractère linéaire pour la résolution des équations, c'est la convexité qui est le caractère distinguant un problème d'optimisation « résoluble » (minima locaux et globaux, temps polynomial de calcul, critères d'arrêt, sélection de point de départ, enjeux numériques) et d'un problème « difficile ». Par ailleurs, beaucoup de problèmes d'optimisation après changement de variable ou changement d'échelle de la fonction à optimiser se ramènent à des problèmes d'optimisation convexe : c'est dire la prééminence de la classe des programmes convexes, classe contenant en particulier celle des programmes linéaires et celle des programmes quadratiques. Enfin certaines opérations (telle le max de fonctions) font disparaître la différentiabilité tout en maintenant la convexité : c'est dire l'importance des programmes convexes non différentiables.

Commençons par la définition d'une fonction convexe.

DÉFINITION 3.1: *Soit E un espace vectoriel réel. La fonction $U : E \to \mathbb{R} \cup \{+\infty\}$ est dite* convexe *si*

$$U(\lambda x + (1-\lambda)y) \leq \lambda U(x) + (1-\lambda)U(y), \quad x, y \in C, \lambda \in (0,1). \tag{49}$$

La fonction U est strictement convexe *si l'inégalité (3.1) est stricte pour $x \neq y$ et $\lambda \in (0,1)$.*

Une fonction U dont l'opposée $-U$ est (strictement) convexe est dite (strictement resp.) concave *si l'inégalité (3.1) est stricte pour $x \neq y$ et $\lambda \in (0,1)$.*

Soit $m > 0$. La fonction U est dite m-fortement *convexe si*

$$U(\lambda x + (1-\lambda)y) \leq \lambda U(x) + (1-\lambda)U(y) - \frac{m}{2}\lambda(1-\lambda)\|x-y\|^2, \quad x, y \in C, \lambda \in (0,1).$$

ce qui est équivalent à la convexité de la fonction $x \mapsto U(x) - \frac{m}{2}\|x\|_2^2$.

L'arithmétique mise à l'œuvre dans cette définition pour $+\infty$ est simple :

$$t + (+\infty) = +\infty \text{ si } t \in \mathbb{R} \cup \{0\}, \quad a(+\infty) = +\infty \text{ si } a > 0.$$

Une fonction m-fortement convexe assure une minoration quadratique de la fonction, alors que la simple convexité n'assure qu'une minoration linéaire (par le plan tangent)

△ REMARQUE 3.1: La définition est adaptable à toute fonction d'un espace affine A de direction E et de point-origine O, où tout point M s'écrit comme $M = O + \overrightarrow{OM}$ pour un vecteur unique $\overrightarrow{OM} \in E$ et le point $P_{\alpha,\beta} = \alpha M + \beta N$, avec $\alpha + \beta = 1, \alpha, \beta \geq 0$, est caractérisé par

$$\overrightarrow{OP_{\alpha,\beta}} = \alpha\overrightarrow{OM} + \beta\overrightarrow{ON}, \quad O \in A.$$

En effet, la donnée d'une fonction $U : x \in E \mapsto U(x)$ est équivalent à la donnée de la fonction $V : M \in A \to U(\overrightarrow{OM})$ et réciproquement U est obtenue de V en posant $U(x) = V(O+x), x \in E$. La fonction $V : A \to \mathbb{R}$ est dite convexe si elle vérifie les inégalités de convexité

$$V(\alpha M + \beta N) \leq \alpha V(M) + \beta V(N), \quad \alpha + \beta = 1, \alpha, \beta \geq 0$$

qui sont équivalents à celles pour U vu que $V(M) = V(O + \overrightarrow{OM}) = U(\overrightarrow{OM})$, $V(N) = V(O + \overrightarrow{ON})$ et que le point $\alpha M + \beta N$ est obtenu en rajoutant à l'origine O le vecteur $\alpha\overrightarrow{OM} + \beta\overrightarrow{ON}$, ce qui donne, pour des α, β comme *supra*,

$$U(\alpha\overrightarrow{OM} + \beta\overrightarrow{ON}) = V(\alpha M + \beta N) \leq \alpha V(M) + \beta V(N) = \alpha U(\overrightarrow{OM}) + \beta U(\overrightarrow{ON}),$$

soit donc la convexité de U, qui entraîne celle de V inversement. Cette définition de la convexité de V est indépendante du choix de l'origine vu que les fonctions $U, \widetilde{U}$ associées à V avec le choix d'origines $O, \widetilde{O}$ respectifs vérifient $U(x) = \widetilde{U}(\overrightarrow{O\widetilde{O}} + x), x \in E$.

Comme exemple de ces équivalences, la fonction distance au point origine O dans un espace affine euclidien sera notée $M \in A \mapsto OM = d(O, M) = \|\overrightarrow{OM}\|$, fonction convexe comme son application associée la norme $x \in E \mapsto \|x\|$: la convexité de la norme dans E induit celle la distance dans A (cf. 3.3.5). D'ailleurs plus généralement, on a $PM = d(P, M) = \overrightarrow{OM} - \overrightarrow{OP}$, distance associée à la fonction $x \to \|x - \overrightarrow{OP}\|$. ▽

La définition de partie convexe $C(\subset E)$ est en fait un cas particulier de fonction convexe.

Définition 3.2: *La partie $C \subset E$ est dite convexe si sa fonction indicatrice I_C définie par $I_C(x) = 0$ si $x \in C$, $+\infty$ sinon, est convexe.*

L'ensemble vide est *stricto sensu* convexe. On essaiera au possible de ne pas trop raisonner sur la convexité de l'ensemble vide. Cette définition de partie convexe a une version plus géométrique :

Lemme 3.1: *La partie $C(\subset E)$ est* convexe *si et seulement si $x + \lambda(y - x)$ appartient à C pour tout x, y dans C et $\lambda \in (0, 1)$ ou, de manière équivalente, si tout segment*

$$[x, y] = \{\lambda x + (1 - \lambda)y, \lambda \in [0, 1]\} = \{x + \theta(y - x), \theta \in [0, 1]\}$$

d'extrémités x, y dans C est inclus dans la partie C.

Démonstration. Si I_C est convexe, pour $x, y \in C$, la convexité de I_C s'exprime suivant $I_C(\lambda x + (1-\lambda)y) \leq \lambda \cdot 0 + (1-\lambda) \cdot 0$, ce qui exprime l'appartenance de $\lambda x + (1-\lambda)y$ à C et donc la convexité de C. Réciproquement, si C est convexe, les inégalités de convexité de I_C à vérifier sont seulement celles où leur membre de droite est fini : de la forme $\lambda I_C(x) + (1 - \lambda)I_C(y)$, celui-ci est nul, majorant les termes du type $I_C(\lambda x + (1 - \lambda)y)$, qui est aussi nul vu la convexité de C et l'appartenance de $I_C(\lambda x + (1 - \lambda)y$ à C. □

Deux types de parties convexes sont déterminées par une fonction convexe : tout d'abord ses domaines de sous-niveau :

Définition 3.3: *Soit U convexe sur E. Son domaine* $\mathbf{dom}\, U = \{x \in E; U(x) < +\infty\}$ *est la partie de E des points x d'image un nombre réel.*

Plus généralement les domaines de sous-niveau $S_{U,\alpha} = \{x \in E; U(x) \leq \alpha\}$ et $\widetilde{S}_{U,\alpha} = \{x \in E; U(x) < \alpha\}$ sont convexes.

Démonstration. L'inégalité de convexité pour la fonction U

$$U(\lambda x + (1 - \lambda)y) \leq \lambda U(x) + (1 - \lambda)U(y), \quad x, y \in C, \lambda \in (0, 1)$$

entraîne les convexités des différents domaines de sous-niveau. □

Sauf menton du contraire, on considérera dans la suite des fonctions convexes U avec un domaine $\mathbf{dom}\, U$ non vide. Il y a correspondance entre fonctions convexes sur E à valeurs finies ou $+\infty$ et fonctions convexes définies sur une partie convexe à valeurs réelles : à $U : E \to \mathbb{R} \cup \{+\infty\}$ on lui associe sa restriction à son domaine $\mathbf{dom}\, U$ et

inversement à $U : C \to \mathbb{R}$ on lui associe son prolongement à E avec valeur $+\infty$ dans le complémentaire de C.

D'autre part l'épigraphe

$$\mathbf{epi}\, U = \{(x,t) \in E \times \mathbb{R}, U(x) \leq t\}$$

d'une fonction $U : E \to \mathbb{R}\{+\infty\}$ détecte la convexité de U d'après le lemme suivant

LEMME 3.2: *L'application $U : E \to \mathbb{R} \cup \{+\infty\}$ est convexe si et seulement si son épigraphe* $\mathbf{epi}\, U$ *est convexe.*

DÉMONSTRATION. Si $\mathbf{epi}\, U$ est convexe, alors vu que $(x, U(x)), (x', U(x')) \in \mathbf{epi}\, U$, $\lambda(x,t) + (1-\lambda)(x',t') = (\lambda x + (1-\lambda)x', \lambda U(x) + (1-\lambda)U(x'))$ est dans $\mathbf{epi}\, U$, soit l'inégalité de convexité assurée par la définition de $\mathbf{epi}\, U$

$$U(\lambda x + (1-\lambda)x') \leq \lambda U(x) + (1-\lambda)U(x')$$

qui assure la convexité de la fonction U. Réciproquement, si U est convexe, son inégalité de convexité

$$U(\lambda x + (1-\lambda)x') \leq \lambda U(x) + (1-\lambda)U(x')$$

ce qui induit pour les éléments $(x,t), (x',t') \in \mathbf{epi}\, U$ qui vérifient $U(x) \leq t$ et $U(x') \leq t'$, la majoration

$$U(\lambda x + (1-\lambda)x') \leq \lambda t + (1-\lambda)t'$$

ce qui assure l'appartenance de $(\lambda x + (1-\lambda)x', \lambda t + (1-\lambda)t' = \lambda(x,t) + (1-\lambda)(x',t')$ dans $\mathbf{epi}\, U$, qui est donc convexe. □

DÉFINITION 3.4: *Une fonction convexe est dite* fermée *si son épigraphe est fermé.*

Une fonction convexe et continue sur un domaine fermé est une fonction fermée. Ainsi, toute norme est fermée sur $\mathbb{R}^n$, ainsi que toute fonction différentiable. La fonction $x \in (0, +\infty) \mapsto 1/x$ est convexe fermée, bien que son domaine $(0, +\infty)$ soit ouvert.

▷ EXEMPLES 3.1:

??.2 $U : x \mapsto -\log(1-x^2)$ avec domaine $\{|x| < 1\}$,

??.3 $U : x \mapsto x \log x$ avec domaine $\mathbb{R}^+$ avec $U(0) = 0$.

??.4 la fonction indicatrice d'un convexe (si la partie est non convexe, I_C n'est pas fermée),

??.5 $U : x \mapsto x \log x$ avec domaine $\mathbb{R}^{+*}$ ou avec domaine $\mathbb{R}^+$ et $U(0) = 1$. avec $U(0) = 0$.

◁

1. Parties convexes

La liste suivante présente quelques convexes usuels, ainsi que des constructions préservant ou engendrant des parties convexes. Les vérifications simples sont laissées aux soins du lecteur.

▷ EXEMPLES 3.2:

3.2.1 Soit H espace de Hilbert, $h \in H$ et $c \in \mathbb{R}$. Les demi-espaces ouvert $E_{h,c} = \{v \in H, \langle v, h\rangle > c\}$ et fermé $\overline{E}_{h,c} = \{v \in H, \langle v, h\rangle \geq c\}$ sont convexes.

3.2.2 Soit E avec une norme $\|\ \|$: les boules ouverte $B_{x_0,r} = \{x, \|x - x_0\| < r\}$ et fermée $\overline{B}_{x_0,r} = \{x, \|x - x_0\| \leq r\}$ sont convexes.

3.2.3 L'intersection d'une famille (finie ou infinie) de convexes est, si elle n'est pas vide, une partie convexe. L'union de deux parties convexes n'est pas en général convexe.

3.2.4 L'intersection finie de demi-espaces est convexe : ces parties sont des polyèdres. Un polyèdre compact convexe est appelé *polytope*. Un *polygone* (convexe) est un polyèdre du plan.

3.2.5 L'image directe ou inverse d'une partie convexe par une application affine $x \in \mathbb{R}^n \mapsto Ax+b$ $(A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m)$ est convexe.

3.2.6 Les parties convexes de $\mathbb{R}$ sont les intervalles (ouverts, fermés, bornés, semi-ouverts,...) : tout convexe C est égal à $]\inf(x \in C), \sup(x \in C)[$ aux extrémités (incluses dans C ou pas) près.

3.2.7 Soit, pour p réel au moins égal à 1, la partie $P_p = \{(x,y) \in \mathbb{R}^2, y \geq |x|^p\}$. Cette partie est convexe : cela résulte pour $p = 2$ de l'identité

$$\lambda y + (1-\lambda)y' - (\lambda x + (1-\lambda)x')^2 = \lambda(y-x^2) + (1-\lambda)(y'-x'^2) + \lambda(1-\lambda)(x-x')^2.$$

Pour les autres valeurs de p, cela sera établi ultérieurement.

3.2.8 La partie $Q_3 = \{(x,y) \in \mathbb{R}^2, y \geq x^3\}$ n'est pas convexe, puisque le segment $[(0,0),(-x,-x^3)]$ pour $x \geq 0$ a son intérieur en dehors de Q_3 : pour $\lambda \in (0,1)$, il n'est pas vrai que

$$-(1-\lambda)^3x^3 \leq -(1-\lambda)x^3, \quad \lambda \in (0,1), x > 0.$$

3.2.9 Pour $\alpha, \beta, A > 0$, on verra ci-dessous que la partie $H_{\alpha,\beta,A} = \{(x,y) \in \mathbb{R}^2_+, x^\alpha y^\beta \geq A\}$ est convexe.

3.2.10 Une partie d'un espace vectoriel est dite *conique* si elle est stable par l'action du groupe $\mathbb{R}^*_+$ *i. e.* telle que $tx \in C$ si $x \in C, t \in \mathbb{R}_+$. Si C est convexe, l'union $\cup_{t>0} tC$ est conique et convexe. Le cône C est convexe si et seulement si il est stable par addition (si u, v sont dans C, il en est de même pour $u+v$).

3.2.11 La partie conique $C = \{(u,x) \in \mathbb{R}\times\mathbb{R}^n, \|x\| \leq u\}$ est convexe : si $(u,x),(v,y) \in C$, alors $\|\lambda x + (1-\lambda)y\| \leq \lambda u + (1-\lambda)v\|$, *i. e.*

$$(\lambda u + (1-\lambda)v, \lambda x + (1-\lambda)y) = \lambda(u,x) + (1-\lambda)(v,y) \in C.$$

3.2.12 Le cône $\mathcal{S}_n^+$ des matrices S symétriques positives d'ordre n (*i. e.* S telle que ${}^{\mathsf{T}}S = S$ et $\langle Sx, x\rangle \geq 0$ pour tout $x \in \mathbb{R}^n$) est convexe, comme intersection d'hyperplans

$$\mathcal{S}_n^{++} = \bigcap_{x\in E\setminus\{0\}} \{\langle Sx, x\rangle \geq 0\}.$$

En particulier, la partie $\{x \geq 0, y \geq 0, xy - z^2 \geq 0\}$ ($\{x > 0, y > 0, xy - z^2 > 0\}$ resp.) est convexe : elle s'identifie à la partie des matrices symétriques positives (définies positives resp.) dans l'espace des matrices symétriques $\begin{pmatrix} x & z \\ z & y \end{pmatrix}$. On peut le montrer directement en vérifiant la stabilité par addition (cf. Ex. 10)

$$(x+x')(y+y') = xy + x'y' + xy' + x'y \geq xy + x'y' + 2\sqrt{xy'x'y} \geq (z+z')^2$$

De manière analogue, la partie constituée des matrices symétriques définies positives (négatives resp.) est un cône convexe ouvert. ◁

THÉORÈME 3.1: *Soit C partie convexe fermée de l'espace de Hilbert E. Alors C est l'intersection des demi-espaces qui le contiennent. Explicitement*

$$C = \cap_{v\in E, \|v\|=1} \{x | \langle x, v\rangle \leq I_C^*(v)\}. \tag{50}$$

où on a introduit la fonction[1] *I_C^* définie par $I_C^*(v) = \sup_{x\in C}\langle v, x\rangle$ pour $v \in E$.*

1. I_C^* est la transformée de Fenchel-Nielsen de l'indicatrice I_C, cf. 3.5.

DÉMONSTRATION. Soit $\widehat{C} = \cap_{v\in E,\|v\|=1}\{\langle x, v\rangle \leq I_C^*(v)\}$. D'après la définition de I_C^*, le convexe C est inclus dans tout demi-espace $\{\langle x, v\rangle \leq I_C^*(v)\}$; ainsi $C \subset \widehat{C}$.

Soit $y \notin C$. Considérons la projection y_C de y sur C (cf. théorème 3.7 ci-dessous). Si $v_y = (y - y_C)/\|y - y_C\|$, alors $\langle x - y_C, v_y\rangle \leq 0$ pour tout $x \in C$. Ainsi $I_C^*(v_y) = \sup_{x\in C}\langle x, v_y\rangle \leq \langle y_C, v_y\rangle$ et par suite

$$I_C^*(v_y) \leq \langle y_C, v_y\rangle < \langle y_C, v_y\rangle + \|y - y_C\| = \langle y_C, v_y\rangle + \langle y - y_C, \frac{y - y_C}{\|y - y_C\|}\rangle = \langle y, v_y\rangle,$$

soit $I_C^*(v_y) < \langle y, v_y\rangle$ et donc $y \notin \widehat{C}$, ce qui conclut la preuve. □

2. Fonctions convexes

Commençons par le constat que la convexité est une propriété en dernier ressort de dimension 1. En effet, la fonction U est (strictement) convexe sur C si et seulement si la restriction de U à tout segment inclus dans C est (strictement) convexe. Pour une fonction U d'une variable, la convexité se lit sur le graphe (plan) de la fonction : pour tout couple de valeurs x, y, la courbe $t \in [x, y] \mapsto U(t)$ est en dessous de la corde $[(x, U(x)), (y, U(y))]$, cf. Fig. III.1.

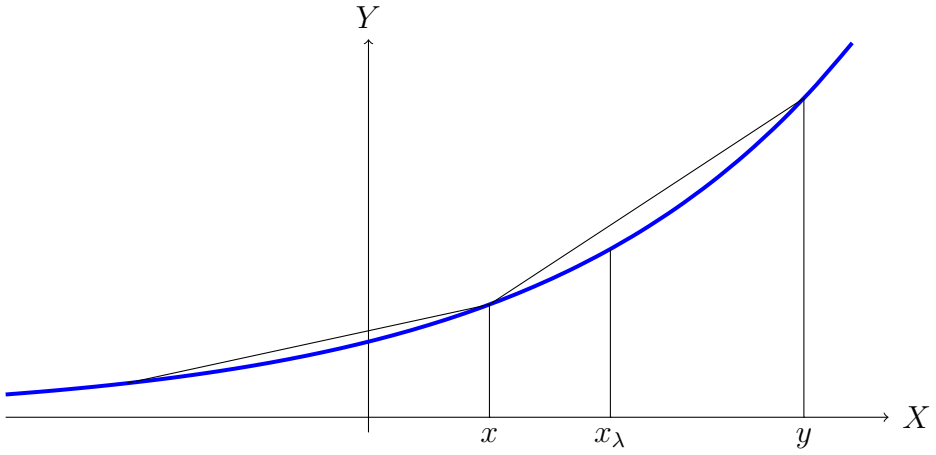


FIGURE III.1 . Le graphe d'une fonction convexe $U : \mathbb{R} \to \mathbb{R}$: la fonction est au-dessous de la corde $[(x, U(x), (y, U(y)]$, *i. e.* pour $x_\lambda = \lambda x + (1-\lambda)y$, $U(x_\lambda) \leq \lambda U(x) + (1 - \lambda)U(y)$.

▷ EXEMPLES 3.3:

3.3.1 L'inégalité (3.1) est une égalité pour toute fonction affine, qui est donc convexe (non strictement convexe) et concave.

3.3.2 Toute norme $\|\ \|$ est convexe. En particulier, la fonction $t \in \mathbb{R} \mapsto |t|$ est convexe.

3.3.3 Si $\Phi \in \mathcal{L}(F, E)$ est linéaire, $e \in E$, $a, b \in \mathbb{R}$ avec $a \geq 0$, alors pour toute fonction convexe $x \in E \mapsto U(x)$, la composée $y \in F \to aU(\Phi(y) + e) + b$ est aussi convexe.

3.3.4 La somme, la multiplication par un nombre positif et le sup de fonctions convexes est convexe. Ainsi la fonction de Leontiev $U_L(x_1, \ldots, x_n) = \max_{i=1}^n |\alpha_i x_i|$ est convexe.

3.3.5 Pour $m > 0$, la fonction $q_m : x \in \mathbb{R} \mapsto \frac{m}{2}x^2$ est strictement convexe, puisque

$$(\lambda x + (1 - \lambda)y)^2 - \left[\lambda x^2 + (1 - \lambda)y^2\right] = -\lambda(1 - \lambda)(x - y)^2.$$

Elle est m-convexe (suivant la définition **??** à la fin de cette section). Plus généralement, la fonction $x \in E \mapsto \|x\|^2$ (où la norme est induite par le produit scalaire) est strictement convexe : en effet sa restriction à toute droite $x + \mathbb{R}d$, avec $\|d\| = 1$, est de la forme $t \mapsto \|x + td\|^2 = \|x\|^2 + 2\langle x, d\rangle t + t^2$; elle est par définition (heureusement compatible) fortement convexe et 1-convexe.

Plus généralement, si A est un opérateur symétrique positif sur l'espace E, alors la fonction $U_A : x \in E \mapsto \langle Ax, x\rangle$ est convexe, strictement convexe si A est défini positif. En effet, la restriction de U_A à la droite $x + \mathbb{R}d$ est de la forme $t \mapsto \langle Ax, x\rangle + 2\langle Ad, x\rangle t + \langle Ad, d\rangle t^2$, convexe (vu que A est positif) et strictement convexe seulement si $\langle Ad, d\rangle > 0$

3.3.6 La somme des r-plus grandes composantes de $x \in \mathbb{R}^n$, soit $f(x) = x_{[1]} + x_{[2]} + \cdots + x_{[r]}$ est convexe, vu que

$$f(x) = \max(x_{i_1} + x_{i_2} + \ldots + x_{i_r} | 1 \leq i_1 < i_2 < \ldots < i_r \leq n)$$

3.3.7 Si g est croissante convexe et U convexe, alors $g \circ U$ est convexe. Ainsi les normes au carré $x \in E \mapsto \|x\|^2$ sont convexes : elle s'obtiennent en composant la norme $\|\ \|$ convexe et l'application croissante convexe $u \in \mathbb{R}_+ \mapsto u^2 \in \mathbb{R}$.

3.3.8 Si C_1, C_2 sont deux convexes et $U : (x, y) \in C_1 \times C_2 \mapsto U(x, y)$ convexe sur $C_1 \times C_2$, alors $V(x) = \inf_{y \in C_2} U(x, y)$ est convexe. En effet

$$\begin{aligned} V(\theta x + (1-\theta)x') &= \inf_{z \in C_2} U(\theta x + (1-\theta)x', z) \\ &\leq U(\theta x + (1-\theta)x', \theta y + (1-\theta)y') \\ &\leq \theta U(x, y) + (1-\theta)U(x', y') \end{aligned}$$

soit en prenant l'inf sur y, puis sur y'

$$V(\theta x + (1-\theta)x') \leq \theta V(x) + (1-\theta)V(x').$$

Ainsi, si C est un convexe de l'espace normé E, $d_C(x) = \inf_{y \in C} \|x - y\|$ est convexe sur E.

3.3.9 Soit $(\Omega, \mathcal{P}, P)$ un espace de probabilités, X une variable aléatoire et U une fonction convexe. Alors

$$U(\mathbb{E}[X]) \leq \mathbb{E}[U(X)],$$

inégalité attribuée à Jensen et dont les éléments de preuve ci-dessous reprennent des propriétés des fonctions convexes exposées ci-dessous.

Supposons tout d'abord U différentiable. Pour $\omega \in \Omega$, la fonction $U_\omega : t \in [0, 1] \mapsto U(tX(\omega) + (1-t)\mathbb{E}[X])$ est convexe dérivable à droite en $t = 0$: la caractérisation des fonctions convexes de la proposition 3.1 fournit l'inégalité $U_\omega(1) - U_\omega(0) \geq U'_\omega(0)$ et par suite

$$U(X(\omega)) - U(\mathbb{E}[X]) \geq \langle \nabla U(\mathbb{E}[X]), X(\omega) - \mathbb{E}[X]\rangle \tag{51}$$

qu'on s'empresse d'intégrer sur Ω pour obtenir $\mathbb{E}(U(X)) - U(\mathbb{E}[X]) \geq 0$.

Si U n'est pas supposée dérivable, on admet l'existence d'un hyperplan d'appui H_X dans $\mathbb{R}^n_x \times \mathbb{R}_y$ passant par le point $(\mathbb{E}[X], U(\mathbb{E}[X]))$ et déterminant un demi-espace H_X^+ contenant l'épigraphe $\mathbf{epi}\, U$: ce demi-espace H_X^+ est, pour a, ν convenables avec $a^2 + \|\nu\|^2 > 0$, d'équation $ay + \langle \nu, x\rangle \geq C$. Vu que $(x, y) \in \mathbf{epi}\, U \subset H_X^+$ pour tout $y \geq U(x)$ le réel a est positif non nul. On peut donc prendre comme équation du demi-plan $h_X(x, y) \geq 0$ où on a posé $h_X(x, y) = y - U(\mathbb{E}(X)) - \langle n, x - \mathbb{E}[X]\rangle$ (on a pris $n = \nu/a$; si U est fonction d'une variable,

n est dans l'intervalle $[U'_g(\mathbb{E}[X]), U'_d(\mathbb{E}[X])]$). On a donc l'analogue de (51) pour $(X(\omega), U(X(\omega))) \in \mathbf{epi}\, U$

$$U(X(\omega)) - U(\mathbb{E}[X]) \geq \langle n, X(\omega) - \mathbb{E}[X]\rangle,$$

ce qui permet de conclure comme précédemment. Si Ω est fini avec $P = \sum_i p_i \delta_{x_i}$, l'inégalité de Jensen prend la forme classique des inégalités de convexité

$$U\left(\sum_i p_i x_i\right) \leq \sum_i p_i U(x_i).$$

Si $\Omega = \mathbb{R}^n$ et P est absolument continue par rapport à la mesure de Lebesgue dx de densité $f(x)$ (avec $\int f(x)dx = 1$), l'inégalité de Jensen prend la forme

$$U\left(\int X(x)f(x)dx\right) \leq \int U(X(x))f(x)dx$$

avec le cas particulier pour X l'identité

$$U\left(\int xf(x)dx\right) \leq \int U(x)f(x)dx.$$

On peut aussi faire appel à l'idempotence de la conjuguée de Legendre (dont une preuve repose sur l'existence d'hyperplans d'appui)

$$U(\mathbb{E}[X]) \leq \sup_\lambda [\langle \lambda, \mathbb{E}[X]\rangle - U^*(\lambda) \leq \sup_\lambda \mathbb{E}[\langle \lambda, X\rangle - U^*(\lambda)] \leq \mathbb{E}[\sup_\lambda [\langle \lambda, X\rangle - U^*(\lambda)] = \mathbb{E}[U(X)]$$

◁

La notion de *fonction conjuguée* d'une fonction quelconque fournit une foultitude de fonctions convexes :

Définition 3.5 (Conjuguée de Fenchel-Legendre): *Soit E un espace de Hilbert et U fonction définie sur $\mathbf{dom}\, U \subset E$. Sa* fonction conjuguée[2] *est la fonction U^* définie sur E suivant*

$$U^*(\lambda) = \sup_{x \in \mathbf{dom}\, U} [\langle \lambda, x\rangle - U(x)], \quad \lambda \in E.$$

Pour une fonction U d'une variable, l'inégalité $U(x) \geq \lambda x - U^*(\lambda)$ indique que la droite d'équation $y = \lambda x - U^*(\lambda)$ est celle de pente λ de hauteur maximale qui soit en dessous du graphe de U. Par ailleurs, $U^*(0) = -\inf_{x \in \mathbf{dom}\, U}[-U(x)]$.

La définition de la transposée U^* de U implique l'inégalité de Fenchel

$$U(x) + U_*(y) \geq \langle x, y\rangle,$$

généralisation aux U non quadratique de l'inégalité

$$\frac{\langle x, x\rangle}{2} + + \frac{\langle y, y\rangle}{2} \geq \langle x, y\rangle.$$

△ Remarque 3.2: Si J est une fonction mesurant le coût de production du *panier de biens* x et $\mathbf{p}$ un vecteur de prix, alors $J^*(\mathbf{p}) = \sup_x(\langle x, \mathbf{p}\rangle - J(x))$ est le profit maximum atteignable associé aux prix $\mathbf{p}$. ▽

▷ Exemples 3.4:

3.4.1 La conjuguée de la forme linéaire $\Phi_\ell : x \mapsto \langle \ell, x\rangle$ est la fonction valant $+\infty$ partout sauf en $\lambda = \ell$ où elle est nulle : en termes de fonction indicatrice, $\Phi^*_\ell = I_\ell$.

3.4.2 Pour $a > 0$, la conjuguée de $x \mapsto ax^2/2$ est $\lambda \mapsto \lambda^2/(2a)$.

2. Cette conjuguée de Fenchel-Legendre est aussi appelée *transformée de Young* ou de *Nielsen*.

3.4.3 Plus généralement, si $U_{A,b,c} : x \mapsto \langle Ax, x\rangle/2 + \langle b, x\rangle + c$, alors $U^*_{A,b,c}(y) = \langle A^{-1}(y-b), y-b\rangle/2 - c$ si A est définie positive. Si A est positive, en écrivant $U_{A,b,c}(x) = \langle A_\pm x_\pm, x_\pm\rangle/2 + \langle b_\pm, x_\pm\rangle + \langle b_0, x_0\rangle + c$ où on a utilisé les décompositions $b = b_0 + b_\pm$ et $x = x_0 + x_\pm$ introduites dans la remarque 0.4, on montre que U^* a pour domaine $\Im m\, A + b$ avec $U^*(y) = \langle A^\dagger(y-b), (y-b)\rangle - c$ où $A^\dagger$ est le pseudo-inverse de A (*i. e.* l'inverse de la restriction de A opérant dans $\Im m\, A = (\ker A)^\perp$).

3.4.4 La conjuguée de $x \mapsto \mathrm{e}^x$ est de domaine $\mathbb{R}_+$, avec $(\mathrm{e}^x)^*(\lambda) = \lambda \log(\lambda/\mathrm{e})$ pour $\lambda \geq 0$, où $\lambda \to -\lambda \log \lambda$ est la fonction d'entropie de Boltzmann-Shannon.

3.4.5 La transformée I^*_C (dite parfois fonction support de la fonction indicatrice I_C du convexe C) a été utilisée dans le théorème 3.1 :

$$I^*_C(v) = \sup_{x\in C}\langle v, x\rangle$$

qui a pour domaine $\mathbb{R}^n$ si C est borné. Si C est un cône convexe, on démontre que $I^*_C = I_{C^+}$ où C^+ est le cône dual de C : $C^+ = \{u, \langle u, v\rangle \geq 0, v \in C\}$.

3.4.6 Soit S une partie de E et U_S valant $+\infty$ sauf sur son domaine S où $U_S(x) = \|x\|^2/2$. Alors, en écrivant $2\langle x, \lambda\rangle - \|x\|^2 = \|\lambda\|^2 - \|x-\lambda\|_2^2$, on obtient $U^*_S(\lambda) = 1/2(\|\lambda\|_2^2 - d_S(\lambda)^2)$.

3.4.7 Si $U : X \in S_n^{++} = -\log \det X$, alors $U^*(X) = -\log \det(-Y) - n$ avec domaine $\mathbf{dom}\, U^* = S_n^{++}$. ◁

Lemme 3.3: *(1) Si $U(x,y) = V(x) + W(y)$, alors $U^*(x,y) = V^*(x) + W^*(y)$.*

(2) Soit $a > 0$ et $U(x) = aV(x)$. Alors $U^*(y) = aV^*(y/a)$.

(3) Si $U(x) = V(x) + \langle a, x\rangle + b$, alors $U^*(y) = V^*(y-a) - b$. Si $U(x) = V(x-b)$, alors $U^*(y) = \langle b, y\rangle + V^*(y)$. Si A est régulière et $U : x \mapsto U(x) = V(Ax)$, alors $U^*(y) = V^*({}^{\intercal}A^{-1}y)$

(4) [convolution infimale] Si $U(x) = \inf_u(V(u) + W(x-u))$, alors $U^*(y) = V^*(y) + W^*(y)$.

Démonstration. À rédiger... □

Le théorème suivant (établi dans le cas U différentiable seulement) assure ce qui a été vérifié dans les exemples précédents, *i. e.* le caractère involutif de la transformée de Legendre

Théorème 3.2: *Si l'application $U : \mathbb{R}^n \to \mathbb{R}$ est convexe fermée, alors $U = (U^*)^*$.*

Démonstration. Commençons par le cas où la fonction U est différentiable. On a

$$\begin{aligned}(U^*)^*(z) = \sup_y(\langle z, y\rangle - U^*(y)) &= \sup_y(\langle z, y\rangle - \sup_x(\langle y, x\rangle - U(x)) \\ &= \sup_y \inf_x(U(x) - \langle y, x - z\rangle),\end{aligned}$$

de telle sorte que, en choisissant $x = z$ pour majorer l'inf, $(U^*)^*(z) \leq U(z)$.

Par ailleurs, vu que U est convexe et si U est différentiable[3]

$$U(x) \geq U(z) + \langle \nabla U(z), z - x\rangle$$

d'où

$$\inf_x[U(x) - \langle \nabla U(z), z - x\rangle] \geq U(z)$$

3. Si U n'est pas différentiable, on a utilisera un w tel que l'hyperplan $f(z) - \langle w, x\rangle$ sépare $f(z)$ de l'épigraphe de U.

$U(x) = V^*(x)$	$\mathbf{dom}\, U$	$V(\xi) = U^*(\xi)$	$\mathbf{dom}\, V$
$W(ax) \quad (a \neq 0)$		$W^*(\xi/a)$	
$W(b+x)$		$W^*(\xi) - b\xi$	
$aW(x) \quad (a > 0)$		$aW^*(\xi/a)$	
0	$\mathbb{R}^n$	0	$\{0\}$
0	$\mathbb{R}_+$	0	$-\mathbb{R}_+$
0	$[-1,1]$	$\lvert\xi\rvert$	$\mathbb{R}$
0	$[0,1]$	ξ^+	$\mathbb{R}$
$\lvert x\rvert^p/p \quad (1 < p \in \mathbb{R})$	$\mathbb{R}$	$\lvert\xi\rvert^q/q \quad (p^{-1} + q^{-1} = 1)$	$\mathbb{R}$
$\lvert x\rvert^p/p \quad (1 < p \in \mathbb{R})$	$\mathbb{R}_+$	$\lvert\xi^+\rvert^q/q \quad (p^{-1} + q^{-1} = 1)$	$\mathbb{R}$
$-x^p/p \quad (p \in (0,1)$	$\mathbb{R}_+$	$-(-\xi)^q/q \quad (p^{-1} + q^{-1} = 1)$	$-\mathbb{R}_{++}$
$\sqrt{1+x^2}$	$\mathbb{R}$	$-1\sqrt{1-\xi^2}$	$[-1,1]$
$-\log x$	$\mathbb{R}_{++}$	$-1 - \log(-\xi)$	$-\mathbb{R}_{++}$
$\operatorname{ch} x$	$\mathbb{R}$	$\operatorname{sh}^{-1}(\xi) - \sqrt{1+\xi^2}$	$\mathbb{R}$
$\operatorname{sh} \lvert x\rvert$	$\mathbb{R}$	$\xi \operatorname{argch} \xi - \sqrt{\xi^2-1} \quad (\lvert\xi\rvert > 1), 0 \quad \text{sinon}$	$\mathbb{R}$
$-\log\cos x$	$(-\pi/2, \pi/2)$	$\xi \operatorname{arctg} \xi + \frac{1}{2}\log(1+\xi^2)$	$\mathbb{R}$
$\log \operatorname{ch} x$	$\mathbb{R}$	$\xi \operatorname{arcth} \xi + \frac{1}{2}\log(1-\xi^2)$	$(-1,1)$
e^x	$\mathbb{R}$	$\xi\log\xi - \xi$	$\mathbb{R}_+$
$\log(1+\mathrm{e}^x)$	$\mathbb{R}$	$\xi\log\xi + (1-\xi)\log(1-\xi)$	$[0,1]$
$-\log(1-\mathrm{e}^x)$	$-\mathbb{R}_+$	$\xi\log\xi - (1+\xi)\log(1+\xi)$	$\mathbb{R}_+$

TABLE 1. Exemples de conjuguées de Fenchel-Legendre

et

$$(U^*)^*(z) = \sup_y \inf_x [U(x) - \langle y, z - x\rangle] \geq U(z)$$

Ainsi vient d'être démontrée l'égalité $(U^*)^* = U$. On a utilisé le fait que $\nabla U(z)$ donne un hyperplan en-dessous du graphe de U : l'existence d'un tel hyperplan d'appui est garantie pour des fonctions convexes non nécessairement différentiables. □

△ REMARQUE 3.3: Une fonction U convexe sur $\mathbb{R}$ n'est pas nécessairement continue. Ainsi, la fonction constante égale à 1 sur $(-1,1)$, valant 2 pour $|x| = 1$ et infinie sur le complémentaire de $[-1,1]$. ▽

3. Convexité et régularité

THÉORÈME 3.3: *Soit C ouvert convexe de $\mathbb{R}^n$. La fonction U convexe sur C et à valeurs finies est continue sur C.*

DÉMONSTRATION. Soit $x_0 \in C$. Soit K_0 un cube (fermé) contenant x_0 en son intérieur : la fonction U est majorée sur K_0 par un majorant de U sur l'ensemble (fini) des sommets de K_0 multiplié par 2^n (raisonner par récurrence sur n). Soit B_0 une boule ouverte centrée en x_0 de rayon r_0 et contenue dans K_0 : en écrivant $x_0 = (x_0+u)/2+(x_0-u)/2$, on a $2U(x_0) \leq U(x_0+u)+U(x_0-u)$, soit $U(x_0+u) \geq 2U(x_0) - \sup_{v\in K_0} |U(v)|$ sur B_0. Ainsi U est bornée sur la boule B_0 et par suite continue sur un voisinage de x_0, d'après le lemme suivant valable dans tout espace vectoriel normé (de dimension finie ou pas). □

LEMME 3.4: *Soit U une fonction convexe définie sur un ouvert convexe C. Si u est bornée sur C, elle y est localement lipschitzienne.*

DÉMONSTRATION. Soient m, M des bornes de U : $m \leq U(x) \leq M$ pour $x \in C$. Soit $x_0 \in C$ et $\delta > 0$ tel que $B(x_0, 2\delta) \subset C$. Soient $y, y' \in B(x_0, \delta)$ distincts. Alors

$$y'' = y' + \delta \frac{y'-y}{\|y'-y\|}$$

est dans la boule $B(x_0, 2\delta) \subset C$ et y' est dans le segment $[y, y'']$

$$y' = \frac{\|y'-y\|}{\delta + \|y'-y\|} y'' + \frac{\delta}{\delta + \|y'-y\|} y$$

d'où par convexité (cf. Fig. III.2 pour une vision géométrique de cette inégalité sur la droite passant par y et y')

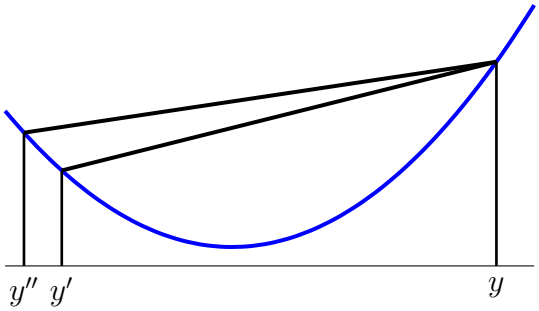


FIGURE III.2 . La pente de la corde $[U(y'), U(y)]$ majore celle de $[U(y''), U(y)]$: $(U(y)-U(y'))/(y-y') \geq (U(y)-U(y''))/(y-y'')$, soit $U(y')-U(y) \leq C|y-y'|$ avec $C = 2\sup|U(u)|/\delta$ vu que $y'' = y' - \delta$.

$$U(y')-U(y) \leq \frac{\|y'-y\|}{\delta+\|y'-y\|}[U(y'')-U(y)] \leq \frac{\|y'-y\|}{\delta}(M-m).$$

et par symétrie

$$|U(y')-U(y)| \leq \frac{M-m}{\delta}\|y'-y\|$$

Ainsi la fonction U est continue en x_0 [4] et sur C tout entier. □

△ REMARQUES 3.4:

(1) Une forme linéaire est convexe : si E est un espace vectoriel de dimension infinie, il existe des formes linéaires non continues. Par ailleurs, l'application U définie sur $[0,1]$ par $U(x) = x^2$ si $x < 1$ et $U(1) = 2$ est convexe, mais non continue en $x = 1$: en dimension finie, le défaut de continuité d'une fonction convexe n'a lieu qu'au bord du domaine de définition.

4. On a en fait démontré que la fonction U est localement lipschitzienne.

(2) En fait, on a montré qu'une fonction convexe U est localement lipschitzienne. On citera le théorème de Rademacher qui assure la différentiabilité presque partout d'une fonction localement lipschitzienne ▽

Une fonction convexe n'est pas nécessairement différentiable en un point intérieur de son domaine, comme par exemple la fonction valeur absolue en $t = 0$. Néanmoins, il y a dérivabilité partielle pour les fonctions d'une variable

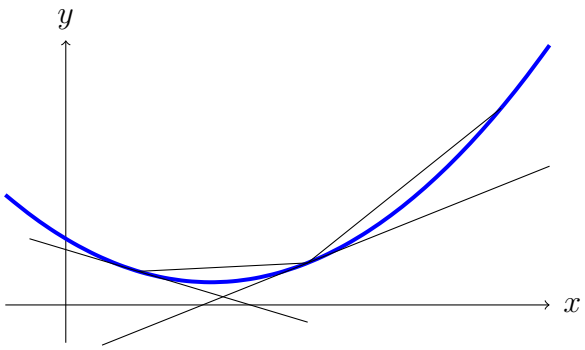


FIGURE III.3 . La convexité est équivalente à la croissance des pentes des cordes, tangentes comprises.

LEMME 3.5: *Soit U définie sur l'intervalle ouvert I. La fonction U est convexe si et seulement si la fonction U est dérivable à droite et à gauche en tout point de I et $U'_g(s) \leq U'_d(s) \leq U'_g(t) \leq U'_d(t)$ pour $s < t$.*

DÉMONSTRATION. Soit $x_0 \in I$. Si U est convexe, alors la fonction de pente $t \mapsto m_{t,x_0} = (U(t) - U(x_0))/(t - x_0)$ est croissante sur $I \setminus \{x_0\}$. En effet, pour $x_0 < t < s$, on a

$$t = \frac{t-s}{x_0-s}x_0 + \frac{x_0-t}{x_0-s}s$$

et par suite

$$U(t) \leq \frac{t-s}{x_0-s}U(x_0) + \frac{x_0-t}{x_0-s}U(s)$$

soit

$$\frac{U(t)-U(x_0)}{t-x_0} \leq \frac{U(s)-U(x_0)}{s-x_0},$$

d'où la croissance sur $(x_0, \infty) \cap I$, la croissance sur $(-\infty, x_0) \cap I$ découlant d'inégalités analogues. En outre, si il y a croissance de ces pentes, la fonction est convexe.

Ainsi, pour U convexe, il y a existence des limites à droite et à gauche

$$U'_d(x_0) = \lim_{t \mapsto x_0^+} \frac{U(t)-U(x_0)}{t-x_0} = \inf_{t>x_0} \frac{U(t)-U(x_0)}{t-x_0},$$
$$U'_g(x_0) = \lim_{t \mapsto x_0^-} \frac{U(t)-U(x_0)}{t-x_0} = \sup_{t<x_0} \frac{U(t)-U(x_0)}{t-x_0}.$$

avec de plus $U'_g(x_0) \leq U'_d(x_0)$.

Réciproquement, si ces limites existent avec croissance des dérivées à droite/gauche, en utilisant les inégalités des accroissements finis, on établit la croissante de la pente $t \mapsto m_{t,x_0}$. □

PROPOSITION 3.1: *Soit U définie sur un intervalle ouvert I de $\mathbb{R}$.*

(1) Si U est différentiable, sont équivalentes

(a) La fonction U est convexe sur I,

(b) $U(t) \geq U(s) + U'(s)(t-s)$ pour $s, t \in I$,

(c) La fonction U' est croissante sur I.

La fonction U est m-fortement convexe si et seulement si pour tout $x, y \in C$, $\langle \nabla U(x) - \nabla U(y), x - y\rangle \geq m\|x-y\|_1^2$.

(2) Si U est dérivable deux fois sur I, U est convexe si et seulement si $U'' \geq 0$ sur I. Si la dérivée seconde U'' est définie positive en tout point de I, alors la fonction U est strictement convexe sur I.

La fonction U est m-fortement convexe si et seulement si $\operatorname{Hess} U \geq m$.

Démonstration. Si U est convexe différentiable, alors la pente $m_{s,t}$ de la corde $[(s, U(s)), (t, U(t))]$ est fonction croissante de t, avec $\lim_{t\to s} = U'(s)$: c'est exactement l'inégalité $U(t) \geq U(s) + U'(s)(t-s)$. Si cette dernière égalité vaut pour tout s et t, on a $U'(s) \leq m_{s,t} \leq U'(t)$ pour $s < t$: c'est la croissance de la dérivée U'. Enfin, si U' est croissante, alors, l'égalité des accroissements finis donne

$$m_{s,t} = U'(x_{s,t}) \leq U'(x_{t,u}) = m_{t,u}, \quad s \leq x_{s,t} \leq t \leq x_{t,u} \leq u,$$

la croissance $m_{s,t} \leq m_{t,u}$ des pentes des cordes avec $s \leq t \leq u$ assurant donc la convexité de U.

La croissance de U' est équivalente à la positivité de U''. L'assertion finale sur la convexité stricte est obtenue en reprenant le développement précédent avec attention sur les inégalité. □

△ Remarque 3.5: La fonction $t \in \mathbb{R} \mapsto t^4$ est strictement convexe, sans que sa hessienne ne soit partout définie : la dérivée seconde $(t^2)''$ est nulle en $t = 0$. ▽

En plusieurs variables et avec des hypothèses de différentiabilité, la convexité admet des caractérisations exprimées en terme de dérivées :

Proposition 3.2: *Soit U définie sur un ouvert convexe C de $\mathbb{R}^n$.*

(1) Si U est différentiable sur C, les propriétés suivantes sont équivalentes

(a) U est convexe sur C

(b) $U(w) \geq U(v) + \langle \nabla U(v), w - v\rangle$ pour $v, w \in C$;

(c) $\langle \nabla U(w) - \nabla U(v), w - v\rangle \geq 0$ pour $v, w \in C$;

Pour la stricte convexité, on a des inégalités strictes.

(2) Si U est de classe $\mathcal{C}^2$ sur C, alors
- *U est convexe sur C si et seulement si* $\operatorname{Hess} U$ *est positive sur C.*
- *si* $\operatorname{Hess} U$ *est définie positive sur C, alors U est strictement convexe sur C, localement fortement convexe.*

Ainsi, pour une fonction convexe, l'approximation linéaire de U en est un minorant global.

Démonstration. Supposons U convexe. Alors, la fonction $U_{v,w}$ définie par $U_{v,w}(t) = U(s_{v,w}(t))$ avec $s_{v,w}(t) = v + t(w-v)$ est convexe et vérifie l'inégalité $U'_{v,w}(0) \leq U_{v,w}(1) - U_{v,w}(0)$ entre les pentes de la tangente en $t = 0$ et de la droite passant par $(0, U_{v,w}(0))$ et

$(1, U_{v,w}(1))$, soit l'inégalité de (i).2. L'inégalité de (i).3 s'obtient en additionnant les inégalités de (i).2 pour les couples (v, w) et (w, v). Enfin, si les inégalités (i).3 sont vérifiées, la fonction $U_{v,w}$ a une dérivée croissante : pour $t > s$

$$\begin{aligned} U'_{v,w}(t) - U'_{v,w}(s) &= \langle \nabla U(s_{v,w}(t)) - \nabla U(s_{v,w}(s)), w - v \rangle \\ &= \frac{\langle \nabla U(s_{v,w}(t)) - \nabla U(s_{v,w}(s)), s_{v,w}(t) - s_{v,w}(s) \rangle}{t - s} \geq 0 \end{aligned}$$

Pour la partie (ii), il suffit de remarquer que la dérivée seconde de $U_{v,v+w}$ en $t = 0$ est la hessienne en v appliquée au vecteur w : $U''_{v,v+w}(0) = \operatorname{Hess} U(v)(w)$. □

△ Remarque 3.6: L'identité (i).2 énonce que le graphe de la fonction affine $w \to U(v) + \langle \nabla U(v), w - v \rangle$ est au-dessous de celui de la fonction U, ces deux graphes coïncidant en $w = v$. L'identité (i).3 est une expression multidimensionnelle de la monotonie (croissance) de la dérivée. ▽

▷ Exemples 3.5:

3.5.1 Les fonctions e^x, $|x|^\alpha$ avec $\alpha \geq 1$ sur $\mathbb{R}$, $-x^\alpha$ avec $0 \leq \alpha \leq 1$ sur $(0, 1)$, $-\log x$, $x^{-\alpha}$ avec $\alpha \geq 0$, $x \log x$ sur $\mathbb{R}_+$ sont convexes.

3.5.2 La fonction $S \in \mathcal{S}_n^{++} \mapsto -\log \det S$ sur le cône ouvert $\mathcal{S}_n^{++}$ des matrices définies positives est convexe. Il suffit de montrer que l'application $t \in (0, 1) \mapsto -\log \det((1-t)S + tT)$ l'est pour tout $S, T \in \mathcal{S}_n^{++}$. Soit $R \in \mathcal{S}_n^{++}$ telle que $S = R^2$ [5]. Alors

$$\begin{aligned} \log \det(S + tT) &= \log \det((1-t)R^2 + tT) = \log \det(R^2) + \log \det(1 - t + tR^{-1}TR^{-1}) \\ &= \log \det(R^2) + \sum_{i=1}^{n} \log(1 - t + t\lambda_i) \end{aligned}$$

où (λ_i) est le spectre des valeurs propres de la matrice $R^{-1}TR^{-1} \in \mathcal{S}_n^{++}$: le dernier terme est concave, comme somme de fonctions concaves vu le calcul de la dérivée seconde de $t \mapsto \log(1 - t + t\lambda_i)$ qui vaut $-(1 - \lambda_i)^2/(1 - t + t\lambda_i)^2$.

On peut calculer la hessienne de $S \mapsto \log \det S$. On munit l'espace des matrices symétriques du produit scalaire $\langle A, B \rangle = \operatorname{tr}(AB)$. On a $\det(1 + h) = 1 + \operatorname{tr} h + \|h\|\varepsilon(h)$. Par suite $\nabla(\log \det)(S) = S^{-1}$ vu que

$$\log \det(S + h) = \log \det S + \log \det(1 + S^{-1}h) = \log \det S + \operatorname{tr}(S^{-1}h) + \|h\|\varepsilon(h).$$

La différentielle seconde est la dérivée de la fonction $S \mapsto S^{-1}$:

$$(S + h)^{-1} = (1 + S^{-1}h)^{-1}S^{-1} = (1 - S^{-1}h + \|h\|\varepsilon(h))S^{-1} = S^{-1} - S^{-1}hS^{-1} + \|h\|\varepsilon(h)$$

soit $-D_S^2(\log \det)(h) = S^{-1}hS^{-1}$. Cette application est bien symétrique

$$\langle S^{-1}hS^{-1}, k \rangle = \operatorname{tr}(S^{-1}hS^{-1}k) = \operatorname{tr}(hS^{-1}kS^{-1}) = \langle h, S^{-1}kS^{-1} \rangle$$

et définie positive

$$\langle S^{-1}hS^{-1}, h \rangle = \operatorname{tr}(S^{-1}hS^{-1}h) = \operatorname{tr}(R^{-1}hR^{-1}R^{-1}hR^{-1}) = \operatorname{tr}((R^{-1}hR^{-1})^2)$$

où R est la racine carrée positive de S.

3.5.3 La fonction *quadratique/linéaire* définie sur $\mathbb{R} \times \mathbb{R}_*$ par

$$f(x, y) = x^2/y, \quad (x, y) \in \mathbb{R} \times \mathbb{R}_+$$

5. Si (σ_j) est le spectre de S, il existe une matrice orthogonale P telle que $S = {}^{\mathsf{T}}P \operatorname{diag}(\sigma_j)P$: la matrice $T = {}^{\mathsf{T}}P \operatorname{diag}(\sqrt{\sigma_j})P$ convient.

est convexe sur $\{y > 0\}$, vu la positivité de sa matrice hessienne

$$\operatorname{Hess} F(x,y) = \frac{2}{y^3}{}^{\mathsf{T}}(y,-x)(y,-x),$$

concave sinon.

3.5.4 La fonction *log/somme/exp* définie par $f(x) = \log \sum_{k=1}^n \mathrm{e}^{x_k}$ est convexe, comme il résultera de la positivité de sa hessienne [6]

$$\operatorname{Hess} f(\log(z)) = \frac{1}{\operatorname{sum}(z)} \operatorname{diag}(z) - \frac{1}{\operatorname{sum}(z)^2} z^{\mathsf{T}} z, \quad z \in \mathbb{R}^k_{*+}$$

vu la positivité de

$$\operatorname{Hess} f(\log(z))(v) = \frac{\operatorname{sum}(z * v^2)\operatorname{sum}(z) - \operatorname{sum}(z * v)^2}{\operatorname{sum}(z)^2}$$

en raison de l'inégalité de Cauchy-Schwarz. La concavité de la moyenne géométrique $f(x) = (\prod_{k=1}^n x_k)^{1/n}$ résulte de calculs analogues. ◁

Vu qu'une fonction convexe d'une variable admet des dérivées à droite, la définition suivante est bien justifiée.

DÉFINITION 3.6: *La* dérivée directionnelle $U'(x;d)$ *de la fonction convexe* U *est l'application définie suivant*

$$(x,d) \in C \times \mathbb{R}^n \mapsto U'(x;d) = \lim_{t \to 0^+} \frac{U(x+td) - U(x)}{t}$$

LEMME 3.6: *Soit* x *intérieur au domaine de la fonction convexe* U*. L'application* $U'(x;\cdot) : d \in \mathbb{R}^n \mapsto U'(x;d)$ *est positivement homogène de degré 1 et convexe. Elle vérifie* $U(y) - U(x) \geq U'(x; y-x)$ *pour* $x, y \in C$*.*

DÉMONSTRATION. L'homogénéité résulte de l'identité

$$U'(x;\rho d) = \rho \lim_{t \to 0^+} \frac{U(x+t\rho d) - U(x)}{\rho t} = \rho U'(x;d),$$

et la convexité de l'identité

$$\begin{aligned} \frac{U(x+t(\alpha_1 d_1 + \alpha_2 d_2)) - U(x)}{t} &= \frac{U(\alpha_1(x+td_1) + \alpha_2(x+td_2)) - U(x)}{t} \\ &\leq \frac{\alpha_1(U(x+td_1) - U(x)) + \alpha_2(U(x+td_2) - U(x))}{t} \end{aligned}$$

suivie d'un passage à la limite en $t \to 0^+$

$$U'(x; \alpha_1 d_1 + \alpha_2 d_2) \leq \alpha_1 U'(x;d_1) + \alpha_2 U'(x;d_2).$$

Enfin, la fonction $U_{x,y} : \lambda \in [0,1] \to U(x + \lambda(y-x))$ est convexe : la pente $P_{x,y}s \in (0,1) \mapsto (U(x+s(y-x)) - U(x))/s$ est croissante, on a donc $U'(x; y-x) = P'_{x,y}(0^+) \leq P_{x,y}(1) = U(y) - U(x)$, ce qui conclut la preuve. □

Le théorème suivant, dû à Alexandroff, énonce une régularité presque $\mathcal{C}^2$ pour une fonction convexe [40].

6. On utilise ici le calcul de R sur les matrices et sa fonction sum.

THÉORÈME 3.4: *Soit C convexe ouvert de $\mathbb{R}^n$ et $U : C \to \mathbb{R}$ convexe. Il existe un ensemble de mesure pleine $\widetilde{C}$ tel que U ait en tout point x de $\widetilde{C}$ un développement de Taylor à l'odre 2,* i. e. *l'existence de $v_x \in \mathbb{R}^n$ et d'une matrice A_x symétrique d'ordre n tels que*

$$U(x+h) = U(x) + \langle v_x, h\rangle + \frac{1}{2}A_x[h] + \varepsilon(\|h\|^2).$$

Terminons sur les liens entre ellipticité/coercivité (cf. définition 2.3) et convexité :

LEMME 3.7: *(i) Si J est de classe $\mathcal{C}^2$, J est α-elliptique si et seulement si*

$$\text{Hess}\, J(x)(v) \geq \alpha\|v\|^2, \quad x \in V, u \in \mathbb{R}^n. \tag{52}$$

(ii) Si J est elliptique, alors J est strictement convexe et coercive ($J(x) \to \infty$ si $\|x\| \to \infty$) avec

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

DÉMONSTRATION. Si J est α-elliptique

$$\langle \nabla J(x+t(y-x)) - \nabla J(x), t(x-y)\rangle \geq \alpha t^2\|x-y\|^2$$

d'où après division par t^2 et $t \to 0+$ l'inégalité (52). Réciproquement, on minore simplement, après avoir écrit la formule de Taylor avec reste intégral de la fonction $t \in [0,1] \mapsto \langle \nabla J(x+t(y-x)), (y-x)\rangle$,

$$\langle \nabla J(y) - \nabla J(x), (y-x)\rangle = \int_0^1 \text{Hess}\, J(x+t(y-x))(y-x)dt \geq \alpha\|y-x\|^2.$$

Pour (ii), la stricte convexité vaut car Hess J est définie positive. En outre, avec la formule de Taylor

$$J(y) = J(x) + \langle \nabla J(x), y-x\rangle + \int_0^1 (1-t)\langle \text{Hess}\, J(x+t(y-x))(y-x)dt$$

on obtient la minoration $J(y) - J(x) \geq \langle \nabla J(x), y-x\rangle + \alpha/2\|y-x\|^2$. □

4. Programmation convexe

DÉFINITION 3.7: *Soient C un convexe de $\mathbb{R}^n$, U une fonction convexe définie sur C, A une matrice d'ordre (m,n), b un vecteur $\mathbb{R}^m$ et $h_j, j = 1,\dots,p$ des fonctions concaves. Le programme convexe associé est le problème d'optimisation*

$$\inf_{\substack{Ax=b \\ h_j(x)\geq 0, j=1,\dots,p}} U(x). \tag{PC}$$

Si U est concave, le problème suivant, équivalent à la recherche du minimum de la fonction convexe $-U$,

$$\sup_{\substack{Ax=b \\ h_j(x)\geq 0, j=1,\dots,p}} U(x)$$

est dit aussi convexe.

Par concavité des h_j, la partie $R = C \cap \cap_{j=1}^p h_j^{-1}(\mathbb{R}_+) \cap \{Ax = b\}$ des points réalisables du programme *(PC)* est fermée convexe : un problème d'optimisation convexe est donc de la forme générale $\inf_{x\in C} U(x)$ où la fonction U et la partie C sont convexes. Le théorème 3.1 a énoncé comment tout convexe fermé C est égal à l'intersection de demi-espaces $\{\langle v, x\rangle \geq C_v\}$ avec $v \in E \setminus \{0\}$: ainsi, le problème d'optimisation $\inf_{x\in C} U(x)$ peut être présenté comme $\inf_{f_a(x)\geq 0, a\in A} U(x)$ avec une famille $(f_a)_{a\in A}$ de formes affines. En fait, il y

a beaucoup de problèmes d'optimisation qui peuvent être présentés comme des problème d'optimisation convexe !

▷ Exemples 3.6:

3.6.1 Les deux programmes

$$\inf_{\substack{x/(1+y^2)\geq 0\\(x+y)^2\leq 0}}[x^2+y^2] \qquad \inf_{\substack{x\geq 0\\x+y=0}}[x^2+y^2]$$

sont différents, mais néanmoins équivalents, le second étant convexe.

3.6.2 Le programme $\min_x(\max_{j=1}^p[\langle a_j,x\rangle+b_j])$ est équivalent au programme convexe (en fait linéaire) $\min_{\langle a_j,x\rangle+b_j\leq t,j=1,\dots,m}[t]$. ◁

△ Remarques 3.7:

(1) Le problème $\max_{x\in C} U(x)$ où U est convexe n'est pas un problème d'optimisation convexe : cela peut être un problème difficile, tant pour l'analyse des solutions que pour leur calcul numérique.

(2) La classe importante des programmes linéaires correspond aux programmes avec fonction d'objectifs linéaire et contraintes affines en égalités ou inégalités

Lemme 3.8: *Quitte à rajouter des variables, un programme linéaire peut être mis sous l'une ou l'autre des deux formes*

$$\inf_{Ax\leq b}\langle h,x\rangle, \qquad \inf_{B(y',y'')=v,y'\geq 0}\langle h,(y',y'')\rangle,$$

où A et B sont des opérateurs linéaires.

Démonstration. Un programme linéaire est de la forme

$$\inf_{\substack{\langle g_i,x\rangle=\alpha_i,i=1,\dots,k\\\langle h_j,x\rangle\leq\beta_j,j=1,\dots,\ell}}\langle h,x\rangle$$

Chaque égalité $\langle g_i,x\rangle=\alpha_i$ est équivalente à la conjonction des deux inégalités $-\langle g_i,x\rangle\leq-\alpha_i, \langle g_i,x\rangle\leq\alpha_i$, l'ensemble des inégalités se rassemblant ainsi dans l'inégalité vectorielle, composante par composante, $Ax-b\geq 0$, où A est une matrice à $m=2k+\ell$ lignes et n colonnes : cela donne une présentation suivant la première forme. Pour obtenir la seconde forme à partir de la première, on introduit des variables, dites d'*ajustement* ou d'*écart*, $s_i, i=1,\dots,m$ telles que l'inégalité $Ax\leq b$ soit équivalente à $\begin{pmatrix}A & \mathrm{Id}\end{pmatrix}^{\mathsf{T}}(x \quad s)=Ax+s=b$ avec $s=(s_i)\geq 0$: c'est la deuxième forme. Cette deuxième forme est équivalente à la première d'après l'argument précédent. □

(3) Le programme en moindres carrés $\inf_{x\in C}\|Ax-b\|_2^2$ est un exemple de programme convexe quadratique. Le programme $\inf_x[\|Ax-b\|_2^2+\|x\|_1]$ est convexe, avec une fonction d'objectif non lisse.

(4) Définissant la partie convexe $C_g=\{x\geq 0, y\geq 0, g(x,y)\geq 0\}$, la fonction $g(x,y)=xy-1$ n'est ni convexe ni concave vu que $\mathrm{Hess}(xy)$ n'est ni positive ni négative. Cependant les fonctions $\overline{g}(x,y)=\sqrt{xy}-1$ et $\widetilde{g}(x,y)=\log x+\log y$ sont concaves et peuvent être prises à la place de g pour définir la partie C_g. Par ailleurs, la fonction g est quasi-concave.

(5) La meilleure approximation en norme uniforme de la fonction $t \in [0,1] \mapsto \operatorname{sh} t$ par des fonctions linéaires $t \mapsto \alpha t$ est un programme linéaire

$$\min_{\alpha} \|\alpha t - \operatorname{sh} t\|_{\mathcal{C}([0,1])} = \min_{\substack{\alpha t - \operatorname{sh} t \le \lambda \\ \alpha t - \operatorname{sh} t \ge -\lambda \\ 0 \le \alpha \le 1}} [\lambda]$$

avec une infinité de contraintes paramétrées par $t \in [0,1]$. ▽

THÉORÈME 3.5: *Soit une fonction U définie sur le convexe C et convexe.*

(i) L'ensemble des points minima de U est, s'il n'est pas vide, convexe. Si la fonction U est strictement convexe, elle a au plus un minimum.

(ii) Tout minimum local de U est un minimum global.

(iii) [Inéquation d'Euler] Si U est différentiable en x_* (au sens où U est la restriction à C d'une fonction convexe différentiable sur un voisinage ouvert de C), le point x_* est point de minimum si et seulement si $\langle \nabla U(x_*), x - x_* \rangle \ge 0$ pour $x \in C$.

(iv) Si x_* est intérieur à C, U différentiable en x_* et $\nabla U(x_*) = 0$, alors x_* est un minimum de U.

DÉMONSTRATION. Soit $U_* = \min_{c \in C} U(c)$ et $x, y \in \operatorname{argmin}(U)$. Alors $U(\theta x + (1-\theta)y) \le \theta U(x) + (1-\theta)U(y) \le U_*$ pour $\theta \in [0,1]$, *i. e.* l'inclusion du segment $[x,y]$ dans $\operatorname{argmin} U$, qui est donc convexe. Si la fonction U est strictement convexe avec deux minima distincts x et y, tout point intérieur au segment $[x,y]$ a une valeur strictement inférieure à $U(x) = U(y)$, ce qui est contradictoire avec la propriété de minima de x et y.

Soit x_* minimum local et supposons l'existence de $y \in C$ tel que $U(y) < U(x_*)$. Alors le point $x_\lambda = x_* + \lambda(y - x_*)$ est arbitrairement près de x_* lorsque $\lambda \to 0^+$ et

$$U(x_\lambda) \le (1-\lambda)U(x_*) + \lambda U(y) < (1-\lambda)U(x_*) + \lambda U(x_*) \le U(x_*),$$

ce qui est contradictoire avec x_* minimum local.

Si x_* est minimum local, alors le point $x_* + t(x - x_*)$ dans le convexe C pour $t \in (0,1)$ vérifie l'inégalité $U(x_* + t(x - x_*)) - U(x_*) \ge 0$ qui implique, en prenant la limite de son quotient par $t \to 0^+$, l'inégalité $\langle \nabla U(x_*), x - x_* \rangle \ge 0$. Si celle-ci vaut, alors l'inégalité (i).2 de la proposition 3.2 indique que x_* est un minimum de U. La propriété (iii) en découle. □

THÉORÈME 3.6: *Soit U une fonction définie sur le convexe C fermé borné d'un espace de Hilbert. Si la fonction U est continue, convexe et minorée, alors U atteint son infimum.*

▷ EXEMPLES 3.7:

3.7.1 La fonction $x \in \mathbb{R} \mapsto \mathrm{e}^x$ est convexe minorée, mais n'atteint pas cependant son minimum : $\mathbb{R}$ n'est pas borné !

3.7.2 La fonction $x \in \ell^2(\mathbb{N}) \mapsto (\|x\|^2 - 1)^2 + \sum_{k=0}^{\infty} x_k^2/(1+k)$ est positive, mais n'atteint pas sa borne inférieure sur la boule unité : cette fonction n'est pas convexe.

3.7.3 L'hypothèse de convexité est essentielle. Soit $E = H^1$ l'espace de Sobolev muni de la norme $\|v\| = (\int_0^1 (v'(x)^2 + v(x)^2)dx)^{1/2}$ et J la fonction définie par $J(v) = \int_0^1 ((|v'(x)| - 1)^2 + v(x)^2)dx$: la fonction J est continue, coercive mais le problème $\inf_{v \in H^1} J(v)$ n'atteint pas son minimum : cet inf est nul, avec suite minimisante (u_n) où u_n est de dérivée constante ± 1 sur les intervalles $[k/n, (k+1)/n], k = 0, \ldots, n$, alternativement croissante et décroissante avec $J(u_n) = 1/(4n)$. ◁

DÉMONSTRATION. Soient $U_0 = \inf_{x \in C} U(x)$, $C_n = \{x : U(x) \le U_0 + 1/n\}$, $\delta_n = \inf_{x \in C_n} \|x\|$ et $\delta_\infty = \sup \delta_n$. Vu que U est convexe, les parties C_n sont convexes : la suite (C_n) est décroissante alors que la suite (δ_n) est croissante bornée de limite finie

$\delta_\infty \le \sup_{x\in C}\|x\|$. On choisit $x_n \in C_n$ tel que $\delta_n \le \|x_n\| \le \delta_n + 1/n$. Dans l'égalité du parallélogramme

$$\|x_n - x_m\|^2 + 4\left\|\frac{x_n + x_m}{2}\right\|^2 = 2\|x_n\|^2 + 2\|x_m\|^2,$$

les termes du membre de droite convergent vers $2\delta_\infty^2$. Par convexité des C_n, $(x_n + x_m)/2$ est dans $C_m \cap C_n$ et donc $\|(x_n + x_m)/2\| \ge \sup(\delta_n, \delta_m)$ Ainsi

$$\limsup \|x_n - x_m\|^2 \le 4\delta_\infty^2 - 4\delta_\infty^2 = 0,$$

i. e. (x_n) est une suite de Cauchy. Si x_* est sa limite, on a $U(x_*) = \lim_n U(x_n) = \inf_x U(x)$, ce qui achève la preuve. □

Ce théorème a comme corollaire l'existence de point de minimum pour une fonction coercive, continue et convexe : il suffit de considérer le convexe borné $C_R = C \cap \{\|u\| \le R\}$ avec R suffisamment grand.

COROLLAIRE 3.1: *Soit U une fonction définie sur la partie convexe fermée non vide C d'un Hilbert E. Si U est convexe, continue et coercive, alors U atteint son minimum sur C.*

Ce théorème a aussi comme corollaire l'important résultat de projection d'un point sur un convexe fermé d'un espace de Hilbert [7].

THÉORÈME 3.7: *Soit C un convexe fermé d'un espace de Hilbert H. Pour tout $v \in H$, il existe un unique vecteur $v_C \in C$ tel que $\|v - v_C\| = \inf_{u\in C}\|v - u\|$. De plus ce vecteur v_C, le* projeté *de v sur C, est caractérisé par la propriété*

$$\langle u - v_C, v - v_C\rangle \le 0, \quad u \in C. \tag{53}$$

L'application, dite de projection, $v \in H \mapsto v_C \in C$ est continue et vérifie $\|w_C - v_C\| \le \|w - v\|$.

▷ EXEMPLE 3.8: Le projeté v_+ de $v = (v_i) \in \mathbb{R}^n$ sur l'orthant $C = \mathbb{R}^n_+$ est donné par $v_+ = (\max(v_i, 0))$. Le projeté v_S sur la boule $\{\|x\| \le 1\}$ d'un v hors la boule est donné par $v_S = v/\|v\|$. ◁

DÉMONSTRATION. Dans la liste d'exemples de fonctions convexes, on a vu que l'application $u \in C \mapsto \|u - v\|$ est convexe : coercive et positive, donc minorée, elle atteint son minimum en $v_C \in C$ d'après le corollaire précédent. Celui-ci est unique, car l'application $u \mapsto \|v - u\|^2$ est strictement convexe. En divisant par θ l'inégalité

$$\|v_C + \theta(u - v_C) - v\|^2 - \|v_C - v\|^2 \ge 0, \quad u \in C \tag{54}$$

la limite lorsque $\theta \to 0^+$ donne l'inégalité (53). Réciproquement, si celle-ci vaut, alors le développement (54) montre que v_C est un minimum de $u \in C \mapsto \|v - u\|$.

D'autre part,

$$\begin{aligned}\|v - w\|^2 - \|v_C - w_C\|^2 = \|v - v_C\|^2 + \|w - w_C\|^2 + 2\langle w_C - w, v - v_C\rangle \\ + 2\langle w_C - w, v_C - w_C\rangle + 2\langle v - v_C, v_C - w_C\rangle\end{aligned}$$

où les deux derniers termes du membre de droite sont positifs d'après (53), ainsi que la somme des trois premiers (égale à $\|v - v_C - w + w_C\|^2$) : par suite, $\|v_C - w_C\| \le \|v - w\|$. □

7. Ce théorème est fondamental pour établir dans le cadre des espaces euclidiens de dimension infinie que sont les espaces de Hilbert les propriétés familières de la dimension finie : somme directe d'un sous-espace vectoriel fermé et de son orthogonal, existence de bases hilbertiennes, faits qui sont admis dans l'appendice **??** et qui se trouvent ainsi établis.

THÉORÈME 3.8 (KKT convexe): *Soit un programme convexe*

$$\min_{\substack{Ax-b=0\\ h_j(x)\geq 0, j=1,\dots,p}} U(x), \tag{55}$$

avec $U, h_1, \dots, h_p$ différentiables. Soit x_* un point qualifié du programme vérifiant les conditions KKT : d'une part $Ax_* = b$ et $h_j(x_*) \geq 0$ avec $S_{x_*} = \{j, h_j(x_*) = 0\}$ l'ensemble des j avec contrainte h_j active en x_*, d'autre part il existe des multiplicateurs de Lagrange $\Lambda_*, (\mu_{j*})_{j\in S_{x_*}}$ avec $\mu_{j*} \geq 0$ tels que le lagrangien

$$\mathcal{L}_{x_*}(x, \Lambda, (\mu_j)_{j\in S_{x_*}}) = U(x) - \langle \Lambda, Ax - b\rangle - \sum_{j\in S_{x_*}} \mu_j h_j(x)$$

vérifie en $x_*, \Lambda^* \in \mathbb{R}^m, (\mu_{j*}) \in \mathbb{R}_+^{S_{x*}}$ la condition

$$\nabla \mathcal{L}_{x_*}(x_*, \Lambda^*, (\mu_{j*})_{j\in S_{x_*}}) = 0. \tag{KKT}$$

Alors x_* est un minimum global du programme (55).

DÉMONSTRATION. Soit le convexe $C = \{Ax = b, h_j(x) \geq 0, j = 1, \dots, p\}$. La fonction $x \in C \mapsto \mathcal{L}_{x_*}(x, \Lambda^*, (\mu_{j*})_{j\in S_{x_*}})$ est convexe, de même que la restriction au segment $[x_*, x] \subset C$ et sa paramétrisation $U_{xx_*} : t \in [0, 1] \mapsto \mathcal{L}_{x_*}(x_* + t(x - x_*), \Lambda^*, (\mu_{j*})_{j\in S_{x_*}})$ qui est de dérivée nulle en $t = 0$ puisque $\mathcal{L}_{x_*}$ est critique en $(x_*, \Lambda_*, M_{S_{x_*}})$: en résulte d'après le (iii) du théorème 3.5 que $t = 0$ est un minimum global de U_{xx_*} sur $[0, 1]$ soit

$$\mathcal{L}_{x_*}(x, \Lambda^*, (\mu_{j*})_{j\in S_{x_*}}) = U_{xx_*}(1) \geq U_{xx_*}(0) = \mathcal{L}_{x_*}(x_*, \Lambda^*, (\mu_{j*})_{j\in S_{x_*}}).$$

Ainsi,

$$U(x) - \sum_{j\in S_{x_*}} \mu_{j*} h_j(x) \geq U(x_*)$$

soit finalement, vue la positivité des h_j sur C et celle des multiplicateurs μ_{j*}, l'inégalité $U(x) \geq U(x_*)$, et ce pour tout $x \in C$. □

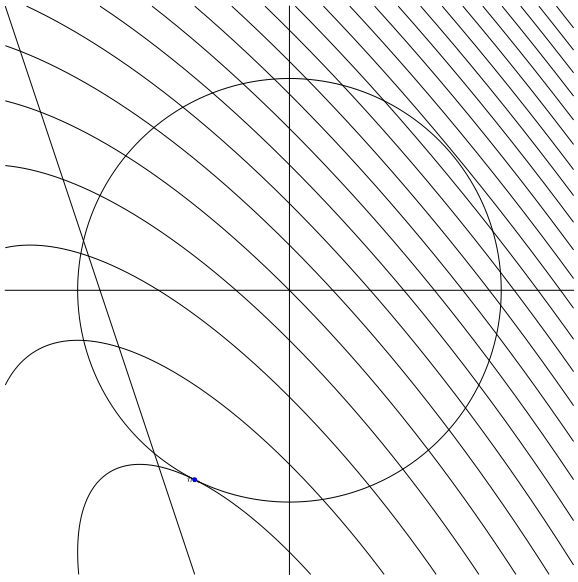


FIGURE III.4 . Les courbes de niveau de la fonction d'objectifs $U(x, y) = 2x^2 + 2xy + y^2 + 10x + 10y$ et le domaine des réalisables $\{x^2 + y^2 \leq 5, 3x + y + 6 \geq 0\}$.

▷ Exemple 3.9: La fonction $(x,y) \in \mathbb{R}^2 \mapsto U(x,y) = 2x^2 + 2xy + y^2 + 10x + 10y$, de hessienne $\begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}$ est strictement convexe sur $\mathbb{R}^2$: son minimum sur tout convexe compact est donc unique. Pour le convexe compact $C = \{x^2 + y^2 \leq 5, -6 \leq 3x + y\}$, le lagrangien à étudier pour les points sur l'arc circulaire de la frontière de C est $\mathcal{L}(x,y,\mu) = U(x,y) - \mu(5 - x^2 - y^2)$ avec gradient

$$\nabla\mathcal{L}(x,y,\mu) = (4x + 2y + 10 + 2\mu x, 2x + 2y + 10 + 2\mu y).$$

Le point $m_* = (-1,-2)$ avec $\mu = 1$ vérifie les conditions KKT : c'est donc l'unique minimum de U restreint à C. La recherche sur l'arc de cercle des points critiques du lagrangien vérifiant KKT passe par la résolution du système

$$(2-\mu)x + y = -5, x + (\mu - 1)y = -5,$$

soit $[(2-\mu)(\mu-1)-1]y = 5(\mu-1), [(2-\mu)(\mu-1)-1]x = 5(2-\mu)$: substituant ces valeurs dans la contrainte $x^2+y^2 = 5$ fournit une équation du quatrième degré pour le coefficient de Lagrange μ avec 1 comme racine évidente et une seule autre racine (négative : elle correspond en fait à un maximum local de U). ◁

La convexité m-forte de J permet d'assurer la convergence des méthodes de descente. Indiquons quelques majorations préparant à la preuve de cette convergence (cf. [7]). L'égalité des accroissements finis en une variable indique, x, y étant donnés, l'existence d'un $z_\theta = \theta x + (1-\theta)y$ tel que

$$\begin{aligned} J(y) &= J(x) + \langle \nabla J(x), y - x\rangle + \frac{1}{2}\operatorname{Hess} J(z_\theta)[y-x] \\ &\geq J(x) + \langle \nabla J(x), y - x\rangle + \frac{m}{2}||y-x||^2 \end{aligned}$$

Le membre de droite est, comme fonction de y, minimum en $y_* = x - \nabla J(x)/m$ avec valeur $J(x) - (2m)^{-1}\|\nabla J(x)\|^2$. Ainsi

$$J(x_*) \geq J(x) - (2m)^{-1}\|\nabla J(x)\|^2$$

Si $\|\nabla J(x)\| \leq \sqrt{2m\varepsilon}$, alors x est proche du minimum x_* : d'une part $J(x) \leq J(x_*) + \varepsilon$, d'autre part,

$$J(x_*) \geq J(x) + \langle \nabla J(x), x_* - x\rangle + \frac{m}{2}\|x - x_*\|^2 \geq J(x) - \|\nabla J(x)\|\|x_* - x\| + \frac{m}{2}\|x - x_*\|^2$$

et donc

$$\|x - x_*\| \leq \frac{2}{m}\|\nabla J(x)\|.$$

5. Sous-gradient et sous-différentiel

Définition 3.8: *Soit U une fonction convexe sur C* [8].

Le vecteur $\xi \in E'$ est dit sous-gradient *de U en $x \in \mathbf{dom}\, U$ si*

$$U(y) \geq U(x) + \langle \xi, y - x\rangle, \quad y \in \mathbf{dom}\, U.$$

Le sous-différentiel *$\partial U(x)$ de U en $x \in \mathbf{dom}\, U$ est la partie constituée de tous les sous-gradients de U en x.*

8. Autrement dit, soit $U : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ convexe avec $x \in C = \mathbf{dom}\, U$ si et seulement si $U(x)$ est fini.

Le sous-différentiel en x de U est donc l'ensemble des pentes ξ de toutes les minorantes affines $\langle \xi, y\rangle + c_x$ de U qui coïncident avec $U(x)$ en x. Le sous-gradient $\xi \in \partial U(x)$ ne donne pas une approximation linéaire valable de U au voisinage de x : il se limite à donner un minorant affine de U, maximal en x.

Si U est différentiable et convexe, alors on a l'inégalité (cf. 3.2.1.b)

$$U(y) \geq U(x) + \langle \nabla U(x), y - x\rangle, \quad y \in \mathbf{dom}\, U.$$

La notion de sous-gradient est une généralisation du gradient des fonctions lisses.

Si U est convexe, il est parfois aisé de calculer les sous-différentiels, justement en raison de cette convexité.

▷ Exemples 3.10:

3.10.1 Pour la fonction v valeur absolue, $\partial_0 v = [-1, 1]$ et $\partial_x v = \mathrm{sign}(x)$ pour x non nul. Pour la fonction partie positive $P(x) = (A(x) + x)/2$, $\partial_0 P = [0, 1]$ et $\partial_x P = (\mathrm{sign}(x) + 1)/2$ pour x non nul.

3.10.2 Soit $U(x) = \max(x^2 + 2|x| - 3, 0)$. Son sous-différentiel est ...

3.10.3 Soit C convexe de $\mathbb{R}^n$. Le sous-différentiel en x de la fonction indicatrice I_C (valant 0 sur C, $+\infty$ sinon) est le cône normal en x à C : $\partial I_C(x) = \{\xi \in \mathbb{R}^n; \langle \xi, x - y\rangle \geq 0, y \in C\}$. ◁

△ Remarque 3.8: La notion de sous-différentiel est très liée à l'analyse convexe, même si la définition n'impose pas ce cadre. Une fonction U ayant un sous-différentiel en tout point est convexe : étant donnés $x, y \in E$, $z_\alpha = \alpha x + (1 - \alpha)y$, si ξ_α est un sous-gradient en z_α, les deux inégalités

$$U(x) \geq U(z_\alpha) + \langle \xi_\alpha, x - z_\alpha\rangle, \quad U(y) \geq U(z_\alpha) + \langle \xi_\alpha, y - z_\alpha\rangle$$

ajoutées l'une à l'autre avec des poids $\alpha, 1 - \alpha$ resp. donnent l'inégalité de convexité. ▽

Proposition 3.3: *Soit U une fonction convexe sur C. Le sous-différentiel $\partial U(x)$ est convexe (éventuellement vide) et fermé.*

Si x est un point intérieur de C et si U est dérivable en x, alors $\partial U(x) = \{\nabla U(x)\}$.

Démonstration. D'après la définition, le sous-différentiel $\partial U(x)$ est intersection de demi-espaces fermés indexés par $y \in \mathbf{dom}\, U$:

$$\partial U(x) = \cap_{y \in E} \{\xi | \langle \xi, y - x\rangle \leq U(y) - U(x)\}.$$

Si x est un point de différentiabilité de U et $\xi \in \partial U(x)$, alors pour tout h, pour $\varepsilon_h > 0$ suffisamment petit, $U(x + th) - U(x) \geq t\langle \xi, h\rangle$ si $t \in]0, \varepsilon_h[$ et donc en passant à la limite $t \to 0^+$ après avoir divisé par $t > 0$, $\langle \nabla U(x), h\rangle \geq \langle \xi, h\rangle$. Vu que h est arbitraire, on a $\xi = \nabla U(x)$. □

On confondra souvent l'ensemble $\partial U(x)$ réduit à un vecteur et le vecteur lui-même.

▷ Exemples 3.11:

3.11.1 Si U est convexe définie sur un intervalle I de $\mathbb{R}$, alors pour x à l'intérieur de I, le sous-différentiel $\partial U(x)$ est égal au segment $[U'_g(x), U'_d(x)]$ d'extrémités les dérivées à droite et à gauche de U. En effet, $\xi \in \partial U(x)$ impose $U(y) - U(x) \geq \xi(y - x)$, ce qui, en faisant $y \to x^-$ ou $\to x^+$, implique les inégalités $U'_g(x) \leq \xi$ et $U'_d \geq \xi$ et donc $\partial U(x) \subset [U'_g(x), U'_d(x)]$. Réciproquement, pour $\xi \in [U'_g(x), U'_d(x)]$, les inégalités

$$\frac{U(x) - U(x_-)}{x - x_-} \leq U'_g(x) \leq \xi \leq U'_d(x) \leq \frac{U(x) - U(x_+)}{x - x_+}, \quad x_- \leq x \leq x_+,$$

impliquent, que y soit supérieur ou inférieur à x, l'inégalité $U(y)-U(x)\leq\xi(y-x)$, soit $\xi\in\partial U(x)$. On a donc établi l'égalité $\partial U(x)\subset[U'_g(x),U'_d(x)]$.

3.11.2 La fonction $x\in[0,\infty)\mapsto-\sqrt{x}$ a un sous-différentiel vide en $x=0$.

3.11.3 La norme euclidienne est lisse, de gradient $x/\|x\|_2$, en dehors de l'origine. En outre, vu Cauchy-Schwarz, l'inégalité $\|y\|_2\geq\langle\xi,y\rangle$ vaut pour tout y si $\|\xi\|_2\leq1$ et ne vaut pas pour $y=\xi$ si $\|\xi\|_2>1$. Ainsi

$$\partial(\|\quad\|_2)(x)=\begin{cases}x/\|x\|_2 & \text{si } x\neq0\\ \{\xi\in\mathbb{R}^n,\|\xi\|_2\leq1\} & \text{si } x=0.\end{cases}$$

3.11.4 À la norme $\|\ \|$ sur E est associée la norme duale $\|\xi\|_d=\sup_{\|x\|\leq1}\langle\xi,x\rangle$ sur son dual E'. Ainsi $\langle\xi,x\rangle\leq\|x\|\,\|\xi\|_d$, avec la majoration (optimale) $\langle\xi,x\rangle\leq\|x\|$ si $\|\xi\|_d\leq1$. Le sous-différentiel de la norme sur E est décrit par

$$\partial(\|\ \|)(x)=\{\xi\in E',\|\xi\|_d=1,\langle x,\xi\rangle=\|x\|\},\quad x\in\mathbf{dom}\,U.$$

En effet, si ξ est un sous-gradient en x, alors pour $t\in(0,1)$, $\|(1\pm t)x\|-\|x\|=\pm t\|x\|\geq\langle(1\pm t)x-x,\xi\rangle=\langle\pm tx,\xi\rangle$, soit $\pm\|x\|\geq\langle\pm x,\xi\rangle$ et donc $\|x\|=\langle x,\xi\rangle$; en outre pour $t>0$ et $u\in E$, l'inégalité $\|x/t+u\|-\|x/t\|\geq\langle\xi,u\rangle$ donne pour $t\to+\infty$ l'inégalité $\|u\|\geq\langle\xi,u\rangle$, soit $\|\xi\|_d\leq1$. Réciproquement, si $\|\xi\|_d=1$, alors $\|y\|\geq\langle y,\xi\rangle$ et si $\langle x,\xi\rangle=\|x\|$, alors $\|y\|-\|x\|=\|y\|-\langle x,\xi\rangle\geq\langle y-x,\xi\rangle$: ξ est un sous gradient en x, ce qui conclut la caractérisation des sous-gradients de la norme en x. ◁

THÉORÈME 3.9: *Soit U convexe fermée propre. Pour x dans l'intérieur de $\mathbf{dom}\,U$, le sous-différentiel $\partial U(x)$ est une partie non vide bornée.*

DÉMONSTRATION. Soit x intérieur au domaine de U. Le point $(x,U(x))$ appartient au bord du convexe $\mathbf{epi}\,U$. Il existe donc un hyperplan d'appui dans $\mathbb{R}^{n+1}$ passant par $(x,U(x))$ et soutenant $\mathbf{epi}\,U$, *i. e.* $\xi_0\in\mathbb{R}^n$ et $\alpha_0\in\mathbb{R}$ avec $\|\xi_0\|_2^2+\alpha_0^2=1$ tels que

$$-\alpha_0t+\langle\xi_0,y\rangle\leq-\alpha_0U(x)+\langle\xi_0,x\rangle,\quad(y,t)\in\mathbf{epi}\,U. \tag{56}$$

Vu que pour tout $t\geq U(x)$ le point (x,t) appartient à $\mathbf{epi}(U)$, on a $\alpha_0\geq0$.

Une fonction convexe est localement lipschitzienne dans l'intérieur de son domaine (cf. la preuve du théorème 3.3). Il existe donc $\varepsilon>0$ et $M>0$ tels que $\overline{B_2(x,\varepsilon)}$ soit incluse dans l'intérieur du domaine de U et $U(y)-U(x)\leq M\|y-x\|$ pour $y\in\overline{B_2(x,\varepsilon)}$ et par suite

$$\langle\xi_0,y-x\rangle\leq\alpha_0(U(y)-U(x))\leq\alpha_0M\|y-x\|_2,\quad y\in B_2(x,\varepsilon).$$

Posant $y=x+\varepsilon\xi_0$, on obtient $\varepsilon\|\xi_0\|^2\leq\alpha_0M\|\xi_0\|$ et donc $\alpha_0>0$. Par suite, avec $\widetilde{\xi}_0=\xi_0/\alpha_0$, on obtient

$$U(y)\geq U(x)+\langle\widetilde{\xi}_0,y-x\rangle,\quad y\in\mathbf{dom}\,U,$$

i. e. l'appartenance de $\widetilde{\xi}_0$ à $\partial U(x)$, qui est donc non vide. Finalement, pour $\xi\in\partial U(x)$ non nul, choisissant $y=x+\varepsilon\xi/\|\xi\|_2$, on a

$$\varepsilon\|\xi\|_2=\langle\xi,y-x\rangle\leq U(y)-U(x)\leq M\|y-x\|_2=M\varepsilon,$$

et donc $\|\xi\|_2\leq M$ ce qui assure que $\partial U(x)$ est borné. □

On peut généraliser un peu ce caractère borné du sous-différentiel

THÉORÈME 3.10: *Soit $U:\mathbb{R}^n\to\mathbb{R}$ convexe et soit K un compact de $\mathbb{R}^n$. Alors $\cup_{x\in K}\{\partial U(x)\}$ est borné.*

Démonstration. Supposons qu'il existe une suite (ξ_k) de sous-gradients non bornés avec $\xi_k \in \partial_{x_k} U$ et $x_k \in K$. Vu que K est compacte, et à suite extraite près, on peut supposer que les suites (x_k) et $(d_k = \xi_k/\|\xi_k\|)$ sont convergentes vers x_∞ et d_∞ resp.. Le caractère sous-gradient de ξ_k permet d'écrire

$$U(x_k + d_k) \geq U(x_k) + \langle \xi_k, d_k \rangle = U(x_k) + \|\xi_k\|.$$

Alors, vu la continuité de U,

$$U(x_\infty + d_\infty) - U(x_\infty) \geq \limsup_{k\to\infty} \|\xi_k\|.$$

est fini, impliquant que la suite $(\|\xi_k\|)$ est bornée, ce qui est contradictoire avec l'hypothèse. □

Théorème 3.11: *Soit U convexe fermé, x point intérieur de son domaine et $d \in \mathbb{R}^n$. Alors $\partial_d U'(x;d)(0) = \partial U(x)$ et $U'(x;d) = \max_{\xi\in\partial U(x)} \langle \xi, d\rangle = (\iota_{\partial U})^*(d)$.*

Ainsi la dérivée directionnelle $d \mapsto U'(x;d)$ apparaît comme la fonction support du sous-différentiel $\partial U(x)$, qui est donc caractérisé par l'ensemble des dérivées directionnelles en x.

Démonstration. On a pour tout $\xi \in \partial U(x)$

$$U'(x;d) = \lim_{t\to 0^+} (U(x+td) - U(x))/t \geq \langle \xi, d\rangle, \quad d \in \mathbb{R}^n. \tag{57}$$

Alors le sous-différentiel $\partial_d U'(x;\cdot)(0)$ de la fonction $d \mapsto U'(x;d)$ en $d = 0$ n'est pas vide et $\partial_x U(x) \subset \partial_d U'(x;\cdot)(0)$.

Réciproquement, soit $\xi_0 \in \partial_d U'(x;\cdot)(0)$. Vu que $d \mapsto U'(x;d)$ est convexe, la croissance des cordes pour une fonction convexe d'une variable donne l'inégalité

$$U(y) - U(x) \geq U'(x; y-x)$$

puis

$$U(y) \geq U(x) + U'(x;y-x) \geq U(x) + \langle \xi_0, y-x\rangle,$$

où la dernière inégalité provient de l'hypothèse $\xi_0 \in \partial_d U'(x;\cdot)(0)$ et l'inégalité entre les extrêmes exprime $\xi_0 \in \partial U(x)$: on a donc montré $\partial_d U'(x;d)(0) \subset \partial U(x)$ et donc l'égalité $\partial_d U'(x;d)(0) = \partial U(x)$.

Soit $\xi_d \in \partial_d U'(x;d)$, sous-différentiel de la fonction convexe $d \mapsto U'(x;d)$ qui n'est pas vide d'après le théorème 3.9 précédent. Alors, pour tout $v \in \mathbb{R}^n$ et $t > 0$

$$tU'(x;v) = U'(x;tv) \geq U'(x;d) + \langle \xi_d, tv - d\rangle.$$

En faisant $t \to +\infty$, on a $U'(x;v) \geq \langle \xi_d, v\rangle$ pour tout $v \in \mathbb{R}^n$, ce qui assure $\xi_d \in \partial_d U'(x;\cdot)(0) = \partial U(x_0)$, alors que la limite quand $t \to 0^+$ donne l'inégalité $0 \geq U'(x;d) - \langle \xi_d, d\rangle$. Cette inégalité, couplée avec l'inégalité (57), entraîne donc l'égalité $U'(x;d) = \langle \xi_d, d\rangle$ et le fait que le max de l'énoncé est bien atteint (en ξ_d !). □

Corollaire 3.2:
Pour $U : \mathbb{R}^m \to \mathbb{R}$ convexe, $A : \mathbb{R}^n \to \mathbb{R}^m$ linéaire et $b \in \mathbb{R}^m$, alors

$$\partial[U \circ (A\cdot + b)](x) = {}^T\!A\,[\partial U(Ax+b)]\,.$$

En particulier, si U est une fonction convexe d'une variable et U_i est définie par $U_i : x \in \mathbb{R}^n \mapsto U(x_i)$, alors $\partial U_i(x) = 0_{i-1} \times \partial U(x_i) \times 0_{n-i}$ $(\subset \mathbb{R}^n)$.

DÉMONSTRATION. Pour $\xi \in \partial U(Ax+b) \subset \mathbb{R}^m$

$$U(Ay+b) - U(Ax+b) \geq \langle \xi, Ay - Ax\rangle = \langle {}^{\mathsf{T}}A\xi, y-x\rangle,$$

soit l'inclusion ${}^{\mathsf{T}}A\,[\partial U(Ax+b)] \subset \partial[U\circ(A\cdot+b)](x)$. L'autre inclusion est plus subtile. Soit $U_{A,b} : x \in \mathbb{R}^n \mapsto U(Ax+b) \in \mathbb{R}$. Alors

$$\max_{\eta\in\partial U_{A,b}(x)} \langle \eta, d\rangle = U'_{A,b}(x;d) = U'(Ax+b;Ad) = \max_{\xi\in\partial U(Ax+b)} \langle \xi, Ad\rangle = \max_{\eta\in {}^{\mathsf{T}}A\partial U(Ax+b)} \langle \eta, d\rangle$$

Du fait qu'un convexe fermé est caractérisé par sa fonction caractéristique, on a donc $\partial U_{A,b}(x) = {}^{\mathsf{T}}A\partial U(Ax+b)$.

La dernière assertion correspond au cas particulier où A est la projection π_i (linéaire) sur la $i$ème variable. □

Le sous-différentiel de U est lié à la transformée de Fenchel-Legendre U^* :

THÉORÈME 3.12: *Soit U convexe sur $\mathbb{R}^n$. Alors $\xi \in \partial U(x)$ si et seulement si $U(x) + U^*(\xi) = \langle \xi, x\rangle$.*

DÉMONSTRATION. Le vecteur ξ est dans $\partial U(x)$ si et seulement si $U(y)-U(x) \geq \langle \xi, y-x\rangle$ pour tout y, soit $\langle \xi, x\rangle - U(x) = U^*(\xi)$ d'après la définition de la transformée U^*. □

PROPOSITION 3.4: *Soient $U_j, j = 1,\dots,m$ des fonctions convexes et $\alpha_1 \geq 0$. Alors, pour x intérieur*[9] *à tous les domaines* $\mathbf{dom}\, U_j, j = 1,\dots,m$,

$$\alpha_1 \partial U_1(x) = \partial(\alpha_1 U_1)(x), \quad \partial\left[\sum_j U_j\right](x) = \sum_j \partial U_j(x).$$

Si $S_x = \{i \in [\![1,\dots m]\!], U_i(x) = U(x)\}$ où $U(x) = \max_j U_j(x)$, alors

$$\partial\left[\max_j U_j\right](x) = \mathrm{Conv}\left[\cup_{j\in S_x} \partial U_j(x)\right].$$

DÉMONSTRATION. La première identité pour $\partial(\alpha U)$ résulte immédiatement de la définition du sous-différentiel. Pour le sous-différentiel de la somme $\sum_j U_j$, il suffit de le montrer pour la somme de deux fonctions. Par additivité des dérivées directionnelles, on a

$$\begin{aligned}(i_{\partial(U+V)(x)})^*(d) = (U+V)'(x;d) = U'(x;d) + V'(x;d) &= \max_{\xi\in\partial U(x)} \langle \xi, d\rangle + \max_{\eta\in\partial V(x)} \langle \eta, d\rangle \\ = \max_{\substack{\xi\in\partial U(x)\\ \eta\in\partial V(x)}} \langle \xi+\eta, d\rangle = \max_{\psi\in\partial U(x)+\partial V(x)} \langle \psi, d\rangle &= (i_{\partial U(x)+\partial V(x)})^*(d).\end{aligned}$$

Le théorème 3.11 énonce que la dérivée directionnelle est la fonction d'appui du sous-différentiel et le caractérise, la dernière égalité donne donc $\partial(U+V)(x) = \partial U(x) + \partial V(x)$.

Montrons maintenant la formule du sous-différentiel du max U des fonctions $U_1,\dots,U_m$. Par continuité des U_j en x (intérieur à tous les domaines), il existe un voisinage V_x de x

9. L'énoncé vaut pour x intérieur à $\mathbf{dom}\, U_i$ dans le sous-espace engendré par $\mathbf{dom}\, U_i$, *i. e.* $x \in \mathrm{relint}(\mathbf{dom}\, U_i)$.

où $U(y) = \max_{j\in S(x)} U_j(y)$. Alors

$$\begin{aligned}(i_{\partial U(x)})^*(d) = U'(x;d) &= \lim_{t\to 0^+} \frac{U(x+td)-U(x)}{t} = \lim_{t\to 0^+} \frac{\max_{j\in S_x} U_j(x+td)-U(x)}{t}\\ &= \lim_{t\to 0^+} \max_{j\in S_x} \frac{U_j(x+td)-U_j(x)}{t} = \max_{j\in S_x} U_j'(x;d)\\ &= \max_{j\in S_x} \max_{\xi_j\in\partial U_j(x)} \langle \xi_j, d\rangle = \max_{\substack{\sigma\in\mathfrak{S}_x\\ \xi_j\in\partial U_j(x)}} \sum_{j\in S_x} \sigma_j\langle \xi_j, d\rangle = \max_{\xi\in \mathrm{Conv}(\partial U_j(x), j\in S_x)} \langle \xi, d\rangle\\ &= (i_{\mathrm{Conv}(\partial U_j(x), j\in S_x)})^*(d)\end{aligned}$$

On a introduit le simplexe

$$\mathfrak{S}_x = \left\{ \sigma = (\sigma_j)_{j\in S_x} \in \mathbb{R}^{S_x}, \sigma_j \geq 0, \sum_{j\in S_x} \sigma_j = 1 \right\}$$

pour lequel $\sum_{j\in S_x} \sigma_j a_j = \max_{j\in S_x} a_j$ si $a = (a_j)_{j\in S_x}$ et $\mathrm{Conv}(\partial U_j(x), j\in S_x)$ est l'enveloppe convexe des parties $\partial U_j(x)$ où $j \in S_x$. □

▷ Exemples 3.12:

3.12.1 Soit U la fonction indicatrice de $\mathbb{R}^-$ et V la fonction de domaine $\mathbb{R}^+$ et y valant $V(x) = -\sqrt{x}$. La somme $U+V$ est la fonction indicatrice de $\{0\}$: $\partial(U+V)(0) = \mathbb{R}$ avec $\partial V(0) = \emptyset$.

3.12.2 La norme $\| \ \|_1$ apparaît comme la somme de fonctions dont les sous-différentiels sont aisément donnés :

$$\begin{aligned}\partial\| \ \|_1(x) &= \sum_{j=1}^n 0_{j-1}\partial|\ |(x_j)0_{n-j}\\ &= \{\xi = (\xi_j) \in \mathbb{R}^n; \xi_j = \mathrm{sign}\, x_j \text{ si } x_j \neq 0, \xi_j \in [-1,1] \text{ sinon}\}.\end{aligned}$$

3.12.3 La norme $\| \ \|_1$ dans $\mathbb{R}^n$ est réalisée aussi comme le max des formes linéaires $\ell_s : x \mapsto \langle s, x\rangle$ avec $s \in \{-1,1\}^n \subset \mathbb{R}^n$. On a alors $S_x = \{s \in \{-1,1\}^n; s_i = \mathrm{sign}\, x_i \text{ si } x_i \neq 0, s_i = \pm 1 \text{ sinon}\}$ et

$$\begin{aligned}\partial\| \ \|_1(x) &= \mathrm{Conv}(\partial\ell_s, s\in S_x)\\ &= \{\xi \in \mathbb{R}^n; \xi_i = \mathrm{sign}\, x_i \text{ si } x_i \neq 0, \xi_i \in [-1,1] \text{ sinon}\}.\end{aligned}$$

On retrouve la description de $\partial\| \ \|_1$.

3.12.1 Pour la norme infinie $\| \ \|_\infty$ et x non nul,

$$\partial(\| \ \|_\infty)(x) = \mathrm{Conv}((\mathrm{sign}\, x_j)e_j, j\in S_x)$$

où $S_x = \{j : |x_j| = \|x\|_\infty\}$ alors que le sous-différentiel $\partial(\| \ \|_\infty)(0)$ est le simplexe engendré par les $\pm e_j, j = 1,\dots,n$, *i. e.* le cube $[-1,1]^n$. ◁

La caractérisation suivante d'un minimum d'une fonction convexe est simple

Lemme 3.9: *Soit $U : C \to \mathbb{R}$ une fonction convexe. Le point x_* est un point de minimum pour U si et seulement si le vecteur nul est dans le sous-différentiel $\partial U(x_*)$.*

Si U est fermée, alors x_* est un minimum global de U si et seulement[10] *si $x_* \in \partial U^*(0)$.*

10. Minimiser une fonction revient à analyser le sous-différentiel de la fonction conjuguée en 0.

DÉMONSTRATION. Si $0 \in \partial U(x_*)$, alors $U(y) - U(x_*) \geq \langle 0, y - x_* \rangle$ et donc $x_* \in \operatorname{argmin} U$. Réciproquement, si $U(y) \geq U(x_*)$, alors $U(y) - U(x_*) \geq 0 = \langle 0, x_* - y \rangle$ et donc $0 \in \partial U(x_*)$.

Le théorème 3.12 établit l'équivalence de $\xi \in \partial U(x)$ et de l'égalité $U(x) + U^*(\xi) = \langle \xi, x \rangle$. Ainsi $0 \in \partial U(x_*)$ si et seulement si $U(x_*) + U^*(0) = 0$ si et seulement si $U^{**}(x_*) + U^*(0) = 0$ (car U est supposée fermée) si et seulement si $x_* \in \partial U^*(0)$. □

△ REMARQUE 3.9: Le point x_* est dans $\operatorname{argmin}_{x \in C} U(x)$ si et seulement si il existe $\xi \in \partial U(x_*)$ tel que $\langle \xi, x - x_* \rangle \geq 0$. ▽

On a utilisé le caractère involutif de la transformée de Legendre-Fenchel, qui s'énonce bien dans le cadre des fonctions convexes à valeurs dans $\mathbb{R} \cup \{+\infty\}$.

THÉORÈME 3.13: *Soit U convexe fermée. Alors $(U^*)^* = U$.*

THÉORÈME 3.14 (KKT convexe): *Soit U convexe sur $\mathbb{R}^n$, $h : \mathbb{R}^n \to \mathbb{R}^m$ à composantes $h_1, \ldots, h_m : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ concaves telles qu'il existe un point de faisabilité $\overline{x}$, dit qualifié*[11]*, vérifiant $h(\overline{x}) > 0$.*

Alors le point x_* est solution de $\min_{h(x) \geq 0} U(x)$ si et seulement si $h(x_*) \geq 0$ et il existe $v \in \partial U(x_*)$, $v_j \in \partial(-h_j)(x*)$, $\lambda_j \geq 0$ pour $j = 1, \ldots, m$ tels que

$$v = \sum_{j \in S_{x_*}} \lambda_{*j} v_j$$

*où S_{x_*} est l'ensemble des indices de contraintes actives en x_* (*i. e. *des j tels que $h_j(x_*) = 0$).*

DÉMONSTRATION. Soit $U_* = \min_{h \geq 0} U$ et Φ la fonction convexe définie par

$$\Phi(x) = \max(U(x) - U_*, -h_1(x), \ldots, -h_m(x)), \quad x \in \mathbb{R}^n$$

et S_x l'ensemble des indices de contraintes actives pour x.

Le point x_* est un point de minimum de U sous la contrainte $h \geq 0$ si et seulement si $x_* \in \operatorname{argmin} \Phi$ avec $\Phi(x_*) = 0$. En effet, si $x_* \in \operatorname{argmin}_{h \geq 0} U$, alors $\Phi(x_*) \geq U(x_*) - U_* = 0$ et vu que $-h(x_*) \leq 0$, on a $\Phi(x_*) = 0$. Pour un x quelconque, soit il est contraint et alors $\Phi(x) \geq U(x) - U_* \geq U(x_*) - U_* = 0$, soit il ne l'est pas et il existe i tel que $h_i(x) < 0$ et donc $\Phi(x) \geq -h_i(x) > 0$. Réciproquement, si $x_* \in \operatorname{argmin} \Phi$ avec $\Phi(x_*) = 0$, on a $-h(x_*) \leq 0$, x_* est contraint, puis $U(x_*) - U_* \leq 0$; par ailleurs, un x contraint avec $\Phi(x) \geq 0$ vérifie $U(x) - U_* \geq 0$, ainsi $U_* = U(x_*)$ et tout x contraint vérifie $U(x) \geq U_*$, soit $x_* \in \operatorname{argmin}_{h \geq 0} U$.

Ainsi $x_* \in \operatorname{argmin}_{h \geq 0} U = \operatorname{argmin} \Phi$ *i. e.* si et seulement si

$$0 \in \partial\Phi(x_*) = \operatorname{Conv}(\partial U(x_*), \partial(-h_j)(x_*) \text{ pour les } j \in S_{x_*}).$$

soit si et seulement il existe $(\alpha_j)_{j \in \{0\} \cup S_{x_*}}$ dans $[0, 1]$, $v \in \partial U(x_*)$, $-v_j \in \partial(-h_j)(x_*)$ tels que

$$0 = \alpha_0 v + \sum_{j \in S_{x_*}} \alpha_j (-v_j).$$

Vu que $-v_j \in \partial(-h_j)(x_*)$, on a $\langle v_j, x_* - y \rangle \leq h_j(x_*) - h_j(y)$ pour tout y et tout $j \in S_{x_*}$. Si $\alpha_0 = 0$, les α_j ne sont pas tous nuls alors que $h_j(x_*) = 0$ si $j \in S_{x_*}$, ainsi

$$0 = \sum_{j \in S_{x_*}} \alpha_j \langle v_j, x_* - \overline{x} \rangle \leq \sum_{j \in S_{x_*}} \alpha_j (h_j(x_*) - h_j(\overline{x})) = - \sum_{j \in S_{x_*}} \alpha_j h_j(\overline{x})$$

11. C'est une condition dite de Slater : elle est remplie si le système de contraintes est régulier en x_*. Une telle condition peut aussi apparaître avec des fonctions affines, sans qu'il y ait régularité (par exemple un cône de sommet x_* s'appuyant sur un polygône).

contredisant l'inégalité stricte $h(\overline{x}) > 0$. Ainsi α_0 est non nul et il suffit de poser $\lambda_j = \alpha_j/\alpha_0$ pour $j \in S_{x_*}$. □

On s'intéresse aux problèmes de minimisation

$$\text{argmin}_{x\in\mathbb{R}} \left[\frac{\|x-y\|_2^2}{2} + \lambda\|x\|_1\right], \quad \text{argmin}_{||x||_1\leq t} \|x-y\|_2^2.$$

Le lemme suivant résout ces programmes en dimension 1.

LEMME 3.10: *Soit $\lambda > 0$ et $y \in \mathbb{R}$. Alors*

$$\text{argmin}_{x\in\mathbb{R}} \left[\frac{(x-y)^2}{2} + \lambda|x|\right] = \begin{cases} y-\lambda & \text{si } y \geq \lambda \\ 0 & \text{si } |y| \leq \lambda \\ y+\lambda & \text{si } y \leq -\lambda \end{cases} = \text{sign}(y)(|y|-\lambda)_+.$$

avec

$$\min_{x\in\mathbb{R}} \left[\frac{(x-y)^2}{2} + \lambda|x|\right] = \begin{cases} y^2/2 & \text{si } |y| \leq \lambda, \\ \lambda(|y|-\lambda/2) & \text{sinon.} \end{cases}$$

Soit $t > 0$. Alors

$$\text{argmin}_{|x|\leq t}(x-y)^2 = \begin{cases} y & \text{si } |y| \leq t, \\ \text{sign}(y)t & \text{sinon.} \end{cases}$$

avec

$$\min_{|x|\leq t}(x-y)^2 = \begin{cases} 0 & \text{si } |y| \leq t, \\ (|y|-t)^2 & \text{sinon.} \end{cases}$$

DÉMONSTRATION. Le graphe de la fonction $x \mapsto \frac{(x-y)^2}{2} + \lambda|x|$ est l'union de deux arcs de paraboles de sommets d'abscisses $y-\lambda, y+\lambda$: le point de minimum de la fonction dépend de la position relative de l'origine (singularité de $|x|$) et de ces abscisses.

Pour le deuxième problème de minimisation, la symétrie en x implique pour les points de minimum $x_*(|y|,t) = -x_*(-|y|,t)$: si $|y| \leq t$, le minimum absolu de $(x-y)^2$ est dans le domaine contraint et pour $y > t$, le point de minimum est au bord du domaine contraint en $x = t$. □

Pour le programme analogue [12] en dimension $n > 1$, on a une solution similaire dont une preuve fait appel aux sous-gradients.

LEMME 3.11: *Soit $\lambda > 0$ et $y \in \mathbb{R}^n$. Alors*

$$\text{argmin}_{x\in\mathbb{R}^n} \left[\frac{\|x-y\|_2^2}{2} + \lambda\|x\|_1\right] = (\text{sign}(y_i)(y_i-\lambda)_+)$$

avec

$$\min_{x\in\mathbb{R}^n} \left[\frac{\|x-y\|_2^2}{2} + \lambda\|x\|_1\right] = \sum_{|y_i|\leq\lambda} y_i^2/2 + \sum_{|y_i|>\lambda} \lambda(|y_i|-\lambda/2).$$

L'opérateur $S_\lambda : y \mapsto (\text{sign}(y_i)(y_i-\lambda)_+)$ est appelé opérateur de seuillage ou de troncature lisse. Pour $y \in \mathbb{R}^n$ et $t > 0$, il existe $\lambda(y,t) > 0$ tel que

$$\text{argmin}_{\|x\|_1\leq t} \|x-y\|_2^2 = \begin{cases} y & \text{si } \|y\|_1 \leq t, \\ \{(\text{sign}(y_i)(y_i-\lambda(x,t))_+)\} & \text{sinon.} \end{cases}$$

12. Ce programme de minimisation pénalisé par une norme de type ℓ_1 est dit *LASSO* pour *Least Absolute Selection and Shrinkage Operator*.

DÉMONSTRATION. Les points de minimum existent : dans le premier cas, la fonction est coercive vu le terme $\lambda\| \ \|_1$, dans le second la boule $\{\|x\|_1 \leq t\}$ est compacte. Vu la stricte convexité de la norme euclidienne $\| \ \|_2$, les minima considérés ici sont tous uniques : ils seront notés $x_*(y,\lambda)$ et $x_*(y,t)$.

Soit $U_\lambda : (x,y) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto U_\lambda(x,y) = \|x-y\|_2^2 + \lambda\|x\|_1$ et $U_{y,\lambda} : x \in \mathbb{R}^n \mapsto U_{y,\lambda}(x) = U_\lambda(x,y)$. Pour tout $\varepsilon \in \{-1,1\}^n$, on a $U_\lambda(\varepsilon * x, \varepsilon * y) = U_\lambda(x,y)$ et par suite $x_*(\varepsilon * y, \lambda) = \varepsilon * x_*(y,\lambda)$. On peut donc se limiter à étudier $x_*(y,\lambda)$ pour y avec toutes ses coordonnées positives. On a alors

$$\begin{aligned} x_*(y,\lambda) &= \operatorname{argmin}_{x\in\mathbb{R}^n} \sum_i \left[\frac{(x_i-y_i)^2}{2} + \lambda|x_i|\right] \\ &= \left(\operatorname{argmin}_{x_i\in\mathbb{R}} \left[\frac{(x_i-y_i)^2}{2} + \lambda|x_i|\right]\right) = ((y_i - \lambda/2)_+) \end{aligned}$$

où on a utilisé le lemme 3.10.

On peut aussi chercher le point de minimum x_* (qui est unique) dont le sous-différentiel $\partial U_{y,\lambda}(x_*)$ contient le vecteur nul. On a le calcul de sous-différentiel

$$\partial U_{y,\lambda}(x) = x - y + \lambda\{s \in \mathbb{R}^n ; s_i = \operatorname{sign}(y_i) \text{ si } x_i \neq 0, s_i \in [-1,1] \text{ sinon}\}.$$

Cherchons x_* qui ait le vecteur nul comme sous-gradient dans $\partial U_{y,\lambda}(x_*)$, *i. e.* un vecteur $\xi \in \partial(\| \ _1)(x_*)$ tel que

$$x_{*i} - y_i + \lambda\xi_i, \quad i = 1,\dots,n$$

avec ξ_i valant $\operatorname{sign} x_i$ ou un réel dans $[-1,1]$ suivant les valeurs des coordonnées de x_* :

- si $y_i > \lambda$, alors $x_{*i} = y_i - \lambda\xi_i$: $x_{*i} \geq y_i - \lambda$ est positif non nul, donc $\xi_i = 1$ et $x_{*i} = y_i - \lambda$,
- si $y_i < -\lambda$, par un raisonnement analogue on obtient $x_{*i} = y_i + \lambda$,
- si $y_i \in [-\lambda,\lambda]$, alors si $x_{*i} > 0$, on a $\xi_i = 1$ et $x_{*i} = y_i - \lambda \leq 0$, ce qui est contradictoire : l'hypothèse $x_{*i} > 0$ est à rejeter. De même, l'hypothèse $x_{*i} < 0$ est à rejeter et on a $x_{*i} = 0$, avec $x_{*i} - y_i + \lambda\xi_i = 0$, soit $\xi_i = y_i/\lambda$, qu'on vérifie bien dans $[-1,1]$ vu l'hypothèse $y_i \in [-\lambda,\lambda]$.

Pour le second programme, comme précédemment, on peut se limiter à étudier le cas y à coordonnées toutes positives. Si $\|y\|_1 \leq t$, alors $x_*(y,t) = y$ est le point de minimum (en lequel la fonction de décision est critique). Si $\|y\|_1 > t$, alors $\|x_*(y,t)\|_1 = t$ et KKT assure l'existence de $w = (w_i) \in \partial\| \ \|_1(x_*(y,t))$ et de $\lambda(y,t) \geq 0$ tels que

$$2(x_*(y,t)-y)_i = \begin{cases} -\lambda(y,t)w_i & \text{si } y_i = 0, \\ -\lambda(y,t) & \text{sinon} \end{cases} \tag{58}$$

avec

$$\begin{aligned} \|x_*(y,t)\|_1 &= \sum_i |x_*(y,t)_i| = \sum_{i:y_i\neq 0} (y_i - \lambda(y,t)/2) + \sum_{i:y_i=0} (y_i - \lambda(y,t)w_i/2) \\ &= \|y\|_1 - \frac{\lambda(y,t)}{2}\left[n - \#S_{x_*(y,t)} + \sum_{i:y_i=0} w_i\right]. \end{aligned}$$

Vu l'unicité des points de minimum et l'équation (58), $x_*(y,t) = x_*(y,\lambda(y,t))$ et donc

$$x_*(y,t) = x_*(y,\lambda(y,t)) = (\operatorname{sign}(y_i)(y_i - \lambda(y,t)/2)_+).$$

Ainsi le $x_*(y,t)$ a possiblement des termes nuls, mais les équations KKT (58) ne permettent pas de préciser le $\lambda(y,t)$ en fonction de y et t : on ne peut expliciter $x_*(y,t)$

comme on l'a fait aisément pour $x_*(y,\lambda)$, il faut développer des méthodes itératives particulières ! □

Rappelons l'analogue du lemme 3.11 pour la norme euclidienne

LEMME 3.12: *Soit $\lambda, t > 0$ et $y \in \mathbb{R}^n$.*

$$\operatorname{argmin}_x \left[\|x-y\|_2^2 + \lambda\|x\|_2^2\right] = y/(1+\lambda),$$

$$\operatorname{argmin}_{\|x\|_2\le t} \|x-y\|_2 = \begin{cases} y & \text{si } \|y\|_2 \le t, \\ t\frac{y}{\|y\|_2} & \text{sinon.} \end{cases}$$

DÉMONSTRATION. Pour le problème sans contrainte, mais avec une pénalité, le gradient $\nabla_x(\|x-y\|_2^2 + \lambda\|x\|_2^2) = 2(x-y) + 2\lambda x$ s'annule si et seulement si $x - y + \lambda x = 0$, soit $x = y/(1+\lambda)$.

Le point y est le point de minimum (global) de $\|x-y\|^2$: si $\|y\|_2 \le t$, c'est aussi celui sur la boule $\{\|x\|_2 \le t\}$. Sinon, celui-ci est sur la sphère $\|x\|_2 = t$ et les conditions de KKT (différentiables) assurent de l'existence de $\lambda \ge 0$ tel que le lagrangien $\|x - y\|^2 - \lambda(t^2 - \|x\|^2)$ est critique en $(x_*, \lambda x_*)$, *i. e.* $2(x_* - y) + 2\lambda x_* x_* = 0$, soit $x_* = y/(1+\lambda_*)$ et $t = \|x_*\|_2 = \|y\|_2/(1+\lambda_*)$ et finalement $x_* = ty/\|y\|_2$ (ce qui est clair géométriquement). □

6. Optimisation avec sous-gradient

On commence par deux itérations convergentes usant de sous-gradients, analogues à des méthodes de descente sans néanmoins partager la propriété de descente stricte à chaque itération comme lorsque la fonction à optimiser est régulière.

THÉORÈME 3.15: *Soit $U : \mathbb{R}^n \to \mathbb{R}$ convexe minorée, $(a_k)_{k\in\mathbb{N}}$ une suite convergente de réels positifs convergente vers 0 et dont la série $\sum_{k\in\mathbb{N}} a_k$ est divergente et la suite $(x_k)_{k\in\mathbb{N}}$ vérifiant la relation de récurrence*

$$x_{k+1} = x_k - a_k\xi_k/\|\xi_k\|, \quad k \ge 1$$

où ξ_k est un sous-gradient non nul de U en x_k. La suite décroissante $U_k = \min_{j=1}^k U(x_j)$ converge vers la valeur minimum $U_* = \min_{x\in\mathbb{R}^n} U(x)$.

Si la série $\sum_{k\ge0} a_k^2$ converge, alors la suite $(x_k)_{k\ge0}$ converge vers un point de minimum de U.

△ REMARQUE 3.10: Pour le choix de ξ_k, il y a besoin de savoir calculer un point du sous-différentiel $\partial_x U(x_k)$, ce qui est souvent plus aisé que de décrire et caractériser l'ensemble des sous-gradients (*i. e.* le sous-différentiel). ▽

DÉMONSTRATION. Raisonnons par l'absurde : il existe $\underline{U} > U_*$ tel que $U_k \ge \underline{U}$ pour tout $k \in \mathbb{N}$. Soit $\underline{x}$ tel que $U(\underline{x}) < \underline{U}$ et $\rho > 0$ tel que $U(x) < \underline{U}$ sur la boule $B(\underline{x}, \rho)$ dont l'existence est assurée par la continuité de U. Alors,

$$\underline{U} \ge U(\underline{x} + \rho\xi_k/\|\xi_k\|) \ge U(x_k) + \langle \xi_k, \underline{x} + \rho\xi_k/\|\xi_k\| - x_k\rangle \ge \underline{U} + \langle \xi_k, \underline{x} - x_k\rangle + \rho\|\xi_k\|$$

d'où résulte l'inégalité $\rho \le \langle \xi_k, x_k - \underline{x}\rangle/\|\xi_k\|$. Par ailleurs

$$\|x_{k+1} - \underline{x}\|^2 = \|x_k - \underline{x}\|^2 - 2a_k\langle \xi_k/\|\xi_k\|, x_k - \underline{x}\rangle + a_k^2 \le \|x_k - \underline{x}\|^2 - 2a_k\rho + a_k^2 \tag{59}$$

Ainsi, vu que $a_k \to 0$, il existe $K > 0$ tel que $a_k \le \rho$ pour $k \ge K$, d'où l'inégalité

$$\|x_{k+1} - \underline{x}\|^2 \le \|x_k - \underline{x}\|^2 - a_k\rho, \quad k \ge K.$$

Sommant les inégalités précédentes de K à $K+n$, on obtient

$$\|x_{K+n+1} - \underline{x}\|^2 \leq \|x_K - \underline{x}\|^2 - \rho \sum_{k=K}^{K+n} a_k$$

ce qui implique que les sommes partielles $\sum_{k=K}^{K+n} a_k$ sont bornées, contredisant la divergence de la série $\sum_{k\geq 0} a_k$. On a donc établi que $U_k \to U_*$.

Considérons l'inégalité de sous-gradient $\langle \xi_k, x_* - x_k\rangle \leq U(x_k) - U(x_*)$ dans un développement similaire à celui de (59) pour un point de minimum x_*

$$\begin{aligned}\|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - 2a_k\langle \xi_k/\|\xi_k\|, x_k - x_*\rangle + a_k^2 \\ &\leq \|x_k - x_*\|^2 + 2a_k(U(x_*) - U(x_k))/\|\xi_k\| + a_k^2 \leq \|x_k - x_*\|^2 + a_k^2\end{aligned}$$

qui induit l'inégalité

$$\|x_{K+1} - x_*\|^2 \leq \|x_1 - x_*\|^2 + \sum_{k=1}^{K} a_k^2 \tag{60}$$

et le fait que la suite $(x_k)_{k\geq 0}$ est bornée. Soit x_{**} un point d'adhérence de (x_k) qui est nécessairement un point de minimum de U vu que $\min_{n\leq k} U(x_n)$ converge vers U_{x_*}. La convergence de toute la suite $(x_k)_{k\geq 0}$ vers x_{**} (qui vérifie bien sûr l'inégalité (**??**)) résulte du lemme suivant. □

LEMME 3.13: *Soit $(u_k)_{k\geq 0}$, $(\delta_k)_{k\geq 0}$ des suites de réels positifs telles que $\sum \delta_k < +\infty$ et*

$$u_{k+1} \leq u_k + \delta_k, \quad k \geq 0 \tag{61}$$

Si la suite $(u_k)_{k\geq 0}$ a 0 comme valeur d'adhérence, alors la suite $(u_k)_{k\geq 0}$ converge vers 0.

DÉMONSTRATION. Soit $\varepsilon > 0$. Il existe N tel que $u_N \leq \varepsilon/2$ et $\sum_{k\geq N} \delta_k < \varepsilon/2$. En sommant les inégalités (61), on obtient

$$u_{N+K+1} \leq u_N + \sum_{k=N}^{N+K} \delta_k \leq \varepsilon, \quad K \geq 0.$$

La convergence $u_k \to 0$ en résulte. □

△ REMARQUE 3.11: La direction $-\xi_k$ n'est pas nécessairement une direction de descente (stricte) en x_k : par exemple, si $U(x,y) = |x| + |y|$, le sous-gradient $\xi_\pm = (1, \pm 1)$ en $m = (1,0)$ laisse la ligne de niveau $U = 1$ invariante. En général, la suite $(U(x_k))_{k\geq 0}$ n'est pas nécessairement décroissante, néanmoins la suite $(\min_{k=1}^n U(x_k))$ l'est, avec convergence vers la valeur minimum.

La convergence de x_n ne peut être rapide, pas plus rapide que celle du pas a_k, qui est lente vu la divergence de la série $\sum a_k$. ▽

THÉORÈME 3.16: *Soit $U : \mathbb{R}^n \to \mathbb{R}$ convexe avec un point de minimum x_* et valeur minimum U_*. Alors la suite $(x_k)_{k\geq 0}$ vérifiant la récurrence*

$$x_{k+1} = x_k - (U(x_k) - U_*)\xi_k/\|\xi_k\|^2, \quad k \geq 1$$

où ξ_k est un sous-gradient non nul de U en x_k, converge vers le point de minimum x_*, avec

$$\liminf_{k\to\infty} \sqrt{k}(U(x_k) - U_*) = 0$$

Si le point de minimum x_* est anguleux au sens où il existe une constante C telle que $U(x) - U_* \geq C\|x - x_*\|$, alors la convergence de U_k est de type linéaire.

DÉMONSTRATION. Soit x_* un point de minimum. Alors

$$\|x_{k+1}-x_*\|^2 = \|x_k-x_*\|^2 - 2(U(x_k)-U(x_*))\langle \xi_k, x_k-x_*\rangle/\|\xi_k\|^2 + (U(x_k)-U(x_*))^2/\|\xi_k\|^2$$

$$\leq \|x_k-x_*\|^2 - (U(x_k)-U(x_*))^2/\|\xi_k\|^2 = \|x_k-x_*\|^2 - (U(x_k)-U_*)^2/\|\xi_k\|^2 \tag{62}$$

où l'inégalité provient de celle caractérisant le gradient $\xi_k \in \partial U(x_k)$. Ainsi

$$\|x_{k+1}-x_*\|^2 + (U(x_k)-U_*)^2/\|\xi_k\|^2 \leq \|x_k-x_*\|^2 \tag{63}$$

la suite $(\|x_k-x_*\|^2)_{k\geq 0}$ est décroissante convergente (car minorée) et la suite $((U(x_k)-U_*)/\|\xi_k\|)_{k\geq 0}$ tend vers 0. La suite $(x_k)_{k\geq 0}$ est bornée, car la suite $(\|x_k-x_*\|)_{k\geq 0}$ l'est : pour un R_* majorant tous les $\|x_k-x_*\|$, alors $\|x_k\| \leq R$ avec $R = R_* + \|x_*\|$. Par suite t l'ensemble des gradients $\cup_{\|x\|\leq R}\|\partial U(x)\|$ (reprendre la preuve du théorème 3.9) est borné : $\sup\{(\|\xi\| \,|\xi \in \partial_x U, \|x\| \leq R\} = R_1 < +\infty$. Alors, $U(x_k)-U_*$ tend vers 0. On peut extraire une sous-suite $(x_{k_n})_{n\geq 0}$ convergente vers x_{**}, qui est donc aussi un minimum de U, vu que $U(x_k) \to U_*$. Considérant l'inégalité (63) avec x_{**} remplaçant x_*, on obtient que la suite (décroissante) $(\|x_k-x_{**}\|)_{k\geq 0}$ toute entière converge vers 0 (comme le fait la sous-suite $(\|x_{k_n}-x_{**}\|)_{n\geq 0}$), ce qui établit la convergence de $(x_k)_{k\geq 0}$ vers un point de minimum de U.

Reprenant l'inégalité (63), on a la convergence de la série $\sum_{k\geq 0}(U(x_k)-U_*)^2/\|\xi_k\|^2$, soit celle de $\sum_{k\geq 0}(U(x_k)-U_*)^2$ vu que l'ensemble des sous-gradients $\{\xi_k\}$ est borné par R_1. Si on suppose $\liminf \sqrt{k}(U(x_k)-U_*) > 0$, on a pour k suffisamment grand $U(x_k)-U_* > C/\sqrt{k}$, ce qui contredit la convergence de la série $\sum_{k\geq 0}(U(x_k)-U_*)^2$. Ainsi $\liminf \sqrt{k}(U(x_k)-U_*) = 0$.

Si x_{**} est un minimum anguleux, *i. e.* $U(x)-U_* \geq C_1\|x-x_{**}\|$ pour une constante C_1, alors l'inégalité (63) donne

$$\|x_{k+1}-x_*\|^2 \leq \|x_k-x_*\|^2 - C_1\|x_k-x_*\|^2/R_1 \leq (1-C_1/R_1)\|x_k-x_*\|^2$$

où R_1 est un majorant des sous-gradients ξ et donc la convergence de type linéaire avec $\theta = \sqrt{1-C_1/R_1}$. □

△ REMARQUE 3.12: Les programmes $\min_x[\lambda\|x\|_1 + \|Ax-y\|_2]^2$ ou $\min_x[\lambda\|x\|_\infty + \|Ax-y\|_2^2]$ n'ont pas de solution explicite... ▽

7. Fonctions quasi-convexes

La notion de fonction quasi-convexe généralise celle de fonction convexe.

DÉFINITION 3.9: *Une fonction U est dite* quasi-convexe *si pour tout A le sous-domaine de niveau $S_A = \{x \in \mathbf{dom}\, U, U(x) \leq A\}$ est convexe. La fonction U est* quasi-concave *si $-U$ est quasi-convexe. La fonction U est* quasi-linéaire *et U et $-U$ sont quasi-convexes.*

△ REMARQUES 3.13:

(1) L'inégalité (3.1) de convexité implique la quasi-convexité d'une fonction convexe : si $U(x), U(y) \leq A$, alors

$$U(\theta x + (1-\theta)y) \leq \theta U(x) + (1-\theta)U(y) \leq \theta A + (1-\theta)A \leq A$$

La réciproque est fausse : toute fonction monotone (par exemple $t \in \mathbb{R} \mapsto t^3$) est quasi-convexe.

(2) Comme la convexité, la quasi-convexité est caractérisée par le comportement sur les droites. ▽

▷ EXEMPLES 3.13:

3.13.1 Une fonction monotone est quasi-linéaire.

3.13.2 La fonction de Cobb-Douglas $U_{CD}^{\alpha\beta} : (K,L) \in \mathbb{R}_+^2 \mapsto K^\alpha L^\beta$ est quasi-concave si $\alpha, \beta \in [0,1]$, elle est concave si et seulement si $\alpha, \beta, \alpha+\beta \in [0,1]$. En effet, si $K^\alpha L^\beta \geq A, \widetilde{K}^\alpha \widetilde{L}^\beta \geq A$, vu la concavité de $x \mapsto x^\alpha$ et l'inégalité $x + x^{-1} \geq 2$,

$$\begin{aligned}(\lambda K + (1-\lambda)\widetilde{K})^\alpha (\lambda L + (1-\lambda)\widetilde{L})^\beta &\geq (\lambda K^\alpha + (1-\lambda)\widetilde{K}^\alpha)(\lambda L^\beta + (1-\lambda)\widetilde{L}^\beta) \\ &\geq \lambda^2 A + \lambda(1-\lambda)[K^\alpha \widetilde{L}^\beta + \widetilde{K}^\alpha L^\beta] + (1-\lambda)^2 A \\ &\geq A[\lambda^2 + \lambda(1-\lambda)[L^{-\beta}\widetilde{L}^\beta + \widetilde{L}^{-\beta} L^\beta] + (1-\lambda)^2] \\ &\geq A[\lambda^2 + 2\lambda(1-\lambda) + (1-\lambda)^2] \geq A.\end{aligned}$$

Par ailleurs, la Hessienne de $U_{CD}^{\alpha\beta}$ est

$$\operatorname{Hess} U_{CD}^{\alpha\beta}(K,L) = \begin{pmatrix} \alpha(\alpha-1)K^{\alpha-2}L^\beta & \alpha\beta K^{\alpha-1}L^{\beta-1} \\ \alpha\beta K^{\alpha-1}L^{\beta-1} & \beta(\beta-1)K^\alpha L^{\beta-2} \end{pmatrix},$$

avec déterminant

$$\text{dét}\operatorname{Hess} U_{CD}^{\alpha\beta}(K,L) = \alpha\beta(1-\alpha-\beta)K^{2\alpha-2}L^{2\beta-2}.$$

Ainsi la fonction $U_{CD}^{\alpha\beta}$ est concave si et seulement si $\alpha, \beta \in [0,1], \alpha+\beta \leq 1$.

3.13.3 La fonction $x \in E \mapsto (\langle a,x\rangle + b)/(\langle c,x\rangle + d)$ est quasi-convexe et quasi-concave sur le domaine $\{\langle c,x\rangle + d > 0\}$ vu que

$$\begin{aligned}S_A &= \{\langle c,x\rangle + d > 0, (\langle a,x\rangle + b)/(\langle c,x\rangle + d) \leq A\} \\ &= \{\langle c,x\rangle + d > 0\} \cap \{\langle a,x\rangle + b - A(\langle c,x\rangle + d) \leq 0\}.\end{aligned}$$

3.13.4 La fonction $\operatorname{ceil} x = \inf\{z \in \mathbb{Z}, z \geq x\}$ est quasi-linéaire, de même que la fonction $\log x$ sur $\mathbb{R}_+$.

3.13.5 Soit, pour U définie sur $\mathbb{R}$, la fonction $\widehat{U}$ définie par $\widehat{U}(t) = U(-t)$. Alors U est quasi-convexe si et seulement si $\widehat{U}$ l'est.

3.13.6 Soit $x = (x_0, x_1, \ldots, x_n) \in \mathbb{R}^{n+1}$ représentant un flux de trésorerie sur une période de n unités temporelles (mois, années,...), avec des opérations de débit $(x_i < 0)$ ou de crédit $(x_i > 0)$. On définit la valeur courante VC de la trésorerie avec taux d'intérêt $r \geq 0$ par

$$VC(x,r) = \sum_{i=0}^{n} (1+r)^{-i} x_i$$

Soit $T \subset \mathbb{R}^{n+1}$ l'ensemble des flux de trésoreries x telles que $x_0 < 0$ (on démarre avec un investissement de $|x_0|$) et $x_0 + x_1 + \ldots + x_n > 0$ (la trésorerie restante $x_1 + \ldots + x_n$ excède l'investissement initial). On a $VC(x,0) > 0$ et $VC(x,+\infty) = x_0 < 0$: on définit le taux de rentabilité interne T_{ri} sur T suivant

$$T_{ri}(x) = \inf\{r \geq 0, VC(x,r) = 0\}, \quad x \in T.$$

La fonction T_{ri} est quasi-concave vu que

$$\{T_{ri}(x) \geq \rho\} = \{VC(x,r) > 0, 0 \leq r < \rho\} = \cap_{0 \leq r < \rho}\{VC(x,r) > 0\}$$

est convexe comme intersection de demi-espaces.

3.13.7 Si une fonction d'utilité est quasi-concave, l'ensemble des préférences $\{U \geq A\}$ d'utilité au moins A est convexe.

3.13.8 Les fonctions $U_1(x) = x$ et $U_2(x) = \inf(x^2, 1)$ sont quasi-convexes, alors que leur somme de l'est pas.

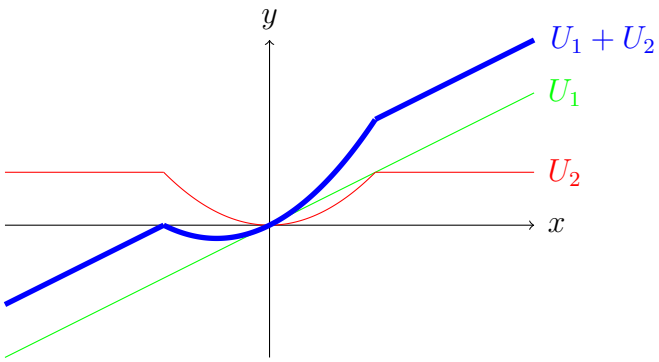


FIGURE III.5 . Les fonctions U_1, U_2 sont quasi-convexes, mais pas $U_1 + U_2$.

3.13.9 Pour $\alpha_i > 0$ et U_i quasi-convexes, le maximum pondéré $\sup_i(\alpha_i U_i)$ est quasi-convexe.

3.13.10 Si h est croissante et U quasi convexe, $h \circ U$ est quasi-convexe.

3.13.11 Pour A linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^m$, $b \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ et $c \in \mathbb{R}$, l'application $x \mapsto U((Ax+b)/(\langle v, x\rangle + c)$ est quasi-convexe sur $\langle v, x\rangle + c > 0$ si U l'est, avec le cas particulier $v = 0$.

3.13.12 Si $U(x, y)$ est quasi-convexe et C convexe, alors $\inf_{y\in C} U(x, y)$ est quasi-convexe.

3.13.13 L'application $x \in \{\|x - a\|_2 \leq \|x - b\|_2\} \to \|x - a\|_2/\|x - b\|_2$ est quasi-convexe.

3.13.14 $x \in \mathbb{R} \mapsto \mathrm{e}^{-x^2}$ est quasi-concave, mais non concave. ◁

PROPOSITION 3.5: *Une fonction U est quasi-convexe sur C si et seulement si*

$$U(x + \lambda(y - x)) \leq \max\left[U(x), U(y)\right], \quad x, y \in C, \lambda \in [0, 1].$$

DÉMONSTRATION. Si U est quasi-convexe, les points x et y, pour lesquels on supposera $U(x) \leq U(y)$ à l'échange de x, y près, sont dans $U^{-1}((-\infty, U(y)])$ qui, convexe, contient le segment $[x, y]$: ainsi, $U(x + \lambda(y - x)) \leq U(y) = \max(U(x), U(y))$ pour $\lambda \in [0, 1]$).

Pour la réciproque, soit $A = \max(U(x,), U(y))$ et $C_A = f^{-1}((-\infty, A])$: vu que $U(x + \lambda(y - x)) \leq \max(U(x), U(y)) = A$, le segment $[x, y]$ est inclus dans C_A, qui est donc convexe. □

▷ EXEMPLE 3.14: Notons $\mathrm{card}(x)$ le nombre de composantes non nulles de x : la fonction card est quasi-concave sur $\mathbb{R}^n_+$ (mais pas sur $\mathbb{R}^n$!) vu

$$\mathrm{card}(x + y) \geq \min(\mathrm{card}(x), \mathrm{card}(y)), \quad x, y \in \mathbb{R}^n_+$$ ◁

Comme pour la convexité, la quasi-convexité de U est caractérisée par le comportement de U en restriction aux intersections de droites avec le domaine de U.

PROPOSITION 3.6: *Soit U continue sur $\mathbb{R}$. La fonction U est quasi-convexe si et seulement si U est monotone ou si U est unimodale avec un minimum (*i. e. *il existe $x_* \in \mathbf{dom}\, U$ tel que U est décroissante sur $(-\infty, x_*] \cap \mathbf{dom}\, U$ et croissante sur $[x_*, \infty) \cap \mathbf{dom}\, U$).*

DÉMONSTRATION. Si U est croissante, $U^{-1}((-\infty, A]) = (-\infty, x_A]$ avec $x_A = \sup\{x, U(x) \leq A\}$: U est donc quasi-convexe. Si U est décroissante, alors $\widehat{U}$ définie par $\widehat{U}(x) = U(-x)$ est croissante, donc quasi-convexe : il en est de même pour U.

Soit U avec valeur minimum m prise sur l'intervalle maximal $[x_-, x_+]$, décroissante sur $(-\infty, x_-]$ et croissante sur $[x_+, \infty)$: $U^{-1}(-\infty, A])$ est vide si $A \leq m$, sinon égal à l'intervalle $[x_-(A), x_+(A)]$ avec $x_-(A) = \inf\{x \leq x_-, U(x) \leq A\}$ et $x_+(A) = \sup\{x \geq x_+, U(x) \leq A\}$. Ainsi, U est quasi-convexe.

Si U n'est pas d'un des trois types précédents, quitte à considérer $\widehat{U}$, il existe $x_- < x_0 < x_1 <$ tels que $U(x_-) < U(x_0)$ et $U(x_0) > U(x_1)$: si $A \in (U(x_-), U(x_0) \cap (U(x_1), U(x_0)$, alors $U^{-1}((-\infty, A))$ n'est pas connexe (et par suite non convexe) : d'après le théorème des valeurs intermédiaires appliqué à la fonction U continue, il contient un point dans chaque intervalle (x_-, x_0) et (x_1, x_0), mais pas x_1. □

Pour les fonctions différentiables, la quasi-convexité est exprimable en termes analogues à la convexité

PROPOSITION 3.7: *(i) Soit U différentiable définie sur un ouvert convexe C de E. La fonction U est quasi-convexe si et seulement si*

$$U(y) \leq U(x) \Longrightarrow \langle \nabla U(x), y - x \rangle \leq 0, \quad x, y \in C.$$

(ii) Soit U deux fois différentiable définie sur un ouvert convexe C de E avec $\nabla U(x) \neq 0$ pour $x \in C$. Si la fonction U est quasi-convexe, alors la hessienne $\operatorname{Hess} U(x)$ *est positive sur l'hyperplan orthogonal à $\nabla U(x)$ pour tout $x \in C$; réciproquement si la hessienne* $\operatorname{Hess} U(x)$ *est strictement positive sur l'hyperplan orthogonal à $\nabla U(x)$ pour tout $x \in C$, alors U est quasi-convexe.*

DÉMONSTRATION. Il suffit de montrer la propriété au premier ordre pour des fonctions d'une variable. Supposons U quasi-convexe et soient x, y avec $U(y) \leq U(x)$. On a donc $U(x + t(y - x)) \leq U(x)$ pour $t \in (0, 1)$, d'où, en dérivant en $t = 0$ l'inégalité $\langle \nabla U(x), y - x \rangle \leq 0$. Réciproquement, supposons x, y avec $U(y) \leq U(x)$, $\pm(x - y) \geq 0$ et $\langle \nabla U(x), y - x \rangle > 0$. Il existe alors $\varepsilon > 0$ tel que $U(x \pm \varepsilon) > U(x \pm \varepsilon/2) > U(x)$: la partie $U^{-1}((-\infty, U(x \pm \varepsilon/2])$ contient x et y, sans contenir $x + \varepsilon$; ainsi U n'est pas quasi-convexe.

Pour (ii), remarquons que si $\dim E = 1$, la condition de non annulation de ∇U implique que U est monotone : U est donc quasi convexe. On suppose dans la suite $\dim E \geq 2$.

Supposons tout d'abord U quasi-convexe. D'après le (i), si $\langle \nabla U(x), z \rangle > 0$, alors $U(x) < U(x + z)$: ainsi par continuité $U(x) \leq U(x + z)$ pour tout z orthogonal à $\nabla U(x)$: x est un minimum de $y \mapsto U(y)$ sous la condition $\langle \nabla U(x), y - x \rangle = 0$; il en résulte la positivité de la restriction de $\operatorname{Hess} U(x)$ à l'orthogonal de $\nabla U(x)$ (égale à la hessienne de la restriction de U à l'orthogonal de $\nabla U(x)$).

Réciproquement supposons la hessienne de U strictement positive sur l'orthogonal de $\nabla U(x)$ passant par x, soit l'hyperplan $H_x = \{\langle \nabla U(x), Y - x \rangle = 0\}$. Alors le point x est minimum local de la restriction de U à $H_x \cap C$ d'après le théorème 2.11.(ii). Montrons que x est un minimum local de U dans le demi-espace $E_x = \{\langle \nabla U(x), y - x \rangle \geq 0\}$ (*i. e.* minimum global sur $E_x \cap B(x, \rho)$ avec ρ assez petit) : soit en effet $y \in (E_x \setminus H_x) \cap C$, $y(\theta) = \theta y + (1 - \theta)x$ et $F_y(\theta) = U(y(\theta))$. Soit θ_0 le sup des θ tels que $F_y(\theta)$ soit le maximum de F_y sur $[0, 1]$. Vu que $F_y'(0) = \langle \nabla U(x), y - x \rangle > 0$, $U(y(\theta)) > U(x)$ pour $\theta > 0$ au voisinage de $\theta = 0$: ainsi $\theta_0 > 0$. Si $\theta_0 < 1$, on a $0 = F_y'(\theta_0) = \langle \nabla U(y(\theta_0)), y - x \rangle = 0$ et donc $\langle \nabla U(y(\theta_0)), y(\theta_0 + h) - y(\theta_0) \rangle = 0$. Par hypothèse, $y(\theta_0)$ est un minimum strict de la restriction de U à l'hyperplan $H_{y(\theta_0)}$ qui contient $y(\theta_0 + h)$: ainsi pour $h > 0$ suffisamment petit, $U(y(\theta_0)) < U(y(\theta_0 + h))$, ce qui contredit le caractère maximal de θ_0

dans $[0,1]$. On a donc $\theta_0 = 1$, et par suite $U(y) = F_y(1) \geq U(x)$ et x est un minimum local de U sur E_x.

Soit $x, y \in C$ et $z \in [x,y]$. Le point z minimise localement U sous la contrainte $\langle \nabla U(z), y - z\rangle \geq 0$, *i. e.* dans le demi-espace E_z. Si $z = \theta x + (1-\theta)y$,

$$\langle \nabla U(z), z\rangle = \theta\langle \nabla U(z), x\rangle + (1-\theta)\langle \nabla U(z), y\rangle,$$

ainsi a-t-on $\langle \nabla U(z), z\rangle \leq \max(\langle \nabla U(z), x\rangle, \langle \nabla U(z), y\rangle)$: soit [13] $\langle \nabla U(z), z\rangle \leq \langle \nabla U(z), y\rangle$, c'est à dire $y \in E_z$ et alors $U(z) \leq U(y)$, soit $\langle \nabla U(z), z\rangle \leq \langle \nabla U(z), x\rangle$ et alors $U(z) \leq U(x)$: dans tous les cas, on a $U(z) \leq \max(U(x), U(y))$, ce qui démontre la quasi-convexité de U. □

△ Remarque 3.14: Ainsi, si $E = \mathbb{R}^n$, d'après le théorème A.4 de l'appendice, la condition du deuxième ordre de quasi-convexité pour U se lit sur les $n-1$ mineurs principaux dominants d'ordre $3, \ldots, n, n+1$ de la matrice hessienne $\operatorname{Hess} U$ bordée par le gradient ∇U

$$\begin{pmatrix} 0 & {}^{\mathsf{T}}\nabla U(x) \\ \nabla U(x) & \operatorname{Hess} U(x) \end{pmatrix}$$

où il a été supposé la première coordonnée de $\nabla U(x)$ non nulle. ▽

▷ Exemple 3.15: La fonction de Cobb-Douglas $U^{\alpha,\beta}_{CD}(x,y) = x^\alpha y^\beta$ a pour hessienne bordée par $\nabla U^{\alpha,\beta}_{CD}$

$$\begin{pmatrix} \alpha(\alpha-1)x^{\alpha-2}y^\beta & \alpha\beta x^{\alpha-1}y^{\beta-1} & \alpha x^{\alpha-1}y^\beta \\ \alpha\beta x^{\alpha-1}y^{\beta-1} & \beta(\beta-1)x^\alpha y^{\beta-2} & \beta x^\alpha y^{\beta-1} \\ \alpha x^{\alpha-1}y^\beta & \beta x^\alpha y^{\beta-1} & 0 \end{pmatrix}$$

de déterminant

$$d_3(x,y) = \alpha\beta(\alpha+\beta)(xy)^{-2}(x^\alpha y^\beta)^3.$$

elle est donc quasi-concave pour $\alpha, \beta > 0$. Pour la fonction de Cobb-Douglas $U^{\alpha,\beta,\gamma}_{CD}(x,y,z) = x^\alpha y^\beta z^\gamma$, sa hessienne bordée par $\nabla U^{\alpha,\beta,\gamma}_{CD}$ a pour déterminant $d_4(x,y,z) = -\alpha\beta\gamma(\alpha+\beta+\gamma)(xyz)^{-2}(U^{\alpha,\beta,\gamma}_{CD})^4$ et son mineur principal dominant d'ordre 3 égal à $d_3(x,y,z) = d_3(x,y)z^{3\gamma}$. ◁

△ Remarque 3.15: Malgré cette similarité des caractérisations de la convexité et quasi-convexité à l'ordre 1, il y a des différences importantes : un minimum local d'une fonction convexe en est un global, ce n'est pas nécessairement le cas pour une fonction quasi-convexe : par ex., la fonction sur $\mathbb{R}$ constante sur l'intervalle $(x_0-\alpha, x_0+\alpha)$ et strictement croissante en dehors admet x_0 comme minimum local non global. ▽

Proposition 3.8: *Soit $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n_+$ et U^α_{CD} la fonction de Cobb-Douglas définie par $U^\alpha_{CD}(x) = \prod_{j=1}^n x_j^{\alpha_j}$. La fonction U^α_{CD} est quasi-concave, concave si $\alpha_1 + \cdots + \alpha_n \leq 1$.*

Démonstration. Vu que la fonction log est croissante, U^α_{CD} et $\log U^\alpha_{CD}$ sont simultanément quasi-concaves : $\log U^\alpha_{CD} = \sum_{j=1}^n \alpha_j \log x_j$ est clairement concave, donc quasi-concave.

Démontrons le lemme suivant

Lemme 3.14: *Si U est quasi-concave sur $\mathbb{R}^n_+$ à valeurs positives et homogène de degré $d \in [0,1]$, alors U est concave.*

13. Géométriquement, $\nabla U(z)$ fait un angle aigu avec $x - z$ ou avec $y - z$.

Preuve du Lemme. Posons $C_{x,y} = (U(y)/U(x))^{1/d}$. Par d-homogénéité et la concavité de $t \in \mathbb{R}_+ \mapsto t^d$ puisque $d \leq 1$, on a,

$$\alpha U(x) + (1-\alpha)U(y) = \alpha U(x) + (1-\alpha)U(C_{x,y}x) \leq U((\alpha + (1-\alpha)C_{x,y})x)$$

et de manière analogue

$$\alpha U(x) + (1-\alpha)U(y) \leq U((1-\alpha+\alpha C_{y,x})y)$$

Ainsi

$$\begin{aligned}\alpha U(x) + (1-\alpha)U(y) &\leq \inf\{U((\alpha+(1-\alpha)C_{x,y})x), U((1-\alpha+\alpha C_{y,x})y)\}\\ &\leq U(\theta(\alpha+(1-\alpha)C_{x,y})x + (1-\theta)(1-\alpha+\alpha C_{y,x})y)\\ &= U(\alpha x + (1-\alpha)y)\end{aligned}$$

où on a utilisé la quasi-concavité de U, puis introduit θ tel que $\theta(\alpha+(1-\alpha)C_{x,y}) = \alpha$ qui vérifie aussi, vu $C_{x,y} = C_{y,x}^{-1}$,

$$\begin{aligned}(1-\theta)(1-\alpha+\alpha C_{y,x}) &= \left(1 - \frac{\alpha}{\alpha+(1-\alpha)C_{x,y}}\right)(1-\alpha+\alpha C_{y,x})\\ &= \frac{(1-\alpha)C_{x,y}}{\alpha+(1-\alpha)C_{x,y}}(1-\alpha+\alpha C_{y,x}) = 1-\alpha.\end{aligned}$$ □

Soit $|\alpha| = \sum_{j=1}^n \alpha_j$. La fonction U_{CD}^α est homogène de degré $|\alpha|$: vu que son log est concave (donc quasi-concave) et que toute fonction croissante d'une fonction quasi-concave est quasi-concave, elle est donc quasi-concave. Homogène de degré $|\alpha| \leq 1$, elle est concave d'après le lemme précédent. □

△ Remarque 3.16: La concavité de U_{CD}^α avec les $\alpha_i \geq 0$ et la somme $|\alpha| = 1$ peut aussi être prouvée en utilisant la transformation conique : si $U_0 : C \to \mathbb{R}$ est convexe (concave resp.) sur le cône convexe C de sommet 0, alors la fonction U dite *transformée conique de* U_0 et définie suivant $U : (x,t) \in C \times \mathbb{R}^+ \mapsto tU_0(x/t)$ est convexe (concave resp.). Il suffit d'écrire [14]

$$U_0\left(\frac{\theta_1 x_1 + \theta_2 x_2}{\theta_1 t_1 + \theta_2 t_2}\right) = U_0\left(\frac{\theta_1 t_1 \frac{x_1}{t_1} + \theta_2 t_2 \frac{x_2}{t_2}}{\theta_1 t_1 + \theta_2 t_2}\right) \leq \frac{\theta_1 t_1}{\theta_1 t_1 + \theta_2 t_2} U_0\left(\frac{x_1}{t_1}\right) + \frac{\theta_2 t_2}{\theta_1 t_1 + \theta_2 t_2} U_0\left(\frac{x_2}{t_2}\right)$$

d'où résulte l'inégalité de convexité (concavité resp.) pour U.

La relation

$$x_n\left(\left(\frac{x_1}{x_n}\right)^{\alpha_1} \cdots \left(\frac{x_{n-1}}{x_n}\right)^{\alpha_{n-1}}\right) = x_1^{\alpha_1} \ldots x_{n-1}^{\alpha_{n-1}} x_n^{1-\alpha_1-\cdots-\alpha_{n-1}}$$

implique que la fonction $U_{CD}^{\alpha_1\ldots\alpha_n}$ de n variables où $\alpha_n = 1-\alpha_1-\cdots-\alpha_{n-1}$ soit $|\alpha| = 1$, est la transformée conique de la fonction $U_{CD}^{\alpha_1\ldots\alpha_{n-1}}$ de $n-1$ variables avec $\alpha_1+\cdots+\alpha_{n-1} \leq 1$.

Avant de montrer par récurrence que U_{CD}^α à n variables avec $|\alpha| \leq 1$ est concave, rappelons que si g est croissante concave et V concave, alors $g \circ V$ est concave, vu que pour $\theta_1, \theta_2 \in [0,1]$ avec $\theta_1 + \theta_2 = 1$,

$$\begin{aligned}\theta_1 g\circ V(x_1) + \theta_2 g\circ V(x_2) &\leq g(\theta_1 V(x_1) + \theta_2 V(x_2))\\ &\leq g(V(\theta_1 x_1 + \theta_2 x_2))) \leq g\circ V(\theta_1 x_1 + \theta_2 x_2)\end{aligned}$$

où on a utilisé la concavité de g, puis sa croissance et la concavité de V.

14. On peut aussi calculer sa hessienne si U_0 est de classe $\mathcal{C}^2$: $\operatorname{Hess} U(x,t)[X,T] = t^{-1}\operatorname{Hess} U_0(x/t)\left(X - \frac{T}{t}x\right)$ qui est positive si U_0 l'est.

L'hypothèse de récurrence est vérifiée en 1 variable : $U_{CD}^{\alpha_1} : x_1 \in \mathbb{R}^+ \mapsto x_1^{\alpha_1}$ est concave si $\alpha_1 \in [0,1]$. Supposons la propriété vraie pour les fonctions U_{CD}^{α} de $n-1$ variables. Soit $\alpha \in [0,1]^n$ avec $\sigma = |\alpha| \leq 1$. Alors, la fonction de Cobb-Douglas $U_{CD}^{\alpha/\sigma}$ de n variables pour laquelle $\sigma = |\alpha| \in [0,1]$ est concave d'après l'hypothèse de récurrence : la fonction $U_{CD}^{\alpha} = U_{CD}^{\sigma} \circ U_{CD}^{\alpha/\sigma}$ est concave vu $|\alpha/\sigma| = 1$ d'après l'hypothèse de récurrence et le rappel sur les propriétés de concavité de la composée $g \circ U$ où on aura pris $g = U_{CD}^{\sigma}$ et $U = U_{CD}^{\alpha/\sigma}$. ▽

La proposition précédente indique la concavité de la moyenne géométrique $M_g = U_{CD}^{(\frac{1}{n},\dots,\frac{1}{n})}$ de n variables : on peut l'établir suivant deux autres manières.

PROPOSITION 3.9: *Soit M_g la fonction* moyenne géométrique $x = (x_1, \dots, x_n) \in \mathbb{O}_{++}^n \mapsto (x_1 \dots x_n)^{1/n}$. *La fonction M_g est concave.*

DÉMONSTRATION. En posant $y_i = x_i^{-1}$, la hessienne de M_g est donnée par

$$[\mathrm{Hess}\, M_g](x) = \frac{M_g(x)}{n^2}\left[(1 - n\delta_{ij})y_i y_j\right]$$

soit

$$[\mathrm{Hess}\, M_g](x)[h] = \frac{M_g(x)}{n^2}\left[\langle h, y\rangle^2 - n\|y * h\|^2\right], \quad h \in \mathbb{R}^n$$

qui est négative d'après Cauchy-Schwarz, d'où la concavité de M_g.

Autre méthode : D'après l'inégalité des moyennes arithmétique et géométrique, on a pour $\xi \in \mathbb{O}_{++}^n$

$$\frac{1}{n}\langle \xi, x\rangle \geq \left(\prod_{i=1}^n \xi_i x_i\right)^{1/n} = \left(\prod_{i=1}^n \xi_i\right)^{1/n}\left(\prod_{i=1}^n x_i\right)^{1/n} = M_g(\xi) M_g(x)$$

ce qui implique

$$\langle \xi, x\rangle \geq \left(\prod_{i=1}^n x_i\right)^{1/n} = M_g(x), \tag{64}$$

si $M_g(\xi) \geq n^{-1}$, et par suite

$$\min_{\substack{\xi_i > 0, i=1,\dots,n \\ M_g(\xi) \geq n^{-1}}} \langle \xi, x\rangle = M_g(x), \quad x \in \mathbb{O}_{++}$$

vu que la moyenne $M_g(x)$ minore le membre de gauche en vertu de l'inégalité précédente, cette inégalité étant une égalité pour $\xi_* = (M_g(x)/(nx_i))$ qui vérifie $\prod_{i=1}^n \xi_{*i} = n^{-n}$. La fonction M_g, min de fonctions linéaires, est donc concave. □

8. Dualité et point selle

L'idée de base de la dualité lagrangienne est de tenir compte des contraintes en prolongeant la fonction d'objectifs U à un espace incluant les multiplicateurs lagrangiens, qui se transforment en *variables duales*, la fonction prolongée incluant fonction d'objectifs et fonctions de contrainte. Pour simplifier on suppose U, ainsi que les fonctions de contrainte, définies sur $\mathcal{D} = \mathbb{R}^n$.

Au problème dit *primal*

$$p^* = \inf_{x \in C} U(x) \tag{65}$$

avec $C = \{g_i(x) = 0, i = 1, \dots, m, h_j(x) \geq 0, j = 1, \dots, p\}$, est associé le lagrangien

$$\mathcal{L}(x, \Lambda, M) = U(x) - \langle \Lambda, g(x) \rangle - \langle M, h(x) \rangle, \quad x \in \mathbb{R}^n, \Lambda \in \mathbb{R}^m, M \in \mathbb{R}^p_+.$$

Remarquons

$$\sup_{\Lambda, M \geq 0} \mathcal{L}(x, \Lambda, M) = \begin{cases} U(x), & \text{si } g(x) = 0, h(x) \geq 0, \\ +\infty, & \text{sinon.} \end{cases}$$

Si $\widetilde{x}$ est un point réalisable du problème primal (65), *i. e.* $\widetilde{x} \in C$, alors, vu que les multiplicateurs $M = (\mu_j)_{j=1,\dots,p}$ sont à coordonnées positives, on a

$$\mathcal{L}(\widetilde{x}, \Lambda, M) \leq U(\widetilde{x})$$

et les minorations pour le problème primal (65)

$$V(\Lambda, M) \leq \inf_{x \in C} U(x), \quad (\Lambda, M) \in \mathbb{R}^m \times \mathbb{R}^n_+, \tag{66}$$

où la fonction V, dite *fonction duale du problème primal (65)*, est définie suivant

$$V(\Lambda, M) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \Lambda, M) \in \mathbb{R} \cup \{-\infty\} \tag{67}$$

avec $\mathbf{dom}\, V$ constitué des Λ, M avec l'inf dans (67) fini. D'après (66), la fonction duale V donne des minorants au minimum de la fonction de coût sur le domaine C des réalisables du programme primal : le meilleur de ces minorants est donné par le *programme dual* du programme primal (65), défini comme le problème d'optimisation

$$d^* = \sup_{\substack{\Lambda, M \geq 0 \\ (\Lambda, M) \in \mathbf{dom}\, V}} V(\Lambda, M) \tag{68}$$

Le problème dual est toujours convexe, que U soit convexe ou pas : la fonction V duale est concave comme inf de fonctions affines relativement aux *variables duales* Λ, M. Il est souvent plus simple que le problème primal. Par ailleurs, le programme dual est sans contraintes, autres que celles de positivité ($M \geq 0$) qui sont aisément traitables algorithmiquement (toute solution approchée M_k est projetée sur l'orthant positif $\mathbb{R}^p_+$) et les conditions implicites naturelles $(\Lambda, M) \in \mathbf{dom}\, V$ (l'exemple du programme linéaire n'est pas bon, car $\mathbf{dom}\, V$ est dans ce cas là restreint à un sous-espace linéaire, ce qui n'est pas le cas des programmes quadratiques ou non linéaires plus généralement !)

▷ Exemples 3.16:

3.16.1 Le programme linéaire $\inf_{Ax=b, x \geq 0} \langle c, x \rangle$ a pour fonction duale

$$V(\Lambda, M) = \inf_{x \in \mathbb{R}^n} [\langle c, x \rangle - \langle \Lambda, Ax - b \rangle - \langle M, x \rangle] = \langle \Lambda, b \rangle + \inf_{x \in \mathbb{R}^n} [\langle c - {}^{\mathsf{T}}A\Lambda - M, x \rangle].$$

Vu qu'une fonction linéaire est inférieurement bornée seulement si elle est nulle, on a

$$V(\Lambda, M) = \begin{cases} \langle \Lambda, b \rangle & \text{si } {}^{\mathsf{T}}A\Lambda + M = c, \\ -\infty, & \text{sinon} \end{cases}$$

soit le problème dual

$$\max_{{}^{\mathsf{T}}A\Lambda + M = c, M \geq 0} \langle \Lambda, b \rangle.$$

3.16.2 Le programme linéaire $\inf_{Ax \leq b} \langle c, x \rangle$ a pour fonction duale

$$V(M) = \inf_{x \in \mathbb{R}^n} [\langle c, x \rangle + \langle M, Ax - b \rangle] = \begin{cases} -\langle M, b \rangle & \text{si } c + {}^{\mathsf{T}}AM = 0 \\ -\infty & \text{sinon} \end{cases}$$

soit le problème dual

$$\max_{c+{}^{\mathsf{T}}AM=0, M\geq 0} -\langle M, b\rangle.$$

3.16.3 Le programme quadratique $\inf_{Ax=b}\|x\|^2$ a pour fonction duale

$$V(\Lambda)=\inf_{x\in\mathbb{R}^n}[\|x\|^2-\langle\Lambda, Ax-b\rangle]$$

Le x minimisant le problème précédent est solution de

$$0=\nabla_x(\|x\|^2-\langle\Lambda, Ax-b\rangle)=2x-{}^{\mathsf{T}}A\Lambda,$$

ainsi $x=-\frac{1}{2}{}^{\mathsf{T}}A\Lambda$ et $V(\Lambda)=-\|{}^{\mathsf{T}}A\Lambda\|^2/4+\langle\Lambda, b\rangle$. Le programme dual est sans contrainte

$$\sup_{\Lambda\in\mathbb{R}^m}[-\|{}^{\mathsf{T}}A\Lambda\|^2/4+\langle\Lambda, b\rangle].$$

3.16.4 Le lagrangien du programme

$$\inf_{\sum_{i=1}^n x_i=1}\left[\sum_{i=1}^n f_i(x_i)\right]$$

est $\mathcal{L}(x,\lambda)=\sum_{i=1}^n[f_i(x_i)+\lambda x_i]-\lambda$. Vu la définition 3.5 de la conjuguée de Fenchel-Nielsen de F

$$F^*(\mu)=\sup_{x\in\mathbb{R}^n}[\langle x,\mu\rangle-F(x)]=-\inf_{x\in\mathbb{R}^n}[F(x)-\langle x,\mu\rangle],\quad \mu\in\mathbb{R}^n,$$

la fonction duale V est donc

$$V(\lambda)=\inf_x\mathcal{L}(x,\lambda)=-\lambda+\sum_{i=1}^n\inf_{x_i}[f_i(x_i)+\lambda x_i]=-\lambda-\sum_{i=1}^n f_i^*(-\lambda).$$

On a donc remplacé un problème convexe (à n variables) par n problèmes convexes (à une variable). De plus, cet exemple exhibe le lien entre la fonction de profit du programme dual et la transformée de Fenchel-Nielsen de la fonction de coût du programme primal. ◁

On a toujours l'*inégalité de dualité faible*

$$d^*\leq p^*$$

vu l'inégalité de min-max

$$d^*=\sup_{\Lambda,M\geq 0}\inf_{x\in\mathbb{R}^n}\mathcal{L}(x,\Lambda,M)\leq p^*=\inf_{x\in\mathbb{R}^n}\sup_{\Lambda,M\geq 0}\mathcal{L}(x,\Lambda,M), \tag{69}$$

cas particulier du lemme suivant

Lemme 3.15: *Soit* $\mathbf{L}$ *une fonction numérique définie sur le produit* $A\times B$. *Alors*

$$\sup_{b\in B}\inf_{a\in A}\mathbf{L}(a,b)\leq\inf_{a\in A}\sup_{b\in B}\mathbf{L}(a,b), \tag{70}$$

Démonstration. Ce résultat découle des inégalités

$$\inf_{\widetilde{a}\in A}\mathbf{L}(a,\widetilde{b})\leq L(a,b)\leq\sup_{\widetilde{b}\in B}\mathbf{L}(\widetilde{a},b),$$

puis

$$\sup_{b\in B}\inf_{\widetilde{a}\in A}\mathbf{L}(a,\widetilde{b})\leq\inf_{a\in A}\sup_{\widetilde{b}\in B}\mathbf{L}(\widetilde{a},b).$$

□

DÉFINITION 3.10: *Le couple* $(\overline{a},\overline{b})$ *est appelé* point selle *ou* point de min-max *de (70) s'il vérifie*

$$\mathbf{L}(\overline{a},b)\le\mathbf{L}(\overline{a},\overline{b})\le\mathbf{L}(a,\overline{b}),\quad a\in A,b\in B. \tag{71}$$

La fonction de min-max $\mathbf{L}$ (pas nécessairement construite comme fonction lagrangienne d'un problème de minimisation avec contraintes) donne lieu à un couple de problèmes primal/dual

$$\inf_{a\in A}J(a),\qquad \sup_{b\in B}V(b), \tag{72}$$

avec $J(a)=\sup_{b\in B}\mathbf{L}(a,b)$ et $V(b)=\inf_{a\in A}\mathbf{L}(a,b)$

PROPOSITION 3.10: *Le couple* $(\overline{a},\overline{b})$ *est un point-selle de* L *sur* $A\times B$ *si et seulement si* $J(\overline{b})=\min_{b\in B}J(b)=\max_{a\in A}V(a)=V(\overline{a})=\mathbf{L}(\overline{a},\overline{b})$.

DÉMONSTRATION. D'après l'inégalité de dualité faible et la définition de J et V, on a

$$V(\overline{b})=\inf_{a\in A}\mathbf{L}(a,\overline{b})\le\sup_{b\in B}\inf_{a\in A}\mathbf{L}(a,b)\le\inf_{a\in A}\sup_{b\in B}\mathbf{L}(a,b)\le\sup_{b\in B}\mathbf{L}(\overline{a},b)=J(\overline{a}).$$

Si $(\overline{a},\overline{b})$ est un point selle, alors les membres extrêmes de l'inégalité précédente sont égaux à $\mathbf{L}(\overline{a},\overline{b})$, et par suite avec tous les membres de cette inégalité. La réciproque s'en déduit pareillement. □

▷ EXEMPLES 3.17:

3.17.1 Étant donné un programme de minimisation de fonction d'objectif J, il n'y a pas unicité pour le choix de la fonction $\mathbf{L}$ de min-max qui donne ce programme comme programme primal. Ainsi, pour le problème

$$\min_{\substack{x\ge0\\ b\le Ax}}\langle c,x\rangle$$

on peut introduire la fonction de min-max $\mathcal{L}(x,y)$ définie par

$$\mathcal{L}(x,y)=\langle c,x\rangle+\langle b,y\rangle-\langle Ax,y\rangle,\quad (x,y)\in\mathbb{R}^n_+\times\mathbb{R}^m_+.$$

On a

$$\max_{y\ge0}\mathcal{L}(x,y)=\begin{cases}\langle c,x\rangle & \text{si } b-Ax\le0,\\ +\infty & \text{sinon,}\end{cases}\qquad \inf_{x\ge0}\mathcal{L}(x,y)=\begin{cases}\langle b,y\rangle & \text{si } c-{}^{\mathsf{T}}Ay\ge0,\\ -\infty & \text{sinon.}\end{cases}$$

avec les programmes (primal et son dual) correspondants

$$\min_{\substack{x\ge0\\ b\le Ax}}\langle c,x\rangle,\qquad \max_{\substack{y\ge0\\ c\ge{}^{\mathsf{T}}Ay}}\langle b,y\rangle;$$

3.17.2 Soit $U(x)$ la fonction de coût pour une production modélisée par x, avec des contraintes $h(x)=(h_j(x))\ge0$: la minimisation $p^*=\inf_{h(x)\ge0}U(x)$ du coût correspond à la maximisation du profit $-U(x)$. La contrainte $h(x)\ge0$ exprime une limitation de divers types : des ressources humaines, un espace de production, une norme environnementale,... Supposons une situation avec possibilité de passer outre les contraintes avec un coût additionnel linéaire $-\mu_jh_j(x)$ (éventuellement négatif, en cas de subvention au producteur ou une contrainte non effective, comme un espace non utilisé qui puisse être loué à un tiers) : le scalaire μ_j (supposé positif) est à interpréter comme le prix à acquitter par unité de la j-ème

contrainte. Le coût total pour la production x et l'ensemble de prix de contrainte $M = (\mu_j)$ est

$$\mathcal{U}(x, M) = U(x) - \sum_{j=1}^{p} \mu_j h_j(x) = U(x) - \langle M, h(x)\rangle$$

que le producteur souhaite minimiser suivant

$$V(M) = \inf_x \mathcal{U}(x, M).$$

La fonction duale $V(M)$ représente le coût optimal en fonction des coûts M de contrainte et $d^* = \max_{M\geq 0} V(M)$ le coût optimal pour l'ensemble de coûts M_* le moins favorable.

La dualité faible indique que le coût optimal $V(M)$ pour une production négligeant les contraintes (le coût du dépassement d'icelles étant indiqué par le vecteur de prix unitaires M) est majoré par le coût de production p^* respectant strictement les contraintes. L'écart $p^* - d^*$ est l'avantage minimum pour le producteur à dépasser les contraintes exprimées par $-h_j \leq 0$: si $d^* = p^*$ avec d^* atteint en M_*, ce vecteur de coûts unitaires M_* est celui pour lequel l'entreprise n'a pas avantage à être autorisée à payer pour ce non respect des contraintes. On appelle parfois ce vecteur M_* ensemble de *prix virtuels* ou *prix d'équilibre*[15]. ◁

La différence $p^* - d^*$, toujours positive, est appelée *saut de dualité optimum* : si elle est nulle, on dit que il y a *dualité forte* entre les problèmes dual et primal.

▷ Exemple 3.18: Le problème $\inf_{Ax=b} \|x\|^2$ (cf. exemple 3.16.4) vérifie la dualité forte.

En effet, soit $b \notin \operatorname{Im} A$, auquel cas $p^* = +\infty$ (par convention : l'ensemble des réalisables est vide) et $d^* = -\infty$, car la fonction V en restriction à $\mathbb{R}\Lambda_0$ avec $\Lambda_0 \in \ker {}^{\mathsf{T}}A = (\operatorname{Im} A)^{\perp}$ et $\Lambda \in (\operatorname{Im} A)^{\perp} \mapsto \langle \Lambda, b\rangle$ est linéaire non nulle vu que $b \notin \operatorname{Im} A$.

Soit $b \in \operatorname{Im} A$ et alors $d^* = p^*$ comme le calcul explicite va l'attester. Les équations de Lagrange du problème $\inf_{Ax=b} \|x\|^2$ sont $2x_* - {}^{\mathsf{T}}A\Lambda^* = 0, Ax_* = b$. Vu que $\operatorname{Im} A^{\mathsf{T}}A = (\ker A^{\mathsf{T}}A)^{\perp} = (\ker {}^{\mathsf{T}}A)^{\perp}$, la restriction A_2 de $A^{\mathsf{T}}A$ à $\operatorname{Im} A$, soit $A_2 x \in \operatorname{Im} A \mapsto A^{\mathsf{T}}Ax \in \operatorname{Im} A$ est bijective. Introduisons le vecteur $b_A \in \operatorname{Im} A$ défini par $b_A = A_2^{-1}b$: $\Lambda^* = 2b_A$ et le minimum est $x_* = {}^{\mathsf{T}}Ab_A$ avec

$$\inf_{Ax=b} \|x\|^2 = \|x_*\|^2 = \langle {}^{\mathsf{T}}Ab_A, {}^{\mathsf{T}}Ab_A\rangle = \langle A^{\mathsf{T}}Ab_A, b_A\rangle = \langle b, b_A\rangle = \langle b, A_2^{-1}b\rangle$$

Par ailleurs, on a

$$V(\Lambda) = -1/4\langle {}^{\mathsf{T}}A\Lambda, {}^{\mathsf{T}}A\Lambda\rangle - \langle \Lambda, (A^{\mathsf{T}}A)b_A\rangle = -\|{}^{\mathsf{T}}A(\Lambda/2 + b_A)\|^2 + \|{}^{\mathsf{T}}Ab_A\|^2$$

ce qui implique, vu

$$\|{}^{\mathsf{T}}Ab_A\|^2 = \langle {}^{\mathsf{T}}Ab_A, {}^{\mathsf{T}}Ab_A\rangle = \langle A^{\mathsf{T}}Ab_A, b_A\rangle = \langle b, b_A\rangle,$$

la dualité forte $\sup_{\Lambda} V(\Lambda) = \inf_{Ax=b} \|x\|^2$. ◁

Théorème 3.17 (Dualité forte de Slater): *On suppose le problème primal (65) convexe avec un point réalisable $\widetilde{x}$ dans l'intérieur (relatif) du convexe C de minimisation vérifiant $h_j(\widetilde{x}) > 0$ pour $j = 1, \ldots, p$ et $A\widetilde{x} = b$. Alors les valeurs optimales du problème primal (65) et du problème dual (68) coïncident*

$$d^* = \sup_{\Lambda, M\geq 0} V(\Lambda, M) = \inf_{g(x)=0, h(x)\geq 0} U(x) = p^*.$$

15. *shadow prices* en anglais.

De plus, si x_* est un point de minimisation du programme primal et (Λ^*, M^*) un point de maximum du problème dual, alors, (x_*, Λ^*, M^*) est un point selle du lagrangien $\mathcal{L}$

$$d^* = \mathcal{L}(x_*, \Lambda^*, M^*) = p^*.$$

▷ Exemples 3.19:

3.19.1 On a vu dans l'exemple 3.18 comment la dualité forte valait pour le problème primal $\inf_{Ax=b} \|x\|^2$ et son dual $\sup_{\Lambda\in\mathbb{R}^m}[-1/4\|^{\mathsf{T}}A\Lambda\|^2 - \langle\Lambda, b\rangle]$.

3.19.2 Il y a dualité forte, avec un point selle, pour un programme linéaire (et son dual qui aussi linéaire) dès que celui-ci a un point réalisable. Le seul cas, où il n'y a pas dualité forte est celui où le problème primal et le problème dual n'ont pas de points réalisables : on a dans ce cas $d^* = -\infty$ et $p^* = +\infty$.

3.19.3 Soit le problèmes convexe

$$\inf_{\substack{x,y>0\\ x^2/y\leq 0}} [\mathrm{e}^{-x}]$$

Il est équivalent au problème $\inf_{x=0} \mathrm{e}^{-x} = 1$. Par ailleurs, le lagrangien est $\mathcal{L}(x, y, \lambda) = \mathrm{e}^{-x} + \lambda x^2/y$ défini sur $\mathbb{R} \times \mathbb{R}_+^2$ et fonction duale

$$V(\lambda) = \inf_{x,y>0} [\mathrm{e}^{-x} + \lambda x^2/y] = 0$$

Le saut de dualité est donc $p^* - d^* = 1$. ◁

△ Remarques 3.17:

(1) De nombreux algorithmes fournissent des points effectifs x_k du problème primal et (Λ_k, M_k) du problème dual. On a $V(\Lambda_k, M_k) \leq U(x_k)$: si de plus $U(x_k) - V(\Lambda_k, M_k) \to 0$, alors $\lim_k U(x_k) = \lim_k V(\Lambda_k, M_k) = p^*$, avec possibilité de test d'arrêt $U(x_k) - V(\Lambda_k, M_k) \leq \varepsilon$. Le problème dual étant convexe est parfois résoluble de manière approchée "relativement" facilement.

(2) Le point selle (x_*, Λ^*, M^*) du théorème 3.17 vérifie les conditions KKT pour le lagrangien $\mathcal{L}$: l'étude des points selle du lagrangien $\mathcal{L}$ est étroitement liée à celle des points vérifiant KKT pour le problème primal. ▽

Chapitre 4

Programmation linéaire

1. Programmes linéaires

Un programme linéaire est de la forme

$$\min_{\substack{A_1x=b_1\\A_2x\geq b_2}} [\langle c,x\rangle + d]$$

où la variable de décision x dans $\mathbb{R}^n$ et $c\in\mathbb{R}^n$, $d\in\mathbb{R}$ (aucun dommage à supposer $d=0$!) tandis que $b_k\in\mathbb{R}^{m_k}, A_i\in\mathbb{R}^{n\times m_k}$ pour $k=1,2$. L'équation $A_1x=b_1$ condense des contraintes (linéaires) d'égalités $\langle a_{1j},x\rangle = b_{1j}$ avec $a_{1j}\in\mathbb{R}^n, j=1,\ldots,m_1$ des vecteurs de $\mathbb{R}^n$ et l'inégalité $A_2x\geq b_2$ condense des contraintes (linéaires) d'inégalités $\langle a_{2j},x\rangle\geq b_{2j}$ avec $a_{2j}\in\mathbb{R}^n, j=1,\ldots,m_2$ des vecteurs de $\mathbb{R}^n$. Les vecteurs $a_{kj}\in\mathbb{R}^n, j=1,\ldots,m_k$ correspondent aux lignes de la matrice A_k.

▷ Exemples 4.1:

4.1.1 Considérons la recherche de la distance entre deux sommets s et t d'un graphe $G=(S,A)$ dont chaque arête a est pondérée par un poids $w_a\geq 0$ [1] (w_a représente la longueur, le temps,... pour aller entre les deux nœuds/extrémités de l'arête). La distance $d(s,t)$ de s à t, définie comme le minimum des poids $w_C = w_{a_1} + w_{a_2} + \ldots + w_{a_k}$ des chemins $C = a_1a_2\ldots a_k$ de s à t, est donnée comme l'optimum du programme linéaire de variables $(x_u)_{u\in S}\in\mathbb{R}^S$

(73)
$$\max_{\substack{x_v\leq x_u+w_{(u,v)},(u,v)\in A\\x_s=0}} x_t$$

En effet, la fonction $u\in S\mapsto d(s,u)$ est un réalisable du programme : si C_u est un chemin de s à u de longueur minimale, $C_u\cup(u,v)$ est un chemin de s à v et donc $d(s,v)\leq d(s,u)+w_{(u,v)}$. Ainsi, le maximum x_t^* est minoré par la distance de s à t. Réciproquement, on considère un chemin $C=(s=u_1,u_2,\ldots,u_k,u_{k+1}=t)$ de s à t de longueur $d=\sum_{j=1}^k w_{(u_j,u_{j+1})}$ minimale : alors pour (x_u) réalisable de (73),

$$\begin{aligned} x_{u_2} &\leq x_s + w_{(s,u_2)} \leq w_{(s,u_2)}\\ x_{u_3} &\leq x_{u_2} + w_{(u_2,u_3)} \leq w_{(s,u_2)} + w_{(u_2,u_3)}\\ &\ldots\\ x_t &\leq w_C = d(s,t)\end{aligned}$$

ce qui établit $d(s,t)$ comme optimum du programme (73).

4.1.2 Des modélisations incorporent souvent, par pur réalisme, des données incertaines (incarnées par des variables aléatoires sur un ensemble fini et à distribution connue). Par exemple, un plan d'alimentation en eau d'une ville dépend de la pluviométrie, de la demande en eau de consommateurs et des industriels. Un

1. Si w_a est de signe quelconque, il faut faire l'hypothèse de non-existence de cycles $a_1a_2\ldots a_k$ de poids $w_{a_1}w_{a_2}\ldots w_{a_k}$ négatif

programme stochastique comprend une variable de décision présente x et des variables de décision futures $y_1, \dots, y_K$, chacune incarnant un scénario probable. La variable de décision x^* maximise le bénéfice espéré sur tous les scénarios du modèle, qui se développe en plusieurs étapes (deux, présent et futur, pour cet exemple introductif).

On considère donc un programme linéaire $\max_{\substack{Ax=b\\x\geq 0}} \langle c, x\rangle$ où b est un vecteur aléatoire avec différentes valeurs $b_1, \dots, b_K$ de probabilités respectives $\pi_1, \dots, \pi_K$. Le vecteur b étant connu, une deuxième décision y (dite décision de recours) est supposée ajuster la décision x à travers le programme $\max_{\substack{My=b-Ax\\y\geq 0}} \langle q, y\rangle$, où la matrice M incorpore l'interaction entre la décision x et la décision de recours y. Le problème est de déterminer la décision x qui optimise la fonction objectif $\langle c, x\rangle$ augmentée des coûts de recours $\langle q, y_k\rangle$ attendus pour chaque scénario de type k en moyenne (soit $\mathbb{E}[\langle q, y\rangle]$).

$$\max_{\substack{x\geq 0\\Ax=b}} [\langle c, x\rangle + \mathbb{E}[\langle q, y\rangle | My_k = b_k - Ax, y_k \geq 0, k = 1, \dots, K]$$

soit la maximixation de $\langle c, x\rangle + \sum_{k=1}^{K} \pi_k \langle q, y_k\rangle$ avec les contraintes

$$\begin{array}{rcccccl} Ax & + & My_1 & & & = & b_1 \\ Ax & + & & My_2 & & = & b_2 \\ \vdots & & & & & & \\ Ax & + & & & My_K & = & b_K \\ x & \geq & 0, \quad y_k & \geq & 0 \quad \text{pour } k = 1, \dots, K & & \end{array}$$

La taille du programme linéaire est importante, vu la grande variété possibles des scénarios aléatoires : néanmoins, le programme stochastique a une formulation en un programme linéaire déterministe, que les calculateurs contemporains n'ont aucune difficulté désormais à résoudre (d'autant plus que le programme a une forme particulière, avec des matrices creuses).

4.1.3 Soit $P = \{\langle a_i, x\rangle \leq b_i, i = 1, \dots, n\}$ un polygone : un centre x_C de Chebyshev est le centre de toute boule $B(x_C, R)$ de rayon maximal R incluse dans P. La boule $B(x_C, R)$ est incluse dans P si et seulement si $\langle a_i, x_C + Ru\rangle \leq b_i$ pour $i = 1, \dots, m$ et $\|u\| \leq 1$, soit, vu Cauchy-Schwarz, $\langle a_i, x_C\rangle + \|a_i\|R \leq b_i$ pour $i = 1, \dots, m$. Ainsi est-on ramené au programme linéaire

$$\max_{\substack{R\geq 0, x\in\mathbb{R}^2\\ \langle a_i, x_C\rangle + \|a_i\|R \leq b_i, i=1,\dots,m}} R.$$

◁

Un programme de la forme

$$\max_{Ax\geq b} [\langle c, x\rangle + d] \tag{74}$$

est dit à forme *canonique* (ou *à inégalités*), alors que

$$\max_{\substack{Ax=b\\x\geq 0}} [\langle c, x\rangle + d] \tag{75}$$

est dit de forme *standard* ;

Pour se ramener à une forme canonique, on remplace chaque contrainte d'égalité $\langle a, x\rangle = b$ par deux contraintes d'inégalité $\langle a, x\rangle \geq b$ et $\langle -a, x\rangle \geq -b$. Pour se ramener

à une forme standard, on introduit une variable d'*écart*[2] s pour chaque contrainte d'inégalité $\langle a, x\rangle \leq b$ remplacée par l'égalité $\langle a, x\rangle + s = b$ et en complétant chaque variable sans signe x par deux variables x', x'' positives et la contrainte d'égalité $x = x' - x''$. On peut transformer tout programme linéaire sous la forme (75).

△ Remarque 4.1: Dans la forme standard, on s'autorise parfois un ensemble réduit d'inégalités $x_i \geq 0, i \in I$ avec $I \subset \{1, \ldots, n\}$. Dans ce cadre, en raisonnant en coordonnées ou bien en rajoutant une variable X et l'équation $X - \langle c, x\rangle = 0$ dans les contraintes, on peut même supposer que c est un vecteur de base, autrement dit que la fonction d'objectif est une coordonnée.

▷ Exemple 4.2: Le programme

$$\min_{\substack{x_1-3x_2+2x_3\leq 3\\ -x_1+2x_2\geq 2\\ x_2,x_3\geq 0}} [-2x_1 + 3x_2]$$

est équivalent à

$$-\max_{\substack{x_1'-x_1''-3x_2+2x_3+s_1=3\\ -x_1'+x_1''+2x_2-s_2=2\\ x_1',x_1'',x_2,x_3,s_1,s_2\geq 0}} [2x_1' - 2x_1'' - 3x_2]. \qquad \triangleleft$$

2. Hyperplans de séparation

Dans $\mathbb{R}^n$ avec $n \geq 2$, le segment $[x, y]$ est d'intérieur vide, bien qu'il soit d'intérieur non vide s'il est considéré relativement à la droite qui le contient. De manière générale, on parlera de l'*intérieur relatif* $\mathrm{ir}(C)$ d'un convexe C, et de sa *frontière relative*, en considérant C comme une partie de l'espace affine $\mathrm{Aff}(C)$ qu'il engendre : l'enveloppe $\mathrm{Aff}(C)$ est le plus petit sous-espace affine contenant C, soit l'intersection de tous les sous-espaces affines contenant C, ou encore l'ensemble des barycentres $\sum_i \lambda_i M_i$ ($\sum \lambda_i = 1, \lambda_i \in \mathbb{R}$) construits à partir de points de C. La *dimension* (affine) de C est celle de son enveloppe affine.

Définition 4.1: *Les parties S et T sont dites* proprement séparées *s'il existe un hyperplan H tels que S et T sont inclus dans les demi-espaces opposés associés à H et au moins une des parties n'est pas incluse dans H.*

△ Remarque 4.2: Si H est du type $H = H_{d,C} = \{\langle d, u\rangle = C\}$, alors, sauf à échanger d, C en $-d, -C$, la séparation de S et T signifie que $\sup_{u\in S}\langle u, d\rangle \leq C \leq \inf_{u\in T}\langle u, d\rangle$ et $\inf_{u\in S}\langle u, d\rangle < \sup_{u\in T}\langle u, d\rangle$. ▽

Le théorème suivant est admis :

Théorème 4.1 (Théorème de séparation): *Deux ensembles convexes non vides S et T dans $\mathbb{R}^n$ peuvent être séparés proprement si et seulement si leurs intérieurs relatifs sont disjoints.*

Définition 4.2: *Soit C fermé convexe et x un point de sa frontière relative. L'hyperplan $H_{d,\langle d,x\rangle} = \{y | \langle d, y\rangle = \langle d, x\rangle\}$ est un hyperplan d'appui (ou de support) de C si $\sup_{y\in C}[\langle d, y\rangle] = \langle d, x\rangle$ et C n'est pas inclus dans l'hyperplan $H_{d,\langle d,x\rangle}$.*

Théorème 4.2: *Soit x un point dans la frontière relative du convexe fermé C.*

(i) Il existe un hyperplan d'appui en x.

(ii) Si H est un hyperplan d'appui en x, alors $C \cap H$ est de dimension affine strictement inférieure à celle de C

2. *slack variables* en anglais

Démonstration. La première assertion résulte du théorème de séparation : il suffit de prendre comme hyperplan d'appui l'hyperplan séparant le point x et l'intérieur relatif de C.

Pour la seconde, supposons H de la forme $H_{d,\langle d,x\rangle}$: vu que H est hyperplan d'appui, la forme linéaire $u \mapsto \langle u, d\rangle$, non constante sur H, l'est sur l'intersection $C_H = C \cap H$, donc sur $\mathrm{Aff}(C_H)$: ainsi $\mathrm{Aff}(C_H)$ est inclus strictement dans $\mathrm{Aff}(C)$ et $\mathrm{Aff}(C_H)$ est de dimension strictement inférieure à celle de $\mathrm{Aff}(C)$. □

3. Points extrêmaux

Définition 4.3: *Soit C convexe non vide. Le point $x \in C$ est dit* point extrémal *de C s'il n'existe aucun segment $[u, v]$ inclus dans C et contenant x en son intérieur relatif.*

▷ Exemple 4.3: Les points extrémaux d'un segment (resp. triangle, disque fermé ou ouvert) sont ses deux extrémités (resp. ses trois sommets, les points de son bord ou inexistants). Un sous-espace vectoriel strict ne contient aucun point extrémal. ◁

Le lemme suivant a sa preuve faite par le lecteur.

Lemme 4.1: *Soit x point d'un convexe C. Le point x est extrémal si et seulement si $C \setminus \{x\}$ est convexe.*

Lemme 4.2: *Soit C convexe fermé. Si C contient une demi-droite issue de x et de direction d, alors pour tout point $y \in C$, la demi-droite issue de y et de direction d est contenue dans C.*

Démonstration. Soit $\tau > 0$. Alors, pour $\varepsilon \in (0, 1)$, le point $m_\varepsilon = (1-\varepsilon)y + \varepsilon(x + \tau\varepsilon^{-1}d)$ est dans C ainsi que sa limite $y + \tau d$ lorsque $\varepsilon \to 0^+$: ainsi $y + \mathbb{R}d \subset C$. Géométriquement, le segment $[y, m_\varepsilon]$ converge vers le segment $[y, y + \tau d]$. □

Théorème 4.3: *Soit C convexe fermé.*

(i) C ne contient aucun point extrémal si et seulement si C contient une droite.

(ii) (Minkowski) Si C est borné, C est l'enveloppe convexe de l'ensemble de ses points extrémaux : $C = \mathrm{Conv}(\mathrm{Ext}(C))$.

Démonstration. Si C contient une droite de direction d, alors, d'après le lemme précédent, par tout point de C passe une droite de même direction d : aucun point n'est donc extrémal. Pour la réciproque, admettons un instant le lemme suivant :

Lemme 4.3: *Soit C convexe fermé non vide, x un point de la frontière relative de C et H un hyperplan d'appui en x. Alors tous les points extrémaux de $C_H = C \cap H$ sont extrémaux dans C.*

On va établir par récurrence sur $n = \dim(C)$ qu'un convexe fermé ne contenant aucune droite contient des points extrémaux. Pour $n = 0$ ou 1, c'est clair. Si $n > 1$, soit $x_0 \in \mathrm{Aff}(C) \setminus C$ (non vide, sinon C serait un espace affine qui contient des droites) et $\overline{x}$ un point de la frontière relative de C (qui existe sur tout segment $[x, x_0]$ avec $x \in C$) avec H comme hyperplan d'appui. Le convexe $C_H = C \cap H$ ne contient pas de droites (sinon C en contiendrait une), donc admet un point extrémal (d'après l'hypothèse de récurrence) qui est aussi extrémal relativement à C (d'après le lemme 4.3 prouvé ci-dessous).

Pour la seconde partie, vu que C est convexe, alors $C \supset \mathrm{Conv}(\mathrm{Ext}(C))$. Pour l'inclusion inverse, on fait à nouveau une récurrence sur la dimension. Soit $x \in C$ et une direction d dans l'espace de directions de l'espace affine $\mathrm{Aff}(C)$: la droite $x + \mathbb{R}d$ a pour intersection avec C le segment $[x_-, x_+]$. Le point $x_\pm$ est dans la frontière relative de C : si $H_\pm$ est un hyperplan d'appui à C en $x_\pm$, alors $C_{H_\pm} = C \cap H_\pm$ est convexe fermé borné

de dimension strictement inférieure à celle de C : d'après l'hypothèse de récurrence et le lemme 4.3,

$$x_\pm \in \mathrm{Conv}(\mathrm{Ext}(C_{H_\pm})) \subset \mathrm{Conv}(\mathrm{Ext}(C))$$

et par suite $x \in [x_-, x_+] \subset \mathrm{Conv}(\mathrm{Ext}(C))$, soit $C \subset \mathrm{Conv}(\mathrm{Ext}(C))$, ce qui achève la démonstration. □

Preuve du Lemme 4.3. La partie $C \cap H$ est fermée convexe. Soit $\overline{x}$ extrémal dans $C \cap H$. Supposons $\overline{x} = \lambda u + (1-\lambda)v$ avec $\lambda \in (0,1)$ et $u, v \in C$. On a, si $H = H_{d,\langle x,d\rangle}$,

$$\langle d, x\rangle = \langle d, \overline{x}\rangle = \lambda\langle d, u\rangle + (1-\lambda)\langle d, v\rangle$$

Par suite, vu que $\langle x, d\rangle \geq \max(\langle u, d\rangle, \langle v, d\rangle)$, $\langle d, x\rangle = \langle d, u\rangle = \langle d, v\rangle$, soit u, v dans $H \cap C$ et $y = u = v$ car y extrémal dans $C \cap H$: on vient de montrer que y est extrémal dans C. □

4. Polyèdres

Définition 4.4: *Un ensemble polyédral convexe de l'espace vectoriel E est toute partie $P_{A,b}$ de E définie par les inégalités $Ax \geq b$ avec A un morphisme linéaire de E dans F et $b \in F$.*

Si E_a est un espace affine avec espace vectoriel de directions l'espace E, on considérera un point base $O \in E_a$, induisant la bijection $M \in E_a \mapsto \overrightarrow{OM}$: la partie $P_{A,b}$ se transporte dans l'espace affine E_a sur la partie $\{M \in E_a | A(\overrightarrow{OM}) \geq b\}$.

△ Remarque 4.3: Si $E = \mathbb{R}^n$ et $F = \mathbb{R}^m$, on peut considérer l'équation $Ax \geq b$ avec A matrice à m lignes et n colonnes. Cette inégalité vectorielle est équivalente aux m inégalités scalaires $\langle a_i, x\rangle \geq b_i$ où a_i sont les vecteurs ligne de la matrice A. ▽

Définition 4.5: *Soient S et D deux parties finies, éventuellement vides, de (l'espace affine) E et de (l'espace vectoriel des directions) $E\backslash\{0\}$ resp.. La partie convexe* $\mathrm{Conv}(S)$ *engendrée par S est définie par*

$$\mathrm{Conv}(S) = \left\{ \sum_{s\in S} \lambda_s s, \lambda_s \geq 0, \sum_{s\in S} \lambda_s = 1 \right\}$$

et la partie conique engendrée par D est définie par

$$\mathrm{Cône}(D) = \left\{ \sum_{d\in D} \mu_d d, \mu_d \geq 0 \right\}$$

La somme des parties convexes $\mathrm{Conv}(S)$ *et* $\mathrm{Cône}(D)$ *est notée $M(S,D)$ et vérifie*

$$M(S,D) = \mathrm{Conv}(S) + \mathrm{Cône}(D) = \left\{ \sum_{s\in S} \lambda_s s + \sum_{d\in D} \mu_d d, \lambda_s \geq 0, \sum_{s\in S} \lambda_s = 1, \mu_d \geq 0 \right\}$$

Si D est vide, la partie $M(S) = M(S,\emptyset)$ est bornée et est appelée polytope.

Le lemme suivant est facilement établi :

Lemme 4.4: *La partie $M(S,D)$ est convexe.*

Démonstration. Les parties $\mathrm{Conv}(S)$ et $\mathrm{Cône}(D)$ sont convexes : par ex., pour $\mathrm{Conv}(S)$, étant donnés deux points $x, \widetilde{x}$ correspondants à des familles $(\lambda_s)_{s\in S}$ et $(\widetilde{\lambda}_s)_{s\in S}$, la combinaison convexe $\alpha x + (1-\alpha)\widetilde{x}$ correspond à la famille $(\alpha\lambda_s + (1-\alpha)\widetilde{\lambda}_s)_{s\in S}$ et est donc un point de $M(S)$. La partie $M(S,D)$, somme de deux convexes, est convexe. □

Le résultat majeur de cette partie est le théorème suivant de caractérisation des ensembles polyédraux convexes (sa démonstration est développée dans la section 6)

THÉORÈME 4.4: *Il y a identité entre la classe des ensembles polyédraux convexes $P_{A,b}$ et celle des parties du type $M(S,D)$.*

△ REMARQUE 4.4: L'ensemble $M(S,\emptyset)$ est un ensemble polyédral convexe borné et cette présentation caractérise les parties polyédrales convexes bornées : un cône Cône(D) (non vide) n'est pas borné. Les cônes polyédraux convexes sont les parties $M(\emptyset,D)$, décrites par des inégalités homogènes $Ax \geq 0$. Si l'ensemble polyédral convexe P ne contient pas de droite, alors la partie des points extrémaux est finie et $P = M(\text{Ext}(P))$. La partie conique Cône(D) d'une partie polyédrale convexe P vérifie Cône$(D) = \{x + td | x + \mathbb{R}^+ d \subset C\}$ (sans que la partie D ne soit uniquement déterminée : Cône(D) = Cône$(D/2)$). ▽

L'importance du théorème de caractérisation précédent est illustrée par les corollaires suivants sur des opérations concernant des parties polyédrales convexes.

COROLLAIRE 4.1: *Soient P, P_1, P_2 des parties polyédrales convexes*

(i) Si $\Phi : E \to F$ est affine, alors $\Phi(P)$ est polyédrale convexe,

(ii) Si $\Psi : G \to E$ est affine, alors $\Psi^{-1}(P)$ est polyédrale convexe,

(iii) L'intersection $P_1 \cap P_2$ est polyédrale convexe.

(iv) La somme convexe $P_1 + P_2$ est polyédrale convexe.

DÉMONSTRATION. Cela résulte des identités suivantes, ou pour chaque cas le choix de la représentation des parties polyédrales convexes est effectué de manière appropriée.

$$\Phi(M(S,D)) = M(\Phi(S), \varphi(D)),$$
$$\Psi^{-1}(\{Ax \geq b\}) = \{(A \circ \psi)(y) \geq b - Au\},$$
$$\{A_1 x \geq b_1\} \cap \{A_2 x \geq b_2\} = \left\{ \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} x \geq \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right\},$$
$$M(S_1, D_1) + M(S_2, D_2) = M(S_1 + S_2, D_1 + D_2)$$

Si φ est la partie linéaire de Φ, on a

$$\Phi\left(\sum_{s\in S} \lambda_s s + \sum_{d\in D} \mu_d d\right) = \sum_{s\in S} \lambda_s \Phi(s) + \sum_{d\in D} \mu_d \varphi(d),$$

ce qui donne la première identité. Pour la seconde, on écrit $\Psi(M) = \Psi(O) + \psi(\overrightarrow{OM})$ où ψ est la partie linéaire de Ψ et O le point base de l'espace affine E_a : l'inéquation $Ax \geq b$ appliquée à $\Psi(M)$ prend la forme est $b \geq A\overrightarrow{O\Psi(M)} = A[\overrightarrow{O\Psi(O)} + \psi(\overrightarrow{OM})]$, soit

$$A \circ \psi(\overrightarrow{OM}) \geq b - A\overrightarrow{O\Psi(O)}.$$

Pour la dernière égalité, on a

$$\sum_{s_1\in S_1} \lambda_{s_1} s_1 + \sum_{s_2\in S_2} \lambda_{s_2} s_2 = \sum_{(s_1,s_2)\in S_1\times S_2} \lambda_{s_2}\lambda_{s_1} s_1 + \sum_{(s_1,s_2)\in S_1\times S_2} \lambda_{s_1}\lambda_{s_2} s_2$$
$$= \sum_{(s_1,s_2)\in S_1\times S_2} \lambda_{s_1}\lambda_{s_2}(s_1 + s_2),$$

i. e. $M(S_1)+M(S_2) = M(S_1+S_2)$. On a de même Cône$(D_1)$+Cône$(D_2)$ = Cône(D_1+D_2) et la relation mentionnée sur $M(S,D)$. □

5. Résolution de problèmes linéaires

On considère ici le programme linéaire (PL) sous la forme canonique

$$\text{(PL)} \qquad \inf_{Ax \geq b} \langle c, x \rangle$$

i. e. on cherche à minimiser la forme linéaire $x \mapsto \langle c, x \rangle$ sur le polyèdre $P = \{Ax \geq b\}$

DÉFINITION 4.6: *Soit (PL) un programme linéaire.*

(i) Le problème (PL) est dit admissible *si le polyèdre P est non vide.*

(ii) Le problème (PL) est dit borné [3] *s'il est admissible et si* $\inf_{x \in P} \langle c, x \rangle > -\infty$.

(iii) Le problème (PL) est dit soluble *s'il existe x_* dans P tel que* $\langle c, x_* \rangle = \inf_{x \in P} \langle c, x \rangle$. *Le point x_* est appelé solution du programme (PL).*

▷ EXEMPLES 4.4:

4.4.1 Le programme

$$\max_{\substack{x_1 - x_2 \leq 1 \\ 2x_1 + x_2 \geq 6 \\ x_1 \geq 3, -x_2 \geq 2 \\ x_1, x_2 \geq 0}} 2x_1 - x_2$$

est sans solution (infaisable), *i. e.* non admissible.

4.4.2 Le programme

$$\max_{\substack{x_1 - x_2 \leq 1 \\ 2x_1 + x_2 \geq 6 \\ x_1, x_2 \geq 0}} 2x_1 - x_2$$

est non borné ◁

THÉORÈME 4.5: *(i) Le problème (PL) est soluble si et seulement il est borné.*

(ii) Si (PL) est soluble et le polyèdre P ne contient pas de droite, alors il existe un point extrémal de P qui est solution.

DÉMONSTRATION. Par définition, un problème soluble est borné. Réciproquement, soit (PL) borné. Le polyèdre P des états admissibles est de la forme $P = M(S, D)$. On a $\langle c, r \rangle \geq 0$ pour tout $d \in D$: sinon, on aurait pour $s \in S$ et un $d_0 \in D$ la convergence $\langle c, s + tr_0 \rangle \to -\infty$ si $t \to +\infty$ contredisant le caractère supposé borné du programme. Vu que la partie S est finie, il existe $s_* \in S$ tel que $\langle c, s_* \rangle = \min_{s \in S} \langle c, s \rangle$. Alors, pour des coefficients λ_s, μ_d positifs avec $\sum_s \lambda_s = 1$,

$$\langle c, \sum_s \lambda_s s + \sum_d \mu_d d \rangle \geq \sum_s \lambda_s \langle c, s \rangle + \sum_d \mu_d \langle c, d \rangle \geq \sum_s \lambda_s \langle c, s_* \rangle = \langle c, s_* \rangle$$

ce qui établit s_* comme solution du programme.

Pour la deuxième assertion, on va établir la propriété par récurrence : en dimension $n = 0$, il n'y a rien à démontrer. Si le polyèdre P ne contient pas de droite, il contient un point extrémal $\overline{x}$ d'après le théorème 4.3. Si la forme linéaire $\langle c, x \rangle$ est constante sur P, alors ce point $\overline{x}$ convient. Sinon, soit x_* une solution de (PL) : l'hyperplan $H_{c, \langle c, x_* \rangle} = \{u | \langle c, u \rangle = \langle c, x_* \rangle\}$ est un hyperplan de support de P en x_*. Ainsi le convexe $P \cap H_{c, \langle c, x_* \rangle}$, partie polyédrale convexe ne contenant pas de droite, contient d'après l'hypothèse de récurrence, un point extrémal x_{**} solution de (PL) restreint à $H_{c, \langle c, x_* \rangle}$, qui est encore un programme linéaire du type polyédral et auquel on aura appliqué l'hypothèse de récurrence. □

3. sous-entendu *inférieurement* s'agissant ici d'un problème de minimisation

6. Preuve du théorème de représentation des polyèdres

PROPOSITION 4.1: *Soit x_* un point du polyèdre $P = \{x \in \mathbb{R}^n, \langle a_i, x\rangle \geq b_i, i = 1, \ldots, m\}$. Le point x_* est extrémal si et seulement si n des inégalités du système $\langle a_i, x\rangle \geq b_i$ sont des égalités, avec la famille des a_i correspondants constituant une base de $\mathbb{R}^n$.*

DÉMONSTRATION. Soit I_* la famille d'indices i tels que $\langle a_i, x_*\rangle = b_i$ et $L = \mathrm{Vect}((a_i)_{i\in I_*})$. Si $L \neq \mathbb{R}^n$, soit d non nul dans $L^\perp$. Alors, les points du segment $[x_* - \varepsilon d, x_* + \varepsilon d]$ vérifient les égalités $\langle a_i, x\rangle = b_i$ pour $i \in I_*$ et les inégalité $\langle a_i, x\rangle > b_i$ pour $i \notin I_*$ si $\varepsilon > 0$ est assez petit, ainsi x_* n'est pas extrémal.

Réciproquement, supposons que la famille $(a_i)_{i\in I_*}$ soit une base de $\mathbb{R}^n$. Si x_* n'est pas extrémal, il existe h non nul tel que $x \pm h \in P$. Ainsi, pour $i \in I_*$, on a

$$b_i \leq \langle a_i, x \pm h\rangle = b_i \pm \langle a_i, h\rangle, \quad i \in I_*$$

soit $\pm\langle a_i, h\rangle \geq 0$, ce qui n'est possible que si $h = 0$, contredisant le caractère extrémal de x_*. □

COROLLAIRE 4.2: *L'ensemble des points extrémaux d'un polyèdre est fini.*

DÉMONSTRATION. D'après la proposition précédente, le nombre des points extrémaux est majoré par le nombre $\binom{n}{m}$ des sous-matrices carrées d'ordre n de A. □

Introduisons la notion de *partie polaire* d'un ensemble convexe

DÉFINITION 4.7: *Soit C convexe fermé contenant l'origine 0. La partie polaire* $\mathrm{Pol}(C)$ *de C est définie suivant*

$$\mathrm{Pol}(C) = \{u | \langle u, x\rangle \leq 1, x \in C\}$$

△ REMARQUE 4.5: Si C est un sous-espace linéaire, son polaire $\mathrm{Pol}(C)$ est son orthogonal $C^\perp$. ▽

LEMME 4.5: *Soit C convexe fermé contenant l'origine 0. Alors son double polaire* $\mathrm{Pol}(\mathrm{Pol}(C))$ *coïncide avec C.*

DÉMONSTRATION. Le polaire $\mathrm{Pol}(C)$ est convexe fermé et contient l'origine. De manière similaire à la bidualité, il est clair que $C \subset \mathrm{Pol}(\mathrm{Pol}(C))$. Supposons l'existence d'un $z \in \mathrm{Pol}(\mathrm{Pol}(C))$ hors de C. Il existe un hyperplan $H_{d,\langle d,z\rangle}$ séparant fortement C et $\{z\}$: $\sup_{x\in C}\langle d, x\rangle < \langle d, z\rangle$. Vu que $0 \in C$, $\langle d, z\rangle$ est strictement positif, et quitte à diviser d par un nombre positif, on peut supposer

$$\sup_{x\in C}\langle d, x\rangle < 1 < \langle d, z\rangle,$$

ce qui implique $d \in \mathrm{Pol}(C)$ et amène une contradiction : l'inégalité $1 < \langle d, z\rangle$ est incompatible avec z dans le double polaire de C. □

LEMME 4.6: *Soit C un convexe fermé de E contenant le point origine. Alors l'origine est intérieure à C si et seulement si* $\mathrm{Pol}(C)$ *est borné.*

DÉMONSTRATION. Si 0 est intérieur à C, alors il existe $\varepsilon > 0$ tel que tout h avec $\|h\| \leq \varepsilon$ soit dans C. Alors tout $x \in \mathrm{Pol}(C)$ vérifie $\varepsilon\|x\| = \langle x, \varepsilon x/\|x\|\rangle \leq 1$, *i. e.* $\|x\| \leq \varepsilon^{-1}$, ce qui exprime le fait que $\mathrm{Pol}(C)$ est borné. Réciproquement, si $\mathrm{Pol}(C)$ est borné, il existe M tel que $\|f\| \leq M$ et on a $\langle x, f\rangle \leq 1$ pour tout x avec $\|x\| \leq M^{-1}$: ainsi un tel x appartient au double polaire de C, soit C lui-même et 0 est intérieur à C. □

On peut désormais montrer le théorème 4.4 pour les polyèdres bornés.

PROPOSITION 4.2: *Un polyédre convexe borné P est exactement égal à l'enveloppe convexe de ses points extrémaux.*

DÉMONSTRATION. Le théorème 4.3 énonce que P est l'enveloppe convexe de ses points extrémaux (en nombre fini d'après le corollaire 4.2). Il s'agit de montrer que l'enveloppe convexe $\mathrm{Conv}(S)$ d'une partie S finie de points est un polyèdre : on peut supposer que E est affinement engendré par S et que $\mathrm{Conv}(S)$ est d'intérieur non vide. Au besoin après une translation, on peut supposer que $\mathrm{Conv}(S)$ contient l'origine comme point intérieur. Ainsi, le polaire $\mathrm{Pol}(\mathrm{Conv}(S))$ est borné d'après la proposition 4.6. Soit $f \in \mathrm{Pol}(\mathrm{Conv}(S))$: il est caractérisé par les inégalités $\langle f, s\rangle \leq 1, s \in S$, ce qui prouve que $C^* = \mathrm{Pol}(\mathrm{Conv}(S))$ est un polyèdre borné, donc de la forme $C^* = \mathrm{Conv}(\mathrm{Ext}(C^*))$. Il est d'intérieur non vide comme polaire d'un convexe borné : nous venons de démontrer que le polaire d'un ensemble du type $\mathrm{Conv}(S^*)$ avec 0 dans son intérieur était un polyèdre : c'est donc le cas de $C = \mathrm{Pol}(C^*)$. □

Preuve du théorème 4.4. Soit $P = \{Ax \geq b\}$ un polyèdre. Si la droite $\mathbb{R}d$ est incluse dans P, alors nécessairement $Ad = 0$ et réciproquement toute direction du noyau de A induit une droite contenue dans P.

Soit $K = \ker A$, $K^\perp$ son orthogonal, π_K la projection orthogonale sur K et $P_{K^\perp} = P \cap K^\perp$: $P_{K^\perp}$ est polyédral convexe vu que $P_{K^\perp} = \{Ax \geq b, \pi_K(x) = 0\}$ et non vide (toute projection orthogonale d'un point de P sur $K^\perp$ est contenue dans $P_{K^\perp}$), il ne contient pas de droite (vu que $A_{|K^\perp}$ est injectif) et $P = P_{K^\perp} + K$ où $K = \mathrm{Cône}(\{\mathbf{b}, -\mathbf{b}\})$ avec $\mathbf{b}$ base de K. On est donc ramené à étudier les polyèdres convexes ne contenant pas de droite et non bornés.

Nous allons établir par récurrence sur la dimension que tout polyèdre convexe P sans droite et non borné est de la forme $M(S, D)$. Commençons par montrer l'existence d'une direction (dite *récessive*) d_∞ de pôle $x \in C$: soit x_i une suite de points de P telle que $\|x_i\| \to \infty$: au choix d'une sous-suite près on peut supposer que $(x - x_i)/\|x - x_i\|$ converge vers d_∞ : les points $x + t(x - x_i)/\|x - x_i\|$, appartenant à C si $\|x - x_i\| \geq t$, convergent vers $x + td_\infty$ lorsque $i \to \infty$. On a établi donc que $P + \mathrm{Cône}(\{d_\infty\}) = P$.

Si la condition vectorielle $Ax \geq b$ est considéré comme l'ensemble des inégalités $\langle a_i, x\rangle \geq b_i$ pour $i = 1, \ldots m$, notons par P_i le polyèdre $P \cap \{\langle a_i, x\rangle = b_i\}$ et I_1 l'ensemble des indices $i \in \{1, \ldots, m\}$ lesquels P_i est non vide. Ces polyèdres convexes P_i, non vides, sont sans droite et de dimension strictement inférieure à celle de P : l'hypothèse de récurrence assure d'une représentation $P_i = M(S_i, D_i)$. On a donc

$$M(\cup_{i\in I_1} S_i, \cup_{i\in I_1} D_i \cup \{d_\infty\}) \subset M. \tag{76}$$

Il s'agit donc de montrer que cette inclusion est une égalité. Soit $x \in P$. Vu que P ne contient pas de droite, $P \cap (x + \mathbb{R}d_\infty) = x - [t_x, +\infty)d_\infty$ avec t_x fini. Le point $x - t_x d_\infty$ vérifie une des égalités $\langle a_i, x\rangle = b_i$ (sinon on pourrait prendre un $\widetilde{t_x} > t_x$ tel que $x - \widetilde{t_x} d_\infty \in P$). Ainsi, il existe $i_x \in I_1$ tel que $x - t_x d_\infty \in P_{i_x}$ et donc dans l'union du membre de droite de (76), ce qui achève la preuve du fait que tout polyèdre convexe peut se représenter suivant $P = M(S, D)$.

Il reste à montrer que tout ensemble du type $M(S, D)$ est un polyèdre convexe. Comme dans la proposition 4.2 où cela a été établi pour les $M(S, \emptyset)$, on va utiliser la polarité. Supposons 0 dans $M(S, D)$ et soit M^* le polaire de $M(S, D)$. L'ensemble des conditions $\langle x, f\rangle \leq 1$ avec $x \in M(S, D)$ est équivalente à la finitude de conditions $\langle s, f\rangle \leq 1$ avec $x \in S$ et $\langle d, f\rangle \leq 0$ avec $d \in D$: ainsi M^* est un polyèdre convexe. D'après ce qui précède, la partie M^* se représente suivant $M^* = M(S^*, D^*)$. Reprenant

pour $M^* = M(S^*, D^*)$ ce qui vient d'être fait pour $M(S, D)$, on en déduit que le polaire $M(S, D) = \mathrm{Pol}(M^*)$ est un polyèdre convexe, ce qui achève la preuve. □

Annexe A

Formes quadratiques

1. Matrices symétriques et formes quadratiques

Une matrice $A = (A_{ij})$ carrée d'ordre n est dite symétrique si $a_{ij} = a_{ji}$ ou encore $\langle Ax, y\rangle = \langle x, Ay\rangle$ pour tout vecteur x, y de $\mathbb{R}^n$.

À toute matrice symétrique réelle A d'ordre n est associée la forme quadratique Q_A sur $\mathbb{R}^n$ définie suivant

$$Q_A(x) = \sum_{i,j=1}^{n} A_{ij}x_ix_j = {}^{\mathsf{T}}XAX, \quad x = \sum_{i=1}^{n} x_ic_i, X = {}^{\mathsf{T}}[x_1x_2\dots x_n],$$

où $\mathbf{c} = (c_i)$ est la base canonique de $\mathbb{R}^n$ et X le vecteur colonne des coordonnées de x dans la base $\mathbf{c}$. Cette correspondance $A \mapsto Q_A$ est bijective : la matrice $A(Q)$ associée à la forme Q est

$$A_{ii} = Q(c_i), \quad A_{ij} = \frac{Q(c_i + c_j) - Q(c_i) - Q(c_j)}{2}$$

De manière générale, on notera $A(Q, \mathbf{b})$ la matrice de la forme quadratique Q relativement à la base $\mathbf{b}$: $A(Q, \mathbf{b})_{ii} = Q(b_i), \dots$.

On rappelle les résultats sur la diagonalisation des matrices symétriques réelles et des formes quadratiques

THÉORÈME A.1: *Soit A une matrice symétrique d'ordre n. Il existe une base orthonormée $\mathbf{b} = (b_i)_{i=1}^n$ et des scalaires $\lambda_1, \dots, \lambda_n$ tels que*

(i) cette base est une base de vecteurs propres de A : $Ab_i = \lambda_i b_i$ pour $i = 1, \dots, n$,

(ii) la forme Q_A est une pure somme de carrés pondérés

$$Q_A(x) = \sum_{i=1}^{n} \lambda_i y_i^2, \quad x = \sum_{i=1}^{n} y_i b_i. \tag{77}$$

DÉMONSTRATION. Nous allons montrer par récurrence la propriété suivante : si E est un espace isomorphe à $\mathbb{R}^n$ avec son produit scalaire et A un opérateur symétrique de E, alors E admet une base orthonormée de vecteurs propres de A. C'est clair pour $n = 1$. La sphère $S = \{\|x\|^2 = 1\}$ est un compact de $\mathbb{R}^n$: ainsi la fonction continue $P_A : x \mapsto \langle Ax, x\rangle$ atteint son minimum m_0 en x_0, ce x_0 résolvant le problème de minimisation

$$\min_{\|x\|^2=1} \langle Ax, x\rangle,$$

il existe un multiplicateur de Lagrange λ_0 tel que $\nabla P_A(x_0) = \lambda_0 \nabla(\|x\|^2)(x_0)$, soit $Ax_0 = \lambda_0 x_0$: le vecteur x_0 (de norme 1) est un vecteur propre de A. On a une décomposition orthogonale $\mathbb{R}^n = \mathbb{R}x_0 \oplus (\mathbb{R}x_0)^\perp$, dont le dernier terme $\{\langle y, x_0\rangle = 0\}$ (de dimension $n-1$) est stable par A :

$$\langle Ay, x_0\rangle = \langle y, Ax_0\rangle = \langle y, \lambda_0 x_0\rangle = \lambda_0\langle y, x_0\rangle = 0, \quad y \in (\mathbb{R}x_0)^\perp.$$

On applique alors l'hypothèse de récurrence à l'opérateur $A_{(\mathbb{R}x_0)^\perp}$, restriction de A à $(\mathbb{R}x_0)^\perp$, ce qui permet de compléter x_0 en une base orthonormée de vecteurs propres diagonalisant A.

Pour la seconde partie du théorème, on prend cette base orthonormée de vecteurs propres $\mathbf{b} = (b_i)$. Ainsi

$$\langle Ax, x\rangle = \sum_{i,j} y_i y_j \langle Ab_i, b_j\rangle = \sum_{i,j} y_i y_j \lambda_j \delta_{i,j} = \sum_i \lambda_i y_i^2$$

ce qui conclut la démonstration. □

On a le corollaire suivant

PROPOSITION A.1: *Le signe de* $\det A(Q, \mathbf{b})$ *ne dépend pas de la base* $\mathbf{b}$ *: c'est celui du produit des valeurs propres de la matrice* $A(Q, \mathbf{b})$.

DÉMONSTRATION. Soit $\mathbf{b} = (b_i), \widetilde{\mathbf{b}} = (\widetilde{b}_i)$ deux bases et P la matrice de changement de base : $\widetilde{b}_i = \sum_j P_{ij} b_j$. Ainsi

$$x = \sum_i \widetilde{x}_i \widetilde{b}_i = \sum_i \widetilde{x}_i \sum_j P_{ij} b_j = \sum_j \left(\sum_i P_{ij} \widetilde{x}_i \right) b_j$$

soit $X = {}^{\mathsf{T}}P\widetilde{X}$ et

$$Q(x) = {}^{\mathsf{T}}X A(Q, \mathbf{b}) X = {}^{\mathsf{T}}({}^{\mathsf{T}}P\widetilde{X}) A(Q, \widetilde{\mathbf{b}}) {}^{\mathsf{T}}P\widetilde{X} = {}^{\mathsf{T}}\widetilde{X} P A(Q, \widetilde{\mathbf{b}}) {}^{\mathsf{T}}P\widetilde{X}$$

soit $A(Q, \mathbf{b}) = P A(Q, \widetilde{\mathbf{b}}) {}^{\mathsf{T}}P$ et par suite $\det A(Q, \mathbf{b}) = (\det P)^2 \det A(Q, \widetilde{\mathbf{b}})$, ce qui établit la proposition. □

2. Formes définies et hyperboliques

La forme Q est dite *définie positive* (resp. *définie négative*, *semi-définie positive*, *semi-définie négative*) si $Q(x) > 0$ (resp. $< 0, \geq 0, \leq 0$) pour $x \in \mathbb{R}^n \setminus \{0\}$. Cela correspond dans (77) aux λ_i tous strictement positifs (strictement négatifs, positifs, négatifs resp.). S'il y a des λ non nuls de signes différents dans (77), la forme Q est dite *hyperbolique* : sa figure de courbes de niveau au voisinage de l'origine (point critique) présente un point selle ou un col, au contraire de celles des formes définies faites d'ellipses centrées à l'origine, cf. la figure A.1

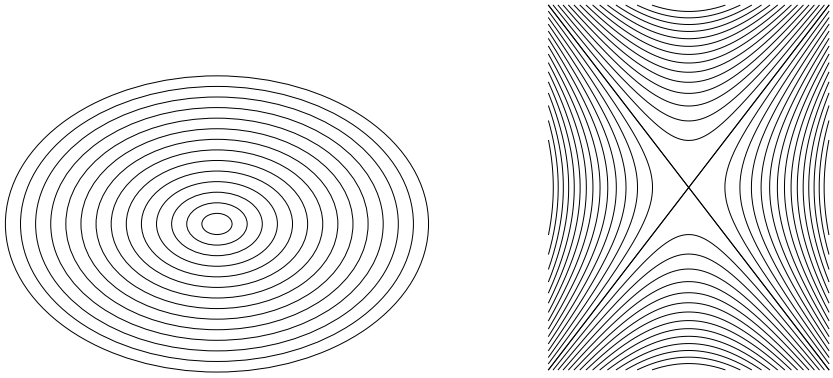


FIGURE A.1 . Les courbes de niveau en dimension 2 pour une forme définie et une forme hyperbolique.

COROLLAIRE A.1: *Soit A une matrice définie positive.*
(i) Il existe une matrice B symétrique définie positive telle que $A = B^2$.
(ii) Il existe une constante $C = C_A$ telle que

$$\langle Ax, x\rangle \geq C\|x\|^2.$$

DÉMONSTRATION. Si A est diagonale, de la forme $A_d = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$, la matrice $B_d = \mathrm{diag}(\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_n})$ et la constante $C = \min_i \alpha_i$ répondent à la question.

En général, et avec les notations du dernier paragraphe, le théorème précédent énonce l'existence d'une matrice P unitaire et d'une matrice diagonale A_d telle que $A = {}^{\mathsf{T}}PA_dP$: vu $P{}^{\mathsf{T}}P = I$, la matrice $B = {}^{\mathsf{T}}PB_dP$ vérifie $A = B^2$, alors que

$$\langle Ax, x\rangle = \langle {}^{\mathsf{T}}PA_dPx, x\rangle \geq \langle A_dPx, Px\rangle \geq \min_i(\alpha_i)\|Px\|^2 = \min_i(\alpha_i)\|x\|^2. \qquad \square$$

Un mineur m_A d'ordre r de la matrice A d'ordre (p, q) est le déterminant d'une sous-matrice M_A obtenue de A en enlevant $p - r$ lignes et $q - r$ colonnes. La matrice mineure M_A est dit *principal* si ses éléments diagonaux sont des éléments diagonaux de A. Le mineur d'ordre r est dit *principal primaire* ou *principal dominant* si la suite des éléments diagonaux de M_A sont les r premiers éléments diagonaux de A, *i. e.* il a été enlevé les $p - r$ dernières lignes et les $q - r$ dernières colonnes.

THÉORÈME A.2: *La forme quadratique Q est définie positive (négative resp.) si et seulement si les mineurs principaux primaires de sa matrice $A(Q)$ sont non nuls positifs (resp. du signe de $(-1)^r$ si r est l'ordre du mineur).*

DÉMONSTRATION. Le cas *défini négatif* découle du cas *défini positif* en considérant $-Q$.

Soit Q définie positive. Le mineur principal primaire d'ordre r est le déterminant de la forme quadratique $Q_{|V_r}$ obtenue par restriction au sous-espace V_r engendré par les r premiers vecteurs de la base canonique. La forme $q_{|V_r}$ est définie positive : une forme définie positive n'a que des valeurs propres positives non nulles, donc son déterminant est positif non nul.

Pour la réciproque, on raisonne par récurrence. D'après l'hypothèse de récurrence, la forme $Q_{|V_{n-1}}$ est définie positive : si on considère le Q-orthogonal K_n de V_{n-1} dans $V_n \simeq \mathbb{R}^n$, on a $\det A(Q) = \det A(Q_{|V_{n-1}}) \det A(Q_{|K_n})$ et par suite $\det A(Q_{|K_n}) > 0$ et $Q_{|K_n}$ définie positive. L'égalité $Q = Q_{|V_{n-1}} + Q_{|K_n}$ issue de la somme directe orthogonale $V_n = V_{n-1} \oplus K_n$ implique alors que $Q(X) > 0$ pour $X \in V_n$: Q est définie positive sur $\mathbb{R}^n$.

Autre voie Soit A avec ses n mineurs principaux dominants $d_1, \ldots, d_n$ positifs non nuls. Supposons que A ne soit pas définie positive. Ainsi, la matrice a au moins une valeur propre négative, et donc deux vu que $\det A = d_n > 0$. Soient deux vecteurs propres u, v orthogonaux associés à ces valeurs propres et la combinaison linéaire $w = \alpha u + \beta v$ dont la dernière coordonnée est nulle. Alors $\langle w, Aw\rangle = \alpha^2\langle u, Au\rangle + \beta\langle v, Av\rangle < 0$. Ainsi la restriction de la forme quadratique $w \mapsto \langle w, Aw\rangle$ au sous-espace engendré par les $n - 1$ premiers vecteurs de la base canonique n'est pas définie positive, alors que les $n - 1$ mineurs principaux dominants sont positifs non nuls, ce qui contredit l'hypothèse de récurrence $\square$

PROPOSITION A.2: *La forme quadratique Q sur $\mathbb{R}^n$ est positive (négative resp.) si et seulement si les mineurs principaux sont tous positifs ou nuls (resp. du signe de $(-1)^r$ si r est l'ordre du mineur).*

DÉMONSTRATION. Comme précédemment il suffit d'examiner le cas des formes positives. Soit $V_{i_1,\ldots,i_k}$ le sous-espace engendré par $c_{i_1}, \ldots, c_{i_k}$, vecteurs de la base canonique $\mathbf{c}$. La forme Q étant positive, sa restriction à tout $V_{i_1,\ldots,i_k}$ est positive, et donc le mineur correspondant $\det A(Q_{|V_{i_1,\ldots,i_k}})$ est positif ou nul.

Réciproquement, on introduit la forme Q_ε de matrice $A(Q_\varepsilon) = A(Q) + \varepsilon\,\mathrm{Id}$, dont le déterminant est donné par

$$\det A(Q_\varepsilon) = \sum_{k=0}^{n} \varepsilon^{n-k} \sum_{i_1<\ldots<i_k} \det A(Q_{|V_{i_1,\ldots,i_k}}).$$

le coefficient de ε^n étant 1. Ainsi $\det A(Q_\varepsilon)$ est un polynôme à coefficients positifs, positif non nul pour $\varepsilon > 0$. Les mineurs principaux dominants de $A(Q_\varepsilon)$ sont pareillement positifs non nuls : d'après la proposition précédente, la forme Q_ε est définie positive et $Q = \lim_{\varepsilon\to 0^+} Q_\varepsilon$ est positive. □

△ Remarque A.1: Pour la semi-posivité d'une matrice, il ne suffit pas de regarder les mineurs principaux dominants. Des contre exemples en témoignent : en dimension 2, avec $\begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}$ ou en dimension 3 avec $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & a \end{pmatrix}$ qui correspond à la forme quadratique $(u+v+x)^2 + (a-1)w^2$. ▽

3. Formes quadratiques sous contraintes

Dans cette section, A est une matrice carrée d'ordre n, B une matrice d'ordre (n,m), K_B le sous-espace $\ker B$ de $\mathbb{R}^n$, Q_A la forme quadratique de matrice A, $Q_{A,B}$ la forme quadratique sur K_B obtenue par restriction de Q_A et $\overline{Q}_{A,B}$ la forme quadratique de matrice $C_{A,B} = \begin{pmatrix} A & {}^{\mathsf{T}}B \\ B & 0_m \end{pmatrix}$. On suppose le mineur principal d'ordre maximal $|B_m|$ non nul.

Théorème A.3: *Si H_m est la forme quadratique hyperbolique $\begin{pmatrix} 0_m & I_m \\ I_m & 0_m \end{pmatrix}$, les formes $Q_{A,B} \oplus H_m$ et $\overline{Q}_{A,B}$ sont isomorphes.*

Démonstration. Soit W_m le sous-espace $W_m = \{{}^{\mathsf{T}}(0,Y), Y \in \mathbb{R}^m\}$ de $\mathbb{R}^{n+m}$. Identifiant $\mathbb{R}^{n+m}$ (et ses sous-espaces) à $\mathbb{R}^n \oplus \mathbb{R}^m$, on a

$$\mathbb{R}^{n+m} = \ker B \oplus \mathrm{Im}\,{}^{\mathsf{T}}B \oplus W_m.$$

Soit $\mathbf{b} = (b_1, \ldots, b_{n-m})$ une base orthonormée de K_B diagonalisant Q_B. Alors, $Q_B(b_i, b_j) = \delta_{ij}\beta_j$ et $C_{A,B}b_j = {}^{\mathsf{T}}(Ab_j, 0_m)$ est dans $W_m^\perp$, de la forme $C_{A,B}b_j = \sum_i x_{ij}b_i + {}^{\mathsf{T}}Bf_j = \beta_j b_j + C_{A,B}f_j$ avec $f_j \in W_m$: en effet

$$\beta_k\delta_{jk} = Q_{A,B}(b_j,b_k) = \langle C_{A,B}b_j, b_k\rangle = \langle \sum_i x_{ij}b_i + {}^{\mathsf{T}}Bf_j, b_k\rangle = \sum_i x_{ij}\delta_{ik} + \langle f_j, Bb_k\rangle = x_{kj}$$

Posons $F = \mathrm{Vect}(b_j - f_j, j = 1, \ldots, n-m)$. Les sous-espaces F et $\mathrm{Im}\,{}^{\mathsf{T}}B \oplus W$ sont $\overline{Q}_{A,B}$-orthogonaux : en effet $C(b_j - f_j) = \beta_j b_j \in \ker B$, par suite $C(F)$ est inclus dans $\ker B$ orthogonal à $\mathrm{Im}\,{}^{\mathsf{T}}B \oplus W$. La forme $\overline{Q}_{A,B}$ sur F est isomorphe à $Q_{A,B}$, vu que $\overline{Q}_{A,B}(b_j - f_j, b_i - f_i) = \langle (b_i - f_i), C_{A,B}(b_j - f_j)\rangle = \beta_j\delta_{i,j}$.

Si $\mathbf{d}$ est une base de $\mathrm{Im}\,{}^{\mathsf{T}}B$, la forme $\overline{Q}_{A,B}$ en restriction à $\mathrm{Im}\,{}^{\mathsf{T}}B \oplus W$ a la forme $\begin{pmatrix} \widetilde{A} & I_m \\ I_m & 0 \end{pmatrix}$, équivalente à la forme hyperbolique H_m vu qu'elle est non dégénérée identiquement nulle sur W_m. □

Corollaire A.2: *Le déterminant $\det Q_{A,B}$ a même signe que $(-1)^m \det \overline{Q}_{A,B}$.*

Nous allons préciser ce corollaire pour en déduire un critère analogue à celui du théorème **??** pour tester le caractère défini de $Q_{A,B}$ en terme des signes de $n-m$ déterminants

LEMME A.1: *La forme $Q_{A,B}$ est définie positive si et seulement si il existe C_0 tel que la forme $Q_A(x)+C(Bx)^2$ est définie positive sur $\mathbb{R}^n$ pour tout $C>C_0$.*

DÉMONSTRATION. Supposons $Q_{A,B}$ définie positive. Si S est la sphère $\{\|x\|=1\}$, la fonction $x\mapsto Q_{A,B}(x)/|Bx|^2$ tend vers $+\infty$ au bord de $S\setminus\{x:Bx=0\}$: elle admet donc un minimum sur cet ensemble, soit m_0. Ainsi, pour $m>\max(0,-m_0)$ et x non nul

$$Q_{A,B}(x)+m\|Bx\|^2\geq\max((m+m_0)\|Bx\|^2,\quad Q_{A,B}(x))>0$$

et donc $Q_{A,B}+m\|B\cdot\|^2$ est définie positive. □

LEMME A.2: *Le déterminant $|A+\lambda^{\mathsf{T}}BB|$ est un polynôme en λ de degré au plus m et dont le terme (éventuellement nul) de plus haut de degré est $(-1)^m\begin{vmatrix}A & {}^{\mathsf{T}}B\\ B & 0_m\end{vmatrix}$.*

DÉMONSTRATION. De l'identité

$$\begin{pmatrix}A & \lambda^{\mathsf{T}}B\\ B & -I_m\end{pmatrix}\begin{pmatrix}I_n & 0_{nm}\\ B & I_m\end{pmatrix}=\begin{pmatrix}A+\lambda^{\mathsf{T}}BB & \lambda^{\mathsf{T}}B\\ 0_{mn} & -I_m\end{pmatrix}$$

découle celle sur les déterminants

$$\begin{vmatrix}A & \lambda^{\mathsf{T}}B\\ B & -I_m\end{vmatrix}=(-1)^m\left|A+\lambda^{\mathsf{T}}BB\right|.$$

Pour obtenir le terme en λ^m dans le développement du déterminant de gauche, on considère le seul coefficient λ dans chacune des m dernières colonnes : sans affecter la valeur de ce terme, on peut donc remplacer la matrice I_m par la matrice nulle 0_m et on conclut en remarquant $\begin{vmatrix}A & \lambda^{\mathsf{T}}B\\ B & 0_m\end{vmatrix}=\lambda^m\begin{vmatrix}A & {}^{\mathsf{T}}B\\ B & 0_m\end{vmatrix}$. □

THÉORÈME A.4: *Soit A carrée d'ordre n symétrique, B d'ordre (m,n) avec le mineur principal d'ordre maximal $|B_m|$ non nul. La forme Q_{AB} induite par A sur $K_B=\ker B$ est définie positive si et seulement si $(-1)^m\begin{vmatrix}A_r & {}^{\mathsf{T}}B_{rm}\\ B_{rm} & 0\end{vmatrix}>0$ pour $r=m+1,\dots n$.*

DÉMONSTRATION. Supposons Q_{AB} définie positive. Le déterminant $\begin{vmatrix}A & {}^{\mathsf{T}}B\\ B & 0\end{vmatrix}$ est non nul, puisque le système $Ax+{}^{\mathsf{T}}By=0, Bx=0$ n'a pas de solution non triviale : une solution (x,y) vérifie $Q_A(x)+\langle x,{}^{\mathsf{T}}By\rangle=0$ et donc $Q_{AB}(x)=0$, puis $x=0$ (Q_{AB} est définie) et $y=0$ puisque B_m est inversible. D'après le théorème **??** et le lemme A.1, on a pour λ suffisamment grand, $|A+\lambda^{\mathsf{T}}BB|>0$ et par suite, grâce au lemme A.2, $(-1)^m\begin{vmatrix}A & {}^{\mathsf{T}}B\\ B & 0\end{vmatrix}>0$. Cet argument vaut pareillement pour $r=n-1,\dots,m+1$ en considérant les restrictions $(Q_A)_{|V_r}=Q_{A_r}$ et $B_{|V_r}$.

Réciproquement, on va montrer que le coefficient du terme de plus haut degré dans $|A_r+\lambda^{\mathsf{T}}B_{rm}B_{rm}|$ est positif pour $r=1,\dots,n$, impliquant que $A+\lambda^{\mathsf{T}}BB$ est défini positif pour λ assez grand et Q_{AB} défini positif d'après le lemme A.1. Cette positivité vaut par hypothèse pour $r>m$. Si $r\leq m$, comme dans la preuve du lemme A.2, écrivons

$$(-1)^m|A_r+\lambda^{\mathsf{T}}B_{rm}B_{rm}|=\begin{vmatrix}A & \lambda^{\mathsf{T}}B_{rm}\\ B_{rm} & -I_m\end{vmatrix}.$$

Dans un développement par rapport aux lignes, le terme λr provient des r premières lignes, indépendamment des autres coefficients de ces lignes. Ainsi le coefficient de λ^r est $\begin{vmatrix} 0 & {}^{\mathsf{T}}B_{rm} \\ B_{rm} & -I_m \end{vmatrix}$. La matrice $\begin{vmatrix} 0 & {}^{\mathsf{T}}B_{rm} \\ B_{rm} & -I_m \end{vmatrix}$ est inversible : comme précédemment, il suffit de vérifier que le système ${}^{\mathsf{T}}B_{rm}y = 0, B_{rm}x - y = 0$ n'a pas de solution non triviale. Une solution (x, y) vérifie $\langle y, B_{rm}x\rangle - \langle y, y\rangle = 0$, d'où $y = 0$ (vu que $B_{rm}y = 0$)), puis $B_{rm}x = 0$ qui donne $x = 0$ (la non nullité de $|B_m|$ assure de l'existence d'un mineur principal d'ordre r non nul). En invoquant les lemmes A.1 et A.2, on obtient la positivité de $(-1)^m \begin{vmatrix} 0 & {}^{\mathsf{T}}B_{rm} \\ B_{rm} & I_m \end{vmatrix}$. □

Corollaire A.3: *Avec les mêmes notations que dans le théorème A.4, la forme Q_{AB} est définie positive si et seulement si chaque mineur principal dominant de taille au moins $m+1$ de $\begin{pmatrix} 0 & B \\ {}^{\mathsf{T}}B & A \end{pmatrix}$ est du signe de $(-1)^m$.*

En considérant $-q$, on obtient un analogue du théorème A.4 pour des formes définies négatives.

Théorème A.5: *Soit A carrée d'ordre n symétrique, B d'ordre (m, n) avec le mineur principal d'ordre maximal $|B_m|$ non nul. La forme Q_{AB} induite par A sur $K_B = \ker B$ est définie négative si et seulement si $(-1)^r \begin{vmatrix} A_r & {}^{\mathsf{T}}B_{rm} \\ B_{rm} & 0 \end{vmatrix} > 0$ pour $r = m+1, \dots n$.*

△ Remarque A.2: Si A est la matrice diagonale $\mathrm{diag}(\alpha_1, \dots, \alpha_n)$ et B est défini suivant $B(x_1, \dots, x_n) = (x_1, \dots, x_m)$, la matrice $D_{A,B}$ est donnée par

$$D_{A,B} = \begin{pmatrix} \alpha_1 & 0 & \dots & \dots & \dots & 0 & 1 & & \\ 0 & \alpha_2 & & & & \vdots & & \ddots & \\ \vdots & & \ddots & & & \vdots & & & 1 \\ \vdots & & & \alpha_{m+1} & & \vdots & 0 & \dots & 0 \\ \vdots & & & & \ddots & 0 & \vdots & & \vdots \\ 0 & \dots & \dots & \dots & 0 & \alpha_n & 0 & \dots & 0 \\ 1 & & & 0 & \dots & 0 & 0 & \dots & 0 \\ & \ddots & & \vdots & & \vdots & \vdots & & \vdots \\ & & 1 & 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix} \tag{78}$$

Pour calculer son déterminant, on remplace par 0 les $\alpha_1, \dots, \alpha_m$ en soustrayant par des multiples convenables des m dernières lignes : le développement [1] suivant les m dernières

1. La matrice (78) est symétrique : sa forme quadratique associée est $\sum_{i=1}^{m} x_i x_{n+i} + \sum_{i=m+1}^{n} \alpha_i x_i^2$, de déterminant $(-1)^m \alpha_{m+1} \dots \alpha_n$.

lignes fournit

$$|D_{A,B}| = (-1)^{m(n-m)} \begin{vmatrix} 0 & \dots & \dots & \dots & \dots & 0 & 1 & & \\ \vdots & & & & & \vdots & & \ddots & \\ 0 & \dots & \dots & \dots & \dots & 0 & & & 1 \\ \alpha_{m+1} & 0 & \dots & \dots & \dots & 0 & 0 & \dots & 0 \\ 0 & \alpha_{m+2} & & & & \vdots & \vdots & & \vdots \\ \vdots & & \ddots & & & \vdots & \vdots & & \vdots \\ \vdots & & & \ddots & & \vdots & \vdots & & \vdots \\ \vdots & & & & \ddots & 0 & \vdots & & \vdots \\ 0 & \dots & \dots & \dots & 0 & \alpha_n & 0 & \dots & 0 \end{vmatrix}$$

qu'on développe relativement aux m dernières colonnes pour obtenir

$$(-1)^{m(n-m)}(-1)^{m(n-m+1)} \begin{vmatrix} \alpha_{m+1} & 0 & \dots & \dots & \dots & 0 \\ 0 & \alpha_{m+2} & & & & \vdots \\ \vdots & & \ddots & & & \vdots \\ \vdots & & & \ddots & & 0 \\ 0 & \dots & \dots & & 0 & \alpha_n \end{vmatrix} = (-1)^m \alpha_{m+1}\alpha_{m+2}\dots\alpha_n$$

Un calcul analogue donne les déterminants

$$d_{m+r} = \begin{vmatrix} (D_{A,B})_r & {}^{\mathsf{T}}B_{mr} \\ B_{mr} & 0_m \end{vmatrix} = (-1)^m \alpha_{m+1}\alpha_{m+2}\dots\alpha_r, \quad r \in [m+1, n].$$

La forme quadratique $Q_{A,B}$ restreinte à $\ker B = \{x_1 = \dots = x_m = 0\}$ est donnée par

$$Q_{A,B}(x) = \alpha_{m+1}x_{m+1}^2 + \dots + \alpha_n x_n^2, \quad x = (0, \dots, 0, x_{m+1}, \dots, x_n) \in \ker B$$

et les $n-m$ conditions de positivité $\alpha_{m+1}, \dots, \alpha_n \geq 0$ sont bien équivalentes à la positivité des $(-1)^m d_{m+r}$ pour $r = m+1, \dots, n$. ▽

Annexe A

Maximum de vraisemblance

Commençons par l'énoncé général qui précise comment les maxima de vraisemblance sont des estimateurs tout à fait intéressants

THÉORÈME A.1: *Soient $(X_k)_{k\geq 0}$ des variables aléatoires indépendantes de même distribution $dP(\theta) = p(x|\theta)dx$ vérifiant un lot (convenable...) d'hypothèses techniques. Alors il existe une suite $\theta_n = \theta(X_1, \ldots, X_n)$ de maxima locaux de la fonction de vraisemblance*

$$p_n(X_1, \ldots, X_n|\theta) = \prod_{k=1}^{n} p(X_k, \theta)$$

qui vérifie

$$\theta_n \xrightarrow{P} \theta,$$

et même

$$\sqrt{n}(\theta_n - \theta) \xrightarrow{L} \mathcal{N}(0, I(\theta)^{-1})$$

où $I(\theta) = E(\partial_\theta \log p(x|\theta))^2$.

Ce théorème présuppose qu'on puisse calculer des maxima de vraisemblance. Le calcul explicite en est plutôt rare, malgré les quelques exemples ci-dessous : l'algorithme EM permet d'avoir des approximations de tels maxima, et donc *in fine* des paramètres.

▷ EXEMPLES A.1:

??.3 Soit $(X_k)_{k=1}^n$ un échantillon de données indépendantes de loi normale $\mathcal{N}(\mu, v)$. La vraisemblance

$$\prod_{k=1}^{n} \frac{\mathrm{e}^{-(X_k-\mu)^2/(2v)}}{\sqrt{2\pi v}}$$

a même maxima que la fonction de log-vraisemblance (à des constantes additives près)

$$\sum_{k=1}^{n} \left[-\frac{(X_k - \mu)^2}{2v} \right] - \frac{n}{2} \log v$$

soit,

$$\mu_{\max} = n^{-1} \sum_{k=1}^{n} X_k, \quad v_{\max} = n^{-1} \sum_{k=1}^{n} (X_k - \mu_{\max})^2.$$

La loi des grands nombres, et le théorème central limite, indiquent bien la convergence décrite dans le théorème.

??.4 Pour une loi de Poisson $P(\lambda)$, on a comme vraisemblance pour les observations (X_k)

$$\prod_{k=1}^{n} \frac{\lambda^{X_k} \mathrm{e}^{-\lambda}}{X_k!}$$

avec log-vraisemblance (à des termes constants près)

$$\sum_{k=1}^{n} x_k \log \lambda - n\lambda$$

de maxima

$$\lambda_{\max} = n^{-1} \sum_{k=1}^{n} X_k.$$

??.5 La vraisemblance de la loi multinomiale $M(n; p_1, \ldots, p_L)$, à des constantes indépendantes des paramètres $\mathbf{p} = (p_1, \ldots, p_L)$ près,

$$V_X(\mathbf{p}) = \prod_{\ell=1}^{L} p_k^{X_\ell}, \quad \mathbf{p} \in \{p_\ell \geq 0, p_1 + \ldots + p_L = 1\}$$

est une fonction de Cobb-Douglas $U_{CD}^{\alpha}(\mathbf{u} = (u_\ell)) = \prod_{\ell=1}^{L} u_\ell^{\alpha_\ell}$, fonction modélisant des productions en économétrie : si les α_ℓ sont positifs, elle est quasi-concave, concave si et seulement si $\sum_\ell \alpha_\ell \in [0, 1]$. La condition d'Euler-Lagrange du premier ordre donne comme point critique $\mathbf{p}_X^* = (X_\ell / \sum_{\ell=1}^{L} X_\ell) = (X_\ell / N)$ vu que $\sum_{\ell=1}^{L} X_\ell = N$, unique maximum de V_X sur le compact $\{p_k \geq 0, p_1 + \ldots + p_L = 1\}$. Pour un échantillon $\mathcal{X} = (X_1, \ldots, X_M)$ de données multinomiales indépendantes de même loi $M(N; p_1, \ldots, p_L)$, la vraisemblance $V_{\mathcal{X}} = \prod_{m=1}^{M} V_{X_m}$ est maximale au point $\mathbf{p}_{\mathcal{X}}^* = (\overline{\mathcal{X}_k} / \sum_{\ell=1}^{L} \overline{\mathcal{X}_\ell})$ avec $\overline{\mathcal{X}_k} = \sum_{m=1}^{M} X_{k,m}/M$ et $\sum_{\ell=1}^{L} \overline{\mathcal{X}_k} = \sum_{m=1}^{M} \sum_{\ell=1}^{L} X_{k,m}/M = N$, soit $p_L^* = (\overline{\mathcal{X}_k}/N)$. ◁

Bibliographie

[1] X. Antoine, P. Dreyfuss, Y. Privat, Introduction à l'optimisation : aspects théoriques, numériques et algorithmes, 2007

[2] D. H. Ballard, C.O. Jelinek, R. Schinzinger, *An algorithm for the solution of constrained generalized polynomial programming problems*, Computer J., 261-266 **17**, 1974.

[3] Minimizing Finite Sums with the Stochastic Average Gradient, F. Bach, N. Le Roux, M. Schmidt, arxiv:1309.2388, mai 2016.

[4] G. Barles, Optimisation dans R^n (et ailleurs ?) : quelques résultats de base, dernier accès fév. 2012.

[5] L. Bottou, F.E. Curtis, J. Nocedal, Optimization Methods for Large-Scale Machine Learning, arxiv:1606.04838, 2016.

[6] L. Bottou, Large-scale machine learning with stochastic gradient descent, In Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010), 177–187, Paris, France, Springer,

[7] A. Boyd, L. Vandenberghe, Convex optimization, Cambridge University Press, 2004.

[8] Carpentier, Stochastic optimisation, numerical methods perso.ensta-paristech.fr/ pcarpent/MNOS/

[9] A. Cauchy, Méthode générale pour la résolution des systèmes d'équations simultanées, C. R., t. XXV, p. 536

[10] J.-C. Culioli, *Introduction à l'optimisation*, Ellipses, 1994.

[11] Dempster, N. M. Laird, D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Royal Stat. Soc., B **39**#1 (1-38), 1977.

[12] J. E. Dennis, R. B. Schnabel, Numerical methods for unconstrained optimization and nonlinear equations, SIAM, 1996.

[13] J. C. Dodu, M. Goursat, A. Hertz, J.-P. Quadrat, M. Viot, Méthodes de gradient stochastique pour l'optimisation des investissements dans un réseau électrique. EDF, Bulletin de la Direction des études et recherches Série C, Mathématiques, Informatique, n° 2, p.133-167, 1981.

[14] C. C. Y. Dorea, *Expected number of steps of a random optimization method*, J. Optimization th., 3 (1983) 165-171.

[15] G. Ewald, Combinatorial convexity and algbraic geometry Graduate texts in math. 168, Springer 1996

[16] J. Gentle, The EM method, www.scs.gmu.edu/ jgentle/csi991/03f/EMIntro.pdf

[17] A. Guimier, Modélisation d'algorithmes d'optimisation à stratégie aléatoire. Calcolo 23 (1986), 21–34.

[18] A. Iouditski, Convex optimization

[19] A. Juditsky, G. Lan, A. Nemirovski, A. Shapiro, Stochastic approximation approach to stochastic programming, SIAM J. Optim., 19(4), 1574–1609, 2009

[20] W. Karush. Minima of functions of several variables with inequalities as side constraints. M.Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois, 1939.

[21] C. T. Kelley. Iterative Methods for Optimization. SIAM, 1999.

[22] H. W. Kuhn, A. W. Tucker. Nonlinear programming. Proceedings of 2nd Berkeley Symposium. Berkeley : University of California Press. pp. 481–492, 1951

[23] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the nelder-mead simplex algorithm in low dimensions. SIAM Journal on Optimization, 9(1) :112–147, 1998.

[24] K. Lange, The EM Algorithm, www.ipam.ucla.edu/publications/inv2003/inv2003_3986.pdf

[25] D. Luenberger, Linear and nonlinear programming.

[26] H. Moulin, F. Fogelman-Soulié, Convexité dans les mathématiques de la décision, Hermann, 1979.

[27] J. A. Nelder and R. Mead. A simplex method for function minimization. Computer Journal, 7 (1965), 308–313.

[28] S. K. Ng, T. Krishman, G. J. McLachlan, *The EM algorithm*, 1-34.

[29] Y. Nesterov, Introductory lectures on convex optimization : a basic course, Kluwer, 2004

[30] J. R. Norris, Markov chains, Cambridge, 1997.

[31] J. Nocedal, S. Wright, Numerical optimization, Springer, 1999.

[32] Y. Pawitan, In all likelihood : statistical modelling and inference using likelihood, Clarendon Press, Oxford, 2001.

[33] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, Numerical recipes : the art of scientific computing, Cambridge University Press, 2007. (cf. www.nr.com/)

[34] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Stat., 22 (1951), pp. 400–407.

[35] S. M. Ross, Applied probability models with optimisation applications, Dover, 1992.

[36] M. S. Sarma, *On the convergence of the Baba and Dorea random optimization methods*, J. Optimization Th. Appl. 337-343 **66**#2, 1990.

[37] F. J. Solis, R. J.-B. Wets, Minimization by random search techniques, Math. Operations Research, 6 (1981), 19-30.

[38] J. Spall, Introduction to stochastic search and optimisation, Wiley, 2003.

[39] R. Tibshirani, Subgradient Method, consulté le 20 déc. 2016.

[40] C. Villani, Optimal transport, Springer, 2009.

[41] A. Zhigljavsky, A. Zilinskas, Stochastic global optimization, Springer, 2008.

Index

Index général

Index des noms