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. Within this context, my research contributions have started from matrix recovery [16] to developing Hessian-based measures to understand learning in neural networks [6, 5], and developing both theoretical understanding [22] and new algorithms [10] for multitask learning.

I draw on techniques from optimization, learning theory, probability, and algorithmic game theory in our research. In addition, my students and I have extensively evaluated our research in applied domains, including language models [29, 9, 27] and transportation [17, 28].

Our goal is to advance the understanding of machine learning in new regimes beyond conventional settings, and we draw on large-scale experiments to enable new discoveries. We have open-sourced our empirical work to ensure it is reproducible and accessible to the community. Next, I elaborate on my research in three directions. I conclude with a discussion of several open questions.

Matrix Recovery and Hessian-based Measurements of Neural Networks

Modern neural networks are typically designed with far more parameters than training examples, a regime in which models learn by memorizing the training data. 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. My contribution involves developing an algorithmic framework to estimate Hessian spectral statistics and using this framework to measure generalization in neural networks. First, I describe my work on gradient-descent dynamics in overparameterized matrix sensing with a small initialization [16], and apply the techniques to ultra-sparse matrix completion and data augmentation. Second, I describe how the Hessian perspective arises, leading to non-vacuous bounds for supervised fine-tuning [6], implications for designing optimization algorithms, and new generalization bounds for graph neural networks [5].

Matrix recovery. In an earlier work at COLT'18 [16], 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 and ultimately converge to the minimum nuclear-norm matrix 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 [24]. Recent work [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. 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.

As an application of these techniques, we consider data augmentation, which is commonly used in learning from limited data [21]. We measure the bias and variance of both invariant and mixup transformations, and formalize our findings in the over-parameterized regression setting.

The connection from (overparameterized) matrix recovery to the Hessian perspective stems from the fact that when the training loss is nearly zero, the Hessian trace of the loss is approximately equal to the nuclear norm of $MM^\top$ [23], and it is well-known that the minimum nuclear-norm solution that achieves zero training loss recovers the ground-truth matrix [19] (provided a certain restricted isometry property holds). This connection motivates us to examine the *spectrum of the loss Hessian matrix* in settings beyond matrix recovery.

Hessian-based measurements for fine-tuning. 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 [12, 6, 23], 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 Hessian spectrum, such as the maximum trace of the loss Hessian over the entire data distribution, times the radius squared. In an ICML'22 paper [6], we derive sharper generalization bounds by setting anisotropic priors in the PAC-Bayes bound. The prior leverages layer-wise Hessian structures while ignoring cross-layer interactions. A particularly interesting property of these 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 [23]. (2) The noise-injection can also accelerate quantization-aware training around saddle points [9]. (3) In preliminary work, we find that the Hessian trace regularization also mitigates grokking in modular arithmetic tasks.

Improved generalization bounds on 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. In an AISTATS'23 paper [5], we use this technique to improve the state-of-the-art generalization bounds for graph neural networks. 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.

Multitask Learning and Foundation Models: Theoretical Understanding and New Algorithms

The problem of multitask learning is as follows: Given k 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? How can we identify the most helpful tasks to train alongside another task? My contributions in this area include developing a theoretical understanding of information transfer [22] and designing new clustering algorithms for multitask learning [10].

Our more recent work points to several new directions, including multi-objective reinforcement learning [26], and a new analysis of linear surrogate modeling [27].

Theoretical understanding of information transfer in multitask learning. We formally study transfer by relating multi-headed neural networks—a common architecture for conducting multitask learning—to two-layer neural networks [16]. Our paper is among 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 with Chris Ré and Weijie Su [22], we improve on the initial result and provide a precise quantification of transfer in high-dimensional linear regression. We formulate hard-parameter sharing for two linear regression tasks in the high-dimensional, proportional regime. Using random matrix theory, we show a phase transition from positive to negative transfer as the number of source-task samples increases in the case of two (parametric) linear-regression tasks with different ground-truth model parameters. A key technical result of this paper is the derivation of the high-dimensional limit for various combinations of two sample covariance matrices with different population covariances. This regime is also known as *covariate shift* in the transfer learning literature.

Clustering algorithms for multitask learning. A key insight from the above theoretical analysis is that negative transfer becomes inherent in a shared neural network when tasks have severe distribution shifts. This issue becomes especially acute as the number of tasks k increases, leading to 2^k subset combinations. To address this issue, we draw inspiration from the data attribution literature [3] and develop a surrogate modeling approach to predict the performance of subset task training outcomes. In a TMLR'23 paper with Huy Nguyen [10], we estimate a simple statistical model from any subset $S \subseteq \{1, 2, \ldots, k\}$ to their trained outcomes, and show a linear in k sample complexity bound on the estimation procedure. Surprisingly, this stylized approach works well in a variety of multitask learning applications, including weak supervision [18].

This procedure provides a task affinity matrix computed based on the estimated subset outcomes. Then, we design convex relaxation algorithms to find approximate task partitions that maximize intra-cluster affinity scores within each partition [8, 11, 14, 15, 29]. The optimization program is based on an affinity matrix that captures task relationships and is estimated via linear surrogate models. This yields a random-forest-style algorithm that captures higher-order correlations, and can be efficiently implemented on language models via a linearization technique reminiscent of neural tangent kernels. With Aneesh Sharma, we validate this approach to overlapping community detection [8, 11] by treating the classical problem as a multi-label node classification task. In another line of collaboration with Lu Wang, we further extend this approach to language models, including low-rank adaptation [14, 15], and in-context learning [29].

Multi-objective reinforcement learning. A related problem that is amenable to the above techniques is in finding optimal policies in MDPs that involve multiple competing objectives (such as honesty vs. helpfulness). This problem has broad applications in modern AI, such as in alignment and in robotics. In a recent paper with my student Zhenshuo Zhang [26], 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.

Understanding linear surrogate models via a Hessian analysis. Modern AI models are trained on diverse tasks, leading to the fundamental 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 measuring the local geometry, or sharpness, of loss landscapes, which can be captured by the Hessian matrix of the loss function. In a very recent paper with Zhenshuo [27], we seek to rigorously understand linear surrogate models through this Hessian analysis. We rigorously show that, assuming that the second-order interactive effects are negligible, the coefficients of linear surrogate models [10] are approximately equal to influence functions.

Learning and Algorithmic Reasoning on Large-Scale Networks

My work in this area spans learning and algorithmic reasoning on large-scale networks such as road networks and mobility networks. This direction provides a domain for applying the preceding ideas and rich connections to classical algorithms and tools in spectral graph theory. My earlier contributions include the *first running-time analysis* of local push (a deterministic algorithm for computing personalized PageRank) on dynamic graphs [25], in which we propose and analyze natural dynamic versions of known local variations of power methods for solving linear systems [1]. Another contribution is an optimization algorithm for reducing epidemic diffusion on weighted graphs by minimizing the sum of the k largest eigenvalues [7], generalizing earlier work in this literature that reduces the top eigenvalue [20]. Below, I highlight two ongoing directions.

Traffic accident prediction. Road accidents present a persistent threat to our society and result in massive economic loss annually around the world. With my student Michael Zhang and also with Haris Koutsopolous [17, 28], we collect traffic accident records from the Department of Transportation websites and construct 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). In addition, we have collected high-resolution satellite imagery of the road segments. Our analysis shows that combining graph neural networks with satellite image embeddings can predict accident occurrences with an AUROC over 90%.

Algorithmic reasoning. Can neural networks learn to follow the execution of an algorithm? We study this question using 12 classical algorithms, including breadth-first search (BFS), depth-first search (DFS), and the Floyd-Warshall algorithm [13] (with Edgar Dobriban). We design graph neural networks to predict the intermediate executions and the final outputs of these algorithms *simultaneously* on Erdős–Rényi random graph distributions. This problem maps to a multi-label node classification setup, where each intermediate step is encoded as a node-label vector for each algorithm on a given input graph.

We develop a hierarchical architecture that learns all the algorithms simultaneously [13]. We find that simple algorithms such as BFS can be learned with near-perfect accuracy, whereas for more complex algorithms such as Floyd-Warshall, the accuracy drops to 63%. The prediction accuracy for DFS is less than 40%, suggesting that its recursive search process is difficult to learn. This work raises several questions about the learnability of an algorithm using neural networks and presents a formal testbed to better understand reasoning.

Discussions

The research above suggests that measuring the Hessian spectrum provides insights into the learnability of large models. There are several fundamental open questions in this direction. What is the computational complexity for accurately estimating the Hessian trace and Hessian spectral density, in terms of the number of Hessian vector product computations? Our latest findings suggest that SGD converges to a stationary point with a significant portion of both positive and negative eigenvalues in language models with quantized training [9]. How do we reconcile these findings with the literature that noise-perturbed SGD converges to a second-order stationary point [4]? Another direction is to apply our research on supervised fine-tuning to study catastrophic forgetting in safety alignment, where models trained with high-safety procedures and later fine-tuned on another task tend to forget safety standards and perform poorly again on jailbreaking tests.

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 (WAW)*. 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] A. Ilyas, S. M. Park, L. Engstrom, G. Leclerc, and A. Madry. "Datamodels: Predicting Predictions from Training Data". In: *International Conference on Machine Learning (ICML)*. 2022.

[4] 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.

[5] 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.

[6] 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.

[7] D. Li, T. Eliassi-Rad, and H. R. Zhang. "Optimal Intervention on Weighted Networks via Edge Centrality". In: *SIAM International Conference on Data Mining (SDM)*. 2023, pp. 424–432.

[8] 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.

[9] D. Li, Z. Liu, K. Yi, Z. Zhang, C. Zhao, R. Krishnamoorthi, H. Khaitan, H. R. Zhang, and S. Li. "WinQ: Accelerating Quantization-Aware Training of Language Models around Saddle Points". In: *International Conference on Machine Learning (ICML)*. 2026.

[10] 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).

[11] 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.

[12] 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.

[13] 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.

[14] D. Li, Z. Zhang, L. Wang, and H. R. Zhang. "Scalable fine-tuning from multiple data sources: A first-order approximation approach". In: *Findings of the Association for Computational Linguistics (EMNLP)*. 2024, pp. 5608–5623.

[15] 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.

[16] Y. Li, T. Ma, and H. R. Zhang. "Algorithmic Regularization in Over-parameterized Matrix Sensing and Neural Networks with Quadratic Activations". In: *Conference On Learning Theory (COLT)*. 2018, pp. 2–47.

[17] 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.

[18] A. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré. "Training complex models with multi-task weak supervision". In: *AAAI conference on artificial intelligence (AAAI)*. 2019, pp. 4763–4771.

[19] B. Recht, M. Fazel, and P. A. Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM review* 52.3 (2010), pp. 471–501.

[20] H. Tong, B. A. Prakash, T. Eliassi-Rad, M. Faloutsos, and C. Faloutsos. "Gelling, and melting, large graphs by edge manipulation". In: *ACM International Conference on Information and Knowledge Management (CIKM)*. 2012, pp. 245–254.

[21] 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.

[22] 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).

[23] 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).

[24] 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).

[25] H. R. 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.