MANCHESTER 1824

M\cr NA
Manchester Numerical Analysis

The University of Manchester

Mixed Precision Algorithms for High Performance Scientific Computing–Part II

- Three-Precision GMRES-Based Iterative Refinement for Least Squares Problems – Sri Pranesh
- How NVIDIA Tensor Cores can Help HPC Scientific Application Unleash the Power of GPUs using Mixed Precision Solvers – Azzam Haidar
- Iterative Refinement in up to Five Precisions for the Solution of Large Sparse Linear Systems – Bastien Vieublé
- Compressed Basis GMRES on High Performance GPUs – Thomas Grützmacher
- DGEMM using Tensor Cores – Daichi Mukunoki

MANCHESTER 1824

The University of Manchester

M\cr NA
Manchester Numerical Analysis

Three Precision GMRES-Based Iterative Refinement for Least Squares Problems

Srikara Pranesh
Department of Mathematics
The University of Manchester

srikara.pranesh@manchester.ac.uk

04-03-2021

Joint work with

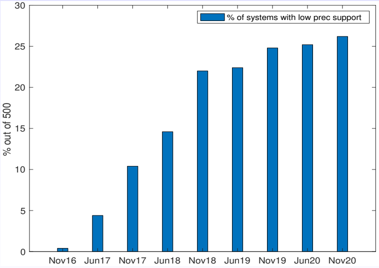
Erin Carson and Nick Higham

Slides Available at http://bit.ly/sri_cse21

Introduction

Type	bits Signif(t)	Exp	Range	$u = 2^{-t}$
bfloat16	8	8	$10^{\pm 38}$	3.9×10^{-3}
fp16	11	5	$10^{\pm 5}$	4.9×10^{-4}
tf32	11	8	$10^{\pm 38}$	4.9×10^{-4}
fp32	24	8	$10^{\pm 38}$	6.0×10^{-8}
fp64	53	11	$10^{\pm 308}$	1.1×10^{-16}
fp128	113	45	$10^{\pm 4932}$	9.6×10^{-35}

Type	bits Signif(t)	Exp	Range	$u = 2^{-t}$
bfloat16	8	8	$10^{\pm 38}$	3.9×10^{-3}
fp16	11	5	$10^{\pm 5}$	4.9×10^{-4}
tf32	11	8	$10^{\pm 38}$	4.9×10^{-4}
fp32	24	8	$10^{\pm 38}$	6.0×10^{-8}
fp64	53	11	$10^{\pm 308}$	1.1×10^{-16}
fp128	113	45	$10^{\pm 4932}$	9.6×10^{-35}

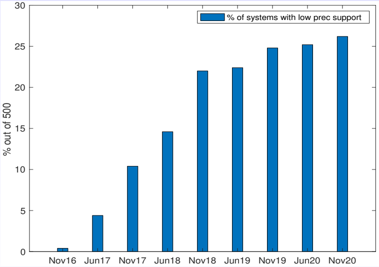


https://www.top500.org/statistics/list/

Introduction

Type	bits Signif(t)	Exp	Range	$u = 2^{-t}$
bfloat16	8	8	$10^{\pm 38}$	3.9×10^{-3}
fp16	11	5	$10^{\pm 5}$	4.9×10^{-4}
tf32	11	8	$10^{\pm 38}$	4.9×10^{-4}
fp32	24	8	$10^{\pm 38}$	6.0×10^{-8}
fp64	53	11	$10^{\pm 308}$	1.1×10^{-16}
fp128	113	45	$10^{\pm 4932}$	9.6×10^{-35}

- We need algorithms ....
  - that use low precision.
  - that are provably robust.
- **Mixed Precision Algorithms**



https://www.top500.org/statistics/list/

GMRES-IR [Carson, Higham 2018]

Given A and b in precision u, and $u_f \geq u \geq u_r$

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.

GMRES-IR [Carson, Higham 2018]

Given A and b in precision u, and $u_f \geq u \geq u_r$

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 $x_1 = \text{fl}(x_0 + d)$ in precision u.

- For a GPU implementation Four times speedup, and 80% reduction in energy consumption. **[Haidar et.al, 2018]**
- Vendor and open source implementations available.
- Extension to SPD systems as well **[Higham, P, 2021]**. **MS21** '***Mixed Precision Numerical Linear Algebra for Statistics Computations***'. Talk by **N. Higham**.

IR for least squares

Björck's Idea

Augmented system

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

$$\widetilde{A}\widetilde{x} = \widetilde{b}$$

- Invoke iterative refinement of a linear system.
- Linear system can be solved using $A = QR$.

IR for least squares

Björck's Idea

Augmented system

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

$$\widetilde{A}\widetilde{x} = \widetilde{b}$$

- Invoke iterative refinement of a linear system.
- Linear system can be solved using $A = QR$.

Question

What is the behavior of a mixed precision iterative refinement algorithm for solving $\widetilde{A}\widetilde{x} = \widetilde{b}$?

Recall

$$u_f \geq u \geq u_r$$

Given $\widetilde{A}$ and $\widetilde{b}$ in precision u.

Solve $\widetilde{A}\widetilde{x}_0 = \widetilde{b}$ in precision u_f.

- $r = \widetilde{b} - \widetilde{A}\widetilde{x}_0$, in u_r.
- Solve $\widetilde{A}d = r$, at precision u.
- Update $x_1 = \mathrm{fl}(\widetilde{x}_0 + d)$ in precision u.

Recall

$$u_f \geq u \geq u_r$$

Given $\widetilde{A}$ and $\widetilde{b}$ in precision u.
Solve $\widetilde{A}\widetilde{x}_0 = \widetilde{b}$ in precision u_f.

- $r = \widetilde{b} - \widetilde{A}\widetilde{x}_0$, in u_r.
- Solve $\widetilde{A}d = r$, at precision u.
- Update $x_1 = \mathrm{fl}(\widetilde{x}_0 + d)$ in precision u.

Variants

- **Type 1** – Solve the augmented system using u_f precision QR factors.
- **Type 2** – GMRES with an appropriate preconditioner.

Convergence Conditions

Properties of correction equation solver to ensure convergence of IR

- Solutions of some relative accuracy. (f'err $\leq$ 1)
- Backward stable.

Convergence Conditions

Properties of correction equation solver to ensure convergence of IR

- Solutions of some relative accuracy. (f'err ≤ 1)
- Backward stable.

Householder QR is assumed

- **Type 1**
  - f'err $\leq c_{m,n} u_f \kappa_\infty(A)$
  - u_f =half (fp16), $\kappa_\infty(A) \leq 2 \times 10^3$.

Preconditioner

Scaling

$$\widetilde{A}_\alpha = \begin{bmatrix} \alpha I & A \\ A^T & 0 \end{bmatrix}, \quad \kappa_2(\widetilde{A}_\alpha) = \mathcal{O}(\kappa_2(A)).$$

if $\alpha = 2^{-1/2}\sigma_{\min}(A)$.

Preconditioner

$$M = \begin{bmatrix} \alpha I & QR \\ R^T Q^T & 0 \end{bmatrix}, \quad M^{-1} = \begin{bmatrix} \frac{1}{\alpha}(I - QQ^T) & QR^{-T} \\ R^{-1}Q^T & -\alpha R^{-1}R^{-T} \end{bmatrix}.$$

Condition number

$$\kappa_\infty(M^{-1}\widetilde{A}) \lesssim (1 + (\cdot)u_f\kappa_\infty(A))^2$$

Type 2

- f'err $\leq \boldsymbol{f}(\boldsymbol{m}+\boldsymbol{n})u\kappa_\infty(M^{-1}\widetilde{A})$.
- f'err$\leq$ 1, if $\kappa_\infty(A) \leq u^{-1/2}u_f$.
- For $u =$**fp64**, $u_f =$ **fp16**, then $\kappa_\infty(A) \lesssim 10^{11}$.

- f'err $\leq \boldsymbol{f}(\boldsymbol{m}+\boldsymbol{n}) u \kappa_{\infty}(M^{-1}\widetilde{A})$.
- f'err ≤ 1, if $\kappa_{\infty}(A) \leq u^{-1/2} u_f$.
- For $u =$ **fp64**, $u_f =$ **fp16**, then $\kappa_{\infty}(A) \lesssim 10^{11}$.

Summary

IR-Type	u_f	u	u_r	$\kappa_{\infty}(A)$	Backward error Norm.	Backward error Comp.	Forward error
Type 1	half	single	double	$2 \cdot 10^3$	single	single	single
	half	double	quad	$2 \cdot 10^3$	double	double	double
Type 2	half	single	double	10^7	single	single	single (**10^7**)
	half	double	quad	10^{16}	double	double	double (**10^{11}**)

Numerical Experiments

- 100×10, mode 3, `randsvd` matrices.
- SuiteSparse: $20 \leq m \leq 2000$, $n \leq 400$, $n < m$.
- (half,double,quad)
- Advanpix for quad precision.
- `chop` for fp16. **[Higham, P, 2019]** (**https://github.com/higham/chop**)
- IR terminated for f'err $\leq nu$.
- $\alpha = 2^{-1/2}\sigma_{\min}(A)$ is computed using low precision $\widehat{R}$ factor.

randsvd

$\kappa_2(A)$	**Type 1**	**Type 2**
1.00e+02	13	2 (16)
1.00e+04	–	2 (24)
1.00e+07	–	2 (49)
1.00e+09	–	4 (153)
1.00e+10	–	5 (292)
1.00e+11	–	7 (491)

SuiteSparse

Name	Type 1	Type 2
divorce	7	1 (6)
Cities	9	2 (14)
ash219	5	1 (5)
WorldCities	11	2 (16)
ash331	5	1 (5)
robot24c1_mat5	–	6 (1811)
ash608	5	1 (5)
ash958	5	1 (5)
illc1033	–	2 (90)
well1033	–	2 (28)
photogrammetry	–	4 (5428)

Remarks

- **Underflow** and **overflow** can be addressed using **column scaling**.
- Similar results for **(half,single,double)**, and **(single,double,quad)**.
- Other saddle point preconditioners can be tried.

$$\begin{bmatrix} \alpha I & 0 \\ 0 & \frac{1}{\alpha}\widehat{R}^T\widehat{R} \end{bmatrix}$$

- Analysis does not apply for two sided preconditioning.
- **MINRES** as the iterative solver for corrections equation.
- Backward stability of **MINRES** cannot be guaranteed.

Conclusion

- GMRES-LSIR is a provably robust mixed precision algorithm for solving unconstrained least squares problem.
- GMRES-LSIR is suitable for solving ill conditioned problems.
- If IR converges quickly, most of the work is done in low precision matrix factorization.
- MINRES works well in practice.
- Two sided preconditioning works well in practice.

Conclusion

- GMRES-LSIR is a provably robust mixed precision algorithm for solving unconstrained least squares problem.
- GMRES-LSIR is suitable for solving ill conditioned problems.
- If IR converges quickly, most of the work is done in low precision matrix factorization.
- MINRES works well in practice.
- Two sided preconditioning works well in practice.

Carson, Higham, **P**, *'Three Precision GMRES-Based Iterative Refinement for Least Squares Problems'*, SISC, 42(6), A4063–A4083.

Slides Available at `http://bit.ly/sri_cse21`

Thank you! Questions?