Mixed precision implicit numerical schemes for solving large systems of ordinary differential equations

Mouhamad AL SAYED ALI, IRMAR Univ. Rennes, Rennes, 35000, France - Rennes
Samuel BERNARD, Université Claude Bernard Lyon 1, ICJ UMR5208, Inria - Villeurbanne
Arsène MARZORATI, Inria Lyon, 69100, Villeurbanne, France - Villeurbanne
Jonathan ROUZAUD-CORNABAS, CITI, INSA Lyon, CNRS, Inria, Lyon, France - Lyon

On modern architectures, the performance of 32-bit (single precision) operations is often at least twice as fast as the performance of 64-bit (double precision) operations (see e.g. [2]). By using a combination of 32-bit and 64-bit floating point arithmetic (mixed precision), we can design numerical schemes that run faster and use less memory while limiting the loss of arithmetic precision due to the use of less precise numerical formats. Furthermore, future architectures will add more floating point precisions such as different types of 16-bit half precision. These new formats should be used in the future to further improve the performance of numerical schemes.
Here we study the use of mixed precision in solving large systems of ordinary differential equations (ODEs) using implicit schemes. These schemes, such as the implicit Euler, the Crank–Nicolson, and implicit Runge–Kutta, require at each integration step, the solution of a large nonlinear system. The nonlinear system is solved by the Newton method, which leads to a set of linear systems involving the Jacobian matrix of the ODE and which are solved by Krylov subspace methods. The convergence of the whole process relies on the quality of the initial solutions for both the Newton method and the linear systems. To improve global convergence, line search and trust region algorithms can be used to improve initial solutions.
We explore several mixed approaches for reducing the arithmetic precision in the resolution of the nonlinear system in order to accelerate the numerical solution of the ODE. These approaches combine the performance of lower precision arithmetic with the accuracy of higher precision arithmetic.
We have tested(see [1]) our results on several models, including the Neural Field model and a multiscale mathematical model for the regulation of the cell cycle by the circadian clock, which is relatively stiff. Numerical experiments show that our approach, running in either sequential or in parallel with MPI, is up to twice as fast as the double precision approach with same level of accuracy. These results also show that the implicit schemes running in single precision are up to two times faster than those in double precision, but they either fail to provide sufficient accuracy or diverge. However, our mixed precision schemes consistently remain convergent.

[1] M. Al-Sayed-Ali, S. Bernard, A. Marzorati, J. Rouzaud-Cornabas. *Mixed-precision implicit numerical schemes for systems of ordinary differential equations.* Numerical Algorithms, **1(1)**, 2025.

[2] N. J. Higham, T. Mary. *Mixed precision algorithms in numerical linear algebra.* Acta Numerica, **32**, 2022.

Contact : mouhamad.alsayedali@univ-rennes.fr