

Research Statement — Hongyang R. Zhang (May, 2026)

As machine learning models and systems grow in scale and complexity, understanding when and why they work becomes a central challenge. My research is organized around three connected themes: second-order perspectives on generalization and optimization for modern machine learning, multitask learning and foundation models, and learning/reasoning on large-scale networks.

I draw on techniques from optimization, learning theory, probability, and mechanism design to advance the foundations of machine learning. My students and I also extensively evaluate these methods in applied settings, including natural language processing and transportation. Our goal is to develop theoretical tools that lead to practical algorithms; in particular, we are committed to open-source development to further the impact of our work on the broader community.

Generalization and Optimization of Neural Nets: A Second-Order Perspective

Modern neural networks are typically designed with far more parameters than training examples, a regime known as *over-parameterization*. Classical generalization theory, e.g., from the VC dimension or Rademacher complexity perspectives, often yields vacuous bounds that do not explain why gradient descent finds solutions that generalize. From an optimization perspective, while it is empirically observed that gradient descent converges to flat, low-rank solutions in this regime, the theoretical understanding of this implicit bias has been limited beyond simplified linear models. Further, existing generalization bounds for architectures such as graph neural networks are scaled with coarse graph statistics (e.g., maximum degree), giving loose bounds for real-world graphs.

My contribution in this area involves studying generalization and optimization from a *second-order perspective*. First, we formally characterize the implicit regularization effect of gradient descent in over-parameterized matrix sensing, from small initialization [18]. We apply the techniques to ultra-sparse matrix completion [26]. Second, we develop a Hessian-based generalization framework for fine-tuning that yields non-vacuous bounds and practical regularization and quantization algorithms [15, 8, 25, 12], as well as state-of-the-art generalization bounds for graph neural networks [7].

Matrix recovery. In a COLT’18 paper [18], we consider gradient descent dynamics in over-parameterized matrix sensing, where we are given linear measurements of an unknown matrix M . We provide a detailed convergence analysis, starting from a small initialization, for recovering M when the number of parameters exceeds the number of training data points. A key insight is that the gradient descent iterates remain close to a low-rank subspace, ultimately converging to the minimum-nuclear-norm solution among all interpolating solutions.

Building on these techniques, we consider matrix completion in the ultra-sparse sampling regime: each entry of the unknown $n \times d$ matrix M is observed with probability $p = C/d$ for some fixed constant C . Cao, Liang, and Valiant [2] established the open problem of whether it is possible to recover one side, or the second-moment matrix $M^\top M/n$ accurately, in this ultra-sparse sampling regime. Our paper resolves it in the affirmative [26]. A key technique is self-normalization, commonly known as the Hájek estimator. We show that this estimator is unbiased for the second-moment matrix and, moreover, reduces variance, yielding more accurate estimates in practice.

Moving from matrix recovery to neural networks, I have considered data augmentation, a practical method for enhancing generalization. In joint work with Greg Valiant and Chris Ré at ICML’20 [22],

we measure the bias and variance of both invariant and mixup transformations, and formalize our findings on augmentations in the over-parameterized matrix regression setting.

Fine-tuning and Hessian-based generalization. Next, I consider supervised fine-tuning, in which gradient descent begins from a pretrained model rather than a random initialization, introducing new challenges that the above results alone cannot address. In a line of work with my students Haotian Ju and Dongyue Li [15, 8, 25], we formally show generalization bounds for fine-tuning. We assume that the fine-tuned model lies within a certain radius from the pretrained model, and show that the generalization gap can be upper-bounded by the (maximum) trace of the Hessian of the loss function over the entire data distribution, times the radius squared. A particularly interesting property of the Hessian-based measurements is that they are non-vacuous, meaning that they can match the scale of empirically observed generalization gaps. This is crucial because the Hessian-based framework can yield meaningful measurements for downstream applications.

Here are several algorithmic implications from this computational framework: (1) We design a noise-injection algorithm with a regularization effect on the Hessian trace of the loss surface and empirically validate this algorithm in a variety of practical settings [25]. (2) We discover that this type of noise-injection can also help accelerate quantization-aware training, which often gets stuck in saddle points [12]. (3) In preliminary work [29], we find empirical evidence that the Hessian trace regularization also mitigates grokking in modular arithmetic tasks.

Graph neural networks. The Hessian trace bound extends naturally to graph-structured data, where we also uncover a connection between this loss curvature and the spectral properties of the graph diffusion matrix. This connection allows us to improve the state-of-the-art generalization bound in graph neural networks [7]. Previous work has shown generalization bounds for graph neural networks that scale with the graph structure, specifically the maximum degree of all vertices. We show a generalization bound that instead scales with the largest singular value of the *graph diffusion matrix*. For example, consider a node classification problem where each node represents a user or a video on a social network, and the goal is to predict each node’s label for a recommendation system. In graph convolutional networks, the largest singular value of the normalized graph Laplacian is at most one. These bounds are numerically much smaller than prior bounds for real-world graphs.

Theoretical Understanding and New Algorithms for Multitask Learning

The problem of multitask learning is as follows: Given several related tasks, how can we train a neural network to make accurate predictions on all of them simultaneously? This line of work is influenced by the development of foundation models, which are often trained on diverse datasets. When different tasks are trained in a network with shared parameters, how does information from one task transfer to another task? Relatedly, how can we identify the most helpful tasks to train alongside another downstream task? My contribution includes (i) theoretical understanding of information transfer in multitask networks [21, 23], (ii) practical algorithms for task/data partitioning [11, 14, 27] and selection [17, 31], and (iii) second-order analysis of task attribution that rigorously connects influence functions to linear surrogate modeling [13, 28].

Theoretical understanding of information transfer. In an ICLR’20 paper with Chris Ré [21], we formally study transfer by relating multi-headed neural networks—a common architecture for conducting multitask learning—to two-layer neural networks [18]. Our paper represents one of the first in-depth analyses of *negative transfer* in two-layer neural networks. With this connection, questions regarding how one task affects another, etc., become amenable to statistical analysis. In a JMLR’25 paper [23], we improve on the initial result and provide a precise quantification of transfer

in high-dimensional linear regression. We formulate hard-parameter sharing estimation for two linear regression tasks in the high-dimensional, proportional regime. Compared with single-task learning, we show a phase transition from positive transfer to negative transfer as the number of source-task samples increases.

New algorithms for multitask learning. A key insight from this work is that negative transfer becomes inherent when tasks have severe distribution shifts. To scale these insights to foundation models, we develop a surrogate modeling approach to predict the performance of learning multiple tasks simultaneously [13]. In a series of papers with my student Dongyue Li, in collaboration with Aneesh Sharma and Lu Wang, we design convex relaxation algorithms to find approximate partitioning of tasks in multitask learning [11, 14, 17, 31]. The optimization program is based on an affinity matrix that captures task relationships and is estimated using a surrogate modeling approach. This yields a random-forest-style algorithm that captures higher-order correlations among tasks more accurately than existing methods, and can be efficiently implemented on top of foundation models using a linearization technique, reminiscent of neural tangent kernels. We have extensively validated this approach in a variety of downstream applications, including *overlapping community detection* [11, 14], *ensemble low-rank adaptation* of language models [17], and *demonstration selection for in-context learning* [31].

A related problem that is amenable to the above techniques is *multi-objective reinforcement learning*, which involves balancing multiple conflicting objectives in RL. This problem has broad applications in modern AI, such as in alignment and in robotics. In an AAAI'26 paper [27], we apply the above approach on top of proximal policy gradient to partition similar trajectories into groups. A key insight is to design a routing mechanism that directs a trajectory to the partition that yields the highest reward.

Task attribution and Hessian analysis in language models. Modern AI models are trained on diverse tasks, leading to the question of quantifying the influence of individual tasks upon a model, a problem we refer to as *task attribution*. This problem is closely related to the local geometry of loss landscapes, which can be captured by the Hessian matrix of the loss function. Prior work has captured this connection through influence functions [9] and Hessian eigenmaps [4]. In an ICLR'26 paper with my students Zhenshuo Zhang and Minxuan Duan [28], we rigorously connect influence functions to linear surrogate modeling [13], an empirical procedure practitioners often use to attribute the influence of data on the trained model.

Learning and Reasoning on Large-Scale Networks

My work in this area spans learning and algorithmic reasoning on large-scale network data. This direction provides both a domain for testing the preceding ideas and a rich source of connections to classical algorithms and tools in spectral graph theory. Earlier contributions include the first running-time analysis of local push (a deterministic algorithm for computing personalized PageRank) on dynamic graphs [24], in which we propose and analyze natural dynamic versions of known local variations of power methods for solving linear systems [1]. Another contribution is a spectral algorithm for reducing epidemic diffusion on weighted graphs by minimizing the sum of the k largest eigenvalues [10], generalizing earlier work in this literature that reduces the top eigenvalue [20].

Traffic accident prediction on road networks: We have also contributed to this space through collecting new graph datasets (with my students Ziniu Zhang and Abhinav Nippani) [19, 30]. We collected traffic accident records from the Department of Transportation websites and constructed a large-scale graph dataset representing road networks in eight states in the US, including Massachusetts. Using this dataset, we study traffic accident prediction on road networks using graph neural networks

(GNNs). We have also collected high-resolution satellite images taken at the road segments. Our results suggest that combining graph neural networks and satellite image embeddings can predict accident occurrences with an AUROC over 90%.

Algorithmic reasoning on graphs: Can neural networks learn to follow the behavior of an algorithm? We study this question using inputs from twelve classical algorithms, such as BFS, DFS, Dijkstra, Kruskal, Strongly Connected Components (SCC), and Floyd-Warshall [16]. We train graph neural networks to predict the intermediate executions and final answers of these algorithms on Erdős–Rényi graphs with 16 nodes. Building on the multitask learning framework from the previous section, our contribution is a hierarchical, branching GNN that simultaneously learns multiple algorithmic reasoning tasks. We find that GNNs can learn simple algorithms such as BFS with over 99% accuracy and SCC with over 90% accuracy, while struggling with more complex algorithms such as Floyd-Warshall, achieving just 63% accuracy. The accuracy for DFS is less than 40%, as with other baseline architectures, suggesting that the recursive search process in DFS is particularly difficult to learn. This work raises several questions regarding the learnability of an algorithm via neural networks and length generalization to larger graphs/sequences.

Future Directions

Estimation of Hessian spectral statistics, large models, and safety. The work above suggests that taking a second-order perspective can provide valuable insights into the generalization, optimization, and interpretation of large models. There are several fundamental questions in this direction. For a modern computational environment, what is the complexity for accurately estimating the Hessian trace (and Hessian spectral density), in terms of the number of Hessian vector product calls one would need? Our latest findings suggest that SGD tends to converge to a stationary point with a significant portion of both positive and negative eigenvalues in language models [12]. How do we reconcile these findings with the common belief that SGD often converges to a local minimum [6]?

Another direction is to apply this approach to study catastrophic forgetting in safety alignment, where models trained with high-safety procedures and later fine-tuned on another task tend to “forget” their safety standards and start to perform poorly again on safety metrics. More broadly, there is an open area in which one formulates various metrics, including safety and honesty, into models through robust optimization. Beyond language models, it may also be interesting to examine this second-order perspective in fundamentally different settings from classical computing.

Alignment, social choice, and decision-making. This direction studies alignment through the lens of social choice and mechanism design. For example, how should platform designers align individual preferences with collective outcomes such as social welfare? We formalize this question in reinforcement learning from human feedback (RLHF), focusing on group-strategyproof aggregation mechanisms that are robust to coalitional misreporting [5] (building on my earlier work [3]). This work connects robust statistics, approximate mechanism design, and social choice. In particular, it extends classical mechanism-design ideas from facility-location to ranking aggregation for alignment. Further questions involve the role of synthetic data in alignment and richer models of preference elicitation, such as pairwise comparisons and partial orderings.

References

- [1] R. Andersen, C. Borgs, J. Chayes, J. Hopcraft, V. S. Mirrokni, and S.-H. Teng. “Local computation of pagerank contributions”. In: *International Workshop on Algorithms and Models for the Web-Graph*. Springer. 2007, pp. 150–165.

- [2] S. Cao, P. Liang, and G. Valiant. “One-sided Matrix Completion from Two Observations Per Row”. In: *International Conference on Machine Learning (ICML)* (2023).
- [3] N. Chen, X. Deng, B. Tang, H. Zhang, and J. Zhang. “Incentive ratio: A game theoretical analysis of market equilibria”. In: *Information and Computation* 285 (2022).
- [4] D. L. Donoho and C. Grimes. “Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data”. In: *Proceedings of the National Academy of Sciences* 100.10 (2003), pp. 5591–5596.
- [5] M. Duan, D. Li, H. R. Zhang, and Z. Zhang. “Approximate Group-Strategyproof Ranking Aggregation Using the Tukey Pessimistic Median”. Manuscript in preparation. 2026.
- [6] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. “How to escape saddle points efficiently”. In: *International Conference on Machine Learning (ICML)*. 2017, pp. 1724–1732.
- [7] H. Ju, D. Li, A. Sharma, and H. R. Zhang. “Generalization in Graph Neural Networks: Improved PAC-Bayesian Bounds on Graph Diffusion”. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2023, pp. 6314–6341.
- [8] H. Ju, D. Li, and H. R. Zhang. “Robust Fine-tuning of Deep Neural Networks with Hessian-based Generalization Guarantees”. In: *International Conference on Machine Learning (ICML)*. 2022, pp. 10431–10461.
- [9] P. W. Koh and P. Liang. “Understanding black-box predictions via influence functions”. In: *International conference on machine learning (ICML)*. 2017, pp. 1885–1894.
- [10] D. Li, T. Eliassi-Rad, and H. R. Zhang. “Optimal Intervention on Weighted Networks via Edge Centrality”. In: *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. 2023, pp. 424–432.
- [11] D. Li, H. Ju, A. Sharma, and H. R. Zhang. “Boosting multitask learning on graphs through higher-order task affinities”. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 2023, pp. 1213–1222.
- [12] D. Li, Z. Liu, K. Yi, C. Zhao, Z. Zhang, R. Krishnamoorthi, H. Khaitan, H. R. Zhang, and S. Li. “WinQ: Accelerating Quantization-Aware Training of Large Language Models around Saddle Points”. In: *International Conference on Machine Learning (ICML)*. 2026.
- [13] D. Li, H. L. Nguyen, and H. R. Zhang. “Identification of Negative Transfers in Multitask Learning using Surrogate Models”. In: *Transactions on Machine Learning Research (TMLR). Featured Certification* (2023).
- [14] D. Li, A. Sharma, and H. R. Zhang. “Scalable Multitask Learning Using Gradient-based Estimation of Task Affinity”. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 2024, pp. 1542–1553.
- [15] D. Li and H. R. Zhang. “Improved Regularization and Robustness for Fine-tuning in Neural Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2021), pp. 27249–27262.
- [16] D. Li, Z. Zhang, M. Duan, E. Dobriban, and H. R. Zhang. “Efficiently Learning Branching Networks for Multitask Algorithmic Reasoning”. In: *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 2026.
- [17] D. Li, Z. Zhang, L. Wang, and H. R. Zhang. “Efficient ensemble for fine-tuning language models on multiple datasets”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*. 2025, pp. 25347–25364.

- [18] Y. Li, T. Ma, and H. Zhang. “Algorithmic Regularization in Over-parameterized Matrix Sensing and Neural Networks with Quadratic Activations”. In: *Conference On Learning Theory (COLT)*. 2018, pp. 2–47.
- [19] A. Nippani, D. Li, H. Ju, H. Koutsopoulos, and H. R. Zhang. “Graph Neural Networks for Road Safety Modeling: Datasets and Evaluations for Accident Analysis”. In: *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*. 2023.
- [20] H. Tong, B. A. Prakash, T. Eliassi-Rad, M. Faloutsos, and C. Faloutsos. “Gelling, and melting, large graphs by edge manipulation”. In: *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM)*. 2012, pp. 245–254.
- [21] S. Wu, H. R. Zhang, and C. Ré. “Understanding and Improving Information Transfer in Multi-Task Learning”. In: *International Conference on Learning Representations (ICLR)* (2020).
- [22] S. Wu, H. R. Zhang, G. Valiant, and C. Ré. “On the Generalization Effects of Linear Transformations in Data Augmentation”. In: *International Conference on Machine Learning (ICML)*. 2020, pp. 10410–10420.
- [23] F. Yang, H. R. Zhang, S. Wu, C. Ré, and W. J. Su. “Precise High-dimensional Asymptotics for Quantifying Heterogeneous Transfers”. In: *Journal of Machine Learning Research (JMLR)* (2025).
- [24] H. Zhang, P. Lofgren, and A. Goel. “Approximate Personalized PageRank on Dynamic Graphs”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2016, pp. 1315–1324.
- [25] H. R. Zhang, D. Li, and H. Ju. “Noise Stability Optimization for Finding Flat Minima: A Hessian-based Regularization Approach”. In: *Transactions on Machine Learning Research (TMLR)* (2024).
- [26] H. R. Zhang, Z. Zhang, H. Nguyen, and G. Lan. “One-Sided Matrix Completion from Ultra-Sparse Samples”. In: *Transactions on Machine Learning Research (TMLR). Featured Certification* (2026).
- [27] Z. Zhang, M. Duan, Y. Ye, and H. R. Zhang. “Scalable Multi-Objective and Meta Reinforcement Learning via Gradient Estimation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2026, pp. 28609–28617.
- [28] Z. Zhang, M. Duan, and H. R. Zhang. “Efficient Estimation of Kernel Surrogate Models for Task Attribution”. In: *International Conference on Learning Representations (ICLR)*. 2026.
- [29] Z. Zhang, J. W. Liu, C. Re, and H. R. Zhang. “A Hessian View of Grokking in Mathematical Reasoning”. In: *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS’24*. 2024.
- [30] Z. Zhang, M. Duan, H. N. Koutsopoulos, and H. R. Zhang. “Learning Multimodal Embeddings for Traffic Accident Prediction and Causal Estimation”. In: *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 2026.
- [31] Z. Zhang, Z. Zhang, D. Li, L. Wang, J. Dy, and H. R. Zhang. “Linear-Time Demonstration Selection for In-Context Learning via Gradient Estimation”. In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2025, pp. 16470–16488.