WENZHI FANG

BHEE 051, Electrical Engineering Building, Northwestern Avenue, West Lafayette, IN
◇ fang375@purdue.edu ◇ Phone: (765)694-5334 ◇ Google Scholar

EDUCATION

Purdue University *West Lafayette, IN, US*
Elmore Family School of Electrical and Computer Engineering
Ph. D. Candidate in Electrical and Computer Engineering *Aug. 2023 – 2027*

ShanghaiTech University *Shanghai, China*
School of Information and Science Technology
M.S. in Communication and Information Systems *Sept. 2020 – Jul. 2023*

Shanghai University *Shanghai, China*
School of Communication and Information Engineering
B.S. in Communication Engineering *Sept. 2016 – Jul. 2020*

MAJOR COURSE

- Machine Learning, Deep Learning, Reinforcement Learning
- Bayesian Data Analysis, Convex Optimization, Matrix Computation

RESEARCH INTERESTS

- **LLM.** Efficient Fine-Tuning of LLM & RL-based Post-training
- **FL.** Federated Learning & Distributed Optimization

TECHNICAL STRENGTHS

Technical Skills Python, Torch, vLLM, VERL, Git

SELECTED WORKS

Work on Large Language Models

[1] W. Fang, D-J. Han, L. Yuan, and C. G. Brinton, **Collaborative Device-Cloud LLM Inference through Reinforcement Learning**, 2025, *Under Review* [Paper]

[2] W. Fang, D-J. Han, L. Yuan, S. Hosseinalipour, and C. G. Brinton, **Federated Sketching LoRA: On-Device Collaborative Fine-Tuning of Large Language Models**, 2025, *Under Review* [Paper]

[3] J. Lee, W. Fang, D-J. Han, S. Hosseinalipour, and C. G. Brinton, **TAP: Two-Stage Adaptive Personalization of Multi-task and Multi-Modal Foundation Models in Federated Learning**, 2025, *Under Review* [Paper]

Work on Federated Learning and Optimization

[4] W. Fang, D-J. Han, E, Chen, S. Wang, and C. G. Brinton, **Hierarchical Federated Learning with Multi-Timescale Gradient Correction**, *Neural Information Processing Systems (NeurIPS) 2024*. [Paper]

[5] W. Fang, D-J. Han, and C. G. Brinton, **Federated Learning over Hierarchical Wireless Networks: Training Latency Minimization via Submodel Partitioning**, *IEEE/ACM Transactions On Networking (ToN) 2025* [Paper]

[6] W. Fang, Z. Yu, Y. Jiang, Y. Shi, C. Jones, and Y. Zhou, **Communication-Efficient Stochastic Zeroth-Order Optimization for Federated Learning**, *IEEE Transactions on Signal Processing (TSP) 2022.* [Paper]

Highlight

- In [1], we propose a unified RL-based post-training framework for device–cloud LLM collaborative inference that integrates routing optimization, enabling on-device LLMs to simultaneously enhance problem-solving, acquire effective routing strategies, and narrow the performance gap to full cloud LLMs.
- In [2], we propose federated sketching LoRA (FSLoRA), a theoretically-grounded methodology that retains LoRA's flexibility while adapting to the communication and computational capabilities of individual devices.
- In [3], we propose TAP (Two-Stage Adaptive Personalization), a federated learning framework that personalizes multi-task and multi-modal foundation models by integrating selective local adaptation with post-FL knowledge distillation to balance task-specific performance and generalization.
- In [4], we proposed an algorithm to address multi-level data heterogeneity in hierarchical federated learning (HFL), deriving strong theoretical results without relying on additional data heterogeneity assumptions. This work fills a critical gap in the existing HFL literature.
- In [5], we investigated the idea of model partitioning on some classical models, such as FCNs and CNNs, and on the modern transformer architecture, to reduce the training consumption.
- In [6], we proposed a federated zeroth-order algorithm (FedZO) with a convergence guarantee. This algorithm makes the training process forward-only, eliminating the memory overhead of backward propagation, which has since inspired numerous works in LLMs.

WORKING EXPERIENCE

Optimization for Machine Learning Lab *Aug., 2022 - Feb. 2023*

Summer Intern *Advisor: Prof. Peter Richtarik* KAUST

ION Lab *Aug., 2023 - Present*

Research Assistant *Advisor: Prof. Christopher G. Brinton* Purdue University

TEACHING EXPERIENCE

SI263: Distributed Optimization *Spring, 2022, ShanghaiTech*

ACEDEMIC SERVICE

Reviewer of NeurIPS, ICML, ICLR, AISTAT, TMLR, AAAI

CONTESTS AND AWARDS

China National Scholarship (Top 0.2% Nationwide), *2021*

First prize of China National Undergraduate Electronic Design Competition, *2019*

First prize of Chinese Mathematics Competitions, Shanghai, *2017*