Research Statement

Tianyang Hu

My research interests lie at the intersection of **Statistics** and **Artificial Intelligence (AI)**, with the goal of advancing the theoretical understanding of AI and developing theory-inspired algorithms. I am particularly interested in Statistical Machine Learning, Trustworthy AI, Representation Learning, Deep Generative Models, and Large Language Models (LLMs).

AI is taking the world by storm. From vanilla neural networks to transformers, from supervised learning to generative modeling, numerous empirical studies have demonstrated remarkable capabilities in handling increasingly complex tasks. However, much of this progress is driven by empirical, trial-and-error methods, especially in the LLM era. This reliance on heuristics has led to two significant challenges: (1) the resulting models often function as black boxes, lacking robustness and trustworthiness; (2) as models grow in size and complexity, empirical validation becomes increasingly difficult and resource-intensive. These challenges underscore the need for a *theory-driven* approach to AI development. With seamless cooperation between theory and practice, and collaboration across academia and industry, AI can advance in a more sustainable way. To this end, my research seeks to bridge the gap between theory and practice by formulating new theoretical problems from practical challenges and rigorously investigating them. This results in a deeper theoretical understanding of AI and the development of practically impactful, theory-inspired algorithms. My research follows three interconnected directions, each contributing progressively to AI advancement.

Statistical Foundations of Deep Learning

The starting point of my research is contributing to the statistical foundation of deep learning, with the goal of providing statistical guarantees for the empirical performance of deep learning, e.g., whether they can achieve the optimal convergence rate and circumvent the curse of dimensionality. I began investigating in this direction right after the seminal works of Schmidt-Hieber [1], which first showed that ReLU neural networks can achieve optimal convergence rates in the classical smooth regression setting. I extended this line of work to classification and derived the first minimax-optimal convergence rates for deep neural network (DNN) classifiers under traditional smooth boundary settings [2] and a novel teacher-student setting [3]. In both cases, DNN classifiers are shown to be statistically optimal while adapting to underlying low-dimensional structures to circumvent the curse of dimensionality. Key technical contributions include deriving complexity bounds for DNN classifiers and adapting existing empirical process bounds to neural networks, whose covering number is always in the logarithmic order but with network-size-dependent constants. Following [3], I further developed the concept of boundary complexity to better characterize the classifier robustness [4].

Inspired by the seminal work of Neural Tangent Kernel [5], I later extended previous statistical analysis to incorporate the optimization process of neural networks, deriving the first convergence results for overparameterized ReLU networks trained with gradient descent and weight decay, both in regression and classification settings [6, 7]. Motivated by data augmentation in contrastive learning, collaborators and I introduced random smoothing in kernel gradient descent and derived optimal convergence rates that adapt to various low-dimensional data assumptions [8]. These contributions form a solid statistical foundation for understanding neural networks, serving as the basis for subsequent investigations.

Representation Learning

Reassured that basic AI models can be statistically optimal, I took a deeper look into the representation learning process, aiming to uncover what features the models learned, how they depend on the training data, and how to utilize them for downstream tasks. As a statistician, the first question is how to characterize the optimal feature in various learning tasks. With the blooming of self-supervised representation learning, I derived a unified view of representation learning from the perspective of preserving the distance between distributions across different dimensions [9, 10]. Such a measure is non-trivial due to the dimension disparity. Existing work has primarily two remedies: considering pairwise relationships such as Gromov-Wasserstein distance or aligning dimensions by projection or embedding. Interestingly, they correspond to contrastive learning [9] and latent space generative modeling [10] respectively. With this enhanced understanding, I developed new methods to utilize learned features more effectively in downstream tasks. This direction also extends to the concept of the “model zoo”, involving the transfer of knowledge from pretrained models to various downstream applications. Collaborators and I demonstrated the efficacy of tweaking features and utilizing pretrained models in out-of-distribution classification [11, 12, 13, 14, 15] as well as vision generation [16, 17, 18, 19, 20]. This is a prime example of the theory-practice cycle in action, where theoretical insights guide the design of more effective learning paradigms.

Generative Modeling

Generative modeling is a natural and important extension of previous supervised and self-supervised learning scenarios. After gaining insights into representation learning, I expanded my research to generative models, focusing on three key aspects of artificial intelligence generated content (AIGC).

Modeling efficiency Data such as images often reside in ultra-high dimensions, making direct modeling computationally inefficient. Numerous successful approaches leverage a low-dimensional latent space induced by an encoder and generate images through a paired decoder. While the selection of the latent space is empirically pivotal, determining the optimal choice remains unclear. My formulation of representation learning as preserving the distance between distributions naturally provides a characterization for the *optimal latent space*, in the sense that it will minimize the required model complexity [10]. This line of work is later extended in [21].

Sampling efficiency Diffusion models are dominant in AIGC, but sampling typically requires dozens of score function evaluations (NFEs), which is time-consuming. By examining the score network architecture and the sampling time schedule, collaborators and I proposed two *training-free* methods that significantly improve sampling efficiency [22, 23], reducing the cost to under 10 NFEs. To achieve extreme sampling acceleration, we further investigated *training-based* methods that distill a diffusion model to a single-step GAN-style generator. Viewing diffusion distillation as a sampling from an un-normalized density problem, collaborators and I extended my previous Stein Neural Sampler [24] and proposed Diff-Instruct [16], a universal approach for transferring knowledge from pre-trained diffusion models, which is theoretically grounded in minimizing a novel distance termed Integral KL divergence. The one-step sampling performance was further strengthened by incorporating adversarial training [25].

Controllability Ensuring adherence to given conditions is another significant challenge in AIGC. As the conditions get stronger, the paradigm gradually shifts from sampling to regression. Collaborators and I elucidated the design space of conditional generation across a wide range of scenarios. Based on my prior work on understanding trained classifiers [7], we demonstrated the key role of classifier calibration on noisy images and proposed several theory-inspired modifications that significantly outperformed existing guidance methods [17]. Such a training-free guidance method with off-the-shelf pretrained classifiers was also applied to 3D generation [20]. For text-to-image generation, we took a step further, showing that conditional generation with strong conditions can be reformulated as a model inversion problem. By inverting discriminative models (e.g., classifiers), we can achieve better text-image alignment while maintaining competitive sample quality [18]. This direction represents the culmination of my efforts to extend theoretical findings to more complex and practical AI scenarios.

Future Work

In my future work, I plan to continue advancing the theoretical understanding of AI and developing theory-inspired new algorithms, with a special focus on LLMs. Despite the unprecedented progress of LLMs and AGI, this emerging area is where the mismatch between theory and practice is the most severe and could benefit the most from rigorous formulations and theoretical understanding. My goal is to contribute to a theory-inspired next-generation LLM that is *statistically optimal* with *explainable learning mechanisms*.

Statistical estimation of algorithms The new era of AGI poses new challenges for the statistical understanding of LLMs. Beyond traditional machine learning, where the object to learn is some simple function (e.g., decision boundary in classification) and the data are usually independent and identically distributed, LLMs handle sequential data and can learn to implement algorithms and handle logical tasks. Our earlier work showed that transformers with enough layers can implement various non-trivial algorithms [26]. However, a statistical foundation is yet to be built. To this end, I formulate a toy case reminiscent of deductive reasoning based on Directed Acyclic Graphs (DAGs), where each edge represents a deductive logic [27]. The estimation target is inherently an algorithm, and the task difficulty can be conveniently characterized by the DAG size. In this well-defined setting, I aim to (1) establish the optimal statistical learning efficiency (lower bound), particularly how the sample complexity depends on DAG size – an exponential dependence would indicate fundamental limitations in the long-chain reasoning ability of current LLMs; (2) analyze the model learning efficiency (upper bound) to determine whether next-token prediction can be statistically optimal and ways to improve sample efficiency. The aforementioned investigations from the angle of statistical analysis provide not only theoretical reassurance beyond empirical performance, but also practical guidelines, such as the optimal data format (augmentations), and test-time scaling laws, among others.

Duality between prompt and LLM weights Both in-context learning (ICL) and fine-tuning can significantly alter model behavior, but their relationship remains underexplored. My recent work [28] has shown that prompts can equivalently be viewed as query-dependent logit biases in each attention layer, which provides significant insights into understanding how prompts guide LLMs in next-token prediction. A key component in the proposed prompt-to-weight conversion method is attention kernel *semi-linearization*, a novel approximation setting where half the input, i.e., keys, are given and fixed. This formulation opens new research directions: (1) ICL-guided fine-tuning: a more scalable, permanent solution than ICL, while being more robust and lightweight compared to traditional fine-tuning, improving "plasticity" of LLMs. [29]. (2) New LLM architectures: incorporating recurrent memory into the attention mechanism through logit biases that specifically handle long-term and common knowledge, promoting pre-training efficiency and explainability. These directions help us better understand the duality between data and model weights and contribute to a more transparent LLM learning process.

Broader Impact Impactful advancements in AI require a synergistic partnership between engineering and scientific disciplines, e.g., mathematics, statistics, and computer science. My unique combination of academic and industrial experiences positions me well to conduct a theory-driven approach to practical improvement. I am eager to collaborate with faculty across disciplines and industry partners, exploring new research directions and formulating new problems to study. My research aims to advance theoretical foundations while addressing real-world challenges, ensuring AI systems are robust, transparent, and trustworthy. These contributions have the potential to make AI more reliable and effective across sectors like healthcare, finance, and autonomous systems, fostering a collaborative effort between academia and industry for advancing AI.

References

[1] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 2020.

[2] **Tianyang Hu**, Ruiqi Liu, Zuofeng Shang, and Guang Cheng. Minimax optimal deep neural network classifiers under smooth decision boundary. *Journal of Machine Learning Research (JMLR) (under minor revision)*, 2022.

[3] **Tianyang Hu**, Zuofeng Shang, and Guang Cheng. Sharp rate of convergence for deep neural network classifiers under the teacher-student setting. *arXiv preprint*, 2020.

[4] Paweł Piwek, Adam Klukowski, and **Tianyang Hu**. Exact count of boundary pieces of relu classifiers: Towards the proper complexity measure for classification. *Uncertainty in Artificial Intelligence (UAI)*, 2023.

[5] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[6] **Tianyang Hu***, Wenjia Wang*, Cong Lin, and Guang Cheng. Regularization matters: A nonparametric perspective on overparametrized neural network. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[7] **Tianyang Hu***, Jun Wang*, Wenjia Wang*, and Zhenguo Li. Understanding square loss in training overparametrized neural network classifiers. *Advances in Neural Information Processing Systems (NeurIPS) (Spotlight)*, 2022.

[8] Liang Ding*, **Tianyang Hu***, Jiahang Jiang*, Donghao Li*, Wenjia Wang*, and Yuan Yao*. Random smoothing regularization in kernel gradient descent learning. *Journal of Machine Learning Research (JMLR)*, 2024.

[9] **Tianyang Hu**, Liu Zhili, Fengwei Zhou, Wenjia Wang, and Weiran Huang. Your contrastive learning is secretly doing stochastic neighbor embedding. *International Conference on Learning Representations (ICLR)*, 2022.

[10] **Tianyang Hu**, Fei Chen, Haonan Wang, Jiawei Li, Wenjia Wang, Jiacheng Sun, and Zhenguo Li. Complexity matters: Rethinking the latent space for generative modeling. *Advances in Neural Information Processing Systems (NeurIPS) (Spotlight)*, 2023.

[11] Qishi Dong*, Awais Muhammad*, Fengwei Zhou*, Chuanlong Xie, **Tianyang Hu**, Yongxin Yang, Sung-Ho Bae, and Zhenguo Li. Zood: Exploiting model zoo for out-of-distribution generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[12] Yimeng Chen, **Tianyang Hu**, Fengwei Zhou, Zhenguo Li, and Zhi-Ming Ma. Explore and exploit the diverse knowledge in model zoo for domain generalization. *International Conference on Machine Learning (ICML)*, 2023.

[13] Longhui Yu, **Tianyang Hu**, Hong Lanqing, Zhen Liu, Adrian Weller, and Weiyang Liu. Continual learning by modeling intra-class variation. *Transactions on Machine Learning Research (TMLR)*, 2023.

[14] Xuantong Liu, Jianfeng Zhang, **Tianyang Hu**, He Cao, Yuan Yao, and Lujia Pan. Inducing neural collapse in deep long-tailed learning. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[15] Jiajun Ma, **Tianyang Hu**, and Wenjia Wang. Deciphering the projection head: Representation evaluation self-supervised learning. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[16] Weijian Luo, **Tianyang Hu**, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[17] Jiajun Ma, **Tianyang Hu**, Wenjia Wang, and Jiacheng Sun. Elucidating the design space of

classifier-guided diffusion generation. *International Conference on Learning Representations (ICLR)*, 2024.

[18] Xuantong Liu, **Tianyang Hu**, Wenjia Wang, Kenji Kawaguchi, and Yuan Yao. Referee can play: An alternative approach to conditional generation via model inversion. *International Conference on Machine Learning (ICML)*, 2024.

[19] Hao Yang, Lanqing Hong, Aoxue Li, **Tianyang Hu**, Zhenguo Li, Gim Hee Lee, and Liwei Wang. Contranerf: Generalizable neural radiance fields for synthetic-to-real novel view synthesis via contrastive learning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[20] Chenhan Jiang*, Yihan Zeng*, **Tianyang Hu**, Songcun Xu, Wei Zhang, Hang Xu, and Dit-Yan Yeung. Jointdreamer: Ensuring geometry consistency and text congruence in text-to-3d generation via joint score distillation. *European Conference on Computer Vision (ECCV)*, 2024.

[21] Xuantong Liu, Shaozhe Hao, Xianbiao Qi, **Tianyang Hu**, Jun Wang, Rong Xiao, and Yuan Yao. Elucidating the design space of language models for image generation. *Under review at ICLR 2025.*

[22] Jiajun Ma*, Shuchen Xue*, **Tianyang Hu**, Wenjia Wang, Zhaoqiang Liu, Zhenguo Li, Zhi-Ming Ma, and Kenji Kawaguchi. The surprising effectiveness of skip-tuning in diffusion sampling. *International Conference on Machine Learning (ICML)*, 2024.

[23] Shuchen Xue, Zhaoqiang Liu, Fei Chen, Shifeng Zhang, **Tianyang Hu**, Enze Xie, and Zhenguo Li. Accelerating diffusion sampling with optimized time steps. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[24] **Tianyang Hu**, Zixiang Chen, Hanxi Sun, Jincheng Bai, Mao Ye, and Guang Cheng. Stein neural sampler. *arXiv preprint*, 2018.

[25] Yihong Luo, Xiaolong Chen, **Tianyang Hu**, and Jing Tang. You only sample once: Taming one-step text-to-image synthesis by self-cooperative diffusion gans. *Under review at ICLR 2025.*

[26] Yihang Gao*, Chuanyang Zheng*, Enze Xie, Han Shi, **Tianyang Hu**, Yu Li, Michael K Ng, Zhenguo Li, and Zhaoqiang Liu. On the expressive power of a variant of the looped transformer. *arXiv preprint*, 2024.

[27] Zhiwei Wang, Yunji Wang, Zhongwang Zhang, Zhangchen Zhou, Hui Jin, **Tianyang Hu**, Jiacheng Sun, Zhenguo Li, Yaoyu Zhang, and Zhi-Qin John Xu. Towards understanding how transformer perform multi-step reasoning with matching operation. *arXiv preprint*, 2024.

[28] Brian K Chen, **Tianyang Hu**, Hui Jin, Hwee Kuan Lee, and Kenji Kawaguchi. Exact conversion of in-context learning to model weights in linearized-attention transformers. *International Conference on Machine Learning (ICML)*, 2024.

[29] Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 2024.