Preconditioning without a preconditioner: faster ridge-regression and Gaussian sampling with randomized block Krylov methods

Tyler Chen, Caroline Huber, Ethan Lin, Hajar Zaid

Abstract

One of the most important tasks in numerical linear algebra is solving the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is symmetric positive definite with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$. Krylov subspace methods (KSMs) such as the conjugate gradient method are among the most powerful methods for this problem and are guaranteed to converge extremely rapidly if the system is well-conditioned; i.e. if $\lambda_1 \approx \lambda_d$. For ill-conditioned systems, *preconditioning* can greatly accelerate the convergence of KSMs. When $\mathbf{A}$ has a rapidly decaying spectrum, a technique called Nyström preconditioning has proven effective [1].

Consider the Nyström approximation

$$\mathbf{A}\langle\mathbf{K}_s\rangle := (\mathbf{A}\mathbf{K}_s)(\mathbf{K}_s^\mathsf{T}\mathbf{A}\mathbf{K}_s)^\dagger(\mathbf{K}_s^\mathsf{T}\mathbf{A}), \tag{2}$$

where $\mathbf{\Omega} \in \mathbb{R}^{d\times(r+2)}$ is a matrix of independent standard normal random variables and $\mathbf{K}_s := [\mathbf{\Omega}\ \mathbf{A}\mathbf{\Omega}\ \cdots\ \mathbf{A}^{s-1}\mathbf{\Omega}] \in \mathbb{R}^{d\times s(r+2)}$. It can be guaranteed that if $s = O(\log(d))$, then with high probability, $\mathbf{A}\langle\mathbf{K}_s\rangle$ approximates $\mathbf{A}$ with spectral-norm error comparable to the best-possible rank-r approximation to $\mathbf{A}$; i.e. $\|\mathbf{A} - \mathbf{A}\langle\mathbf{K}_s\rangle\| = O(\lambda_{r+1})$ [3]. Define a preconditioner

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

where $\mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}$ is the eigendecomposition of $\mathbf{A}\langle\mathbf{K}_s\rangle$. Following the approach of [1], we show that if $\theta \in [\lambda_d, \lambda_{r+1}]$ and $s = O(\log(d))$, then with high probability, then

$$\kappa(\mathbf{P}^{-1/2}\mathbf{A}\mathbf{P}^{-1/2}) = O(\lambda_{r+1}/\lambda_d). \tag{4}$$

As a result, preconditioned-CG with the preconditioner (3) converges at a rate depending on $\sqrt{\lambda_{r+1}/\lambda_d}$ [2]. If $\mathbf{A}$ has just r large eigenvalues, the convergence of preconditioned-CG will be extremely rapid.

One downside to Nyström preconditioning is the need to choose hyperparameters such as θ and s. Our observation is that, after t iterations, block-CG with a starting block $[\mathbf{b}\ \mathbf{\Omega}]$ has error at most that of Nyström preconditioned CG after $t - s - 1$ iterations. Thus, *block-CG enjoys the effects of (Nyström) preconditioning, without the need for constructing a preconditioner or choose parameters.*[1] This allows us to prove the following convergence guarantee.[2]

Theorem 1. *Fix a value $r \geq 0$ and let $\mathbf{b}_2, \ldots, \mathbf{b}_{r+2}$ be independent standard Gaussian vectors. Then after t iterations the block-CG iterate $\mathbf{x}_t^{\text{b-CG}}$ corresponding to a starting block $[\mathbf{b}\ \mathbf{b}_2\ \cdots\ \mathbf{b}_{r+2}]$ satisfies, with probability at least $99/100$,*

$$\frac{\|\mathbf{A}^{-1}\mathbf{b} - \mathbf{x}_t^{\text{b-CG}}\|_\mathbf{A}}{\|\mathbf{A}^{-1}\mathbf{b}\|_\mathbf{A}} \leq 2\exp\left(-\frac{t - (3 + \log(d)/2)}{3\sqrt{\lambda_{r+1}/\lambda_d}}\right).$$

[1] We are assuming iterations, not matrix-vector products, are the dominant cost.

[2] This bound is reminiscent of the "killing off the top eigenvalues" bounds for CG. However, instead of a burn-in period of r iterations, we require a burn-in period of $O(\log(d))$ iterations (independent of r).

More generally, for any $\mu \geq 0$, block-CG (and Nyström preconditioned CG) can be used to solve the regularized linear system

$$(\mathbf{A} + \mu\mathbf{I})\mathbf{x} = \mathbf{b}. \tag{5}$$

Systems of the form (5) arise in a variety of settings, but we are particularly motivated by two critical tasks in machine learning and data science: solving ridge-regression problems and sampling Gaussian vectors. By adapting our bound Theorem 1 for block-CG, we obtain state-of-the-art convergence guarantees for existing Lanczos-based methods used to solve these tasks.

References

[1] Z. Frangella, J. Tropp, and M. Udell (2023). Randomized Nyström Preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 44(2), 718–752.

[2] A. Greenbaum (1997). *Iterative Methods for Solving Linear Systems*. Society for Industrial and Applied Mathematics.

[3] J. Tropp and R. Webber (2023). Randomized algorithms for low-rank matrix approximation: Design, analysis, and applications.