Chapitre 2

Systèmes linéaires

2.1 Introduction

Résoudre un système linéaire est une opération très courante :

— Un grand nombre de problèmes d'ingénierie conduisent, après modélisation puis éventuellement une procédure de discrétisation, à résoudre des systèmes linéaires : calculs sur des réseaux électriques ou hydrauliques en régime stationnaire, calculs de structures, etc.

— Cette opération intervient aussi comme "sous-méthode" dans de nombreux algorithmes (problèmes d'interpolation, problèmes de moindres carrés, résolution de systèmes non-linéaires, optimisation, automatique, traitement du signal, etc.).

On cherche donc à résoudre le problème suivant : "étant donnés une matrice A et un vecteur b, trouver le vecteur x solution de" :

$$Ax = b \quad \text{avec} \quad \left\{ \begin{array}{l} A \in \mathcal{M}_{nn}(\mathbb{R}) = \mathbb{R}^{n\times n} \\ x \in \mathbb{R}^n \\ b \in \mathbb{R}^n \end{array} \right. \tag{2.1}$$

Remarque : ici les nombres utilisés sont des nombres réels mais ce que nous allons voir dans ce cours est aussi valable pour des nombres complexes.

Notations :

— Si A est une matrice (n, m), A_i désigne la matrice $(1, m)$ (appelée aussi vecteur ligne) formée par la i ème ligne de A et A^j désigne la matrice $(n, 1)$ (appelée aussi vecteur colonne) formée par la j ème colonne de A.

— On remarque que pour tout vecteur $x \in \mathbb{R}^n$ (on considèrera toujours chaque vecteur comme une matrice unicolonne) :

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1 \underbrace{\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{e^1} + x_2 \underbrace{\begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}}_{e^2} + \cdots + x_n \underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}}_{e^n} = \sum_{k=1}^{n} x_k e^k$$

qui fait apparaître la base $(e^1, e^2, \ldots, e^n)$ appelée base canonique de $\mathbb{R}^n$ (cette famille est génératrice par définition mais on montre très facilement qu'elle est libre). On a $(e^j)_i = \delta_{i,j}$ (le symbole de Kronecker).

— Si A est une matrice (n, m) et B une matrice (m, p) alors le produit matriciel AB est bien défini et donne une matrice de taille (n, p) avec :

$$(AB)_{i,j} = \sum_{k=1}^{m} a_{i,k} b_{k,j},\ \forall (i,j) \in [\![1, n]\!] \times [\![1, p]\!]$$

— Si A est une matrice (n, m) alors $A^\top$, la transposée de A est une matrice (m, n) avec $(A^\top)_{i,j} = a_{j,i}$ (les lignes de A forment les colonnes de $A^\top$).

— Si x et y sont deux vecteurs de $\mathbb{R}^n$ alors :

$$(x|y) := x_1 y_1 + \cdots + x_n y_n = \sum_{k=1}^{n} x_k y_k$$

est le produit scalaire canonique de $\mathbb{R}^n$. Il permet de généraliser la notion d'orthogonalité usuelle. On remarque que $(x|y) = x^\top y = y^\top x$.

2.1.1 Rappels théoriques

On connaît une condition qui assure l'existence (quelque soit le second membre b) et l'unicité d'une solution de (2.1), c'est à dire :

$$\forall b \in \mathbb{R}^n, \exists! x \in \mathbb{R}^n\ :\ Ax = b$$

qui est $\det(A) \neq 0$. On peut "voir" un système linéaire de deux façons :

Vue géométrique comme intersection d'hyperplans affines (via les lignes de A)

En dimension 2 (2.1) s'écrit :

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 &= b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 &= b_2 \end{aligned}$$

il s'agit de deux équations de droite (on suppose que les coefficients de chaque droite ne sont pas tous nuls, c'est à dire que $\forall i, \exists j : a_{i,j} \neq 0$) et on cherche donc l'intersection de ces 2 droites (on cherche x qui appartient à la droite 1 et à la droite 2). On sait que :

— si les deux droites sont confondues, il y a une infinité de solutions ;

— si les droites sont non confondues mais parallèles il n'y a pas de solution ;

— sinon (droites non parallèles) il y a un et un seul point d'intersection, c'est à dire une et une seule solution.

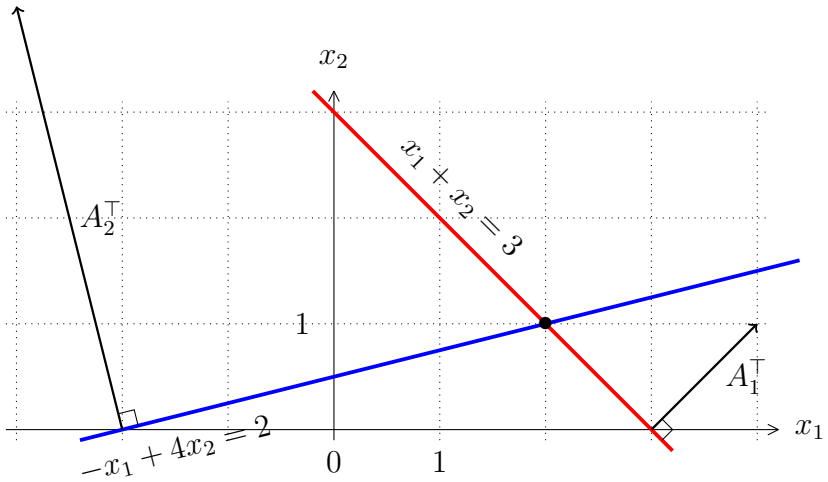
On peut remarquer que ce dernier cas semble générique et en fait ne dépend pas du second membre b :

— si les coefficients de chaque droite étaient choisis au hasard (selon une certaine loi à expliciter), il y aurait vraiment très peu de chance (en fait une probabilité nulle) que les droites soient parallèles ;

— dans ce cas de droites non parallèles, on voit bien que les coefficients b_1 et b_2 ne jouent aucun rôle dans l'existence et l'unicité d'une solution (mais bien sûr la solution change avec b).

Considérons par exemple le système linéaire $Ax = b$ suivant :

$$\underbrace{\begin{bmatrix} 1 & 1 \\ -1 & 4 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 3 \\ 2 \end{bmatrix}}_{b} \iff \begin{cases} x_1 + x_2 &= 3 \\ -x_1 + 4x_2 &= 2 \end{cases}$$



Les deux droites ne sont pas parallèles, leur intersection est le point $[2, 1]^\top$ qui est donc l'unique solution de ce système linéaire.

Remarque : dans cette figure apparaît le fait géométrique suivant bien connu : une droite d'équation $\alpha_1 x_1 + \alpha_2 x_2 = \beta$ est perpendiculaire au vecteur $\alpha := [\alpha_1, \alpha_2]^\top$ constitué par ses coefficients. En effet pour obtenir un vecteur v parallèle à cette droite il suffit de faire la différence entre deux points quelconques qui sont sur la droite, par exemple x et x' dont les coordonnées vérifient l'équation de la droite :

$$\begin{aligned} \alpha_1 x_1 + \alpha_2 x_2 &= \beta \\ \alpha_1 x'_1 + \alpha_2 x'_2 &= \beta \end{aligned}$$

En faisant la différence de ces deux équations, il vient :

$$\alpha_1 \underbrace{(x_1 - x'_1)}_{=v_1} + \alpha_2 \underbrace{(x_2 - x'_2)}_{=v_2} = 0$$

ce qui s'écrit encore $(\alpha|v) = 0$, tout vecteur v parallèle à la droite est orthogonal avec le vecteur α.

Cette interprétation permet de comprendre à quoi correspond une équation linéaire de la forme suivante dans $\mathbb{R}^n$ (où au moins un des coefficients appelé α_{i^*} est non nul) :

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = \beta \tag{2.2}$$

Il s'agit d'un hyperplan affine (dans un espace vectoriel à n dimensions c'est un sous espace affine [1] de dimension $n-1$) passant par [2] :

$$x^* = \frac{\beta}{\alpha_{i^*}} e^{i^*}$$

et orthogonal au vecteur $\alpha = [\alpha_1, \ldots, \alpha_n]^\top$. En effet d'après cette interprétation géométrique, un point x appartiendrait à l'"hyperplan si et seulement si : $(\alpha|x - x^*) = 0$. En développant ce produit scalaire il vient :

$$(\alpha|x) - \underbrace{(\alpha|x^*)}_{=\beta} = 0 \iff \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n = \beta$$

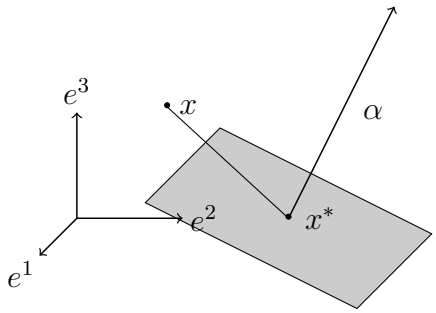
1. Dans $\mathbb{R}^n$ un sous-espace affine est une droite, un plan, etc, qui ne "passe" pas forcément en $0 = [0, \ldots, 0]^\top$: un sous-espace vectoriel passe toujours en 0 ; un sous-ensemble E de $\mathbb{R}^n$ est un sous-espace affine si et seulement si il est stable par combinaison barycentrique : $\forall x, y \in E$, $\forall \alpha, \beta \in \mathbb{R}$ tels que $\alpha + \beta = 1$ on doit avoir $\alpha x + \beta y \in E$.

2. Il est clair que x^* vérifie bien l'équation (2.2).

et on retrouve bien l'équation (2.2). Une notion qui sera utilisée plus tard (en particulier pour le cours de GRO) est celle de demi-espace : un tel hyperplan permet de séparer l'espace en 2 parties. Par exemple les points qui sont "au-dessus" de l'hyperplan ("au-dessus" au sens de la direction du vecteur α) et comprenant l'hyperplan lui même seront caractérisés par :

$$(\alpha|x - x^*) \geq 0 \iff \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \geq \beta$$

Exemple en 3D, avec l'hyperplan $0x_1 + x_2 + 2x_3 = 2$:



Le point x dessiné vérifie $(\alpha|x - x^*) > 0$ (c'est à dire $x_2 + 2x_3 > 2$) il est bien strictement "au-dessus" du plan vis à vis de l'orientation donnée par le vecteur $\alpha = [0, 1, 2]^\top$ (le point du plan x^* choisi a pour coordonnée $x^* = [0, 2, 0]^\top$).

Ainsi pour un système linéaire $Ax = b$ a n équations et n inconnues, l'interprétation géométrique se généralise de la façon suivante : on cherche l'intersection des n hyperplans d'équation $A_i x = b_i \iff a_{i,1}x_1 + a_{i,2}x_2 + \cdots + a_{i,n}x_n = b_i$, $i = 1, \ldots, n$.

Vue algébrique (via les colonnes de A)

Reprenons notre système linéaire 2×2 mais en l'écrivant de la façon suivante :

$$\left\{ \begin{array}{rcl} a_{1,1}x_1 + a_{1,2}x_2 & = & b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 & = & b_2 \end{array} \right. \iff x_1 \underbrace{\begin{bmatrix} a_{1,1} \\ a_{2,1} \end{bmatrix}}_{A^1} + x_2 \underbrace{\begin{bmatrix} a_{1,2} \\ a_{2,2} \end{bmatrix}}_{A^2} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}}_{b} \iff x_1 A^1 + x_2 A^2 = b$$

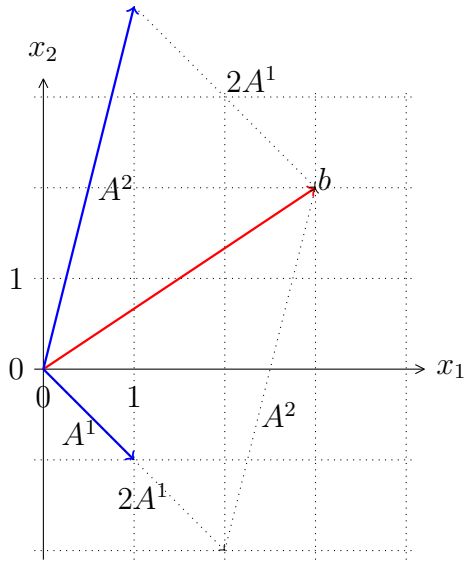
Le problème apparaît alors comme la recherche d'une combinaison linéaire des colonnes de la matrice A égale au second membre b. Pour un système $n \times n$:

$$Ax = b \iff x_1 A^1 + x_2 A^2 + \ldots x_n A^n = b \iff \sum_{j=1}^{n} x_j A^j = b$$

Pour le système 2×2 vu précédemment :

$$\left\{ \begin{array}{rcl} x_1 + x_2 & = & 3 \\ -x_1 + 4x_2 & = & 2 \end{array} \right. \iff x_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

On peut aussi dessiner cela mais la résolution (trouver la bonne combinaison linéaire) n'est pas si évidente, il faut projeter b sur A^1 parallèlement à A^2 (ou l'inverse) (on retrouve bien sûr la solution $[2, 1]^\top$ obtenue précédemment par intersection des deux droites) :



Sur le déterminant d'une matrice

Quelques propriétés du déterminant d'une matrice carrée :

— On peut définir le déterminant de $A \in \mathbb{R}^{n\times n}$ comme l'application multi-linéaire alternée des colonnes de A dans $\mathbb{R}$ qui vaut 1 sur la matrice identité. C'est à dire que $\det(A) := \det(A^1, A^2, \dots, A^n)$ où :

— multilinéaire veut dire linéaire en chacun des arguments :

$$\det(\dots, \alpha u + \beta v, \dots) = \alpha \det(\dots, u, \dots) + \beta \det(\dots, v, \dots)$$

— alternée veut dire que le déterminant est nul si deux arguments sont identiques ;

— et enfin $\det(e^1, \dots, e^n) = 1$.

Nous n'aurons pas vraiment besoin de cette définition mais les propriétés multilinéaire et alternée permettent de simplifier certains déterminants. *Exercice : montrer que si on échange deux colonnes d'un déterminant, il change de signe* $\det(\dots, u, \dots, v, \dots) = -\det(\dots, v, \dots, u, \dots)$. Aide : partir de $\det(\dots, u+v, \dots, u+v, \dots) = 0$.

— Un déterminant peut se calculer récursivement par la formule de Laplace, en notant $[A]_{(-i,-j)}$ la matrice obtenue en supprimant la ligne i et la colonne j de A, on a :

— développement par rapport à la colonne j de la matrice :

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+j} a_{i,j} \det([A]_{(-i,-j)})$$

— développement par rapport à la ligne i de la matrice :

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} \det([A]_{(-i,-j)})$$

(cf exercice sur les matrices triangulaires). On se ramène alors à des calculs de déterminants de matrices $(n-1, n-1)$, et, sachant que l'on sait calculer le déterminant d'une matrice $(1,1)$ [3] on peut en déduire une méthode de calcul (très inefficace) d'un déterminant quelconque.

3. $\det(a) = a$; en effet $\det(a) = a\det(1)$ par linéarité par rapport au premier (et seul) argument puis $\det(1) = 1$ car le déterminant vaut 1 sur la "matrice" identité $(1,1)$). *Exercice : en utilisant une formule de Laplace retrouver la formule pour un déterminant* 2×2.

— $\det(A^1, A^2, \ldots, A^n)$ est aussi égal à la mesure signée du parallélépipède généralisé défini par les n vecteurs dans $\mathbb{R}^n$: pour $n = 2$ c'est l'aire (orienté) du parallélogramme A^1, A^2, pour $n = 3$ c'est le volume (orienté) du parallélépipède A^1, A^2,A^3, etc. Ainsi pour $n = 2$ si A^1 et A^2 sont colinéaires, le parallélogramme correspondant est dégénéré et son aire est nulle (donc $\det(A) = 0$). De même pour $n = 3$, si les 3 vecteurs sont liés, le parallélépipède est aussi dégénéré et son volume est nul, etc.
— On a $\det(A^\top) = \det(A)$ (en particulier cela veut dire que l'aire du parallélogramme A_1, A_2 est égale à l'aire du parallélogramme A^1, A^2, cf dessins précédents).
— Si A et B sont 2 matrices (n, n), $\det(AB) = \det(A)\det(B)$.

Systèmes linéaires généraux

On peut aussi s'intéresser à des systèmes linéaires n'ayant pas forcément le même nombre d'équations (m) que d'inconnues (n) : étant donné $A \in \mathbb{R}^{m\times n}$ et $b \in \mathbb{R}^m$ on cherche $x \in \mathbb{R}^n$ vérifiant $Ax = b$. Les deux points de vue précédents sont toujours valables :
— l'ensemble des solutions est l'intersection (peut être vide) des m hyperplans affines d'équation $A_i x = b_i$ de $\mathbb{R}^n$;
— l'ensemble des solutions est l'ensemble (peut être vide) des $x \in \mathbb{R}^n$ tels qu'on obtient b par combinaison linéaire des colonnes de A ($\sum_{j=1}^n x_j A^j = b$).

Existence d'une solution : l'ensemble des $y \in \mathbb{R}^m$ tels qu'il existe $x \in \mathbb{R}^n$ avec $y = Ax$ est appelée ImA : on montre facilement que c'est un sous-espace vectoriel de l'espace but $\mathbb{R}^m$ et (comme $Ax = \sum_j x_j A^j$) on peut aussi le voir comme le sous-espace vectoriel formés des combinaisons linéaires des colonnes de A, soit :

$$ImA = Vect(A^1, \ldots, A^n)$$

L'existence d'au moins une solution pour b donné s'écrit donc $b \in ImA$. Il est clair que si $ImA = \mathbb{R}^m$ on a donc existence pour tout second membre b ; par contre lorsque $rangA := \dim(ImA) < m$, on aura pas forcément existence d'une solution. Si b est choisi au hasard, il y a très peu de chance qu'il y en ait une : supposons par exemple $rangA = 2$ et $m = 3$, ImA est donc un plan vectoriel et b choisi au hasard dans $\mathbb{R}^3$ a peu de chance de se trouver sur ce plan.

Unicité d'une (éventuelle) solution : on a unicité si et seulement si $KerA = \{0\}$ [4]. En effet soit 2 solutions x et x', vérifiant donc $Ax = b$ et $Ax' = b$. Par différence $Ax - Ax' = A(x - x') = b - b = 0$ et donc $x - x' \in KerA$. Ainsi si $KerA$ ne contient que le vecteur nul, $x = x'$. Réciproquement si $dim(KerA) \geq 1$ [5], et si x vérifie $Ax = b$ alors tout vecteur $x + u$ avec $u \in KerA$ est aussi solution du système linéaire :

$$A(x + u) = \underbrace{Ax}_{=b} + \underbrace{Au}_{=0} = b$$

dans un tel cas ($b \in ImA$ et $dim(KerA) \geq 1$) on a donc une infinité de solutions.

Théorème du rang : on a la relation suivante :

$$dim(KerA) + \underbrace{dim(ImA)}_{=rangA} = dim(\mathbb{R}^n) = n$$

4. Rappelons la définition du noyau de A : $KerA = \{x \in \mathbb{R}^n : Ax = 0\}$; on montre facilement que c'est un sous-espace vectoriel de l'espace de départ ($\mathbb{R}^n$) (exercice).
5. Si $KerA$ n'est pas restreint au vecteur nul c'est au moins une droite vectorielle.

Si on revient maintenant au cas carré ($n = m$) on voit que si $KerA = \{0\}$ alors $rangA = n$, l'unicité implique l'existence (pour tout second membre). Inversement si $rangA = n$ alors $KerA = \{0\}$, l'existence (pour tout second membre) implique l'unicité. Pour vérifier qu'une matrice est inversible il suffit donc de vérifier l'une de ces deux conditions.

2.1.2 Dans la pratique...

On veut donc résoudre un système linéaire carré. Soit $\det(A) \neq 0$ (ou $KerA = \{0\}$ ou $rangA = n$) et on a existence et unicité pour tout second membre, soit $\det(A) = 0$ (ce qui est exceptionnel si les coefficients de A sont choisis au hasard) et alors on a existence que si $b \in ImA$ (ce qui est exceptionnel si b est choisi au hasard) et dans ce cas on a une infinité de solutions.

Quelques questions et remarques :

— Vous avez appris en 1er cycle comment résoudre un système linéaire par une méthode de type Gauss. Est-ce la bonne méthode ?

Réponse : oui ! Voilà une bonne nouvelle. Attention en premier cycle on apprend plutôt la méthode de Gauss-Jordan qui permet d'obtenir directement la solution ; dans ce cours on va utiliser la méthode de Gauss qui conduit, partant de $Ax = b$ à obtenir un système linéaire équivalent $Ux = b'$ où U est une matrice triangulaire supérieure (il y a donc une étape supplémentaire) ; néanmoins on peut montrer (malgré l'étape supplémentaire de la résolution du système triangulaire) que c'est la méthode la plus économique et c'est donc elle qui est dans les bonnes bibliothèques d'algèbre linéaire. Il existe cependant d'autres méthodes (dite itératives comme le gradient conjugué, GMRES, etc.) qui sont utilisées pour résoudre certains grands systèmes linéaires creux (creux veut dire que la matrice comporte une très grande majorité de coefficients nuls).

— Doit-on d'abord vérifier que $\det(A) \neq 0$ (ou par exemple que $KerA = \{0\}$) avant de tenter un calcul ?

Réponse : non on découvre l'inversibilité par la méthode de Gauss (le déterminant de A s'obtient comme un sous-produit de cette méthode, dans la plupart des cas c'est la méthode la plus rapide pour le calculer !).

— On a souvent plusieurs systèmes linéaires à résoudre avec la même matrice : $Ax^{(k)} = b^{(k)}, k = 1, \ldots, K$. Les différents seconds membres $b^{(k)}$ n'étant pas forcément connus au même moment (par exemple $b^{(2)}$ peut dépendre de $x^{(1)}$, etc. Comment optimiser les calculs ?

Réponse : en réinterprétant la méthode de Gauss comme une factorisation de la matrice A (cf troisième partie de ce chapitre).

— En arithmétique flottante on sait que chaque calcul élémentaire peut être entaché d'une (généralement très petite) erreur ; si A est inversible (auquel cas on devrait pouvoir calculer la solution) quels critères pourraient nous permettre d'avoir une indication sur la précision de la solution obtenue ?

Réponse : on pourrait penser que la magnitude de $|det(A)|$ pourrait donner une indication : si $|det(A)| >> 1$ est grand alors (avec une bonne méthode de Gauss) les petites erreurs ne s'amplifie pas trop... Il n'en est rien, la bonne notion est celle de conditionnement de la matrice $A = \|A\|\|A^{-1}\|$ qui sera abordée dans la dernière partie de ce chapitre.

2.2 La méthode de Gauss

2.2.1 Résoudre un système triangulaire

Résoudre $Tx = b$ lorsque la matrice T est triangulaire supérieure (c'est à dire lorsque $t_{i,j} = 0$ pour $j < i$) est assez simple ; en tenant compte des coefficients nuls de la matrice, $Tx = b$ s'écrit :

$$\left\{\begin{array}{ccccccccccc}
t_{1,1}x_1 + & t_{1,2}x_2 & + \ldots & \ldots & \ldots & \ldots & + t_{1,n}x_n & = & b_1 \\
 & t_{2,2}x_2 & + \ldots & \ldots & \ldots & \ldots & + t_{2,n}x_n & = & b_2 \\
 & & \ddots & \ldots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & & & t_{i,i}x_i & + \ldots & \ldots & + t_{i,n}x_n & = & b_i \\
 & & & & \ddots & \vdots & \vdots & \vdots & \vdots \\
 & & & & & t_{n-1,n-1}x_{n-1} & + t_{n-1,n}x_n & = & b_{n-1} \\
 & & & & & & t_{n,n}x_n & = & b_n
\end{array}\right.$$

On peut montrer que $\det(T) = \prod_{k=1}^{n} t_{k,k}$ (cf exercice 2 feuille 2) et par conséquent T est inversible si et seulement si tous ses coefficients diagonaux sont non nuls. Pour résoudre un tel système il suffit de commencer par résoudre la dernière équation (étape 1) ce qui nous donne x_n puis (étape 2) connaissant x_n l'avant dernière équation nous donne x_{n-1}, etc. L'étape $n - i + 1$ utilise la i ème équation et à ce moment de l'algorithme on connait $x_{i+1}, \ldots, x_n$, on en déduit donc x_i (sachant que $t_{i,i} \neq 0$) :

$$x_i = \left(b_i - \sum_{j=i+1}^{n} t_{i,j}x_j\right) / t_{i,i}$$

D'où l'algorithme :

pour $i = n, n-1, \ldots, 1$
  $x_i \leftarrow \left(b_i - \sum_{j=i+1}^{n} t_{i,j}x_j\right) / t_{i,i}$

ou encore en écrivant la somme en pseudo-code [6] :

pour $i = n, n-1, \ldots, 1$
  $temp \leftarrow b_i$
  pour $j = i+1, \ldots, n$
    $temp \leftarrow temp - t_{i,j}x_j$
  $x_i \leftarrow temp/t_{i,i}$

Pour cet algorithme (parfois appelé "remontée") il est assez simple de calculer le nombre d'opérations arithmétiques (cf exercice 5 feuille 2), on obtient :

$$C_{remontee} : \frac{n(n-1)}{2} \text{ multiplications}, \quad \frac{n(n-1)}{2} \text{ additions}, \quad n \text{ divisions}$$

De même il est tout aussi facile de résoudre un système triangulaire inférieur (exercice).

6. On utilise la variable *temp* mais on aurait pu tout aussi bien utiliser directement x_i.

2.2.2 "Triangulariser" un système linéaire

La méthode de Gauss consiste à effectuer des transformations successives ($n-1$ en tout) sur (2.1) qui conduisent à un système linéaire équivalent dont la matrice est triangulaire supérieure : on a vu précédement qu'il était très simple de le résoudre !

Écrivons l'équation matricielle (2.1) comme système d'équations linéaires :

$$\left\{\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + \ldots\cdots + a_{1n}x_n &=& b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots\cdots + a_{2n}x_n &=& b_2 \\ \vdots & & \vdots \\ \vdots & & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots\cdots + a_{nn}x_n &=& b_n \end{array}\right. \tag{2.3}$$

— Si $a_{11} = 0$, on cherche alors $a_{k1} \neq 0$ et l'on échange les équations 1 et k : ceci est toujours possible puisque si tous les coefficients de la première colonne de A sont nuls alors $\det(A) = 0$.

— Supposons donc $a_{11} \neq 0$ (ce qui est donc toujours possible modulo un échange de lignes), l'idée de la méthode est de soustraire la première équation multipliée par le "bon" coefficient à la deuxième équation, de façon à éliminer la variable x_1 dans la nouvelle (deuxième) équation ainsi obtenue. On recommence ensuite cette manip sur l'équation 3, puis sur la 4, etc, et enfin sur la dernière. Le coefficient qui permet d'éliminer la variable x_1 dans l'équation i se calcule par :

$$coef_i^{(1)} = \frac{a_{i1}}{a_{11}}$$

et la nouvelle équation i obtenue $equ_i^{(1)} = equ_i - coef_i^{(1)} \times equ_1$ est :

$$\underbrace{(a_{i1} - \frac{a_{i1}}{a_{11}}a_{11})}_{=0}x_1 + (a_{i2} - \frac{a_{i1}}{a_{11}}a_{12})x_2 + \cdots + (a_{in} - \frac{a_{i1}}{a_{11}}a_{1n})x_n = b_i - \frac{a_{i1}}{a_{11}}b_1$$

Nous venons d'effectuer la première étape de la méthode de Gauss, le nouveau système linéaire obtenu est :

$$A^{(1)}x = b^{(1)} : \left\{\begin{array}{rcl} a_{11}x_1 \quad + a_{12}x_2 + \ldots\cdots + a_{1n}x_n &=& b_1 \\ 0 \quad + a_{22}^{(1)}x_2 + \ldots\cdots + a_{2n}^{(1)}x_n &=& b_2^{(1)} \\ \vdots \qquad\qquad \vdots & & \vdots \\ \vdots \qquad\qquad \vdots & & \vdots \\ 0 \quad + a_{n2}^{(1)}x_2 + \ldots\cdots + a_{nn}^{(1)}x_n &=& b_n^{(1)} \end{array}\right. \tag{2.4}$$

avec :

$$a_{ij}^{(1)} = a_{ij} - a_{1j}\frac{a_{i1}}{a_{11}} \text{ pour } 2 \leq i,j \leq n \text{ et } b_i^{(1)} = b_i - b_1\frac{a_{i1}}{a_{11}} \text{ pour } 2 \leq i \leq n.$$

Remarques :

1. la transformation qui permet d'obtenir le nouveau système (2.4) est inversible [7] et correspond à la multiplication par une matrice inversible $M^{(1)}$, on a $A^{(1)} = M^{(1)}A$ et $b^{(1)} = M^{(1)}b$;
2. les manipulations sur les équations du système sont équivalentes à des manipulations sur les lignes de la matrice A et du vecteur b : ainsi éliminer l'inconnue x_1 des équations $2, 3, \ldots, n$ correspond à faire apparaître des zéros sous la diagonale dans la première colonne de la matrice $(A^{(1)})$ avec les opérations :

$$ligne_i^{(1)} = ligne_i - coef_i^{(1)} \times ligne_1 \text{ pour } 2 \leq i \leq n.$$

7. Avec $equ_i^{(1)} + coef_i^{(1)} \times equ_1$ on retrouve les équations initiales.

Il faut bien sûr appliquer la même transformation sur le vecteur b.

Et la suite ? La deuxième étape revient à faire exactement la même chose sur le sous-système constitué par les équations $2, 3, \ldots, n$. Comme ce sous-système ne dépend pas de la variable x_1, c'est un système de $n-1$ équations à $n-1$ inconnues, sur lequel on applique la même technique :

— si le coefficient $a_{22}^{(1)} \neq 0$ on procède tout de suite à la phase d'élimination de l'inconnue x_2 dans les équations $3, 4, \ldots, n$;

— sinon on cherche un coefficient non nul :

$$a_{i2}^{(1)} \neq 0, \qquad 2 < i \leq n$$

puis on échange les équations 2 et i et l'on procède alors à la phase d'élimination.

Remarque : on peut toujours trouver $i \in [\![2, n]\!]$ tel que $a_{i2}^{(1)} \neq 0$; en effet comme la matrice $A^{(1)}$ a le découpage par blocs suivant :

$$A^{(1)} = \left(\begin{array}{c|c} a_{11} & a_{12} \ldots a_{1n} \\ \hline 0 & \\ \vdots & [A^{(1)}]_{[-1,-1)} \\ 0 & \end{array}\right)$$

le calcul de son déterminant (en développant par rapport à la première colonne) donne :

$$\det(A^{(1)}) = a_{11} \det([A^{(1)}]_{(-1,-1)}).$$

D'autre part comme $A^{(1)}$ est obtenue à partir de A sur laquelle on a appliqué une transformation linéaire inversible ($A^{(1)} = M^{(1)}A$), son déterminant est non nul. Ainsi, comme $a_{11} \neq 0$, la première colonne de la matrice $[A^{(1)}]_{(-1,-1)}$ ne peut être nulle (sinon son déterminant serait nul et par conséquent celui de $A^{(1)}$ aussi), c-a-d qu'il existe bien $i \in [\![2, n]\!]$ tel que $a_{i2}^{(1)} \neq 0$. De la même façon, on montre que ce résultat est aussi vrai pour la suite : si A est inversible on peut trouver, à chaque étape k de Gauss, un coefficient non nul :

$$a_{ik}^{(k-1)} \neq 0, \quad \text{avec } i \in [\![k, n]\!]$$

Ainsi la méthode de Gauss marche à tous les coups sur une matrice inversible (en arithmétique exacte !).

Et la fin ? Au bout de $n-1$ étapes (qui consistent donc toutes à faire la même chose sur des systèmes linéaires de plus en plus petits), on obtient le système linéaire équivalent $A^{(n-1)}x = b^{(n-1)}$, où la matrice $A^{(n-1)}$ est triangulaire supérieure :

$$A^{(n-1)} = \begin{pmatrix} a_{11} & \times & \ldots & \times \\ 0 & a_{22}^{(1)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \times \\ 0 & \ldots & 0 & a_{nn}^{(n-1)} \end{pmatrix}$$

avec ses éléments diagonaux (les pivots) $a_{ii}^{(i-1)}$ tous non nuls. Il suffit alors d'utiliser l'algorithme vu dans la section précédente pour résoudre ce système triangulaire.

2.3 La factorisation LU : écriture matricielle de la méthode de Gauss

2.3.1 Action d'une matrice de la forme $M = I - z(e^k)^\top$

Soit B une matrice (n, m), on regarde la transformation opérée par la multiplication MB où $z \in \mathbb{R}^n$, e^k est le k ème vecteur de la base canonique de $\mathbb{R}^n$ et I la matrice identité. La matrice M correspond à une matrice identité dans laquelle on a ajouté le vecteur $-z$ dans la k ème colonne (faire le calcul !).

$$MB = (I - z(e^k)^\top)B = B - z(e^k)^\top B = B - zB_k$$

En décomposant la matrice MB ligne par ligne, on obtient :

$$MB = \begin{pmatrix} \frac{B_1 - z_1 B_k}{B_2 - z_2 B_k} \\ \hline \vdots \\ \hline B_n - z_n B_k \end{pmatrix}$$

c'est à dire qu'à chaque ligne i de la matrice B initiale, on a soustrait la ligne k multipliée par le coefficient z_i (i ème composante du vecteur z). D'autre part si $z_k = 0$ alors la matrice M est nécessairement inversible d'inverse :

$$L = (M^{-1}) = I + z(e^k)^\top$$

En effet :

$$\begin{aligned} LM &= (I + z(e^k)^\top)(I - z(e^k)^\top) \\ &= I \underbrace{+z(e^k)^\top - z(e^k)^\top}_{0} - z(e^k)^\top z(e^k)^\top \\ &= I - z\left((e^k)^\top z\right)(e^k)^\top \\ &= I - z\underbrace{(z|e^k)}_{z_k=0}(e^k)^\top \\ &= I \end{aligned}$$

Avec cet outil, on obtient facilement les matrices qui permettent de passer d'une étape à l'autre dans la méthode de Gauss (l'échange des lignes pour trouver un pivot non nul, est lui obtenu par une matrice de permutation élémentaire ; nous en parlerons plus loin.) :

— pour l'étape 1 : $A^{(1)} = M^{(1)}A$ et $b^{(1)} = M^{(1)}b$ avec $M^{(1)} = I - z^{(1)}(e^1)^\top$ le vecteur $z^{(1)}$ étant :

$$z^{(1)} = \begin{pmatrix} 0 \\ a_{21}/a_{11} \\ \vdots \\ a_{n1}/a_{11} \end{pmatrix}$$

En effet $M^{(1)}$ effectue bien les opérations attendues sur les lignes :

$$\begin{cases} A_1^{(1)} = A_1 - 0 \times A_1 \\ A_i^{(1)} = A_i - z_i \times A_1 \ \text{ pour } 2 \leq i \leq n \end{cases}$$

(et on a le même effet sur b bien sûr).

— pour une étape k quelconque ($1 \leq k \leq n-1$), la matrice $A^{(k-1)}$ étant de la forme :

$$\begin{pmatrix} a_{11} & \times & \times & \times & \times & \times \\ 0 & a_{22}^{(1)} & \times & \times & \times & \times \\ \vdots & \ddots & \ddots & \times & \times & \times \\ \vdots & \dots & 0 & a_{kk}^{(k-1)} & \times & \times \\ \vdots & \dots & \vdots & \vdots & \times & \times \\ 0 & \dots & 0 & a_{nk}^{(k-1)} & \times & \times \end{pmatrix}$$

si $a_{kk}^{(k-1)} \neq 0$ [8] on peut procéder à l'élimination (de la variable x_k dans les équations $k+1, \dots, n$) en utilisant la matrice $M^{(k)} = I - z^{(k)}(e^{(k)})^\top$ avec :

$$z^{(k)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ a_{k+1,k}^{(k-1)}/a_{kk}^{(k-1)} \\ \vdots \\ a_{n,k}^{(k-1)}/a_{kk}^{(k-1)} \end{pmatrix}$$

Remarquons qu'avec ces vecteurs $z^{(k)}$ particuliers, les matrices $M^{(k)}$ sont inversibles d'inverse $I + z^{(k)}(e^{(k)})^\top$, et qu'elles sont aussi triangulaires inférieures puisque les k premières composantes des $z^{(k)}$ sont nulles (cf exercice 2 feuille 2). De plus comme les coefficients diagonaux de ces matrices triangulaires sont tous égaux à 1, on a aussi $\det(M^{(k)}) = 1$ (de même pour leurs inverses, cf exercice 2 feuille 2).

2.3.2 Théorème 1 : La factorisation $A = LU$

Soit une matrice A (n, n) inversible et telle qu'à chaque étape de la méthode de Gauss on ait (sans procéder à des échanges de lignes) :

$$a_{kk}^{(k-1)} \neq 0 \qquad \forall k \in [\![1, n-1]\!].$$

Alors il existe une unique factorisation de la matrice A de la forme :

$$A = LU$$

où L est une matrice triangulaire inférieure à diagonale unité (c-a-d que $l_{ii} = 1$, $\forall i$) et U est une matrice triangulaire supérieure inversible U.

Preuve abrégée : On pose $U = A^{(n-1)}$ la dernière matrice obtenue (qui possède donc les propriétés attendues, sauf qu' il reste à montrer que $a_{nn}^{(n-1)}$ est non nul), on a :

$$U = A^{(n-1)} = M^{(n-1)}A^{(n-2)} = M^{(n-1)}M^{(n-2)}A^{(n-3)} = M^{(n-1)}M^{(n-2)} \dots M^{(1)}A \qquad (2.5)$$

où les matrices $M^{(k)}$ sont de la forme $M^{(k)} = I - z^{(k)}(e^k)^\top$ avec $z_i^{(k)} = 0$ pour $i \in [\![1, k]\!]$. Si on prend les déterminants de l'équation ci-dessus :

$$\det(U) = \prod_{i=1}^{n} a_{ii}^{(i-1)} = \left(\prod_{i=1}^{n-1} \det(M^{(i)}) \right) \det(A) = \det(A)$$

8. Dans le cas contraire il faut au préalable, échanger la ligne k avec une ligne $i > k$ de façon à obtenir un pivot non nul.

car $\det(M^{(i)}) = 1$, par conséquent on a nécessairement $a_{nn}^{(n-1)} \neq 0$. Les inverses de ces matrices sont appelées $L^{(k)}$ et donc $L^{(k)} = I + z^{(k)}(e^k)^\top$ d'après un résultat précédent. En multipliant (2.5) à gauche par successivement $L^{(n-1)}$, $L^{(n-2)}$, ..., $L^{(1)}$, on obtient :

$$A = L^{(1)}L^{(2)} \dots L^{(n-1)}U \tag{2.6}$$

On pose alors $L = L^{(1)}L^{(2)} \dots L^{(n-1)}$. Comme produit de matrices triangulaires inférieures à diagonale unité, L est aussi triangulaire inférieure à diagonale unité (cf exercice 2 feuille 2) L'unicité fait l'objet de l'exercice 4 de la feuille 2. □

Pour le calcul effectif de la matrice L on a un résultat assez fort : elle s'obtient sans aucun calcul supplémentaire puisque :

$$L = I + \sum_{k=1}^{n-1} z^{(k)}(e^k)^\top = I + \left(\begin{array}{c|c|c|c|c} z^{(1)} & z^{(2)} & \dots & z^{(n-1)} & 0 \end{array} \right)$$

ce résultat se montrant en généralisant le calcul suivant : le produit de deux matrices $L^{(k)}L^{(k')}$ avec $k < k'$ est égal à :

$$L^{(k)}L^{(k')} = I + z^{(k)}(e^k)^\top + z^{(k')}(e^{k'})^\top$$

(faire le calcul).

Remarques :

1. Cette première factorisation repose sur l'hypothèse suivante : à chaque étape de la méthode de Gauss on suppose que $a_{kk}^{(k-1)} \neq 0$ (sans avoir à effectuer des échanges de lignes). Par exemple la matrice :
$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$
ne convient pas pour cette méthode. Cependant pour certaines matrices qui interviennent dans la pratique (matrices à diagonale dominante, matrices définies positives), cette condition est vérifiée, la factorisation étant de plus stable vis à vis des erreurs d'arrondi numérique dues aux calculs avec les flottants.
2. Cependant comme *a priori* il semble que l'on ait très peu de chance de tomber pile sur un zéro, cette méthode parait convenir dans la plupart des cas (il suffit de faire un test pour vérifier si le pivot est nul et arrêter la méthode au cas où, en gérant l'exception (message d'erreur, positionnement d'une variable booléenne, etc...)). En fait pour une matrice générale cet algorithme peut être très mauvais car un petit pivot peut amplifier les erreurs d'arrondi. La méthode standard consiste, à chaque étape k, à rechercher d'abord le pivot maximum en valeur absolue (ou module dans le cas complexe) dans la colonne k (à partir de la ligne k) :
$$\max_{i \in [\![k,n]\!]} \left| a_{ik}^{(k-1)} \right|$$
(que l'on suppose atteind en i_0 par exemple) et à échanger les lignes k et i_0. Matriciellement cela revient à multiplier en premier par une matrice de permutation $P^{(k)}$ avant de procéder à la phase d'élimination (correspondant à la multiplication par la matrice $M^{(k)}$) :
$$A^{(k)} = M^{(k)}P^{(k)}A^{(k-1)}.$$
Cette stratégie appelée "méthode de Gauss à pivot partiel" conduit à une factorisation du type $PA = LU$ où P est une matrice de permutation.

2.3.3 Utilisation d'une factorisation pour résoudre un système linéaire

Lorsque l'on a obtenu une factorisation $A = LU$, résoudre un système linéaire consiste à résoudre deux systèmes triangulaires : $Ax = b \Leftrightarrow LUx = b$, on pose $Ux = y$ et l'on résoud :

(i) $Ly = b$ on obtient y (résolution d'un système triangulaire inférieur)
(ii) $Ux = y$ on obtient x (résolution d'un système triangulaire supérieur)

Le coût est donc d'environ n^2 multiplications et n^2 additions.

Dans le cas d'une factorisation $PA = LU$, la méthode ci-dessus est précédée de l'application de la permutation sur le second membre : $Ax = b \Leftrightarrow PAx = Pb \Leftrightarrow LUx = Pb$.

2.3.4 Algorithme et coût de la méthode de Gauss/LU

Commençons par expliquer ce qui diffère entre la méthode de Gauss et la factorisation $A = LU$ ou $PA = LU$ d'une matrice. En fait presque rien : pour obtenir une factorisation ($A = LU$ ou $PA = LU$) on procède avec une méthode de Gauss qui agit uniquement sur la matrice (pas d'opérations sur un second membre) et qui **enregistre les coefficients d'élimination successifs** pour former la matrice L (qui s'obtient directement avec ces coefficients), et, dans le cas d'une méthode avec échange de lignes (pivot partiel qui conduit à une factorisation $PA = LU$), on enregistre aussi les permutations successives (en fait un seul tableau d'entiers de dimension n suffit). Dans la méthode de Gauss classique on procède aussi à la fin à la résolution du système triangulaire supérieur.

Voici un algorithme basique pour cette factorisation $A = LU$. Celui-ci utilise un tableau `LU` qui est initialisé par copie du tableau `A`. La factorisation est calculée "en place" dans le tableau `LU`, la place laissée par les 0 de l'étape k est utilisée pour stocker les coefficients d'élimination : ainsi la partie strictement triangulaire inférieure de L correspond alors à la partie strictement triangulaire inférieure du tableau `LU`, la partie triangulaire supérieure étant occupée par la matrice triangulaire supérieure U. Il n'y a pas de perte d'information car on sait que les coefficients de la diagonale de la matrice L sont tous égaux à 1. On obtient ainsi un stockage "compact" de la factorisation.

Initialisation du tableau `LU` *par copie :*
$\texttt{LU} \leftarrow \texttt{A}$
Boucle des $n-1$ *étapes :*
pour k de 1 à $n-1$
  si $\texttt{LU}_{k,k} = 0$
    arrêt de l'algorithme (gérer l'exception...)
  pour i de $k+1$ à n les opérations sur les lignes
    $\texttt{LU}_{i,k} \leftarrow \texttt{LU}_{i,k}/\texttt{LU}_{k,k}$ calcul et stockage du coef d'élimination $z_i^{(k)}$
    pour j de $k+1$ à n mise à jour de la ligne i
      $\texttt{LU}_{i,j} \leftarrow \texttt{LU}_{i,j} - \texttt{LU}_{i,k} \times \texttt{LU}_{k,j}$

Le nombre d'opérations arithmétiques de cet algorithme (cf exercice 5 feuille 2) s'élève à :

$$C_{LU} \ : \ \frac{n(n-1)(2n-1)}{6} \text{ additions et multiplications et } \frac{n(n-1)}{2} \text{ divisions}$$

Pour simplifier un peu, on ne prend en compte que les termes en $O(n^3)$, d'où :

$$C_{LU} \simeq \frac{1}{3}n^3 \text{ additions et multiplications}$$

Le coût de la factorisation $PA = LU$ avec la stratégie du pivot partiel n'est que légèrement supérieur car la recherche du coefficient maximum intervient avec un coût en $O(n^2)$ (de même que les opérations d'échanges de lignes, et le coût total de la mise à jour de la permutation peut se faire avec seulement $O(n)$ opérations). De même le coût de la méthode de Gauss classique, est juste un peu plus élevé puisque les opérations sur le second membre et la résolution finale du système triangulaire supérieur rajoutent simplement environ n^2 multiplications et additions.

2.3.5 A quoi sert la formalisation de la méthode de Gauss en factorisation sur la matrice ?

Dans la pratique, on a souvent à résoudre des systèmes linéaires qui comportent tous la même matrice :

$$Ax^{(i)} = b^{(i)} \qquad 1 \leq i \leq m \tag{2.7}$$

et où les vecteurs $b^{(i)}$ ne sont pas tous connus au même moment [9], par exemple $b^{(i+1)}$ s'obtient avec un calcul qui utilise $x^{(i)}$. Cette situation est très courante dans les problèmes d'ingénierie : souvent la matrice A correspond à la modélisation d'un système, le second membre est l'entrée imposée à ce système et la solution $x = A^{-1}b$ est la réponse du système pour l'entrée b. On a bien sûr envie de calculer les réponses $x^{(i)}$ qui correspondent à diverses entrées $b^{(i)}$.

Si on utilise bêtement une méthode de Gauss classique à chaque fois, on refait toujours les mêmes opérations sur la matrice A alors que c'est inutile ! On obtient donc un coût calcul d'environ :

$$C_{meth1} = m \times \left(C_{fact} + n^2\right)$$

où $C_{fact} \simeq \frac{1}{3}n^3$ est le nombre d'opérations (multiplications et additions/soustraction) correspondant au travail sur A (identique à celui de la factorisation donc), la partie n^2 correspond à la somme du travail d'élimination sur le second membre et du travail pour résoudre le système triangulaire supérieur final.

Si on procède d'abord par une factorisation, puis par m "descentes-remontées" on obtient :

$$C_{meth2} = C_{fact} + m \times n^2$$

Et :

$$\frac{C_{meth1}}{C_{meth2}} = \frac{m(C_{fract} + n^2)}{C_{fact} + mn^2} = \frac{m(C_{fract} + n^2)}{C_{fact} + n^2 + (m-1)n^2} = \frac{m}{1 + \frac{(m-1)n^2}{C_{fact}+n^2}}$$

et en utilisant l'approximation $C_{fact} \simeq \frac{1}{3}n^3$, on obtient :

$$\frac{C_{meth1}}{C_{meth2}} \simeq \frac{m}{1 + \frac{3(m-1)}{n}}$$

Ainsi pour $n >> m$ la méthode 2 tend à être presque m fois plus rapide que la méthode 1 et pour $m >> n$ la méthode 2 est environ $n/3$ plus rapide que la 1.

Le fait de pouvoir résoudre des systèmes linéaires assez rapidement (en n^2) une fois la factorisation calculée à d'autres avantages comme celui de pouvoir estimer $\|A^{-1}\|$ sans avoir à calculer explicitement A^{-1}.

9. Car il est simple de généraliser la méthode de Gauss classique pour qu'elle s'adapte à plusieurs seconds membres.

2.4 Notions sur les normes vectorielles et matricielles

2.4.1 Normes vectorielles

Exemples et définition

Le concept de norme généralise la notion de longueur d'un vecteur : grâce à un seul nombre on obtient une propriété importante d'un vecteur. Une norme permet de définir aussi une notion de "distance" entre deux vecteurs x et y en calculant la norme de la différence $\|x-y\|$. Dans cette partie du cours on utilise des vecteurs de $\mathbb{R}^n$ [10] et la norme habituelle est la norme euclidienne (ou norme 2) :

$$\|x\|_2 = \left(\sum_{k=1}^{n} x_k^2\right)^{\frac{1}{2}}$$

mais on utilisera aussi la norme infinie :

$$\|x\|_\infty = \max_{k\in[\![1,n]\!]} |x_k|$$

et la norme 1 :

$$\|x\|_1 = \sum_{k=1}^{n} |x_k|$$

D'une manière générale, étant donné un réel $p \geq 1$, la quantité suivante (appelée norme p de x) :

$$\|x\|_p = \left(\sum_{k=1}^{n} |x_k|^p\right)^{1/p}$$

définie bien une norme (mais nous n'aurons besoin essentiellement que des 3 normes précédentes).

Une norme est en fait une application de l'espace vectoriel réel E dans $\mathbb{R}^+$ qui doit vérifier les 3 axiomes suivants :

- (i) $\|x\| = 0 \iff x = 0$ (la norme est nulle uniquement pour le vecteur nul) ;
- (ii) $\|\alpha x\| = |\alpha|\|x\|, \forall x \in E, \forall \alpha \in \mathbb{R}$ (on dit que la norme est homogène) ;
- (iii) $\|x+y\| \leq \|x\| + \|y\|, \forall x, y \in E$. C'est l'inégalité triangulaire qui traduit une propriété géométrique attendue d'une norme comprise comme "longueur" d'un vecteur.

Avec une norme on peut faire de la "topologie" : en posant $d(x,y) := \|x-y\|$, on définit une distance [11] entre deux vecteurs. Par exemple, la boule ouverte de centre x et de rayon r :

$$B(x,r) := \{y \in E : \|y-x\| < r\}$$

est l'ensemble des vecteurs se situant à une "distance" strictement inférieure à r du vecteur x, la boule fermée étant :

$$\bar{B}(x,r) := \{y \in E : \|y-x\| \leq r\}$$

Ces boules dépendent du type de la norme utilisée et deux normes différentes vont définir des distances différentes (cf exercice 1 feuille 3).

Les normes vectorielles sont utiles dans les algorithmes travaillant sur des vecteurs où l'on doit arrêter l'algorithme sur un critère de proximité entre deux vecteurs. Quelques exemples :

10. Dans le prochain chapitre on verra aussi des normes de fonctions qui sont aussi des vecteurs !
11. Une distance répond aussi à des propriétés mathématiques précises que nous n'énoncerons pas ici.

1. arrêt de l'algorithme lorsque la distance (absolue) est inférieure à une tolérance ϵ (qui peut être précisée par l'utilisateur) :
$$\|x - y\| \leq \epsilon$$
2. souvent on préfère une sorte de distance "relative" tenant compte de la magnitude des vecteurs : on peut alors utiliser :
$$\|x - y\| \leq \epsilon \ \max\{\|x\|, \|y\|\}$$
mais si les vecteurs x et y peuvent être parfois très petits, on prèfère alors souvent utiliser un test absolu dans ce cas, ce qui peut se traduire par :
$$\|x - y\| \leq \epsilon \max\{s, \|x\|, \|y\|\}$$
par exemple avec $s = 10^{-3}$ on a un test relatif lorsque $\max\{\|x\|, \|y\|\} > 10^{-3}$ et un test absolu avec la tolérance $10^{-3}\epsilon$ quand $\max\{\|x\|, \|y\|\} \leq 10^{-3}$

Les normes sont aussi utiles pour tenir compte de l'effet des erreurs d'arrondi numérique. Par exemple si un vecteur x est codé dans la machine sans problème d'underflow ou d'overflow (on appelle $fl(x)$ le vecteur codé), on peut montrer pour toutes les normes p ainsi que la norme infinie (cf feuille 3) que :

$$\|fl(x) - x\| \leq \mathbf{u}\|x\|, \text{ ce qui s'écrit aussi } \frac{\|fl(x) - x\|}{\|x\|} \leq \mathbf{u} \text{ pour } x \neq 0.$$

On généralise ainsi le cas scalaire : sans underflow ni overflow l'erreur relative sur le codage d'un vecteur est aussi bornée par le epsilon machine.

Une propriété utile des normes : $\forall x, y \in E$, on a :

$$\|x - y\| \geq |\|x\| - \|y\||$$

Preuve : on a $\|x\| = \|(x - y) + y\| \leq \|x - y\| + \|y\|$ par l'inégalité triangulaire d'où :

$$\|x - y\| \geq \|x\| - \|y\|.$$

En inversant les rôles de x et y on obtient $\|y - x\| \geq \|y\| - \|x\|$ et comme

$$\|y - x\| = \| - 1(x - y)\| = | - 1|\|x - y\| = \|x - y\|$$

par homogénéité d'une norme, on obtient bien l'inégalité annoncée. □

Notions sur les suites convergentes et les suites de Cauchy

Une suite de vecteurs $x^{(0)}, x^{(1)}, \ldots$ d'un espace vectoriel E est une application de $\mathbb{N}$ dans E et est souvent notée $(x^{(k)})_{k \geq 0}$. Cependant comme l'indice de départ n'est pas forcément 0, on note plus généralement une suite en entourant le terme générique par des parenthèses $(x^{(k)})$.

Soit une suite de vecteurs $(x^{(k)})$ (d'un espace vectoriel normé E), on dit que cette suite converge vers $x^* \in E$ et on note $\lim_{k \to +\infty} x^{(k)} = x^*$ si et seulement si la suite de réels positifs $\|x^{(k)} - x^*\|$ tend vers 0 (quand k tend vers $+\infty$) :

$$\lim_{k \to +\infty} x^{(k)} = x^* \iff \lim_{k \to +\infty} \|x^{(k)} - x^*\| = 0$$

ce qui s'exprime aussi par :

$$\forall \epsilon > 0, \exists N(\epsilon) \in \mathbb{N} \text{ tel que } k \geq N(\epsilon) \Rightarrow \|x^{(k)} - x^*\| \leq \epsilon$$

En termes plus verbeux cela veut dire que quelque soit le rayon ϵ aussi petit que l'on veut (10^{-3},10^{-9},...), on peut trouver un entier N (qui dépend de ϵ d'où l'écriture $N(\epsilon)$ et en général plus ϵ est petit, plus $N(\epsilon)$ doit être grand) tel que tous les termes $x^{(k)}$ de la suite à partir du rang $N(\epsilon)$ (i.e. $k \geq N(\epsilon)$) se trouvent dans une boule de centre x^* et de rayon ϵ (tous les termes $x^{(k)}$ pour $k \geq N(\epsilon)$ sont à une distance au plus ϵ de la limite x^*).

On montre facilement que :

— si une suite est convergente, sa limite est unique ;

— si $(x^{(k)})$ est une suite qui converge vers x^* et (α_k) une suite de réels qui converge vers α alors la suite $(y^{(k)})$ où $y^{(k)} := \alpha_k x^{(k)}$ converge vers αx^* ;

— si $(x^{(k)})$ converge vers x^* et $(y^{(k)})$ une autre suite, converge vers y^* alors la suite $(z^{(k)})$ où $z^{(k)} := x^{(k)} + y^{(k)}$ converge vers $x^* + y^*$.

Suites de Cauchy

Parfois on aimerait décider de la convergence d'une suite sans forcément connaître sa limite. Si $(x^{(k)})$ est une suite convergente, on remarque que pour $k, n \geq N(\frac{\epsilon}{2})$ on a :

$$\|x^{(k)} - x^{(n)}\| = \|(x^{(k)} - x^*) + (x^* - x^{(n)})\| \leq \underbrace{\|x^{(k)} - x^*\|}_{\leq \frac{\epsilon}{2}} + \underbrace{\|x^* - x^{(n)}\|}_{\leq \frac{\epsilon}{2}} \leq \epsilon$$

Donc une suite convergente a aussi la propriété suivante :

$$\forall \epsilon > 0, \exists M(\epsilon) \in \mathbb{N} \text{ tel que } k, n \geq M(\epsilon) \Rightarrow \|x^{(k)} - x^{(n)}\| \leq \epsilon$$

appelée propriété de Cauchy (on dit que $(x^{(k)})$ est une suite de Cauchy). On remarque que la limite x^* de la suite n'intervient pas dans cette propriété qui reflète juste la condensation des termes de la suite : on choisit un seuil ϵ aussi petit que l'on veut et la distance entre deux termes quelconques de la suite de rang supérieur ou égal à $M(\epsilon)$ est toujours inférieure ou égale à ϵ.

Intuitivement une suite qui possède cette propriété de Cauchy semble évidemment convergente. Ceci est vrai dans un espace vectoriel normé complet [12], mais en dehors de ce cas la suite ne converge pas forcément vers un élément de l'espace de départ. Par exemple vous savez sans doute tous qu'il existe des suites de nombres rationnels qui convergent vers un nombre irrationnel. Un exemple classique est le suivant :

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right)$$

où $a > 0$. Partant de $x_0 > 0$ la suite [13] converge vers $\sqrt{a}$. On remarque facilement que si on choisit $a \in \mathbb{Q}$ et $x_0 \in \mathbb{Q}$ il est clair que tous les termes de la suite sont des nombres rationnels. Or avec $a = 2$ par exemple, la limite de la suite $(\sqrt{2})$ n'est pas un nombre rationnel. Notons en passant que la construction des nombres réels peut se faire en partant des nombres rationnels auxquels on rajoute la limite des suites de Cauchy de rationnels et que ce même procédé est utilisé pour rendre complet des e.v.n. qui ne le sont pas. Les espaces $\mathbb{R}^n$ sont des e.v.n. complets.

12. En fait un espace vectoriel normé complet est un e.v.n. dans lequel toutes les suites de Cauchy sont convergentes ! On a donc pas ajouté de nouvelles propriétés qui permettraient aux suites de Cauchy de converger systématiquement...

13. Il s'agit de la méthode de Héron.

Equivalence entre deux normes

Si dans un espace vectoriel E (e.v.) on dispose de plusieurs normes, par exemple $\|.\|_{N_1}$ et $\|.\|_{N_2}$ on fabrique alors a priori deux espaces vectoriels normés (e.v.n.) différents que l'on pourrait appeler $(E, \|.\|_{N_1})$ (E muni de $\|.\|_{N_1}$) et $(E, \|.\|_{N_2})$ (E muni de $\|.\|_{N_2}$). Une question naturelle est de savoir si elles induisent la même topologie, c'est à dire est-ce que finalement $(E, \|.\|_{N_1}) = (E, \|.\|_{N_2})$? Avoir la "même topologie" se résume aux deux questions suivantes :

— si une suite converge pour la norme N_1, converge-t-elle aussi la norme N_2 ?

— et inversement si une suite converge pour la norme N_2 converge-t-elle pour la norme N_1 ?

Définition : On dit que deux normes $\|.\|_{N_1}$ et $\|.\|_{N_2}$ définies sur un même e.v. E sont équivalentes si et seulement si il existe deux constantes $\gamma > 0$ et $\beta > 0$ telles que :

$$\gamma\|x\|_{N_1} \leq \|x\|_{N_2} \leq \beta\|x\|_{N_1},\ \forall x \in E$$

Remarquons que l'on alors l'encadrement suivant pour $\|x\|_{N_1}$:

$$\frac{1}{\beta}\|x\|_{N_2} \leq \|x\|_{N_1} \leq \frac{1}{\gamma}\|x\|_{N_2},\ \forall x \in E$$

L'équivalence des normes $\|.\|_{N_1}$ et $\|.\|_{N_2}$ permet de répondre oui à la question posée. En effet soit $(x^{(k)})$ une suite convergente (vers x^*) pour $\|.\|_{N_1}$. Comme :

$$0 \leq \|x^{(k)} - x^*\|_{N_2} \leq \beta\|x^{(k)} - x^*\|_{N_1}$$

il est clair que $\|x^{(k)} - x^*\|_{N_1} \to 0$ quand $k \to +\infty$ implique que $\|x^{(k)} - x^*\|_{N_2} \to 0$ quand $k \to +\infty$. De même l'inégalité $\|x\|_{N_1} \leq \frac{1}{\gamma}\|x\|_{N_2}$ permet de montrer que la convergence selon N_2 implique la convergence selon N_1.

Equivalence entre toutes les normes de $\mathbb{R}^n$

Montrons que les normes infinie et 1 sont équivalentes dans $\mathbb{R}^n$. On a en effet :

$$\|x\|_\infty = \max_{k\in[\![1,n]\!]} |x_k| \leq \sum_{k=1}^{n} |x_k| = \|x\|_1$$

et :

$$\|x\|_1 = \sum_{k=1}^{n} |x_k| \leq \sum_{k=1}^{n} \|x\|_\infty = n\|x\|_\infty$$

On a donc bien $\gamma\|x\|_\infty \leq \|x\|_1 \leq \beta\|x\|_\infty$, $\forall x \in \mathbb{R}^n$ avec $\gamma = 1$ et $\beta = n$.

En fait on peut montrer que toutes les normes sur $\mathbb{R}^n$ sont équivalentes [14], ce que nous admettrons ici.

2.4.2 Notions sur les normes matricielles

Il est assez facile de vérifier que l'ensemble des matrices réelles à m lignes et n colonnes (noté $\mathcal{M}_{m,n}(\mathbb{R})$ ou $\mathbb{R}^{m\times n}$) muni de l'opération interne $(A + B)_{i,j} = a_{i,j} + b_{i,j}$ et de l'opération externe $(\alpha A)_{i,j} = \alpha a_{i,j}$ est un espace vectoriel réel de dimensions $m \times n$. Si on peut considérer les matrices comme des vecteurs, on doit pouvoir munir un espace vectoriel de matrices de normes ce qui va

14. Ce résultat reste vrai pour tout espace vectoriel de dimension finie.

nous permettre de définir des "voisinages" et des "boules" de matrices, par exemple si A est une matrice donnée, $r > 0$ et si $\|.\|$ est une norme matricielle, on va pouvoir considérer :

$$B(A, r) := \{M \in \mathbb{R}^{m\times n} : \|M - A\| < r\}$$

et si r est suffisamment petit, on s'attend à ce que les matrices de $B(A, r)$ aient un comportement proche de celui de A. On montrera en TD que si A est inversible [15] on peut alors choisir $r > 0$ (dépendant de A) tel que les matrices de $B(A, r)$ soient aussi toutes inversibles.

Enfin concernant la résolution de systèmes linéaires $Ax = b$ du fait des erreurs de codage et des erreurs d'arrondi numérique commises lors des calculs, le système vraiment résolu par l'ordinateur peut correspondre en fait à la résolution exacte d'un système linéaire perturbé :

$$(A + \delta A)(x + \delta x) = b + \delta b$$

et on aimerait borner l'erreur relative sur la solution $\|\delta x\|/\|x\|$ en fonction des erreurs relatives sur les données $\|\delta A\|/\|A\|$ et $\|\delta b\|/\|b\|$.

Remarque : pour que la notion de norme matricielle soit la plus intéressante possible on rajoute une quatrième propriété :

$$\forall x \in \mathbb{R}^n, \|Ax\|_\beta \leq \|A\|\|x\|_\alpha$$

On dit alors que la norme matricielle $\|.\|$ est compatible avec la norme vectorielle $\|.\|_\alpha$ de l'espace de départ ($\mathbb{R}^n$) et la norme vectorielle $\|.\|_\beta$ de l'espace but ($\mathbb{R}^m$).

Il existe une manière générique de construire des normes matricielles à partir de normes vectorielles par le procédé suivant :

1. on choisit une norme vectorielle $\|.\|_\alpha$ de l'espace de départ $\mathbb{R}^n$ et une norme vectorielle $\|.\|_\beta$ de l'espace but $\mathbb{R}^m$ (si A est carrée, on choisit le plus souvent la même norme) ;
2. on définit alors :
$$\|A\|_{\alpha,\beta} := \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|_\beta}{\|x\|_\alpha}$$

De cette façon on définit bien une norme matricielle dite subordonnée (ou induite) par les normes vectorielles utilisées :

Théorème 2 : $\forall A \in \mathbb{R}^{m\times n}$ on a $\|A\|_{\alpha,\beta} < +\infty$ et l'application $A \mapsto \|A\|_{\alpha,\beta}$ est bien une norme sur l'e.v. $\mathbb{R}^{m\times n}$ qui vérifie aussi la quatrième propriété spécifique aux normes matricielles.

Preuve : La quatrième propriété s'obtient par la définition même de $\|A\|_{\alpha,\beta}$. En effet :

$$\|A\|_{\alpha,\beta} := \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|_\beta}{\|x\|_\alpha} \geq \frac{\|Ax\|_\beta}{\|x\|_\alpha}, \forall x \neq 0.$$

On obtient donc :

$$\|Ax\|_\beta \leq \|A\|_{\alpha,\beta}\|x\|_\alpha, \forall x \neq 0.$$

et cette inégalité reste vraie aussi pour $x = 0$.

Remarque : la suite de la preuve est donnée ci-après mais peut être omise... La démonstration du premier point ($\forall A \in \mathbb{R}^{m\times n}$ $\|A\|_{\alpha,\beta} < +\infty$) est faite toute à la fin (de deux façons différentes).

15. Il s'agit donc ici de matrices carrées, cad $m = n$.

Pour la propriété (i) :

$$\|A\|_{\alpha,\beta} = 0 \iff \frac{\|Ax\|_\beta}{\|x\|_\alpha} = 0, \forall x \neq 0 \iff \|Ax\|_\beta = 0, \forall x \iff Ax = 0, \forall x \iff A = 0$$

Pour la propriété (ii) : soit λ un réel quelconque :

$$\|\lambda A\|_{\alpha,\beta} = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|\lambda Ax\|_\beta}{\|x\|_\alpha} = \sup_{x\in\mathbb{R}^n, x\neq 0} |\lambda| \frac{\|Ax\|_\beta}{\|x\|_\alpha} = |\lambda| \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|_\beta}{\|x\|_\alpha} = |\lambda| \|A\|_{\alpha,\beta}$$

Pour la propriété (iii) : soit A et B deux matrices quelconques de $\mathbb{R}^{m\times n}$:

$$\|A+B\|_{\alpha,\beta} = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|(A+B)x\|_\beta}{\|x\|_\alpha} = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax+Bx\|_\beta}{\|x\|_\alpha}$$

En appliquant l'inégalité triangulaire $\|Ax+Bx\|_\beta \leq \|Ax\|_\beta + \|Bx\|_\beta$, il vient :

$$\sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax+Bx\|_\beta}{\|x\|_\alpha} \leq \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|_\beta + \|Bx\|_\beta}{\|x\|_\alpha} \leq \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|_\beta}{\|x\|_\alpha} + \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Bx\|_\beta}{\|x\|_\alpha}$$

d'où :

$$\|A+B\|_{\alpha,\beta} \leq \|A\|_{\alpha,\beta} + \|B\|_{\alpha,\beta}$$

Pour le premier point voici une première démonstration qui utilise des notions plus avancées d'analyse. Elle est suivie d'une démonstration utilisant des notions plus élémentaires.

Dem 1. Montrons d'abord qu'on peut aussi utiliser la définition suivante pour $\|A\|_{\alpha,\beta}$:

$$\|A\|_{\alpha,\beta} := \max_{\|x\|_\alpha = 1} \|Ax\|_\beta$$

En effet soit donc $x \neq 0$, et $y = \lambda x$ avec $\lambda \neq 0$. En faisant varier λ dans $\mathbb{R}^*$ on parcourt toute la droite vectorielle engendrée par x (sauf 0). Or :

$$\frac{\|Ay\|_\beta}{\|y\|_\alpha} = \frac{\|A\lambda x\|_\beta}{\|\lambda x\|_\alpha} = \frac{\|\lambda Ax\|_\beta}{|\lambda|\|x\|_\alpha} = \frac{|\lambda|\|Ax\|_\beta}{|\lambda|\|x\|_\alpha} = \frac{\|Ax\|_\beta}{\|x\|_\alpha}$$

ainsi ce rapport donne la même valeur sur toute la droite vectorielle $Vect\{x\}$ (sauf 0 où il n'est pas défini) et on peut se restreindre à le calculer pour $y \in Vect\{x\}$ tel que $\|y\|_\alpha = 1$ (on a 2 tels vecteurs mais cette légère redondance n'est pas bien grave). Finalement on obtient donc :

$$\|A\|_{\alpha,\beta} = \sup_{\|x\|_\alpha = 1} \|Ax\|_\beta$$

Dans les cours d'analyse un peu plus avancés on apprend que la sphère unité d'un e.v.n. de dimension finie est un ensemble compact [16] et qu'une fonction continue [17] définie sur un compact "atteint" sa borne supérieure et sa borne inférieure, d'où l'utilisation du max à la place du sup de départ et de la finitude de $\|A\|_{\alpha,\beta}$:

$$\|A\|_{\alpha,\beta} = \max_{\|x\|_\alpha = 1} \|Ax\|_\beta = \|Ax^*\|_\beta < +\infty$$

où x^* est un vecteur qui réalise le max de $\|Ax^*\|_\beta$ sur la sphère unité de $\mathbb{R}^n$ selon la norme $\|.\|_\alpha$. *Rmq :* on peut aussi montrer que :

$$\|A\|_{\alpha,\beta} = \sup_{\|x\|_\alpha \leq 1} \|Ax\|_\beta$$

16. Dans un e.v.n. de dimension finie E, un compact est un sous-ensemble fermé et borné de E.
17. On peut montrer que la fonction $x \mapsto \|Ax\|_\beta$ est bien continue.

Dem 2. Comme $Ax = \sum_{j=1}^{n} x_j A^j$, en utilisant l'inégalité triangulaire ($\|.\|_\beta$), il vient :

$$\|Ax\|_\beta \leq \sum_{j=1}^{n} \|x_j A^j\|_\beta$$

Puis par homogénéité (de $\|.\|_\beta$) :

$$\sum_{j=1}^{n} \|x_j A^j\|_\beta = \sum_{j=1}^{n} |x_j| \|A^j\|_\beta$$

On continue en utilisant la majoration $|x_j| \leq \max_k |x_k| = \|x\|_\infty$ ce qui nous donne :

$$\sum_{j=1}^{n} |x_j| \|A^j\|_\beta \leq \underbrace{\max_k |x_k|}_{=\|x\|_\infty} M \text{ avec } M = \sum_j \|A^j\|_\beta$$

Enfin on utilise l'équivalence entre les normes de $\mathbb{R}^n$:

$$\|Ax\|_\beta \leq M\|x\|_\infty \leq M\gamma\|x\|_\alpha \text{ soit } \frac{\|Ax\|_\beta}{\|x\|_\alpha} \leq M\gamma, \forall x \neq 0$$

d'où :

$$\|A\|_{\alpha,\beta} := \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|_\beta}{\|x\|_\alpha} \leq M\gamma < +\infty$$

Quelques résultats sur les normes matricielles subordonnées

Le plus souvent les normes matricielles subordonnées utilisées seront définies en utilisant la même norme vectorielle p ($p = 1, 2$ ou ∞) pour l'espace de départ et l'espace but. On notera $\|A\|_p$ cette norme plutôt que $\|A\|_{p,p}$. En TD on montrera que :

$$\|A\|_\infty := \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_i \sum_j |a_{i,j}|$$

et que :

$$\|A\|_1 := \sup_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_j \sum_i |a_{i,j}|$$

et donc ces deux normes matricielles sont très simples à calculer avec peu d'opérations (d'ordre le nombre d'éléments $m \times n$ de la matrice). On peut aussi montrer que :

$$\|A\|_2 := \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\rho(A^\top A)}$$

où $\rho(M)$ désigne le rayon spectral de la matrice M c'est à dire :

$$\rho(M) = \max_{\lambda \in Spec(M)} |\lambda|$$

Sans entrer dans les détails le calcul de $\|A\|_2$ est beaucoup plus coûteux que celui de $\|A\|_\infty$ et $\|A\|_1$ ce qui explique l'importance de ces deux dernières normes dans les calculs pratiques.

Résultat 1 : Soit une $\|.\|$ une norme matricielle subordonnée à la norme vectorielle $\|.\|$ de $\mathbb{R}^n$, alors $\|I\| = 1$.

Preuve :

$$\|I\| = \sup_{x\neq 0} \frac{\|Ix\|}{\|x\|} = \sup_{x\neq 0} \frac{\|x\|}{\|x\|} = 1$$

Résultat 2 : Soit $A \in \mathbb{R}^{m\times n}$ et $B \in \mathbb{R}^{p\times m}$ et $\|.\|_n$, $\|.\|_m$, $\|.\|_p$, un choix de normes vectorielles sur respectivement $\mathbb{R}^n$, $\mathbb{R}^m$ et $\mathbb{R}^p$, alors :

$$\|BA\|_{n,p} \leq \|B\|_{m,p}\|A\|_{n,m}$$

Preuve : On a :

$$\|BA\|_{n,p} := \sup_{x\neq 0} \frac{\|BAx\|_p}{\|x\|_n}$$

mais $\|BAx\|_p = \|B(\underbrace{Ax}_{y\in\mathbb{R}^m})\|_p \leq \|B\|_{m,p}\|Ax\|_m \leq \|B\|_{m,p}\|A\|_{m,n}\|x\|_n$ d'où :

$$\|BA\|_{n,p} := \sup_{x\neq 0} \frac{\|BAx\|_p}{\|x\|_n} \leq \sup_{x\neq 0} \frac{\|B\|_{m,p}\|A\|_{n,m}\|x\|_n}{\|x\|_n} = \|B\|_{m,p}\|A\|_{n,m}$$

Un corollaire facile de ce résultat est que pour une matrice carrée A on aura $\|A^k\| \leq \|A\|^k$ (attention ici $A^k = \underbrace{A \times A \cdots \times A}_{k \text{ fois}}$).

Définition : soit $A \in \mathbb{R}^{n\times n}$ une matrice inversible. On appelle conditionnement de A pour la norme matricielle $\|.\|$, le nombre $\kappa(A) = \|A\|\|A^{-1}\|$.

Résultat 3 : On a toujours $\kappa(A) \geq 1$.

Preuve : D'après le résultat 1 on a $1 = \|I\|$, et comme $I = AA^{-1}$, on obtient en utilisant le résultat 2 :

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = \kappa(A).$$

Remarque : une matrice avec $\kappa(A)$ pas trop grand sera dite bien conditionnée et ne posera pas de problème pour la résolution d'un système linéaire. A l'inverse une matrice avec $\kappa(A) \simeq 1/\mathbf{u}$ ne pourra pas être traitée convenablement dans le système flottant utilisé (cf TD). En bref le conditionnement est la bonne mesure pour l'inversibilité numérique d'une matrice en arithmétique flottante.

Théorème 3 : Soit $B \in \mathbb{R}^{n\times n}$ telle que $\|B\| < 1$ pour une norme matricielle subordonnée alors :

(i) $\quad I - B$ est inversible
(ii) $\quad (I-B)^{-1} = I + B + B^2 + \cdots = \sum_{k\geq 0} B^k$
(iii) $\quad \|(I-B)^{-1}\| \leq \frac{1}{1-\|B\|}$

Rmq : de même ici B^k désigne la k ème puissance de la matrice B.

Preuve : Pour l'inversibilité montrons que $Ker(I-B) = \{0\}$:

$$(I-B)x = 0 \iff x = Bx \Rightarrow \|x\| = \|Bx\| \Rightarrow \|x\| \leq \|B\|\|x\|$$

Comme $\|B\| < 1$ cette dernière inégalité n'est possible que pour $x = 0$. Donc $I - B$ est bien une matrice inversible.

Pour (ii), posons $S^{(K)} = \sum_{k=0}^{K} B^k$, la convergence de la série $I + B + B^2 + \dots$ correspond à la convergence de la suite $(S^{(K)})$. On a :

$$(I-B)S^{(K)} = (I-B)(I+B+\cdots+B^K) = I-B+B-B^2+\cdots-B^K+B^K-B^{K+1} = I-B^{K+1}$$

et comme $I - B$ est inversible :

$$S^{(K)} = (I-B)^{-1}(I-B^{K+1}) \iff S^{(K)} - (I-B)^{-1} = -(I-B)^{-1}B^{K+1}$$

en utilisant le résultat 2 :

$$\|S^{(K)} - (I-B)^{-1}\| = \|(I-B)^{-1}B^{K+1}\| \leq \|(I-B)^{-1}\|\|B^{K+1}\| \leq \|(I-B)^{-1}\|(\underbrace{\|B\|}_{<1})^{K+1} \to 0 \text{ qd } K \to \infty$$

d'où :

$$S := \lim_{K\to+\infty} S^{(K)} = (I-B)^{-1}$$

Pour le dernier point en utilisant l'inégalité triangulaire et le résultat 2, il vient :

$$\|S\| \leq \underbrace{\|I\|}_{=1} + \|B\| + \|B^2\| + \|B^3\| + \cdots \leq 1 + \|B\| + \|B\|^2 + \|B\|^3 + \cdots = \frac{1}{1-\|B\|} \quad \square$$

Pour terminer ce chapitre voici un résultat sur la méthode de Gauss en arithmétique flottante. Il utilise la notation $|A|$ pour désigner la matrice de coefficients $|a_{i,j}|$.

Théorème 4 : Soit $A \in \mathbb{R}^{n\times n}$. On suppose que la méthode de Gauss (avec ou sans échanges d'équations) conduite en arithmétique flottante sans overflow ni underflow, a permis d'obtenir des facteurs $\hat{L}$ et $\hat{U}$ et une solution $\hat{x}$ du système linéaire $Ax = b$, alors on a :

$$(A + \delta A)\hat{x} = b, \text{ avec } |\delta A| \leq 2\gamma_n |\hat{L}||\hat{U}| \text{ et } \gamma_n = \frac{n\mathbf{u}}{1-n\mathbf{u}}$$

Commentaires :

1. Ce résultat nous dit donc que la solution numérique obtenue est la solution exacte du système perturbé $(A + \delta A)\hat{x} = b$ et nous permet a posteriori de calculer une majoration de la perturbation δA [18].

   Pour obtenir une majotation sur l'erreur relative sur la solution, il faut :
   — se choisir une norme matricielle subordonnée $\|.\|$ (norme 1 ou infinie) ;
   — utiliser le résultat suivant (on montrera une inégalité de ce type en TD) sur la solution d'un système perturbé :
   $$\frac{\|x - \hat{x}\|}{\|\hat{x}\|} \leq \kappa(A)\frac{\|\delta A\|}{\|A\|}$$
   — calculer la norme de la matrice qui majore $|\delta A|$ (on obtient alors une majoration de $\|\delta A\|$) ;

18. Pour cela il faut calculer le produit matriciel $|\hat{L}||\hat{U}|$ ce qui peut être coûteux ou alors on majore par le produit des deux normes mais ce n'est pas optimal car $\||\hat{L}|\|\|\hat{U}|\|$ peut être bien plus grand que $\||\hat{L}||\hat{U}|\|$.

— calculer $\|A\|$ puis une estimation [19] de $\|A^{-1}\|$ de manière à obtenir une estimation de $\kappa(A)$.

2. Si on utilise la stratégie du pivot partiel alors il est facile de voir que la matrice triangulaire inférieure à diagonale unité $\hat{L}$ est telle que $|\hat{l}_{i,j}| \leq 1$. On a donc en norme 1 ou en norme infinie $\||\hat{L}|\| \leq n$. On peut alors se passer d'effectuer le produit matriciel $|\hat{L}||\hat{U}|$ pour obtenir une majoration de $\|\delta A\|$.

   Enfin on peut aussi montrer que cette stratégie permet de limiter la croissance des coefficients des matrices successives $A^{(k)}$ et donc d'obtenir une matrice $\hat{U}$ "raisonnable". Plus précisément en posant :

$$\rho_n = \frac{\max_{i,j,k} |a_{i,j}^{(k)}|}{\max_{i,j} |a_{i,j}|}$$

   le facteur de croissance des coefficients, on a :

$$\|\delta A\|_\infty \leq 2n^2 \gamma_n \rho_n \|A\|_\infty$$

   et Wilkinson (un des plus grands contributeurs sur l'influence des erreurs d'arrondi numérique dans les méthodes numériques) n'aurait jamais vu de facteur ρ_n dépassant 16 sur les matrices intervenant dans les applications ! Cependant on peut construire des matrices très spéciales pour lesquelles ce facteur de croissance peut attendre 2^n et on ne peut donc pas dire que la méthode de Gauss à pivot partiel est stable pour toutes les matrices.

19. Connaissant une factorisation de A il existe des algorithmes rapide (en $O(n^2)$) permettant d'estimer $\|A^{-1}\|$ sans calculer A^{-1}.