Squeezing a Matrix Into Half Precision, with an Application to Solving Linear Systems

Srikara Pranesh
School of Mathematics
The University of Manchester

srikara.pranesh@manchester.ac.uk

29-05-2019

Joint work with Prof. Nick Higham and Mawussi Zounon
Motivation

Low precision floating-point formats are increasingly supported by computer hardware.

<table>
<thead>
<tr>
<th></th>
<th>( u )</th>
<th>( x_{\min}^S )</th>
<th>( x_{\min} )</th>
<th>( x_{\max} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>bfloat16</td>
<td>( 3.91 \times 10^{-3} )</td>
<td>( 9.18 \times 10^{-41} )</td>
<td>( 1.18 \times 10^{-38} )</td>
<td>( 3.39 \times 10^{38} )</td>
</tr>
<tr>
<td>fp16</td>
<td>( 4.88 \times 10^{-4} )</td>
<td>( 5.96 \times 10^{-8} )</td>
<td>( 6.10 \times 10^{-5} )</td>
<td>( 6.55 \times 10^{4} )</td>
</tr>
</tbody>
</table>

- **fp16** –
  - Current – NVIDIA since P100, AMD M125 GPU.
  - Future – Fujitsu A64FX Arm processor, IBM.

- **bfloat16** –
  - Current – Google TPU.
  - Future – Intel Nervana Neural Network Processor, Intel Cooper Lake.
Main driver for these new generation of architectures is machine learning.

- Applications in scientific computing
  - In climate science to resolve low scale features. Tim Palmer et al.

Applications in scientific computing

In climate science to resolve low scale features. Tim Palmer et al.
Main driver for these new generation of architectures is machine learning.

- Applications in scientific computing
  - In climate science to resolve low scale features. Tim Palmer et.al
  - In numerical linear algebra.
    - For the solution of linear systems GMRES-based Iterative refinement (GMRES-IR). (Carson and Higham 2018).
Main driver for these new generation of architectures is machine learning.

- Applications in scientific computing
  - In climate science to resolve low scale features. Tim Palmer et.al
  - In numerical linear algebra.
    - For the solution of linear systems GMRES-based Iterative refinement (GMRES-IR). (Carson and Higham 2018).

Part of a broader picture in the context of algorithms for extreme scale computing. J. Dongarra et.al. classify these multi precision algorithms as ‘Responsibly Reckless’.
Main driver for these new generation of architectures is machine learning.

- Applications in scientific computing
  - In climate science to resolve low scale features. Tim Palmer et.al
  - In numerical linear algebra.
    - For the solution of linear systems GMRES-based Iterative refinement (GMRES-IR). (Carson and Higham 2018).

Part of a broader picture in the context of algorithms for extreme scale computing. J. Dongarra et.al. classify these multi precision algorithms as ‘Responsibly Reckless’.

GMRES-IR is the focus of this talk
Given $A$ and $b$ in precision $u$.

solve $Ax_0 = b$ using the LU factors of precision $u_f > u$

$\bullet \quad r = b - Ax_0$, in $u_r < u$.

$\bullet$ Solve $\tilde{A}d \equiv \hat{U}^{-1}\hat{L}^{-1}A = \hat{U}^{-1}\hat{L}^{-1}r$, at precision $u$ using GMRES.

$\bullet$ Update $x_1 = fl(x_0 + d)$ in precision $u$.

<table>
<thead>
<tr>
<th>$u_f$</th>
<th>$u$</th>
<th>$u_r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>half</td>
<td>single</td>
<td>double</td>
</tr>
<tr>
<td>half</td>
<td>double</td>
<td>quad</td>
</tr>
<tr>
<td>single</td>
<td>double</td>
<td>quad</td>
</tr>
</tbody>
</table>
Features

- Backward and Forward errors of the order of $u$ if
  \[ \kappa_\infty(\tilde{A})u \ll 1. \]
- Speedup of 4 and energy reduction of 80% in NVIDIA V100. J. Dongarra et.al.
- Implementation available in MAGMA since 2.5.0 version.
Issues

- Range of fp16 number: $[5.96 \times 10^{-8}, 6.55 \times 10^4]$. 
- GMRES-IR involves conversion to fp16, which can cause
  - Undeflow
  - Overflow
  - Numbers in the range $[10^{-8}, 10^{-5}]$ are subnormal
Range of fp16 number: $[5.96 \times 10^{-8}, 6.55 \times 10^{4}]$. GMRES-IR involves conversion to fp16, which can cause
- Undeflow
- Overflow
- Numbers in the range $[10^{-8}, 10^{-5}]$ are subnormal

An algorithms to squeeze a matrix into the range of fp16, whilst using its complete range.
Inf. Round and replace Infinities

1: $A^{(h)} = fl_h(A)$
2: For every $i$ and $j$ such that $|a_{ij}^{(h)}| \geq \theta x_{\text{max}}$, set $a_{ij}^{(h)} = \text{sign}(a_{ij}) \theta x_{\text{max}}$.

Large perturbation

Scale. Scale and then round

1: $a_{\text{max}} = \max_{i,j} |a_{ij}|$
2: $\mu = \theta x_{\text{max}} / a_{\text{max}}$
3: $A^{(h)} = fl_h(\mu A)$

Underflow or subnormal numbers if $a_{\text{max}} \gg \theta x_{\text{max}}$
Two-sides Diagonal Scaling

**2DS.** Rounds $A \in \mathbb{R}^{n \times n}$ to the fp16 matrix $A^{(h)}$, scaling all elements to avoid overflow. \( \theta \in (0, 1] \) is a parameter.

1: Apply any two-sided diagonal scaling algorithm to $A$, to obtain diagonal matrices $R$, $S$.
2: Let $\beta = \max_{i,j} |RAS|_{ij}$.
3: $\mu = \theta \times_{\text{max}} / \beta$
4: $A^{(h)} = \text{fl}_h(\mu(RAS))$

---

Row and Column equilibriation

1: $r_i = \|A(i,:)\|_{-1}^{-1}$, \( i = 1 : n \)
2: $R = \text{diag}(r)$
3: $\tilde{A} = RA \quad \% \tilde{A}$ is row equilibrated.
4: $s_j = \|\tilde{A}(:,j)\|_{-1}^{-1}$, \( i = 1 : n \)
5: $S = \text{diag}(s)$
θ – headroom for further computation.

In $PA = LU$,

$$|l_{ij}| \leq 1, \quad |u_{ij}| \leq \rho_n \max_i |a_{ij}|.$$ 

If $\theta = 0.1$ (say), we can show that the pivot underflows if

$$\kappa_{\infty}(A) \geq \frac{\theta x_{\max}}{x_{\min}^s}.$$ 

For $\text{fp16}$ $\kappa_{\infty}(A) \geq 1.09 \times 10^{11}$. 

13 badly scaled matrices with $\max_{ij} |a_{ij}| \geq x_{\max}$ for fp16 are chosen from SuiteSparse Matrix Collection.

- $\kappa_\infty(A) \leq 10^{14}$
- $\theta = 0.1$.

Precisions, (half,single,double) and (half,double,quad).

For fp16 MATLAB class by Moler, and Advanpix for quad precision.

$M = \mu S \hat{U}^{-1} \hat{L}^{-1} R$ is used as the preconditioner to avoid the change of norm.

Iterative refinement is terminated when $b'err \leq nu$. 
#GMRES iterations (#IR steps)

<table>
<thead>
<tr>
<th>Index</th>
<th>(half, single, double)</th>
<th>(half, double, quad)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inf</td>
<td>Scale</td>
</tr>
<tr>
<td>1</td>
<td>4 (1)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>2</td>
<td>3 (1)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>3</td>
<td>45 (3)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>4</td>
<td>31 (2)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>5</td>
<td>95 (2)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>6</td>
<td>10 (1)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>7</td>
<td>0 (0)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>8</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>9</td>
<td>94 (3)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>10</td>
<td>409 (5)</td>
<td>1 (1)</td>
</tr>
<tr>
<td>11</td>
<td>212 (2)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>12</td>
<td>0 (0)</td>
<td>– (–)</td>
</tr>
<tr>
<td>13</td>
<td>0 (0)</td>
<td>– (–)</td>
</tr>
</tbody>
</table>
## Two sided diagonal scaling – 2DS

<table>
<thead>
<tr>
<th>Index</th>
<th>(half, single, double)</th>
<th>(half, double, quad)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 (0)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>2</td>
<td>0 (0)</td>
<td>4 (2)</td>
</tr>
<tr>
<td>3</td>
<td>2 (1)</td>
<td>6 (2)</td>
</tr>
<tr>
<td>4</td>
<td>0 (0)</td>
<td>16 (2)</td>
</tr>
<tr>
<td>5</td>
<td>0 (0)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>6</td>
<td>0 (0)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>7</td>
<td>0 (0)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>8</td>
<td>0 (0)</td>
<td>2 (1)</td>
</tr>
<tr>
<td>9</td>
<td>0 (0)</td>
<td>9 (3)</td>
</tr>
<tr>
<td>10</td>
<td>1 (1)</td>
<td>11 (3)</td>
</tr>
<tr>
<td>11</td>
<td>0 (0)</td>
<td>36 (3)</td>
</tr>
<tr>
<td>12</td>
<td>0 (0)</td>
<td>9 (2)</td>
</tr>
<tr>
<td>13</td>
<td>0 (0)</td>
<td>7 (2)</td>
</tr>
</tbody>
</table>
Remarks

- Purpose of two sided diagonal scaling is to squeeze the matrix into fp16 range.
- GMRES-IR with 2DS is mathematically equivalent to the unscaled system if the pivot sequence does not change.
- Numerically equivalent if scaling factors are powers of two.
- Pivot sequence may change after diagonal scaling.
- Important to work with unscaled problems as scaling changes norms!
Overflow and/or underflow issues are crucial in the context of fp16.

Two-sided diagonal scaling works well compared to simple remedies.

Multiplication by $\theta x_{\text{max}}$ makes complete use of the fp16 range.

2DS algorithm expands the range of problems which can be solved using GMRES-IR.

Further details “N.J. Higham, S. Pranesh, and M. Zounon. Squeezing a Matrix into Half Precision, with an Application to Solving Linear Systems.”
Overflow and/or underflow issues are crucial in the context of fp16.

Two-sided diagonal scaling works well compared to simple remedies.

Multiplication by $\theta x_{\text{max}}$ makes complete use of the fp16 range.

2DS algorithm expands the range of problems which can be solved using GMRES-IR.

Further details “N.J. Higham, S. Pranesh, and M. Zounon. Squeezing a Matrix into Half Precision, with an Application to Solving Linear Systems.”

Thank You.

Questions ???