# General and Scalable Optimization for Robust AI

**Sijia Liu**

Department of Computer Science & Engineering, Michigan State University, MI, USA
MIT-IBM Watson AI Lab, IBM Research, USA
liusiji5@msu.edu

## Abstract

Deep neural networks (DNNs) can easily be manipulated (by an adversary) to output drastically different predictions and can be done so in a controlled and directed way. This process is known as adversarial attack and is considered one of the major hurdles in using DNNs in high-stakes and real-world applications. Although developing methods to secure DNNs against adversaries is now a primary research focus, it suffers from limitations such as *lack of optimization generality* and *lack of optimization scalability*. My research highlights will offer a holistic understanding of optimization foundations for robust AI, peer into their emerging challenges, and present recent solutions developed by my research group.

## Research Highlights Description

I plan to review and foster new trends and advances in general and scalable optimization for robust AI across the full theory/algorithm/application stack.

**Enhancing optimization generality for robust AI.** I will investigate the challenges of the lack of algorithmic generality in robust AI. Taking adversarial robustness as an example, nearly all existing work adopted min-max optimization (MMO) as the algorithmic backbone of robust learning. Yet, the MMO principle requests the defender and the adversary to share the same objective function type, leading to a poor generality of robust learning against diverse adversarial attacks. By contrast, my recent work (Zhang et al. 2022d) has shown that bi-level optimization (BLO) can facilitate us to build a general and scalable robust training framework as it allows us to customize its lower-level objectives to incorporate different types of adversaries. The strength of BLO, resulting in a principled hierarchical learning framework, also occurs in tackling the model efficiency challenge in AI, supported by our work in (Zhang et al. 2022c).

**Enhancing optimization scalability for robust AI.** In this part, I will investigate the scalability challenges of robust AI in the algorithm, model, and data planes. These challenges include computationally-intensive robust training (with difficulty scaling to large foundational models) and restricted white-box model assumption (with the incapability of robustifying private, black-box AI models). To address these challenges, I will introduce my contributions to scalable optimization for robust AI, including distributed robust learning (DRL) (Zhang et al. 2022a) and zeroth-order optimization (ZOO) (Liu et al. 2020). Although DRL can speed up training by making full use of the computing capability of multiple data-locality (distributed) machines, it suffers from a large-batch optimization challenge (*i.e.*, the lack of stochasticity makes the robust learner challenging to escape from a sharp local minimum). Thus, I will delve into DRL's theory and practice when integrated with large-batch optimization approaches. In addition, I will review ZOO techniques to scale robust learning to black-box models using only input-output model queries. I will illustrate a promising connection between ZOO and black-box attack and defense and introduce practical ZOO algorithms with graceful scalability in high dimensions (Zhang et al. 2022b).



## Biography

**Sijia Liu** is currently an Assistant Professor at the CSE department of Michigan State University, and an Affiliated Professor at the MIT-IBM Watson AI Lab, IBM Research. His research spans the areas of machine learning and computer vision, with a focus on trustworthy and scalable AI. He received the Best Paper Runner-Up Award at UAI'22. He also received the Best Student Paper Award at ICASSP'17.

## References

Liu, S.; Chen, P.-Y.; Kailkhura, B.; Zhang, G.; Hero III, A. O.; and Varshney, P. K. 2020. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5): 43–54.

Zhang, G.; Lu, S.; Zhang, Y.; Chen, X.; Chen, P.-Y.; Fan, Q.; Martie, L.; Horesh, L.; Hong, M.; and Liu, S. 2022a. Distributed adversarial training to robustify deep neural networks at scale. In *UAI*, 2353–2363.

Zhang, Y.; Yao, Y.; Jia, J.; Yi, J.; Hong, M.; Chang, S.; and Liu, S. 2022b. How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective. In *ICLR*.

Zhang, Y.; Yao, Y.; Ram, P.; Zhao, P.; Chen, T.; Hong, M.; Wang, Y.; and Liu, S. 2022c. Advancing Model Pruning via Bi-level Optimization. In *NeurIPS*.

Zhang, Y.; Zhang, G.; Khanduri, P.; Hong, M.; Chang, S.; and Liu, S. 2022d. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *ICML*, 26693–26712.