Research statement

Jie Wang

School of Industrial and Systems Engineering, Georgia Institute of Technology, jwang3163@gatech.edu

We are living in an era of ever-growing availability of big data across diverse fields, including computer science, electrical engineering, medicine, marketing, and beyond. Traditional decision-making algorithms, which often depend on rigid and restrictive assumptions, may fail to perform effectively in complex, real-world settings. To overcome these limitations, it is essential to adopt data-driven approaches that can capture the inherent dynamics and uncertainties of these environments, leading to better decisions. It is equally important to ensure these decisions reflect our societal and ethical values, such as robustness, interpretability, privacy, fairness, and energy efficiency. In my research, I leverage tools from machine learning, optimization, and statistics to design data-driven decision-making algorithms that are not only effective but also aligned with those core principles. The diagram in Figure 1 illustrates the general roadmap of my research.

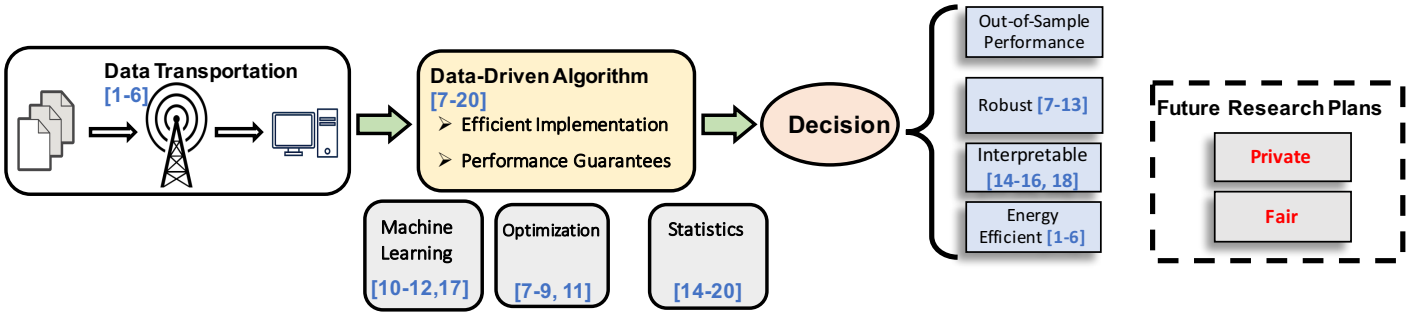

Figure 1: Illustration of my research roadmap for data-driven decision-making (the blue brackets highlight my related reference papers)

In Figure 1, I divide the data-driven decision-making procedure into three key stages:

(1) **Data Transportation**: This stage involves transporting data at unprecedented scales, often characterized by long transport distances, short delays, and limited transmission power. These challenges are common in modern data acquisition environments, such as wireless networks, low-earth orbiters, underwater acoustic networks, and deep space networks.

(2) **Algorithm Development and Decision-Making**: From methodological standpoint, this stage requires designing computationally efficient algorithms that can be practically implemented. From theoretical perspective, it involves analyzing computational complexity and establishing generalization guarantees to ensure it performs well on unseen testing data.

(3) **Deployment in Real-World Environments**: The final stage focuses on deploying decisions that achieve superior out-of-sample performance while adhering to our societal and ethical values.

In the following sections, I will highlight how my research addresses the challenges posed by each stage and outline my future research plans.

Energy-Efficient Data Transportation. In a long-distance transportation task, digital data is transmitted sequentially from a source node to several intermediate nodes before arriving at its destination, referred to as *multi-hop line network communication*. Each transmission link may suffer from distortions like errors and malicious modifications, and each node is allowed to perform coding to improve the transport efficiency. According to the report from International Energy Agency, current data transportation accounts for 2-3% of global electricity consumption and 1% of greenhouse gas emissions. It is urgent to develop new techniques that simultaneously achieve effective data transportation and low energy consumption.

In [1, 2], we studied the optimal throughput of this problem. Routing, a classical baseline in transportation, is not a good solution: Consider a simple instance where each time of transmission has loss probability 0.2, then the throughput of the line network with L hops using routing is 0.8^L, which delays very fast even for a small L. This is known as the "curse of multihop" in the industry. To combat this issue, we proposed a batched network coding communication protocol, which leverages optimization and information theory techniques to find the optimal coding strategies at each node to maximize the throughput subject to certain transmission power constraints. We showed that when the number of hops L is large, this protocol achieves high throughput $O(1)$ with low transmission power $O(\ln L)$ and low storage cost $O(\ln \ln L)$, which breaks the common sense in this area that low transmission power and high throughput cannot be achieved simultaneously.

Following from this work, I collaborated with scholars from the information theory and communication area (Urbashi mitra, Shenghao Yang, and Hoover H.F.Yin) to extend the batched network coding protocol in two aspects. First, we utilize robust optimization technique that incorporates the errors arising from data transmission to improve the throughput [3–5]. Second, we leverage discrete optimization to design coding strategies that only depend on the sparse and interpretable representation of input data [6].

Distributionally Robust Optimization (DRO). Many decision-making problems involve key parameters drawn from unknown probability distributions, which could be indirectly observed only from training data. Data-driven approaches usually extract information from these training samples and develop decision-making strategies that perform well when deployed in new environments that contain unseen testing samples. Unfortunately, due to the limited sample size of collected training data, distributional shift between the training population and the testing population (e.g., resulting from measurement error or adversarial perturbations), and the high dimension of the data uncertainty, conventional non-robust approaches often fail to provide reliable solutions. DRO presents a promising approach to tackle these challenges, by finding an optimal decision that performs well even under the most adverse distributions within an prescribed distributional uncertainty set (called the ambiguity set). Thus, DRO often under-promises during the training phase but achieves superior out-of-sample performance when deployed in testing environments. My research has focused on advancing theory, algorithm, and applications in this area.

Selecting an appropriate ambiguity set is crucial for DRO, which should simultaneously enable computational tractability, flexibility, and interpretability. In [7], we introduced a new DRO model using the entropic regularized optimal transport distance (the Sinkhorn DRO model) to strike a balance of those requirements: We contributed a first-order method that efficiently identifies near-optimal solutions with low computation and storage costs. Surprisingly, the resulting problems in Sinkhorn DRO can generally be solved with a complexity level same to that of non-robust training. By solving the Sinkhorn DRO model, we obtain both the robust decision and the most adverse distribution with a well-defined continuous probability density. A continuous worst-case distribution is often desirable and has good generalization performance since the testing population distribution may also be continuous. In industrial applications such as self-driving, healthcare, and power systems, it is convenient to leverage the continuous worst-case distribution for "stress-test".

Building on this foundation, I collaborated with scholars from the optimization area (Daniel Kuhn, Andreas Krause, Xin Chen, Niao He, and Yifan Hu) to extend the first-order algorithm for solving Sinkhorn DRO into general applications: We consider the generic setup in which the optimizer uses first-order algorithms to reduce risk, whereas unbiased gradient estimators are no longer accessible. This occurs frequently when solving problems related to robustness (DRO), personalization (end-to-end contextual learning), privacy (meta-learning), and beyond. In those applications, the optimizer can construct gradient estimators with small biases but large computational costs. In [8, 9], we designed optimization algorithms that balance the trade-off between those quantities, aiming to reduce the total computational costs of obtaining the final decision.

DRO has broad applications in operations research and machine learning, and I explored some of them in my research projects. When leveraging data-driven algorithms to make a new decision, it is critical to provide its performance estimate in the testing environment before deploying it in the

real-world system. A bad decision in practice will be highly risky and even cause unethical issues. In [10], we leveraged DRO and statistics to provide the confidence interval estimate of the decision's performance for both safe exploration and optimistic planning concerns: In high-stake mission-critical environments, providing a confidence lower bound enables safety guarantees and thus help to reduce the risk and circumvent catastrophic events; For risk-seeking environments, providing an optimistic upper bound of performance helps to balance the exploration and exploitation trade-off. In [11], we leveraged DRO and bilevel optimization to train artificial intelligence models with robust performance under adversarial data perturbations. We also adopted DRO to provide reliable decisions in healthcare and pricing applications. In [12], we leveraged our proposed Sinkhorn DRO framework for trustworthy sepsis prediction with real datasets collected from Emory University Hospital and Grady Hospital. In [13], we provided a distributionally robust pricing strategy for Blue Summit Supplies eCommerce products.

Two-Sample Testing. Two-sample testing has long been a challenge in statistics, involving the decision to determine whether the difference between two populations is significant based on the collected observations from two groups. This problem is foundational for fairness assessments in machine learning, especially in detecting biases across demographic groups. Classical approaches often struggle to make reliable decisions due to the parametric assumptions on target distributions and the high-dimensional and complex nature of the big data era. I have developed various two-sample tests in my research to resolve fundamental challenges in this topic.

In [14–16], we proposed efficient testing frameworks for small-sample and high-dimensional data, often encountered in deep learning applications. We tackled this challenge by employing a linear (or nonlinear) dimensionality reduction operator that projects data distributions onto 1-dimensional spaces with maximum separability and next performs hypothesis testing on the univariate data. Our framework requires no assumption about the form of the underlying distributions, and the univariate dimensionality reduction operator gives us interpretable tools to distinguish the differences between high-dimensional data samples. The training phase of this framework requires solving a highly non-convex optimization problem, for which we developed several approximation algorithms, including the first-order manifold optimization method with convergence analysis and the convex relaxation with approximation ratio guarantees. We also leveraged tools from high-dimensional statistics to analyze the theoretical performance of our algorithm, demonstrating its flexibility for distinguishing general populations and near-optimal performance.

In [17], we bridged two-sample testing with deep learning and provided the statistical foundations for trustworthy machine learning. Inspired by the empirical success and flexible function representation of neural networks, we designed a neural-net-based two-sample test for high-dimensional samples supported on a low-dimensional manifold, which is often the case for popular machine learning datasets (MNIST, SVHN, CIFAR-10, and ImageNet). Our statistical theory showed that the performance of this approach only depends on the intrinsic dimension, which is much smaller than the feature dimension, indicating it does not have a curse of dimensionality issue.

In [18], we studied the variable selection for two-sample testing, aiming to select the most informative variables to distinguish samples from the two groups. In practical applications, it is crucial to identify interpretable variables that contribute to the inherent differences between populations. For instance, in gene expressions and biological indicators, only a small subset of variables may account for the disparities between normal and abnormal data samples. We adopt the kernel-based method to distinguish samples flexibly, but the corresponding optimization formulation is inherently NP-hard to solve (because it involves sparsity constraints for variable selection). To tackle the challenge, we leverage modern mixed-integer optimization techniques to develop efficient exact and approximation algorithms with strong performance guarantees. We also derived the statistical performance guarantees of our framework for different kernel choices. Finally, we showed that it achieves superior performance and returns interpretable models, enabling effective sparse data selections for two-sample testing.

Future Research Directions. My research will continue to advance the data-driven decision-making framework described above. In addition to furthering the directions I have explored, I am eager to undertake several new research initiatives.

Sequential and Contextual Learning. I am particularly interested in extending Distributionally Robust Optimization (DRO) to modern machine learning areas. One promising avenue is developing robust models with tractable algorithms for sequential decision-making, an area that has not been extensively explored. I plan to begin by investigating the quickest change-point detection problem, which aims to identify abnormal changes in sequential data as quickly as possible. In our initial work, we provided robust detection algorithms assuming that both normal and abnormal samples are independent and identically distributed [19, 20]. A key challenge moving forward will be extending this framework to handle non-stationary data. Another exciting direction is the development of robust decision-making models that account for side information, such as contextual variables. Since side information can influence both the data distribution and the loss function, constructing a flexible uncertainty set that maintains computational tractability presents a unique challenge. I aim to develop data-driven DRO models for contextual learning and create tractable optimization algorithms to address these complex scenarios.

Responsible AI. I am also committed to establishing responsible AI models that promote social good. Currently, AI model performance often varies significantly across subpopulations. For example, medical treatment policies may show degraded performance for underrepresented groups, such as non-white populations, and speech recognition systems frequently underperform for individuals with minority accents. More broadly, machine learning models tend to exhibit differing performance based on demographic attributes like race, gender, or age in applications such as facial recognition, video captioning, language identification, and academic recommendation systems. To address these disparities, it is critical to train machine learning models that optimize for worst-case performance across these subpopulations. This challenge aligns closely with the minimax nature of DRO, making it a promising approach to improving fairness and equity in AI systems. Furthermore, inspired by DRO's focus on finding worst-case distributions, I intend to explore privacy issues in machine learning by generating worst-case adversarial noises that ensure compliance with differential privacy standards.

References

[1] **Jie Wang**, Shenghao Yang, Yanyan Dong, and Yiheng Zhang. On achievable rates of line networks with generalized batched network coding. ***IEEE Journal on Selected Areas in Communications***, 2024.

[2] Yanyan Dong, Shenghao Yang, **Jie Wang**, and Fan Cheng. Throughput and latency analysis for line networks with outage links. ***IEEE Journal on Selected Areas in Information Theory***, 2024.

[3] **Jie Wang**, Zhiyuan Jia, Hoover HF Yin, and Shenghao Yang. Small-sample inferred adaptive recoding for batched network coding. In *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021.

[4] **Jie Wang**, Talha Bozkus, Yao Xie, and Urbashi Mitra. Reliable adaptive recoding for batched network coding with burst-noise channels. In *2023 57th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2023.

[5] Hoover HF Yin, **Jie Wang**, and Sherman SM Chow. Distributionally robust degree optimization for bats codes. In *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024.

[6] Hoover HF Yin and **Jie Wang**. Sparse degree optimization for bats codes. In *2024 Information Theory Workshop (ITW)*, 2024.

[7] **Jie Wang**, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, 2021. Major Revision at Operations Research.
- Winner of INFORMS 2022 Best Poster Award.

[8] Yifan Hu, **Jie Wang**, Yao Xie, Andreas Krause, and Daniel Kuhn. Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36, 2024. **Journal Version to be Submitted to Operations Research**.

[9] Yifan Hu, **Jie Wang**, Xin Chen, and Niao He. Multi-level monte-carlo gradient methods for stochastic optimization with biased oracles. *arXiv preprint arXiv:2408.11084*, 2024. **Under Reviewed**.

[10] **Jie Wang**, Rui Gao, and Hongyuan Zha. Reliable off-policy evaluation for reinforcement learning. ***Operations Research***, 72(2), 2024.

[11] **Jie Wang**, Rui Gao, and Yao Xie. Regularization for adversarial robust learning. *arXiv preprint arXiv:2408.09672*, 2024. **Journal Version to be Submitted to Operations Research**
• Winner of Best Theoretical Paper Competition at INFORMS 2023 Workshop on Data Mining and Decision Analytics.

[12] **Jie Wang**, Ronald Moore, Yao Xie, and Rishikesan Kamaleswaran. Improving sepsis prediction model generalization with optimal transport. In *Machine Learning for Health*. PMLR, 2022.

[13] **Jie Wang**. Reliable offline pricing in ecommerce decision-making: A distributionally robust viewpoint. `https://drive.google.com/file/d/14Z1Verq3lBJOONL3CR921YeIreSRF9jH/view?usp=sharing`, 2023.
•Finalist for Data Challenge Competition in INFORMS 2023 Workshop on Data Mining and Decision Analytics.

[14] **Jie Wang**, Rui Gao, and Yao Xie. Two-sample test using projected wasserstein distance. In *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021.

[15] **Jie Wang**, Rui Gao, and Yao Xie. Two-sample test with kernel projected wasserstein distance. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, 2022.
• Selected as Oral Presentation (Acceptance rate 44/1685=2.6%).

[16] **Jie Wang**, March Boedihardjo, and Yao Xie. Statistical and computational guarantees of kernel max-sliced wasserstein distances. *arXiv preprint arXiv:2405.15441*, 2024.
• Finalist of INFORMS 2024 Data Mining Best Paper Award Competition.

[17] **Jie Wang**, Minshuo Chen, Tuo Zhao, Wenjing Liao, and Yao Xie. A manifold two-sample test study: integral probability metric with neural networks. ***Information and Inference: A Journal of the IMA***, 12(3), 2023.

[18] **Jie Wang**, Santanu S Dey, and Yao Xie. Variable selection for kernel two-sample tests. *arXiv preprint arXiv:2302.07415*, 2023. **Under Reviewed**
• Runner-up of the INFORMS 2024 Computing Society (ICS) Student Paper Award.

[19] **Jie Wang**, Rui Gao, and Yao Xie. Non-convex robust hypothesis testing using sinkhorn uncertainty sets. In *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024.

[20] **Jie Wang** and Yao Xie. A data-driven approach to robust hypothesis testing using sinkhorn uncertainty sets. In *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022.