A composable abstraction of hierarchical methods for matrix-vector product acceleration

Antoine Gicquel*[1]

[1]LABRI – Inria Bordeaux – France

Abstract

Matrix-vector multiplication is an essential mathematical operation in various fields of scientific computing and acts as the cornerstone of numerous iterative algorithms. As a consequence, accelerating this operation can significantly improve the computational efficiency of multiple solvers. When dealing with data-sparse matrices characterized by block low-rank structures, hierarchical methods are often employed to reduce the computational cost and memory requirements from quadratic to log-linear or even linear complexity with controlled accuracy. Common examples of such methods include the Fast Multipole Method (FMM), ${H}$ matrices, ${H}ˆ2$ matrices, Hierarchically Semiseparable (HSS) matrices, Hierarchically Off-Diagonal Low-Rank (HODLR) matrices, and Block Low-Rank (BLR) matrices, which are sometimes referred to as their flat counterparts. All of these methods rely on similar core components. First, the matrix is partitioned into blocks. Then, an admissibility criterion determines which off-diagonal blocks can be approximated. Next, a compression strategy is used to obtain low-rank approximations of these (admissible) blocks, which can be combined with a nested-basis strategy. Despite these similarities, each method defines its own matrix-vector product algorithm. To avoid this redundancy and the additional implementation costs, this work proposes a composable abstraction that encompasses and generalizes all these hierarchical methods. From this abstraction, we derive a single, generic matrix-vector product algorithm that is modular; each component (e.g., partitioning strategy, admissibility criterion, compression technique) can be specialized to address different scenarios. Experimental validation demonstrates that our abstraction achieves comparable numerical results to those of the original algorithms.

*Speaker