UNIVERSITY of HOUSTON

Department of Mathematics

Scientific Computing Seminar

Dr. Eric C. Cyr

Exposing structure in neural networks to accelerate training: Towards multilevel algorithms

Thursday, October 9, 2025
1 PM- 2 PM
Room 646 PGH

Abstract: The recent explosion of large-language models (LLMs) in the commercial space has created unprecedented energy demands driven by the growth of data centers. One aspect of this is the need to train large scale neural network models on massive amounts of data. Recent work has demonstrated that pretraining a LLM on the Frontier supercomputer would require two years with ideal parallelism [1]. Yet despite advances in optimizers the training algorithms remain largely the same even as the neural networks scale to trillions of parameters and suffer from quadratic scaling as a result. This motivates our aspirational hypothesis that multilevel (or hierarchical) methodologies can dramatically accelerate training algorithms. To this end we consider structural choices in neural network design that improve training performance and may ultimately be used in developing an efficient multilevel strategy.

This talk proceeds in three parts. In the first part, we discuss an adaptive basis perspective that has proved fruitful in Scientific Machine Learning (SciML). Considering a neural network used for regression, the final linear layer can be viewed separately from the prior nonlinear layers. This naturally leads to a two-step algorithm. In the first step, the linear layer is treated as basis coefficients that are determined by a least squares algorithm. The second step adapts the basis to the data using a step of gradient descent. To ensure a good basis is possible, an initialization algorithm is proposed based on this adaptive basis viewpoint that ensures the initial stability of the network.

Building on this development, the second part of the talk presents layer-parallel training of neural ODEs. This multilevel algorithm exposes additional parallelism in forward and backward propagation leading to parallel acceleration. A nested iteration strategy naturally provides further

improvements. However, more general multilevel approaches have proven difficult as the "downcycle" in multigrid is destructive to training progress.

The third part of the talk presents one approach for introducing additional structure that is more amenable to a full multilevel treatment. We present a view of Kolmogorov-Arnold networks (KANs) that reformulate the activation function as a spline that can be naturally adapted. We further explore how this approach can be related to a multi-channel ReLU network, with a specifically chosen preconditioner to accelerate convergence. Finally, we present preliminary results that motivate how this architecture can be used in a multilevel algorithm.

References:

(1) Dash, S., Lyngaas, I.R., Yin, J., Wang, X., Egele, R., Ellis, J.A., Maiterth, M., Cong, G., Wang, F. and Balaprakash, P., 2024, May. Optimizing distributed training on frontier for large language models. In ISC High Performance 2024 Research Paper Proceedings (39th International Conference).

Speaker Bio:

Eric C. Cyr has been at Sandia Albuquerque since 2009 and is a Principal Member of the Technical staff. Prior to that he received a BS in Computer Science from Clemson University in 2002, and a PhD in Computer Science from University of Illinois at Urbana-Champaign in 2008. His formal training is in numerical methods, and scientific computing. As a Sandian he has worked in a range of diverse areas including high-performance computing, preconditioners for multi-physics, discretizations for computational plasma simulation, sensitivity analysis for PDEs, numerical optimization, and software development for science applications. In 2018, Eric received the DOE Early Career Award to fund a focus on developing layer-parallel methods for training deep neural networks. This work has broadened his exposure to machine learning. To this end, Eric is working to extend the applicability of HPC to deep learning, understand the behavior and approximation properties of neural networks, and develop methodologies where machine learning can be used to enhance the impact of computation on science and engineering disciplines.