Chapitre I

L'optimisation linéaire et la méthode du simplexe

1. Notations matricielles

On note $\mathcal{L}(\mathbf{R}^m, \mathbf{R}^n)$ l'espace des applications linéaires de $\mathbf{R}^m$ dans $\mathbf{R}^n$. On l'identifiera à l'espace des matrices à n lignes et m colonnes, à coefficients réels.

Soit $A \in \mathcal{L}(\mathbf{R}^m, \mathbf{R}^n)$. Pour tout entier i $(1 \leq i \leq m)$ et tout entier j $(1 \leq j \leq n)$, on note

A_i la i-ème colonne de A; c'est un vecteur-colonne à n composantes;

A^j la j-ème ligne de A; c'est un vecteur-ligne à m composantes;

A_i^j le coefficient de A situé dans la i-ème colonne et la j-ème ligne.

Plus généralement, soit γ une application de l'ensemble $\{1, 2, \ldots, k\}$ dans $\{1, 2, \ldots, m\}$. On note A_γ la matrice à n lignes et k colonnes dont les colonnes sont $A_{\gamma(1)}, \ldots, A_{\gamma(k)}$.

De même, soit λ une application de l'ensemble $\{1, 2, \ldots, h\}$ dans $\{1, 2, \ldots, n\}$ On note A^λ la matrice à h lignes et m colonnes dont les lignes sont $A^{\lambda(1)}, \ldots, A^{\lambda(h)}$.

Soit $A \in \mathcal{L}(\mathbf{R}^m, \mathbf{R}^n)$, $B \in \mathcal{L}(\mathbf{R}^p, \mathbf{R}^m)$. On note $AB \in \mathcal{L}(\mathbf{R}^p, \mathbf{R}^n)$ l'application composée. On a la formule bien connue

$$(AB)_i^j = A^j B_i = \sum_{k=1}^{m} A_k^j B_i^k \,, \quad (1 \leq i \leq p \,,\ 1 \leq j \leq n) \,.$$

Cette formule se généralise au cas où γ est une application de $\{1, \ldots, k\}$ dans $\{1, \ldots, p\}$ et λ une application de $\{1, \ldots, h\}$ dans $\{1, \ldots, n\}$:

$$(AB)_\gamma^\lambda = A^\lambda B_\gamma \,.$$

En particulier, si λ est l'application identique de $\{1, \ldots, n\}$,

$$(AB)_\gamma = AB_\gamma \,,$$

et si γ est l'application identique de $\{1, \ldots, p\}$,

$$(AB)^\lambda = A^\lambda B \,.$$

2. Problèmes d'optimisation linéaire et géométrie des polytopes

2.1. Définition. On appelle *problème d'optimisation linéaire sous forme standard* le problème suivant:

Trouver un élément x de $\mathbf{R}^m$ vérifiant

$$Ax = b\,, \qquad x \geq 0\,,$$

et rendant maximum la fonction

$$f(x) = cx\,.$$

Dans ce qui précède, $A \in \mathcal{L}(\mathbf{R}^m, \mathbf{R}^n)$, $b \in \mathbf{R}^n$ et $c \in (\mathbf{R}^m)^* = \mathcal{L}(\mathbf{R}^m, \mathbf{R})$ sont donnés. On a écrit

$$x \geq 0$$

pour exprimer que toutes les composantes de x vérifient

$$x^i \geq 0 \quad \text{pour tout } i \in \{1, \ldots, m\}\,.$$

2.2. Autres problèmes d'optimisation linéaire. On rencontre d'autres problèmes d'optimisation linéaire qui, par un changement de variables, se ramènent au problème sous forme standard. On en donnera deux exemples.

1. On suppose donnés $A_1 \in \mathcal{L}(\mathbf{R}^p, \mathbf{R}^n)$, $b \in \mathbf{R}^n$ et $c_1 \in (\mathbf{R}^p)^* = \mathcal{L}(\mathbf{R}^p, \mathbf{R})$. On doit trouver $x_1 \in \mathbf{R}^p$ vérifiant

$$A_1 x_1 \leq b\,, \qquad x_1 \geq 0\,,$$

et rendant maximum la fonction

$$f_1(x_1) = c_1 x_1\,.$$

On a écrit

$$A_1 x_1 \leq b$$

pour exprimer que

$$(A_1 x_1)^i \leq b^i \quad \text{pour tout } i \in \{1, \ldots, n\}\,.$$

Afin de ramener ce problème à la forme standard, on introduit la variable d'écart $x_2 \in \mathbf{R}^n$. On pose:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbf{R}^{n+p}\,, \qquad Ax = A_1 x_1 + x_2\,.$$

Le problème s'exprime alors:

Trouver $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbf{R}^{n+p}$ vérifiant

$$Ax = b\,, \qquad x \geq 0\,,$$

et rendant maximum la fonction

$$f(x) = c_1 x_1\,.$$

On l'a donc ramené à la forme standard.

2. Les notations étant les mêmes qu'en 2.1, on peut rencontrer le cas où l'on n'impose pas à toutes les composantes de l'inconnue $x \in \mathbf{R}^m$ d'être ≥ 0. On ramène le problème à la forme standard en *dédoublant* ces composantes. Par exemple, si la première composante x^1 peut être de signe quelconque, on posera

$$x^1 = y^1 - z^1 ,$$

et on imposera à y^1 et à z^1 d'être ≥ 0. La nouvelle inconnue aura alors $m+1$ composantes $y^1, z^1, x^2, \ldots, x^m$. La relation imposée $Ax = b$ s'écrit, avec les nouvelles inconnues,

$$A_1^i(y^1 - z^1) + \sum_{j=2}^{m} A_j^i x^j = b^i , \quad (1 \leq i \leq n) ,$$

et la fonction à maximiser f s'écrit

$$f(y^1 - z^1, x^2, \ldots, x^m) = c_1(y^1 - z^1) + \sum_{j=2}^{m} c_j x^j .$$

Le problème est ainsi ramené à la forme standard.

Dans ce qui suit, on revient au problème sous forme standard et on suppose $A \in \mathcal{L}(\mathbf{R}^m, \mathbf{R}^n)$ de rang n.

2.3. Remarques.

1. On peut tout aussi bien chercher à minimiser la fonction linéaire f (il suffit de changer c en $-c$).

2. Ce problème est d'importance pratique considérable. On s'y réfère parfois par le terme de "programmation linéaire". Le mot programmation étant à prendre au sens ancien de programme d'activité que l'on modélise par des données linéaires.

3. Donnons un exemple (un peu trop simple pour correspondre à une situation réelle, mais significatif) de problème concret relevant de la programmation linéaire. Une usine fabrique divers produits P_i $(1 \leq i \leq m)$, en utilisant des matières premières Q_j $(1 \leq j \leq n)$. La quantité de matière première Q_j disponible est supposée fixée: on la note b^j. La fabrication d'une unité du produit P_i nécessite une quantité A_i^1 de la matière première Q_1, une quantité A_i^2 de la matière première Q_2, ..., une quantité A_i^n de la matière première Q_n. Chaque unité de produit P_i vendue procure au fabricant un revenu net de c_i unités monétaires. Soit x^i la quantité de produit P_i qui est fabriquée. Le revenu du fabricant est

$$f(x^1, \ldots, x^n) = \sum_{i=1}^{m} c_i x^i .$$

Le fabricant cherche à maximiser ce revenu, en respectant les contraintes

$$x^i \geq 0 ,$$

exprimant que les quantités produites ne peuvent être négatives, et

$$\sum_{i=1}^{m} A_i^j x^i \leq b^j \,, \quad (1 \leq j \leq n) \,,$$

exprimant que la quantité consommée de chaque matière première ne peut pas être supérieure à la quantité disponible.

On reconnaît un problème d'optimisation du type de celui décrit en 2.2.1.

2.4. Notions sur les convexes, les polytopes et les polyèdres. Ainsi qu'on le verra bientôt, la résolution du problème d'optimisation linéaire 2.1 fait intervenir quelques propriétés des parties convexes de $\mathbf{R}^m$, en particulier des polytopes et des polyèdres. On va tout d'abord rappeler quelques notions à ce propos.

Une partie A de $\mathbf{R}^m$ est dite *convexe* si, pour tout couple de points x et y de A et tout réel λ vérifiant $0 \leq \lambda \leq 1$, le point $\lambda x + (1-\lambda)y$ est élément de A. Géométriquement, cela signifie que le segment de droite joignant deux points x et y de A est entièrement contenu dans A.

Par exemple, soit $f \in (\mathbf{R}^m)^* = \mathcal{L}(\mathbf{R}^m, \mathbf{R})$ une forme linéaire non identiquement nulle sur $\mathbf{R}^m$. Pour tout k réel, l'ensemble des points $x \in \mathbf{R}^m$ tels que $f(x) \leq k$ est appelé *demi-espace fermé* d'équation $f(x) \leq k$. On voit immédiatement que c'est une partie convexe de $\mathbf{R}^m$.

On vérifie aussi immédiatement que l'intersection d'une famille quelconque de convexes est convexe.

On appelle *polytope convexe* de $\mathbf{R}^m$ l'intersection (supposée non vide) d'une famille finie de demi-espaces fermés de $\mathbf{R}^m$. Un polytope convexe est évidemment convexe et fermé.

Par exemple, l'ensemble des points x de $\mathbf{R}^m$ qui vérifient $x \geq 0$ (c'est-à-dire dont toutes les composantes x^i sont ≥ 0) est un polytope convexe, appelé *cône positif*.

Autres exemples: pour toute forme linéaire non identiquement nulle f sur $\mathbf{R}^m$ et tout réel k, les demi-espaces fermés d'équations $f \leq k$ et $f \geq k$ sont des polytopes convexes. Leur intersection est l'hyperplan affine d'équation $f = k$. C'est aussi un polytope convexe. Cet exemple montre qu'un polytope convexe de $\mathbf{R}^m$ peut être de dimension strictement inférieure à m (la dimension d'un polytope étant, par définition, la dimension du plus petit sous-espace affine de $\mathbf{R}^m$ qui contient ce polytope).

On appelle *polyèdre convexe* de $\mathbf{R}^m$ un polytope convexe borné de $\mathbf{R}^m$. Un polyèdre convexe est fermé et borné, donc compact.

Soient $x_1, \ldots, x_k$ des points de $\mathbf{R}^m$. Un point x de $\mathbf{R}^m$ est dit *combinaison convexe* des points x_i $(1 \leq i \leq k)$ s'il peut s'écrire sous la forme

$$x = \sum_{i=1}^{k} \lambda_i x^i \,,$$

les λ_i étant des réels vérifiant

$$\lambda_i \geq 0 \,, \ (1 \leq i \leq k) \,, \quad \sum_{i=1}^{k} \lambda_i = 1 \,.$$

On vérifie aisément que l'ensemble des combinaisons convexes des points x_i $(1 \leq i \leq k)$ est convexe et fermée. C'est le plus petit convexe contenant les x_i. On dit que c'est *l'enveloppe convexe* de la famille de points $(x_1, \ldots, x_k)$. On montre que c'est un polyèdre convexe.

Soit A une partie convexe et fermée de $\mathbf{R}^m$. Un point x de A est dit *extrémal* si l'égalité

$$x = ty + (1-t)z\,, \quad y \text{ et } z \in A\,, \quad 0 < t < 1\,,$$

n'a pas d'autre solution que $y = z = x$. Géométriquement, cela signifie qu'il n'existe pas de segment de droite non réduit à un point, contenu dans A et contenant le point x dans son intérieur (c'est-à-dire de manière telle que x ne soit pas une des extrémités de ce segment).

Lorsque A est un polytope (ou un polyèdre) convexe, ses points extrémaux sont aussi appelés *sommets*.

On montre qu'un polytope (ou un polyèdre) convexe A de $\mathbf{R}^m$ peut être décomposé en *faces* de diverses dimensions, comprises entre 0 et la dimension du polytope (elle-même inférieure ou égale à m). Les sommets sont les faces de dimension 0. La face de plus grande dimension (égale à $\dim A$, la dimension du polytope) est unique; c'est *l'intérieur* de ce polytope. Les faces de dimension 1 sont aussi appelées *arêtes*. Une face de dimension n $(0 \leq n \leq \dim A \leq m)$ est un ouvert connexe d'un sous-espace affine de dimension n de $\mathbf{R}^m$; son adhérence est un polytope convexe de dimension n; c'est la réunion de cette face et de certaines autres faces de A de dimension $< n$.

2.5. Définition. On appelle *ensemble admissible* pour le problème d'optimisation linéaire sous forme standard 2.1, l'ensemble

$$M = \left\{ x \in \mathbf{R}^m \;\middle|\; Ax = b\,,\; x \geq 0 \right\}.$$

L'ensemble admissible M est l'intersection du sous-espace affine (de dimension $m - n$, puisque A a été supposée de rang n) $\left\{ x \in \mathbf{R}^m \;\middle|\; Ax = b \right\}$, et du cône convexe fermé $\left\{ x \in \mathbf{R}^m \;\middle|\; x \geq 0 \right\}$. L'ensemble M est donc convexe et fermé. Ainsi qu'on l'a vu en 2.4, c'est un *polytope* convexe. Si de plus M est borné, donc compact, c'est un *polyèdre* convexe.

2.6. Proposition. *Soit M l'ensemble des états admissibles du problème d'optimisation 2.1, et $x \in M$. On note $I(x)$ la suite ordonnée des indices $i \in \{1, \ldots, m\}$ tels que $x^i > 0$. Les deux propriétés suivantes sont équivalentes:*

(i) *x est un sommet de M,*

(ii) *les vecteurs A_i, $i \in I(x)$, sont linéairement indépendants.*

Démonstration. Supposons (ii) faux. Il existe alors un élément $z \in \mathbf{R}^m$, $z \neq 0$, tel que $z^i = 0$ pour $i \notin I(x)$, vérifiant $Az = 0$. Pour ϵ réel de module $|\epsilon|$ assez petit, on a

$$x \pm \epsilon z \geq 0 \quad \text{et} \quad A(x \pm \epsilon z) = b\,,$$

donc $x \pm \epsilon z \in M$. Par suite

$$x = \frac{1}{2}\big((x + \epsilon z) + (x - \epsilon z)\big)\,,$$

ce qui prouve que (i) est faux.

Réciproquement supposons (i) faux. Il existe alors y et $z \in M$, $y \neq z$, et $\lambda \in]0,1[$, tels que

$$x = \lambda y + (1-\lambda)z\,.$$

Soit $i \in \{1,\dots,m\}$. On a $x^i = \lambda y^i + (1-\lambda)z^i$, avec $x^i \geq 0$, $y^i \geq 0$, $z^i \geq 0$, et $0 < \lambda < 1$. Par suite, x^i est nul si et seulement si y^i et z^i sont tous deux nuls. Le vecteur $y - z$, non identiquement nul, vérifie $y^i - z^i = 0$ pour tout $i \notin I(x)$, et il est dans le noyau de A puisque $A(y-z) = 0$, ce qui prouve que (ii) est faux. □

2.7. Proposition. *Soit M l'ensemble des états admissibles du problème d'optimisation 2.1. On suppose M non vide. Alors il existe au moins un sommet de M.*

Démonstration. Pour tout point z de M, on note $I(z)$ l'ensemble des indices $i \in \{1,\dots,m\}$ tels que $z^i > 0$, et $|I(z)|$ le nombre d'éléments de $I(z)$. Soit $I = \inf_{z\in M} |I(z)|$. Il existe $x \in M$ tel que $|I(x)| = I$. Montrons que x est un sommet de M. Si $I = 0$, $x = 0$ et c'est évidemment un sommet. Supposons $I > 0$. Si x n'était pas un sommet de M, les I vecteurs A_i, $i \in I(x)$, ne seraient pas linéairement indépendants. Il existerait donc des réels λ^i, $i \in I(x)$, non tous nuls, tels que $\sum_{i\in I(x)} \lambda^i A_i = 0$. Pour $j \in \{1,\dots,m\}\backslash I(x)$, on pose $\lambda^j = 0$. Le vecteur $\lambda \in \mathbf{R}^m$ ainsi défini vérifie $A\lambda = 0$. Pour tout réel t, on pose $z(t) = x + t\lambda$. On a

$$A\big(z(t)\big) = b\,.$$

L'ensemble des réels t tels que $z(t) \geq 0$ est un intervalle fermé, contenant un intervalle ouvert contenant l'origine, nécessairement borné au moins d'un côté. Soit t_m une extrémité (finie) de cet intervalle. Alors $z(t_m)$ est élément de M et $I\big(z(t_m)\big) \subset I(x)$, l'inclusion étant stricte car il existe nécessairement un élément i de $I(x)$ tel que $z^i(t_m) = 0$. Ceci étant en contradiction avec la définition de I, x est un sommet de M. □

2.8. Proposition. *L'ensemble convexe (polytope ou polyèdre)*

$$M = \big\{\, x \in \mathbf{R}^m \;\big|\; Ax = b\,,\ x \geq 0 \,\big\}\,,$$

a un nombre fini de sommets $\nu(M) \leq C_m^n$ (rappelons que C_m^n est le nombre de façons de choisir n colonnes parmi les m colonnes de A).

Cette proposition est bien sûr un corollaire immédiat de 2.6.

2.9. Théorème. *Soit M l'ensemble des états admissibles du problème d'optimisation 2.1. On suppose M non vide et borné, et on note $E(M)$ l'ensemble des sommets de M. Alors M est l'ensemble des combinaisons convexes d'éléments de $E(M)$; cela signifie qu'un point $z \in \mathbf{R}^m$ est élément de M si et seulement s'il existe des éléments $x_1,\dots,x_k$ de $E(M)$ et des réels $\lambda_1,\dots,\lambda_k$, vérifiant*

$$\lambda_i \geq 0\,, \quad \sum_{i=1}^{k} \lambda_i = 1\,, \quad z = \sum_{i=1}^{k} \lambda_i x_i\,.$$

Démonstration. Soit $z \in M$ et $r = |I(z)|$. Si $r = 0$, $z = 0$ et c'est un sommet.

Faisons l'hypothèse de récurrence: pour $r \leq l$, z est une combinaison convexe d'éléments de $E(M)$. Supposons $r = l + 1$. Si z n'est pas un sommet, il existe u et v, éléments de M, $u \neq v$, et $t_0 \in]0,1[$, tels que $z = t_0 u + (1 - t_0)v$.

Pour tout réel t, posons

$$z(t) = tu + (1 - t)v\,.$$

L'ensemble M étant supposé borné, l'ensemble des $t \in \mathbf{R}$ tels que $z(t) \in M$ est un intervalle fermé et borné $[\alpha, \beta]$. On a $\big|I\big(z(\alpha)\big)\big| < |I(z)|$ et $\big|I\big(z(\beta)\big)\big| < |I(z)|$, sans quoi α, ou β, ne serait pas extrémité de l'ensemble des t pour lesquels $z(t) \in M$. D'après l'hypothèse de récurrence, $z(\alpha)$ et $z(\beta)$ sont combinaisons convexes d'éléments de $E(M)$. Mais $z = z(t_0)$ est lui-même combinaison convexe de $z(\alpha)$ et de $z(\beta)$, donc combinaison convexe d'éléments de $E(M)$. Réciproquement, puisque $E(M) \subset M$ et que M est convexe, toute combinaison convexe d'éléments de $E(M)$ est élément de M. □

2.10. Théorème. *Soit M l'ensemble des états admissibles du problème d'optimisation 2.1. On suppose M non vide et borné. Alors f atteint son maximum sur M en un sommet.*

Démonstration. La fonction f est continue sur M, qui est compact puisque fermé et borné. Donc f atteint son maximum en un point z de M. Si z n'est pas un sommet, il existe d'après le théorème 2.9 des sommets $x_1, \ldots, x_k$ de M et des réels $\lambda_1, \ldots, \lambda_k$ vérifiant

$$\lambda_i \geq 0\,, \quad \sum_{i=1}^{k} \lambda_i = 1\,, \quad \sum_{i=1}^{k} \lambda_i x_i = z\,.$$

On en déduit

$$f(z) = cz = \sum_{i=1}^{k} \lambda_i f(x_i)\,,$$

d'où

$$\sum_{i=1}^{k} \lambda_i \big(f(z) - f(x_i)\big) = 0\,.$$

Les λ_i étant tous ≥ 0 et les $f(z) - f(x_i)$ aussi (car f atteint son maximum sur M au point z), on a nécessairement $f(z) = f(x_i)$ pour tout $i \in \{1, \ldots, k\}$ tel que $\lambda_i \neq 0$. Puisque $\sum_{i=1}^{k} \lambda_i = 1$, l'ensemble des $i \in \{1, \ldots, k\}$ tels que $\lambda_i \neq 0$ est non vide, ce qui montre que f atteint son maximum sur M en un sommet (au moins). □

Nous allons maintenant considérer la géométrie des polytopes non bornés.

2.11. Définition. On dit qu'un vecteur $y \in \mathbf{R}^m$, $y \geq 0$, est un *rayon infini* du polytope M si pour tout $x \in M$ et tout $\lambda \geq 0$, $x + \lambda y \in M$.

Remarquons que y est un rayon infini si et seulement si

$$Ay = 0 \quad \text{et} \quad y \geq 0\,.$$

L'ensemble

$$Y = \big\{\, y \in \mathbf{R}^m \;\big|\; Ay = 0 \text{ et } y \geq 0 \,\big\}$$

des rayons infinis est un cône convexe.

Considérons l'hyperplan H d'équation

$$\sum_{j=1}^{m} y^j = 1\,.$$

Son intersection $H \cap Y$ avec l'ensemble des rayons infinis est un polyèdre convexe. D'après la proposition 2.8, il admet un nombre fini de sommets $y_1, \ldots, y_k$, et d'après 2.9, tout point de $H \cap Y$ est combinaison convexe des y_i. Comme tout point de Y se déduit par une homothétie de rapport positif d'un point de $H \cap Y$, on en déduit que tout point de Y est combinaison linéaire à coefficients ≥ 0 des y_i, $1 \leq i \leq k$. Les vecteurs $y_1, \ldots, y_k$ sont appelés *rayons extrémaux* de M.

En procédant comme dans la démonstration de 2.9, on obtient:

2.12. Théorème. *Tout point d'un polytope convexe $M \subset \mathbf{R}^m$ est somme d'une combinaison convexe de ses point extrémaux et d'une combinaison linéaire à coefficients ≥ 0 de ses rayons extrémaux.*

3. La méthode du simplexe: aperçu général

3.1. Le problème.

La méthode du simplexe, introduite par G.B. Dantzig en 1947, permet de résoudre le problème d'optimisation linéaire 2.1, que nous rappelons ici:

– Trouver un élement $x \in \mathbf{R}^m$, vérifiant

$$Ax = b\,, \quad x \geq 0\,,$$

et rendant maximum la fonction

$$f(x) = cx\,,$$

où $A \in \mathcal{L}(\mathbf{R}^m, \mathbf{R}^n)$, $b \in \mathbf{R}^n$, $c \in (\mathbf{R}^m)^* = \mathcal{L}(\mathbf{R}^m, \mathbf{R})$ sont donnés.

Dans ce qui suit, on supposera A de rang n. Comme précédemment,

$$M = \left\{ x \in \mathbf{R}^m \;\middle|\; Ax = b\,,\ x \geq 0 \right\}$$

désigne l'ensemble admissible du problème, et $E(M)$ désigne l'ensemble des sommets de M.

3.2. Principe de la méthode.

Les grandes lignes de la méthode du simplexe sont les suivantes.

Première étape. On détermine un sommet x de M.

Deuxième étape. On applique à x un critère permettant de savoir s'il existe un point z de M tel que $f(z) > f(x)$. Si ce n'est pas le cas, f atteint son maximum en x et le problème est résolu. Si c'est le cas, on passe à la troisième étape.

Troisième étape. On détermine un autre sommet x' de M, voisin de x en un sens qu'on précisera, tel que $f(x') > f(x)$.

Quatrième étape. On revient à la deuxième étape, en remplaçant x par x'.

La méthode donne lieu à un algorithme qui, dans le cas où M est non vide et borné, conduit à la solution effective du problème en un nombre fini d'opérations. On laissera de côté pour le moment la première étape (la détermination d'un sommet de M), car on verra que lorsqu'on ne dispose pas d'un sommet en évidence, cette détermination peut se faire en résolvant un problème d'optimisation linéaire auxiliaire dans lequel l'ensemble admissible admet l'origine pour sommet. On va donc étudier d'abord en détail la deuxième, puis la troisième étape, et on reviendra ensuite sur la première.

Donnons pour commencer une définition.

3.3. Définition. Dans les hypothèses et avec les notations de 3.1, un sommet x de M est dit *non dégénéré* si le nombre de composantes non nulles de x est égal à n.

3.4. Remarques.

1. Soit x un sommet de M, $I(x)$ l'ensemble des indices $i \in \{1, \ldots, m\}$ tels que $x^i > 0$. D'après 2.5, les vecteurs A_i de $\mathbf{R}^n$, avec $i \in I(x)$, sont linéairement indépendants. Leur nombre est donc $\leq n$. Le sommet x est non dégénéré si et seulement si ce nombre est égal à n, donc le plus grand possible.

2. Soit γ une partie de $\{1, \ldots, m\}$ comportant n éléments, telle que les vecteurs A_i $(i \in \gamma)$ soient linéairement indépendants. Une telle partie de $\{1, \ldots, m\}$ existe toujours car la matrice A est supposée de rang n. On va montrer que s'il existe un sommet x de M tel que $I(x) \subset \gamma$, celui-ci est unique. En ordonnant γ d'une manière quelconque (par exemple l'ordre naturel) on peut considérer γ comme une application injective de $\{1, \ldots, n\}$ dans $\{1, \ldots, m\}$. Dire que les vecteurs A_i $(i \in \gamma)$ sont linéairement indépendants équivaut à dire que la matrice A_γ (notations du paragraphe 1) est inversible. Si x est un sommet de M tel que $I(x) \subset \gamma$, on a (toujours avec les notations du paragraphe 1)

$$A_\gamma x^\gamma = b\,,$$

donc

$$x^\gamma = (A_\gamma)^{-1} b\,.$$

Cette relation détermine x de manière unique, puisqu'on doit avoir

$$\begin{cases} x^{\gamma(i)} = \left((A_\gamma)^{-1} b\right)^i\,, \\ x^j = 0 \quad \text{pour } j \notin \gamma\,. \end{cases}$$

Le point x de $\mathbf{R}^m$ défini par ces relations est un sommet de M si et seulement si les $x^{\gamma(i)}$ $(1 \leq i \leq n)$ sont tous ≥ 0.

3. Soit x un sommet de M (pouvant éventuellement être dégénéré) et $I(x)$ l'ensemble des indices $i \in \{1, \dots, m\}$ tels que $x^i > 0$. On a vu (remarque 1 ci-dessus) que le nombre d'éléments de $I(x)$ est nécessairement $\leq n$. Puisque A est de rang n et que les A_i $(i \in I(x))$ sont linéairement indépendants, il existe une partie γ de $\{1, \dots, m\}$ comportant n éléments, contenant $I(x)$ et telle que les A_j $(j \in \gamma)$ soient linéairement indépendants. Lorsque le sommet x est non dégénéré, γ est déterminé de manière unique (puisque alors $\gamma = I(x)$). Lorsque x est dégénéré, γ est en général non unique.

4. L'ensemble des parties γ de $\{1, \dots, m\}$ comportant n éléments est fini (il a $\dfrac{m!}{n!(m-n)!}$ éléments). Par suite le nombre de sommets de M est fini.

4. Critère de maximalité: cas d'un sommet non dégénéré

4.1. Quelques formules. Les hypothèses et notations étant celles de 3.1, soit x un sommet non dégénéré de M, et $\gamma = I(x)$ l'ensemble des indices $i \in \{1, \dots, m\}$ tels que $x^i > 0$. Puisque x est non dégénéré, γ a n éléments. On l'ordonne d'une manière quelconque (par exemple selon l'ordre naturel); on peut alors considérer γ comme une application injective de $\{1, \dots, n\}$ dans $\{1, \dots, m\}$. Puisque x est un sommet, la matrice $n \times n$ A_γ (notations du paragraphe 1) est inversible.

Soit δ le complémentaire de γ dans $\{1, \dots, m\}$. On ordonne δ de manière quelconque (par exemple selon l'ordre naturel).

Pour tout élément z de M on peut écrire, avec les notations du paragraphe 1,

$$Az = A_\gamma z^\gamma + A_\delta z^\delta = b\,.$$

En particulier, si on remplace z par x, on a $x^\delta = 0$, donc

$$A_\gamma x^\gamma = b\,.$$

On en déduit

$$A_\gamma z^\gamma = A_\gamma x^\gamma - A_\delta z^\delta\,,$$

ou encore puisque A_γ est inversible,

$$z^\gamma = x^\gamma - (A_\gamma)^{-1} A_\delta z^\delta\,. \qquad (*)$$

Calculons maintenant $f(z)$.

$$\begin{aligned} f(z) = cz &= c_\gamma z^\gamma + c_\delta z^\delta \\ &= c_\gamma x^\gamma + \big(c_\delta - c_\gamma (A_\gamma)^{-1} A_\delta\big) z^\delta\,. \end{aligned}$$

Mais comme d'autre part

$$f(x) = c_\gamma x^\gamma\,,$$

on peut écrire

$$f(z) = f(x) + \big(c_\delta - c_\gamma (A_\gamma)^{-1} A_\delta\big) z^\delta\,. \qquad (**)$$

On peut énoncer:

4.2. Proposition. *Les hypothèses et notations étant celles précisées ci-dessus, la fonction f atteint son maximum sur M au sommet non dégénéré x si et seulement si toutes les composantes du vecteur-ligne (à $m-n$ composantes)*

$$c_\delta - c_\gamma (A_\gamma)^{-1} A_\delta \qquad (***)$$

sont ≤ 0.

Démonstration. Si toutes les composantes du vecteur-ligne $(***)$ sont ≤ 0, l'expression $(**)$ ci-dessus montre que f atteint son maximum sur M au point x, puisque toutes les composantes du vecteur-colonne z^δ sont ≥ 0. Réciproquement, si f atteint son maximum au point x, cette même expression montre que toutes les composantes du vecteur-ligne $(***)$ sont ≤ 0, car si l'une de ces composantes, d'indice $\delta(i)$, était > 0, on obtiendrait en prenant pour z^δ un vecteur-colonne ayant toutes ses composantes nulles sauf celle d'indice $\delta(i)$, celle-ci étant choisie strictement positive mais assez petite pour que les composantes de z^γ (données par la formule $(*)$) restent positives, un élément z de M tel que $f(z) > f(x)$. □

4.3. Remarque. Pour pouvoir appliquer effectivement le critère de maximalité 4.2, on doit connaître explicitement la matrice $(A_\gamma)^{-1} A_\delta$. On verra dans le paragraphe suivant qu'on peut déterminer cette matrice, à chaque remplacement d'un sommet par un autre.

5. Pivotement à partir d'un sommet non dégénéré

Les hypothèses et notations étant celles du paragraphe 4, on suppose de plus que l'application du critère de maximalité 4.2 a montré que f n'atteignait pas son maximum en x. Certaines composantes du vecteur-ligne $c_\delta - c_\gamma (A_\gamma)^{-1} A_\delta$ sont donc strictement positives. On choisit l'une de celles-ci, par exemple la plus grande. Soit k son indice ($k \in \{1, \ldots, m-n\}$). On a donc par hypothèse

$$c_{\delta(k)} - \left(c_\gamma (A_\gamma)^{-1} A_\delta\right)_k > 0\,.$$

On va déterminer un sommet y de M, voisin de x, tel que $f(y) > f(x)$. Soit v^δ le vecteur-colonne à $m-n$ composantes, tel que

$$v^{\delta(i)} = \begin{cases} 0 & \text{pour } i \neq k\,,\ 1 \leq i \leq m-n, \\ 1 & \text{pour } i = k. \end{cases}$$

Pour tout réel $t \geq 0$, on pose

$$\begin{cases} y^\gamma(t) = x^\gamma - t(A_\gamma)^{-1} A_\delta v^\delta\,, \\ y^\delta(t) = t v^\delta\,. \end{cases}$$

Ces relations déterminent un vecteur-colonne $y(t)$ à m composantes, si l'on pose, pour $1 \leq i \leq m$,

$$y^i(t) = \begin{cases} y^{\gamma(i)} & \text{si } i \in \gamma\,, \\ y^{\delta(i)} & \text{si } i \in \delta\,. \end{cases}$$

On a

$$\begin{aligned} Ay(t) &= A_\gamma y^\gamma(t) + A_\delta y^\delta(t) = A_\gamma x^\gamma = b\,, \\ f\big(y(t)\big) &= c_\gamma y^\gamma(t) + c_\delta y^\delta(t) \\ &= f(x) + t\Big(c_{\delta(k)} - \big(c_\gamma (A_\gamma)^{-1} A_\delta\big)_k\Big)\,. \end{aligned}$$

On voit que $y(t)$ est élément de M si et seulement si toutes ses composantes sont ≥ 0. Mais puisque les composantes de v^δ sont toutes ≥ 0 (celle d'indice k est égale à 1, les autres sont nulles) et que toutes les composantes de x^γ sont, par hypothèse, strictement positives, on voit que l'ensemble des réels $t \geq 0$ tels que $y(t) \in M$ est:

- la demi-droite $[0, +\infty[$ si toutes les composantes du vecteur-colonne (à n composantes) $(A_\gamma)^{-1} A_\delta v^\delta$ sont ≤ 0,
- un intervalle $[0, t_{max}]$, avec $t_{max} > 0$, si certaines composantes du vecteur-colonne $(A_\gamma)^{-1} A_\delta v^\delta$ sont > 0; dans ce cas, t_{max} est donné par

$$t_{max} = \inf_{i \in J} \frac{x^{\gamma(i)}}{\big((A_\gamma)^{-1} A_\delta v^\delta\big)^i}\,,$$

avec

$$J = \Big\{\, i \in \{1, \ldots, m\} \;\Big|\; \big((A_\gamma)^{-1} A_\delta v^\delta\big)^i > 0 \,\Big\}\,.$$

On peut énoncer:

5.1. Proposition. *On se place dans les hypothèse précisées ci-dessus.*

1. *Si toutes les composantes du vecteur-colonne $(A_\gamma)^{-1} A_\delta v^\delta$ sont ≤ 0, M n'est pas borné et la fonction f n'est pas majorée sur M; dans ce cas, le problème d'optimisation 3.1 n'a pas de solution.*

2. *Si une composante au moins du vecteur-colonne $(A_\gamma)^{-1} A_\delta v^\delta$ est > 0, le point $y(t_{max})$ est un sommet de M qui vérifie*

$$f\big(y(t_{max})\big) > f(x)\,.$$

Le sommet $y(t_{max})$ est non dégénéré si et seulement si l'ensemble des indices $i \in \{1, \ldots, n\}$ tels que $\big((A_\gamma)^{-1} A_\delta v^\delta\big)^i > 0$ et que

$$t_{max} = \frac{x^{\gamma(i)}}{\big((A_\gamma)^{-1} A_\delta v^\delta\big)^i}$$

a un seul élément.

Démonstration. Si toutes les composantes de $(A_\gamma)^{-1} A_\delta v^\delta$ sont ≤ 0, $y(t)$ appartient à M pour tout réel $t \geq 0$. L'ensemble M, contenant une demi-droite affine, n'est pas borné . D'autre part, l'expression de $f\big(y(t)\big)$ donnée ci-dessus montre que $f\big(y(t)\big)$ tend vers $+\infty$ lorsque $t \to +\infty$, donc que f n'est pas majorée sur M.

Si une composante au moins de $(A_\gamma)^{-1} A_\delta v^\delta$ est > 0, l'ensemble des $t \geq 0$ tels que $y(t) \in M$ est l'intervalle fermé borné $[0, t_{max}]$. L'expression de t_{max} donnée ci-dessus

montre que $t_{max} > 0$. Par suite, $y(t_{max})$ est élément de M et $f\big(y(t_{max})\big) > f(x)$. Montrons que $y(t_{max})$ est un sommet de M. D'après la proposition 2.5, il suffit de montrer que les vecteurs A_i, avec $i \in I\big(y(t_{max})\big)$, sont linéairement indépendants. On a posé

$$I\big(y(t_{max})\big) = \big\{\, i \in \{1, \ldots, m\} \;\big|\; y(t_{max})^i > 0 \,\big\}\,.$$

Soit j un élément de $\{1, \ldots, n\}$ tel que $\big((A_\gamma)^{-1} A_\delta v^\delta\big)^j > 0$ et que

$$\frac{x^{\gamma(j)}}{\big((A_\gamma)^{-1} A_\delta v^\delta\big)^j} = t_{max}\,,$$

(on sait qu'un tel élément j existe). On a alors $y^{\gamma(j)}(t_{max}) = 0$. Les composantes éventuellement non nulles de y sont donc $y^{\gamma(i)}$ pour $i \in \{1, \ldots, n\}$, $i \neq j$, et $y^{\delta(k)}(t_{max}) = t_{max} > 0$. Il suffit donc de prouver que les vecteurs $A_{\gamma(i)}$ ($i \in \{1, \ldots, n\}$, $i \neq j$) et $A_{\delta(k)}$ sont linéairement indépendants. Supposons qu'ils ne le soient pas. Il existe alors des scalaires $\lambda^{\gamma(i)}$ ($1 \leq i \leq n$, $i \neq j$) et $\lambda^{\delta(k)}$, non tous nuls, tels que

$$\sum_{1 \leq i \leq n,\ i \neq j} \lambda^{\gamma(i)} A_{\gamma(i)} + \lambda^{\delta(k)} A_{\delta(k)} = 0\,. \qquad (*)$$

Les vecteurs $A_{\gamma(i)}$ ($1 \leq i \leq n$) étant linéairement indépendants, $\lambda^{\delta(k)}$ est nécessairement non nul. On peut donc se ramener au cas où $\lambda^{\delta(k)} = 1$. D'autre part, on peut poser

$$\lambda^{\gamma(j)} = 0\,,$$

et on remarque que

$$A_{\delta(k)} = A_\delta v^\delta\,,$$

puisque $v^{\delta(i)} = 0$ si $i \neq k$ et $v^{\delta(k)} = 1$. L'égalité $(*)$ ci-dessus s'écrit donc, sous forme matricielle,

$$A_\gamma \lambda^\gamma + A_\delta v^\delta = 0\,.$$

En multipliant à gauche par $(A_\gamma)^{-1}$ (c'est possible, A_γ étant inversible) on en déduit

$$\lambda^\gamma + (A_\gamma)^{-1} A_\delta v^\delta = 0\,.$$

Comme on sait que $\lambda^{\gamma(j)} = 0$, on a donc

$$\big((A_\gamma)^{-1} A_\delta v^\delta\big)^j = 0\,,$$

ce qui est en contradiction avec la définition même de j. Les vecteurs $A_{\gamma(i)}$ ($1 \leq i \leq n$, $i \neq j$) et $A_{\delta(k)}$ sont donc linéairement indépendants, et $y(t_{max})$ est un sommet de M.

Enfin $y(t_{max})$ est non dégénéré si et seulement si le nombre de ses composantes strictement positives est n, c'est-à-dire si et seulement si

$$\Big\{\, i \in \{1, \ldots, n\} \;\Big|\; \big((A_\gamma)^{-1} A_\delta v^\delta\big)^i > 0 \quad \text{et} \quad t_{max} = x^{\gamma(i)} / \big((A_\gamma)^{-1} A_\delta v^\delta\big)^i \,\Big\}$$

a un seul élément. □

5.2. Remarques.

1. Partant d'un sommet non dégénéré x de M où la fonction f n'atteint pas son maximum, la proposition 5.1 permet, dans le cas où l'ensemble admissible M est borné, de déterminer un autre sommet y de M (précédemment noté $y(t_{max})$), tel que $f(y) > f(x)$. La proposition 4.2 permet alors de savoir si f atteint son maximum sur M au point y. Si ce n'est pas le cas, et si y est non dégénéré, on peut appliquer la construction décrite dans la proposition 5.1 pour déterminer un nouveau sommet de M où f prend une valeur strictement plus grande qu'en y. On répètera ces opérations autant de fois qu'il le faudra pour aboutir à un sommet où f atteint son maximum. Si M est borné, et si au cours des constructions effectuées on ne rencontre aucun sommet dégénéré, on est assuré d'aboutir après un nombre fini d'opérations. En effet, le nombre de sommets de M est fini et tant que le sommet courant (supposé non dégénéré) n'est pas un sommet où f atteint son maximum, la méthode indiquée dans la proposition 5.1 permet de déterminer un autre sommet où f prend une valeur strictement supérieure.

La suite de constructions décrite ci-dessus peut ne pas aboutir à la détermination d'un sommet où la fonction f atteint son maximum dans deux cas:

- lorsqu'on aboutit à un sommet x de M tel que, avec les notations de la proposition 5.1, toutes les composantes de $(A_\gamma)^{-1}A_\delta v^\delta$ soient ≤ 0; on sait alors que M n'est pas borné et que la fonction f n'est pas majorée sur M, donc que le problème d'optimisation étudié n'a pas de solution;
- lorsqu'on aboutit à un sommet de M dégénéré; on étudiera plus loin les difficultés liées aux sommets dégénérés, et on verra que la construction décrite ci-dessus peut être adaptée afin de ne pas être interrompue par la rencontre d'un sommet dégénéré.

2. Avec les notations de la proposition 5.1, le sommet $y(t_{max})$ est voisin du sommet x au sens suivant: il existe un seul indice $k \in \{1, \ldots, m\}$ tel que $y^k > 0$ et $x^k = 0$. Du point de vue géométrique, cela signifie que ces deux sommets sont les deux extrémités d'une *arête* du polytope M.

3. Pour effectuer les constructions indiquées dans la remarque 1 ci-dessus, on doit, pour chaque sommet x rencontré, déterminer la matrice $(A_\gamma)^{-1}A_\delta$, où on a noté $\gamma = I(x)$ l'ensemble des indices $i \in \{1, \ldots, m\}$ tels que $x^i > 0$, et δ le complémentaire de γ dans $\{1, \ldots, m\}$. On a supposé γ et δ ordonnés de manière quelconque. La proposition ci-après permet la détermination effective de cette matrice, pour chaque sommet rencontré.

5.3. Proposition. *Les hypothèses et notations étant celles de la proposition 5.1, soit* $j \in \{1, \ldots, n\}$ *tel que* $\left((A_\gamma)^{-1}A_\delta v^\delta\right)^j > 0$ *et que* $\dfrac{x^{\gamma(j)}}{\left((A_\gamma)^{-1}A_\delta v^\delta\right)^j} = t_{\max}$. *Soient* γ' *et* δ' *les applications injectives, respectivement de* $\{1, \ldots, n\}$ *et de* $\{1, \ldots, m-n\}$, *dans* $\{1, \ldots, m\}$ *définies par:*

$$\gamma'(r) = \begin{cases} \gamma(r) & \text{pour } r \neq j\,,\ 1 \leq r \leq n, \\ \delta(k) & \text{pour } r = j, \end{cases}$$

$$\delta'(i) = \begin{cases} \delta(i) & \text{pour } i \neq k\,,\ 1 \leq i \leq m-n, \\ \gamma(j) & \text{pour } i = k. \end{cases}$$

Alors les images de γ' et de δ' sont deux parties de $\{1,\dots,m\}$ complémentaires l'une de l'autre, et la matrice $A_{\gamma'}$ est inversible. Posons, pour alléger l'écriture,

$$\alpha = (A_\gamma)^{-1}A_\delta\,,\quad \beta = (A_{\gamma'})^{-1}A_{\delta'}\,.$$

Alors le coefficient α_k^j est non nul et les coefficients de β s'expriment, au moyen de ceux α, par les formules:

$$\beta_i^r = \begin{cases} \alpha_i^r - \dfrac{\alpha_i^j\alpha_k^r}{\alpha_k^j} & \text{pour } i\neq k\,,\ 1\leq i\leq m-n \text{ et } r\neq j\,,\ 1\leq r\leq n,\\ \dfrac{\alpha_i^j}{\alpha_k^j} & \text{pour } i\neq k\,,\ 1\leq i\leq m-n \text{ et } r=j,\\ \dfrac{1}{\alpha_i^j} & \text{pour } i=k \text{ et } r=j. \end{cases}$$

Démonstration. Les images de γ' et de δ' sont, par définition même, deux parties complémentaires de $\{1,\dots,m\}$, tout comme les images de γ et de δ. Les colonnes de la matrice $A_{\gamma'}$ sont les vecteurs-colonnes $A_{\gamma(i)}$ pour $1\leq i\leq n$, $i\neq j$, et le vecteur $A_{\delta(k)}$. On a vu, lors de la démonstration de la proposition 5.1, qu'ils sont linéairement indépendants. La matrice $A_{\gamma'}$ est donc inversible. Le coefficient α_k^j n'est autre que $\big((A_\gamma)^{-1}A_\delta v^\delta\big)^j$, puisque la seule composante non nulle de v^δ est $v^{\delta(k)}=1$.

Puisque $\alpha=(A_\gamma)^{-1}A_\delta$ et $\beta=(A_{\gamma'})^{-1}A_{\delta'}$, on a

$$A_\gamma\alpha = A_\delta\,,\quad A_{\gamma'}\beta = A_{\delta'}\,,$$

ou, en explicitant les produits matriciels,

$$\sum_{r=1}^n \alpha_i^r A^l_{\gamma(r)} = A^l_{\delta(i)}\,,\quad 1\leq l\leq n\,,\quad 1\leq i\leq m-n\,, \tag{$*$}$$

$$\sum_{r=1}^n \beta_i^r A^l_{\gamma'(r)} = A^l_{\delta'(i)}\,,\quad 1\leq l\leq n\,,\quad 1\leq i\leq m-n\,. \tag{$**$}$$

En tenant compte des expressions de γ' et de δ' au moyen de γ et δ, on peut mettre l'égalité $(*)$ sous les formes suivantes:

– pour $i=k$,

$$\sum_{1\leq r\leq n\,,\,i\neq j} \alpha_k^r A^l_{\gamma'(r)} + \alpha_k^j A^l_{\delta'(k)} = A^l_{\gamma'(j)}\,,$$

ou en divisant par α_k^jen en réordonnant les termes

$$\sum_{1\leq r\leq n\,,\,r\neq j}\left(-\frac{\alpha_k^r}{\alpha_k^j}\right)A^l_{\gamma'(r)} + \frac{1}{\alpha_k^j}A^l_{\gamma'(j)} = A^l_{\delta'(k)}\,,\quad 1\leq l\leq n\,; \tag{$***$}$$

– pour $i \neq k$, $1 \leq i \leq m-n$,

$$\sum_{1 \leq r \leq n\,,\, r \neq j} \alpha_i^r A^l_{\gamma'(r)} + \alpha_i^j A^l_{\delta'(k)} = A^l_{\delta'(i)}\,.$$

Remplaçons $A^l_{\delta'(i)}$ par son expression (***). On obtient:

$$\sum_{1 \leq r \leq n\,,\, r \neq j} \left(\alpha_i^r - \frac{\alpha_i^j \alpha_k^r}{\alpha_k^j} \right) A^l_{\gamma'(r)} + \frac{\alpha_i^j}{\alpha_k^j} A^l_{\gamma'(j)} = A^l_{\delta'(i)}\,, \quad 1 \leq i \leq m-n\,,\ i \neq k\,,\ 1 \leq l \leq n\,. \tag{****}$$

En identifiant (***) et (****) avec (**) on en déduit, la matrice $A_{\gamma'}$ étant inversible, les expressions des coefficients β_i^r indiquées dans l'énoncé. □

5.4. Remarques.

1. Les formules indiquées dans la proposition 5.3 s'obtiennent en remplaçant la base de $\mathbf{R}^n$ formée par les vecteurs $A_{\gamma(i)}$, par une autre base, formée par les vecteurs $A_{\gamma'(i)}$, $1 \leq i \leq n$. Ces deux bases ne diffèrent que par un seul vecteur: $A_{\gamma(j)}$ est remplacé par $A_{\delta(k)}$. Un changement de base de ce type est appelé *pivotement*, et le coefficient non nul α_k^j est appelé *pivot*.

2. L'image de γ' contient l'ensemble $I\big(y(t_{max})\big)$ des indices $i \in \{1, \ldots, m\}$ tels que $y^i(t_{max}) > 0$, et lui est égale lorsque le sommet $y(t_{max})$ est non dégénéré.

5.5. Exemple. Soit P l'ensemble des points de $\mathbf{R}^3$ dont les coordonnées x, y, z vérifient

$$x \geq 0\,, \quad y \geq 0\,, \quad z \geq 0\,,$$
$$x - y \leq 4\,, \quad 3x - 4y + 3z \leq 12\,, \quad z \leq 2\,, \quad y \leq 3\,.$$

On se propose de trouver un point de P où la fonction

$$f = x + y + 2z$$

atteint son maximum.

On reconnaît un problème d'optimisation linéaire de la forme décrite en 2.2.1. Afin de le ramener à la forme standard, on introduit des variables d'écart t_1, t_2, t_3, t_4. Soit M l'ensemble des points de $\mathbf{R}^7$ dont les coordonnées x, y, z, t_1, t_2, t_3, t_4 vérifient

$$x \geq 0\,, \quad y \geq 0\,, \quad z \geq 0\,, \quad t_1 \geq 0\,, \quad t_2 \geq 0\,, \quad t_3 \geq 0\,, \quad t_4 \geq 0\,, \tag{1}$$
$$x - y + t_1 = 4\,, \quad 3x - 4y + 3z + t_2 = 12\,, \quad z + t_3 = 2\,, \quad y + t_4 = 3\,. \tag{2}$$

On veut déterminer un point de M où la fonction

$$f = x + y + 2z \tag{3}$$

atteint son maximum. On a maintenant à résoudre un problème d'optimisation linéaire sous forme standard. Une fois celui-ci résolu, il suffira de projeter sur $\mathbf{R}^3$ le point

$(x, y, z, t_1, t_2, t_3, t_4)$ de M trouvé, c'est-à-dire de garder seulement ses trois premières coordonnées (x, y, z), pour obtenir un point de P où f atteint son maximum.

Dans le problème sous forme standard que nous avons à résoudre, $m = 7$, $n = 4$. La matrice A (à 4 lignes et 7 colonnes), le vecteur-colonne b (à 4 composantes) et le vecteur-ligne c (à 7 composantes) ont pour expressions:

$$A = \begin{pmatrix} 1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 3 & -4 & 3 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} ; \quad b = \begin{pmatrix} 4 \\ 12 \\ 2 \\ 3 \end{pmatrix} ; \quad c = (1 \quad 1 \quad 2 \quad 0 \quad 0 \quad 0 \quad 0) .$$

Le point m_0 de $\mathbf{R}^7$ de coordonnées $x = 0$, $y = 0$, $z = 0$, $t_1 = 4$, $t_2 = 12$, $t_3 = 2$, $t_4 = 3$ est élément de M. Les composantes non nulles de ce point sont celles d'indices 4, 5, 6 et 7. Les colonnes de la matrice A d'indices 4, 5, 6 et 7 étant linéairement indépendantes, la proposition 2.6 montre que m_0 est un sommet de M. C'est un sommet non dégénéré, car le nombre de ses composantes non nulles est 4. On prendra ce point comme point de départ pour l'application de la méthode du simplexe.

Compte tenu des inégalités (1), les troisième et quatrième équations (2) montrent que tout point de M a ses coordonnées z et t_3 comprises entre 0 et 2 et ses coordonnées y et t_4 comprises entre 0 et 3. La première équation (2) montre alors que ses coordonnées x et t_1 sont comprises entre 0 et 7. Enfin la seconde équation (2) montre que sa coordonnée t_2 est comprise entre 0 et 24. Ceci prouve que M est borné; c'est donc un polyèdre convexe.

Au paragraphe 4 , nous avions noté x le sommet non dégénéré utilisé comme point de départ, z un point courant de M et, au début du paragraphe 5, nous avions noté y un autre sommet obtenu par pivotement. Comme maintenant x, y et z désignent les trois premières coordonnées dans $\mathbf{R}^7$, nous noterons m_0 le sommet pris comme point de départ (c'est le point de coordonnées $x = y = z = 0$, $t_1 = 4$, $t_2 = 12$, $t_3 = 2$, $t_4 = 3$). Nous noterons m un point courant de M, m_1, m_2, ..., les sommets obtenus par pivotements successifs. Les autres notations seront les mêmes qu'au paragraphe 4.1.

La partie γ de $\{1, 2, 3, 4, 5, 6, 7\}$ associée au sommet non dégénéré m_0 est $\{4, 5, 6, 7\}$. La partie complémentaire δ est $\{1, 2, 3\}$. Les matrices A_γ et A_δ sont

$$A_\gamma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} , \quad A_\delta = \begin{pmatrix} 1 & -1 & 0 \\ 3 & -4 & 3 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} .$$

On a donc $(A_\gamma)^{-1} A_\delta = A_\delta$.

Les formules $(*)$ du paragraphe 4.1, qui expriment la partie m^γ des coordonnées d'un point courant m de M au moyen de la partie ${m_0}^\delta$ des coordonnées du sommet de départ m_0 et de la partie m^δ des coordonnées de m,

$$m^\gamma = {m_0}^\gamma - (A_\gamma)^{-1} A_\delta m^\delta ,$$

s'explicitent sous la forme suivante, $(x, y, z, t_1, t_2, t_3, t_4)$ désignant les coordonnées de m,

$$\begin{cases} t_1 = 4 - x + y , \\ t_2 = 12 - 3x + 4y - 3z , \\ t_3 = 2 - z , \\ t_4 = 3 - y . \end{cases} \tag{4}$$

Pour obtenir ces expressions, on peut bien entendu utiliser l'expression théorique (∗) du paragraphe 4.1 et remplacer $(A_\gamma)^{-1}A_\delta$, x^γ (maintenant noté ${m_0}^\gamma$) et z^δ (maintenant noté m^δ) par leurs expressions. Plus simplement, on peut remarquer que ces expressions s'obtiennent en résolvant par rapport à t_1, t_2, t_3 et t_4 les équations (2) ci-dessus qui définissent (avec les inégalités (1)) l'ensemble admissible M.

De même, la formule (∗∗) du paragraphe 4.1,

$$f(m) = f(m_0) + \big(c_\delta - c_\gamma(A_\gamma)^{-1}A_\delta\big)m^\delta\,,$$

qui exprime la valeur de la fonction f au point courant m au moyen de sa valeur au sommet de départ m_0 et de la partie m^δ des composantes de m, s'écrit maintenant

$$f = x + y + 2z\,.$$

C'est, tout simplement, l'expression (3) de f donnée ci-dessus.

Les composantes de $\big(c_\delta - c_\gamma(A_\gamma)^{-1}A_\delta\big)$ sont les coefficients de x, y et z dans l'expression de f ci-dessus, c'est-à-dire 1, 1 et 2. Elles sont strictement positives; la proposition 4.2 montre donc que f n'atteint pas son maximum au point m_0.

On effectue un pivotement en choisissant une des composantes strictement positives de $\big(c_\delta - c_\gamma(A_\gamma)^{-1}A_\delta\big)$, par exemple la troisième (qui vaut 2 et qui correspond au coeffficient de z dans l'expression de f). Ainsi qu'on l'a vu au début du paragraphe 5, le nouveau sommet de M auquel conduit ce pivotement s'obtient en conservant à x et à y les valeurs 0, en posant $z = t$ et en donnant à t la valeur la plus grande possible t_{max} pour laquelle t_1, t_2, t_3 et t_4 sont ≥ 0. Les équations (4) montrent que $t_{max} = 2$, car t_3 s'annule pour $z = 2$ (et deviendrait strictement négatif si z prenait une valeur strictement supérieure à 2). Le nouveau sommet m_1 auquel mène ce pivotement est donc le point de coordonnées $x = 0$, $y = 0$, $z = 2$, $t_1 = 4$, $t_2 = 6$, $t_3 = 0$, $t_4 = 3$. Les deux parties complémentaires γ' et δ' de $\{1, \ldots, 7\}$ définies dans la proposition 5.3 sont, respectivement, $\{3, 4, 5, 7\}$ et $\{1, 2, 6\}$. Pour calculer les coefficients de la matrice $(A_{\gamma'})^{-1}A_{\delta'}$, on pourrait bien entendu utiliser les formules théoriques données dans la proposition 5.3. Il est plus simple de remarquer que les coefficients de cette matrice apparaissent tout naturellement lorsqu'on résoud les équations (4) par rapport aux coordonnées z, t_1, t_2 et t_4, c'est-à-dire par rapport aux coordonnées d'indices éléments de γ'. On obtient ainsi

$$\begin{cases} z = 2 - t_3\,, \\ t_1 = 4 - x + y\,, \\ t_2 = 6 - 3x + 4y + 3t_3\,, \\ t_4 = 3 - y\,. \end{cases} \tag{5}$$

On a donc

$$(A_{\gamma'})^{-1}A_{\delta'} = \begin{pmatrix} -1 & 1 & 0 \\ -3 & 4 & 3 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}\,.$$

Quant à l'expression de f en un point courant m de M, au moyen de la partie $m^{\delta'}$ des coordonnées de ce point, on l'obtient en remplaçant, dans l'expression (3) de f, z par son expression au moyen de x, y et t_3 donnée par la première équation (5). On obtient

$$f = 4 + x + y - 2t_3\,. \tag{6}$$

Les coefficients de x et de y dans le second membre de l'expression ci-dessus sont > 0; donc d'après la proposition 4.2, f n'atteint pas son maximum sur M au sommet m_1. On doit effectuer encore un pivotement. On choisit cette fois le coefficient de x (qui vaut 1). On fait donc $y = t_3 = 0$, $x = t$, et on donne à t la valeur la plus grande pour laquelle z, t_1, t_2 et t_4, tirés des équations (5), sont ≥ 0. On voit que la plus grande valeur possible pour t est 2, car t_2 s'annule pour $x = 2$. Le nouveau sommet obtenu est donc le point m_2 de coordonnées $x = 2$, $y = 0$, $z = 2$, $t_1 = 2$, $t_2 = 0$, $t_3 = 0$, $t_4 = 3$. Les parties complémentaires γ'' et δ'' de $\{1, \ldots, 7\}$ sont donc $\{1, 3, 4, 7\}$ et $\{2, 5, 6\}$. Comme ci-dessus, au lieu de déterminer l'expression de la matrice $(A_{\gamma''})^{-1} A_{\delta''}$ au moyen des formules théoriques de la proposition 5.3, on résoud les équations (5) par rapport aux variables x, z, t_1, t_4. On obtient

$$\begin{cases} x = 2 + \dfrac{4}{3}y - \dfrac{1}{3}t_2 + t_3\,, \\ z = 2 - t_3\,, \\ t_1 = 2 - \dfrac{1}{3}y + \dfrac{1}{3}t_2 - t_3\,, \\ t_4 = 3 - y\,. \end{cases} \tag{7}$$

On exprime f en fonction des coordonnées y, t_2 et t_3 en remplaçant x dans (6) par l'expression donnée par la première équation (7). On obtient:

$$f = 6 + \frac{7}{3}y - \frac{1}{3}t_2 - t_3\,. \tag{8}$$

Le coefficient de y dans le membre de gauche de l'expression ci-dessus étant > 0, on doit faire encore un pivotement. On doit cette fois faire $t_2 = t_3 = 0$ et donner à y la valeur la plus grande possible pour laquelle x, z, t_1 et t_4, donnés par les expressions (7), sont ≥ 0. On voit que cette valeur de y, la plus grande possible, est 3, car pour cette valeur t_4 est nul. Le nouveau sommet m_3 obtenu par ce pivotement est le point de coordonnées $x = 6$, $y = 3$, $z = 2$, $t_1 = 1$, $t_2 = 0$, $t_3 = 0$, $t_4 = 0$.

On résoud (7) par rapport à x, y, z et t_1. On obtient

$$\begin{cases} x = 6 - \dfrac{1}{3}t_2 + t_3 - \dfrac{4}{3}t_4\,, \\ y = 3 - t_4\,, \\ z = 2 - t_3\,, \\ t_1 = 1 + \dfrac{1}{3}t_2 - t_3 + \dfrac{1}{3}t_4\,. \end{cases} \tag{9}$$

On exprime f au moyen de t_2, t_3, t_4 en remplaçant y dans (8) par son expression donnée par la seconde formule (9). On obtient

$$f = 13 - \frac{1}{3}t_2 - t_3 - \frac{7}{3}t_4\,. \tag{10}$$

Cette fois, les coefficients de t_2, t_3 et t_4 dans le membre de gauche de (10) sont tous < 0. La fonction f atteint donc son maximum sur M au point m_3, et la valeur de ce maximum est 13.

6. Pivotement à partir d'un sommet dégénéré

Les hypothèses et notations sont celles de 3.1. Soit x un sommet dégénéré de M, et $I(x)$ l'ensemble des indices $i \in \{1,\dots,m\}$ tels que $x^i \neq 0$. Le nombre d'éléments de $I(x)$ est strictement inférieur à n. Mais puisque les vecteurs-colonnes A_i, $i \in I(x)$, sont linéairement indépendants et que le rang de la matrice A est n, il existe une partie γ de $\{1,\dots,m\}$ comportant exactement n éléments, contenant $I(x)$ et telle que les vecteurs-colonnes A_i, $i \in \gamma$, soient linéairement indépendants. On peut ordonner γ de manière quelconque, donc la considérer comme une application injective de $\{1,\dots,n\}$ dans $\{1,\dots,m\}$. On note δ le complémentaire de γ dans $\{1,\dots,m\}$, que l'on ordonne de manière quelconque, et que l'on considère comme une application injective de $\{1,\dots,m-n\}$ dans $\{1,\dots,m\}$. On sait que la matrice A_γ (notations du paragraphe 1) est inversible. On supposera, dans ce qui suit, que le sommet x est explicitement connu, ainsi que γ, δ et la matrice $(A_\gamma)^{-1}A_\delta$.

Les calculs effectués au paragraphe 4 sont encore valables, et montrent que pour tout élément z de M, on a

$$z^\gamma = x^\gamma - (A_\gamma)^{-1}A_\delta z^\delta\,, \tag{$*$}$$

$$f(z) = f(x) + \big(c_\delta - c_\gamma(A_\gamma)^{-1}A_\delta\big)z^\delta\,. \tag{$**$}$$

Le critère de maximalité 4.1 n'est plus applicable. On peut cependant énoncer:

6.1. Proposition. *On se place dans les hypothèses précisées ci-dessus.*

1. *Si les $m-n$ composantes du vecteur-colonne $c_\delta - c_\gamma(A_\gamma)^{-1}A_\delta$ sont toutes ≤ 0, la fonction f atteint son maximum en x.*

2. *S'il existe un indice $k \in \{1,\dots,m-n\}$ tel que $c_{\delta(k)} - \big(c_\gamma(A_\gamma)^{-1}A_\delta\big)_k > 0$, et si de plus, pour tout $j \in \{1,\dots,n\}$ vérifiant $x^{\gamma(j)} = 0$, on a $\big((A_\gamma)^{-1}A_{\delta(k)}\big)^j \leq 0$, la fonction f n'atteint pas son maximum au point x.*

Démonstration.

1. Dans ces hypothèses, l'expression $(**)$ de $f(z)$ montre que f atteint son maximum en x, puisque toutes les composantes de z^δ sont ≥ 0.

2. Soit v^δ le vecteur-colonne à $m-n$ composantes défini par

$$v^{\delta(i)} = \begin{cases} 0 & \text{pour } i \neq k\,,\ 1 \leq i \leq m-n, \\ 1 & \text{pour } i = k. \end{cases}$$

Posons, pour tout réel $t \geq 0$,

$$\begin{cases} y^\gamma(t) = x^\gamma - t(A_\gamma)^{-1}A_\delta v^\delta\,, \\ y^\delta(t) = tv^\delta\,. \end{cases}$$

Ces relations déterminent un vecteur-colonne $y(t)$ à m composantes, si l'on pose, pour chaque i $(1 \leq i \leq m)$,

$$y^i(t) = \begin{cases} y^{\gamma(i)}(t) & \text{si } i \in \gamma\,, \\ y^{\delta(i)}(t) & \text{si } i \in \delta\,. \end{cases}$$

Des calculs identiques à ceux du paragraphe 5 montrent alors que

$$Ay(t) = b\,,$$
$$f\big(y(t)\big) = f(x) + t\Big(c_{\delta(k)} - \big(c_\gamma (A_\gamma)^{-1} A_\delta\big)_k\Big)\,.$$

Pour $t > 0$ et assez petit, on voit que $y(t)$ est élément de M car toutes ses composantes sont ≥ 0, et que $f\big(y(t)\big) > f(x)$. □

6.2. Proposition. *On se place dans les hypothèses de 6.1.2. Soit $k \in \{1, \ldots, m-n\}$ tel que $c_{\delta(k)} - \big(c_\gamma (A_\gamma)^{-1} A_\delta\big)_k > 0$ et que, pour tout $j \in \{1, \ldots, n\}$ vérifiant $x^{\gamma(j)} = 0$, on ait $\big((A_\gamma)^{-1} A_{\delta(k)}\big)^j \leq 0$. On définit v^δ et, pour tout $t \geq 0$, $y(t)$, comme dans la démonstration de 6.1.2.*

1. *Si les n composantes du vecteur-colonne $(A_\gamma)^{-1} A_\delta v^\delta$ sont toutes ≤ 0, M n'est pas borné, la fonction f n'est pas majorée sur M et le problème d'optimisation considéré n'a pas de solution.*

2. *Si l'ensemble d'indices*

$$J = \big\{\, i \in \{1, \ldots, n\} \;\big|\; \big((A_\gamma)^{-1} A_\delta v^\delta\big)^i > 0 \,\big\}$$

est non vide, posons

$$t_{max} = \inf_{i \in J} \frac{x^{\gamma(i)}}{\big((A_\gamma)^{-1} A_\delta v^\delta\big)^i}\,.$$

Le point $y(t_{max})$ est alors un sommet de M qui vérifie $f\big(y(t_{max})\big) > f(x)$.

Démonstration. Elle est pratiquement identique à celle de 5.1. Le lecteur fera lui-même aisément les adaptations nécessaires. □

6.3. Remarques. Supposons qu'on applique la méthode du simplexe pour résoudre le problème d'optimisation linéaire 3.1, et qu'à une certaine étape, on rencontre un sommet x dégénéré. Plusieurs cas sont alors possibles.

1. Si ce sommet, et la partie ordonnée γ de $\{1, \ldots, m\}$ qui lui correspond, vérifient les hypothèses de 6.1.1, on a résolu le problème puisque f atteint son maximum en x.

2. Si ce sommet, et la partie ordonnée γ de $\{1, \ldots, m\}$ qui lui correspond, vérifient les hypothèses de 6.1.2 et de 6.2.1, on sait que le problème considéré n'a pas de solution, f n'étant pas majorée sur M.

3. Si ce sommet et la partie γ qui lui correspond vérifient les hypothèses de 6.1.2 et de 6.2.2, cette dernière proposition permet de déterminer un autre sommet $y = y(t_{max})$ de M vérifiant $f(y) > f(x)$. De plus, la proposition 5.3 reste applicable et permet la détermination effective de la matrice $(A_{\gamma'})^{-1} A_{\delta'}$ associée au nouveau sommet y. On pourra donc poursuivre l'application de la méthode, comme dans le cas où tous les sommets rencontrés sont non dégénérés.

4. On peut cependant rencontrer aussi le cas où les hypothèses de 6.1.1 et 6.1.2 ne sont pas vérifiées, c'est-à-dire le cas où il existe un indice $k \in \{1, \dots, m-n\}$, pas nécessairement unique, tel que

$$c_{\delta(k)} - \left(c_\gamma (A_\gamma)^{-1} A_\delta\right)_k > 0\,,$$

mais que, pour chacun de ces indices k, il existe un indice $j \in \{1, \dots, n\}$ vérifiant

$$x^{\gamma(j)} = 0 \quad \text{et} \quad \left((A_\gamma)^{-1} A_\delta\right)^j > 0\,.$$

Supposons ces indices k et j choisis, et voyons ce qui se passe lorsqu'on essaie d'appliquer la méthode de pivotement décrite au paragraphe 5. Posons, comme dans la démonstration de 6.1,

$$v^{\gamma(i)} = \begin{cases} 0 & \text{pour } i \neq k\,,\ 1 \leq i \leq m-n\,, \\ 1 & \text{pour } i = k\,, \end{cases}$$

puis, pour tout $t \geq 0$,

$$y^\gamma(t) = x^\gamma - t(A_\gamma)^{-1} A_\delta v^\delta\,,$$
$$y^\delta(t) = t v^\delta\,.$$

Enfin, pour chaque indice i $(1 \leq i \leq m)$, posons

$$y^i(t) = \begin{cases} y^{\gamma(i)}(t) & \text{si } i \in \gamma\,, \\ y^{\delta(i)}(t) & \text{si } i \in \delta\,. \end{cases}$$

L'ensemble des réels $t \geq 0$ tels que $y(t) \in M$ est maintenant réduit à $\{0\}$, de sorte qu'on doit poser

$$t_{max} = 0\,.$$

Par suite, le sommet $y(t_{max})$, qui doit remplacer x, coïncide avec x. Cependant, la suite ordonnée γ' associée à ce sommet, et la suite complémentaire δ', ne coïncident pas avec γ et δ, respectivement. Elles sont données, comme dans la proposition 5.3, par

$$\gamma'(r) = \begin{cases} \gamma(r) & \text{pour } r \neq j\,,\ 1 \leq r \leq n, \\ \delta(k) & \text{pour } r = j, \end{cases}$$
$$\delta'(i) = \begin{cases} \delta(i) & \text{pour } i \neq k\,,\ 1 \leq i \leq m-n, \\ \gamma(j) & \text{pour } i = k. \end{cases}$$

On vérifie aisément que les vecteurs A_i, $i \in \gamma'$, sont linéairement indépendants et que les formules donnant l'expression de la matrice $(A_{\gamma'})^{-1} A_{\delta'}$, indiquées dans la proposition 5.3, restent valables. Le "pivotement" consistera alors à remplacer γ par γ', δ par δ', le sommet x restant inchangé. On pourra alors être dans l'un des cas envisagés en 1, 2 ou 3 ci-dessus, et poursuivre l'application de la méthode du simplexe. Si l'on est encore dans le cas où les hypothèses de 6.1.1 et 6.1.2 ne sont pas vérifiées, on effectuera un nouveau "pivotement", consistant à remplacer γ' par γ'', δ' par δ'', le sommet x restant toujours inchangé. On répètera cette opération autant de fois qu'il sera nécessaire pour que la situation se débloque. On devra veiller à ce que les parties à n éléments $\gamma, \gamma', \dots, \gamma^{(p)}, \dots$ contenant $I(x)$ successivement choisies ne forment pas un "cycle", c'est-à-dire soient toutes distinctes, sans quoi l'algorithme construit risquerait de se poursuivre indéfiniment (en "tournant en rond") sans jamais aboutir.

7. La détermination d'un sommet

Pour mettre en œuvre effectivement la méthode du simplexe, il reste encore à montrer comment trouver un sommet particulier de l'ensemble admissible M, et déterminer explicitement la matrice correspondante $(A_\gamma)^{-1}A_\delta$ (pour un choix de γ et de δ adapté à ce sommet). On considèrera tout d'abord un cas particulier.

7.1. Un cas particulier. Considérons le problème d'optimisation linéaire décrit au paragraphe 2.2. Les éléments A_1 de $\mathcal{L}(\mathbf{R}^p, \mathbf{R}^n)$, b de $\mathbf{R}^n$, c_1 de $(\mathbf{R}^p)^* = \mathcal{L}(\mathbf{R}^p, \mathbf{R})$ sont donnés, et on doit déterminer un élément x_1 de $\mathbf{R}^p$ vérifiant

$$A_1 x_1 \leq b\,, \qquad x_1 \geq 0\,,$$

et rendant maximum la fonction

$$f_1(x_1) = c_1 x_1\,.$$

Ainsi qu'on l'a vu en 2.2, on ramène ce problème à la forme standard en introduisant la variable d'écart $x_2 \in \mathbf{R}^n$ et en posant

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbf{R}^{n+p}\,, \quad Ax = A_1 x_1 + x_2\,.$$

Le problème devient alors: trouver $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbf{R}^{n+p}$ vérifiant

$$Ax = b\,, \quad x \geq 0\,,$$

et rendant maximum la fonction

$$f(x) = c_1 x_1\,.$$

Supposons $b \geq 0$. Le point $x = \begin{pmatrix} 0 \\ b \end{pmatrix} \in \mathbf{R}^{n+p}$ est alors un sommet de l'ensemble admissible du problème. Ce sommet est non dégénéré si et seulement si toutes les composantes de b sont > 0. Prenons $\gamma = \{p+1, \ldots, p+n\}$, $\delta = \{1, \ldots, p\}$, et ordonnons ces parties de $\{1, \ldots, p+n\}$ selon l'ordre naturel. La matrice A_γ est alors la matrice unité (à n lignes et n colonnes) et les matrices A_δ et $(A_\gamma)^{-1}A_\delta$ coïncident avec la matrice A_1. On a donc toutes les données nécessaires pour appliquer la méthode du simplexe.

7.2. Cas général. Revenons au problème d'optimisation linéaire sous forme standard, exposé aux paragraphes 2.1 et 3.1. Afin de déterminer un sommet de l'ensemble admissible M, on peut procéder comme suit.

On commence par changer en leurs opposés les composantes de b strictement négatives, ainsi que les lignes de même indice de la matrice A. On se ramène ainsi au cas où toutes les composantes de b sont ≥ 0.

On introduit alors la variable auxiliaire $y \in \mathbf{R}^n$, et on étudie le problème d'optimisation linéaire suivant: trouver $\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbf{R}^{m+n}$ vérifiant

$$Ax + y = b\,, \qquad x \geq 0\,, \qquad y \geq 0\,,$$

et rendant maximum la fonction

$$g = -\sum_{i=1}^{n} y^i \,.$$

On appellera ce problème "problème auxiliaire" pour le distinguer de celui initialement donné. Si l'ensemble admissible du problème initial

$$M = \left\{ x \in \mathbf{R}^n \;\middle|\; Ax = b\,, \quad x \geq 0 \right\}$$

est non vide, l'ensemble admissible du problème auxiliaire

$$M' = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbf{R}^{m+n} \;\middle|\; Ax + y = b\,,\ x \geq 0\,,\ y \geq 0 \right\}$$

contient $M \times \{0\}$ (où $\{0\}$ désigne l'origine de $\mathbf{R}^n$); le maximum de la fonction g sur M' est 0, et il est atteint en tout point de $M \times \{0\}$. D'autre part, M' admet pour sommet le point $\begin{pmatrix} 0 \\ b \end{pmatrix}$. On peut donc appliquer la méthode du simplexe, en prenant ce point pour point de départ, pour résoudre le problème d'optimisation auxiliaire. On obtiendra ainsi un sommet de M', nécessairement de la forme $\begin{pmatrix} x \\ 0 \end{pmatrix}$, car en ce point g atteint son maximum (qui est nul). Le point correspondant x de $\mathbf{R}^m$ est un sommet de M, que l'on pourra utiliser comme point de départ pour l'application de la méthode du simplexe au problème initial.

On observera que le problème d'optimisation auxiliaire est du type particulier étudié en 7.1, et qu'en déterminant le sommet x de M par la méthode indiquée ci-dessus, on obtiendra en même temps l'expression de la matrice correspondante $(A_\gamma)^{-1}A_\delta$.

8. Disposition des calculs en tableaux

Lorsqu'on travaille à la main (ce qui est possible tant que n et m ne sont pas trop grands) la tâche peut être sensiblement facilitée par une disposition astucieuse des calculs, sous forme de tableaux. Cette disposition peut également aider à mieux comprendre le fonctionnement de la méthode du simplexe, et constituer un guide pour la réalisation d'un programme d'ordinateur effectuant le travail automatiquement.

Nous allons exposer cette disposition en tableaux sur l'exemple traité au paragraphe 5.5. Après introduction des variables d'écart t_1, t_2, t_3 et t_4, on était parvenu au problème suivant. On note M l'ensemble des points de $\mathbf{R}^7$ dont les coordonnées x, y, z, t_1, t_2, t_3, t_4 vérifient

$$\begin{gathered} x \geq 0\,, \quad y \geq 0\,, \quad z \geq 0\,, \quad t_1 \geq 0\,, \quad t_2 \geq 0\,, \quad t_3 \geq 0\,, \quad t_4 \geq 0\,, \\ x - y + t_1 = 4\,, \quad 3x - 4y + 3z + t_2 = 12\,, \quad z + t_3 = 2\,, \quad y + t_4 = 3\,. \end{gathered}$$

On veut déterminer un point de M où la fonction

$$f = x + y + 2z$$

atteint son maximum.

Le tableau ci-dessous regroupe les données du problème.

x	y	z	t_1	t_2	t_3	t_4	
1	-1	0	1	0	0	0	4
3	-4	3	0	1	0	0	12
0	0	1	0	0	1	0	2
0	1	0	0	0	0	1	3
1	1	2	0	0	0	0	f
			$*$	$*$	$*$	$*$	

La ligne supérieure de ce tableau (que nous numéroterons 0) indique simplement le nom des variables x, y, z, t_1, t_2, t_3, t_4. Lorsqu'on sera suffisamment familiarisé avec la disposition en tableaux, on pourra l'omettre car elle ne changera pas tout au long des opérations qui seront effectuées.

Les cinq lignes suivantes (que nous numéroterons 1, 2, 3, 4 et 5) contiennent, dans les sept colonnes de gauche, les coefficients des variables x, y, z, t_1, t_2, t_3 et t_4, dans le système d'équations linéaires

$$\begin{array}{rrrrrrrcl}
x & -y & & +t_1 & & & & = & 4, \\
3x & -4y & +3z & & +t_2 & & & = & 12, \\
 & & z & & & +t_3 & & = & 2, \\
 & y & & & & & +t_4 & = & 3; \\
x & +y & +2z & & & & & = & f.
\end{array}$$

Parmi ces cinq lignes, on a séparé par un trait horizontal les quatre premières de la dernière, car leurs natures sont différentes: les lignes 1 à 4 correspondent aux quatre équations linéaires qui servent à définir le polytope M, tandis que la ligne 5 correspond à l'expression de la fonction à maximiser f. Dans la dernière colonne des lignes 1 à 4 on a fait figurer les membres de droite des équations linéaires qui servent à définir M. Dans la dernière colonne de la ligne 5, on a mis f pour rappeler que cette ligne donne l'expression de la fonction f.

Enfin la dernière ligne contient seulement des marques $*$ placées dans certaines colonnes. On va voir tout de suite lesquelles et pourquoi.

On sait que le polytope M comporte un sommet en évidence, de coordonnées $x = 0$, $y = 0$, $z = 0$, $t_1 = 4$, $t_2 = 6$, $t_3 = 2$, $t_4 = 3$, qui sert de point de départ dans l'application de la méthode du simplexe. On notera S_1 ce sommet. On remarque que les colonnes qui sont marquées d'un signe $*$ dans la dernière ligne sont précisément celles qui correspondent aux coordonnées non nulles t_1, t_2, t_3 et t_4 de S_1. On remarque aussi que les coefficients de la cinquième ligne du tableau (celle qui donne l'expression de f) qui sont dans les colonnes non marquées par un signe $*$ sont nuls. Cela restera vrai pour tous les tableaux qui seront calculés successivement.

Pour savoir si la fonction f atteint son maximum sur M en ce sommet, il suffit de regarder le signe des coefficients qui figurent dans les colonnes non marquées d'un $*$ de la cinquième ligne du tableau. Si un au moins de ces coefficients est > 0, la proposition 4.2 permet d'affirmer que f n'atteint pas son maximum en ce sommet. En effet, ces coefficients sont tout

simplement les composantes du vecteur-ligne noté, dans le paragraphe4, $c_\delta - c_\gamma (A_\gamma)^{-1} A_\delta$. Dans le cas présent, ces coefficients sont 1, 2 et 1; ils sont > 0, donc f n'atteint pas son maximum au sommet S_1. On comprend d'ailleurs immédiatement pourquoi, en remarquant que cette cinquième ligne ne fait qu'exprimer la relation

$$x + y + 2z = f\,;$$

comme x, y et z sont nuls au sommet S_1, on peut en augmentant une de ces coordonnées, faire croître la valeur de f, tout en restant dans M.

On choisit une des colonnes dont le coefficient dans la cinquième ligne est > 0. Par exemple, la colonne correspondant à la variable z. Effectuer un pivotement consiste à augmenter le plus possible la valeur de la variable z (les autres variables x et y dont les colonnes ne sont pas marquées d'un $*$ restant nulles). Si aucun des coefficients situés dans les quatre premières lignes de cette colonne n'était > 0, la proposition 5.1 montrerait que M n'est pas borné et que f peut prendre des valeurs arbitrairement grandes: en effet ces coefficients sont les composantes du vecteur-colonne que nous avions noté, dans cette proposition, $(A_\gamma)^{-1} A_\delta v^\delta$. Dans notre exemple ce n'est pas le cas: deux de ces coefficients, ceux de la seconde et de la troisième ligne, sont > 0. Chacune des lignes correspondantes indique une valeur que z ne doit pas dépasser. La seconde ligne exprime en effet l'égalité

$$3x - 4y + 3z + t_2 = 12\,.$$

Les variables x et y restant nulles, on fait croître z, depuis la valeur 0, et simultanément décroître t_2 depuis la valeur 12 de manière telle que cette égalité reste toujours satisfaite. Puisque t_2 doit rester ≥ 0, on voit que z doit rester ≤ 4. De même, la troisième ligne exprime l'égalité

$$z + t_3 = 2\,.$$

Comme t_3 doit rester ≥ 0, z doit rester ≤ 2.

On prend la plus petite des deux valeurs ainsi trouvées: c'est 2, et la ligne qui a servi à déterminer cette valeur maximale est la troisième. Le coefficient qui se trouve dans cette troisième ligne et dans la colonne de z est notre *pivot*. Dans notre exemple, sa valeur est 1. Le nouveau sommet S_2 auquel on va aboutir après pivotement aura z pour composante non nulle, et t_3 pour composante nulle. Dans la dernière ligne du tableau, on effacera donc le signe $*$ de la colonne de t_3, et on en placera un dans la colonne de z. Pour obtenir les coefficients du nouveau tableau, on retranche, de chaque ligne du tableau autre que celle du pivot (dans cet exemple, la troisième) un multiple convenablement choisi de la ligne du pivot, afin d'annuler tous les coefficients situés dans la colonne du pivot, autres que le pivot lui-même. Ainsi, dans notre exemple, on laisse la première ligne inchangée car son coefficient dans la colonne du pivot est nul; on retranche de la seconde ligne 3 fois la troisième; on laisse inchangée la troisième ligne (celle du pivot), ainsi que la quatrième puisque son coefficient dans la colonne du pivot est déjà nul. On retranche enfin de la

cinquième ligne 2 fois la troisième ligne. On obtient ainsi le tableau

x	y	z	t_1	t_2	t_3	t_4	
1	-1	0	1	0	0	0	4
3	-4	0	0	1	-3	0	6
0	0	1	0	0	1	0	2
0	1	0	0	0	0	1	3
1	1	0	0	0	-2	0	$f-4$
		*	*	*		*	

Ce tableau ne fait qu'exprimer le système d'équations linéaires, obtenu à partir du précédent en retranchant des autres équations des multiples convenablement choisis de la troisième équation,

$$\begin{array}{rrrrrrrl} x & -y & & +t_1 & & & = & 4, \\ 3x & -4y & & & +t_2 & -3t_3 & = & 6, \\ & & z & & & +t_3 & = & 2, \\ & y & & & & & +t_4 = & 3; \\ x & +y & & & & -2t_3 & = & f-4. \end{array}$$

Le nouveau sommet S_2 auquel nous avons abouti est le point de coordonnées $x=0$, $y=0$, $z=2$, $t_1=4$, $t_2=6$, $t_3=0$, $t_4=3$. On remarque qu'on obtient ces valeurs simplement en regardant les colonnes marquées d'un $*$; chacune d'elles contient un seul coefficient non nul, en fait égal à 1, et le coefficient situé dans la même ligne que ce coefficient non nul, dans la colonne de droite, donne la valeur de la coordonnée correspondante.

Partant de ce tableau, on effectue les mêmes opérations que celles faites précédemment. On trouve successivement les tableaux suivants.

x	y	z	t_1	t_2	t_3	t_4	
0	1/3	0	1	$-1/3$	1	0	2
1	$-4/3$	0	0	1/3	-1	0	2
O	0	1	0	0	1	0	2
0	1	0	0	0	0	1	3
0	7/3	0	0	$-1/3$	-1	0	$f-6$
*		*	*			*	

A cette étape, le sommet S_3 auquel on est arrivé est le point de coordonnées $x=2$, $y=0$, $z=2$, $t_1=2$, $t_2=0$, $t_3=0$, $t_4=3$; la valeur de f en ce point est 6. On remarque qu'on a divisé la logne du pivot par 3, afin de donner au pivot (qui est le coefficient sitié colonne 1, ligne 2) la valeur 1. La ligne 5 comportant un coefficient 7/3, strictement positif, dans

la colonne de y, un nouveau pivotement est nécessaire et conduit au tableau

x	y	z	t_1	t_2	t_3	t_4	
0	0	0	1	$-1/3$	1	$-1/3$	1
1	0	0	0	$1/3$	-1	$4/3$	6
O	0	1	0	0	1	0	2
0	1	0	0	0	0	1	3
0	0	0	0	$-1/3$	-1	$-7/3$	$f - 13$
		*	*	*		*	

Le sommet S_4 auquel on est arrivé est le point de coordonnées $x = 6$, $y = 3$, $z = 2$, $t_1 = 1$, $t_2 = 0$, $t_3 = 0$, $t_4 = 0$; la valeur de f en ce point est 13. La dernière ligne du tableau ayant, dans ses sept premières colonnes, des coefficients tous ≤ 0, la fonction f atteint son maximum sur M au point S_4.

Chapitre II

Méthodes directes

1. Quelques remarques sur les systèmes linéaires

1.1 Systèmes linéaires. Une partie du présent chapitre traite de la résolution de systèmes linéaires de la forme

$$\begin{aligned} a_1^1 x^1 + a_2^1 x^2 + \cdots + a_n^1 x^n &= b^1 \,, \\ a_1^2 x^1 + a_2^2 x^2 + \cdots + a_n^2 x^n &= b^2 \,, \\ \cdots \qquad & \cdots \\ a_1^m x^1 + a_2^m x^2 + \cdots + a_n^m x^n &= b^m \,. \end{aligned} \tag{1}$$

Dans ce système les inconnues sont les x^i, $1 \leq i \leq n$, tandis que les a_i^j et les b^j, $1 \leq i \leq n$, $1 \leq j \leq m$, sont des scalaires donnés. On supposera dans la suite que le corps des scalaires considéré est le corps des réels $\mathbf{R}$, mais on pourrait traiter de même le cas où c'est le corps des complexes $\mathbf{C}$.

1.2. Écriture matricielle. On écrit souvent ce système sous forme matricielle

$$AX = B \,, \tag{2}$$

avec

$$A = \begin{pmatrix} a_1^1 & a_2^1 & \dots & a_n^1 \\ a_1^2 & a_2^2 & \dots & a_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^m & a_2^m & \dots & a_n^m \end{pmatrix} , \qquad X = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^n \end{pmatrix} , \qquad B = \begin{pmatrix} b^1 \\ b^2 \\ \vdots \\ b^m \end{pmatrix} .$$

On dit que X est un vecteur (on précise parfois vecteur-colonne) élément de $\mathbf{R}^n$, B un vecteur élément de $\mathbf{R}^m$ et A une matrice $m \times n$ (à m lignes et n colonnes). On considèrera A comme un élément de l'espace $\mathcal{L}(\mathbf{R}^n, \mathbf{R}^m)$ des applications linéaires de $\mathbf{R}^n$ dans $\mathbf{R}^m$; on identifiera en effet systématiquement les applications linéaires entre espaces numériques et les matrices qui les représentent.

1.3. Point de vue géométrique. L'application linéaire $A \in \mathcal{L}(\mathbf{R}^n, \mathbf{R}^m)$ et le vecteur $B \in \mathbf{R}^m$ étant donnés, résoudre le système linéaire (2) consiste à trouver le (ou les) vecteur(s) de $\mathbf{R}^n$ que l'application A envoie sur B. On rappelle que l'image $A(\mathbf{R}^n)$ de

A (ensemble des vecteurs AX, lorsque X parcourt $\mathbf{R}^n$) est un sous-espace vectoriel de $\mathbf{R}^m$, dont la dimension p est appelée *rang* de l'application linéaire A. Ce sous-espace est engendré par les vecteurs formés par les colonnes de la matrice A, puisque ceux-ci sont les images (par A) des vecteurs de la base canonique de $\mathbf{R}^n$. On rappelle également que le noyau $\ker A$ de l'application A (ensemble des $X \in \mathbf{R}^n$ tels que $AX = 0$) est un sous-espace vectoriel de $\mathbf{R}^n$, et qu'on a entre le rang de A, la dimension n de l'espace sur lequel A est définie, et la dimension du noyau de A, la relation

$$n = p + \dim \ker A \,,$$

ce qui implique, puisque $\dim \ker A \geq 0$,

$$p \leq n \,.$$

D'autre part, le rang p de A vérifie évidemment

$$p \leq m \,,$$

puisque l'image de A est un sous-espace de $\mathbf{R}^m$. Par suite,

$$p \leq \inf(m, n) \,.$$

Ces quelques considérations géométriques rendent très facile la discussion de l'existence de solutions du système linéaire (2):

– Supposons le rang p de A égal à la dimension m de l'espace dans lequel A prend ses valeurs, ce qui implique $n \geq m$. Le système (2) admet alors des solutions quel que soit le choix du vecteur $B \in \mathbf{R}^m$.

– Supposons le rang p de A strictement inférieur à la dimension m de l'espace dans lequel A prend ses valeurs. Le système (2) admet alors des solutions si et seulement si le vecteur B est élément de l'image de A.

– Dans un cas comme dans l'autre, lorsque l'ensemble des solutions de (2) est non vide, cet ensemble est un sous-espace affine de $\mathbf{R}^n$, se déduisant du noyau de A par une translation, donc de même dimension que ce noyau. Il comporte un élément unique si et seulement si le noyau de A est réduit à $\{0\}$, c'est-à-dire si et seulement si A est injective. On reconnaît qu'on est dans ce cas lorsque les vecteurs formés par les colonnes de A sont linéairement indépendants.

– En particulier, lorsqu'on a $p = n = m$, l'application A est un isomorphisme de $\mathbf{R}^n$ sur lui-même. Pour chaque choix du vecteur $B \in \mathbf{R}^n$, le système (2) admet alors une solution unique. On reconnaît qu'on est dans ce cas lorsque la matrice A est carrée ($n = m$) et de déterminant non nul.

1.4. Systèmes augmentés. Dans certaines applications on doit résoudre plusieurs systèmes linéaires de la forme (2), la matrice A étant la même pour tous ces systèmes, tandis que le vecteur $B \in \mathbf{R}^m$ prend plusieurs valeurs différentes. Il est commode de

considérer tous ces systèmes comme formant un seul système, dont l'inconnue est exprimée sous forme matricielle. Les algorithmes que nous étudierons plus loin permettront en effet de les résoudre simultanément.

Ainsi par exemple, le système

$$\begin{pmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ a_1^3 & a_2^3 & a_3^3 \end{pmatrix} \begin{pmatrix} x_1^1 & x_2^1 & y_1^1 & y_2^1 & y_3^1 \\ x_1^2 & x_2^2 & y_1^2 & y_2^2 & y_3^2 \\ x_1^3 & x_2^3 & y_1^3 & y_2^3 & y_3^3 \end{pmatrix} = \begin{pmatrix} b_1^1 & b_2^1 & 1 & 0 & 0 \\ b_1^2 & b_2^2 & 0 & 1 & 0 \\ b_1^3 & b_2^3 & 0 & 0 & 1 \end{pmatrix} \tag{3}$$

est équivalent à la famille de systèmes

$$\begin{pmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ a_1^3 & a_2^3 & a_3^3 \end{pmatrix} \begin{pmatrix} x_i^1 \\ x_i^2 \\ x_i^3 \end{pmatrix} = \begin{pmatrix} b_i^1 \\ b_i^2 \\ b_i^3 \end{pmatrix}, \qquad 1 \leq i \leq 2\,, \tag{4}$$

et

$$\begin{pmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ a_1^3 & a_2^3 & a_3^3 \end{pmatrix} \begin{pmatrix} y_1^1 & y_2^1 & y_3^1 \\ y_1^2 & y_2^2 & y_3^2 \\ y_1^3 & y_2^3 & y_3^3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{5}$$

qui exprime que la matrice $Y = (y_j^i)$ est l'inverse de la matrice $A = (a_j^i)$.

On remarque qu'après avoir calculé l'inverse Y de la matrice A, on peut aisément obtenir la solution du système (2) pour tout vecteur B, puisqu'elle est donnée par

$$X = YB\,. \tag{6}$$

On peut donc se demander s'il est bien utile de résoudre un système tel que (3), dont la solution donne, à la fois, l'inverse de A et la solution de (2) pour plusieurs valeurs différentes de B. En pratique, la précision obtenue pour la résolution de (2) par calcul de Y puis application de la formule (6) est en général moins bonne que celle donnée par une résolution directe de (2), car les imprécisions lors du calcul de Y sont amplifiées lorsqu'on fait le produit matriciel de Y par B à droite. C'est pourquoi la résolution de systèmes de la forme (3) est utile.

On considèrera donc, dans la suite, des systèmes tels que (3), de la forme

$$AX = B \tag{7}$$

formellement identique à celle du système (2), où B et X sont à valeurs dans des espaces de matrices, et non plus dans $\mathbf{R}^m$ et $\mathbf{R}^n$, respectivement: $B \in \mathcal{L}(\mathbf{R}^p, \mathbf{R}^m)$ est une matrice $m \times p$, $X \in \mathcal{L}(\mathbf{R}^p, \mathbf{R}^n)$ est une matrice $n \times p$, tandis que $A \in \mathcal{L}(\mathbf{R}^n, \mathbf{R}^m)$ est une matrice $m \times n$.

1.5. Effet de quelques transformations. Considérons le système (7), dans lequel A, B et X sont des matrices, respectivement, $m \times n$, $m \times p$ et $n \times p$, l'inconnue étant X.

1. Remplaçons la matrice A par une autre matrice A' dont la i-ème ligne A'^i est une combinaison linéaire des lignes de A, de la forme

$$A'^i = \sum_j \lambda_j A^j\,,$$

les autres lignes restant inchangées,

$$A'^j = A^j \qquad \text{pour} \quad j \neq i\,.$$

Simultanément, remplaçons la matrice B par B', dont la i-ème ligne B'^i est donnée par la combinaison linéaire de lignes de B de mêmes coefficients λ_j,

$$B'^i = \sum_j \lambda_j B^j\,,$$

les autres lignes restant inchangées,

$$B'^j = B^j \qquad \text{pour} \quad j \neq i\,.$$

On voit aisément que toute solution X de (7) est aussi solution de

$$A'X = B'\,. \tag{8}$$

Réciproquement, si $\lambda_i \neq 0$, les transformations qui font passer de A à A' et de B à B' sont inversibles, de sorte que toute solution X de (8) est solution de (7). Dans ce cas, les systèmes (7) et (8) sont équivalents.

On a en effet

$$A' = HA\,, \qquad B' = HB\,,$$

où H est la matrice $n \times n$ dont la i-ème ligne est

$$(\,\lambda_1 \quad \lambda_2 \quad \ldots \quad \lambda_n\,)$$

et dont les autres lignes comportent un seul terme non nul, égal à 1, en position diagonale. On passe donc de (7) à (8) en multipliant les deux membres de (7) à gauche par H. Si $\lambda_i \neq 0$, la matrice H est inversible. Son inverse H^{-1} est en effet la matrice dont la i-ème ligne est

$$\left(-\frac{\lambda_1}{\lambda_i} \quad -\frac{\lambda_2}{\lambda_i} \quad \cdots \quad -\frac{\lambda_{i-1}}{\lambda_i} \quad \frac{1}{\lambda_i} \quad -\frac{\lambda_{i+1}}{\lambda_i} \quad \cdots \quad -\frac{\lambda_n}{\lambda_i}\right)$$

et dont toutes les autres lignes comportent un seul terme non nul, égal à 1, en position diagonale. On passe de (8) à (7) en multipliant les deux membres de (8) à gauche par H^{-1}.

2. Soit σ une permutation de $\{1, \ldots, n\}$. On note A^σ et B^σ les matrices obtenues en permutant, selon la permutation σ, l'ordre des lignes de A et de B, respectivement. On veut dire par là que A^σ est la matrice dont la première ligne est $A^{\sigma(1)}$, la seconde $A^{\sigma(2)}$, etc...

Toute solution X de (7) est aussi solution de

$$A^\sigma X = B^\sigma\,, \tag{9}$$

et réciproquement, puisque $(A^\sigma)^{\sigma^{-1}} = A$, $(B^\sigma)^{\sigma^{-1}} = B$. Lorsqu'on les exprime sous forme scalaire, les systèmes (7) et (9) ne diffèrent en effet que par l'ordre des équations.

On peut aussi remarquer que

$$A^\sigma = \mathbf{1}^\sigma A\,, \qquad B^\sigma = \mathbf{1}^\sigma B\,,$$

où $\mathbf{1}^\sigma$ est la matrice déduite de la matrice unité $n \times n$ en appliquant la permutation σ à l'ordre des lignes. Le système (9) s'obtient donc par multiplication des deux membres de (7) par $\mathbf{1}^\sigma$ à gauche.

3. Soit τ une permutation de $\{1, \ldots, m\}$. On note A_τ la matrice obtenue en appliquant la permutation τ à l'ordre des colonnes de A (on veut dire par là que la première colonne de A_τ est $A_{\tau(1)}$, la seconde $A_{\tau(2)}$, etc...), et X^τ la matrice obtenue en appliquant la permutation τ à l'ordre des lignes de X. On voit aisément que X est solution de (7) si et seulement si X^τ est solution de

$$A_\tau X^\tau = B\,. \tag{10}$$

En effet, lorsqu'on les exprime sous forme scalaire, les systèmes (7) et (10) ne diffèrent que par l'ordre des termes dans chaque équation.

On peut remarquer aussi que

$$A_\tau = A\mathbf{1}_\tau\,,$$

où $\mathbf{1}_\tau$ est la matrice déduite de la matrice unité $n \times n$ par application de la permutation τ à l'ordre des colonnes; L'équivalence des systèmes (7) et (10) résulte des égalités

$$A_\tau X^\tau = A\mathbf{1}_\tau \mathbf{1}^\tau X = AX\,,$$

car $\mathbf{1}_\tau$ et $\mathbf{1}^\tau$ sont deux matrices inverses l'une de l'autre.

2. La méthode de Gauss-Jordan

2.1. Cas d'un système régulier. On considère le système

$$AX = B\,, \tag{11}$$

où $A \in \mathcal{L}(\mathbf{R}^n, \mathbf{R}^n)$ est une matrice $n \times n$, B et $X \in \mathcal{L}(\mathbf{R}^p, \mathbf{R}^n)$ des matrices $n \times p$, X étant l'inconnue. On remarquera que la matrice A considérée maintenant est carrée. On la supposera de plus régulière, c'est-à-dire de déterminant non nul. On verra plus loin que la méthode de Gauss-Jordan peut être appliquée aussi lorsque A et singulière et n'a pas le même nombre de lignes que de colonnes.

Pour résoudre ce système, la *méthode de Gauss-Jordan* (aussi appelée *méthode de substitution*, ou *méthode des pivots*) consiste à appliquer aux matrices A, B et X une suite de transformations du type de celles décrites dans le paragraphe 1.5. Chacune de ces transformations donne un système équivalent à (11). On obtient ainsi une suite de systèmes

$$\begin{aligned} A^{(1)}X^{(1)} &= B^{(1)}\,,\\ A^{(2)}X^{(2)} &= B^{(2)}\,,\\ \cdots \quad &\cdots\\ A^{(n+1)}X^{(n+1)} &= B^{(n+1)}\,, \end{aligned} \tag{12}$$

tous équivalents à (11). A chaque étape, on choisit les transformations effectuées afin que la nouvelle matrice $A^{(k)}$ ait un plus grand nombre de coefficients non diagonaux nuls, et un plus grand nombre de coefficients diagonaux égaux à 1, que la précédente $A^{(k-1)}$.

Par convention, on a posé $A^{(1)} = A$, $B^{(1)} = B$, $X^{(1)} = X$, de sorte que le premier système de la suite (12) est le système de départ (11). D'autre part, on placera dans ce paragraphe et les suivants, les indices de ligne et de colonne d'une matrice tous deux en position basse: ainsi $a_{i\,j}^{(q)}$ est le coefficient de la matrice $A^{(q)}$ situé dans la i-ème ligne et la j-ème colonne. Cette convention est faite pour des raisons de commodité typographique, la position haute étant occupée par l'indice (q), placé entre parenthèses pour éviter toute confusion, qui indexe la suite de matrices construite.

On montrera plus loin que cette suite de transformations aboutit, au bout d'un nombre fini d'étapes (en général égal à n), à un système

$$A^{(n+1)} X^{(n+1)} = B^{(n+1)} \tag{13}$$

avec

$$A^{(n+1)} = \mathbf{1}\,, \tag{14}$$

matrice unité $n \times n$. Ce système se résoud immédiatement:

$$X^{(n+1)} = B^{(n+1)}\,. \tag{15}$$

Une fois $X^{(n+1)}$ obtenu, X s'obtient en permutant l'ordre des lignes de $X^{(n+1)}$. Bien entendu, lors de chaque transformation effectuée pour construire la suite de systèmes (12), on devra noter la permutation appliquée à l'ordre des colonnes de la matrice $A^{(i)}$. La permutation à appliquer à l'ordre des lignes de $X^{(n)}$ pour obtenir X est l'inverse de la composée de toutes ces permutations.

On va examiner de plus près les étapes successives de la construction.

1. **Première étape.**

On rappelle que par convention, la première étape est celle qui fait passer de $A^{(1)} = A$, $B^{(1)} = B$ et $X^{(1)} = X$ à $A^{(2)}$, $B^{(2)}$ et $X^{(2)}$.

Si le coefficient $a_{1\,1}$ de la matrice A est non nul, on déduit $A^{(2)}$ de A en divisant la première ligne de cette matrice par $a_{1\,1}$, et en retranchant aux lignes suivantes la première ligne, multipliée par un facteur choisi afin d'annuler le coefficient situé dans la première colonne. On déduit $B^{(2)}$ de B par la même transformation. On a donc

$$a_{i\,j}^{(2)} = \begin{cases} \dfrac{a_{1\,j}}{a_{1\,1}} & \text{si } i = 1\,, \quad 1 \le j \le n\,, \\ a_{i\,j} - \dfrac{a_{i\,1}}{a_{1\,1}} a_{1\,j} & \text{si } 2 \le i \le n\,, \quad 1 \le j \le n\,, \end{cases} \tag{16}$$

$$b_{i\,j}^{(2)} = \begin{cases} \dfrac{b_{1\,j}}{a_{1\,1}} & \text{si } i = 1\,, \quad 1 \le j \le p\,, \\ b_{i\,j} - \dfrac{a_{i\,1}}{a_{1\,1}} b_{1\,j} & \text{si } 2 \le i \le n\,, \quad 1 \le j \le p\,. \end{cases} \tag{17}$$

La première colonne de $A^{(2)}$ a donc un seul coefficient non nul, $a_{1\,1}^{(2)} = 1$, en position diagonale.

Si $a_{1\,1}$ est nul, on commence par choisir dans la première colonne de A un coefficient non nul $a_{s\,1}$. Il en existe au moins un, car dans le cas contraire la première colonne de la matrice A serait identiquement nulle et cette matrice serait singulière, contrairement à ce qui a été supposé. Il y a une part d'arbitraire dans le choix de ce coefficient. En pratique, on choisit souvent le coefficient le plus grand en valeur absolue dans la première colonne de A. Soit σ la permutation de $\{1,\dots,n\}$ qui échange 1 et s et laisse les autres éléments inchangés. Ainsi qu'on l'a vu au paragraphe 1.5, le système (11) est équivalent au système

$$A^\sigma X = B^\sigma\,, \tag{18}$$

où A^σ et B^σ sont les matrices déduites, respectivement, de A et de B par échange des lignes 1 et s. On dit que cette transformation est un *pivotement*, et que le coefficient $a_{s\,1}$ en est le *pivot*.

Le coefficient situé dans la première ligne et la première colonne de A^σ est maintenant $a_{s\,1}$; comme il est non nul, on peut appliquer à (18) la transformation décrite ci-dessus. On définira donc $A^{(1)}$ en posant:

$$a^{(2)}_{i\,j} = \begin{cases} \dfrac{(a^\sigma)_{1\,j}}{(a^\sigma)_{1\,1}} & \text{si } i = 1\,, \quad 1 \le j \le n\,, \\ (a^\sigma)_{i\,j} - \dfrac{(a^\sigma)_{i\,1}}{(a^\sigma)_{1\,1}}(a^\sigma)_{1\,j} & \text{si } 2 \le i \le n\,, \quad 1 \le j \le n\,, \end{cases} \tag{19}$$

ou, compte tenu de

$$(a^\sigma)_{i\,j} = a_{\sigma(i)\,j} = \begin{cases} a_{s\,j} & \text{si } i = 1\,, \quad 1 \le j \le n\,, \\ a_{1\,j} & \text{si } i = s\,, \quad 1 \le j \le n\,, \\ a_{i\,j} & \text{si } i \ne 1\,, \quad i \ne s\,, \quad 1 \le i, j \le n\,, \end{cases}$$

$$a^{(2)}_{i\,j} = \begin{cases} \dfrac{a_{s\,j}}{a_{s\,1}} & \text{si } i = 1\,, \quad 1 \le j \le n\,, \\ a_{1\,j} - \dfrac{a_{1\,1}}{a_{s\,1}}a_{s\,j} & \text{si } i = s\,, \quad 1 \le j \le n\,. \\ a_{i\,j} - \dfrac{a_{i\,1}}{a_{s\,1}}a_{s\,j} & \text{si } i \ne 1\,, \quad i \ne s\,, \quad 1 \le i, j \le n\,. \end{cases} \tag{20}$$

De même, la matrice $B^{(1)}$, est définie par

$$b^{(2)}_{i\,j} = \begin{cases} \dfrac{b_{s\,j}}{a_{s\,1}} & \text{si } i = 1\,, \quad 1 \le j \le p\,, \\ b_{1\,j} - \dfrac{a_{1\,1}}{a_{s\,1}}b_{s\,j} & \text{si } i = s\,, \quad 1 \le j \le p\,. \\ b_{i\,j} - \dfrac{a_{i\,1}}{a_{s\,1}}b_{s\,j} & \text{si } i \ne 1\,, \quad i \ne s\,, \quad 1 \le i \le n\,, \quad 1 \le j \le p\,. \end{cases} \tag{21}$$

2. **k-ième étape.**

Supposons qu'après la $k-1$-ième étape, on soit arrivé au système

$$A^{(k)} X = B^{(k)}\,, \tag{22}$$

la matrice $A^{(k)}$ ayant, dans chacune de ses $k-1$ premières colonnes, un seul coefficient non nul, égal à 1, situé en position diagonale:

$$a_{i\,j}^{(k)} = \begin{cases} 0 & \text{pour } j \leq k-1 \text{ et } i \neq j\,, \\ 1 & \text{pour } i = j \leq k-1\,. \end{cases} \tag{23}$$

La k-ième étape est celle qui fait passer de $A^{(k)}$ et $B^{(k)}$ à $A^{(k+1)}$ et $B^{(k+1)}$. Si le coefficient $a_{k\,k}^{(k)}$ de la matrice $A^{(k)}$ est non nul, on déduit $A^{(k+1)}$ de $A^{(k)}$ en divisant la k-ième ligne de cette matrice par $a_{k\,k}^{(k)}$ et en retranchant aux autres lignes la k-ième, multipliée par un facteur choisi afin d'annuler le coefficient situé dans la k-ième colonne. On déduit $B^{(k+1)}$ de $B^{(k)}$ par la même transformation. On a donc

$$a_{i\,j}^{(k+1)} = \begin{cases} \dfrac{a_{k\,j}^{(k)}}{a_{k\,k}^{(k)}} & \text{si } i = k\,, \quad 1 \leq j \leq n\,, \\ a_{i\,j}^{(k)} - \dfrac{a_{i\,k}^{(k)}}{a_{k\,k}^{(k)}} a_{k\,j}^{(k)} & \text{si } 1 \leq i \leq n\,, \quad i \neq k\,, \quad 1 \leq j \leq n\,, \end{cases} \tag{24}$$

$$b_{i\,j}^{(k+1)} = \begin{cases} \dfrac{b_{k\,j}^{(k)}}{a_{k\,k}^{(k)}} & \text{si } i = k\,, \quad 1 \leq j \leq p\,, \\ b_{i\,j}^{(k)} - \dfrac{a_{i\,k}^{(k)}}{a_{k\,k}^{(k)}} b_{k\,j}^{(k)} & \text{si } 1 \leq i \leq n\,, \quad i \neq k\,, \quad 1 \leq j \leq p\,. \end{cases} \tag{25}$$

On remarque que chacune des k premières colonnes de $A^{(k+1)}$ a un seul coefficient non nul, égal à 1, situé en position diagonale. On peut donc passer à l'étape suivante.

Si $a_{k\,k}^{(k)}$ est nul, on effectue un pivotement avant de procéder à l'élimination, comme indiqué dans la description de la première étape. Il existe dans la k-ième colonne de $A^{(k)}$ un coefficient non nul $a_{s\,k}^{(k)}$ dont l'indice de ligne vérifie $s \geq k$. Si ce n'était pas le cas la matrice $A^{(k)}$ serait de la forme

$$\begin{pmatrix} 1 & 0 & \dots & 0 & a_{1\,k} & a_{1\,k+1} & \dots & a_{1\,n} \\ 0 & \ddots & & \vdots & \vdots & \vdots & & \vdots \\ \vdots & & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 & a_{k-1\,k} & a_{k-1\,k+1} & \dots & a_{k-1\,n} \\ 0 & 0 & \dots & 0 & 0 & a_{k\,k+1} & \dots & a_{k\,n} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & 0 & a_{n\,k+1} & \dots & a_{n\,n} \end{pmatrix}$$

(pour alléger l'écriture, on a omis les indice (k) en position supérieure). Elle aurait donc un déterminant nul. Par suite, la matrice de départ A aurait aussi un déterminant nul, ce qui est contraire à l'hypothèse. On choisit pour pivot l'un de ces coefficients non nuls $a_{s\,k}^{(k)}$, avec $s \geq k$. En pratique, on choisit souvent celui qui est le plus grand en valeur absolue. On transforme les matrices $A^{(k)}$ et $B^{(k)}$ par échange des lignes k et s. Puis on applique aux matrices ainsi déduites de $A^{(k)}$ et de $B^{(k)}$ la transformation, par combinaison de lignes,

décrite ci-dessus. On laisse au lecteur le soin d'établir, en adaptant (24) et (25), les formules qui donnent le matrices transformées $A^{(k+1)}$ et $B^{(k+1)}$.

On remarquera que lors de la dernière étape ($k = n$) il n'y a pas lieu de faire de pivotement, le seul pivot possible étant $a_{n\,n}^{(n)}$.

Les considérations qui précèdent montrent qu'après n étapes (chaque étape comportant une transformation par combinaison de lignes, et pouvant comporter aussi un pivotement, c'est-à-dire une permutation de lignes), on aboutit à un système de la forme (13) dont la matrice $A^{(n+1)}$ est la matrice unité. On a donc $X = B^{(n+1)}$.

Les résultats de ce paragraphe permettent d'énoncer:

2.2 Théorème. *La méthode de Gauss-Jordan permet la résolution effective de tout système linéaire régulier.*

2.3. Commentaire sur le choix des pivots. Dans le paragraphe 2.1, lors de la description de la méthode, on a indiqué qu'à la k-ième étape, lorsque le coefficient $a_{k\,k}^{(k)}$ était non nul, on n'effectuait pas de pivotement et on déduisait $A^{(k+1)}$ de $A^{(k)}$ et $B^{(k+1)}$ de $B^{(k)}$ par les formules (24) et (25). Dans la pratique, il est préférable d'effectuer un pivotement même lorsque le coefficient $a_{k\,k}^{(k)}$ est non nul, chaque fois que ce coefficient n'est pas parmi les plus grands, en valeur absolue, des termes des lignes de rang $\geq k$ situés dans la k-iéme colonne de $A^{(k)}$. En effet, diviser par un coefficient très petit peut amplifier considérablement les erreurs dues à la précision, toujours finie, des calculs numériques en nombres réels (que ces calculs soient faits "à la main" ou par ordinateur).

D'autre part, la méthode décrite dans les paragraphes 2.1 et 2.2 n'utilise que deux des trois types de transformations décrits au paragraphe 1.5: combinaison de lignes de A et, simultanément, des lignes correspondantes de B (paragraphe 1.5.1); permutation de lignes de A et, simultanément, des lignes correspondantes de B (paragraphe 1.5.2). On n'a pas eu besoin d'utiliser de transformation du type décrit en 1.5.3: permutation de colonnes de A et, simultanément, des lignes de même rang de X.

On peut modifier légèrement la méthode de Gauss-Jordan en utilisant aussi des transformations du type décrit en 1.5.3. On indique ci-dessous comment procéder, en supposant qu'on en est à la k-ième étape; il faudra, bien entendu, procéder de même à chaque étape, dès la première. On est arrivé à un système de la forme

$$A^{(k)}X^{(k)} = B^{(k)}\,,$$

dont la matrice $A^{(k)}$ a, dans chacune de ses $k-1$ premières colonnes, un seul coefficient non nul, égal à 1, situé en position diagonale. Ses coefficients vérifient les relations (23). On cherche alors le plus grand des coefficients en valeur absolue, non plus seulement dans la k-ième colonne de $A^{(k)}$ et les lignes de rang $\geq k$, mais dans toutes les colonnes et les lignes de $A^{(k)}$ de rang $\geq k$; notons-le $a_{r\,s}^{(k)}$. On a donc $r \geq k$, $s \geq k$. On choisit ce coefficient pour pivot. La $k+1$-ième étape comportera alors trois transformations:

- permutation des lignes d'indices k et r dans les matrices $A^{(k)}$ et $B^{(k)}$,
- permutation des colonnes d'indices k et s dans la matrice déduite de $A^{(k)}$ par la transformation précédente, et permutation des lignes d'indices k et s dans la matrice $X^{(k)}$,

– combinaison de lignes comme indiqué dans la description de la k-ième étape au paragraphe 2.1.

Cette variante de la méthode, qui d'après certains auteurs améliore la stabilité et la précision, est appelée *méthode de Gauss-Jordan avec pivotement total.* Par contraste, la méthode décrite au paragraphe 2.2 sera dite *avec pivotement partiel.* Le pivotement total a l'inconvénient de modifier l'inconnue X. Si on l'utilise, on obtiendra, à la dernière étape, non pas l'inconnue X elle-même, mais une inconnue $X^{(n)}$, déduite de X par une certaine permutation des lignes. En pratique, ce n'est pas un inconvénient bien grave: il suffit de déterminer, à chaque étape k $(1 \leq k \leq n-1)$, quelle est la permutation de lignes qui fait passer de $X^{(k+1)}$ à X, et de la conserver en mémoire jusqu'à l'étape suivante; on l'obtient tout simplement en composant la transposition de lignes qui fait passer de $X^{(k+1)}$ à $X^{(k)}$ (connue dès que le pivot a été choisi) avec la permutation de lignes qui fait passer de $X^{(k)}$ à X, elle même déterminée à l'étape précédente. Après la $(n-1)$-ième étape, une fois $X^{(n)}$ obtenu, il suffira de lui appliquer la permutation de lignes ainsi trouvée pour obtenir X.

Que l'on applique la méthode de Gauss-Jordan sous sa forme originale ou avec pivotement total, il subsiste une part d'arbitraire dans le choix des pivots. Supposons en effet que l'on ait pris pour règle de prendre pour pivot le coefficient le plus grand en valeur absolue, dans les lignes ou colonnes où l'on peut le choisir. En multipliant chaque ligne de A et la ligne de même rang de B par un même facteur, on ne modifie pas la solution, alors que cette transformation, si on l'effectue avant application de la méthode de Gauss-Jordan, modifiera probablement le choix des pivots. Pour cette raison, certais auteurs recommandent de "normaliser" les équations avant application de la méthode de Gauss-Jordan, en multiplant chaque ligne de A et la ligne correspondante de B par un même facteur, ce facteur étant choisi de manière telle qu'après transformation, la somme des valeurs absolues des termes de chaque ligne de A (par exemple, ou toute autre norme facile à calculer) soit la même pour toutes les lignes.

Par ailleurs, la méthode de Gauss-Jordan avec pivotement total rend beaucoup plus aisé le traitement des systèmes linéaires singuliers, ainsi qu'on va le voir dans le paragraphe qui suit.

2.4. Cas d'un système singulier. Lorsqu'on applique la méthode de Gauss-Jordan avec pivotement total à un système linéaire de la forme (11) dont la matrice A est carrée, mais à déterminant nul, on arrive, à une certaine étape de la construction, à un système

$$A^{(k)}X^{(k)} = B^{(k)} \tag{26}$$

dont la matrice $A^{(k)}$ a toutes ses colonnes de rang $\leq k-1$ comportant chacune un seul coefficient non nul, égal à 1, situé en position diagonale, et a toutes ses lignes de rang $\geq k$ identiquement nulles. En effet, si cela ne se produisait pas, on pourrait poursuivre l'application de la méthode et on aboutirait à un système de la forme (13) dont la matrice $A^{(n)}$ serait diagonale régulière, ce qui contredirait le fait que A est singulière. La méthode permet donc de reconnaître si le système auquel on l'applique est singulier.

Deux cas sont alors possibles:

Premier cas. Les lignes de $B^{(k)}$ de rang $\geq k$ ne sont pas toutes identiquement nulles. On peut alors affirmer que le système considéré n'admet aucune solution.

Deuxième cas. Les lignes de $B^{(k)}$ de rang $\geq k$ sont toutes identiquement nulles. On voit alors qu'on peut donner aux composantes $x_{i\,j}^{(k)}$ de $X^{(k)}$ dont l'indice de ligne vérifie $i \geq k$ des valeurs quelconques. Le système (27) permet de déterminer les composantes $x_{i\,j}^{(q)}$ d'indice de ligne $i \leq k-1$ en fonction des précédentes, de manière unique, par des formules explicites que le lecteur établira aisément.

Considérons maintenant le cas où la matrice A du système de départ est rectangulaire, à m lignes et n colonnes. Il n'est pas nécessaire de distinguer les cas où le nombre de lignes est plus grand ou plus petit que le nombre de colonnes. Le lecteur vérifiera aisément que la méthode de Gauss-Jordan avec pivotement total s'applique encore, comme dans le cas où la matrice A est carrée.

Les résultats des deux paragraphes précédents permettent d'énoncer:

2.5. Théorème. *La méthode de Gauss-Jordan avec pivotement total permet la résolution effective de tout système linéaire, que celui-ci soit régulier ou singulier.*

2.6. Remarque à propos des systèmes singuliers. La méthode de Gauss-Jordan avec pivotement total nous a donné un critère permettant de reconnaître si le système linéaire étudié était singulier: à une certaine étape de la construction, on obtient une matrice $A^{(k)}$ dont toutes les lignes de rang $\geq k$ sont identiquement nulles. On a obtenu en même temps un critère permettant, lorsqu'on rencontre ce cas, de savoir si le système admet ou non des solutions: on regarde si les lignes de $B^{(k)}$ de rang $\geq k$ sont toutes identiquement nulles.

Ces critères doivent être adaptés dans la pratique numérique, car le résultat d'un calcul n'est que très exceptionnellement rigoureusement nul, du fait de la précision finie de ce calcul. Un résultat non nul, mais très petit en valeur absolue, inférieur à un certain seuil, devra être considéré comme non significativement différent de zéro, et assimilé à un résultat nul. Le choix de ce seuil dépend de la précision avec laquelle sont effectuées les opérations arithmétiques.

2.7. L'algorithme de Gauss-Jordan. Soit à résoudre le système

$$AX = B\,, \tag{27}$$

où A est une matrice $n \times n$ donnée, l'inconnue X une matrice $n \times p$, et B une matrice $n \times p$ donnée. La méthode de Gauss-Jordan, exposée en détail dans les paragraphes précédents, se formalise en l'algorithme décrit ci-dessous. Pour simplifier, on présentera cet algorithme dans le cas où la matrice du système est carrée, et où l'on applique la variante de pivotement total. L'algorithme est décrit ci-dessous en langage ordinaire. Le lecteur adepte de la programmation ne devrait avoir que peu de mal à le traduire dans son langage préféré (Pascal, Modula 2, C, Ada, Fortran, ...).

On utilise un indice k, qui variera de 1 à n, pour indexer les étapes successives de la construction. On note $A^{(k)}$ et $B^{(k)}$ les matrices qui interviennent dans le système transformé obtenu après l'étape $k-1$. On note τ_k la permutation de $\{1, \ldots, n\}$ telle que

$$(X^{(k)})^{\tau_k} = X\,.$$

Deux remarques permettent de simplifier l'écriture de l'algorithme:

1. Les formules qui font passer de $A^{(k)}$ à $A^{(k+1)}$ et de $B^{(k)}$ à $B^{(k+1)}$ sont les mêmes. On pourra donc les faire exécuter par les mêmes instructions en groupant $A^{(k)}$ et $B^{(k)}$ en une seule matrice $n \times (n+p)$:

$$D^{(k)} = \left(A^{(k)} \mid B^{(k)} \right)$$

obtenue en juxtaposant les colonnes de $A^{(k)}$ et de $B^{(k)}$.

2. Après la $(k-1)$-ième étape, on connaît $D^{(k)} = \left(A^{(k)} \mid B^{(k)} \right)$ et τ_k, et on n'a pas besoin, pour les étapes suivantes, des $D^{(j)}$ et τ_j pour $j < k$. On n'utilisera donc que deux variables pour tout l'algorithme, une matrice $n \times (n+p)$, notée D, et une permutation τ de $\{1, \ldots, n\}$. On modifiera les valeurs affectées à ces deux variables après chaque étape.

Pour $k = 0$, avant la première étape, on initialise l'algorithme en attribuant à D et τ les valeurs

$$D = \left(A \mid B \right), \qquad \tau = \mathrm{id}_{\{1,\ldots,n\}} .$$

Après l'étape $k-1$ $(1 \leq k \leq n-2)$, les valeurs de D et de τ sont connues, et sont

$$D = \left(A^{(k)} \mid B^{(k)} \right), \qquad \tau = \tau_k .$$

On recherche alors un pivot dans les lignes et colonnes de D de rang compris entre k et n (le rang de colonne ne doit pas être supérieur à n afin de rester dans la partie de D qui correspond à $A^{(k)}$); ce sera le coefficient le plus grand en valeur absolue, parmi ceux de ces lignes et colonnes.

Pour le déterminer, on prend, provisoirement, le coefficient $d_{k\,k}$ et on compare sa valeur absolue avec celle du coefficient suivant $d_{k+1\,k}$. On retient celui dont la valeur absolue est la plus grande, et on le compare au suivant $d_{k+2\,k}$, et ainsi de suite jusquà ce qu'on atteigne le coefficient $d_{n\,k}$. On passe ensuite au coefficient de la colonne suivante $d_{k\,k+1}$, et on continue ainsi, colonne par colonne, jusqu'à ce que l'on atteigne le coefficient $d_{n\,n}$.

Une fois le pivot $d_{r\,s}$ déterminé, on compare sa valeur absolue au seuil au dessous duquel un nombre réel est considéré comme non significativement différent de zéro:

- Si cette valeur absolue est inférieure à ce seuil, on examine successivement tous les termes des lignes de rang $\geq k$ de la matrice D, dans les colonnes de rang $\geq n+1$, en procédant comme indiqué ci-dessus pour le choix du pivot. Si l'on en rencontre un dont la valeur absolue dépasse le seuil choisi, on déclare que le système n'a pas de solution, et on arrête le déroulement du programme. Si tous les termes ont une valeur absolue inférieure au seuil fixé, on déclare que le système, bien que singulier, admet des solutions, que l'on peut choisir arbitrairement les coefficients des lignes de $X^{(k)}$ de rang $\geq k$. On détermine une base de l'espace des solutions grâce aux formules donnant les coefficients de $X^{(k)}$ situés dans les lignes de rang $\leq k$ en fonction de ceux qu'on peut choisir arbitrairement. On passe ensuite à la dernière étape pour déterminer X.
- Si cette valeur absolue est supérieure au seuil fixé, on prend l'élément $d_{r\,s}$ pour pivot. En effectuant les transformations indiquées aux paragraphes 2.1 et 2.2, on peut affecter à D sa nouvelle valeur

$$D = \left(A^{(k+1)} \mid B^{(k+1)} \right),$$

Quant à la nouvelle valeur qui doit être affectée à la permutation τ, elle s'obtient en composant l'ancienne valeur de τ_k avec la transposition qui échange les indices k et s.

Si le système est régulier, on atteint l'étape n. On connaît alors $X^{(n+1)} = B^{(n+1)}$, (formé par les p dernières colonnes de la valeur de D à cette étape, ainsi que τ_{n+1}, qui est la valeur de τ à cette étape. On obtient X en appliquant la permutation τ_{n+1} à l'ordre des colonnes de $X^{(n+1)}$.

On laisse au lecteur le soin de voir comment terminer l'algorithme lorsque le système est singulier mais admet des solutions. Cela présente d'ailleurs un intérêt limité, car il existe d'autres algorithmes mieux adaptés que celui de Gauss-Jordan à la résolution des systèmes singuliers.

2.8. Exercice. Déterminer le nombre d'opérations (additions ou soustractions, multiplications, divisions) nécessaires pour résoudre le système (27) par la méthode de Gauss-Jordan, dans le cas où A est une matrice $n \times n$ régulière, B et X des matrices $n \times p$.

On trouve:

$(n+1)p + \dfrac{n(n+1)}{2}$ divisions,

$(n-1)(n+1)p + \dfrac{(n-1)n(n+1)}{2}$ multiplications,

$(n-1)(n+1)p + \dfrac{(n-1)n(n+1)}{2}$ additions ou soustractions.

3. Matrices triangulaires

Avant d'aborder l'étude de la méthode de Gauss, on va donner quelques compléments sur les matrices triangulaires.

3.1. Définition. Une matrice carrée $A = (a_{i\,j})$ à n colonnes et n lignes est dite

- *triangulaire supérieure* si $a_{i\,j} = 0$ pour $i > j$,
- *triangulaire inférieure* si $a_{i\,j} = 0$ pour $i < j$.

(On rappelle que le premier indice indexe les lignes et le second les colonnes).

3.2. Quelques propriétés.

1. Le produit de deux matrices triangulaires supérieures (resp., inférieures) est une matrice triangulaire supérieure (resp., inférieure). Montrons-le par exemple pour des matrices triangulaires supérieures $A = (a_{i\,j})$, $B = (b_{i\,j})$ Soit $C = (c_{i\,j}) = AB$. On a

$$c_{i\,j} = \sum_k a_{i\,k} b_{k\,j}\,.$$

Supposons $i > j$. Si $k < i$, on a $a_{i\,k} = 0$; si $k \geq i$, cela implique $k > j$ donc $b_{k\,j} = 0$. Par suite $c_{i\,j} = 0$.

2. La transposée d'une matrice triangulaire supérieure est triangulaire inférieure.

3. Une matrice triangulaire A est inversible si et seulement si tous ses termes diagonaux son non nuls. Son déterminant est en effet

$$\det(A) = \prod_i a_{i\,i}\,.$$

3.3. Systèmes linéaires à matrice triangulaire. Considérons le système linéaire

$$AX = B\,, \tag{28}$$

où $A = (a_{i\,j})$ est une matrice $n \times n$ triangulaire supérieure, $X = {}^t(x_1, \ldots, x_n)$ et $B = {}^t(b_1, \ldots, b_n)$ deux vecteurs-colonnes à n composantes. Sous forme scalaire, ce système s'écrit

$$\begin{aligned} a_{1\,1}x_1 + a_{1\,2}x_2 + \cdots + a_{1\,n}x_n &= b_1\,,\\ a_{2\,2}x_2 + \cdots + a_{2\,n}x_n &= b_2\,,\\ \cdots \qquad & \cdots\\ a_{n\,n}x_n &= b_n\,. \end{aligned}$$

On suppose A inversible donc tous les $a_{i\,i} \neq 0$. La résolution du système est très facile, car on peut tirer x_n de la dernière équation,

$$x_n = \frac{b_n}{a_{n\,n}}\,,$$

puis x_{n-1} de l'avant-dernière, et ainsi de suite jusqu'à x_1 qu'on tire de la première équation. La formule exprimant x_k (en supposant $x_{k+1}, \ldots, x_n$ précédemment déterminés) est

$$x_k = \frac{1}{a_{k\,k}} \left(b_k - \sum_{i=1}^{n-k} a_{k\,k+i}\, x_{k+i} \right)\,.$$

Cette méthode de résolution sera appelée *substitution en retour*.

Lorsqu'on explicite les expressions de $x_1, \ldots, x_n$ en effectuant les substitutions, on obtient des expressions de la forme

$$\begin{aligned} c_{1\,1}b_1 + c_{1\,2}b_2 + \cdots + c_{1\,n}b_n &= x_1\,,\\ c_{2\,2}b_2 + \cdots + c_{2\,n}b_n &= x_2\,,\\ \cdots \qquad & \cdots\\ c_{n\,n}b_n &= x_n\,. \end{aligned}$$

Les $c_{i\,j}$ sont les coefficients de la matrice $C = A^{-1}$, inverse de A. On remarque que

$$c_{i\,j} = 0 \quad \text{si} \quad i > j\,, \quad c_{i\,i} = \frac{1}{a_{i\,i}}\,.$$

On a donc prouvé le résultat:

Proposition. *L'inverse d'une matrice triangulaire supérieure inversible est triangulaire supérieure.*

Considérons maintenant le cas où A est triangulaire inférieure. Le système (28) s'explicite maintenant en

$$\begin{aligned} a_{1\,1}x_1 &= b_1\,, \\ a_{2\,1}x_1 + a_{2\,2}x_2 &= b_2\,, \\ \cdots \qquad & \cdots \\ a_{n\,1}x_1 + a_{n\,2}x_2 + \cdots + a_{n\,n}x_n &= b_n\,. \end{aligned}$$

Ce système se résoud en tirant x_1 de la première équation

$$x_1 = \frac{b_1}{a_{1\,1}}\,,$$

puis x_2 de la seconde équation, et ainsi de suite jusqu'à x_n qu'on tire de la dernière équation. L'expression de x_k (supposant $x_1, \ldots, x_{k-1}$ précédemment déterminés) est

$$x_k = \frac{1}{a_{k\,k}} \left(b_k - \sum_{i=1}^{k-1} a_{k\,i}x_i \right)\,.$$

Cette méthode de résolution sera appelée *substitution directe*.

Comme ci-dessus, on peut énoncer le résultat:

Proposition. *L'inverse d'une matrice triangulaire inférieure inversible est triangulaire inférieure.*

4. La méthode de Gauss

4.1. Description de la méthode. La *méthode de Gauss*, ou *méthode de simple élimination*, est une méthode de résolution des systèmes linéaires très voisine, dans son principe, de celle de Gauss-Jordan. La méthode de Gauss est historiquement antérieure à celle de Gauss-Jordan. Nous la présentons cependant après celle-ci car elle est peut-être un peu plus subtile, et permet de mieux comprendre les méthodes de décomposition triangulaire qui seront exposées ensuite.

Soit à résoudre le système

$$AX = B\,, \tag{29}$$

où A est une matrice $n \times n$ donnée, X et B des matrices $n \times p$, X étant l'inconnue et B étant donnée. Comme la méthode de Gauss-Jordan, la méthode de Gauss consiste à former une suite de systèmes

$$\begin{aligned} A^{(1)}X^{(1)} &= B^{(1)}\,, \\ A^{(2)}X^{(2)} &= B^{(2)}\,, \\ \cdots \quad & \cdots \\ A^{(n)}X^{(n)} &= B^{(n)}\,, \end{aligned} \tag{30}$$

en appliquant aux matrices A, B et X certaines transformations des types décrits au paragraphe 1.5. Mais au lieu de chercher à transformer la matrice A en la matrice unité, on se contente ici de la transformer en une matrice triangulaire supérieure. On a vu, au paragraphe précédent, qu'il sera alors facile de résoudre le système par des formules explicites.

La première étape est presque identique à celle de la méthode de Gauss-Jordan, décrite au paragraphe 2.1; la seule différence est que dans la méthode de Gauss, on laisse la première ligne inchangée, on ne la divise pas par le pivot; les transformations appliquées aux autres lignes sont les mêmes que dans la méthode de Gauss-Jordan.

Décrivons la $(k+1)$-ième étape. Après la k-ième étape, on est arrivé au système

$$A^{(k)}X^{(k)} = B^{(k)}\,,$$

les coefficients de la matrice $A^{(k)}$ vérifiant

$$a_{i\,j}^{(k)} \begin{cases} \neq 0 & \text{pour } i = j \leq k-1\,, \\ = 0 & \text{pour } 1 \leq i \leq k-1\,, \quad 1 \leq j < i\,. \end{cases} \tag{31}$$

La matrice $A^{(k)}$ est donc de la forme

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \dots & a_{1\,k-1} & a_{1\,k} & \dots & a_{1\,n} \\ 0 & a_{2\,2} & \dots & a_{2\,k-1} & a_{2\,k} & & \vdots \\ \vdots & & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & a_{k-1\,k-1} & a_{k-1\,k} & \dots & a_{k-1\,n} \\ 0 & 0 & \dots & 0 & a_{k\,k} & \dots & a_{k\,n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & a_{n\,k} & \dots & a_{n\,n} \end{pmatrix}$$

(pour alléger l'écriture, on a omis les indice (k) en position supérieure).

Si le coefficient $a_{k\,k}^{(k)}$ est non nul, on retranche la k-ième ligne, multipliée par un facteur convenable, des lignes suivantes, de rangs $k+1$, $k+2, \dots, n$, le facteur étant choisi afin d'annuler le coefficient situé dans la k-ième colonne. On laisse les lignes de rang $\leq k$ inchangées (c'est ce qui distingue la méthode de Gauss de celle de Gauss-Jordan). On pose donc

$$a_{i\,j}^{(k+1)} = \begin{cases} a_{i\,j}^{(k)} & \text{si } 1 \leq i \leq k\,, \quad 1 \leq j \leq n\,, \\ 0 & \text{si } k+1 \leq i \leq n\,, \quad 1 \leq j \leq k\,, \\ a_{i\,j}^{(k)} - \dfrac{a_{i\,k}^{(k)}}{a_{k\,k}^{(k)}} a_{k\,j}^{(k)} & \text{si } k+1 \leq i \leq n\,, \quad k+1 \leq j \leq n\,, \end{cases} \tag{32}$$

On déduit $B^{(k+1)}$ de $B^{(k)}$ par la même combinaison de lignes:

$$b_{i\,j}^{(k+1)} = \begin{cases} b_{i\,j}^{(k)} & \text{si } 1 \leq i \leq k\,, \quad 1 \leq j \leq p\,, \\ b_{i\,j}^{(k)} - \dfrac{a_{i\,k}^{(k)}}{a_{k\,k}^{(k)}} b_{k\,j}^{(k)} & \text{si } k+1 \leq i \leq n\,, \quad 1 \leq j \leq p\,, \end{cases} \tag{33}$$

Si le coefficient $a_{k\,k}^{(k)}$ est nul, on effectue un pivotement, comme dans la méthode de Gauss-Jordan, avant la combinaison de lignes décrite ci-dessus. On pourra rechercher le pivot soit seulement parmi les éléments de la k-ième colonne, dont le rang de ligne est $\geq k$, soit parmi tous les éléments dont les rangs de ligne et de colonne sont tous deux $\geq k$. Dans le premier cas, on dit qu'on effectue un *pivotement partiel*, et dans le second un *pivotement total*.

En pratique, comme dans la méthode de Gauss-Jordan, il est recommandé (voire même indispensable pour assurer la stabilité de la méthode) d'effectuer un pivotement même lorsque le coefficient $a_{k\,k}^{(k)}$ est non nul, et de prendre pour pivot un coefficient aussi grand que possible en valeur absolue.

Lorsque le système de départ est régulier on aboutit, après $n-1$ étapes, à un système

$$A^{(n)}X^{(n)} = B^{(n)} \tag{34}$$

dont la matrice $A^{(n)}$ est triangulaire supérieure et à coefficients diagonaux tous non nuls: $a_{i\,j}^{(n)} = 0$ si $i > j$, $\neq 0$ si $i = j$. Le système (34) se résoud alors comme indiqué au paragraphe 3.3.

Lorsque le système de départ est singulier et qu'on applique la variante de pivotement total, on aboutit, à une certaine étape, à un système

$$A^{(q)}X^{(q)} = B^{(q)}$$

dont la matrice $A^{(q)}$ est triangulaire supérieure, a ses coefficients diagonaux $a_{i\,i}^{(q)}$ non nuls pour $1 \leq i \leq q-1$, et a ses lignes de rang $\geq q$ identiquement nulles. Comme dans la méthode de Gauss-Jordan, si les lignes de $B^{(q)}$ de rang $\geq q$ ne sont pas identiquement nulles, le système n'a pas de solution. Si elles sont identiquement nulles, le système admet des solutions; on peut choisir arbitrairement les coefficients de $X^{(q)}$ situés dans les lignes de rang $\geq q$; les autres composantes s'obtiennent, en fonction de celles-ci, par la méthode de substitution en retour décrite au paragraphe 3.3.

Une fois déterminé $X^{(n)}$, on en déduit X par permutation de lignes, comme dans la méthode de Gauss-Jordan.

On remarquera qu'en général la méthode de Gauss comporte $n-1$ étapes, alors que celle de Gauss-Jordan en comporte n: dans la méthode de Gauss-Jordan on obtient, après la $n-1$-ième étape, une matrice A^n diagonale jusqu'à sa $(n-1)$-ième colonne et dont tous les termes diagonaux sont égaux à 1, sauf le dernier, $a_{n\,n}^{(n)}$. Il faut une dernière étape consistant à diviser la dernière ligne de $A^{(n)}$ par $a_{n\,n}^{(n)}$, et à la retrancher, après multiplication par un facteur convenable, des lignes précédentes, afin d'annuler les termes non diagonaux de la dernière colonne. Cette dernière étape n'existe pas dans la méthode de Gauss puisqu'on veut obtenir une matrice triangulaire, ce qu'on a déjà après la $(n-1)$-ième étape.

On laisse au lecteur le soin de décrire plus complètement l'algorithme de Gauss pour la résolution d'un système linéaire (il pourra s'inspirer du paragraphe 2.7 qui décrit l'algorithme de Gauss-Jordan, et devra ajouter les opérations supplémentaires nécessaires pour la résolution du système à matrice triangulaire obtenu).

4.2. Exercice. Déterminer le nombre d'opérations (additions ou soustractions, multiplications, divisions) nécessaires pour résoudre un système linéaire par la méthode de Gauss. Comparer avec l'algorithme de Gauss-Jordan.

On trouve, pour la mise du système sous forme triangulaire:

$$\frac{(n-1)n}{2} + (n-1)p \text{ divisions,}$$

$$\frac{(n-1)n(2n-1)}{6} + \frac{(n-1)n}{2}p \text{ multiplications,}$$

$$\frac{(n-1)n(2n-1)}{6} + \frac{(n-1)n}{2}p \text{ additions ou soustractions.}$$

Il faut encore ajouter à cela les nombres d'opérations nécessaires pour résoudre les p systèmes linéaires à matrice triangulaire obtenus. On trouve:

$$np \text{ divisions,}$$

$$\frac{(n-1)n}{2}p \text{ multiplications,}$$

$$\frac{(n-1)n}{2}p \text{ additions ou soustractions.}$$

Au total, cela fait donc

$$\frac{(n-1)n}{2} + (2n-1)p \text{ divisions,}$$

$$\frac{(n-1)n(2n-1)}{6} + (n-1)np \text{ multiplications,}$$

$$\frac{(n-1)n(2n-1)}{6} + (n-1)np \text{ additions ou soustractions.}$$

On remarquera que la méthode de Gauss est, en général, plus avantageuse que celle de Gauss-Jordan: pour n grand, le nombre de multiplications (et aussi d'additions ou soustractions) à effectuer est de l'ordre de $n^3/2$ pour la méthode de Gauss-Jordan, et de l'ordre de $n^3/3$ pour la méthode de Gauss.

5. Décomposition triangulaire

5.1. Intérêt de la décomposition triangulaire. Soit à résoudre le système linéaire

$$AX = B\,, \tag{35}$$

où A est une matrice $n \times n$, X et B des matrices $n \times p$, X étant l'inconnue.

Définition. On dira que la matrice A admet une décomposition triangulaire si elle peut s'exprimer sous forme d'un produit

$$A = GD\,,$$

où G est triangulaire inférieure et D triangulaire supérieure.

On a utilisé les lettres G et D pour évoquer "gauche" et "droite", plutôt que I pour "inférieur" et S pour "supérieur", la lettre I risquant de causer des confusions avec la matrice unité.

Si A admet une décomposition triangulaire connue, il est facile de résoudre le système (35) en résolvant successivement deux systèmes,

$$GY = B\,, \quad \text{puis} \quad DX = Y\,,$$

le premier à matrice triangulaire supérieure, le second à matrice triangulaire inférieure. On utilisera pour cela les méthodes de substitution directe et en retour exposées au paragraphe 3.3.

5.2. Quelques questions. On est naturellement conduit à se poser les questions suivantes.

1. Toute matrice admet-elle une décomposition triangulaire? Si ce n'est pas le cas, peut-on donner des conditions nécessaires et suffisantes d'existence d'une telle décomposition?

2. Lorsqu'une matrice A admet une décomposition triangulaire, celle-ci est-elle unique?

3. Comment trouver explicitement cette décomposition lorsqu'elle existe?

La réponse à la première question est négative (même lorsqu'on impose à A d'être inversible). Considérons en effet la matrice

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

et essayons de l'identifier au produit des deux matrices

$$G = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}, \qquad D = \begin{pmatrix} d & e \\ 0 & f \end{pmatrix}.$$

On obtient les relations

$$\begin{cases} ad = 0\,, \\ cf = 0\,, \end{cases} \qquad \begin{cases} ae = 1\,, \\ bd = 1\,, \end{cases}$$

qui sont incompatibles (le second système impose à a et à d d'être non nuls, tandis que le premier impose à l'un des deux au moins d'être nul).

On donnera plus loin une condition nécessaire et suffisante pour qu'une matrice inversible admette une décomposition triangulaire. On montrera aussi que toute matrice régulière peut être transformée, par une permutation de ses lignes, en une matrice admettant une décomposition triangulaire. Pour l'application à la résolution de systèmes linéaires, ce résultat est intéressant car le système (35) est équivalent au système

$$A^\sigma X = B^\sigma\,,$$

où σ est la permutation qui, appliquée à l'ordre des lignes de A, la transforme en la matrice A^σ admettant une décomposition triangulaire.

La réponse à la seconde question est négative: si $A = GD$ est une décomposition triangulaire de A, $A = G'D'$, avec $G' = GK$ et $D' = K^{-1}D$ en est une autre, K étant une matrice diagonale inversible quelconque. Cependant, la proposition suivante montre qu'en imposant une condition supplémentaire à G, on peut rendre la décomposition unique.

Proposition. *Soit A une matrice inversible admettant une décomposition triangulaire. Il existe une unique décomposition triangulaire*

$$A = GD$$

de A telle que la matrice G ait tous ses coefficients diagonaux égaux à 1.

Démonstration. S'il existe une décomposition triangulaire

$$A = \Gamma\Delta$$

de A, les coefficients diagonaux de Γ sont non nuls, puisque A (donc aussi Γ) est inversible. En multipliant Γ à gauche par la matrice diagonale K dont les coefficients sont les inverses des coefficients diagonaux de Γ, on obtient une matrice triangulaire inférieure $G = \Gamma K$ dont tous les coefficients diagonaux sont égaux à 1. Il suffit de poser $D = K^{-1}\Delta$ pour obtenir l'autre terme de la décomposition triangulaire cherchée.

Supposons qu'il existe deux décompositions triangulaires

$$A = GD = G'D'$$

de la matrice A. On a donc

$$G'^{-1}G = D'D^{-1}\,,$$

où le membre de gauche est une matrice triangulaire inférieure, et le membre de droite une matrice triangulaire supérieure. Les deux membres sont donc égaux à une matrice diagonale K, et on a

$$G' = GK^{-1}\,, \quad D' = KD\,.$$

Si G et G' ont toutes deux leurs coefficients diagonaux égaux à 1, K est la matrice unité et on a $G = G'$, $D = D'$. □

Les paragraphes suivants répondent à la question 3: comment déterminer explicitement la décomposition triangulaire d'une matrice?

6. Méthode de Gauss et décomposition triangulaire

6.1. Cas où la méthode de Gauss s'applique sans pivotement. On reprend les étapes successives de la méthode de Gauss, appliquée au système

$$AX = B\,,$$

où A est une matrice $n \times n$ qu'on supposera inversible. On s'intéressera plus particulièrement aux transformations successives qui donnent $A^{(2)}$, $A^{(3)}$, ..., $A^{(n)}$ à partir de $A^{(1)} = A$. On supposera qu'à chaque étape, le pivot peut être choisi dans la diagonale, et qu'on n'a donc pas à effectuer de pivotement.

À la première étape, la matrice $A^{(2)}$ se déduit de la matrice donnée $A^{(1)} = A$ en laissant la première ligne inchangée et en retranchant, de chacune des lignes suivantes, d'indice i, le produit de la première ligne par le facteur $a_{i\,1}/a_{1\,1}$. Cette transformation consiste en fait à poser

$$A^{(2)} = P^{(1)}A^{(1)} = P^{(1)}A\,,$$

où $P^{(1)}$ est la matrice dont la première colonne a pour transposée (on effectue cette transposition pour des raisons typographiques):

$$\left(1 \quad -\frac{a_{2\,1}}{a_{1\,1}} \quad -\frac{a_{3\,1}}{a_{1\,1}} \quad \cdots \quad -\frac{a_{n\,1}}{a_{1\,1}}\right),$$

et dont toutes les autres colonnes ont un seul élément non nul, égal à 1, en position diagonale.

Juste avant la k-ième étape, on a obtenu une matrice $A^{(k)} = \left(a_{i\,j}^{(k)}\right)$, partiellement triangulaire, dont les coefficients situés dans les $k-1$ premières colonne, au dessous de la diagonale, sont nuls:

$$a_{i\,j}^{(k)} = 0 \qquad \text{pour} \quad i > j \text{ et } 1 \leq j \leq k-1\,.$$

Comme ci-dessus, on voit que $A^{(k+1)}$ se déduit de A^k par la transformation

$$A^{(k+1)} = P^{(k)}A^{(k)}\,,$$

où $P^{(k)}$ est la matrice dont la k-ième colonne a pour transposée

$$\left(0 \quad \cdots \quad 0 \quad 1 \quad -\frac{a_{k+1\,k}^{(k)}}{a_{k\,k}^{(k)}} \quad \cdots \quad -\frac{a_{k+1\,n}^{(k)}}{a_{k\,k}^{(k)}}\right)$$

(le 1 occupe la k-ième position) et dont les autres colonnes comportent chacune un seul terme non nul, égal à 1, en position diagonale.

On remarque que $P^{(k)}$ est inversible et a pour inverse la matrice $Q^{(k)}$ dont la k-ième colonne a pour transposée

$$\left(0 \quad \ldots \quad 0 \quad 1 \quad \frac{a_{k+1\,k}^{(k)}}{a_{k\,k}^{(k)}} \quad \cdots \quad \frac{a_{k+1\,n}^{(k)}}{a_{k\,k}^{(k)}}\right)$$

et dont les autres colonnes comportent chacune un seul terme non nul, égal à 1, en position diagonale. On peut écrire

$$A^{(k+1)} = P^{(k)}A^{(k)} = P^{(k)}P^{(k-1)}A^{(k-1)} = \cdots = P^{(k)}P^{(k-1)}\dots P^{(1)}A\,,$$

ou encore

$$A = R^{(k)}A^{(k+1)}\,, \quad \text{avec} \quad R^{(k)} = Q^{(1)}Q^{(2)}\dots Q^{(k)}\,.$$

On vérifie aisément que $R^{(k)}$ est la matrice dont la première colonne est identique à la première colonne de $Q^{(1)}$, la seconde identique à la seconde colonne de $Q^{(2)}$, ..., la k-ième identique à la k-ième colonne de $Q^{(k)}$, et dont les autres colonnes comportent chacune un seul terme non nul, égal à 1, en position diagonale.

À la $(n-1)$-ième étape, on obtient

$$A = GD\,,$$

où $G = R^{(n-1)}$ est une matrice triangulaire inférieure dont tous les coefficients diagonaux sont égaux à 1, et $D = A^{(n)}$ une matrice triangulaire supérieure. On remarquera que l'algorithme de Gauss permet, à chaque étape k, la détermination explicite de la k-ième colonne de G.

On voit donc que lorsque l'algorithme de Gauss est applicable sans pivotement à une matrice A inversible, cette matrice admet une décomposition triangulaire dont l'algorithme de Gauss permet la détermination explicite. Avant d'établir la réciproque, on introduit la définition:

Définition. Soit $A = (a_{i\,j})$, $(1 \le i, j \le n)$ une matrice $n \times n$. Pour tout entier k $(1 \le k \le n)$, on appelle *mineur principal de rang k de A* la matrice $M_k = (a_{i\,j})$ $(1 \le i, j \le k)$ formée par les termes des k premières lignes et des k premières colonnes de A.

On peut maintenant énoncer:

Théorème. *Soit $A = (a_{i\,j})$ une matrice $n \times n$ inversible. Les trois propriétés suivantes sont équivalentes:*

1. *L'algorithme de Gauss est applicable à A sans pivotement.*
2. *La matrice A admet une décomposition triangulaire.*
3. *Les mineurs principaux M_k de A $(1 \le k \le n)$ ont tous un déterminant non nul.*

Lorsque ces propriétés équivalentes sont satisfaites, les pivots $a_{k\,k}^{(k)}$ obtenus par application de l'algorithme de Gauss ont pour valeurs:

$$a_{k\,k}^{(k)} = \frac{\det M_k}{\det M_{k-1}}\,, \tag{36}$$

et l'algorithme de Gauss permet la détermination explicite de la décomposition triangulaire de A.

Démonstration. On a déjà prouvé que 1 implique 2. Supposons 2 vérifié, et soit $A = GD$ la décomposition triangulaire de A (unique d'après la proposition du paragraphe 5.2) telle que $G = (g_{i\,j})$ ait tous ses coefficients diagonaux égaux à 1. On vérifie que G peut s'écrire

$$G = G_1 G_2 \cdots G_n \,,$$

où G_k est la matrice dont la k-ième colonne est égale à la k-ième colonne de G et dont toutes les autres colonnes contiennent chacune un seul coefficient non nul, égal à 1, en position diagonale. Pour tout k $(1 \leq k \leq n)$ et toute matrice Δ, la matrice $G_k\Delta$ est la matrice dont les k premières lignes sont les mêmes que celles de Δ et dont les autres lignes, de rang $k+i$ $(1 \leq i \leq n-k)$ sont, chacune, somme de la $(k+i)$-ème ligne de Δ et du produit par $g_{k+i\,k}$ de la k-ième ligne de Δ. Par suite, les déterminants des mineurs principaux de même rang de Δ et de $G_k\Delta$ sont égaux.

En appliquant ce résultat au cas où $\Delta = G_{k+1} \cdots G_n D$, et en donnant successivement à k les valeurs n, $n-1$, ..., 1, on voit que les déterminants des mineurs principaux de même rang de D et de $A = GD$ sont égaux. Mais comme D est triangulaire supérieure, son k-ième mineur principal a pour déterminant le produit $d_{1\,1} \cdots d_{k\,k}$. On a donc

$$\det M_k = d_{1\,1} \cdots d_{k\,k} \,.$$

Comme A est inversible, les $d_{i\,i}$ sont tous non nuls. Par suite, les mineurs principaux de A ont tous un déterminant non nul, et on a prouvé que 2 implique 3.

Supposons 3 vérifié (les mineurs principaux de A ont tous un déterminant non nul). Supposons qu'on ait pu appliquer à A l'algorithme de Gauss jusqu'à l'étape k, sans pivotement. Le même raisonnement que ci-dessus montre que les déterminants des mineurs principaux de même rang de A et de $A^{(k)}$ sont égaux. Or $A^{(k)}$ est partiellement triangulaire, son mineur principal d'ordre k est une matrice triangulaire supérieure, dont le déterminant est

$$a_{1\,1}^{(k)} a_{2\,2}^{(k)} \cdots a_{k\,k}^{(k)} \,.$$

Le coefficient $a_{k\,k}^{(k)}$ est donc non nul, et on peut effectuer la $(k+1)$-ième étape sans pivotement.

On a donc prouvé l'équivalence des propriétés 1, 2 et 3. De plus, on a vu au cours du raisonnement que les pivots sont bien donnés par la formule (36). Enfin, les considérations présentées au début du paragraphe 6.1 montrent que l'algorithme de Gauss, lorsqu'on peut l'appliquer sans pivotement, donne explicitement la décomposition triangulaire de A. □

6.2. Cas où des pivotements sont nécessaires. On considère maintenant le cas où l'application de l'algorithme de Gauss à la matrice A (toujours supposée inversible) nécessite des pivotements. On n'utilisera que des pivotements partiels (modification de l'ordre des lignes de A); on sait que cela suffit pour appliquer l'algorithme de Gauss à toute matrice inversible.

À la première étape, on appliquera donc une permutation σ_1 à l'ordre des lignes de $A^{(1)} = A$, avant d'effectuer la transformation par combinaison de colonnes. On remarquera que σ_1 est en fait une transposition échangeant 1 avec un autre élément s_1 de $\{1, \ldots, n\}$, et laissant les autres éléments inchangés. On note A^{σ_1} la matrice ainsi déduite de A. On sait que

$$A^{\sigma_1} = \mathbf{1}^{\sigma_1} A \,,$$

où $\mathbf{1}^{\sigma_1}$ est la matrice déduite de la matrice unité en appliquant la transposition σ_1 à l'ordre des lignes. La matrice $A^{(2)}$ est alors définie par

$$A^{(2)} = P^{(1)} A^{\sigma_1} = P^{(1)} \mathbf{1}^{\sigma_1} A \,,$$

où $P^{(1)}$ est la matrice définie comme dans le paragraphe précédent, mais en remplaçant A par A^{σ_1}. En remarquant que $(\mathbf{1}^{\sigma_1})^{-1} = \mathbf{1}^{\sigma_1}$, et en notant comme précédemment $Q^{(1)}$ l'inverse de P^1, on peut écrire

$$\mathbf{1}^{\sigma_1} Q^{(1)} A^{(2)} = A^{(1)} = A \,.$$

De même, le passage de $A^{(k)}$ à $A^{(k+1)}$, à l'étape k, se traduit par

$$\mathbf{1}^{\sigma_k} Q^{(k)} A^{(k+1)} = A^{(k)} \,,$$

où σ_k désigne la transposition de lignes qu'on fait subir à $A^{(k-1)}$ (échange de la k-ième ligne avec une ligne de rang $s_k \geq k$) avant d'effectuer la combinaison de lignes qui donne $A^{(k+1)}$. On peut donc écrire, pour tout k $(1 \leq k \leq n-1)$:

$$A = \mathbf{1}^{\sigma_1} Q^{(1)} \mathbf{1}^{\sigma_2} Q^{(2)} \cdots \mathbf{1}^{\sigma_k} Q^{(k)} A^{(k+1)} \,. \tag{37}$$

On veut modifier l'ordre des termes du membre de droite de l'égalité ci-dessus, afin de placer tous les $\mathbf{1}^{\sigma_i}$ à gauche des $Q^{(j)}$. Ces matrices ne commutent pas. Cependant, on va voir qu'elles vérifient une propriété remarquablement simple, qui permet d'effectuer ces commutations à condition de modifier convenablement les $Q^{(j)}$. On remarque tout d'abord que Q^j est une matrice dont la j-ième colonne a pour transposée

$$(\, 0 \quad \ldots \quad 0 \quad 1 \quad \alpha_{j+1} \quad \ldots \quad \alpha_n \,)$$

(le 1 occupant la j-ème place), et dont les autres colonnes ont chacune un seul terme non nul, égal à 1, en position diagonale. Les α_{j+k} sont des scalaires qu'il n'est pas utile d'expliciter.

D'autre part, σ_i est la transposition qui échange l'élément i de $\{1, \ldots, n\}$ avec un autre élément $s \geq i$, les places des autres éléments restant inchangées.

On peut alors énoncer:

Lemme. *Pour $i > j$, on a, avec les notations précisées ci-dessus,*

$$Q^{(j)} \mathbf{1}^{\sigma_i} = \mathbf{1}^{\sigma_i} Q'^{(j)} \,,$$

où $Q'^{(j)}$ est la matrice dont la j-ème colonne a pour transposée

$$(\, 0 \quad \ldots \quad 0 \quad 1 \quad \alpha_{\sigma_i(j+1)} \quad \ldots \quad \alpha_{\sigma_i(n)} \,)$$

et dont les autres colonnes ont un seul terme non nul, égal à 1, en position diagonale.

En d'autres termes, on peut inverser l'ordre de $Q^{(j)}$ et de $\mathbf{1}^{\sigma_i}$ dans le produit qui figure au membre de droite de (37), à condition d'échanger en même temps les termes de rang i et de rang s dans la j-ème colonne de $Q^{(j)}$ (si σ_i est la transposition qui échange i et s).

Démonstration. On rappelle que σ_i est la transposition de $(1,\dots,n)$ qui échange les termes i et s_i, avec $s_i > i$, les autres termes restant inchangés. On remarque que les matrices $\mathbf{1}^{\sigma_i}$ et $\mathbf{1}_{\sigma_i}$, déduites de la matrice unité par échange, respectivement, des *colonnes* d'indices i et s_i et des *lignes* d'indices i et s_i, sont égales. Pour toute matrice Q carrée $n \times n$, les matrices $Q\mathbf{1}^{\sigma_i}$ et $\mathbf{1}^{\sigma_i}Q$ se déduisent de Q par échange, respectivement, des colonnes d'indices i et s_i et des lignes d'indices i et s_i. Supposons Q du même type que $Q^{(j)}$; sa j-ème colonne a des coefficients nuls au dessus de la diagonale, un coefficient égal à 1 dans la diagonale, et au dessous de la diagonale, des coefficients $q_{j+1},\dots,q_n$. Toutes ses autres colonnes ont chacune un seul coefficient non nul, égal à 1, situé dans la diagonale. Puisque $j < i < s_i$, la matrice $Q\mathbf{1}^{\sigma_i}$ diffère de Q uniquement par le fait que dans ses lignes de rangs i et s_i, le coefficient 1, au lieu d'être dans la diagonale, se trouve, respectivement, aux places s_i et i. La matrice $\mathbf{1}^{\sigma_i}Q$ diffère elle aussi de Q uniquement par ses lignes de rangs i et s_i; mais en plus de la place des coefficients 1, il y a une autre différence: le j-ème terme de la ligne i est q_{s_i} au lieu de q_i, et le j-ème terme de la ligne s_i est q_i au lieu de q_{s_i}. Moyennant cette simple remarque, la propriété indiquée dans l'énoncé du lemme devient évidente. □

On peut donc écrire:

$$A = \mathbf{1}^{\sigma_1}\mathbf{1}^{\sigma_2}\dots\mathbf{1}^{\sigma_k}S^{(1)}S^{(2)}\dots S^{(k)}A^{(k+1)}\,,$$

où $S^{(1)}, S^{(2)},\dots,S^{(k)}$ sont les matrices qui se déduisent de $Q^{(1)}, Q^{(2)},\dots,Q^{(k)}$ par application du lemme.

On en déduit:

$$\mathbf{1}^{\tau_k}A = G^{(k)}A^{(k+1)}\,,$$

où τ_k est la permutation composée

$$\tau_k = \sigma_k\sigma_{k-1}\dots\sigma_1\,,$$

et où

$$G^{(k)} = S^{(1)}S^{(2)}\dots S^{(k)}\,.$$

Pour $k = n-1$, on obtient

$$\mathbf{1}^{\tau_n}A = GD\,,$$

où $G = G^{(n-1)}$ est une matrice triangulaire inférieure dont tous les coefficients diagonaux sont égaux à 1, $D = A^{(n)}$ une matrice triangulaire supérieure et τ_n la permutation de $\{1,\dots,n\}$ obtenue en composant les transpositions correspondant aux pivotements effectués lors de l'application de l'algorithme de Gauss. On peut donc énoncer:

Théorème. *Pour toute matrice $n \times n$ inversible A, il existe une permutation τ de $\{1,\dots,n\}$ telle que $\mathbf{1}^{\tau}A$ (matrice déduite de A par application de la permutation τ à l'ordre des lignes) admette une décomposition triangulaire. L'algorithme de Gauss permet la détermination explicite de τ et de cette décomposition.*

6.3. L'algorithme de Crout. Soit A une matrice $n \times n$ inversible. On désire déterminer une permutation τ telle que $\mathbf{1}^{\tau}A$ admette une décomposition triangulaire, et la décomposition triangulaire de $\mathbf{1}^{\tau}A$. Ainsi qu'on l'a vu dans les deux paragraphes précédents,

l'algorithme de Gauss permet de résoudre ce problème, pourvu qu'on lui apporte quelques aménagements. Ces aménagements consistent essentiellement à noter, après chaque étape, non seulement la matrice $A^{(k)}$ (comme on le faisait lorsque l'on appliquait l'algorithme de Gauss pour résoudre un système linéaire), mais aussi ce qui est nécessaire pour obtenir finalement la permutation τ et la décomposition triangulaire de $\mathbf{1}^\tau A$. Ainsi aménagé, l'algorithme de Gauss est appelé *algorithme de Crout*.

On a vu au paragraphe 6.2 qu'après la $(k-1)$-ième étape de l'application de l'algorithme de Gauss (k prenant successivement les valeurs $1, 2, \ldots, n$) on avait pu déterminer une permutation τ_{k-1} de $\{1, \ldots, n\}$, une matrice triangulaire inférieure $G^{(k-1)}$ et une matrice $A^{(k)}$, telles que

$$\mathbf{1}^{\tau_{k-1}} A = G^{(k-1)} A^{(k)} .$$

La matrice $A^{(k)}$ n'est que partiellement triangulaire supérieure: les coefficients de ses $k-1$ premières colonnes situés au desssous de la diagonale sont nuls (mais ses colonnes de rang $\geq k$ sont quelconques). D'autre part, la matrice $G^{(k-1)}$ a ses colonnes de rang $\geq k$ identiquement nulles; quant à ses colonnes de rang $\leq k-1$, leurs coefficients situés sur la diagonale sont égaux à 1 et ceux situés au dessus de la diagonale sont nuls. En résumé, les matrices $A^{(k)}$ et $G^{(k-1)}$ sont de la forme

$$A^{(k)} = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,k-1} & a_{1\,k} & \ldots & a_{1\,n} \\ 0 & a_{2\,2} & \ldots & a_{2\,k-1} & a_{2\,k} & & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & a_{k-1\,k-1} & a_{k-1\,k} & \ldots & a_{k-1\,n} \\ 0 & 0 & \ldots & 0 & a_{k\,k} & \ldots & a_{k\,n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 & a_{n\,k} & \ldots & a_{n\,n} \end{pmatrix} ,$$

$$G^{(k-1)} = \begin{pmatrix} 1 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ g_{2\,1} & 1 & 0 & \ldots & & & & \vdots \\ \vdots & \ddots & \ddots & 0 & \ldots & & & \vdots \\ g_{k-1\,1} & g_{k-1\,2} & \ddots & 1 & 0 & \ldots & & \vdots \\ g_{k\,1} & g_{k\,2} & \ldots & g_{k\,k-1} & 1 & 0 & & \vdots \\ \vdots & \vdots & & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \ddots & 0 \\ g_{n\,1} & g_{n\,2} & \ldots & g_{n\,k-1} & 0 & \ldots & 0 & 1 \end{pmatrix} .$$

On a omis les indices (k) et $(k-1)$ en position haute pour alléger l'écriture.

On remarque que toute l'information contenue dans $A^{(k)}$ et dans $G^{(k-1)}$ peut être rassemblée en une seule matrice $n \times n$ notée $H^{(k)} = (h_{i\,j}^{(k)})$. Les $k-1$ premières colonnes de $H^{(k)}$ auront pour coefficients situés au dessus de la diagonale et sur la diagonale, les coefficients de $A^{(k)}$, et au dessous de la diagonale, les coefficients de $G^{(k-1)}$. Les colonnes de $H^{(k)}$ de rang $\geq k$ seront les mêmes que celles de $A^{(k)}$. En d'autres termes:

$$h_{i\,j}^{(k)} = \begin{cases} a_{i\,j}^{(k)} & \text{pour } i \leq j\,, \\ & \text{et pour } k \leq j\,,\ 1 \leq i \leq n\,, \\ g_{i\,j}^{(k-1)} & \text{pour } 1 \leq i \leq n\,,\ j < i\,,\ j \leq k-1\,. \end{cases}$$

En groupant ainsi $A^{(k)}$ et $G^{(k-1)}$, on peut réaliser un algorithme nécessitant moins d'espace mémoire. De plus (et c'est là une remarque très astucieuse, due sans doute à Crout), les permutations à appliquer à $A^{(k)}$ et à $G^{(k-1)}$ lorsqu'on effectue un pivotement, à la k-ième étape, sont les mêmes. Ce groupement permet donc de les faire effectuer par les mêmes instructions.

L'algorithme utilisera donc seulement deux variables, une matrice $n \times n$, notée H, et une permutation τ de $\{1, \ldots, n\}$. On initialise l'algorithme en attribuant à ces variables, à l'étape zéro, les valeurs

$$H = A\,, \qquad \tau = \mathrm{id}_{\{1\ldots,n\}}\,.$$

En suivant les indications données au paragraphe 6.2, on établit aisément les formules servant à transformer les valeurs attribuées à H et à τ lors des différentes étapes de l'algorithme. Après la $(n-1)$-ième étape, la valeur de τ est celle cherchée, et celle de H donne à la fois les deux termes G et D de la décomposition triangulaire de $\mathbf{1}^\tau A$: G s'obtient en prenant la partie de H strictement au dessous de la diagonale, et en ajoutant des 1 dans la diagonale; D s'obtient en prenant la partie de H située au dessus de la diagonale et sur la diagonale.

6.4. Exercices.

1. Déterminer le nombre d'opérations (divisions, multiplications, additions ou soustractions) nécessaires pour l'application de l'algorithme de Crout.

2. Retrouver directement les formules donnant la décomposition triangulaire d'une matrice inversible (ou d'une matrice qui s'en déduit par permutation de lignes) en écrivant les équations identifiant cette matrice au produit GD d'une matrice triangulaire inférieure G dont les termes diagonaux sont égaux à 1, et d'une matrice triangulaire supérieure D.

6.5. Remarque. À l'issue de l'algorithme de Crout, on a pu construire une permutation τ de $\{1, \ldots, n\}$ telle que $\mathbf{1}^\tau A$ admette une décomposition triangulaire

$$\mathbf{1}^\tau A = GD\,.$$

On en déduit

$$\det \mathbf{1}^\tau \, \det A = \det G \, det D$$

ou, puisque $\det \mathbf{1}^\tau = \epsilon(\tau)$ (signature de la permutation τ, égale à 1 si τ est paire et à -1 si τ est impaire) et que $\det G = 1$ (car G est triangulaire et a ses coefficients diagonaux égaux à 1),

$$\det A = \epsilon(\tau) \det D\,.$$

Le calcul de $\det D$ est facile: c'est le produit des coefficients diagonaux. L'algorithme de Crout permet donc le calcul du déterminant de A avec un nombre d'opérations de l'ordre de n^3. Cette méthode est bien préférable en pratique au calcul du déterminant par développement par rapport à une ligne ou une colonne, qui nécessite un nombre d'opérations de l'ordre de $n!$.

7. Matrices symétriques définies positives

On exposera plus loin une autre méthode de décomposition triangulaire applicable aux matrices d'un type particulier, les matrices symétriques définies positives. Ce cas particulier est important car on rencontre fréquemment des systèmes linéaires dont la matrice est symétrique définie positive. Supposons par exemple que l'on cherche à minimiser une fonction convexe $\Phi : \mathbf{R}^n \to \mathbf{R}$, et que l'on approche cette fonction, au voisinage de l'origine, au moyen de la formule de Taylor à l'ordre 2, par une expression quadratique de la forme

$$\Phi(X) = \Phi(0) + \langle Y, X\rangle + \frac{1}{2}\langle X, AX\rangle\,.$$

Dans cette expression, $Y = D\Phi(0)$ et $A = D^2\Phi(0)$ sont les différentielles première et seconde de Φ à l'origine. On sait que A est symétrique; de plus elle est souvent définie positive (lorsque Φ est convexe). Cette expression approchée de Φ atteint son minimum au point X solution du système linéaire

$$AX = -Y\,.$$

On donne dans le présent paragraphe quelques indications sur les matrices symétriques définies positives.

7.1. Notations et conventions. Soit E un espace vectoriel de dimension finie. On l'identifiera souvent à l'espace $\mathcal{L}(\mathbf{R}, E)$ des applications linéaires de $\mathbf{R}$ dans E, grâce à l'isomorphisme qui associe, à chaque élément x de E, l'application linéaire de $\mathbf{R}$ dans E: $t \mapsto tx$ $(t \in \mathbf{R})$. L'isomorphisme inverse est l'application qui associe, à chaque application linéaire de $\mathbf{R}$ dans E, la valeur de cette application linéaire au point 1 de $\mathbf{R}$.

On notera E^* le dual de E, c'est-à-dire l'espace $\mathcal{L}(E, \mathbf{R})$ des formes linéaires sur E (applications linéaires de E dans $\mathbf{R}$). On notera $\langle \alpha, x\rangle$, ou $\alpha(x)$, ou tout simplement αx la valeur prise par un élément α de E^* en un point x de E. L'application de $E^* \times E$ dans $\mathbf{R}$: $(\alpha, x) \mapsto \alpha x$ est appelée *couplage par dualité*.

On rappelle que le dual E^{**} de E^* s'identifie canoniquement à E.

Dans le cas où $E = \mathbf{R}$, la convention indiquée ci-dessus, consistant à identifier E à $\mathcal{L}(\mathbf{R}, E)$, nous conduit à identifier $\mathbf{R}$ à son dual $\mathbf{R}^*$: on associe à chaque $x \in \mathbf{R}$ l'application linéaire de $\mathbf{R}$ dans lui-même: $t \mapsto tx$. Le couplage par dualité de $\mathbf{R}$ avec lui-même est tout simplement le produit ordinaire $(t, x) \mapsto tx$.

7.2. Formes bilinéaires. Soient E et F deux espaces vectoriels de dimension finie. Une *forme bilinéaire* sur $E \times F$ est une application $\varphi : E \times F \to \mathbf{R}$ linéaire par rapport à chaque argument.

À toute forme bilinéaire φ sur $E \times F$, on peut associer, de manière naturelle, une application linéaire $\widetilde{\varphi} : E \to F^*$ de E dans le dual de F, en posant

$$\big\langle \widetilde{\varphi}(x), y\big\rangle = \varphi(x, y)\,, \quad x \in E\,,\ y \in F\,.$$

L'application $\varphi \mapsto \widetilde{\varphi}$ est un isomorphisme de l'espace des formes bilinéaires sur $E \times F$, sur l'espace des applications linéaires de E dans le dual F^* de F. On identifie souvent ces deux

espaces au moyen de cette application; on utilisera donc la même notation φ pour désigner une forme bilinéaire sur $E \times F$ et l'application linéaire de E dans F^* qui lui correspond.

Soit φ une forme bilinéaire sur $E \times F$. Lorsqu'on a choisi des bases $(e_1, \ldots, e_n)$ de E, $(f_1, \ldots, f_m)$ de F, on associe à la forme bilinéaire φ la matrice $A = (A_{i\,j})$, à m lignes et n colonnes, définie par

$$A_{i\,j} = \varphi(e_j, f_i)\,.$$

On remarque que A est tout simplement la matrice de φ, considérée comme application linéaire de E dans F^*, relativement à la base $(e_1, \ldots, e_n)$ de E et à la base de F^* duale de la base $(f_1, \ldots, f_m)$.

Lorsque $E = F$, une forme bilinéaire φ sur $E \times E$ est souvent appelée, par abus de langage, forme bilinéaire sur E. On dit que φ est *symétrique* si

$$\varphi(x, y) = \varphi(y, x) \quad \text{pour tous } x \text{ et } y \in E\,.$$

Cette forme bilinéaire, supposée symétrique, est dite *positive* si

$$\varphi(x, x) \geq 0 \quad \text{pourtout } x \in E\,;$$

elle est dite *définie positive* si

$$\varphi(x, x) > 0 \quad \text{pour tout } x \in E\,,\ x \neq 0\,.$$

On a, dans le paragraphe précédent, identifié $\mathbf{R}$ et son dual. Plus généralement, pour tout entier $n \geq 1$, on peut identifier $\mathbf{R}^n$ à son dual $(\mathbf{R}^n)^*$. En convenant de noter x^i $(1 \leq i \leq n)$ les composantes de chaque élément x de $\mathbf{R}^n$, cette identification est obtenue en associant à chaque élément x de $\mathbf{R}^n$ l'application linéaire de $\mathbf{R}^n$ dans $\mathbf{R}$:

$$y \mapsto \sum_{i=1}^{n} x^i y^i\,.$$

Le couplage par dualité de $\mathbf{R}^n$ avec lui-même ainsi défini, noté

$$(x, y) \mapsto (x|y) = \sum_{i=1}^{n} x^i y^i\,,$$

est appelé *produit scalaire euclidien usuel* sur $\mathbf{R}^n$. On remarquera que c'est une forme bilinéaire symétrique, et que pour tout $x \in \mathbf{R}^n$, $x \neq 0$, $(x|x) > 0$.

L'application de $\mathbf{R}^n$ dans $\mathbf{R}$:

$$x \mapsto \|x\| = (x|x)^{1/2} = \left(\sum_{i=1}^{n} (x^i)^2\right)^{1/2}$$

est appelée *norme euclidienne usuelle* sur $\mathbf{R}^n$.

7.3. Définition. Soient E et F deux espaces vectoriels réels de dimension finie et $f \in \mathcal{L}(E,F)$ une application linéaire de E dans F. On appelle *transposée* de f et on note ${}^t f$ l'application linéaire du dual F^* de F dans le dual E^* de E, définie par

$$\langle {}^t f\alpha, x\rangle = \langle \alpha, f(x)\rangle\,, \quad \alpha \in F^*\,, \quad x \in E\,.$$

7.4. Proposition. *Soient E, F et G trois espaces vectoriels réels de dimension finie, $f \in \mathcal{L}(E,F)$, $g \in \mathcal{L}(F,G)$.*

1. *On a*
$${}^t({}^t f) = f\,.$$

2. *Si f est un isomorphisme, ${}^t f$ est aussi un isomorphisme et on a*
$${}^t(f^{-1}) = ({}^t f)^{-1}\,.$$

3. *On a*
$${}^t(g \circ f) = {}^t f \circ {}^t g\,.$$

Ces propriétés, conséquences directes de la définition, pourront être démontrées par le lecteur comme exercice.

7.5. Transposée d'une matrice. Considérons le cas où $E = \mathbf{R}^m$, $F = \mathbf{R}^n$. Une application linéaire de $\mathbf{R}^m$ dans $\mathbf{R}^n$ est alors une matrice A à n lignes et m colonnes. Sa transposée ${}^t A$ est l'application linéaire de $(\mathbf{R}^n)^*$ dans $(\mathbf{R}^m)^*$, définie par

$$\langle {}^t A\alpha, x\rangle = \langle \alpha, Ax\rangle\,, \quad \alpha \in (\mathbf{R}^n)^*\,,\ x \in \mathbf{R}^m\,.$$

Conformément aux conventions indiquées en 7.2, on identifie $\mathbf{R}^m$ et $\mathbf{R}^n$ à leurs duaux respectifs. L'égalité ci-dessus s'écrit, en utilisant le produit scalaire euclidien usuel sur ces espaces:

$$({}^t A\alpha | x) = (\alpha | Ax)\,, \quad \alpha \in \mathbf{R}^n\,,\ x \in \mathbf{R}^m\,.$$

L'application ${}^t A \in \mathcal{L}(\mathbf{R}^n, \mathbf{R}^m)$ est alors une matrice à m lignes et n colonnes. Ses coefficients $({}^t A)_{i\,j}$ (l'indice i désignant la ligne et j la colonne) sont données par

$$({}^t A)_{i\,j} = ({}^t A f_j | e_i) = (f_j | Ae_i) = A_{j\,i}\,,$$

où on a noté $(f_1, \ldots, f_n)$ et $(e_1, \ldots, e_m)$ les bases canoniques, respectivement, de $\mathbf{R}^n$ et de $\mathbf{R}^m$.

On peut donc énoncer le résultat (qu'on pourrait d'ailleurs prendre pour définition):

7.6. Proposition. *La transposée d'une matrice A à n lignes et m colonnes est la matrice ${}^t A$, à m lignes et n colonnes, déduite de A par échange des lignes et des colonnes.*

7.7. Proposition. *Soient A une matrice $n \times m$ et B une matrice $m \times p$.*

1. *On a*
$$^t(^tA) = A\,.$$

2. *Si $n = m$ et si A est inversible, ^tA est aussi inversible et on a*
$$^t(A^{-1}) = (^tA)^{-1}\,.$$

3. *On a*
$$^t(AB) = {}^tB\,{}^tA\,.$$

Cette proposition est un corollaire évident de 7.4.

7.8. Produit scalaire euclidien et transposition. Lorsqu'on utilise les notations matricielles, un élément x de $\mathbf{R}^n$ est considéré comme une matrice à n lignes et 1 colonne

$$x = \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^n \end{pmatrix}\,.$$

On dit que c'est un *vecteur-colonne* à n composantes. Comme indiqué en 7.1, on l'assimile à l'application linéaire de $\mathbf{R}$ dans $\mathbf{R}^n$: $t \mapsto tx$. L'application transposée, notée ^tx, est la matrice à 1 ligne et n colonnes

$$^tx = (\,x^1 \quad x^2 \quad \dots \quad x^n\,)\;;$$

c'est une application linéaire de $\mathbf{R}^n$ dans $\mathbf{R}$ (identifiés à leurs duaux respectifs). Le produit scalaire euclidien usuel sur $\mathbf{R}^n$ peut donc s'exprimer, au moyen de la transposition, par

$$(x|y) = {}^txy = (\,x^1 \quad x^2 \quad \dots \quad x^n\,) \begin{pmatrix} y^1 \\ y^2 \\ \vdots \\ y^n \end{pmatrix} = \sum_{i=1}^{n} x^i y^i\,.$$

7.9. Définition. Soit $A \in \mathcal{L}(\mathbf{R}^n, \mathbf{R}^n)$ une matrice $n \times n$. On dit que A est:

- *symétrique* si $A = {}^tA$,
- *symétrique positive* si elle est symétrique et vérifie
$$(x|Ax) = {}^txAx \geq 0 \quad \text{pour tout } x \in \mathbf{R}^n\,,$$
- *symétrique définie positive* si elle est symétrique et vérifie
$$(x|Ax) = {}^txAx > 0 \quad \text{pour tout } x \in \mathbf{R}^n\,,\ x \neq 0\,.$$

7.10. Exemple. Soit B une matrice $n \times n$ quelconque. La matrice

$$A = {}^tBB$$

est symétrique, puisque d'après 7.7

$${}^tA = {}^t({}^tBB) = {}^tB\,{}^t({}^tB) = {}^tBB\,.$$

Elle est aussi positive puisque pour tout $x \in \mathbf{R}^n$:

$${}^txAx = {}^tx\,{}^tBBx = {}^t(Bx)(Bx) = (Bx|Bx) \geq 0\,.$$

On voit qu'elle est définie positive si et seulement si B est inversible.

7.11. Proposition. *Toute matrice $n \times n$ A symétrique définie positive est inversible et son inverse A^{-1} est symétrique définie positive. De plus, pour tout entier k $(1 \leq k \leq n)$, le mineur principal M_k de A est une matrice $k \times k$ symétrique définie positive.*

Démonstration. Puisque pour tout $x \in \mathbf{R}^n$, $x \neq 0$, on a ${}^txAx > 0$, A est injective donc inversible. D'après 7.7, A^{-1} est symétrique. Soit $z \in \mathbf{R}^n$, $z \neq 0$, et $x = A^{-1}z$. On a

$${}^tzA^{-1}z = {}^tx\,{}^tAA^{-1}Ax = {}^txAx > 0\,,$$

ce qui prouve que A^{-1} est définie positive.

Rappelons qu'on appelle mineur principal de A d'ordre k la matrice $k \times k$ M_k, formée par les coefficients a_{ij} de A dont les indices de ligne et de colonne sont $\leq k$. Les mineurs principaux de A sont évidemment des matrices symétriques. Soit

$$x = {}^t\,(\,x^1 \quad x^2 \quad \ldots \quad x^k\,)$$

un élément non nul de $\mathbf{R}^k$, et

$$y = {}^t\,(\,x^1 \quad x^2 \quad \ldots \quad x^k \quad 0 \quad \ldots \quad 0\,)$$

le vecteur élément de $\mathbf{R}^n$ dont les k premières composantes sont égales à celles de x et dont les $n-k$ autres composantes sont nulles. On remarque que x étant non nul, y aussi est non nul. Puisque A est définie positive

$${}^txM_kx = {}^tyAy > 0\,,$$

donc M_k est aussi définie positive. □

8. La méthode de Cholesky

Etablissons d'abord un lemme.

8.1. Lemme. *Soit A une matrice réelle symétrique. Toutes ses valeurs propres sont réelles. Si de plus A est définie positive, toutes ses valeurs propres et son déterminant sont strictement positifs.*

Démonstration. En prolongeant si nécessaire A en un endomorphisme de $\mathbf{C}^n$ (ce qui est toujours possible en posant, pour tous u et $v \in \mathbf{R}^n$, $A(u+iv) = A(u) + iA(v)$), on est assuré de l'existence de valeurs propres (éventuellement complexes) de A et, pour chaque valeur propre, d'un vecteur propre associé (éventuellement élément de $\mathbf{C}^n$). Soit alors λ une valeur propre de A, et $w = u + iv$ (avec u et $v \in \mathbf{R}^n$) un vecteur propre associé à la valeur propre λ. On a

$$ {}^t(u - iv)A(u + iv) = \lambda({}^t uu + {}^t vv) = \lambda(\|u\|^2 + \|v\|^2)\,. $$

Prenons les transposés conjugués (changement de i en $-i$) des deux membres. On obtient, puisque A est réelle et symétrique,

$$ {}^t(u - iv)A(u + iv) = \overline{\lambda}(\|u\|^2 + \|v\|^2)\,, $$

d'où puisque $\|u\|^2 + \|v\|^2 \neq 0$:

$$ \lambda = \overline{\lambda}\,. $$

Par suite, λ est réel, u et v vérifient tous deux

$$ Au = \lambda u\,, \quad Av = \lambda v\,, $$

et on a

$$ {}^t uAu = \lambda\|u\|^2\,, \quad {}^t vAv = \lambda\|v\|^2\,. $$

Supposons maintenant de plus A définie positive. Comme u et v ne sont pas tous deux nuls, les égalités ci-dessus montrent que

$$ \lambda > 0\,. $$

Toutes les valeurs propres de A sont donc strictement positives. Mais ce sont les racines du polynôme caractéristique de A, c'est-à-dire du déterminant de $A - \lambda I$, où I est la matrice $n \times n$ unité. D'après les relations bien connues entre coefficients d'un polynôme et fonctions symétriques des racines, le déterminant de A est le produit des valeurs propres de A (chacune apparaissant un nombre de fois égal à sa multiplicité). Le déterminant de A est donc positif. □

La méthode de Cholesky, pour la décomposition triangulaire d'une matrice symétrique définie positive, est basée sur le théorème suivant.

8.2. Théorème. *Soit A une matrice $n \times n$ symétrique. Les trois propriétés suivantes sont équivalentes.*

1. *La matrice A est définie positive.*
2. *Les déterminants des mineurs principaux M_k de A $(1 \leq k \leq n)$ sont tous > 0.*
3. *Il existe une matrice $n \times n$ inversible B telle que*

$$A = {}^tBB\,.$$

Lorsque ces trois conditions équivalentes sont satisfaites, on peut imposer de plus à la matrice B d'être triangulaire supérieure et d'avoir tous ses éléments diagonaux > 0. Moyennant quoi, la matrice B est unique.

Démonstration. Supposons A définie positive. D'après 7.11, ses mineurs principaux sont tous des matrices symétriques définies positives. D'après le lemme 8.1, leurs déterminants sont strictement positifs. On a prouvé que 1 implique 2.

Supposons les déterminants des mineurs principaux de la matrice symétrique A tous strictement positifs. D'après le théorème du paragraphe 6.1, la matrice A admet une décomposition triangulaire

$$A = GD\,,$$

avec G triangulaire inférieure, à coefficients diagonaux égaux à 1, et D triangulaire supérieure, à coefficients diagonaux donnés par les formules

$$d_{1\,1} = \Delta_1 = a_{1\,1}\,, \quad d_{k\,k} = \frac{\Delta_k}{\Delta_{k-1}} \text{ pour } 2 \leq k \leq n\,.$$

On a désigné par Δ_k le déterminant du mineur principal de A d'ordre k. Les coefficients diagonaux de D sont donc tous strictement positifs.

Soit S la matrice diagonale ayant pour coefficients

$$s_{k\,k} = \frac{1}{\sqrt{d_{k\,k}}}\,.$$

Posons

$$G' = GS^{-1}\,, \qquad B = SD\,.$$

Les matrices G' et B sont triangulaires, respectivement inférieure et supérieure. Elles vérifient

$$A = G'B\,, \tag{$*$}$$

et ont pour coefficients diagonaux

$$g'_{k\,k} = b_{k\,k} = \sqrt{d_{k\,k}}\,.$$

En prenant les transposés des deux membres de $(*)$ on obtient, puisque A est symétrique,

$$A = G'B = {}^tB\,{}^tG'\,,$$

d'où puisque G' et B sont inversibles

$$({}^tB)^{-1}G' = {}^tG'({}^tB)^{-1}\,.$$

Mais la matrice au premier membre de l'égalité ci-dessus est triangulaire inférieure, celle au second membre triangulaire supérieure, donc toutes deux sont diagonales. L'égalité des termes diagonaux de G' et de B montre que cette matrice diagonale est en fait la matrice unité. On en déduit

$$G' = {}^tB\,.$$

On a ainsi prouvé que 2 implique 3.

D'après l'exemple 7.10, 3 implique 1. On a donc prouvé l'équivalence des propriétés 1, 2 et 3.

Enfin on a vu, en démontrant que 2 impliquait 3, que la matrice B pouvait être choisie triangulaire supérieure, à coefficients diagonaux tous strictement positifs. Le même raisonnement que celui fait pour montrer que $G' = {}^tB$ montre que la matrice B est alors unique. □

8.3. L'algorithme de Cholesky. Soit $A = (a_{ij})$ une matrice $n \times n$ définie positive. D'après le théorème 8.1, il existe une matrice triangulaire supérieure unique B à coefficients diagonaux strictement positifs telle que

$$A = {}^tBB\,. \tag{$*$}$$

L'algorithme de Cholesky permet la détermination explicite de B. Compte tenu de

$$a_{ij} = a_{ji}$$

et de

$$b_{ij} = 0 \quad \text{si } i > j\,,$$

l'égalité $(*)$ s'écrit, en explicitant les composantes,

$$a_{ij} = \sum_{k=1}^{i} b_{ki} b_{kj}\,, \quad 1 \le i \le j \le n\,. \tag{$**$}$$

Dans une première étape, on fait $i = 1$. Les équations $(**)$ nous donnent d'abord, pour $j = 1$,

$$a_{11} = (b_{11})^2\,,$$

d'où puisque $b_{11} > 0$,

$$b_{11} = \sqrt{a_{11}}\,.$$

Puis ces équations nous donnent, pour $2 \le j \le n$,

$$a_{1j} = b_{11}\, b_{1j}\,,$$

d'où, puisque b_{11} a déjà été déterminé,

$$b_{1j} = \frac{a_{1j}}{b_{11}}\,.$$

On procède ainsi étape par étape, chaque étape correspondant à une valeur de l'indice i. Après la $i-1$-ième étape, les $b_{k\,j}$ ont été déterminés pour $1 \leq k \leq i-1$, $k \leq j \leq n$. A la i-ème étape, les équations $(**)$ nous donnent, pour $j = i$,

$$a_{i\,i} = \sum_{k=1}^{i} (b_{k\,i})^2 \,,$$

d'où puisque les $b_{k\,i}$ sont déjà connus pour $1 \leq k \leq i-1$, et que $b_{i\,i} > 0$,

$$b_{i\,i} = \sqrt{a_{i\,i} - \sum_{k=1}^{i-1} (b_{k\,i})^2} \,.$$

Le théorème 8.2 nous permet d'affirmer que l'expression figurant sous le radical est strictement positive. Puis les équations $(**)$ nous donnent, pour $i+1 \leq j \leq n$,

$$a_{i\,j} = \sum_{k=1}^{i} b_{k\,i} b_{k\,j} \,,$$

d'où, puisque dans cette expression le seul terme pas encore connu est $b_{i\,j}$,

$$b_{i\,j} = \frac{a_{i\,j} - \sum_{k=1}^{i-1} b_{k\,i} b_{k\,j}}{b_{i\,i}} \,.$$

On remarque que l'algorithme permet en même temps de voir si la matrice A est définie positive: si elle ne l'est pas, on obtient, à une certaine étape, une expression pour le calcul de $b_{i\,i}$ comportant la racine carrée d'une quantité non strictement positive (nulle ou négative). On ne peut alors plus appliquer l'algorithme.

8.4. Exercice. Calculer le nombre d'opérations (additions ou soustractions, multiplications, extractions de racines carrées) nécessaires pour obtenir une décomposition triangulaire d'une matrice $n \times n$ symétrique définie positive par la méthode de Cholesky.

On trouve:

$\dfrac{(n-1)n(n+1)}{6}$ additions ou soustractions,

$\dfrac{(n-1)n(n+1)}{6}$ multiplications,

$\dfrac{(n-1)n}{2}$ divisions,

n extractions de racines carrées.

9. Fonctions symétriques des racines d'un polynôme

Les méthodes de Leverrier et de Souriau, exposées plus loin, permettent le calcul des coefficients du polynôme caractéristique d'une matrice à n lignes et n colonnes. On va rappeler d'abord quelques propriétés des fonctions symétriques des racines d'un polynôme, qui seront utilisées lors de l'exposé de ces méthodes.

9.1. Fonctions symétriques élémentaires. On considère un polynôme f, de degré n, à coefficients réels ou complexes,

$$f(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n\,. \qquad (*)$$

On a $a_0 \neq 0$, puisque f est supposé de degré n.

Soient $x_1, x_2, \ldots, x_n$ les racines de ce polynôme, chaque racine étant répétée un nombre de fois égal à sa multiplicité; grâce à cette convention, f a effectivement n racines. On a

$$f(x) = a_0(x - x_1)(x - x_2)\cdots(x - x_n)\,. \qquad (**)$$

Pour chaque entier k $(1 \leq k \leq n)$, on note σ_k la somme des produits de toutes les sous-familles de k termes pris parmi $x_1, x_2, \ldots, x_n$. On a donc

$$\begin{aligned}
&\sigma_1 = x_1 + x_2 + \cdots + x_n\,,\\
&\sigma_2 = x_1x_2 + x_1x_3 + \cdots + x_{n-1}x_n\,,\\
&\cdots \quad \cdots\\
&\sigma_n = x_1x_2\cdots x_n\,.
\end{aligned}$$

Les σ_k sont appelées *fonctions symétriques élémentaires* des racines du polynôme f.

Il existe des relations simples entre les coefficients a_k du polynôme f et les fonctions symétriques élémentaires, qu'on détermine aisément en effectuant le produit qui figure au second membre de $(**)$ et en identifiant l'expression ainsi trouvée avec le second membre de $(*)$. Ces relations sont:

$$\sigma_k = (-1)^k\,\frac{a_k}{a_0}\,.$$

D'autre part, pour tout entier k, on note S_k la somme des puissances k-ièmes des racines de f:

$$S_k = (x_1)^k + (x_2)^k + \cdots + (x_n)^k\,.$$

9.2. Les relations de Newton. Les coefficients a_k du polynôme f et les sommes S_k des puissances k-ièmes des racines de ce polynôme sont liés par des relations, appelées *relations de Newton*. Les n premières relations de Newton s'écrivent:

$$\begin{array}{llllllllllllll}
1 & a_1 & + & a_0 & S_1 & = & 0\,, & & & & & & & \\
2 & a_2 & + & a_1 & S_1 & + & a_0 & S_2 & = & 0\,, & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & & & & & \\
k & a_k & + & a_{k-1} & S_1 & + & a_{k-2} & S_2 & + & \cdots & + & a_0 & S_k & = & 0\,, \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & & & & & \\
n & a_n & + & a_{n-1} & S_1 & + & a_{n-2} & S_2 & + & \cdots & + & a_0 & S_n & = & 0\,.
\end{array}$$

Pour $k > n$, les relations de Newton s'écrivent:

$$a_n S_{k-n} + a_{n-1} S_{k-n+1} + \cdots + a_0 S_k = 0\,.$$

On remarque que pour tout entier $k \geq 1$, k pouvant être $\leq n$ ou $> n$, le système formé par les k premières relations de Newton permet très facilement la détermination explicite de S_1, S_2, ..., S_k en fonction des coefficients du polynôme f (le système linéaire correspondant est à matrice triangulaire). Inversement, ces mêmes équations (mais pour $1 \leq k \leq n$) permettent très facilement la détermination explicite de a_1, a_2, ..., a_k en fonction de a_0, S_1, S_2, ..., S_k.

Rappelons une démonstration des relations de Newton. On a

$$\begin{aligned} f(x) &= a_0 x^n + a_1 x^{n-1} + \cdots + a_n \\ &= a_0 (x - x_1)(x - x_2) \cdots (x - x_n)\,. \end{aligned}$$

Dérivons la première expression de f. On obtient

$$f'(x) = n a_0 x^{n-1} + (n-1) a_1 x^{n-2} + \cdots + a_{n-1}\,.$$

En utilisant la seconde expression de f, et en calculant sa dérivée logarithmique, on obtient

$$\frac{f'(x)}{f(x)} = \frac{1}{x - x_1} + \frac{1}{x - x_2} + \cdots + \frac{1}{x - x_n}\,,$$

d'où

$$f'(x) = \frac{f(x)}{x - x_1} + \frac{f(x)}{x - x_2} + \cdots + \frac{f(x)}{x - x_n}\,.$$

Les quotients $\dfrac{f(x)}{x - x_i}$ peuvent être calculés en effectuant la division euclidienne suivant les puissances décroissantes de la variable. On obtient

$$\frac{f(x)}{x - x_i} = a_0 x^{n-1} + (a_1 + a_0 x_i) x^{n-2} + \big(a_2 + (a_1 + a_0 x_i) x_i\big) x^{n-3} + \cdots$$

En ajoutant les termes pour i allant de 1 à n, on obtient

$$\begin{aligned} f'(x) = n a_0 x^{n-1} &+ (n a_1 + a_0 S_1) x^{n-2} + (n a_2 + a_1 S_1 + a_0 S_2) x^{n-3} + \cdots \\ &+ (n a_{n-1} + a_{n-2} S_1 + a_{n-3} S_2 + \cdots + a_0 S_{n-1})\,. \end{aligned}$$

En identifiant les termes de même degré dans les deux expressions de $f'(x)$ obtenues ci-dessus, on obtient

$$\begin{aligned} n a_0 &= n a_0\,, \\ (n-1) a_1 &= n a_1 + a_0 S_1\,, \\ (n-2) a_2 &= n a_2 + a_1 S_1 + a_0 S_2\,, \\ \cdots &\quad \cdots \\ a_{n-1} &= n a_{n-1} + a_{n-2} S_1 + \cdots + a_0 S_{n-1}\,. \end{aligned}$$

La première de ces égalités est une identité. Les autres sont les $n-1$ premières relations de Newton. Pour obtenir la k-ième relation de Newton, avec $k \geq n$, on écrit

$$(x_i)^{k-n}\big(a_n + a_{n-1}x_i + a_{n-2}(x_i)^2 + \cdots + a_0(x_i)^n\big) = 0\,.$$

En développant et en ajoutant, pour i allant de 1 à n, on obtient

$$a_n S_{k-n} + a_{n-1}S_{k-n+1} + \cdots + a_0 S_k = 0\,.$$

Ce sont les relations de Newton de rang $k \geq n$.

10. La méthode de Leverrier

Soit $A = (a_{i\,j})$ une matrice $n \times n$. On rappelle que le *polynôme caractéristique* de A est le déterminant de la matrice $A - xI$, I désignant la matrice $n \times n$ unité, et x la variable dans ce polynôme. On le notera

$$P_A(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_n\,.$$

C'est un polynôme de degré n. En évaluant le coefficient du terme de plus haut degré en x, on voit aisément que

$$a_0 = (-1)^n\,.$$

On notera $x_1, \ldots, x_n$ les racines de $P_A(x)$.

La méthode de Leverrier pour la détermination de $P_A(x)$ repose sur le lemme suivant.

10.1. Lemme. *Pour tout entier $k \geq 1$, la somme*

$$S_k = (x_1)^k + (x_2)^k + \cdots + (x_n)^k$$

est égale à la trace de la matrice A^k.

Démonstration. La matrice A est un endomorphisme de $\mathbf{R}^n$, qui se prolonge de manière naturelle en un endomorphisme de $\mathbf{C}^n$, qu'on notera $\mathcal{A}$ pour éviter les confusions. Soit P une matrice $n \times n$ inversible, dont les coefficients peuvent éventuellement être complexes. Les colonnes de P forment une nouvelle base de $\mathbf{C}^n$, dans laquelle l'endomorphisme $\mathcal{A}$ a pour matrice

$$B = P^{-1}AP\,.$$

On sait que quelle que soit la matrice inversible P, on a

$$\mathrm{Trace}(B) = \mathrm{Trace}(A)\,,$$

et que, pour un choix convenable de P, la matrice B est triangulaire inférieure, et a pour coefficients diagonaux les valeurs propres de A, c'est-à-dire les racines $x_1, \ldots, x_n$ de son polynôme caractéristique P_A. Pour tout entier $k \geq 1$, on a

$$B^k = P^{-1}A^kP\,, \quad \mathrm{Trace}(B^k) = \mathrm{Trace}(A^k)\,.$$

Mais B étant triangulaire inférieure, B^k l'est aussi, et ses termes diagonaux sont $(x_1)^k, \ldots,$ $(x_n)^k$. Par suite,

$$\mathrm{Trace}(A^k) = \mathrm{Trace}(B^k) = S_k\,. \qquad \square$$

10.2. L'algorithme de Leverrier. On connaît déjà le coefficient a_0 du polynôme P_A:

$$a_0 = (-1)^n\,.$$

A la première étape, on calcule

$$S_1 = \mathrm{Trace}(A)\,.$$

La première relation de Newton donne alors

$$a_1 = -a_0 S_1 = -(-1)^n \mathrm{Trace}(A)\,.$$

Après la $(k-1)$-ième étape, on connaît $S_1, \ldots, S_{k-1}$, A^{k-1}, ainsi que $a_0, a_1, \ldots, a_{k-1}$. On calcule alors

$$A^k = AA^{k-1}\,,$$

puis

$$S_k = \mathrm{Trace}(A^k)\,.$$

La k-ième relation de Newton donne alors

$$a_k = -\frac{1}{k}(a_{k-1}S_1 + a_{k-2}S_2 + \cdots + a_0 S_k)\,.$$

10.3. Exercice. Calculer le nombre d'opérations (additions ou soustractions, multiplications, divisions) nécessaires pour déterminer le polynôme caractéristique d'une matrice $n \times n$ par la méthode de Leverrier.

11. La méthode de Souriau

Cette méthode permet le calcul, entre autres choses, du polynôme caractéristique d'une matrice $n \times n$ et, lorsque cette matrice est inversible, de son inverse. Nous allons d'abord rappeler quelques propriétés de la matrice des cofacteurs utilisées par cette méthode.

11.1. Rappel sur la matrice des cofacteurs. Soit $A = (a_{i\,j})$ une matrice $n \times n$. Pour tout couple (i, j), on désigne par $\Delta_{i\,j}$ le déterminant de la matrice $(n-1) \times (n-1)$ obtenue à partir de A en supprimant la i-ème ligne et la j-ème colonne.

Développons le déterminant de A par rapport à la i-ème ligne de cette matrice. On obtient (i étant fixé):

$$\det A = \sum_{k=1}^{n} (-1)^{i+k} a_{i\,k} \Delta_{i\,k} \,. \tag{1}$$

On en déduit, pour tout couple (i, j) fixé $(1 \leq i, j \leq n)$:

$$\sum_{k=1}^{n} (-1)^{j+k} a_{i\,k} \Delta_{j\,k} = \det A \delta_{i\,j} = \begin{cases} \det A & \text{si } i = j \,, \\ 0 & \text{si } i \neq j \,. \end{cases} \tag{2}$$

En effet, pour $i = j$, c'est l'expression de $\det A$ donnée en (1); pour $i \neq j$, c'est le développement, par rapport à la i-ème ligne, du déterminant d'une matrice déduite de A en remplaçant la i-ème ligne de A par sa j-ème ligne; mais une telle matrice, ayant ses i-ème et j-ème lignes identiques, a un déterminant nul.

De même, en développant le déterminant de A par rapport à la j-ème colonne de cette matrice, on montre que pour tout couple (i, j) fixé, $(1 \leq i, j \leq n)$,

$$\sum_{k=1}^{n} (-1)^{j+k} a_{k\,i} \Delta_{k\,j} = \det A \, \delta_{i\,j} = \begin{cases} \det A & \text{si } i = j \,, \\ 0 & \text{si } i \neq j \,. \end{cases} \tag{3}$$

Afin de donner une forme plus commode aux expressions (2) et (3), on pose

$$\widehat{a}_{i\,j} = (-1)^{i+j} \Delta_{j\,i} \,.$$

On remarquera l'interversion des indices i et j entre les membres de droite et de gauche de cette égalité.

La matrice $\widehat{A} = (\widehat{a}_{i\,j})$, où i est l'indice de ligne et j l'indice de colonne, est appelée *matrice des cofacteurs* de A.

Grâce à cette définition, les formules (2) et (3) s'écrivent, au moyen du produit des matrices A et $\widehat{A}$,

$$A\widehat{A} = \widehat{A}A = \det A \, I_n \,,$$

où I_n désigne la matrice $n \times n$ unité.

On en déduit immédiatement que si $\det A \neq 0$, la matrice A est inversible et a pour inverse

$$A^{-1} = \frac{1}{\det A} \, \widehat{A} \,.$$

La matrice des cofacteurs de A intervient aussi dans l'expression de la différentielle de l'application "déterminant", qui, à une matrice A, associe son déterminant $\det A$. Pour le voir, reprenons la formule (1):

$$\det A = \sum_{k=1}^{n} (-1)^{i+k} a_{i\,k} \Delta_{i\,k} \,.$$

Chaque $\Delta_{i\,k}$ est un polynôme homogène de degré $n-1$ en les coefficients $a_{r\,s}$ de A, avec $r \neq i$ et $s \neq k$ puisque c'est le déterminant d'une matrice obtenue en supprimant la i-ème ligne et la k-ième colonne de A. On a donc

$$\frac{\partial(\det A)}{\partial a_{i\,j}} = (-1)^{i+j} \Delta_{i\,j} \,,$$

qu'on peut aussi écrire, en faisant intervenir le cofacteur $\widehat{a}_{i\,j}$,

$$\frac{\partial(\det A)}{\partial a_{i\,j}} = \widehat{a}_{j\,i} \,.$$

La différentielle de l'application "déterminant" est donc:

$$\begin{aligned} d(\det A) &= \sum_{(i,j)} \frac{\partial(\det A)}{\partial a_{i\,j}} \, da_{i\,j} = \sum_{(i,j)} \widehat{a}_{j\,i} \, da_{i\,j} \\ &= \mathrm{Trace}(\widehat{A}\, dA) = \mathrm{Trace}(dA\, \widehat{A}) \,. \end{aligned}$$

[On rappelle que la *trace* d'une matrice A, notée $\mathrm{Trace}\, A$, est la somme de ses coefficients diagonaux.]

11.2. Application à la détermination du polynôme caractéristique. Soit s une variable (qu'on pourra supposer élément de $\mathbf{R}$ ou de $\mathbf{C}$). A désigne comme précédemment une matrice $n \times n$ donnée. On sait que la matrice $(sI_n - A)$, fonction de la variable s, a pour déterminant $(-1)^n P_A(s)$, où $P_A(s)$ est le polynôme caractéristique de la matrice A. On sait que $(-1)^n P_A(s)$ est un polynôme en s de degré n, dont le terme de plus haut degré est s^n. On l'écrira:

$$(-1)^n P_A(s) = \det(sI_n - A) = s^n + d_1 s^{n-1} + \cdots + d_n \,,$$

les coefficients $d_1, \ldots, d_n$ étant des scalaires. Avec les notations du paragraphe 10, on a $d_k = (-1)^n a_k$, $1 \leq k \leq n$.

Soit d'autre part

$$Q_A(s) = \widehat{(sI_n - A)}$$

la matrice des cofacteurs de la matrice $(sI_n - A)$. On peut considérer $Q_A(s)$ comme un polynôme de degré $n-1$ en s dont les coefficients sont des matrices $n \times n$. On l'écrit:

$$Q_A(s) = B_0 s^{n-1} + B_1 s^{n-2} + \cdots + B_{n-1} \,,$$

où $B_0, \ldots, B_{n-1}$ sont des matrices $n \times n$.

D'après les résultats du paragraphe précédent, appliqués à $(sI_n - A)$, et non plus à A, on a

$$(sI_n - A)Q_A(s) = Q_A(s)(sI_n - A) = P_A(s)I_n\,.$$

On rappelle que le *spectre* de A est l'ensemble des racines de son polynôme caractéristique $P_A(s)$. Donc, si s n'est pas élément du spectre de A, $P_A(s)$ est non nul, la matrice $(sI_n - A)$ est inversible et a pour inverse

$$(sI_n - A)^{-1} = \frac{1}{P_A(s)}\,Q_A(s)\,.$$

11.3. Théorème (formules de Souriau). *Les coefficients $B_0, \ldots, B_{n-1}$ du polynôme à coefficients matriciels $Q_A(s)$, et les coefficients $d_1, \ldots, d_n$ du polynôme $(-1)^n P_A(s)$, où $P_A(s)$ est le polynôme caractéristique de la matrice A, sont liés par les relations:*

$$\begin{aligned}
B_0 &= I_n\,, & \operatorname{Trace}(B_0A) &= -d_1\,,\\
B_1 &= B_0A + d_1I_n\,, & \operatorname{Trace}(B_1A) &= -d_2\,,\\
B_2 &= B_1A + d_2I_n\,, & \operatorname{Trace}(B_2A) &= -3d_3\,,\\
&\cdots\cdots & &\cdots\cdots\\
B_k &= B_{k-1}A + d_kI_n\,, & \operatorname{Trace}(B_kA) &= -(k+1)d_{k+1}\,,\\
&\cdots\cdots & &\cdots\cdots\\
B_{n-1} &= B_{n-2}A + d_{n-1}I_n\,, & \operatorname{Trace}(B_{n-1}A) &= -nd_n\,,\\
0 &= B_{n-1}A + d_nI_n\,.
\end{aligned}$$

Les $2n$ premières formules permettent de calculer successivement B_0, d_1, B_1, d_2, …, B_k, d_{k+1}, …, B_{n-1}, d_n. La dernière formule $0 = B_{n-1}A + d_nI_n$ n'est pas nécessaire au calcul des coefficients B_i, d_j; elle peut être employée comme vérification.

Démonstration. On sait que

$$Q_A(s)(sI_n - A) = (-1)^n P_A(s)I_n\,,$$

ou, en tenant compte des expressions de $Q_A(s)$ et de $(-1)^n P_A(s)$,

$$(B_0s^{n-1} + B_1s^{n-2} + \cdots + B_{n-1})(sI_n - A) = (s^n + d_1s^{n-1} + \cdots + d_n)I_n\,.$$

En développant le membre de gauche et en identifiant les termes de même degré des deux membres, on obtient:

$$\begin{aligned}
B_0 &= I_n\,,\\
B_1 &= B_0A + d_1I_n\,,\\
B_2 &= B_1A + d_2I_n\,,\\
&\cdots\cdots\\
B_k &= B_{k-1}A + d_kI_n\,,\\
&\cdots\cdots\\
B_{n-1} &= B_{n-2}A + d_{n-1}I_n\,,\\
0 &= B_{n-1}A + d_nI_n\,.
\end{aligned}$$

Ce sont les formules de Souriau indiquées dans l'énoncé dans la colonne de gauche.

Prenons les traces des deux membres de chacune de ces formules. Nous obtenons:

$$\begin{aligned}
\operatorname{Trace} B_0 &= n\,,\\
\operatorname{Trace} B_1 &= \operatorname{Trace}(B_0 A) + nd_1\,,\\
\operatorname{Trace} B_2 &= \operatorname{Trace}(B_1 A) + nd_2\,,\\
&\cdots\cdots\\
\operatorname{Trace} B_k &= \operatorname{Trace}(B_{k-1} A) + nd_k\,,\\
&\cdots\cdots\\
\operatorname{Trace} B_{n-1} &= \operatorname{Trace}(B_{n-2} A) + nd_{n-1}\,,\\
0 &= \operatorname{Trace}(B_{n-1} A) + nd_n\,.
\end{aligned} \tag{4}$$

D'autre part, on a établi au paragraphe 11.1 la formule donnant l'expression de la différentielle de l'application "déterminant":

$$d(\det A) = \operatorname{Trace}(\widehat{A}\, dA)\,.$$

Appliquons cette formule, non pas à la matrice A, mais à la matrice $(sI_n - A)$, considérée comme fonction de s. Son déterminant étant $(-1)^n P_A(s)$, et la matrice de ses cofacteurs étant $Q_A(s)$, on obtient

$$\begin{aligned}
(-1)^n \frac{d}{ds} P_A(s) &= \operatorname{Trace}\big(Q_A(s) \frac{d}{ds}(sI_n - A)\big)\\
&= \operatorname{Trace}\big(Q_A(s) I_n\big)\\
&= \operatorname{Trace}\big(Q_A(s)\big)\,.
\end{aligned}$$

En remplaçant $(-1)^n P_A(s)$ et $Q_A(s)$ par leurs expressions, on obtient

$$ns^{n-1} + (n-1)d_1 s^{n-2} + \cdots + d_{n-1} = \operatorname{Trace} B_0 s^{n-1} + \operatorname{Trace} B_1 s^{n-2} + \cdots + \operatorname{Trace} B_{n-1}\,.$$

En égalant les termes de même degré en s des deux membres, on obtient

$$\begin{aligned}
\operatorname{Trace} B_0 &= n\,,\\
\operatorname{Trace} B_1 &= (n-1)d_1\,,\\
&\cdots\cdots\\
\operatorname{Trace} B_{n-1} &= d_{n-1}\,.
\end{aligned}$$

En portant ces expressions des $\operatorname{Trace} B_i$ dans les formules (4), on obtient les formules de Souriau figurant dans la colonne de droite de l'énoncé. □

11.4. Remarque. La méthode de Souriau permet la détermination, non seulement du polynôme caractéristique $P_A(s)$ de la matrice A, mais aussi celle de tous les coefficients matriciels du polynôme $Q_A(s)$ (matrice des cofacteurs de $(sI_n - A)$). On en déduit aisément l'inverse de $(sI_n - A)$ pour tout s n'appartenant pas au spectre de A et, en particulier, l'inverse de A (lorsque A est inversible).

11.5. Exercice. Déterminer le nombre d'opérations (additions ou soustractions, multiplications, divisions) nécessaires pour l'application de la méthode de Souriau à une matrice $n \times n$. On supposera que les produits de matrices sont effectués par la méthode de la "force brutale", lignes par colonnes, sans que l'on cherche à réduire le nombre de multiplications par des regroupements de termes (ce qui est possible, au prix d'un plus grand nombre d'additions et de soustractions, mais c'est un autre exercice).

On trouve

$n(n-1)\big(2+n(n-1)\big)$ additions ou soustractions,

$(n-1)n^3$ multiplications,

$n-1$ divisions.

Si l'on utilise la dernière équation pour faire une vérification des calculs, on doit faire encore

n additions,

et de plus comparer à 0 les n^2 termes d'une matrice.

On remarque que dans cette méthode le nombre d'opérations croît comme n^4.

12. Le théorème de Sturm

Le théorème de Sturm est à la base d'une méthode de détermination du nombre de racines réelles d'une équation algébrique contenues dans un intervalle donné de $\mathbf{R}$. Il utilise comme ingrédient essentiel l'algorithme d'Euclide pour la détermination du PGCD (plus grand commun diviseur) de deux polynômes, étudié en premier cycle, que nous allons brièvement rappeler.

12.1. L'algorithme d'Euclide. Cet algorithme est l'un des plus anciennement connus, puisqu'on le trouve dans les fameux "Éléments" d'Euclide, datant probablement du troisième siècle avant J. C. Initialement créé pour la détermination du PGCD (plus grand commun diviseur) de deux entiers positifs, il s'applique sans changement à la détermination du PGCD de deux polynômes. C'est le cas qui nous intéresse ici.

Soient P_0 et P_1 deux polynômes en la variable x, dont les degrés, notés $\deg P_0$ et $\deg P_1$, vérifient $\deg P_0 \geq \deg P_1$. On effectue la division de P_0 par P_1, suivant les puissances décroissantes de la variable. On a

$$P_0 = P_1 Q + R\,, \quad \text{avec } \deg R < \deg P_1\,,$$

où Q désigne le quotient de P_0 par P_1 et R le reste. Si $R = 0$, le PGCD de P_0 et P_1 est P_1. Si $R \neq 0$, on remarque que tout polynôme qui divise à la fois P_0 et P_1 divise aussi R, et que tout polynôme qui divise à la fois P_1 et R divise aussi P_0. Le PGCD de P_0 et P_1 est donc le même que le PGCD de P_1 et R. On peut répéter, pour P_1 et R, le même raisonnement que celui présenté ci-dessus pour P_0 et P_1. On est ainsi conduit à la construction suivante, appelée *algorithme d'Euclide*. On définit une suite finie de polynômes P_i, i entier ≥ 0, dont les deux premiers termes sont les polynômes donnés P_0 et P_1, dont les degrés, à partir du troisième terme, forment une suite strictement décroissante. En supposant que les termes

de cette suite ont été définis jusqu'à P_i, on effectue la division de P_{i-1} par P_i suivant les puissances décroissantes de la variable. On désigne le quotient par Q_{i+1} et le reste par R_{i+1}. On écrit donc

$$P_{i-1} = P_i Q_{i+1} + R_{i+1}\,, \quad \text{avec } \deg R_{i+1} < \deg P_i\,.$$

Si $R_{i+1} = 0$, la suite s'arrête à P_i. Si $R_{i+1} \neq 0$, on pose $P_{i+1} = R_{i+1}$, et on répète pour P_i et P_{i+1} le opérations décrites ci-dessus pour P_{i-1} et P_i. L'algorithme ainsi défini s'arrête nécessairement, car les degrés des polynômes successivement définis forment une suite strictement décroissante. Le dernier terme de la suite, P_k, est le PGCD de P_0 et de P_1.

Exercice. On note P_k le PGCD de P_0 et P_1. Montrer que l'algorithme d'Euclide permet de déterminer non seulement P_k, mais aussi deux polynômes S_0 et S_1 tels que

$$P_k = P_0 S_0 + P_1 S_1\,.$$

En particulier, si P_0 et P_1 sont premiers entre eux, P_k est un scalaire non nul, et il existe deux polynômes U et V tels que

$$1 = P_0 U + P_1 V\,.$$

L'algorithme d'Euclide permet la détermination effective de U et V.

12.2. La suite de Sturm d'un polynôme. Soit P un polynôme de degré n en la variable x, à coefficients réels. On le note:

$$P(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_n\,,$$

avec, par hypothèse, $a_0 \neq 0$. Soit P' son polynôme dérivé:

$$P'(x) = n a_0 x^{n-1} + (n-1) a_1 x^{n-2} + \cdots + a_{n-1}\,.$$

On applique l'algorithme d'Euclide pour la détermination du PGCD de P et de P'. On va donc définir une suite finie de polynômes $(P_0, P_1, \ldots, P_k)$, dont les deux premiers sont $P_0 = P$, $P_1 = P'$. Pour des raisons qui apparaîtront plus loin, on introduit cependant une légère modification de cet algorithme, que l'on va décrire. À la première étape, on effectue la division de $P_0 = P$ par $P_1 = P'$, suivant les puissances décroissantes de la variable. On a donc

$$P_0 = P_1 Q + R\,, \quad \text{avec } \deg R < \deg P_1\,.$$

Si $R = 0$, la suite s'arrête à P_1. Si $R \neq 0$, on pose

$$P_2 = -R\,,$$

au lieu de poser, comme dans l'algorithme d'Euclide standard, $P_2 = R$. De même, à la i-ème étape, les termes de la suite ayant été définis jusqu'à P_i, on effectue la division de P_{i-1} par P_i suivant les puissances décroissantes de la variable. On écrit donc

$$P_{i-1} = P_i Q_{i+1} + R_{i+1}\,, \quad \text{avec } \deg R_{i+1} < \deg P_i\,.$$

Si le reste R_{i+1} de cette division est nul, la suite s'arrête à P_i. S'il est non nul, on pose

$$P_{i+1} = -R_{i+1}\,,$$

et non pas R_{i+1} comme dans l'algorithme d'Euclide standard. On a donc, pour tout i, $1 \leq i \leq k-1$,

$$P_{i-1} = P_i Q_{i+1} - P_{i+1}\,, \quad \text{avec } \deg P_{i+1} < \deg P_i\,.$$

La suite de polynômes ainsi construite, ($P_0 = P$, $P_1 = P'$, ..., P_k), sera appelée *suite de Sturm* du polynôme P. Son dernier terme, P_k, est le PGCD de P et de P': les changements de signe qu'on a faits dans la variante de l'algorithme d'Euclide utilisée ici n'empêchent pas qu'on trouve quand même ainsi le PGCD, qui d'ailleurs n'est défini qu'à un facteur scalaire multiplicatif non nul près. Le dernier terme P_k est une constante non nulle si et seulement si P et son polynôme dérivé P' sont premiers entre eux, c'est-à-dire si et seulement si toutes les racines du polynôme P sont simples.

Pour tout point x de $\mathbf{R}$ qui n'est pas racine de P, on appelle *nombre de changements de signes* de la suite de Sturm de P au point x, le nombre de changements de signes entre deux termes consécutifs de la suite $\big(P_0(x), P_1(x), \ldots, P_k(x)\big)$, lorsqu'on considère les termes de cette suite en commençant par $P_0(x)$, qui est non nul par hypothèse. Afin que ce nombre de changements de signes soit bien défini, même lorsque certains des $P_i(x)$ (pour $i \geq 1$) sont nuls, on convient de considérer que lorsque $P_i(x)$ est nul, son signe est le même que celui du terme précédent $P_{i-1}(x)$.

12.3. Théorème de Sturm. *Soit P un polynôme en la variable x, à coefficients réels. On suppose que toutes les racines de P sont simples. Pour tout $x \in \mathbf{R}$ tel que $P(x) \neq 0$, on note $v(x)$ le nombre de changements de signes de la suite de Sturm de P au point x. Soient a et b deux réels vérifiant $a < b$, aucun des deux n'étant racine de P. Le nombre de racines de P contenues dans l'intervalle ouvert $]a, b[$ est égal à $v(a) - v(b)$.*

Démonstration. Soit $(P_0, P_1, \ldots, P_k)$ la suite de Sturm de P. Comme on a supposé que toutes les racines de P sont simples, P_k est une constante non nulle, et deux polynômes consécutifs P_{i-1} et P_i de cette suite, pour $1 \leq i \leq k-1$, sont premiers entre eux. Lorsque x varie dans un intervalle ne contenant pas de racines des polynômes P_i ($0 \leq i \leq k-1$), les termes de la suite de Sturm gardent des signes constants, donc $v(x)$ reste constant. Pour étudier les variations de $v(x)$, il suffit donc d'examiner ce qui se passe lorsque x traverse, en croissant, un réel r qui est racine d'un ou de plusieurs termes de la suite de Sturm.

Supposons r racine de $P_0 = P$. Alors r n'est pas racine de $P_1 = P'$, puisque P_0 et P_1 sont premiers entre eux. Considérons successivement deux cas:

(i) Si $P_1(r) > 0$, P est croissant au voisinage de r. Donc pour $\epsilon > 0$ assez petit, $P_0(x-\epsilon) < 0$ et $P_0(x+\epsilon) > 0$. En $x - \epsilon$, la suite de Sturm de P comporte donc un changement de signe entre les deux premiers termes P_0 et P_1, tandis qu'en $x + \epsilon$, les deux premiers termes de cette suite sont de même signe.

(ii) Si $P_1(r) < 0$, P est décroissant au voisinage de r. Donc pour $\epsilon > 0$ assez petit, $P_0(x-\epsilon) > 0$ et $P_0(x+\epsilon) < 0$. Donc comme dans le cas précédent, entre les deux premiers termes P_0 et P_1, il y a un changement de signe en $x - \epsilon$ et il n'y en a pas en $x + \epsilon$.

Supposons maintenant r racine de P_i, pour un certain i vérifiant $1 \leq i \leq k-1$, (r pouvant être éventuellement aussi racine d'autres P_j, $0 \leq j \leq k-1$). On a, d'après la définition de P_{i+1},

$$P_{i-1} = P_i Q_{i+1} - P_{i+1}\,,$$

donc puisque $P_i(r) = 0$,

$$P_{i-1}(r) = -P_{i+1}(r)\,.$$

Mais d'autre part, deux polynômes consécutifs de la suite de Sturm étant premiers entre eux, r n'est racine ni de P_{i-1}, ni de P_{i+1}. Par suite, $P_{i-1}(r)$ et $P_{i+1}(r)$ sont non nuls, et opposés l'un de l'autre. Par continuité, pour x assez voisin de r, $P_{i-1}(x)$ et $P_{i+1}(x)$ sont non nuls et de signes contraires l'un de l'autre. Supposons par exemple $P_{i-1}(x) > 0$ et $P_{i+1}(x) < 0$, pour x assez voisin de r. Plusieurs cas sont à considérer:

(i) Supposons, pour $\epsilon > 0$ assez petit, $P_i(r-\epsilon) < 0$ et $P_i(r+\epsilon > 0$. En $r-\epsilon$, la suite de Sturm présente un changement de signe entre P_{i-1} et P_i, et n'a pas de changement de signe entre P_i et P_{i+1}; soit au total, 1 changement de signe dans la partie de la suite $(P_{i-1}, P_i, P_{i+1}$. En $r+\epsilon$, la suite de Sturm n'a pas de changement de signe entre P_{i-1} et P_i, et un changement de signe entre P_i et P_{i+1}; soit, au total, 1 changement de signe dans la partie de la suite $(P_{i-1}, P_i, P_{i+1}$. Dans cette partie de la suite, il y a donc le même nombre, 1, de changements de signe en $r-\epsilon$ et en $r+\epsilon$ (et aussi en r, puisque par convention $P_i(r)$, qui est nul, est considéré comme de même signe que $P_{i-1}(r)$; on n'a donc pas de changement de signe entre $P_{i-1}(r)$ et $P_i(r)$, et un changement de signe entre $P_i(r)$ et $P_{i+1}(r)$).

(ii) Supposons, pour $\epsilon > 0$ assez petit, $P_i(r-\epsilon) > 0$ et $P_i(r+\epsilon) < 0$. Comme ci-dessus, on voit que dans la partie de la suite (P_{i-1}, P_i, P_{i+1}), il y a 1 changement de signe, aussi bien en $r-\epsilon$ qu'en $r+\epsilon$ et qu'en r.

(iii) Supposons, pour $\epsilon > 0$ assez petit, $P_i(r-\epsilon)$ et $P_i(r+\epsilon)$ de même signe (cela peut se produire si r est racine de P_i de multiplicité paire). Alors encore, on voit comme ci-dessus que dans la partie de la suite de Sturm $(P_{i-1}, P_i, P_{i+1}$, il y a 1 changement de signe, aussi bien en $r-\epsilon$, qu'en r, ou en $r+\epsilon$.

On a donc prouvé que $v(x)$ augmente exactement d'une unité lorsque x traverse en croissant une racine de P, et ne varie pas lorsque x traverse une racine d'un autre polynôme P_i $(1 \leq i \leq k-1)$ de la suite de Sturm qui n'est pas racine de P. Le résultat indiqué dans le théorème en découle immédiatement. □

12.4. Remarque. Le théorème de Sturm reste valable lorsque l'intervalle $]a, b[$ n'est pas borné. En particulier, $v(-\infty) - v(+\infty)$ est le nombre de racines réelles de P.

12.5. Exercice. Appliquer la méthode de Sturm pour déterminer le nombre de racines réelles du polynôme

$$P(x) = x^{3600} + bx + 1$$

où b est une constante réelle. On fera une discussion suivant la valeur de b.

On remarque, sur cet exemple, que le nombre de racines réelles d'un polynôme peut être beaucoup plus petit que le nombre total de racines, réelles ou complexes (dans cet exemple, il y a 3600 racines réelles ou complexes, dont au plus deux racines réelles).