Kaiwen Hu

5 Yiheyuan Rd., Beijing, 100871, China https://kaotty.github.io kaiwenhu@stu.pku.edu.cn

Education

School of EECS, Peking University Beijing, China
Bachelor of Science in Computer Science *Sep. 2021 - July 2025 (Expected)*

- GPA: **3.73/4.00**
- TOEFL: Reading 30 Listening 28 Speaking 27 Writing 29 Total 114
- Highlighted Courses: Mathematical Analysis I, II (94, 87), Advanced Algebra I, II (87, 91.5), Practice of Programming in C&C++ (91), Introduction to Artificial Intelligence (91), Probability Theory and Statistics(A) (88), Set Theory and Graph Theory (89), Algebraic Structure and Combinatorial Mathematics (96), Mathematical Logic (96), Quantum Computing (86), Operating System (89), Computer Architectures (88)

Research Interests

- Representation Learning: Revisiting the existing deep learning (e.g., self-supervised learning) paradigms and uncover their underlying mechanisms theoretically.
- Reinforcement Learning: Improving the sample efficiency of multiple settings with guarantees of robustness.

Publications and Preprints

* indicates equal contributions

- **Understanding the Role of Equivariance in Self-supervised Learning**
  Yifei Wang*, **Kaiwen Hu***, Sharut Gupta, Ziyu Ye, Yisen Wang, Stefanie Jegelka
  The 38th Annual Conference on Neural Information Processing Systems (**NeurIPS**), 2024, Acceptance Rate 25.8%.
- **Projection Head is Secretly an Information Bottleneck**
  Zhuo Ouyang*, **Kaiwen Hu***, Qi Zhang, Yifei Wang, Yisen Wang
  Under review.

Research Experiences

Equivariant Self-supervised Learning Theory (Accepted at NeurIPS 2024) Oct. 2023-May 2024
Supervised by **Prof. Yisen Wang (School of Artificial Intelligence, Peking University)**

- **Summary:** In this work, we seek to establish a theoretical explanation for the principle of equivariant self-supervised learning (E-SSL) using the information theory. We utilize the explaining-away effect to analyze the mutual information between augmentation and class information, which indicates that E-SSL can indeed encourage the model to extract class-relevant features from data and benefit downstream tasks. We also propose three principles for the design of E-SSL and explain how advanced E-SSL designs echo with our theory.

  Contributions:
- Use a toy model to investigate how to maximize the mutual information between the augmentation information and the class information when given the encoder feature.
- Discover that a balanced mixture of augmentation information and class information makes a good feature and that a larger action space is preferable for learning a better feature.
- Verify that acquiring class information can help E-SSL by explicitly injecting and eliminating class information during pretraining.
- Compare the model's performance on different equivariant prediction tasks and their corresponding downstream accuracies and find out that rotation prediction turns out to be the most effective pretraining task, which implies that global transformations tend to yield suitable pretraining tasks.

- Verify that aggressive base transformations can effectively prune shortcuts like learning from colors in E-SSL, encouraging the model to rely more on extracting class information during pretraining.
- Compare the performance of a regular network and an equivariant one and discover that the latter improves feature equivariance and achieves much higher downstream accuracy across different datasets and base transformations.

Contrastive Learning Theory (Under review with positive scores) June 2024-Oct. 2024

Supervised by **Prof. Yisen Wang (School of Artificial Intelligence, Peking University)**

- **Summary:** In this work, we aim to investigate the role of the projection head in contrastive learning from an information theory perspective. We establish theoretical lower and upper bounds for the downstream performance, suggesting that the projection head should act as an information bottleneck. In addition, we optimize the design of the projection head by means of training and structural regularization.

  Contributions:
- Establish the theoretical lower and upper bounds on downstream performance and propose that the key point is to control the mutual information between the encoder and projector features.
- Justify our theoretical guarantees by computing the correlation between the lower and upper bounds and the downstream task accuracy.
- Add the matrix mutual information between the encoder and projector features as a regularization term to the contrastive loss, which increases downstream task accuracy and aligns well with theory.
- Discover that discretizing the projector feature improves downstream performance across different datasets and propose a brief theoretical explanation for this method.
- Finetune the classifier to verify that the bottleneck of downstream performance is not the capability of the classifier but the mutual information between encoder and projector features.

Low-rank MDPs in the Constrained Setting (Plan to submit to ICML 2025) June 2024-Present

Supervised by **Prof. Lin Yang (Electrical and Computer Engineering Department, UCLA)**

- **Summary:** In this work, we aim to investigate low-rank MDPs in the constrained setting. Our goal is to figure out the sample complexity in the constrained setting for both relaxed feasibility and strict feasibility. Our analysis is currently based on the offline setting and we plan to extend to the online setting and improve our current results in the following two months.

  Contributions:
- Adopt the MLE guarantee and the specially designed bonus term from previous works to bound the distance between the value functions derived from the ground truth and the estimated transition models.
- Use the primal-dual approach to obtain an approximately optimal reward value while satisfying the constraint.
- Solve the sample complexity of the relaxed feasibility, and prove that it is the same with that of the unconstrained setting.
- Checking the sample complexity of the strict feasibility obtained by considering a linear combination of the dual regrets of the ground truth optimal policies for the primal and constraint reward values.

Skills

- Programming Languages: C, C++, Python (Pytorch).
- Tools: VS code, Visual Studio, Github, Latex, Notions, Markdown.

Honors and Awards

Third Prize in the Peking University Programming Contest	2023
Peking University Excellent Study Award	2022
First Prize in the National High School Mathematics Competition (Shanghai Provincial)	2020
First Prize in the National High School Mathematics Competition (Shanghai Provincial)	2019