Accelerating Randomized Tensor Decompositions using Structured Random Matrices

Grey Ballard

Abstract

Tensor decompositions are generalizations of low-rank matrix approximations to higher dimensional data. They have become popular for their utility across applications—including blind source separation, dimensionality reduction, compression, anomaly detection—where the original data is represented as a multidimensional array. In this talk, I will focus on randomized methods for computing Tucker and Tensor Train (TT) decompositions. The kernel computation is computing a sketch of an unfolding/matricization of the tensor, and we impose structure on the random matrix in order to exploit structure in the tensor unfolding and reduce computational cost. I will discuss theoretical results on the accuracy of these approaches, their accuracy in practice, and the performance improvement they achieve over deterministic methods.

The TT format is useful in particular for tensors of many modes, including representing approximations to the solution of certain types of parametrized partial differential equations. The fundamental operation used to maintain feasible memory and computational time is called rounding, which truncates the internal ranks of a tensor already in TT format. I will present multiple randomized algorithms for this task that are generalizations of randomized low-rank matrix approximation algorithms and provide significant reduction in computation compared to deterministic TT rounding algorithms. We achieve computational efficiency by using random matrix sketches that mirror the TT format of the input tensor. Randomization is particularly effective in the case of rounding a sum of TT tensors, which is the bottleneck computation in the adaptation of GMRES to vectors in TT format. In this talk, I will present the randomized algorithms, compare their empirical accuracy and computational time with deterministic alternatives (including results from [1]), and discuss recent progress on probabilistic error analysis of the algorithms.

I will present two Tucker decomposition algorithms that scale to large data (and many processors), significantly reduce both computation and communication cost compared to previous deterministic and randomized approaches, and obtain nearly the same approximation errors. The key idea in our algorithms is to perform randomized sketches with Kronecker-structured random matrices, which reduces computation compared to unstructured random matrices and can be implemented using a fundamental tensor computational kernel. I will state probabilistic error analysis of our algorithms and present a new parallel algorithm for the structured randomized sketch. Our experimental results demonstrate that our combination of randomization and parallelization achieves accurate Tucker decompositions much faster than alternative approaches. We observe up to a $16 \times$ speedup over the fastest deterministic parallel implementation on 3D simulation data [2].

References

- Hussam Al Daas, Grey Ballard, Paul Cazeaux, Eric Hallman, Agnieszka Miedlar, Mirjeta Pasha, Tim W. Reid, and Arvind K. Saibaba. Randomized algorithms for rounding in the tensor-train format. SIAM Journal on Scientific Computing, 45(1):A74–A95, 2023.
- [2] Rachel Minster, Zitong Li, and Grey Ballard. Parallel randomized Tucker decomposition algorithms. SIAM Journal on Scientific Computing, 46(2):A1186–A1213, 2024.