Institute of Mathematics CAS, v. v. i.

Žitná 25, 115 67 Praha 1
Czech Republic

Review of the PhD thesis
Eda Oktay: Mixed-Precision Computations in Numerical Linear Algebra

The thesis studies an interesting topic of combining different precisions of floating points numbers in order to accelerate selected algorithms of numerical linear algebra. Motivation for this research comes from the fact that modern computer processors and especially accelerators are equipped with dedicated units for single and half precision floating point arithmetic which can be significantly faster than computing in the standard double precision due to faster arithmetic operations as well as moving data between processor and memory or in exchanging messages. On the other hand, reducing the precision naturally decreases the accuracy of computations. Consequently, it is crucial to understand the role of precision in our algorithms and to reduce (or increase) precision only in the appropriate parts of the algorithms.

The PhD candidate selected six algorithms from numerical linear algebra and she analyses them for the effect of using different precision in different parts of the algorithms. Apart of reducing the precision, the analysis leads to increasing precision in certain situations, for example, to delay the loss of orthogonality in the classical Gram-Schmidt algorithm.

The thesis is written in English on a very high level of language and with minimal amount of typos (I think there should not be epsilon in the denominator of the second equation on page 9 and there should be R^{-1} rather than U^{-1} under the first equation on page 50). After the first chapter of Introduction, three following chapters present the concepts of using mixed precision in linear algebra, iterative methods, and block orthogonalization methods. The next six chapters present results of the candidate, with each chapter corresponding to one algorithm. Results of these chapters have been already published in two high-quality journal papers and two papers in conference proceedings. Consequently, these chapters have their own lists of references and are structured in a paper-like style. The results are finally summarized in the last chapter of Conclusions.

I consider the topic as well as the results as very important. Understanding and insights into the behaviour of the algorithms with respect to used precision is critical to accompany experimental studies more or less limiting themselves to switching different precisions at different places of algorithms and evaluating its effects on accuracy a posteriori. I was especially pleased by the result in Table 7.4, in which the approach of Krylov subspace recycling in connection to GMRES led to significant savings in the number of iterations. What I slightly miss in the thesis is an analysis of the algorithms on real hardware, including time measurements. At least in certain parts of the thesis, such results would help in making conclusions with respect to run times. However, I understand a large room for potential external collaboration in this regard, and the author also plans such testing as suggested in the Conclusions.

I consider Eda Oktay's PhD thesis as an excellent work in numerical linear algebra, and **I recommend it for the defense**. In addition, I would like the candidate to answer the following four questions.

Questions:

1. Could you explain why it is important to consider u^2 precision for matrix vector products in step 6 of Algoritm 19?

2. On page 118, you mention that *"It is not clear why for the steam3 matrix, MSIR performs two SGMRES-IR steps and one GMRES-IR step when SGMRES-IR alone only requires one step."* Do you have an explanation for this behaviour now? Unless I miss something, it must be related to fulfilling one of the four criteria on page 106, right?

3. Do you know what is the adoption of the presented algorithms in numerical libraries, e.g. Ginkgo or MAGMA?

4. While single and half precisions are implemented in hardware nowadays, how widespread is the hardware implementation of quadruple precision that you propose to use in certain parts of the algorithms?

June 5, 2024

Ing. Jakub Šístek, Ph.D.
Institute of Mathematics of the Czech Academy of Sciences
Czech Republic