Adaptive Sketching Based Construction of $\mathcal{H}^2$ Matrices on GPUs

Sherry Li, Wajih Boukaram, Yang Liu, Pieter Ghysels

Abstract

We present a novel linear-complexity bottom-up sketching-based algorithm for constructing a $\mathcal{H}^2$ matrix and its high performance GPU implementation. The construction algorithm requires both a black-box sketching operator and an entry evaluation function. The novelty of our GPU approach centers around the design and implementation of the above two operations in batched mode on GPU with accommodation for variable-size data structures in a batch. The batch algorithms minimize the number of kernel launches and maximize the GPU throughput. When applied to covariance matrices, volume IE matrices and $\mathcal{H}^2$ update operations, our proposed GPU implementation achieves up to 13× speedup over our CPU implementation, and up to 1000× speedup over an existing GPU implementation of the top-down sketching-based algorithm from the H2Opus library. This is the first GPU implementation of the class of bottom-up sketching-based $\mathcal{H}^2$ construction algorithms.

Reference

W. Boukaram, Y. Liu, P. Ghysels, X.S. Li, "Adaptive Sketching Based Construction of $\mathcal{H}^2$ Matrices on GPUs", Proc. of IPDPS Workshop Parallel and Distributed Scientific and Engineering Computing, Milano, Italy, June 3-7, 2025.