Institut de Recherche
en Informatique de Toulouse

Equipe APO
IRIT-CNRS
ENSEEIHT
2 rue C. Camichel
31071 TOULOUSE CEDEX

Mixed precision algorithms for low-rank approximations

Background

Modern applications in data analysis and scientific computing are characterized by significant volumes of data and extremely high operational complexity which make the use of supercomputers inevitable. Following recent technological developments, the architecture of high performance computers is becoming more and more heterogeneous, that is to say that the processors are often accompanied by *specialized* computing units capable of achieving very high performance but only on particular operations. A form of specialization is represented by low precision computing units such as *half precision* allowing to carry out calculations more efficiently (less time, memory and energy) but with a lower arithmetic precision. In this context, it is extremely important to design algorithms with low operational complexity and, at the same time, capable of using the computational power of modern computers by exploiting their heterogeneity. This stage will be interested in the study of mixed precision approximate linear algebra algorithms, that is, efficient algorithms, with low operational complexity based on the simultaneous use of several floating point arithmetics in order to take advantage of low precision units while providing guarantees on the overall accuracy of the solution.

Internship topic

Low-rank approximation methods [4] are widely used techniques for reducing the operational complexity and memory consumption of numerous algorithms in data analysis and scientific computing. These methods take advantage of the fast decay of singular values in various types matrices to efficiently approximate these as the product of two matrices of low rank:

$$B \approx XY^T, \quad B \in \mathcal{R}^{n\times n}, \quad X, Y \in \mathcal{R}^{n\times r}, \quad r \ll n, \quad \|B - XY^T\| < \varepsilon \qquad (1)$$

Although the most accurate method for computing such an approximation is the singular value decomposition (SVD), in practice other methods are preferred which are less costly; one such method is the QR factorization with column pivoting (QRCP). Recent work [1] has demonstrated that the memory consumption and the execution time of algorithms can be even further reduced if low-rank approximations are stored using multiple arithmetic precisions. In this new format, p different precisions can be used simultaneously: highest precisions are used to store the data carrying the most important information and the lowest to store less important data:

$$B \approx \sum_{i=1}^{p} X_i Y_i^T, \qquad X_i, Y_i \in \mathcal{R}^{n \times r_i}, \qquad \sum_{i=1}^{p} r_i = r \tag{2}$$

A rigorous rounding error analysis shows that the use of multiple precision does not degrade the overall quality of the approximation with respect to the case where only the highest precision is used to represent all the data.

Objective of the internship

Mixed-precision low-rank approximations can be computed in two steps: first, a low-rank approximation is computed using only high precision, resulting in equation (1), and then the X and Y matrices are partitioned in p groups and each group converted into the corresponding precision leading to equation (2). The objective of this internship is, instead, to develop a QRCP algorithm for computing directly the representation in equation (2) starting from matrix B without the intermediate step of equation (1). Multiple techniques can be combined to achieve this objective. First, the developed algorithm can, itself, use multiple computing precisions depending on how "important" is the data being computed; this will reduce the execution time and memory consumption for computing equation (2). Furthermore, approximate pivoting techniques, such as randomized pivoting [3] or tournament pivoting [2], can be combined with mixed precision to achieve higher computation efficiency and better parallelism.

The internship can address multiple topics related to the above objective ranging from the design of such algorithms, conducting their rounding error analysis to provide bounds on the overall accuracy of the mixed-precision low-rank approximation and the actual implementation and experimental evaluation on parallel computers possibly equipped with GPUs.

Context

This internship will be conducted within the NumPEx project (https://numpex.irisa.fr/) which is a large national project on high performance computing (HPC) which is related to the upcoming installation of an Exascale (i.e., 10^{18} floating-point operations per second) supercomputer in France. This offers a very vibrant context where numerous interactions are possible with some of the most talented researchers in HPC in and outside of France.

It will be possible to pursue the work of the internship in a PhD in the context of the NumPEx project.

Organisation

Team	:	equipe IRIT-APO
Supervision	:	in collaboration with the LIP6 laboratory (Paris)
Duration	:	5-6 months
Remuneration	:	around 600€ / month
Provisional starting date	:	February/March 2023
Place	:	ENSEEIHT-IRIT, 2 rue Claude Camichel, 31000 Toulouse
Contact	:	Alfredo Buttari <alfredo.buttari@irit.fr>

References

[1] Patrick Amestoy, Olivier Boiteau, Alfredo Buttari, Matthieu Gerest, Fabienne Jézéquel, Jean-Yves L'Excellent, and Theo Mary. Mixed precision low-rank approximations and their application to block low-rank LU factorization. *IMA Journal of Numerical Analysis*, 08 2022. drac037.

[2] James W. Demmel, Laura Grigori, Ming Gu, and Hua Xiang. Communication avoiding rank revealing qr factorization with column pivoting. *SIAM Journal on Matrix Analysis and Applications*, 36(1):55–89, 2015.

[3] Per-Gunnar Martinsson, Gregorio Quintana Ortí, Nathan Heavner, and Robert van de Geijn. Householder qr factorization with randomization for column pivoting (hqrrp). *SIAM Journal on Scientific Computing*, 39(2):C96–C115, 2017.

[4] Theo Mary. *Block Low-Rank multifrontal solvers: complexity, performance, and scalability.* PhD thesis, EDMITT, Université Paul Sabatier, Toulouse, France, November 2017.