Mixed Precision Randomized Nyström Approximation

Erin C. Carson

Charles University

ILAS 2023

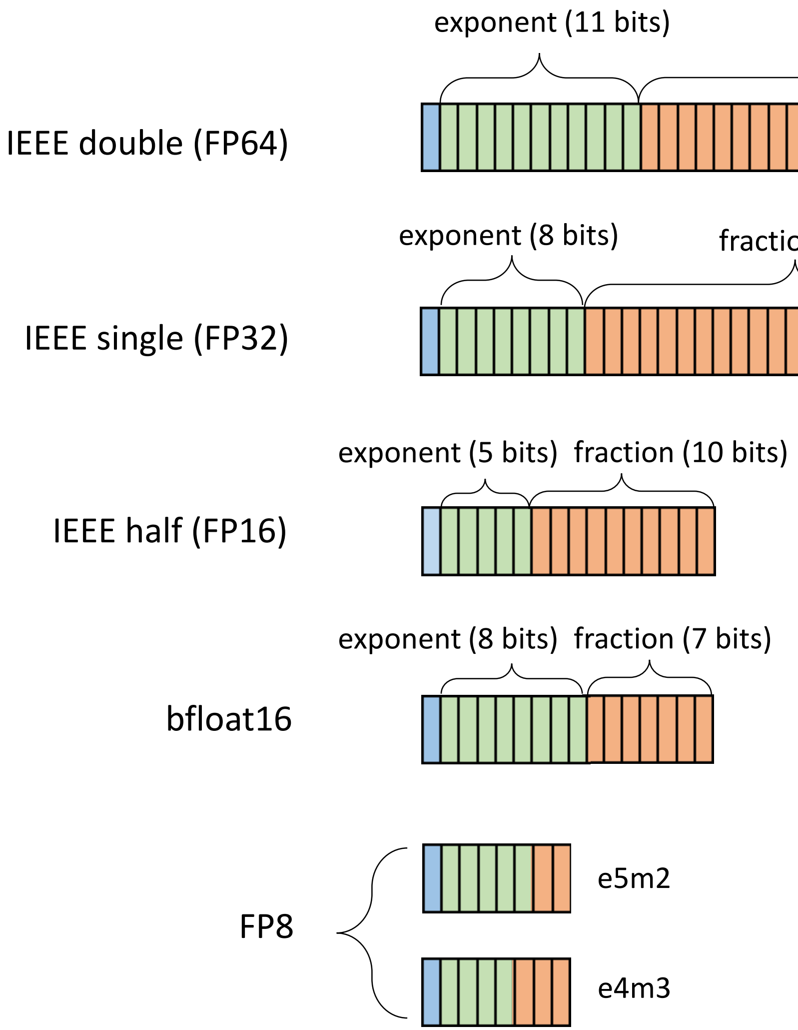
FACULTY OF MATHEMATICS AND PHYSICS Charles University

Co-funded by the European Union

We acknowledge funding from ERC Starting Grant No. 101075632 and the Exascale Computing 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.

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]

For survey and references, see [Abdelfattah et al., IJHPC, 2021], [Higham and Mary, 2022]

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 $\boldsymbol{U} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{k}}$ **are** $\boldsymbol{k}$ **approximate eigenvectors of** $\boldsymbol{A}$ and $U^T U = I$, $\boldsymbol{\Theta}$ **is diagonal with approximations to eigenvalues of** $\boldsymbol{A}$, 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:

$$A_N = (A\Omega)(\Omega^T A\Omega)^\dagger (A\Omega)^T$$

where Ω is an $n \times k$ sampling matrix

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.

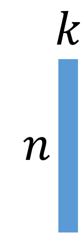
Randomized Nyström Approximation

[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

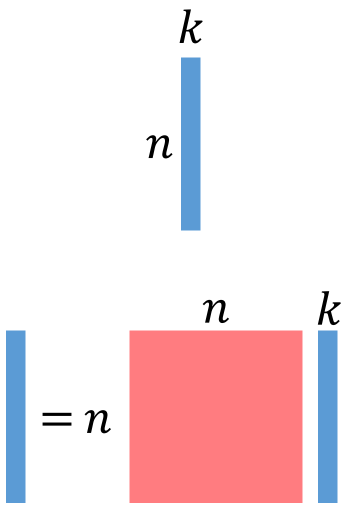
[Tropp et al., 2017]

Given sym. PSD matrix A, target rank k

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

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

$\boldsymbol{Y} = \boldsymbol{AQ}$



Randomized Nyström Approximation

[Tropp et al., 2017]

Given sym. PSD matrix A, target rank k

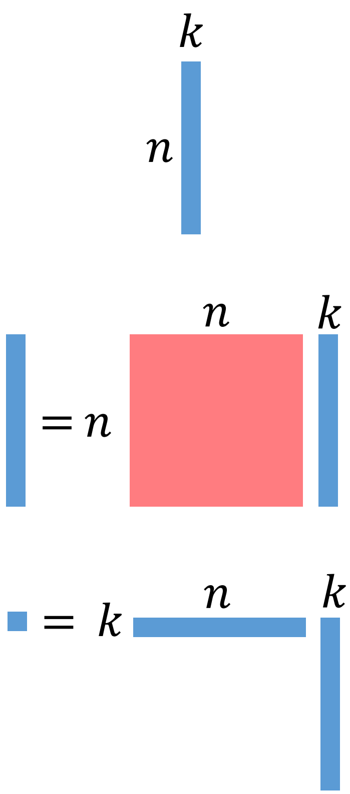
$G = \text{randn}(n, k)$

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

$\boldsymbol{Y = AQ}$

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

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

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

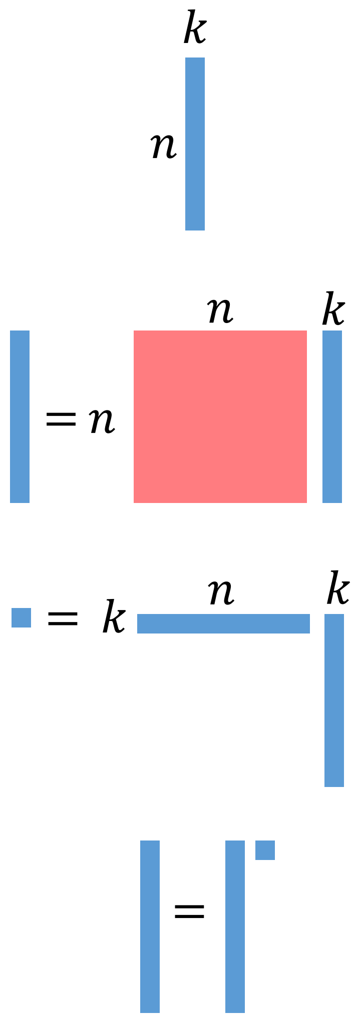
$\boldsymbol{Y} = \boldsymbol{AQ}$

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

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

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

$\boldsymbol{Y = AQ}$

Compute shift ν; $Y_\nu = Y + \nu Q$

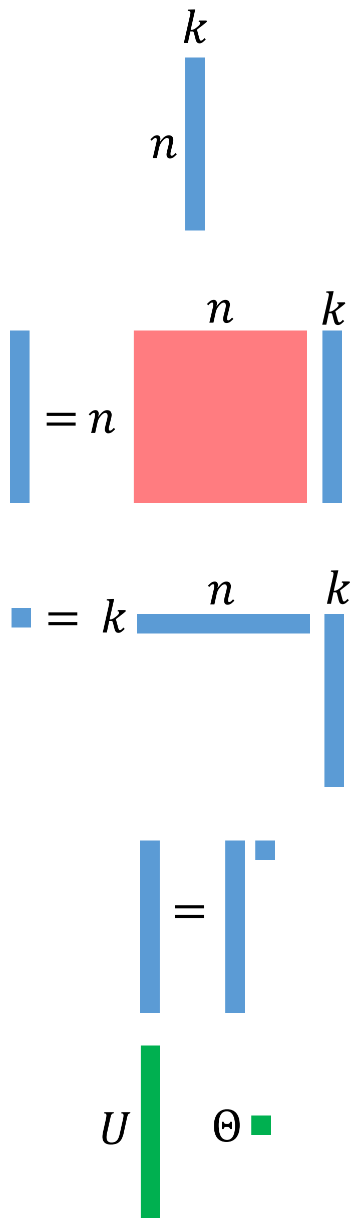
$B = Q^T Y_\nu$

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

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

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

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

$\boldsymbol{Y = AQ}$

Compute shift ν; $Y_\nu = Y + \nu Q$

$B = Q^T Y_\nu$

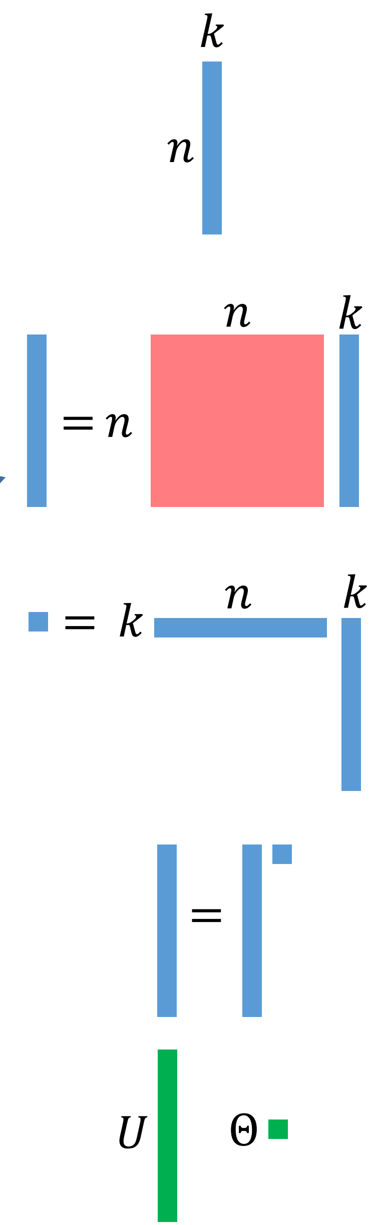
$C = \text{chol}((B + B^T)/2)$

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?



Randomized Nyström Approximation

[Tropp et al., 2017]

Given sym. PSD matrix A, target rank k

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

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

$\boldsymbol{Y = AQ}$ **(precision $\boldsymbol{u_p}$)**

Compute shift ν; $Y_\nu = Y + \nu Q$ (precision u)

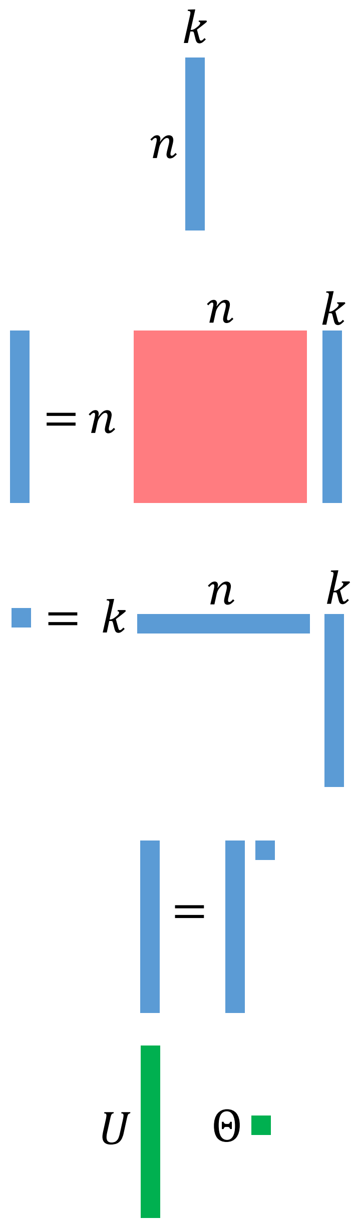
$B = Q^T Y_\nu$ (precision u)

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

Solve $F = Y_\nu / C$ (precision u)

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

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



Error Bounds

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

exact Nyström approximation (A_N)

Nyström approximation computed in finite precision ($\hat{A}_N$)

Error Bounds

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

Error Bounds

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

Deterministic bound [Gittens, Mahoney, 2016]:

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

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]:

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

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

Goal: 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 $u \ll u_p$

Preliminaries

[Theorem 2.13, Davidson and Szarek, 2001]:

Let G be an $n \times k$ Gaussian matrix with $n > k$. Then for every $t \geq 0$,

$$P\left\{\|G\|_2 \geq n^{1/2} + k^{1/2} + t\right\} \leq e^{-t^2/2}.$$

Preliminaries

[Theorem 2.13, Davidson and Szarek, 2001]:

Let G be an $n \times k$ Gaussian matrix with $n > k$. Then for every $t \geq 0$,

$$P\left\{\|G\|_2 \geq n^{1/2} + k^{1/2} + t\right\} \leq e^{-t^2/2}.$$

[Theorem 1.2, Szarek, 1991]:

If G is an $n \times n$ Gaussian matrix,

$$P\left\{\sigma_{min}(G) < \frac{\alpha}{n^{1/2}}\right\} \leq c_1 \alpha$$

for universal constant c_1.

Finite Precision Error

The computed Y satisfies

$$\hat{Y} = AQ + \Delta, \qquad \text{where } \|\Delta\|_2 \leq n^{\frac{1}{2}} \tilde{\gamma}_n^{(p)} \|A\|_2$$

where $\tilde{\gamma}_n^{(p)} = cnu_p/(1 - cnu_p)$ for a small constant c independent of n.

Finite Precision Error

The computed Y satisfies

$$\hat{Y} = AQ + \Delta, \qquad \text{where } \|\Delta\|_2 \leq n^{\frac{1}{2}} \tilde{\gamma}_n^{(p)} \|A\|_2$$

where $\tilde{\gamma}_n^{(p)} = cnu_p/(1 - cnu_p)$ for a small constant c independent of n.

$$\hat{A}_N = (AQ + \Delta)\left(Q^T AQ + \tilde{\Delta}\right)^{-1}(AQ + \Delta)^T \qquad \text{where } \tilde{\Delta} = Q^T \Delta + \Delta^T Q$$

Finite Precision Error

The computed Y satisfies

$$\hat{Y} = AQ + \Delta, \qquad \text{where } \|\Delta\|_2 \leq n^{\frac{1}{2}}\tilde{\gamma}_n^{(p)}\|A\|_2$$

where $\tilde{\gamma}_n^{(p)} = cnu_p/(1 - cnu_p)$ for a small constant c independent of n.

$$\begin{aligned}\hat{A}_N &= (AQ + \Delta)\left(Q^T AQ + \widetilde{\Delta}\right)^{-1}(AQ + \Delta)^T \qquad \text{where } \widetilde{\Delta} = Q^T\Delta + \Delta^T Q \\ &= A_N + AQ(Q^T AQ)^{-1}\Delta^T + \Delta(Q^T AQ)^{-1}(AQ)^T \\ &\quad -\frac{1}{2}\left(AQ(Q^T AQ)^{-1}\widetilde{\Delta}(Q^T AQ)^{-1}(AQ)^T\right)\end{aligned}$$

Finite Precision Error

The computed Y satisfies

$$\hat{Y} = AQ + \Delta, \qquad \text{where } \|\Delta\|_2 \leq n^{\frac{1}{2}} \tilde{\gamma}_n^{(p)} \|A\|_2$$

where $\tilde{\gamma}_n^{(p)} = cnu_p/(1 - cnu_p)$ for a small constant c independent of n.

$$\begin{aligned}\hat{A}_N &= (AQ + \Delta)\left(Q^T AQ + \widetilde{\Delta}\right)^{-1}(AQ + \Delta)^T \qquad \text{where } \widetilde{\Delta} = Q^T\Delta + \Delta^T Q \\ &= A_N + AQ(Q^T AQ)^{-1}\Delta^T + \Delta(Q^T AQ)^{-1}(AQ)^T \\ &\quad -\tfrac{1}{2}\left(AQ(Q^T AQ)^{-1}\widetilde{\Delta}(Q^T AQ)^{-1}(AQ)^T\right)\end{aligned}$$

Finite precision error:

$$\left\|\hat{A}_N - A_N\right\|_2 \leq (2\,\|AQ(Q^T AQ)^{-1}\|_2 + \|AQ(Q^T AQ)^{-1}\|_2^2)\|\Delta\|_2$$

Finite Precision Error

The computed Y satisfies

$$\hat{Y} = AQ + \Delta, \qquad \text{where } \|\Delta\|_2 \leq n^{\frac{1}{2}} \tilde{\gamma}_n^{(p)} \|A\|_2$$

where $\tilde{\gamma}_n^{(p)} = cnu_p/(1 - cnu_p)$ for a small constant c independent of n.

$$\begin{aligned} \hat{A}_N &= (AQ + \Delta)\big(Q^T AQ + \tilde{\Delta}\big)^{-1}(AQ + \Delta)^T \qquad \text{where } \tilde{\Delta} = Q^T\Delta + \Delta^T Q \\ &= A_N + AQ(Q^TAQ)^{-1}\Delta^T + \Delta(Q^TAQ)^{-1}(AQ)^T \\ &\quad -\frac{1}{2}\big(AQ(Q^TAQ)^{-1}\tilde{\Delta}(Q^TAQ)^{-1}(AQ)^T\big) \end{aligned}$$

Finite precision error:

$$\big\|\hat{A}_N - A_N\big\|_2 \leq (2\,\|AQ(Q^TAQ)^{-1}\|_2 + \|AQ(Q^TAQ)^{-1}\|_2^2)\|\Delta\|_2$$

weighted pseudoinverse

Weighted Pseudoinverse

$$X_D^{\dagger} = DX(X^T DX)^{\dagger}$$

[Stewart, 1989]: Let X have full column rank, let U be an orthonormal basis for the column space of X, and let D be diagonal with positive diagonal elements. Then

$$\sup_{D \in \mathcal{D}_+} \|X_D^{\dagger}\| \leq \rho^{-1} \|X^{\dagger}\|$$

and

$$\rho \leq \min \inf_{+}(U_I)$$

where U_I denotes any submatrix formed from a set of rows of U.

Weighted Pseudoinverse

$$X_D^\dagger = DX(X^T DX)^\dagger$$

[Stewart, 1989]: Let X have full column rank, let U be an orthonormal basis for the column space of X, and let D be diagonal with positive diagonal elements. Then

$$\sup_{D \in \mathcal{D}_+} \|X_D^\dagger\| \leq \rho^{-1} \|X^\dagger\|$$

and

$$\rho \leq \min \inf_+(U_I)$$

where U_I denotes any submatrix formed from a set of rows of U.

[O'Leary, 1990]: $\rho = \min \inf_+(U_I)$

Other related work, e.g., [Forsgren, 1996].

Weighted Pseudoinverse

Lemma: Let A be an $n \times n$ symmetric positive semidefinite matrix and let $X_A^\dagger = AX(X^T AX)^\dagger$ where X is $n \times k$ with full column rank. Then

$$\left\| X_A^\dagger \right\|_2 \leq \frac{\lambda_{max}^{1/2}(A)}{\sigma_{min}(X^T A^{1/2})}.$$

Weighted Pseudoinverse

Lemma: Let A be an $n \times n$ symmetric positive semidefinite matrix and let $X_A^\dagger = AX(X^TAX)^\dagger$ where X is $n \times k$ with full column rank. Then

$$\|X_A^\dagger\|_2 \leq \frac{\lambda_{max}^{1/2}(A)}{\sigma_{min}(X^TA^{1/2})}.$$

If X has standard Gaussian entries, then

$$\|X_A^\dagger\|_2 \leq \frac{\kappa(A_k)^{1/2}k^{1/2}}{\alpha}$$

with failure probability at most $c_1\alpha$.

Back to Finite Precision Error Bound

$$\left\|\hat{A}_N - A_N\right\|_2 \le (2\,\|AQ(Q^TAQ)^{-1}\|_2 + \|AQ(Q^TAQ)^{-1}\|_2^2)\|\Delta\|_2$$

Let $G = QR$. Then

$$\|AQ(Q^TAQ)^{-1}\|_2 \le \frac{\kappa(A_k)^{1/2}\|G\|_2}{\sigma_{min}(W_1^TG)}$$

where the columns of W_1 are the leading k eigenvectors of A.

Back to Finite Precision Error Bound

$$\left\|\hat{A}_N - A_N\right\|_2 \leq (2\,\|AQ(Q^TAQ)^{-1}\|_2 + \|AQ(Q^TAQ)^{-1}\|_2^2)\|\Delta\|_2$$

Let $G = QR$. Then

$$\|AQ(Q^TAQ)^{-1}\|_2 \leq \frac{\kappa(A_k)^{1/2}\|G\|_2}{\sigma_{min}(W_1^TG)}$$

where the columns of W_1 are the leading k eigenvectors of A.

Using results on Gaussian matrices gives:

[C., Daužickaitė, 2022]: With failure probability at most $e^{-t^2/2} + c_1\alpha$,

$$\left\|A_N - \hat{A}_N\right\|_2 \lesssim \alpha^{-1} n^{1/2} k \left(n^{1/2} + k^{1/2} + t\right)^2 u_p \|A\|_2 \kappa(A_k).$$

Back to Finite Precision Error Bound

$$\left\|\hat{A}_N - A_N\right\|_2 \leq (2\,\|AQ(Q^TAQ)^{-1}\|_2 + \|AQ(Q^TAQ)^{-1}\|_2^2)\|\Delta\|_2$$

Let $G = QR$. Then

$$\|AQ(Q^TAQ)^{-1}\|_2 \leq \frac{\kappa(A_k)^{1/2}\|G\|_2}{\sigma_{min}(W_1^T G)}$$

where the columns of W_1 are the leading k eigenvectors of A.

Using results on Gaussian matrices gives:

[C., Daužickaitė, 2022]: With failure probability at most $e^{-t^2/2} + c_1\alpha$,

$$\left\|A_N - \hat{A}_N\right\|_2 \lesssim \alpha^{-1} n^{1/2} k \left(n^{1/2} + k^{1/2} + t\right)^2 u_p \|A\|_2 \kappa(A_k).$$

Interpretation: Likely that $\left\|A_N - \hat{A}_N\right\|_2 \gtrsim \|A - A_N\|_2$ when

$$\frac{\lambda_{k+1}}{\lambda_1} \lesssim \sqrt{n}\, u_p$$

Back to Finite Precision Error Bound

$$\left\|\hat{A}_N - A_N\right\|_2 \leq (2\,\|AQ(Q^TAQ)^{-1}\|_2 + \|AQ(Q^TAQ)^{-1}\|_2^2)\|\Delta\|_2$$

Let $G = QR$. Then

$$\|AQ(Q^TAQ)^{-1}\|_2 \leq \frac{\kappa(A_k)^{1/2}\|G\|_2}{\sigma_{min}(W_1^T G)}$$

where the columns of W_1 are the leading k eigenvectors of A.

Using results on Gaussian matrices gives:

[C., Daužickaitė, 2022]: With failure probability at most $e^{-t^2/2} + c_1\alpha$,

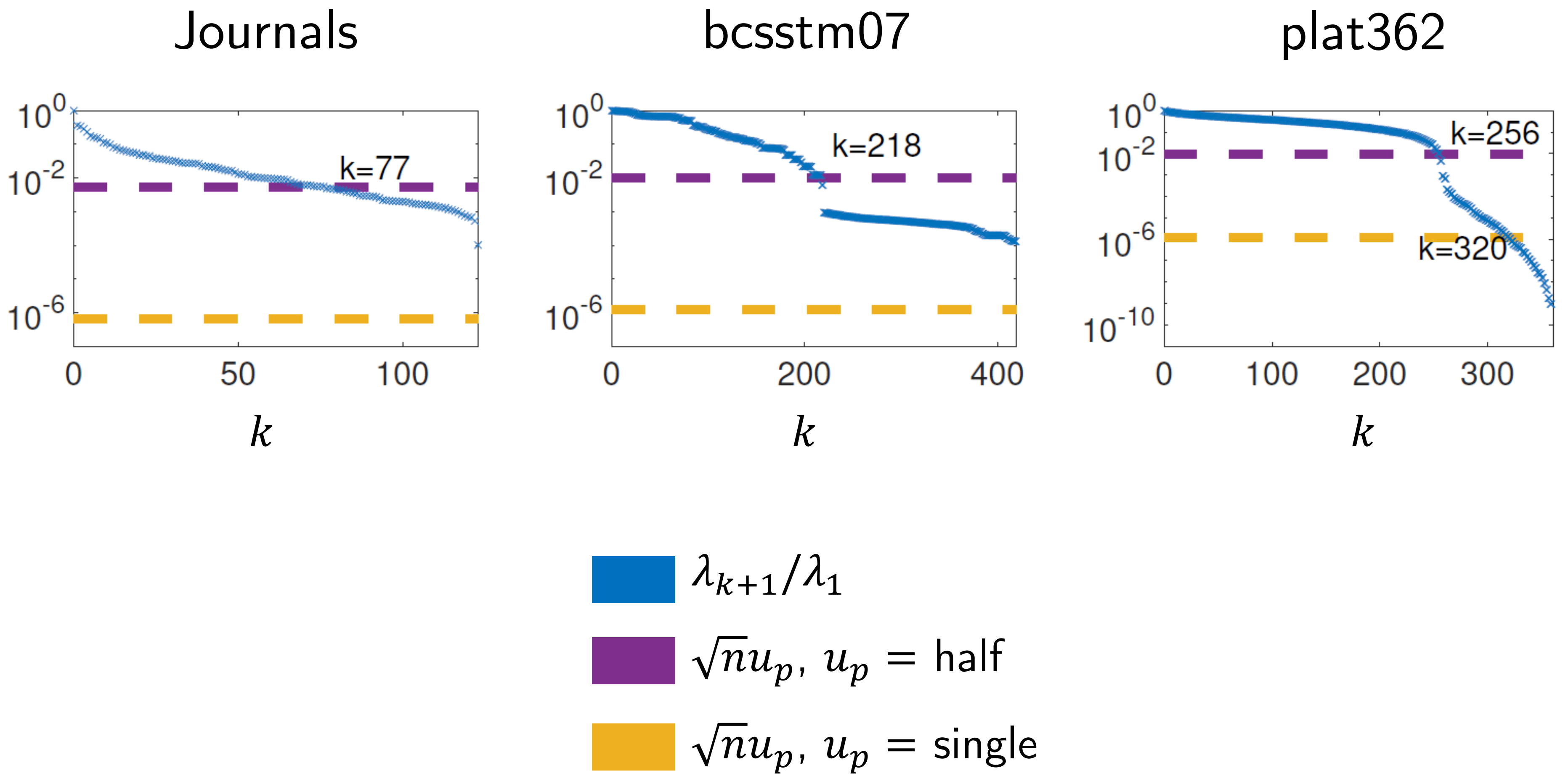
$$\left\|A_N - \hat{A}_N\right\|_2 \lesssim \alpha^{-1} n^{1/2} k \left(n^{1/2} + k^{1/2} + t\right)^2 u_p \|A\|_2 \kappa(A_k).$$

Interpretation: Likely that $\left\|A_N - \hat{A}_N\right\|_2 \gtrsim \|A - A_N\|_2$ when

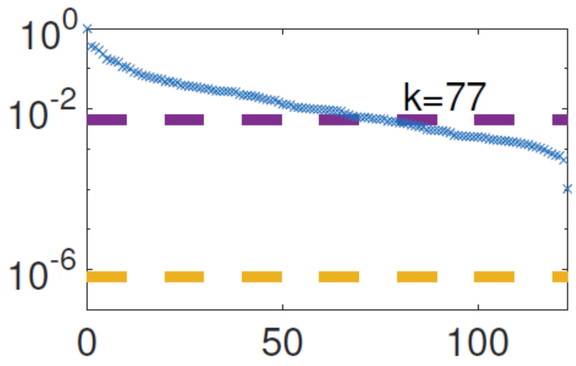
$$\frac{\lambda_{k+1}}{\lambda_1} \lesssim \sqrt{n}\, u_p$$

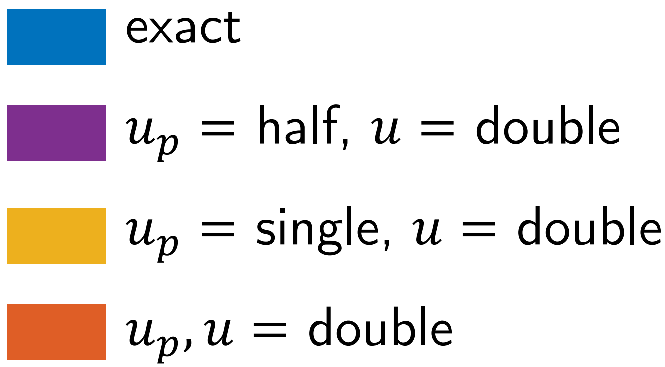
The worse the low-rank representation, the lower the precision we can use!

Numerical Experiment

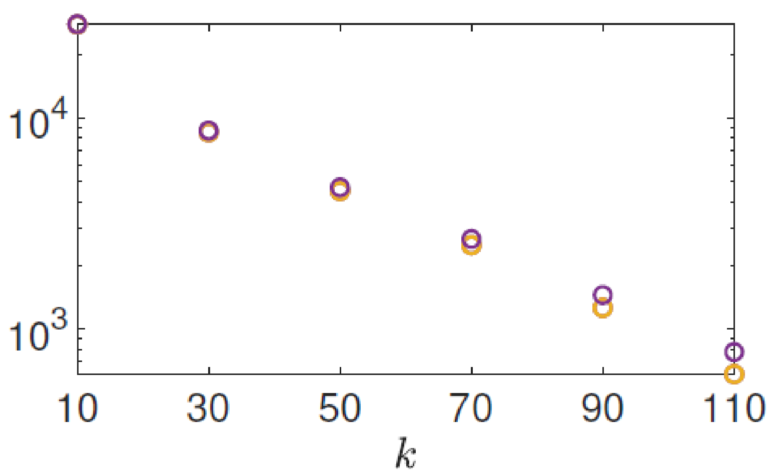


Numerical Experiment: Journals

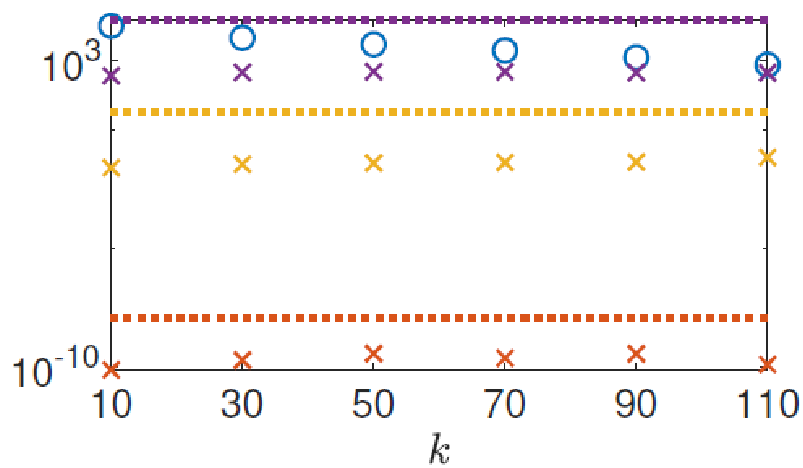




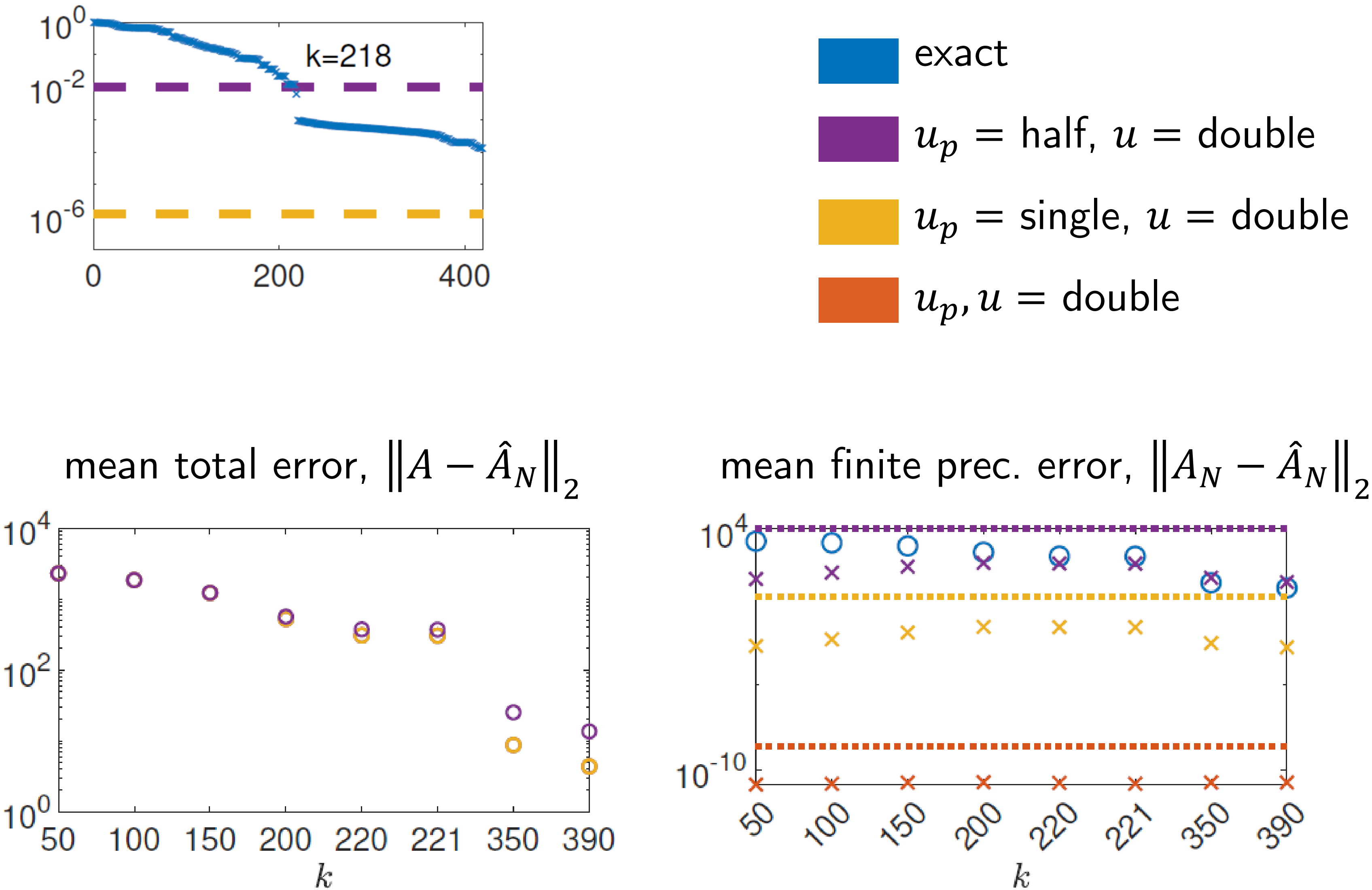
mean total error, $\|A - \hat{A}_N\|_2$



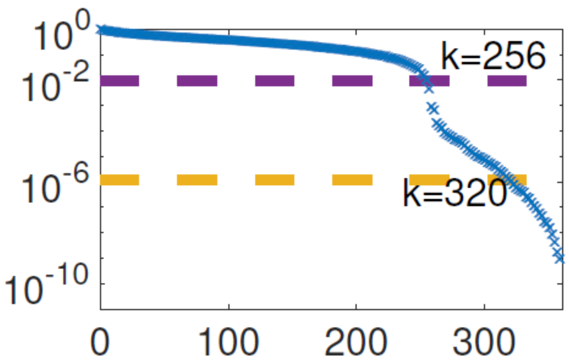
mean finite prec. error, $\|A_N - \hat{A}_N\|_2$



Numerical Experiment: bcsstm07



Numerical Experiment: plat362



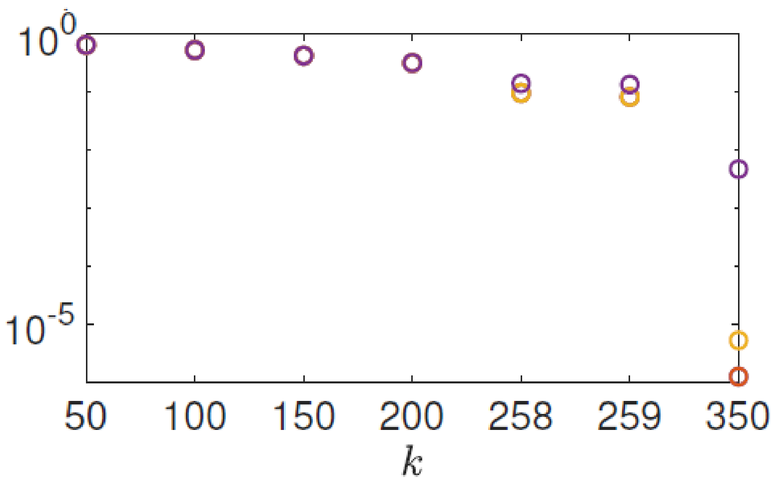
exact

u_p = half, u = double

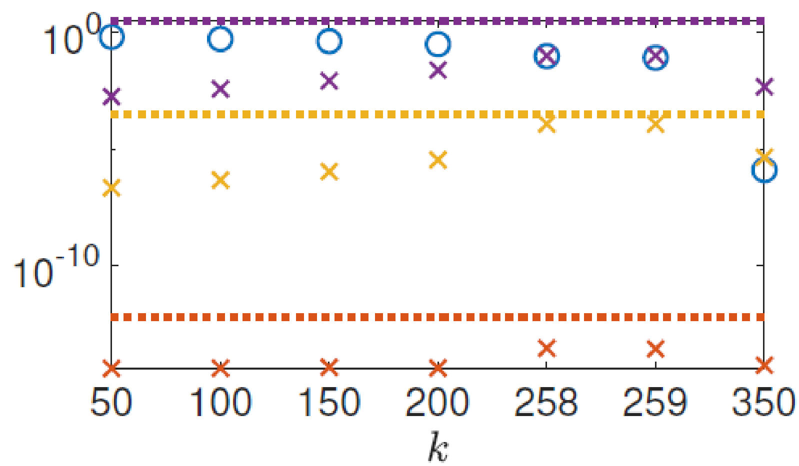
u_p = single, u = double

u_p, u = double

mean total error, $\|A - \hat{A}_N\|_2$



mean finite prec. error, $\|A_N - \hat{A}_N\|_2$



Condition Number Bounds

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

Let

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

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

Condition Number Bounds

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

Let

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

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

Then

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

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

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

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

Condition Number Bounds

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

Let

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

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

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

Then

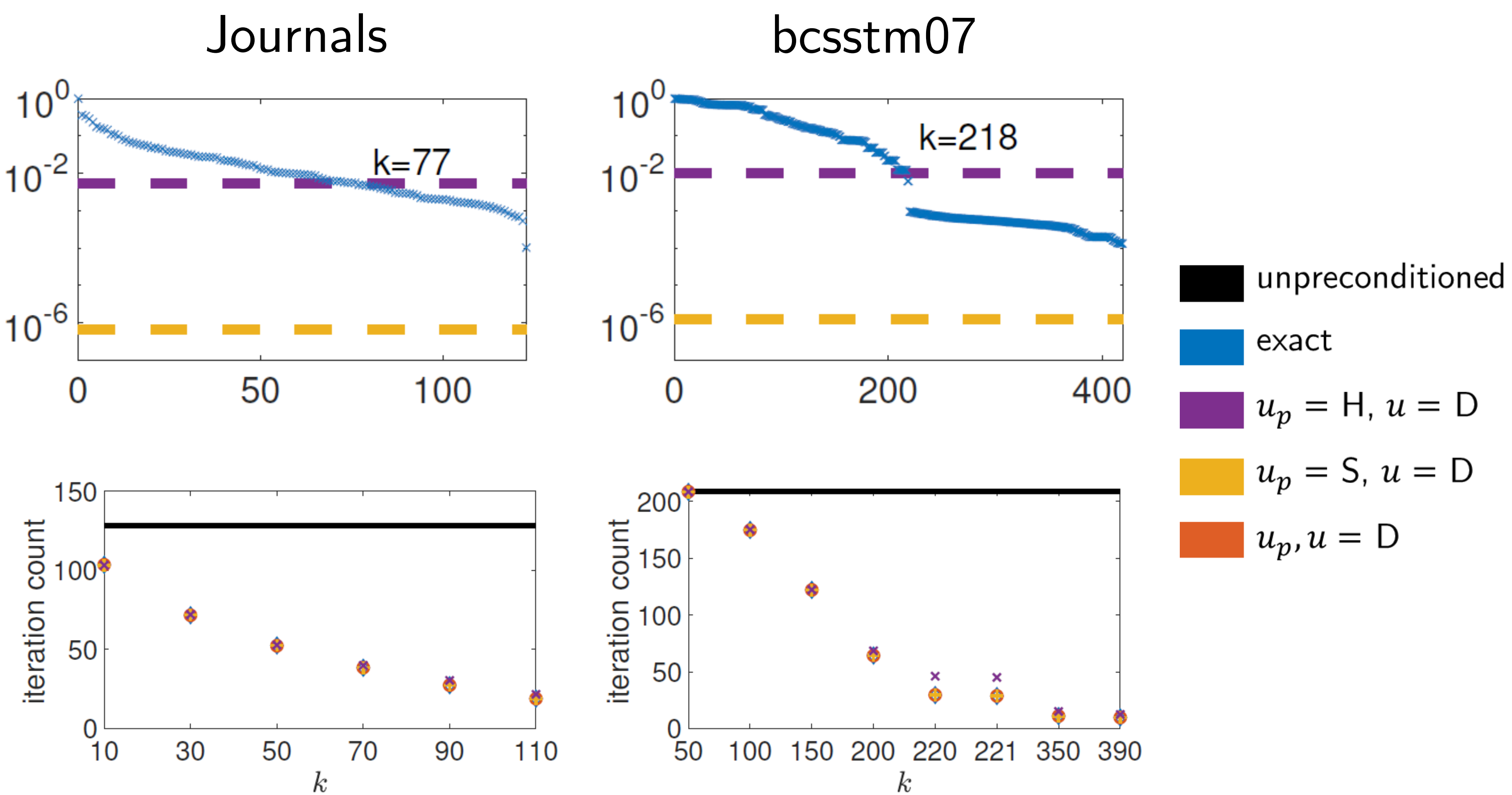
$$\max\left\{1, \frac{\hat{\lambda}_k + \mu - \|\mathcal{E}\|_2}{\mu + \lambda_{min}(A)}\right\} \leq \kappa\left(\hat{P}^{-1/2}(A + \mu I)\hat{P}^{-1/2}\right) \leq 1 + \frac{\hat{\lambda}_k + \|E\|_2 + 2\|\mathcal{E}\|_2}{\mu - \|\mathcal{E}\|_2}$$

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

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

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

Numerical Experiment



https://github.com/dauzickaite/mpNystrom

Takeaway

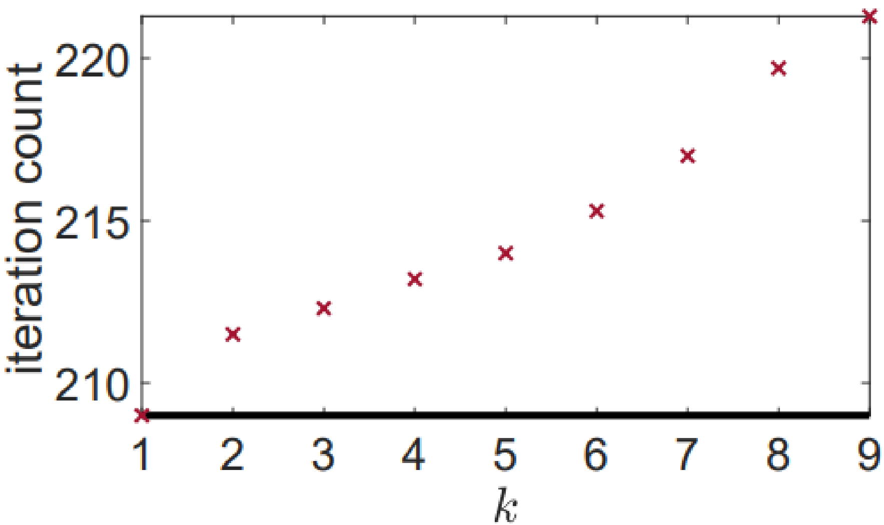
- Lots of excited related and ongoing work
  - [Connolly, Higham, Pranesh, 2022]: Mixed precision randomized SVD
  - [Meier, Nakatsukasa, Townsend, Webb, 2023]: Finite precision analysis of Blendenpik-type preconditioning in LSQR
  - [Georgiou, Boutsikas, Drineas, Anzt, 2023]: Mixed precision randomized preconditioner for LSQR on GPUs
  - Ongoing with Ieva Daužickaitė: Analysis of randomized preconditioners for GMRES-based iterative refinement for least squares

- In general, big opportunity for *combining forms of inexactness* (e.g., low rank approximation + low precision computation)

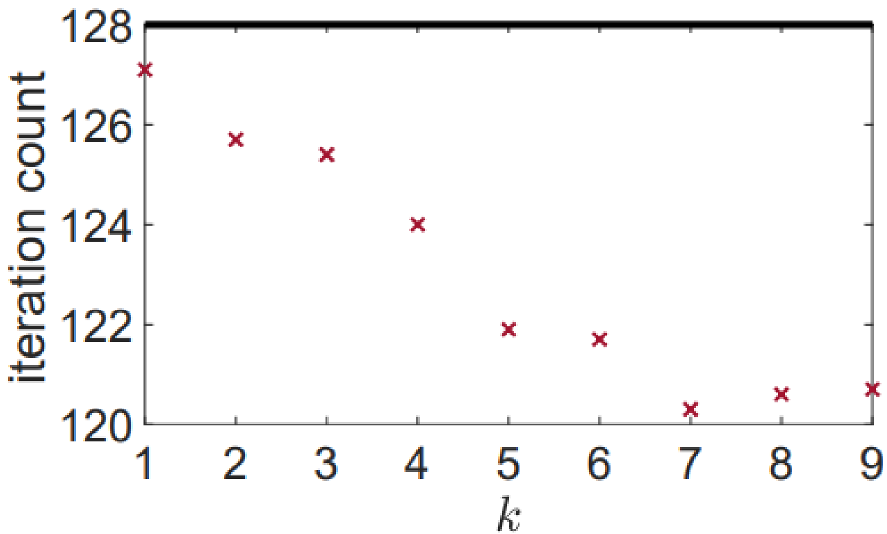
Thank You!

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

Quarter precision?



bcsstm07, iteration count



Journals, iteration count

https://github.com/dauzickaite/mpNystrom