Randomized Householder-Cholesky QR Factorization with Multisketching

Andrew J. Higgins, Daniel B. Szyld, Erik G. Boman, Ichitaro Yamazaki

Abstract

Computing the QR factorization of tall-and-skinny matrices is a critical component of many scientific and engineering applications, including the solution of least squares problems, block orthogonalization kernels for solving linear systems and eigenvalue problems within block or s-step Krylov methods, dimensionality reduction methods for data analysis like Principal Component Analysis, and many others. Two of the most popular high performance QR algorithms for tall-and-skinny matrices are the CholeskyQR2 and shifted CholeskyQR3 algorithms [3, 4], thanks to their communication-avoiding properties along with their exploitation of vendor provided highly-optimized dense linear algebra subroutines, allowing them to achieve high performance on rapidly evolving modern computer architectures. However, CholeskyQR2 may fail to accurately factorize a matrix V when its condition number $\kappa(V) \gtrapprox \mathbf{u}^{-1/2}$, where $\mathbf{u}$ is unit roundoff [12]. Shifted CholeskyQR3 is numerically stable as long as $\kappa(V) \lessapprox \mathbf{u}^{-1}$, but it requires over 50% more computational and communication cost than CholeskyQR2 [3]. Although TSQR [2] is a more stable communication-avoiding algorithm than the aforementioned Cholesky-based methods, it relies on a non-standard reduction operator, which can make it substantially slower than CholeskyQR2 in practice [4], and is significantly harder to implement efficiently on high performance GPUs. Hence, Cholesky-based QR methods remain popular on modern architectures.

Random sketching has become a popular dimension reduction technique in the fields of numerical linear algebra and data analysis. The central premise of random sketching is to embed a set $\mathcal{V} \subset \mathbb{R}^n$ into a lower-dimensional space via some random projection $S : \mathbb{R}^n \to \mathbb{R}^s$, with $s \ll n$. In numerical linear algebra applications, the random sketch matrix $S \in \mathbb{R}^{s\times n}$ is often selected to be an (ε, d, m) *oblivious subspace embedding*, i.e., for any m-dimensional subspace $\mathcal{V} \subset \mathbb{R}^n$ and $x \in \mathcal{V}$, there is some $\varepsilon \in [0, 1)$ such that

$$\sqrt{1-\varepsilon}\ \|x\|_2 \le \|Sx\|_2 \le \sqrt{1+\varepsilon}\ \|x\|_2,$$

with probability at least $1 - d$ [8, 9]. Such (ε, d, m) oblivious subspace embeddings S are attractive in numerical linear algebra, because if one chooses the subspace $\mathcal{V} \subset \mathbb{R}^n$ to be the column space of a matrix $V \in \mathbb{R}^{n\times m}$, the embeddings can be shown to approximately preserve singular values,

$$(1+\varepsilon)^{-1/2}\ \sigma_{min}(SV) \le \sigma_{min}(V) \le \sigma_{max}(V) \le (1-\varepsilon)^{-1/2}\ \sigma_{max}(SV),$$

and therefore approximately preserve condition numbers,

$$\kappa(V) \le \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\ \kappa(SV),$$

with probability at least $1 - d$. In the context of QR factorizations, one can factorize the small sketched matrix $QR = SV$, and use the triangular factor R as a preconditioner for the large unsketched matrix V, which is effective because

$$\kappa(VR^{-1}) \le \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\ \kappa(SVR^{-1}) = \sqrt{\frac{1+\varepsilon}{1-\varepsilon}} = O(1),$$

for ε sufficiently below 1. This approach is known as the *sketch-and-precondition* framework [7].

In this talk, we present the results from our recent work [5], which analyzes a randomized tall-skinny QR algorithm called randomized Householder-Cholesky QR (`rand_cholQR`). The algorithm uses the sketch-and-precondition framework with Householder QR as a preprocessing step before following up with a pass of CholeskyQR to fully orthogonalize the preconditioned matrix with little computational and communication cost. In order to reduce the cost of the computations even further, we propose to use "multisketching," i.e., the use of two consecutive random sketch matrices, within the sketch-and-precondition framework. Our approach is general in the sense that our analysis applies to any two oblivious subspace embedding sketching matrices, but is specifically motivated by the use of a large sparse sketch followed by a smaller dense sketch, such as a Gaussian or Radamacher sketch [1], as this particular strategy significantly reduces the complexity of applying the sketch operator. Our analysis applies in particular to Count-Gauss (one application of CountSketch followed by a Gaussian sketch), as described in [6, 10, 11].

We prove that with high probability, the orthogonality error of `rand_cholQR` is on the order of unit roundoff for any numerically full-rank matrix V (i.e., $\kappa(V) \lessapprox \mathbf{u}^{-1}$) and hence it is as stable as shifted CholeskyQR3 and it is significantly more numerically stable than CholeskyQR2. Our numerical experiments ilustrate the theoretical results and suggest that `rand_cholQR` often succeeds for numerically rank-deficient problems as well, unlike either CholeskyQR2 or shifted CholeskyQR3. In addition, the `rand_cholQR` algorithm may be implemented using the same basic linear algebra kernels as CholeskyQR2. Therefore, it is simple to implement and has the same communication-avoiding properties. We perform a computational study on a state-of-the-art GPU to demonstrate that `rand_cholQR` can perform up to 4% faster than CholeskyQR2 and 56.6% faster than shifted CholeskyQR3, while significantly improving the robustness of CholeskyQR2.

In summary, our primary contribution consists of a new error analysis of a multisketched randomized QR algorithm, proving it can be safely used for matrices of larger condition number than CholeskyQR2 can handle. Numerical experiments confirm and illustrate the theory. Our secondary contribution is a computational study on a state-of-the-art GPU that tangibly demonstrates that the multisketched algorithm has superior performance over the single sketch algorithms and similar performance to the high performance but less stable CholeskyQR2 algorithm.

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson- Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66:671–687, 2003. Special Issue on PODS 2001.

[2] James W. Demmel, Laura Grigori, Mark Hoemmen, and Julien Langou. Communication-optimal parallel and sequential QR and LU factorizations. *SIAM Journal on Scientific Computing*, 34:A206–A239, 2012.

[3] Takeshi Fukaya, Ramaseshan Kannan, Yuji Nakatsukasa, Yusaku Yamamoto, and Yuka Yanagisawa. Shifted Cholesky QR for computing the QR factorization of ill-conditioned matrices. *SIAM Journal on Scientific Computing*, 42:477-503, 2020.

[4] Takeshi Fukaya, Yuji Nakatsukasa, Yuka Yanagisawa, and Yusaku Yamamoto. CholeskyQR2: A simple and communication-avoiding algorithm for computing a tall-skinny QR factorization on a large-scale parallel system. In *Proceedings of ScalA: 5th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems*, pages 31-38, Los Alamitos, CA, 2014. IEEE Computer Society.

[5] Andrew J. Higgins, Daniel B. Szyld, Erik G. Boman, and Ichitaro Yamazaki. Analysis of Randomized Householder-Cholesky QR Factorization with Multisketching, 2024. arXiv:2309.05868.

[6] Michael Kapralov, Vamsi Potluru, and David Woodruff. How to fake multiply by a Gaussian matrix. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 2101–2110. Proceedings of Machine Learning Research, 2016.

[7] Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical linear algebra: foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

[8] Yuji Nakatsukasa and Joel A. Tropp. Fast & accurate randomized algorithms for linear systems and eigenvalue problems, 2021. arXiv:2111.00113.

[9] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152, Los Alamitos, CA, 2006. IEEE Computer Society.

[10] Aleksandros Sobczyk and Efstratios Gallopoulos. Estimating leverage scores via rank revealing methods and randomization. *SIAM Journal on Matrix Analysis and Applications*, 42:199–1228, 2021.

[11] Aleksandros Sobczyk and Efstratios Gallopoulos. pylspack: Parallel algo- rithms and data structures for sketching, column subset selection, regres- sion, and leverage scores. *ACM Transactions on Mathematical Software*, 48:1–27, 2022.

[12] Yusaku Yamamoto, Yuji Nakatsukasa, Yuka Yanagisawa, and Takeshi Fukaya. Roundoff error analysis of the Cholesky QR2 algorithm. *Electronic Transactions on Numerical Analysis*, 44:306–326, 2015.