MANCHESTER 1824

The University of Manchester

M\cr NA
Manchester Numerical Analysis

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.

	u	$x_{\min}^{s}$	$x_{\min}$	$x_{\max}$
bfloat16	3.91×10^{-3}	9.18×10^{-41}	1.18×10^{-38}	3.39×10^{38}
fp16	4.88×10^{-4}	5.96×10^{-8}	6.10×10^{-5}	6.55×10^{4}

- 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.

Applications

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

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).

Applications

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*'.

Applications

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

GMRES-IR

Given A and b in precision u.

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

- $r = b - Ax_0$, in $u_r < u$.
- Solve $\widetilde{A}d \equiv \widehat{U}^{-1}\widehat{L}^{-1}A = \widehat{U}^{-1}\widehat{L}^{-1}r$, at precision u using GMRES.
- Update $x_1 = \mathrm{fl}(x_0 + d)$ in precision u.

u_f	u	u_r
half	single	double
half	double	quad
single	double	quad

Features

- Backward and Forward errors of the order of u if $\kappa_\infty(\widetilde{A})u \ll 1$.
- 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

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

An algorithms to squeeze a matrix into the range of fp16, whilst using its complete range.

Simple remedies

Inf. Round and replace Infinities

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

Large peturbation

Scale. Scale and then round

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

Underflow or subnormal numbers if $a_{\max} \gg \theta x_{\max}$

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: $\mu = \theta x_{\max}/\beta$
4: $A^{(h)} = \mathrm{fl_h}(\mu(RAS))$

Row and Column equilibriation

1: $r_i = \|A(i,:)\|_\infty^{-1}$, $i = 1: n$
2: $R = \mathrm{diag}(r)$
3: $\widetilde{A} = RA$ % $\widetilde{A}$ is row equilibrated.
4: $s_j = \|\widetilde{A}(:,j)\|_\infty^{-1}$, $i = 1: n$
5: $S = \mathrm{diag}(s)$

Choice of θ

- θ – headroom for further computation.
- In $PA = LU$,

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

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

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

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

Numerical Experiments

- 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.
- $M = \mu S \widehat{U}^{-1} \widehat{L}^{-1} R$ is used as the preconditioner to avoid the change of norm.
- Iterative refinement is terminated when b'err $\leq nu$.

Simple methods

#GMRES iterations (#IR steps)

Index	(half, single, double)		(half, double, quad)	
	Inf	Scale	Inf	Scale
1	4 (1)	2 (1)	12 (2)	5 (2)
2	3 (1)	2 (1)	9 (2)	6 (2)
3	45 (3)	2 (1)	65 (3)	6 (2)
4	31 (2)	0 (0)	233 (3)	15 (2)
5	95 (2)	0 (0)	258 (3)	4 (2)
6	10 (1)	0 (0)	158 (4)	5 (2)
7	0 (0)	1 (1)	4 (2)	4 (2)
8	0 (0)	0 (0)	21 (3)	7 (2)
9	94 (3)	2 (1)	119 (3)	9 (3)
10	409 (5)	1 (1)	330 (3)	6 (2)
11	212 (2)	2 (1)	562 (3)	6 (2)
12	0 (0)	– (–)	9 (2)	– (–)
13	0 (0)	– (–)	8 (2)	– (–)

Two sided diagonal scaling – 2DS

Index	(half, single, double)	(half, double, quad)
1	0 (0)	2 (1)
2	0 (0)	4 (2)
3	2 (1)	6 (2)
4	0 (0)	16 (2)
5	0 (0)	2 (1)
6	0 (0)	2 (1)
7	0 (0)	2 (1)
8	0 (0)	8 (2)
9	0 (0)	9 (3)
10	1 (1)	11 (3)
11	0 (0)	36 (3)
12	0 (0)	9 (2)
13	0 (0)	7 (2)

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!

Conclusion

- Overflow and/or underflow issues are crucial in the context of fp16.
- Two-sided diagonal scaling works well compared to simple remedies.
- Multiplication by $\theta x_{\max}$ 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.*"

Conclusion

- Overflow and/or underflow issues are crucial in the context of fp16.
- Two-sided diagonal scaling works well compared to simple remedies.
- Multiplication by θx_{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 ???