YUNAN YANG, Cornell University

Training Distribution Optimization in the Space of Probability Measures

A central question in data-driven modeling is: from which probability distribution should training samples be drawn to most effectively approximate a target function or operator? This work addresses this question in the setting where "effectiveness" is measured by out-of-distribution (OOD) generalization accuracy across a family of downstream tasks. We formulate the problem as minimizing the expected OOD generalization error, or an upper bound thereof, over the space of probability measures. The optimal sampling distribution depends jointly on the model class (e.g., kernel regressors, neural networks), the evaluation metric, and the target map itself. Building on this characterization, we propose two adaptive, target-dependent data selection algorithms based on bilevel and alternating optimization. The resulting surrogate models exhibit significantly improved robustness to distributional shifts and consistently outperform models trained with conventional, non-adaptive, or target-independent sampling across benchmark problems in function approximation, operator learning, and inverse modeling.