# CASPR: Combining Axis Preconditioners using Kronecker Sums/Products for Training Large Neural Networks

*Inderjit S. Dhillon, Sai S. Duvvuri*

Abstract

Deep Neural Networks (DNNs) have transformed fields like computer vision, natural language processing, and scientific research by enabling systems to learn complex patterns, make high-level predictions, and analyze large data sets. DNNs have driven advancements in material sciences, chemistry, and physics, significantly aiding scientific discovery. However, they are difficult to optimize due to their large parameter spaces and can require extensive computational resources, and thus effectively training DNNs is a contemporary challenge.

Most DNNs, including Large Language Models, are trained using adaptive regularization methods such as Adam, which can be regarded as diagonally preconditioned stochastic gradient descent. This diagonal preconditioner comes from a diagonal approximation of the gradient outer product matrix. However, a recent open competition called "AlgoPerf: Training Algorithms benchmark competition" [1] revealed an intriguing discovery: a non-diagonal preconditioning method called Shampoo [2], which uses a Kronecker product approximation of the outer-product matrix, was found to be the best method on a varied suite of benchmark problems.

In this talk, I will introduce adaptive methods and show how Kroencker products can be used to get a computationally efficient preconditioner. I will then talk about a general technique called Combining AxeS PReconditioners (CASPR) [3], which optimizes matrix-shaped DNN parameters by finding different preconditioners for each mode/axis of the parameter and combining them using a Kronecker-sum based approximation. The Kronecker-sum based combination allows us to show that CASPR is ordered between the Kronecker product based combination, Shampoo, and full-matrix "Adagrad" preconditioners in Loewner order, and as a result it is nearer to full-matrix Adagrad than Shampoo. Experimental results demonstrate that CASPR approximates the gradient second-moment matrix more accurately, and shows improvement in training and generalization performance compared to the existing practical adaptive regularization methods in a variety of tasks including graph neural network on OGBG-molpcba, Transformer on a universal dependencies dataset and auto-regressive large language modeling on the C4 dataset.

## References

[1] https://mlcommons.org/2024/08/mlc-algoperf-benchmark-competition, 2024.

[2] V. Gupta, T. Koren and Y. Singer. Shampoo: Preconditioned Stochastic Tensor Optimization. *Proceedings of The 35th International Conference on Machine Learning (ICML)*, 2018.

[3] S. S. Duvvuri, F. Devvrit, R. Anil, C. Hsieh and I. S Dhillon. Combining Axes Preconditioners through Kronecker Approximation for Deep Learning. *Proceedings of The Twelfth International Conference on Learning Representations (ICLR)*, 2024.