MULTIFIDELITY LINEAR REGRESSION FOR SCIENTIFIC MACHINE LEARNING FROM SCARCE DATA

Elizabeth Qian,[1] Anirban Chaudhuri,[2] Dayoung Kang,[1] and Vignesh Sella[2]

[1]Georgia Institute of Technology
[2]University of Texas at Austin

ABSTRACT

Machine learning (ML) methods have garnered significant interest as potential methods for learning surrogate models for complex engineering systems for which traditional simulation is expensive. However, in many scientific and engineering settings, training data are scarce due to the cost of generating data from traditional high-fidelity simulations. ML models trained on scarce data have high variance and are sensitive to vagaries of the training data set. We propose a new multifidelity training approach for scientific machine learning that exploits the scientific context where data of varying fidelities and costs are available; for example high-fidelity data may be generated by an expensive fully resolved physics simulation whereas lower-fidelity data may arise from a cheaper model based on simplifying assumptions. We use the multifidelity data to define new multifidelity Monte Carlo estimators for the unknown parameters of linear regression models, and provide theoretical analyses that guarantee accuracy and improved robustness to small training budgets. Numerical results show that multifidelity learned models achieve order-of-magnitude lower expected error than standard training approaches when high-fidelity data are scarce.