Journée Mathématique d'Amiens

Résolution numérique de systèmes linéaires : une longue histoire

Jean-Paul Chehab

LAMFA, Université de Picardie Jules Verne, Amiens

Amiens, 4 juin 2008

Ce n'est pas si simple ...

Soit $A \in \mathcal{M}(\mathbb{R}^N)$,

$$\exists! x \in \mathbb{R}^N / Ax = b \iff det(A) \neq 0.$$

Une fois ce problème résolu, les ennuis commencent. La méthode de Cramer permet (en théorie) de calculer explicitement la solution du système. Donnons-nous un ordinateur effectuant $100\ 10^6$ opérations à la seconde

- Calcul d'un déterminant d'un système 20×20 : 15 400 ans
- Résolution système 100×100 : $3\ 10^{146}$ ans ($\simeq 10^{135}$ âge de l'univers)

Calcul d'un déterminant $\simeq n(n!)$ opérations

Plan

- Origine des problèmes
- Pleines ou creuses
- Les méthodes directes
- Une (petite) histoire des méthodes itéatives
  - Les méthodes de type relaxation
  - Les méthodes de Krylov
- Le préconditionnement
- Solveurs rapides

Problèmes issus de la physique

démarche

Modèle physique

$\Downarrow$

EDP non linéaires

$\Downarrow$

Discrétisation (différences finies, éléments finis)

$\Downarrow$

Linéarisation

$\Downarrow$

Systèmes creux $Ax = b$

Mécanique des fluides, mécanique des structures, électromagnétisme, réseaux électriques, problèmes de réservoirs donnent lieu à des systèmes de plusieurs millions d'inconnues !

Exemples

- Problème de Poisson

$$-\Delta u = f \text{ dans } \Omega,\ u = g \text{ sur } \partial\Omega$$

- Problème de Stokes

$$-\nu\Delta u + \nabla p = f \text{ dans } \Omega, \nabla . u = 0 \text{ dans } \Omega,\ u = g \text{ sur } \partial\Omega$$

- Vibration d'une poutre ou d'une plaque encastrée

$$\Delta^2 u = f \text{ dans } \Omega,\ u = 0, \frac{\partial u}{\partial n} = 0 \text{ sur } \partial\Omega$$

Autres problèmes

- Contrôle

  Equation de Sylvester $AX + XB = C$

- Traitement des images

  Régularisation de Tikhonov $\min \|Au - b\|^2 + \|\Gamma u\|^2$

- Approximation, maximum de vraisemblance, ajustement

$$\min_{a_j} \sum_{i=1}^{N} (y_i - \sum_{k=1}^{M} a_k f_k(x_i))^2 \Rightarrow 0 = \sum_{i=1}^{N} (y_i - \sum_{=1}^{M} a_j f_j(x_i)) f_k(x_i)$$

$$\sum_{j=1}^{M} \alpha_{k,j} a_j = \beta_k.$$

Système $M \times M$ avec $\alpha_{k,j} = \sum_{i=1}^{N} f_j(x_i) f_k(x_i)$ et $\beta_k = \sum_{i=1}^{N} y_i f_k(x_i)$.

Matrices creuses et matrices pleines

Matrices pleines

La plupart des coefficients sont non nuls. On stocke systématiquement par un tableau à deux entrées $A_{i,j}$

Matrices creuses

La plupart des coefficients sont nuls. On stocke seulement les éléments non nuls (gain de place mémoire). Différentes techniques : stockage diagonal (matrices bandes), stockage morse ...
Matrices de discrétisation d'opérateurs locaux (EDP) ; typiquement nb coeffs $\neq 0$ $\mathcal{O}(n)$ au lieu de n^2.

Méthodes directes

Méthode directe : résolution effective en un nombre fini d'opérations (dépend de la dimension)

idée

Se ramener à la résolution effective d'un sytème triangulaire équivalent
$Ax = b \iff Mx = b'$ (M triangulaire supérieure)
nb opérations $\mathcal{O}(n^3)$

Gauss

Méthode de Gauss (1800), connues des chinois pour les systèmes 3x3 au 3ième sciècle AC, nb opérations $\mathcal{O}(n^3)$

IDÉE

Ecrire A sous la forme d'un produit de deux matrices faciles à inverser (triangulaire, orthogonale)

- Cholesky (André-Louis Cholesky, 1910) : A symétrique définie positive, $A = S^T S$ avec S triangulaire supérieure

$$\text{résolution } Ax = b \iff S^T y = b \text{ puis } Sx = y$$

nb opérations $\mathcal{O}(\frac{n^3}{3})$

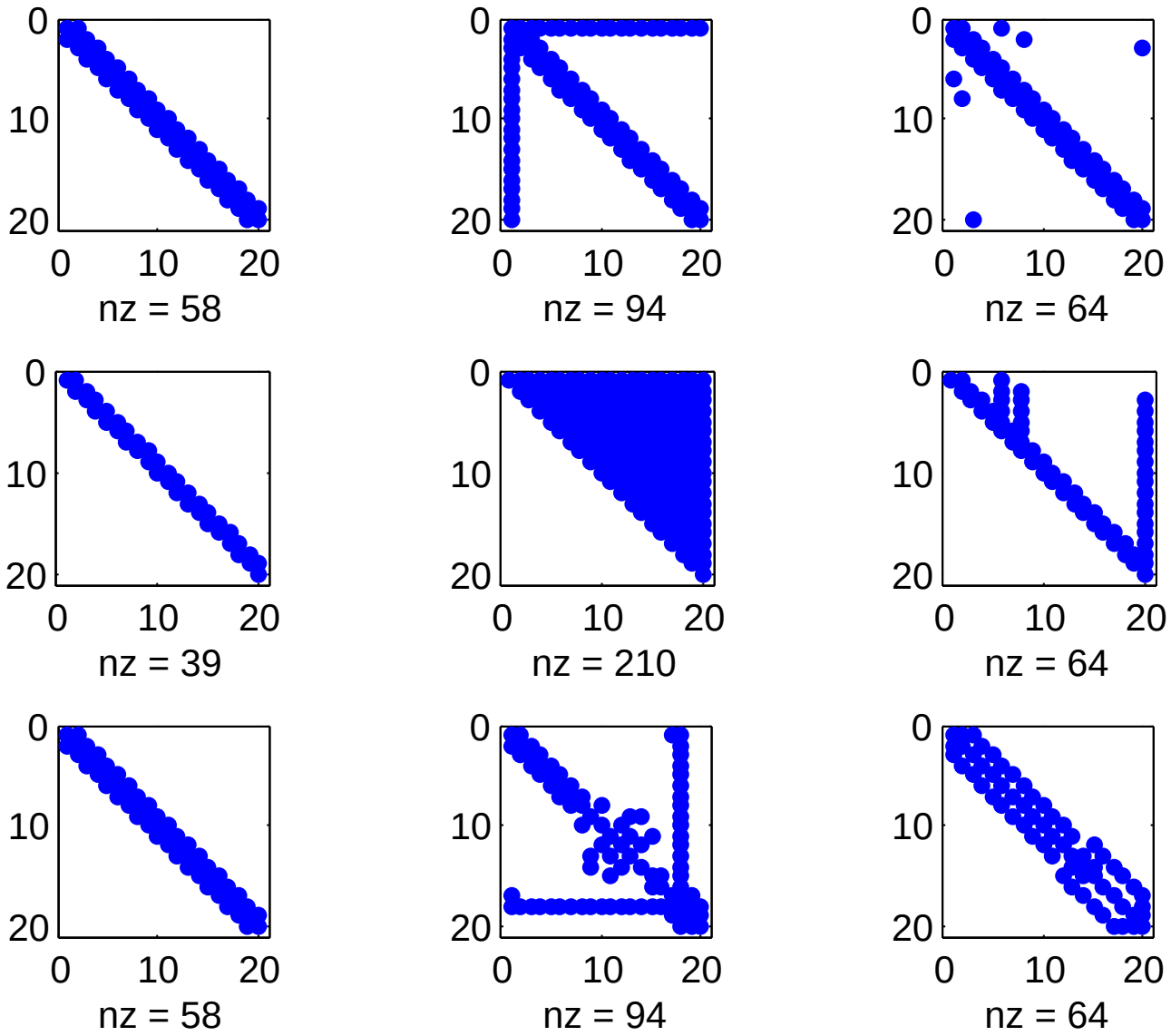
- Méthode LU (Alan Turing, 1948) A avec mineurs principaux non nuls, $A = LU$ avec L et U triangulaire inférieure (resp. supérieure)

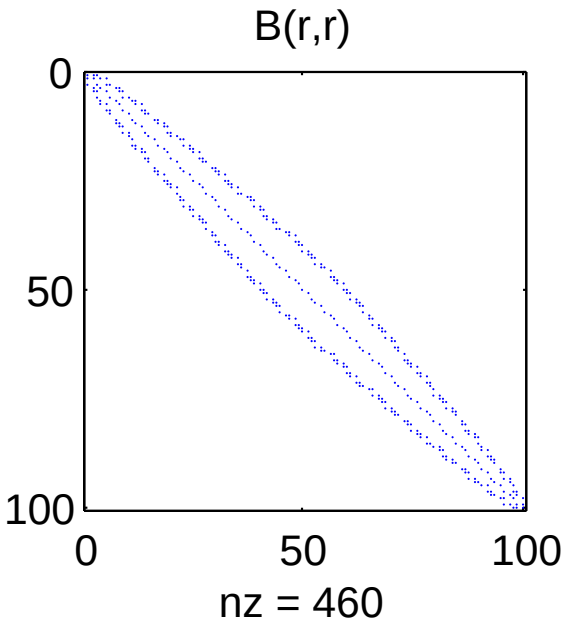
$$\text{résolution } Ax = b \iff Ly = b \text{ puis } Ux = y$$

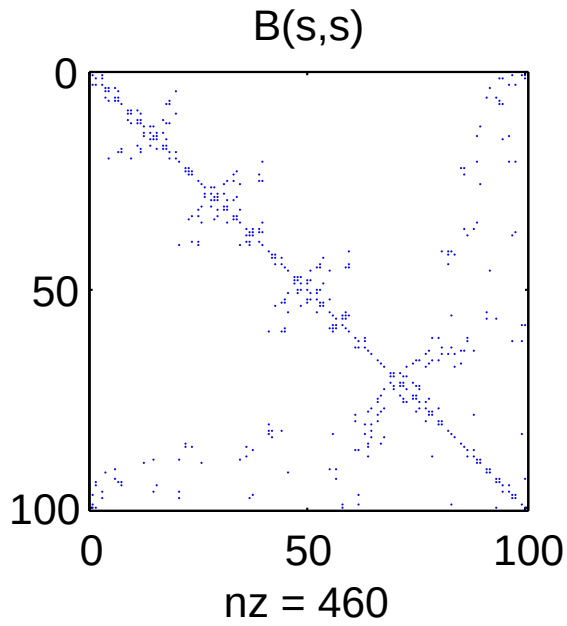
nb opérations $\mathcal{O}(\frac{2n^3}{3})$

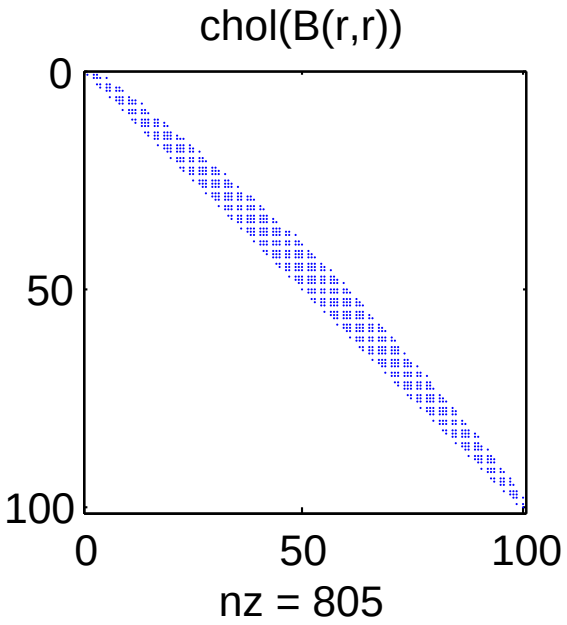
- Factorisation QR, (Wilkinson 1958), $A = QR$ avec Q orthogonale, R triangulaire supérieure.

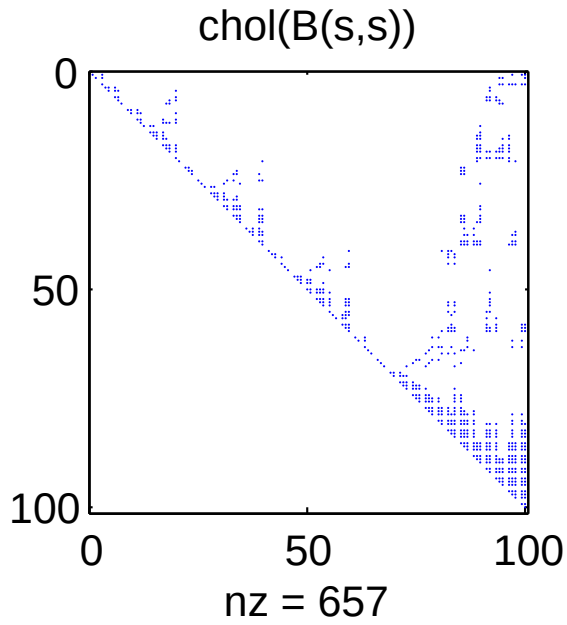
$$\text{résolution } Ax = b \iff Rx = Q^T b$$


nz = 58
nz = 94
nz = 64
nz = 39
nz = 210
nz = 64
nz = 58
nz = 94
nz = 64


B(r,r)
0
50
100
0
50
100
nz = 460


B(s,s)
0
50
100
0
50
100
nz = 460


chol(B(r,r))
0
50
100
0
50
100
nz = 805


chol(B(s,s))
0
50
100
0
50
100
nz = 657

Idée : construire $u^{(k)}$ telle que $\lim_{k\to+\infty} u^{(k)} = u = A^{-1}b$.

$$\text{On écrit } A = M - N$$

avec M inversible et facile à inverser (diagonale, triangulaire) :
$Au = b \iff Mu = Nu + b$

Résolution par point fixe de Picard $u^{(k+1)} = M^{-1}Nu^{(k)} + M^{-1}b$

Théorème

$$\lim_{k\to+\infty} u^{(k)} = u = A^{-1}b \iff \rho(M^{-1}N) < 1.$$

Remarque

On peut écrire $u^{(k+1)} = u^{(k)} - M^{-1}(Au^{(k)} - b)$

En particulier si $M = A$, *on converge en une itération. Idée : prendre* M *"proche" de* $A \Rightarrow$ *notion de préconditionnement.*

Quelques méthodes classiques

$$A = \begin{pmatrix} \ddots & & & \\ & \ddots & -F & \\ & D & & \\ -E & & \ddots & \\ & & & \ddots \end{pmatrix} \quad \text{avec} \quad \begin{array}{rl} D & \text{part. diagonale de } A \\ -E & \text{" triang. inf. stricte de } A \\ -F & \text{" triang. sup. stricte de } A \end{array}$$

- $M = D$: méthode de Jacobi (1846)
  (système d'ordre 7)
- $M = D - E$: méthode Gauss-Seidel (Gauss 1823, Seidel 1874)
- $M = I - P^{-1}A$ avec $P = (D - \omega E)D^{-1}(D - \omega F)$: SSOR (Varga 1960)
  (résolution d'un système 20 000 inconnues (1960) et à 100 000 inconnues en 1965 (pb phys nucléaire).
- $M = \frac{1}{\alpha} Id$: méthode de Richardson (Richardson 1910)

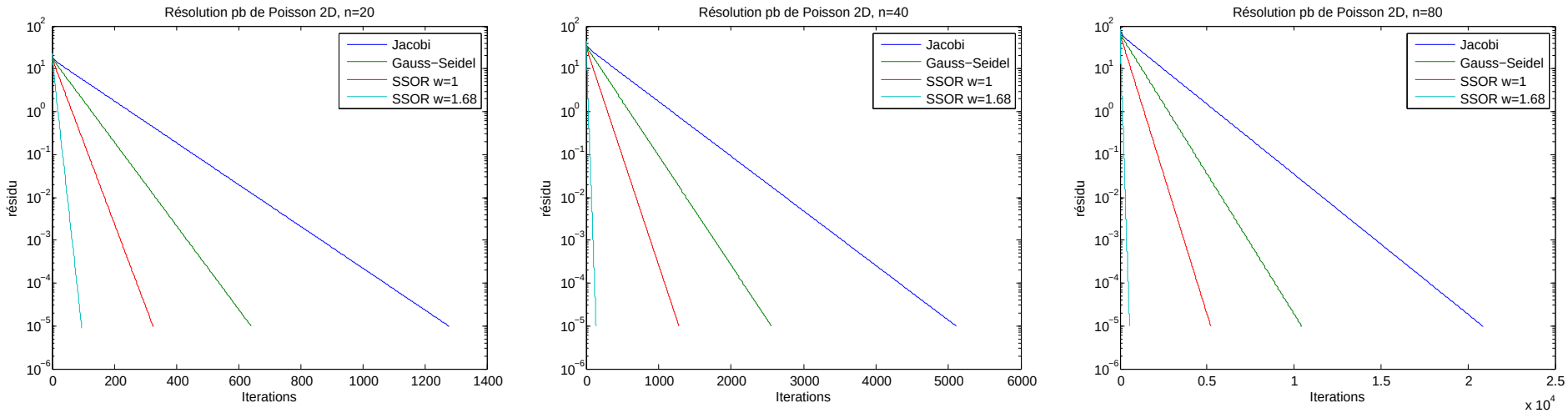


FIGURE: Comparaison des méthodes Jacobi, Gauss-Seidel et SSOR (n=20, n=40, n=80)

Méthode de projection 1

Soit A SDP. La méthode de Richardson (1910) est obtenue pour $M = \frac{1}{\alpha} Id$

$$u^{(k+1)} = u^{(k)} + \alpha(b - Au^{(k)}) = u^{(k)} + \alpha r^{(k)} \rightarrow r^{(k)} = (Id - \alpha A)^k r^{(0)}$$

Théorème

La méthode converge SSI $0 < \alpha < \frac{2}{\lambda_{max}}$ *et* $\rho(I - \alpha A)$ *est minimisé pour* $\alpha^* = \frac{2}{\lambda_{min} + \lambda_{max}}$.

Généralisations pour α variable : $u^{(k+1)} = u^{(k)} + \alpha_k r^{(k)}$ (D. Young, 1954). On a

$$r^{(k+1)} = \left(\prod_{k=0}^{k} (Id - \alpha_k A) \right) r^{(0)} = P_{k+1}(A) r^{(0)}.$$

Remarque

La méthode de la plus profonde descente due à L-A Cauchy (1847) consiste à prendre α_k *de sorte à minimiser* $\|r^{(k+1)}\|$.
Le problème de ces méthodes est qu'elles convergent lentement

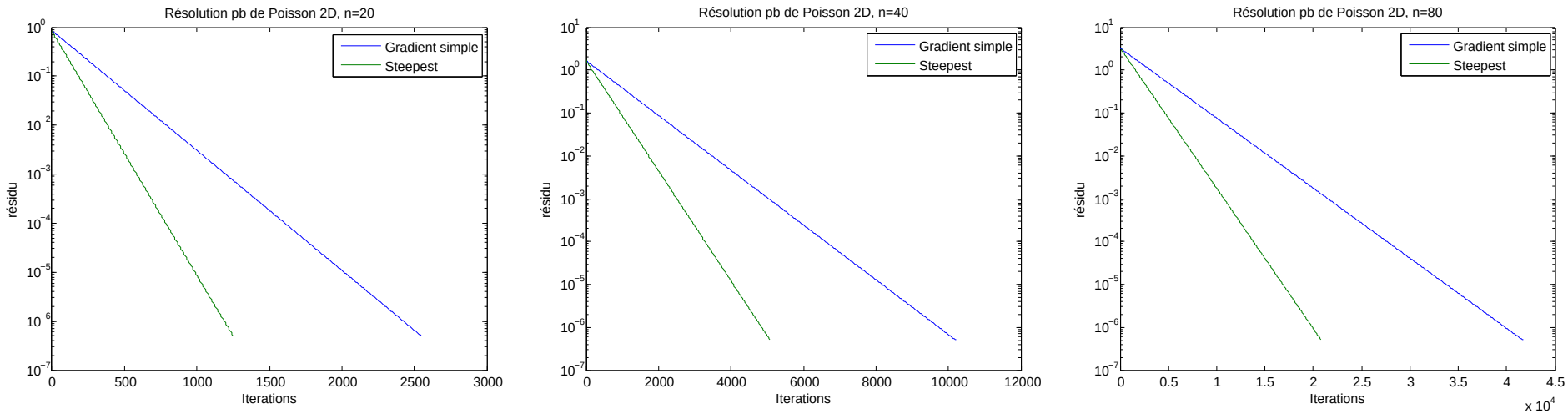


FIGURE: Comparaison des méthodes de gradient simple et de plus profonde descente (n=20, n=40, n=80)

Définition

Le conditionnement (introduit par J. von Neumann en 1948) d'une matrice en norme d'opérateur p est

$$Cond_p(A) = \|A\|_p \|A^{-1}\|_p$$

avec $\max_{\|u\|=1} \|Au\|_p$ *où* $\|.\|_p$ *est la norme* ℓ^p *vectorielle.*
On a toujours $cond_p(Id) = 1$.

Théorème

Pour la méthode de plus profonde descente, on a

$$\|e_k\|_A \leq \left(\frac{cond_2(A) - 1}{cond_2(A) + 1}\right)^k \|e_0\|_A$$

où $e_k = u_k - u$ *et* $\|e\|_A^2 =< e, Ae >$.

Définition

$$\textit{Méthode de descente } u^{(k+1)} = u^{(k)} + \alpha_k p_k$$

avec α_k *: paramètre de descente,* p_k *direction de descente.*

Remarque

Lorsque A *est SPD,*
$Au = b \iff u = \textit{argmin}\frac{1}{2} < Av, v > - < b, v >= \textit{argmin}\Phi(v)$. *On cherche* α_k *et* p_k *de sorte à ce que*

$$\Phi(u^{(k+1)}) < \Phi(u^{(k)}).$$

Par exemple $p_k = -\nabla\Phi(u^{(k)})$ *et* $0 < \alpha_k < \frac{2}{\rho(A)}$.
Modification de la direction de descente (Frankel 1950)

$$u^{(k+1)} = u^{(k)} + \beta_k p_k,\ p_k = r_k - \alpha_k p_{k-1}, p_{-1} = r_{-1} = 0.$$

On cherche α_k *et* β_k *de sorte à minimiser* $\phi(u^{(k+1)})$.

Méthode du gradient conjugué (Stiefel 1952)

CG

poser $r^0 = b - Au^0,\ p^0 = r^0$
pour $k = 0, \cdots$
poser $z^k = Ap^k$
poser $\alpha_k = \frac{<p^k, r^k>}{<p^k, z^k>}$
poser $u^{k+1} = u^k + \alpha_k p^k$
poser $r^{k+1} = r^k - \alpha_k z^k$
poser $\beta_{k+1} = \frac{<z^k, r^{k+1}>}{<p^k, z^k>}$
poser $p^{k+1} = r^{k+1} - \beta_{k+1} p^k$

Théorème

La méthode CG converge en au plus N itérations. De plus

$$\|e^k\|_A \leq \frac{2c^k}{1+2c^{2k}}\|e^0\|_A,\ \text{avec } c = \frac{\sqrt{K_2(A)}-1}{\sqrt{K_2(A)}+1}.$$

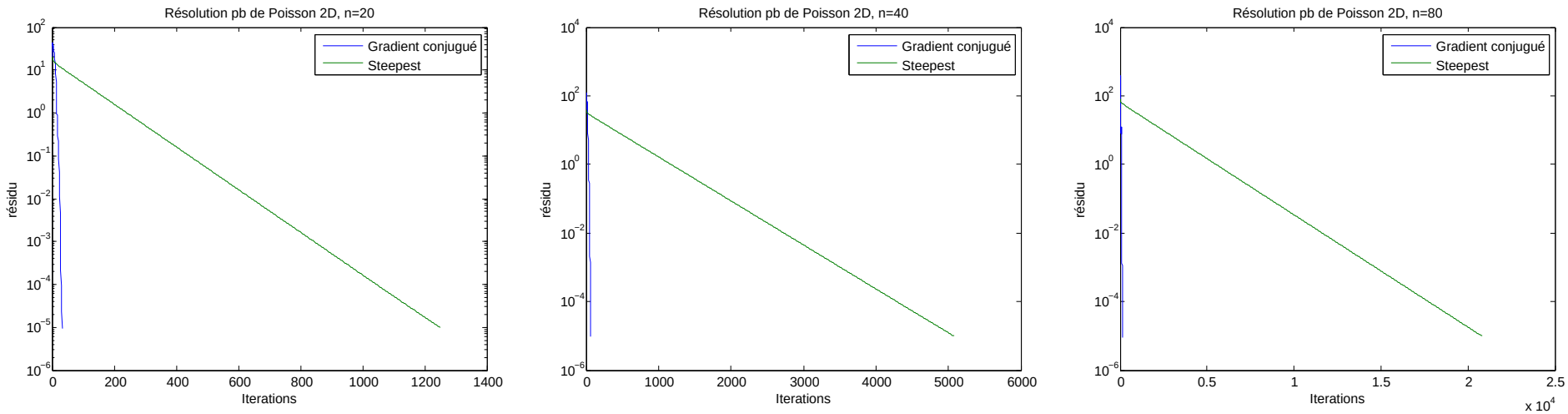


Figure: Comparaison des méthodes de gradient conjugué et de plus profonde descente (n=20, n=40, n=80)

Problèmes non symétriques

Problème

Jusqu'au début des années 1980, seul le gradient conjugué (éventuellement préconditionné) est efficace pour résoudre les pbs symétriques. Quand A n'est pas symétrique

$$Au = b \iff A^T Au = A^T b$$

On applique le gradient conjugué MAIS

- $cond(A^T A) >> cond(A)$
- Plus d'opérations.

Idée

Soit l'espace de Krylov $K^k = Vect(r^0, Ar^0, \cdots, A^{k-1}r^0)$. On cherche u^k sous la forme

$u^{k+1} = u^0 + z^k$ où $z^k \in K^k$ tel que $\|r^{k+1}\|_2^2 = \|r^0 - Az^k\|_2^2 = \min_{z \in K^k} \|r^0 - Az\|_2^2$

Si $dim(K^k) = k$ on considère $(v^1, \cdots, v^k)$ une base de K^k. On cherche à exprimer z^k dans cette base, sous la forme

$$z^k = \sum_{j=1}^{k} y_j v^j \text{ avec } \frac{\partial \|r^0 - \sum_{j=1}^{k} y_j A v^j\|_2^2}{\partial y_j} = 0, \ j = 1, \cdots, k$$

$$\sum_{j=1}^{k} < v^i, A^*Av^j > y_j =< v^i, A^*r^0 >, \ i = 1, \cdots, k.$$

$MY = F$ Pb : choix de la base v : orthonormée de K^k (Gramm-Schmidt)

Lemme

$$Av^i = \sum_{j=1}^{i} < v^i, A^* v^j > v^j \iff Au^k = b.$$

GMRES (Saad Schultz 1983)

`Calculer` $r^0 = b - Au^0$
`poser` $v^1 = \frac{r^0}{\|r^0\|_2}$
`pour k=1,`$\cdots$
`calculer` V^{k+1}
`calculer` $= H^k$
`résoudre` $(H^k)^* H^k Y = \beta (H^k)^* e^1$
`poser` $u^k = u^0 + QY$

C'est la méthode la plus utilisée actuellement : très robuste

GMRES (DÉTAIL)

Calculer $r^0 = b - Au^0$

poser $v^1 = r^0 / \|r^0\|_2$

pour k=1,$\cdots$

calculer v^{k+1}

calculer H_i^k $i = 1, \cdots, k+1$ par Arnoldi

Appliquer les rotations de Givens

pour $i = 1, \cdots k-1$

$$\begin{pmatrix} H_{i,k}^k \\ H_{i+1,k}^k \end{pmatrix} = \begin{pmatrix} c_i & s_i \\ -s_i & c_i \end{pmatrix} \begin{pmatrix} H_{i,k}^k \\ H_{i+1,k}^k \end{pmatrix}$$

calculer c_k et s_k

résoudre $HY = (\beta(0, \cdots, 0, 1)^T$

poser $u^k = u^0 + QY$

Formulation générale des méthodes de projection

Cadre

On se donne K et L deux sous-espaces de $\mathbb{R}^N$. On définit le problème approché

$$\text{Trouver } \bar{x} \in K \text{ tel que } b - A\bar{x} \perp L$$

Auterment dit

$$\text{Trouver } \bar{x} \in x_0 + K \text{ tel que } b - A\bar{x} \perp L$$

En posant $\bar{x} = x_0 + \delta$ et $r_0 = b - Ax_0$, le problème approché s'écrit

$$\text{Trouver } \delta \in K \text{ tel que } r_0 - A\delta \perp L$$

Prototype

Jusqu'à la convergence faire

1. Choisir deux sous espaces K et L
2. Choisir des bases $V = [v_1, \cdots, v_m]$ pour K et $W = [w_1, \cdots, w_m]$ pour L
3. Calculer
   - 3.1 $r = b - Ax$
   - 3.2 $y = (W^T AV)^{-1} W^T r$
   - 3.3 $x = x + Vy$

Deux cas importants

2cas

1. $L = AK$, alors $\|b - A\bar{x}\|_2 = \min_{z \in K} \|b - Az\|$
   Classe des méthodes de résidu minimal : C,CGR, GMRES, ORTHOMIN,CGNR
   Petrov-Galerkin
2. $L = K$: classe des méthodes de Galerkin ou de projection orthogonale. Quand A est SDP

$$\|x^* - x\|_A = \min_{z \in K} \|x^* - z\|_A$$

$$-\Delta u + \gamma \left(e^{xy} \frac{\partial u}{\partial x} + e^{-xy} \frac{\partial u}{\partial y} \right) + \alpha u = f$$

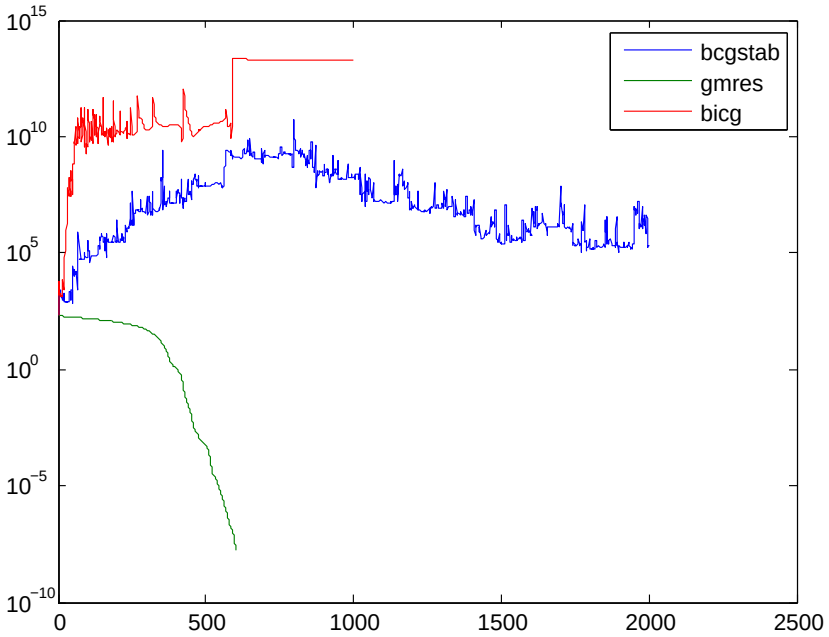


Figure: Comparaison gmres, bcgstab,bicg, n=200, système 40 000 $\times$ 40 000

Converger plus rapidement

Le Gradient s'est imposé (Reid, 1972) lorsqu'on a utilisé le préconditionnement (D.J. Evans, 1968)
Idée : résoudre le système équivalent

$$M^{-1}Au = M^{-1}b$$

avec $cond(M^{-1}A) << cond(A)$ et M facile à inverser

- Préconditionnement gauche $M^{-1}Au = M^{-1}b$
- Préconditionnement droit $AM^{-1}v = b$, avec $M^{-1}v = u$
- Préconditionnement décomposé $M_L^{-1}AM_R^{-1}v = M_L^{-1}b$ avec $M_R^{-1}v = u$ et $M = M_L M_R$.

En pratique, cela revient à résoudre un système supplémentaire $Kz = y$ avec $K \simeq A$.

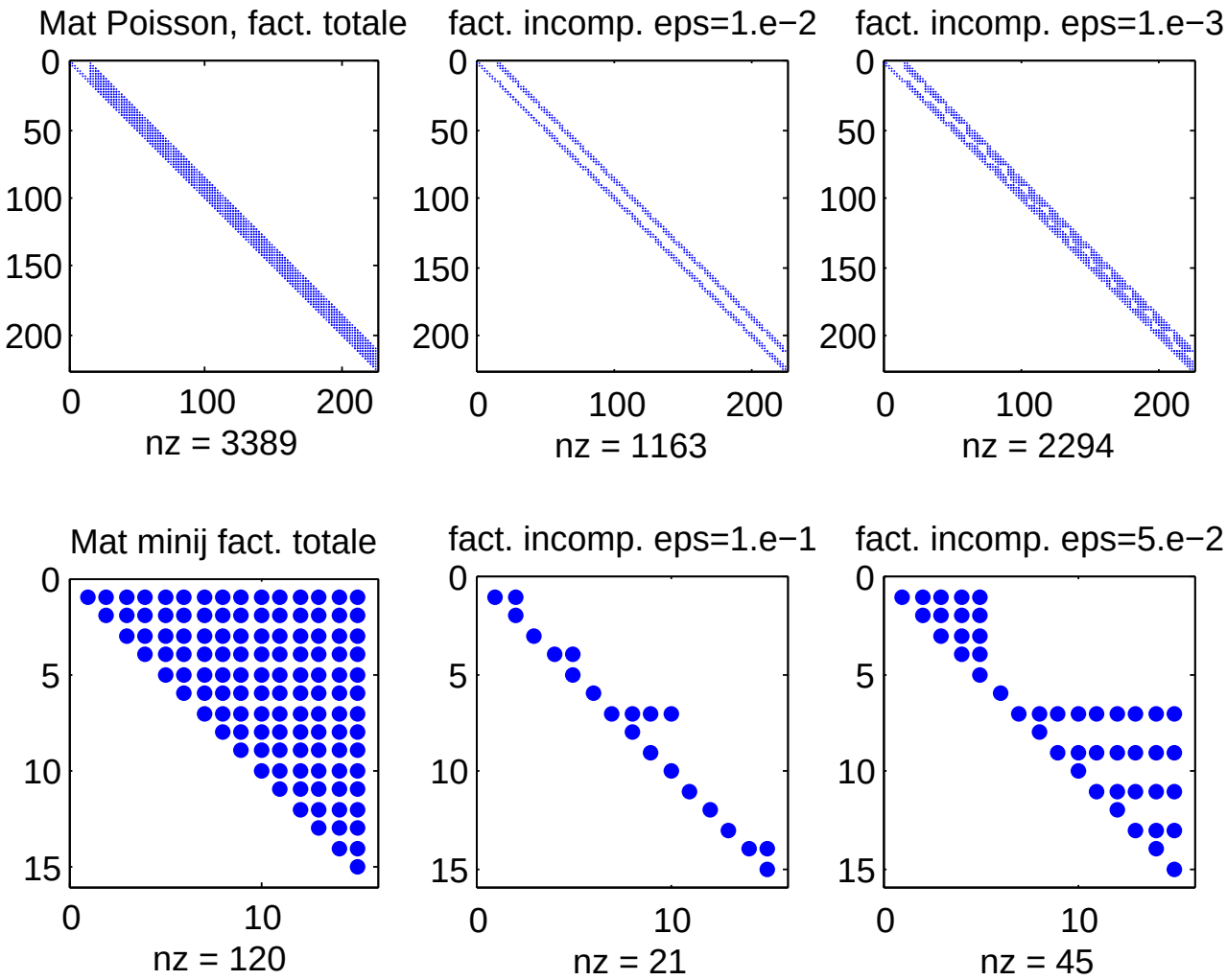
- Factorisations incomplètes (LU, Cholesky)

$$A = LU; A = L_0U_0 + R \text{ avec } L_0 \text{ et } U_0 \text{ ayant peu déléments}$$

- Minimisation de norme de Frobenius (approximations d'inverses creuses)

$$\min \|Id - AM|\|_F^2 = \min \sum_{i=1}^{N} \sum_{j=1}^{N} (\delta_{i,j} - \sum_{k=1}^{N} A_{ik} M_{kj})^2$$

  se calcule numériquement par une méthode de descente, le profil de M peut être prescrit (Benson, 1973).
  Flots d'inverses approchés creux (C, 2002, 2007)
- Préconditionneurs hiérarchiques (H. Yserentant, 1986), Multiniveaux recursifs ARMS (Saad, 2002)
- Décomposition de domaines
- De manière générale, tout solveur partiel du système


Mat Poisson, fact. totale
nz = 3389
fact. incomp. eps=1.e−2
nz = 1163
fact. incomp. eps=1.e−3
nz = 2294
Mat minij fact. totale
nz = 120
fact. incomp. eps=1.e−1
nz = 21
fact. incomp. eps=5.e−2
nz = 45

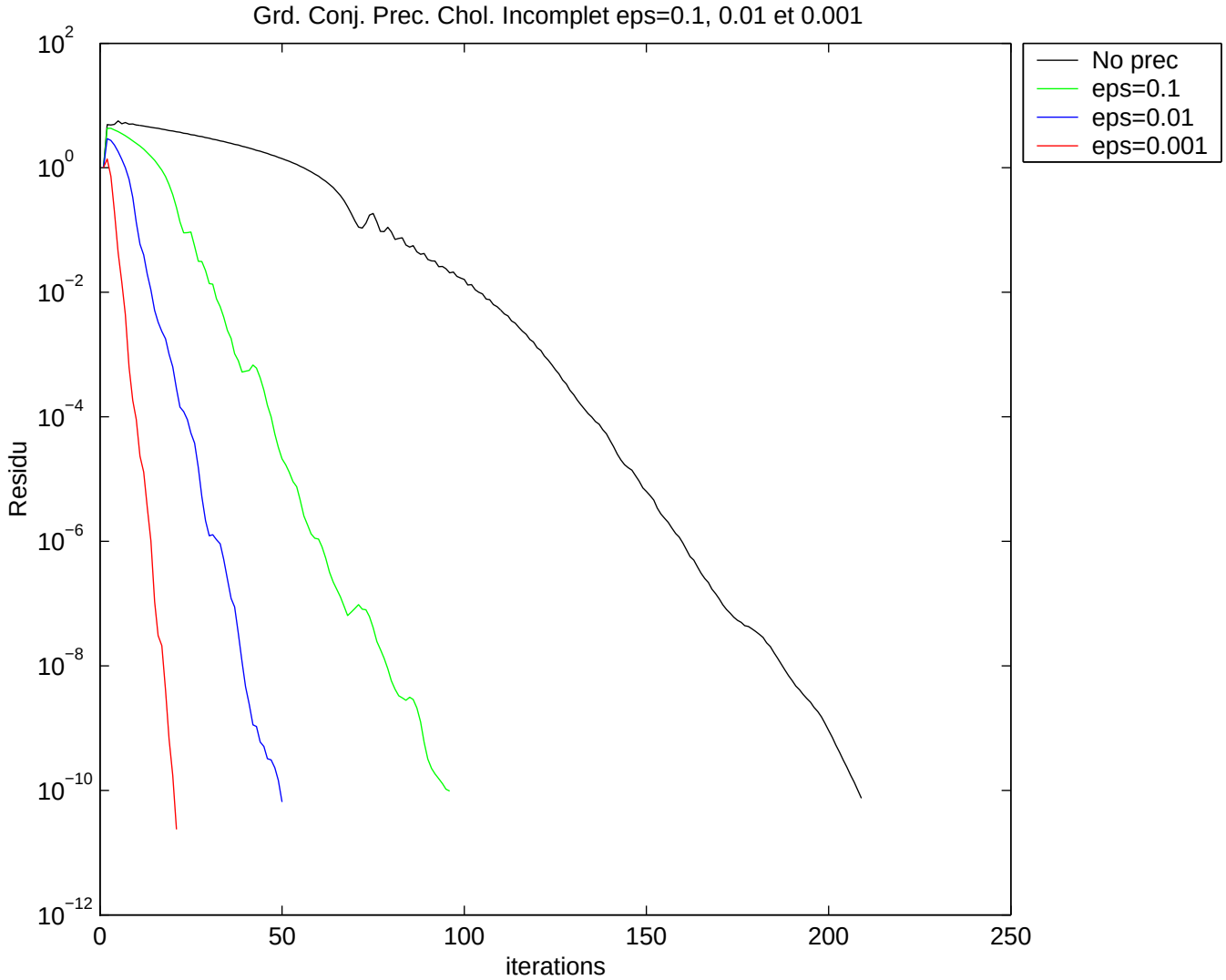


FIGURE: Gradient conjugué préconditionné par Cholesky incomplet pour plusieurs valeurs de seuillage (a)

$$-\Delta u + \gamma \left(e^{xy} \frac{\partial u}{\partial x} + e^{-xy} \frac{\partial u}{\partial y} \right) + \alpha u = f$$

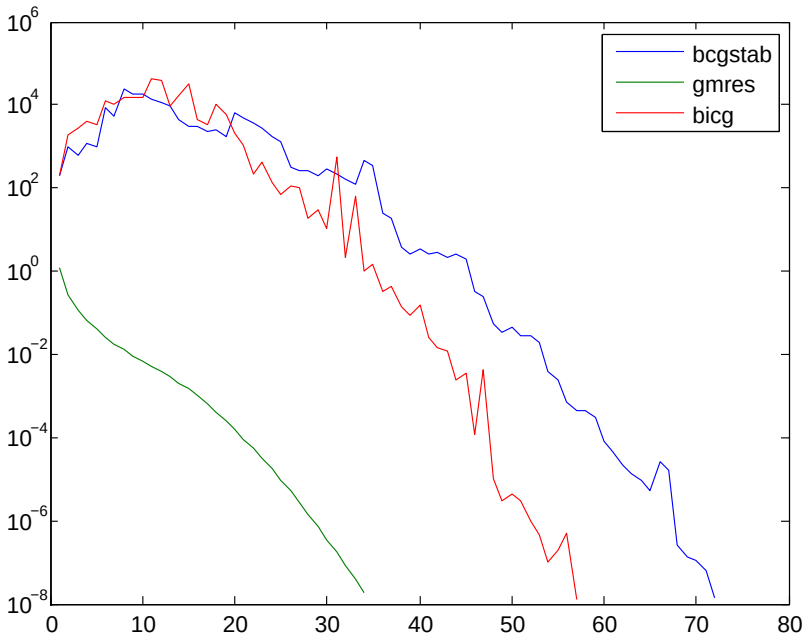
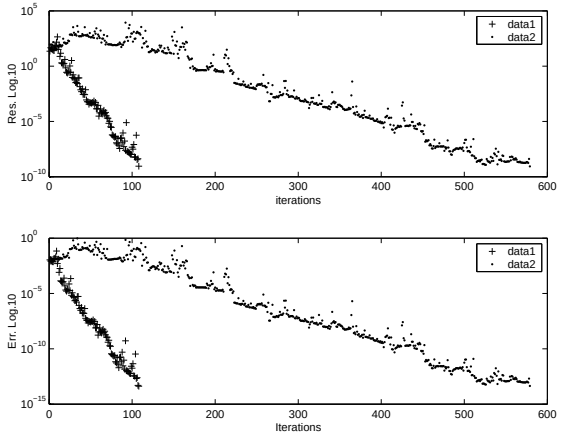
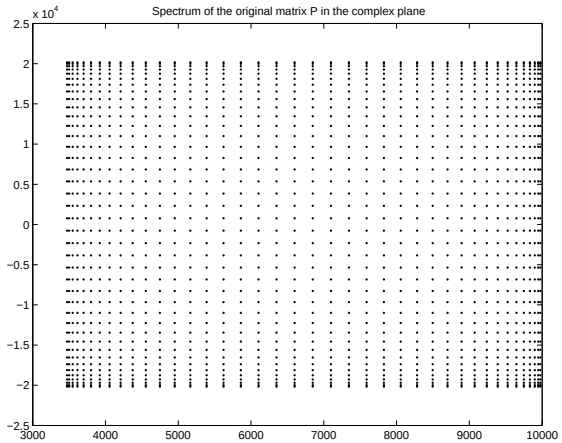


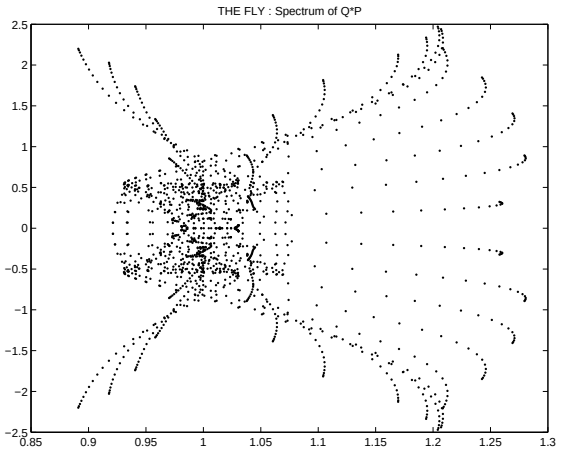
Figure: Comparaison gmres, bcgstab,bicg, n=200, système 40 000 $\times$ 40 000 péconditionné par LU incomplet seuil=0.1



(a)



(b)



(c)

FIGURE: Precond. Discr. matrix $-\Delta + 500\partial_x + 20\partial_y$ +HDBC

Précond. ARMS (Saad - Suchomel 2001)

Factorisation LU et compléments de Schur

$$\begin{pmatrix} B & F \\ E & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}$$

se réécrit comme

$$\begin{pmatrix} L & 0 \\ EU^{-1} & L \end{pmatrix} \times \begin{pmatrix} U & L^{-1}F \\ 0 & S \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \begin{pmatrix} f \\ g \end{pmatrix}$$

où $S = C_E B^{-1} F$ est le complément de Schur et $B = LU$

Idée

1. Effectuer cette décomposition récursivement sur S
2. Simplifier en prenant L et U appochées (creuses)
3. Approcher S par une matrice creuse et recommencer.

Algebraic Recursive Multilevel Solver (ARMS)

Algorithme

$$A_k = \begin{pmatrix} B_k & F_k \\ E_k & C_k \end{pmatrix} \simeq \begin{pmatrix} L_k & 0 \\ E_k U_k^{-1} & I \end{pmatrix} \times \begin{pmatrix} I & 0 \\ 0 & A_{k+1} \end{pmatrix} \times \begin{pmatrix} U_k & L_k^{-1} F_k \\ 0 & I \end{pmatrix}$$

Péconditionneurs très robustes pour des problèmes de grandes échelles (plusieurs dizaines de millions d'inconnues)

Solveurs rapides (dédiés à des problèmes précis)

Problème type

$$-\Delta u = f \text{ dans }]0,1[^2$$

- Multigrille $\mathcal{O}(N)$ (Southwell, 1935, deux grilles) Fedorenko 1972 (premier multigrille), popularisé par Brandt et Hackebush 70's Utilisation de plusieurs niveaux de discrétisation pour lisser les composantes hautes fréquences et se ramener à la résolution effective de petits systèmes.
- FFT $\mathcal{O}(N\ Log(N))$: permet de passer rapidement dans la base des vecteurs propres, de résoudre le système (diagonal) et de revenir dans la base initiale.

Remarque

Une méthode de descente très bien préconditionnée peut être compétitive (en nb d'itérations) avec le Multigrille.

Actuellement

- Augmentation de la capacité de calcul des machines permet d'aborder de nouveau problèmes
- Les méthodes directes OK pour les pbs en dimension 2 mais trop coûteux en dimension 3
- Les problèmes de plus en plus difficiles (très loins des exemples académiques, Poisson) et les méthodes itératives convergent lentement. Ex : pb de chimie quantique, de physique nucléaire (matrices denses, très mal conditionnées)
- Compromis : utiliser des méthodes directes (fact. incomplètes) pour accélérer la convergence des méthodes itératives : préconditionnement

Remarque

On dispose de méthodes robustes (GMRES) et aucun algorithme vraiment nouveau n'a été proposé depuis (il faut changer d'approche pour cela !). Le préconditionnement est devenu un des thèmes principaux.

1. M. Benzi, Preconditioning technique for large linear systems: a survey, J. Comput. Phys. 182 (2002), no. 2, 418–477
2. A. Cauchy [1847], Méthodes générales pour la résolution des systèmes d'équations simultanées, *C. R. Acad. Sci. Par.* 25, pp. 536–538.
3. M. R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, J. research Nat. Bur Standards, vol 49, 1952
4. A. Turing, Rounding-off errors in matrix processes, Quart. J. Mech. Appl. Math. 1, (1948). 287–308
5. Y. Saad, H. Van der Vorst, Iterative Solution of Linear systems in the 20-th Century, Numerical analysis 2000, Vol. III. Linear algebra. J. Comput. Appl. Math. 123 (2000), no. 1-2, 1–33