Analyse numérique MIGS 1re Année

Franz Chouly, Xavier Dupuis, Killian Vuillemot

M1 MIGS

Université de Bourgogne

Analyse numérique

Franz Chouly – Xavier Dupuis – Killian Vuillemot

Année 2020–2021

Table des matières

Chapitre 1

Introduction

Le but de ce cours est de vous présenter des notions d'analyse numérique, à la fois "avancées", et aussi utiles pour de nombreux problèmes qu'on rencontre, presque, quotidiennement, en mathématique appliquée.

Une première partie traite de la résolution numérique de systèmes linéaires symétriques et de grande taille. Il vient en complément du cours d'analyse numérique de L3 que vous avez peut être suivi, où vous avez vu les méthodes les plus élémentaires (décomposition LU, de Cholesky et méthodes itératives simples comme Jacobi ou Gauss-Seidel). On présentera ici une méthode itérative plus "sophistiquée", mais qui est d'usage courant et qui disponible maintenant dans la plupart des librairies et logiciels de mathématique appliquée. Il s'agit de la méthode du *gradient conjugué*.

C'est une méthode qui sert par exemple pour résoudre des systèmes linéaires construits pour calculer les solutions approchées de certaines *équations aux dérivées partielles (EDP)* qui permettent de modéliser des problèmes de physique, mécanique, biologie, économie, etc. On verra en 2e année, en Calcul Scientifique 2, comment cette méthode se combine par exemple avec la *méthode des éléments finis* qui sert, justement, à construire des approximations pour les solutions d'une classe très large d'EDP. Vous verrez aussi ce semestre, en Calcul Scientifique 1, la *méthode des différences finies*, qui joue un rôle similaire.

Par ailleurs, on (re–)verra qu'il y a une équivalence entre la résolution d'un système linéaire avec une matrice symétrique et la minimisation d'une fonctionnelle quadratique. En conséquence, la méthode du gradient conjugué est aussi particulièrement efficace en *optimisation*, dès qu'une fonctionnelle quadratique intervient.

Dans une deuxième partie, on s'intéresse à la résolution de systèmes linéaires *mal posés*, autrement dit qui n'admettent pas de solution, ou, pire, qui en admettent beaucoup trop. C'est une situation qui intervient très fréquemment et dans un grand nombre de problèmes en sciences appliquées. En particulier c'est quelque chose qu'on retrouve fréquemment en *statistique* et *analyse de données* : le cas d'école, que vous connaissez sans doute, est celui de la *régression linéaire*. On verra une méthode générale, qui est la méthode des *moindres carrés*, qui permet de reformuler un problème mal posé en un problème bien posé, qui n'admet plus qu'une seule solution, ou un ensemble raisonnable de solutions, qui est (sont) la (les) meilleure(s) dans un certain sens (le sens des "moindres carrés"). On en profitera pour (re–)voir à ce sujet la *décomposition en valeur singulières (SVD)* d'une matrice, qui est un résultat qui possède aussi un intérêt pour lui-même, et qui est largement utilisé en statistique, pour l'analyse en composantes principales par exemple, ou même aussi pour effectuer de la compression d'images. On terminera en étudiant une technique efficace de résolution d'un problème de moindres carrés.

De nombreux documents et ouvrages existent, où ces deux sujets, très classiques maintenant, sont traités de façon détaillée : voir par exemple [3, 5, 8, 9, 10], ou encore [2, 7] pour un point de vue plus optimisation [1], ou même [6] pour un point de vue plus statistique. Parmi tous ceux-ci, deux ont largement servi à la préparation de ce cours, et vous pouvez vous y référer si vous souhaitez approfondir encore d'avantage. Il y a tout d'abord l'ouvrage de Grégoire Allaire et Sidi M. Kaber intitulé 'Numerical Linear Algebra' [1], et aussi l'ouvrage de Martin Gander et al. intitulé 'Scientific Computing' [4].

Les prérequis nécessaires sont essentiellement des bonnes bases en algèbre linéaire et des notions en analyse des fonctions à plusieurs variables. Une connaissance du cours d'Analyse Numérique de L3 est un plus, mais n'est pas indispensable (il vous est quand même vivement conseillé de vous faire une idée sur son contenu). Avoir suivi le cours de Statistique de L3 et le cours d'Optimisation 1 vous sera aussi utile.

1. Voir également https://who.rocq.inria.fr/Jean-Charles.Gilbert/ensta/optim.html.

Chapitre 2

Résolution de systèmes linéaires symétriques de grande dimension

Nous allons étudier dans ce chapitre la méthode du *gradient conjugué*, qui est devenue au fil du temps la méthode incontournable pour résoudre des systèmes linéaires, à partir du moment où la matrice est symétrique, de grande taille, et creuse. Bien sûr, résoudre un système linéaire avec une matrice symétrique, cela revient à minimiser une fonctionnelle quadratique : on approfondira cette équivalence.

On va commencer par motiver la méthode, puis donner explicitement l'algorithme du gradient conjugué (GC, en abrégé). Ensuite, nous procéderons à son analyse mathématique : comme il s'agit d'une méthode itérative, on va étudier sa convergence vers la solution du problème linéaire qui nous intéresse. On terminera avec des considérations d'ordre pratique.

I Introduction

Commençons par voir ce qui suit :

Exemple :

Soit le *Problème aux Limites (BVP)* suivant :

$$(S)\begin{cases} -\dfrac{d^2u}{dx^2} = f \quad \text{sur} \quad]0,1[\\ u(0) = u(1) = 0. \end{cases}$$

Dans (S), $f : x \mapsto f(x)$ et $u : x \mapsto u(x)$ sont deux fonctions de $]0;1[$ à valeurs réelles. Il s'agit d'une équation différentielle sur un domaine borné (l'intervalle $]0;1[$) auquel on a rajouté des conditions aux limites sur les bords $\{0\}$ et $\{1\}$. Ce système modélise une corde élastique qui, au repos, est confondue avec l'intervalle $[0;1]$. Soumise à une force, elle va se déformer pour occuper une autre position de l'espace. La fonction f modélise cette force (c'est une densité linéique de force), elle est supposée connue. La fonction u représente le champ de déplacement horizontal : le graphe de u représente la nouvelle position de la corde, une fois que les efforts f ont été appliqués. C'est cette fonction l'inconnue du problème, qu'on cherche à trouver une fois que f est appliquée. La première équation de (S) provient en fait du Principe Fondamental de la Statique appliqué à un fragment infinitésimal de la corde : la tension de la corde doit équilibrer la force externe f. Les deux équations suivantes $u(0) = 0$ et $u(1) = 0$ signifient simplement que la corde est attachée à ses deux extrêmités (déplacement nul sur ces extrêmités).

On peut montrer qu'il est possible d'approcher la solution de (S) par :

$$\underbrace{\begin{pmatrix} 2 & -1 & & & & & 0 \\ -1 & 2 & -1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -1 & 2 & -1 \\ 0 & & & & & -1 & 2 \end{pmatrix}}_{A \in \mathcal{M}_n(\mathbb{R})} \underbrace{\begin{pmatrix} u_1 \\ \vdots \\ u_{j-1} \\ u_j \\ u_{j+1} \\ \vdots \\ u_n \end{pmatrix}}_{x \in \mathbb{R}^n} = \underbrace{\begin{pmatrix} f_1 \\ \vdots \\ f_{j-1} \\ f_j \\ f_{j+1} \\ \vdots \\ f_n \end{pmatrix}}_{b \in \mathbb{R}^n} \qquad f_j \simeq f(x_j)\Delta x^2.$$

Autrement dit, la résolution du système linéaire ci-dessus va nous permettre d'obtenir n valeurs $u_1, \ldots, u_n$ qui sont des valeurs approchées de la solution u de (S), dans le sens où $u_j \simeq u(x_j)$, où les x_j sont des points uniformément répartis dans l'intervalle $]0;1[$: $x_j = j\Delta x$ où Δx est le pas de discrétisation.

En fait, ce système linéaire est obtenu à partir de (S) à l'aide de la *méthode des différences finies*, où on a utilisé la formule

$$\frac{d^2u}{dx^2}(x_j) \simeq \frac{u_{j-1} - 2u_j + u_{j+1}}{\Delta x^2}$$

pour avoir une approximation de la première équation de (S).

On remarque qu'il est illusoire de vouloir résoudre le système linéaire manuellement : pour avoir une approximation précise de la solution u, il faut prendre un grand nombre n de points sur l'intervalle $]0;1[$ (et donc Δx le plus petit possible). Il faut donc utiliser l'ordinateur .

Bien sûr, c'est un cas particulier, mais en fait, une large classe de phénomènes physiques (ou chimiques, biologiques, etc) peut être modélisé par des problèmes aux limites comme celui-ci, où interviennent une équation différentielle ou aux dérivées partielles complétée par des conditions de bord. Ces problèmes n'admettent en général pas de solution analytique, et on utilise des techniques d'approximation numérique comme les différences finies ou les éléments finis pour effectuer une résolution approchée. Si l'équation différentielle ou aux dérivées partielles est linéaire, on aboutit au final à la résolution d'un système linéaire.

Remarques. Terminons par les remarques suivantes :

i) La matrice A est symétrique définie positive.

ii) La matrice A est creuse : une large majorité de ses coefficients sont nuls. En fait, ici, elle est même tridiagonale.

On va tirer partie de la structure particulière de A pour introduire une méthode efficace de résolution. Bien sûr, si A est de petite taille, on peut utiliser une méthode directe comme une décomposition de Cholesky pour la résolution. Néanmoins, pour une matrice de grande taille, ce n'est pas idéal, ni en termes de temps de calcul, et encore moins en terme de stockage en mémoire. En fait, on tirer partie du caractère symétrique de A, qui permet de reformuler le problème $Ax = b$ comme un problème de minimisation. Aussi, on va tirer partie du fait que A est creuse.

II La méthode du gradient conjugué

On introduit d'abord la méthode, puis viennent ensuite quelques premières observations. On termine en l'appliquant sur un exemple.

II. 1) Description générale

Revenons maintenant à un problème plus général, et on se propose maintenant tout simplement de résoudre le système suivant :

$$Ax = b, \tag{2.1}$$

où $x, b \in \mathbb{R}^n$, et où la matrice $A \in \mathcal{M}_n(\mathbb{R})$ $(n \geq 1)$ est :

— symétrique : $A^t = A$,

— positive : pour tout $z \in \mathbb{R}^n$, $\langle Az, z\rangle \geq 0$,

— définie : pour tout $z \in \mathbb{R}^n$, si $\langle Az, z\rangle = 0$, alors $z = 0$.

Ci-dessus, la notation $\langle \cdot, \cdot \rangle$ désigne le produit scalaire euclidien dans $\mathbb{R}^n$. On rappelle qu'en conséquence, la forme bilinéaire $\langle A\cdot, \cdot\rangle$ définit un produit scalaire sur $\mathbb{R}^n$. On rappelle aussi que A est diagonalisable dans $\mathbb{R}$ et que toutes ses valeurs propres sont réelles et strictement positives.

On va approcher la solution de (2.1) par itérations successives. On construit donc une suite $x^0, \ldots, x^k, \ldots$ de vecteurs de $\mathbb{R}^n$ qui, si tout va bien, va approcher la solution x de (2.1) lorsque k va tendre vers l'infini. On définit de plus une suite associée de résidus

$$r^k = b - Ax^k,$$

pour tout entier k. Ce résidu (ou plus exactement sa norme) permet de mesurer à quel point l'approximation x^k de x résoud le système (2.1). On remarque en particulier que le résidu devient nul si $x^k = x$. On va donc faire de sorte à ce que la suite des résidus tende vers 0 lorsque k va tendre vers l'infini.

On se donne comme point de départ un vecteur $x^0 \in \mathbb{R}^n$ et le résidu correspondant est $r^0 = b - Ax^0$. On introduit une troisième variable auxilliaire $p^0 = r^0 (= b - Ax^0)$.

La méthode du gradient conjugué consiste alors à calculer la suite de triplets

$$(x^k, r^k, p^k)_{k\in\mathbb{N}}$$

à partir de (x^0, r^0, p^0) à l'aide des relations de récurrence suivantes :

$$\begin{cases} x^{k+1} = x^k + \alpha_k\, p^k \\ r^{k+1} = r^k - \alpha_k\, Ap^k \\ p^{k+1} = r^{k+1} + \beta_k\, p^k \end{cases}$$

où les coefficients intermédiaires α_k et β_k sont obtenus en utilisant les formules suivantes

$$\alpha_k = \frac{\left\|r^k\right\|^2}{\langle Ap^k, p^k\rangle}, \qquad \beta_k = \frac{\left\|r^{k+1}\right\|^2}{\left\|r^k\right\|^2},$$

où $\langle \cdot, \cdot \rangle$ désigne toujours le produit scalaire euclidien dans $\mathbb{R}^n$ et $\|\cdot\|$ la norme euclidienne dans $\mathbb{R}^n$.

On donne l'algorithme, tel qu'il pourrait être implémenté sous Octave, Scilab, Python, etc, dans l'encart (Algorithm 1). Cet algorithme peut être encapsulé dans une fonction, qui prendrait comme paramètres d'entrée la matrice A et le vecteur b du système linéaire. Comme paramètres optionnels, on pourrait rajouter éventuellement la valeur initiale de x (x^0), avec 0 comme valeur par défaut, et la tolérance ε. L'algorithme retourne en sortie la valeur finale de x, solution approchée du système linéaire. Notez aussi le critère d'arrêt pour la boucle WHILE, qui se fait sur le résidu.

Algorithm 1 Algorithme du gradient conjugué

Require: $A \in \mathcal{M}_n(\mathbb{R}), b \in \mathbb{R}^n$ (optionally $x \in \mathbb{R}^n$ and $\varepsilon \in \mathbb{R}$)
  $x \in \mathbb{R}^n$
  $r = b - Ax$
  $p = r$
  $\gamma = \|r\|^2$
  while $\gamma > \varepsilon \ \ (\approx 10^{-12})$ **do**
    $y = Ap$
    $\alpha = \dfrac{\gamma}{< y, p >}$
    $x = x + \alpha p$
    $r = r - \alpha y$
    $\beta = \dfrac{\|r\|^2}{\gamma}$
    $\gamma = \|r\|^2$
    $p = r + \beta p$
  end while
  return $x \in \mathbb{R}^n$

II. 2) Premières observations

On remarquera que dans cet algorithme, A n'est utilisée que pour faire des produits matrice-vecteur. L'algorithme devient alors particulièrement intéressant si on dispose d'une implémentation efficace du produit matrice-vecteur. En particulier, il n'est pas forcément obligatoire de stocker les coefficients de A pour effectuer ce produit, ce qui est un plus si la matrice est de très grande taille.

On va détailler par la suite la construction et la justification de cet algorithme, mais, de façon purement heuristique, on peut voir p^k comme une direction de descente à l'itération k. On montrerera d'ailleurs que ces directions de descentes successives sont

orthogonales dans un certain sens. La dernière itération, qui met à jour les directions de descente, est faite de sorte à rendre les résidus de plus en plus petits.

II. 3) Un premier exemple

On peut essayer d'appliquer l'Algorithme 1 sur des systèmes linéaires simples et de petite taille pour se faire une idée de son fonctionnement et de son comportement. Ceci est l'objet de l'exercice suivant.

Exercice :

1) Soit $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \in \mathcal{M}_n(\mathbb{R})$ et soit $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.
   a) Vérifier que A est symétrique définie positive.
   b) Trouver x solution de $Ax = b$.
   c) Calculer $(x^k)_k$ pour (A, b) et $x^0 = (0, 0)$.
   d) Recommencer avec $b = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$.

2) Soit $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$ et soient $b_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ et $b_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$.
   Calculer $(x^k)_k$ pour b_1 et b_2 avec $x^0 = (0, 0, 0)$.

Une fois ceci fait, passons à l'étude de cette méthode de résolution.

III Minimisation de fonctionnelles quadratiques

Pour étudier le comportement de l'algorithme de gradient conjugué, et, comprendre comment il est construit, il faut commencer par se refaire une idée sur la minimisation de fonctionnelles quadratiques en dimension finie. On va rappeler ici des résultats que vous avez vu en Optimisation 1, en particulier lorsque vous avez vu les méthodes de descente de gradient. On va revoir les choses sous un angle un peu plus 'algèbre linéaire'.

III. 1) Caractérisation de minima globaux

On va commencer déjà par donner des résultats sur la caractérisation de minima globaux de fonctionnelles quadratiques. Donnons déjà un premier résultat qui explicite le lien entre résolution d'un système linéaire symétrique et minimisation d'une fonctionnelle quadratique.

Proposition III.1. Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique et soit la fonctionnelle :

$$\begin{array}{rccc} J: & \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \end{array}.$$

Alors J est différentiable sur $\mathbb{R}^n$ et son gradient est

$$\nabla J(x) = Ax - b.$$

Une condition nécessaire pour que x soit un minimum de J est $Ax = b$.

Exercice : Démontrer le résultat ci-dessus.

Une méthode pour résoudre $Ax = b$ serait alors de faire une descente de gradient sur la fonctionnelle J et on pourrait du coup écrire la récurrence suivante :

$$\begin{cases} x^0 \in \mathbb{R}^n \\ x^{k+1} = x^k - \alpha_k \nabla J(x^k) = x^k - \alpha_k [Ax^k - b], \end{cases}$$

avec α_k un pas de descente donné, fixe ou variable (pour du gradient à pas optimal). Notez que le pas de descente α_k ici est différent de l'expression donnée plus haut pour le gradient conjugué.

Remarques. On peut noter en particulier :

i) La descente de gradient définie ci-dessus est en fait une méthode de Richardson : Rappelons qu'une méthode de Richardson se base sur la décomposition suivante :

$$A = M - N, \qquad \begin{cases} Ax = b \\ (M - N)x = b \end{cases}.$$

Pour retrouver une méthode de Richardson, il suffit donc de remarquer qu'on peut écrire :

$$\frac{1}{\alpha_k}x^{k+1} = \frac{1}{\alpha_k}x^k + (b - Ax^k)$$
$$\underbrace{\left(\frac{1}{\alpha_k}I\right)}_{M} x^{k+1} = \underbrace{\left(\frac{1}{\alpha_k}I - A\right)}_{N} x^k + b.$$

ii) Comme pour le gradient conjugué, on a seulement besoin de connaître le produit matrice-vecteur Ax^k.

iii) Cette méthode n'est pas très judicieuse car elle demande trop d'itérations en pratique afin d'avoir une bonne approximation.

Donnons maintenant une caractérisation complète de l'ensemble des minima d'une fonctionnelle quadratique en dimension finie.

Théorème III.2. *(minima d'une fonction quadratique)*
Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique. Soit :

$$\begin{array}{rccc} J: & \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \end{array}.$$

Alors :

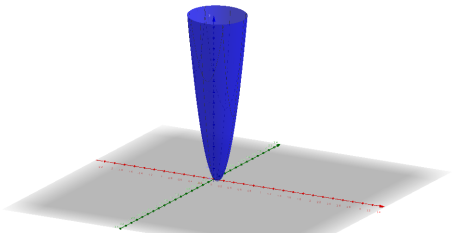
i) Si A est définie positive, J admet un unique minimum sur $\mathbb{R}^n$, solution de $Ax = b$.

ii) Si A est positive indéfinie et $b \in \mathrm{Im}(A)$ alors l'ensemble des minima de J est non-vide et ces minima sont les solutions de $Ax = b$.

iii) Si A est non positive ou si A est positive indéfinie, avec, en plus, $b \notin \mathrm{Im}(A)$, alors J n'admet pas de minimum ($\inf_{\mathbb{R}^n} J = -\infty$).

Exemple :

Pour i), on pose par exemple $A = \begin{pmatrix} 10 & 1 \\ 1 & 10 \end{pmatrix}$ et $b = \begin{pmatrix} 1 & 1 \end{pmatrix}$.



Représentation graphique de $J(x)$

Pour ii), on pose par exemple $A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ et $b = \begin{pmatrix} 1 & 0 \end{pmatrix}$.

Pour iii), on pose par exemple $A = \begin{pmatrix} 10 & 0 \\ 0 & -10 \end{pmatrix}$ et $b = \begin{pmatrix} 1 & 1 \end{pmatrix}$.

Démonstration. La matrice A est symétrique réelle donc elle est diagonalisable avec valeurs propres réelles. Soient $\lambda_1, \ldots, \lambda_n$ ses valeurs propres et $\hat{e}_1, \ldots, \hat{e}_n$ les vecteurs propres associés.
Décomposons x et b dans la base de vecteurs propres :

$$x = \sum_{i=1}^{n} \hat{x}_i \hat{e}_i \quad , \quad b = \sum_{i=1}^{n} \hat{b}_i \hat{e}_i.$$

Alors la fonctionnelle J s'écrit

$$\begin{aligned} J(x) &= \frac{1}{2} \sum_{i=1}^{n} \lambda_i \hat{x}_i^2 - \sum_{i=1}^{n} \hat{b}_i \hat{x}_i \\ (\text{si } \lambda_i \neq 0\ ) \quad &= \frac{1}{2} \sum_{i=1}^{n} \left[\lambda_i \left(\hat{x}_i - \frac{\hat{b}_i}{\lambda_i} \right)^2 - \frac{\hat{b}_i^2}{\lambda_i} \right]. \end{aligned}$$

i) Si A est définie positive alors $\lambda_i > 0$ pour $i = 1, \ldots, n$.
Pour minimiser J il faut et il suffit de minimiser chaque terme

$$\lambda_i \left(\hat{x}_i - \frac{\hat{b}_i}{\lambda_i} \right)^2$$

pour tout $i = 1, \ldots, n$.

On obtient alors que J admet un unique minimum $x = \begin{pmatrix} \frac{\hat{b}_1}{\lambda_1} \\ \vdots \\ \frac{\hat{b}_n}{\lambda_n} \end{pmatrix}$ qui est l'unique solution de $Ax = b$.

ii) Si A est indéfinie positive alors $\lambda_i \geqslant 0$ pour $i = 1, \dots, n$ et il existe $i_0 \in \{1, \dots, n\}$ tel que $\lambda_{i_0} = 0$. De plus, $b \in \mathrm{Im}(A)$. On appelle alors $\tilde{A}$ la restriction de A à $\mathrm{Im}(A)$ et on écrit $Ax = b$ comme suit :

$$\begin{pmatrix} \lambda_{i_1} & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & & & & \vdots \\ \vdots & & \lambda_{i_m} & & & \vdots \\ \vdots & & & 0 & & \vdots \\ \vdots & & & & \ddots & \vdots \\ 0 & \dots & \dots & \dots & \dots & 0 \end{pmatrix} \begin{pmatrix} x_{i_1} \\ \vdots \\ x_{i_m} \\ x_{i_{m+1}} \\ \vdots \\ x_{i_n} \end{pmatrix} = \begin{pmatrix} b_{i_1} \\ \vdots \\ b_{i_m} \\ b_{i_{m+1}} \\ \vdots \\ b_{i_n} \end{pmatrix}.$$

Avec : $m = \mathrm{rg}\,(A)$, $\tilde{A} = \begin{pmatrix} \lambda_{i_1} & & \\ & \ddots & \\ & & \lambda_{i_m} \end{pmatrix}$ et $b_{i_{m+1}} = \dots = b_{i_n} = 0$ car $b \in \mathrm{Im}(A)$.

On peut alors utiliser i) sur la matrice $\tilde{A}$ et on a alors que J admet un unique minimum $x \in \mathrm{Im}(A)$. En conséquence, tout vecteur $x + z$, avec $z \in \ker(A)$ est solution de $Ax = b$ et donc minimum de J.

iii) Si A est indéfinie positive, et $b \notin \mathrm{Im}(A)$, alors cela signifie que $\lambda_i \geqslant 0$ pour $i = 1, \dots, n$ et qu'il existe un indice $i_0 \in \{1, \dots, n\}$ tel que $\lambda_{i_0} = 0$ et $b_{i_0} \neq 0$.
On prend alors $x = t\hat{e}_{i_0}$ avec $t \in \mathbb{R}$. D'où : $J(x) = -b_{i_0} t$.
Alors, avec $t \to \pm\infty$(en fonction du signe de b_{i_0}), on a $\inf_{\mathbb{R}^n} J = -\infty$.
Si maintenant A est non positive, alors il existe au moins $i_0 \in \{1, \dots, n\}$ tel que $\lambda_{i_0} < 0$.
On prend : $x = t\hat{e}_{i_0}, t \in \mathbb{R}$. D'où :

$$J(x) = \frac{1}{2}\lambda_{i_0} t^2 - \hat{b}_{i_0} t.$$

Et donc, $\lim_{t \to \pm\infty} J(t\hat{e}_{i_0}) = -\infty$.

□

La première idée du gradient conjugué est que l'on va chercher à minimiser J sur un sous-espace vectoriel de petite dimension.

On va donc voir ce qui se passe si on cherche à minimiser une fonctionnelle quadratique sur un sous espace.

III. 2) Minima sur un sous-espace

Commençons avec ce premier résultat :

Corollaire III.3. Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive et soit J la forme quadratique associée. Soit F un sous-espace vectoriel de $\mathbb{R}^n$.

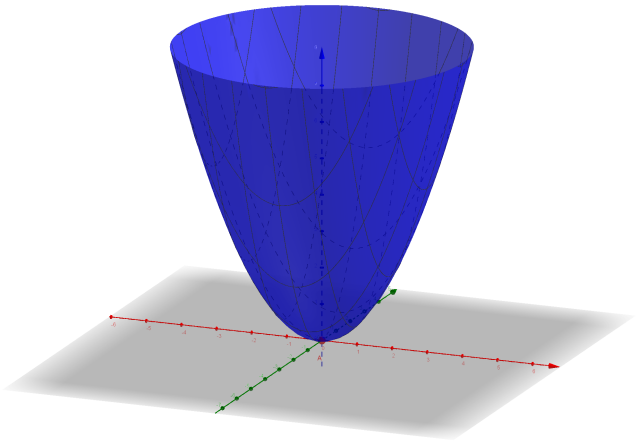
i) Il existe un unique $x_0 \in F$ qui est le minimum global de J (sur F).

ii) Ce même x_0 est l'unique vecteur solution de l'équation variationnelle :

$$\langle Ax_0 - b, y\rangle = 0 \qquad \forall y \in F.$$

Exemple :

Représentation graphique de $J(x) = \dfrac{x_1^2 + x_2^2}{2}$ et du minimum $(0,0,0)$ sur $\mathbb{R}^2$.



Démonstration. Soit $P : \mathbb{R}^n \to F$ une projection orthogonale de $\mathbb{R}^n$ sur F. On a alors :

$$\min_{x\in F} J(x) = \min_{y\in\mathbb{R}^n} \tilde{J}(y) \quad \text{avec} \quad \tilde{J} = J \circ P.$$

Écrivons :

$$\begin{aligned}\tilde{J}(y) &= J(Py)\\ &= \frac{1}{2}\langle APy, Py\rangle - \langle b, Py\rangle\\ &= \frac{1}{2}\langle P^tAPy, y\rangle - \langle P^tb, y\rangle.\end{aligned}$$

On observe que P^tAP est symétrique et positive (mais pas forcément définie). On applique alors le théorème III.2 page 12.
On a $P^tb \in \mathrm{Im}(P^tAP)$.
En effet, dans le cas contraire, on aurait : $\inf_{\mathbb{R}^n} \tilde{J}(y) = -\infty$, ce qui est impossible car :

$$\min_{\mathbb{R}^n} \tilde{J}(y) = \min_{F} J(x) \geqslant \min_{\mathbb{R}^n} J(x) > -\infty.$$

Donc, $\tilde{J}$ admet au moins un minimum sur $\mathbb{R}^n$ solution de $P^tAPy = P^tb$.
Soient maintenant y_1 et y_2 deux solutions de $P^tAPy = P^tb$.
Alors on vérifie $P^tAP(y_1 - y_2) = 0$ et donc :

$$\begin{aligned}&\langle P^tAP(y_1 - y_2), y_1 - y_2\rangle = 0\\ &\langle AP(y_1 - y_2), P(y_1 - y_2)\rangle = 0\\ &Py_1 = Py_2. \qquad (\star)\end{aligned}$$

Pour la dernière ligne, on a utilisé le fait que A est symétrique, définie et positive, et donc $\langle A\cdot, \cdot\rangle$ est un produit scalaire. Soit alors $x_0 = Py_1$, on a bien $x_0 \in F$.

Il vient ensuite que x_0 est l'unique minimum de J sur F. En effet supposons que $\tilde{x}$ est un minimum de J sur F , on a alors :

$$P\tilde{x} = \tilde{x}$$

car $\tilde{x}$ est dans F. Mais alors $\tilde{J}(\tilde{x}) = J(P(\tilde{x})) = J(\tilde{x})$ et donc $\tilde{x}$ est un minimum, sur $\mathbb{R}^n$, de $\tilde{J}$. Il s'en suit alors, en utilisant la caractérisation $(\star)$:

$$\tilde{x} = P\tilde{x} = Py_1 = x_0.$$

En conséquence, x_0 est bien le seul minimum de J sur F.
Soit maintenant $y \in F$, on peut écrire :

$$\begin{aligned}\langle Ax_0 - b, y\rangle &= \langle Ax_0 - b, Py\rangle\\ &= \langle P^tAx_0 - P^tb, y\rangle\\ &= \langle \underbrace{P^tAPx_0 - P^tb}_{=0}, y\rangle.\end{aligned}$$

Si on avait de plus $\tilde{x} \in F$ tel que :

$$\begin{aligned} & \langle A\tilde{x} - b, y\rangle = 0 \quad \forall y \in F \\ \text{(et on aurait alors)} \qquad & \langle A(\tilde{x} - x_0), y\rangle = 0 \\ \text{puis en prenant } y = \tilde{x} - x_0 \ & \in F \ : \ \tilde{x} - x_0 = 0. \end{aligned}$$

Donc x_0 est l'unique vecteur de F qui vérifie $\langle Ax_0 - b, y\rangle = 0$ pour tout $y \in F$. □

Une fois ceci démontré, il reste maintenant à bien choisir les sous-espaces de dimension finie sur lesquels on souhaite minimiser J.

IV Espaces de Krylov

La notion d'espace de Krylov est la notion fondamentale qui permet de construire des méthodes itératives efficaces pour résoudre des systèmes linéaires : gradient conjugué pour une matrice symétrique, et GMRES par exemple pour des matrices quelconques. Il faut bien garder à l'esprit que la dimension du système linéaire à résoudre, n, est très grande. Il faut donc à la fois avoir des espaces de dimension toute petite par rapport à n, mais aussi construits judicieusement à partir de la matrice A et du vecteur b. Commençons déjà par définir ces espaces.

Définition IV.1. Soient $r \in \mathbb{R}^n$, $r \neq 0$, $k \geqslant 0$ et $A \in \mathcal{M}_n(\mathbb{R})$.
L'espace de Krylov associé à (r, A) est :

$$K_k = K_k(A, r) = \text{vect}\{r, Ar, A^2r, \ldots, A^kr\} \subset \mathbb{R}^n.$$

Le Lemme suivant va jouer un rôle important dans la construction et l'analyse de la méthode de gradient conjugué.

Lemme IV.1. La suite $(K_k)_{k\geqslant 0}$ est croissante :

$$K_k \subset K_{k+1}.$$

De plus, il existe un unique $k_0 \in \{0, \dots, n-1\}$ tel que :

$$\begin{cases} \forall k \in \{0, \dots, k_0\} & , \ \dim K_k = k+1 \\ \forall k \in \{k_0+1, \dots, n-1\} & , \ \dim K_k = k_0+1. \end{cases}$$

L'indice k_0 est appelé dimension critique de Krylov.

Démonstration. Tout d'abord, comme $r \neq 0$ on a $\dim K_0 = 1$.
De plus, $K_k \subset \mathbb{R}^n$ donc $\dim K_k \leqslant n \ , \ \forall k \in \mathbb{N}$.
Soit $k_0 = \max \underbrace{\{k \in \mathbb{N}, \dim K_j = j+1 \ , \ \forall j \leqslant k\}}_{\neq \emptyset \qquad \subset \{0,\dots,n-1\}}$.
Par définition de l'entier k_0 :

$$\dim K_{k_0+1} < (k_0+1)+1.$$

D'où : $\dim K_{k_0+1} \leqslant k_0+1$.
De plus : $K_{k_0} \subset K_{k_0+1}$ et $\dim K_{k_0} = k_0+1$.
On a donc nécessairement :

$$\dim K_{k_0+1} = k_0+1.$$

Ce qui entraîne :

$$A^{k_0+1}r \in \text{vect}(r, Ar, \dots, A^{k_0}r).$$

Et alors :

$$A^{k_0+2}r \in \text{vect}(Ar, A^2r, \dots, A^{k_0+1}r) \subset \text{vect}(r, Ar, \dots, A^{k_0}r).$$

Par induction sur k :

$$\forall k \geqslant k_0 \ , \ A^k r \in \text{vect}(r, Ar, \dots, A^{k_0}r) = K_{k_0}.$$

Finalement on en déduit que pour $k \geqslant k_0$: $K_k = K_{k_0}$, ce qui termine la preuve. □

Remarque. On va maintenant faire une première observation sur la méthode de descente de gradient à pas variable.

Proposition IV.2. Soit la suite :

$$\begin{cases} x^0 \in \mathbb{R}^n, \\ x^{k+1} = x^k + \alpha_k(b - Ax^k). \end{cases}$$

Soit : $r^k = b - Ax^k$ (résidus). On vérifie que :

i) $r^k \in K_k(A, r^0)$,

ii) $x^{k+1} \in [x^0 + K_k]$.

Exercice : Démontrer le résultat ci-dessus.

On réalise ainsi que, à chaque itération, on cherche une nouvelle approximation de la solution dans un espace de Krylov. On voit alors qu'une amélioration évidente de cette méthode de descente de gradient consisterait à chercher le minimum de J dans cet espace. On va voir qu'en fait, c'est ce que fait le gradient conjugué. Commençons auparavant par une première observation.

Proposition IV.3. Soit $(x^k)_{k\geqslant 0}$ une suite de $\mathbb{R}^n$. Soient $r^0 = b - Ax^0$ et $K_k = K_k(A, r^0)$.
Si $x^{k+1} \in [x^0 + K_k]$ alors $r^{k+1} = b - Ax^{k+1} \in K_{k+1}$.

Exercice : Démontrer le résultat ci-dessus.

V Reformulation et analyse du gradient conjugué

Nous allons maintenant reformuler la méthode de gradient conjugué, en s'appuyant sur ce qu'on a vu précédemment, en particulier la reformulation en un problème de minimisation puis son approximation en sous-problèmes qui font intervenir des espaces de Krylov.

V. 1) Reformulation de la méthode

On considère toujours une matrice $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive.
Le point de départ est le suivant : soient $x^0 \in \mathbb{R}^n$ et $r^0 = b - Ax^0$.
On suppose qu'on a obtenu à l'itération $k \geqslant 0$ le couple (x^k, r^k).
On va chercher $x^{k+1} \in [x^0 + K_k]$ et on calcule ensuite :

$$r^{k+1} = b - Ax^{k+1}.$$

On se ramène ainsi au problème suivant à chaque itération : comment bien choisir x^{k+1} ? (afin d'approcher au mieux x). En fait, il y a deux façons équivalentes de choisir l'itérée x^{k+1} :

Définition V.1. On prend x^{k+1} tel que : $r^{k+1} \perp K_k$.

Définition V.2. On prend x^{k+1} qui minimise :

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

sur $[x^0 + K_k]$.

On voit déjà que les deux possibilités permettront sans doute d'aboutir à un algorithme efficace : avec la première définition, on oblige le résidu à être orthogonal à des sous-espaces de plus en plus 'gros' ; avec la deuxième définition, on minimise directement J sur cette même suite de sous-espaces. Voyons déjà un premier résultat très important.

Théorème V.1. Soit la suite (x^k, r^k) définie ci-dessus.

i) Pour chaque définition V.1 et V.2 , on obtient un unique vecteur qui vérifie les propriétés souhaitées.

ii) Les deux définitions coïncident exactement.

Démonstration. On définit la fonction auxiliaire suivante :

$$\begin{array}{rccc} \Phi : & K_k & \longrightarrow & \mathbb{R} \\ & y & \longmapsto & J(x^0 + y) - J(x^0) \end{array}$$

Remarquons déjà que :

$$\begin{aligned}\Phi(y) &= \frac{1}{2}\langle A(x^0+y), x^0+y\rangle - \langle b, x^0+y\rangle - J(x^0)\\ &= \frac{1}{2}\langle Ax^0, x^0\rangle - \langle b, x^0\rangle - J(x^0) + \frac{1}{2}\langle Ay, y\rangle + \langle Ax^0, y\rangle - \langle b, y\rangle\\ &= J(x^0) - J(x^0) + \frac{1}{2}\langle Ay, y\rangle - \langle r^0, y\rangle\\ &= \frac{1}{2}\langle Ay, y\rangle - \langle r^0, y\rangle.\end{aligned}$$

Ensuite, il faut garder en tête que minimiser J sur $[x^0 + K_k]$ revient à minimiser Φ sur K_k.
Alors, d'après le corollaire III.3 page 15, Φ admet un unique minimiseur y^{k+1} sur K_k et donc $x^0 + y^{k+1}$ est l'unique minimiseur de J sur $[x^0 + K_k]$.
On peut donc poser : $x^{k+1} = x^0 + y^{k+1}$.
Le corollaire III.3 page 15 nous assure de plus que $y^{k+1} \in K_k$ est l'unique solution de l'équation variationnelle :

$$\langle Ay^{k+1} - r^0, y\rangle = 0 \qquad \forall y \in K_k.$$

Alors, pour $y \in K_k$:

$$\begin{aligned}\langle r^{k+1}, y\rangle &= \langle b - Ax^{k+1}, y\rangle\\ &= \langle b - Ax^0 - Ay^{k+1}, y\rangle\\ &= \langle r^0 - Ay^{k+1}, y\rangle\\ &= 0.\end{aligned}$$

Donc $x^{k+1} \in [x^0 + K_k]$ est bien le seul vecteur de $[x^0 + K_k]$ tel que $r^{k+1} \perp K_k$. □

V. 2) La méthode converge

Le résultat suivant est le clou du spectacle sur cette partie. Il garantit qu'en procédant comme ci-dessous, on atteint exactement la solution en un nombre fini d'itérations.

Théorème V.2. En n itérations au plus, la suite (x^k, r^k) permet d'obtenir exactement $x \in \mathbb{R}$, solution unique de $Ax = b$.

Remarque. La méthode ainsi définie est une méthode directe. Elle permet d'obtenir la solution du système linéaire $Ax = b$ (ou le minimum de J) en un nombre fini d'itérations. Ce résultat se vérifie rarement en pratique, car il suppose qu'on puisse travailler en arithmétique exacte.

Démonstration. On reprend la notation k_0 pour désigner la dimension critique de la suite $(K_k)_{k\geqslant 0}$. On va distinguer en fait deux cas de figure. En effet, soit la suite emboîtée (K_k) finit par 'envahir' tout l'espace $\mathbb{R}^n$, soit au bout d'un nombre fini d'itérations, elle 'stagne' comme sous-espace propre de $\mathbb{R}^n$.

Cas 1 : Si $k_0 = n - 1$.
Alors $\dim K_{k_0} = n$ et donc $[x^0 + K_k] = \mathbb{R}^n$. Avec la définition V.2 page 20 on a que x^{k+1} est le minimum de J sur $\mathbb{R}^n$. D'après le théorème III.2 page 12 on a alors $x = x^{k_0+1} = x^n$ qui est l'unique solution de $Ax = b$ (on a alors convergé en au plus n itérations).

Cas 2 : Si $k_0 < n - 1$.
Alors : $\forall k \geqslant k_0$, $\dim K_k = k_0 + 1$ (d'après le lemme IV.1 page 18) et donc $A^{k_0+1}r^0 \in K_{k_0}$. On peut alors écrire :

$$A^{k_0+1}r^0 = \sum_{i=0}^{k_0} \alpha_i A^i r^0, \qquad (\alpha_i \in \mathbb{R}). \tag{2.2}$$

On remarque de plus que $\alpha_0 \neq 0$ car sinon, en multipliant 2.2 par A^{-1}, on aurait $A^{k_0}r^0 \in K_{k_0-1}$ et $k_0 - 1$ serait la dimension critique.
On écrit donc :

$$A^{k_0+1}r^0 = \sum_{i=1}^{k_0} \alpha_i A^i r^0 + \alpha_0 (\underbrace{b - Ax^0}_{r^0})$$

$$\frac{1}{\alpha_0} A^{k_0+1}r^0 - \frac{1}{\alpha_0} \sum_{i=1}^{k_0} \alpha_i A^i r^0 + Ax^0 = b$$

$$\underbrace{A(\frac{1}{\alpha_0} A^{k_0} r^0 - \frac{1}{\alpha_0} \sum_{i=1}^{k_0} \alpha_i A^{i-1} r^0 + x^0)}_{=x \text{ car } Ax=b \text{ n'admet qu'une solution}} = b.$$

Mais alors, avec x ainsi caractérisé, on a aussi :

$$x \in [x^0 + K_{k_0}].$$

En conséquence x est aussi l'unique minimiseur de J sur $[x^0 + K_k]$. Comme d'après la définition V.2 page 20 nous avons aussi que x^{k_0+1} est l'unique minimiseur de J sur $[x^0 + K_k]$, on a forcément :

$$x = x^{k_0+1}.$$

Ce qui permet de conclure (et on a alors convergé en en au plus k_0+1 itérations).

On a donc bien démontré dans les deux cas qu'on obtenait la solution exacte en au plus n itérations. □

V. 3) On retrouve bien la méthode de départ

La dernière étape consiste maintenant à se convaincre qu'en ayant procédé de la sorte, on a obtenu en fait l'algo décrit à la Section II. De toutes façons, la suite (x^k, r^k) ainsi construite, même si elle est parfaitement bien définie, n'est pas évidente à calculer d'un point de vue pratique. On va donc reformuler l'algorithme que nous avons construit afin d'aboutir à une définition plus directe et facile à implémenter de x^{k+1}. On va alors se rendre compte qu'on débouche exactement sur l'Algorithme 1 donné en II.

On va commencer déjà à voir la proposition suivante :

Proposition V.3. Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive. Soit (x^k, r^k) la suite définie précédemment. Définissons, pour $k \in \mathbb{N}$: $d^k = x^{k+1} - x^k$.
Alors :

i) Les espaces de Krylov K_k sont caractérisés par

$$\begin{aligned} K_k(A, r^0) &= \text{vect}(r^0, \dots, r^k) \\ &= \text{vect}(d^0, \dots, d^k). \end{aligned}$$

ii) $(r^k)_{0 \leqslant k \leqslant n-1}$ est une famille orthogonale :

$$(k \neq l) \Longrightarrow \langle r^k, r^l \rangle = 0.$$

iii) $(d^k)_{0 \leqslant k \leqslant n-1}$ est une famille conjuguée par rapport à A :

$$(k \neq l) \Longrightarrow \langle Ad^k, d^l \rangle = 0.$$

Remarque. C'est de ce dernier point iii) que vient le nom de gradient conjugué.

Exercice : Démontrer le résultat ci-dessus.

Maintenant, l'idée va être d'utiliser une variable auxiliaire (le fameux p^k qu'on avait vu en Section 1).

Théorème V.4. Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive, soit $(x^k)_k$ la suite du gradient conjugué et soit $(r^k = b - Ax^k)_k$ la suite des résidus associée.
Alors, il existe une suite $(p^k)_k$, A–conjuguée, donnée par :

$$\begin{cases} p^0 = r^0 = b - Ax^0, \\ \forall k \in \{0, \dots, k_0\} \ : \ \begin{cases} x^{k+1} = x^k + \alpha_k p^k, \\ r^{k+1} = r^k - \alpha_k A p^k, \\ p^{k+1} = r^{k+1} + \beta_k p^k, \end{cases} \end{cases}$$

avec :

$$\alpha_k = \frac{\left\| r^k \right\|^2}{\langle Ap^k, p^k \rangle}, \qquad \beta_k = \frac{\left\| r^{k+1} \right\|^2}{\left\| r^k \right\|^2}.$$

Démonstration. On se limite à $k \leqslant k_0$.
Pour chaque indice $k \in \{0, \dots, k_0\}$ on construit p^k orthogonale par rapport au produit scalaire $\langle A\cdot, \cdot \rangle$ par application de Gram-Schmidt à la suite $(r^k)_k$:

$$p^k = r^k + \sum_{j=0}^{k-1} \beta_{kj} p_j.$$

On veut pour $i < k$:

$$\langle Ap^k, p^i \rangle = \langle Ar^k, p^i \rangle + \sum_{j=0}^{k-1} \beta_{kj} \underbrace{\langle Ap^j, p^i \rangle}_{=0 \text{ si } i \neq j} = 0$$
$$\implies \langle Ar^k, p^i \rangle + \beta_{ik} \langle Ap^i, p^i \rangle = 0.$$

Il suffit donc de prendre : $\beta_{ik} = -\dfrac{\langle Ar^k, p^i \rangle}{\langle Ap^i, p^i \rangle}$.
On a vu précédemment que d^k est A–conjuguée :

$$\langle Ad^k, d^i \rangle = 0 \quad \text{si} \quad k \neq i$$

On a également vu que : $K_k = \text{vect}(d^0, \ldots, d^k)$.
De plus, $K_k = \text{vect}(p^0, \ldots, p^k)$ car $K_k = \text{vect}(r^0, \ldots, r^k)$.
L'unicité (au signe près) de la construction par la méthode de Gram-Schmidt implique alors la relation suivante entre les d^k et les p^k :

$$\frac{d^k}{\|d^k\|} = \pm \frac{p^k}{\|p^k\|}.$$

En se souvenant que $d^k = x^{k+1} - x^k$ on a alors qu'il existe $\alpha_k \in \mathbb{R}$ tel que :

$$x^{k+1} = x^k + \alpha_k p^k.$$

Voyons maintenant :

$$\begin{aligned}
\langle Ar^k, p^j \rangle &= \langle r^k, Ap^j \rangle \quad j \leqslant k-1 \\
&= \langle r^k, \frac{1}{\alpha_j} A(x^{j+1} - x^j) \rangle \\
&= \frac{1}{\alpha_j} \langle r^k, (Ax^{j+1} - b) + (b - Ax^j) \rangle \\
&= \frac{1}{\alpha_j} \left(- \underbrace{\langle r^k, r^{j+1} \rangle}_{=0 \quad \text{pour } j \leqslant k-2} + \underbrace{\langle r^k, r^j \rangle}_{=0} \right) \qquad (\mathcal{R}).
\end{aligned}$$

On reprend $p^k = r^k + \sum_{j=0}^{k-1} \beta_{jk} p^j$, ainsi que l'expression :

$$\beta_{jk} = -\frac{\langle Ar^k, p^j \rangle}{\langle Ap^j, p^j \rangle}.$$

En combinant avec la relation $(\mathcal{R})$ établie précédemment on aboutit à :

$$\beta_{jk} = \begin{cases} 0 \quad \text{si} \quad j \leqslant k-2 \\ -\dfrac{\langle Ar^k, p^{k-1} \rangle}{\langle Ap^{k-1}, p^{k-1} \rangle} = \dfrac{1}{\alpha_{k-1}} \dfrac{\left\| r^k \right\|^2}{\langle Ap^{k-1}, p^{k-1} \rangle} \quad (j = k-1) \end{cases}.$$

Et donc aussi à $p^k = r^k + \beta_{k-1,k} p^{k-1}$.
On peut poser finalement :

$$\beta_{k-1} = \beta_{k-1,k}$$

et on remarque aussi

$$\begin{aligned} r^{k+1} &= b - Ax^{k+1} \\ &= b - Ax^k - \alpha_k Ap^k \\ &= r^k - \alpha_k Ap^k. \end{aligned}$$

On a bien obtenu les relations de récurrence voulues pour le triplet (x^k, r^k, p^k). Pour terminer, il faut trouver les valeurs souhaitées de α_k et de β_k. Commençons par utiliser les relations :

$$r^k = p^k - \beta_{k-1} p^{k-1}, \qquad \langle Ap^k, p^{k-1} \rangle = 0.$$

Donc :

$$\begin{aligned} \langle Ap^k, r^k \rangle &= \langle Ap^k, p^k \rangle - \beta_{k-1} \underbrace{\langle Ap^k, p^{k-1} \rangle}_{=0} \\ &= \langle Ap^k, p^k \rangle. \end{aligned}$$

Souvenons nous maintenant que :

$$0 = \langle r^{k+1}, r^k \rangle = \langle r^k, r^k \rangle - \alpha_k \langle Ap^k, r^k \rangle.$$

On en déduit :

$$\begin{aligned} \langle Ar^k, p^k \rangle &= \frac{1}{\alpha_k} \left\| r^k \right\|^2, \\ \alpha_k &= \frac{\left\| r^k \right\|^2}{\langle Ar^k, p^k \rangle} = \frac{\left\| r^k \right\|^2}{\langle Ap^k, p^k \rangle}. \end{aligned}$$

Finalement :

$$\beta_{k-1} = \frac{\langle Ap^{k-1}, p^{k-1} \rangle}{\left\| r^{k-1} \right\|^2} \frac{\left\| r^k \right\|^2}{\langle Ap^{k-1}, p^{k-1} \rangle} = \frac{\left\| r^k \right\|^2}{\left\| r^{k-1} \right\|^2}.$$

□

Remarque. On peut démontrer la réciproque de ce théorème, autrement dit qu'en vérifiant les relations de récurrence du Gradient Conjugué données dans l'énoncé, on retrouve la suite (x^k) où à chaque étape, J est minimisée dans l'espace de Krylov $[x^0 + K_k]$ (voir le livre de Grégoire Allaire et Sidi M. Kaber pour le détail de la preuve).

Déterminons maintenant le nombre d'opérations de cet algorithme. Dans le pire cas possible (cas défavorable), on aura $k_0 + 1 = n$ itérations. A chaque itération, l'opération la plus couteuse est le produit matrice-vecteur ($y = Ap$), qui demande n^2 opérations (si la matrice n'est pas creuse). On a donc un coût total (pessimiste) $N_{op} \simeq n^3$. Dans ce cas, on est moins bons que Cholesky ($n^3/6$).

Pour que l'algorithme soit efficace, il faut donc déjà avoir une implémentation efficace du produit matrice-vecteur, en particulier pour des matrices creuses. Aussi, il faut pouvoir arrêter l'algorithme au bout d'un nombre k d'itérations bien inférieur à n, ce qui est possible en pratique car, pour des matrices bien conditionnées, on peut atteindre une très bonne précision sur x en quelques itérations seulement. Bien sûr, on n'obtient plus la solution exacte alors, mais de toutes façons cette solution est rarement atteinte à cause de la représentation des réels en arithmétique finie et des erreurs d'arrondis qu'elle provoque.

VI Vitesse de convergence

Dans cette section, on va considérer le gradient conjugué comme une méthode itérative. On va voir que cette méthode converge rapidement (on parle de convergence quadratique), ce qu'on va établir via le théorème suivant, qui est le dernier gros théorème de ce chapitre :

Théorème VI.1. Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive et soit $x \in \mathbb{R}^n$ solution de $Ax = b$.
Soit $(x^k)_k$ la suite de gradient conjugué.
Alors :

$$\left\|x^k - x\right\| \leqslant 2\sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \left\|x^0 - x\right\|, \quad \text{avec} \quad \kappa = \mathrm{cond}_2(A)\left(= \frac{\lambda_n}{\lambda_1}\right).$$

Dans la formule ci-dessus, $\mathrm{cond}_2(A)$ désigne le conditionnement matriciel (dans la norme matricielle induite par la norme euclidienne) de A, et λ_1, λ_n, les plus petite et plus grandes valeurs propres de A.

Pour prouver ce théorème, nous aurons besoin du résultat suivant, qui sera admis (voir le livre de Grégoire Allaire et Sidi M. Kaber pour la preuve).

Proposition VI.2. Soient $a, b \in \mathbb{R}$ avec $a < b$. Soit $\mathbb{P}_k^\beta$ l'ensemble des polynômes Q de degré k qui vérifient $Q(\beta) = 1$ avec $\beta \notin [a, b]$. Soit :

$$\begin{array}{rccc} \psi : & \mathbb{P}_k^\beta & \longrightarrow & \mathbb{R} \\ & Q & \longmapsto & \max\limits_{x\in[a,b]} |Q(x)| \end{array}.$$

La fonction ψ admet un unique minimiseur sur $\mathbb{P}_k^\beta$ qui est

$$\frac{T_k\left(\dfrac{2x-(a+b)}{b-a}\right)}{T_k\left(\dfrac{2\beta-(a+b)}{b-a}\right)}$$

avec T_k le polynôme de Chebyshev de degré k.

Rappelons que le polynôme de Chebyshev de degré k est l'unique polynôme de degré k qui coïncide sur $[-1, 1]$ avec la fonction

$$t \to \cos(k \arccos t).$$

Démonstration. (Théorème VI.1 page 27)
Rappelons que x^k est le minimum sur $[x^0 + K_{k-1}]$ de :

$$\begin{aligned} J(z) &= \frac{1}{2}\langle Az, z\rangle - \langle b, z\rangle \\ &= \frac{1}{2}\langle A(z-x), z-x\rangle - \langle b, z\rangle - \frac{1}{2}\langle Ax, x\rangle + \langle \underbrace{Ax}_{=b}, z\rangle \\ &= \frac{1}{2}\langle A(z-x), z-x\rangle - \frac{1}{2}\langle Ax, x\rangle \\ &= \frac{1}{2}\|z-x\|_A^2 - \frac{1}{2}\langle Ax, x\rangle \quad \text{avec} \quad \|z-x\|_A = \langle A(z-x), z-x\rangle^{\frac{1}{2}}. \qquad (\star\star) \end{aligned}$$

Nous avons d'abord ajouté et retranché x, puis simplifié.

Ensuite rappelons-nous des relations : $\begin{cases} x^{k+1} &= x^k + \alpha_k p^k, \\ r^{k+1} &= r^k - \alpha_k A p^k, \\ p^{k+1} &= r^{k+1} + \beta_k p^k. \end{cases}$

Avec une récurrence immédiate, on déduit de la première relation :

$$x^k = x^0 + \sum_{j=0}^{k-1} \alpha_j p^j. \tag{2.3}$$

Et en utilisant les deux relations suivantes :

$$\begin{aligned} p^1 &= r^1 + \beta_0 p^0 \\ &= r^0 - \alpha_0 A p^0 + \beta_0 p^0. \end{aligned}$$

Par induction sur k, on vérifie alors qu'on a p^k qui est un polynôme de degré k en A par rapport à p^0. On peut donc écrire (2.3) comme :

$$x^k = x^0 + q_{k-1}(A) p^0$$

avec q_{k-1} un polynôme de degré inférieur ou égal à $k-1$.
De plus on peut écrire :

$$p^0 = r^0 = b - Ax^0 = A(x - x^0).$$

On appelle $e^k = x^k - x$ l'erreur à l'itération k. On obtient alors en utilisant les relations ci-dessus :

$$\begin{aligned} e^k = x^k - x &= x^0 + q_{k-1}(A) p^0 - x \\ &= x^0 + q_{k-1}(A) A (x - x^0) - x \\ &= (I - A q_{k-1}(A)) e^0 \\ &= Q_k(A) e^0, \end{aligned}$$

avec Q_k un polynôme de degré inférieur ou égal à k.

Maintenant on prend $u_1, \ldots, u_n$ une base orthonormée de vecteurs propres de A avec $\lambda_1, \ldots, \lambda_n$ les valeurs propres associées. Décomposons e^0 dans cette base, puis

calculons sa norme :

$$\begin{aligned} e^0 &= \sum_{j=0}^{n} e_j^0 u_j \\ \left\|e^0\right\|_A^2 &= \langle Ae^0, e^0\rangle \\ &= \langle \sum_{j=0}^{n} \lambda_j e_j^0 u_j, \sum_{l=0}^{n} e_l^0 u_l\rangle \\ &= \sum_{j=0}^{n} \lambda_j |e_j^0|^2. \end{aligned}$$

Ici, nous avons utilisé la relation :

$$Ae^0 = A\left(\sum_{j=0}^{n} e_j^0 u_j\right) = \sum_{j=0}^{n} e_j^0 (Au_j) = \sum_{j=0}^{n} (\lambda_j e_j^0) u_j.$$

On fait de même avec e^k. On a d'abord :

$$e^k = \sum_{j=0}^{n} e_j^k u_j,$$

et puis on utilise encore les relations

$$e^k = Q_k(A)e^0 = Q_k(A)\left(\sum_{j=0}^{n} e_j^0 u_j\right) = \sum_{j=0}^{n} e_j^0 (Q_k(A)u_j) = \sum_{j=0}^{n} (Q_k(\lambda_j) e_j^0) u_j,$$

et (en procédant de même)

$$Ae^k = \sum_{j=0}^{n} (\lambda_j Q_k(\lambda_j) e_j^0) u_j,$$

ces relations découlant du fait que la matrice $Q_k(A)$ est aussi diagonale dans la base de vecteurs propres de A, avec pour valeurs propres associées $Q_k(\lambda_1), \ldots, Q_k(\lambda_n)$ (et ceci est vrai pour n'importe quel polynôme en la matrice A). On peut alors calculer

$$\begin{aligned} \left\|e^k\right\|_A^2 &= \langle Ae^k, e^k\rangle \\ &= \langle \sum_{j=0}^{n} (\lambda_j Q_k(\lambda_j) e_j^0) u_j, \sum_{l=0}^{n} (Q_k(\lambda_l) e_l^0) u_l\rangle \\ &= \sum_{j=0}^{n} \lambda_j \left(Q_k(\lambda_j) e_j^0\right)^2. \end{aligned}$$

Soit maintenant $z \in [x^0 + K_{k-1}]$, on a

$$x - z = (x - x^0) + (x^0 - z) = e^0 + (x^0 - z)$$

et comme $z - x^0 \in K_{k-1}$, on peut écrire

$$z - x^0 = q(A)r^0 = q(A)A(x - x^0) = -q(A)Ae^0,$$

avec q un polynôme quelconque de degré inférieur ou égal à $k-1$ (on a utilisé de nouveau la relation $r^0 = A(x - x^0)$. Il en découle alors

$$x - z = (I - q(A)A)e^0,$$

et on note $Q(A) = I - q(A)A$. On remarque que $Q(0) = 1 - q(0)0 = 1$. Et donc lorsque q décrit $\mathbb{P}_{k-1}$, alors Q décrit

$$\mathbb{P}_k^0 (= \{Q \in \mathbb{P}_k \mid Q(0) = 1\}).$$

En décomposant alors $x - z$ dans la base orthonormée de vecteurs propres $u_1, \ldots, u_k$, on obtient alors

$$\|x - z\|_A^2 = \sum_{j=1}^{n} \lambda_j \left(Q(\lambda_j)e_j^0\right)^2.$$

Comme a vu en début de preuve ($\star\star$) que le gradient conjugué minimise $\|x - z\|_A^2$ sur $[x^0 + K_{k-1}]$ on obtient finalement :

$$\left\|e^k\right\|_A^2 = \left\|x - x^k\right\|_A^2 = \min_{Q \in \mathbb{P}_k^0} \sum_{j=1}^{n} \lambda_j \left(Q(\lambda_j)e_j^0\right)^2.$$

On peut alors majorer :

$$\begin{aligned}
\left\|e^k\right\|_A^2 &\leqslant \min_{Q \in \mathbb{P}_k^0} \left[\max_{1 \leqslant l \leqslant n} (Q(\lambda_l))^2\right] \sum_{j=1}^{n} \lambda_j (e_j^0)^2 \\
&\leqslant \left\|e^0\right\|_A^2 \min_{Q \in \mathbb{P}_k^0} \left[\max_{\lambda_1 \leqslant x \leqslant \lambda_n} (Q(x))^2\right] \\
&\leqslant \left\|e^0\right\|_A^2 \left(\min_{Q \in \mathbb{P}_k^0} \left[\max_{\lambda_1 \leqslant x \leqslant \lambda_n} |Q(x)|\right]\right)^2.
\end{aligned}$$

On rappelle que le polynôme Q qui vérifie :

$$\min_{Q \in \mathbb{P}_k^0} \left[\max_{a \leqslant x \leqslant b} |Q(x)|\right]$$

et $Q(0) = 1$ est :

$$\frac{T_k\left(\dfrac{2x-(a+b)}{b-a}\right)}{T_k\left(\dfrac{2\cdot 0-(a+b)}{b-a}\right)}.$$

On vérifie d'ailleurs que $0 \notin [\lambda_1, \lambda_n]$ car $\lambda_1 > 0$ (A est symétrique, définie et positive). On a alors, en utilisant le fait que $|T_k|$ a pour valeur maximale 1 :

$$\max_{a\leqslant x\leqslant b}\left|\frac{T_k\left(\dfrac{2x-(a+b)}{b-a}\right)}{T_k\left(\dfrac{2\cdot 0-(a+b)}{b-a}\right)}\right| = \frac{1}{\left|T_k\left(\dfrac{2\cdot 0-(a+b)}{b-a}\right)\right|}.$$

Dans notre cas, $\beta = 0$, $a = \lambda_1$, $b = \lambda_n$. D'où :

$$\min_{Q\in\mathbb{P}_k^0}\left[\max_{\lambda_1\leqslant x\leqslant\lambda_n}|Q(x)|\right] = \frac{1}{\left|T_k\left(\dfrac{\lambda_1+\lambda_n}{\lambda_n-\lambda_1}\right)\right|}.$$

En utilisant la relation entre le conditionnement κ de A et le ratio des valeurs propres extrêmes, on remarque :

$$\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1} = \frac{\dfrac{\lambda_n}{\lambda_1}+1}{\dfrac{\lambda_n}{\lambda_1}-1} = \frac{\kappa+1}{\kappa-1}.$$

On utilise maintenant la relation suivante sur les polynômes de Chebyshev (voir le livre de Grégoire Allaire et Sidi Khaber pour les détails) :

$$2T_k(x) = (x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k \geqslant (x+\sqrt{x^2-1})^k,$$

pour $x > 1$.
On prend alors : $x = \dfrac{\kappa+1}{\kappa-1}$ et on a donc :

$$\begin{aligned}
\left|2T_k\left(\frac{\kappa+1}{\kappa-1}\right)\right|^{\frac{1}{k}} &\geqslant \frac{\kappa+1}{\kappa-1} + \sqrt{\left(\frac{\kappa+1}{\kappa-1}\right)^2 - 1} \\
&= \frac{\kappa+1+\sqrt{(\kappa+1)^2-(\kappa-1)^2}}{\kappa-1} \qquad \text{[remarquer ensuite :} (\kappa+1)^2-(\kappa-1)^2 = 4\kappa] \\
&= \frac{\kappa+1+2\sqrt{\kappa}}{\kappa-1} \\
&= \frac{(\sqrt{\kappa}+1)^2}{(\sqrt{\kappa}-1)(\sqrt{\kappa}+1)} \\
&= \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}.
\end{aligned}$$

Reprenons alors la majoration :

$$\left\|e^k\right\|_A^2 \leqslant \left\|e^0\right\|_A^2 \left(\min_{Q\in\mathbb{P}_k^0}\left[\max_{\lambda_1\leqslant x\leqslant\lambda_n}|Q(x)|\right]\right)^2 \leqslant \left\|e^0\right\|_A^2 \left(\frac{2}{\left|2T_k\left(\frac{\kappa+1}{\kappa-1}\right)\right|}\right)^2.$$

Avec le calcul ci-dessus, on obtient :

$$\left\|e^k\right\|_A^2 \leqslant \left\|e^0\right\|_A^2 \left(\frac{2}{\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k}\right)^2.$$

On aboutit finalement à :

$$\left\|e^k\right\|_A^2 \leqslant 4\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\left\|e^0\right\|_A^2$$

et puis on utilise alors l'équivalence des normes :

$$\lambda_1\|x\|^2 \leqslant \|x\|_A^2 \leqslant \lambda_n\|x\|^2$$

ce qui donne

$$\left\|e^k\right\| \leqslant \frac{1}{\sqrt{\lambda_1}}\left\|e^k\right\|_A \leqslant \frac{2}{\sqrt{\lambda_1}}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\left\|e^0\right\|_A \leqslant \frac{\sqrt{\lambda_n}}{\sqrt{\lambda_1}}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\left\|e^0\right\|,$$

ce qui est bien le résultat voulu. □

Chapitre 3

Moindres carrés

On va maintenant aborder un sujet un peu plus général, et voir comment on peut procéder lors de la résolution de systèmes linéaires mal posés. Par cela, on entend un système linéaire qui fait intervenir une matrice qui n'est pas inversible. Il peut donc ne pas y avoir de solution, ou y en avoir beaucoup trop (ensemble de solutions qui est un espace affine de dimension supérieure ou égale à 1).

On va déjà commencer par motiver ce problème en présentant quelques applications. En fait, c'est un problème omniprésent en sciences appliquées, que ce soit pour de l'analyse de données, des sciences pour l'ingénieur, etc. Ensuite, on va donner une définition précise et voir comment on peut définir des solutions pour ce type de problème, qui soient les "plus moins pires" possibles (solutions au sens des moindres carrés). Nous verrons alors un résultat important concernant les matrices : la décomposition en valeurs singulières (SVD), qui non seulement à un intérêt en soi, mais en plus qui permet de résoudre élégamment le problème général des moindres carrés. Au final, nous verrons une technique de résolution pratique, ou on utilise une décomposition QR via la méthode de Householder.

I Introduction et motivation

D'une façon très générale, on va introduire le problème suivant : trouver $x \in \mathbb{R}^n$ solution du système linéaire

$$Ax = b \qquad \text{avec } A \in \mathcal{M}_{mn}(\mathbb{R}) \ , b \in \mathbb{R}^m, \quad m, n \in \mathbb{N}.$$

Le vecteur b est un vecteur colonne à m lignes, et la matrice A possède m lignes et n colonnes. On cherche donc un vecteur colonne x à n lignes. La matrice A est supposée absolument quelconque. On ne suppose pas en particulier que la matrice est carrée et on peut avoir alors $m \neq n$. Même si on essayera de rester assez général, le cas de figure qui nous intéressera fréquemment est $m > n$ ("plus de données que de paramètres"). Bien sûr, A n'étant pas forcément inversible, le système $Ax = b$ peut ne pas avoir de solutions. Il suffit pour cela tout simplement que b ne soit pas dans l'image de A. D'un autre côté, si b appartient à l'image de A, mais si le noyau de A est de dimension supérieure ou égale à 1, il y alors "trop de solutions" : l'ensemble des solutions est un espace affine porté par une solution particulière et le noyau de A. On peut souhaiter alors récupérer une seule solution parmi celles-ci.

L'idée des moindres carrés est de "relaxer" un peu le problème $Ax = b$ afin d'obtenir toujours une solution, au moins, qui de plus, est la meilleure (ou, disons, la moins mauvaise), dans un certain sens. Plus précisément, on souhaite alors trouver x qui minimise sur $\mathbb{R}^n$ la quantité

$$\frac{1}{2} \|Ax - b\|^2$$

où $\|\cdot\|$ désigne la norme euclidienne usuelle sur $\mathbb{R}^n$. On va voir d'abord quelques exemples pour lesquels cette approche est utile. Ensuite, on va reprendre le problème sous sa forme générale, et formuler une théorie qui va répondre aux questions suivantes : déjà, est-ce qu'on peut s'assurer que le problème au sens des moindres carrés a toujours une solution ? Comment alors caractériser précisément l'ensemble des solutions "au sens des moindres carrés" ? Si ces solutions restent encore "trop nombreuses", peut-on en extraire une seule particulièrement ? Aussi, nous allons essayer de voir une méthode pratique qui permette de calculer sur ordinateur la solution d'un problème au sens des moindres carrés.

Exemple :

1) La régression "polynômiale".

Données : un nuage de points $(x_i, y_i)_{1 \leqslant i \leqslant m}$ et $x_1 < x_2 < \cdots < x_m$.

On veut un polynôme de degré $n - 1$ qui approche au mieux le nuage de points. On cherche alors le pôlynôme p sous la forme :

$$p(x) = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1}.$$

Une formulation naturelle pour que le graphe de ce polynôme p approche au mieux le nuage de point, consiste à trouver les coefficients qui minimisent la quantité :

$$\frac{1}{2}\sum_{i=1}^{m}\left(y_i - p(x_i)\right)^2.$$

Ici, typiquement, on peut supposer qu'on a beaucoup de points (par exemple car on a beaucoup de mesures expérimentales) et qu'ils se superposent à une courbe simple, à ceci près qu'ils sont entâchés de "bruit" (au sens statistique). Ceci correspond alors au cas $m > n$.

Remarque. Lors d'une régression polynômiale, on a deux cas particuliers :

1) Si $n = 2$ alors c'est une régression linéaire (sous R : $lm(y \sim x)$).
2) Si $n = m$ alors on fait de l'interpolation polynômiale.

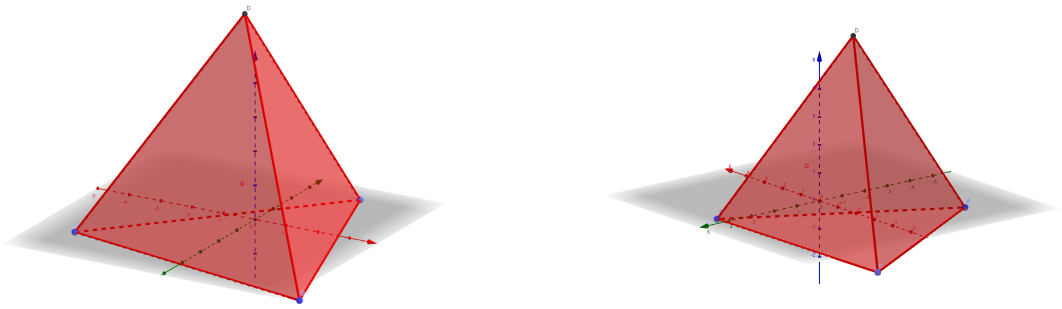
Exemple :

2) Problème de recalage dit "problème de Procruste" [1].

Soit Ω un solide de $\mathbb{R}^3$, qui subit une transformation rigide.
On appelle Ω_0 le domaine occupé par le solide avant transformation puis Ω_1 le solide le domaine occupé après transformation. On suppose que les paramètres de la transformation rigide sont inconnus, mais que, dans un repère donné, on est capable de mesurer précisément la position de différents points en correspondance sur les domaines avant et après transformation. On se donne alors :
Une famille $\vec{p_i} = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}$ de m points qui sont sur la surface de Ω_0, et aussi une famille $\vec{\hat{p}_i} = \begin{pmatrix} \hat{x}_i \\ \hat{y}_i \\ \hat{z}_i \end{pmatrix}$ de m points sur Ω_1, en correspondance.
On veut alors trouver une transformation rigide (R, t), avec :

— R la matrice de rotation,
— t le vecteur de translation,

1. Terminologie issue du mythe grec : voir https://mythologiegrecque.fandom.com/fr/wiki/Procruste ou encore http://remacle.org/bloodwolf/historiens/diodore/livre4b.htm.



(a) *Solide* Ω_0 (b) *Solide* Ω_1

FIGURE 3.1 – *Représentation graphique du solide qui occupe successivement les positions* Ω_0 *et* Ω_1.

qui minimisent la quantité :

$$\frac{1}{2}\sum_{i=1}^{m}\left\|\vec{\hat{p}}_i - (R\vec{p}_i + t)\right\|^2.$$

A nouveau dans cet exemple, pour déterminer sans ambiguïté et avec précision la transformation rigide, on va prendre $m \gg n$.

Exemple :

3) Problèmes inverses.

On prend un système dont le comportement dépend d'un petit nombre de paramètres $p \in \mathbb{R}^n$ inconnus, par exemple un matériau dont le comportement mécanique est donné par deux paramètres (le module d'Young E, qui traduit sa rigidité, et le coefficient de Poisson ν qui traduit sa compressibilité). On effectue m observations (mesures), par exemple en sollicitant le matériau et en mesurant des déformations, des contraintes, etc. On obtient ainsi $y \in \mathbb{R}^m$ le vecteur des observations.
On a la relation : $Ap = y$ et on veut trouver p à partir de y. Ici A est une matrice qui traduit le comportement "physique" du système, sous l'hypothèse que les observations dépendent linéairement des paramètres.

II Principe général des moindres carrés

On revient au cadre général :
Soit $A \in \mathcal{M}_{mn}(\mathbb{R})$ et soit $b \in \mathbb{R}^m$.
On veut trouver $x \in \mathbb{R}^n$ qui résolve :

$$Ax \simeq b. \tag{3.1}$$

Ici, on dira que $Ax = b$ est (possiblement) *mal posé* : il peut ne pas y avoir de solution x ou il peut y en avoir plusieurs. Il faut alors définir différemment une solution, ce qui motive la définition suivante :

Définition II.1. On appelle formulation par moindres carrés de (3.1) le problème de minimisation :
Trouver $x \in \mathbb{R}^n$ qui minimise :

$$\mathcal{J}_{\text{MC}} : z \to \frac{1}{2} \|Az - b\|^2 \ , \quad z \in \mathbb{R}^n,$$

avec $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^n$.
On dit alors que x est solution de (3.1) au sens des moindres carrés.

On va commencer déjà par montrer que ce problème admet au moins une solution, et caractériser l'ensemble des solutions.

Théorème II.1. On a l'équivalence entre les trois points suivants :

i) $x \in \mathbb{R}^n$ minimise $\mathcal{J}_{\text{MC}}$,

ii) $\langle Ax - b, Ay \rangle = 0 \ \ \forall y \in \mathbb{R}^n$,

iii) x est solution de l'*équation normale*

$$A^t Ax = A^t b.$$

De plus, $\mathcal{J}_{\text{MC}}$ admet au moins un minimiseur sur $\mathbb{R}^n$.

Démonstration.

i) $\Rightarrow$ ii) Supposons que x minimise $\mathcal{J}_{\text{MC}}$. Alors, pour tous $t \in \mathbb{R}$, $y \in \mathbb{R}^n$ (avec $z = x + ty$) :

$$\|Az - b\|^2 \geqslant \|Ax - b\|^2 .$$

Et :

$$\begin{aligned}\|Az - b\|^2 &= \langle Az - b, Az - b\rangle \\ &= \langle A(x + ty) - b, A(x + ty) - b\rangle \\ &= \langle (Ax - b) + tAy, (Ax - b) + tAy\rangle \\ &= \|Ax - b\|^2 + 2t\langle Ax - b, Ay\rangle + t^2 \|Ay\|^2 .\end{aligned}$$

On a donc :

$$t^2 \|Ay\|^2 + 2t\langle Ax - b, Ay\rangle \geqslant 0.$$

Alors, pour $t > 0$:

$$t \|Ay\|^2 + 2\langle Ax - b, Ay\rangle \geqslant 0.$$

Et donc, en faisant tendre $t \to 0$: $\langle Ax - b, Ay\rangle \geqslant 0$.
Ensuite, pour $t < 0$:

$$t \|Ay\|^2 + 2\langle Ax - b, Ay\rangle \leqslant 0.$$

Et donc, en faisant tendre $t \to 0$: $\langle Ax - b, Ay\rangle \leqslant 0$.
Finalement :

$$\langle Ax - b, Ay\rangle = 0 \ \forall y \in \mathbb{R}^n.$$

ii) $\Rightarrow$ iii) On a de plus :

$$\langle A^t(Ax - b), y\rangle = 0 \quad \forall y \in \mathbb{R}^n.$$

Donc

$$A^t(Ax - b) = 0$$

et donc :

$$A^t Ax = A^t b.$$

ii) $\Rightarrow$ i) Si x vérifie ii) on a alors en posant $y = z - x$ ($t = 1$), pour tout $z \in \mathbb{R}^n$:

$$\begin{aligned}\|Az - b\|^2 &= \|Ax - b\|^2 + \underbrace{2\langle Ax - b, Ay\rangle}_{=0} + \underbrace{\|Ay\|^2}_{\geqslant 0} \\ &\geqslant \|Ax - b\|^2 \quad \text{donc } x \text{ minimise } \mathcal{J}_{\text{MC}}.\end{aligned}$$

iii) $\Rightarrow$ ii)

$$\langle A^tAx - A^tb, y\rangle = 0 \quad \forall y \in \mathbb{R}^n$$
$$\langle Ax - b, Ay\rangle = 0.$$

Les équivalences sont donc prouvées. Reste maintenant à montrer que $\mathcal{J}_{\text{MC}}$ admet au moins un minimiseur sur $\mathbb{R}^n$. On utilise le théorème de projection sur un sous-espace vectoriel fermé, qui assure l'existence et l'unicité d'un vecteur $\tilde{b} \in \text{Im}\, A$ solution de :

$$\langle \tilde{b} - b, v\rangle = 0 \quad \forall v \in \text{Im}\, A.$$

On prend ensuite x tel que

$$Ax = \tilde{b}$$

(comme $\tilde{b} \in \text{Im}\, A$, il existe forcément un x solution). Soit maintenant $z \in \mathbb{R}^n$. Posons $v = Az \in \text{Im}\, A$. On vérifie alors :

$$\langle Ax - b, Az\rangle = 0.$$

En conséquence x vérifie la condition ii), et est donc un minimiseur de $\mathcal{J}_{\text{MC}}$. □

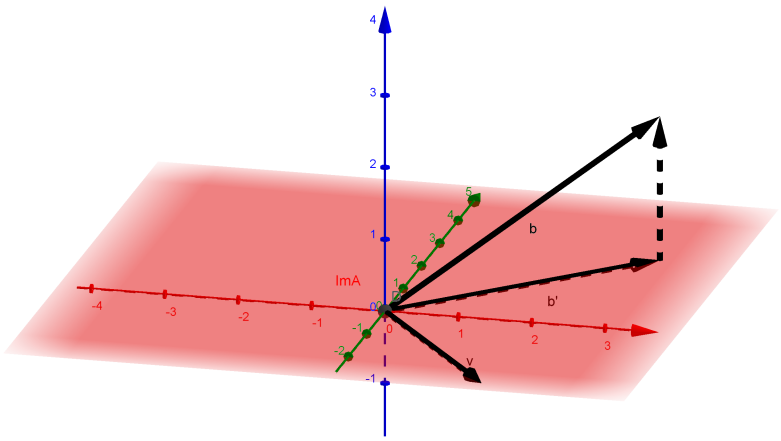


FIGURE 3.2 – *Représentation graphique de la situation.*

Remarque. On a l'existence d'un seul vecteur $\tilde{b} \in \operatorname{Im} A$ tel que :

$$\langle \tilde{b} - b, Az \rangle = 0 \quad , \ \forall z \in \mathbb{R}^n.$$

Mais si $\ker A \neq \{0\}$, on peut avoir plusieurs x tels que $Ax = \tilde{b}$.

Remarque. Pour résoudre le problème des moindres carrés, on peut se ramener à la résolution du système d'équations normales :

$$A^t A x = A^t b.$$

C'est un système qui fait intervenir une matrice carrée, de taille $n \times n$.

Exercice :

On va voir le cas particulier de la régression linéaire.

Soient $(t_i, y_i)_{1 \leqslant i \leqslant m}$ m points et $p(t) = \alpha_0 + \alpha_1 t$ avec $\alpha_0,\ \alpha_1$ à déterminer. On cherche $(\alpha_0, \alpha_1) \in \mathbb{R}^2$ tel que :

$$\frac{1}{2} \sum_{i=1}^{m} (y_i - (\alpha_0 + \alpha_1 t_i))^2 \quad \text{soit minimale.}$$

1) Montrer qu'on peut reformuler ce problème sous la forme :

$$\mathcal{J}_{\mathrm{MC}}(x) = \frac{1}{2} \|Ax - b\|^2 \qquad \text{avec} \quad \left(x = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} \right).$$

2) Expliciter et résoudre le système normal correspondant :

$$A^t A x = A^t b.$$

3) Décrire une situation où $\ker A \neq \{0\}$.

III SVD pour les moindres carrés

La SVD, ou décomposition en valeurs singulières, est un outil très puissant d'analyse matricielle. C'est une transformation qui, dans un certain sens, généralise la diagonalisation. Elle permet en quelque sorte de 'passer la matrice aux rayons X'. Il y a de

très nombreuses applications de la SVD : on peut l'appliquer directement pour faire de la compression d'image, ou elle sert aussi en statistique pour faire de l'analyse en composantes principales (ACP ou PCA). On va d'abord voir le résultat principal, puis ensuite quelques conséquences directes. Un résultat important concerne la définition de la pseudo-inverse d'une matrice, qui généralise la notion d'inverse pour une matrice quelconque. La pseudo-inverse se définit directement à partir de la SVD. On verra alors finalement que la SVD et la pseudo-inverse permettent de caractériser, et calculer, directement les solutions d'un problème aux moindres carrés.

III. 1) La SVD

Donnons d'emblée le résultat principal :

Théorème III.1. (SVD)
Soit $A \in \mathcal{M}_{mn}(\mathbb{R})$ avec $m \geqslant n$.
Il existe $U \in \mathcal{M}_m(\mathbb{R})$ et $V \in \mathcal{M}_n(\mathbb{R})$ orthogonales et il existe

$$\Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ 0 & \ddots & 0 \\ 0 & & \sigma_n \\ & 0 & \end{pmatrix} \in \mathcal{M}_{mn}(\mathbb{R})$$

avec $\sigma_1 \geqslant \cdots \geqslant \sigma_n \geqslant 0$ et telles que :

$$A = U\Sigma V^t.$$

On appelle alors :

- $U = [u_1 \dots u_m]$: vecteurs singuliers gauches,
- $V = [v_1 \dots v_n]$: vecteurs singuliers droits,
- σ_i : valeurs singulières de A.

L'hypothèse $m \geqslant n$ est surtout effectuée ici pour simplifier la présentation et la preuve, mais on peut sans problème s'en affranchir. On va voir une preuve constructive, mais qui n'est pas utilisée en pratique pour calculer la SVD (la description des algorithmes les plus aboutits dépasse très largement le cadre de ce cours).

Démonstration. Comme $\|A\| = \max_{x \in \mathbb{R}^n,\ \|x\|=1} \|Ax\|$, il existe $x \in \mathbb{R}^n$ tel que $\|x\| = 1$ et $\|Ax\| = \|A\|$.

On pose alors $\sigma = \|A\|$ et $z = Ax \in \mathbb{R}^n$. Soit $y = \dfrac{z}{\|z\|}$ alors :

$$Ax = z = \underbrace{\|z\|}_{=\|Ax\|} y = \sigma y.$$

Posons ensuite $V = [x\ v_1]$ et $U = [y\ u_1]$ où on choisit v_1 (respectivement u_1) de sorte que $(x\ v_1)$ (respectivement $(y\ u_1)$) soit une famille de vecteurs orthonormés. On a donc :

$$v_1{}^t x = 0, \qquad u_1{}^t y = 0.$$

Calculons :

$$\begin{aligned} A_1 = U^t A V &= \begin{pmatrix} y^t \\ u_1{}^t \end{pmatrix} A \begin{pmatrix} x & v_1 \end{pmatrix} \\ &= \begin{pmatrix} y^t \\ u_1{}^t \end{pmatrix} \begin{pmatrix} Ax & Av_1 \end{pmatrix} = \begin{pmatrix} y^t Ax & y^t Av_1 \\ u_1{}^t Ax & u_1{}^t Av_1 \end{pmatrix} \\ &= \begin{pmatrix} \sigma & \omega^t \\ 0 & B \end{pmatrix}. \end{aligned}$$

On a utilisé

$$y^t Ax = y^t \sigma y = \sigma, \qquad u_1{}^t Ax = \sigma(u_1{}^t y) = 0,$$

et on a défini $\omega := (Av_1)^t y$ et $B := u_1{}^t A v_1$. Reste à montrer que : $\omega^t = 0$. Pour cela, calculons

$$A_1 \begin{pmatrix} \sigma \\ \omega \end{pmatrix} = \begin{pmatrix} \sigma^2 + \|\omega\|^2 \\ B\omega \end{pmatrix}.$$

Alors :

$$\begin{aligned} \left\| A_1 \begin{pmatrix} \sigma \\ \omega \end{pmatrix} \right\|^2 &= (\sigma^2 + \|\omega\|^2)^2 + \|B\omega\|^2 \\ &\geqslant (\sigma^2 + \|\omega\|^2)^2. \end{aligned}$$

Comme U et V sont orthonormées :

$$\|A_1\| = \left\| U^t A V \right\| = \|A\| = \sigma.$$

Et donc :

$$\sigma^2 = \|A_1\|^2 = \max_{x \in \mathbb{R}^n} \frac{\|A_1 x\|^2}{\|x\|^2}.$$

En particulier, pour $x = \begin{pmatrix} \sigma \\ \omega \end{pmatrix}$:

$$\sigma^2 \geqslant \frac{\left\| A_1 \begin{pmatrix} \sigma \\ \omega \end{pmatrix} \right\|^2}{\left\| \begin{pmatrix} \sigma \\ \omega \end{pmatrix} \right\|^2}$$
$$\sigma^2 \geqslant \frac{(\sigma^2 + \|\omega\|^2)^2}{(\sigma^2 + \|\omega\|^2)}$$
$$\sigma^2 \geqslant \sigma^2 + \|\omega\|^2$$
$$\|\omega\|^2 = 0 \ : \ \omega = 0.$$

Finalement on obtient :

$$U^t A V = \begin{pmatrix} \sigma & 0 \\ 0 & B \end{pmatrix}.$$

On termine alors la démonstration en reproduisant $(n-1)$ fois le même procédé. □

Exercice :

Calculer la SVD pour la matrice suivante :

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

On pourra éventuellement démontrer les deux résultats ci-dessous à titre d'exercice.

Remarque. Soit r le rang de la matrice A. On a $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$.

Proposition III.2. Soit U, Σ, V la SVD de $A \in \mathcal{M}_{mn}(\mathbb{R})$.
Alors : $V_1, \ldots V_n$ sont les vecteurs propres de $A^t A$ et $\sigma_1^2, \ldots, \sigma_n^2$ sont les valeurs propres associées.

III. 2) La pseudo-inverse

Voyons maintenant une notion qui va nous servir pour caractériser la solution du problème des moindres carrés :

Définition III.1. Soit $A \in \mathcal{M}_{mn}(\mathbb{R})$ et soit $A = U\Sigma V^t$ sa SVD.
On écrit :

$$\Sigma = \begin{bmatrix} \Sigma_n \\ 0 \end{bmatrix} \text{ où } \Sigma_n = \text{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0) \in \mathcal{M}_n(\mathbb{R}).$$

On a $\sigma_1 \geq \ldots \geq \sigma_r > 0$. On peut donc définir :

$$\Sigma_n^+ = \text{diag}((\sigma_1)^{-1}, \ldots, (\sigma_r)^{-1}, 0, \ldots, 0)$$

Soit maintenant :

$$\Sigma^+ = \begin{bmatrix} \Sigma_n^+ & 0 \end{bmatrix} \in \mathcal{M}_{nm}(\mathbb{R}).$$

La pseudo-inverse de A est alors :

$$A^+ = V\Sigma^+ U^t.$$

On va voir maintenant que la pseudo-inverse est utile pour définir les projections orthogonales sur les espaces fondamentaux associés à la matrice A. Rappelons avant cela un résultat utile d'algèbre linéaire :

Proposition III.3. (*Relations entre espaces fondamentaux*)
Soit $A \in \mathcal{M}_{mn}(\mathbb{R})$, on a alors :

$$\begin{aligned} (\text{Im}(A))^\perp &= \ker(A^t), \\ \ker(A) &= (\text{Im}(A^t))^\perp. \end{aligned}$$

Avec ce résultat, on va voir maintenant que la pseudo-inverse permet de construire très simplement les projecteurs orthogonaux sur les noyaux et images de A (et de sa transposée).

Théorème III.4. (*Projection sur les sous-espaces fondamentaux*)
Soit $A \in \mathcal{M}_{mn}(\mathbb{R})$ et soit $A^+ \in \mathcal{M}_{nm}(\mathbb{R})$ sa pseudo-inverse. Alors, la matrice associée à la projection orthogonale sur $\mathrm{Im}(A)$ est :

$$P_{\mathrm{Im}(A)} = AA^+.$$

On a, de plus :

$$P_{\mathrm{Im}(A^t)} = A^+A, \quad P_{\ker(A)} = I - A^+A, \quad P_{\ker(A^t)} = I - AA^+.$$

Démonstration. On va montrer seulement la première relation, les autres s'obtiennent de la même façon.
Commençons par calculer, en utilisant d'abord la définition de la pseudo-inverse et la SVD de A :

$$\begin{aligned}(AA^+)^t &= (A^+)^t A^t = (V\Sigma^+U^t)^t(U\Sigma V^t)^t \\ &= U(\Sigma^+)^t \underbrace{V^tV}_{=I}\Sigma^tU^t \\ &= U(\Sigma^+)^t\Sigma^tU^t \\ &= UI_rU^t,\end{aligned}$$

avec :

$$I_r := \mathrm{diag}(\underbrace{1,\dots,1}_{r},\underbrace{0,\dots,0}_{m-r}) = \Sigma\Sigma^+ = {\Sigma^+}^t\Sigma^t \in \mathcal{M}_m(\mathbb{R}).$$

D'un autre côté, en procédant à l'identique :

$$\begin{aligned}AA^+ &= U\Sigma V^tV\Sigma^+U^t \\ &= UI_rU^t.\end{aligned}$$

On en déduit que AA^+ est symétrique.
En utilisant à nouveau les propriétés d'orthogonalité de U et V, puis la relation $I_r\Sigma = \Sigma$, calculons maintenant :

$$\begin{aligned}AA^+A &= (U\Sigma V^t)(V\Sigma^+U^t)(U\Sigma V^t) \\ &= U\Sigma\Sigma^+\Sigma V^t \\ &= UI_r\Sigma V^t \\ &= U\Sigma V^t \\ &= A.\end{aligned}$$

Ensuite on voit que

$$AA^+AA^+ = (AA^+A)A^+ = AA^+,$$

ce qui prouve que AA^+ est idempotent : c'est donc une projection. De plus, AA^+ est symétrique donc c'est une projection orthogonale.
Soit maintenant $y \in \operatorname{Im} A$. Alors :

$$\exists x \in \mathbb{R}^n \ : \ y = Ax.$$

Donc :

$$AA^+y = AA^+(Ax) = Ax = y.$$

Soit maintenant $z \in \mathbb{R}^m$ avec $z \perp \operatorname{Im}(A)$ $(A^tz = 0)$.
Alors :

$$AA^+z = (AA^+)^tz = A^{+t}A^tz = 0.$$

Donc on a bien vérifié que :

$$AA^+ = P_{\operatorname{Im}(A)}.$$

□

III. 3) Solution des moindres carrés

On peut maintenant caractériser toutes les solutions du problème des moindres carrés. C'est l'objet du résultat suivant, qui est le clou du spectacle pour ce chapitre.

Théorème III.5. Soit $A \in \mathcal{M}_{mn}(\mathbb{R})$ avec A^+ sa pseudo-inverse et soit $b \in \mathbb{R}^m$. Alors

$$x_b = A^+b$$

minimise sur $\mathbb{R}^n$ la quantité $\|Ax - b\|^2$, et est donc une solution du problème des moindre carrés. On peut donc caractériser et écrire toutes les solutions comme suit :

$$x_b + x, \ x \in \ker A.$$

De plus, x_b est la solution de norme minimale, c'est à dire :

$$\text{Si } \|Ax - b\|^2 = \|Ax_b - b\|^2 \text{ avec } x_b \neq x \text{ alors } \|x_b\| < \|x\|.$$

Démonstration. Soit $x \in \mathbb{R}^n$. Écrivons :

$$\begin{aligned} Ax - b &= Ax - Ax_b + Ax_b - b \\ &= A(x - x_b) - (I - AA^+)b \quad (x_b = A^+b). \end{aligned}$$

La décomposition est orthogonale car $A(x - x_b) \in \mathrm{Im}(A)$ et :

$$\begin{aligned} (I - AA^+)b &= P_{\ker(A^t)}b \\ &= P_{(\mathrm{Im}\, A)^\perp}b \in (\mathrm{Im}\, A)^\perp. \end{aligned}$$

Avec Pythagore on obtient donc :

$$\|Ax - b\|^2 = \|A(x - x_b)\|^2 + \left\|(I - AA^+)b\right\|^2.$$

Comme :

$$(I - AA^+)b = b - AA^+b = b - Ax_b,$$

on obtient finalement :

$$\|b - Ax_b\|^2 \leqslant \|b - Ax\|^2.$$

Et donc x_b est une solution du problème des moindres carrés.
Soit maintenant $x \in \mathbb{R}^n, x \neq x_b$ et qui vérifie :

$$\|b - Ax\|^2 = \|b - Ax_b\|^2.$$

Décomposons :

$$x = x_b + (x - x_b).$$

On a : $x - x_b \in \ker A$ car $A(x - x_b) = P_{\mathrm{Im}\, A}b - P_{\mathrm{Im}\, A}b = 0$ (on se souvient que pour x une solution du problème des moindres carrés, on a Ax qui est la projection orthogonale de b sur l'image de A, voir le Théorème II.1). De plus : $x_b = A^+b \in (\ker A)^\perp$.
En effet, soit $z \in \ker A$, $Az = 0$ (donc $U\Sigma V^t z = 0$). Alors :

$$\begin{aligned} \langle A^+b, z\rangle &= \langle V\Sigma^+U^tb, z\rangle \\ &= \langle \Sigma^+U^tb, V^tz\rangle \\ &= \langle U^tb, (\Sigma^+)^tV^tz\rangle \\ &= \langle b, U(\Sigma^+)^tV^tz\rangle = 0. \end{aligned}$$

Appliquons à nouveau Pythagore :

$$\|x\|^2 = \|x_b\|^2 + \underbrace{\|x - x_b\|^2}_{>0}.$$

Finalement :

$$\|x_b\| < \|x\|.$$

□

IV Résolution pratique du problème des moindres carrés

Il y a plusieurs possibilités pour résoudre en pratique le problème des moindres carrés. Voyons-en déjà une, qui repose sur ce que nous venons de voir dans la section précédente.

IV. 1) Algorithme par SVD

Si on dispose d'une façon simple et rapide d'obtenir la décomposition en valeurs singulières de A, on peut alors résoudre les moindres carrés grâce à la pseudo-inverse de A. Ceci est en quelque sorte un sous-produit du Théorème III.5. Cela donne l'algorithme 2 page 49.

Algorithm 2 Résolution des moindres carrés par SVD

Require: $A \in \mathcal{M}_{mn}(\mathbb{R}),\ b \in \mathbb{R}^m$.
SVD : $A = U\Sigma V^t$.
Trouver $r \in \mathbb{N}$ tel que $\sigma_r > 0$ et $\sigma_j = 0,\ \forall j > r$.
Décomposer

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} {V_1}^t \\ {V_2}^t \end{bmatrix}$$

avec $\Sigma_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$.
Calculer $x_b = V_1 \Sigma_r^{-1} {U_1}^t b$.
return $x_b \in \mathbb{R}^n$.

On voit qu'ici on a renvoyé une version "réduite" de la pseudo-inverse, où on a récupéré uniquement les composantes non-nulles, ce qui fait gagner de l'espace mémoire et du temps de calcul. Une difficulté réside quand même à trouver l'indice r à partir duquel les valeurs singulières seront nulles. En effet, comme l'ordinateur ne travaille pas en arithmétique exacte, les valeurs singulières nulles seront approchées par une valeur très petite. Il faut alors fixer, sans se tromper, un seuil qui va permettre de les différencier des autres.

En bonus, on voit que toutes les solutions sont données par :

$$x = x_b + V_2 c \quad \forall c \in \mathbb{R}^{n-r},$$

car V_2 permet d'obtenir les éléments du noyau de A.

IV. 2) Solution des équations normales

Comme on n'a pas toujours un algorithme de SVD sous la main (et de fait, la SVD est difficile à programmer), il est intéressant d'avoir des techniques alternatives. Pour certains problèmes aussi, le calcul de la SVD peut être trop coûteux en ressources de calcul, et il n'est pas nécessaire pour obtenir une solution (on obtient en fait beaucoup plus d'informations que le strict nécessaire avec la SVD). Un autre procédé pour résoudre le problème des moindres carrés consiste alors à résoudre le système d'équations normales obtenu au Théorème II.1 :

$$A^t Ax = A^t b.$$

Pour cela, on peut faire une décomposition de Cholesky de la matrice $A^t A$, et on obtient $A^t A = R^t R$ avec R triangulaire supérieure. Il est ensuite très facile de résoudre les équations.

On a vu que la matrice $A^t A$ est de taille $n \times n$, et ce procédé est donc pertinent si n est petit. Dans l'exemple de la régression linéaire vu précédemment, on pouvait faire de la sorte, et, comme on avait $n = 2$, on pouvait même trouver une solution analytique. Pour n qui devient grand, cette technique peut poser problème. Nous allons voir pourquoi.

Précisons d'abord quelques notions sur le conditionnement matriciel. Pour la matrice A, qui n'est plus forcément une matrice carrée, on étend la notion de conditionnement comme suit :

$$\kappa(A) = \|A\|\|A^+\|,$$

où $\|\cdot\|$ est la norme matricielle induite par les normes euclidiennes sur $\mathbb{R}^m$ et sur $\mathbb{R}^n$, et où A^+ désigne la pseudo-inverse de A.

Exercice :

Montrer que

$$\kappa(A) = \frac{\sigma_1(A)}{\sigma_r(A)},$$

avec $\sigma_1(A)$ la plus grande valeur singulière de A, et $\sigma_r(A)$ la plus petite valeur singulière (strictement positive).

Exercice :

Montrer que si A est une matrice carrée inversible, on retrouve la notion de conditionnement déjà vue pour les matrices carrées.

On va montrer ici que la résolution du problème de moindres carrés via solution directe des équations normales n'est pas idéale en termes de conditionnement.

Exercice : Supposons que le rang de A est n. En utilisant la décomposition en valeurs singulières (SVD) de A, montrer que

$$\kappa(A^tA) = \kappa(A)^2.$$

Le résultat ci-dessus peut s'interpréter comme suit : la résolution via les équations normales est bien moins conditionnée que pour un système linéaire "classique" (où le conditionnement est en $\kappa(A)$). En conséquence, cette méthode sera beaucoup moins robuste aux erreurs d'arrondi. Voyons cela sur l'exemple suivant (le "fameux exemple de P. Läuchli") :

Exercice : Soit $\delta > 0$ et la matrice

$$A = \begin{bmatrix} 1 & 1 \\ \delta & 0 \\ 0 & \delta \end{bmatrix}.$$

Quel est le rang de A ? Quelle est l'expression de la matrice A^tA ? Supposons que la précision avec laquelle on peut représenter les réels en machine est ε, que devient A^tA si $\delta < \sqrt{\varepsilon}$? □

Voyons donc maintenant une méthode efficace pour résoudre les moindres carrées, et qui est moins désastreuse en termes de conditionnement.

IV. 3) Résolution par décomposition QR

La décomposition QR est la méthode la plus utilisée en pratique pour résoudre des problèmes de moindres carrés. Elle a en outre un intérêt en soi et elle sert pour d'autres problèmes (problèmes aux valeurs propres par exemple). Ici nous expliquons déjà pourquoi cette méthode est intéressante pour les moindres carrés, et ensuite nous en donnons une implémentation efficace, qui est la méthode de Householder. C'est cette implémentation qui est effectuée dans la plupart des bons codes de calcul.

Il existe une autre méthode, plus élémentaire, basée sur Gramm-Schmidt, que certains d'entre vous ont peut-être vu en 3e année de Licence, mais qui est moins robuste et performante. Elle est décrite dans le chapitre 7 du livre de Grégoire Allaire et de Sidi M. Kaber. On supposera dans cette section que le rang de A est n.

IV. 3) a) Principe

La décomposition QR de la matrice A est donnée par

$$A = QR,$$

avec Q une matrice carrée à m lignes et m colonnes, qui est orthogonale ($Q^tQ = I$), et R une matrice à m lignes et n colonnes qui est triangulaire supérieure (ses coefficients (r_{ij}) sont nuls dès que $i > j$).

L'utilisation de la décomposition QR repose sur l'observation suivante : la solution du x du problème de moindres carrés reste inchangée si on multiplie A et b par une même matrice orthogonale O.

Exercice : Démontrer l'observation ci-dessus.

Choisissons maintenant $O = Q^t$, et il vient :

$$Q^t(Ax - b) = Q^tAx - Q^tb = Q^t(QR)x - Q^tb = Rx - Q^tb$$

où on a utilisé la décomposition QR de la matrice A, puis la propriété d'orthogonalité de la matrice Q.

Nous allons noter R_n la sous-matrice de R de taille n qui contient les n premières lignes de R, $y = Q^tb$, ainsi que y_n le vecteur colonne à n lignes, qui contient les n premières lignes de y. La solution du problème des moindres carrés est alors obtenue par simple résolution du système :

$$R_nx = y_n$$

avec une solution qui peut être obtenue par algorithme de remontée, compte-tenu du fait que R_n est triangulaire supérieure.

Exercice : Justifier que la solution du système $R_nx = y_n$ donne effectivement la solution du problème de moindres carrés.

L'autre avantage, ici, est que le conditionnement de la matrice R est égal à celui de la matrice A, donc à ce niveau, la méthode est indubitablement meilleure que celle qui consiste à passer par le système d'équations normales.

Exercice : Montrer que $\kappa(R) = \kappa(A)$.

On voit ici que tout repose finalement sur la décomposition QR. Nous avons supposé qu'il était possible d'écrire cette décomposition : il nous faut maintenant le démontrer. On va le faire en utilisant la méthode de Householder, qui est en plus une technique efficace utilisée en pratique. Avant de donner la méthode, nous allons déjà étudier des matrices aux propriétés particulières, qui sont à l'origine de sa construction.

IV. 3) b) Les matrices de Householder

Tout d'abord, la matrice de Householder associée au vecteur nul est la matrice identité de $\mathbb{R}^m$:

$$H(0) = I_m.$$

Ensuite, pour chaque vecteur $v \in \mathbb{R}^m$ non nul, on définit la matrice de Householder $H(v)$ comme suit

$$H(v) = I_m - \frac{2vv^t}{\|v\|^2}.$$

Dans l'exercice suivant, on va établir quelques propriétés intéressantes des matrices de Householder, qui vont être utiles pour la suite :

Exercice : Soit $H(v)$ une matrice de Householder. Montrer que :

1. $H(v)$ est symétrique.
2. $H(v)^2 = I_m$.
3. $H(v)$ est orthogonale.
4. Pour tout vecteur unitaire e :

$$H(v + \|v\|e)v = -\|v\|e, \qquad H(v - \|v\|e)v = +\|v\|e.$$

5. Montrer que $H(v)$ est une symétrie orthogonale par rapport à l'hyperplan orthogonal à v.

IV. 3) c) La méthode de Householder

L'idée de base de l'algorithme est d'utiliser les matrices de Householder vues précédemment, et de construire une suite de matrices, qui, multipliées avec A, permettra d'aboutir progressivement à une matrice triangulaire supérieure.

On introduit donc, pour $k = 1, \ldots, n$, une suite de matrices de Householder H^k, qui sont des matrices carrées de taille m, et de matrices A^{k+1}, à m lignes et n colonnes, qui vérifient :

$$A^1 = A, \quad A^{k+1} = H^k A^k, \quad A^{n+1} = R,$$

avec R qui est triangulaire supérieure.

On choisit les matrices de Householder H^k de telle sorte à ce que, sur les $(k-1)$ premières colonnes, il n'y ait que des zéros sous la diagonale. L'algorithme détaillé est alors décrit ci-après.

Initialisation et première itération : On définit $A^1 = A$, et on note a^1 le premier vecteur colonne de A. Si celui-ci a toutes ses composantes nulles, sauf la première, il suffit de prendre $H^1 = I_m = H(0)$. Sinon on définit :

$$H^1 = H(a^1 + \|a^1\| e_1),$$

avec e_1 le premier vecteur de la base canonique, et puis $A^2 = H^1 A^1$.

Exercice : Montrer qu'avec ce choix de H^1, on a bien la première colonne de A^2 qui a toutes ses composantes nulles, à l'exception, éventuellement, de la première.

Etape k : On suppose qu'on a réussi à construire A^k de telle sorte que, sur les $(k-1)$ premières colonnes, tous les termes en dessous de la diagonale sont nuls. On définit cette fois-ci le vecteur a^k en prenant seulement les dernières composantes de la k-ème colonne de A^k : on prend a^k comme un vecteur de taille $(m+1-k)$ qui contient les $(m+1-k)$ derniers termes de la k-ème colonne.

Si a^k a tous ses coefficients nuls, excepté (éventuellement) le premier, il n'y a rien à faire, et on prend $H^k = I_m = H(0)$, sinon on définit H^k comme suit :

$$H^k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & H(a^k + \|a^k\| e_1) \end{bmatrix}.$$

Exercice : Montrer qu'avec ce choix de H^k, on a bien la matrice $A^{k+1} = H^k A^k$ qui a, sur les k premières colonnes, tous les termes en dessous de la diagonale qui sont nuls.

Etape n, et fin : Avec le procédé décrit auparavant, on a obtenu une matrice A^{n+1}, avec sur les $(n+1)-1 = n$ premières colonnes, tous les termes qui sont nuls en

dessous de la diagonale : A^{n+1} est donc une matrice triangulaire supérieure. De plus, on a

$$A^{n+1} = H^n A^n = \ldots = H^n \ldots H^1 A^1 = H^n \ldots H^1 A.$$

On pose alors $A^{n+1} = R$, triangulaire supérieure et $Q = H_1 \ldots H_n$, orthogonale. On a obtenu ainsi la décomposition QR de A.

$\star$

Bibliographie

[1] G. Allaire and S. M. Kaber, *Numerical linear algebra*, vol. 55 of Texts in Applied Mathematics, Springer, New York, 2008.

[2] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical optimization*, Universitext, Springer-Verlag, Berlin, second ed., 2006.

[3] P. G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, Paris, 1982.

[4] W. Gander, M. J. Gander, and F. Kwok, *Scientific computing*, vol. 11 of Texts in Computational Science and Engineering, Springer, Cham, 2014.

[5] G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, fourth ed., 2013.

[6] K. Lange, *Numerical analysis for statisticians*, Statistics and Computing, Springer, New York, second ed., 2010.

[7] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, second ed., 2006.

[8] A. Quarteroni, F. Saleri, and P. Gervasio, *Scientific computing with MATLAB and Octave*, vol. 2 of Texts in Computational Science and Engineering, Springer, Heidelberg, 2014. Fourth edition.

[9] M. Schatzman, *Numerical analysis : a mathematical introduction*, Oxford University Press, 2002.

[10] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, vol. 12 of Texts in Applied Mathematics, Springer-Verlag, New York, third ed., 2002. Translated from the German by R. Bartels, W. Gautschi and C. Witzgall.