Analyse numérique, Matmeca 1ere année
Corrigé de la feuille 2[1]

1. Méthode d'élimination de Gauss, factorisation LU et LDU

1.a On résout le système suivant par la méthode du pivot de Gauss :

$$(\Sigma)\left\{\begin{array}{rcrcrcrl} 2x_1 & + & x_2 & & & + & x_4 = 1 & \mathrm{L}_1 \\ 4x_1 & + & 3x_2 & + & 2x_3 & + & 4x_4 = 0 & \mathrm{L}_2 \\ & & x_2 & + & 5x_3 & + & 2x_4 = 2 & \mathrm{L}_3 \\ 6x_1 & + & 3x_2 & + & 6x_3 & + & 8x_4 = 0 & \mathrm{L}_4 \end{array}\right.$$

$$(\Sigma) \iff \left\{\begin{array}{rcrcrcrl} 2x_1 & + & x_2 & & & + & x_4 = 1 & \mathrm{L}_1 \\ & & x_2 & + & 2x_3 & + & 2x_4 = -1 & \mathrm{L}_2 \leftarrow \mathrm{L}_2 - 2\mathrm{L}_1 \\ & & x_2 & + & 5x_3 & + & 2x_4 = 2 & \mathrm{L}_3 \\ & & & & 6x_3 & + & 5x_4 = -3 & \mathrm{L}_4 \leftarrow \mathrm{L}_4 - 3\mathrm{L}_1 \end{array}\right.$$

$$(\Sigma) \iff \left\{\begin{array}{rcrcrcrl} 2x_1 & + & x_2 & & & + & x_4 = 1 & \mathrm{L}_1 \\ & & x_2 & + & 2x_3 & + & 2x_4 = -1 & \mathrm{L}_2 \\ & & & & 3x_3 & + & = 3 & \mathrm{L}_3 \leftarrow \mathrm{L}_3 - \mathrm{L}_2 \\ & & & & 6x_3 & + & 5x_4 = -3 & \mathrm{L}_4 \end{array}\right.$$

$$(\Sigma) \iff \left\{\begin{array}{rcrcrcrl} 2x_1 & + & x_2 & & & + & x_4 = 1 & \mathrm{L}_1 \\ & & x_2 & + & 2x_3 & + & 2x_4 = -1 & \mathrm{L}_2 \\ & & & & 3x_3 & & = 3 & \mathrm{L}_3 \leftarrow \mathrm{L}_3 - \mathrm{L}_2 \\ & & & & & & 5x_4 = -9 & \mathrm{L}_4 \end{array}\right.$$

$$(\Sigma) \iff \left\{\begin{array}{l} x_1 = \dfrac{11}{10} \\ x_2 = \dfrac{3}{5} \\ x_3 = 1 \\ x_4 = -\dfrac{9}{5} \end{array}\right.$$

1.b L'algorithme du pivot de Gauss consiste à trianguler le système que l'on cherche à résoudre, puis à appliquer un algorithme de remontée. À l'étape d'élimination de l'inconnue i, on commence par chercher le plus grand pivot, puis on échange éventuellement la ligne i avec celle contenant le plus grand pivot, et enfin, on élimine la variable i des lignes $i+1$ à n.

[1] généré avec $\LaTeX\ 2_\varepsilon$. Tous les commentaires, compléments, insultes et remarques désobligeantes sont les bienvenus à `perrier@math.u-bordeaux1.fr`

```
#on commence par trianguler le systeme
POUR i de 1 a n-1 FAIRE
     #recherche du plus grand pivot
     i0=i
     max=abs(u(i,i))
     POUR j de i+1 a n FAIRE
          SI (abs(u(j,i))>max) ALORS
               i0=j
               max=abs(u(j,i))
          FIN SI
     FIN POUR

     #echange de lignes (le cas echeant)
     SI (i<>i0) ALORS
          POUR j de i a n FAIRE
               aux=u(i,j)
               u(i,j)=u(i,i0)
               u(i,i0)=aux
          FIN POUR
     FIN SI



     #faire pivoter
     POUR j de 1 a n FAIRE
          alpha=u(j,i)/u(i,i)
          POUR k de i+1 a n FAIRE
               u(j,k)=u(j,k)-alpha*u(i,k)
          FIN POUR k
          y(j)=y(j)-alpha*y(i)
     FIN POUR j
fin POUR i

#Algorithme de remontee
POUR i de n a 1 FAIRE
     x(i)=y(i)
     POUR j de i+1 a n FAIRE
          x(i)=x(i)-u(i,j)*x(j)
     FIN POUR j
     x(i)=x(i)/u(i,i)
FIN POUR i
```

$\boxed{\textbf{1.c}}$ En reprenant les étapes de l'algorithme de Gauss de la question (a), mais avec un second membre général ${}^t(y_1, y_2, y_3, y_4)$, on aboutit à

$$\begin{pmatrix} 2 & 1 & 0 & 1 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 2 & -1 & 1 & 0 \\ -7 & 2 & -2 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$$

on se retrouve avec, à gauche, une matrice triangulaire supérieure, et à droite une matrice triangulaire inférieure avec des 1 sur la diagonale. En fait, à gauche, on a U, et à droite, L^{-1}. Ainsi, en inversant la matrice L, on trouve

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 3 & 0 & 2 & 1 \end{pmatrix} \qquad U = \begin{pmatrix} 2 & 1 & 0 & 1 \\ 0 & 1 & 2 & 2 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}$$

1.d.i Montrons que la factorisation LU d'une matrice, où L est une matrice triangulaire inférieure avec des 1 sur la diagonale, et U est une matrice triangulaire supérieure, est unique. Pour cela, on suppose que l'on dispose de deux décompositions L_1U_1 et L_2U_2 d'une même matrice A. Alors $L_1U_1 = L_2U_2$, que l'on peut réecrire

$$L_2^{-1}L_1 = U_2U_1^{-1} \tag{1}$$

dans (1), le membre de gauche est une matrice triangulaire inférieure, avec des 1 sur la diagonale, et le membre de droite est une matrice triangulaire supérieure. On en déduit que $L_2^{-1}L_1 = U_2U_1^{-1} = \mathrm{Id}$, c'est à dire que $L_1 = L_2$, et $U_1 = U_2$.

De la même manière, on suppose qu'on dispose de deux décompositions LDU pour une matrice A donnée : $A = L_1D_1U_1 = L_2D_2U_2$. Alors d'après l'unicité de la décomposition LU, on voit que $L_1 = L_2$, et que $D_1U_1 = D_2U_2$. En identifiant les coefficients de la diagonale de $D_1U_1 = D_2U_2$, on voit que $D_1 = D_2$, et donc que $U_1 = U_2$.

1.d.ii Soit A une matrice symétrique telle que $A = LDU$. Alors en transposant cette égalité, on trouve ${}^tA = {}^tU\,{}^tD\,{}^tL = {}^tU\,D\,{}^tL$. Comme ${}^tA = A$, on en déduit que $LDU = {}^tU\,D\,{}^tL$, donc par unicité de la décomposition LDU, $U = {}^tL$.

Dans le cas où la matrice A est symétrique définie positive, montrons que les coefficients de D sont positifs ; la matrice A étant symétrique définie positive, on sait que pour tout vecteur X non nul, ${}^tX\,AX > 0$, c'est à dire que ${}^tX\,{}^tL\,DLX > 0$, que l'on peut écrire sous la forme ${}^t(LX)\,DLX > 0$. En choisissant pour X les vecteurs X_i tels que $LX_i = e_i$, où e_i est le i^{eme} vecteur de la base canonique, on voit que les coefficients de D sont tous positifs. Si les d_i sont les coefficients diagonaux de D, on note D' la matrice diagonale dont les coefficients sont les $\sqrt{d_i}$. On a alors $A = {}^tL\,D'^2L$. En posant $B = D'L$, on trouve bien $A = {}^tB\,B$.

1.e.i Afin de trouver U, on commence par effectuer l'élimination de Gauss sur la matrice A. En la faisant sur les premières lignes, on s'aperçoit que comme A est tridiagonale, l'élimination laisse inchangés les coefficients de la surdiagonale. Lors de l'élimination, on suppose être arrivé à

$$\begin{pmatrix}
d_1 & c_1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\
0 & \ddots & \ddots & \ddots & & & & & \vdots \\
\vdots & \ddots & \ddots & \ddots & \ddots & & & & \vdots \\
\vdots & & 0 & d_i & c_i & 0 & & & \vdots \\
0 & \dots & 0 & a_{i+1} & b_{i+1} & c_{i+1} & 0 & & \vdots \\
\vdots & & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & & & \ddots & \ddots & \ddots & \ddots & 0 \\
\vdots & & & & & \ddots & a_{n-1} & b_{n-1} & c_{n-1} \\
0 & \dots & \dots & \dots & \dots & \dots & 0 & a_n & b_n
\end{pmatrix}$$

On effectue l'élimination de Gauss sur la ligne $i+1$, c'est à dire qu'on fait l'opération

$$\mathrm{L}_{i+1} \leftarrow \mathrm{L}_{i+1} - \frac{a_{i+1}}{d_i}\,\mathrm{L}_i$$

on en déduit la formule de récurrence suivante pour les d_i :

$$\begin{cases} d_1 = b_1 \\ d_{i+1} = b_{i+1} - \dfrac{a_{i+1}c_i}{d_i} \end{cases}$$

et la matrice U est de la forme

$$\begin{pmatrix}
d_1 & c_1 & 0 & \dots & 0 \\
0 & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & 0 \\
\vdots & & \ddots & d_{n-1} & c_{n-1} \\
0 & \dots & \dots & 0 & d_n
\end{pmatrix}$$

Pour déterminer L, on note $l_{i,k}$ ses coefficients de L, et on sait que le produit de L par U doit être égal à A, soit

$$\forall i,j \in [\![\, 0\,;\, n \,]\!] \qquad a_{i,j} = \sum_{k=1}^{n} l_{i,k}u_{k,j}$$

Dans cette dernière somme, seuls deux coefficients sont non nuls : $u_{j,j}$ et $u_{j-1,j}$, soit

$$a_{i,j} = l_{i,j}u_{j,j} + l_{i,j-1}u_{j-1,j} \tag{2}$$

dans l'équation (2), on suppose que $j = i$:

$$a_{i,i} = l_{i,i}u_{i,i} + l_{i,i-1}u_{i-1,i}$$

et on sait que $a_{i,i} = b_i$, $l_{i,i} = 1$, $u_{i-1,i} = c_{i-1}$ et on a noté $u_{i,i} = d_i$, donc on a

$$b_i = d_i + l_{i,i-1}c_{i-1}$$

d'où

$$l_{i,i-1} = \frac{b_i - d_i}{c_{i-1}}$$

on a ainsi déterminé tous les coefficients de la sous–diagonale de L. Montrons à présent que tous les autres coefficients sont nuls. En écrivant l'équation (2), pour $j = i - 1$, on trouve

$$a_{i,i-1} = l_{i,i-1}u_{i-1,i-1} + l_{i,i-2}u_{i-2,i-1}$$

soit

$$a_i = \frac{b_i - d_i}{c_{i-1}}\, d_{i-1} + l_{i,i-2}u_{i-2,i-1} \tag{3}$$

or on sait que

$$d_i = b_i - \frac{a_i c_{i-1}}{d_{i-1}}$$

on voit donc immédiatement que

$$a_i = \frac{b_i - d_i}{c_{i-1}}\, d_{i-1}$$

donc, d'après (3)

$$l_{i,i-2}u_{i-2,i-1} = 0$$

et on en déduit que $l_{i,i-2} = 0$.

Supposons maintenant que les coefficients de la forme $l_{i,i-k}$ soient tous nuls ($k \geqslant 2$). Alors en écrivant (2) pour $j = i_k$, on trouve

$$a_{i,i-k} = l_{i,i-k}u_{i-k,i-k} + l_{i,i-k-1}u_{i-k-1,i-k}$$

alors $l_{i,i-k} = 0$ par hypothèse et $a_{i,i-k} = 0$, car $k > 2$. On en déduit que $l_{i,i-k-1} = 0$.

Concernant l'algorithme, on suppose que l'on connait trois tableaux `A`, `B`, `C` de dimensions n. On cherche à calculer les tableaux `D`, où les `D(i)` sont les coefficients de la diagonale de U, et `E`, où les `E(i)` sont les coefficients de la sous diagonale de L.

> Dans un soucis d'optimalité, il suffirait en fait de prendre des vecteurs `B`, `C`, `D` et `E` de dimension $n - 1$. On a gardé ici des vecteurs de dimension n dans le seul soucis de lisibilité des indices.

```
D(1)=B(1)

POUR i de 1 a n-1 FAIRE
     D(i+1)=B(i+1)-A(i+1)*C(i)/D(i)
     E(i+1)=(B(i+1)-D(i+1))/C(i)
FIN POUR
```

Cet algorithme nécessite $5(n-1)$ opérations.

Concernant l'algorithme de descente–remontée, on cherche à résoudre le système LUX = Y en résolvant d'abord le système LY′ = Y, puis le système UX = Y′. On suppose connu en début d'algorithme le tableau du second membre `Y`.

```
#Algorithme de descente (on résout L*Yprime=Y)
Yprime(1)=Y(1)

POUR i de 2 a n FAIRE
```

```
     Yprime(i)=Y(i)-E(i)*Yprime(i-1)
FIN POUR

#Algorithme de remontée (on résout U*X=Yprime)
X(n)=Yprime(n)/D(n)

POUR i de n-1 à 1 FAIRE
     X(i)=(Yprime(i)-C(i)*X(i+1))/D(i)
FIN POUR
```

L'algorithme de descente nécessite $2(n-1)$ opérations, et l'algorithme de remontée $3(n-1)$ opérations.

> L'algorithme proposé n' « écrase » pas le tableau `Y` des variables données. En pratique, si le seul but recherché est de résoudre le système sans se soucier de conserver les données, il suffirait de remplacer `Yprime` par `Y` puis `X` par `Y`.

2. Factorisation de Choleski, systèmes creux.

On part de la matrice carrée suivante

$$\begin{cases} a_{i,i} = \alpha \\ a_{i,i+1} = a_{i+1,i} = -1 \text{ pour } 1 \leqslant i \leqslant n-1 \\ a_{n,1} = a_{1,n} = -1 \end{cases}$$

Ce système intervient par exemple dans la discrétisation de

$$\begin{cases} \forall x \in \mathbb{R} & -u''(x) = f(x) \\ \forall x \in \mathbb{R} & u(x+1) = u(x) \end{cases}$$

où f est une fonction 1-périodique donnée. On se donne une subdivision $(x_0, x_1, \dots x_n)$ régulière de $[\,0\,;1\,]$ (que l'on prolonge sur $\mathbb{R}$ par periodicité), avec $x_i = \dfrac{i}{n}$. En posant $u_i = u(x_i)$ et $f_i = f(x_i)$ et en discrétisant la partie gauche par différences finies, on trouve, pour tout i entier

$$(\mathbf{E_i}) \qquad \frac{-u_{i+1} + 2u_i - u_{i-1}}{h^2} = f_i$$

La fonction que l'on cherche à approcher étant périodique, on va se contenter de chercher les u_i pour $i \in [\![\, 1\,;\, n \,]\!]$, et on déduit les autres u_i avec la relation de périodicité

$$\forall i \in \mathbb{Z} \qquad u_{i+n} = u_i$$

Dans l'équation $(\mathbf{E_i})$, les u_j intervenant sont bien dans $[\![\, 1\,;\, n \,]\!]$, à l'exception de l'équation $(\mathbf{E_1})$ et de l'équation $(\mathbf{E_n})$:

$$\begin{array}{ll} (\mathbf{E_1}) & \dfrac{-u_0 + 2u_1 - u_2}{h^2} = f_1 \\ (\mathbf{E_n}) & \dfrac{-u_{n-1} + 2u_n - u_{n+1}}{h^2} = f_n \end{array}$$

par périodicité, on remplace u_0 par u_n et u_{n+1} par u_1, pour trouver

$$\begin{array}{ll} (\mathbf{E_1}) & \dfrac{-u_n + 2u_1 - u_2}{h^2} = f_1 \\ (\mathbf{E_n}) & \dfrac{-u_{n-1} + 2u_n - u_1}{h^2} = f_n \end{array}$$

Finalement, le problème initial se ramène à

$$\mathrm{A}u = h^2 f$$

avec

- A la matrice de l'énoncé avec $\alpha = 2$
- u le vecteur colonne ${}^t(u_1, u_2, \dots, u_n)$
- f le vecteur colonne ${}^t(f_1, f_2, \dots, f_n)$.

2.a On sait que la matrice A est symétrique. Elle sera donc définie positive si et seulement si ses valeurs propres sont toutes strictement positives. Afin de localiser les valeurs propres de A, utilisons le théorème de Gershgörin

> **Rappel : Théorème de Gershgörin**
> Soit A ue matrice $n \times n$. On pose
>
> $$\Lambda_i = \sum_{j=1, j \neq i}^{n} |a_{i,j}|$$
>
> Alors
>
> 1. $\mathrm{Sp}(\mathrm{A}) \subset \bigcup_{i=1}^{n} \mathrm{B}(a_{i,i}, \Lambda_i)$ (où B est une **boule fermée**)
> 2. Si de plus A est irréductible (c'est à dire si le graphe orienté associé à A est connexe), alors si λ est une valeur propre de A qui se trouve également sur la frontière de $\bigcup_{i=1}^{n} \mathrm{B}(a_{i,i}, \Lambda_i)$, alors $\lambda \in \bigcap_{i=1}^{n} \partial \mathrm{B}(a_{i,i}, \Lambda_i)$.

Pour tout i, on a dans notre cas $\Lambda_i = 2$. Donc si $\alpha > 2$, alors $\mathrm{B}(\alpha, 2) \cap \mathbb{R} \subset \mathbb{R}_+^*$. On en déduit que la matrice A est définie positive.

2.b

> Rappelons les principaux outils et résultats permettant d'étudier le creux lors de la décomposition $\mathrm{A} = \mathrm{LD}\,{}^t\mathrm{L}$ (présentés du plus grossier au plus fin)
>
> **Largeur de bande/Bande**
> La largeur de bande de la matrice A est l'entier
>
> $$\beta(\mathrm{A}) = \max \left\{ |i-j| \quad a_{i,j} \neq 0 \right\}$$
>
> On définit alors
>
> $$\mathrm{Bande}(\mathrm{A}) = \left\{ (i,j) \quad 0 \leqslant i-j \leqslant \beta(\mathrm{A}) \right\}$$
>
> On a alors le résultat suivant concernant le remplissage de A lors de sa décomposition $\mathrm{LD}\,{}^t\mathrm{L}$:
>
> $$\mathrm{Bande}(\mathrm{L} + \mathrm{D} + {}^t\mathrm{L}) = \mathrm{Bande}(\mathrm{A})$$
>
> Si l'on ne tient compte que des coefficients nuls n'appartenant pas à $\mathrm{Bande}(\mathrm{L} + \mathrm{D} + {}^t\mathrm{L}) = \mathrm{Bande}(\mathrm{A})$, alors le nombre d'opérations nécessaires pour factoriser A est égal à
>
> $$\frac{1}{2}\beta(\beta+3)n - \frac{\beta^3}{3} - \beta^2 - \frac{2\beta}{3}$$
>
> et le nombre d'opérations pour la résolution est égal à
>
> $$2(\beta+1)n - \beta(\beta+1)$$
>
> **Profil**

Pour un indice i donné, on note

$$\begin{aligned} &f_i(\mathrm{A}) = f_i = \min\left\{j \quad a_{i,j} \neq 0\right\} \\ &\beta_i = i - f_i \end{aligned}$$

On définit alors le profil de A par

$$\mathrm{Prof}(\mathrm{A}) = \left\{(i,j) \quad 0 \leqslant i - j \leqslant \beta_i(\mathrm{A})\right\}$$

Et on a le résultat suivant

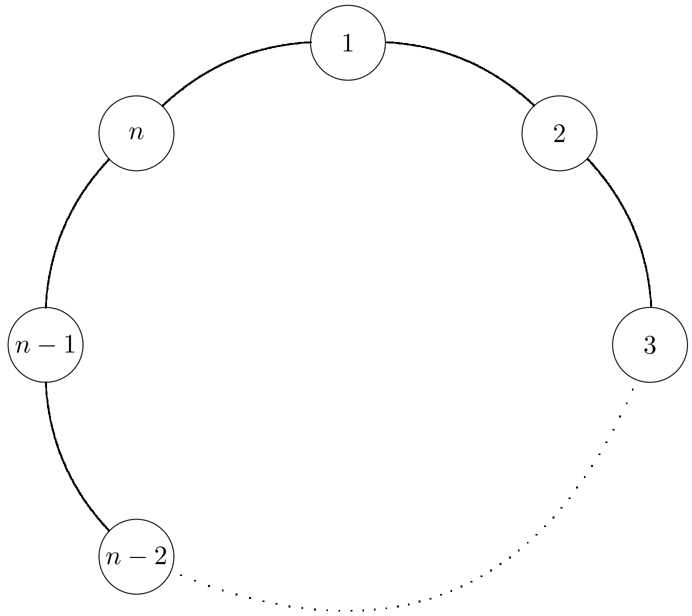
$$\mathrm{Prof}(\mathrm{A}) = \mathrm{Prof}(\mathrm{L} + \mathrm{D} + {}^t\mathrm{L})$$

Théorème de Rose-Tarjan-Lucker
Soit $\mathrm{G}_\alpha = (\mathrm{V}, \mathrm{E})$ le graphe associé à une matrice symétrique A, où E est l'ensemble des segments de G_α, et α est une numérotation des sommets. Alors $(v, w) \in \mathrm{E}^\star$ si et seulement si

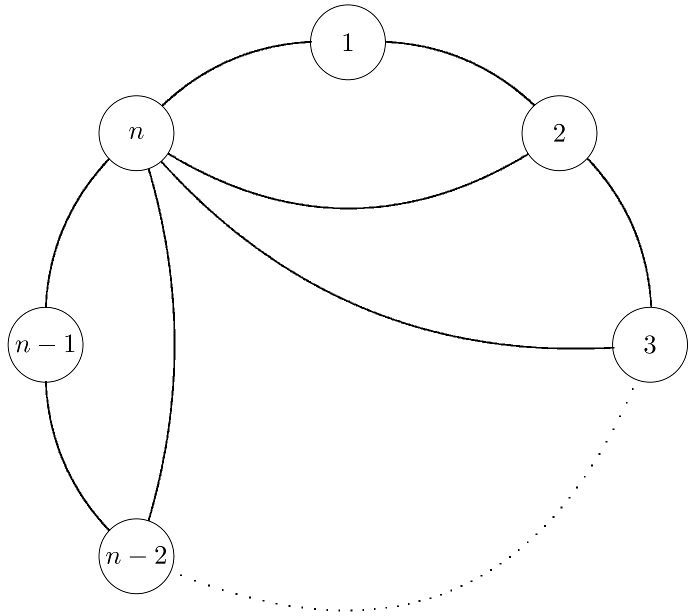
- ou bien $(v, w) \in \mathrm{E}$
- ou bien il existe un chemin dans G_α que l'on note $z^{(1)}, z^{(2)}, \ldots, z^{(k+1)}$, avec $v = z^{(1)}$, $w = z^{(k+1)}$ et tel que pour tout $i = 2 \ldots k$

$$\alpha^{-1}(z^{(i)}) < \min(\alpha^{-1}(v), \alpha^{-1}(w))$$

Le graphe de A est le suivant



En appliquant le théorème de Rose-Tarjan-Lucker, on trouve le graphe $\star$ suivant



Explication
Tout d'abord, vu le profil de la matrice A, on sait dès le départ que les seuls segments qui peuvent apparaître pour le graphe * de G sont les segments de la forme (i, n), pour $1 \leqslant i \leqslant n$. Ensuite, étant donné un sommet i, on peut le relier au sommet n *via* le chemin

$$(i, i-1)\ (i-1, i-2)\ (i-2, i-3)\ \ldots\ (3, 2)\ (2, 1)\ (1, n)$$

qui est bien tel que le numéro des sommets visités est strictement inférieur à i et à n.

$\boxed{\mathbf{2.c}}$ La matrice LD contient $3n-3$ coefficients non nuls : les éléments de la diagonale, les éléments de la sous diagonale, et les élements de la dernière ligne.

Commençons par chercher les relations de récurrence vérifiées par les coefficients de L et de D. On note $l_{i,j}$ les coefficients de la matrice L (ils sont donc nuls si $j > i$), et d_i les coefficients de D. En multipliant L par D, puis par ${}^t\mathrm{L}$, on trouve la matrice de coefficients

$$(\mathrm{LD}\,{}^t\mathrm{L})_{i,j} = \sum_{k=1}^{\min(i,j)} l_{i,k} d_k l_{j,k}$$

Comme la matrice est symétrique, on ne s'intéresse qu'à la partie triangulaire inférieure de la matrice $\mathrm{A} = \mathrm{LD}\,{}^t\mathrm{L}$, c'est à dire qu'on prend les indices (i, j) tels que $i \leqslant j$. On a alors

$$(\mathrm{LD}\,{}^t\mathrm{L})_{i,j} = \sum_{k=1}^{i} l_{i,k} d_k l_{j,k}$$

Pour $i = j$, on trouve la formule de récurrence sur les d_i :

$$d_i = a_{i,i} - \sum_{k=1}^{i-1} d_k {l_{i,k}}^2 \quad \text{pour} \quad j < i$$

de même pour les $l_{i,j}$, on trouve

$$l_{i,j} = \frac{1}{d_i}\left(a_{i,j} - \sum_{k=1}^{i-1} d_k l_{i,k} l_{j,k}\right)$$

Dans la récurrence définissant les d_i, et pour $i < n$, on sait que les $l_{i,k}$ sont tous nuls, à l'exception de $l_{i,i-1}$. La relation de récurrence sur les d_i devient donc

$$d_i = \alpha - {l_{i,i-1}}^2 d_{i-1}$$

Concernant les $l_{i,j}$, on trouve

$$l_{i+1,i} = \frac{1}{d_i} a_{i+1,i} = -\frac{1}{d_i}$$

Enfin, pour les coefficients de la dernière ligne, on trouve

$$l_{n,i} = \frac{1}{d_i}\left(a_{n,j} - d_{i-1} l_{n,j-1} l_{j,j-1}\right) = \frac{1}{d_i}\left(-1 - d_{i-1}\right)$$

Pour l'algorithme, on ne stocke que trois vecteurs : le vecteur des coefficients diagonaux de D que l'on note `D`, le vecteur des coefficients sous-diagonaux de L, que l'on note `A`, et le vecteur de la dernière ligne, que l'on note `B`.

```
###############################
#Algorithme de Factorisation
###############################

D(1)=alpha

POUR i de 2 a N FAIRE
     A(i)=-1/D(i-1)
     D(i)=alpha-A(i)*D(i-1)
     B(i)=-(1+D(i-1))/D(i)
FIN POUR

###############################
#Algorithmes de Resolution
###############################

#Algorithme de descente
X(1)=Y(1)

POUR i de 2 a (n-1) FAIRE
     X(i)=Y(i)-A(i)*X(i-1)
FIN POUR

X(n)=Y(n)
POUR i de 1 a n FAIRE
     X(n)=X(n)-B(i)*X(i)
FIN POUR


#Resolution de DX'=X
POUR i de 1 a n FAIRE
     X(i)=X(i)/D(i)
FIN POUR
```

```
#Algorithme de remontee
X(n-1)=X(n-1)-B(n-1)*X(n)
POUR i de 2 a n-1 FAIRE
     X(n-i)=X(n-i)-A(i)*X(i+1)-B(i)*X(n)
FIN POUR
```

3. Discrétisation du Laplacien en dimension 2 et méthode de Choleski

$\boxed{\textbf{3.a}}$ On s'intéresse au problème continu suivant : on note D de carré unité fermé $\mathrm{D} = [\,0\,;1\,] \times [\,0\,;1\,]$, et $\partial\mathrm{D}$ le bord de D

$$\begin{cases} -\dfrac{\partial^2 u}{\partial x^2}(x,y) - \dfrac{\partial^2 u}{\partial y^2}(x,y) = f(x) \quad \text{pour } (x,y) \in \mathrm{D} \\ u = 0 \quad \text{sur } \partial\mathrm{D} \end{cases}$$

Vérifions que $u_{p,q}\,:\,(x,y) \mapsto \sin(p\pi x)\sin(q\pi y)$ est fonction propre de l'opérateur $-\Delta$ sur D :

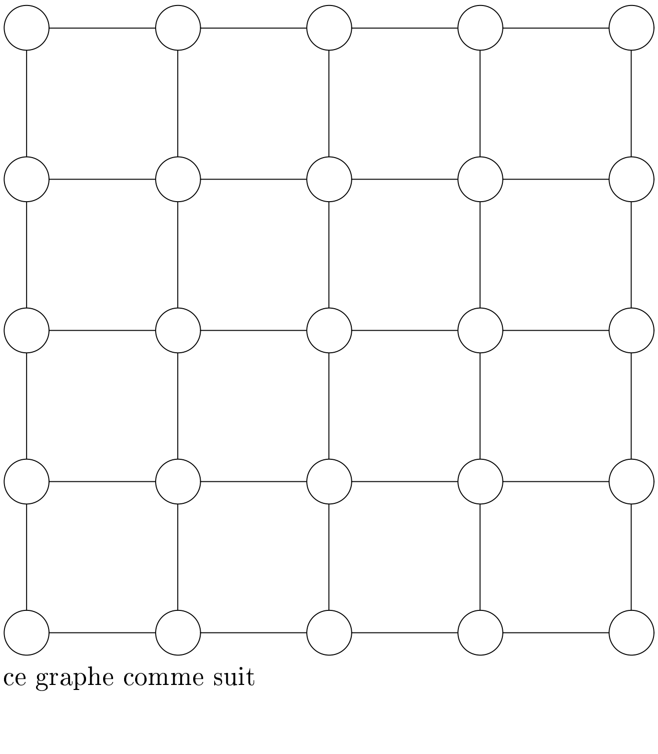
$$\begin{aligned} \forall (x,y) \in \mathrm{D} \qquad -\Delta u_{p,q}(x,y) &= (p\pi)^2 \sin(q\pi y)\sin(p\pi x) + (q\pi)^2 \sin(q\pi y)\sin(p\pi x) \\ &= (p^2+q^2)\pi^2 \sin(q\pi y)\sin(p\pi x) \\ -\Delta u_{p,q}(x,y) &= (p^2+q^2)\pi^2 u_{p,q}(x,y) \end{aligned}$$

On en déduit que les fonctions $u_{p,q}$ sont fonctions propres, associées à la valeur propre $\pi^2(p^2+q^2)$ du Laplacien sur D.

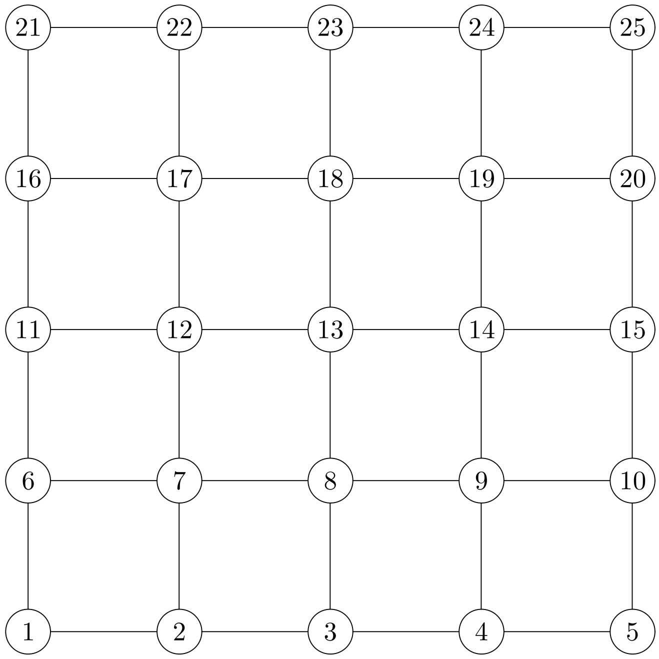
$\boxed{\textbf{3.b.i}}$ Le graphe de la matrice A peut s'obtenir immédiatement, en prenant pour sommets du graphe les noeuds du maillage, et en remarquant que la relation

$$\frac{-u_{i+1,j} + 2u_{i,j} - u_{i-1,j}}{h^2} + \frac{-u_{i,j+1} + 2u_{i,j} - u_{i,j-1}}{h^2} = f_{i,j}$$

signifie que chaque noeud est en relation avec les noeuds directement voisins, horizontalement et verticalement.



On numérote ce graphe comme suit



on obtient la matrice par blocs suivante

$$\mathrm{A} = \begin{pmatrix} \mathrm{B} & -\mathrm{id}_5 & 0 & 0 & 0 \\ -\mathrm{id}_5 & \mathrm{B} & -\mathrm{id}_5 & 0 & 0 \\ 0 & -\mathrm{id}_5 & \mathrm{B} & -\mathrm{id}_5 & 0 \\ 0 & 0 & -\mathrm{id}_5 & \mathrm{B} & -\mathrm{id}_5 \\ 0 & 0 & 0 & -\mathrm{id}_5 & \mathrm{B} \end{pmatrix}$$

où B est la matrice

$$\mathrm{B} = \begin{pmatrix} 4 & -1 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & -1 & 4 \end{pmatrix}$$

A est symétrique définie positive

Afin de montrer que A est symétrique, définie positive, on va utiliser la théorème de Gershgörin

On voit ici que les B_i sont de centre 4, et de rayon $2, 3$ ou 4 selon que la ligne i correspond, dans le graphe, à un coin, à un point du bord, ou à un point du milieu. On en déduit que dans notre cas,

$$\bigcup_{i=1}^{n} \mathrm{B}(a_{i,i}, \Lambda_i) = \mathrm{B}(4,4)$$

on voit donc que les valeurs propres de A sont comprises entre 0 et 8. Supposons que 0 soit valeur propre de A. Alors d'après le théorème de Gershgörin, comme 0 est une valeur extrêmale de $\bigcup_{i=1}^{n} \mathrm{B}(a_{i,i}, \Lambda_i) = \mathrm{B}(4,4)$, et que la matrice A est irréductible, on en déduit que 0 est dans $\bigcap_{i=1}^{n} \partial\mathrm{B}(a_{i,i}, \Lambda_i)$. Or cet ensemble est vide dans notre cas. On en déduit que 0 n'est pas valeur propre de A, donc que les valeurs propres de A sont toutes strictement positives, ce qui veut dire que A est définie positive.

Conditionnement de A

Intuitivement, si on se souvient du Laplacien en dimension 1 avec conditions de Dirichlet, les vecteurs propres du problème discrétisé correspondaient aux valeurs des vecteurs propres, évalués sur les points du maillage. Soient p et q fixés. Cherchons donc des vecteurs propres de la forme

$$u_{i,j} = \sin\left(\frac{p\pi i}{l+1}\right)\sin\left(\frac{q\pi j}{l+1}\right)$$

on trouve alors

$$\begin{aligned}
&\frac{-u_{i+1,j} + 2u_{i,j} - u_{i-1,j}}{h^2} + \frac{-u_{i,j+1} + 2u_{i,j} - u_{i,j-1}}{h^2} \\
= \ & 4\sin^2\left(\frac{p\pi}{2(l+1)}\right)\sin\left(\frac{p\pi i}{l+1}\right)\sin\left(\frac{q\pi j}{l+1}\right) \\
& \qquad +4\sin^2\left(\frac{q\pi}{2(l+1)}\right)\sin\left(\frac{p\pi i}{l+1}\right)\sin\left(\frac{q\pi j}{l+1}\right) \\
= \ & 4\left(\sin^2\left(\frac{p\pi}{2(l+1)}\right) + \sin^2\left(\frac{q\pi}{2(l+1)}\right)\right)\sin\left(\frac{p\pi i}{l+1}\right)\sin\left(\frac{q\pi j}{l+1}\right)
\end{aligned}$$

On a donc trouvé l^2 vecteurs propres indépendants, qui correspondent aux valeurs propres de A qui sont

$$\lambda_{p,q} = 4\left(\sin^2\left(\frac{p\pi}{2(l+1)}\right) + \sin^2\left(\frac{q\pi}{2(l+1)}\right)\right)$$

on en déduit que

$$\lambda_{\max} = 8\sin^2\left(\frac{l\pi}{2(l+1)}\right)$$

et que

$$\lambda_{\min} = 8\sin^2\left(\frac{\pi}{2(l+1)}\right)$$

Finalement, on trouve le conditionnement de A par la formule

$$\mathrm{Cond}_2(\mathrm{A}) = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{\sin^2\left(\dfrac{l\pi}{2(l+1)}\right)}{\sin^2\left(\dfrac{\pi}{2(l+1)}\right)}$$

3.b.ii

> **Rappel : Algorithme de Cuthill–Mac Kee**
> On se donne un graphe non orienté. On appelle *degré* d'un sommet le nombre de voisins d'un sommet de ce graphe
> Les étapes sont les suivantes

- Choisir un sommet de départ et lui affecter le numéro 1
- Pour $i = 1, 2, \dots n$, trouver tous les voisins non renumérotés du sommet x_i et les numéroter par ordre croissant de degré
- La numérotation « Reverse Cuthill–Mac Kee » est donnée par : $y_1, y_2 \dots y_n$ où $y_i = x_{n-i+1}$.

Afin d'évaluer le profil de A, on calcule successivement les f_i, les β_i, et les ω_i. Ceux-ci sont résumés dans le tableau suivant.

Ligne	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
f_i	1	1	2	3	4	6	6	7	8	9	11	11	12	13	14
β_i	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1

Ligne	16	17	18	19	20	21	22	23	24	25
f_i	1	2	3	4	5	6	7	8	9	10
β_i	15	15	15	15	15	15	15	15	15	15

On trouve au total 187 termes dans le profil de A.
Regardons à présent ce qui se passe pour la seconde numérotation de A.

Ligne	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
f_i	1	1	1	2	2	3	4	4	5	6	7	7	8	9	10
β_i	0	1	2	2	3	3	3	4	4	4	4	5	5	5	5

Ligne	16	17	18	19	20	21	22	23	24	25
f_i	11	12	13	14	16	17	18	20	21	23
β_i	5	5	5	5	4	4	4	3	3	2

On trouve au total 115 termes dans le profil de A.

D'une manière générale, on numérote « en diagonale » le graphe de la matrice $l \times l$, comme on a numéroté le graphe de la matrice 25×25 (voir page suivante). On numérote également les diagonales (en gras sur le dessin). Celles-ci contiennent :

- k termes sur la k^{ieme} diagonale si $k \leqslant l$.
- $2l - k$ termes sur la diagonale si $l + 1 \leqslant k \leqslant 2l - 1$.

Plaçons-nous sur une diagonale numérotée i avec $i \leqslant l$. Le plus petit terme de la diagonale est égal à la somme des termes sur les diagonales précédentes, plus 1 :

$$\left(\sum_{k=1}^{i-1} k\right) + 1 = \frac{i(i-1)}{2} + 1$$

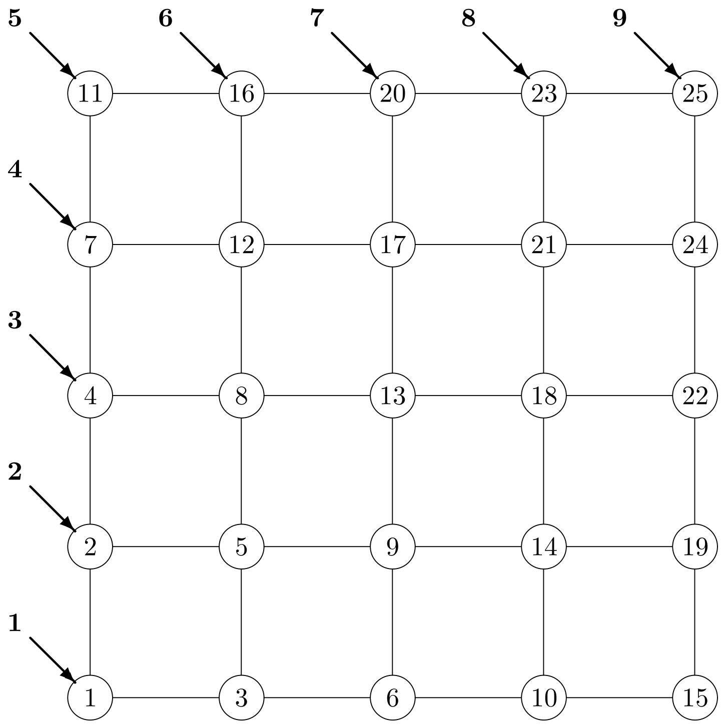
Le voisin de plus petit numéro du premier élément d'une diagonale est celui qui est situé en dessous, et qui a donc le numéro $\dfrac{(i-1)(i-2)}{2}+1$ (par exemple, sur le dessin, le plus petit voisin du sommet 7 est le sommet 4). Pour un sommet qui n'est pas le premier de la diagonale, le plus petit voisin est celui situé directement à gauche.

On en déduit que pour le premier terme de la diagonale i, la bande contient

$$\frac{i(i-1)}{2} + 1 - \left(\frac{(i-1)(i-2)}{2} + 1\right) + 1 = \frac{(i-1)}{2}\,(i - i + 2) + 1 = i$$

Pour les autres termes, la bande contient un terme de plus, soit $i + 1$. Finalement, le nombre total d'éléments se trouvant dans le profil de L, sur les lignes $\leqslant l$ est égal à

$$\sum_{i=1}^{l} (i + (i-1) \times (i+1)) = \sum_{i=1}^{l} \left(i + i^2 - 1\right)$$



Tripotons un peu cette somme :

$$\begin{aligned}\sum_{i=1}^{l}\left(i+i^2-1\right) &= \frac{l(l+1)}{2}+\frac{l(l+1)(2l+1)}{6}-l\\ &= \frac{l(l+1)(3+2l+1)}{6}-l\\ &= \frac{l(l+1)(l+2)}{3}-l\\ \sum_{i=1}^{l}\left(i+i^2-1\right) &= \frac{l(l^2+3l-1)}{3}\end{aligned}$$

On se place à présent sur une diagonale $i > l$. Le premier terme de cette diagonale a pour numéro

$$\sum_{k=1}^{l} k + \sum_{k=l+1}^{i-1}(2l-k)+1$$

On simplifie cette somme :

$$\begin{aligned}\sum_{k=l+1}^{i-1}(2l-k) &= 2l(i-1-(l+1)+1)-\sum_{k=l+1}^{i-1}k\\ &= 2l(i-l-1)-\sum_{k=1}^{i-1}k+\sum_{k=1}^{l}k\\ &= 2l(i-l-1)-\frac{i(i-1)}{2}+\frac{l(l+1)}{2}\end{aligned}$$

Finalement, le premier terme de la diagonale a pour numéro

$$l(l+1)+2l(i-l-1)-\frac{i(i-1)}{2}+1$$

Le long de cette diagonale, la largeur de bande est constante, égale à la différence entre le premier sommet de la diagonale et le premier sommet de la diagonale précédente, plus 1. Le long de la diagonale i, la largeur de bande est donc égale à

$$l(l+1)+2l(i-l-1)-\frac{i(i-1)}{2}+1-\left(l(l+1)+2l(i-1-l-1)-\frac{(i-1)(i-2)}{2}+1\right)$$

soit, après simplification $2l-i+1$. Ainsi, étant donné une ligne de la matrice, dont le numéro se trouve sur la diagonale i, le nombre d'éléments de cette ligne se trouvant dans le profil est égal à $2l-i+2$ Il reste à multiplier cette quantité par le nombre d'éléments se trouvant sur la diagonale i, c'est à dire $(2l-i)$, et à sommer entre $l+1$ et $2l-1$:

$$\sum_{i=l+1}^{2l-1}(2l-i+2)(2l-i)$$

dans cette dernière somme, on effectue le changement d'indice $l+1+j=i$, et on trouve

$$\sum_{i=l+1}^{2l-i}(2l-i+2)=\sum_{j=0}^{l-2}(l-j-1)(l-j+1)=\sum_{j=0}^{l-2}\left((l-j)^2-1\right)$$

On calcule cette dernière somme

$$\begin{aligned}\sum_{j=0}^{l-2}\left((l-j)^2-1\right)&=\left(\sum_{i=1}^{l}i^2\right)-1-(l-1)\\&=\frac{l(l+1)(2l+1)}{6}-l\\&=\frac{l(2l^2+3l-5)}{6}\\\sum_{j=0}^{l-2}\left((l-j)^2-1\right)&=\frac{l(l-1)(2l+5)}{6}\end{aligned}$$

Finalement, on trouve, dans le profil de A un nombre de termes égal à

$$\frac{l(4l^2+9l-7)}{6}$$

qui est équivalent à

$$\frac{2l^3}{3}$$

Cela représente un gain d'environ 1/3 par rapport à la première numérotation proposée.

4. Méthode de Househölder

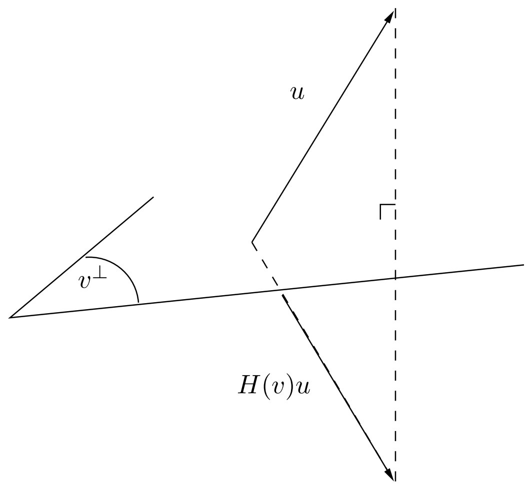
> Dans tout l'exercice, il faudra bien se rappeler que ${}^t v\, v$ est un *scalaire*.

$\boxed{\textbf{4.a}}$ On note $\mathrm{H}(v) = \mathrm{Id} - \frac{2}{{}^t v\, v} v\,{}^t v$. Montrons que $\mathrm{H}(v)$ est une matrice symétrique :

$$\begin{aligned}
{}^t\mathrm{H}(v) &= {}^t\left(\mathrm{Id} - \frac{2}{{}^t v\, v} v\,{}^t v\right) \\
&= {}^t\mathrm{Id} - \frac{2}{{}^t v\, v}\,{}^t\left(v\,{}^t v\right) \\
&= \mathrm{Id} - \frac{2}{{}^t v\, v} v\,{}^t v \\
{}^t\mathrm{H}(v) &= \mathrm{H}(v)
\end{aligned}$$

on en déduit que la matrice $\mathrm{H}(v)$ est une matrice symétrique. Montrons à présent que $\mathrm{H}(v)$ est orthogonale. Pour cela, il suffit de montrer que ${}^t\mathrm{H}(v)\,\mathrm{H}(v) = \mathrm{Id}$ ou encore que $\mathrm{H}(v)^2 = \mathrm{Id}$, puisque $\mathrm{H}(v)$ est symétrique.

$$\begin{aligned}
\mathrm{H}(v)^2 &= \left(\mathrm{Id} - \frac{2}{{}^t v\, v}\, v\,{}^t v\right)^2 \\
&= \mathrm{Id} - \frac{4}{{}^t v\, v}\, v\,{}^t v + \frac{4}{{}^t v\, v\,{}^t v\, v}\, v\,{}^t v\, v\,{}^t v \\
&= \mathrm{Id} - \frac{4}{{}^t v\, v}\, v\,{}^t v + \frac{4}{{}^t v\, v}\, v\,{}^t v \\
\mathrm{H}(v)^2 &= \mathrm{Id}
\end{aligned}$$

> $\mathrm{H}(v)$ est une *symétrie orthogonale*, par rapport à l'hyperplan $\mathrm{Ker}\,(\mathrm{H}(v) - \mathrm{Id}\,) = \mathrm{Ker}\,(v\,{}^t v) = v^\perp$, et sur $\mathrm{Ker}\,(\mathrm{H}(v) + \mathrm{Id}\,) = \mathbb{R}v$.
>
> 
>

$\boxed{\textbf{4.b}}$

$$\begin{aligned}
\mathrm{H}(a+\|a\|e_1)a = & \ \mathrm{H}(a+\|a\|e_1)(a+\|a\|e_1) - \|a\|\mathrm{H}(a+\|a\|e_1)e_1 \\
= & \ a+\|a\|e_1 - \|a\|e_1 \\
& -\|a\|\frac{2}{{}^t(a+\|a\|e_1)(a+\|a\|e_1)}\,(a+\|a\|e_1)\,{}^t(a+\|a\|e_1)\,e_1
\end{aligned}$$

par ailleurs on a

$${}^t(a+\|a\|e_1)(a+\|a\|e_1) = \|a\|^2 + 2\|a\|a_1 + \|a\|^2 = 2\|a\|(\|a\|+a_1)$$

et

$${}^t(a+\|a\|e_1)\,e_1 = a_1 + \|a\|$$

on a donc

$$\begin{aligned}
& \|a\|\frac{2}{{}^t(a+\|a\|e_1)(a+\|a\|e_1)}\,(a+\|a\|e_1)\,{}^t(a+\|a\|e_1)\,e_1 \\
& = \|a\|\frac{2(a_1+\|a\|)}{2\|a\|(\|a\|+a_1)}(a+\|a\|e_1) = a+\|a\|e_1
\end{aligned}$$

Finalement, on trouve

$$\mathrm{H}(a+\|a\|e_1)a = -\|a\|e_1$$

On montre de même que

$$\mathrm{H}(a-\|a\|e_1)a = \|a\|e_1$$

$\boxed{\textbf{4.c}}$ Soit A une matrice $n\times n$, régulière, et b un vecteur de $\mathbb{R}^n$. On s'intéresse (une fois de plus...) au système $\mathrm{A}x = b$. Soit C_1 la première colonne de A. En multipliant à gauche le système par la matrice $\mathrm{H}(\mathrm{C}_1+\|\mathrm{C}_1\|e_1)$, et en posant $\mathrm{A}^{(2)} = \mathrm{H}(\mathrm{C}_1+\|\mathrm{C}_1\|e_1)\mathrm{A}$ et $b^{(2)} = \mathrm{H}(\mathrm{C}_1+\|\mathrm{C}_1\|e_1)b$, on se ramène au système équivalent $\mathrm{A}^{(2)}x = b^{(2)}$. D'après la question précédente, la première colonne de $\mathrm{A}^{(2)}$ est égale à $-\|\mathrm{C}_1\|e_1$, on en déduit que $a_{i,1}^{(2)}$ est nul pour tout $i \geqslant 2$.

Comme

$$\mathrm{A}^{(2)} = \mathrm{H}(\mathrm{C}_1+\|\mathrm{C}_1\|e_1)\mathrm{A}$$

et que H est une matrice orthogonales, on sait que le conditionnement en norme 2 de A et de $\mathrm{A}^{(2)}$ sont égaux.

$\boxed{\textbf{4.d}}$ On suppose être arrivés à

$$\begin{pmatrix}
* & \cdots & \cdots & \cdots & \cdots & \cdots & * \\
0 & \ddots & & & & & \vdots \\
\vdots & \ddots & \ddots & & & & \vdots \\
\vdots & & 0 & * & \cdots & \cdots & * \\
\vdots & & \vdots & \vdots & & & \vdots \\
\vdots & & \vdots & \vdots & & & \vdots \\
0 & \cdots & 0 & * & \cdots & \cdots & *
\end{pmatrix}$$

que l'on note formellement par blocs

$$\left(\begin{array}{c|c}
\mathrm{A}_1^{(k)} & \mathrm{A}_2^{(k)} \\
\hline
0 & \mathrm{A}_3^{(k)}
\end{array}\right)$$

où $A_1^{(k)}$ est triangulaire supérieure. On note $C^{(k)}$ la première colonne de $A_3^{(k)}$. En multipliant $A^{(k)}$ à gauche par la matrice

$$\left(\begin{array}{c|c} \mathrm{Id} & 0 \\ \hline 0 & H(C^{(k)} + \|C^{(k)}\| e_1) \end{array}\right)$$

on trouve bien une matrice $A^{(k+1)}$ telle que $a_{i,j}^{(k+1)} = 0$ pour $1 \leqslant j \leqslant k$ et $j+1 \leqslant i$.

On remarque facilement que vu que $H(C^{(k)} + \|C^{(k)}\| e_1)$ est orthogonale, la matrice

$$\left(\begin{array}{c|c} \mathrm{Id} & 0 \\ \hline 0 & H(C^{(k)} + \|C^{(k)}\| e_1) \end{array}\right)$$

est également orthogonale.

4.e D'après la question précédente, on peut construire une suite finie de matrices orthogonales $Q^{(i)}$ telle que la matrice R définie par

$$Q^{(n)} Q^{(n-1)} \ldots Q^{(1)} A = R$$

soit une matrice triangulaire supérieure. Il reste donc à poser

$$\begin{aligned} Q &= (Q^{(1)})^{-1} (Q^{(2)})^{-1} \ldots (Q^{(n-1)})^{-1} (Q^{(n)})^{-1} \\ &= {}^t Q^{(1)} \, {}^t Q^{(2)} \ldots {}^t Q^{(n)} \end{aligned}$$

pour trouver une décomposition QR à A, où Q est une matrice orthogonale, et R une matrice triangulaire supérieure. Le gros avantage de cette méthode est que le conditionnement en norme 2 de A est le même que celui de R. Ainsi, si l'on désire résoudre le système $Ax = b$, qui se ramène au système $QRx = b$, il suffit de résoudre le système $Rx = {}^t Q \, b$, qui est triangulaire, donc qui se résout facilement, et qui est bien conditionné pour peu que A le soit.