Error Bounds for Iterative Refinement in Three Precisions

Erin C. Carson, New York University

Nicholas J. Higham, University of Manchester

SIAM Annual Meeting
Portland, Oregon
July 13, 2018

Hardware Support for Multiprecision Computation

Use of low precision in machine learning has driven emergence of low-precision capabilities in hardware:

- Half precision (FP16) defined as storage format in 2008 IEEE standard
- ARM NEON: SIMD architecture, instructions for 8x16-bit, 4x32-bit, 2x64-bit
- AMD Radeon Instinct MI25 GPU, 2017:
  - single: 12.3 TFLOPS, half: 24.6 TFLOPS
- NVIDIA Tesla P100, 2016: native ISA support for 16-bit FP arithmetic
- NVIDIA Tesla V100, 2017: tensor cores for half precision;

  4x4 matrix multiply in one clock cycle
  - double: 7 TFLOPS, half+tensor: 112 TFLOPS (**16x!**)
- Google's Tensor processing unit (TPU): quantizes 32-bit FP computations into 8-bit integer arithmetic
- Aurora Exascale supercomputer: (2021) Expected extensive support for reduced-precision arithmetic (32/16/8-bit)

Iterative Refinement for $Ax = b$

Iterative refinement: well-established method for improving an approximate solution to $Ax = b$

A is $n \times n$ and nonsingular; u is unit roundoff

Solve $Ax_0 = b$ by LU factorization

for $i = 0\colon \mathrm{maxit}$

$r_i = b - Ax_i$

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

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

Iterative Refinement for $Ax = b$

Iterative refinement: well-established method for improving an approximate solution to $Ax = b$

A is $n \times n$ and nonsingular; u is unit roundoff

Solve $Ax_0 = b$ by LU factorization (in precision u)

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

$r_i = b - Ax_i$ (in precision u^2)

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

$x_{i+1} = x_i + d_i$ (in precision u)

"Traditional" (high-precision residual computation)

[Wilkinson, 1948] (fixed point), [Moler, 1967] (floating point)

Iterative Refinement for $Ax = b$

Iterative refinement: well-established method for improving an approximate solution to $Ax = b$

A is $n \times n$ and nonsingular; u is unit roundoff

Solve $Ax_0 = b$ by LU factorization (in precision u)

for $i = 0\colon \mathrm{maxit}$

$r_i = b - Ax_i$ (in precision u)

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

$x_{i+1} = x_i + d_i$ (in precision u)

"Fixed-Precision"

[Jankowski and Woźniakowski, 1977], [Skeel, 1980], [Higham, 1991]

Iterative Refinement for $Ax = b$

Iterative refinement: well-established method for improving an approximate solution to $Ax = b$

A is $n \times n$ and nonsingular; u is unit roundoff

Solve $Ax_0 = b$ by LU factorization (in precision $u^{1/2}$)

for $i = 0\colon \mathrm{maxit}$

$r_i = b - Ax_i$ (in precision u)

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

$x_{i+1} = x_i + d_i$ (in precision u)

"Low-precision factorization"

[Langou et al., 2006], [Arioli and Duff, 2009], [Hogg and Scott, 2010], [Abdelfattah et al., 2016]

Iterative Refinement in 3 Precisions

Existing analyses only support at most two precisions

Can we combine the performance benefits of low-precision factorization IR with the accuracy of traditional IR?

Iterative Refinement in 3 Precisions

Existing analyses only support at most two precisions

Can we combine the performance benefits of low-precision factorization IR with the accuracy of traditional IR?

$\Rightarrow$ **3-precision iterative refinement**

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

$$u_f \geq u \geq u_r$$

Iterative Refinement in 3 Precisions

Existing analyses only support at most two precisions

Can we combine the performance benefits of low-precision factorization IR with the accuracy of traditional IR?

$\Rightarrow$ **3-precision iterative refinement**

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

$$u_f \geq u \geq u_r$$

- New analysis **generalizes** existing types of IR:

[C. and Higham, SIAM SISC 40(2), 2018]

Traditional	$u_f = u, u_r = u^2$
Fixed precision	$u_f = u = u_r$
Lower precision factorization	$u_f^2 = u = u_r$

(and **improves** upon existing analyses in some cases)

Iterative Refinement in 3 Precisions

Existing analyses only support at most two precisions

Can we combine the performance benefits of low-precision factorization IR with the accuracy of traditional IR?

$\Rightarrow$ **3-precision iterative refinement**

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

$$u_f \geq u \geq u_r$$

- New analysis **generalizes** existing types of IR:

[C. and Higham, SIAM SISC 40(2), 2018]

Traditional	$u_f = u, u_r = u^2$
Fixed precision	$u_f = u = u_r$
Lower precision factorization	$u_f^2 = u = u_r$

(and **improves** upon existing analyses in some cases)

- Enables **new** types of IR: (half, single, double), (half, single, quad), (half, double, quad), etc.

Key Analysis Innovations I

Obtain tighter upper bounds:

Typical bounds used in analysis: $\|A(x - \hat{x}_i)\|_\infty \le \|A\|_\infty \|x - \hat{x}_i\|_\infty$

Key Analysis Innovations I

Obtain tighter upper bounds:

Typical bounds used in analysis: $\|A(x - \hat{x}_i)\|_\infty \leq \|A\|_\infty \|x - \hat{x}_i\|_\infty$

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

Key Analysis Innovations I

Obtain tighter upper bounds:

Typical bounds used in analysis: $\|A(x-\hat{x}_i)\|_\infty \leq \|A\|_\infty \|x-\hat{x}_i\|_\infty$

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

For a stable refinement scheme, in early stages we expect

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

Key Analysis Innovations I

Obtain tighter upper bounds:

Typical bounds used in analysis: $\|A(x-\hat{x}_i)\|_\infty \leq \|A\|_\infty \|x-\hat{x}_i\|_\infty$

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

For a stable refinement scheme, in early stages we expect

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

But close to convergence,

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

Key Analysis Innovations I

$$\|r_i\|_2 = \mu_i^{(2)} \|A\|_2 \|x - \hat{x}_i\|_2$$

$$x - \hat{x}_i \quad = \quad V\Sigma^{-1}U^T r_i \quad = \quad \sum_{j=1}^{n} \frac{(u_j^T r_i) v_j}{\sigma_j} \qquad\qquad (A = U\Sigma V^T)$$

Key Analysis Innovations I

$$\|r_i\|_2 = \mu_i^{(2)} \|A\|_2 \|x - \hat{x}_i\|_2$$

$$x - \hat{x}_i \quad = \quad V\Sigma^{-1}U^T r_i \quad = \quad \sum_{j=1}^{n} \frac{(u_j^T r_i) v_j}{\sigma_j} \qquad\qquad (A = U\Sigma V^T)$$

$$\|x - \hat{x}_i\|_2^2 \quad \geq \quad \sum_{j=n+1-k}^{n} \frac{(u_j^T r_i)^2}{\sigma_j^2} \quad \geq \quad \frac{1}{\sigma_{n+1-k}^2} \sum_{j=n+1-k}^{n} (u_j^T r_i)^2 \quad = \quad \frac{\|P_k r_i\|_2^2}{\sigma_{n+1-k}^2}$$

where $P_k = U_k U_k^T, U_k = [u_{n+1-k}, \ldots, u_n]$

Key Analysis Innovations I

$$\|r_i\|_2 = \mu_i^{(2)} \|A\|_2 \|x - \hat{x}_i\|_2$$

$$x - \hat{x}_i \quad = \quad V\Sigma^{-1}U^T r_i \quad = \quad \sum_{j=1}^{n} \frac{(u_j^T r_i) v_j}{\sigma_j} \qquad (A = U\Sigma V^T)$$

$$\|x - \hat{x}_i\|_2^2 \quad \geq \sum_{j=n+1-k}^{n} \frac{\left(u_j^T r_i\right)^2}{\sigma_j^2} \quad \geq \quad \frac{1}{\sigma_{n+1-k}^2} \sum_{j=n+1-k}^{n} \left(u_j^T r_i\right)^2 \quad = \quad \frac{\|P_k r_i\|_2^2}{\sigma_{n+1-k}^2}$$

where $P_k = U_k U_k^T, U_k = [u_{n+1-k}, \dots, u_n]$

$$\mu_i^{(2)} \leq \frac{\|r_i\|_2}{\|P_k r_i\|_2} \frac{\sigma_{n+1-k}}{\sigma_1}$$

Key Analysis Innovations I

$$\|r_i\|_2 = \mu_i^{(2)} \|A\|_2 \|x - \hat{x}_i\|_2$$

$$x - \hat{x}_i \quad = \quad V\Sigma^{-1}U^T r_i \quad = \quad \sum_{j=1}^{n} \frac{(u_j^T r_i) v_j}{\sigma_j} \qquad\qquad (A = U\Sigma V^T)$$

$$\|x - \hat{x}_i\|_2^2 \quad \geq \quad \sum_{j=n+1-k}^{n} \frac{\left(u_j^T r_i\right)^2}{\sigma_j^2} \quad \geq \quad \frac{1}{\sigma_{n+1-k}^2} \sum_{j=n+1-k}^{n} \left(u_j^T r_i\right)^2 \quad = \quad \frac{\|P_k r_i\|_2^2}{\sigma_{n+1-k}^2}$$

where $P_k = U_k U_k^T, U_k = [u_{n+1-k}, \dots, u_n]$

$$\mu_i^{(2)} \leq \frac{\|r_i\|_2}{\|P_k r_i\|_2} \frac{\sigma_{n+1-k}}{\sigma_1}$$

- $\mu_i^{(2)} \ll 1$ if r_i contains significant component in span(U_k) for any k s.t. $\sigma_{n+1-k} \approx \sigma_n$

Key Analysis Innovations I

$$\|r_i\|_2 = \mu_i^{(2)} \|A\|_2 \|x - \hat{x}_i\|_2$$

$$x - \hat{x}_i \quad = \quad V\Sigma^{-1}U^T r_i \quad = \quad \sum_{j=1}^{n} \frac{(u_j^T r_i) v_j}{\sigma_j} \qquad (A = U\Sigma V^T)$$

$$\|x - \hat{x}_i\|_2^2 \quad \geq \sum_{j=n+1-k}^{n} \frac{(u_j^T r_i)^2}{\sigma_j^2} \quad \geq \quad \frac{1}{\sigma_{n+1-k}^2} \sum_{j=n+1-k}^{n} (u_j^T r_i)^2 \quad = \quad \frac{\|P_k r_i\|_2^2}{\sigma_{n+1-k}^2}$$

where $P_k = U_k U_k^T, U_k = [u_{n+1-k}, \dots, u_n]$

$$\mu_i^{(2)} \leq \frac{\|r_i\|_2}{\|P_k r_i\|_2} \frac{\sigma_{n+1-k}}{\sigma_1}$$

- $\mu_i^{(2)} \ll 1$ if r_i contains significant component in span(U_k) for any k s.t. $\sigma_{n+1-k} \approx \sigma_n$
- Expect $\mu_i^{(2)} \ll 1$ when r_i is "typical", i.e., contains sizeable components in the direction of each left singular vector
- In that case, $x - \hat{x}_i$ is not "typical", i.e., it contains large components in right singular vectors corresponding to small singular values of A

Key Analysis Innovations I

$$\|r_i\|_2 = \mu_i^{(2)} \|A\|_2 \|x - \hat{x}_i\|_2$$

$$x - \hat{x}_i \quad = \quad V\Sigma^{-1}U^T r_i \quad = \quad \sum_{j=1}^{n} \frac{(u_j^T r_i) v_j}{\sigma_j} \qquad (A = U\Sigma V^T)$$

$$\|x - \hat{x}_i\|_2^2 \quad \geq \sum_{j=n+1-k}^{n} \frac{\left(u_j^T r_i\right)^2}{\sigma_j^2} \quad \geq \quad \frac{1}{\sigma_{n+1-k}^2} \sum_{j=n+1-k}^{n} \left(u_j^T r_i\right)^2 \quad = \quad \frac{\|P_k r_i\|_2^2}{\sigma_{n+1-k}^2}$$

where $P_k = U_k U_k^T, U_k = [u_{n+1-k}, \dots, u_n]$

$$\mu_i^{(2)} \leq \frac{\|r_i\|_2}{\|P_k r_i\|_2} \frac{\sigma_{n+1-k}}{\sigma_1}$$

- $\mu_i^{(2)} \ll 1$ if r_i contains significant component in $\text{span}(U_k)$ for any k s.t. $\sigma_{n+1-k} \approx \sigma_n$
- Expect $\mu_i^{(2)} \ll 1$ when r_i is "typical", i.e., contains sizeable components in the direction of each left singular vector
- In that case, $x - \hat{x}_i$ is not "typical", i.e., it contains large components in right singular vectors corresponding to small singular values of A
- Wilkinson (1977), comment in unpublished manuscript: $\mu_i^{(2)}$ increases with i

Key Analysis Innovations II

Allow for general solver:

Let u_s be the *effective precision* of the solve, with $u \leq u_s \leq u_f$

Key Analysis Innovations II

Allow for general solver:

Let u_s be the *effective precision* of the solve, with $u \le u_s \le u_f$

example: LU solve:

$u_s = u_f$

Key Analysis Innovations II

Allow for general solver:

Let $\boldsymbol{u_s}$ be the *effective precision* of the solve, with $\boldsymbol{u} \le \boldsymbol{u_s} \le \boldsymbol{u_f}$

example: LU solve:

$\boldsymbol{u_s} = \boldsymbol{u_f}$

Assume computed solution $\hat{d}_i$ to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + \boldsymbol{u_s} E_i) d_i, \quad \boldsymbol{u_s} \|E_i\|_\infty < 1$

   → normwise relative forward error is bounded by multiple of u_s and is less than 1

Key Analysis Innovations II

Allow for general solver:

Let $\boldsymbol{u_s}$ be the *effective precision* of the solve, with $\boldsymbol{u} \leq \boldsymbol{u_s} \leq \boldsymbol{u_f}$

example: LU solve:

$\boldsymbol{u_s} = \boldsymbol{u_f}$

Assume computed solution $\hat{d}_i$ to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + \boldsymbol{u_s} E_i) d_i, \quad \boldsymbol{u_s} \|E_i\|_\infty < 1$

   → normwise relative forward error is bounded by multiple of u_s and is less than 1

$$\boldsymbol{u_s} \|E_i\|_\infty \leq 3n \boldsymbol{u_f} \left\| |A^{-1}| |\hat{L}| |\hat{U}| \right\|_\infty$$

Key Analysis Innovations II

Allow for general solver:
Let $\boldsymbol{u_s}$ be the *effective precision* of the solve, with $\boldsymbol{u} \leq \boldsymbol{u_s} \leq \boldsymbol{u_f}$

example: LU solve:
$\boldsymbol{u_s} = \boldsymbol{u_f}$

Assume computed solution $\hat{d}_i$ to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + \boldsymbol{u_s} E_i) d_i, \quad \boldsymbol{u_s} \|E_i\|_\infty < 1$

   → normwise relative forward error is bounded by multiple of u_s and is less than 1

$$\boldsymbol{u_s} \|E_i\|_\infty \leq 3n \boldsymbol{u_f} \left\| |A^{-1}| |\hat{L}| |\hat{U}| \right\|_\infty$$

2. $\left\| \hat{r}_i - A\hat{d}_i \right\|_\infty \leq \boldsymbol{u_s} (c_1 \|A\|_\infty \left\| \hat{d}_i \right\|_\infty + c_2 \|\hat{r}_i\|_\infty)$

   → normwise relative backward error is at most $\max(c_1, c_2)\, u_s$

Key Analysis Innovations II

Allow for general solver:

Let $\boldsymbol{u_s}$ be the *effective precision* of the solve, with $\boldsymbol{u} \le \boldsymbol{u_s} \le \boldsymbol{u_f}$

example: LU solve:

$\boldsymbol{u_s} = \boldsymbol{u_f}$

Assume computed solution $\hat{d}_i$ to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + \boldsymbol{u_s} E_i) d_i, \quad \boldsymbol{u_s} \|E_i\|_\infty < 1$

   → normwise relative forward error is bounded by multiple of u_s and is less than 1

$$\boldsymbol{u_s} \|E_i\|_\infty \le 3n\boldsymbol{u_f} \left\| |A^{-1}| |\hat{L}| |\hat{U}| \right\|_\infty$$

2. $\left\| \hat{r}_i - A\hat{d}_i \right\|_\infty \le \boldsymbol{u_s} (c_1 \|A\|_\infty \left\| \hat{d}_i \right\|_\infty + c_2 \|\hat{r}_i\|_\infty)$

   → normwise relative backward error is at most $\max(c_1, c_2)\, u_s$

$$\max(c_1, c_2)\, \boldsymbol{u_s} \le \frac{3n\boldsymbol{u_f} \left\| |\hat{L}| |\hat{U}| \right\|_\infty}{\|A\|_\infty}$$

Key Analysis Innovations II

Allow for general solver:
Let $\boldsymbol{u_s}$ be the *effective precision* of the solve, with $\boldsymbol{u} \le \boldsymbol{u_s} \le \boldsymbol{u_f}$

example: LU solve:
$\boldsymbol{u_s} = \boldsymbol{u_f}$

Assume computed solution $\hat{d}_i$ to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + \boldsymbol{u_s} E_i) d_i, \quad \boldsymbol{u_s} \|E_i\|_\infty < 1$

   → normwise relative forward error is bounded by multiple of u_s and is less than 1

$$\boldsymbol{u_s} \|E_i\|_\infty \le 3n\boldsymbol{u_f} \left\| |A^{-1}| |\hat{L}| |\hat{U}| \right\|_\infty$$

2. $\left\| \hat{r}_i - A\hat{d}_i \right\|_\infty \le \boldsymbol{u_s} (c_1 \|A\|_\infty \left\| \hat{d}_i \right\|_\infty + c_2 \|\hat{r}_i\|_\infty)$

   → normwise relative backward error is at most $\max(c_1, c_2)\, u_s$

$$\max(c_1, c_2)\, \boldsymbol{u_s} \le \frac{3n\boldsymbol{u_f} \left\| |\hat{L}| |\hat{U}| \right\|_\infty}{\|A\|_\infty}$$

3. $\left| \hat{r}_i - A\hat{d}_i \right| \le \boldsymbol{u_s} G_i |\hat{d}_i|$

   → componentwise relative backward error is bounded by a multiple of u_s

Key Analysis Innovations II

Allow for general solver:
Let $\boldsymbol{u_s}$ be the *effective precision* of the solve, with $\boldsymbol{u} \leq \boldsymbol{u_s} \leq \boldsymbol{u_f}$

example: LU solve:
$\boldsymbol{u_s} = \boldsymbol{u_f}$

Assume computed solution $\hat{d}_i$ to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + \boldsymbol{u_s} E_i) d_i, \quad \boldsymbol{u_s} \|E_i\|_\infty < 1$

   → normwise relative forward error is bounded by multiple of u_s and is less than 1

   $\boldsymbol{u_s} \|E_i\|_\infty \leq 3n\boldsymbol{u_f} \left\| |A^{-1}||\hat{L}||\hat{U}| \right\|_\infty$

2. $\left\|\hat{r}_i - A\hat{d}_i\right\|_\infty \leq \boldsymbol{u_s}(c_1\|A\|_\infty \left\|\hat{d}_i\right\|_\infty + c_2\|\hat{r}_i\|_\infty)$

   → normwise relative backward error is at most $\max(c_1, c_2)\, u_s$

   $\max(c_1, c_2)\, \boldsymbol{u_s} \leq \dfrac{3n\boldsymbol{u_f} \left\| |\hat{L}||\hat{U}| \right\|_\infty}{\|A\|_\infty}$

3. $\left|\hat{r}_i - A\hat{d}_i\right| \leq \boldsymbol{u_s} G_i |\hat{d}_i|$

   → componentwise relative backward error is bounded by a multiple of u_s

   $\boldsymbol{u_s} \|G_i\|_\infty \leq 3n\boldsymbol{u_f} \left\| |\hat{L}||\hat{U}| \right\|_\infty$

Key Analysis Innovations II

Allow for general solver:
Let $\boldsymbol{u_s}$ be the *effective precision* of the solve, with $\boldsymbol{u} \leq \boldsymbol{u_s} \leq \boldsymbol{u_f}$

example: LU solve:
$\boldsymbol{u_s} = \boldsymbol{u_f}$

Assume computed solution $\hat{d}_i$ to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + \boldsymbol{u_s} E_i) d_i, \quad \boldsymbol{u_s} \|E_i\|_\infty < 1$

   → normwise relative forward error is bounded by multiple of u_s and is less than 1

   $\boldsymbol{u_s} \|E_i\|_\infty \leq 3n \boldsymbol{u_f} \left\| |A^{-1}| |\hat{L}| |\hat{U}| \right\|_\infty$

2. $\left\| \hat{r}_i - A\hat{d}_i \right\|_\infty \leq \boldsymbol{u_s} (c_1 \|A\|_\infty \left\| \hat{d}_i \right\|_\infty + c_2 \|\hat{r}_i\|_\infty)$

   → normwise relative backward error is at most $\max(c_1, c_2)\, u_s$

   $\max(c_1, c_2)\, \boldsymbol{u_s} \leq \dfrac{3n \boldsymbol{u_f} \left\| |\hat{L}| |\hat{U}| \right\|_\infty}{\|A\|_\infty}$

3. $\left| \hat{r}_i - A\hat{d}_i \right| \leq \boldsymbol{u_s} G_i |\hat{d}_i|$

   → componentwise relative backward error is bounded by a multiple of u_s

   $\boldsymbol{u_s} \|G_i\|_\infty \leq 3n \boldsymbol{u_f} \left\| |\hat{L}| |\hat{U}| \right\|_\infty$

E_i, c_1, c_2, and G_i depend on A, $\hat{r}_i$, n, and $\boldsymbol{u_s}$

Forward Error for IR3

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

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

$$\text{cond}(A) = \| \, |A^{-1}||A| \, \|_\infty$$

$$\text{cond}(A, x) = \| \, |A^{-1}||A||x| \, \|_\infty / \|x\|_\infty$$

Forward Error for IR3

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

$$\kappa_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty$$
$$\text{cond}(A) = \|\,|A^{-1}||A|\,\|_\infty$$
$$\text{cond}(A,x) = \|\,|A^{-1}||A||x|\,\|_\infty / \|x\|_\infty$$

Theorem [C. and Higham, SISC 40(2), 2018]

For IR in precisions $\boldsymbol{u_f} \geq \boldsymbol{u} \geq \boldsymbol{u_r}$ and effective solve precision $\boldsymbol{u_s}$, if

$$\phi_i \equiv 2\boldsymbol{u_s} \min(\text{cond}(A), \kappa_\infty(A)\mu_i) + \boldsymbol{u_s}\|E_i\|_\infty$$

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

$$\frac{\|x - \hat{x}_i\|_\infty}{\|x\|_\infty} \lesssim 4N\boldsymbol{u_r}\, \text{cond}(A,x) + \boldsymbol{u},$$

where N is the maximum number of nonzeros per row in A.

Forward Error for IR3

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

$$\kappa_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty$$
$$\text{cond}(A) = \|\, |A^{-1}||A| \,\|_\infty$$
$$\text{cond}(A,x) = \|\, |A^{-1}||A||x| \,\|_\infty / \|x\|_\infty$$

Theorem [C. and Higham, SISC 40(2), 2018]

For IR in precisions $\boldsymbol{u_f} \geq \boldsymbol{u} \geq \boldsymbol{u_r}$ and effective solve precision $\boldsymbol{u_s}$, if

$$\phi_i \equiv 2\boldsymbol{u_s} \min(\text{cond}(A), \kappa_\infty(A)\mu_i) + \boldsymbol{u_s}\|E_i\|_\infty$$

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

$$\frac{\|x - \hat{x}_i\|_\infty}{\|x\|_\infty} \lesssim 4N\boldsymbol{u_r}\, \text{cond}(A,x) + \boldsymbol{u},$$

where N is the maximum number of nonzeros per row in A.

Analogous traditional bounds: $\phi_i \equiv 3n\boldsymbol{u_f}\kappa_\infty(A)$

Normwise Backward Error for IR3

Theorem [C. and Higham, SISC 40(2), 2018]

For IR in precisions $\boldsymbol{u_f} \geq \boldsymbol{u} \geq \boldsymbol{u_r}$ and effective solve precision $\boldsymbol{u_s}$, if

$$\phi_i \equiv (c_1 \kappa_\infty(A) + c_2)\boldsymbol{u_s}$$

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

$$\|b - A\hat{x}_i\|_\infty \lesssim N\boldsymbol{u}(\|b\|_\infty + \|A\|_\infty \|\hat{x}_i\|_\infty),$$

where N is the maximum number of nonzeros per row in A.

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

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

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

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

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

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

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\text{cond}(A,x) \cdot 10^{-8}$
	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\text{cond}(A,x) \cdot 10^{-16}$
	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
Fixed	S	S	S	10^8	10^{-8}	10^{-8}	$\text{cond}(A,x) \cdot 10^{-8}$
Trad.	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\text{cond}(A,x) \cdot 10^{-16}$
	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

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

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\text{cond}(A,x) \cdot 10^{-8}$
New	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\text{cond}(A,x) \cdot 10^{-16}$
New	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
Fixed	S	S	S	10^8	10^{-8}	10^{-8}	$\text{cond}(A,x) \cdot 10^{-8}$
Trad.	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\text{cond}(A,x) \cdot 10^{-16}$
New	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

$\Rightarrow$ Benefit of IR3 vs. "LP fact.": no $\text{cond}(A,x)$ term in forward error

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

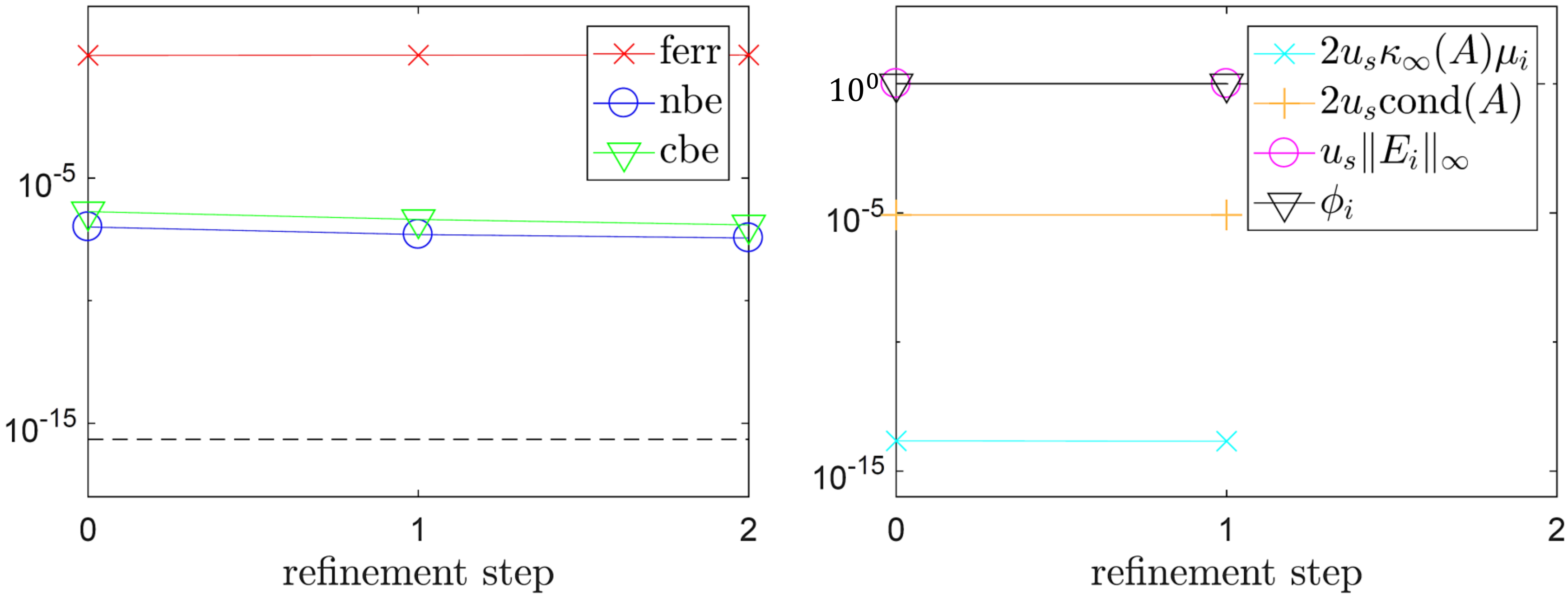
	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error: norm	Backward error: comp	Forward error
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\text{cond}(A,x) \cdot 10^{-8}$
New	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\text{cond}(A,x) \cdot 10^{-16}$
New	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
Fixed	S	S	S	10^8	10^{-8}	10^{-8}	$\text{cond}(A,x) \cdot 10^{-8}$
Trad.	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\text{cond}(A,x) \cdot 10^{-16}$
New	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

$\Rightarrow$ Benefit of IR3 vs. traditional IR: As long as $\kappa_\infty(A) \leq 10^4$, can use lower precision factorization w/no loss of accuracy!

```
A = gallery('randsvd', 100, 1e9, 2)
b = randn(100,1)
```

$\kappa_\infty(A) \approx$ 2e10, $\text{cond}(A,x) \approx$ 5e9

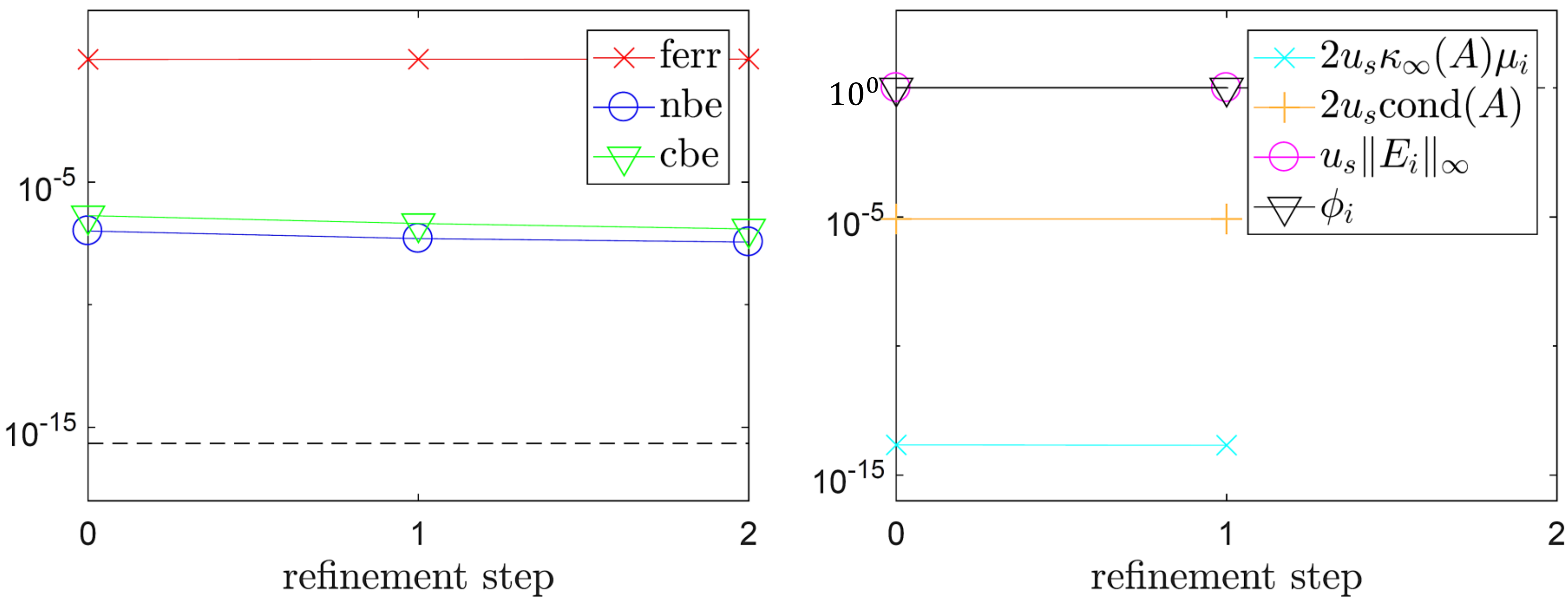
Standard (LU-based) IR with u_f: single, u: double, u_r: double



```
A = gallery('randsvd', 100, 1e9, 2)
b = randn(100,1)
```

$\kappa_\infty(A) \approx$ 2e10, $\text{cond}(A,x) \approx$ 5e9

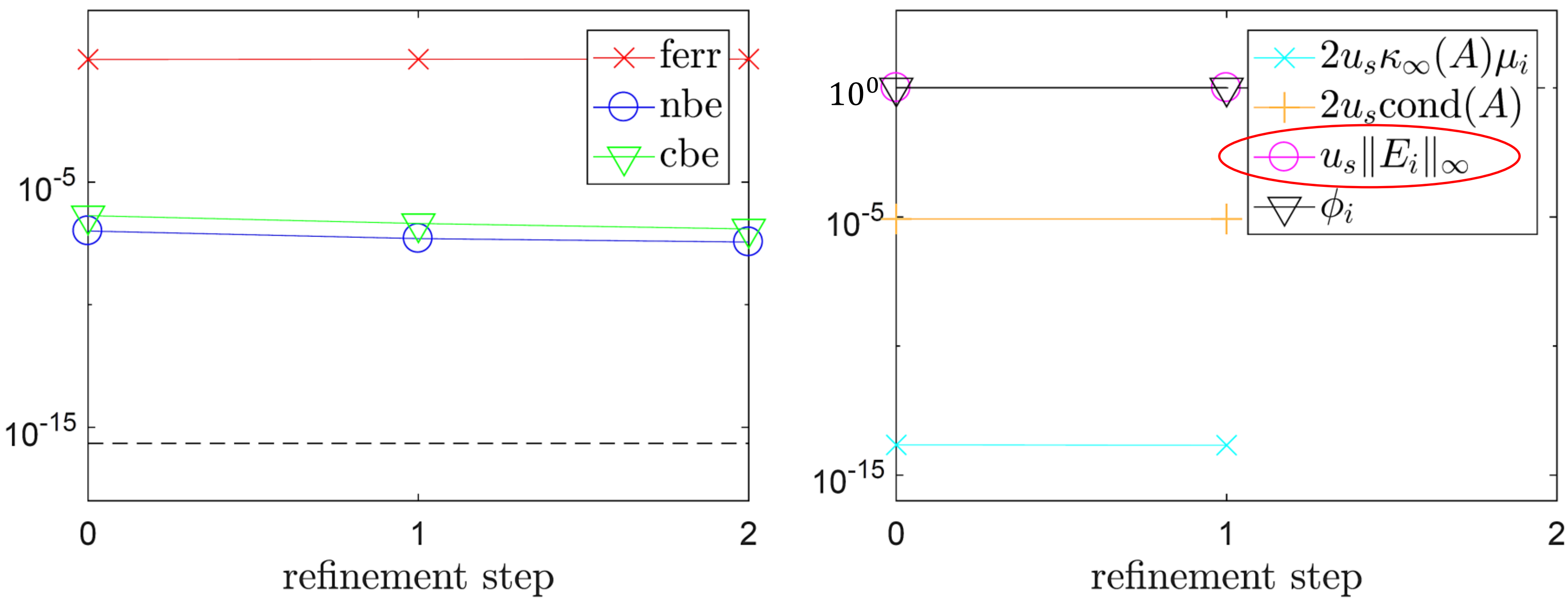
Standard (LU-based) IR with u_f: single, u: double, u_r: quad



```
A = gallery('randsvd', 100, 1e9, 2)
b = randn(100,1)
```

$\kappa_\infty(A) \approx$ 2e10, $\text{cond}(A,x) \approx$ 5e9

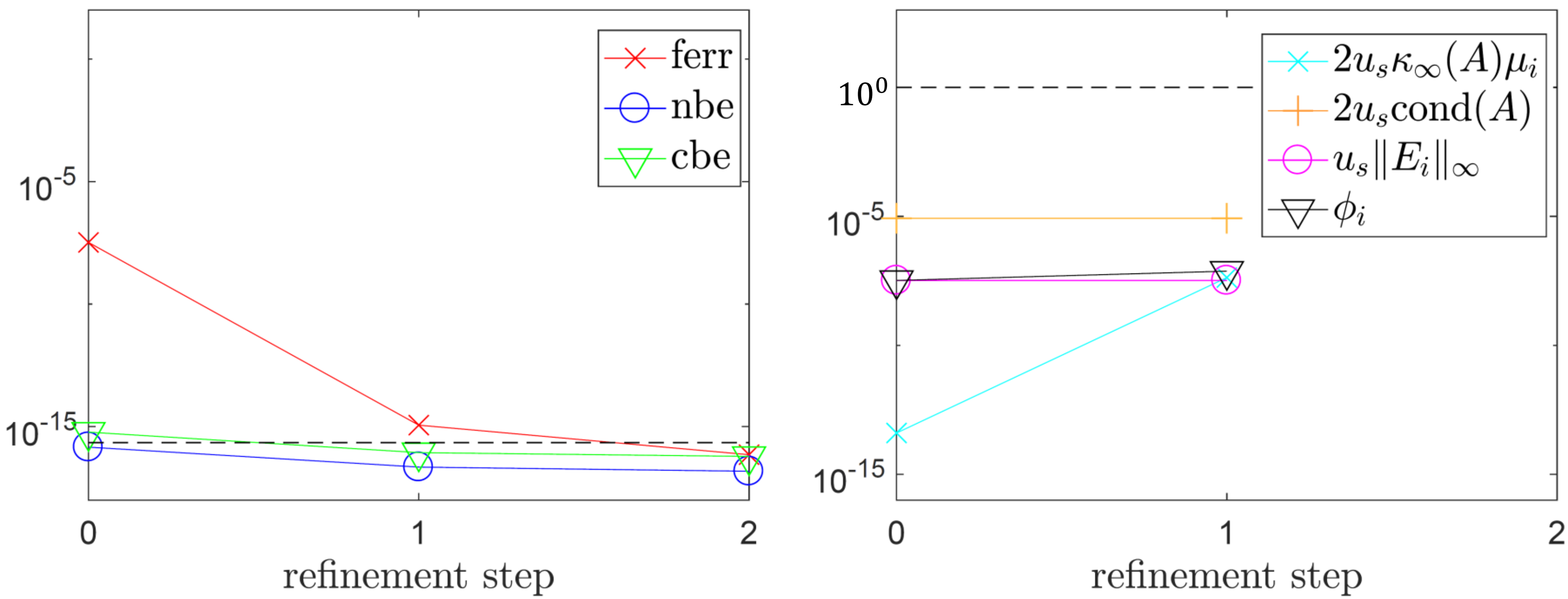
Standard (LU-based) IR with u_f: single, u: double, u_r: quad



```
A = gallery('randsvd', 100, 1e9, 2)
b = randn(100,1)
```

$\kappa_\infty(A) \approx$ 2e10, $\text{cond}(A,x) \approx$ 5e9

Standard (LU-based) IR with u_f: double, u: double, u_r: quad



GMRES-Based Iterative Refinement

- Observation [Rump, 1990]: if $\hat{L}$ and $\hat{U}$ are computed LU factors of A in precision u_f, then

$$\kappa_\infty\left(\hat{U}^{-1}\hat{L}^{-1}A\right) \approx 1 + \kappa_\infty(A)u_f,$$

even if $\kappa_\infty(A) \gg u_f^{-1}$.

GMRES-Based Iterative Refinement

- Observation [Rump, 1990]: if $\hat{L}$ and $\hat{U}$ are computed LU factors of A in precision u_f, then

$$\kappa_\infty\left(\hat{U}^{-1}\hat{L}^{-1}A\right) \approx 1 + \kappa_\infty(A)u_f,$$

even if $\kappa_\infty(A) \gg u_f^{-1}$.

<u>GMRES-IR</u> [C. and Higham, SISC 39(6), 2017]

- 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-Based Iterative Refinement

- Observation [Rump, 1990]: if $\hat{L}$ and $\hat{U}$ are computed LU factors of A in precision u_f, then

$$\kappa_\infty\left(\hat{U}^{-1}\hat{L}^{-1}A\right) \approx 1 + \kappa_\infty(A)u_f,$$

even if $\kappa_\infty(A) \gg u_f^{-1}$.

<u>GMRES-IR</u> [C. and Higham, SISC 39(6), 2017]

- To compute the updates d_i, apply GMRES to $\underbrace{\hat{U}^{-1}\hat{L}^{-1}A}_{\tilde{A}}d_i = \underbrace{\hat{U}^{-1}\hat{L}^{-1}r_i}_{\tilde{r}_i}$

Solve $Ax_0 = b$ by LU factorization

for $i = 0$: maxit

$r_i = b - Ax_i$

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

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

GMRES-Based Iterative Refinement

- Observation [Rump, 1990]: if $\hat{L}$ and $\hat{U}$ are computed LU factors of A in precision u_f, then

$$\kappa_\infty\left(\hat{U}^{-1}\hat{L}^{-1}A\right) \approx 1 + \kappa_\infty(A)u_f,$$

even if $\kappa_\infty(A) \gg u_f^{-1}$.

GMRES-IR [C. and Higham, SISC 39(6), 2017]

- To compute the updates d_i, apply GMRES to $\underbrace{\hat{U}^{-1}\hat{L}^{-1}A}_{\tilde{A}}d_i = \underbrace{\hat{U}^{-1}\hat{L}^{-1}r_i}_{\tilde{r}_i}$

Solve $Ax_0 = b$ by LU factorization

for $i = 0\colon \text{maxit}$

$r_i = b - Ax_i$

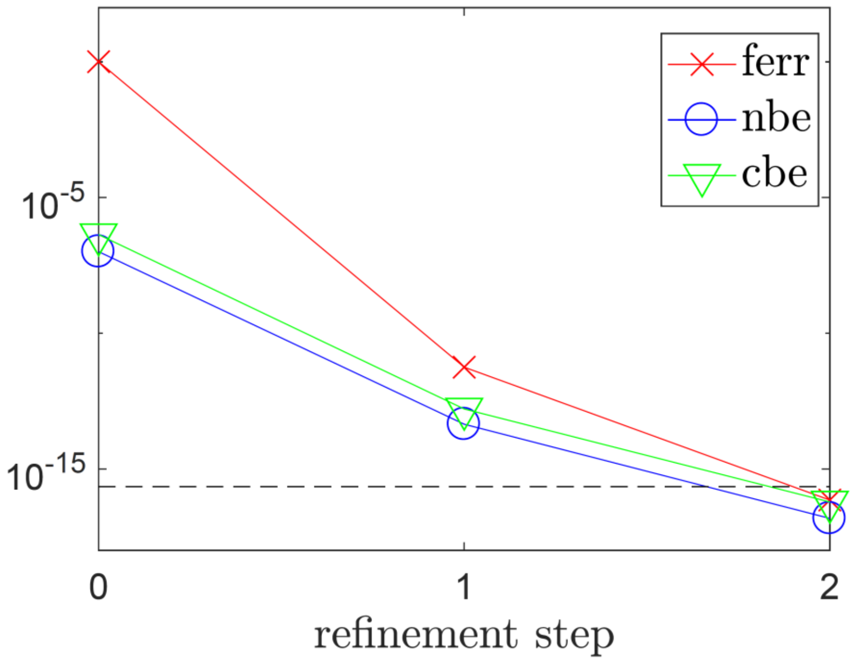
Solve $Ad_i = r_i$ via GMRES on $\tilde{A}d_i = \tilde{r}_i$ → $u_s = u$

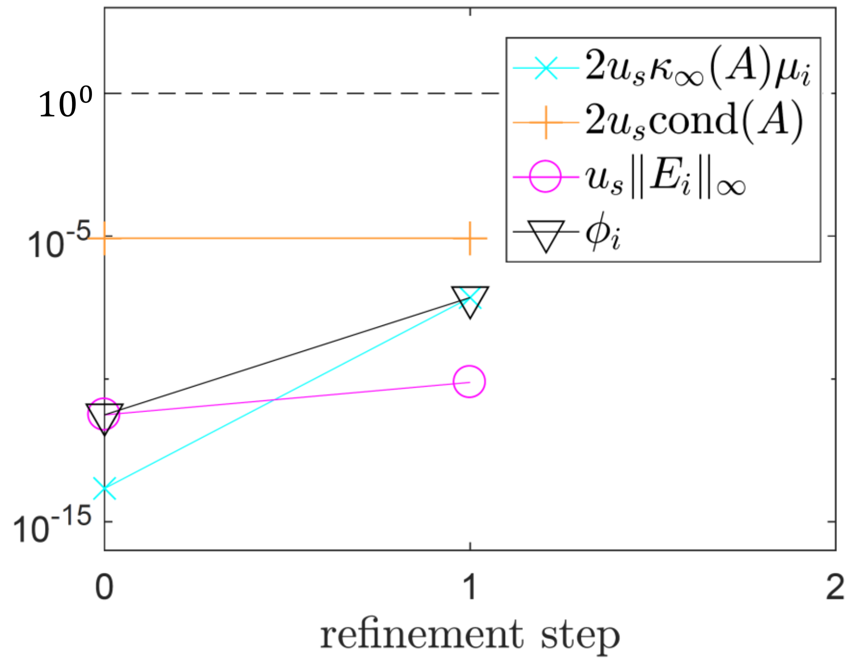
$x_{i+1} = x_i + d_i$

```
A = gallery('randsvd', 100, 1e9, 2)
b = randn(100,1)
```

$\kappa_\infty(A) \approx$ 2e10, $\mathrm{cond}(A,x) \approx$ 5e9, $\kappa_\infty(\tilde{A}) \approx$ 2e4

GMRES-IR with u_f: single, u: double, u_r: quad





GMRES-IR: Summary

Benefits of GMRES-IR:

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

GMRES-IR: Summary

Benefits of GMRES-IR:

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

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

GMRES-IR: Summary

Benefits of GMRES-IR:

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

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

$$\kappa_\infty(A) \leq u^{-1/2}\, u_f^{-1}$$

GMRES-IR: Summary

Benefits of GMRES-IR:

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

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

GMRES-IR: Summary

Benefits of GMRES-IR:

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

Try IR3! MATLAB codes available at: https://github.com/eccarson/ir3

Performance results? Stay tuned for the next talk by Azzam Haidar;
3-precision approach on NVIDIA V100

Comments and Caveats

- Convergence tolerance τ for GMRES?
  - Smaller $\tau \rightarrow$ more GMRES iterations, potentially fewer refinement steps
  - Larger $\tau \rightarrow$ fewer GMRES iterations, potentially more refinement steps

Comments and Caveats

- Convergence tolerance τ for GMRES?
  - Smaller $\tau \rightarrow$ more GMRES iterations, potentially fewer refinement steps
  - Larger $\tau \rightarrow$ fewer GMRES iterations, potentially more refinement steps
- Convergence rate of GMRES?

Comments and Caveats

- Convergence tolerance τ for GMRES?
  - Smaller $\tau \rightarrow$ more GMRES iterations, potentially fewer refinement steps
  - Larger $\tau \rightarrow$ fewer GMRES iterations, potentially more refinement steps
- Convergence rate of GMRES?
  - If A is ill conditioned and LU factorization is performed in very low precision, it can be a poor preconditioner
    - e.g., if $\tilde{A}$ still has cluster of eigenvalues near origin, GMRES can stagnate until n^{th} iteration, regardless of $\kappa_\infty(A)$ [Liesen and Tichý, 2004]

Comments and Caveats

- Convergence tolerance τ for GMRES?
    - Smaller $\tau \rightarrow$ more GMRES iterations, potentially fewer refinement steps
    - Larger $\tau \rightarrow$ fewer GMRES iterations, potentially more refinement steps

- Convergence rate of GMRES?
    - If A is ill conditioned and LU factorization is performed in very low precision, it can be a poor preconditioner
        - e.g., if $\tilde{A}$ still has cluster of eigenvalues near origin, GMRES can stagnate until n^{th} iteration, regardless of $\kappa_\infty(A)$ [Liesen and Tichý, 2004]
    - Potential remedies: deflation, Krylov subspace recycling, using additional preconditioner

Comments and Caveats

- Convergence tolerance τ for GMRES?
  - Smaller $\tau \rightarrow$ more GMRES iterations, potentially fewer refinement steps
  - Larger $\tau \rightarrow$ fewer GMRES iterations, potentially more refinement steps

- Convergence rate of GMRES?
  - If A is ill conditioned and LU factorization is performed in very low precision, it can be a poor preconditioner
    - e.g., if $\tilde{A}$ still has cluster of eigenvalues near origin, GMRES can stagnate until n^{th} iteration, regardless of $\kappa_\infty(A)$ [Liesen and Tichý, 2004]
  - Potential remedies: deflation, Krylov subspace recycling, using additional preconditioner

- Depending on conditioning of A, applying $\tilde{A}$ to a vector must be done accurately (precision u^2) in each GMRES iteration

Comments and Caveats

- Convergence tolerance τ for GMRES?
  - Smaller $\tau \rightarrow$ more GMRES iterations, potentially fewer refinement steps
  - Larger $\tau \rightarrow$ fewer GMRES iterations, potentially more refinement steps

- Convergence rate of GMRES?
  - If A is ill conditioned and LU factorization is performed in very low precision, it can be a poor preconditioner
    - e.g., if $\tilde{A}$ still has cluster of eigenvalues near origin, GMRES can stagnate until n^{th} iteration, regardless of $\kappa_\infty(A)$ [Liesen and Tichý, 2004]
  - Potential remedies: deflation, Krylov subspace recycling, using additional preconditioner

- Depending on conditioning of A, applying $\tilde{A}$ to a vector must be done accurately (precision u^2) in each GMRES iteration

- Why GMRES?
  - Theoretical purposes: existing analysis and proof of backward stability [Paige, Rozložník, Strakoš, 2006]
  - In practice, use any solver you want!

Thank You!

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