

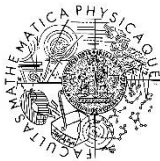
Recent Progress in Mixed Precision Numerical Linear Algebra

Erin Carson
Charles University

NJH60 2022

Manchester, UK

July 8, 2022



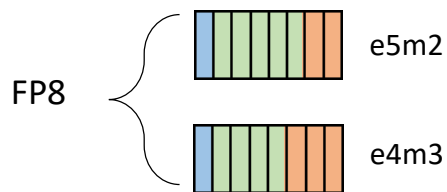
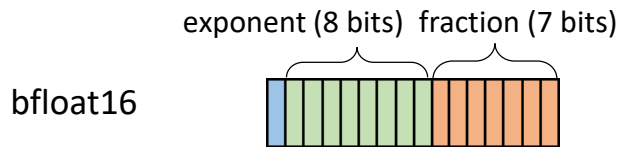
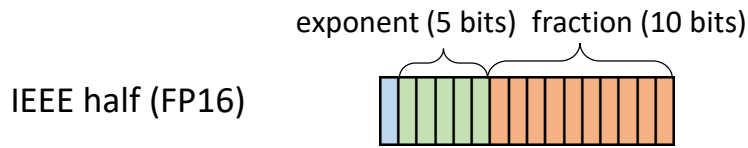
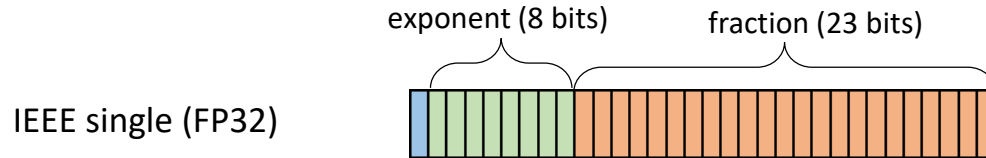
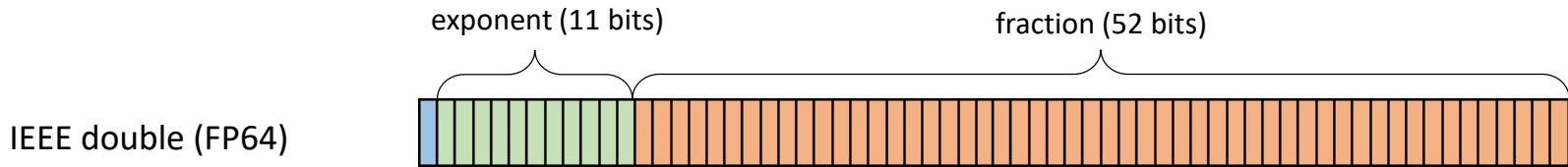
FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

History

- April 2016: Nick gives a talk at NASC Seminar at NYU
 - GMRES-based iterative refinement
[C. and Higham. "A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems." *SIAM SISC* 39.6 (2017): A2834-A2856.]
- Summer 2017: Visit to Manchester
 - 3-precision iterative refinement
[C. and Higham. "Accelerating the solution of linear systems by iterative refinement in three precisions." *SIAM SISC* 40.2 (2018): A817-A847.]
- Summer 2018: Visit to Manchester
 - 3-precision iterative refinement for least squares problems
[C., Higham, and Pranesh. "Three-precision GMRES-based iterative refinement for least squares problems." *SIAM SISC* 42.6 (2020): A4063-A4083.]

Floating Point Formats

$$(-1)^{\text{sign}} \times 2^{(\text{exponent}-\text{offset})} \times 1.\text{fraction}$$



| | size (bits) | range | u | perf. (NVIDIA H100) |
|----------|-------------|----------------|---------------------|---------------------|
| FP64 | 64 | $10^{\pm 308}$ | 1×10^{-16} | 60 Tflops/s |
| FP32 | 32 | $10^{\pm 38}$ | 6×10^{-8} | 1 Pflop/s |
| FP16 | 16 | $10^{\pm 5}$ | 5×10^{-4} | 2 Pflops/s |
| bfloat16 | 16 | $10^{\pm 38}$ | 4×10^{-3} | |
| FP8-e5m2 | 8 | $10^{\pm 5}$ | 1×10^{-1} | 4 Pflops/s |
| FP8-e4m3 | 8 | $10^{\pm 2}$ | 6×10^{-2} | |

Mixed precision in NLA

- **BLAS**: cuBLAS, MAGMA, [Agullo et al. 2009], [Abdelfattah et al., 2019], [Haidar et al., 2018]
- **Iterative refinement**:
 - Long history: [Wilkinson, 1963], [Moler, 1967], [Stewart, 1973], ...
 - More recently: [Langou et al., 2006], [C., Higham, 2017], [C., Higham, 2018], [C., Higham, Pranesh, 2020], [Amestoy et al., 2021]
- **Matrix factorizations**: [Haidar et al., 2017], [Haidar et al., 2018], [Haidar et al., 2020], [Abdelfattah et al., 2020]
- **Eigenvalue problems**: [Dongarra, 1982], [Dongarra, 1983], [Tisseur, 2001], [Davies et al., 2001], [Petschow et al., 2014], [Alvermann et al., 2019]
- **Sparse direct solvers**: [Buttari et al., 2008]
- **Orthogonalization**: [Yamazaki et al., 2015]
- **Multigrid**: [Tamstorf et al., 2020], [Richter et al., 2014], [Sumiyoshi et al., 2014], [Ljungkvist, Kronbichler, 2017, 2019]
- **(Preconditioned) Krylov subspace methods**: [Emans, van der Meer, 2012], [Yamagishi, Matsumura, 2016], [C., Gergelits, Yamazaki, 2021], [Clark, 2019], [Anzt et al., 2019], [Clark et al., 2010], [Gratton et al., 2020], [Arioli, Duff, 2009], [Hogg, Scott, 2010]

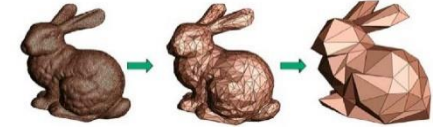
Mixed precision in NLA

- **BLAS**: cuBLAS, MAGMA, [Agullo et al. 2009], [Abdelfattah et al., 2019], [Haidar et al., 2018]
- **Iterative refinement**:
 - Long history: [Wilkinson, 1963], [Moler, 1967], [Stewart, 1973], ...
 - More recently: [Langou et al., 2006], [C., Higham, 2017], [C., Higham, 2018], [C., Higham, Pranesh, 2020], [Amestoy et al., 2021]
- **Matrix factorizations**: [Haidar et al., 2017], [Haidar et al., 2018], [Haidar et al., 2020], [Abdelfattah et al., 2020]
- **Eigenvalue problems**: [Dongarra, 1982], [Dongarra, 1983], [Tisseur, 2001], [Davies et al., 2001], [Petschow et al., 2014], [Alvermann et al., 2019]
- **Sparse direct solvers**: [Buttari et al., 2008]
- **Orthogonalization**: [Yamazaki et al., 2015]
- **Multigrid**: [Tamstorf et al., 2020], [Richter et al., 2014], [Sumiyoshi et al., 2014], [Ljungkvist, Kronbichler, 2017, 2019]
- **(Preconditioned) Krylov subspace methods**: [Emans, van der Meer, 2012], [Yamagishi, Matsumura, 2016], [C., Gergelits, Yamazaki, 2021], [Clark, 2019], [Anzt et al., 2019], [Clark et al., 2010], [Gratton et al., 2020], [Arioli, Duff, 2009], [Hogg, Scott, 2010]

Inexact computations

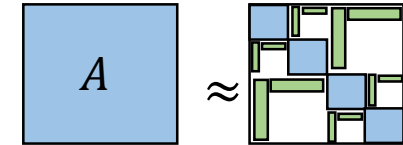
- In real computations we have many sources of inexactness
 - Imperfect data, measurement error
 - Modeling error, discretization error
 - Intentional approximation to improve performance
 - Reduced models, Low-rank representations, sparsification, randomization

Model Reduction



[Schilders, van der Vorst, Rommes, 2008]

Low-rank (hierarchical) approximation



Sparsification, Randomized algorithms

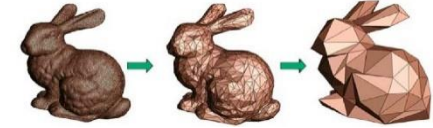


[Sinha, 2018]

Inexact computations

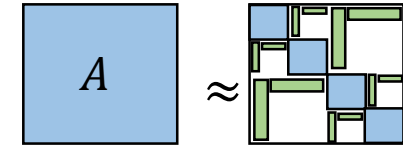
- In real computations we have many sources of inexactness
 - Imperfect data, measurement error
 - Modeling error, discretization error
 - Intentional approximation to improve performance
 - Reduced models, Low-rank representations, sparsification, randomization
- Given that we are already working with so much inexactness, do we need to be using double precision?

Model Reduction



[Schilders, van der Vorst, Rommes, 2008]

Low-rank (hierarchical) approximation



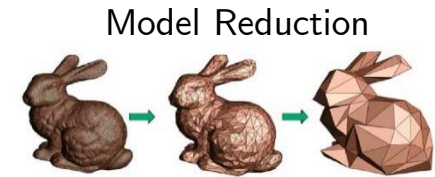
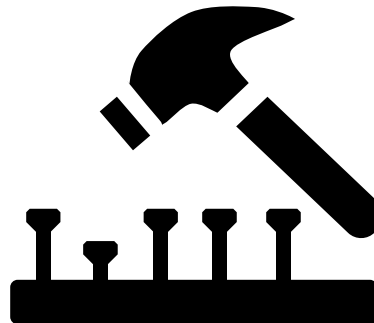
Sparsification, Randomized algorithms



[Sinha, 2018]

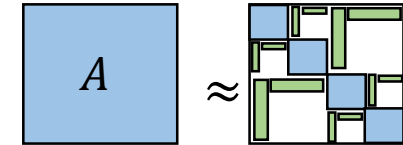
Inexact computations

- In real computations we have many sources of inexactness
 - Imperfect data, measurement error
 - Modeling error, discretization error
 - Intentional approximation to improve performance
 - Reduced models, Low-rank representations, sparsification, randomization
- Given that we are already working with so much inexactness, do we need to be using double precision?



[Schilders, van der Vorst, Rommes, 2008]

Low-rank (hierarchical) approximation



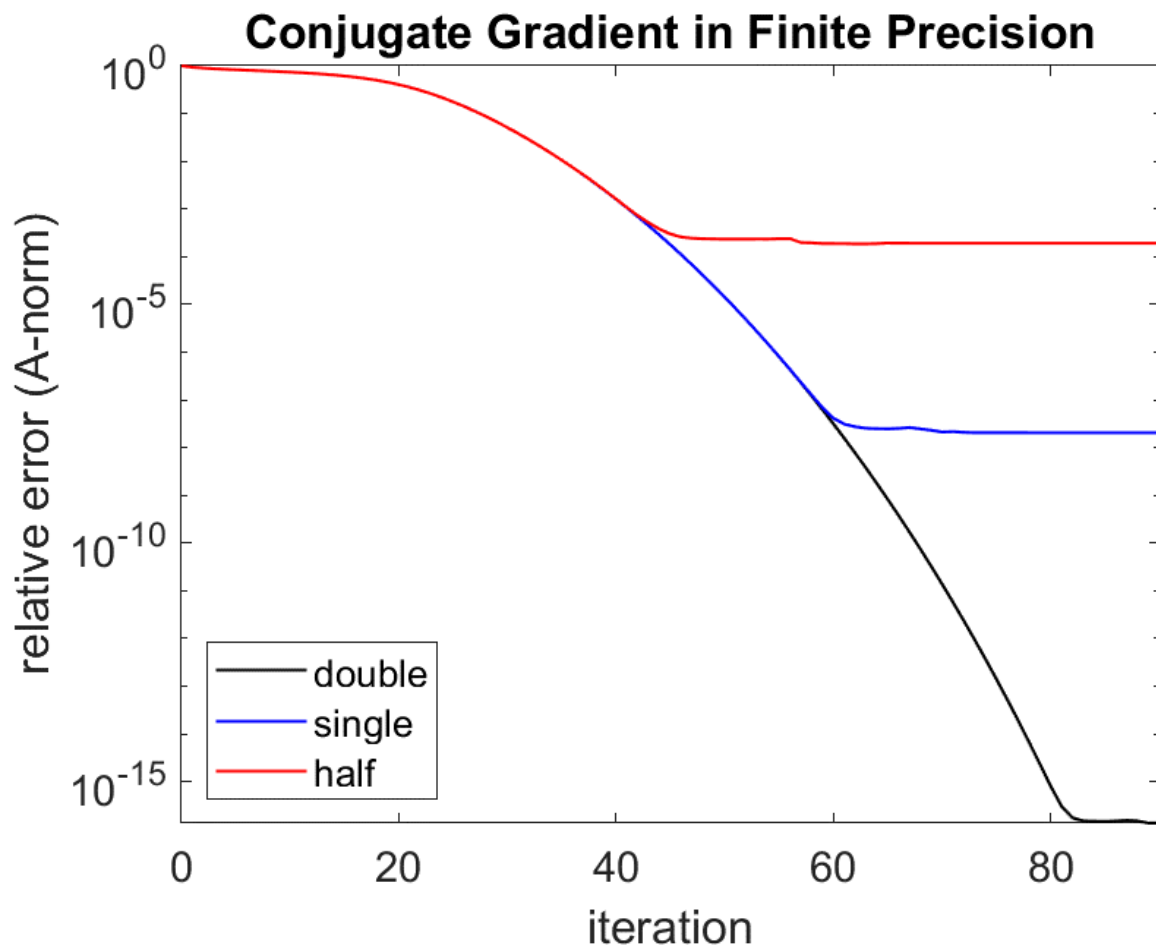
Sparsification, Randomized algorithms



[Sinha, 2018]

Caution: Iterative Methods

```
A = diag(linspace(.001,1,100));  
b = ones(n,1);
```

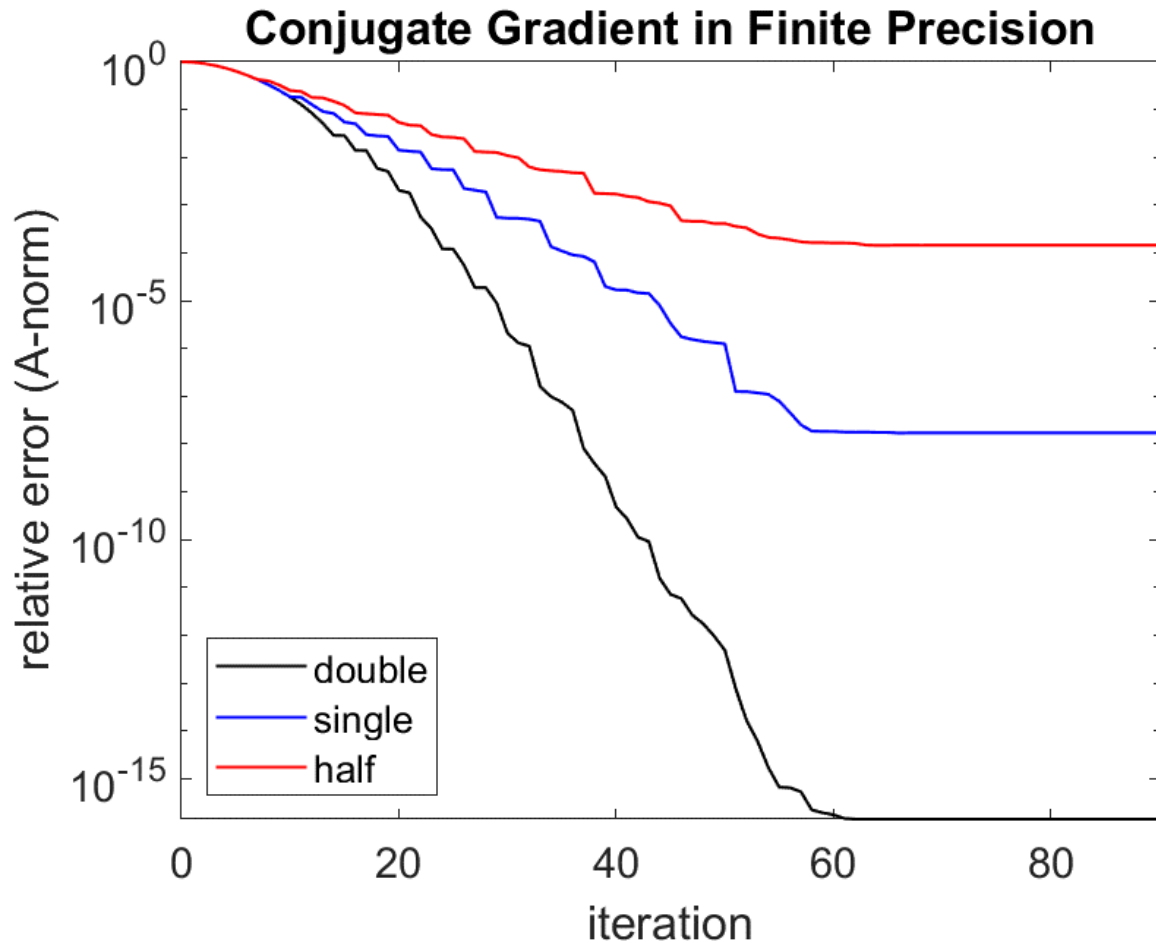


Caution: Iterative Methods

$$n = 100, \lambda_1 = 10^{-3}, \lambda_n = 1$$

$$\lambda_i = \lambda_1 + \left(\frac{i-1}{n-1}\right)(\lambda_n - \lambda_1)(0.65)^{n-i}, \quad i = 2, \dots, n-1$$

b = ones(n, 1);



Our setting

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive semidefinite matrix. Want to solve

$$(A + \mu I)x = b$$

where $\mu \geq 0$ is set so that $A + \mu I$ is positive definite.

Assume A has rapidly decreasing eigenvalues or cluster of large eigenvalues.

Many applications, e.g., ridge regression.

Limited Memory Preconditioners

Want to solve using PCG using **spectral limited memory preconditioner** [Gratton, Sartenaer, Tshimanga, 2011], [Tshimanga et al., 2008]:

$$P = I - UU^T + \frac{1}{\alpha + \mu} U(\Theta + \mu I)U^T$$
$$P^{-1} = I - UU^T + (\alpha + \mu)U(\Theta + \mu I)^{-1}U^T$$

where columns of $U \in \mathbb{R}^{n \times k}$ are k approximate eigenvectors of A and $U^T U = I$, Θ is diagonal with approximations to eigenvalues of A , and $\alpha \geq 0$.

Used in data assimilation [Laloyaux et al., 2018], [Mogensen, Alonso Balmaseda, Weaver, 2012], [Moore et al., 2011], [Daužickaitė, Lawless, Scott, van Leeuwen, 2021]

Randomized Nyström Approximation

Want to compute a rank- k approximation $A \approx U\Theta U^T$ via the randomized Nyström method.

Nyström approximation:

$$A_N = (AQ)(Q^T AQ)^+(AQ)^T$$

where Q is an $n \times k$ test matrix (random projection).

In the case that A is very large, **matrix-matrix products with A are the bottleneck.**

This motivates the **single-pass version** of the Nyström method.

Randomized Nyström Approximation

[Tropp et al., 2017]

Given sym. PSD matrix A , target rank k

$$G = \text{randn}(n, k)$$

$$[Q, \sim] = \text{qr}(G, 0)$$



Randomized Nyström Approximation

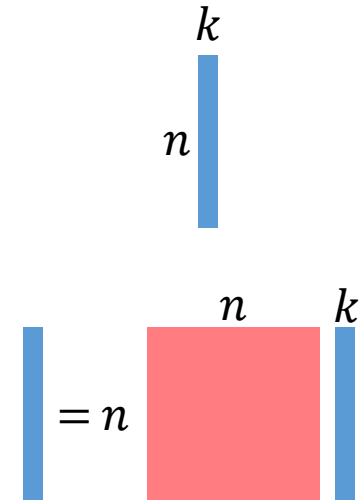
[Tropp et al., 2017]

Given sym. PSD matrix A , target rank k

$$G = \text{randn}(n, k)$$

$$[Q, \sim] = \text{qr}(G, 0)$$

$$Y = AQ$$



Randomized Nyström Approximation

[Tropp et al., 2017]

Given sym. PSD matrix A , target rank k

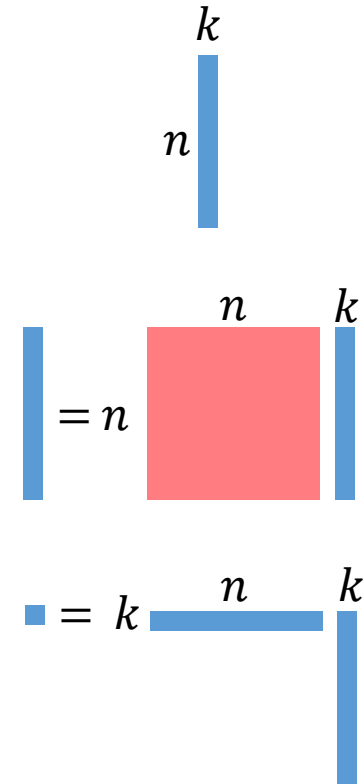
$$G = \text{randn}(n, k)$$

$$[Q, \sim] = \text{qr}(G, 0)$$

$$Y = AQ$$

Compute shift ν ; $Y_\nu = Y + \nu Q$

$$B = Q^T Y_\nu$$



Randomized Nyström Approximation

[Tropp et al., 2017]

Given sym. PSD matrix A , target rank k

$$G = \text{randn}(n, k)$$

$$[Q, \sim] = \text{qr}(G, 0)$$

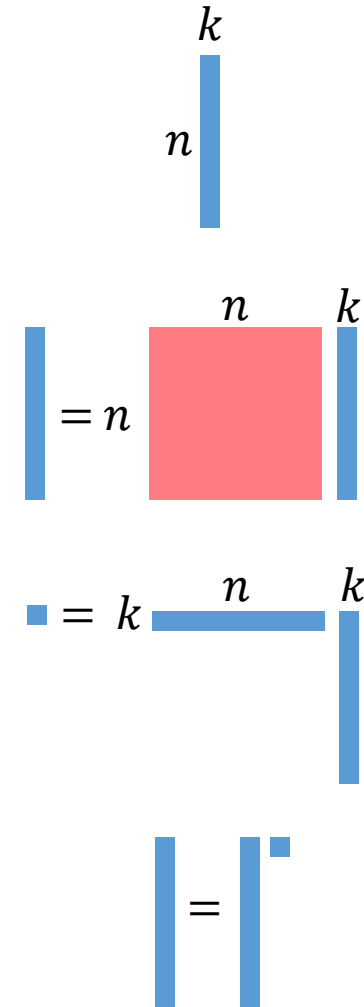
$$Y = AQ$$

Compute shift ν ; $Y_\nu = Y + \nu Q$

$$B = Q^T Y_\nu$$

$$C = \text{chol}((B + B^T)/2)$$

Solve $F = Y_\nu / C$



Randomized Nyström Approximation

[Tropp et al., 2017]

Given sym. PSD matrix A , target rank k

$$G = \text{randn}(n, k)$$

$$[Q, \sim] = \text{qr}(G, 0)$$

$$Y = AQ$$

Compute shift ν ; $Y_\nu = Y + \nu Q$

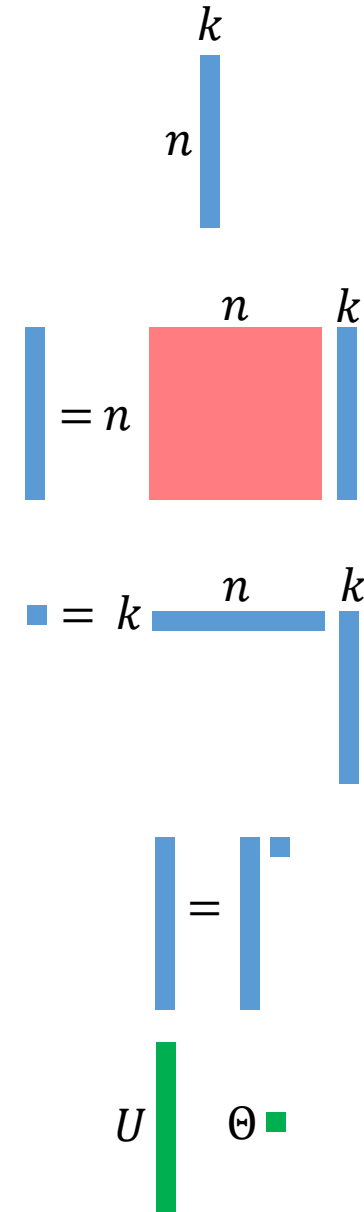
$$B = Q^T Y_\nu$$

$$C = \text{chol}((B + B^T)/2)$$

Solve $F = Y_\nu / C$

$$[U, \Sigma, \sim] = \text{svd}(F, 0)$$

$$\Theta = \max(0, \Sigma^2 - \nu I)$$



Randomized Nyström Approximation

[Tropp et al., 2017]

Given sym. PSD matrix A , target rank k

$$G = \text{randn}(n, k)$$

$$[Q, \sim] = \text{qr}(G, 0)$$

$$Y = AQ$$

Compute shift ν ; $Y_\nu = Y + \nu Q$

$$B = Q^T Y_\nu$$

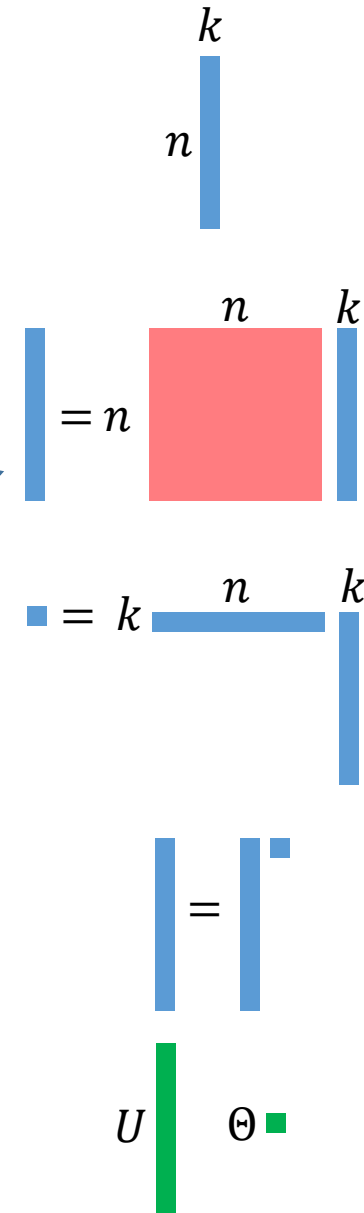
$$C = \text{chol}((B + B^T)/2)$$

Solve $F = Y_\nu / C$

$$[U, \Sigma, \sim] = \text{svd}(F, 0)$$

$$\Theta = \max(0, \Sigma^2 - \nu I)$$

Can we further reduce the cost of the matrix-matrix product with A by using low precision?




Error Bounds

$$\|A - \hat{A}_N\|_2 = \|A - A_N + A_N - \hat{A}_N\|_2 \leq \|A - A_N\|_2 + \|A_N - \hat{A}_N\|_2$$

exact Nyström
approximation




Nyström approximation
computed in
finite precision



Error Bounds

$$\|A - \hat{A}_N\|_2 = \|A - A_N + A_N - \hat{A}_N\|_2 \leq \underbrace{\|A - A_N\|_2}_{\substack{\text{exact} \\ \text{approximation} \\ \text{error}}} + \underbrace{\|A_N - \hat{A}_N\|_2}_{\substack{\text{finite precision} \\ \text{error}}}$$

Error Bounds

$$\|A - \hat{A}_N\|_2 = \|A - A_N + A_N - \hat{A}_N\|_2 \leq \underbrace{\|A - A_N\|_2}_{\substack{\text{exact} \\ \text{approximation} \\ \text{error}}} + \underbrace{\|A_N - \hat{A}_N\|_2}_{\substack{\text{finite precision} \\ \text{error}}}$$


Error Bounds

$$\|A - \hat{A}_N\|_2 = \|A - A_N + A_N - \hat{A}_N\|_2 \leq \underbrace{\|A - A_N\|_2}_{\substack{\text{exact} \\ \text{approximation} \\ \text{error}}} + \underbrace{\|A_N - \hat{A}_N\|_2}_{\substack{\text{finite precision} \\ \text{error}}}$$

Deterministic bound [Gittens, Mahoney, 2016]:

$$\|A - A_N\|_2 \leq \lambda_{k+1} + \left\| \Sigma_2^{1/2} U_2^T Q (U_1 Q)^+ \right\|_2^2$$

with $A = [U_1 \ U_2] \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} [U_1 \ U_2]^T$.

Error Bounds

$$\|A - \hat{A}_N\|_2 = \|A - A_N + A_N - \hat{A}_N\|_2 \leq \underbrace{\|A - A_N\|_2}_{\text{exact approximation error}} + \underbrace{\|A_N - \hat{A}_N\|_2}_{\text{finite precision error}}$$

Deterministic bound [Gittens, Mahoney, 2016]:

$$\|A - A_N\|_2 \leq \lambda_{k+1} + \left\| \Sigma_2^{1/2} U_2^T Q (U_1 Q)^+ \right\|_2^2$$

with $A = [U_1 \ U_2] \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} [U_1 \ U_2]^T$.

Expected value bound [Frangella, Tropp, Udell, 2021]:

$$\mathbb{E} \|A - A_N\|_2 \leq \min_{2 \leq p \leq k-2} \left(\left(1 + \frac{2(k-p)}{p-1} \right) \lambda_{k-p+1} + \frac{2e^2 k}{p^2 - 1} \sum_{j=k-p+1}^n \lambda_j \right)$$

where $\lambda_i \geq \lambda_{i+1}$ are the eigenvalues of A .

Finite Precision Error Bound

Finite precision error: $A_N - \hat{A}_N$

Assumptions:

- A is stored in precision u_p and matrix-matrix product AQ is computed in precision u_p
- All other quantities stored and computed in precision $u \ll u_p$

Finite Precision Error Bound

Finite precision error: $A_N - \hat{A}_N$

Assumptions:

- A is stored in precision u_p and matrix-matrix product AQ is computed in precision u_p
- All other quantities stored and computed in precision $u \ll u_p$

[C., Daužickaitė, 2022]: With high probability,

$$\|A_N - \hat{A}_N\|_2 \leq O(u_p)n^{5/2}\|A\|_2$$

Finite Precision Error Bound

Finite precision error: $A_N - \hat{A}_N$

Assumptions:

- A is stored in precision u_p and matrix-matrix product AQ is computed in precision u_p
- All other quantities stored and computed in precision $u \ll u_p$

[C., Daužickaitė, 2022]: With high probability,

$$\|A_N - \hat{A}_N\|_2 \leq O(u_p)n^{5/2}\|A\|_2$$

Interpretation: $\|A_N - \hat{A}_N\|_2 \gtrsim \|A - A_N\|_2$ when

$$\frac{\lambda_{k+1}}{\lambda_1} \lesssim \sqrt{nu_p}$$

Finite Precision Error Bound

Finite precision error: $A_N - \hat{A}_N$

Assumptions:

- A is stored in precision u_p and matrix-matrix product AQ is computed in precision u_p
- All other quantities stored and computed in precision $u \ll u_p$

[C., Daužickaitė, 2022]: With high probability,

$$\|A_N - \hat{A}_N\|_2 \leq O(u_p)n^{5/2}\|A\|_2$$

Interpretation: $\|A_N - \hat{A}_N\|_2 \gtrsim \|A - A_N\|_2$ when

$$\frac{\lambda_{k+1}}{\lambda_1} \lesssim \sqrt{nu_p}$$

The more approximate the low-rank representation, the lower the precision we can use!

Condition Number Bounds

Let $E = A - A_N$, $\mathcal{E} = A_N - \hat{A}_N$, and assume $(A + \mu I)$ is SPD.

Let

$$\hat{P}^{-1} = I - \hat{U}\hat{U}^T + (\hat{\lambda}_k + \mu)\hat{U}(\hat{\Theta} + \mu I)^{-1}\hat{U}^T$$

be the LMP preconditioner constructed using the mixed precision Nyström approximation $\hat{A}_N = \hat{U}\hat{\Theta}\hat{U}^T$.

Condition Number Bounds

Let $E = A - A_N$, $\mathcal{E} = A_N - \hat{A}_N$, and assume $(A + \mu I)$ is SPD.

Let

$$\hat{P}^{-1} = I - \hat{U}\hat{U}^T + (\hat{\lambda}_k + \mu)\hat{U}(\hat{\Theta} + \mu I)^{-1}\hat{U}^T$$

be the LMP preconditioner constructed using the mixed precision Nyström approximation $\hat{A}_N = \hat{U}\hat{\Theta}\hat{U}^T$.

Then

$$\max \left\{ 1, \frac{\hat{\lambda}_k + \mu - \|\mathcal{E}\|_2}{\mu + \lambda_{\min}(A)} \right\} \leq \kappa(\hat{P}^{-1/2}(A + \mu I)\hat{P}^{-1/2}) \leq 1 + \frac{\hat{\lambda}_k + \|E\|_2 + 2\|\mathcal{E}\|_2}{\mu - \|\mathcal{E}\|_2}$$

where the upper bound holds if $\mu > \|\mathcal{E}\|_2$.

Regardless of this constraint, if A is positive definite, then

$$\kappa(\hat{P}^{-1/2}(A + \mu I)\hat{P}^{-1/2}) \leq (\hat{\lambda}_k + \mu + \|E\|_2 + \|\mathcal{E}\|_2) \left(\frac{1}{\hat{\lambda}_k + \mu} + \frac{\|\mathcal{E}\|_2 + 1}{\lambda_{\min}(A) + \mu} \right).$$

Condition Number Bounds

Let $E = A - A_N$, $\mathcal{E} = A_N - \hat{A}_N$, and assume $(A + \mu I)$ is SPD.

Let

$$\hat{P}^{-1} = I - \hat{U}\hat{U}^T + (\hat{\lambda}_k + \mu)\hat{U}(\hat{\Theta} + \mu I)^{-1}\hat{U}^T$$

be the LMP preconditioner constructed using the mixed precision Nyström approximation $\hat{A}_N = \hat{U}\hat{\Theta}\hat{U}^T$.

Then

If $\mathcal{E} = 0$, reduces to bounds of [Frangella, Tropp, Udell, 2021] for exact case.

$$\max \left\{ 1, \frac{\hat{\lambda}_k + \mu - \|\mathcal{E}\|_2}{\mu + \lambda_{\min}(A)} \right\} \leq \kappa(\hat{P}^{-1/2}(A + \mu I)\hat{P}^{-1/2}) \leq 1 + \frac{\hat{\lambda}_k + \|E\|_2 + 2\|\mathcal{E}\|_2}{\mu - \|\mathcal{E}\|_2}$$

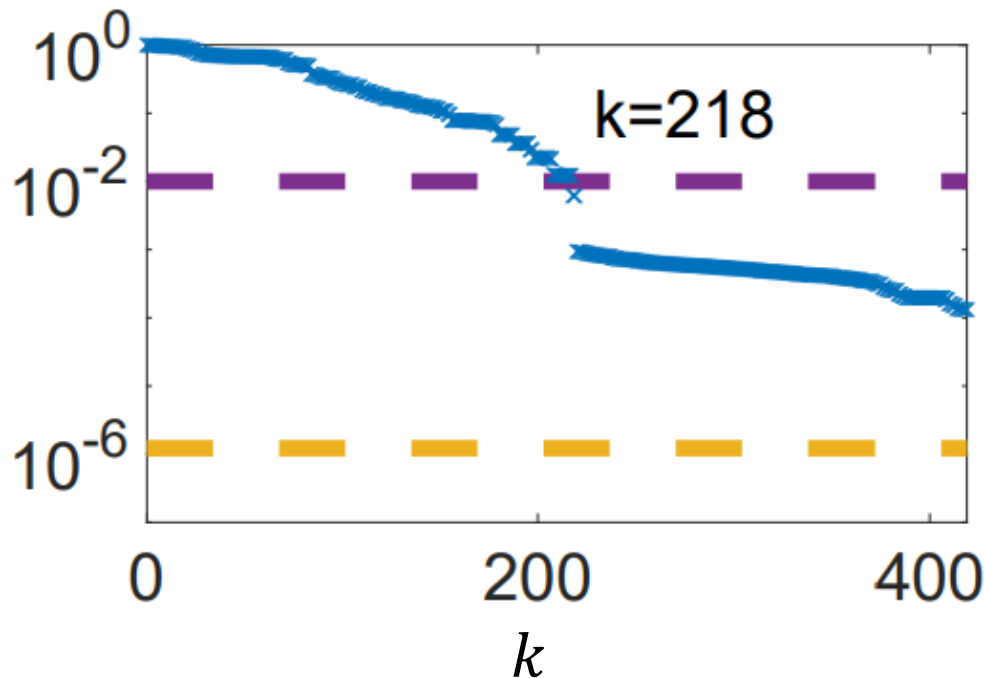
where the upper bound holds if $\mu > \|\mathcal{E}\|_2$.

Regardless of this constraint, if A is positive definite, then

$$\kappa(\hat{P}^{-1/2}(A + \mu I)\hat{P}^{-1/2}) \leq (\hat{\lambda}_k + \mu + \|E\|_2 + \|\mathcal{E}\|_2) \left(\frac{1}{\hat{\lambda}_k + \mu} + \frac{\|\mathcal{E}\|_2 + 1}{\lambda_{\min}(A) + \mu} \right).$$

Numerical Experiment

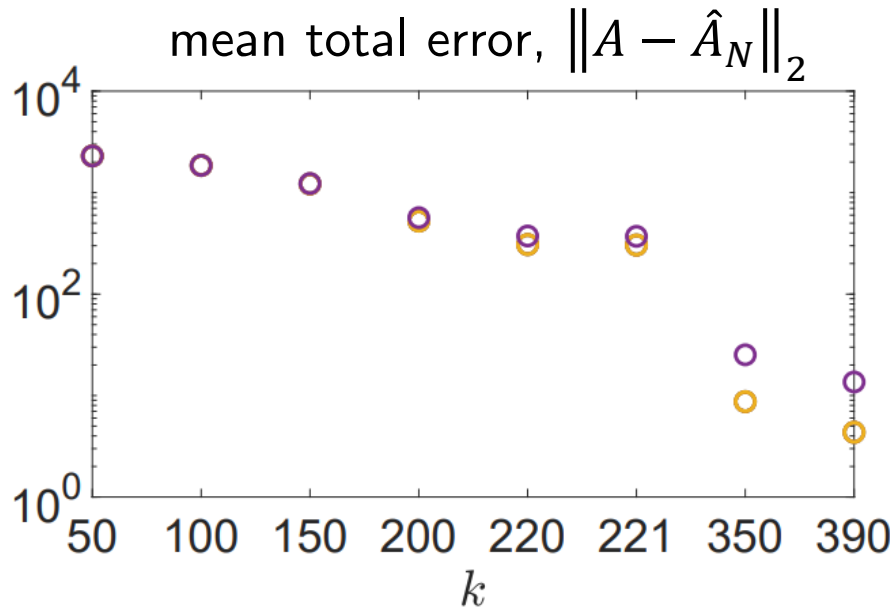
Matrix: bcsstm07, $n = 420$



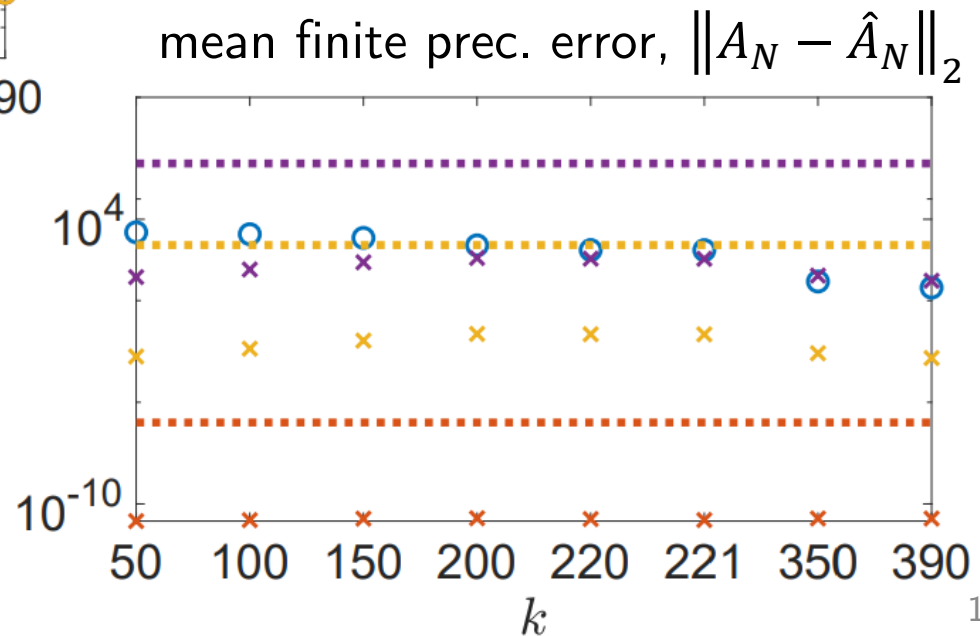
- λ_{k+1}/λ_1
- $\sqrt{n}u_p, u_p = \text{half}$
- $\sqrt{n}u_p, u_p = \text{single}$

Numerical Experiment

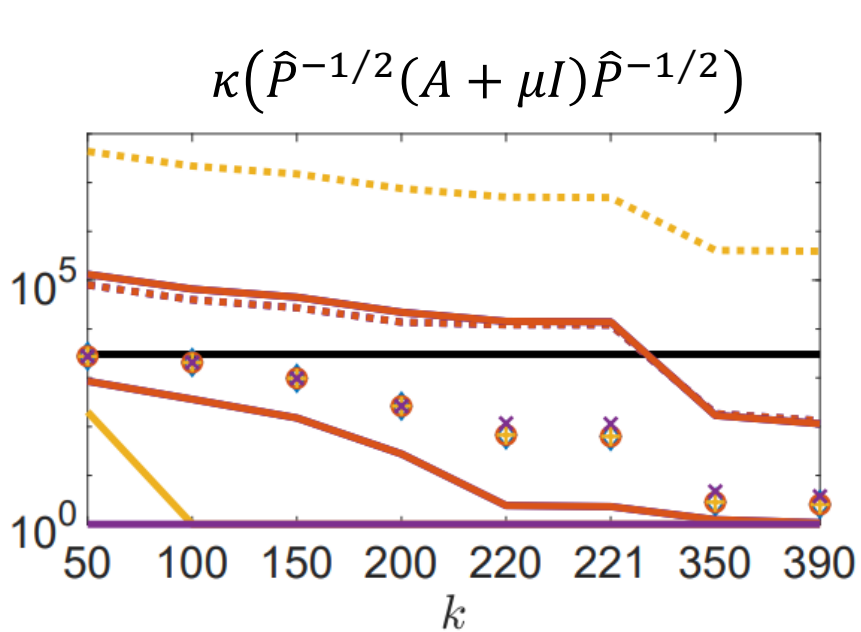
Matrix: bcsstm07, $n = 420$



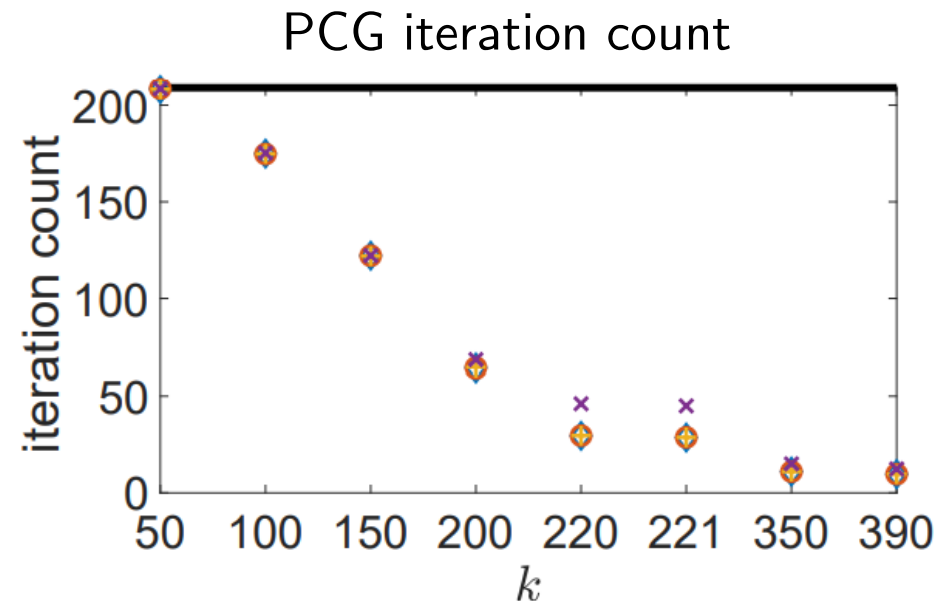
- exact
- mixed, $u_p = \text{half}$
- mixed, $u_p = \text{single}$
- mixed, $u_p = \text{double}$



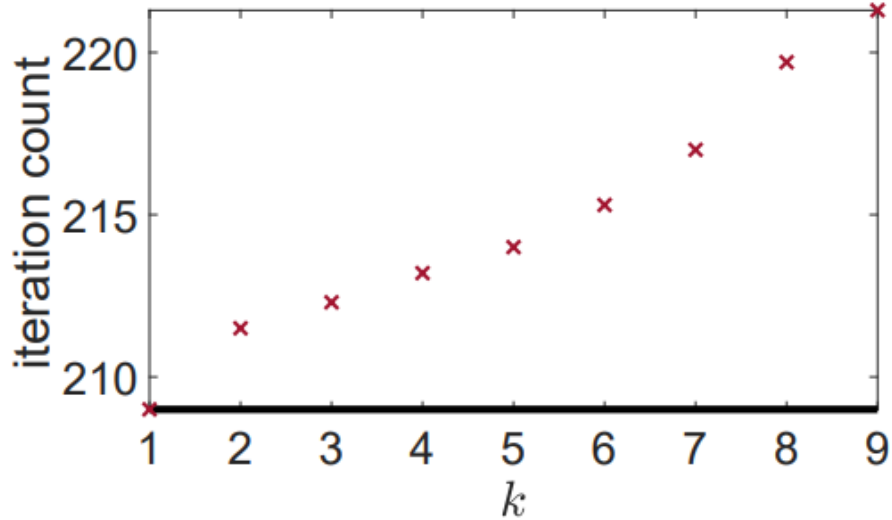
Numerical Experiment



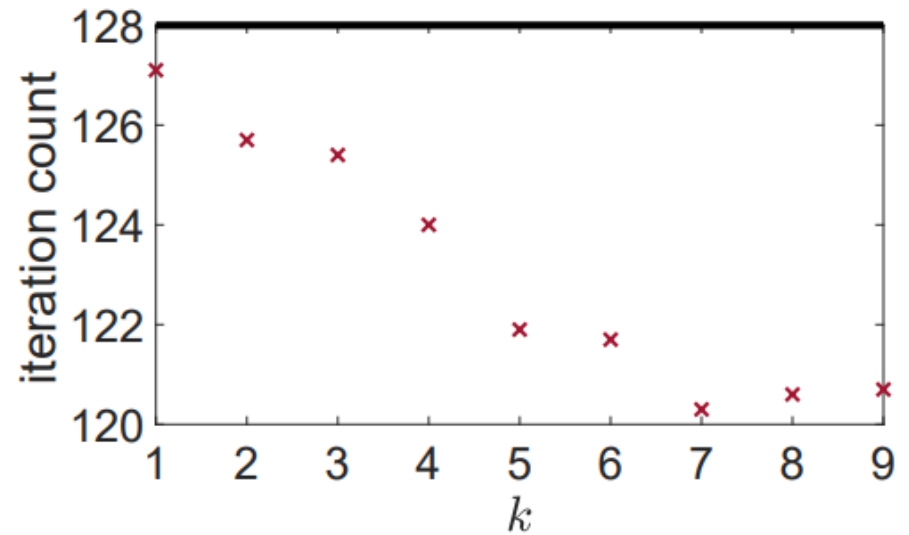
- unpreconditioned
- exact
- mixed, $u_p = \text{half}$
- mixed, $u_p = \text{single}$
- mixed, $u_p = \text{double}$



Quarter precision?



bcsstm07, iteration count



Journals, iteration count

Other Recent Work in Mixed Precision NLA

[Oktaý and C. "Multistage mixed precision iterative refinement." *NLAA* 29.4 (2022): e2434]

[Oktaý and C. "Mixed Precision GMRES-based Iterative Refinement with Recycling." *arXiv preprint arXiv:2201.09827* (2022)]

[C. and Khan. "Mixed Precision Iterative Refinement with Sparse Approximate Inverse Preconditioning." *arXiv preprint arXiv:2202.10204* (2022)]

[C., Gergelits, and Yamazaki. "Mixed precision s-step Lanczos and conjugate gradient algorithms." *NLAA* 29.3 (2022): e2425]

[C, Kelley, and Yamazaki. "Mixed Precision s-step Conjugate Gradient with Residual Replacement on GPUs", In Proc. IPDPS (2022)]

Summary and Takeaway

- We now have a multi-precision ecosystem
- Huge opportunities for using mixed precision in matrix computations
- But also big challenges!

Happy

$$(-1)^0 \times 2^{(132-127)} \times 1.875^{\text{th}}$$

Birthday, Nick!

Contact: carson@karlin.mff.cuni.cz
www.karlin.mff.cuni.cz/~carson/