Sample Past COS 514 Projects

Optimization

Learning Single-Index Models Using Stochastic Weighted Averaging

We consider the setting of learning a single-index model $\sigma(\theta^\star \cdot x)$ when the data distribution is the d-dimensional isotropic Gaussian, and the link function σ is known. Existing literature on learning single index models for such settings relate the sample complexity of this class to the information exponent $k^\star$ of the link function σ, which is determined by its Hermite expansion. Arous et al. (2021) showed that online SGD requires $n \gtrsim d^{k^\star-1}$ samples, while Damian et al. (2023) proved that a smoothed SGD algorithm achieves the optimal $n \gtrsim d^{k^\star/2}$ rate, albeit with impractical smoothing steps. This project proposes a practical mini-batch SGD algorithm with stochastic weighted averaging (SWA), matching the optimal sample complexity while remaining simple and implementable, thus partially resolving a conjecture by Abbe et al. (2023).

References

- Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. *Online Stochastic Gradient Descent on Non-Convex Losses from High-Dimensional Inference*, 2021.
- Alex Damian, Eshaan Nichani, Rong Ge, and Jason D. Lee. *Smoothing the Landscape Boosts the Signal for SGD: Optimal Sample Complexity for Learning Single Index Models*, 2023.
- Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. *SGD Learning on Neural Networks: Leap Complexity and Saddle-to-Saddle Dynamics*, 2023.

Generalization

Does Flatness Always Predict Generalization?

This project examines training stochasticity as a means to disentangle flatness from generalization. It investigates whether flatness predicts which of two neural networks—trained with different initializations or data subsets—generalizes better. Results are compared to other predictors of generalization from Jiang et al. (2019), using the Kendall rank correlation coefficient. Empirical findings are inconclusive: while flatness occasionally correlates with better generalization, it does not consistently outperform other measures.

Reference

- Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. *Fantastic Generalization Measures and Where to Find Them.* arXiv preprint arXiv:1912.02178, 2019.

Data Valuation and Interpretability

Data Valuation with Shapley Values

The Shapley value, originating in cooperative game theory, measures each participant's contribution to the collective utility. Recent work extends this concept to data valuation in machine learning, assigning importance scores to training examples based on their impact on model performance. This project investigates the robustness of Shapley values across training regimes and explores their use for identifying mislabeled data, detecting overfitting, and understanding dataset composition.

Architectures

Learnable Expressive Power Separation of Transformers over Fully-Connected Networks

Self-attention units—the core of transformer architectures—are central to modern deep learning. Despite their success, theoretical understanding of their advantages over simpler architectures remains limited. Focusing on the sparse averaging task (Sanford et al., 2023), this project shows that fully-connected networks require polynomial memory in sequence length, whereas transformers achieve logarithmic scaling. We extend lower bounds to expected loss and prove that a one-layer transformer can efficiently learn this task through gradient descent under appropriate parametrization and encoding.

Reference

- C. Sanford, D. Hsu, and M. Telgarsky. *Representational Strengths and Limitations of Transformers.* arXiv preprint arXiv:2306.02896, 2023.

Diffusion Models

Provable Optimization by Thompson Sampling-Guided Conditional Diffusion Models

Guided diffusion models are a powerful tool for generating structured data. This project proposes a conditional diffusion approach combined with Thompson Sampling (TS) to address protein design problems, aiming to generate high-quality sequences that balance novelty and biochemical fitness. The method integrates TS for reward learning and conditional diffusion for guided generation. Under a linear reward assumption, a Bayesian regret bound of $\tilde{O}(d\sqrt{MT})$ is established, where d is the data dimension, M population size, and T the number of iterations. Synthetic experiments show that TS-guided diffusion outperforms Directed Evolution baselines in convergence and regret.

Language Models

Exploring Data Selection Methods for Language Model Training

This project explores methods for selecting “high-quality” training data for large language models (LLMs) in a Python code generation task. Four selection methods are compared: log-likelihood scores, contrastive log-likelihood scores, pointwise mutual information (PMI), and a classifier-based approach using ChatGPT annotations. Models with 350M parameters are trained on curated data subsets and evaluated on a simplified HumanEval benchmark. The classifier-based selection method, leveraging human-like ChatGPT judgments, achieves the best performance. The PMI-based approach also outperforms random baselines, suggesting that information-theoretic signals can serve as interpretable indicators of data quality.