

Lei Wu

Room 372, Zhihua Building, Peking University, No. 5 Yiheyuan Road, Beijing, China
leiwu@math.pku.edu.cn • leiwu0.github.io • Google Scholar

ACADEMIC POSITIONS

- **PEKING UNIVERSITY**, Beijing, China 2021.12 – present
Assistant Professor, School of Mathematical Sciences & Center for Machine Learning Research
- **UNIVERSITY OF PENNSYLVANIA**, Philadelphia, USA 2021.08 – 2021.11
Postdoctoral Researcher, Department of Statistics and Data Science
Mentor: Weijie Su
- **PRINCETON UNIVERSITY**, Princeton, USA 2018.11 – 2021.07
Postdoctoral Researcher, Program in Applied and Computational Mathematics
Mentor: Weinan E

EDUCATION

- **PEKING UNIVERSITY**, Beijing, China 2012.09 – 2018.07
Ph.D. in Computational Mathematics
Advisor: Weinan E
- **NANKAI UNIVERSITY**, Tianjin, China 2008.09 – 2012.06
B.S. in Mathematics

RESEARCH INTERESTS

- **Scientific foundations of deep learning**
Scaling laws, learning dynamics, approximation, and generalization.
- **Large-scale pretraining**
Optimizers, hyperparameters, architectures, and data.

GRANTS

- **NSFC Young Scientists Fund (B)** 2026 – 2028
Mathematical Theory of Deep Learning; PI, CNY 2,000,000.
- **NSFC Major Research Plan (Fostering Program)** 2025 – 2027
Stability-Based Training Algorithms for Deep Learning; PI, CNY 800,000.
- **Start-up Grant, Peking University** 2022 – 2024
PI, CNY 500,000.

PUBLICATIONS REFEREED*

- [38] Z. Wang[†], B. Li[†], J.-N. Wang[†], L. Wang[†], J. Wang[†], **L. Wu**, “When and How signSGD Outperform SGD: A Functional Scaling Law Perspective,” *NeurIPS 2026 (Spotlight)*.
- [37] B. Li[†], Z. Wang[†], F. Chen, S. Zhao, R. Zheng, **L. Wu**, “Optimal Learning-Rate Schedules under Functional Scaling Laws: Power Decay and Warmup-Stable-Decay,” *COLT 2026*.
- [36] J. Wang[†], M. Wang[†], J. Zhang, W. Wang, P. Pei, X. Cai, W. E, **L. Wu**, “GradPower: Powering Gradients for Faster Language Model Pre-Training,” *ICML 2026*.
- [35] J. Wang[†], B. Li[†], Z. Zhou, M. Wang, Y. Sun, J. Zhang, X. Cai, **L. Wu**, “Fast Catch-Up, Late Switching: Optimal Batch Size Scheduling via Functional Scaling Laws,” *ICLR 2026*.
- [34] M. Wang, **L. Wu**, “The Noise Geometry of Stochastic Gradient Descent,” *Transactions on Machine Learning Research*, 2026.
- [33] B. Li[†], F. Chen[†], Z. Huang[†], L. Wang[†], **L. Wu**, “Functional Scaling Laws in Kernel Regression: Loss Dynamics and Learning Rate Schedules,” *NeurIPS 2025 (Spotlight)*.
- [32] C. Zeng, Y. Zhang, J. Zhou, Y. Wang, Z. Wang, Y. Liu, **L. Wu**, D.Z. Huang, “Point Cloud Neural Operator for Parametric PDEs on Complex and Variable Geometries,” *Comput. Methods Appl. Mech. Eng.*, 2025.
- [31] J. Wang[†], M. Wang[†], Z. Zhou[†], J. Yan, W. E, **L. Wu**, “The Sharpness Disparity Principle in Transformers for Accelerating Language Model Pre-Training,” *ICML 2025*.

*† indicates alphabetical ordering or equal contribution.

- [30] K. Zhan[†], P. Li[†], **L. Wu**, “Analyzing the role of permutation invariance in linear mode connectivity,” *AISTATS* 2025.
- [29] H. Chen[†], J. Long[†], **L. Wu**[†], “A duality framework for analyzing random feature and two-layer neural networks,” *Annals of Statistics*, 2025.
- [28] M. Wang, J. Wang, H. He, Z. Wang, G. Huang, F. Xiong, Z. Li, **L. Wu**, “Improving generalization and convergence by enhancing implicit regularization,” *NeurIPS* 2024.
- [27] Z. Liu, M. Wang, H. Li, **L. Wu**, “Parameter symmetry and noise equilibrium of stochastic gradient descent,” *NeurIPS* 2024.
- [26] M.A. Mohamadi, Z. Li, **L. Wu**, D.J. Sutherland, “Why do you grok? A theoretical analysis on grokking modular addition,” *ICML* 2024.
- [25] K. Liu[†], Z. Wang[†], **L. Wu**, “The local landscape of phase retrieval under limited samples,” *IEEE Transactions on Information Theory*, 2024.
- [24] M. Wang, Z. Min, **L. Wu**, “Achieving margin maximization exponentially fast via progressive norm rescaling,” *ICML* 2024.
- [23] **L. Wu**, “Embedding inequalities for Barron-type spaces,” *Journal of Machine Learning*, 2023.
- [22] Z. Wang, **L. Wu**, “Theoretical analysis of inductive biases in deep convolutional networks,” *NeurIPS* 2023.
- [21] **L. Wu**, W.J. Su, “The implicit regularization of dynamical stability in stochastic gradient descent,” *ICML* 2023.
- [20] **L. Wu**, M. Wang, W.J. Su, “The alignment property of SGD noise and how it helps select flat minima: A stability analysis,” *NeurIPS* 2022.
- [19] C. Ma, D. Kunin, **L. Wu**, L. Ying, “Beyond the quadratic approximation: the multiscale structure of neural network loss landscapes,” *Journal of Machine Learning*, 2022.
- [18] **L. Wu**, J. Long, “A spectral-based analysis of the separation between two-layer neural networks and linear methods,” *Journal of Machine Learning Research*, 2022.
- [17] **L. Wu**, “Learning a single neuron for non-monotonic activation functions,” *AISTATS* 2022.
- [16] C. Ma[†], **L. Wu**[†], W. E, “A qualitative study of the dynamic behavior of adaptive gradient algorithms,” *Mathematical and Scientific Machine Learning (MSML)*, 2021.
- [15] W. E[†], C. Ma[†], **L. Wu**[†], “The Barron space and flow-induced function spaces for neural network models,” *Constructive Approximation*, 2021.
- [14] W. E[†], C. Ma[†], S. Wojtowytsch[†], **L. Wu**[†], “Towards a mathematical understanding of neural network-based machine learning: what we know and what we don’t,” *CSIAM Transactions on Applied Mathematics*, 2020.
- [13] H. Lei, **L. Wu**, W. E, “Machine learning based non-Newtonian fluid model with molecular fidelity,” *Physical Review E*, 2020.
- [12] W. E[†], C. Ma[†], **L. Wu**[†], “Machine learning from a continuous viewpoint, I,” *Science China Mathematics*, 2020.
- [11] C. Ma[†], **L. Wu**[†], W. E, “The slow deterioration of the generalization error of the random feature model,” *Mathematical and Scientific Machine Learning (MSML)*, 2020.
- [10] W. E[†], C. Ma[†], **L. Wu**[†], “A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics,” *Science China Mathematics*, 2020.
- [9] W. E[†], C. Ma[†], **L. Wu**[†], “The generalization error of minimum-norm solutions for over-parameterized neural networks,” *Journal of Pure and Applied Functional Analysis*, 2020.
- [8] **L. Wu**, Z. Zhu, “Towards understanding and improving the transferability of adversarial examples in deep neural networks,” *ACML*, 2020.
- [7] **L. Wu**[†], Q. Wang[†], C. Ma, “Global convergence of gradient descent for deep linear residual networks,” *NeurIPS* 2019.
- [6] W. E[†], C. Ma[†], **L. Wu**[†], “A priori estimates of the population risk for two-layer neural networks,” *Communications in Mathematical Sciences*, 2019.
- [5] Z. Zhu, J. Wu, B. Yu, **L. Wu**, J. Ma, “The anisotropic noise in stochastic gradient descent: Its behavior of escaping from minima and regularization effects,” *ICML* 2019.
- [4] **L. Wu**, C. Ma, W. E, “How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective,” *NeurIPS* 2018.

- [3] Y.-A. Ma, E.B. Fox, T. Chen, **L. Wu**, “Irreversible samplers from jump and continuous Markov processes,” *Statistics and Computing*, 2018.
- [2] **L. Wu**, Z. Zhu, W. E, “Towards understanding generalization of deep learning: perspective of loss landscapes,” *Workshop on Principled Approaches to Deep Learning, ICML 2017*.
- [1] H. Lei, N.A. Baker, **L. Wu**, G.K. Schenter, C.J. Mundy, A.M. Tartakovsky, “Smoothed dissipative particle dynamics model for mesoscopic multiphase flows in the presence of thermal fluctuations,” *Physical Review E*, 2016.

PREPRINTS

- [9] J.-N. Wang[†], Z. Huang[†], K. Li[†], Z. Wang, **L. Wu**, “Momentum in large-batch training: Polyak enlarges the critical batch size, Nesterov improves data efficiency,” *arXiv preprint*, 2026.
- [8] Z. Liu[†], R. Zheng[†], S. Zhang, C. Tian, K. Chen, Z. Zhang, **L. Wu**, “Effective Learning Rate Governs Loss Dynamics in Language Model Pretraining,” *arXiv preprint*, 2026.
- [7] Y. Liu[†], Z. Wang[†], **L. Wu**[†], S. Zhang[†], “Smoothness Adaptivity in Constant-Depth Neural Networks: Optimal Rates via Smooth Activations,” *arXiv preprint*, 2026.
- [6] Z. Lin[†], P. Li[†], **L. Wu**, “Exploring neural network landscapes: star-shaped and geodesic connectivity,” *arXiv preprint*, 2024.
- [5] J. Long[†], X. Peng[†], **L. Wu**[†], “Optimal Rates and Saturation for Noiseless Kernel Ridge Regression,” *arXiv preprint*, 2024.
- [4] H. Chen, J. Long, **L. Wu**, “The L^∞ Learnability of Reproducing Kernel Hilbert Spaces,” *arXiv preprint*, 2023.
- [3] C. Ma[†], **L. Wu**[†], W. E, “The quenching-activation behavior of the gradient descent dynamics for two-layer neural network models,” *arXiv preprint*, 2020.
- [2] Z. Li[†], C. Ma[†], **L. Wu**[†], “Complexity measures for neural networks with general activation functions using path-based norms,” *arXiv preprint*, 2020.
- [1] W. E[†], C. Ma[†], **L. Wu**[†], “Analysis of the gradient descent algorithm for a deep neural network model with skip-connections,” *arXiv preprint*, 2019.

TEACHING

PEKING UNIVERSITY

- Topics in Deep Learning Theory 2024 – present
- Mathematical Introduction to Machine Learning 2023 – present
- Advanced Mathematics (B) 2023
- Introduction to Data Science 2022

OTHER TEACHING

- Mathematical Introduction to Machine Learning, Princeton University (co-instructor) 2021
- Summer schools on deep learning:
 - Fudan University 2026
 - Peking University 2018, 2019, 2021
 - Shanghai Jiao Tong University 2020

ACADEMIC

ORGANIZATION

- Organizer, Foundations of Modern Intelligence (FoMI) Seminar (ongoing).
- Co-organizer, PKU Workshop on Mathematical Foundations of Modern Machine Learning, 2026.
- Co-organizer, the 3rd China Conference on Scientific Machine Learning (CSML), 2025.

REVIEWING

- *Conferences*: NeurIPS, ICML, ICLR, AISTATS.
- *Journals*: JMLR; SIAM J. Numer. Anal.; SIAM J. Math. Data Sci.; SIAM J. Math. Anal.; Constructive Approximation; J. Sci. Comput.; Ann. Appl. Probab.; Mathematical Programming.