MÉTHODES ITÉRATIVES DE RÉSOLUTION DE SYSTÈMES LINÉAIRES

David Ryckelynck

Centre des Matériaux, Mines ParisTech

8 octobre 2015

Plan du cours

Motivations

- Le passage à l'échelle pour le calcul parallèle est plus facile avec les méthodes itératives qu'avec les méthodes directes.
- Pas de problème de numérotation des inconnues à optimiser.

Mais :

- La convergence en un nombre ≪raisonnable≫ d'itérations, n'est pas toujours acquise, elle dépend de la structure de la matrice, du point de départ, du critère d'arrêt... A étudier au cas par cas.
- Peu adapté à la résolution de problèmes à second membres multiples, si ceux-ci sont très différents les uns des autres.
- Il existe des méthodes de factorisation **L**.**U** multifrontales très performantes en calcul parallèle (voir MUMPS `http://mumps.enseeiht.fr/index.php?page=doc`).

La méthode de Gauss Seidel

On cherche $\mathbf{q} \in \mathbb{R}^N$ tel que :

$$\mathbf{K}.\mathbf{q} = \mathbf{F}$$

avec la séparation de **K** en sa diagonale **D**, sa partie inférieure **L** et sa partie supérieure **U** :

$$\mathbf{K} = \mathbf{L} + \mathbf{D} + \mathbf{U}$$

On choisit $\mathbf{q}^{(o)}$ et on construit l'itéré $\mathbf{q}^{(i+1)}$ tel que :

$$(\mathbf{L} + \mathbf{D}).\mathbf{q}^{(i+1)} = \mathbf{F} - \mathbf{U}.\mathbf{q}^{(i)}$$

Il s'agit d'un système linéaire à matrice triangulaire inférieure. Sa résolution se fait progressivement de l'indice 1 à l'indice N. c'est une opération dite de **descente**.

Le nombre d'opérations que l'on peut traiter en parallèle n'est pas très important. Surtout au début du calcul de $\mathbf{q}^{(i+1)}$. Il y a peu d'additions et de multiplications à réaliser en parallèle.

C'est une méthode de point fixe associée au problème $\mathbf{x} = (\mathbf{L} + \mathbf{D})^{-1}.(\mathbf{F} - \mathbf{U}.\mathbf{x})$. Il faut que le rayon spectral de $(\mathbf{L} + \mathbf{D})^{-1}.\mathbf{U}$ soit inférieur à 1 pour qu'il y ait convergence. Si **K** est symétrique définie positive, ou strictement à diagonale dominante ($|K_{ii}| > \sum_{j \neq i} |K_{ij}|$), la méthode converge.

Les méthodes de descente

On cherche $\mathbf{q} \in \mathbb{R}^N$ tel que :

$$\mathbf{K}.\mathbf{q} = \mathbf{F}$$

On suppose $\mathbf{K}$ symétrique définie positive (SPD en Anglais).

$$\mathbf{K}.\mathbf{q} = \mathbf{F} \quad \Leftrightarrow \quad \mathbf{q} = \arg\min_{\mathbf{q}^\star} J(\mathbf{q}^\star)$$

avec

$$J(\mathbf{q}^\star) = \left(\frac{1}{2}\mathbf{q}^{\star T}.\mathbf{K}.\mathbf{q}^\star - \mathbf{q}^{\star T}.\mathbf{F}\right)$$

Convexité :

$$J(\mathbf{q}^b) \geq J(\mathbf{q}^a) + (\mathbf{q}^b - \mathbf{q}^a)^T.\nabla J(\mathbf{q}^a)$$

$$\nabla J(\mathbf{q}^a) = \mathbf{K}.\mathbf{q}^a - \mathbf{F}$$

Preuve :

$$J(\mathbf{q}^b) = J(\mathbf{q}^a) + (\mathbf{q}^b - \mathbf{q}^a)^T.\nabla J(\mathbf{q}^a) + \frac{1}{2}(\mathbf{q}^b - \mathbf{q}^a)^T.\mathbf{K}.(\mathbf{q}^b - \mathbf{q}^a)$$

Donc le minimum local est le minimum global :

$$\nabla J(\mathbf{q}) = 0$$

La plus forte pente (Steepest Descent)

En $\mathbf{q}^a$ la plus forte pente est dans la direction du gradient :

$$\mathbf{r} = -\nabla J(\mathbf{q}^a)$$

donc $\mathbf{r} = \mathbf{F} - \mathbf{K}.\mathbf{q}^a$, c'est le résidu du système linéaire à résoudre.

Le problème de minimisation dans la direction $\mathbf{r}$ s'écrit : trouver $\alpha \in \mathbb{R}$ tel que

$$\alpha = \arg\min_{\alpha^\star} J(\mathbf{q}^a + \alpha^\star \, \mathbf{r})$$

On obtient :

$$\alpha = \frac{\mathbf{r}^T.\mathbf{r}}{\mathbf{r}^T.\mathbf{K}.\mathbf{r}}$$

Méthode de la plus forte pente

A chaque étape on minimise J dans la direction de la plus forte pente :

Initialisation avec $\mathbf{q}^{(o)}$ donné. On en déduit $\mathbf{r}^{(o)} = \mathbf{F} - \mathbf{K}.\mathbf{q}^{(o)}$.

- $\mathbf{y}^{(i)} = \mathbf{K}.\mathbf{r}^{(i)}$
- $\alpha^{(i)} = \frac{\mathbf{r}^{(i)\ T}.\mathbf{r}^{(i)}}{\mathbf{r}^{(i)\ T}.\mathbf{y}^{(i)}}$
- $\mathbf{q}^{(i+1)} = \mathbf{q}^{(i)} + \alpha^{(i)}\ \mathbf{r}^{(i)}$
- $\mathbf{r}^{(i+1)} = \mathbf{r}^{(i)} - \alpha^{(i)}\ \mathbf{y}^{(i)}$

Arrêt des itérations si $\|\mathbf{r}^{(i+1)}\| < \epsilon_{tol}$.

Périodiquement, le résidu exacte est recalculé pour réduire la propagation d'erreurs d'arrondi.

Toutes ces opérations sont facilement parallélisables sur plusieurs coeurs de processeurs.

Propriété de la méthode de la plus forte pente

Les directions de descente successives sont orthogonales :

$$\mathbf{r}^{(i+1)\,T}.\mathbf{r}^{(i)} = \mathbf{r}^{(i)\,T}.\mathbf{r}^{(i)} - \frac{\mathbf{r}^{(i)\,T}.\mathbf{r}^{(i)}}{\mathbf{r}^{(i)\,T}.\mathbf{y}^{(i)}}.\mathbf{y}^{(i)\,T}.\mathbf{r}^{(i)} = 0$$

Méthode du gradient conjugué

On souhaite une orthogonalité des directions de descente au sens de la matrice **K** (directions K-conjuguées) pour éviter des effets zig-zag lors de la convergence.

Initialisation avec $\mathbf{q}^{(o)}$ donné. On en déduit $\mathbf{r}^{(o)} = \mathbf{F} - \mathbf{K}.\mathbf{q}^{(o)}$, $\mathbf{d}^{(o)} = \mathbf{r}^{(o)}$.

- $\mathbf{y}^{(i)} = \mathbf{K}.\mathbf{d}^{(i)}$
- $\alpha^{(i)} = \frac{\mathbf{d}^{(i)\,T}.\mathbf{r}^{(i)}}{\mathbf{d}^{(i)\,T}.\mathbf{y}^{(i)}}$
- $\mathbf{q}^{(i+1)} = \mathbf{q}^{(i)} + \alpha^{(i)}\,\mathbf{d}^{(i)}$
- $\mathbf{r}^{(i+1)} = \mathbf{r}^{(i)} - \alpha^{(i)}\,\mathbf{y}^{(i)}$
- $\beta^{(i+1)} = \frac{\|\mathbf{r}^{(i+1)}\|^2}{\|\mathbf{r}^{(i)}\|^2}$
- $\mathbf{d}^{(i+1)} = \mathbf{r}^{(i+1)} + \beta^{(i+1)}\,\mathbf{d}^{(i)}$

Arrêt des itérations si $\|\mathbf{r}^{(i+1)}\| < \epsilon_{tol}$.

Propriété :

$$\mathbf{r}^{(i+1)T}.\mathbf{d}^{(i)} = 0$$

Propriété de convergence de la méthode du gradient conjugué

Après N itérations, deux cas de figure se présentent :

- Soit le résidu est nul $\mathbf{r}^{(N)} = 0$, donc il y a convergence.
- Soit il est orthogonal aux N directions de descente $\mathbf{d}^{(o)}, \ldots, \mathbf{d}^{(N-1)}$. Or ces directions sont K-orthogonales. Elles constituent une base de $\mathbb{R}^N$. Donc le résidu est nul. Il y a convergence.
- La convergence est plus rapide selon les modes ”locaux” que selon les modes ”globaux” (à grande longueur de variation).

Méthode du gradient conjugué préconditionné

Initialisation avec $\mathbf{q}^o$ donné. On en déduit $\mathbf{r}^{(o)} = \mathbf{F} - \mathbf{K}.\mathbf{q}^{(o)}$, $\mathbf{d}^{(o)} = \mathbf{M}^{-1}\,\mathbf{r}^{(o)}$, $\mathbf{g}^{(o)} = \mathbf{d}^{(o)}$.

- $\mathbf{z}^{(i)} = \mathbf{K}.\mathbf{d}^{(i)}$
- $\alpha^{(i)} = \frac{\mathbf{r}^{(i)\,T}.\mathbf{g}^{(i)}}{\mathbf{d}^{(i)\,T}.\mathbf{z}^{(i)}}$
- $\mathbf{q}^{(i+1)} = \mathbf{q}^{(i)} + \alpha^{(i)}\,\mathbf{d}^{(i)}$
- $\mathbf{r}^{(i+1)} = \mathbf{r}^{(i)} - \alpha^{(i)}\,\mathbf{z}^{(i)}$
- $\mathbf{g}^{(i+1)} = \mathbf{M}^{-1}.\mathbf{r}^{(i+1)}$
- $\beta^{(i+1)} = \frac{\mathbf{r}^{(i+1)\,T}.\mathbf{g}^{(i+1)}}{\mathbf{r}^{(i)\,T}.\mathbf{g}^{(i)}}$
- $\mathbf{d}^{(i+1)} = \mathbf{g}^{(i+1)} + \beta^{(i+1)}\,\mathbf{d}^{(i)}$

Arrêt des itérations si $\|\mathbf{r}^{(i+1)}\| < \epsilon_{tol}$.

Préconditionneur de Jacobi : $\mathbf{M} = \mathrm{diag}(K_{ii})$.

Préconditionneur par factorisation incomplète de Cholesky : On cherche $\mathbf{G}$ triangulaire inférieure aussi creuse que possible tel que :

$$\|\mathbf{K} - \mathbf{G}.\mathbf{G}^T\| < \Delta_{tol}, \qquad \mathbf{M} = \mathbf{G}.\mathbf{G}^T$$

On peut choisir $\mathbf{G}^{-1}$ tel que la structure creuse de cette matrice soit celle de la matrice $\mathbf{M}$.

La méthode GMRES (Generalized Minimum Residual Saad & Schultz 1986)

L'objectif est d'étendre l'algorithme de gradient conjugué au traitement des systèmes à matrice non symétrique.
La solution approchée donnée par l'algorithme du Gradient Conjugué appartient à un **sous-espace de Krylov** :

$$\widetilde{\mathbf{q}} - \mathbf{q}^{(o)} \in \text{span}\{\mathbf{r}^{(o)},\ \mathbf{K}.\mathbf{r}^{(o)},\ \mathbf{K}^2.\mathbf{r}^{(o)}, \ldots, \mathbf{K}^m.\mathbf{r}^{(o)}\}$$

Pour GMRES, on cherche le minimum de la norme des résidus dans un sous-espace :

$$\widetilde{\mathbf{q}} = \arg \min_{\mathbf{q}^\star \in \mathbf{q}^{(o)} + \text{span}\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(m)}\}} \|\mathbf{F} - \mathbf{K}.\mathbf{q}^\star\|$$

où $(\mathbf{v}^{(k)})_{k=1}^m$ est une base orthogonale du sous-espace de Krylov $\text{span}\{\mathbf{r}^{(o)},\ \mathbf{K}.\mathbf{r}^{(o)},\ \mathbf{K}^2.\mathbf{r}^{(o)}, \ldots, \mathbf{K}^{m-1}.\mathbf{r}^{(o)}\}$. Cette base est construite par la méthode d'Arnoldi : $\mathbf{v}^{(1)} = \mathbf{r}^{(o)}/\|\mathbf{r}^{(o)}\|$. Puis, de $j = 1$ à $m-1$

- $H_{ij} = (\mathbf{K}.\mathbf{v}^{(j)})^T.\mathbf{v}^{(i)},\ i = 1, \ldots, j$
- $\widehat{\mathbf{v}}^{(j+1)} = \mathbf{K}.\mathbf{v}^{(j)} - \sum_{i=1}^{j} \mathbf{v}^{(i)}\, H_{ij}$
- $H_{j+1\,j} = \|\widehat{\mathbf{v}}^{(j+1)}\|$
- Si $H_{j+1\,j} = 0$ alors on arrête les itérations, sinon $\mathbf{v}^{(j+1)} = \widehat{\mathbf{v}}^{(j+1)}/H_{j+1\,j}$

Propriété :

$$\mathbf{K}.[\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(m)}] = [\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(m+1)}].\mathbf{H}$$

Quelques remarques sur la convergence des modes propres à grande valeur propre

Considérons le cas d'une matrice exprimée dans la base de ses modes propres :

$$\mathbf{K} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_N \end{bmatrix}$$

Soit $\overline{\mathbf{q}}$ la solution exacte de $\mathbf{K}.\overline{\mathbf{q}} = \mathbf{F}$.

Alors

$$\mathbf{r}^{(o)} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_N \end{bmatrix} .(\overline{\mathbf{q}} - \mathbf{q}^{(o)})$$

et

$$\alpha^{(1)} = \frac{\sum_{i=1}^{N} \lambda_i^2 (\overline{q}_i - q_i^{(o)})^2}{\sum_{i=1}^{N} \lambda_i^3 (\overline{q}_i - q_i^{(o)})^2}$$

Les modes aux valeurs propres élevées jouent un rôle plus important que les autres dans le calcul de α. On constate que la vitesse de convergence de la solution est plus rapide pour les composantes portées par ces modes.

Méthodes multigrilles

On associe au problème à résoudre un problème à grille grossière, ne contenant pas les modes à valeurs propres élevées du problème d'origine. On note $\widehat{\mathbf{K}}$, $\widehat{\mathbf{q}}$ et $\widehat{\mathbf{F}}$, la matrice, la solution et le second membre du problème grossier. Ce problème a $\widehat{N}$ inconnues, avec $\widehat{N} < N$.

On définit un opérateur de restriction $\mathcal{R} \in \mathbb{R}^{\widehat{N} \times N}$ tel que :

$$\widehat{\mathbf{F}} = \mathcal{R}.\mathbf{F}$$

On définit un opérateur de prolongement $\mathcal{P} \in \mathbb{R}^{N \times \widehat{N}}$ tel que :

$$\widetilde{\mathbf{q}} = \mathcal{P}.\widehat{\mathbf{q}}$$

avec, pour $\mathbf{F}$ et $\widehat{\mathbf{q}}$ quelconques :

$$\widetilde{\mathbf{q}}^T.\mathbf{F} = \widehat{\mathbf{q}}^T.\widehat{\mathbf{F}} \Rightarrow \mathcal{R} = \mathcal{P}^T$$

Dans le cadre de la méthode des éléments finis $\mathcal{P}$ est construit par une méthode d'interpolation.

Méthodes multigrilles

Après quelques étapes du gradient conjugué on peut améliorer la convergence sur les modes à valeurs propres basses en traitant un problème directe à l'échelle grossière :

$$\mathbf{r} = \mathbf{F} - \mathbf{K}.\mathbf{q}^{(\nu)}$$

$$\widehat{\mathbf{K}}.\delta\widehat{\mathbf{q}} = \mathcal{R}.\mathbf{r}$$

$$\mathbf{q}^{(\nu+1)} = \mathbf{q}^{(\nu)} + \mathcal{P}.\delta\widehat{\mathbf{q}}$$

Puis refaire ν étapes de gradient conjugué. Il s'agit d'un cycle en V.

On peut aussi commencer par une étape directe sur le problème grossier. Il s'agit alors d'une approche Full-Multigrid.

Exercice

Découverte de :

- scipy.sparse.linalg.cg(A, b, x0=None, tol=1e-05)
- scipy.sparse.linalg.gmres(A, b, x0=None, tol=1e-05)

`http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.sparse.linalg.cg.html`

Sur quel type de matrice peut-on comparer ces deux méthodes ?

Comparer la précision pour un nombre d'itérations fixées (à l'aide du maximum d'itération).

Comparer le temps d'exécution pour une précision donnée.

Comparer la forme des résidus pour une précision donnée.

Retrouver dans le code source