MVA, Projet optimisation : Inégalité linéaires

Valentin DE BORTOLI, Louis THIRY

30 janvier 2017

Table des matières

1 Présentation du problème

On s'intéresse aux problèmes d'égalité et d'inégalité aux moindres carrés.

Définition 1. *Soient $A \in \mathcal{M}_{n,m}(\mathbb{R})$ avec $(n,m) \in \mathbb{N}^2$ et $b \in \mathbb{R}^n$. On appelle problème d'égalité aux moindres carrés et on note $\mathcal{P}_=(A,b)$ le probleme de minimisation :*

$$\mathcal{P}_=(A,b): \quad \hat{x} = \underset{x \in \mathbb{R}^m}{argmin} \, ||Ax - b||^2 \tag{1}$$

Résoudre le problème d'égalité aux moindres carrés permet de trouver la meilleure solution possible au système linéaire $Ax = b$:

— lorsqu'il existe une solution à ce système linéaire, les deux problèmes sont équivalents
— si b n'est pas dans l'images de A , où bien lorsque l'on a plus d'équations que d'inconnues ($n > m$) et est inconsistant, il n'y a aucune solutions. Dans ce cas, une solution x_0 du problème d'égalité aux moindres carrés est la meilleure approximations telle que $Ax_0 = b$ au sens des moindres carrés. C'est une solution approximative au système linéaire $Ax = b$

Définition 2. *On appelle problème d'inégalité aux moindres carrés et on note $\mathscr{P}_{\leq}(A,b)$ le problème de minimisation :*

$$\mathscr{P}_{\leq}(A,b): \quad \hat{x} = \underset{x\in\mathbb{R}^m}{argmin}\, ||(Ax-b)_+||^2 \tag{2}$$

De même que pour le cas d'égalité, résoudre le problème d'inégalité aux moindres carrés permet de trouver la meilleure solution possible au système d'inégalité linéaires $Ax \leq b$:

— lorsqu'il existe une solution à ce système d'inégalité, les deux problèmes sont équivalents

— Dans le cas où ils n'y a pas de solutions aux système d'inégalité, c'est une solution approximative.

Le problème d'égalité est bien compris : on connait la forme des solutions. Ce problème est traité dans la première partie du rapport.

Pour le problème d'inégalité, on a simplement un résultat d'existence.

Le but de ce rapport est de proposer un algorithme pour trouver une solution au problème d'inégalité aux moindres carrés $\mathscr{P}_{\leq}(A,b)$.

2 Équations linéaires

Les équations du type $Ax = b$ avec $A \in \mathscr{M}_{n,m}(\mathbb{R})$ et $b \in \mathbb{R}^n$ ont été largement étudiées. Des conditions nécessaires et suffisantes simples ont été exprimées et des méthodes pour trouver ces solutions lorsqu'elles existent ont été développées (méthode du pivot de Gauss, décomposition LU...). Lorsqu'aucune solution n'existe, on considère le problème $\mathscr{P}_{=}(A,b)$, 1. Dans cette section on traite le problème des égalités linéaires aux moindres carrés.

2.1 Rappels sur la pseudo-inverse

On introduit la notion de matrice pseudo inverse de Moore-Penrose [5] qui joue un rôle particulier dans l'étude des problèmes aux moindres carrés.

Théorème 1. *Soit $A \in \mathscr{M}_{n,m}(\mathbb{R})$ avec $(n,m) \in \mathbb{N}^2$. Il existe (M,N) deux matrices orthogonales de $\mathscr{M}_m(\mathbb{R})$, respectivement $\mathscr{M}_n(\mathbb{R})$ et $\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$ avec D diagonale strictement positive telles que $A = N\Sigma M$.*

Démonstration : On considère la matrice $A^TA \in \mathscr{M}_m(\mathbb{R})$. Celle-ci est symétrique, semi-définie positive et on peut appliquer le théorème spectral. On a donc une matrice orthogonale P qui vérifie $P^TA^TAP = \begin{pmatrix} D^2 & 0 \\ 0 & 0 \end{pmatrix}$ avec D^2 diagonale strictement positive. Il est à noter que D est une matrice de $\mathscr{M}_r(\mathbb{R})$ où r est le rang de la matrice A (et donc plus petit que m et n). On note $P = [P_1\ P_2]$ où $P_1 \in \mathscr{M}_{m,r}(\mathbb{R})$ et $P_2 \in \mathscr{M}_{m,m-r}(\mathbb{R})$. On a alors :

$$\begin{cases} P_1^TA^TAP_1 = D^2 \\ P_1^TA^TAP_2 = 0 \\ P_2^TA^TAP_1 = 0 \\ P_2^TA^TAP_2 = 0 \end{cases} \tag{3}$$

Posons $Q_1 = AP_1D^{-1}$. On a bien $Q_1^TQ_1 = D^{-1}P_1^TA^TAP_1D^{-1} = Id_r$. On complète Q_1 en une matrice orthogonale de $\mathscr{M}_n(\mathbb{R})$ et on a $Q^TAP = \begin{pmatrix} DP_1^TA^TAP_1 & DP_1^TA^TAP_2 \\ Q_2^TAP_1 & Q_2^TA^TAP_2 \end{pmatrix}$. Or $P_1^TA^TAP_2 = 0$ et $Q_2^TQ_1 = DQ_2^TAP_1 = 0$ et puisque $P_2^TA^TAP_2 = 0$, $AP_2 = 0$. Donc on trouve bien $Q^TAP = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$.

On peut alors définir la pseudo-inverse.

Définition 3 (Pseudo-inverse de Moore-Penrose, 1920)**.** *On appelle pseudo-inverse de Moore-Penrose de A et on note A^+ la matrice $A^+ = M^T\begin{pmatrix} D^{-1} & 0 \\ 0 & 0 \end{pmatrix}N^T \in \mathscr{M}_{m,n}(\mathbb{R})$ (où M,N sont les matrices orthogonales qui apparaissent dans la décomposition en valeurs singulières).*

Proposition 1. *A^+ vérifie les propriétés suivantes :*

— $A^+AA^+ = A^+$

— $AA^+A = A$

— AA^+ *est symétrique*

— A^+A *est symétrique*

On peut alors montrer la propriété suivante :

Proposition 2. *Il existe une unique matrice qui vérifie ces quatre propriétés, la pseudo-inverse de Moore-Penrose.*

Démonstration : Soient (B,C) deux matrices satisfaisant les propriétés énoncées. On a alors :

$$\begin{aligned} AB &= ACAB \\ &= (AC)^T(AB)^T \\ &= (ABAC)^T \\ &= ABAC \\ &= AC \end{aligned} \tag{4}$$

De la même manière, $BA = CA$. Ensuite on a :

$$\begin{aligned} C &= CAC \\ &= CAB \\ &= BAB \\ &= B \end{aligned} \tag{5}$$

On établit maintenant une proposition concernant des décompositions orthogonales de l'espace.

Proposition 3. *On a les égalité suivantes :*

— $kerA^T \overset{\perp}{\oplus} ImA = \mathbb{R}^m$

— $kerA \overset{\perp}{\oplus} ImA^T = \mathbb{R}^n$

— $kerA^+ \overset{\perp}{\oplus} ImA = \mathbb{R}^m$ *et* AA^+ *projecteur orthogonal sur* ImA *de noyau* $kerA^+$

— $kerA \overset{\perp}{\oplus} ImA^+ = \mathbb{R}^n$ *et* A^+A *projecteur orthogonal sur* ImA^+ *de noyau* $kerA$

Démonstration : concernant les deux premiers points, on vérifie rapidement l'orthogonalité de ces deux espaces. Puis on peut conclure sur le fait que leur somme directe correspond bien à tout l'espace en utilisant un argument de dimension (notamment $\text{rg}A = \text{rg}A^T$).
Pour les deux autres points on ne démontre que le premier (le second s'effectue exactement de la même manière). $(AA^+)^2 = AA^+$ donc c'est un projecteur et la condition de symétrie assure son orthogonalité. $\ker A^+$ est trivialement inclus dans son noyau. Soit un élément x tel que $AA^+x = 0$, alors $A^+AA^+x = A^+x = 0$ donc c'est un élément du noyau de A^+. On en déduit l'égalité des noyaux. De plus $\text{Im}A \subset \text{Im}AA^+$. Soit $y = Ax$. Alors $y = AA^+Ax$ donc $y \in \text{Im}AA^+$. On a donc l'égalité des images. On peut donc conclure.

Proposition 4. *Une conséquence de cette proposition est* $kerA^T = kerA^+$ *et* $ImA^T = ImA^+$.

Remarque : cette identification des supplémentaires est vraie seulement parce que les sommes directes sont orthogonales.

2.2 Solutions au problème d'égalité aux moindres carrés

Revenons au problème des moindres carrés.

Proposition 5. *Un élément* x *est solution du problème* $\mathscr{P}_=(A,b)$ *si et seulement si* $A^T(Ax-b) = 0$.

Démonstration : Le problème d'égalité aux moindre carrés consiste en la minimisation de la fonction F : $x \mapsto ||Ax-b||^2$. F étant convexe, ses minimums sont ses points stationnaires, qui vérifient $\nabla F = 0$, i.e. $A^T(Ax-b) = 0$.

Proposition 6. *L'ensemble des solutions du problème* $\mathscr{P}_=(A,b)$ *est l'ensemble* $A^+b + kerA$.

Démonstration : Une solution x du problème vérifie $A^T(Ax-b)=0$ donc est solution du système linéraire $A^TAx=A^Tb$. L'ensemble des solutions de ce système linéaire est $x_0+\ker A^TA$ ou x_0 est une solution particulière ce ce système.

— A^+b est une solution particulière de ce système :

$$\begin{aligned} A^TAA^+ &= A^T(AA^+)^T \\ &= (AA^+A)^T \\ &= A^T \\ \implies A^TA(A^+b) &= A^Tb \end{aligned} \tag{6}$$

— $\ker A = \ker A^TA$.

$$\begin{aligned} &\ker A \subset \ker A^TA \\ &\ker A^TA \subset \ker A : \quad x \in \ker A^TA \implies A^TAx=0 \implies x^TA^TAx=0=||Ax||^2 \implies Ax=0 \implies x\in\ker A \end{aligned} \tag{7}$$

Donc l'ensemble des solutions du problème $\mathscr{P}_=(A,b)$ est $A^+b+\ker A$.

Proposition 7. *A^+b est la solution de $\mathscr{P}_=(A,b)$ de norme minimale.*

Démonstration : Toute solutions x_0 de $\mathscr{P}_=(A,b)$ s'écrit $x_0=A^+b+y$, $y\in\ker A$. Comme $\ker A\perp \mathrm{Im}A^+$, par le théorème de Pythagore $\|x_0\|^2=\|A^+b\|^2+\|y\|^2$ donc $\|x_0\|\leq\|A^+b\|$.

Le problème des égalités linéaires étant traité on s'intéresse maintenant au cas des inégalités linéaires.

3 Inégalités linéaires : le problème de la consistance

3.1 Introduction et quelques définitions

On s'intéresse désormais à 2, c'est-à-dire au problème $\mathscr{P}_\leq(A,b)$. Si la théorie est bien comprise dans le cas des égalités linéaires (théorie de la dimension) et que les solutions sont bien décrites (sous-espaces affines) ce n'est pas le cas pour les inégalités linéaires. En effet, l'ensemble des solutions (lorsqu'il est non vide) forme un polyèdre convexe, plus difficile à manipuler qu'un simple espace vectoriel. Dans cette section on donne des éléments de réponse concernant l'unicité et l'existence de solutions via les théorèmes de l'alternative. On présente également deux méthodes pour trouver des solutions dans le cas consistant :

— l'élimination de Fourier-Motzkin
— l'algorithme du simplexe.

Définition 4 (Consistance)**.** *Soient $A\in\mathscr{M}_{n\times m}(\mathbb{R})$ et $b\in\mathbb{R}^n$ On dit que le système d'inéquations linéaires $Ax\leq b$ est dit consistant si $\exists x\in\mathbb{R}^m$ tel que $Ax\leq b$.*

Remarques :
— Dans le cas consistant, la résolution de $Ax\leq b$ est équivalente à la résolution de $\mathscr{P}_\leq(A,b)$
— Trouver une solution au système $Ax\leq b$ n'est pas un problème facile bien que son expression soit très simple.

Définition 5 (Forte consistance)**.** *On dit qu'un système d'inéquations linéaires (A,b) est fortement consistant si $\exists x\in\mathbb{R}^m$ tel que $Ax<b$.*

Proposition 8. *Soit (A,b) un système fortement consistant alors (A,b) admet une infinité de solutions.*

Démonstration : On considère $(1+\epsilon)x$ avec ϵ assez petit pour que les inégalité linéaires soient toujours vérifiées (vrai par continuité). Puisque l'ensemble des solutions est convexe et qu'on a deux solutions distinctes, celui-ci est infini.

3.2 Consitance et théorèmes de l'alternative

Définition 6 (Homogène)**.** *Un système d'inégalité $Ax\leq b$ est dit homogène si $b=0$.*

Nous allons énoncer des théorèmes d'alternative qui nous permettront d'établir une condition nécessaire est suffisante à la consistance d'un système d'inégalité homogène.

Définition 7 (Théorème d'alternative)**.** *Un théorème d'alternative consiste en un ensemble de deux assertions P et Q telles que l'une est fausse quand l'autre est vraie :*

$$P \implies \neg Q \text{ and } \neg P \implies Q$$

Théorème 2 (Gordan, 1873)**.** *On note* $(a_i)_{i\in[\![1,m]\!]} \in (\mathbb{R}^m)^n$ *les lignes de A* $(\mathscr{P})$ $\exists x \in \mathbb{R}^m,\ \forall i \in [\![1,n]\!],\ \langle a_i, x\rangle < 0$ *est équivalent à la négation de la proposition suivante :* $(\mathscr{Q})$ $\exists (\lambda_i)_{i\in[\![1,n]\!]} \in {\mathbb{R}_+}^n \setminus \{0\},\ \sum_{i=1}^{n} \lambda_i a_i = 0$

Remarques :
- On peut donner une interprétation géométrique de ce théorème : aucun des a_i ne doit être dans le cône engendré par les vecteurs $(-a_j)_{j\in[\![1,n]\!]\setminus\{i\}}$. Cette remarque géométrique peut être comprise en termes de condition pour que les demi-hyperplans portés par les a_i s'intersectent.
- Ce théorème se rapproche d'une condition nécessaire et suffisante pour que le système d'inégalité $Ax \leq 0$ admette une solution consistante. Malheureusement les inégalité sont strictes.

Deux démonstrations du théorème de Gordan On va citer ici deux démonstrations du théorème de Gordan. La première est totalement géométrique [2], la seconde utilise à profit l'existence de solutions au sens des moindres carrés. Il est à noter que l'on ne considère que l'implication $(\neg\mathscr{Q} \Rightarrow \mathscr{P})$ que l'on démontre par contraposée. L'autre implication $(\mathscr{Q} \Rightarrow \neg\mathscr{P})$ est triviale.

Démonstration 1 : On considère Π_0 la projection de 0 sur le polyèdre convexe fermé formé par la famille $(-a_i)_{i\in[\![1,n]\!]}$, noté Δ. On écrit alors la caractérisation de cette projection :

$$\forall x \in \Delta \langle x - \Pi_0, -\Pi_0\rangle \leq 0 \tag{8}$$

Appliqué en $(-a_i)_{i\in[\![1,n]\!]}$ on trouve que $\forall i \in [\![1,n]\!],\ \langle a_i, \Pi_0\rangle + \|\Pi_0\|^2 \leq 0$. Or on a $\neg\mathscr{P}$ donc Π_0 ne peut être de norme différente de 0 (sinon on a une solution stricte aux inégalité). Ainsi $\Pi_0 = 0$. Donc 0 est combinaison convexe des $(-a_i)_{i\in[\![1,n]\!]}$ donc des $(a_i)_{i\in[\![1,n]\!]}$ également et on en déduit le théorème de Gordan.

Démonstration 2 : on considère le problème $\langle a_i, x\rangle \leq -\epsilon$ avec $\epsilon \in \mathbb{R}_+^*$ et ce $\forall i \in [\![1,n]\!]$. Le problème n'admet pas de solution (usage de $\neg\mathscr{P}$) mais en admet une au sens des moindres carrés. Elle vérifie alors $A^T(Ax-\epsilon)_+ = 0$. Mais on peut réécrire cela $\sum_{i=1}^{n} a_i((Ax-\epsilon)_+)_i = 0$. Enfin, on remarque que $\exists i \in [\![1,n]\!],\ ((Ax-\epsilon)_+)_i > 0$ sinon on a une solution consistante. En posant $\lambda_i = ((Ax-\epsilon)_+)_i$ on a donc montré $\mathscr{Q}$.

Le lemme de Farkas permet de passer aux inégalité larges.

Lemme 1 (Farkas, 1902)**.** $(\mathscr{P})$ $\exists x \in \mathbb{R}^m,\ \forall i \in [\![1,n]\!],\ \langle a_i, x\rangle \leq 0$ *et* $\langle a_1, x\rangle < 0$ *est équivalent à la négation de la proposition suivante :* $(\mathscr{Q})$ $\exists (\lambda_i)_{i\in[\![1,n]\!]} \in {\mathbb{R}_+}^n \setminus \{\lambda_1 = 0\},\ \sum_{i=1}^{n} \lambda_i a_i = 0$

Démonstration : Encore une fois $\mathscr{Q} \Rightarrow \neg\mathscr{P}$ est évidente. L'implication réciproque $\neg\mathscr{P} \Rightarrow \mathscr{Q}$ est plus délicate. On va raisonner par récurrence sur la dimension de l'espace. L'initialisation est triviale. On constate que si $\neg\mathscr{P}$ alors on est dans les conditions du théorème de Gordan et $\exists (\lambda_i)_{i\in[\![1,n]\!]} \in \mathbb{R}_+^n \setminus \{0\}$ telle que $\sum_{i=1}^{n} \lambda_i a_i = 0$. Si $\lambda_1 > 0$ alors on a terminé. Plaçons-nous désormais dans le cadre où $\lambda_1 = 0$. On a donc $\sum_{i=1}^{n} \lambda_i a_i = 0$ et au moins un des λ_i (avec $i \in [\![2,n]\!]$) est non nul, on le note λ_j. Plaçons-nous sur $a_j^\perp$ qui est de dimension strictement inférieure à celle de l'espace ambiant (on a supposé tous les $a_i \neq 0$, l'extension au cas où l'un d'entre eux ou plusieurs sont nuls est immédiate). On peut appliquer le lemme de Farkas sur $a_j^\perp$ et on a $(\alpha_i)_{i\in[\![1,n]\!]\setminus\{j\}} \in \mathbb{R}_+^{n-1}$ avec $\alpha_1 > 0$ tel que $\sum_{i=1, i\neq j}^{n} \alpha_i a_i = \mu a_j$ avec $\mu \in \mathbb{R}$ (en effet il a fallu projeter les a_i dans $a_j^\perp$ pour pouvoir appliquer le lemme, c'est-à-dire, il a fallu appliquer le lemme aux $a_i - \langle a_i, a_j\rangle a_j$). Si μ est négatif alors on a terminé. Si μ est positif alors on remplace a_j par $\frac{-1}{\lambda_j} \sum_{i=1, i\neq j}^{n} \lambda_i a_i$ et on conclut.

On peut maintenant citer le théorème de Motzkin, qui bien que semblant plus fort que le lemme de Farkas en est une conséquence directe [1].

Théorème 3 (Motzkin, 1936)**.** $(\mathscr{P})$ $\exists x \in \mathbb{R}^m,\ \forall i \in [\![1,n]\!],\ \langle a_i, x\rangle \leq 0$ *et* $\forall i \in [\![1,p]\!],\ \langle a_i, x\rangle < 0$ *est équivalent à la négation de la proposition suivante :* $(\mathscr{Q})$ $\exists (\lambda_i)_{i\in[\![1,n]\!]} \in {\mathbb{R}_+}^n \setminus \{(\lambda_1, \ldots, \lambda_p) = 0\},\ \sum_{i=1}^{n} \lambda_i a_i = 0$

1. Tout comme le théorème de Brouwer paraît plus simple que le théorème de Schauder!

Démonstration : Encore une fois $\mathscr{Q} \Rightarrow \neg\mathscr{P}$ est triviale. On considère seulement le sens $\neq \mathscr{P} \Rightarrow \mathscr{Q}$. On considère $C_i = \{x,\ \langle a_i, x\rangle \leq 0,\ \langle a_j, x\rangle \leq 0\}$. Si tous les C_i sont non vides pour $i \in [\![1,p]\!]$ alors on peut considérer un élément $x = \sum_{i=1}^{p} x_i$ avec $x_i \in C_i$. Cet élément est alors une solution de notre système d'inéquations linéaires et cela rentre en contradiction avec $\neg\mathscr{P}$. Ainsi, au moins un des C_i est vide. On peut alors appliquer le lemme de Farkas pour ce problème et donc on obtient une famille $(\lambda_i)_{i\in[\![1,n]\!]}$ avec λ_i non nul et donc $(\lambda_1,\dots,\lambda_p)$ non nul.

Le cas des inégalité larges est donc bien compris. Il s'agit maintenant d'entériner le cas où $b \neq 0$.

3.3 Retour aux inégalités linéaires

Pour cela on va se servir des équivalences suivantes :

$$\exists x \in \mathbb{R}^m,\ Ax \leq b \Leftrightarrow \exists x \in \mathbb{R}^{m+1} (A|b)x \leq 0, \text{ et } x_{m+1} = -1 \Leftrightarrow \exists x \in \mathbb{R}^{m+1} (A|b)x \leq 0, \text{ et } x_{m+1} < 0 \tag{9}$$

où $(A|b)$ est la matrice complétée de A par le vecteur b, c'est-à-dire : $(A|b) = [A\ b]$. La dernière équivalence est obtenue en utilisant la linéarité des équations.

Proposition 9. *Le système $Ax \leq b$ n'admet pas de solutions si et seulement si e_{m+1} est un élément du cône engendré par $([-a_i - b_i])_{i\in[\![1,n]\!]}$. On remarque que $-(A|b)^T$ est la matrice dont les colonnes sont les $[-a_i - b_i]$.*

Démonstration : c'est une simple conséquence du lemme de Farkas.
Désormais on note $\tilde{A} = -(A|b)^T$.

Proposition 10. *Soit (A,b) un système d'inéquations linéaires consistant et si $n \leq m$. Alors il admet une infinité de solutions.*

Démonstration : Si le système est fortement consistant alors on se sert de la proposition démontrée plus haut. Sinon, puisque le système est consistant sans être fortement consistant on peut utiliser le lemme de Farkas puis le théorème de Gordan pour écrire :

$$\begin{cases} \neg\left(\exists (\lambda_i)_{i\in[\![1,n+1]\!]} \in \mathbb{R}_+^{n+1}\setminus\{\lambda_{n+1} \neq 0\},\ \sum_{i=1}^{n} \lambda_i [a_i\, b_i] - \lambda_{n+1} e_{n+1} = 0\right) \\ \exists (\lambda_i)_{i\in[\![1,n+1]\!]} \in \mathbb{R}_+^{n+1}\setminus\{0\},\ \sum_{i=1}^{n} \lambda_i [a_i\, b_i] - \lambda_{n+1} e_{n+1} = 0 \end{cases} \tag{10}$$

Donc on peut rassembler ces deux informations pour écrire :

$$\exists (\lambda_i)_{i\in[\![1,n]\!]}, \in \mathbb{R}_+^{n}\setminus\{0\},\ \sum_{i=1}^{n} \lambda_i [a_i\, b_i] = 0 \tag{11}$$

Cela donne a fortiori $\mathrm{rg}(A^T) < n$. Or $\mathrm{rg}(A) = \mathrm{rg}(A^T)$. Donc puisque $m = \dim(\mathrm{Ker}A) + \mathrm{rg}A$, $\ker A \neq \{0\}$. Donc on peut ajouter n'importe quel élément du noyau et on a toujours une solution de notre système d'inéquations linéaires. On a donc une infinité de solutions

Remarque : le cas $n < m$ est trivial. Seul le cas $n = m$ est intéressant.

Proposition 11. *L'intersection non vide de deux demi-espaces de $\mathbb{R}^n$ est un convexe de dimension au pire $n-1$. Si c'est le cas alors c'est l'hyperplan affine séparateur.*

Ainsi pour obtenir une réduction de dimension, si la première équation était $\langle a, x\rangle \leq b$ on doit avoir $\langle -a, x\rangle \leq -b$ pour la seconde. Malheureusement, ce résultat ne persiste pas pour un polyèdre convexe (c'est-à-dire l'intersection de plusieurs demi-espaces).

Proposition 12. *Soit $\mathscr{P}$ un polyèdre convexe de $\mathbb{R}^n$. Alors, l'intersection non vide d'un demi-espace avec ce polyèdre convexe est un convexe de dimension au plus n. De plus celui-ci est de dimension k si et seulement si l'intersection du polyèdre convexe avec le demi-espace est une face de dimension k du polyèdre convexe.*

Pour déterminer les faces du polyèdre convexe on doit déterminer les sommets. Pour déterminer les faces il convient ensuite de considérer les combinaisons convexes des sommets qui préservent des égalité. Ce procédé peut être extrêmement couteux. En effet, le nombre de sommets est possiblement bien supérieur à la dimension. Par exemple, le cube $[0,1]^n$ a pour sommets 2^n points. Néanmoins pour des petits problèmes on peut toujours raisonner via des considérations géométriques.

Ayant obtenu des résultats sur le nombre de solutions lorsque la consistance est acquise on se tourne désormais vers le problème de déterminer si oui ou non il existe des solutions consistantes. On rappelle qu'on s'intéresse toujours au problème $Ax \leq b$ avec $A \in \mathcal{M}_{n,m}(\mathbb{R})$ et $b \in \mathbb{R}^n$.

Proposition 13. *Résoudre le problème* $(\mathscr{P}_1)$: $Ax \leq b$ *est consistant, revient à résoudre les problème* $(\mathscr{P}_2)$: $-Bx \leq \tilde{A}^+ e_{m+1}$ *est consistant et* $(\mathscr{P}_2)'$: $\tilde{A}x = e_{m+1}$ *où* B *est une matrice dont les colonnes forment une base de* $\ker\tilde{A}$.

Démonstration : On sait que $\mathscr{P}_1$ n'admet pas de solutions si et seulement si il existe un élément $c \geq 0$ tel que $\tilde{A}c = e_{m+1}$. L'ensemble des solutions de l'équation $\tilde{A}x = e_{m+1}$ (résolution du problème $\mathscr{P}_2'$) peut facilement être trouvé par élimination de Gauss. Si ce problème n'admet pas de solution alors $\mathscr{P}_1$ admet des solutions. Supposons maintenant que ce problème admette des solutions. L'ensemble de celle-ci est donné par $\tilde{A}^+ e_{m+1} + \ker\tilde{A}$. Il est à signaler qu'une base de $\ker\tilde{A}$ peut être identifiée lors de la recherche de solutions de $\mathscr{P}_2'$. Il s'agit alors de trouver un élément de cet ensemble dont toutes les coordonnées sont positives. Il faut donc résoudre le problème $\mathscr{P}_2$.

Remarque 1 : l'ajout du problème $\mathscr{P}_2'$ n'est pas un problème car celui-ci est facile à résoudre.

Remarque 2 : $B \in \mathcal{M}_{n,n-\text{rang}(A|b)}(\mathbb{R})$.

Remarque 3 : le problème des inégalité linéaires est très simple à résoudre si $\text{rang}(A) = 1$ ou $\text{rang}(A) = 0$. On dira alors que le problème est facile à résoudre.

Ainsi, via la proposition précédente on peut construire un nouveau problème de taille $(n, n - \text{rang}(A|b))$. Il peut être utile de considérer ce problème si $\text{rg}(A) = n$. Dans le cas général, il est compliqué de déterminer si le système est consistant ou non. On ne dispose que depuis récemment d'algorithmes qui convergent en temps polynomiaux et beaucoup de questions sont encore ouvertes à ce sujet.

3.4 Algorithmes et consistance

On va ici décrire deux algorithmes pour la résolution du problème de consistance :
— l'élimination de Fourier-Motzkin
— l'algorithme du simplexe

3.4.1 L'élimination de Fourier-Motzkin

On présente ici un premier algorithme trouvé par Joseph Fourier et redécouvert par Théodore Motzkin, [4]. L'idée ici est de réduire le nombre de variables, c'est-à-dire de diminuer m la taille de l'espace d'entrée. Malheureusement, le nombre d'équations à vérifier explose dans ce cas.

Proposition 14. *Soit un système* $Ax \leq b$ *avec* $b \in \mathbb{R}^n$ *et* $A \in \mathcal{M}_{n,m}(\mathbb{R})$. *On peut trouver* $B \in \mathcal{M}_{\lceil \frac{n^2}{4} \rceil, m-1}$ *et* $b' \in \mathbb{R}^{\lceil \frac{n^2}{4} \rceil}$ *tel que la résolution de* $Ax \leq b$ *soit équivalent à celle de* $Bx' \leq b'$.

Démonstration : On considère la dernière coordonnée de x, x_m. On note :
— n_A le nombre d'équations où l'on peut isoler x_n et qui sont de la forme $x_n \leq A_i(x_1, \ldots, x_{n-1})$ avec $i \in [\![1, n_a]\!]$.
— n_B le nombre d'équations où l'on peut isoler x_n et qui sont de la forme $x_n \geq B_i(x_1, \ldots, x_{n-1})$ avec $i \in [\![1, n_b]\!]$.
Une condition nécessaire de consistance est alors

$$\exists (x_1, \ldots, x_{n-1}) \max_{i \in [\![1, n_a]\!]} B_i(x_1, \ldots, x_{n-1}) \leq \min_{i \in [\![1, n_b]\!]} A_i(x_1, \ldots, x_{n-1}) \tag{12}$$

Ce qui se traduit par $n_a n_b$ équations du type $B_j(x_1, \ldots, x_{n-1}) \leq A_i(x_1, \ldots, x_{n-1})$ avec $j \in [\![1, n_b]\!]$ et $i \in [\![1, n_a]\!]$.

Ainsi le nombre d'équations à vérifier pour avoir consistance est $n - n_a - n_b + n_a n_b$. En maximisant cette quantité (on peut d'abord remarquer que le maximum est atteint pour $n_a + n_b = n$ si $n \geq 2$ puis étudier le polynôme de degré 2 en n_a) on trouve le résultat.

Remarque : cette technique n'est pas satisfaisante. En effet, pour un système dont l'entrée possède m coordonnées on obtient une résolution de l'ordre de $4\left(\frac{n}{4}\right)^{2^m}$ étapes. La complexité est donc doublement exponentielle.

3.4.2 L'algorithme du simplexe

On va maintenant décrire un algorithme développé par George Dantzig en 1947, on ne donnera pas de preuve de convergence de l'algorithme. La description complète de celui-ci est en effet assez technique et demande une attention particulière pour éviter de boucler (règles d'anti-cyclage). On présente simplement l'idée générale.

Tout d'abord il convient de se placer dans un cadre plus large, celui de la programmation linéaire qui cherche à résoudre le problème suivant :

$$\begin{cases} \inf\limits_{x \in \mathbb{R}^m} \langle c, x \rangle \\ \text{sous la contrainte } \ Ax \leq b \end{cases} \tag{13}$$

Un théorème de programmation linéaire assure que si une solution existe alors il existe une solution en un des sommets du polyèdre convexe défini par $Ax \leq b$. Ainsi l'algorithme procède de la manière suivante :

— on commence par se placer en un sommet
— on se déplace dans la direction de minimisation sur une arête
— on aboutit à un autre sommet. Soit celui-ci est solution, soit il ne l'est pas et on recommence.

Remarque : deux points importants sont à noter :

— l'algorithme converge mais la vitesse de convergence est a priori exponentielle. Néanmoins, l'algorithme est très efficace dans la plupart des cas.
— Si le problème est non borné, c'est-à-dire si le minimum n'est pas atteint, l'algorithme s'arrête après un nombre fini d'étapes.
— on pourrait penser que l'on se place ici dans un cadre bien éloigné de notre problème de départ qui consistait à chercher une solution au problème de consistance des inégalités linéaires. En effet le problème de la consistance des inégalités linéaires est aussi compliqué que celui de la programmation linéaire.

La proposition suivante illustre comment l'algorithme du simplexe répond au problème de consistance.

Proposition 15. *Soit le problème de programmation linéaire suivant :*

$$\begin{cases} min \sum\limits_i z_i \\ Ax + Dz = b \\ x \geq 0, z \geq 0 \end{cases} \tag{14}$$

Avec D une matrice diagonale avec D_{ii} qui vaut 1 si $b_i \geq 0$ et -1 sinon. On peut toujours débuter l'algorithme du simplexe de $(0, Db)$. Si la solution donnée par l'algorithme du simplexe est $(\hat{x}, \hat{z})$ avec $\hat{z} \neq 0$ alors le système d'inégalités n'est pas consistant. Si $\hat{z} = 0$ alors le système d'inégalités est consistant et $\hat{x}$ est une solution de $A\hat{x} = b$ avec $\hat{x} \geq 0$

On détaille dans 5.2.1 pourquoi ce problème d'égalités linéaires sous contraintes peut être mis en correspondance avec un problème d'inégalités linéaires.

4 Inégalités linéaires aux moindres carrés

4.1 Existence de solutions et recherche de solutions

Théorème 4. *Quelque soit la matrice $A \in \mathcal{M}_{n \times m}(\mathbb{R})$ et le vecteur $b \in \mathbb{R}^n$, le problème d'inégalité aux moindres carrés $\mathscr{P}_{\leq}(A, b)$ admet une solution $x_0 \in \mathbb{R}^m$. Quelque soit la solution x_0, le vecteur résidu $z = (Ax_0 - b)_+$ est unique est x est une solution du problème si et seulement si $(Ax - b)_+ = z$.*

Démonstration : Pour l'existence, trois preuves différentes sont proposées dans l'article [1]. Il est à noter que $\|(Ax-b)_+\|^2$ peut s'interpréter très facilement comme la distance de $Ax-b$ à $K = \{x \in \mathbb{R}^m,\ \forall i \in [\![1,m]\!],\ x_i \geq 0\}$. Le problème peut-alors s'écrire comme un problème de projection sur $\mathrm{Im}A + K$ de b. Le théorème de projection sur un convexe fermé dans un espace Hilbertien assure l'existence et l'unicité de ce projeté.

Remarques : Il n'existe pas pour l'instant de théorème donnant la forme des solutions au problème $\mathscr{P}_{\leq}(A,b)$. C'est pourquoi nous allons nous intéresser plus particulièrement aux algorithmes qui proposent de trouver une solution au problème.

La fonction $F(x) = \|(Ax-b)_+\|^2$ que l'on cherche à minimiser est différentiable et convexe. Malheureusement elle n'est pas strictement convexe en général. On peut le voir en remarquant que si le minimum n'est pas unique on n'a aucune chance d'aboutir à une stricte convexité. On n'a donc aucune assurance que les algorithmes "classiques" de minimisation de fonctions convexes (gradient, gradient conjugué...) convergent vers une solution.

4.2 L'algorithme de Han

L'algorithme de Han [3] qui permet de résoudre le problème des inégalité linéaires au moindre carré en temps fini.

4.2.1 Description

On se donne $x_0 \in \mathbb{R}^m$. On définit $I = \{i \in [\![1,n]\!],\ (Ax_k-b)_i \geq 0\}$. On considère la formule d'itération suivante (en omettant les indices sur l'ensemble I pour les notations) :

$$\begin{aligned} d_k &= A_I^+(A_I x_k - b) \\ \widehat{\lambda} &= \underset{\lambda \in \mathbb{R}^+}{\operatorname{argmin}}\ f(x-\lambda d_k) \\ x_{k+1} &= x_k - \widehat{\lambda} d_k \end{aligned} \tag{15}$$

Remarques :

— si l'on suppose que $I = [\![1,n]\!]$ et que $A_I^T A_I = A^T A$ inversible, notre fonction F a minimiser est deux fois différentiable de Hessienne inversible. L'algorithme de Newton s'applique et donne comme direction de descente :

$$\begin{aligned} d_k &= (A^TA)^{-1}A^T(Ax_k-b) \\ A^TAd_k &= A^T(Ax_k-b) \end{aligned} \tag{16}$$

— on peut essayer d'appliquer cette stratégie dans le cas général en cherchant d_k qui résolve au sens des moindres carrés le système :

$$A_I^T A_I d_k = A_I^T(A_I x_k - b_I) \tag{17}$$

équivalent à :

$$A_I d_k = (A_I x_k - b_I) \tag{18}$$

dont on connait les solutions : $d_k \in A_I^+(A_I x_k - b_I) + \ker A_I$. On prend d_k de norme minimale, c'est à dire $d_k = A_I^+(A_I x_k - b_I)$. C'est la direction de descente proposée par l'algorithme de Han.

4.2.2 Preuve de convergence

Commençons par montrer que d_k ainsi défini est tel que $-d_k$ est une direction de descente. On note $d(x) = A_I^+(A_I x - b_I)$.

Proposition 16. *On a* $\nabla f(x) = A_I^T A_I d(x)$ *et donc* $d(x)^T \nabla f(x) = \|A_I d(x)\|^2$

Démonstration : On a :

$$\begin{aligned} A_I^T A_I d(x) &= A_I^T A_I A_I^+ A_I x - A_I^T A_I A_I^+ b_I \\ &= A_I^T A_I x - A_I^T (A_I A_I^+)^T b_I \\ &= A_I^T A_I x - (A_I A_I^+ A_I)^T b_I \\ &= A_I^T A_I x - A_I^T b_I \\ &= \nabla f(x) \end{aligned} \tag{19}$$

Proposition 17. $(f(x_k))_{k\in\mathbb{N}}$ *est décroissante et convergente, et* $\nabla f(x_k) \underset{k\to+\infty}{\to} 0$.

Démonstration :

— $(f(x_k))_{k\in\mathbb{N}}$ décroit.
En effet, comme $d_k^T \nabla f(x_k) \leq 0$, d_k est une direction de descente. Comme $\widehat{\lambda} = \underset{\lambda}{\operatorname{argmin}}\, f(x_k - \lambda d_k)$, $f(x_{k+1}) \leq f(x_k)$. La suite étant décroissante majorée, elle converge.

— $\nabla f(x_k) \underset{k\to+\infty}{\to} 0$

$$\begin{aligned} f(x-\lambda d) - f(x) &= -\int_0^1 d^T \nabla(x - \lambda t d)\lambda \mathrm{d}t \\ &= -\int_0^1 d^T \nabla f(x-\lambda t d)\lambda \mathrm{d}t + \int_0^1 d^T \nabla f(x)\lambda \mathrm{d}t - d^T \nabla f(x)\lambda \\ &= \int_0^1 \lambda d^T (\nabla f(x) - \nabla f(x - \lambda t d)) - \|A_I d\|^2 \lambda \\ &\leq \frac{\|A_I\|^2 \|d\|}{2}\lambda^2 - \|A_I d\|^2 \lambda \end{aligned} \tag{20}$$

On s'est servi du caractère Lipschitz du gradient. $\|\nabla f(x) - \nabla f(y)\| \leq \|A\|^2 \|x - \|y\|$. On s'est également servi du fait que λ est positif (puisqu'on a une direction de descente et la fonction est convexe on peut se restreindre à ce cas). En minimisant le terme de droite en λ on trouve :

$$f(x-\lambda d) - f(x) \leq -\frac{1}{2}\frac{\|A_I d\|^4}{\|A_I\|^2 \|d\|^2} \tag{21}$$

Ainsi puisque la série de terme générale $f(x_k) - f(x_{k+1})$ est convergente, $\frac{\|A_I d_k\|^2}{\|d_k\|} \to 0$. Soit $d_k \to 0$ et dans ce cas $\nabla f(x_k) \to 0$. Sinon $A_I d_k \to 0$ et la conclusion est la même.

Il s'agit maintenant de montrer que le résidu $(Ax_k - b)_+$ converge vers le résidu (unique) d'une solution du problème.

Lemme 2. *Soit $(b_k, c_k)_{k\in\mathbb{N}}$ deux vecteurs tels que le systèmes $Ax \leq b_k$ et $Bx = c_k$ soit consistant pour tout $k \in \mathbb{N}$. On suppose que nos deux suites convergent vers b^*, respectivement c^*. Le système $Ax \leq b^*$ et $Bx = c^*$ est consistant.*

Proposition 18. *Soit $z_k = (Ax_k - b)_+$ alors $z_k \to z^*$ où z^* est le résidu optimal.*

Démonstration : Par décroissance de $f(x_k)$ la suite z_k est contenue dans un compact. Soit une suite extraite qui converge vers $\overline{z}$. On la note encore z_k. Dans ce cas l'ensemble $I = \{i \in [\![1, n]\!],\ \langle a_i, x\rangle > b_i\}$ est constant à partir d'un certain rang. On estime que la suite z_k commence à partir de ce rang. Le système $A_I x = b_I + z_k$ est consistant quel que soit k. Donc via le lemme : $A_I x = b_I + \overline{z}$ est consistant. Donc on peut trouver un $\overline{x}$ tel que $\overline{z} = (A\overline{x} - b)_+$. Mais $\nabla f(\overline{x}) = A^T \overline{z}$ et est limite de $\nabla f(x_k)$ (donc vaut 0). On a alors $\overline{x}$ qui est solution aux moindres carrés et donc $\overline{z} = z^*$. On conclut dans le cas général en supposant que z_k ne tend pas vers z^*. Dans ce cas on extrait une sous-suite qui est toujours à distance ϵ de z^* mais on peut extraire une sous-suite convergente de cette sous-suite. Celle-ci converge vers z^* en vertu de ce qui a été dit au dessus. D'où l'absurdité et on a la convergence attendue.
On n'a toujours aucune information sur la suite x_k et sa convergence est dure à évaluer. On va montrer que la suite converge en un nombre fini d'étapes en s'intéressant au nombre d'indices $I(x_k) = \{i \in [\![1, n]\!],\ \langle a_i, x_k\rangle \geq b_i\}$. Si celui-ci stagne alors on aura une solution de notre problème, sinon il ne peut être que décroissant. Puisque c'est une suite d'entiers naturels, on pourra conclure sur la convergence.

Lemme 3. *Il existe $\epsilon \in \mathbb{R}_+^*$ tel que $\|(Ax - b)_+ - z^*\| \leq \epsilon$ implique que $A_{I(x)} y = b_{I(x)} + z^*_{I(x)}$ est consistant.*

Démonstration : On suppose que la proposition est fausse on a donc une suite x_k telle que $(Ax_k - b)_+$ tend vers z^* mais le système $A_{I(x_k)} y = b_{I(x_k)} + z^*_{I(x_k)}$ n'est jamais consistant. Puisqu'on a une infinité d'éléments de la suite et un nombre fini d'indices possibles on fixe I qui contient une infinité de termes. On renomme notre suite x_k en conséquence. On a toujours $A_I y = b_I + (z_k)_I$ qui est consistant. Donc $A_I y = b_I + z_I^*$ est consistant (par passage de la consistance à la limite). C'est absurde donc la proposition est vraie.

Lemme 4. *Il existe $\epsilon \in \mathbb{R}_+^*$ tel que si $\|(Ax - b)_+ - z^*\| \leq \epsilon$ alors $I(x - d) \subset I(x)$.*

Démonstration : On choisit $\epsilon \in \mathbb{R}_+^*$ tel que l'on soit dans les hypothèses du lemme précédent et qu'en plus $P = \{i \in [\![1,n]\!],\ z_i^* > 0\} \subset I(x)$. Dans ce cas $A_I y = b_I + z_I^*$ est consistant. On a $A^T z^* = A_I^T z_I^* = 0$ (puisque z^* est associé à une solution du problème aux moindres carrés et puisque $P \subset I$ on peut se limiter à z_I^* au lieu de z^*). On pose $\overline{y}$ une solution du système $A_I y = b_I + z_I^*$. C'est alors également une solution de $A_I^T A_I y = A_I^T b_I$. Donc $\overline{y}$ est une solution du problème aux moindres carrés $A_I y = b_I$. On a la même propriété sur $x-d$. Par égalité des résidus : $A_I(x-d) = b_I + z_I^*$. Par positivité de z_I^* on conclut sur l'inclusion.

Lemme 5. *Si $I(x) = I(x-d)$ alors $x - \lambda d$ est une solution aux moindres carrés. Si on a inclusion stricte alors l'inclusion est stricte également de $I(x)$ dans $I(x-\lambda d)$.*

Démonstration : Il est aisé de constater que si $I(x) = I(x-d)$ alors en se servant de $A_I^T A_I d = \nabla f(x)$ on a une solution de l'inégalité au sens des moindres carrés. Donc, $x - \lambda d$ est également solution (puisque $f(x - \lambda d) \leq f(x-d)$). On montre que $\lambda \in]0,1]$ et donc que $x - \lambda d = \lambda(x-d) + (1-\lambda)x$. Ainsi $I(x) \subset I(x-\lambda d)$ (le fait que $\lambda \in]0,1]$, vient en étudiant $f(x-\lambda d)$ en 0 et 1. On ne peut pas avoir égalité car dans ce cas $\lambda = 1$ (toujours en étudiant $g'(t) = \nabla f(x - td)$ en λ ici). Dans ce cas $x - \lambda d = x - d$ et c'est absurde au vu de l'hypothèse d'inclusion stricte. Donc on a inclusion stricte.
Ces trois lemmes permettent d'énoncer (sans démonstration, il s'agit simplement d'appliquer les deux derniers lemmes...) le théorème suivant.

Théorème 5. *L'algorithme de Han converge en un nombre fini d'étapes.*

Remarque : néanmoins, on ne dispose pas de borne sur le temps de convergence de l'algorithme, ni même sur la distance à une solution.

4.2.3 Expériences

On a implémenté l'algorithme du Han pour vérifier les propriétés de convergence. On a choisi une matrice de taille 100×1000 avec ses coefficients choisis de manière aléatoire uniforme sur $[0,1]$ et un vecteur de taille 100 dont les coefficients sont choisis uniformément dans $[-\frac{1}{2}, \frac{1}{2}]$. On a vérifié que le système était consistant (la matrice étant inversible).

On vérifie que l'algorithme de Han converge en temps fini. Néanmoins, un des problèmes de cet algorithme est que même si il converge en temps fini ce temps peut être arbitrairement long. On observe que ce temps est raisonnable en procédant à l'expérience suivante : on fait tourner l'algorithme de Han sur plusieurs tailles de problèmes à chaque fois plusieurs fois. On moyenne le nombre d'itérations nécessaires à la convergence pour chaque taille de problème. Ici on a pris :

— $n_{it} = 10$ (nombre d'itérations à taille fixée)
— $n_v = [10, 50, 100, 500, 1000]$

On obtient les résultats suivants :

Taille du système	10	50	100	500	1000
Nombre moyen d'itérations	2.8	30.7	58.8	143.8	138.2

Dans les prochaines sections, on envisage d'autres manières d'aborder le problème menant à des algorithmes moins compliqués que celui trouvé par Han.

5 D'autres approches...

5.1 Régularisation quadratique

5.1.1 Description

Dans cette sous section on considère le problème des inégalité linéaires au moindre carré avec régularisation L^2. Le problème devient alors de trouver l'argument minimum de la fonction suivante :

$$F(x) = \|(Ax-b)_+\|^2 + \alpha\|x\|^2 \tag{22}$$

Avec $\alpha \in \mathbb{R}_+^*$. Cette fonction définie sur $\mathbb{R}^m$ est différentiable, convexe. De plus, elle est coercive donc l'existence du minimum est assuré (cette existence représentait une véritable difficulté sans la régularisation L^2). Tout argument minimum $\hat{x}$, celui-ci doit vérifier :

$$A^T(A\hat{x}-b)_+ + \alpha\hat{x} = 0 \tag{23}$$

Il s'agit donc de résoudre un problème de point fixe :

$$x = -\frac{1}{\alpha}\left(A^T(Ax-b)_+\right) \tag{24}$$

On pose $f_\alpha(x) = -\frac{1}{\alpha}\left(A^T(Ax-b)_+\right)$.

On peut aussi envisager une résolution du problème via une méthode de descente de gradient. En effet, l'ajout de la norme au carré de x permet de rendre le problème strictement convexe. En effet, on a :

$$\langle \nabla F(x) - \nabla F(y), x-y\rangle \geq 2(\alpha - \lambda_M)\|x-y\|^2 \tag{25}$$

Avec λ_M la plus grande valeur propre de A^TA. Ainsi on a stricte convexité si $\alpha > \lambda_M$. Dans la pratique on peut prendre α bien plus petit comme on le verra dans les expériences.

5.1.2 Expériences

Si α est assez grand alors la fonction f_α est contractante. Dans ce cas, on peut définir la suite $x_n = f_\alpha(x_{n-1})$ avec initialisation quelconque. Le minimum est alors unique et limite de cette suite. Dans les figures suivantes on présente une courte étude numérique de ce phénomène. Dans la suite on a choisi une matrice dont les coefficients sont choisis selon une loi uniforme entre 0 et 1. On prend cette matrice de taille 100×100. De la même manière on choisit un vecteur b dont les coefficients sont choisis selon une loi uniforme entre $-\frac{1}{2}$ et $\frac{1}{2}$. N correspond au nombre d'itérations de l'algorithme. La matrice et les vecteurs utilisés ici pour nos expériences définissent un système consistant (pour vérifier cela on a utilisé 13).

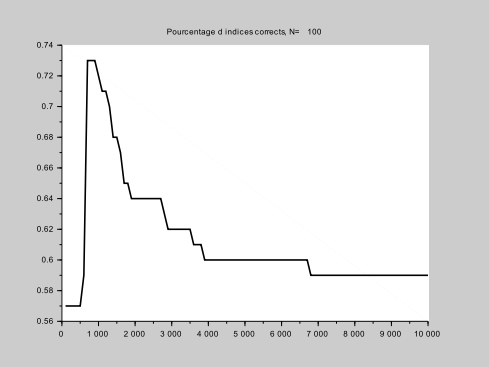

FIGURE 1 – Pourcentage d'indices valides pour $N = 100$

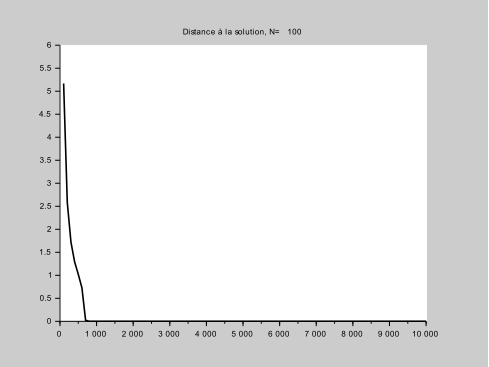

FIGURE 2 – Distance entre la dernière itération et l'avant-dernière pour $N = 100$

Ici on étudie l'évolution du nombres d'indices corrects, c'est-à-dire tels que $(Ax)_i \leq b_i$. On constate que les résultats se situent d'abord autour d'une moitié d'indices corrects ce qui n'est pas satisfaisant. On constate que pour ces α trop faibles l'algorithme n'a pas convergé. Ensuite, on observe une forte augmentation puis une décroissance stricte du nombre d'indices corrects. Cela est dû au fait qu'on cherche de plus en plus à minimiser la norme de x sans chercher à minimiser la norme de $(Ax-b)_+$. Conformément à ce qu'on attendait, il s'agit de prendre α le plus petit possible pour que le plus d'inégalité soient satisfaites.

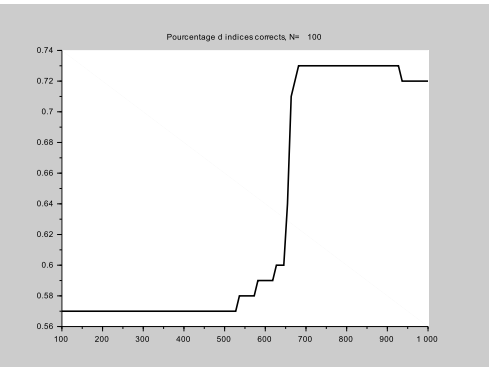

FIGURE 3 – Pourcentage d'indices valides pour $N = 100$

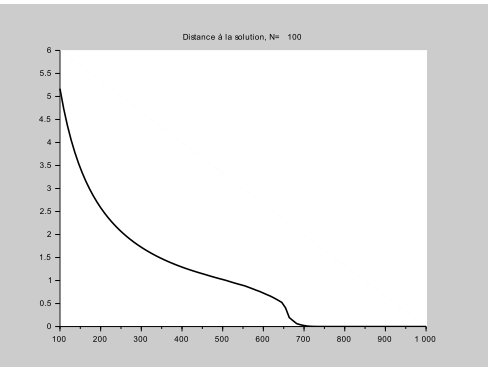

FIGURE 4 – Distance entre la dernière itération et l'avant-dernière pour $N = 100$

Ici, on a détaillé le comportement de l'algorithme lorsque α est petit. Il est clair que l'augmentation du nombre d'indices corrects est dû à la convergence de l'algorithme. Néanmoins, au maximum, seuls 74 indices sur 100 vérifient les contraintes. Pour faire mieux il serait souhaitable de diminuer la valeur de α. Malheureusement on perd alors la convergence.

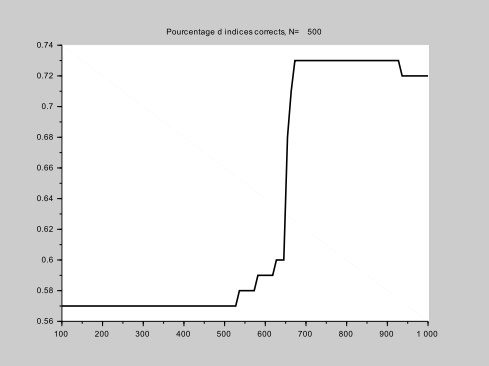

FIGURE 5 – Pourcentage d'indices valides pour $N = 500$

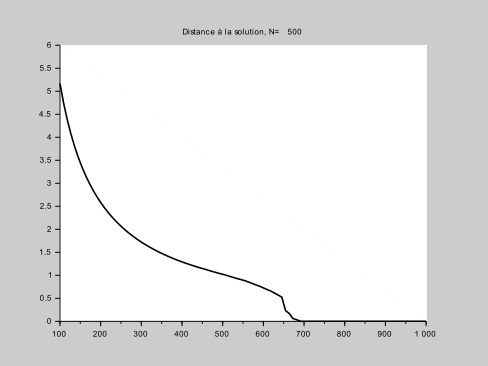

FIGURE 6 – Distance entre la dernière itération et l'avant-dernière pour $N = 500$

On a ici tenté d'augmenter le nombre d'itérations de l'algorithme afin de déterminer si le problème de la non-convergence est dû à un α trop petit ou à un nombre d'itérations trop faible. Le fait que les résultats soient quasiment lorsque $N = 500$ nous incite à pencher vers la première hypothèse d'un α trop petit.

On présente maintenant les résultats obtenus via une autre méthode de minimisation : l'algorithme du gradient à pas fixe.

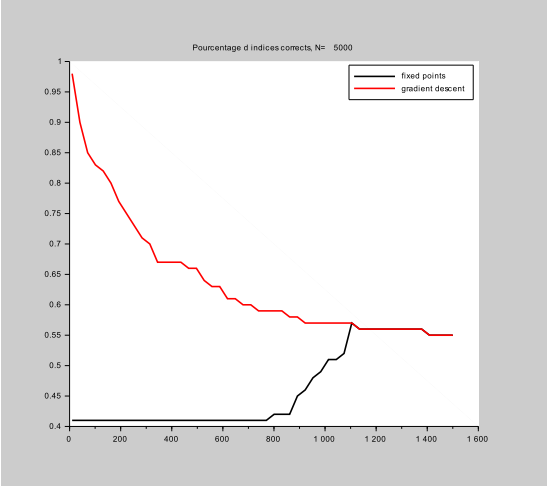

FIGURE 7 – Pourcentage d'indices valides pour $N = 5000$

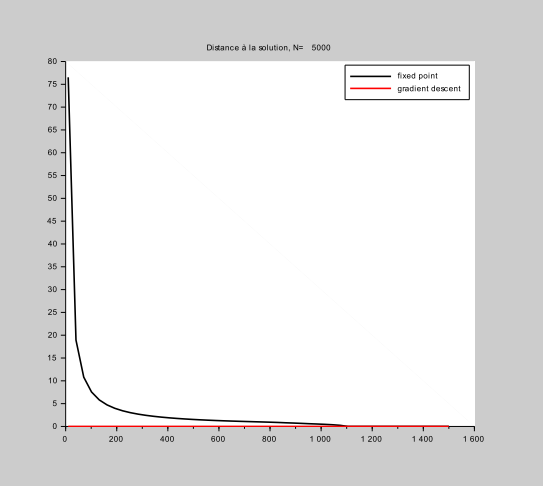

FIGURE 8 – Distance entre la dernière itération et l'avant-dernière pour $N = 5000$

On a ici comparé les résultats obtenus avec la méthode du gradient et celle du point fixe. Le pas choisi pour l'algorithme du gradient à pas fixe est $\gamma = 10^{-5}$. Ainsi, même si on doit toujours avoir un α assez grand pour obtenir une convergence, la condition est plus facilement vérifiée et on peut considérer des α bien plus petits que dans le cadre de la résolution avec une méthode de point fixe.

Remarque : même si cette méthode semble donner de meilleurs résultats que la méthode de point fixe, on doit fixer un paramètre supplémentaire par rapport à la méthode des itérées. On peut supprimer ce choix arbitraire de paramètre en procédant à une étape de *linesearch* mais l'algorithme devient alors bien plus long.

5.2 Un nouveau problème aux moindres carrés

5.2.1 Du cas d'inégalité au cas d'égalité sous contrainte

Théorème 6. *Soient $A \in \mathscr{M}_{n,m}(\mathbb{R})$ avec $(n,m) \in \mathbb{N}^2$ et $b \in \mathbb{R}^n$. Il existe $(n',m') \in \mathbb{N}^2$ et $A' \in \mathscr{M}_{n',m'}(\mathbb{R})$ et $b' \in \mathbb{R}^{n'}$ tels que le système d'inégalités linéaires*

$$Ax \leq b \tag{26}$$

soit équivalent au système d'égalités linéaires avec contrainte :

$$\begin{aligned} A'x &= b' \\ x &\geq 0 \end{aligned} \tag{27}$$

Démonstration : On pose $z = Ax - b$ de sorte que le système $Ax \leq b$ est équivalent à :

$$\begin{aligned} z &\geq 0 \\ Ax + z &= b \end{aligned} \tag{28}$$

On pose $x_+ = max(x,0)$ et $x_- = max(-x,0)$ de sorte que :

$$\begin{aligned} x_+ &\geq 0 \\ x_+ &\geq 0 \\ \tilde{A} &= [A, -A, Id_n] \\ \tilde{x} &= \begin{bmatrix} x_+ \\ x_- \\ z \end{bmatrix} \end{aligned} \tag{29}$$

et on a bien :

$$\begin{aligned} \tilde{A}\tilde{x} &= Ax + z = b \\ \tilde{x} &\geq 0 \end{aligned} \tag{30}$$

qui est équivalent au système $Ax \leq b$. Il est à noter que $\tilde{A} \in \mathscr{M}_{n,2m+n}$.

Ce problème d'égalité sous contrainte que nous avons énoncé peut être résolu au sens des moindres carrés. En effet, on considère le nouveau problème :

$$\mathscr{P}'_{\leq}(A,b) : \ \hat{\tilde{x}} = \underset{\tilde{x}}{argmin}(\|\tilde{A}\tilde{x} - b\|^2) \text{ sachant } \tilde{x} \geq 0 \tag{31}$$

Puisque que le problème est convexe les conditions de Karush, Kuhn et Tucker (KKT) donnent une condition nécessaire et suffisante d'optimalité :

$$\begin{aligned} \tilde{A}^T\left(\tilde{A}\tilde{x} - b\right) &= \Lambda \\ \Lambda &\geq 0 \end{aligned} \tag{32}$$

C'est-à-dire :

$$\begin{aligned} \tilde{A}^T(Ax + z - b) &= \Lambda \\ \Lambda &\geq 0 \end{aligned} \tag{33}$$

Avec $\Lambda = [\Lambda_1 \ \Lambda_2 \ \Lambda_3]^T$ avec $(\Lambda_1, \Lambda_2, \Lambda_3) \in ({\mathbb{R}_+}^m)^3$. On obtient le jeu d'équations suivant :

$$\begin{cases} A^T(Ax + z - b) = \Lambda_1 \\ -A^T(Ax + z - b) = \Lambda_2 \\ Ax + z - b = \Lambda_3 \end{cases} \tag{34}$$

On trouve alors que $\Lambda_1 = -\Lambda_2$. La condition de positivité impose que $\Lambda_1 = \Lambda_2 = 0$. On a l'équivalence suivante :

$$\tilde{x} = (x_+, x_-, z) \text{ est solution du problème } \mathscr{P}'_{\leq} \ \Leftrightarrow \ A^T(Ax + z - b) = 0 \text{ et } Ax + z - b \geq 0 \tag{35}$$

De cela on peut donner le théorème suivant.

Théorème 7. *Toute solution de $\mathscr{P}_{\leq}$ permet de construire une solution de $\mathscr{P}'_{\leq}$.*

Preuve : Soit x une solution du problème des inégalités aux moindres carrés au sens défini par Urruty. On pose x_+, x_- les parties positives et négatives correspondantes. Enfin on pose $z \in \mathbb{R}^m$ tel que $z_i = 0$ si $(Ax-b)_i > 0$ et $z_i = -(Ax-b)_i$ sinon. Dans ce cas : $Ax-b+z = (Ax-b)_+$ et $A^T(Ax-b)_+ = 0$. De plus la condition $Ax+z-b \geq 0$ est vérifiée.

Ainsi en résolvant le problème aux moindres carrés $\mathscr{P}'_{\leq}$ on trouve des solutions dont on peut vérifier si elles appartiennent à l'ensemble des solutions du problème aux moindres carrés original, $\mathscr{P}_{\leq}$. La résolution de ce problème est envisagé par un algorithme de type gradient projeté. Il est à préciser qu'a priori on n'a pas d'assurance de convergence de l'algorithme. En effet, la stricte convexité requiert $\ker\tilde{A} = \{0\}$ ce qui n'est jamais acquis. On pourrait malgré tout envisager de décomposer x sur $\ker\tilde{A} \overset{\perp}{\oplus} \mathrm{Im}\tilde{A}^T$. En posant : $z = x - \tilde{A}^T(\tilde{A}x - b)$ et en décomposant z sur $\ker\tilde{A} \overset{\perp}{\oplus} \mathrm{Im}\tilde{A}^T$. On constate que z n'évolue pas sur $\ker\tilde{A}$ et que l'on applique simplement l'itération sur $\mathrm{Im}\tilde{A}$. Notre problème vient du fait que l'on applique une non-linéarité (i.e $(.)_+$) qui rebat les cartes au moment de passer à l'itération suivante.

5.2.2 Algorithme de gradient projeté

Conscient de ces remarques on a tout de même essayé notre algorithme de gradient projeté sur des données. Ici on présente les résultats obtenus pour une matrice aléatoire de taille 10×10 avec coefficients choisis de manière uniforme dans $[0,1]$ et b vecteur de taille 10×1 avec coefficients choisis de manière uniforme dans $[-\frac{1}{2}, \frac{1}{2}]$. Le pas choisi est ici fixé à 0.005. On affiche tout d'abord l'évolution en fonction du nombre d'itérations de la différence entre deux itérées (en valeur absolue et en échelle semi-logarithmique).

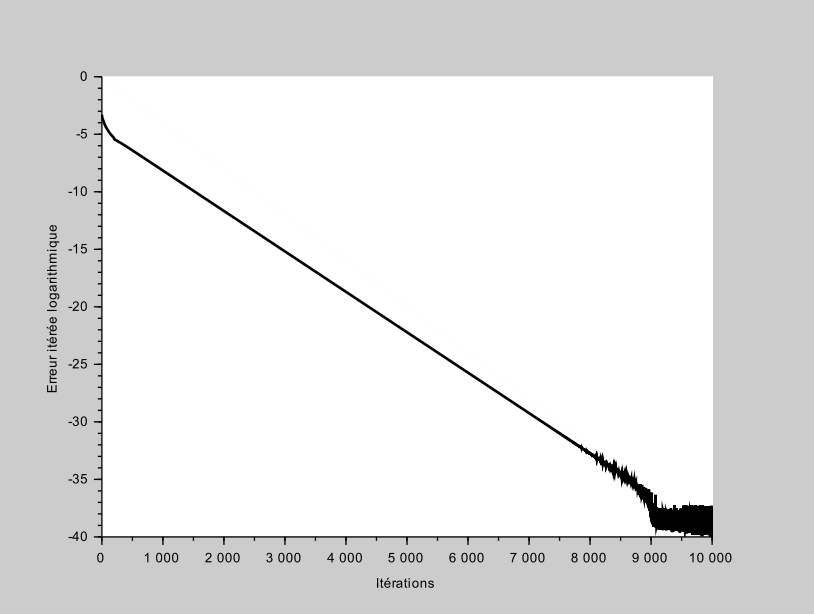

FIGURE 9 – Différence logarithmique entre deux itérées

On observe bien une stabilisation de l'algorithme. On arrête l'algorithme après 5000 itérations et on observe les résultats. On affiche ici deux courbes correspondant à deux coordonnées du résidu (ces deux coordonnées ont été choisies sans raison particulière, on aurait pu afficher toutes les autres coordonnées mais on choisit de ne pas le faire afin de gagner en lisibilité).

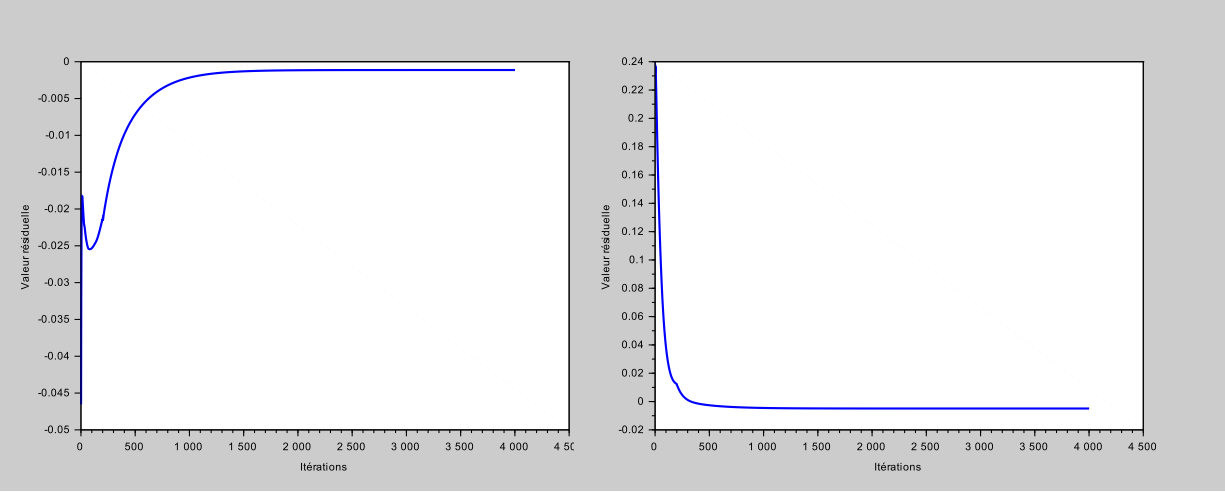

FIGURE 10 – Coordonnée 9 du vecteur résiduel

FIGURE 11 – Coordonnée 8 du vecteur résiduel

Toutes les coordonnées se stabilisent autour d'une valeur qui, dans le cadre de notre expérience, correspondait à une solution au problème de la consistance (en effet le système choisi aléatoirement est de rang 10 et donc consistant). La coordonnée 8 a un comportement particulièrement intéressant puisqu'elle reste toujours positive mais tend vers 0 (pour $N = 10000$, elle est de l'ordre de 10^{-14}).

Références

[1] Penot Jean-Paul Contesse Luis, Hiriart-Urruty Jean-Baptiste. Least-squares solution of linear inequality systems : a pedestrian approach.

[2] George Bernard Dantzig. *Linear programming and extensions*. Princeton university press, 1998.

[3] Han. Least-squares solution of linear inequalities. 1980.

[4] Theodor Motzkin. The theory of linear inequalities. Technical report, DTIC Document, 1952.

[5] Roger Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge Univ Press, 1955.