Preconditioned GMRES-based Iterative Refinement

for the Solution of Sparse, Ill-Conditioned Linear Systems

Erin Carson and Nicholas J. Higham

New York University University of Manchester

August 2, 2017

Preconditioning 2017, Vancouver, BC

Iterative Refinement for $Ax = b$

A is $n \times n$, nonsingular

Solve $Ax_0 = b$ by LU factorization

for $i = 0{:}\,\mathrm{maxit}$

$\quad r_i = b - Ax_i$

$\quad$ Solve $Ad_i = r_i$

$\quad x_{i+1} = x_i + d_i$

Iterative Refinement for $Ax = b$

A is $n \times n$, nonsingular

Solve $Ax_0 = b$ by LU factorization

for $i = 0{:}\,\mathrm{maxit}$

$$r_i = b - Ax_i$$

Solve $Ad_i = r_i$ via $d_i = U^{-1}(L^{-1}r_i)$

$$x_{i+1} = x_i + d_i$$

Notation/Setting

- Assume standard floating point arithmetic
  - u denotes unit roundoff
  - "Gamma notation": $\gamma_k = \frac{ku}{1-ku}$

- Condition numbers
  - $|A| = |(a_{ij})|$
  - $\kappa_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty$
  - $\mathrm{cond}(A, x) = \frac{\| \, |A^{-1}||A||x| \, \|_\infty}{\|x\|_\infty}$
  - $\mathrm{cond}(A) = \mathrm{cond}(A, e) = \| \, |A^{-1}||A| \, \|_\infty$
  - $1 \leq \mathrm{cond}(A, x) \leq \mathrm{cond}(A) \leq \kappa_\infty(A)$

Error Bounds ("Traditional" IR)

Solve $Ax_0 = b$ by LU factorization

for $i = 0{:}\,\mathrm{maxit}$

$$r_i = b - Ax_i$$

$$d_i = U^{-1}(L^{-1}r_i)$$

$$x_{i+1} = x_i + d_i$$

Error Bounds ("Traditional" IR)

Solve $Ax_0 = b$ by LU factorization	precision u
for $i = 0{:}\,\mathrm{maxit}$	
$r_i = b - Ax_i$	precision u^2
$d_i = U^{-1}(L^{-1}r_i)$	precision u
$x_{i+1} = x_i + d_i$	precision u

Error Bounds ("Traditional" IR)

Solve $Ax_0 = b$ by LU factorization	precision u
for $i = 0{:}\,\mathrm{maxit}$	
$r_i = b - Ax_i$	precision u^2
$d_i = U^{-1}(L^{-1}r_i)$	precision u
$x_{i+1} = x_i + d_i$	precision u

- Early analyses by Wilkinson (1963), Moler (1967)
- If $\kappa_\infty(A)u < 1$, then error contracts (at a rate depending on $\kappa_\infty(A)$) until

$$\frac{\|x - x_i\|_\infty}{\|x\|_\infty} \approx u$$

Information in $\widehat{L}\widehat{U} \approx A$

- Empirically observed by Rump (1990) that if $\widehat{L}$ and $\widehat{U}$ are computed LU factors of A from GEPP, then $\kappa\left(\widehat{U}^{-1}\widehat{L}^{-1}A\right) \approx 1 + \kappa(A)u$

Information in $\widehat{L}\widehat{U} \approx A$

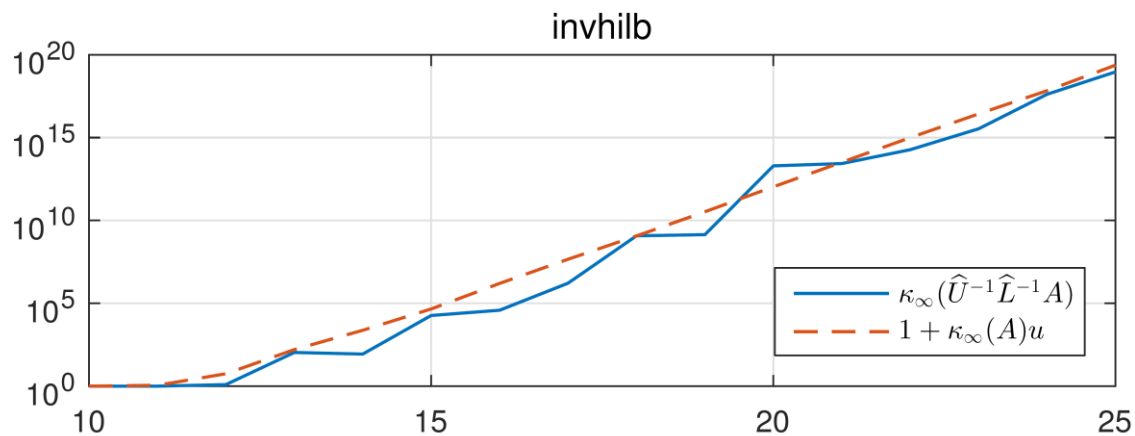
- Empirically observed by Rump (1990) that if $\widehat{L}$ and $\widehat{U}$ are computed LU factors of A from GEPP, then $\kappa\left(\widehat{U}^{-1}\widehat{L}^{-1}A\right) \approx 1 + \kappa(A)u$
  - Even if $\kappa(A) \gg u^{-1}$

Information in $\widehat{L}\widehat{U} \approx A$

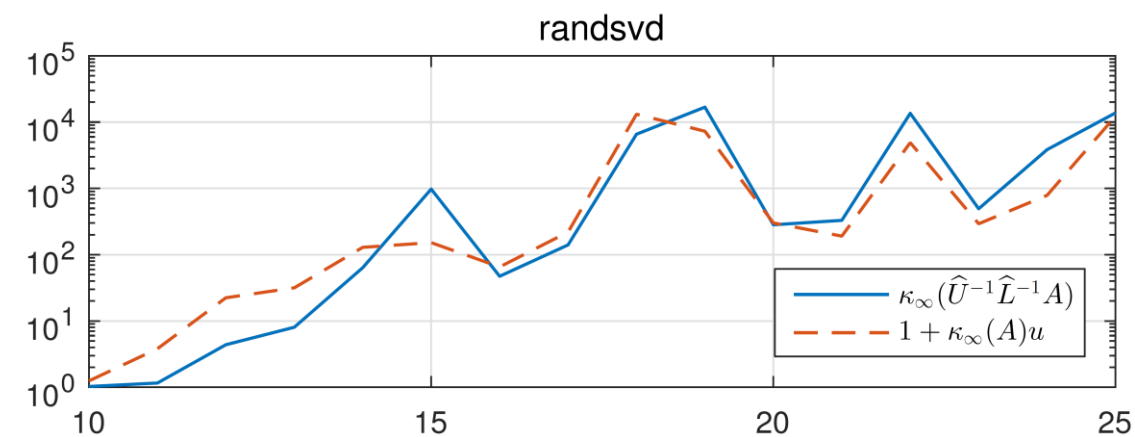
- Empirically observed by Rump (1990) that if $\widehat{L}$ and $\widehat{U}$ are computed LU factors of A from GEPP, then $\kappa\left(\widehat{U}^{-1}\widehat{L}^{-1}A\right) \approx 1 + \kappa(A)u$
  - Even if $\kappa(A) \gg u^{-1}$

Examples: ill-conditioned problems ($10^{13} \leq \kappa_\infty(A) \leq 10^{35}$), u = double

```
A = invhilb(n)
```




```
A =
gallery('randsvd',
n,10^(n+5))
```



New Analysis Summary

- New rounding error analysis of IR
- Identifies a mechanism by which iterative refinement can work when $\kappa_\infty(A) > u^{-1}$

New Analysis Summary

- New rounding error analysis of IR
- Identifies a mechanism by which iterative refinement can work when $\kappa_\infty(A) > u^{-1}$

- Requires that we can solve the equations for the updates d_i with some relative accuracy
  - Accomplished by using existing LU factors as preconditioners in GMRES method $\Longrightarrow$ GMRES-IR

New Analysis Summary

- New rounding error analysis of IR
- Identifies a mechanism by which iterative refinement can work when $\kappa_\infty(A) > u^{-1}$

- Requires that we can solve the equations for the updates d_i with some relative accuracy
  - Accomplished by using existing LU factors as preconditioners in GMRES method $\Longrightarrow$ GMRES-IR

- Even when $\kappa_\infty(A) \gtrsim u^{-1}$, GMRES-IR produces $\hat{x}$ for which

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} \approx u$$

New Analysis Summary

- New rounding error analysis of IR
- Identifies a mechanism by which iterative refinement can work when $\kappa_\infty(A) > u^{-1}$
- Requires that we can solve the equations for the updates d_i with some relative accuracy
  - Accomplished by using existing LU factors as preconditioners in GMRES method $\Longrightarrow$ GMRES-IR
- Even when $\kappa_\infty(A) \gtrsim u^{-1}$, GMRES-IR produces $\hat{x}$ for which

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} \approx u$$

- Need to define a few quantities...

The quantity θ_i

- Assume computed solution to $Ad_i = r_i$ satisfies

$$\frac{\left\|d_i - \hat{d}_i\right\|_\infty}{\|d_i\|_\infty} = \theta_i u$$

- θ_i depends on A, r_i, n, u, and the method of solving $Ad_i = r_i$

The quantity μ_i

- Traditional IR analyses use the bound: $\|A(x - \hat{x}_i)\|_\infty \leq \|A\|_\infty \|x - \hat{x}_i\|_\infty$

The quantity μ_i

- Traditional IR analyses use the bound: $\|A(x - \hat{x}_i)\|_\infty \leq \|A\|_\infty \|x - \hat{x}_i\|_\infty$
- Need a tighter bound; define

$$\|A(x - \hat{x}_i)\|_\infty = \mu_i \|A\|_\infty \|x - \hat{x}_i\|_\infty$$

- Note that $\kappa_\infty(A)^{-1} \leq \mu_i \leq 1$

The quantity μ_i

- Traditional IR analyses use the bound: $\|A(x - \hat{x}_i)\|_\infty \leq \|A\|_\infty \|x - \hat{x}_i\|_\infty$
- Need a tighter bound; define

$$\|A(x - \hat{x}_i)\|_\infty = \mu_i \|A\|_\infty \|x - \hat{x}_i\|_\infty$$

- Note that $\kappa_\infty(A)^{-1} \leq \mu_i \leq 1$

$$\mu_i \|A\|_\infty \|x - \hat{x}_i\|_\infty = \|A(x - \hat{x}_i)\|_\infty = \|b - A\hat{x}_i\|_\infty = \|r_i\|_\infty$$

The quantity μ_i

- Traditional IR analyses use the bound: $\|A(x - \hat{x}_i)\|_\infty \leq \|A\|_\infty \|x - \hat{x}_i\|_\infty$
- Need a tighter bound; define

$$\|A(x - \hat{x}_i)\|_\infty = \mu_i \|A\|_\infty \|x - \hat{x}_i\|_\infty$$

- Note that $\kappa_\infty(A)^{-1} \leq \mu_i \leq 1$

$$\mu_i \|A\|_\infty \|x - \hat{x}_i\|_\infty = \|A(x - \hat{x}_i)\|_\infty = \|b - A\hat{x}_i\|_\infty = \|r_i\|_\infty$$

- For a stable solver, in early stages we expect

$$\frac{\|r_i\|}{\|A\|\|\hat{x}_i\|} \approx u \ll \frac{\|x - \hat{x}_i\|}{\|x\|} \longrightarrow \boxed{\mu_i \ll 1}$$

The quantity μ_i

- Traditional IR analyses use the bound: $\|A(x-\hat{x}_i)\|_\infty \le \|A\|_\infty \|x-\hat{x}_i\|_\infty$
- Need a tighter bound; define

$$\|A(x-\hat{x}_i)\|_\infty = \mu_i \|A\|_\infty \|x-\hat{x}_i\|_\infty$$

- Note that $\kappa_\infty(A)^{-1} \le \mu_i \le 1$

$$\mu_i \|A\|_\infty \|x-\hat{x}_i\|_\infty = \|A(x-\hat{x}_i)\|_\infty = \|b-A\hat{x}_i\|_\infty = \|r_i\|_\infty$$

- For a stable solver, in early stages we expect

$$\frac{\|r_i\|}{\|A\|\|\hat{x}_i\|} \approx u \ll \frac{\|x-\hat{x}_i\|}{\|x\|} \longrightarrow \boxed{\mu_i \ll 1}$$

- But close to convergence,

$$\|r_i\| \approx \|A\|\|x-\hat{x}_i\| \longrightarrow \boxed{\mu_i \approx 1}$$

Theorem (C. & Higham, 2017)

Let IR in precisions u and u^2 be applied to a linear system $Ax = b$ with nonsingular $A \in \mathbb{R}^{n\times n}$ and a given approximate solution x_0. Assume that the solver for the corrective term d_i satisfies $\left\|d_i - \hat{d}_i\right\|_\infty / \|d_i\|_\infty = \theta_i u$. Then for $i \geq 0$, the computed iterate $\hat{x}_{i+1}$ satisfies

$$\begin{aligned}\|x - \hat{x}_{i+1}\|_\infty \leq\ & (2\mu_i \kappa_\infty(A) u + \theta_i u)\, \|x - \hat{x}_i\|_\infty \\ & + n u^2 (1 + \theta_i u) \| \, |A^{-1}| \, (|b| + |A||\hat{x}_i|) \|_\infty + u \|\hat{x}_{i+1}\|\end{aligned}$$

Theorem (C. & Higham, 2017)

Let IR in precisions u and u^2 be applied to a linear system $Ax = b$ with nonsingular $A \in \mathbb{R}^{n\times n}$ and a given approximate solution x_0. Assume that the solver for the corrective term d_i satisfies $\left\|d_i - \hat{d}_i\right\|_\infty / \|d_i\|_\infty = \theta_i u$. Then for $i \geq 0$, the computed iterate $\hat{x}_{i+1}$ satisfies

$$\|x - \hat{x}_{i+1}\|_\infty \leq (2\mu_i \kappa_\infty(A)u + \theta_i u)\,\|x - \hat{x}_i\|_\infty + nu^2(1+\theta_i u)\|\,|A^{-1}|\,(|b| + |A||\hat{x}_i|)\|_\infty + u\|\hat{x}_{i+1}\|$$

As long as for all i,

$$2\mu_i \kappa_\infty(A)u + \theta_i u < 1,$$

the error will contract until a limiting normwise relative error of order

$$2nu^2(1+\theta u)\mathrm{cond}(A,x) + u$$

is achieved, where θ is an upper bound on the θ_i terms.

Theorem (C. & Higham, 2017)

Let IR in precisions u and u^2 be applied to a linear system $Ax = b$ with nonsingular $A \in \mathbb{R}^{n\times n}$ and a given approximate solution x_0. Assume that the solver for the corrective term d_i satisfies $\left\|d_i - \hat{d}_i\right\|_\infty / \|d_i\|_\infty = \theta_i u$. Then for $i \geq 0$, the computed iterate $\hat{x}_{i+1}$ satisfies

$$\|x - \hat{x}_{i+1}\|_\infty \leq (2\mu_i \kappa_\infty(A)u + \theta_i u)\, \|x - \hat{x}_i\|_\infty$$
$$+ nu^2(1 + \theta_i u) \| \, |A^{-1}| \, (|b| + |A||\hat{x}_i|)\|_\infty + u\|\hat{x}_{i+1}\|$$

As long as for all i,

$$2\mu_i \kappa_\infty(A)u + \theta_i u < 1,$$

the error will contract until a limiting normwise relative error of order

$$2nu^2(1 + \theta u)\mathrm{cond}(A, x) + u$$

is achieved, where θ is an upper bound on the θ_i terms.

$\approx u$ if $\mathrm{cond}(A, x)u \lesssim 1$ (essentially indep. of θ as long as $\theta u < 1$)

Theorem (C. & Higham, 2017)

Let IR in precisions u and u^2 be applied to a linear system $Ax = b$ with nonsingular $A \in \mathbb{R}^{n\times n}$ and a given approximate solution x_0. Assume that the solver for the corrective term d_i satisfies $\|d_i - \hat{d}_i\|_\infty / \|d_i\|_\infty = \theta_i u$. Then for $i \geq 0$, the computed iterate $\hat{x}_{i+1}$ satisfies

$$\|x - \hat{x}_{i+1}\|_\infty \leq (2\mu_i \kappa_\infty(A)u + \theta_i u)\, \|x - \hat{x}_i\|_\infty + nu^2(1+\theta_i u)\| \, |A^{-1}| \, (|b| + |A||\hat{x}_i|)\|_\infty + u\|\hat{x}_{i+1}\|$$

$\mu_i \kappa_\infty(A)u < 1$ (condition on iteration)

$\theta_i u < 1$ (condition on solver, data)

As long as for all i,

$$2\mu_i \kappa_\infty(A)u + \theta_i u < 1,$$

the error will contract until a limiting normwise relative error of order

$$2nu^2(1+\theta u)\mathrm{cond}(A,x) + u$$

$\approx u$ if $\mathrm{cond}(A,x)u \lesssim 1$ (essentially indep. of θ as long as $\theta u < 1$)

is achieved, where θ is an upper bound on the θ_i terms.

Standard (LU-based) iterative refinement

- If $\kappa_\infty(A) > u^{-1}$, $\theta_i u < 1$ can not be guaranteed no matter how precision is used in the substitutions

Standard (LU-based) iterative refinement

- If $\kappa_\infty(A) > u^{-1}$, $\theta_i u < 1$ can not be guaranteed no matter how precision is used in the substitutions

- Assume that the solve $\hat{d}_i = \widehat{U}^{-1}\widehat{L}^{-1}\,\hat{r}_i$ is carried out exactly:

$$A + \Delta A = \widehat{L}\widehat{U}, \qquad |\Delta A| \le \gamma_n |\widehat{L}||\widehat{U}|$$

$$\hat{d}_i = \widehat{U}^{-1}\widehat{L}^{-1}\,\hat{r}_i = (A + \Delta A)^{-1}\,\hat{r}_i$$

$$\theta_i u = \frac{\left\|\hat{d}_i - d_i\right\|_\infty}{\|d_i\|_\infty} \approx \frac{\|A^{-1}\Delta A d_i\|_\infty}{\|d_i\|_\infty} \le \gamma_n \left\| |A^{-1}| \left|\widehat{L}\right| \left|\widehat{U}\right| \right\|_\infty$$

Standard (LU-based) iterative refinement

- If $\kappa_\infty(A) > u^{-1}$, $\theta_i u < 1$ can not be guaranteed no matter how precision is used in the substitutions

- Assume that the solve $\hat{d}_i = \widehat{U}^{-1}\widehat{L}^{-1}\,\hat{r}_i$ is carried out exactly:

$$A + \Delta A = \widehat{L}\widehat{U}, \qquad |\Delta A| \le \gamma_n |\widehat{L}||\widehat{U}|$$

$$\hat{d}_i = \widehat{U}^{-1}\widehat{L}^{-1}\,\hat{r}_i = (A + \Delta A)^{-1}\,\hat{r}_i$$

$$\theta_i u = \frac{\|\hat{d}_i - d_i\|_\infty}{\|d_i\|_\infty} \approx \frac{\|A^{-1}\Delta A d_i\|_\infty}{\|d_i\|_\infty} \le \gamma_n \underbrace{\left\| |A^{-1}||\widehat{L}||\widehat{U}| \right\|_\infty}_{\text{at least as large as } \mathrm{cond}(A),\ \text{usually similar size to } \kappa_\infty(A)}$$

GMRES-based iterative refinement

- To compute the updates d_i, apply GMRES to

$$\underbrace{\widehat{U}^{-1}\widehat{L}^{-1}A}_{\tilde{A}}d_i = \underbrace{\widehat{U}^{-1}\widehat{L}^{-1}r_i}_{\tilde{r}_i}$$

GMRES-based iterative refinement

- To compute the updates d_i, apply GMRES to

$$\underbrace{\widehat{U}^{-1}\widehat{L}^{-1}A}_{\tilde{A}}d_i = \underbrace{\widehat{U}^{-1}\widehat{L}^{-1}r_i}_{\tilde{r}_i}$$

Standard IR:

Solve $Ax_0 = b$ by LU factorization

for $i = 0{:}\,\mathrm{maxit}$

$r_i = b - Ax_i$

Solve $Ad_i = r_i$ via $d_i = U^{-1}(L^{-1}r_i)$

$x_{i+1} = x_i + d_i$

GMRES-based iterative refinement

- To compute the updates d_i, apply GMRES to

$$\underbrace{\hat{U}^{-1}\hat{L}^{-1}A}_{\tilde{A}}d_i = \underbrace{\hat{U}^{-1}\hat{L}^{-1}r_i}_{\tilde{r}_i}$$

GMRES-IR:

Solve $Ax_0 = b$ by LU factorization

for $i = 0{:}\,\mathrm{maxit}$

$r_i = b - Ax_i$

Solve $Ad_i = r_i$ via GMRES on $\tilde{A}d_i = \tilde{r}_i$

$x_{i+1} = x_i + d_i$

Extending GMRES backward stability results

- Backward error results for GMRES of Paige, Rozložník, Strakoš (2006) can be extended to the left-preconditioned case

Extending GMRES backward stability results

- Backward error results for GMRES of Paige, Rozložník, Strakoš (2006) can be extended to the left-preconditioned case
- As long as within GMRES, $\tilde{A}$ (not explicitly formed) is applied to a vector with sufficient accuracy,

see (C. & Higham, 2017)

$$\frac{\|d_i - \hat{d}_i\|_\infty}{\|d_i\|_\infty} = \theta_i u \lesssim \gamma_n \kappa_\infty(\tilde{A})$$

Extending GMRES backward stability results

- Backward error results for GMRES of Paige, Rozložník, Strakoš (2006) can be extended to the left-preconditioned case
- As long as within GMRES, $\tilde{A}$ (not explicitly formed) is applied to a vector with sufficient accuracy,

see (C. & Higham, 2017)

$$\frac{\|d_i - \hat{d}_i\|_\infty}{\|d_i\|_\infty} = \theta_i u \lesssim \gamma_n \kappa_\infty(\tilde{A})$$

$$\kappa_\infty(\tilde{A}) \leq \left(1 + \gamma_n \left\| |A^{-1}| |\hat{L}| |\hat{U}| \right\|_\infty\right)^2 \ll \kappa_\infty(A)$$

Extending GMRES backward stability results

- Backward error results for GMRES of Paige, Rozložník, Strakoš (2006) can be extended to the left-preconditioned case
- As long as within GMRES, $\tilde{A}$ (not explicitly formed) is applied to a vector with sufficient accuracy,

see (C. & Higham, 2017)

$$\frac{\|d_i - \hat{d}_i\|_\infty}{\|d_i\|_\infty} = \theta_i u \lesssim \gamma_n \kappa_\infty(\tilde{A})$$

$$\kappa_\infty(\tilde{A}) \leq \left(1 + \gamma_n \left\| |A^{-1}| |\hat{L}| |\hat{U}| \right\|_\infty\right)^2 \ll \kappa_\infty(A)$$

(usually $\kappa_\infty(\tilde{A}) \approx 1 + \kappa_\infty(A)u$)

Extending GMRES backward stability results

- Backward error results for GMRES of Paige, Rozložník, Strakoš (2006) can be extended to the left-preconditioned case
- As long as within GMRES, $\tilde{A}$ (not explicitly formed) is applied to a vector with sufficient accuracy,

see (C. & Higham, 2017)

$$\frac{\|d_i - \hat{d}_i\|_\infty}{\|d_i\|_\infty} = \theta_i u \lesssim \gamma_n \kappa_\infty(\tilde{A})$$

$$\kappa_\infty(\tilde{A}) \leq \left(1 + \gamma_n \left\| |A^{-1}| |\hat{L}| |\hat{U}| \right\|_\infty\right)^2 \ll \kappa_\infty(A)$$

(usually $\kappa_\infty(\tilde{A}) \approx 1 + \kappa_\infty(A)u$)

$\Longrightarrow$ Even if $\kappa_\infty(A) > u^{-1}$, $\theta_i u < 1$

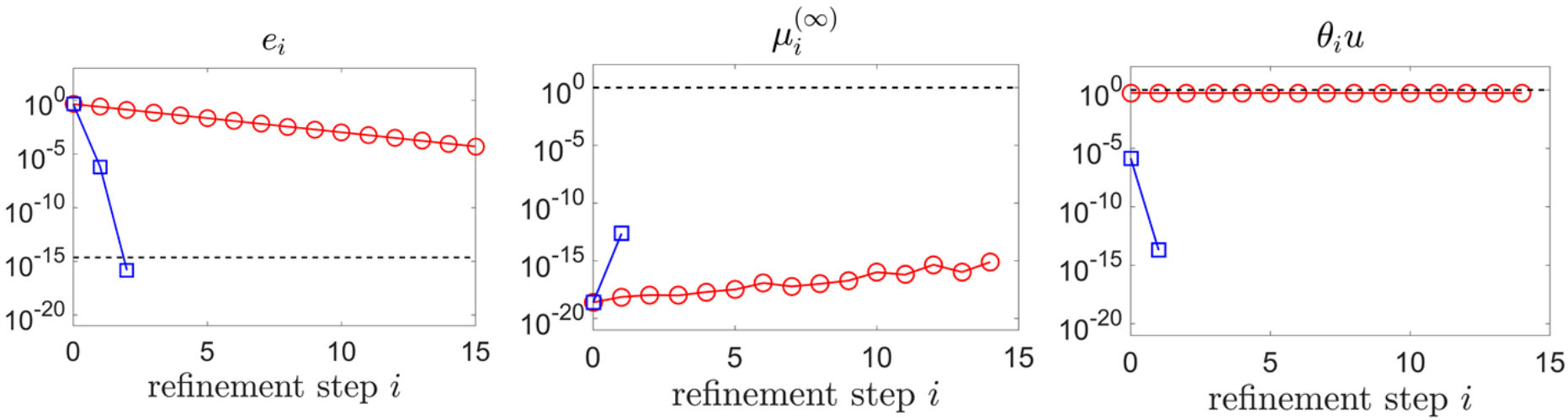
Numerical experiments

$u = 2^{-53}$ (double), $u^2 = 2^{-113}$ (quad)

UFSMC matrix: oscil_dcop_06, $n = 430$

$\mathrm{cond}(A) = 2 \cdot 10^{18}$, $\kappa_\infty(A) = 1 \cdot 10^{21}$, $\kappa(\tilde{A}) = 45$

$b = \mathrm{randn}(n, 1)$



$e_i = \|x - \hat{x}_i\|_\infty / \|x\|_\infty$ $\qquad$ $\mu_i^{(\infty)} = \frac{\|A(x-\hat{x}_i)\|_\infty}{\|A\|_\infty \|x-\hat{x}_i\|_\infty}$ $\qquad$ $\theta_i u = \left\|d_i - \hat{d}_i\right\|_\infty / \|d_i\|_\infty$

Standard IR steps	GMRES-IR steps	GMRES its.
–	2	7 (3,4)

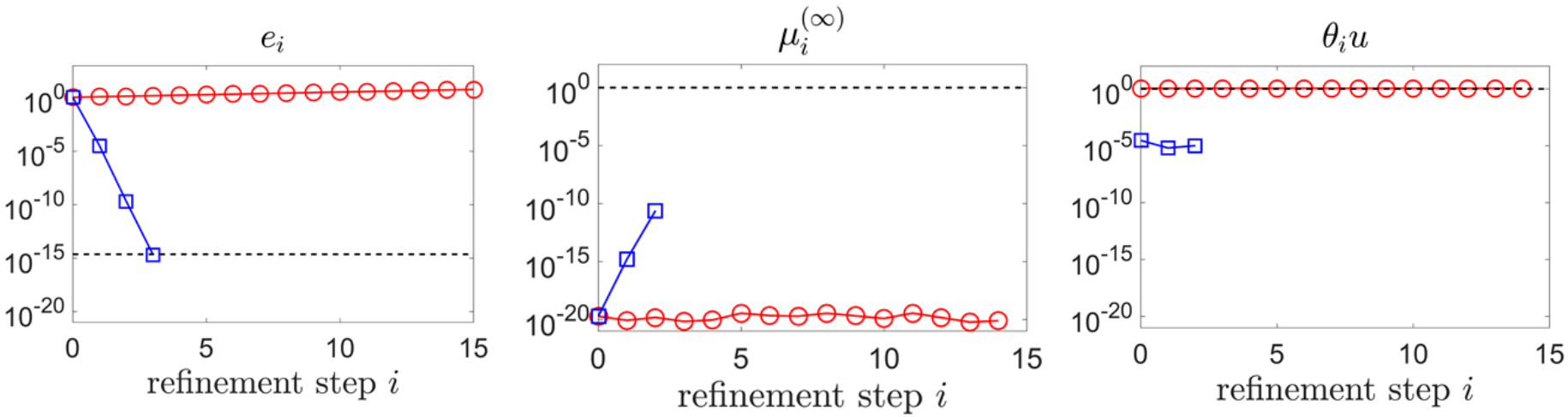
Numerical experiments

$u = 2^{-53}$ (double), $u^2 = 2^{-113}$ (quad)

UFSMC matrix: oscil_dcop_43, $n = 430$

$\mathrm{cond}(A) = 1 \cdot 10^{18}$, $\kappa_\infty(A) = 8 \cdot 10^{20}$, $\kappa(\tilde{A}) = 2.1$

$b = \mathrm{randn}(n, 1)$



$e_i = \|x - \hat{x}_i\|_\infty / \|x\|_\infty$ $\qquad$ $\mu_i^{(\infty)} = \frac{\|A(x-\hat{x}_i)\|_\infty}{\|A\|_\infty \|x-\hat{x}_i\|_\infty}$ $\qquad$ $\theta_i u = \left\|d_i - \hat{d}_i\right\|_\infty / \|d_i\|_\infty$

Standard IR steps	GMRES-IR steps	GMRES its.
–	3	10 (2,4,4)

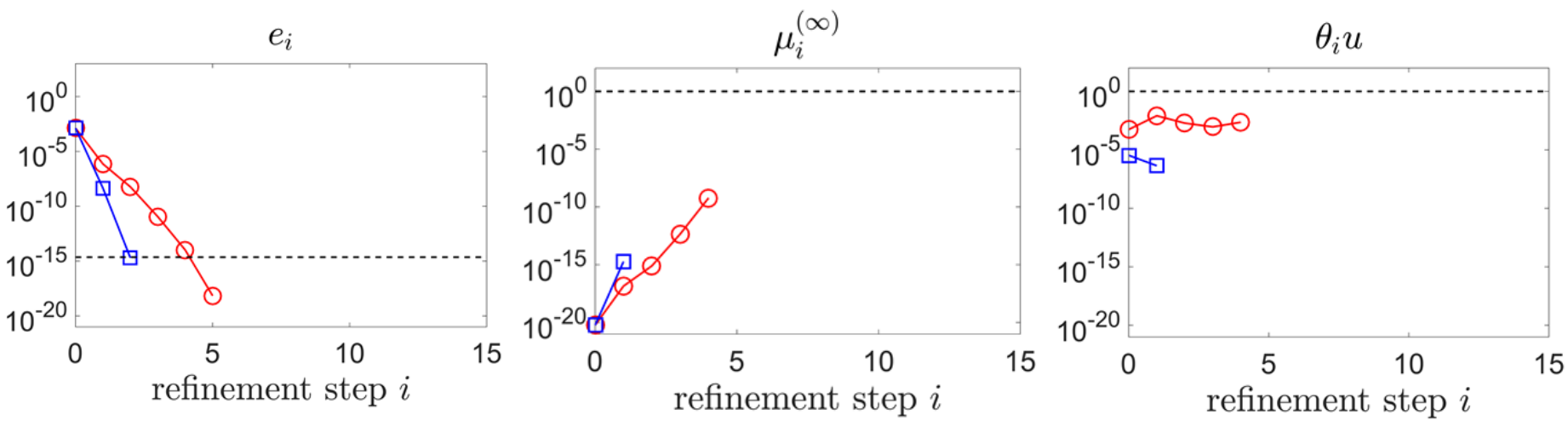
Numerical experiments

$u = 2^{-53}$ (double), $u^2 = 2^{-113}$ (quad)

UFSMC matrix: mhda416, $n = 416$

$\mathrm{cond}(A) = 1 \cdot 10^{19}$, $\kappa_\infty(A) = 2 \cdot 10^{25}$, $\kappa(\tilde{A}) = 7 \cdot 10^{9}$

$b = \mathrm{randn}(n, 1)$



$e_i = \|x - \hat{x}_i\|_\infty / \|x\|_\infty$

$\mu_i^{(\infty)} = \frac{\|A(x-\hat{x}_i)\|_\infty}{\|A\|_\infty \|x-\hat{x}_i\|_\infty}$

$\theta_i u = \left\|d_i - \hat{d}_i\right\|_\infty / \|d_i\|_\infty$

Standard IR steps	GMRES-IR steps	GMRES its.
5	2	3 (1,2)

Two-Stage IR

- Sometimes standard (LU-based) IR converges despite $\kappa_\infty(A) > u^{-1}$
  - Cheaper than GMRES-IR per refinement step
  - But hard to predict

Two-Stage IR

- Sometimes standard (LU-based) IR converges despite $\kappa_\infty(A) > u^{-1}$
  - Cheaper than GMRES-IR per refinement step
  - But hard to predict

- Two-Stage IR
  - Solve $Ax_0 = b$ by LU factorization
  - Attempt standard IR
  - If convergence is slow, or divergence, switch to GMRES-IR (making use of existing LU factorization)

Two-Stage IR

- Sometimes standard (LU-based) IR converges despite $\kappa_\infty(A) > u^{-1}$
  - Cheaper than GMRES-IR per refinement step
  - But hard to predict

- Two-Stage IR
  - Solve $Ax_0 = b$ by LU factorization
  - Attempt standard IR
  - If convergence is slow, or divergence, switch to GMRES-IR (making use of existing LU factorization)

- Decision to switch can be based on, e.g., stopping criteria for forward error of Demmel et al. (2006)

- Future work...

Extensions

- Pivoting
  - common to use pivoting strategy to minimize fill
  - static pivoting, threshold pivoting

Extensions

- Pivoting
  - common to use pivoting strategy to minimize fill
  - static pivoting, threshold pivoting
- Incomplete LU factorizations
  - As long as $\kappa_\infty(\tilde{A})u < 1$, $\theta_i u < 1$, so expect refinement process to converge

Extensions

- Pivoting
  - common to use pivoting strategy to minimize fill
  - static pivoting, threshold pivoting
- Incomplete LU factorizations
  - As long as $\kappa_\infty(\tilde{A})u < 1$, $\theta_i u < 1$, so expect refinement process to converge
- Other solvers
  - Left-preconditioned, unrestarted GMRES used here for theoretical purposes
  - In practice, many potential modifications may improve performance while still resulting in IR convergence
    - Restarted GMRES
    - Right, split preconditioned GMRES, FGMRES
    - Other Krylov subspace methods (not necessarily backward stable)

Extensions II: Iterative refinement in 3 precisions

- Emerging architectures feature built-in support for multiprecision computation, rising interest in low-precision storage and computation (performance and energy savings!)

Extensions II: Iterative refinement in 3 precisions

- Emerging architectures feature built-in support for multiprecision computation, rising interest in low-precision storage and computation (performance and energy savings!)
  - Half precision (FP16) defined as storage format in 2008 IEEE standard
  - Intel Ivy bridge, 2012: supports half precision for storage
  - NVIDIA Tesla P100, 2016: native hardware ISA support for 16-bit FP arithmetic
  - TSUBAME3.0 supercomputer, 2017: projected 12.2 double-precision petaflops, 64.3 half-precision petaflops
  - Intel Xeon Phi (Knights Mill), 2017: will support 16-bit FP
  - Google Tensorflow processor (TPU): quantizes 32-bit FP computations into 8-bit arithmetic

Extensions II: Iterative refinement in 3 precisions

- Emerging architectures feature built-in support for multiprecision computation, rising interest in low-precision storage and computation (performance and energy savings!)
  - Half precision (FP16) defined as storage format in 2008 IEEE standard
  - Intel Ivy bridge, 2012: supports half precision for storage
  - NVIDIA Tesla P100, 2016: native hardware ISA support for 16-bit FP arithmetic
  - TSUBAME3.0 supercomputer, 2017: projected 12.2 double-precision petaflops, 64.3 half-precision petaflops
  - Intel Xeon Phi (Knights Mill), 2017: will support 16-bit FP
  - Google Tensorflow processor (TPU): quantizes 32-bit FP computations into 8-bit arithmetic

- Can we use lower precision in the most expensive part of solving $Ax = b$ using IR (the LU factorization) and still obtain accurate solutions?

Extensions II: Iterative refinement in 3 precisions

- Emerging architectures feature built-in support for multiprecision computation, rising interest in low-precision storage and computation (performance and energy savings!)
  - Half precision (FP16) defined as storage format in 2008 IEEE standard
  - Intel Ivy bridge, 2012: supports half precision for storage
  - NVIDIA Tesla P100, 2016: native hardware ISA support for 16-bit FP arithmetic
  - TSUBAME3.0 supercomputer, 2017: projected 12.2 double-precision petaflops, 64.3 half-precision petaflops
  - Intel Xeon Phi (Knights Mill), 2017: will support 16-bit FP
  - Google Tensorflow processor (TPU): quantizes 32-bit FP computations into 8-bit arithmetic

- Can we use lower precision in the most expensive part of solving $Ax = b$ using IR (the LU factorization) and still obtain accurate solutions?
- Three precisions:

u_f = factorization precision, u = working precision, u_r = residual precision

$$u_f \geq u \geq u_r$$

Extensions II: Iterative refinement in 3 precisions

- Existing analyses:
  - **Wilkinson** (1963): fixed-point arithmetic.
  - **Moler** (1967): floating-point arithmetic.
  - **Higham** (1997, 2002): more general analysis for arbitrary solver.
  - **Langou et al.** (2006): lower precision LU.
- All the above support **at most two precisions** and require $\kappa_\infty(A)u < 1$.

Extensions II: Iterative refinement in 3 precisions

- Existing analyses:
  - **Wilkinson** (1963): fixed-point arithmetic.
  - **Moler** (1967): floating-point arithmetic.
  - **Higham** (1997, 2002): more general analysis for arbitrary solver.
  - **Langou et al.** (2006): lower precision LU.
- All the above support **at most two precisions** and require $\kappa_\infty(A)u < 1$.

New analysis generalizes and extends existing types of IR: (u_f, u, u_r)

Traditional Wilkinson (1948)

SSD DDQ HHS	
HHD HHQ SSQ	

70s/80s Skeel (1980) Jankowski and Wozniakowski (1977)

SSS DDD	
HHH	

2000s Dongarra, Langou et al. (2006)

SDD HSS DQQ	
HDD HQQ SQQ	

New

HSD HSQ HDQ SDQ	

Extensions II: Iterative refinement in 3 precisions

- Three precisions:
  - u_f: factorization precision
  - u: working precision
  - u_r: residual computation precision

Theorem (C. & Higham, 2017)

For IR in precisions $\boldsymbol{u_f} \geq \boldsymbol{u} \geq \boldsymbol{u_r}$, if

$$\phi_i = 2\boldsymbol{u_f} \min(\mathrm{cond}(A), \kappa_\infty(A)\mu_i) + \boldsymbol{u_f}\theta_i$$

is sufficiently less than 1, then the forward error is reduced on the ith iteration by a factor $\approx \phi_i$ until an iterate $\hat{x}$ is produced for which

$$\frac{\|x - \hat{x}\|_\infty}{\|x\|_\infty} \lesssim 4n\boldsymbol{u_r}\mathrm{cond}(A, x) + \boldsymbol{u}.$$

- Analogous standard bounds would have $\mu_i = 1$, $u_f\theta_i = \kappa_\infty(A)u$

Extensions II: Iterative refinement in 3 precisions

Standard (LU-based) IR in three precisions:

				Backward error		
u_f	u	u_r	$\kappa_\infty(A)$	norm	comp	Forward error
H	S	S	10^4	S	S	$\mathrm{cond}(A,x)\cdot 10^{-8}$
H	S	D	10^4	S	S	S
H	D	D	10^4	D	D	$\mathrm{cond}(A,x)\cdot 10^{-16}$
H	D	Q	10^4	D	D	D
S	S	S	10^8	S	S	$\mathrm{cond}(A,x)\cdot 10^{-8}$
S	S	D	10^8	S	S	S
S	D	D	10^8	D	D	$\mathrm{cond}(A,x)\cdot 10^{-16}$
S	D	Q	10^8	D	D	D

Extensions II: Iterative refinement in 3 precisions

Standard (LU-based) IR in three precisions:

u_f	u	u_r	$\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
H	S	S	10^4	S	S	$\mathrm{cond}(A,x)\cdot 10^{-8}$
H	S	D	10^4	S	S	S
H	D	D	10^4	D	D	$\mathrm{cond}(A,x)\cdot 10^{-16}$
H	D	Q	10^4	D	D	D
S	S	S	10^8	S	S	$\mathrm{cond}(A,x)\cdot 10^{-8}$
S	S	D	10^8	S	S	S
S	D	D	10^8	D	D	$\mathrm{cond}(A,x)\cdot 10^{-16}$
S	D	Q	10^8	D	D	D

Extensions II: Iterative refinement in 3 precisions

Standard (LU-based) IR in three precisions:

u_f	u	u_r	$\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
H	S	S	10^4	S	S	$\mathrm{cond}(A,x) \cdot 10^{-8}$
H	S	D	10^4	S	S	S
H	D	D	10^4	D	D	$\mathrm{cond}(A,x) \cdot 10^{-16}$
H	D	Q	10^4	D	D	D
S	S	S	10^8	S	S	$\mathrm{cond}(A,x) \cdot 10^{-8}$
S	S	D	10^8	S	S	S
S	D	D	10^8	D	D	$\mathrm{cond}(A,x) \cdot 10^{-16}$
S	D	Q	10^8	D	D	D

benefit of three precisions vs. $u_f \geq u, u = u_r$: no cond(A,x) term in forward error

Extensions II: Iterative refinement in 3 precisions

Standard (LU-based) IR in three precisions:

u_f	u	u_r	$\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
H	S	S	10^4	S	S	$\mathrm{cond}(A,x)\cdot 10^{-8}$
H	S	D	10^4	S	S	S
H	D	D	10^4	D	D	$\mathrm{cond}(A,x)\cdot 10^{-16}$
H	D	Q	10^4	D	D	D
S	S	S	10^8	S	S	$\mathrm{cond}(A,x)\cdot 10^{-8}$
S	S	D	10^8	S	S	S
S	D	D	10^8	D	D	$\mathrm{cond}(A,x)\cdot 10^{-16}$
S	D	Q	10^8	D	D	D

Extensions II: Iterative refinement in 3 precisions

Standard (LU-based) IR in three precisions:

u_f	u	u_r	$\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
H	S	S	10^4	S	S	$\mathrm{cond}(A,x)\cdot 10^{-8}$
H	S	D	10^4	S	S	S
H	D	D	10^4	D	D	$\mathrm{cond}(A,x)\cdot 10^{-16}$
H	D	Q	10^4	D	D	D
S	S	S	10^8	S	S	$\mathrm{cond}(A,x)\cdot 10^{-8}$
S	S	D	10^8	S	S	S
S	D	D	10^8	D	D	$\mathrm{cond}(A,x)\cdot 10^{-16}$
S	D	Q	10^8	D	D	D

If $\kappa_\infty(A) \leq 10^4$, can use lower precision factorization with no loss of accuracy!

Extensions II: Iterative refinement in 3 precisions

Benefits of GMRES-IR:

	u_f	u	u_r	$\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LU-IR	H	S	D	10^4	S	S	S
GMRES-IR	H	S	D	10^8	S	S	S
LU-IR	S	D	Q	10^8	D	D	D
GMRES-IR	S	D	Q	10^{16}	D	D	D
LU-IR	H	D	Q	10^4	D	D	D
GMRES-IR	H	D	Q	10^{12}	D	D	D

Extensions II: Iterative refinement in 3 precisions

Benefits of GMRES-IR:

	u_f	u	u_r	$\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LU-IR	H	S	D	10^4	S	S	S
GMRES-IR	H	S	D	10^8	S	S	S
LU-IR	S	D	Q	10^8	D	D	D
GMRES-IR	S	D	Q	10^{16}	D	D	D
LU-IR	H	D	Q	10^4	D	D	D
GMRES-IR	H	D	Q	10^{12}	D	D	D

With GMRES-IR, lower precision factorization will work for higher $\kappa_\infty(A)$

Extensions II: Iterative refinement in 3 precisions

Benefits of GMRES-IR:

	u_f	u	u_r	$\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LU-IR	H	S	D	10^4	S	S	S
GMRES-IR	H	S	D	10^8	S	S	S
LU-IR	S	D	Q	10^8	D	D	D
GMRES-IR	S	D	Q	10^{16}	D	D	D
LU-IR	H	D	Q	10^4	D	D	D
GMRES-IR	H	D	Q	10^{12}	D	D	D

Extensions II: Iterative refinement in 3 precisions

Benefits of GMRES-IR:

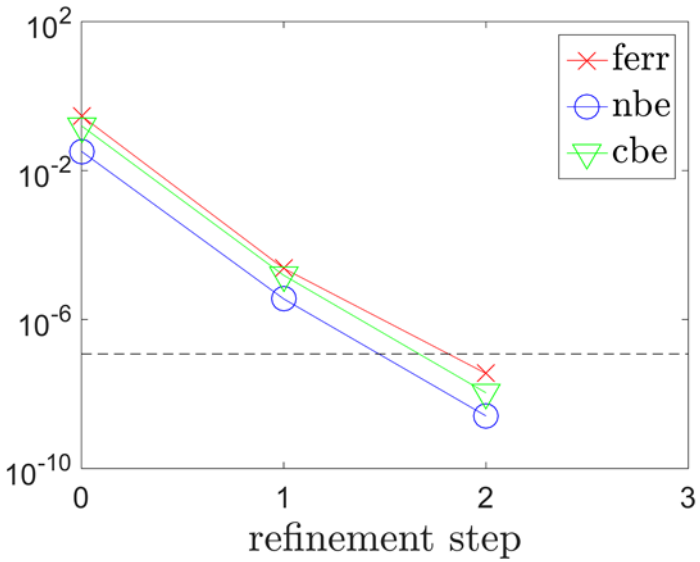
	u_f	u	u_r	$\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LU-IR	H	S	D	10^4	S	S	S
GMRES-IR	H	S	D	10^8	S	S	S
LU-IR	S	D	Q	10^8	D	D	D
GMRES-IR	S	D	Q	10^{16}	D	D	D
LU-IR	H	D	Q	10^4	D	D	D
GMRES-IR	H	D	Q	10^{12}	D	D	D

If $\kappa_\infty(A) \leq 10^{12}$, can use lower precision factorization with no loss of accuracy!

Numerical experiments

$u_f = 2^{-11}$ (half), $u = 2^{-24}$ (single), $u_r = 2^{-53}$ (double)

```
A = gallery('randsvd',100,kappa,2)
b = randn(100,1)
```

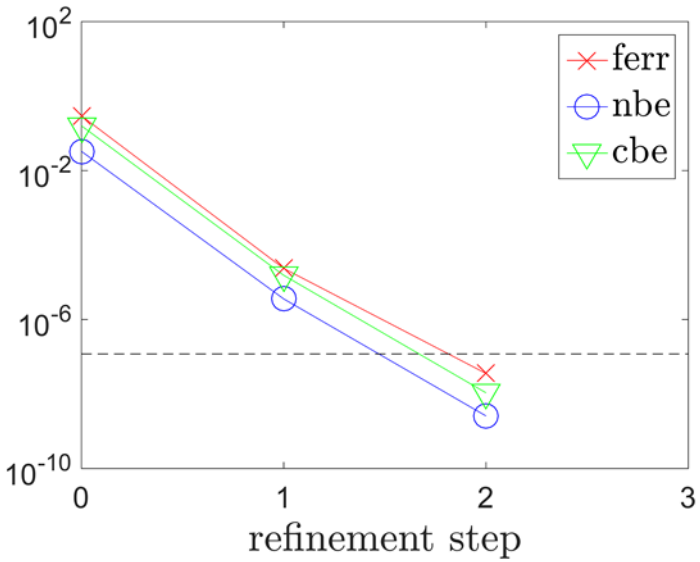


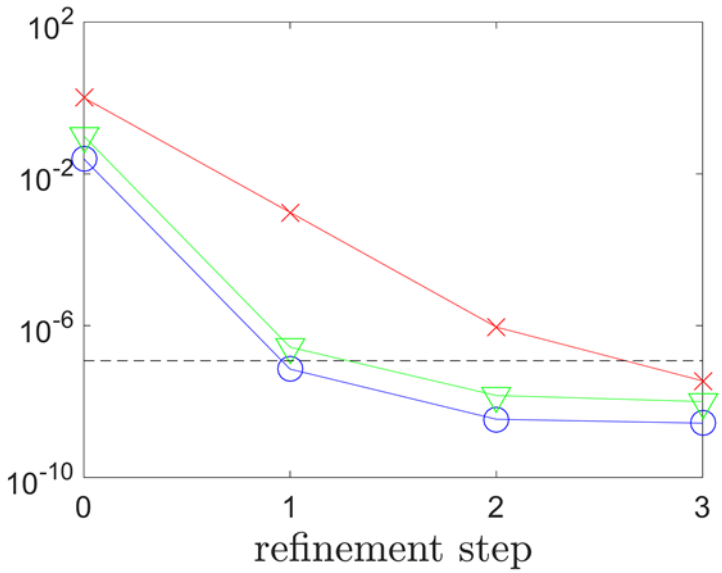
$\kappa_\infty(A) = 2 \cdot 10^2$
$\kappa_\infty(\tilde{A}) = 14$
GMRES its: 11 (5,6)

Numerical experiments

$u_f = 2^{-11}$ (half), $u = 2^{-24}$ (single), $u_r = 2^{-53}$ (double)

```
A = gallery('randsvd',100,kappa,2)
b = randn(100,1)
```





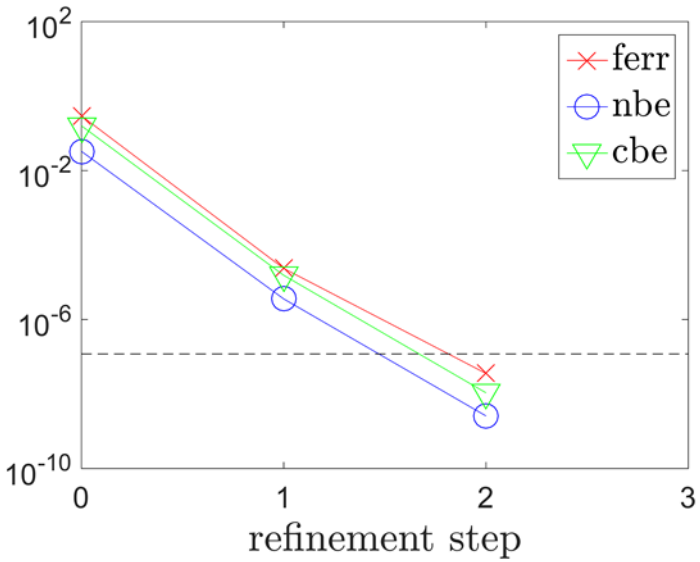
$\kappa_\infty(A) = 2 \cdot 10^2$
$\kappa_\infty(\tilde{A}) = 14$
GMRES its: 11 (5,6)

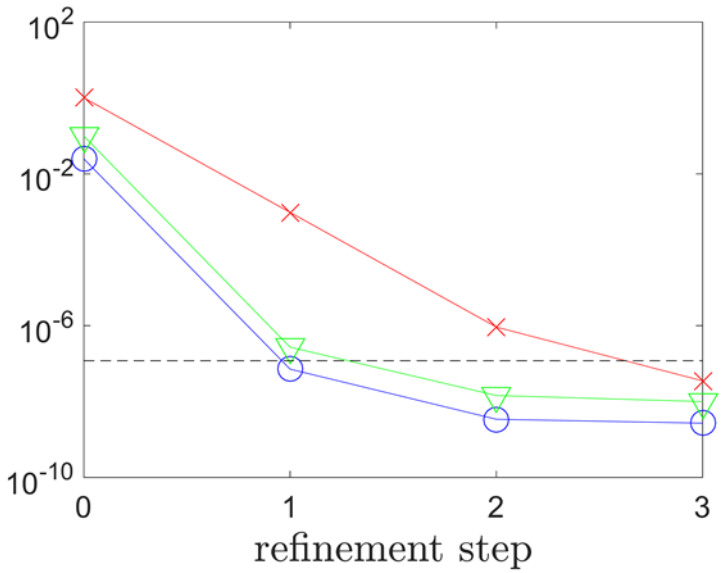
$\kappa_\infty(A) = 2 \cdot 10^6$
$\kappa_\infty(\tilde{A}) = 8 \cdot 10^6$
GMRES its: 40 (7,24,9)

Numerical experiments

$u_f = 2^{-11}$ (half), $u = 2^{-24}$ (single), $u_r = 2^{-53}$ (double)

```
A = gallery('randsvd',100,kappa,2)
b = randn(100,1)
```





$\kappa_\infty(A) = 2 \cdot 10^2$
$\kappa_\infty(\tilde{A}) = 14$
GMRES its: 11 (5,6)

$\kappa_\infty(A) = 2 \cdot 10^6$
$\kappa_\infty(\tilde{A}) = 8 \cdot 10^6$
GMRES its: 40 (7,24,9)

GMRES convergence rate

- If LU factorization is computed in lower precision, can diminish effectiveness as a preconditioner
- Theory only guarantees that GMRES will converge to an accurate solution within n iterations
  - If close to n iterations, no expected performance benefit

GMRES convergence rate

- If LU factorization is computed in lower precision, can diminish effectiveness as a preconditioner
- Theory only guarantees that GMRES will converge to an accurate solution within n iterations
  - If close to n iterations, no expected performance benefit
- If $\tilde{A}$ is nonnormal, spectrum of $\tilde{A}$ is irrelevant to GMRES convergence rate (Greenbaum, Pták, Strakoš, 1996)

GMRES convergence rate

- If LU factorization is computed in lower precision, can diminish effectiveness as a preconditioner
- Theory only guarantees that GMRES will converge to an accurate solution within n iterations
  - If close to n iterations, no expected performance benefit

- If $\tilde{A}$ is nonnormal, spectrum of $\tilde{A}$ is irrelevant to GMRES convergence rate (Greenbaum, Pták, Strakoš, 1996)
- If $\tilde{A}$ is normal, spectrum of $\tilde{A}$ determines GMRES convergence rate (Liesen & Tichý, 2004)
  - But small $\kappa_\infty(\tilde{A})$ may not mean fast GMRES convergence
    - e.g., if $\tilde{A}$ has a cluster of eigenvalues close to the origin

GMRES convergence rate

- If LU factorization is computed in lower precision, can diminish effectiveness as a preconditioner
- Theory only guarantees that GMRES will converge to an accurate solution within n iterations
  - If close to n iterations, no expected performance benefit

- If $\tilde{A}$ is nonnormal, spectrum of $\tilde{A}$ is irrelevant to GMRES convergence rate (Greenbaum, Pták, Strakoš, 1996)
- If $\tilde{A}$ is normal, spectrum of $\tilde{A}$ determines GMRES convergence rate (Liesen & Tichý, 2004)
  - But small $\kappa_\infty(\tilde{A})$ may not mean fast GMRES convergence
    - e.g., if $\tilde{A}$ has a cluster of eigenvalues close to the origin

$\Rightarrow$ Can only make guarantees on fast GMRES convergence in some cases, e.g., normality and no eigenvalue cluster near origin

- Potential fixes for slow GMRES convergence: apply additional preconditioner, deflation, other Krylov subspace methods

Thank you!

erinc@cims.nyu.edu
http://math.nyu.edu/~erinc/

Resources:

- E. Carson and N. J. Higham. A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems. MIMS EPrint 2017.12.
- E. Carson and N. J. Higham. Accelerating the solution of linear systems by iterative refinement in three precisions. MIMS EPrint 2017.24.
- MATLAB code for iterative refinement in 3 precisions: https://github.com/eccarson/ir3/

IEEE Standard 754-1985 and 2008 Revision

Type	Size	Range	$u = 2^{-t}$
half	16 bits	$10^{\pm 5}$	$2^{-11} \approx 4.9 \times 10^{-4}$
single	32 bits	$10^{\pm 38}$	$2^{-24} \approx 6.0 \times 10^{-8}$
double	64 bits	$10^{\pm 308}$	$2^{-53} \approx 1.1 \times 10^{-16}$
quadruple	128 bits	$10^{\pm 4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

- Arithmetic ops ($+, -, *, /, \sqrt{}$) performed *as if* first calculated to infinite precision, then rounded.
- Default: round to nearest, round to even in case of tie.
- Half precision is a *storage format only*.

Summary of the sizes of the quantities in assumptions (2.3)–(2.5) for solution of the correction equation with LU factorization (section 7) and GMRES-IR (section 8). Note that $f(n) = O(n^2)$.

	$u_s\|E_i\|_\infty$	$u_s \max(c_1, c_2)$	$u_s\|G_i\|_\infty$
IR w/LU fact.	$3nu_f\||A^{-1}||\widehat{L}||\widehat{U}|\|_\infty$	$3nu_f \dfrac{\||\widehat{L}||\widehat{U}|\|_\infty}{\|A\|_\infty}$	$3nu_f\||\widehat{L}||\widehat{U}|\|_\infty$
GMRES-IR	$uf(n)(1+\gamma_n^f\kappa_\infty(A))^2$	$O(u)$	$O(u\|A\|_\infty)$

Different scenarios for iterative refinement in IEEE arithmetic. The columns represent different choices for u_f, u, and u_r, where in the notation of Algorithm 1.1 the data is stored at precision u, the solves in steps 1 and 4 are carried out in precision $u_f = u_s$, and residuals are computed at precision u_r. The last column indicates whether any existing backward or forward error analysis is applicable to this situation when LU factorization is used as the solver.

	Precision			
Usage	Half	Single	Double	Existing analysis?
Traditional		data, solve	residual	✓
Traditional			data, solve, residual	✓
2000s		solve	data, residual	✓
New	solve	data, residual		✓
New	solve	data	residual	×
New	solve		data, residual	✓

Summary of existing rounding error analyses for iterative refinement in floating point arithmetic indicating (a) whether the analyses apply to LU factorization only or to an arbitrary solver, (b) whether the backward or forward error analyses are componentwise ("comp") or normwise ("norm"), and (c) the assumptions on the precisions u_f, u_s, u, u_r in Algorithm 1.1 ($u_f = u$ and $u_s = u_f$ unless otherwise stated).

	Year	Solver	Forward error	Backward error	Precisions
Moler [26]	1967	LU	norm	–	$u \geq u_r$
Stewart [33]	1973	LU	norm	–	$u \geq u_r$
Jankowski et al. [21]	1977	arb.	norm	norm	$u = u_r$
Skeel [31]	1980	LU	comp	comp	$u \geq u_r$
Higham [16]	1991	arb.	comp	comp	$u = u_r$
Higham [17], [18]	1997	arb.	comp	comp	$u \geq u_r$
Tisseur [34]	2001	arb.	norm	norm	$u \geq u_r$
Langou et al. [23]	2006	LU	norm	norm	$u_f \geq u = u_r$
Carson and Higham [9]	2017	arb.	comp	–	$u \geq u_r$
This work	2017	arb.	comp	comp, norm	$u_f \geq u_s \geq u \geq u_r$