Statistical Methodologies for Decision-Making and Uncertainty Reduction in Machine Learning

Haofeng Zhang

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY, USA
hz2553@columbia.edu

Abstract

While advances in machine learning and the expansion of massive datasets have significantly improved predictive accuracy, the translation of these predictions into actionable decisions—alongside a robust understanding of associated risks—remains underexplored. My research focuses on developing methodology and theory in data-driven decision-making and uncertainty quantification that effectively address core data challenges. This paper presents two connected pillars of my research: data-driven contextual optimization, uncertainty quantification and reduction.

Data-Driven Contextual Optimization

Data-driven optimization aims to construct a decision by optimizing a cost that involves random outcomes only observed from data, which is a common and important task for many real-world applications such as inventory management. When the data also involves some contexts (or features), this problem is known as data-driven contextual optimization and has recently received a lot of attention due to emerging contextual data and machine learning tools. The classic approach for data-driven optimization uses sample average approximation (SAA), but it does not apply easily to constrained, contextual problems as feasibility and generalization issues may arise. Therefore, in practice, a model-based approach is used where we first fit a distribution over the features (price, weather) and response variable (demand) using machine learning. Then, we solve a stochastic optimization problem using this fitted model for every new context that arises – this approach is known as estimate-then-optimize (ETO). More recently, researchers have proposed integrating the estimation and optimization processes (integrated-estimation-optimization, IEO) by selecting the model based on minimizing decision error instead of prediction error.

Statistical Framework for Comparisons. One natural but important direction is which methodology yields the best results, and how is the performance of the various methods impacted by model specification, sample size, data quality, and type of objective/constraints. In our paper (Elmachtoub et al. 2023), we theoretically compare the mainstream approaches for contextual optimization - ETO and IEO - in the setting with nonlinear objectives, which is most common in practice yet unexplored in the literature. We develop a new regret-based framework based on stochastic dominance to study their statistical performance rigorously. Our main results depend on whether the model class of distributions covers the ground-truth distribution (well-specified or misspecified), and the performance ordering is completely opposite in these two cases. When the model is mis-specified, we show that IEO is better than ETO (potentially explaining the superior performance of IEO reported in several papers). However, our main finding is that when the model is well-specified, ETO actually outperforms IEO asymptotically in the "strong" sense of stochastic dominance of the regret (optimality gap). In other words, the entire regret distribution, not only its mean or other moments, is always better for ETO compared to IEO. This phenomenon shows that ETO can outperform IEO under a wide class of problems, giving an opposite view of the common belief. The mathematical deviations of the results require new decomposition ideas to link stochastic dominance comparisons of asymptotic regret distributions to covariance matrix comparisons, and utilize information-theoretic bounds and multiple novel matrix inequalities to establish our findings. Furthermore, in our subsequent paper (Elmachtoub et al. 2024), we extend these findings and dissect the performance comparisons between these approaches in terms of the degree of model misspecifications in the finite-sample regime. This is achieved by balancing error terms in higher-order expansions of the regret distribution, accounting for degrees of model misspecification, and adapting a Berry-Esseen-type bound on estimation error to derive finite-sample bounds on regret.

Recommendation systems and bandit algorithms. Bandit problems are an important sequential decision-making problem widely used in many real-world applications, such as recommendation systems, causal inference, and clinical trials. Bandits require balancing the tradeoff between exploration and exploitation since the outcome of a decision not chosen can never be observed, which requires measuring the uncertainty in decision-making. Bayesian bandit algorithms with approximate Bayesian inference have been widely used. However, there is a large discrepancy between their practical performance and theory (with only negative theoretical results). To bridge this gap, we propose a theoretical framework that can analyze bandit problems in the presence of approximate inference. In Huang et al. (2023), we study an Enhanced Bayesian Upper Confidence Bound (EBUCB) algorithm and demonstrate that EBUCB can achieve the optimal regret order $O(\log T)$ if the inference error measured by two different α-divergences is less than a constant. Moreover, in Huang, Lam, and Zhang (2024), we demonstrate that a similar phenomenon also holds for Linear Thompson Sampling and Linear Bayesian Upper Confidence Bound with approximate inference in contextual linear bandits.

Uncertainty Quantification and Reduction

Uncertainty quantification (UQ) means measuring, dissecting, and understanding the impacts of different sources of errors. It is critical for enhancing estimation models and providing a necessary component toward reliable decision-making. Generally, predictive models incur two potential types of uncertainty: epistemic uncertainty that arises from the statistical error due to the limited data size, and aleatoric uncertainty from the intrinsic randomness of stochastic outputs. The major challenges of UQ include: 1) identifying and quantifying each source of uncertainty, not only with statistical guarantees, but also computational efficiency especially in the face of large-scale models 2) reducing uncertainty to improve estimation or prediction performance; 3) understanding how uncertainty impacts downstream decision-making problems to make better decisions.

Efficient UQ for neural networks. Neural networks have been increasingly significant for prediction and metamodeling, but the epistemic uncertainty of neural networks is challenging and less investigated. Unlike classical models such as linear regression, neural network models are non-convex and have many local minima. Epistemic uncertainties in neural networks arise not only from data, but also from the training procedure that often injects substantial noise. These two interacting errors hinder the attainment of statistical guarantees using classical statistical theory and, moreover, impose computational challenges on UQ due to the need for repeated network retraining. To address this challenge, we create statistically guaranteed schemes to remove the training procedural uncertainty by using what we call *procedural-noise-correcting* network that is trained on a suitably and artificially labeled data set that mimics the procedural variability (Huang, Lam, and Zhang 2023). By leveraging recent connections of neural networks with kernel regression, our work provides simultaneous frequentist statistical guarantees and low computation overhead in terms of few network retrainings.

Metamodels and learning theory for prediction intervals. Simulation metamodeling aims to construct lower-fidelity models to represent input-output relations using a few simulation runs. Traditional approaches target the mean response surface, while we study an alternative metamodeling approach that creates prediction intervals to capture the predictive uncertainty of simulation outputs. More precisely, we study the generation of prediction intervals as an empirical constrained optimization problem that minimizes the average interval width while maintaining coverage accuracy across data (Chen et al. 2021). Our work provides a general learning theory to characterize the optimality-feasibility tradeoff, and further proposes calibration machinery and the corresponding statistical theory to optimally select the model that manages this tradeoff practically. Then in Lam and Zhang (2024), we extend our approach to stochastic settings, aiming to bypass the computational scalability issues encountered by conventional metamodeling approaches such as stochastic kriging.

Reducing joint aleatoric and epistemic uncertainty in simulation. Stochastic simulation under input uncertainty often runs into both epistemic noises (from the input data when calibrating model parameters) and aleatoric uncertainty (intrinsic system stochasticity exhibited by the Monte Carlo samples). At the same time, to speed up the simulation, one often attends to variance reduction techniques and, in this regard, two recent techniques derived from "kernelizing" Stein's identity, one based on control variate and one on importance sampling, have been proposed to reduce the error in Monte Carlo computation to super-canonical convergence (i.e., faster than square-root convergence in standard Monte Carlo). In our paper (Lam and Zhang 2023), we demonstrate that applying the above two techniques alone to problems with joint aleatoric-epistemic uncertainty will have degenerate performance, and we present a new general framework to encompass both techniques via a *doubly robust* idea that simultaneously performs control variate and importance sampling. We show in particular that our approach can effectively reduce the uncertainty regarding mean squared error rates across different scenarios when these existing methods fail.

References

Chen, H.; Huang, Z.; Lam, H.; Qian, H.; and Zhang, H. 2021. Learning prediction intervals for regression: Generalization and calibration. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 820–828. PMLR.

Elmachtoub, A. N.; Lam, H.; Lan, H.; and Zhang, H. 2024. Dissecting the Impact of Model Misspecification in Data-driven Optimization. Under review.

Elmachtoub, A. N.; Lam, H.; Zhang, H.; and Zhao, Y. 2023. Estimate-Then-Optimize versus Integrated-Estimation-Optimization versus Sample Average Approximation: A Stochastic Dominance Perspective. Under revision in *Operations Research*.

Huang, Z.; Lam, H.; Meisami, A.; and Zhang, H. 2023. Optimal Regret Is Achievable with Bounded Approximate Inference Error: An Enhanced Bayesian Upper Confidence Bound Framework. *Advances in neural information processing systems*.

Huang, Z.; Lam, H.; and Zhang, H. 2023. Efficient Uncertainty Quantification and Reduction for Over-Parameterized Neural Networks. *Advances in neural information processing systems*.

Huang, Z.; Lam, H.; and Zhang, H. 2024. Bayesian Bandit Algorithms with Approximate Inference in Stochastic Linear Bandits. Under review, arXiv preprint *arXiv:2406.14071*.

Lam, H.; and Zhang, H. 2023. Doubly Robust Stein-Kernelized Monte Carlo Estimator: Simultaneous Bias-Variance Reduction and Supercanonical Convergence. *Journal of Machine Learning Research*, 24(85): 1–58.

Lam, H.; and Zhang, H. 2024. Prediction Intervals for Simulation Metamodeling. Under revision in *ACM Transactions on Modeling and Computer Simulation (TOMACS)*.