


CHARLES UNIVERSITY
Faculty of mathematics
and physics

Co-funded by the
European Union

Mixed Precision
FGMRES

Joint with Erin Carson

We acknowledge funding from ERC Starting Grant No. 101075632 and the ExascaleComputing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Admin. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the ERC. Neither the European Union nor the granting authority can be held responsible for them.

dauzickaite@karlin.mff.cuni.cz

Ieva Daužickaitė

Sparse Days, June 18, 2024, Toulouse

PROBLEM

- **Solve**

$$Ax = b,$$

where $A \in \mathbb{R}^{n \times n}$ is indefinite and nonsymmetric, and $x, b \in \mathbb{R}^n$.

PROBLEM

- **Solve**

$$Ax = b,$$

where $A \in \mathbb{R}^{n \times n}$ is indefinite and nonsymmetric, and $x, b \in \mathbb{R}^n$.

- **Use:**
  - **Iterative Krylov subspace solver FGMRES**
  - **Preconditioning**
  - **Mixed precision**

PROBLEM

- **Solve**

$$Ax = b,$$

where $A \in \mathbb{R}^{n \times n}$ is indefinite and nonsymmetric, and $x, b \in \mathbb{R}^n$.

- **Use:**
  - **Iterative Krylov subspace solver FGMRES**
  - **Preconditioning**
  - **Mixed precision**
- **Question: can we do it in a backward stable way?**

$$\frac{\|b - A\bar{x}_k\|}{\|b\| + \|A\|\|\bar{x}_k\|} \leq ?$$

MIXED PRECISION

BITS

	Sign	Exponent	Significand	Range	Roundoff u
IEEE fp128 (Quadruple)	1	15	112	$10^{\pm 4932}$	1×10^{-34}
IEEE fp64 (Double)	1	11	52	$10^{\pm 308}$	1×10^{-16}
IEEE fp32 (Single)	1	8	23	$10^{\pm 38}$	6×10^{-8}
IEEE fp16 (Half)	1	5	10	$10^{\pm 5}$	5×10^{-4}

RANGE PRECISION

Low precision

available on hardware

reduced range

higher performance

larger unit roundoff

Mixed precision framework

- Aims to reduce the cost and preserve suitable accuracy;
- Perform expensive/cheap computations in lower/higher precision.

Reviews: [Abdelfattah et al., 2021], [Higham and Mary, 2022]

PRECONDITIONING

$$Ax = b$$

Left-preconditioning

$$M_L^{-1}Ax = M_L^{-1}b$$

Right-preconditioning

$$AM_R^{-1}\tilde{x} = b,$$
where $\tilde{x} = M_R x$

Split-preconditioning

$$M_L^{-1}AM_R^{-1}\tilde{x} = M_L^{-1}b,$$
where $\tilde{x} = M_R x$

PRECONDITIONING

$$Ax = b$$

Left-preconditioning

$$M_L^{-1}Ax = M_L^{-1}b$$

Right-preconditioning

$$AM_R^{-1}\tilde{x} = b,$$
where $\tilde{x} = M_R x$

Split-preconditioning

$$M_L^{-1}AM_R^{-1}\tilde{x} = M_L^{-1}b,$$
where $\tilde{x} = M_R x$

We consider

- **Split-preconditioning case**
- **General preconditioners M_L and M_R**

FGMRES: MIXED PRECISION, SPLIT PRECONDITIONER

Set x_0 and tolerance τ

$r_0 = M_L^{-1}(b - Ax_0)$

$\beta = \|r_0\|; v_1 = r_0/\beta$

for $k = 1,2,\ldots$

$z_k = M_R^{-1} v_k$

$s = Az_k$

$w = M_L^{-1} s$

for $i = 1,2,\ldots,k$

$h_{i,k} = v_i^T w$

$w = w - h_{i,k} v_i$

end

$h_{k+1,k} = \|w\|$

$H_k = \{h_{i,j}\}_{1 \le i \le j+1; 1 \le j \le k}$

$Z_k = [z_1, \ldots, z_k]$

$y_k = \arg\min_y \|\beta e_1 - H_k y\|$

if $\|\beta e_1 - H_k y_k\| \le \tau\beta$

$x_k = x_0 + Z_k y_k$

else

$v_{k+1} = w/h_{k+1,k};\ V_{k+1} = [v_1, \ldots, v_{k+1}]$

end

end

Solve
$M_L^{-1} A M_R^{-1} \tilde{x} = M_L^{-1} b,$
where $\tilde{x} = M_R x$

Arnoldi's method

Least squares problem

Solution update

FGMRES: MIXED PRECISION, SPLIT PRECONDITIONER

Set x_0 and tolerance τ

$r_0 = M_L^{-1}(b - Ax_0)$

$\beta = \|r_0\|; v_1 = r_0/\beta$

for $k = 1,2,\ldots$

$z_k = M_R^{-1} v_k$

$s = Az_k$

$w = M_L^{-1} s$

for $i = 1,2,\ldots,k$

$h_{i,k} = v_i^T w$

$w = w - h_{i,k} v_i$

end

$h_{k+1,k} = \|w\|$

$H_k = \{h_{i,j}\}_{1 \le i \le j+1; 1 \le j \le k}$

$Z_k = [z_1, \ldots, z_k]$

$y_k = \arg\min_y \|\beta e_1 - H_k y\|$

if $\|\beta e_1 - H_k y_k\| \le \tau\beta$

$x_k = x_0 + Z_k y_k$

else

$v_{k+1} = w/h_{k+1,k};\ V_{k+1} = [v_1, \ldots, v_{k+1}]$

end

end

Solve
$M_L^{-1} A M_R^{-1} \tilde{x} = M_L^{-1} b,$
where $\tilde{x} = M_R x$

Arnoldi's method

Least squares problem

Solution update

Mixed precision

(Unit roundoff)

Matvecs with A : u_A

Applying M_L : u_L

Applying M_R : u_R

Working precision: u

FGMRES: MIXED PRECISION, SPLIT PRECONDITIONER

Set x_0 and tolerance τ

$r_0 = M_L^{-1}(b - Ax_0)$

$\beta = \|r_0\|; v_1 = r_0/\beta$

for $k = 1,2,\ldots$

$z_k = M_R^{-1} v_k$

$s = Az_k$

$w = M_L^{-1} s$

for $i = 1,2,\ldots,k$

$h_{i,k} = v_i^T w$

$w = w - h_{i,k} v_i$

end

$h_{k+1,k} = \|w\|$

$H_k = \{h_{i,j}\}_{1 \le i \le j+1; 1 \le j \le k}$

$Z_k = [z_1, \ldots, z_k]$

$y_k = \arg\min_y \|\beta e_1 - H_k y\|$

if $\|\beta e_1 - H_k y_k\| \le \tau\beta$

$x_k = x_0 + Z_k y_k$

else

$v_{k+1} = w/h_{k+1,k}; \; V_{k+1} = [v_1, \ldots, v_{k+1}]$

end

end

Solve
$M_L^{-1} A M_R^{-1} \tilde{x} = M_L^{-1} b,$
where $\tilde{x} = M_R x$

Arnoldi's method

Least squares problem

Solution update

Analysis is based on Arioli, Duff, Gratton, and Pralet (2007), and Arioli and Duff (2009)

Assumptions

- **Constant M_R throughout iterations**
- **Applying preconditioners (Vieublé, 2022)**

$$fl(M_L^{-1} w_j) = M_L^{-1} w_j + \Delta M_{L,j} w_j, \quad |\Delta M_{L,j}| \le c(n) u_L E_{L,j}$$
$$fl(M_R^{-1} w_j) = M_R^{-1} w_j + \Delta M_{R,j} w_j, \quad |\Delta M_{R,j}| \le c(n) u_R E_{R,j}$$

- **Matvecs with $M_L^{-1}A$ (Vieublé, 2022)**

$$fl(M_L^{-1} A z_j) = (M_{L,j}^{-1} + \Delta M_{L,j})(A + \Delta A_j) z_j$$
$$\approx M_L^{-1} A z_j + f_j,$$

where $\|f_j\| \le (u_A \psi_{A,j} + u_L \psi_{L,j}) \|M_L^{-1} A\| \|z_j\|,$

$$u_A \psi_{A,j} = \frac{\|M_L^{-1} \Delta A_j z_j\|}{\|M_L^{-1} A\| \|z_j\|} \text{ and } u_L \psi_{L,j} = \frac{\|\Delta M_{L,j} A z_j\|}{\|M_L^{-1} A\| \|z_j\|}$$

BACKWARD ERROR BOUND

Under some technical assumptions and if

$$\rho = 1.3\tilde{c}(n,k)\|M_R\|\left(u\|\bar{Z}_k\| + u_R\|E_R\|\right) < 1,$$

Theorem 2.1, Carson and D. (2024)

then the residual for the left-preconditioned system is bounded as

$$\|M_L^{-1}(b - A\bar{x}_k)\| \leq \frac{1.3c(n,k)}{1-\rho}(\zeta_1 + \zeta_2),$$

where

$$\zeta_1 = \left(u + u_L\|E_L M_L\|\right)\|M_L^{-1}b\|,$$

$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right)\|M_L^{-1}A\|\left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right).$$

BACKWARD ERROR BOUND

Under some technical assumptions and if

$$\rho = 1.3\tilde{c}(n,k)\|M_R\|\left(u\|\bar{Z}_k\| + u_R\|E_R\|\right) < 1,$$

Theorem 2.1, Carson and D. (2024)

then the residual for the left-preconditioned system is bounded as

$$\|M_L^{-1}(b - A\bar{x}_k)\| \leq \frac{1.3c(n,k)}{1-\rho}(\zeta_1 + \zeta_2),$$

where

$$\zeta_1 = \left(u + u_L\|E_L M_L\|\right)\|M_L^{-1}b\|,$$
$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right)\|M_L^{-1}A\|\left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right).$$

Then the normwise relative backward error for the original problem is bounded by

$$\frac{\|b - A\bar{x}_k\|}{\|b\| + \|A\|\|\bar{x}_k\|} \leq \frac{1.3c(n,k)}{1-\rho}\frac{\zeta_1 + \zeta_2}{\|b\| + \|A\|\|\bar{x}_k\|}\min\{\kappa(M_L), \kappa(M_L^{-1}A)\}.$$

Corollary 2.2, Carson and D. (2024)

BACKWARD ERROR BOUND

Under some technical assumptions and if

$$\rho = 1.3\tilde{c}(n,k)\|M_R\|\left(u\|\bar{Z}_k\| + u_R\|E_R\|\right) < 1,$$

Theorem 2.1, Carson and D. (2024)

then the residual for the left-preconditioned system is bounded as

$$\|M_L^{-1}(b - A\bar{x}_k)\| \leq \frac{1.3c(n,k)}{1-\rho}(\zeta_1 + \zeta_2),$$

where

$$\zeta_1 = \left(u + u_L\|E_L M_L\|\right)\|M_L^{-1}b\|,$$
$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right)\|M_L^{-1}A\|\left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right).$$

We expect ζ_2 to dominate

Then the normwise relative backward error for the original problem is bounded by

$$\frac{\|b - A\bar{x}_k\|}{\|b\| + \|A\|\|\bar{x}_k\|} \leq \frac{1.3c(n,k)}{1-\rho}\frac{\zeta_1 + \zeta_2}{\|b\| + \|A\|\|\bar{x}_k\|}\min\{\kappa(M_L), \kappa(M_L^{-1}A)\}.$$

Corollary 2.2, Carson and D. (2024)

BACKWARD ERROR BOUND

Under some technical assumptions and if

$$\rho = 1.3\tilde{c}(n,k)\|M_R\|\left(u\|\bar{Z}_k\| + u_R\|E_R\|\right) < 1,$$

then the residual for the left-preconditioned system is bounded as

$$\|M_L^{-1}(b - A\bar{x}_k)\| \leq \frac{1.3c(n,k)}{1-\rho}(\zeta_1 + \zeta_2),$$

where

$$\zeta_1 = \left(u + u_L\|E_L M_L\|\right)\|M_L^{-1}b\|,$$
$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right)\|M_L^{-1}A\|\left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right).$$

Theorem 2.1, Carson and D. (2024)

We expect ζ_2 to dominate

Point I:

u_A and u_L have to be chosen carefully

Then the normwise relative backward error for the original problem is bounded by

$$\frac{\|b - A\bar{x}_k\|}{\|b\| + \|A\|\|\bar{x}_k\|} \leq \frac{1.3c(n,k)}{1-\rho}\frac{\zeta_1 + \zeta_2}{\|b\| + \|A\|\|\bar{x}_k\|}\min\{\kappa(M_L), \kappa(M_L^{-1}A)\}.$$

Corollary 2.2, Carson and D. (2024)

SETTING THE PRECISIONS

Matvecs with A : u_A

Applying M_L : u_L

Applying M_R : u_R

Working precision: u

$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right) \|M_L^{-1}A\| \left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right)$$

› u_A **is set so that** $u_A \approx u/\psi_A$

Driven by $\kappa(M_L)$, Vieublé (2022)

$$u_A\psi_{A,j} = \frac{\|M_L^{-1}\Delta A_j z_j\|}{\|M_L^{-1}A\|\|z_j\|}$$

SETTING THE PRECISIONS

Matvecs with A : u_A

Applying M_L : u_L

Applying M_R : u_R

Working precision: u

$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right) \|M_L^{-1}A\| \left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right)$$

- u_A **is set so that** $u_A \approx u/\psi_A$
- u_L **is set so that** $u_L \approx u_A\psi_A/\psi_L$

Driven by $\kappa(M_L)$, Vieublé (2022)

$$u_A\psi_{A,j} = \frac{\|M_L^{-1}\Delta A_j z_j\|}{\|M_L^{-1}A\|\|z_j\|}$$

$$u_L\psi_{L,j} = \frac{\|\Delta M_{L,j} A z_j\|}{\|M_L^{-1}A\|\|z_j\|}$$

$\psi_L \leq \psi_A$ is likely, Vieublé (2022)

SETTING THE PRECISIONS

Matvecs with A : u_A

Applying M_L : u_L

Applying M_R : u_R

Working precision: u

$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right) \|M_L^{-1}A\| \left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right)$$

- u_A **is set so that** $u_A \approx u/\psi_A$
- u_L **is set so that** $u_L \approx u_A\psi_A/\psi_L$
- u_R **is set so that** $\dfrac{\|E_R\|}{\|M_R^{-1}\|} \leq u_R^{-1}$

Point II:
u_R may be chosen flexibly

Driven by $\kappa(M_L)$, Vieublé (2022)

$$u_A\psi_{A,j} = \frac{\|M_L^{-1}\Delta A_j z_j\|}{\|M_L^{-1}A\|\|z_j\|}$$

$$u_L\psi_{L,j} = \frac{\|\Delta M_{L,j} A z_j\|}{\|M_L^{-1}A\|\|z_j\|}$$

$\psi_L \leq \psi_A$ is likely, Vieublé (2022)

LEFT-, SPLIT- OR RIGHT-PRECONDITIONING?

Left-preconditioning

$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right) \|M_L^{-1}A\| \left(\|\bar{x}_k - \bar{x}_0\| + \|\bar{x}_0\|\right)$$

Split-preconditioning

$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right) \|M_L^{-1}A\| \left(\|\bar{Z}_k\| \|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right)$$

Right-preconditioning

$$\zeta_2 = \left(u + u_A\psi_A\right) \|A\| \left(\|\bar{Z}_k\| \|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right)$$

LEFT-, SPLIT- OR RIGHT-PRECONDITIONING?

Left-preconditioning

$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right) \|M_L^{-1}A\| \left(\|\bar{x}_k - \bar{x}_0\| + \|\bar{x}_0\|\right)$$

Split-preconditioning

$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right) \|M_L^{-1}A\| \left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right)$$

Right-preconditioning

$$\zeta_2 = \left(u + u_A\psi_A\right) \|A\| \left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right)$$

Aim

Small backward error → **Left-preconditioning with u_A and u_L set to possibly high precisions**

LEFT-, SPLIT- OR RIGHT-PRECONDITIONING?

Left-preconditioning

$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right) \|M_L^{-1}A\| \left(\|\bar{x}_k - \bar{x}_0\| + \|\bar{x}_0\|\right)$$

Split-preconditioning

$$\zeta_2 = \left(u + u_A\psi_A + u_L\psi_L\right) \|M_L^{-1}A\| \left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right)$$

Right-preconditioning

$$\zeta_2 = \left(u + u_A\psi_A\right) \|A\| \left(\|\bar{Z}_k\|\|M_R(\bar{x}_k - \bar{x}_0)\| + \|\bar{x}_0\|\right)$$

Point III:
choice of preconditioning strategy depends on application constraints on the precisions

Aim

Small backward error → **Left-preconditioning with u_A and u_L set to possibly high precisions**

Apply preconditioner in low precision → **Right-preconditioning, but be cautious of $\|\bar{Z}_k\|$**

NUMERICS

Dense

Split-preconditioning

- $A = UDV$ **is** 200×200**,** $\kappa(A) = 10^5$
- $M_L = \hat{L}$, $M_R = \hat{U}$ **and** $\hat{L}\hat{U} \approx A$ **(computed to 4 digits of accuracy)**
- $u = u_A$ **set to double**
- **We set** $E_R = |\hat{U}^{-1}||\hat{U}||\hat{U}^{-1}|$**, then**

  $\|E_R\|/\|M_R^{-1}\| = 2048$

  and $\|E_R\|/\|M_R^{-1}\| \leq u_R^{-1}$ **holds for all** u_R **shown here**

NUMERICS

Dense

Split-preconditioning

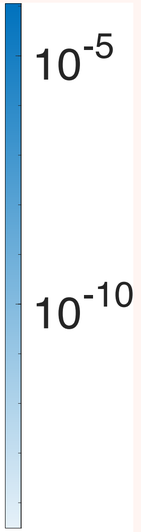
- $A = UDV$ **is** 200×200**,** $\kappa(A) = 10^5$
- $M_L = \hat{L}$, $M_R = \hat{U}$ **and** $\hat{L}\hat{U} \approx A$ **(computed to 4 digits of accuracy)**
- $u = u_A$ **set to double**
- **We set** $E_R = |\hat{U}^{-1}||\hat{U}||\hat{U}^{-1}|$**, then**

  $\|E_R\|/\|M_R^{-1}\| = 2048$

  and $\|E_R\|/\|M_R^{-1}\| \leq u_R^{-1}$ **holds for all** u_R **shown here**

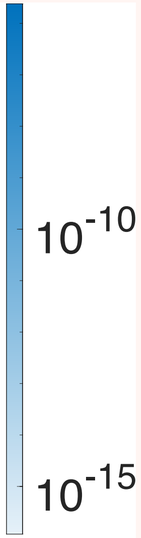
Bound

u_L \ u_R	Half	Single	Double	Quad
Quad	3e-15	5e-15	6e-15	5e-15
Double	3e-15	5e-15	5e-15	6e-15
Single	1e-08	3e-09	3e-09	3e-09
Half	1e-04	3e-05	4e-05	4e-05



BE

u_L \ u_R	Half	Single	Double	Quad
Quad	1e-16	1e-16	1e-16	1e-16
Double	1e-16	1e-16	1e-16	1e-16
Single	2e-10	9e-11	1e-10	1e-10
Half	2e-06	8e-07	8e-07	8e-07



NUMERICS

Dense

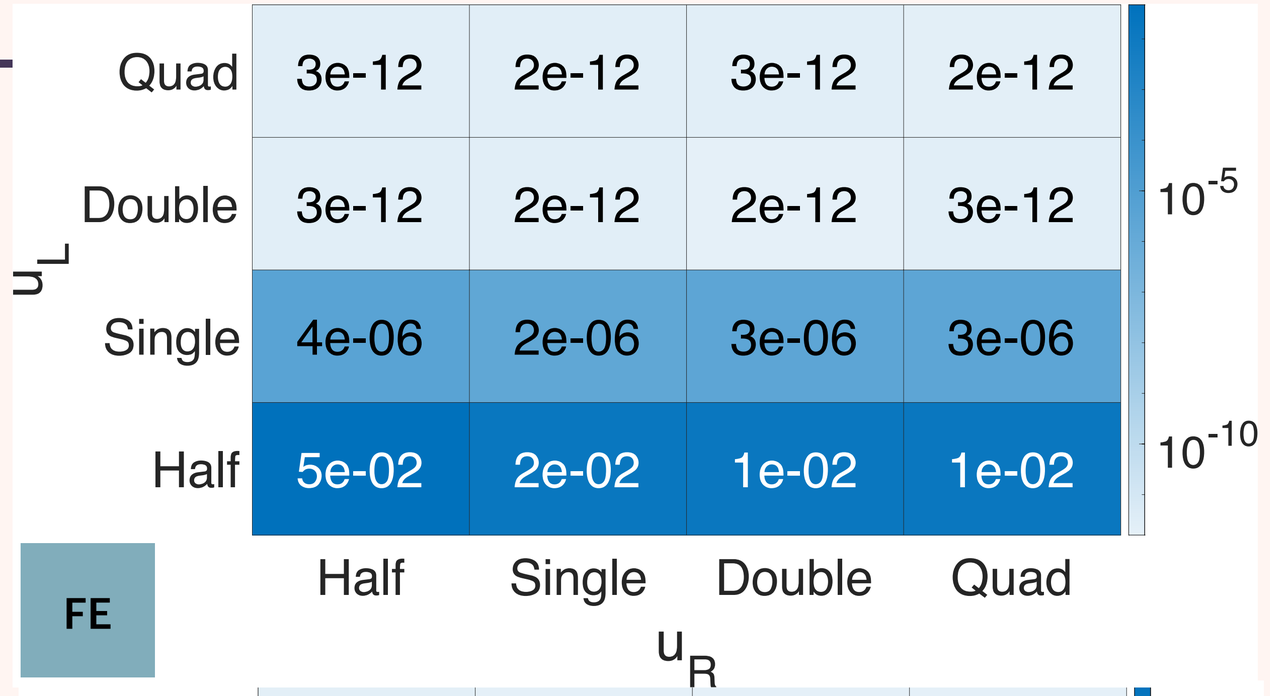
Split-preconditioning

- $A = UDV$ **is** 200×200**,** $\kappa(A) = 10^5$
- $M_L = \hat{L}$, $M_R = \hat{U}$ **and** $\hat{L}\hat{U} \approx A$ **(computed to 4 digits of accuracy)**
- $u = u_A$ **set to double**
- **We set** $E_R = |\hat{U}^{-1}||\hat{U}||\hat{U}^{-1}|$**, then**

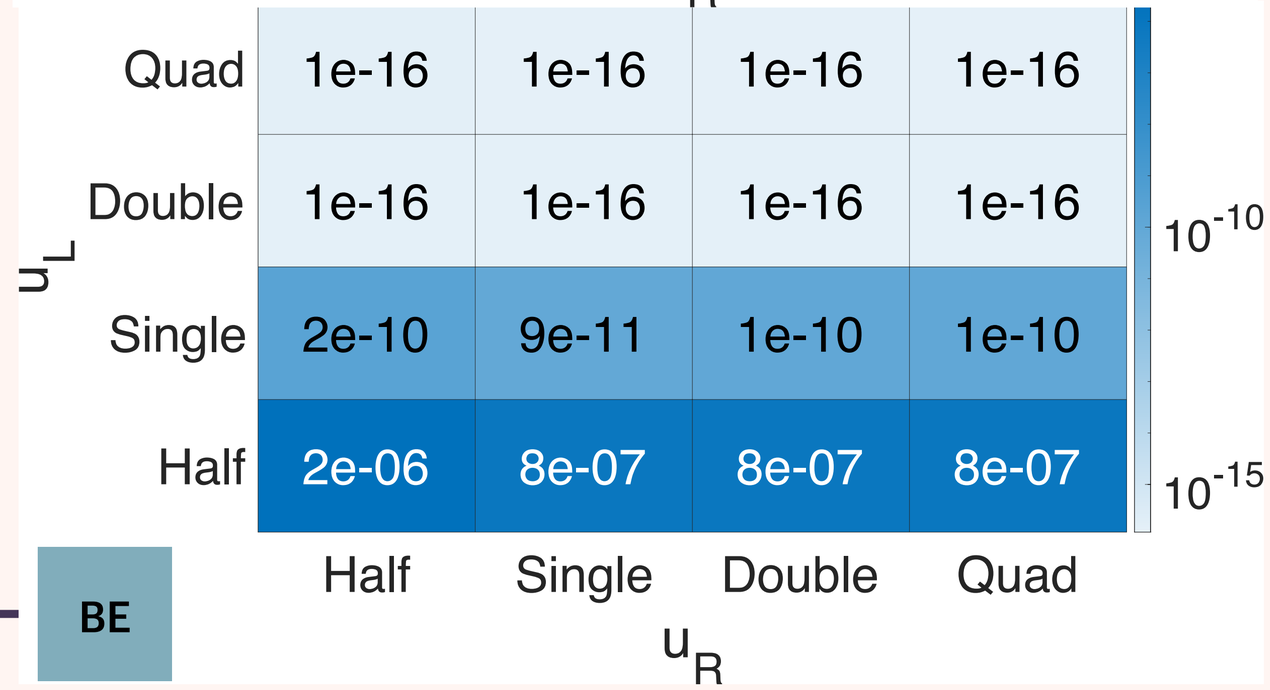
  $\|E_R\|/\|M_R^{-1}\| = 2048$

  and $\|E_R\|/\|M_R^{-1}\| \leq u_R^{-1}$ **holds for all** u_R **shown here**

FE



BE



NUMERICS

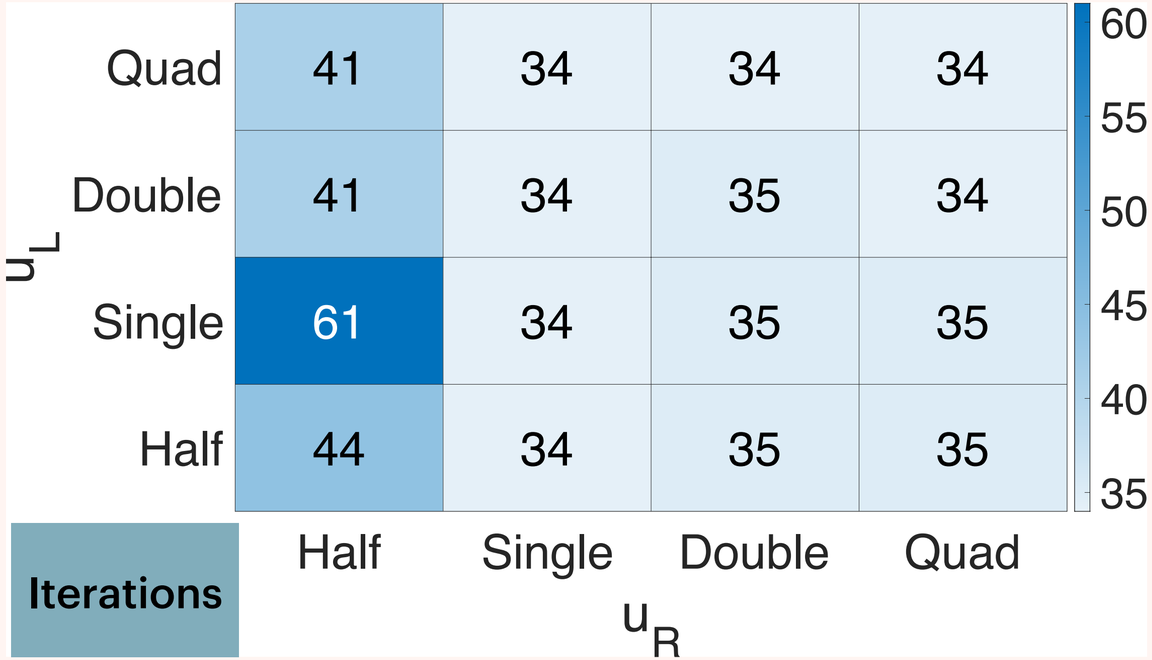
Dense

Split-preconditioning

- $A = UDV$ **is** 200×200**,** $\kappa(A) = 10^5$
- $M_L = \hat{L}$, $M_R = \hat{U}$ **and** $\hat{L}\hat{U} \approx A$ **(computed to 4 digits of accuracy)**
- $u = u_A$ **set to double**
- **We set** $E_R = |\hat{U}^{-1}||\hat{U}||\hat{U}^{-1}|$**, then**

  $\|E_R\|/\|M_R^{-1}\| = 2048$

  and $\|E_R\|/\|M_R^{-1}\| \leq u_R^{-1}$ **holds for all** u_R **shown here**



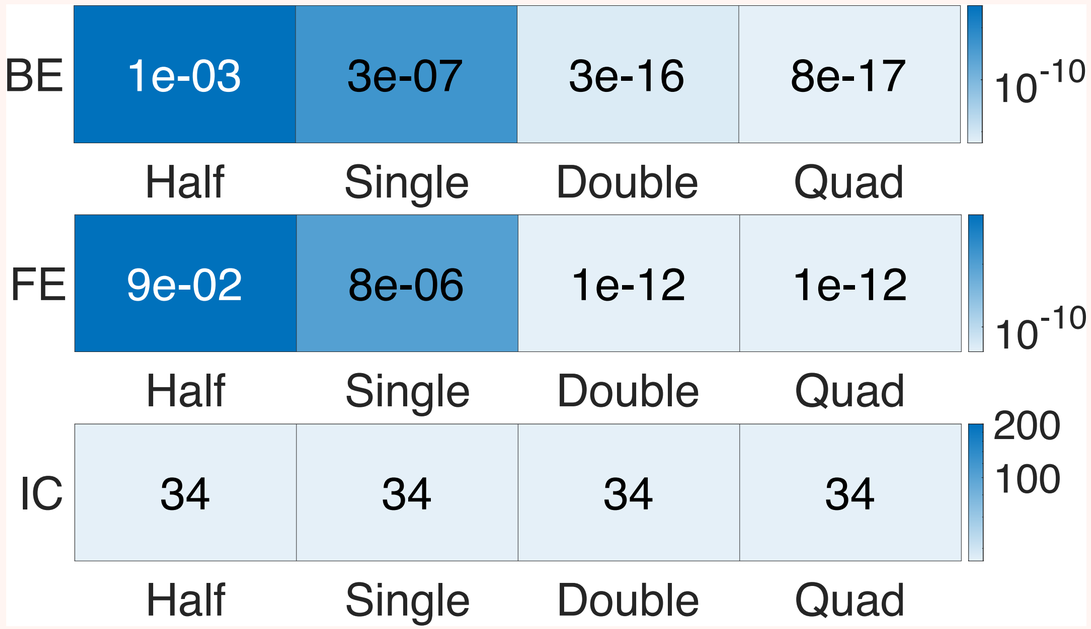
$$\rho = 1.3\tilde{c}(n,k)\|M_R\|\left(u\|\bar{Z}_k\| + u_R\|E_R\|\right) < 1$$

NUMERICS

Dense

Left-preconditioning

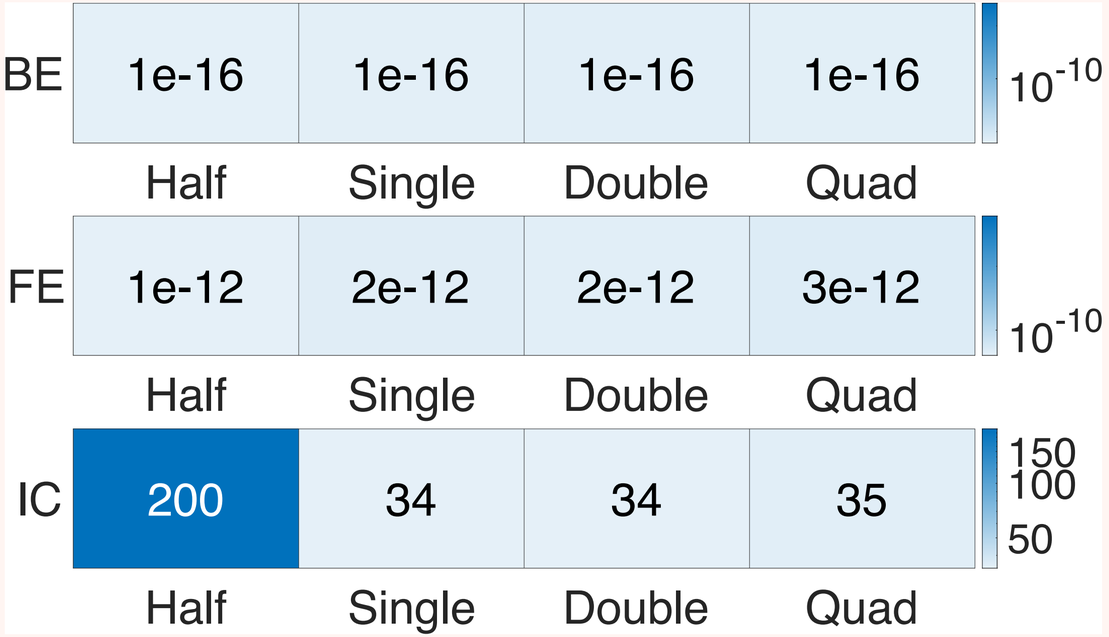
$$M_L = \hat{L}\hat{U}$$



u_L

Right-preconditioning

$$M_R = \hat{L}\hat{U}$$



u_R

SUMMARY

We provide bounds for the normwise backward error of split-preconditioned FGMRES in four precisions.

Roundoffs: Matvecs with A : u_A | Applying M_L : u_L | Applying M_R : u_R | Working precision: u

Point I:
u_A and u_L have to be chosen carefully

Point II:
u_R may be chosen flexibly

Point III:
choice of preconditioning strategy depends on application constraints on the precisions

CARSON, E., & DAUŽICKAITE, I. (2024). THE STABILITY OF SPLIT-PRECONDITIONED FGMRES IN FOUR PRECISIONS. *Electronic Transactions on Numerical Analysis*, *60*, 40-58.

THANK YOU!

dauzickaite@karlin.mff.cuni.cz

REFERENCES

- A. ABDELFATTAH, H. ANZT , E. G. BOMAN, et al. *A survey of numerical linear algebra methods utilizing mixed-precision arithmetic.* The International Journal of High Performance Computing Applications, 35(4) (2021), pp. 44-369.
- P. AMESTOY, A. BUTTARI, N. J. HIGHAM, J.-Y. L'EXCELLENT, T. MARY, AND B. VIEUBLÉ, *Five-precision GMRES-based iterative refinement.* SIAM Journal on Matrix Analysis and Applications, 45(1) (2024), pp. 529-552.
- M. ARIOLI AND I. S. DUFF, *Using FGMRES to obtain backward stability in mixed precision.* Electronic Transactions on Numerical Analysis, 33 (2009), pp. 31–44.
- M. ARIOLI, I. S. DUFF, S. GRATTON, AND S. PRALET, *A note on GMRES preconditioned by a perturbed* LDL^T *decomposition with static pivoting.* SIAM Journal on Scientific Computing, 29 (2007), pp. 2024–2044.
- E. CARSON AND N. J. HIGHAM, *A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems.* SIAM Journal on Scientific Computing, 39 (2017), pp. A2834–A2856.
- N. HIGHAM AND T. MARY, *Mixed precision algorithms in numerical linear algebra.* Acta Numerica, 31 (2022), pp. 347-414.
- B. VIEUBLÉ, *Mixed precision iterative refinement for the solution of large sparse linear systems*, PhD thesis, INP Toulouse, University of Toulouse, France, 2022.

GMRES & FGMRES

Some existing stability results

Solve

$$M_L^{-1} A M_R^{-1} \tilde{x} = M_L^{-1} b,$$

where $\tilde{x} = M_R x$

Normwise relative backward error:

$$\min\{\epsilon : (A + \Delta A)x_k = b + \Delta b, \|\Delta A\| \le \epsilon \|A\|, \|\Delta b\| \le \epsilon \|b\|\} = \frac{\|b - Ax_k\|}{\|b\| + \|A\| \|x_k\|}$$

Roundoffs

Computing M_L and/or M_R : u_f

Matvecs with A : u_A

Matvecs with M_L : u_L

Matvecs with M_R : u_R

Working precision: u

GMRES

- **Arioli et al (2007): NOT backward stable with**
  - $M_L = I$, $M_R = \widehat{L}\,\widehat{D}\,\widehat{L}^T$, $M_R = I$, **where** $\widehat{L}\,\widehat{D}\,\widehat{L}^T \approx A$ **is computed with static pivoting**
  - u
- **Carson and Higham (2017): backward stable with**
  - $M_L = \widehat{L}\,\widehat{U}$, $M_R = I$, **where** $\widehat{L}\,\widehat{U} \approx A$
  - $u_A = u_L = u^2$
- **Amestoy at al (2021): backward stable with**
  - $M_L = \widehat{L}\,\widehat{U}$, $M_R = I$, **where** $\widehat{L}\,\widehat{U} \approx A$ **is computed with roundoff** u_f
  - u_f, $u_A = u_L$, u
- **Vieublé (2022; PhD thesis): backward stable with**
  - **General** M_L **with assumptions on matvec errors,** $M_R = I$,
  - u_A, u_L, u

FGMRES

- **Arioli et al (2007): backward stable with**
  - $M_L = I$, $M_R = \widehat{L}\,\widehat{D}\,\widehat{L}^T$, $M_R = I$, **where** $\widehat{L}\,\widehat{D}\,\widehat{L}^T \approx A$ **is computed with static pivoting**
  - u
- **Arioli and Duff (2009): backward stable with**
  - $M_L = I$, $M_R = \widehat{L}\,\widehat{U}$, **where** $\widehat{L}\,\widehat{U} \approx A$ **is computed in single precision**
  - u **is double,** u_R **is double or single**