Xingjian's Statement of Purposes

During the past decade, the trajectory of my experience has led me to the convergence of algorithms, mathematics, and deep learning. Within this intersecting area lies elegant theories and practical models that, when combined, will create intelligent systems with both spontaneous thinking and logical reasoning abilities, able to perceive, think, and explore in a wide range of settings beyond human-level capability.

Algorithms were my idea of fun in high school: Participating in the Chinese Olympiad of Informatics, advancing to the USACO camp, and winning 1st place in the Canadian Computing Olympiad national team selection, I discussed algorithms and competed with friends all over the world. Later, I studied algorithms with theoretical foundations; through the Fast Fourier Transform, the Miller-Rabin primality test, and Euler's iterative method, I glimpsed a dazzling world behind codes, of complicated analytical and algebraic structures I could not yet understand. To study the theories behind algorithms, I chose to major in mathematics and computer science at Oxford.

My tipping point occurred when I learned about the simulated annealing algorithm. Looking back, the realization I had was remarkably similar to what Prof. Sanjeev Arora described in his blog, *Off the Convex Path*. It was the first algorithm I encountered that was not provably correct, yet possessed the power to tackle NP-hard problems in real settings. Suddenly, all of the classic algorithms collapsed into a plane in my brain – a plane of idealized worlds without errors. However, the points outside of the plane are countless and attractive. Real-world problems are intractable, erroneous, and stochastic in nature. What, then, is the suitable paradigm to solve them? Intrigued by this question, I entered the world of deep learning, seeking ways to combine its strengths into the paradigm of provable algorithms.

Learning-Augmented Sorting. My first exploration came about using inaccurate advice from ML models to speed up sorting, perhaps the most fundamental algorithmic task. This line of work, started by Vassilvitskii[1], seeks to design algorithms with theoretical guarantees that can leverage predictions from ML models. Previous works have proposed sorting algorithms that leverage noisy signals and output correctly with high probability, but no one has proposed provably correct sorting algorithms with erroneous advice. Working with Prof. Christian Coester at Oxford, I proposed a setting where a quick-and-dirty comparison function is available, besides the original clean one. I drew insights from self-balanced data structures and various randomized algorithms, and designed a novel algorithm that can precisely sort using a small number of clean comparisons, if most of the given dirty comparisons are correct. I also proved that my algorithm is optimal in complexity, and experimentally demonstrated its effectiveness of leveraging predictions.

This work, which will appear at **NeurIPS 2023**, has attracted attention from the theoretical and application worlds. While attending the Cargese Workshop on Combinatorial Optimization, a professor told me that he was considering teaching my algorithm in the first algorithmic course at his university. Researchers from a drug discovery lab told me that, leveraging my new results and ML models, they could largely reduce the number of comparative experiments needed to select molecular structures among thousands of candidates, while still producing valid conclusions.

Wasserstein Distributional Adversarial Robustness of Neural Networks. Unlike algorithms, neural networks (NNs) are vulnerable to worst-case attacks. The lack of adversarial robustness prevents NNs from being safely deployed in out-of-sample real-world settings. However, there were no suitable theoretical tools to analyze the adversarial behavior of NNs. Working with Prof. Jan Obloj at Oxford, we have used optimal transport to lift the robustness problem into infinite-dimensional space, and then leveraged sensitivity analysis tools in Distributional Robust Optimization (DRO). As the only team member with an ML background, I translated complicated mathematical tools into tractable algorithms. I proposed a new loss function, Rectified DLR, along with an adversarial attack algorithm, to evaluate the robustness of neural networks. To mitigate the intractability of minimax optimization in robust training, I employed offline learning techniques from RL to propose a new training algorithm. This work will appear at **NeurIPS 2023**.

Neuro-symbolic control on diffusion models. Diffusion models have all the strengths of "System 1 thinking", as defined in the book, *thinking, fast and slow*: they excel at depicting emotions, styles, and scene

[1] Thodoris Lykouris and Sergei Vassilvitskii, "Competitive caching with machine learned advice," *CoRR* abs/1802.05399 (2018), http://arxiv.org/abs/1802.05399.

diversity in the generated images. However, without symbolic control, they fall short in compositionality, consistency, and generalizability, the strengths of logical systems. To address this issue, I worked with Prof. Jiajun Wu at Stanford, focusing on improving the relational compositionality of diffusion models. I proposed a pipeline to decompose the generation task based on scene graphs into the denoising process of single objects and relations, designed a two-stage generative pipeline using object bounding boxes as a bridge, and achieved state-of-the-art performance in generating multiple objects with relations on the CLEVR dataset.

Future Directions. My ultimate aspiration is to create learning systems that combine the rapid and spontaneous capabilities of ML models with the logical and symbolic reasoning of algorithms. Such systems would outperform pure ML models in consistency, interpretability, and compositionality, while also surpassing classic algorithms in adaptability and handling complexity in dynamic and complex environments. This hybrid pipeline would perceive, understand, and interact with the world in both instinctive and logical ways.

Several research directions immediately pique my interest:

Better Interfaces for ML-Augmented Algorithms: I aim to design better interfaces for algorithmic pipelines that include ML models as their components. Currently, most work in learning-augmented algorithms considers predictions with discrete structures, for the convenience of complexity analysis. This leads to a loss of expressive power during communication. Like two powerful machines, running separately and only connected by a thin tube. I aim to redesign the way ML models interact with classic algorithms inside the pipeline, e.g., making it probabilistic and bidirectional, enabling effective information transfer and integration between two sides. More broadly, I'm keen to redesign classic algorithms with ML enhancements in fields like statistics, integer programming, and even physics, where the focus has traditionally been on worst-case analysis. Following my exploration in sorting, I found quick-and-dirty subroutines useful and widely applicable in scientific settings. I aim to design algorithms that can seamlessly leverage dirty subroutines as with clean ones, while maintaining resilience against occasional inaccuracies.

Neural-symbolic, neural-algorithmic Systems: Neural-symbolic models have gained attention in recent years for their enhanced reasoning capabilities compared with end-to-end pipelines. I am excited to extend this line of work to generative models, and to more complicated algorithmic control. I plan to extract executable programs from prompts, which perform reasoning and decompose generation tasks into multiple calls of simple component-wise generative functions. By exploiting the mechanism of recursion or pseudo randomness in the program, such an approach would be consistent in reflecting the hierarchical and probabilistic nature of generated images. In addition, the success in neural-symbolic models has inspired me to replace some other “black boxes” inside machine learning pipelines with algorithmic systems, to achieve stronger theoretical properties for the entire pipeline. For example, in my thesis at Visual Geometry Group, I replaced the denoising component of diffusion models with a fixed-point dynamic system, enabling the reuse of solutions previously impossible for UNets. With my insights into classic algorithms, I will be able to identify exploitable structures in real settings and design suitable algorithms to integrate into ML pipelines.

Why Columbia? Columbia is undoubtedly the optimal place for me to pursue my research aspirations. I’d be excited to continue working on neural-symbolic controls for generative models with Prof. **Carl Vondrik**. His research has a taste of theoretical insights, which have greatly inspired me. I would be interested in enhancing the controlling methods in his work ViperGPT and adopting them to more domains. I am also a good match to work with Prof. **Christos Papadimitriou**. Under his guidance, I would love to use the algorithmic lens to examine deep learning models. In a similar vein, I am also interested in Prof. **Tim Roughgarden**’s work on algorithms. Under his guidance, I can explore the limits of classic algorithms and how to reach beyond them.

Quantifying Goodharting in Reinforcement Learning (RL). In the realm of RL, crafting reward functions that accurately encapsulate real-world tasks is nearly impossible. Researchers have observed *Goodharting* behaviors – over-optimized agents that find unintended ways to increase their rewards. During my project at the Oxford AI Safety Lab, I was on a team that aimed to provide a principled understanding of the behavior of RL agents when a proxy reward is maximized. By modeling the policy trajectory of the Maximal Causal

Entropy method under occupancy measures, we provided a geometric interpretation of why Goodharting emerges in Markov decision processes. Then, I derived two policy optimization methods that provably avoid Goodharting. This research, under review at **ICLR2024**, has attracted considerable interest from the AI Alignment community.