Méthodes numériques : optimisation.
Examen du 13 mai 2016 — Corrigé.

1 Préconditionnement de la méthode du gradient conjugué

1. La ligne 15 du programme initialise la variable `fAx` par `x*diag`, dont la i-ème coordonnée vaut $x_i d_i$, où x_i est la coordonnée numéro i de la variable notée `x` et d_i est la coordonée numéro i du vecteur d (stocké dans la variable *diag*).
   Ensuite on enlève x_i à la coordonnée numéro j de ce vecteur, et on enlève x_j à la coordonnée numéro i, dès qu'une arête (i,j) est dans la liste des arêtes.
   Au final la coordonnée numéro k de la variable `fAx` vaut

$$d_k x_k \quad - \sum_{\text{arêtes de la forme } (i,k)} x_i \quad - \sum_{\text{arêtes de la forme } (k,j)} x_j .$$

   Ou plus simplement, si on note n_{ij} le nombre d'arêtes de la forme (i,j) dans la liste d'arêtes (on peut avoir $n_{ij} = 0$ s'il n'y a pas de telle arête, et dans tous les cas on a $n_{ii} = 0$) on obtient que la coordonnée numéro k de la variable `fAx` vaut

$$d_k x_k - \sum_i n_{ik} x_i - \sum_j n_{kj} x_j = d_k x_k - \sum_j (n_{jk} + n_{kj}) x_j = \sum_j (d_k \delta_{jk} - n_{jk} - n_{kj}) x_j .$$

   On obtient bien une expression de la forme $\sum_j a_{kj} x_j = (Ax)_k$, en posant $A = (a_{ij})_{0 \leqslant i,j < N}$ la matrice de $M_N(\mathbb{R})$ donnée par $a_{kj} = d_k \delta_{jk} - n_{jk} - n_{kj}$.
   On obtient donc

$$\begin{aligned} &a_{ii} = d_i \text{ puisque } n_{ii} = 0, \\ &a_{ij} = -n_{ij} - n_{ji} = \text{nombre d'arêtes de la forme } (i,j) \text{ ou } (j,i), \text{ pour } i \neq j, \end{aligned}$$

   ce qui donne directement que A est une matrice symétrique.
   D'autre part, on peut voir que le vecteur d est initialisé avec toutes les coordonnées valant δ (ligne 8), puis que pour chaque arête de type (i,j) on ajoute 1. Donc au final d_k vaut δ plus le nombre d'arêtes de la forme (i,k) ou (k,j), c'est à dire

$$d_k = \delta + \sum_i n_{ik} + \sum_j n_{kj} = \delta + \sum_i (n_{ik} + n_{ki}).$$

   Autre méthode : initialiser `fAx` par `x*diag`, c'est prendre Dx, où D est la matrice diagonale ayant les d_k sur sa diagonale.
   Le vecteur ne contenant que x_j en position i, est le vecteur $E_{ij}x$ (où $E_{i,j} \in M_n(\mathbb{R})$ n'a que des zéros sauf en ligne i et colonne j). Donc pour chaque arête (i,j), le code retranche à `fAx` le vecteur $(E_{ij} + E_{ji})x$.
   Au final on obtient donc $A = D - \sum_{(i,j)\in\mathcal{L}} (E_{ij} + E_{ji})$, où $\mathcal{L}$ est la liste des arêtes. On voit donc bien que A est symétrique.
   De plus on peut voir de la même manière que $D = \delta \, \mathrm{I_N} + \sum_{(i,j)\in\mathcal{L}} (E_{ii} + E_{jj})$. Donc au final on obtient

$$A = \delta \, \mathrm{I_N} + \sum_{(i,j)\in\mathcal{L}} (E_{ii} + E_{jj} - E_{ij} - E_{ji}).$$

2. Si on voulait stocker A sous la forme d'une matrice, on aurait à stocker N^2 coefficients, dont un très grand nombre de zéros : il y a moins de $10N$ arêtes (voir ligne 3), donc moins de $2 \cdot 10N$ coefficients a_{ij} non nuls. De plus, on voit que le calcul de Ax grâce à la fonction `fA` est peu coûteux : N multiplications de réels pour l'initialisation de `fAx` (ligne 15) et pour chaque arête de la liste, 2 soustractions (soit moins de $2 \cdot 10N$ au total). Alors que l'on sait que le produit matriciel d'une matrice $N \times N$ par un vecteur de taille N nécessite de l'ordre de N^2 multiplications et additions.
On ne va donc pas résoudre le système $Ax + b = 0$ par une méthode directe mais par une méthode d'approximation, par exemple par une méthode de minimisation de $\frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$. Si la matrice A est définie positive, on sait que la méthode du gradient conjuguée est en règle générale bien adaptée aux problèmes de grande dimension, et que ses itérées ne nécessitent pas de connaître la matrice A, mais seulement de savoir calculer Ax. On pourrait aussi utiliser simplement des méthodes de descente de gradient, à pas fixe ou à pas optimal (puisque dans ce cas on peut calculer exactement le pas optimal, la fonction à minimiser est quadratique), mais on sait qu'elles sont moins performantes en pratique que la méthode du gradient conjuguée.

3. (a) On a tout d'abord que E est bien inversible (sinon E^T ne le serait pas non plus, et donc M non plus). Comme $\nabla\hat{f}(\hat{x}) = \hat{A}\hat{x} + \hat{b}$, on a

$$\hat{x} \text{ point critique de } \hat{f} \Leftrightarrow \hat{A}\hat{x} + \hat{b} = 0 \Leftrightarrow E^{-1}A(E^{-1})^T\hat{x} + E^{-1}b = 0 \Leftrightarrow A(E^{-1})^T\hat{x} + b = 0$$

Et donc si $Ax + b = 0$, en posant $\hat{x} = E^T x$, on obtient que $\hat{x}$ est point critique de $\hat{f}$, et réciproquement si $\hat{x}$ est un point critique de $\hat{f}$ et $x = (E^{-1})^T\hat{x}$, alors on obtient $Ax + b = 0$.

(b) On prend $\hat{x}_0 = 0_{\mathbb{R}^n}$, on pose $\hat{r}_0 = \hat{A}\hat{x}_0 + \hat{b} = \hat{b}$ et $\hat{p}_0 = -\hat{r}_0 = -\hat{b}$. Les relations de récurrence sont, pour $k \geqslant 0$ et dès que $\hat{r}_k \neq 0$

$$\hat{x}_{k+1} = \hat{x}_k + \alpha_k\hat{p}_k, \text{ avec } \alpha_k = \frac{\langle \hat{r}_k, \hat{r}_k\rangle}{\langle \hat{p}_k, \hat{A}\hat{p}_k\rangle},$$

$$\hat{r}_{k+1} = \hat{r}_k + \alpha_k\hat{A}\hat{p}_k, \quad \text{et} \quad \hat{p}_{k+1} = -\hat{r}_{k+1} + \frac{\langle \hat{r}_{k+1}, \hat{r}_{k+1}\rangle}{\langle \hat{r}_k, \hat{r}_k\rangle}\hat{p}_k.$$

(c) On obtient donc pour $k \geqslant 0$, dès que $r_k \neq 0$ (ce qui est équivalent à $\hat{r}_k \neq 0$) :

$$E^T x_{k+1} = E^T x_k + \alpha_k E^T p_k,$$
$$E^{-1} r_{k+1} = E^{-1} r_k + \alpha_k E^{-1}A(E^{-1})^T E^T p_k,$$
$$E^T p_{k+1} = -E^{-1} r_{k+1} + \frac{\langle E^{-1}r_{k+1}, E^{-1}r_{k+1}\rangle}{\langle E^{-1}r_k, E^{-1}r_k\rangle} E^T p_k,$$

avec $\alpha_k = \frac{\langle E^{-1}r_k, E^{-1}r_k\rangle}{\langle E^T p_k, E^{-1}A(E^{-1})^T E^T p_k\rangle}$. Comme $\langle E^{-1}r_k, E^{-1}r_k\rangle = \langle r_k, (E^{-1})^T E^{-1} r_k\rangle = \langle r_k, M^{-1}r_k\rangle$ et que $\langle E^T p_k, E^{-1}A(E^{-1})^T E^T p_k\rangle = \langle p_k, EE^{-1}Ap_k\rangle = \langle p_k, Ap_k\rangle$, on obtient bien la valeur donnée pour α_k et en multipliant à gauche les relations de récurrence pour x_k, r_k et p_k par $(E^{-1})^T$, E et $(E^{-1})^T$, on obtient bien les relations voulues.
On a $x_0 = (E^{-1})^T\hat{x}_0 = 0_{\mathbb{R}^n}$, puis

$$r_0 = E\hat{r}_0 = E(\hat{A}\hat{x}_0 + \hat{b}) = A(E^{-1})^T\hat{x}_0 + b = Ax_0 + b = b.$$

Et enfin

$$p_0 = (E^{-1})^T\hat{p}_0 = -(E^{-1})^T\hat{r}_0 = -(E^{-1})^T E^{-1} r_0 = -M^{-1}r_0 = -M^{-1}b.$$

On peut donc calculer r_0 et p_0 seulement avec l'expression de M^{-1} (sans connaître précisément la décomposition particulière $M = EE^T$). Et les relations de récurrence que l'on obtient au final ne font pas non plus intervenir E.

4. (a) La matrice M est la matrice diagonale avec $m_{ii} = d_i$. On ne la stocke pas sous la forme d'une matrice qui contiendrait des zéros inutilement. Le calcul de $M^{-1}x$ se fait simplement en N multiplications, alors que si on la stockait comme une matrice, cela nécessiterait N^2 multiplications et additions, en plus du problème d'espace de stockage.

(b) On peut écrire par exemple :

```
def GCP(b,eps=1e-3,Nmax=100):
    x=zeros(N)
    r=fA(x)+b
    iMr=invM(r)
    p=-iMr
    psriMr=prodscal(r,iMr)
    n2r=prodscal(r,r)
    tol=n2r*eps**2
    listen2=[n2r]
    k=0

    while k<Nmax and n2r>tol:
        k+=1

        Ap=fA(p)
        alpha=psriMr/prodscal(p,Ap)
        x+=alpha*p
        r+=alpha*Ap
        iMr=invM(r)
        nouvpsriMr=prodscal(r,iMr)
        p=-iMr+nouvpsriMr/psriMr*p
        psriMr=nouvpsriMr
        n2r=prodscal(r,r)
        listen2.append(n2r)
    return x,listen2
```

(c) On peut écrire :

```
b=rand(N)
compteurps=0
compteurfA=0
x,listen2=GC(b,1e-8)
print(compteurps,compteurfA)
semilogy(sqrt(listen2))
```

5. On sait que le préconditionnement peut améliorer la vitesse de convergence. D'autre part on sait qu'en grande dimension, lorsque le nombre de conditionnement K de A est grand, la méthode de descente de gradient à pas optimal est bien moins performante (taux de l'ordre de $\frac{L-\ell}{L+\ell} \approx 1 - \frac{2}{K}$) que celle du gradient conjugué (taux asymptotique de l'ordre de $\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}} \approx 1 - \frac{2}{\sqrt{K}}$). On peut donc attribuer la méthode 1 au gradient conjugué (23 itérations pour atteindre une tolérance relative de 10^{-8}), la méthode 2 au gradient à pas optimal (au bout de 100 itérations, on n'a toujours pas atteint la tolérance relative 10^{-5}), et enfin la méthode 3 au gradient conjugué préconditionné (pour lequel on espère que le nombre de conditionnement de $\widehat{A}$ est plus petit que celui de A), qui atteint la tolérance relative 10^{-8} en seulement 14 itérations.

   On observe sur l'échelle semilogarithmiques des courbes ayant l'allure de droites, ce qui indique une convergence linéaire pour les trois méthodes. Le fait que l'erreur augmente initialement est possible : bien que ce soient des méthodes de descente pour la fonction f, on a bien que $f(x_k) - f(x_*)$ diminue à chaque étape, mais c'est une quantité que l'on ne connaît pas puisqu'on ne connaît pas x_*. On n'est pas assuré que la norme des gradients (qui est la quantité à laquelle on a accès) diminue à chaque étape.

6. * Lorsque $\delta = 0$, d'après la question 1, en prenant le vecteur $x = (1, 1, \ldots, 1)$, on obtient que $(Ax)_j = d_j - \sum_j (n_{ij} + n_{ji}) = 0$, donc A n'est pas injective (donc pas définie positive).
Dans le cas général on obtient

$$\begin{aligned}
\langle x, Ax\rangle &= \sum_j \Big(\delta + \sum_i (n_{ij} + n_{ji})\Big) x_j^2 - \sum_{i,j} x_i (n_{ij} + n_{ji}) x_j \\
&= \delta \|x\|^2 + \sum_{i,j} (n_{ij} + n_{ji})(x_j^2 - x_i x_j) \\
&= \delta \|x\|^2 + \frac{1}{2} \sum_{i,j} (n_{ij} + n_{ji})(x_j^2 - x_i x_j) + \frac{1}{2} \sum_{j,i} (n_{ji} + n_{ij})(x_i^2 - x_j x_i) \\
&= \delta \|x\|^2 + \frac{1}{2} \sum_{i,j} (n_{ij} + n_{ji})(x_j^2 - 2 x_i x_j + x_i^2) \\
&= \delta \|x\|^2 + \frac{1}{2} \sum_{i,j} (n_{ij} + n_{ji})(x_j - x_i)^2 \geqslant \delta \|x\|^2.
\end{aligned}$$

Et on a donc, si $\delta > 0$, que $\langle x, Ax\rangle > 0$ dès que $x \neq 0$.
Autre méthode : en utilisant l'autre formule pour A, on a que $\langle x, (E_{ii} + E_{jj} - E_{ij} - E_{ji})x\rangle = x_i x_i + x_j x_j - x_i x_j - x_j x_i = (x_i - x_j)^2$. On obtient donc directement

$$\langle x, Ax\rangle = \delta \|x\|^2 + \sum_{(i,j) \in \mathcal{L}} (x_i - x_j)^2.$$

On obtient donc directement que si tous les x_i sont égaux et $\delta = 0$ alors $\langle x, Ax\rangle = 0$, ce qui donne que A n'est pas définie positive, et que A est définie positive si $\delta > 0$.

2 Précision de la méthode de Newton

1. Formule de récurrence : $x_{k+1} = x_k - H_f(x_k)^{-1} \nabla f(x_k)$.
Soit x tel que $\|x_* - x\| \leqslant 1$. Comme on a $\langle h, H_f(x)h\rangle \geqslant \|h\|^2$ pour tout $h \in \mathbb{R}^n$, alors $H_f(x)$ est symétrique définie positive (donc inversible) et sa plus petite valeur propre est supérieure à 1. Donc la plus grande des valeurs propres de $H_f(x)^{-1}$ (qui est l'inverse de la plus petite de H_f) est donc inférieure à 1. On obtient donc que $\|H_f(x)^{-1}h\| \leqslant \|h\|$.
On pouvait aussi écrire $\|H_f(x)^{-1}h\|^2 \leqslant \langle H_f(x)^{-1}h, H_f(x)H_f(x)^{-1}h\rangle \leqslant \|H_f(x)^{-1}h\|\|h\|$ par l'inégalité de Cauchy-Schwarz, et simplifier par $\|H_f(x)^{-1}h\|$ dès que $h \neq 0$ pour obtenir l'estimation.
La hessienne en x_* est donc symétrique définie positive (puisque $\|x_* - x_*\| = 0 \leqslant 1$), et donc le point critique x_* est un point de minimum local strict. Comme on ne sait pas quel est le comportement de f en dehors des points tels que $\|x - x_*\| \leqslant 1$, on ne peut rien dire sur le fait que ce minimum soit global (il se pourrait même que f n'ait pas de minimum, par exemple pour la fonction $x \to \frac{1}{2}x^2 - \frac{1}{12}x^4$ de $\mathbb{R} \to \mathbb{R}$, qui vérifie bien les hypothèses, mais tend vers $-\infty$ en $\pm\infty$).
2. On fait comme dans le cours (même si l'hypothèse est légèrement différente). Si $\|x - x_*\| \leqslant 1$, on applique la formule de Taylor avec reste intégral à ∇f entre x_* et x :

$$\nabla f(x_*) = \nabla f(x) + \int_0^1 H_f(x + t(x_* - x))(x_* - x) dt.$$

Donc en soustrayant $H_f(x)(x_* - x)$, on obtient

$$\nabla f(x_*) - \nabla f(x) - H_f(x)(x_* - x) = \int_0^1 [H_f(x + t(x_* - x)) - H_f(x)](x_* - x) dt.$$

En posant $y = x + t(x_* - x)$, on a bien y sur le segment joignant x_* à x, et donc (pour $t \neq 0$)

$$\begin{aligned}\|[H_f(x + t(x_* - x)) - H_f(x)](x_* - x)\| &= \|\frac{1}{t}[H_f(y) - H_f(x)](y - x)\| \\ &= \frac{1}{t}\|[H_f(x) - H_f(y)](x - y)\| \leqslant \frac{1}{t}\|x - y\|^2 = t\|x_* - x\|^2.\end{aligned}$$

On a donc

$$\|\nabla f(x_*) - \nabla f(x) - H_f(x)(x_* - x)\| \leqslant \int_0^1 t\|x_* - x\|^2 = \frac{1}{2}\|x - x_*\|^2.$$

On montre la suite par récurrence (exactement comme dans le cours) : si $\|x_k - x_*\| \leqslant 1$, alors

$$\begin{aligned}\|x_{k+1} - x_*\| &= \|x_k - x_* - H_f(x_k)^{-1}\nabla f(x_k)\| = \left\|H_f(x_k)^{-1}\Big[H_f(x_k)(x_k - x_*) - \nabla f(x_k)\Big]\right\| \\ &\leqslant \|H_f(x_k)(x_k - x_*) - \nabla f(x_k)\| = \|\nabla f(x_*) - \nabla f(x_k) - H_f(x_k)(x_* - x_k)\| \\ &\leqslant \frac{1}{2}\|x_k - x_*\|^2,\end{aligned}$$

et donc on obtient aussi $\|x_{k+1} - x_*\| \leqslant \frac{1}{2} \leqslant 1$ (on a utilisé ci-dessus le fait que $\nabla f(x_*) = 0$). Donc comme $\|x_0 - x_*\| \leqslant 1$, on a donc $\|x_k - x_*\| \leqslant 1$ pour tout k, et le calcul ci-dessus montre donc que $\|x_{k+1} - x_*\| \leqslant \frac{1}{2}\|x_k - x_*\|^2$.

3. La fonction f étant de classe C^2, on a que $x \to \|H_f(x)\|$ est continue sur le compact $\overline{B}(x_*, 1)$, donc y est bornée par une constante $K > 0$. On a alors

$$\begin{aligned}\|\nabla f(x)\| = \left\|\nabla f(x_*) + \int_0^1 H_f(x + t(x_* - x))(x_* - x)dt\right\| &\leqslant \int_0^1 \|H_f(x + t(x_* - x))\|\|x_* - x\|dt \\ &\leqslant \int_0^1 K\|x_* - x\|dt = K\|x_* - x\|.\end{aligned}$$

Puis, si $\|\hat{x_k} - x_*\| \leqslant 1$:

$$\begin{aligned}&\|\hat{x}_{k+1} - x_*\| = \|\hat{x}_k - x_* - \hat{N}(\hat{x}_k, \hat{g}(\hat{x}_k))\| \\ &\leqslant \|\hat{x}_k - x_* - H_f(\hat{x}_k)^{-1}\nabla f(\hat{x}_k)\| + \|H_f(\hat{x}_k)^{-1}\nabla f(\hat{x}_k) - \hat{N}(\hat{x}_k, \hat{g}(\hat{x}_k))\| \\ &\leqslant \tfrac{1}{2}\|\hat{x}_k - x_*\|^2 + \|H_f(\hat{x}_k)^{-1}\nabla f(\hat{x}_k) - H_f(\hat{x}_k)^{-1}\hat{g}(\hat{x}_k)\| + \|H_f(\hat{x}_k)^{-1}\hat{g}(\hat{x}_k) - \hat{N}(\hat{x}_k, \hat{g}(\hat{x}_k))\| \\ &\leqslant \tfrac{1}{2}\|\hat{x}_k - x_*\|^2 + \|H_f(\hat{x}_k)^{-1}[\nabla f(\hat{x}_k) - \hat{g}(\hat{x}_k)]\| + \eta_A\|\hat{g}(\hat{x}_k)\| \\ &\leqslant \tfrac{1}{2}\|\hat{x}_k - x_*\|^2 + \|\nabla f(\hat{x}_k) - \hat{g}(\hat{x}_k)\| + \eta_A(\|\nabla f(\hat{x})_k\| + \|\hat{g}(\hat{x}_k) - \nabla f(\hat{x})_k\|) \\ &\leqslant \tfrac{1}{2}\|\hat{x}_k - x_*\|^2 + \eta_g + \eta_A(K\|\hat{x}_k - x_*\| + \eta_g),\end{aligned}$$

ce qui est bien l'estimation demandée.

4. Il y avait une coquille dans l'énoncé, il fallait évidemment mettre des $\hat{x}$ à la place de x (ce qui n'aurait pas vraiment eu de sens), et il fallait lire $\max(4K\eta_A, 4\sqrt{\eta_g}))$ dans les deux expressions (et non pas $\max(4K\eta_A, 2\sqrt{\eta_g}))$ pour la première). Le barème a été très indulgent avec cela (mais en réalité une seule copie est arrivée jusqu'à là), l'important était la méthode pour obtenir les estimations, et la compréhension des ordres de grandeur.
Tant que $1 > \|\hat{x}_k - x_*\| \geqslant \max(4K\eta_A, 4\sqrt{\eta_g}))$, on a donc $\eta_g \leqslant \frac{1}{16}\|\hat{x}_k - x_*\|^2$, puis $K\eta_A \leqslant \frac{1}{4}\|\hat{x}_k - x_*\|$ et on obtient donc au final

$$\|\hat{x}_{k+1} - x_*\| \leqslant \|\hat{x}_k - x_*\|^2(\tfrac{1}{2} + \tfrac{1}{16}(1 + \eta_A) + \tfrac{1}{4}) \leqslant \|\hat{x}_k - x_*\|^2.$$

Et donc par récurrence on obtient bien que la suite des $\|\hat{x}_k - x_*\|$ est décroissante tant que l'on a $\|\hat{x}_k - x_*\| \geqslant \max(4K\eta_A, 4\sqrt{\eta_g}))$. Et dans ce cas on a bien par récurrence que comme on

a $\|\hat{x}_0 - x_*\| = \|\hat{x_0} - x_*\|^{2^0}$, alors si $\|\hat{x}_k - x_*\| \leqslant \|\hat{x_0} - x_*\|^{2^k}$ et $\|\hat{x}_k - x_*\| \geqslant \max(4K\eta_A, 4\sqrt{\eta_g}))$, alors

$$\|\hat{x}_{k+1} - x_*\| \leqslant \|\hat{x}_k - x_*\|^2 \leqslant (\|\hat{x_0} - x_*\|^{2^k})^2 = \|\hat{x_0} - x_*\|^{2^{k+1}}.$$

Comme la suite des $\|\hat{x_0} - x_*\|^{2^{k+1}}$ tend vers zéro (puisque $\|\hat{x_0} - x_*\| < 1$), on ne peut donc pas avoir $\|\hat{x}_k - x_*\| \geqslant \max(4K\eta_A, 4\sqrt{\eta_g}))$ pour tout k, donc il existe un plus petit rang k_0 telle que l'erreur $\|\hat{x}_{k_0} - x_*\|$ est plus petite que $\max(4K\eta_A, 4\sqrt{\eta_g}))$.
Si on a $K\eta_A \leqslant C\sqrt{\eta_g}$ avec C une constante de l'ordre de grandeur de un, alors on obtient que $\|\hat{x}_{k_0} - x_*\| \leqslant 4C\sqrt{\eta_g}$. Puis, en utilisant la question précédente, on obtient

$$\|\hat{x}_{k_0+1} - x_*\| \leqslant \frac{1}{2}(4C\sqrt{\eta_g})^2 + C\sqrt{\eta_g} \cdot 4C\sqrt{\eta_g} + \eta_g(1 + \eta_A) \leqslant C'\eta_g,$$

où $C' = 12C^2 + 1 + \eta_A$, que l'on peut considérer comme de l'ordre de grandeur de un.
Comme dans l'estimation de la question précédente, on observe un η_g seul, on ne peut pas espérer avoir mieux qu'une erreur de l'ordre de grandeur de η_g en utilisant seulement cette estimation.

5. On écrit la formule de Taylor à l'ordre un pour $\partial_j f$ entre x et $x + \varepsilon e_i$.

$$\partial_j f(x + \varepsilon e_i) - \partial_j f(x) = \int_0^1 \langle \nabla(\partial_j f(x + \varepsilon t e_i)), \varepsilon e_i \rangle dt = \varepsilon \int_0^1 \partial_i \partial_j f(x + \varepsilon t) dt.$$

De sorte que

$$\begin{aligned} |\tfrac{1}{\varepsilon}[\partial_j f(x + \varepsilon e_i) - \partial_j f(x)] - \partial_i \partial_j f(x)| &= |\int_0^1 \partial_i \partial_j f(x + \varepsilon t) dt - \int_0^1 \partial_i \partial_j f(x) dt| \\ &\leqslant \int_0^1 |\partial_i \partial_j f(x + \varepsilon t) - \partial_j f(x)| dt \leqslant L\varepsilon \int_0^1 \varepsilon t \, dt = \tfrac{1}{2} L\varepsilon. \end{aligned}$$

Pour $1 \leqslant i \leqslant j \leqslant n$,

$$\begin{aligned} |\hat{h}_{ij}(x) - \partial_i \partial_j f(x)| &= |\tfrac{1}{\varepsilon}[\hat{g}_j(x + \varepsilon e_i) - \hat{g}_j(x)] - \partial_i \partial_j f(x)| \\ &\leqslant |\tfrac{1}{\varepsilon}[\hat{g}_j(x + \varepsilon e_i) - \hat{g}_j(x)] - \tfrac{1}{\varepsilon}[\partial_j f(x + \varepsilon e_i) - \partial_j f(x)]| \\ &\quad + |\tfrac{1}{\varepsilon}[\partial_j f(x + \varepsilon e_i) - \partial_j f(x)] - \partial_i \partial_j f(x)| \\ &\leqslant \tfrac{1}{\varepsilon}|\hat{g}_j(x + \varepsilon e_i) - \partial_j f(x + \varepsilon e_i)| + \tfrac{1}{\varepsilon}|\hat{g}_j(x) - \partial_j f(x)| + \tfrac{1}{2} L\varepsilon \\ &\leqslant \tfrac{1}{\varepsilon}\|\hat{g}(x + \varepsilon e_i) - \nabla f(x + \varepsilon e_i)\| + \tfrac{1}{\varepsilon}\|\hat{g}(x) - \nabla f(x)\| + \tfrac{1}{2} L\varepsilon \\ &\leqslant \tfrac{2}{\varepsilon}\eta_g + \tfrac{1}{2} L\varepsilon. \end{aligned}$$

En minimisant cette somme de deux termes en fonction de ε, on obtient que les deux termes doivent être du même ordre de grandeur (sinon on peut diminuer celui dont l'ordre de grandeur est grand, l'autre étant négligeable), ce qui nous donne que $\frac{\eta_g}{\varepsilon}$ et ε doivent être du même ordre, ou encore que ε doit être de l'ordre de $\sqrt{\eta_g}$ (le calcul exact donne $\varepsilon = 2\sqrt{\frac{\eta_g}{L}}$), et on obtient que la précision de l'approximation est alors aussi de l'ordre de $\sqrt{\eta_g}$.
Pour calculer $\hat{H}(x)$, il faut calculer h_{ij} pour $1 \leqslant i \leqslant j \leqslant n$, c'est à dire calculer $\hat{g}_j(x)$ pour $1 \leqslant j \leqslant n$ (n évaluations) et $\hat{g}_j(x + \varepsilon e_i)$ pour $1 \leqslant i \leqslant j \leqslant n$ (soit $\frac{n(n+1)}{2}$ évaluations). Soit au total $\frac{n(n+1)}{2} + n = \frac{n(n+3)}{2}$ évaluations des fonctions $\hat{g}_i$.
Si on avait choisi les différences finies centrées, il aurait fallu évaluer les $\hat{g}_j(x+\varepsilon e_i)$ et les $\hat{g}_j(x-\varepsilon e_i)$ pour $1 \leqslant i \leqslant j \leqslant n$, soit $n(n+1)$ évaluations. On a donc un coût environ deux fois plus élevé, alors qu'on n'a pas besoin d'une telle précision : d'après la question précédente, on n'a besoin que d'une précision $\sqrt{\eta_g}$ pour la hessienne, le fait d'avoir une meilleure précision n'améliorera pas la précision de la méthode de Newton.

3 Taux de convergence de la méthode de gradient à pas fixe pour la norme $\|\cdot\|_A$

1. On a $x_{k+1} = x_k - \alpha\nabla f(x_k)$. Et on a $\nabla f(x) = Ax + b$. Donc ici

$$r_{k+1} = \nabla f(x_{k+1}) = Ax_{k+1} + b = A(x_k - \alpha r_k) + b = Ax_k + b - \alpha Ar_k = r_k - \alpha Ar_k.$$

2. Il y a un unique point critique x_* donné par $Ax_* + b = 0$ (car A est symétrique définie positive, donc inversible).
On a donc $f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle = \frac{1}{2}\langle x, Ax\rangle - \langle Ax_*, x\rangle$. En particulier $f(x_*) = -\frac{1}{2}\langle x_*, Ax*\rangle$, et donc
$$f(x) - f(x_*) = \frac{1}{2}\langle x, Ax\rangle - \langle Ax_*, x\rangle + \frac{1}{2}\langle x_*, Ax_*\rangle.$$
D'autre part, en utilisant que $\langle x, Ax_*\rangle = \langle Ax, x_*\rangle$ (car A est symétrique) et le fait que le produit scalaire est symétrique :
$$\begin{aligned}\frac{1}{2}\langle x - x_*, A(x - x_*)\rangle &= \frac{1}{2}\langle x, Ax\rangle - \frac{1}{2}\langle x, Ax_*\rangle - \frac{1}{2}\langle x_*, Ax\rangle + \frac{1}{2}\langle x_*, Ax_*\rangle \\ &= \frac{1}{2}\langle x, Ax\rangle - \langle Ax_*, x\rangle + \frac{1}{2}\langle x_*, Ax_*\rangle = f(x) - f(x_*).\end{aligned}$$
Comme A est symétrique définie positive, on obtient donc que $f(x) - f(x_*) \geqslant 0$, et même que $f(x) - f(x_*) > 0$ dès que $x - x_* \neq 0$. Donc x_* est un point de minimum de f (et c'est même l'unique point de minimum).
Enfin on a $A(x_k - x_*) = Ax_k - Ax_* = Ax_k + b = r_k$, et donc $x_k - x_* = A^{-1}r_k$. On obtient donc
$$f(x_k) - f(x_*) = \frac{1}{2}\langle x - x_*, A(x - x_*)\rangle = \frac{1}{2}\langle A^{-1}r_k, r_k\rangle.$$

3. On a $A^{-1}r_k = \sum_{i=1}^n c_{i,k}A^{-1}e_i = \sum_{i=1}^n c_{i,k}\frac{1}{\lambda_i}e_i$. Et comme la base des e_i est orthonormale, on obtient
$$\varepsilon_k = \frac{1}{2}\langle r_k, A^{-1}r_k\rangle = \frac{1}{2}\sum_{i=1}^n \frac{c_{i,k}^2}{\lambda_i}.$$
D'autre part $r_{k+1} = r_k - \alpha Ar_k = \sum_{i=1}^n c_{i,k}(1 - \alpha\lambda_i)e_i$. Donc $c_{i,k+1} = (1 - \alpha\lambda_i)c_{i,k}$. Et on obtient
$$\varepsilon_{k+1} = \frac{1}{2}\sum_{i=1}^n \frac{c_{i,k+1}^2}{\lambda_i} = \frac{1}{2}\sum_{i=1}^n (1 - \alpha\lambda_i)^2\frac{c_{i,k}^2}{\lambda_i} \leqslant \max_{1\leqslant i\leqslant n}|1 - \alpha\lambda_i|^2\varepsilon_k.$$
Comme on a $1 - \alpha L \leqslant 1 - \alpha\lambda_i \leqslant 1 - \alpha\ell$, on obtient que $|1 - \alpha\lambda_i| \leqslant \rho(\alpha)$ pour tout i, ce qui donne l'estimation demandée.
Finalement, si on a $r_0 = \beta e_{i_{\max}}$, où $\lambda_{i_{\max}} = \ell$ ou L suivant que $\rho(\alpha) = |1 - \alpha\ell|$ ou $|1 - \alpha L|$ (de sorte que $\rho(\alpha) = |1 - \alpha\lambda_{i_{\max}}|$), on obtient par récurrence que $r_k = (1 - \alpha\lambda_{i_{\max}})^k r_0$, et donc on a toujours $r_{k+1} = (1 - \alpha\lambda_{i_{\max}})r_k$ et donc pour tout k, on a $\varepsilon_{k+1} = (1 - \alpha\lambda_{i_{\max}})^2\varepsilon_k$.
Autrement dit, si l'on prend $x_0 = A^{-1}(\beta e_{i_{\max}} + b)$ (de sorte que $r_0 = Ax_0 + b = \beta e_{i_{\max}}$), on a toujours $\varepsilon_{k+1} = \rho(\alpha)^2\varepsilon_k$.

4. On a donc $\frac{1}{2}\|x_{k+1} - x_*\|_A^2 = \varepsilon_{k+1} \leqslant \rho(\alpha)^2\varepsilon_k = \rho\frac{1}{2}(\alpha)^2\|x_k - x_*\|_A^2$, soit
$$\|x_{k+1} - x_*\|_A \leqslant \rho(\alpha)\|x_k - x_*\|_A.$$
On en déduit que si $\rho(\alpha) < 1$, on obtient le critère de convergence linéaire des x_k vers x_* en norme $\|\cdot\|_A$, avec un taux de convergence linéaire inférieur ou égal à $\rho(\alpha)$. On a
$$\rho(\alpha) < 1 \Leftrightarrow \begin{cases}|1 - \alpha\ell| < 1\\ |1 - \alpha L| < 1\end{cases} \Leftrightarrow \begin{cases}-1 < 1 - \alpha\ell < 1\\ -1 < 1 - \alpha L < 1\end{cases} \Leftrightarrow \begin{cases}\alpha \in]0, \frac{2}{\ell}[\\ \alpha \in]0, \frac{2}{L}[\end{cases} \Leftrightarrow \alpha \in]0, \frac{2}{L}[,$$

puisque $L > \ell$. Si $\alpha \in]0, \frac{2}{L}[$, on a convergence linéaire à taux $\rho(\alpha) < 1$ des x_k vers x_* en norme $\|\cdot\|_A$.
Enfin pour trouver le α qui minimise $\rho(\alpha)$, on étudie la fonction. Graphiquement on observe que le minimum est atteint lorsque $|1-\alpha\ell| = 1-\alpha\ell = -1+\alpha L = |1-\alpha L|$, c'est à dire pour $\alpha = \frac{2}{L+\ell}$. En effet, si $0 \leqslant \alpha \leqslant \frac{2}{L+\ell}$, alors $1-\alpha\ell \geqslant \alpha L - 1$, et comme on a aussi $1-\alpha\ell \geqslant 1-\alpha L$, on obtient $1-\alpha\ell \geqslant |1-\alpha L|$, donc $|1-\alpha\ell| = 1-\alpha\ell$ et $\rho(\alpha) = 1-\alpha\ell$. Donc ρ est strictement décroissante sur $[0, \frac{2}{L+\ell}]$. Puis, si $\alpha \geqslant \frac{2}{L+\ell}$ alors $\alpha L - 1 \geqslant 1-\alpha\ell$ et d'autre part $\alpha L - 1 \geqslant \alpha\ell - 1$ donc de même $\alpha L - 1 \geqslant |1-\alpha\ell|$ donc $|1-\alpha L| = \alpha L - 1 = \rho(\alpha)$ donc ρ est strictement croissante sur $[\frac{2}{L+\ell}, +\infty]$.
La fonction ρ atteint donc son unique minimum en $\alpha = \frac{2}{L+\ell}$.

5. On obtient donc exactement les mêmes résultats en terme de taux de convergence que pour la méthode de gradient à pas fixe pour la norme euclidienne. Ce n'est pas étonnant, on pouvait déjà obtenir au moins le même taux (les normes étant équivalentes). Mais on a en plus les mêmes estimations précises entre deux itérées successives : $\|x_{k+1} - x_*\|_A \leqslant \rho(\alpha)\|x_k - x_*\|_A$ tout comme on avait $\|x_{k+1} - x_*\| \leqslant \rho(\alpha)\|x_k - x_*\|$.
Lorsque l'on choisit de prendre le α qui minimise $\rho(\alpha)$, on obtient le même taux $\rho(\frac{2}{L+\ell}) = \frac{L-\ell}{L+\ell}$ que dans le cas de la méthode de gradient optimal, pour lequel on avait : $\|\overline{x}_{k+1} - x_*\|_A \leqslant \frac{L-\ell}{L+\ell}\|\overline{x}_k - x_*\|_A$, où les $\overline{x}_k$ sont les itérées de la méthode de gradient à pas optimal.