The Stability of Split-Preconditioned FGMRES in Four Precisions

Erin Carson, Ieva Daužickaitė

Abstract

We consider the problem of solving a linear system of equations $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ is nonsymmetric and $x, b \in \mathbb{R}^n$. When A is large and sparse, the iterative generalized minimal residual method (GMRES) or its flexible variant (FGMRES) are often used. In these and other Krylov subspace methods, preconditioning is an essential ingredient. Given a preconditioner $P = M_L M_R$, the original problem is transformed to

$$M_L^{-1} A M_R^{-1} \tilde{x} = M_L^{-1} b, \qquad \text{where } M_R^{-1}\tilde{x} = x.$$

The emergence of mixed precision hardware has motivated work in developing mixed precision algorithms for matrix computations; see, e.g., the recent survey [4]. Modern GPUs offer double, single, half, and even quarter precision, along with specialized tensor core instructions; see, e.g., [5]. The use of lower precision can offer significant performance improvements, although this comes at a numerical cost. With fewer bits, we have a greater unit roundoff and a smaller range of representable numbers. The goal is thus to selectively use different precisions within algorithms such that performance is potentially improved without adversely affecting the desired numerical properties.

In this talk, based on the published work [3], we consider the split-preconditioned FGMRES method in a mixed precision framework, in which four potentially different precisions can be used for computations with the coefficient matrix A (unit roundoff u_A), left-preconditioner M_L (unit roundoff u_L), right-preconditioner M_R (unit roundoff u_R), and all other computations (unit roundoff u).

Our analysis is applicable to general preconditioners with minimal assumptions. Briefly, following the strategy of [6], we assume that the application of M_L^{-1} and M_R^{-1} can be computed such that

$$\begin{aligned} fl(M_L^{-1} w_j) &= M_L^{-1} w_j + \Delta M_{L,j} w_j, \quad |\Delta M_{L,j}| \le c(n) u_L E_{L,j}, \\ fl(M_R^{-1} w_j) &= M_R^{-1} w_j + \Delta M_{R,j} w_j, \quad |\Delta M_{R,j}| \le c(n) u_R E_{R,j}, \end{aligned}$$

where $fl(\cdot)$ denotes the quantity computed in floating point arithmetic, $E_{L,j}$ and $E_{R,j}$ have positive entries, $w_j \in \mathbb{R}^n$, and $c(n)$ is a constant that depends on n only. Note that a particular strength of FGMRES is that it allows the right preconditioner to change throughout the iterations; for simplicity, we consider the case here where the preconditioners are static, although our results could be extended to allow dynamic preconditioning.

We define $\tilde{A} \equiv M_L^{-1} A$ and $\tilde{b} \equiv M_L^{-1} b$ and assume that matrix-vector products with $\tilde{A}$ can be computed so that

$$fl(\tilde{A} z_j) = (M_L^{-1} + \Delta M_{L,j})(A + \Delta A_j) z_j.$$

Denoting

$$u_A \psi_{A,j} = \frac{\|M_L^{-1} \Delta A_j z_j\|}{\|\tilde{A}\| \|z_j\|} \quad \text{and} \quad u_L \psi_{L,j} = \frac{\|\Delta M_{L,j} A z_j\|}{\|\tilde{A}\| \|z_j\|},$$

where $\|\cdot\|$ denotes the 2-norm, and ignoring the second order terms, we can write

$$fl(\tilde{A} z_j) \approx \tilde{A} z_j + f_j, \qquad \text{where } \|f_j\| \le (u_A \psi_{A,j} + u_L \psi_{L,j}) \|\tilde{A}\| \|z_j\|.$$

We first present general bounds on the backward and forward errors in split-preconditioned FGMRES, which is based on the previous works [1] and [2]. Our analysis provides guidance on how the precisions should be set when the target backward error is of order u. To summarize, the precision for applying M_L must be chosen in relation to u, u_A, and the required backward and forward errors, because u_L heavily influences the achievable backward error. We can be more flexible when choosing u_R as it does not influence the backward error directly. Our analysis holds under a sufficient but not necessary assumption on u_R in relation to M_R. As long as M_R is not singular in precision u_R (note that scaling strategies may be used to ensure this), setting u_R to a low precision is sufficient. Very low precisions u_L and u_R may delay the convergence, but setting $u_L \leq u$ or $u_R \leq u$ does not improve the convergence in general. Note that these conclusions apply to the full left- and right-preconditioning cases as well.

We observe that the forward error is determined by the backward error and the condition number of the left-preconditioned coefficient matrix. This motivates concentrating effort on constructing an appropriate left-preconditioner when aiming for a small forward error: the preconditioner should reduce the condition number sufficiently and needs to be applied in a suitably chosen precision.

We further provide insights on which preconditioning strategy (left, right, or split) may be preferred under certain objectives related to the desired the backward and forward errors. To summarize, if a small backward error is the main concern and A is ill-conditioned, and we have a 'good' preconditioner, so that $\kappa(\tilde{A})$ is small and we can afford setting u_A and u_L to precisions that are high enough to neutralize the ψ_A and ψ_L terms, then left-preconditioning should be used. If however, we cannot afford setting u_A and u_L to high precisions but can construct a split-preconditioner such that $\kappa(M_L)$ is small, then split-preconditioning (note that in this case ψ_A and ψ_L may be smaller too) or full right-preconditioning may be preferential. If our main concern is applying the preconditioner in lower than the working precision (which may be relevant, for example, when A is very sparse and the preconditioner uses some dense factors), the bounds suggest that full left-preconditioning should not be used as $u_A\psi_A$ and $u_L\psi_L$ may be large. Full right-preconditioning may be most suitable in this case.

We present a suite of numerical experiments which support our theoretical results. Essentially, the experiments confirm that the precision in which the left preconditioner is applied has a significant effect on the forward and backward errors, but very little effect on the number of FGMRES iterations required for convergence. Conversely, the precision in which the right preconditioner is applied has almost no effect on the resulting forward and backward errors, but can affect the FGMRES convergence.

References

[1] Mario Arioli and Iain S Duff. Using FGMRES to obtain backward stability in mixed precision. *Electronic Transactions on Numerical Analysis*, 33:31–44, 2009.

[2] Mario Arioli, Iain S Duff, Serge Gratton, and Stéphane Pralet. A note on GMRES preconditioned by a perturbed LDL^T decomposition with static pivoting. *SIAM Journal on Scientific Computing*, 29(5):2024–2044, 2007.

[3] Erin Carson and Ieva Daužickaitė. The stability of split-preconditioned FGMRES in four precisions. *Electronic Transactions on Numerical Analysis*, 60:40–58, 2024.

[4] Nicholas J. Higham and Theo Mary. Mixed precision algorithms in numerical linear algebra. *Acta Numerica*, 31:347–414, 2022.

[5] NVIDIA H100 Tensor Core GPU. NVIDIA, `https://www.nvidia.com/en-us/data-center/h100/`, 2023.

[6] Bastien Vieublé. *Mixed precision iterative refinement for the solution of large sparse linear systems*. PhD thesis, INP Toulouse, University of Toulouse, France, 2022.