Iterative Refinement in Three Precisions

Erin C. Carson
Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University

Parallel Solution Methods for Systems Arising from PDEs
Centre International de Rencontres Mathématiques, Luminy, France
September 16-20, 2019

This research was supported by Charles University Primus program project No. PRIMUS/19/SCI/11.
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
- **Future exascale supercomputers**: (~2021) Expected extensive support for reduced-precision arithmetic (32/16/8-bit)
Performance of LU factorization on an NVIDIA V100 GPU

[Haidar, Tomov, Dongarra, Higham, 2018]
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 $i = 0$: 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 $i = 0$: 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 $Ax = b$

As long as $\kappa_\infty(A) \leq u^{-1}$,
- relative forward error is $O(u)$
- relative normwise and componentwise backward errors are $O(u)$

Solve $Ax_0 = b$ by LU factorization

for $i = 0$: maxit

$$r_i = b - Ax_i$$

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

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

"Traditional" (high-precision residual computation)

[Wilkinson, 1948] (fixed point), [Moler, 1967] (floating point)
Iterative Refinement for $Ax = b$

Solve $Ax_0 = b$ by LU factorization for $i = 0$: 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 $Ax = b$

As long as $\kappa_\infty(A) \leq u^{-1}$,
- relative forward error is $O(u)\text{cond}(A, x)$
- relative normwise and componentwise backward errors are $O(u)$

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

for $i = 0$: 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 $Ax = b$

Solve $Ax_0 = b$ by LU factorization

for $i = 0$: maxit

$$r_i = b - Ax_i$$  (in precision $u^{1/2}$)

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 $u$)

"Low-precision factorization"

[Langou et al., 2006], [Arioli and Duff, 2009], [Hogg and Scott, 2010], [Abdelfattah et al., 2016]
Iterative Refinement for $Ax = b$

As long as $\kappa_\infty(A) \leq u^{-1/2}$,
- relative forward error is $O(u)\text{cond}(A, x)$
- relative normwise and componentwise backward errors are $O(u)$

Solve $Ax_0 = b$ by LU factorization

for $i = 0$: maxit

\[ r_i = b - Ax_i \]  
\[ \text{Solve } Ad_i = r_i \quad \text{via } d_i = U^{-1}(L^{-1}r_i) \]  
\[ x_{i+1} = x_i + d_i \]

"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?

⇒ 3-precision iterative refinement

\( u_f = \text{factorization precision}, \quad u = \text{working precision}, \quad u_r = \text{residual precision} \)

\[ u_f \geq u \geq u_r \]
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?

⇒ 3-precision iterative refinement

\[ u_f = \text{factorization precision}, \quad u = \text{working precision}, \quad u_r = \text{residual precision} \]

\[ u_f \geq u \geq u_r \]

• New analysis generalizes existing types of IR:

<table>
<thead>
<tr>
<th>Type</th>
<th>( u_f = u, u_r = u^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>( u_f = u, u_r = u^2 )</td>
</tr>
<tr>
<td>Fixed precision</td>
<td>( u_f = u = u_r )</td>
</tr>
<tr>
<td>Lower precision factorization</td>
<td>( u_f^2 = u = u_r )</td>
</tr>
</tbody>
</table>

(And improves upon existing analyses in some cases)

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

⇒ 3-precision iterative refinement

\[ u_f = \text{factorization precision}, \quad u = \text{working precision}, \quad u_r = \text{residual precision} \]

\[ u_f \geq u \geq u_r \]

• New analysis generalizes existing types of IR:

<table>
<thead>
<tr>
<th>Type</th>
<th>( u_f )</th>
<th>( u_r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional</td>
<td>( u_f = u, u_r = u^2 )</td>
<td></td>
</tr>
<tr>
<td>Fixed precision</td>
<td>( u_f = u = u_r )</td>
<td></td>
</tr>
<tr>
<td>Lower precision factorization</td>
<td>( u_f^2 = u = u_r )</td>
<td></td>
</tr>
</tbody>
</table>

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

(and improves upon existing analyses in some cases)

• Enables new types of IR: (half, single, double), (half, single, quad), (half, double, quad), etc.
Obtain tighter upper bounds:

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

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

Define \( \mu_i : \|A(x - \hat{x}_i)\|_\infty = \mu_i \|A\|_\infty \|x - \hat{x}_i\|_\infty \)
Obtain tighter upper bounds:

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

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

For a stable refinement scheme, in early stages we expect

\[
\frac{\|r_i\|_\infty}{\|A\|_\infty \|\hat{x}_i\|_\infty} \approx u \ll \frac{\|x - \hat{x}_i\|_\infty}{\|x\|_\infty} \quad \Rightarrow \quad \mu_i \ll 1
\]
Obtain tighter upper bounds:

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

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

For a stable refinement scheme, in early stages we expect

\[
\frac{\|r_i\|}{\|A\|\|\hat{x}_i\|} \approx u \ll \frac{\|x - \hat{x}_i\|}{\|x\|} \Rightarrow \mu_i \ll 1
\]

But close to convergence,

\[
\|r_i\| \approx \|A\|\|x - \hat{x}_i\| \Rightarrow \mu_i \approx 1
\]
Allow for general solver:
Let $u_s$ be the effective precision of the solve, with $u \leq u_s \leq u_f$
Allow for general solver:
Let $u_s$ be the *effective precision* of the solve, with $u \leq u_s \leq u_f$

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

1. $\hat{d}_i = (I + u_s E_i) d_i, \quad u_s \|E_i\|_\infty < 1$
   → normwise relative forward error is bounded by multiple of $u_s$ and is less than 1
Allow for general solver:
Let $u_s$ be the effective precision of the solve, with $u \leq u_s \leq u_f$

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

1. $\hat{d}_i = (I + u_s E_i) d_i$, $u_s \|E_i\|_\infty < 1$
   $\rightarrow$ normwise relative forward error is bounded by multiple of $u_s$ and is less than 1
Allow for general solver:
Let $u_s$ be the effective precision of the solve, with $u \leq u_s \leq u_f$

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

1. $\hat{d}_i = (I + u_s E_i)d_i$, $u_s \|E_i\|_\infty < 1$
   
   → normwise relative forward error is bounded by multiple of $u_s$ and is less than 1

2. $\|\hat{r}_i - A\hat{d}_i\|_\infty \leq u_s(c_1 \|A\|_\infty \|\hat{d}_i\|_\infty + c_2 \|\hat{r}_i\|_\infty)$
   
   → normwise relative backward error is at most $\max(c_1, c_2) u_s$

example: LU solve:

$u_s \|E_i\|_\infty \leq 3n u_f \|A^{-1}\| \|\hat{L}\| \|\hat{U}\|_\infty$
Allow for general solver:
Let \( u_s \) be the effective precision of the solve, with \( u \leq u_s \leq u_f \)

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

1. \( \hat{d}_i = (I + u_sE_i)d_i, \quad u_s\|E_i\|_\infty < 1 \)
   \( \rightarrow \) normwise relative forward error is bounded by multiple of \( u_s \) and is less than 1

2. \( \|\hat{r}_i - A\hat{d}_i\|_\infty \leq u_s(c_1\|A\|_\infty\|\hat{d}_i\|_\infty + c_2\|\hat{r}_i\|_\infty) \)
   \( \rightarrow \) normwise relative backward error is at most \( \max(c_1, c_2) u_s \)

**Example:** LU solve:

\[
u_s\|E_i\|_\infty \leq 3nuf \||A^{-1}||\hat{L}||\hat{U}||_\infty\]

\[
\text{max}(c_1, c_2) u_s \leq \frac{3nuf\||\hat{L}||\hat{U}||_\infty}{\|A\|_\infty}
\]
Allow for general solver:
Let $u_s$ be the effective precision of the solve, with $u \leq u_s \leq u_f$

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

1. $\hat{d}_i = (I + u_sE_i)d_i, \quad u_s\|E_i\|_\infty < 1$
   \quad $\rightarrow$ normwise relative forward error is bounded by multiple of $u_s$ and is less than 1

2. $\|\hat{r}_i - A\hat{d}_i\|_\infty \leq u_s(c_1\|A\|_\infty\|\hat{d}_i\|_\infty + c_2\|\hat{r}_i\|_\infty)$
   \quad $\rightarrow$ normwise relative backward error is at most $\max(c_1, c_2) u_s$

3. $|\hat{r}_i - A\hat{d}_i| \leq u_s G_i |\hat{d}_i|$
   \quad $\rightarrow$ componentwise relative backward error is bounded by a multiple of $u_s$

$E_i, c_1, c_2,$ and $G_i$ depend on $A, \hat{r}_i, n,$ and $u_s$
Allow for general solver:

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

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

1. $\hat{d}_i = (I + u_s E_i) d_i$, $u_s \|E_i\|_\infty < 1$
   $\rightarrow$ normwise relative forward error is bounded by multiple of $u_s$ and is less than 1

2. $\|\hat{r}_i - A\hat{d}_i\|_\infty \leq u_s (c_1 \|A\|_\infty \|\hat{d}_i\|_\infty + c_2 \|\hat{r}_i\|_\infty)$
   $\rightarrow$ normwise relative backward error is at most $\max(c_1, c_2) u_s$

3. $|\hat{r}_i - A\hat{d}_i| \leq u_s G_i |\hat{d}_i|$
   $\rightarrow$ componentwise relative backward error is bounded by a multiple of $u_s$

$E_i, c_1, c_2, \text{ and } G_i$ depend on $A$, $\hat{r}_i$, $n$, and $u_s$
Allow for general solver:
Let $u_s$ be the effective precision of the solve, with $u \leq u_s \leq u_f$

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

1. $\hat{d}_i = (I + u_s E_i)d_i$, $u_s \|E_i\|_\infty < 1$
   → normwise relative forward error is bounded by multiple of $u_s$ and is less than 1

2. $\|\hat{r}_i - A\hat{d}_i\|_\infty \leq u_s (c_1 \|A\|_\infty \|\hat{d}_i\|_\infty + c_2 \|\hat{r}_i\|_\infty)$
   → normwise relative backward error is at most $\max(c_1, c_2) u_s$

3. $|\hat{r}_i - A\hat{d}_i| \leq u_s G_i |\hat{d}_i|$
   → componentwise relative backward error is bounded by a multiple of $u_s$

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

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

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

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

$$
\frac{\|x - \hat{x}_i\|_\infty}{\|x\|_\infty} \leq 4Nu_r \text{cond}(A, x) + u,
$$

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

\[\begin{align*}
\kappa_\infty(A) &= \|A^{-1}\|_\infty \|A\|_\infty \\
\text{cond}(A) &= \| |A^{-1}| |A| \|_\infty \\
\text{cond}(A, x) &= \| |A^{-1}| |A||x| \|_\infty / \|x\|_\infty
\end{align*}\]
Forward Error for IR3

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

Analogous traditional bounds:

$$\phi_i \equiv 3n u_f \kappa_\infty(A)$$

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

\[\text{cond}(A) = \| |A^{-1}| |A| \|_\infty\]

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

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

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

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

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

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

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

Analogous traditional bounds: $\phi_i \equiv 3n u_f \kappa_\infty(A)$
Normwise Backward Error for IR3

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

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

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

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

$$\|b - A\hat{x}_i\|_\infty \approx Nu(\|b\|_\infty + \|A\|_\infty \|\hat{x}_i\|_\infty),$$

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 $\approx 10^{-34}$

<table>
<thead>
<tr>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
<th>$\max \kappa_\infty(A)$</th>
<th>Backward error</th>
<th>Forward error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>norm</td>
<td>comp</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>S</td>
<td>S</td>
<td>10^4</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>H</td>
<td>S</td>
<td>D</td>
<td>10^4</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>H</td>
<td>D</td>
<td>D</td>
<td>10^4</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>10^4</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>S</td>
<td>10^8</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>D</td>
<td>10^8</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>S</td>
<td>D</td>
<td>D</td>
<td>10^8</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>10^8</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
</tbody>
</table>
**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 \(\approx 10^{-34}\)

<table>
<thead>
<tr>
<th>(u_f)</th>
<th>(u)</th>
<th>(u_r)</th>
<th>(\max \kappa_{\infty}(A))</th>
<th>Backward error</th>
<th>Forward error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>norm comp</td>
<td></td>
<td>cond((A, x)) \cdot 10^{-8}</td>
</tr>
<tr>
<td>H</td>
<td>S</td>
<td>S</td>
<td>10^4 10^{-8} 10^{-8}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>S</td>
<td>D</td>
<td>10^4 10^{-8} 10^{-8}</td>
<td></td>
<td>10^{-8}</td>
</tr>
<tr>
<td>H</td>
<td>D</td>
<td>D</td>
<td>10^4 10^{-16} 10^{-16}</td>
<td></td>
<td>cond((A, x)) \cdot 10^{-16}</td>
</tr>
<tr>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>10^4 10^{-16} 10^{-16}</td>
<td></td>
<td>10^{-16}</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>S</td>
<td>10^8 10^{-8} 10^{-8}</td>
<td></td>
<td>cond((A, x)) \cdot 10^{-8}</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>D</td>
<td>10^8 10^{-8} 10^{-8}</td>
<td></td>
<td>10^{-8}</td>
</tr>
<tr>
<td>S</td>
<td>D</td>
<td>D</td>
<td>10^8 10^{-16} 10^{-16}</td>
<td></td>
<td>cond((A, x)) \cdot 10^{-16}</td>
</tr>
<tr>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>10^8 10^{-16} 10^{-16}</td>
<td></td>
<td>10^{-16}</td>
</tr>
</tbody>
</table>
### Standard (LU-based) IR in three precisions ($u_S = u_F$)

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

<table>
<thead>
<tr>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
<th>max $\kappa_\infty(A)$</th>
<th>Backward error</th>
<th>Forward error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>norm</td>
<td>comp</td>
</tr>
<tr>
<td>LP fact.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>S</td>
<td>S</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>LP fact.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>D</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>S</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>LP fact.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>D</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
</tbody>
</table>
### 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 $\approx 10^{-34}$

<table>
<thead>
<tr>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
<th>max $\kappa_\infty(A)$</th>
<th>Backward error</th>
<th>Forward error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>norm</td>
<td>comp</td>
</tr>
<tr>
<td><strong>LP fact.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>S</td>
<td>S</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td><strong>LP fact.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>D</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td><strong>Fixed</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>S</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td><strong>Trad.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td><strong>LP fact.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>D</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
</tbody>
</table>
Standard (LU-based) IR in three precisions ($u_s = u_f$)

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

<table>
<thead>
<tr>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
<th>$\max k_\infty(A)$</th>
<th>Backward error</th>
<th>Forward error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>norm</td>
<td>comp</td>
</tr>
<tr>
<td>LP fact.</td>
<td>New</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>S</td>
<td>S</td>
<td>10^4</td>
<td>10^{-8}</td>
<td>10^{-8}</td>
</tr>
<tr>
<td>LP fact.</td>
<td>New</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>D</td>
<td>D</td>
<td>10^4</td>
<td>10^{-16}</td>
<td>10^{-16}</td>
</tr>
<tr>
<td>Fixed</td>
<td>Trad.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>S</td>
<td>S</td>
<td>10^8</td>
<td>10^{-8}</td>
<td>10^{-8}</td>
</tr>
<tr>
<td>LP fact.</td>
<td>New</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>D</td>
<td>D</td>
<td>10^8</td>
<td>10^{-16}</td>
<td>10^{-16}</td>
</tr>
<tr>
<td>New</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>10^8</td>
<td>10^{-16}</td>
<td>10^{-16}</td>
</tr>
</tbody>
</table>
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 $\approx 10^{-34}$

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

$\Rightarrow$ 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 $\approx 10^{-34}$

<table>
<thead>
<tr>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
<th>$\max \kappa_\infty(A)$</th>
<th>Backward error $|\cdot|$</th>
<th>Forward error $|\cdot|$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>norm</td>
<td>comp</td>
</tr>
<tr>
<td>LP fact.</td>
<td>H</td>
<td>S</td>
<td>S</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>New</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>LP fact.</td>
<td>H</td>
<td>D</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>New</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>Fixed</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>Trad.</td>
<td>S</td>
<td>S</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>LP fact.</td>
<td>S</td>
<td>D</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>New</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
</tr>
</tbody>
</table>

$\Rightarrow$ 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)

\( \kappa_\infty(A) \approx 2e10, \; \text{cond}(A, x) \approx 5e9 \)
A = gallery('randsvd', 100, 1e9, 2)
b = randn(100,1)

\[ \kappa_\infty(A) \approx 2e10, \quad \text{cond}(A, x) \approx 5e9 \]
A = gallery('randsvd', 100, 1e9, 2)
b = randn(100,1)

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

Standard (LU-based) IR with \( u_f \): single, \( u \): double, \( u_r \): quad
A = gallery('randsvd', 100, 1e9, 2)
b = randn(100,1)

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

Standard (LU-based) IR with $u_f$: double, $u$: double, $u_r$: quad
• Observation [Rump, 1990]: if $\hat{L}$ and $\hat{U}$ are computed LU factors of $A$ in precision $u_f$, then

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

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

\[
\kappa_\infty(\hat{U}^{-1}\hat{L}^{-1}A) \approx 1 + \kappa_\infty(A)u_f,
\]

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

\[
\hat{U}^{-1}\hat{L}^{-1}d_i = \hat{U}^{-1}\hat{L}^{-1}\bar{r}_i
\]
GMRES-Based Iterative Refinement

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

\[
\kappa_\infty(\hat{U}^{-1}\hat{L}^{-1}A) \approx 1 + \kappa_\infty(A)u_f,
\]

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 \( \hat{U}^{-1}\hat{L}^{-1}Ad_i = \hat{U}^{-1}\hat{L}^{-1}r_i \)

Solve \( Ax_0 = b \) by LU factorization

for \( i = 0 \): maxit

\[
\begin{align*}
    r_i &= b - Ax_i \\
    \text{Solve } Ad_i &= r_i & \text{via GMRES on } \tilde{A}d_i = \tilde{r}_i \\
    x_{i+1} &= x_i + d_i
\end{align*}
\]
GMRES-Based Iterative Refinement

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

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

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

$$\hat{U}^{-1}\hat{L}^{-1}Ad_i = \hat{U}^{-1}\hat{L}^{-1}r_i$$

Solve $Ax_0 = b$ by LU factorization

for $i = 0$: maxit

$$r_i = b - Ax_i$$

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

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

$$u_s = u$$
A = gallery('randsvd', 100, 1e9, 2)
b = randn(100,1)

\[ \kappa_\infty(A) \approx 2e10, \quad \text{cond}(A,x) \approx 5e9 \]
\( A = \text{gallery('randsvd', 100, 1e9, 2)} \)
\( b = \text{randn}(100,1) \)

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

**GMRES-IR** with \( u_f \): single, \( u \): double, \( u_r \): quad
## GMRES-IR: Summary

### Benefits of GMRES-IR:

<table>
<thead>
<tr>
<th></th>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
<th>max $\kappa_\infty(A)$</th>
<th>Backward error</th>
<th>Forward error</th>
</tr>
</thead>
<tbody>
<tr>
<td>LU-IR</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>LU-IR</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^{16}$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>LU-IR</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^{12}$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
</tbody>
</table>
GMRES-IR: Summary

Benefits of GMRES-IR:

<table>
<thead>
<tr>
<th></th>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
<th>max $\kappa_\infty(A)$</th>
<th>Backward error</th>
<th>Forward error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>max $\kappa_\infty(A)$</td>
<td>norm</td>
<td>comp</td>
</tr>
<tr>
<td>LU-IR</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>LU-IR</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^{16}$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>LU-IR</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^{12}$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
</tbody>
</table>

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

**Benefits of GMRES-IR:**

<table>
<thead>
<tr>
<th></th>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
<th>$\max \kappa_\infty(A)$</th>
<th>Backward error</th>
<th>Forward error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>norm</td>
<td>comp</td>
</tr>
<tr>
<td>LU-IR</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>LU-IR</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^{16}$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>LU-IR</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^{12}$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
</tbody>
</table>

⇒ 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:

<table>
<thead>
<tr>
<th></th>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
<th>max $\kappa_\infty(A)$</th>
<th>Backward error</th>
<th>Forward error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>norm</td>
<td>comp</td>
</tr>
<tr>
<td>LU-IR</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>LU-IR</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^{16}$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>LU-IR</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^{12}$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
</tbody>
</table>

⇒ If $\kappa_\infty(A) \leq 10^{12}$, can use lower precision factorization w/no loss of accuracy!
## GMRES-IR: Summary

Benefits of GMRES-IR:

<table>
<thead>
<tr>
<th></th>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
<th>$\max \kappa_\infty(A)$</th>
<th>Backward error</th>
<th>Forward error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>norm</td>
<td>comp</td>
</tr>
<tr>
<td>LU-IR</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^4$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>H</td>
<td>S</td>
<td>D</td>
<td>$10^8$</td>
<td>$10^{-8}$</td>
<td>$10^{-8}$</td>
</tr>
<tr>
<td>LU-IR</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^8$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>S</td>
<td>D</td>
<td>Q</td>
<td>$10^{16}$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>LU-IR</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^4$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
<tr>
<td>GMRES-IR</td>
<td>H</td>
<td>D</td>
<td>Q</td>
<td>$10^{12}$</td>
<td>$10^{-16}$</td>
<td>$10^{-16}$</td>
</tr>
</tbody>
</table>

Try IR3! MATLAB codes available at: [https://github.com/eccarson/ir3](https://github.com/eccarson/ir3)
Comments and Caveats

• Convergence tolerance $\tau$ for GMRES?
  • Smaller $\tau \rightarrow$ more GMRES iterations, potentially fewer refinement steps
  • Larger $\tau \rightarrow$ fewer GMRES iterations, potentially more refinement steps
• Convergence tolerance $\tau$ 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 $\tau$ 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 $\tau$ 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 $\tau$ 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 $\tau$ 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!
• Want to solve

$$\min_x \| b - Ax \|_2$$

where $A \in \mathbb{R}^{m \times n}$ ($m > n$) has rank $n$

• Commonly solved using QR factorization:

$$A = QR = [Q_1, Q_2] \begin{bmatrix} U \\ 0 \end{bmatrix}$$

where $Q$ is an $m \times m$ orthogonal matrix and $U$ is upper triangular.

$$x = U^{-1}Q_1^Tb, \quad \|b - Ax\|_2 = \|Q_2^Tb\|_2$$
• Want to solve
\[ \min_x \|b - Ax\|_2 \]
where \( A \in \mathbb{R}^{m \times n} \) (\( m > n \)) has rank \( n \)

• Commonly solved using QR factorization:
\[ A = QR = [Q_1, Q_2] \begin{bmatrix} U \\ 0 \end{bmatrix} \]
where \( Q \) is an \( m \times m \) orthogonal matrix and \( U \) is upper triangular.
\[ x = U^{-1}Q_1^Tb, \quad \|b - Ax\|_2 = \|Q_2^Tb\|_2 \]

• As in linear system case, for ill-conditioned problems, iterative refinement often needed to improve accuracy and stability
Extension to Least Squares Problems

• (Björck, 1967): Least squares problem can be written as a linear system with square matrix of size \((m + n)\):

\[
\begin{bmatrix}
I & A \\
A^T & 0
\end{bmatrix}
\begin{bmatrix}
x \\
r
\end{bmatrix}
= 
\begin{bmatrix}
b \\
0
\end{bmatrix}
\]

\(\tilde{A}\tilde{x} = \tilde{b}\)

• Results for linear systems also apply to least squares problems
Extension to Least Squares Problems

• (Björck, 1967): Least squares problem can be written as a linear system with square matrix of size \((m + n)\):

\[
\begin{bmatrix}
I & A \\
A^T & 0
\end{bmatrix}
\begin{bmatrix}
x \\
r
\end{bmatrix}
=
\begin{bmatrix}
b \\
0
\end{bmatrix}
\]

\[
\tilde{A}\tilde{x} = \tilde{b}
\]

• Results for linear systems also apply to least squares problems

• Extension of GMRES-based IR for least squares
Extension to Least Squares Problems

• (Björck, 1967): Least squares problem can be written as a linear system with square matrix of size $(m + n)$:

\[
\begin{bmatrix}
I & A \\
A^T & 0
\end{bmatrix}
\begin{bmatrix}
x \\
r
\end{bmatrix}
=
\begin{bmatrix}
b \\
0
\end{bmatrix}
\]

\[\tilde{A} \tilde{x} = \tilde{b}\]

• Results for linear systems also apply to least squares problems

• Extension of GMRES-based IR for least squares
  • Don’t want to compute LU of $\tilde{A}$...
  • How to define a good preconditioner using QR factors?
    • Constraints: left-preconditioning, $\kappa_\infty(M^{-1}\tilde{A})$ is reasonably small
• (Björck, 1967): Least squares problem can be written as a linear system with square matrix of size $(m + n)$:

$$
\begin{bmatrix}
I & A \\
A^T & 0
\end{bmatrix}
\begin{bmatrix}
x \\
r
\end{bmatrix}
= 
\begin{bmatrix}
b \\
0
\end{bmatrix}
\tilde{A}\tilde{x} = \tilde{b}
$$

• Results for linear systems also apply to least squares problems

• Extension of GMRES-based IR for least squares
  • Don't want to compute LU of $\tilde{A}$...
  • How to define a good preconditioner using QR factors?
    • Constraints: left-preconditioning, $\kappa_\infty(M^{-1}\tilde{A})$ is reasonably small
    • Many possibilities...requirements of theory vs. what works in practice
GMRES-IR for Least Squares

- Ex: block diagonal preconditioner ([Murphy, Golub, Wathen, 2000], [Ipsen, 2001])

\[
\begin{bmatrix}
\alpha I & 0 \\
0 & \frac{1}{\alpha} \hat{R}^T \hat{R}
\end{bmatrix} = \begin{bmatrix}
\sqrt{\alpha} I & 0 \\
0 & \frac{1}{\sqrt{\alpha}} \hat{R}^T
\end{bmatrix} \begin{bmatrix}
\sqrt{\alpha} I & 0 \\
0 & \frac{1}{\sqrt{\alpha}} \hat{R}
\end{bmatrix} \equiv M_1 M_2
\]
GMRES-IR for Least Squares

• Ex: block diagonal preconditioner ([Murphy, Golub, Wathen, 2000], [Ipsen, 2001])

\[
\begin{bmatrix}
\alpha I & 0 \\
0 & \frac{1}{\alpha} \hat{R}^T \hat{R}
\end{bmatrix} = \begin{bmatrix}
\sqrt{\alpha} I & 0 \\
0 & \frac{1}{\sqrt{\alpha}} \hat{R}^T
\end{bmatrix} \begin{bmatrix}
\sqrt{\alpha} I & 0 \\
0 & \frac{1}{\sqrt{\alpha}} \hat{R}
\end{bmatrix} \equiv M_1 M_2
\]

• Assuming QR factorization is exact,

\[
M_2^{-1} M_1^{-1} \tilde{A} = \begin{bmatrix}
I & \frac{1}{\alpha} A \\
\alpha \hat{R}^{-1} \hat{R}^{-T} A^T & 0
\end{bmatrix}
\]

is nonsymmetric, diagonalizable, with eigenvalues \( \left\{ 1, \frac{1}{2} (1 \pm \sqrt{5}) \right\} \).

• However, condition number can still be quite large; unsuitable for proving backward stability of GMRES
GMRES-IR for Least Squares

- Ex: block diagonal preconditioner ([Murphy, Golub, Wathen, 2000], [Ipsen, 2001])
  \[
  \begin{bmatrix}
  \alpha I & 0 \\
  0 & \frac{1}{\alpha} \hat{R}^T \hat{R}
  \end{bmatrix}
  = 
  \begin{bmatrix}
  \sqrt{\alpha} I & 0 \\
  0 & \frac{1}{\sqrt{\alpha}} \hat{R}^T
  \end{bmatrix}
  \begin{bmatrix}
  \sqrt{\alpha} I & 0 \\
  0 & \frac{1}{\sqrt{\alpha}} \hat{R}
  \end{bmatrix}
  \equiv M_1 M_2
  \]

- Assuming QR factorization is exact,
  \[
  M_2^{-1} M_1^{-1} \tilde{A} = \begin{bmatrix}
  I & \frac{1}{\alpha} A \\
  \alpha \hat{R}^{-1} \hat{R}^T A^T & 0
  \end{bmatrix}
  \]

  is nonsymmetric, diagonalizable, with eigenvalues \( \{1, \frac{1}{2} (1 \pm \sqrt{5})\} \).
  - However, condition number can still be quite large; unsuitable for proving backward stability of GMRES

- If we take split preconditioner
  \[
  \begin{bmatrix}
  I & A \hat{R}
  \end{bmatrix}
  \]
  we will have a well-conditioned system
  - However, split-preconditioned GMRES is not backward stable
  - Potentially useful in practice, not but in theory
GMRES-IR for Least Squares

- One option:

\[ M = \begin{bmatrix} \alpha I & \hat{Q}_1 \hat{R} \\ \hat{R}^T \hat{Q}_1^T & 0 \end{bmatrix} \]

- Then we can prove that for the left-preconditioned system,

\[ \kappa(M^{-1} \tilde{A}) \leq \left( 1 + u_f c \kappa(A) \right)^2 \]

where \( c = O(m^{7/2}) \), where we note this bound is pessimistic.

- Thus even if \( \kappa(A) \gg u_f^{-1} \), the preconditioned system can still be reasonably well conditioned
GMRES-IR for Least Squares

• One option:

\[ M = \begin{bmatrix} \alpha I & \hat{Q}_1 \hat{R} \\ \hat{R}^T \hat{Q}_1^T & 0 \end{bmatrix} \]

• Then we can prove that for the left-preconditioned system,

\[ \kappa(M^{-1} \tilde{A}) \leq \left(1 + u_f c \kappa(A)\right)^2 \]

where \( c = O(m^{7/2}) \), where we note this bound is pessimistic.

• Thus even if \( \kappa(A) \gg u_f^{-1} \), the preconditioned system can still be reasonably well conditioned.

• GMRES run on \( \tilde{A} \) with left-preconditioner \( M \) gives

\[ u_s \|E_i\|_\infty \equiv u f (m + n) \kappa_\infty (M^{-1} \tilde{A}) \]

where \( f \) is a quadratic polynomial.
GMRES-IR for Least Squares

• One option:

\[ M = \begin{bmatrix} \alpha I & \hat{Q}_1\hat{R} \\ \hat{R}^T\hat{Q}_1^T & 0 \end{bmatrix} \]

• Then we can prove that for the left-preconditioned system,

\[ \kappa(M^{-1}\tilde{A}) \leq \left( 1 + u_f c \kappa(A) \right)^2 \]

where \( c = O(m^{7/2}) \), where we note this bound is pessimistic.

• Thus even if \( \kappa(A) \gg u_f^{-1} \), the preconditioned system can still be reasonably well conditioned.

• GMRES run on \( \tilde{A} \) with left-preconditioner \( M \) gives

\[ u_s \|E_i\|_\infty \equiv u f (m + n)\kappa_\infty(M^{-1}\tilde{A}) \]

where \( f \) is a quadratic polynomial.

• So for GMRES-based LSIR, \( u_s \equiv u \); expect convergence of forward error when \( \kappa_\infty(A) < u^{-1/2} u_f^{-1} \)
gallery('randsvd', [100,10], kappa(i), 3)
QR factorization computed in half precision; preconditioned system computed exactly
A = gallery('randsvd', [100, 10], kappa, 3)
b = randn(100,1); b = b./norm(b)

GMRES-LSIR and "Standard" LSIR with
\( \mathbf{u}_f \): half, \( \mathbf{u} \): single, \( \mathbf{u}_r \): double

\( \kappa = 1 \times 10^3 \)
\begin{verbatim}
A = gallery('randsvd', [100, 10], kappa, 3)
b = randn(100,1); b = b./norm(b)
\end{verbatim}

GMRES-LSIR and "Standard" LSIR with
\( \mathbf{u}_f \): half, \( \mathbf{u} \): single, \( \mathbf{u}_r \): double

\( \kappa = 1e+04 \)
GMRES-LSIR and "Standard" LSIR with

\( \mathbf{u}_f \): half, \( \mathbf{u} \): single, \( \mathbf{u}_r \): double

\( \kappa = 1e+06 \)

\[
A = \text{gallery}(\text{'randsvd'}, \begin{bmatrix} 100, 10 \end{bmatrix}, \kappaappa, 3)
\]
\[
b = \text{randn}(100,1); \ b = b./\text{norm}(b)
\]
The rise of multiprecision hardware

• Future machines will support a range of precisions: quarter, half, single, double, quad
The rise of multiprecision hardware

• Future machines will support a range of precisions: quarter, half, single, double, quad

• New, non-IEEE compliant floating point formats will appear in commercially-available hardware
  • e.g., bfloat16 (truncated 16-bit version of single precision)
The rise of multiprecision hardware

• Future machines will support a range of precisions: quarter, half, single, double, quad

• New, non-IEEE compliant floating point formats will appear in commercially-available hardware
  • e.g., bfloat16 (truncated 16-bit version of single precision)

• Lower-precision arithmetic is faster and more energy efficient, but the potential for its use depends heavily on the particular problem and algorithm
The rise of multiprecision hardware

- Future machines will support a range of precisions: quarter, half, single, double, quad

- New, non-IEEE compliant floating point formats will appear in commercially-available hardware
  - e.g., bfloat16 (truncated 16-bit version of single precision)

- Lower-precision arithmetic is faster and more energy efficient, but the potential for its use depends heavily on the particular problem and algorithm

- As numerical analysts, we must determine when and where we can exploit lower-precision hardware to improve performance
Thank You!
carson@karlin.mff.cuni.cz
www.karlin.mff.cuni.cz/~carson/