Preconditioning without a preconditioner using randomized block KSMs

Tyler Chen (Global Technology & Applied Research, JPMorganChase)

Background

Solving the regularized linear system

$$\mathbf{A}_\mu \mathbf{x} = \mathbf{b}, \quad \mathbf{A}_\mu := \mathbf{A} + \mu \mathbf{I}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is symmetric positive definite and $\mu \geq 0$ is a critical task across the computational sciences.

Krylov Subspace Methods (KSM) make use of the Krylov subspace

$$\mathcal{K}_t(\mathbf{A}, \mathbf{b}) := \operatorname{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{t-1}\mathbf{b}\}. \tag{2}$$

Nyström preconditioning

If $\mathbf{A}$ is poorly conditioned due to the presence of r eigenvalues much larger than the remaining $n-r$ eigenvalues, then we might hope to learn a good approximation of the top r eigenvalues and ''correct'' this ill-conditioning. In particular, one can form the preconditioner

$$\mathbf{P}_\mu := \frac{1}{\theta + \mu}\mathbf{U}(\mathbf{D} + \mu\mathbf{I})\mathbf{U}^\mathsf{T} + (\mathbf{I} - \mathbf{U}\mathbf{U}^\mathsf{T}), \tag{3}$$

where $\theta > 0$ is a parameter that must be chosen along with the factorization $\mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}$. It is not hard to verify that

$$\mathbf{P}_\mu^{-1} = (\theta + \mu)\mathbf{U}(\mathbf{D} + \mu\mathbf{I})^{-1}\mathbf{U}^\mathsf{T} + (\mathbf{I} - \mathbf{U}\mathbf{U}^\mathsf{T}). \tag{4}$$

In particular, it's reasonable to take $\mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}$ as the eigendecomposition of the Nyström approximation

$$\mathbf{A}\langle \mathbf{K}_s \rangle := (\mathbf{A}\mathbf{K}_s)(\mathbf{K}_s^\mathsf{T}\mathbf{A}\mathbf{K}_s)^\dagger(\mathbf{K}_s^\mathsf{T}\mathbf{A}), \tag{5}$$

where $\mathbf{\Omega} \in \mathbb{R}^{d\times \ell}$ is a matrix of independent standard normal random variables and

$$\mathbf{K}_s := [\mathbf{\Omega}\ \mathbf{A}\mathbf{\Omega}\ \cdots\ \mathbf{A}^{s-1}\mathbf{\Omega}] \in \mathbb{R}^{d\times(s\ell)}. \tag{6}$$

This variant of the Nyström approximation is among the most powerful randomized low-rank approximation algorithms, and can be implemented using s matrix-loads [TW23].

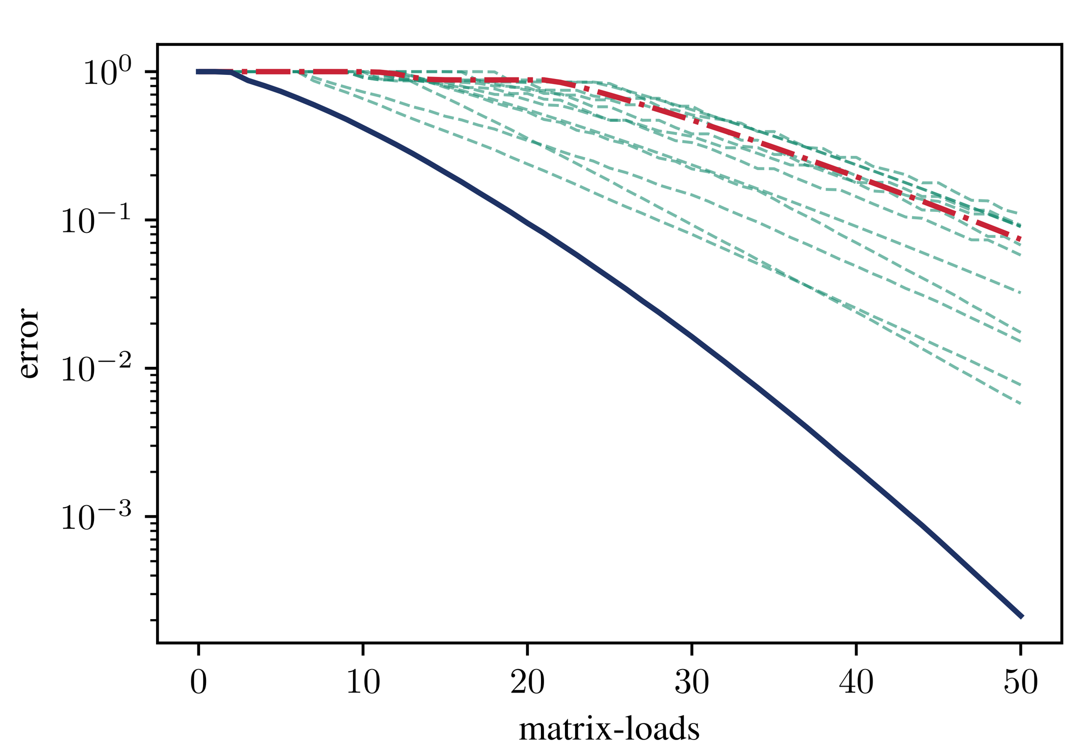

Figure. Convergence (in terms of matrix-loads) of block CG (—) standard CG (-·-) and the state of the art Nyström PCG [FTU23] with various choices of hyperparameters (---). Block CG outperforms these existing methods without the need for selecting hyperparameters, which may be difficult to do effectively in practice.

Our approach: augmented block-CG

Given a matrix $\mathbf{B} \in \mathbb{R}^{d\times m}$ (typically $m \ll d$) with columns $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(m)}$, the block Krylov subspace is

$$\mathcal{K}_t(\mathbf{A}, \mathbf{B}) := \mathcal{K}_t(\mathbf{A}, \mathbf{b}^{(1)}) + \cdots + \mathcal{K}_t(\mathbf{A}, \mathbf{b}^{(m)}). \tag{7}$$

That is, $\mathcal{K}_t(\mathbf{A}, \mathbf{B})$ is the space consisting of all linear combinations of vectors in $\mathcal{K}_t(\mathbf{A}, \mathbf{b}^{(1)}), \ldots, \mathcal{K}_t(\mathbf{A}, \mathbf{b}^{(m)})$.

This naturally gives rise to the block-CG algorithm [OLe80].

Definition. *Let* $\mathbf{B} = [\mathbf{b}^{(1)}\ \cdots\ \mathbf{b}^{(m)}]$. *The t-th block-CG iterates are defined as*

$$\operatorname{bcg}_t^{(i)}(\mu) := \operatorname*{argmin}_{\mathbf{x}\in\mathcal{K}_t(\mathbf{A},\mathbf{B})} \|\mathbf{A}_\mu^{-1}\mathbf{b}^{(i)} - \mathbf{x}\|_{\mathbf{A}_\mu}.$$

The block-CG iterates $\operatorname{bcg}_t^{(1)}(\mu), \ldots, \operatorname{bcg}_t^{(m)}(\mu)$ can be simultaneously computed using $t-1$ block matrix-vector products with $\mathbf{A}$.

Our first main result is the observation that by augmenting $\mathbf{b}$ with $\mathbf{\Omega}$, block-CG implicitly enjoys the benefits of certain classes of preconditioners built using $\mathbf{\Omega}$. In particular, we have the following error guarantee:

Theorem. *Fix any matrix $\mathbf{\Omega} \in \mathbb{R}^{d\times m}$ and let $\mathbf{P} = (\mathbf{I} + \mathbf{X})^{-1}$ be any preconditioner where $\operatorname{range}(\mathbf{X}) \subseteq \mathcal{K}_{s+1}(\mathbf{A}, \mathbf{\Omega})$. Define the augmented starting block $\mathbf{B} = [\mathbf{b}\ \mathbf{\Omega}]$. Then, for any $t \geq s$, the t-th block-CG iterate is related to the $(t-s)$-th preconditioned-CG iterate corresponding to the preconditioner $\mathbf{P}_\mu$ in that*

$$\|\mathbf{A}_\mu^{-1}\mathbf{b} - \operatorname{bcg}_t^{(1)}(\mu)\|_{\mathbf{A}_\mu} \leq \|\mathbf{A}_\mu^{-1}\mathbf{b} - \operatorname{pcg}_{t-s}(\mu)\|_{\mathbf{A}_\mu}.$$

In particular, when $\mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T} = \mathbf{A}\langle\mathbf{K}_s\rangle$, then the deflation preconditioner $\mathbf{P}_\mu$ defined in (3) has the form

$$\mathbf{P}^{-1} = \mathbf{I} + \mathbf{X}, \text{ where } \operatorname{range}(\mathbf{X}) \subseteq \mathcal{K}_s(\mathbf{A}, \mathbf{\Omega}). \tag{8}$$

Corollary. *Let $\mathbf{\Omega} \in \mathbb{R}^{d\times(r+2)q}$, where $q \geq \log(1/\delta)/\log(100)$, be a random Gaussian matrix and define the augmented starting block $\mathbf{B} = [\mathbf{b}\ \mathbf{\Omega}]$. Let*

$$\varepsilon_t(\mu) := 2\exp\left(-\frac{t - (3 + \log(d))}{3\sqrt{(\lambda_{r+1} + \mu)/(\lambda_d + \mu)}}\right).$$

Then the block-CG satisfies, with probability at least $1-\delta$,

$$\left\{\forall \mu \geq 0 : \frac{\|\mathbf{A}_\mu^{-1}\mathbf{b} - \operatorname{bcg}_t^{(1)}(\mu)\|_{\mathbf{A}_\mu}}{\|\mathbf{A}_\mu^{-1}\mathbf{b}\|_{\mathbf{A}_\mu}} \leq \varepsilon_t(\mu)\right\}.$$

Takeaway: *Block CG automatically matches the guarantees/performance of Nyström PCG, without the need to build a preconditioner!*

Numerical Experiments

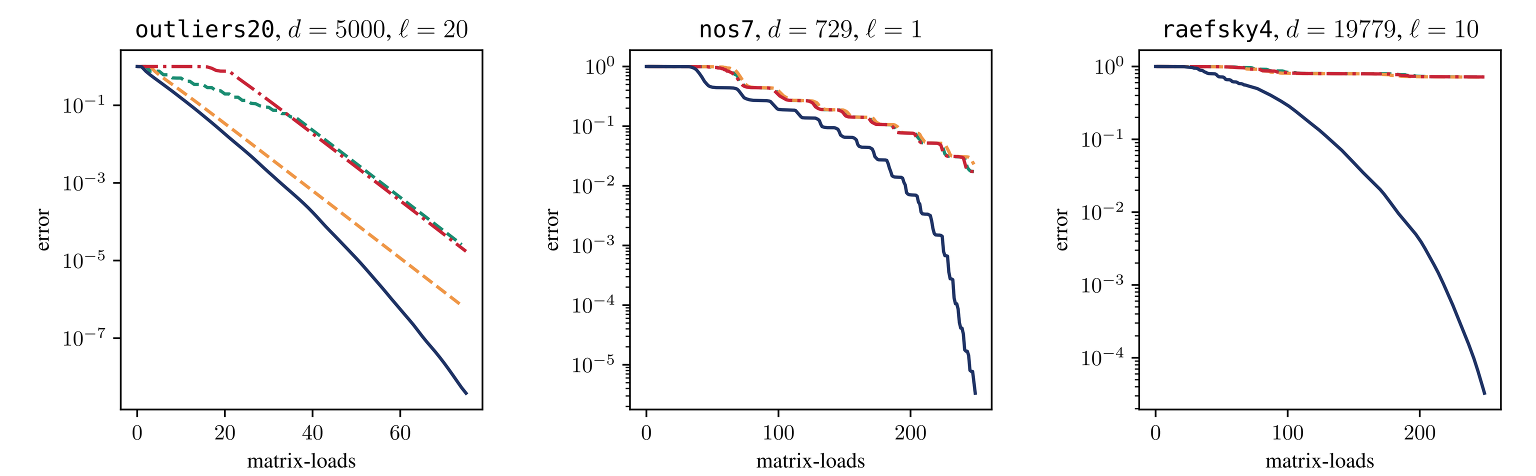

Figure. Error versus matrix-loads for block-CG (—), CG (-·-), and Nyström PCG with $s = 1$ (---) and $s = 3$ (---) on several test problems.

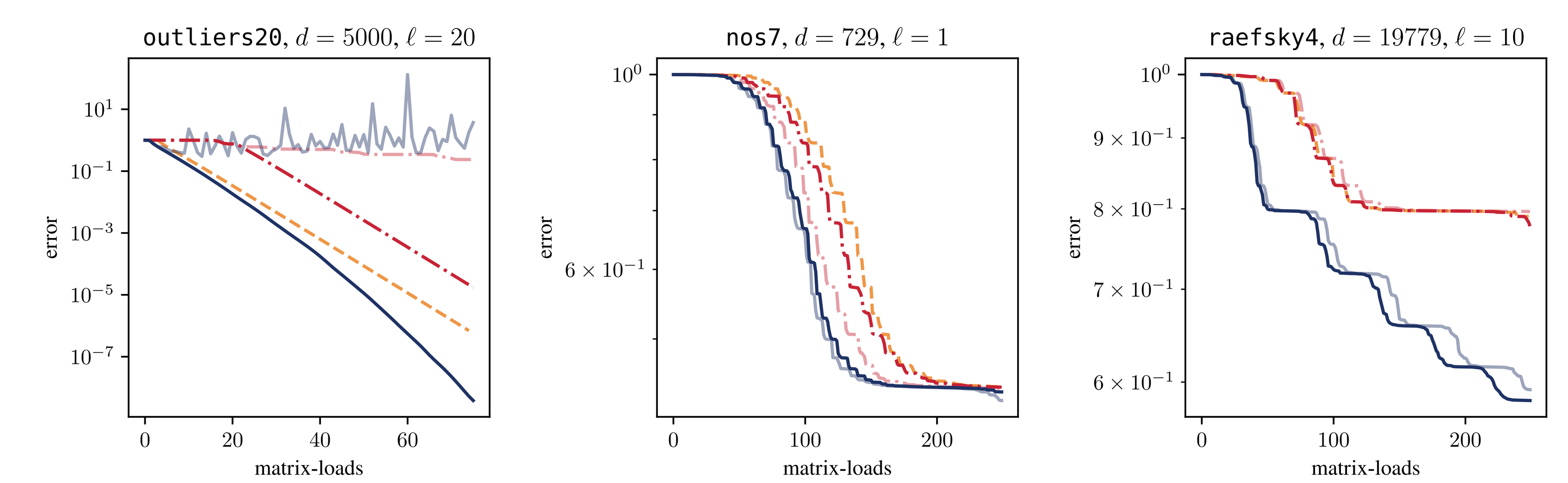

Figure. Error versus matrix-loads for block-CG (—) with reorthogonalization for 3 iterations, CG (-·-) with reorthogonalization for 3ℓ iterations, and Nyström PCG with $s = 3$ (---) without any reorthgonalization. Light curves show convergence of block-CG and CG with no reorthogonalization.

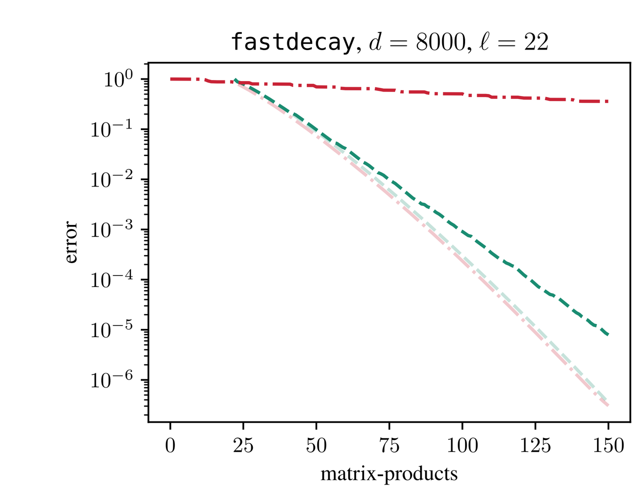

Figure. CG (-·-) and Nyström PCG with $s = 1$ (---) without any reorthgonalization.

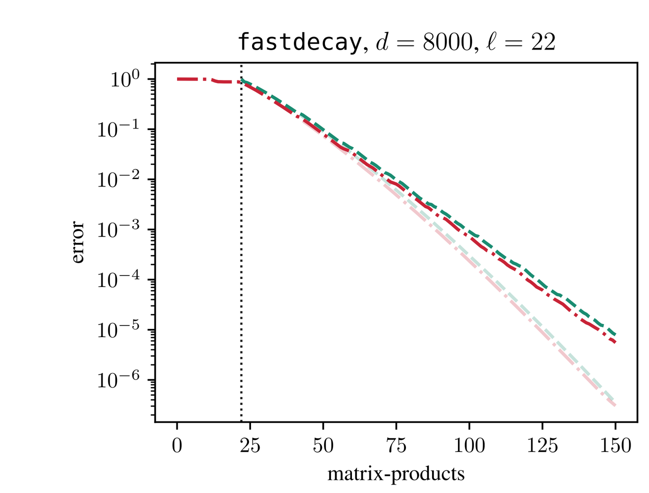

Figure. CG (-·-) with reorthogonalization for 22 iterations and Nyström PCG with $s = 1$ (---) without any reorth.

References

[FTU23] Zachary Frangella, Joel A. Tropp, and Madeleine Udell. ''Randomized Nyström Preconditioning''. In: *SIAM Journal on Matrix Analysis and Applications* 44.2 (May 2023), pp. 718–752. ISSN: 1095-7162.

[OLe80] Dianne P. O'Leary. ''The block conjugate gradient algorithm and related methods''. In: *Linear Algebra and its Applications* 29 (Feb. 1980), pp. 293–322. ISSN: 0024-3795.

[TW23] Joel A. Tropp and Robert J. Webber. *Randomized algorithms for low-rank matrix approximation: Design, analysis, and applications*. 2023. arXiv: 2306.12418 [math.NA].