Systèmes non déterminés

On rencontre souvent des problèmes de moindres-carrés. Mathématiquement, ces problèmes consistent à minimiser le carré de la norme euclidienne d'une fonction à valeurs vectorielles, qui peut être linéaire (on parle alors de moindres-carrés linéaire, voir la section 19.1) ou non linéaire (on parle alors de moindres-carrés non linéaire, voir la section 19.3). Ce chapitre leur est consacré. Étant donné la grande variété de problèmes qui peuvent être modélisés comme problème de moindres-carrés, on les rencontre sous des appellations différentes : *estimation de paramètres*, problème *d'identification*, de *calibration* ou de *régression* en statistiques [199, 333].

19.1 Moindres-carrés linéaire

> *Presque chaque soir, je fais une nouvelle édition du tableau, qu'il est facile d'améliorer n'importe où. Contre la monotonie du travail d'arpentage, c'est toujours une plaisante distraction ; on peut aussi voir immédiatement si quelque chose de douteux s'est glissé, ce qui reste à obtenir, etc. Je vous recommande cette méthode comme modèle. Il ne vous arrivera presque plus jamais de pratiquer une élimination directe, du moins quand vous avez plus de deux inconnues. La procédure indirecte peut se faire à moitié endormi, ou en pensant à autre chose.*
>
> C.F. Gauss, extrait d'une lettre à G.L. Gerling, datée du 26 décembre 1823, dans laquelle il loue les mérites de sa méthode de relaxation, par rapport aux méthodes directes, pour résoudre l'équation normale d'un problème de moindres-carrés linéaire. Traduction de A. Michel-Pajus [112].

19.1.1 Définition du problème

Un *problème de moindres-carrés linéaire* (MCL) est un problème d'optimisation qui s'écrit de la manière suivante :

$$\min_{x\in\mathbb{R}^n} \left(f(x) = \frac{1}{2}\|Ax-b\|_2^2 \right), \tag{19.1}$$

où $\|\cdot\|_2$ est la norme ℓ_2, A est une matrice de type $m \times n$ et $b \in \mathbb{R}^m$. Lorsque $m = n$, on parle de problème *d'interpolation linéaire*. Ce problème peut se voir de différentes manières.

Le premier point de vue est « concret ». On cherche à déterminer des *paramètres* $x \in \mathbb{R}^n$ d'un « système » au moyen de *mesures* $b \in \mathbb{R}^m$ réalisées sur celui-ci. La loi qui relie les paramètres x aux *quantités mesurées* Ax est supposée linéaire. Le problème de moindres-carrés linéaire permet de déterminer les paramètres x qui donnent des quantités mesurées Ax au plus proche des mesures b. De ce point de vue, le problème consiste à projeter b sur l'image de A au moyen du produit scalaire euclidien (voir la section 2.5.1, on pourrait prendre d'autres produits scalaires d'ailleurs).

Le second point de vue est « abstrait ». On s'intéresse à la résolution du système linéaire $Ax = b$. On n'impose pas que $b \in \mathcal{R}(A)$, si bien que ce système n'a peut-être pas de solution. Le problème de moindres-carrés linéaire cherche alors à résoudre ce système linéaire « au mieux », en minimisant le résidu $Ax - b$.

Il s'agit donc d'un problème « fondamental », auquel sont rattachés divers concepts bien connus en algèbre linéaire. On note

$$r = \operatorname{rg} A$$

le rang de A.

19.1.2 Exemple : régression linéaire

En statistiques, un problème de *régression linéaire* consiste à déterminer au mieux la dépendance *affine* supposée entre m mesures et $n-1$ variables. En général $m \gg n$. Pour prendre les notations de la section 19.1.1, la i-ième mesure b_i résulte des valeurs $(a_{i,1}, \dots, a_{i,n-1})$ données au vecteur dont elle dépend. Si la mesure dépend affinement du vecteur, il doit y avoir un vecteur de paramètres $x \in \mathbb{R}^n$, inconnu mais indépendant des mesures prises, tel que le i-ième résidu

$$r_i = b_i - \left(\sum_{j=1}^{n-1} a_{i,j} x_j + x_n \right)$$

soit nul (x_n est le coefficient indépendant des mesures rendant la dépendance *affine* plutôt que *linéaire*). Si l'on note

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m \qquad \text{et} \qquad A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n-1} & 1 \\ \vdots & & \vdots & \vdots \\ a_{m,1} & \cdots & a_{1,n-1} & 1 \end{pmatrix} \in \mathbb{R}^{m \times n},$$

le résidu

$$r = b - Ax \in \mathbb{R}^m$$

doit être nul. En présence d'erreurs de mesure ou d'une relation mesures-données non linéaire, on peut chercher à annuler r au mieux, ce qui conduit à résoudre le problème de moindres-carrés linéaire (19.1).

Dans le cas où $n = 2$, les couples $\{(b_i, a_{i,1})\}_{i=1}^m$ représentent un nuage de m points dans le plan. Pour la solution $x \in \mathbb{R}^2$ du problème de moindres-carrés, l'application affine $a \in \mathbb{R} \mapsto ax_1 + x_2$ donne la relation affine « la plus proche » des couples $\{(b_i, a_{i,1})\}_{i=1}^m$. La figure 19.1 reprise de [619 ; 2013, figure 2.2] donne ainsi la droite passant au plus près, au sens des moindres-carrés, des mesures de l'*indice des problèmes sanitaires et sociaux* en fonction des *inégalités de revenus* ; il y a une mesure (un point) pour chacun des 21 pays considérés dans l'étude.

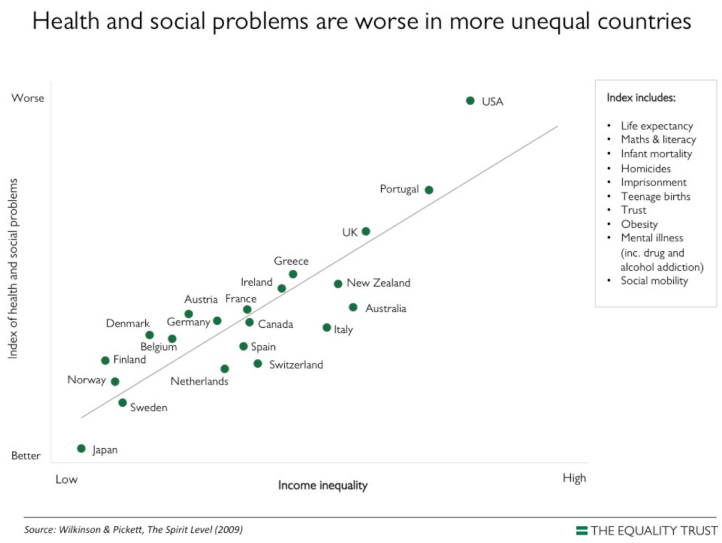


Fig. 19.1. Dépendance affine « la plus proche » de 21 couples (indice des problèmes sanitaires et sociaux, inégalités de revenus), avec un couple par pays considéré.

19.1.3 L'ensemble des solutions

La condition d'optimalité du premier ordre de ce problème s'écrit $\nabla f(x) = 0$ ou encore

$$A^\mathsf{T} Ax = A^\mathsf{T} b. \tag{19.2}$$

Celle-ci porte le nom d'*équation normale* de (19.1).

La proposition suivante règle la question de l'existence et de l'unicité des solutions de (19.1).

Proposition 19.1 (ensemble des solutions) *Le problème* (19.1) *est convexe et admet toujours une solution. Celle-ci est unique si, et seulement si, A est injective. L'ensemble des solutions de* (19.1) *s'écrit $x_p + \mathcal{N}(A)$, où x_p est une solution particulière de* (19.1) *et la valeur optimale est $\frac{1}{2}\|Pb\|_2^2$, où P est le projecteur orthogonal (pour le produit scalaire euclidien) sur $\mathcal{R}(A)^\perp$.*

Démonstration. Le problème (19.1) consiste à projeter b sur l'image de A (voir la section 2.5.1), qui est un convexe fermé non vide. Il existe donc un unique élément $y \in \mathcal{R}(A)$ qui est le plus proche de b (proposition 2.25). Cet élément est de la forme $y = Ax$, où x est une solution de (19.1).

Si A est injective f est strictement convexe ($A^\mathsf{T}A$ est définie positive) et donc (19.1) a une solution unique. Si A n'est pas injective et x est solution, tous les points de $x + \mathcal{N}(A)$ ($\neq \{x\}$) sont aussi solutions ; par ailleurs, si x et x' sont deux solutions de (19.1), elles vérifient l'équation normale et donc $x - x' \in \mathcal{N}(A^\mathsf{T}A) = \mathcal{N}(A)$.

Enfin, $Q = I - P$ étant le projecteur orthogonal sur $\mathcal{R}(A)$, la valeur optimale s'écrit $\frac{1}{2}\|Qb - b\|_2^2 = \frac{1}{2}\|Pb\|_2^2$. □

Il existe de nombreuses démonstrations de l'existence d'une solution de (19.1). Celle donnée ci-dessus considère directement le problème d'optimisation. On peut aussi s'intéresser à son système d'optimalité (19.2) (du fait de la convexité du critère — sa hessienne $A^\mathsf{T}A$ est semi-défini positif — il y a équivalence entre les solutions (19.1) et celles de son système d'optimalité ; théorème 4.9). Pour montrer que ce dernier a toujours une solution, il suffit d'observer que $A^\mathsf{T}b \in \mathcal{R}(A^\mathsf{T}A)$.

L'ensemble des solutions de (19.1) sont donc les solutions de l'équation normale (19.2). On peut s'intéresser à la *solution de norme minimale*, qui est donc définie par

$$\begin{cases} \min \frac{1}{2}\|x\|_2^2 \\ A^\mathsf{T}Ax = A^\mathsf{T}b. \end{cases} \tag{19.3}$$

On peut bien parler de « la » solution de norme minimale, car le critère de ce problème étant strictement convexe, il y a exactement une unique solution de norme minimale. On peut caractériser la solution de ce problème. Comme celui-ci est convexe, ses conditions d'optimalité du premier ordre sont nécessaires et suffisantes (on notera en effet que les contraintes sont qualifiées, car affines). Donc $\hat{x}$ est la solution de norme minimale si, et seulement si, il existe un multiplicateur $\lambda \in \mathbb{R}^n$ (non nécessairement unique) tel que

$$\begin{cases} \hat{x} + A^\mathsf{T}A\lambda = 0 \\ A^\mathsf{T}A\hat{x} = A^\mathsf{T}b. \end{cases} \tag{19.4}$$

La première condition s'écrit aussi $\hat{x} \in \mathcal{R}(A^\mathsf{T}A) = \mathcal{R}(A^\mathsf{T})$. Dès lors, $\hat{x}$ est solution de norme minimale de (19.1), c'est-à-dire solution de (19.3), si, et seulement si,

$$\begin{cases} \hat{x} \in \mathcal{R}(A^\mathsf{T}) \\ A^\mathsf{T}A\hat{x} = A^\mathsf{T}b. \end{cases} \tag{19.5}$$

Comme $\hat{x}$ est univoquement déterminé par le système linéaire d'optimalité (19.4) (il peut cependant avoir de multiples solutions en λ si A n'est pas injective), qui se récrit

$$\begin{pmatrix} I & A^\mathsf{T}A \\ A^\mathsf{T}A & 0 \end{pmatrix} \begin{pmatrix} \hat{x} \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ A^\mathsf{T}b \end{pmatrix},$$

l'application qui à $b \in \mathbb{R}^m$ fait correspondre $\hat{x}$ est linéaire et la matrice qui la représente est appelée le pseudo-inverse (de Moore-Penrose) de A. On note cette matrice $A^\dagger$ et la solution de norme minimale s'écrit

$$\hat{x} = A^\dagger b.$$

On peut voir $\hat{x}$ comme le résultat d'une tentative de trouver une solution au système $Ax = b$, qui n'en n'a pas nécessairement (on se rappelle, en particulier, que A n'est pas carrée), au moyen de deux opérations :

- la première force l'*existence* d'une solution en relaxant « $Ax = b$ » en un problème de minimisation, celui que l'on trouve dans (19.1),
- la seconde force l'*unicité* de la solution par le problème d'optimisation (19.3).

19.1.4 Résolution numérique

On distingue les algorithmes qui utilisent l'équation normale (19.2) (résolution par factorisation de Cholesky ou par gradient conjugué) et ceux qui s'attaquent directement à la formulation originale (19.1) (résolution par factorisation QR ou par factorisation en valeurs singulières).

Résolution de l'équation normale par factorisation

L'approche la plus simple est de faire la *factorisation de Cholesky* de $A^\mathsf{T}A$ et de calculer la solution de l'équation normale en résolvant les deux systèmes triangulaires qui en découlent. Cette approche n'est pas sans inconvénients. D'une part elle oblige de former la matrice $A^\mathsf{T}A$, ce qui demande $O(mn^2)$ opérations (ce n'est souvent pas négligeable). Ensuite, on perd aussi en précision, du fait de l'annulation de l'influence des petits éléments de A (en particulier dans les termes diagonaux $(A^\mathsf{T}A)_{ii} = \sum_j A_{ji}^2$). Enfin, cette approche peut éventuellement détruire la creusité éventuelle de A (si A a une ligne pleine, $A^\mathsf{T}A$ sera en général une matrice pleine).

Une autre possibilité, bien adaptée aux matrices A creuses est de résoudre ce que l'on appelle le *système linéaire augmenté*, équivalent à l'équation normale, que l'on obtient à partir de celle-ci en posant $y = -Ax$:

$$\begin{pmatrix} I & A \\ A^\mathsf{T} & 0 \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} 0 \\ -A^\mathsf{T}b \end{pmatrix}.$$

La matrice K de ce système linéaire est symétrique, mais n'est pas définie positive. D'autre part, l'ordre $n+m$ de K peut être beaucoup plus important que l'ordre n de l'équation normale ; mais si A est creuse, K l'est aussi. On peut alors la factoriser par des méthodes pouvant prendre en compte son caractère creux (comme les solveurs MA27/MA47 de Duff et Reid [181, 182, 183]). On se rappellera que K n'est en général pas définie positive (si A est injective, K a n valeurs propres strictement négatives et m valeurs propres strictement positives), si bien qu'une factorisation de Cholesky ne convient pas.

Résolution de l'équation normale par gradient conjugué

Dans cette approche, on minimise f par l'algorithme du gradient conjugué (GC), ce qui revient à résoudre l'équation normale par le même algorithme. Théoriquement, cet algorithme n'est bien défini que lorsque la matrice du système à résoudre, ici $A^\mathsf{T}A$, est définie positive. Toutefois, grâce à la structure de l'équation normale, le GC peut être utilisé pour résoudre cette équation. C'est ce qu'affirme la proposition suivante. On rappelle que $r = \operatorname{rg} A$.

Proposition 19.2 *L'algorithme du gradient conjugué pour minimiser* (19.1) *est bien défini et converge en au plus r itérations. De plus, si l'itéré initial est pris dans $\mathcal{R}(A^\mathsf{T})$ (par exemple $x_1 = 0$), les itérés convergent vers la solution de norme minimale de* (19.1).

Démonstration. C'est une application directe du point 2 de la proposition 8.11. Le système linéaire à résoudre est l'équation normale (19.2) : on a bien $A^{\mathsf{T}}A \succcurlyeq 0$, $A^{\mathsf{T}}b \in \mathcal{R}(A^{\mathsf{T}}) = \mathcal{R}(A^{\mathsf{T}}A)$ et $\dim \mathcal{R}(A^{\mathsf{T}}A) = \dim \mathcal{R}(A^{\mathsf{T}}) = \dim \mathcal{R}(A) = r$. □

Le principal avantage du GC est d'être utilisable pour les grands problèmes. Cependant, cette approche est sensible aux erreurs d'arrondi et présente des problèmes de stabilité numérique (difficultés si A a des valeurs singulières nulles ou presque nulles). En particulier, *il ne faut pas implémenter l'algorithme du GC standard directement sur l'équation normale*. L'algorithme recommandé dans [63], appelé CGLS ci-dessous, est celui déjà proposé par Hestenes et Steifel [312, 571]. Ces deux algorithmes sont équivalents en arithmétique exacte, mais c'est ce dernier qui a les meilleures propriétés de stabilité numérique.

L'algorithme CGLS génère des itérés x_k, en faisant un pas $\alpha_k > 0$ le long d'une direction d_k : $x_{k+1} = x_k + \alpha_k d_k$. C'est ici le résidu $r_k := b - Ax_k$, dont on cherche à minimiser la norme, qui est mis à jour récursivement (par $r_{k+1} = x_k - \alpha_k p_k$, où $p_k := Ad_k$) et non pas le résidu de l'équation normale, lequel est calculé par $s_k := A^{\mathsf{T}}r_k$. C'est essentiellement par cette mise-à-jour du résidu que CGLS diffère d'une application directe du GC standard.

Algorithme 19.3 (CGLS)

1. On se donne $x_1 \in \mathbb{R}^n$;
2. On calcule le résidu $r_1 = b - Ax_1$, l'opposé du gradient $s_1 = A^{\mathsf{T}}r_1$ et sa norme au carré $\gamma_1 = \|s_1\|_2^2$;
3. Pour $k = 1, 2, \ldots$ faire :
   - 3.1. Si $\gamma_k \simeq 0$, on s'arrête ;
   - 3.2. *Paramètre de conjugaison* : si $k \geqslant 2$, $\beta_k = \gamma_k/\gamma_{k-1}$;
   - 3.3. *Déplacement en x* :

$$d_k = \begin{cases} s_k & \text{si } k = 1 \\ s_k + \beta_k d_{k-1} & \text{si } k \geqslant 2 \,; \end{cases}$$

   - 3.4. *Déplacement en r* : $p_k = Ad_k$;
   - 3.5. *Calcul du pas* : $\alpha_k = \gamma_k/\|p_k\|_2^2$;
   - 3.6. *Nouveau point* : $x_{k+1} = x_k + \alpha_k d_k$;
   - 3.7. *Nouveau résidu* : $r_{k+1} = r_k - \alpha_k p_k$;
   - 3.8. *Nouveau gradient* : $s_{k+1} = A^{\mathsf{T}}r_{k+1}$; $\gamma_{k+1} = \|s_{k+1}\|_2^2$;

On peut encore noter que l'algorithme CGLS évite de faire les produits $A^{\mathsf{T}}(Ad)$ qui, en calcul flottant, peuvent détériorer la performance lorsque le système est mal conditionné ; voir [469 ; section 7.1] et [63 ; section 4.2]. Un autre algorithme à l'efficacité semblable est LSQR de Paige et Saunders [469]. C'est une autre version stable du GC, fondée sur la bidiagonalisation de Lanczos et la factorisation QR.

Résolution par factorisation QR de A

Soit

$$A = QR$$

une factorisation QR de A, où Q est une matrice orthogonale et R est de la forme

$$R = \begin{pmatrix} R_1 \\ 0_{(m-r)\times n} \end{pmatrix}.$$

On a noté $0_{(m-r)\times n}$ la matrice nulle de type $(m-r)\times n$. On sait que R_1 est triangulaire supérieure ($(R_1)_{ij} = 0$ si $i > j$) et que ses éléments diagonaux $(R_1)_{ii}$ $(1 \leqslant i \leqslant r)$ sont non nuls.

On a $\|Q^\mathsf{T} y\|_2 = \|y\|_2$ par orthogonalité de Q, si bien que

$$\begin{aligned} \|Ax - b\|_2^2 &= \|Q^\mathsf{T}(Ax - b)\|_2^2 \\ &= \|Rx - Q^\mathsf{T} b\|_2^2 \\ &= \sum_{i=1}^{r} (Rx - Q^\mathsf{T} b)_i^2 + \sum_{i=r+1}^{m} (Q^\mathsf{T} b)_i^2. \end{aligned}$$

La somme du second terme ne dépend pas de x et la somme du premier terme peut être annulée en résolvant

$$R_1 x = \tilde{b}, \tag{19.6}$$

où $\tilde{b} \in \mathbb{R}^n$ est défini par $\tilde{b}_i = (Q^\mathsf{T} b)_i$ pour $1 \leqslant i \leqslant r$. Du fait de la structure de R_1, cette équation a toujours une solution. Celle-ci est aussi une solution du problème (19.1) puisqu'elle donne au critère f sa valeur minimale $\frac{1}{2}\sum_{i=r+1}^{m} (Q^\mathsf{T} b)_i^2$.

La solution de (19.6) est unique si A est injective, car alors $r = n$ et le système (19.6) est carré. Si A n'est pas injective ($r < n$), ce système permet d'écrire les r premières composantes de x comme fonction de ses $n - r$ dernières composantes, celles-ci pouvant être choisies arbitrairement. On obtiendra la solution de (19.1) avec le plus de zéros en prenant $x_i = 0$ pour $r + 1 \leqslant i \leqslant n$.

Numériquement, la résolution de (19.1) par la factorisation QR de A est très stable. Mais la solution peut être très différente de celle de norme minimale. Mieux vaut utiliser la factorisation SVD (plus coûteuse).

Résolution par factorisation en valeurs singulières (SVD) de A ▲

La décomposition en valeurs singulières de A, de type $m \times n$ et de rang r, s'écrit

$$A = U \Sigma V^\mathsf{T},$$

où U est une matrice de type $m \times r$ dont les colonnes sont orthogonales ($U^\mathsf{T} U = I_r$), V est une matrice de type $n \times r$ dont les colonnes sont orthogonales ($V^\mathsf{T} V = I_r$) et Σ est une matrice d'ordre r diagonale définie positive ($\Sigma_{ij} = 0$ si $i \neq j$ et $\Sigma_{ii} = \sigma_i > 0$). Le pseudo-inverse de A s'écrit

$$A^\dagger = V \Sigma^{-1} U^\mathsf{T}.$$

19.1.5 Estimation numérique de la qualité de la solution ▲

La *matrice de résolution* d'un problème de moindres-carrés linéaire est définie par

$$R = A^\dagger A.$$

Cette formule montre que R est le projecteur orthogonal sur $\mathcal{R}(A^\dagger)$. En particulier, celui-ci projette toute solution x du problème de moindres-carrés sur la solution de norme minimale $\hat{x} = Rx$. On aura donc $R = I_n$ si, et seulement si, le problème de moindres-carrés a une unique solution. D'autre part, plus la colonne j de R est proche de e^j (le j-ième vecteur de base de $\mathbb{R}^m$), plus la j-ième composante des solutions est bien déterminée.

La *matrice de covariance des erreurs a posteriori sur la solution de norme minimale* est définie par

$$C_x = A^\dagger C_b (A^\dagger)^\mathsf{T},$$

où C_b est la *matrice de covariance des erreurs a priori sur les données* b. La matrice C_x spécifie comment des erreurs sur b se propagent à la solution de norme minimale.

Pour les problèmes dont la taille permet de calculer la factorisation SVD de A, disons n et m ne dépassant pas quelques milliers, on peut calculer R et C_x par

$$R = VV^\mathsf{T} \qquad \text{et} \qquad C_x = V\Sigma^{-1}U^\mathsf{T} C_b U \Sigma^{-1} V^\mathsf{T}.$$

Pour les problèmes de très grande taille, le calcul explicite de $A^\dagger$ n'est pas possible. Dans [461], la colonne i de $A^\dagger$, solution de norme minimale de

$$\inf_x \ \|Ax - e^i\|_2$$

est approchée par le point obtenu après une seule itération de gradient démarrant en zéro avec pas optimal. Ceci revient à approcher $A^\dagger$ par $A^\mathsf{T} D$, où D est la matrice diagonale des pas optimaux.

Voir [331, 618] pour des références en géophysique.

19.2 Moindres-carrés polyédrique

Dans les problèmes de moindres-carrés linéaires, on cherche à réaliser au mieux une équation linéaire, éventuellement sans solution. Comment s'y prendre si l'on cherche à réaliser au mieux un système d'équations et d'inéquations linéaires ? C'est à cette question que répond la modélisation sous forme de problème de moindres-carrés polyédrique (MCP).

19.2.1 Définition du problème

Un système d'équations et d'inéquations linéaires peut se représenter de manière compacte par un *problème d'inclusion linéaire dans un intervalle*, lequel s'écrit de la manière suivante

$$Ax \in [l, u], \tag{19.7}$$

où $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$ est l'inconnue, $l \in (\mathbb{R} \cup \{-\infty\})^m$ et $u \in (\mathbb{R} \cup \{+\infty\})^m$. On supposera toujours que $l \leqslant u$. Avec les bornes l et u pouvant avoir des composantes infinies, il est nécessaire de préciser ce que l'on entend par l'intervalle $[l, u]$; celui-ci est défini par

$$[l, u] := \{y \in \mathbb{R}^m : l \leqslant y \leqslant u\}.$$

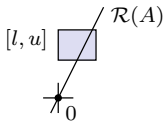
Il ne contient donc pas de valeurs infinies. Le problème d'inclusion (19.7) permet de représenter une égalité linéaire en prenant des bornes égales et une simple inégalité linéaire en prenant une borne infinie :

$$\begin{aligned} A_{i:}x = b_i \quad &\iff \quad A_{i:}x \in [l_i, u_i], \text{ avec } l_i = u_i = b_i, \\ A_{i:}x \leqslant b_i \quad &\iff \quad A_{i:}x \in [l_i, u_i], \text{ avec } l_i = -\infty \text{ et } u_i = b_i, \end{aligned}$$

où $A_{i:}$ désigne la ligne i de A.

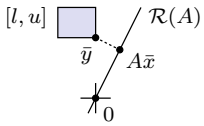
Clairement, on pourra trouver une solution de l'inclusion (19.7), c'est-à-dire un point x tel que Ax soit dans $[l, u]$, si, et seulement si, la condition géométrique suivante a lieu :

$$\mathcal{R}(A) \cap [l, u] \neq \varnothing. \tag{19.8}$$



Si cette condition n'est pas remplie, on pourra essayer de réaliser (19.7) « au mieux » en cherchant x tel que Ax soit le plus proche possible de $[l, u]$, au sens de la norme ℓ_2 (d'où vient le vocable « moindres-carrés »), ce qui s'écrit

$$\min_{x\in\mathbb{R}^n} \mathrm{d}_{[l,u]}(Ax), \tag{19.9}$$



où $\mathrm{d}_{[l,u]} : y \in \mathbb{R}^m \mapsto \mathrm{d}_{[l,u]}(y) := \min\{\|y - y'\|_2 : y' \in [l, u]\}$. On a utilisé l'opérateur « min » dans (19.9) plutôt que « inf », car on verra à la proposition 19.4 que ce problème a toujours une solution (comme les problèmes de moindres-carrés linéaires, voir la proposition 19.1). La valeur optimale de (19.9) est donc la distance euclidienne entre $\mathcal{R}(A)$ et $[l, u]$.

Nous donnerons à (19.9) le nom de *problème de moindres-carrés polyédrique* (MCP) ; le qualificatif *polyédrique* venant du fait que l'ensemble des x vérifiant (19.7) est un polyèdre convexe.

Le problème de moindres-carrés polyédrique (19.9) peut se récrire de différentes manières, qui aident à sa compréhension et à son analyse. En utilisant la définition de la distance $\mathrm{d}_{[l,u]}$, on voit que (19.9) se récrit comme suit

$$(19.9) \quad \iff \quad \min_{(x,y)\in\mathbb{R}^n\times[l,u]} \|y - Ax\|_2. \tag{19.10}$$

Ici aussi l'opérateur « min » est justifié par la proposition 19.4. En introduisant $s := y - Ax$ et en éliminant y, on obtient

$$(19.9) \quad \iff \quad \begin{cases} \min_{(x,s)\in\mathbb{R}^n\times\mathbb{R}^m} \|s\|_2 \\ l \leqslant Ax + s \leqslant u. \end{cases} \tag{19.11}$$

Les contraintes de ce problème sont toujours réalisables (par exemple en prenant $x = 0$ et $s \in [l, u]$ supposé non vide). Un vecteur s qui rend les contraintes de (19.11) réalisables est appelé une *translation réalisable*. L'ensemble des translations réalisables est clairement l'ensemble noté et défini par

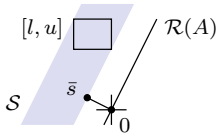
$$\mathcal{S} := [l, u] + \mathcal{R}(A).$$

On peut donc interpréter le problème (19.11) comme celui qui recherche à déterminer la plus petite translation s qui rende l'inclusion $Ax + s \in [l, u]$ réalisable ; on l'appelle la *plus petite translation réalisatible* et on la note $\bar{s}$. En éliminant x du problème (19.11), on obtient

$$(19.9) \quad \Longleftrightarrow \quad \min_{s \in \mathcal{S}} \|s\|_2, \tag{19.12}$$

qui est une expression du problème (19.9) ayant l'avantage d'être coercive. L'utilisation de l'opérateur « min » dans (19.12) est aussi justifié à la proposition 19.4. Ce problème montre aussi que $\bar{s}$ est la projection de zéro sur $\mathcal{S}$:

$$\bar{s} := \arg\min_{s \in \mathcal{S}} \|s\|_2. \tag{19.13}$$



Cette translation $\bar{s}$ est nulle si, et seulement si, (19.7) est réalisable.

19.2.2 Existence et unicité de solution

On note $\mathrm{P}_{[l,u]}$ le projecteur orthogonal sur $[l, u]$, pour le produit scalaire euclidien de $\mathbb{R}^m$.

Proposition 19.4 (existence de solution) *L'ensemble $\mathcal{S}$ est un polyèdre convexe non vide. Les problèmes (19.9), (19.10), (19.11) et (19.12) ont tous une solution. Leurs solutions $(\bar{x}, \bar{y}, \bar{s})$ se correspondent et on a*

$$\bar{s} = \bar{y} - A\bar{x} \qquad et \qquad \bar{y} = \mathrm{P}_{[l,u]}(A\bar{x}). \tag{19.14}$$

La plus petite translation réalisable $\bar{s}$ est déterminée de manière unique.

Démonstration. L'ensemble $\mathcal{S}$ est un polyèdre convexe non vide, comme somme de deux polyèdres convexes non vides.

L'existence et l'unicité de la plus petite translation réalisable $\bar{s}$, solution de (19.13), vient du fait qu'elle est la projection de zéro sur le convexe fermé non vide $\mathcal{S}$ (proposition 2.25).

L'existence de solution pour les problèmes (19.9), (19.10) et (19.11) s'en déduit, de même que la correspondance entre les solutions, $\bar{s} = \bar{y} - A\bar{x}$ et $\bar{y} = \mathrm{P}_{[l,u]}(A\bar{x})$. □

L'unicité du problème de moindres-carrés polyédrique, caractérisée dans la proposition 19.5 ci-dessous a un lien étroit avec l'unicité des solutions d'un système

d'inégalités linéaires : on trouve une caractérisation de l'unicité semblable à celle de la proposition 2.51 et même identique si l'inclusion (19.7) est réalisable. La caractérisation donnée dans la proposition ci-dessous généralise aussi celle de l'unicité de solution du problème de moindres-carrés linéaire, exprimée par l'injectivité de la matrice. Nous reviendrons sur ces deux points après la démonstration de la proposition. Dans cette proposition, on a noté $\mathrm{T}_{[l,u]}(y)$ le cône tangent à $[l,u]$ en $y \in [l,u]$ et $\mathrm{N}_{[l,u]}(y)$ le cône normal à $[l,u]$ en $y \in [l,u]$.

Proposition 19.5 (unicité de solution) *Soit $\bar{x}$ une solution du problème (19.9). Alors les propriétés suivantes sont équivalentes :*
(i) $\bar{x}$ est solution unique de (19.9),
(ii) tout d tel que $Ad \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))$ est nul,
(iii) $A^{\mathsf{T}}(\mathrm{N}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))) = \mathbb{R}^n$.

Démonstration. [$(i) \Rightarrow (ii)$] Soit $d \in \mathbb{R}^n$ tel que $Ad \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))$. Comme l'intervalle $[l,u]$ est polyédrique son cône tangent se confond avec son cône des directions admissibles (exercice 2.43), si bien que $\mathrm{P}_{[l,u]}(A\bar{x}) + tAd \in [l,u]$ pour $t > 0$ assez petit. Ceci a pour conséquence que $\bar{x}_t := \bar{x} + td$, avec $t > 0$ assez petit, vérifie

$$\mathrm{d}_{[l,u]}(A\bar{x}_t) \leqslant \|[\mathrm{P}_{[l,u]}(A\bar{x}) + tAd] - A\bar{x}_t\|_2 = \|\,\mathrm{P}_{[l,u]}(A\bar{x}) - A\bar{x}\|_2 = \mathrm{d}_{[l,u]}(A\bar{x}).$$

Comme $\bar{x}$ est solution de (19.9), $\bar{x}_t$ l'est aussi. L'unicité de la solution affirmée par (i) implique alors que $\bar{x}_t = \bar{x}$, c'est-à-dire que $d = 0$.

[$(ii) \Rightarrow (i)$] Soit $\bar{x}'$ une solution de (19.9) ; on cherche à montrer que $\bar{x}' = \bar{x}$. Par l'unicité de $\bar{s}$ (proposition 19.4), il existe des $\bar{y}$ et $\bar{y}' \in [l,u]$ tels que $\bar{s} = \bar{y} - A\bar{x} = \bar{y}' - A\bar{x}'$. Dès lors, $A(\bar{x}' - \bar{x}) = \bar{y}' - \bar{y}$. Comme $\bar{y} = \mathrm{P}_{[l,u]}(A\bar{x})$ (proposition 19.4) et $\bar{y}' \in [l,u]$, $\bar{y}' - \bar{y} \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))$. Alors $A(\bar{x}' - \bar{x}) \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))$ et (ii) impliquent que $\bar{x}' = \bar{x}$.

[$(ii) \Leftrightarrow (iii)$] On a

$$\begin{aligned}
(ii) &\iff \{0\} = \{d : Ad \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))\} \\
&\iff \mathbb{R}^n = \{d : Ad \in \mathrm{T}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))\}^+ \qquad (19.15)\\
&\iff \mathbb{R}^n = -A^{\mathsf{T}}(\mathrm{N}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x}))) \qquad (19.16)\\
&\iff (iii),
\end{aligned}$$

où en (19.15) on a pris le dual de chaque membre et en (19.16) on a utilisé le lemme de Farkas (proposition 2.45), la relation de dualité reliant le cône tangent et le cône normal (point 4 de la proposition 2.55) et le caractère fermé de $A^{\mathsf{T}}(\mathrm{N}_{[l,u]}(\mathrm{P}_{[l,u]}(A\bar{x})))$ (image linéaire d'un polyèdre convexe ; proposition 2.18 et exercice 2.43). □

Voyons quelles formes prennent les conditions nécessaires et suffisantes d'unicité de la proposition 19.5 dans deux cas particuliers.

- Dans le problème de moindres-carrés linéaire (19.1), on cherche à réaliser au mieux le système linéaire $Ax = b$. On retrouve ce problème avec (19.9) en prenant pour intervalle $[l,u]$ le singleton $\{b\}$. Dans ce cas, le cône tangent $\mathrm{T}_{\{b\}}(\mathrm{P}_{\{b\}}(A\bar{x})) =$

$\mathrm{T}_{\{b\}}(b)$ est réduit à $\{0\}$ et le cône normal $\mathrm{N}_{\{b\}}(\mathrm{P}_{\{b\}}(A\bar{x})) = \mathrm{N}_{\{b\}}(b)$ est $\mathbb{R}^m$. Alors la condition (ii) exprime l'injectivité de A et (iii) la surjectivité de A^T, ce qui est bien la condition nécessaire et suffisante d'unicité révélée par la proposition 19.1.

On notera que la condition nécessaire et suffisante d'unicité de la solution du problème de moindres-carrés linéaire ne dépend que de A, alors que celle de la proposition 19.5 fait intervenir une solution $\bar{x}$.

- Si l'on cherche à réaliser au mieux un système d'inéquations $Ax \leqslant b$ par l'approche décrite à la section 19.2.1, on prendra $[l, u] =]-\infty, b]$ dans (19.7). Alors $\mathrm{d}_{]-\infty,b]}(Ax) = \mathrm{d}_{\mathbb{R}^m_-}(Ax-b) = \|\mathrm{P}_{\mathbb{R}^m_-}(Ax-b) - (Ax-b)\|_2 = \|(Ax-b)^- + (Ax-b)\|_2 = \|(Ax-b)^+\|_2$, si bien que le problème (19.9) s'écrit dans ce cas

$$\min_{x\in\mathbb{R}^n} \|(Ax-b)^+\|_2. \tag{19.17}$$

Ce problème a donc toujours une solution (proposition 19.4).

Pour déterminer les conditions d'unicité de la solution de ce problème au moyen de la proposition 19.5, on observe que $\mathrm{P}_{]-\infty,b]}(A\bar{x}) = b + \mathrm{P}_{\mathbb{R}^m_-}(A\bar{x}-b)$ ou

$$\mathrm{P}_{]-\infty,b]}(A\bar{x}) = b - (A\bar{x}-b)^-. \tag{19.18}$$

Alors $\mathrm{T}_{]-\infty,b]}(\mathrm{P}_{]-\infty,b]}(A\bar{x})) = \mathrm{T}_{\mathbb{R}^m_-}(-(A\bar{x}-b)^-) = \{d \in \mathbb{R}^n : A_I d \leqslant 0\}$, où I est l'ensemble des indices $i \in [1:m]$ tels que $(A\bar{x}-b)_i^- = 0$, c'est-à-dire

$$I = \{i \in [1:m] : (A\bar{x}-b)_i \geqslant 0\}.$$

Dès lors, d'après la proposition 19.5, le problème (19.17) a une solution unique si, et seulement si, l'une des conditions équivalentes suivantes a lieu :

$$\begin{gathered}\text{tout } d \in \mathbb{R}^n \text{ tel que } A_I d \leqslant 0 \text{ est nul,} \\ A_I^\mathsf{T}(\mathbb{R}_+^{|I|}) = \mathbb{R}^n.\end{gathered} \tag{19.19}$$

Si le coût optimal de (19.17) est nul, l'ensemble des solutions du problème est le polyèdre convexe $\{x \in \mathbb{R}^n : Ax \leqslant b\}$. On sait par la proposition 2.51 que celui-ci est le singleton $\{\bar{x}\}$ si, et seulement si, tout vecteur d vérifiant $A_I d \leqslant 0$ est nul, condition dans laquelle $I := \{i \in [1:m] : (A\bar{x}-b)_i = 0\}$. C'est bien ce qu'affirme (19.19), car les deux ensembles d'indices I sont identiques dans ce cas.

Évidemment la condition (ii) de la proposition est plus forte que l'injectivité de A et la condition (iii) de la proposition est plus forte que la surjectivité de A ; ainsi (ii) ou $(iii) \Rightarrow A$ injective ou A^T surjective, mais la réciproque n'est pas garantie. Comme contre-exemple, considérons le problème (19.17) avec $m = n = 1$, $A = 1$ et $b = 0$, qui consiste à minimiser x^+ sur $\mathbb{R}$: A est injective (ou $A^\mathsf{T} = A$ est surjective), mais l'ensemble des solutions $\mathbb{R}_-$ n'est pas un singleton, donc (ii) n'a pas lieu [en effet $\mathbb{R}_- \neq \{0\}$] (ou (iii) n'a pas lieu [en effet $\mathbb{R}_+ \neq \mathbb{R}$]).

19.2.3 Condition d'optimalité

Les problèmes (19.9), (19.10), (19.11) et (19.12) sont convexes. On peut donc en donner des conditions nécessaires et suffisantes d'optimalité. On note $\mathrm{Sol}(P)$ l'ensemble des solutions du problème P, tout en évitant les doubles parenthèses.

Proposition 19.6 (conditions d'optimalité) *Les conditions nécessaires et suffisantes d'optimalité suivantes ont lieu :*

$$\bar{x} \in \mathrm{Sol}\,(19.9) \iff A^{\mathsf{T}}\big(\mathrm{P}_{[l,u]}(A\bar{x}) - A\bar{x}\big) = 0, \tag{19.20}$$

$$(\bar{x},\bar{y}) \in \mathrm{Sol}\,(19.10) \iff A^{\mathsf{T}}(\bar{y} - A\bar{x}) = 0 \text{ et } \bar{y} = \mathrm{P}_{[l,u]}(A\bar{x}), \tag{19.21}$$

$$(\bar{x},\bar{s}) \in \mathrm{Sol}\,(19.11) \iff A^{\mathsf{T}}\bar{s} = 0 \text{ et } \bar{s} = \mathrm{P}_{[l,u]}(A\bar{x}) - A\bar{x}, \tag{19.22}$$

$$\bar{s} \in \mathrm{Sol}\,(19.12) \iff \bar{s} = \mathrm{P}_{[l,u]+\mathcal{R}(A)}(0). \tag{19.23}$$

Démonstration. On ne change pas le problème (19.9) en élevant son critère au carré. Ce nouveau critère est alors convexe différentiable avec un gradient pour le produit scalaire euclidien qui, en $\bar{x}$, vaut $2A^{\mathsf{T}}(\mathrm{P}_{[l,u]}(A\bar{x}) - A\bar{x})$. La condition nécessaire et suffisante d'optimalité (19.20) découle alors de la condition de Fermat (proposition 4.9).

Les autres conditions d'optimalité (19.21) et (19.22) se déduisent de (19.20) et du fait que les solutions des problèmes (19.9), (19.10) et (19.11) se correspondent avec les relations (19.14). La condition d'optimalité (19.23) est immédiate. □

Appliquons ce résultat au problème (19.17). D'après (19.18), $A\bar{x} - \mathrm{P}_{]-\infty,b]}(A\bar{x}) = (A\bar{x} - b)^+$, si bien que

$$\bar{x} \text{ est solution de (19.17)} \iff A^{\mathsf{T}}(A\bar{x} - b)^+ = 0,$$

ce que l'on aurait pu obtenir directement en annulant la dérivée de la fonction différentiable $x \in \mathbb{R}^n \mapsto \|(Ax - b)^+\|_2^2$, comme dans la démonstration de la proposition.

19.3 Moindres-carrés non linéaire

19.3.1 Définition du problème

Soit $r : \mathbb{R}^n \to \mathbb{R}^m$ une fonction régulière, que l'on appelle ci-dessous le *résidu*. On cherche à annuler celui-ci au sens des moindres carrés, c'est-à-dire :

$$\min_{x\in\mathbb{R}^n} \Big(f(x) = \frac{1}{2}\|r(x)\|^2\Big), \tag{19.24}$$

où $\|\cdot\|$ désigne la norme ℓ_2. Un problème de ce type porte le nom de *problème de moindres-carrés non linéaire*, parce qu'on y minimise la somme des carrés des résidus non linéaires r_i. C'est une version non linéaire du problème de moindres-carrés linéaire (19.1), puisque l'on retrouve ce dernier en définissant r en x par $r(x) = Ax - b$. En pratique, on a $m \gg n$, mais, sauf mention contraire, nous ne faisons pas d'hypothèse systématique sur la grandeur relative de m et n ci-dessous.

On note

$$J(x) = r'(x)$$

la jacobienne de r en x, qui est une matrice de type $m \times n$. On calcule aisément le gradient et la hessienne du critère de (19.24) pour le produit scalaire euclidien :

$$\nabla f(x) = J(x)^{\mathsf{T}} r(x) \quad \text{et} \quad \nabla^2 f(x) = J(x)^{\mathsf{T}} J(x) + \sum_{i=1}^{p} r_i(x) \nabla^2 r_i(x). \tag{19.25}$$

Dans un contexte où une suite $\{x_k\} \subseteq \mathbb{R}^n$ est définie, on note encore

$$r_k := r(x_k), \qquad g_k := \nabla f(x_k) \qquad \text{et} \qquad J_k := J(x_k). \tag{19.26}$$

19.3.2 Algorithme de Gauss-Newton

L'algorithme et sa cohérence

L'application de la méthode de Newton pour minimiser la fonction f définie en (19.24) requiert le calcul de la hessienne $\nabla^2 f(x)$ et donc, comme le montre la formule (19.25), le calcul des dérivées secondes du résidu. Pour certains problèmes de moindres-carrés, ce calcul peut être très coûteux, si bien que l'on s'intéresse à l'algorithme qui néglige ces termes de la hessienne et calcule donc en x_k une direction d_k en résolvant le système linéaire suivant :

$$\boxed{\left(J_k^{\mathsf{T}} J_k\right) d_k = -J_k^{\mathsf{T}} r_k.} \tag{19.27}$$

On a noté $J_k = J(x_k)$ et $r_k = r(x_k)$ (il y a une petite ambiguïté de notation — r_k n'est pas la k-ième composant du résidu — mais celle-ci sera toujours facilement levée par le contexte). La perte d'efficacité locale par rapport à l'algorithme de Newton sera donc faible si le terme négligé $\sum_i r_i(x) \nabla^2 r_i(x)$ de la hessienne de f est relativement petit, ce qui sera le cas si r est (presque) linéaire ou si $r(x)$ est petit (cas des problèmes d'identification).

On peut obtenir la même direction en raisonnant comme suit. On considère le modèle quadratique de f obtenu, non pas par son développement au deuxième ordre (ce serait l'algorithme de Newton et l'on cherche à simplifier celui-ci ici), mais en linéarisant le résidu à l'intérieur de la norme dans (19.24). Ceci donne :

$$\min_{d \in \mathbb{R}^n} \frac{1}{2} \|r_k + J_k d\|^2. \tag{19.28}$$

On observe que ce problème définit les mêmes directions d_k que précédemment, car l'équation d'optimalité du premier ordre de (19.28) n'est autre que (19.27). De plus celle-ci est nécessaire et suffisante car le problème (19.28) est convexe.

Le résultat suivant montre qu'il est raisonnable de construire une méthode à direction de descente en prenant la direction d_k définie par (19.27) ou (19.28) comme direction le long desquelles on se déplace.

Proposition 19.7 (direction de descente de Gauss-Newton) *Le système linéaire (19.27) a une solution d_k. Si x_k n'est pas un point stationnaire du problème de moindres-carrés non linéaire (19.24), alors d_k est une direction de descente de f en x_k.*

Démonstration. Le problème (19.28) a toujours au moins une solution car c'est un problème de moindres-carrés linéaire (proposition 19.1). D'autre part, en prenant le produit scalaire des deux membres de (19.27) avec d_k, on trouve

$$g_k^T d_k = -\|J_k d_k\|^2,$$

qui est strictement négatif lorsque x_k n'est pas stationnaire (car alors $J_k^T r_k \neq 0$ et donc $J_k d_k \neq 0$ par (19.27)). □

Ceci nous conduit à l'*algorithme de Gauss-Newton*, dont nous décrivons une itération ci-dessous.

Algorithme 19.8 (Gauss-Newton) L'algorithme calcule l'itéré x_{k+1} à partir de l'étéré x_k de la manière suivante.

1. *Test d'arrêt.* Si $J_k^T r_k \simeq 0$, arrêt de l'algorithme.
2. *Direction de descente.* Calculer une solution d_k de (19.27).
3. *Recherche linéaire.* Trouver un pas $\alpha_k > 0$ en x_k le long de d_k par une règle de recherche linéaire satisfaisant la condition de Zoutendijk (6.19) ou (6.26a).
4. *Nouvel itéré.* $x_{k+1} := x_k + \alpha_k d_k$.

À l'étape 3, une règle de recherche linéaire appropriée est celle d'Armijo ou de Goldstein (voir la section 6.3.3).

Convergence globale et complexité itérative

Des conditions de convergence globale et de complexité itérative globale de cet algorithme sont données dans le théorème qui suit. On y requiert une hypothèse assez forte, qui est celle de l'*injectivité uniforme* de la suite $\{J_k\}$, ce qui veut dire qu'il existe une constante $\alpha_J > 0$ telle que pour tout indice $k \geqslant 1$ et tout vecteur $v \in \mathbb{R}^n$, on ait

$$\|J_k v\| \geqslant \alpha_J \|v\|. \tag{19.29}$$

Dans ce cas, la direction de Gauss-Newton est déterminée de manière unique par (19.27) et la convergence globale *vers un point stationnaire* de f est assurée (point 1) avec une complexité itérative en $O(\varepsilon^{-2})$ (point 2) ; cette complexité ne dépend pas des dimensions n et m du problème.

Les résultats s'améliorent beaucoup s'il s'avère que $\{J_k^T\}$ est uniformément injective, puisqu'alors, l'algorithme assure la convergence de $\{r_k\}$ vers zéro (point 3), avec une complexité itérative en $O(\log \varepsilon^{-1})$ (point 4), ce qui est spectaculairement mieux ; cette complexité ne dépend pas de n et m. D'ailleurs, dans ce cas, on doit avoir $m = n$ (car J_k est déclarée bijective) et l'on cherche donc un zéro du système de n équations à n inconnues $r(x) = 0$.

Théorème 19.9 (convergence et complexité itérative de l'algorithme de Gauss-Newton) *Supposons que f donnée par (19.24) soit $\mathcal{C}^{1,1}$ dans un voisinage de $\mathcal{N}_1 := \{x \in \mathbb{R}^n : f(x) \leqslant f(x_1)\}$. Soit $\{x_k\}$ une suite générée par l'algorithme 19.8 de Gauss-Newton, telle que $\{J_k\}$ soit bornée et uniformément injective. Alors,*

1) $J_k^T r_k \to 0$,
2) *il existe une constante $C > 0$, indépendante de n et m, telle que, pour tout $\varepsilon > 0$, $\|J_k^T r_k\| \leqslant \varepsilon$ pour un indice k inférieur à $\lceil C\varepsilon^{-2} \rceil$,*
3) *si, de plus, $\{J_k^T\}$ est uniformément injective, alors*
   a) *$f(x_k) \to 0$ linéairement,*
   b) *il existe une constante $C' > 0$, indépendante de $n = m$, telle que, pour tout $\varepsilon > 0$, $\|r_k\| \leqslant \varepsilon \|r_0\|$ pour tout indice $k \geqslant \lceil C' \log \varepsilon^{-1} \rceil$.*

Démonstration. 1) Si la condition de Zoutendijk (6.19) a lieu, on utilise la proposition 6.8 et sa conclusion (6.21). Pour cela, on observe que $f_* := \inf_k f(x_k) \in [0, f(x_1)]$ est bien fini et que le cosinus de l'angle θ_k entre d_k et $-g_k$ est minoré par une constante strictement positive :

$$\cos\theta_k := \frac{-g_k^T d_k}{\|g_k\|\,\|d_k\|} = \frac{\|J_k d_k\|^2}{\|J_k^T r_k\|\,\|d_k\|} \geqslant \frac{\alpha_J \|J_k d_k\|}{\|J_k^T J_k d_k\|} \geqslant \frac{\alpha_J}{\beta_J}, \tag{19.30}$$

où l'on a utilisé (19.29) (injectivité uniforme de $\{J_k\}$) et la borne β_J sur $\|J_k^T\| = \|J_k\|$. Alors, (6.21) implique que $g_k = J_k^T r_k \to 0$.

Si la condition (6.26a) a lieu, $|g_k^T d_k| = \|J_k d_k\|^2$ doit tendre vers zéro (car $f(x_k)$ et $f(x_{k+1})$ tendent vers la même valeur) et comme $\|J_k d_k\| \geqslant \alpha_J \|d_k\|$ par (19.29) (injectivité uniforme de $\{J_k\}$), on a $d_k \to 0$. Par (19.27) et la bornitude de $\{J_k\}$, ceci implique aussi $g_k = J_k^T r_k \to 0$.

2) Si la condition de Zoutendijk (6.19) a lieu, le résultat se déduit de la seconde partie de la proposition 6.8 qui affirme que, pour tout $\varepsilon > 0$, $\|J_k^T r_k\| \leqslant \varepsilon$ pour au moins un indice k inférieur au K_ε donné par (6.22), avec $\gamma = \alpha_J/\beta_J$.

Considérons maintenant le cas où la condition (6.26a) a lieu. Par (19.27) et l'existence d'une borne β_J sur $\|J_k^T\| = \|J_k\|$, on a

$$\|g_k\|^2 = \|J_k^T J_k d_k\|^2 \leqslant \beta_J^2 \|J_k d_k\|^2 = \beta_J^2 |g_k^T d_k|.$$

Alors la condition (6.26a) se récrit

$$f(x_{k+1}) \leqslant f(x_k) - C\beta_J^{-2}\|g_k\|^2 \qquad \text{ou} \qquad C\beta_J^{-2}\|g_k\|^2 \leqslant f(x_k) - f(x_{k+1}). \tag{19.31}$$

En sommant de $k = 1$ à K, on trouve

$$\min_{k \in [1:K]} \|g_k\|^2 \leqslant \frac{1}{K} \sum_{k=1}^{K} \|g_k\|^2 \leqslant \frac{f(x_1) - f_*}{C\beta_J^{-2} K}.$$

Donc un des $\|g_k\|$, avec $k \in [1:K]$, est inférieur à $\varepsilon > 0$ si le membre de droite est inférieur à ε^2, c'est-à-dire si $K \geqslant C^{-1}\beta_J^2 (f(x_1) - f_*)\varepsilon^{-2}$.

3) Pour une constante strictement positive C_1, on a

$$f(x_{k+1}) \leqslant f(x_k) - C_1\|g_k\|^2,$$

soit par la condition de Zoutendijk (6.19) et (19.30), soit par la condition (6.26a) et (19.31). Alors, en utilisant $g_k = J_k^T r_k$ et $\|J_k^T r_k\|^2 \geqslant (\alpha'_J)^2\|r_k\|^2 = 2(\alpha'_J)^2 f(x_k)$ par l'injectivité uniforme de J_k^T, on trouve

$$f(x_{k+1}) \leqslant (1 - C_2) f(x_k),$$

pour la constante strictement positive $C_2 = 2C_1(\alpha'_J)^2$. Ceci montre la convergence linéaire de $f(x_k)$ vers zéro (point 3.a). Puis par récurrence,

$$\frac{f(x_k)}{f(x_0)} \leqslant (1 - C_2)^k.$$

Dès lors, pour un $\varepsilon > 0$ donné, $\|r_k\| \leqslant \varepsilon\|r_0\|$ si $f(x_k)/f(x_0) \leqslant \varepsilon^2$ et donc certainement si le membre de droite ci-dessus est inférieur à ε^2 :

$$(1 - C_2)^k \leqslant \varepsilon^2.$$

En prenant les logarithmes et en tenant compte du fait que $\log(1-C_2) < 0$, on obtient comme condition sur k :

$$k \geqslant \frac{2\log\varepsilon^{-1}}{\log(1 - C_2)^{-1}}.$$

On obtient donc le résultat du point 3.b avec la constante $C' := 2/\log(1-C_2)^{-1}$. □

Cette courte analyse montre que, contrairement à la méthode de Newton, l'algorithme de Gauss-Newton est bien défini à tout itéré x_k (la direction d_k existe toujours) et, pourvu que x_k ne soit pas un point stationnaire de la fonction de moindres carrés f définie en (19.24), la direction calculée d_k est de descente pour f. Il demande toutefois de résoudre le système linéaire (19.27), qui est moins bien conditionné que (10.3). Sa convergence est assurée sous des conditions souvent vérifiées, mais pas toujours. L'algorithme de Gauss-Newton peut en effet très bien générer une suite convergeant vers un point auquel f n'est pas stationnaire, a fortiori n'annulant pas le résidu. De plus, la vitesse de convergence peut être assez lente si le résidu n'est pas nul en la solution ou si r y est fortement non linéaire (r'' non nul), car l'algorithme s'écarte alors de la méthode de Newton par un terme significativement non nul $\nabla^2 f(x) - J(x)^T J(x) = \sum_i r_i(x)\nabla^2 r_i(x)$ (voir (19.25)).

L'algorithme de Levenberg-Morrison-Marquardt de la section suivante apporte quelques remèdes à ces défauts ; en particulier, il permet de se défaire de la condition d'injectivité uniforme de $\{J_k\}$.

19.3.3 Algorithme de Levenberg-Morrison-Marquardt

L'algorithme et sa cohérence

L'algorithme de Levenberg-Morrison-Marquardt (LMM) cherche à résoudre le problème de moindres-carrés non linéaire (19.1) en générant une suite $\{x_k\} \subseteq \mathbb{R}^n$

de la manière suivante. Il définit le déplacement $s_k \in \mathbb{R}^n$ de l'itéré courant x_k à l'itéré suivant $x_{k+1} := x_k + s_k$ en résolvant le système linéaire

$$\boxed{\left(J_k^T J_k + \lambda_k M_k\right) s_k = -J_k^T r_k,} \tag{19.32}$$

où l'on a utilisé les notations (19.26). Dans ce système, $M_k \in \mathcal{S}^n_{++}$ est une matrice symétrique définie positive, dont le rôle et les règles qu'elle doit vérifier seront précisés plus loin, et $\lambda_k > 0$ est un paramètre, appelé tantôt *facteur de pénalisation*, tantôt *multiplicateur*, selon l'interprétation que l'on en fait. Ce multiplicateur joue un rôle important dans l'algorithme, qui le détermine lui-même à chaque itération, alors que M_k peut être laissé au choix de l'utilisateur. Ce multiplicateur est l'élément nouveau par rapport à la direction de l'algorithme de Gauss-Newton, calculée comme solution de (19.27), qui n'est autre que (19.32) avec $\lambda_k = 0$. Remarquons qu'ici, grâce à la stricte positivité de λ_k et la définie positivité de M_k, la matrice du système linéaire (19.32) est symétrique définie positive, si bien que s_k y est déterminé de manière unique. Lorsqu'il sera important de mentionner la dépendance en λ_k de la solution de (19.32), on notera celle-ci

$$s_k(\lambda_k).$$

Insistons sur le fait que s_k est le déplacement de x_k à x_{k+1} et pas une direction de déplacement le long de laquelle on ferait un pas pouvant être différent de un ; pour le dire autrement, l'algorithme présenté ci-dessous n'utilise pas de recherche linéaire pour déterminer l'itéré suivant, comme c'est le cas pour l'algorithme de Gauss-Newton de la section 19.3.2. Sur ce plan, l'algorithme s'apparente davantage aux méthodes à régions de confiance. En particulier, comme nous allons le voir, c'est en augmentant λ_k que l'on diminue la grandeur du déplacement $s_k(\lambda_k)$, ce qui explique pourquoi c'est l'algorithme lui-même qui doit prendre en charge la détermination de λ_k. C'est aussi ce réglage du multiplicateur λ_k qui permettra de faire décroître f suffisamment à chaque itération.

On voit que si $g_k = 0$, le déplacement donné par (19.32) est nul, si bien que l'algorithme ne peut pas progresser. Par conséquent, l'algorithme de LMM ne peut pas trouver mieux qu'un point stationnaire de la fonction de moindres-carrés f, mais un tel point peut être un zéro de r si sa jacobienne y est injective. C'est donc en ces termes que se formuleront le test d'arrêt de l'algorithme 19.10 et le résultat de convergence globale du théorème 19.12.

Il est utile de constater que le déplacement s_k de l'algorithme de LMM, solution de l'équation (19.32), est aussi solution du problème de moindres-carrés linéaire *pénalisé* suivant

$$\min_{s \in \mathbb{R}^n} \frac{1}{2} \|r_k + J_k s\|^2 + \frac{\lambda_k}{2} s^T M_k s, \tag{19.33}$$

où $\|\cdot\|$ désigne la norme ℓ_2 dans $\mathbb{R}^m$ (on pourra comparer ce problème d'optimisation quadratique associé à celui (19.28) qui était associé au système de Gauss-Newton (19.27)). En effet, la condition nécessaire et suffisante d'optimalité de ce problème quadratique convexe, c'est-à-dire la nullité du gradient de son critère par rapport à s, n'est autre que (19.32). De ce point de vue, λ est un *facteur de pénalisation*. Il sera

utile d'introduire une notation pour la fonction qui intervient dans le critère de ce problème, qui est la fonction $\varphi_k : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ définie en $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}$ par

$$\varphi_k(s, \lambda) := \frac{1}{2} \|J_k s + r_k\|^2 + \frac{\lambda}{2} s^T M_k s.$$

Au terme $-\lambda\Delta^2/2$ indépendant de s près, cette fonction est le lagrangien du problème

$$\begin{cases} \min_{s \in \mathbb{R}^n} \frac{1}{2} \|J_k s + r_k\|^2 \\ s^T M_k s \leqslant \Delta^2. \end{cases} \tag{19.34}$$

où $\Delta > 0$ est le *rayon de confiance* du problème (voir le chapitre 9). De ce point de vue, λ est un *multiplicateur* associé à la contrainte de région de confiance $\{s \in \mathbb{R}^n : s^T M_k s \leqslant \Delta^2\}$ de ce problème.

On déduit des points de vue présentés dans le paragraphe précédent que, $\|s_k(\lambda)\|$ décroît lorsque λ croît (proposition 13.2). De plus (proposition 19.11 et exercice 19.1), lorsque $\lambda \to \infty$ et $g_k \neq 0$, on a $s_k(\lambda) \to 0$ et $\lambda M_k s_k(\lambda) \to -g_k$, si bien que, dans ces circonstances, $s_k(\lambda)$ s'aligne sur $-M_k^{-1} g_k$, dans le sens où

$$s_k(\lambda) = -\frac{\|s_k(\lambda)\|}{\|M_k^{-1} g_k\|} M_k^{-1} g_k + o(\|s_k(\lambda)\|), \qquad \text{lorsque } \lambda \to \infty.$$

Par la différentiabilité supposée de la fonction f définie en (19.24), on en déduit que

$$f(x_k + s_k(\lambda)) = f(x_k) - \frac{g_k^T M_k^{-1} g_k}{\|M_k^{-1} g_k\|} \|s_k(\lambda)\| + o(\|s_k(\lambda)\|), \quad \text{lorsque } \lambda \to \infty.$$

Dès lors, toujours si $g_k \neq 0$, $f(x_k + s_k(\lambda)) < f(x_k)$ si λ est pris assez grand. C'est en ajustant λ à chaque itération que l'algorithme de LMM assure la décroissance suffisante de f et la convergence de g_k vers zéro.

Il y a plusieurs manières de déterminer λ_k à chaque itération. Nous présentons ci-dessous celle proposée par Osborne [465] qui prend racine sur la méthode des régions de confiance. Son intérêt est de ne pas devoir résoudre le problème de région de confiance (19.34) à chaque itération, lequel peut être coûteux pour certains problèmes. L'idée est similaire à celle utilisée par les régions de confiance et fait entrer en jeu le rapport ρ_k entre la décroissance réelle de f apportée par le déplacement $s_k(\lambda)$, pour un certain $\lambda > 0$, et la décroissance prédite par le modèle φ_k :

$$\rho_k(\lambda) := \frac{1}{2} \frac{f(x_k) - f(x_k + s_k(\lambda))}{f(x_k) - \varphi_k(s_k(\lambda), \lambda)}.$$

Le facteur $1/2$ est placé pour que la variation au premier ordre du numérateur, qui vaut $-g_k^T s_k$, soit identique à celle du dénominateur, si bien que fonction et modèle sont alors « tangents ». Le dénominateur vaut en effet

$$\begin{aligned} 2[f(x_k) - \varphi_k(s_k(\lambda), \lambda)] &= -2\, r_k^T J_k s_k(\lambda) - \|J_k s_k(\lambda)\|^2 - \lambda\, s_k(\lambda)^T M_k s_k(\lambda) \\ &= -g_k^T s_k(\lambda), \end{aligned} \tag{19.35}$$

où l'on a utilisé $g_k = J_k^T r_k$ et (19.32) pour obtenir la dernière égalité. Comme g_k est supposé non nul au cours de l'itération (l'algorithme s'arrête dès qu'il trouve

un point stationnaire de f) et que $s_k(\lambda)$ minimise $\varphi_k(\cdot,\lambda)$, le dénominateur de ρ_k est strictement positif. En effet, par (19.32), $s_k(\lambda)$ est non nul lorsque $g_k \neq 0$ et comme $s_k(\lambda)$ est l'unique point qui minimise le critère de (19.33), on a nécessairement $\varphi_k(s_k(\lambda),\lambda) < \varphi_k(0,\lambda) = f(x_k)$. Au passage, nous avons montré que, quel que soit $\lambda > 0$,

$$g_k^T s_k(\lambda) < 0.$$

Alors, si le rapport $\rho_k(\lambda)$ est supérieur à un seuil préfixé $\eta_1 \in \mathbb{R}_{++}$, l'on a grâce à (19.35)

$$f(x_k + s_k(\lambda)) \leqslant f(x_k) + \eta_1\, g_k^T s_k(\lambda), \tag{19.36}$$

qui rappelle la condition de décroissance suffisante (6.9) utilisée en recherche linéaire, sauf qu'ici le pas α_k est remplacé par le paramètre λ qui ajuste la grandeur de $s_k(\lambda)$. On considère qu'une décroissance suffisante de f est obtenue lorsque cette inégalité est vérifiée pour un certain $\lambda > 0$ (cet considération sera validée par le résultat de convergence ci-dessous) et on accepte $s_k(\lambda)$ comme déplacement de l'itération k. Dans le cas contraire, tant que l'inégalité ci-dessus n'est pas vérifiée, on augmente λ et on résout à nouveau (19.32) ; on montrera par la proposition 19.11 que cette procédure de réglage du λ se conclut en un nombre fini d'étapes.

Nous avons à présent sous la main et à l'esprit, tous les éléments permettant d'énoncer et de comprendre *une itération* de l'algorithme de LMM, dans la version d'Osborne [465]. On suppose qu'au début de l'itération k, l'on dispose d'un itéré $x_k \in \mathbb{R}^n$, d'un multiplicateur $\lambda_{k-1} > 0$ et d'une matrice $M_k \in \mathcal{S}^n_{++}$. L'itéré x_k et le multiplicateur λ_{k-1} sont mis à jour par l'itération et une nouvelle matrice M_k pourra être choisie en fin d'itération. Par constante, la description ci-dessous entend une valeur qui ne dépend pas de l'itération.

Algorithme 19.10 (Levenberg-Morrison-Marquardt revisité)
L'algorithme utilise les constantes suivantes : $0 < \tau_1 < 1 < \tau_2$ pour la mise à jour de λ_{k-1} et $0 < \eta_1 < \eta_2 < 1$ comme seuils de satisfaction de la décroissance de f. On suppose que $M_k \in \mathcal{S}^n_{++}$. Une itération de l'algorithme se déroule comme suit.

1. *Test d'arrêt.* Si $J_k^T r_k \simeq 0$, arrêt de l'algorithme.
2. *Déplacement.* Prendre $\lambda_{k,0} := \lambda_{k-1}$ et répéter les opérations suivantes pour $i \in \mathbb{N}$ jusqu'à satisfaction du test de sortie (19.38).
   2.1. Calculer la solution $s_{k,i}$ du système linéaire
   $$\left(J_k^T J_k + \lambda_{k,i} M_k\right) s_{k,i} = -J_k^T r_k, \tag{19.37}$$
   2.2. Si
   $$f(x_k + s_{k,i}) \leqslant f(x_k) + \eta_1 g_k^T s_{k,i}, \tag{19.38}$$
   sortir de la boucle courante avec $s_k := s_{k,i}$ (on va à l'étape 3), sinon $\lambda_{k,i+1} = \tau_2 \lambda_{k,i}$.
3. *Nouveau facteur de pénalisation.* Si

$$f(x_k + s_k) \leqslant f(x_k) + \eta_2 g_k^\mathsf{T} s_k, \tag{19.39}$$

$\lambda_k := \tau_1 \lambda_{k,i}$, sinon $\lambda_k := \lambda_{k,i}$.

4. *Nouvel itéré.* $x_{k+1} := x_k + s_k$.
5. *Nouvelle matrice.* Choix de $M_{k+1} \in \mathcal{S}^n_{++}$.

Voici quelques remarques sur cet algorithme.

- Le coût de cet algorithme est principalement lié au nombre de systèmes linéaires (19.37) qu'il faut résoudre.
- Typiquement, on prendra $\eta_1 \simeq 10^{-4}$, comme pour la constante ω_1 de (6.9) en recherche linéaire, de manière à rendre le test de deecroissance suffisante (19.38) aussi peu contraignant que possible.
- L'étape 3 et la constante η_2 ne jouent pas de rôle dans la convergence mais sont introduites pour ne pas imposer la croissance de la suite $\{\lambda_k\}$, ce qui ne serait pas adéquat (le premier multiplicateur λ_1 choisi peut être trop grand et donc conduire à des déplacements s_k trop petits, ce qui ralentirait la convergence). Si la décroissance de f est forte au sens de l'inégalité (19.39), alors l'algorithme s'autorise prendre λ_k plus petit que $\lambda_{k,i}$ à l'itération suivante.
- Au lieu de régler λ_k, certains auteurs [597] préfèrent prendre le multiplicateur de la forme $\lambda_k = \mu_k \|r_k\|^\delta$, avec $\delta \geqslant 0$ ($\delta = 0$ dans notre cas), et ajuster μ_k au lieu de λ_k à chaque itération.

Le caractère *bien défini* de l'algorithme de LMM, c'est-à-dire le fait qu'il sorte de la boucle de l'étape 2 en un nombre fini de cycles, est suggéré par la discussion qui précède son énoncé. Avec la proposition suivante, nous le montrons de manière rigoureuse.

Proposition 19.11 (descente suffisante) *Si r est différentiable en x_k, si le gradient $g_k := J_k^\mathsf{T} r_k$ est non nul, si $M_k \in \mathcal{S}^n_{++}$ et si $\eta_1 \in]0,1[$, alors*

1) *$s_k(\lambda) \neq 0$ pour tout $\lambda \geqslant 0$,*
2) *$s_k(\lambda) \to 0$ lorsque $\lambda \to \infty$,*
3) *$s_k(\lambda)/\|s_k(\lambda)\| \to -M_k^{-1} g_k / \|M_k^{-1} g_k\|$ lorsque $\lambda \to \infty$,*
4) *(19.36) est vérifiée pour tout λ suffisamment grand.*

Démonstration. 1) Selon (19.32), $s_k(\lambda)$ est non nul parce que $g_k \neq 0$.

2) La convergence de $s_k(\lambda)$ vers zéro est due au fait que $s_k(\lambda)$ minimise $\varphi_k(\cdot, \lambda)$ et donc

$$0 \leqslant \frac{\lambda}{2} s_k(\lambda)^\mathsf{T} M_k s_k(\lambda) \leqslant \varphi_k(s_k(\lambda), \lambda) \leqslant \varphi_k(0, \lambda) = f(x_k).$$

En divisant certains membres par $\lambda > 0$ et en faisant tendre $\lambda \to \infty$, on voit que $s_k(\lambda)^\mathsf{T} M_k s_k(\lambda) \to 0$ et donc $s_k(\lambda) \to 0$, car $M_k \in \mathcal{S}^n_{++}$.

3) D'après (19.32) et $s_k(\lambda) \to 0$, on voit que $\lambda M_k s_k(\lambda) \to -g_k$, ce qui implique que $\lambda s_k(\lambda) \to -M_k^{-1} g_k$ et donc $s_k(\lambda)/\|s_k(\lambda)\| \to -M_k^{-1} g_k / \|M_k^{-1} g_k\|$.

4) On raisonne par l'absurde en supposant que (19.36) n'est pas vérifiée pour une suite de $\lambda \to \infty$. Alors, pour ces $\lambda \to \infty$, on a

$$\frac{f(x_k + s_k(\lambda)) - f(x_k) - g_k^T s_k(\lambda)}{\|s_k(\lambda)\|} > (1 - \eta_1) \frac{-g_k^T s_k(\lambda)}{\|s_k(\lambda)\|}.$$

Par la différentiabilité de f en x_k, le membre de gauche tend vers zéro. Par le point 3, le membre de droite tend vers $(1-\eta_1) g_k^T M_k^{-1} g_k / \|M_k^{-1} g_k\|$, qui est strictement positif. On a obtenu la contradiction souhaitée. □

Convergence globale et complexité itérative

On pourrait penser que le surcoût de l'algorithme 19.10, dû au besoin de devoir résoudre un nouveau système linéaire (19.37) chaque fois que l'inégalité de décroissance suffisante (19.38) n'est pas vérifiée, est trop important et qu'il serait préférable de faire de la recherche linéaire bien moins coûteuse le long de la première direction $s_{k,0}$ calculée si celle-ci n'est pas acceptée par (19.38). C'est ce que propose de faire certains auteurs. Cependant, l'algorithme décrit a l'avantage d'avoir un résultat de convergence globale, sans requérir l'injectivité uniforme des J_k, alors que celle-ci est requise par l'algorithme de Gauss-Newton (théorème 19.9). Seule intervient la bornitude de la suite $\{(J_k, \lambda_k M_k)\}$ (la bornitude de $\{\lambda_k M_k\}$ peut-être contrôlée par un choix d'implémentation de l'algorithme, mais pas celle de $\{J_k\}$ qui dépend du problème considéré). Comme annoncé dans la description de l'algorithme, sans hypothèse supplémentaire, sa convergence s'exprime en termes du gradient $g_k = J_k^T r_k$.

Le résultat qui demande le moins d'hypothèse est le suivant. Il suppose qu'une suite est générée par l'algorithme de LMM et donc que celui-ci ne trouve pas un point stationnaire de f en un nombre fini d'itérations.

Théorème 19.12 (convergence de l'algorithme LMM) *Supposons que la fonction de moindres-carrés f soit différentiable. Soit $\{(x_k, \lambda_k)\}$ une suite générée par l'algorithme 19.10. Alors,*

1) *$\{f(x_k)\}$ converge,*
2) *pour toute partie infinie $\mathcal{K}$ de $\mathbb{N}$ telle que $\{(J_k, \lambda_k M_k)\}_{k \in \mathcal{K}}$ est bornée, on a $\{J_k^T r_k\}_{k \in \mathcal{K}} \to 0$ lorsque $k \to \infty$ dans $\mathcal{K}$.*

Démonstration. 1) La convergence de la suite $\{f(x_k)\}$ découle de sa décroissance et du fait qu'elle est bornée inférieurement (par zéro).

2) D'après l'inégalité (19.38) et la convergence de $f(x_k)$, on a

$$g_k^T s_k \to 0, \qquad \text{lorsque } k \to \infty. \tag{19.40}$$

Il s'agit à présent de montrer que la convergence dans (19.40) est due à un gradient g_k qui tend vers zéro et pas à la convergence de s_k vers zéro (qui a probablement aussi lieu). D'après (19.32) et la semi-définie positivité de M_k,

$$-g_k^T s_k = \|J_k s_k\|^2 + \lambda_k s_k^T M_k s_k = \|J_k s_k\|^2 + \lambda_k \|M_k^{1/2} s_k\|^2.$$

On déduit alors de (19.40) et de la positivité de λ_k que

$$J_k s_k \to 0 \quad \text{et} \quad \lambda_k^{1/2} M_k^{1/2} s_k \to 0, \qquad \text{lorsque } k \to \infty.$$

Dès lors, si $\{(J_k, \lambda_k M_k)\}_{k\in\mathcal{K}}$ est bornée, on a

$$J_k^T J_k s_k \to 0 \quad \text{et} \quad \lambda_k M_k s_k \to 0, \qquad \text{lorsque } k \to \infty \text{ dans } \mathcal{K}.$$

La définition (19.32) de l'itération implique maintenant que $g_k \to 0$ lorsque $k \to \infty$ dans $\mathcal{K}$. □

Pour avoir des résultats plus forts, on a besoin d'un peu plus de régularité sur f, à savoir le caractère lipschitzien de sa dérivée, et sur les matrices M_k, à savoir leur uniforme définie positivité. On peut alors montrer, dans un premier temps, que la suite $\{\lambda_k\}$ des multiplicateurs est bornée.

Lemme 19.13 (CS pour avoir des multiplicateurs bornés) *Supposons que la fonction de moindres-carrés f soit $\mathcal{C}^{1,1}_L$ et qu'il existe une constante $\beta_{M^{-1}}$ telle que pour tout k on ait $\|M_k^{-1}\| \leqslant \beta_{M^{-1}}$. Alors, pour tout $k \geqslant 1$, on a*

$$\lambda_k \leqslant \beta_\lambda := \max\left(\lambda_0, \frac{\tau_2 \beta_{M^{-1}} L}{1-\eta_1}\right). \tag{19.41}$$

Démonstration. Il suffit de trouver une borne supérieure pour les multiplicateurs λ_k tels que $\hat{\lambda}_k = \lambda_k/\tau_2$ n'a pas été accepté par l'inégalité de décroissance suffisante (19.38). Les autres multiplicateurs sont en effet plus petits qu'un de ces derniers ou plus petit que λ_0. Si l'on note $\hat{s}_k := s_k(\hat{\lambda}_k)$, on a $f(x_k+\hat{s}_k) > f(x_k) + \eta_1 g_k^T \hat{s}_k$ et donc

$$\frac{f(x_k+\hat{s}_k) - f(x_k) - g_k^T\hat{s}_k}{\|\hat{s}_k\|^2} > (1-\eta_1)\frac{-g_k^T\hat{s}_k}{\|\hat{s}_k\|^2}. \tag{19.42}$$

Le théorème des accroissements finis (corollaire C.13) et la L-lipschitzianité de f conduisent à la majoration suivante

$$|f(x_k+\hat{s}_k) - f(x_k) - g_k^T\hat{s}_k| \leqslant \left(\sup_{t\in]0,1[} \|\nabla f(x_k+t\hat{s}_k) - \nabla f(x_k)\|\right)\|\hat{s}_k\| \leqslant L\|\hat{s}_k\|^2.$$

Dès lors, le membre de gauche de l'inégalité (19.42) est majoré par L. Le membre de droite de cette même inégalité (19.42) peut être minoré en observant que $(J_k^T J_k + \hat{\lambda}_k M_k)\hat{s}_k = -g_k$ et donc que

$$-g_k\hat{s}_k = \hat{s}_k^T(J_k^T J_k + \hat{\lambda}_k M_k)\hat{s}_k \geqslant \hat{\lambda}_k \beta_{M^{-1}}^{-1}\|\hat{s}_k\|^2.$$

L'inégalité (19.42) permet donc d'écrire

$$L > (1-\eta_1)\beta_{M^{-1}}^{-1}\hat{\lambda}_k,$$

ce qui donne la majoration (19.41), puisque $\hat{\lambda}_k = \lambda_k/\tau_2$. □

On aurait pu affaiblir une hypothèse du résultat précédent en n'exigeant le caractère $\mathcal{C}^{1,1}_L$ de f que sur un voisinage suffisamment étendu de l'ensemble de sous-niveau $\mathcal{N}_1 := \{x \in \mathbb{R}^n : f(x) \leqslant f(x_1)\}$, auquel appartiennent les itérés x_k, pour que tous les points rejetés $x_k + \hat{s}_k$ (voir la démonstration) y soient contenus. Ce soin apporté à la démonstration est présent dans [596].

En bref, le résultat suivant nous apprend que, sans avoir besoin de l'injectivité uniforme de $\{J_k\}$, ce qui est à comparer avec le théorème 19.9 sur la convergence de l'algorithme de Gauss-Newton, la complexité itérative de l'algorithme 19.10 de LMM est au pire en $O(\varepsilon^{-2})$, qui est la complexité itérative des algorithmes du gradient (proposition 6.8) et de Gauss-Newton (théorème 19.9), ce qui veut dire que l'on peut obtenir un gradient g_k de norme inférieure à un seuil $\varepsilon > 0$ arbitraire en moins de $O(\varepsilon^{-2})$ itérations. C'est une borne très grande, mais elle a l'intérêt de ne pas dépendre de la dimension du problème. Lorsque la suite générée $\{J_k^T\}$ est uniformément injective, la situation devient beaucoup plus favorable, puisqu'alors la suite $\{f(x_k)\}$ converge linéairement vers zéro et la complexité itérative est en $O(\log \varepsilon^{-1})$.

Théorème 19.14 (complexité itérative de l'algorithme LMM) *Supposons que la fonction de moindres-carrés f soit $\mathcal{C}^{1,1}$. Soit $\{(x_k, \lambda_k)\}$ une suite générée par l'algorithme 19.10. On suppose en outre que la suite $\{(J_k, M_k, M_k^{-1})\}$ est bornée. Alors,*

1) *il existe une constante C, indépendante de n et m, telle que, pour tout $\varepsilon > 0$, $\|J_k^T r_k\| \leqslant \varepsilon$ pour un indice k inférieur à $\lceil C\varepsilon^{-2} \rceil$,*
2) *si, de plus, $\{J_k^T\}$ est uniformément injective, alors*
   a) *$f(x_k) \to 0$ linéairement,*
   b) *il existe une constante $C' > 0$, indépendante de n et m, telle que, pour tout $\varepsilon > 0$, $\|r_k\| \leqslant \varepsilon \|r_0\|$ pour tout indice $k \geqslant \lceil C' \log \varepsilon^{-1} \rceil$.*

Démonstration. 1) En sommant les K premières inégalités (19.38), on obtient

$$-\eta_1 \sum_{k=1}^{K} g_k^T s_k \leqslant f(x_1) - f(x_{K+1}) \leqslant f(x_1) - f_*.$$

où $f_* := \min_k f(x_k)$ est positif. Comme dans la démonstration du théorème 19.12, $-g_k^T s_k = \|J_k s_k\|^2 + \lambda_k \|M_k^{1/2} s_k\|^2$ et donc, par la positivité de λ_k, l'inégalité précédente apporte les majorations suivantes

$$\sum_{k=1}^{K} \|J_k s_k\|^2 \leqslant \frac{f(x_1) - f_*}{\eta_1} \qquad \text{et} \qquad \sum_{k=1}^{K} \lambda_k \|M_k^{1/2} s_k\|^2 \leqslant \frac{f(x_1) - f_*}{\eta_1}.$$

Soient β_J un majorant de $\{\|J_k\|\}$, β_M un majorant de $\{\|M_k\|\}$ et β_λ un majorant de $\{\|\lambda_k\|\}$, que existent par hypothèse et par le lemme 19.13 (ces majorants ne dépendent pas n et m). Il vient

$$\sum_{k=1}^{K} \|J_k^T J_k s_k\|^2 \leqslant \frac{\beta_J^2[f(x_1)-f_*]}{\eta_1} \qquad \text{et} \qquad \sum_{k=1}^{K} \lambda_k^2 \|M_k s_k\|^2 \leqslant \frac{\beta_\lambda \beta_M [f(x_1)-f_*]}{\eta_1}.$$

Par ailleurs,

$$\begin{aligned} \|g_k\|^2 &= \|J_k^T J_k s_k + \lambda_k M_k s_k\|^2 \qquad [(19.32)] \\ &\leqslant \left(\|J_k^T J_k s_k\| + \lambda_k \|M_k s_k\|\right)^2 \qquad [\text{inégalité triangulaire}] \\ &\leqslant 2\left(\|J_k^T J_k s_k\|^2 + \lambda_k^2 \|M_k s_k\|^2\right) \qquad [(a+b)^2 \leqslant 2(a^2+b^2)]. \end{aligned} \tag{19.43}$$

Avec les majorations précédentes, on trouve que

$$K\left(\min_{k\in[1:K]} \|g_k\|^2\right) \leqslant \sum_{k=1}^{K} \|g_k\|^2 \leqslant 2(\beta_J^2 + \beta_\lambda \beta_M)\frac{f(x_1)-f_*}{\eta_1}.$$

Dès lors, $\min\{\|g_k\| : k \in [1:K]\}$ est plus petit qu'un $\varepsilon > 0$ arbitraire donné, dès que le membre de droite ci-dessus est inférieur à $K\varepsilon^2$, ce qui s'écrit

$$K \geqslant K_\varepsilon := \left\lceil 2\varepsilon^{-2}(\beta_J^2 + \beta_\lambda \beta_M)\frac{f(x_1)-f_*}{\eta_1} \right\rceil.$$

On a donc montré qu'un des $\|g_k\|$, avec $k \leqslant K_\varepsilon$, est inférieur à ε.

2) On part de (19.38) :

$$f(x_{k+1}) \leqslant f(x_k) + \eta_1 g_k^T s_k \leqslant f(x_k) - C_1 \|g_k\|^2, \tag{19.44}$$

où la dernière inégalité provient de

$$\begin{aligned} &\|g_k\|^2 \\ &\quad\leqslant 2\left(\|J_k^T J_k s_k\|^2 + \lambda_k^2 \|M_k s_k\|^2\right) \qquad [(19.43)] \\ &\quad\leqslant 2\beta_J^2 \|J_k s_k\|^2 + 2\beta_\lambda \beta_M \lambda_k s_k^T M_k s_k \qquad [\|J_k^T\| \leqslant \beta_J,\ \|\lambda_k\| \leqslant \beta_\lambda,\ \|M_k\| \leqslant \beta_M] \\ &\quad\leqslant (\eta_1/C_1)(\|J_k s_k\|^2 + \lambda_k s_k^T M_k s_k) \qquad [C_1 := \eta_1 \max(2\beta_J^2, 2\beta_\lambda\beta_M)^{-1}] \\ &\quad= -(\eta_1/C_1) g_k^T s_k \qquad [(19.32)]. \end{aligned}$$

On poursuit comme dans la démonstration du point 3 du théorème 19.9 : $\|g_k\|^2 = \|J_k^T r_k\|^2 \geqslant (\alpha'_J)^2 \|r_k\|^2 = 2(\alpha'_J)^2 f(x_k)$ par l'injectivité uniforme de J_k^T, si bien que (19.44) devient

$$f(x_{k+1}) \leqslant (1 - C_2) f(x_k),$$

pour la constante strictement positive $C_2 = 2C_1(\alpha'_J)^2$. Ceci montre la convergence linéaire de $f(x_k)$ vers zéro (point 2.a). Puis par récurrence,

$$\frac{f(x_k)}{f(x_0)} \leqslant (1 - C_2)^k.$$

Dès lors, $\|r_k\| \leqslant \varepsilon \|r_0\|$ si $f(x_k)/f(x_0) \leqslant \varepsilon^2$ et donc certainement si le membre de droite ci-dessus est inférieur à ε^2 :

$$(1 - C_2)^k \leqslant \varepsilon^2.$$

En prenant le logarithme des deux membres et en utilisant $\log(1-C_2)<0$, on obtient comme condition sur k :

$$k \geqslant \frac{2\log \varepsilon^{-1}}{\log(1-C_2)^{-1}}.$$

On obtient donc le résultat du point 2.b avec la constante $C' := 2/\log(1-C_2)^{-1}$, qui est indépendante de n et m. □

19.4 Recherche de solution parcimonieuse ▲

19.4.1 Motivation

Supposons que l'on dispose d'une mesure $y := Ms + x$ (M est une matrice $n \times p$) d'un signal $s \in \mathbb{R}^p$, qui est perturbée sur *quelques* composantes par le vecteur $x \in \mathbb{R}^n$, lequel a donc beaucoup de composantes nulles. Dans ce cadre, M est injective : pour que l'opération de récupération de s à partir de y ait quelques chances de succès, il faut qu'il y ait plus de mesures que de composantes au signal recherché. Dans ce cadre, l'on dispose aussi d'un annihilant à gauche A de M : c'est une matrice $m \times n$ telle que $AM = 0$ (par exemple, A peut être le projecteur orthogonal sur $\mathcal{R}(A)^\perp$). Alors on connaît $b := Ay = Ax$ et on cherche à écarter les mesures i corrompues par les perturbations $x_i \neq 0$ en résolvant

$$\begin{cases} \inf \ \|x\|_0 \\ Ax = b, \end{cases} \tag{19.45}$$

où $\|x\|_0 := |\{i \in [1:n] : x_i \neq 0\}|$ est le *compteur de composante non nulle* de x, c'est-à-dire le nombre de ses composantes non nulles (ce n'est pas une norme, par manque de positivité homogène). Ce problème est NP-ardu [448 ; 1995].

Voir la suite, par exemple, chez Candès et Tao [105 ; 2005].

Problème d'*acquisition comprimée* (*compressed sensing*).

19.4.2 Poursuite de base

La *poursuite de base* (de l'anglais *basis pursuit*) est une technique d'optimisation utilisée initialement en *traitement du signal* qui revient à résoudre un problème de la forme

$$(P_1) \quad \begin{cases} \min \ \|x\|_1 \\ Ax = b, \end{cases} \tag{19.46}$$

où l'inconnue est un vecteur $x \in \mathbb{R}^n$, $\|\cdot\|_1$ est la norme ℓ_1, A est une matrice réelle $m \times n$ et $b \in \mathbb{R}^m$. Il s'agit donc de trouver le plus petit vecteur $x \in \mathbb{R}^n$, au sens de la norme ℓ_1, qui vérifie les contraintes affines $Ax = b$. Ce problème est convexe, mais non lisse.

Comme nous le verrons, l'intérêt du problème est de sélectionner une solution du système linéaire $Ax = b$, supposé sous-déterminé (donc avec plus d'une solution), ayant peu d'éléments non nuls. La non-différentiabilité de la norme ℓ_1 joue un rôle-clé dans l'obtention de cette propriété.

L'appellation *poursuite de base* vient de l'algorithme du simplexe qui était proposé dans l'article original [117 ; 1998] pour résoudre le problème ci-dessus, lequel détermine une *base optimale*. Dans la terminologie de cet algorithme, il s'agit d'une sélection de m colonnes de A telle que la sous-matrice B correspondante soit inversible et détermine la solution par $B^{-1}b$.

Motivation

Voici comment on peut être amené à résoudre un problème d'optimisation de la forme (P_1) ci-dessus. Un problème classique en *traitement du signal* consiste à trouver une décomposition *parcimonieuse* (c'est-à-dire formée de peu d'éléments) d'un signal donné dans un *dictionnaire surabondant* de signaux, contenant par exemple des sinusoïdes (décomposition de Fourier), des ondelettes, et bien d'autres signaux. Dans l'écriture ci-dessus, b est le signal à décomposer, les colonnes de A sont les éléments du dictionnaire de signaux et les composantes de x sont les coefficients recherchés pour représenter le signal au moyen des signaux du dictionnaire. On peut donc écrire

$$b = \sum_{j=1}^{m} x_j A^j,$$

où A^j est la colonne j de A. Lorsque le dictionnaire de signaux A^j est *surabondant*, $m > n$ et la décomposition de x comme ci-dessus n'est pas unique. Lorsqu'on cherche une décomposition parcimonieuse, l'on cherche à avoir le moins de coefficients x_j non nuls. C'est ce qui permet d'avoir une représentation compacte du signal (compression de celui-ci).

Annuler le plus de coefficients x_j revient à résoudre le problème

$$(P_0) \quad \begin{cases} \min \ \|x\|_0 \\ Ax = b, \end{cases}$$

où $\|x\|_0 := |\{i \in [1:n] : x_i \neq 0\}|$ est le nombre d'éléments non nuls de x (ce n'est pas une norme, car l'homogénéité n'a pas lieu, mais la limite de $\|x\|_p$ lorsque $p \to 0$, d'où la notation). Ce dernier problème est malheureusement NP-ardu [448], ce qui est aujourd'hui un handicap rédhibitoire lorsqu'on veut résoudre des problèmes de grande taille. Le problème (P_1) peut être vu comme une *approximation traitable* du problème (P_0), pour les raisons suivantes.

- Le problème (P_1) consiste à trouver un point du sous-espace affine $\mathcal{A} := \{x \in \mathbb{R}^n : Ax = b\}$ le plus proche de zéro au sens de la norme ℓ_1. Comme la boule unité B_1 de cette dernière est polyédrique, elle a un nombre fini de sommets et le problème (P_1) a tendance à trouver une solution en un sommet de $\mathrm{val}(P_1)B_1$ (on a noté $\mathrm{val}(P_1)$ la valeur optimale du problème (P_1)) ou sur une face contenant peu de sommets de cette boule. Or les sommets de $\mathrm{val}(P_1)B_1$ sont des multiples des vecteurs de base de $\mathbb{R}^n$, qui ont toutes leurs composantes nulles sauf une ! La solution de (P_1) aura donc tendance à avoir beaucoup d'éléments nuls.
- Par ailleurs, le problème (P_1) est un problème *convexe*, qui peut être récrit comme un problème d'optimisation linéaire (voir ci-dessous) et donc peut être résolu en temps polynomial.

Cette approche de résolution de (P_0) par (P_1) a été proposée par [117 ; 1998].

Analyse du problème

DUALITÉ

Le problème (P_1) s'écrit aussi

$$\inf_{x\in\mathbb{R}^n}\sup_{y\in\mathbb{R}^m}\Big(\|x\|_1 - y^{\mathsf{T}}(Ax-b)\Big).$$

Le dual lagrangien de (P_1) est donc le problème

$$\sup_{y\in\mathbb{R}^m}\inf_{x\in\mathbb{R}^n}\Big(\|x\|_1 - y^{\mathsf{T}}(Ax-b)\Big) = \sup_{y\in\mathbb{R}^m}\left(b^{\mathsf{T}}y - \sup_{x\in\mathbb{R}^n}\big((A^{\mathsf{T}}y)^{\mathsf{T}}x - \|x\|_1\big)\right).$$

En se rappelant l'expression de la conjuguée de la norme (exercice 3.30), le dernier supremum est nul si $\|A^{\mathsf{T}}y\|_\infty \leqslant 1$ et vaut $+\infty$ sinon, si bien que l'on obtient comme problème dual lagrangien

$$(D_1)\quad \begin{cases} \sup\ b^{\mathsf{T}}y \\ \|A^{\mathsf{T}}y\|_\infty \leqslant 1. \end{cases}$$

On a bien sûr le résultat de dualité faible

$$\operatorname{val}(D_1) \leqslant \operatorname{val}(P_1). \tag{19.47}$$

Nous allons montrer qu'il y a dualité forte, dans un sens à préciser, entre le primal (P_1) et le dual (D_1). Une manière de montrer cette dualité forte est de récrire ces problèmes sous la forme des problèmes d'optimisation linéaire (P_1') et (D_1') ci-dessous et d'utiliser la dualité forte entre (P_1') et (D_1') (théorème 17.11). On peut reformuler le problème (P_1) comme un problème d'optimisation linéaire en $(u,v)\in\mathbb{R}^n\times\mathbb{R}^n$ sous forme standard :

$$(P_1')\quad \begin{cases} \min\ e^{\mathsf{T}}u + e^{\mathsf{T}}v \\ \begin{pmatrix} A & -A \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = b, \\ u \geqslant 0, \quad v \geqslant 0. \end{cases}$$

où $e\in\mathbb{R}^n$ est le vecteur dont toutes les composantes valent 1. Le lien entre les variables (u,v) de (P_1') et la variable x de (P_1) est $x = u - v$. Quant au problème dual (D_1), on peut l'écrire comme le problème d'optimisation linéaire suivant

$$(D_1')\quad \begin{cases} \sup\ b^{\mathsf{T}}y \\ -e \leqslant A^{\mathsf{T}}y \leqslant e. \end{cases}$$

D'après la section 17.3.1, (D_1') est le dual lagrangien de (P_1'). On déduit alors du théorème 17.11 une partie du résultat de dualité forte suivant. Il y a une différence toutefois : il n'y a ici jamais de saut de dualité (essentiellement parce que le dual est toujours réalisable et sa valeur optimale ne peut donc pas être $-\infty$).

Proposition 19.15 (dualité forte en poursuite de base)

1) *Les propriétés suivantes sont équivalentes :*
   (*i*) (P_1) *est réalisable,*
   (*ii*) (P_1) *a une solution,*
   (*iii*) (D_1) *a une solution.*
2) *Il n'y a pas de saut de dualité : si les propriétés* (*i*)-(*iii*) *ont lieu* $\mathrm{val}(D_1) = \mathrm{val}(P_1) \in \mathbb{R}$, *sinon* $\mathrm{val}(D_1) = \mathrm{val}(P_1) = +\infty$.
3) *Un point* $\bar{x} \in \mathbb{R}^n$ *est solution de* (P_1) *et un point* $\bar{y} \in \mathbb{R}^m$ *est solution de* (D_1) *si, et seulement si,* $(\bar{x}, \bar{y})$ *est un point-selle du lagrangien* $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mapsto \|x\|_1 - y^\mathsf{T}(Ax - b)$.

Démonstration. 1) Les équivalences entre (*i*)-(*iii*) se déduisent directement du théorème 17.11 et de l'équivalence entre (P_1) et (P_1'), d'une part, et (D_1) et (D_1'), d'autre part.

2) D'après le théorème 17.11, il n'y a pas de saut de dualité si les propriétés (*i*)-(*iii*) ont lieu et dans ce cas les valeurs optimales sont finies.

Supposons maintenant que les propriétés (*i*)-(*iii*) n'ont pas lieu. Alors (P_1) n'a pas de solution et donc $b \notin \mathcal{R}(A)$ (on le verra au point 1 de la proposition 19.16 ci-dessous, mais c'est évident dès à présent), ce qui revient à dire que $\mathrm{val}(P_1) = +\infty$. Mais si $b \notin \mathcal{R}(A)$, on peut écrire $b = Ax_0 + b_1$ avec b_1 non nul dans $\mathcal{R}(A)^\perp = \mathcal{N}(A^\mathsf{T})$. Alors, quel que soit $t \in \mathbb{R}$, $y_t := tb_1 \in \mathcal{N}(A^\mathsf{T})$ est admissible pour (D_1) et on a $b^\mathsf{T} y_t = t\|b_1\|_2^2 \to +\infty$ lorsque $t \to +\infty$. Donc $\mathrm{val}(D_1) = +\infty$.

3) C'est une conséquence des points 1 et 2 et du théorème 14.3. □

Existence de solution, unicité, Conditions d'optimalité

La condition d'existence de solution de (P_1) découlant du théorème de Weierstrass, de nature topologique, est triviale (point 2 de la proposition 19.16 ci-dessous). Mais le problème (P_1) est convexe, si bien que ses conditions d'existence de solution sont équivalentes à ses conditions d'optimalité ; cette remarque conduit aux conditions d'existence de solution (19.48) qui s'expriment, comme toutes conditions d'optimalité, en termes d'un candidat-solution $\bar{x}$.

On notera que la condition (19.49) prend en compte le cas où $b = 0$. En effet dans ce cas $\bar{x} = 0$ est l'unique solution de (P_1). Les conditions (19.49) n'accepte pas d'utres points, car si $\bar{x} \neq 0$, alors $I \neq \varnothing$ et l'injectivité de A_I et $A_I \bar{x}_I = 0$ impliquerait que $\bar{x}_I = 0$; une contradiction.

Proposition 19.16 (existence, unicité, conditions d'optimalité)
1) *L'ensemble des solution de* (P_1) *est un polyèdre convexe, éventuellement vide.*
2) *Le problème* (P_1) *a une solution si, et seulement si,* $b \in \mathcal{R}(A)$.
3) *Un point* $\bar{x} \in \mathbb{R}^n$ *est solution de* (P_1) *si, et seulement si,*

$$\begin{cases} A\bar{x} = b \\ \text{il existe } y \in \mathbb{R}^m \text{ tel que } A_I^\mathsf{T} y = s \text{ et } \|A_{I^c}^\mathsf{T} y\|_\infty \leqslant 1, \end{cases} \tag{19.48}$$

où $I := \{i \in [1:n] : \bar{x}_i \neq 0\}$ *et* $s := \operatorname{sgn}(\bar{x}_I)$.

4) *Un point* $\bar{x} \in \mathbb{R}^n$ *est solution unique de* (P_1) *si, et seulement si,*

$$\begin{cases} A\bar{x} = b \\ A_I \text{ est injective} \\ \text{il existe } y \in \mathbb{R}^m \text{ tel que } A_I^\mathsf{T} y = s \text{ et } \|A_{I^c}^\mathsf{T} y\|_\infty < 1, \end{cases} \tag{19.49}$$

où $I := \{i \in [1:n] : \bar{x}_i \neq 0\}$ *et* $s := \operatorname{sgn}(\bar{x}_I)$.

Démonstration. 1) L'ensemble des solutions de (P_1) est l'image par l'application linéaire $T : (u, v) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto u - v \in \mathbb{R}^n$ de l'ensemble des solutions du problème d'optimisation linéaire (P_1'), qui est un polyèdre convexe. C'est donc un polyèdre convexe (proposition 2.18).

2) Évidemment, si (P_1) a une solution $\bar{x}$, $b = A\bar{x} \in \mathcal{R}(A)$. Inversement, c'est aussi évident : l'ensemble admissible est non vide (par la condition $b \in \mathcal{R}(A)$) et fermé (c'est un sous-espace affine en dimension finie) ; puis la coercivité du critère assure l'existence d'une solution (proposition 1.4).

3) Le problème (P_1) étant convexe, $\bar{x} \in \mathbb{R}^n$ en est solution si, et seulement si, zéro est dans le sous-différentiel de la fonction

$$x \in \mathbb{R}^n \mapsto \|x\|_1 + \mathcal{I}_{\mathcal{A}}(x), \tag{19.50}$$

où $\mathcal{A} := \{x \in \mathbb{R}^n : Ax = b\}$ (corollaire 3.61). Par la proposition 3.71, ceci peut s'écrire

$$0 \in \partial(\|\cdot\|_1)(\bar{x}) + \partial\mathcal{I}_{\mathcal{A}}(\bar{x}).$$

Forcément, $\bar{x} \in \mathcal{A}$, sinon $\mathrm{N}_{\mathcal{A}}(\bar{x}) = \varnothing$. Par ailleurs, $\partial(\|\cdot\|_1)(\bar{x}) = \{z \in \mathbb{R}^n : \|z\|_\infty \leqslant 1, z^\mathsf{T}\bar{x} = \|\bar{x}\|_1\} = \{z \in \mathbb{R}^n : z_I = s \text{ et } \|z_{I^c}\|_\infty \leqslant 1\} =: D$ (exercice 3.30) et $\partial\mathcal{I}_{\mathcal{A}}(\bar{x}) = \mathrm{N}_{\mathcal{A}}(\bar{x})$ (exercice 3.28). On voit facilement que $\mathrm{N}_{\mathcal{A}}(\bar{x}) = \mathcal{N}(A)^\perp = \mathcal{R}(A^\mathsf{T})$, si bien que l'élément de $\partial(\|\cdot\|_1)(\bar{x})$ dans la condition d'optimalité ci-dessus est de la forme $A^\mathsf{T}\bar{y}$ avec $\bar{y} \in \mathbb{R}^m$. Dès lors (19.48) a lieu.

4) La fonction (19.50) étant polyédrique, elle a un minimiseur unique si, et seulement si,

$$0 \in (\partial(\|\cdot\|_1)(\bar{x}) + \partial\mathcal{I}_{\mathcal{A}}(\bar{x}))^\circ = \big(D + \mathcal{R}(A^\mathsf{T})\big)^\circ,$$

où on a utilisé les expressions de $\partial(\|\cdot\|_1)(\bar{x})$ et $\partial\mathcal{I}_{\mathcal{A}}(\bar{x})$ déterminées dans la démonstration du point 3. Ceci peut s'exprimer par les deux conditions suivantes :

$$\operatorname{aff}\big(D + \mathcal{R}(A^\mathsf{T})\big) = \mathbb{R}^n, \tag{19.51}$$

$$0 \in D^\ominus + \mathcal{R}(A^\mathsf{T}). \tag{19.52}$$

La condition (19.51) s'exprime aussi par $\operatorname{aff} D + \mathcal{R}(A^\mathsf{T}) = \mathbb{R}^n$ ou $\{z \in \mathbb{R}^n : z_I = s\} + \mathcal{R}(A^\mathsf{T}) = \mathbb{R}^n$. Il n'est pas difficile de montrer par l'algèbre linéaire que cela équivaut à l'injectivité de A_I. La condition (19.52) est équivalente à l'existence d'un $y \in \mathbb{R}^m$ tel que $A^\mathsf{T} y \in D^\ominus = \{z \in \mathbb{R}^n : z_I = s \text{ et } \|z_{I^c}\|_\infty < 1\}$. On a donc montré que (19.49) équivaut au fait que $\bar{x}$ est la solution unique de (P_1). □

19.4.3 Problème LASSO ▲

Le *problème* LASSO (pour *Least Absolute Shrinkage and Selection Operator*) consiste à trouver un vecteur $x \in \mathbb{R}^n$ solution de

$$\begin{cases} \min_x \frac{1}{2}\|Ax-b\|_2^2 \\ \|x\|_1 \leqslant r, \end{cases} \tag{19.53}$$

où $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $r > 0$, et $\|\cdot\|_p$ désigne la norme ℓ_p. Ce problème est convexe (critère et contrainte convexe), mais l'on prendra garde à la non-différentiabilité de sa contrainte.

Le problème LASSO, dont le nom imagé a fait florès et a sans doute contribué à la notoriété de son initiateur [583], est utilisé (en statistiques, en fouille de données,...) pour résoudre approximativement des problèmes de moindres-carrés linéaires par des vecteurs ayant beaucoup de composantes nulles (c'est la raison d'être de la norme ℓ_1 en contrainte) [466, 467, 584, 585]. L'approche est surtout utile lorsque $m \ll n$ (il y a alors beaucoup de x vérifiant $Ax = b$ approximativement et la norme ℓ_1 est utilisée pour sélectionner ceux ayant beaucoup de composantes nulles), alors que le problème de moindres-carrés linéaire classique (19.1) (sans la contrainte en norme ℓ_1 dans (19.53)) est surtout utilisé lorsque $m \gg n$ (le plus souvent, il n'y a alors pas de x vérifiant $Ax = b$).

Commençons par des propriétés de la norme ℓ_1.

Lemme 19.17 (norme ℓ_1) *Le sous-différentiel $S(x) := \partial(\|\cdot\|_1)(x)$ de la norme ℓ_1 en $x \in \mathbb{R}^n$ s'écrit*

$$\begin{aligned} S(x) = \{x^* \in \mathbb{R}^n : {} & x_i^* = -1 \text{ si } x_i < 0, \\ & x_i^* \in [-1,+1] \text{ si } x_i = 0, \\ & x_i^* = +1 \text{ si } x_i > 0\}. \end{aligned} \tag{19.54}$$

Le cône tangent $\mathrm{T}_{rB_1}(x)$ à la boule rB_1 de rayon $r > 0$ pour la norme ℓ_1 en un point x de norme $\|x\|_1 = r$ ne dépend pas de r et est le cône dual négatif de $S(x)$:

$$\mathrm{T}_{rB_1}(x) = S(x)^-. \tag{19.55}$$

DÉMONSTRATION. Le sous-différentiel $S(x)$ est calculé à l'exercice 3.30.

Pour démontrer (19.55), il est utile d'introduire

$$\begin{aligned} S_0(x) = \{x^* \in \mathbb{R}^n : {} & x_i^* = -1 \text{ si } x_i < 0, \\ & x_i^* = \pm 1 \text{ si } x_i = 0, \\ & x_i^* = +1 \text{ si } x_i > 0\}, \end{aligned}$$

Comme la boule unité pour la norme ℓ_∞ (dont les éléments ont leurs composantes dans $[-1,1]$) est l'enveloppe convexe de ses sommets (dont les composantes valent ± 1), on a $S(x) = \operatorname{co} S_0(x)$. Par ailleurs, la boule rB_1 est un polyèdre convexe qui s'écrit $rB_1 = \{x \in \mathbb{R}^n : Ex \leqslant r\}$, où la matrice E a ses 2^n lignes formées de tous les vecteurs

différents de $\mathbb{R}^n$ dont les composantes valent ± 1. D'après l'exercice 2.43, si l'on note $I := \{i : (Ex)_i = r\}$, son cône tangent s'écrit

$$\begin{aligned} \mathrm{T}_{rB_1}(x) &= \{d \in \mathbb{R}^n : (Ed)_I \leqslant 0\} \\ &= \{d \in \mathbb{R}^n : s^{\mathsf{T}} d \leqslant 0,\ \forall\, s \in S_0(x)\} \\ &= S(x)^-. \end{aligned}$$

Pour la dernière égalité, on a utilisé le fait que le dual négatif de $S_0(x)$ est égal à celui de son enveloppe convexe $S(x)$ (point 3 de la proposition 2.43). □

Exhibons maintenant des conditions d'existence et d'unicité de solution du problème LASSO. Dans leur expression, on utilise le rayon r_0 défini comme suit :

$$X_0 := \arg\min\{\|Ax - b\|_2 : x \in \mathbb{R}^n\} \qquad \text{et} \qquad r_0 := \inf\{\|x\|_1 : x \in X_0\}.$$

Donc X_0 est l'ensemble des minimiseurs sans contrainte (il est non vide par la proposition 19.1) et r_0 est la distance de zéro à X_0 pour la norme ℓ_1 (elle est donc finie).

Proposition 19.18 (existence et unicité de solution) *Le problème* LASSO (19.53) *admet une solution. Une telle solution $\bar{x}$ est unique si, et seulement si, l'une des deux conditions suivantes est vérifiée*
(*i*) *A est injective,*
(*ii*) *$r \leqslant r_0$ et $\mathcal{N}(A) \cap S(\bar{x})^+ = \{0\}$.*

DÉMONSTRATION. Le problème a une solution par le théorème de Weierstrass (théorème 1.2), parce que le critère est continu et l'ensemble admissible est non vide et compact. Soit $\bar{x}$ une solution du problème.

[⇒] Supposons $\bar{x}$ soit l'unique solution du problème. Montrons dans un premier temps que l'on a nécessairement

$$\mathcal{N}(A) \cap S(\bar{x})^+ = \{0\}. \tag{19.56}$$

Soit $h \in \mathcal{N}(A) \cap S(\bar{x})^+$. D'après (19.55), $-h \in \mathcal{N}(A) \cap \mathrm{T}_{\|x\|_1 B_1}(x)$ et donc pour $\alpha > 0$ petit, $\bar{x}_\alpha := \bar{x} - \alpha h$ vérifie $\|A\bar{x}_\alpha - b\|_2 = \|A\bar{x} - b\|_2$ et, par la polyédricité de $\|x\|_1 B_1$, $\|\bar{x}_\alpha\|_1 \leqslant \|\bar{x}\|_1 \leqslant r$. Donc $\bar{x}_\alpha$ est aussi solution. L'unicité de celle-ci implique que $h = 0$.

On peut maintenant établir qu'une des conditions (*i*) ou (*ii*) a lieu. Si (*i*) n'a pas lieu, alors nécessairement $r \leqslant r_0$ (sinon il existe des solutions $\bar{x}'$ de norme $r_0 < r$ et $\bar{x}' + \alpha h$ est encore solution si $h \in \mathcal{N}(A) \setminus \{0\}$ et α est proche de zéro) et donc (*ii*) a lieu. Si (*ii*) n'a pas lieu, alors nécessairement $r > r_0$ (puisque (19.56) est vérifiée) et donc A doit être injective pour que l'unicité ait lieu (même raisonnement que précédemment).

[⇐] Réciproquement, si (*i*) a lieu, le critère est strictement convexe et l'unicité de solution s'en déduit.

Supposons maintenant que (*ii*) ait lieu et que $\bar{x}'$ soit une solution du problème. Il suffit de montrer que $h := \bar{x} - \bar{x}'$ est dans $\mathcal{N}(A) \cap S(\bar{x})^+$, ce qui impliquera que $h = 0$ et donc que $\bar{x}$ est la seule solution.

- Montrons que $h \in \mathcal{N}(A)$. Par convexité de l'ensemble des solutions, $\bar{x}_\alpha := \bar{x} - \alpha h$ est aussi solution pour tout $\alpha \in [0,1]$. De l'égalité $\|A\bar{x} - b\|_2 = \|A\bar{x}_\alpha - b\|_2$, on obtient $\|A\bar{x} - b\|_2^2 = \|(A\bar{x} - b) - \alpha Ah\|_2^2$, qui après développement donne

$$-2\alpha h^\mathsf{T} A^\mathsf{T}(A\bar{x} - b) + \alpha^2 \|Ah\|_2^2 = 0.$$

  Comme cette identité est vraie pour tout $\alpha \in [0,1]$, on en déduit que $h \in \mathcal{N}(A)$.
- Montrons que $h \in S(\bar{x})^+$. Dans ce but, on observe que

$$\|\bar{x}\|_1 = \|\bar{x}'\|_1 = r. \tag{19.57}$$

  En effet, si $\|\bar{x}\|_1 < r$, $\bar{x}$ est solution sans contrainte de (19.53) et donc $\bar{x} \in X_0$, ce qui implique que $\|\bar{x}\|_1 \geqslant r_0$, en contradiction avec $\|\bar{x}\| \leqslant r < r_0$; pour la même raison $\|\bar{x}'\| = r$. Alors $\bar{x}$ et $\bar{x}' \in rB_1$, si bien que $-h \in \mathrm{T}_{rB_1}(\bar{x}) = -S(\bar{x})^+$, par (19.55). □

Proposition 19.19 (conditions d'optimalité) *Soit $\bar{x} \in \mathbb{R}^n$. Les conditions suivantes sont équivalentes*

(i) $\bar{x}$ est solution du problème LASSO (19.53),
(ii) $\exists\, s \in S(\bar{x})$ tel que $-A^\mathsf{T}(A\bar{x} - b) = \|A^\mathsf{T}(A\bar{x} - b)\|_\infty s$,
(iii) $-(A\bar{x} - b)^\mathsf{T} A\bar{x} = \|A^\mathsf{T}(A\bar{x} - b)\|_\infty \Delta$.

Démonstration. □

Montrons que l'on peut prendre comme dual lagrangien du problème LASSO, le problème

$$\sup_{z \in \mathbb{R}^n} \; -\left(\frac{1}{2}\|Az - b\|_2^2 + \Delta \, \|A^\mathsf{T} Az\|_\infty \right). \tag{19.58}$$

Notes

La méthode des moindres-carrés pour l'identification de paramètres a été proposée par Gauss. Cette approche lui permit de déterminer l'orbite de la planète Cérès en 1801 [62], découverte fortuitement la même année par Piazzi. Cette méthode fut introduite indépendamment par Legendre dans ses « *Nouvelles méthodes pour la détermination des orbites des comètes* » en 1806 et celui-ci a un temps revendiqué cette découverte. Il s'est finalement avéré que Gauss fut le premier à en avoir fait l'exposition, mais dans des notes non publiées datant de 1799. La première référence de Gauss traitant de ce sujet est sa « *Theoria motus corporum coelestium* » de 1809 [277]. D'autres éléments historiques sont relatés par Goldstine [265 ; 1977].

Les problèmes de moindres-carrés linéaires et leurs techniques de résolution sont passés en revue dans les livres et monographies de Hanson et Lawson [304] et de Björck [61, 62]. Autres ouvrages sur l'estimation de paramètres par moindres-carrés :

Bard [37 ; 1974] ; Arnold et Beck [22 ; 1977]. Mentionnons aussi la technique de régularisation de ces problèmes due à Tikhonov [586 ; 1963]. Les *problèmes de moindres-carrés linéaires totaux* sont passés en revue par Van Huffel et Vandewalle [322].

Le résultat de complexité en $O(\log \varepsilon^{-1})$ des théorèmes 19.9 et 19.14 a été obtenu en reprenant un argument de [597] pour l'algorithme de Levenberg-Morrison-Marquardt, mais qui peut aussi s'utiliser pour l'analyse de l'algorithme de Gauss-Newton.

L'algorithme de Levenberg-Morrison-Marquardt (LMM) présenté et étudié à la section 19.3.3 a été proposé par Levenberg [389 ; 1944], avec des notations qui rendent son texte difficile à suivre aujourd'hui ; l'article ne donne pas de résultat de convergence. Sans référence à Levenberg, ni à quiconque d'ailleurs, Morrison [443 ; 1960] introduit un algorithme similaire en raisonnant à partir de ce que l'on appelle aujourd'hui la notion de *région de confiance* (chapitre 9) ; il en donne quelques propriétés mais pas de résultat de convergence. L'approche fut ensuite redécouverte et analysée par Marquardt [407 ; 1963]. La détermination du paramètre de pénalisation présentée ici est due à Osborne [465 ; 1976]. Le résultat de convergence globale (théorème 19.12) que nous donnons est plus fort que celui travaillé dans [465], car il ne requiert pas la bornitude de $\{\lambda_k\}$, ni d'hypothèse entraînant celle-ci. La complexité itérative de l'algorithme de LMM a été étudiée dans [596, 597] ; l'analyse que nous en donnons dans la démonstration du théorème 19.14 nous semble plus rapide et est directement inspirée de celle sur la complexité itérative de l'algorithme du gradient (proposition 6.8). Pour d'autres approches et contributions, voir [437, 643] et leurs références.

Exercices

19.1. *Problème de moindres-carrés linéaire régularisé.* Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces vectoriels normés de dimension finie et de normes notées $\|\cdot\|_{\mathbb{E}}$ et $\|\cdot\|_{\mathbb{F}}$ respectivement, $A : \mathbb{E} \to \mathbb{F}$ une application linéaire et $b \in \mathbb{F}$. On considère le problème

$$(P) \quad \inf_{x \in \mathbb{E}} \|Ax - b\|_{\mathbb{F}}$$

et sa version régularisée

$$(P_\varepsilon) \quad \inf_{x \in \mathbb{E}} \|Ax - b\|_{\mathbb{F}}^{\alpha} + \varepsilon \|x\|_{\mathbb{E}}^{\beta},$$

où $\alpha > 0$, $\beta > 0$ et $\varepsilon > 0$.

1) (P) a une solution.
2) (P_ε) a une solution. On en sélectionne une, notée $\bar{x}_\varepsilon$.
3) Lorsque $\varepsilon \to \infty$, $\bar{x}_\varepsilon \to 0$.
4) Lorsque $\varepsilon \downarrow 0$, $\|\bar{x}_\varepsilon\|_{\mathbb{E}}$ croît ; la suite $\{\bar{x}_\varepsilon\}_{\varepsilon \downarrow 0}$ est bornée et ses points d'adhérence sont solutions de $\min\{\|x\|_{\mathbb{E}} : x \in S\}$, où S est l'ensemble des solutions de (P).
5) Si $\mathbb{E}$ et $\mathbb{F}$ sont euclidiens, si $\|\cdot\|_{\mathbb{E}}$ et $\|\cdot\|_{\mathbb{F}}$ sont associées à un produit scalaire, si $\alpha = \beta = 2$, alors
   - (P_ε) a une solution unique,
   - $\varepsilon \bar{x}_\varepsilon$ converge vers $A^* b$ si $\varepsilon \to \infty$ (A^* est l'opérateur adjoint de $A : \mathbb{E} \to \mathbb{F}$) et si, de plus, $A^* b \neq 0$, alors $\bar{x}_\varepsilon / \|\bar{x}_\varepsilon\| \to A^* b / \|A^* b\|$ (où $\|\cdot\|$ est une norme quelconque),
   - $\bar{x}_\varepsilon$ converge vers la solution de norme minimale de (P) si $\varepsilon \downarrow 0$.

19.2. *Problème de moindres-carrés linéaire sous région de confiance.* Soient $A \in \mathbb{R}^{m\times n}$ une matrice, $b \in \mathbb{R}^m$ et $\Delta > 0$. On considère le problème suivant

$$\begin{cases} \min \|Ax - b\| \\ \|x\| \leqslant \Delta. \end{cases} \tag{19.59}$$

dans lequel $\|\cdot\|$ désigne la norme euclidienne sur $\mathbb{R}^m$ ou $\mathbb{R}^n$.

1) Montrez que le problème (19.59) a une solution.
2) *Justifiez* et *écrivez* les conditions d'optimalité de Karush, Kuhn et Tucker de (19.59) (de manière à rendre le calcul plus aisé, on pourra élever certaines quantités au carré). Montrez que ces conditions sont nécessaires et suffisantes pour l'optimalité.
3) Montrez que l'on peut trouver une solution de (19.59) qui soit dans $\mathcal{R}(A^\mathsf{T})$.
4) Montrez que (19.59) a une unique solution si, et seulement si,

$$\bar{B}(0,\Delta) \cap \{x : A^\mathsf{T}(Ax - b) = 0\} \text{ a au plus un élément.}$$

On a noté $\bar{B}(0,\Delta)$ la boule fermée de centre 0 et de rayon Δ pour la norme euclidienne.

19.3. *Approximation de Tchebychev d'un système linéaire surdéterminé* [86 ; ex. 5.6]. Soient A une matrice de type $m \times n$ et $b \in \mathbb{R}^m$. On considère le problème d'*approximation au sens de Tchebychev* ou en norme ℓ_∞ suivant :

$$v_{\text{tch}} := \min_{x\in\mathbb{R}^n} \|Ax - b\|_\infty. \tag{19.60}$$

Le problème (19.60) n'a pas nécessairement une unique solution, mais il en a au moins une, d'où l'utilisation du « min ».

1) Montrez que le problème (19.60) a une solution.

On ne connait pas d'expression analytique de v_{tch}, ni a fortiori l'expression analytique des solutions de (19.60), alors que les solutions du problème de moindres-carrés sont connues. En particulier, on connait la forme de ses solutions x_{mc} et son résidu optimal

$$r_{\text{mc}} := Ax_{\text{mc}} - b.$$

Il est donc naturel de chercher à savoir si ce résidu permet d'estimer v_{tch} et plus précisément d'en donner une borne inférieure (positive bien sûr).

2) Montrez que

$$\frac{1}{\sqrt{m}} \|r_{\text{mc}}\|_\infty \leqslant v_{\text{tch}}. \tag{19.61}$$

On montre maintenant que la dualité permet de resserrer cette estimation de v_{tch}.

3) Montrez dans quel sens on peut considérer que le problème

$$\max_{\substack{y\in\mathbb{R}^m \\ \|y\|_1\leqslant 1 \\ A^\mathsf{T}y=0}} b^\mathsf{T}y \tag{19.62}$$

est dual du problème (19.60). Montrez que (19.62) a une solution.
4) On suppose ici que $r_{\text{mc}} \neq 0$ (dans le cas contraire, $v_{\text{tch}} = 0$ et il n'y a donc pas lieu de trouver un minorant strictement positif de v_{tch}). En choisissant bien un point admissible du problème dual (19.62), montrez que l'on a

$$\frac{\|r_{\text{mc}}\|_2^2}{\|r_{\text{mc}}\|_1} \leqslant v_{\text{tch}}.$$

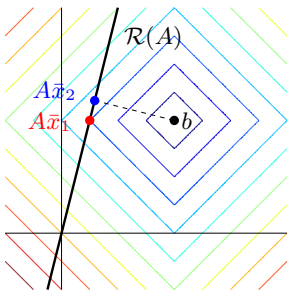
Montrez que ce minorant est meilleur que celui donné en (19.61).

19.4. *Meilleure résolution d'un système linéaire non réalisable.* On considère le problème d'optimisation en $x \in \mathbb{R}^n$ suivant

$$(P) \qquad \inf_{x\in\mathbb{R}^n} \|Ax-b\|,$$

où $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ et $\|\cdot\|$ est une norme *quelconque* sur $\mathbb{R}^m$. On note $f : \mathbb{R}^n \to \mathbb{R}$ la fonction *convexe* définie en $x \in \mathbb{R}^n$ par $f(x) = \|Ax-b\|$.

1) Montrez que le sous-différentiel de f en $x \in \mathbb{R}^n$, pour le produit scalaire euclidien, s'écrit
$$\partial f(x) = \{A^\mathsf{T} y : \|y\|_\mathrm{D} \leqslant 1,\ y^\mathsf{T}(Ax-b) = \|Ax-b\|\}, \tag{19.63}$$
où $\|\cdot\|_\mathrm{D}$ désigne la norme duale de $\|\cdot\|$ pour le produit scalaire euclidien.
2) Montrez dans quel sens on peut considérer le problème (D) ci-dessous comme un problème dual min-max (section 14.1.1) de (P) :
$$(D) \qquad \begin{cases} \sup_{y\in\mathbb{R}^m} b^\mathsf{T} y \\ A^\mathsf{T} y = 0 \\ \|y\|_\mathrm{D} \leqslant 1. \end{cases}$$
3) Montrez que (P) et (D) ont une solution et qu'il n'y a pas de saut de dualité.
4) Montrez que l'ensemble des solutions de (P) s'écrit
$$\mathrm{Sol}(P) = C + \mathcal{N}(A),$$
où C est le convexe $\{x \in \mathcal{R}(A^\mathsf{T}) : \|Ax-b\| = \mathrm{val}(P)\}$ et que C est réduit à un point lorsque $\|\cdot\|$ est la norme ℓ_2 de $\mathbb{R}^m$.
5) Montrez que, lorsque la norme utilisée dans (P) est la norme ℓ_1, le problème (P) peut s'écrire comme un problème d'optimisation linéaire.



Remarque. Si $\|\cdot\|$ est la norme ℓ_1, en résolvant (P), on cherche à satisfaire exactement le plus grand nombre possible d'équations scalaires du système $Ax = b$ (qui est éventuellement non réalisable), tout en minimisant l'erreur $\|Ax-b\|_1$, sans savoir a priori quelles sont ces équations scalaires qui peuvent être résolues exactement. Ceci est illustré en dimension $m = 2$ à la figure ci-jointe. On voit que la solution $\bar{x}_1$, obtenue avec la norme ℓ_1 rend la seconde composante de $A\bar{x}_1$ égale à celle de b, donc la seconde équation du système, c'est-à-dire $(A\bar{x}_1 - b)_2 = 0$, est vérifiee. Par contre, la solution $\bar{x}_2$, obtenue avec la norme ℓ_2, rend toutes les composantes de $A\bar{x}_2$ différentes de celles de b. Par ailleurs, la résolution en norme ℓ_2, qui ne demande que la résolution d'un système linéaire, est moins coûteuse que la résolution en norme ℓ_1, laquelle requiert la résolution d'un problème d'optimisation linéaire.