Computer Science Seminar

The Applications of Alternating Minimization Algorithms on Deep Learning Models

Junxiang Wang
Emory University

Abstract: Gradient Descent(GD) and its variants are the most popular optimizers for training deep learning models. However, they suffer from many challenges such as gradient vanishing and poor conditioning, which prevent their more widespread use. To address these intrinsic drawbacks, alternating minimization methods have attracted attention from researchers as a potential way to train deep learning models. Their idea is to decompose a neural network into a series of linear and nonlinear equality constraints, which generate multiple subproblems and they can be minimized alternately. Their empirical evaluations demonstrate good scalability and high accuracy. They also avoid gradient vanishing problems and allow for non-differentiable activation functions, as well as allowing for complex non-smooth regularization and the constraints that are increasingly important for neural network architectures.¡br¿

This dissertation aims to develop alternating minimization methods to train the Multi-Layer Perceptron(MLP) model. This includes deep learning Alternating Direction Method of Multipliers(dlADMM), monotonous Deep Learning Alternating Minimization(mDLAM), and parallel deep learning Alternating Direction Method of Multipliers(pdADMM). The extended pdADMM-G algorithm and the pdADMM-G-Q algorithms are developed to train the Graph-Augmented Multi-Layer Perceptron(GA-MLP) model.¡br¿

For the dlADMM algorithm, parameters in each layer are updated in a backward and forward fashion. The time complexity is reduced from cubic to quadratic in(latent) feature dimensions for subproblems by iterative quadratic approximations and backtracking. Finally, we provide the convergence guarantee of the dlADMM algorithm under mild conditions.¡br¿

For the mDLAM algorithm, our innovative inequality-constrained formulation infinitely approximates the original problem with non-convex equality constraints, enabling our convergence proof of the proposed mDLAM algorithm regardless of the choice of hyperparameters. Our mDLAM algorithm is shown to achieve a fast linear convergence by the Nesterov acceleration technique.¡br¿

For the pdADMM algorithm, we achieve model parallelism by breaking layer dependency: parameters in each layer of neural networks can be updated independently in parallel. The convergence of the proposed pdADMM to a stationary point is theoretically proven under mild conditions. The convergence rate of the pdADMM is proven to be $o(1/k)$, where k is the number of iterations.¡br¿

For the pdADMM-G algorithm and the pdADMM-G-Q algorithm, in order to achieve model parallelism, we extend the proposed pdADMM algorithm to train the GA-MLP model, named the pdADMM-G algorithm. The extended pdADMM-G-Q algorithm reduces communication costs by introducing the quantization technique. Theoretical convergence to a (quantized) stationary point of two proposed algorithms is provided with a sublinear convergence rate $o(1/k)$, where k is the number of iterations.

Tuesday, November 15, 2022, 11:30 am
https://emory.zoom.us/j/5693008550

COMPUTER SCIENCE
EMORY UNIVERSITY