Quelle précision pour le raffinement itératif ?

Stef Graillat

LIP6 - Université Pierre et Marie Curie (Paris 6)

Rencontres arithmétique de l'informatique mathématique
22 janvier 2006, Montpellier

Introduction

Raffinement itératif : améliorer la précision d'une solution approchée $\widehat{x}$ du système $Ax = b$.

1. Calculer le résidu $r = b - A\widehat{x}$.
2. Résoudre $Ad = r$.
3. Mettre à jour $y = \widehat{x} + d$.
   (Recommencer à l'étape 1 sur nécessaire en remplaçant $\widehat{x}$ par y.)

$\longrightarrow$ méthode de Newton pour l'équation $F(x) = 0$ avec $F(x) = Ax - b$

Stabilité des méthodes directes

Systèmes linéaire $Ax = b$ avec $A \in \mathbb{R}^{n \times n}$ inversible

On suppose que le solveur produit une solution approchée $\widehat{x}$ vérifiant

$$(A + \Delta A)\widehat{x} = b, \quad |\Delta A| \leq \mathbf{u} W,$$

où W est une matrice positive dépendante de A, n et $\mathbf{u}$ (mais pas de b)

C'est le cas pour l'élimination gaussienne avec pivotage partiel (GEPP) :

- matrice quelconque : $\mathbf{u}W = \gamma_{3n}|\widehat{L}||\widehat{U}|$ où $\widehat{L}$ et $\widehat{U}$ sont le résultat calculé de la décomposition LU de A
- matrice triangulaire : $\mathbf{u}W = \gamma_n|A|$

$$\gamma_n = \frac{n\mathbf{u}}{1 - n\mathbf{u}} \text{ avec } \mathbf{u} \text{ unité d'arrondi}$$

Conditionnements

On utilise les conditionnements suivants :

- conditionnement du système linéaire

$$\mathrm{cond}(A, x) := \frac{\||A^{-1}||A||x|\|_\infty}{\|x\|_\infty}$$

- conditionnement « de la matrice »

$$\mathrm{cond}(A) := \||A^{-1}||A|\|_\infty$$

- conditionnement de l'inversion

$$\kappa_\infty = \|A\|_\infty \|A^{-1}\|$$

On a toujours

$$\mathrm{cond}(A, x) \leq \mathrm{cond}(A) \leq \kappa_\infty.$$

Historique

- Wilkinson 1948 : premier programme de raffinement itératif
- Wilkinson 1963 : analyse pour une arithmétique en virgule fixe
- Moller 1967 : analyse pour une arithmétique en virgule flottante
- Jankowski et Woźniakwoski 1977 : stabilité inverse du raffinement itératif
- Rump et Böhm 1983 : évaluation polynomiale précise par raffinement itératif
- Higham 1997 : raffinement itératif avec calcul du résidu en précision doublée
- Tisseur 2001 : algorithme de Newton en précision finie

Précision du raffinement itératif avec précision fixée

1. Calculer le résidu $r = b - A\widehat{x}$.
2. Résoudre $Ad = r$.
3. Mettre à jour $y = \widehat{x} + d$.
   (Recommencer à l'étape 1 sur nécessaire en remplaçant $\widehat{x}$ par y.)

Théorème 1 (Higham)

Appliquons l'algorithme de raffinement itératif au système linéaire $Ax = b$ d'ordre n, en utilisant un solveur stable. Notons $\eta = 2\mathbf{u}\,\mathrm{cond}(A)$. Alors, en supposant que η est suffisamment plus petit que 1, *le raffinement itératif réduit l'erreur directe d'un facteur proche de η à chaque étape, jusqu'à ce que $\|x - \widehat{x}_i\|_\infty / \|x\|_\infty \lesssim 2n\,\mathrm{cond}(A, x)\mathbf{u}$.*

$\longrightarrow$ stabilité inverse du raffinement itératif

Précision du raffinement itératif avec précision mixte

1. Calculer le résidu $r = b - A\widehat{x}$ en précision doublée.
2. Résoudre $Ad = r$.
3. Mettre à jour $y = \widehat{x} + d$.
   (Recommencer à l'étape 1 sur nécessaire en remplaçant $\widehat{x}$ par y.)

Théorème 2 (Higham)

Appliquons l'algorithme de raffinement itératif au système linéaire $Ax = b$ d'ordre n, en utilisant un solveur stable et en calculant le résidu avec une précision doublée. Notons $\eta = 2\mathbf{u}\,\mathrm{cond}(A)$. alors, si η est suffisamment plus petit que 1, le raffinement itératif réduit l'erreur directe d'un facteur proche de η à chaque étape, jusqu'à ce que $\|x - \widehat{x}_i\|_\infty / \|x\|_\infty \approx \mathbf{u}$.

En fait, Higham montre que $\|x - \widehat{x}_i\|_\infty / \|x\|_\infty \lesssim \mathbf{u} + 2n\mathbf{u}\,\mathrm{cond}(A, x)$.
$\longrightarrow$ résultat aussi précis que si on l'avait calculé avec une précision doublée
$\longrightarrow$ stabilité directe du raffinement itératif

Un bon critère d'arrêt ?

2 critères d'arrêt classiques :

- basé sur la norme du résidu $\|r\|_\infty$
- basé sur l'erreur relative direct estimée par $\|x_{i+1} - x_i\|_\infty / \|x_{i+1}\|_\infty$

Bornes d'erreur certifiées $\Rightarrow$ évite le calcul par intervalles

Borne d'erreur absolue certifiée :

$$\text{Trouver } \kappa \in \mathbb{F} \text{ tel que } \|x - \widehat{x}_i\|_\infty \leq \kappa$$

Bornes d'erreur relatives certifiées :

$$\boxed{\frac{\|x - \widehat{x}_i\|_\infty}{\|\widehat{x}_i\|_\infty} \leq \mathsf{fl}\left((\kappa / \|\widehat{x}\|_\infty)/(1 - 2\mathbf{u})\right)}$$

$$\frac{\|x - \widehat{x}_i\|_\infty}{\|x\|_\infty} \leq \mathsf{fl}\left((\kappa/(\|\widehat{x}_i\|_\infty - \kappa))/(1 - 3\mathbf{u})\right)$$

Une étape de raffinement itératif = algorithme compensé

Algorithme 1 (Une étape de raffinement itératif)

function $\texttt{res} = \texttt{OneStepItRaf}(A, b)$

$\widehat{x} = \mathrm{fl}(A \backslash b)$

$\widehat{r} = \mathrm{fl}_e(b - A\widehat{x})$ % calculé en précision doublée

$\widehat{d} = \mathrm{fl}(A \backslash \widehat{r})$

$\texttt{res} = \mathrm{fl}(\widehat{x} + \widehat{d})$

Si $r = b - A\widehat{x}$ et $d = A^{-1}r$ alors $x = \widehat{x} + d$

$\longrightarrow$ transformation exacte entre le système (A, b) et $(\widehat{x}, (A, r))$.

$\longrightarrow$ on va approcher l'erreur $d = A^{-1}r$ par $A \backslash \widehat{r}$.

Borne d'erreur *a priori*

$$\|\mathtt{res}-x\|_\infty \leq \mathbf{u}\|x\|_\infty+(1+\mathbf{u})(\mathbf{u}^2\||A^{-1}|W|x|\|_\infty+2\gamma_{n+1}^2\||A^{-1}||A||x|\|_\infty + \mathbf{u}^2\||A^{-1}|^2W^2|x|\|_\infty)+\mathcal{O}(\mathbf{u}^3)$$

ou encore en relatif

$$\frac{\|\mathtt{res}-x\|_\infty}{\|x\|_\infty} \leq \mathbf{u}+(1+\mathbf{u})\times$$
$$(\mathbf{u}^2\frac{\||A^{-1}|W|x|\|_\infty}{\|x\|_\infty}+2\gamma_{n+1}^2\frac{\||A^{-1}||A||x|\|_\infty}{\|x\|_\infty}+\mathbf{u}^2\frac{\||A^{-1}|^2W^2|x|\|_\infty}{\|x\|_\infty})+\mathcal{O}(\mathbf{u}^3)$$

avec $\mathbf{u}W = \gamma_{3n}|\widehat{L}||\widehat{U}|$ où $\widehat{L}$ et $\widehat{U}$ sont le résultat calculé de la décomposition LU de A

Borne d'erreur *a priori* pour les systèmes triangulaires

Une étape de raffinement itératif pour les systèmes triangulaire = méthode CENA de compensation pour les systèmes triangulaires

Pour les système triangulaires, $\mathbf{u}W = \gamma_n|A|$

$$\frac{\|\mathtt{res} - x\|_\infty}{\|x\|_\infty} \leq \mathbf{u} + (1+\mathbf{u})(\mathbf{u}\gamma_n + 2\gamma_{n+1}^2 + \gamma_n^2\,\mathrm{cond}(A))\,\mathrm{cond}(A,x) + \mathcal{O}(\mathbf{u}^3)$$

$\longrightarrow$ résultat pas aussi précis que si on l'avait calculé avec une précision doublée à cause du $\mathrm{cond}(A)$

Borne d'erreur certifiée (1/3)

$$\|\widehat{x} - A^{-1}b\|_\infty \leq \frac{\|R(A\widehat{x} - b)\|_\infty}{1 - \|I - RA\|_\infty} \qquad \text{(Oishi, Rump)}$$

Algorithme 2 (Ogita, Rump and Oishi)

Étant donnés $A \in \mathbb{F}^{n\times n}$ et R un inverse approché, l'algorithme suivant calcule une borne supérieure α pour $\|RA - I\|_\infty$.

function $\alpha = \texttt{Alpha.Std}(A, R)$
  if $(3n+2)\mathbf{u} \geq 1$, `error`("verification failed"), end
  $\alpha_1 = \mathsf{fl}(\|RA - I\|_\infty)$
  if $\alpha_1 \geq 1$, `error`("verification failed"), end
  $\alpha_1 = \mathsf{fl}(\||R|(|A|e)\|_\infty)$ % $e = (1, \ldots, 1)^T$
  $\alpha = \mathsf{fl}((\alpha_1 + \widetilde{\gamma}_{3n+2}(\alpha_2 + 2))/(1 - 2\mathbf{u}))$

Borne d'erreur certifiée (2/3)

$$\|\widehat{x} - A^{-1}b\|_\infty \leq \frac{\|R(A\widehat{x} - b)\|_\infty}{1 - \|I - RA\|_\infty}$$

Algorithme 3 (Ogita, Rump and Oishi)

Étant donnés $A \in \mathbb{F}^{n\times n}$, $b \in \mathbb{F}^n$ et x une solution approchée de $Ax = b$ et R un inversé approché de A, l'algorithme suivant calcule une borne supérieure β pour $\|R(A\widetilde{x} - b)\|_\infty$.

function $\beta = \texttt{Beta.Std}(A, b, \widetilde{x}, R)$

$r_{\text{mid}} = \text{fl}(A\widetilde{x} - b)$

$r_{\text{rad}} = \text{fl}(\widetilde{\gamma}_{2n+4}(|A||x| + |b|))$

$q = \text{fl}(|R|(\widetilde{\gamma}_{n+1}|r_{\text{mid}}| + r_{\text{rad}})/(1 - (n+3)\mathbf{u}))$

$\beta = \text{fl}((\||Rr_{\text{mid}}| + q\|_\infty)/(1 - 2\mathbf{u}))$

Borne d'erreur certifiée (3/3)

$$\|\mathtt{res} - x\|_\infty \leq \mathsf{fl}\left(\left(\mathbf{u}\|\mathtt{res}\|_\infty + \frac{\beta + \|R\|_\infty \delta}{1-\alpha}\right)/(1-(n+4)\mathbf{u})\right)$$

Simulations numériques

système $Ax = b$ avec $A \in \mathbb{F}^{10\times 10}$ et b tel que $x = (1, \ldots, 1)$ soit solution

conditionnement	erreur relative	erreur relative avec raffinement
10^8	10^{-9}	0
10^{18}	10^{-1}	0
10^{30}	1	10^{-1}
10^{40}	1	1

Raffinement itératif précis

Pour $r \in \mathbb{R}$, $\mathbb{F} \ni f = \square(r)$ est un arrondi fidèle de r (c-à-d l'arrondi vers $+\infty$ ou vers $-\infty$)

Algorithme 4

function $\texttt{res}_N = \texttt{RafIt1}(A, b, N)$
  $\widehat{x}_1 = \mathsf{fl}(A \backslash b)$
  for $i = 2 : N$
    $\widehat{r}_i = \square(b - A(\sum_{k=1}^{i-1} \widehat{x}_k))$
    $\widehat{x}_i = \mathsf{fl}(A \backslash \widehat{r}_i)$
  $\texttt{res}_N = \square(\sum_{k=1}^{N} \widehat{x}_k)$

$\square$ est disponible via `accdot` de Rump, Ogita et Oishi

Raffinement itératif précis

Important : trouver un bon inverse approché R, $\|I - RA\|_\infty \ll 1$

Si $\mathbf{u}\||A^{-1}|W\|_\infty \leq 1/2$ alors $R = \mathtt{inv}(A)$ convient

Théorème 3

Soit l'algorithme `RafIt1` *appliqué au système linéaire régulier* $Ax = b$ *d'ordre* n, *en utilisant un solveur stable. Si* $\mathbf{u}\||A^{-1}|W\|_\infty \leq 1/2$ *et* $\mathbf{u}\,\mathrm{cond}(A) < 1/8$ *alors*

$$|x - \mathtt{res}_N| \leq 2\mathbf{u}|x| + (1 + 2\mathbf{u})G^{N-1}|\widehat{x}_1 - x|$$

avec $\|G\|_\infty < 1$. *Par conséquent,*

$$\frac{\|x - \mathtt{res}_N\|_\infty}{\|x\|_\infty} \leq 2\mathbf{u} + (1 + 2\mathbf{u})\|G\|_\infty^{N-1}\frac{\|\widehat{x}_1 - x\|_\infty}{\|x\|_\infty}$$

Raffinement itératif précis

Si $\texttt{inv}(A)$ ne convient pas, alors `AccInv` de Rump donne une expansion de A^{-1} comme somme de matrices flottantes

Algorithme 5

function $\texttt{res}_N = \texttt{RafIt2}(A, b, N)$

$R = \texttt{AccInv}(A)$ % $R = R_1 + \cdots + R_m$ avec $\|I - RA\|_\infty < 1/2$

$\widehat{x}_1 = \square(Rb)$

for $i = 2 : N$

$\widehat{r}_i = \square(b - A(\sum_{k=1}^{i-1} \widehat{x}_k))$

$\widehat{x}_i = \square(R\widehat{r}_i)$

$\texttt{res}_N = \square(\sum_{k=1}^{N} \widehat{x}_k)$

Raffinement itératif précis

Théorème 4

Soit l'algorithme `RafIt2` *appliqué au système linéaire régulier* $Ax = b$ *d'ordre* n. *Alors*

$$|x - \mathtt{res}_N| \leq 2\mathbf{u}|x| + (1 + 2\mathbf{u})G^{N-1}|\widehat{x}_1 - x|$$

avec $\|G\|_\infty < 1$. *Par conséquent,*

$$\frac{\|x - \mathtt{res}_N\|_\infty}{\|x\|_\infty} \leq 2\mathbf{u} + (1 + 2\mathbf{u})\|G\|_\infty^{N-1} \frac{\|\widehat{x}_1 - x\|_\infty}{\|x\|_\infty}$$

Borne d'erreur certifiée

$$\|\widehat{x} - A^{-1}b\|_\infty \leq \frac{\|R(A\widehat{x} - b)\|_\infty}{1 - \|I - RA\|_\infty}$$

ou

$$|\widehat{x} - A^{-1}b| \leq |R(A\widehat{x} - b)| + \frac{\|R(A\widehat{x} - b)\|_\infty}{1 - \|G\|_\infty} \qquad \text{(Yamamoto)}$$

avec $G = RA - I$, $\|G\|_\infty < 1$ et

$$\left(\sum_{j=1}^{n} |G_{1,j}|, \ldots, \sum_{j=1}^{n} |G_{n,j}| \right)^T \leq t$$

Conclusion et perspectives

Conclusion :

- raffinement itératif = méthode de Newton = algorithme compensé
- raffinement itératif précis $\longrightarrow$ « pseudo-expansion » du résultat

Perspectives :

- Simulations numériques pour la précision et les performances
- Étude du raffinement itératif précis (c-à-d méthode de Newton précise) pour le calcul de zéros de polynômes (univariés plus multivariés) et de valeurs propres (généralisées, etc.)