Hyunin Lee

Email : hyunin@berkeley.edu | linkedin | github | homepage

Education

University of California, Berkeley — CA, United States
Ph.D. in Mechanical Engineering — *Aug. 2022 –*

Seoul National University — Seoul, Rep.of.Korea
B.S in Mechanical Engineering; summa cum laude — *Mar. 2015 – Feb. 2022*

Research Area

Reinforcement Learning, **Multi-Modal Learning**, **AI Safety**

Publications / C: Conference, J: Journal, P: Preprint

[**P2**] Cross-attention Secretly Performs Orthogonal Alignment in Recommendation models.
H. Lee, Y. Zhang, H. Nguyen, X. Liu, N. Park, C. Jung, Y. Wang, Z. Wang, S. Sojoudi, X. Feng. *preprint.* [**pdf/blog**]

[**P1**] A Prospect Theoretic Rationality: Loss Aversion Can Accelerate Reinforcement Learning
H. Lee, C. Park, S. Sojoudi, N. Mehr. *preprint.* [**pdf**]

[**C5**] Position: AI Safety Must Embrace an Antifragile Perspective.
M. Jin, **H. Lee**. *ICML*. 2025. [**pdf**]

[**C4**] A Black Swan Hypothesis: The Role of Human Irrationality in AI Safety.
H. Lee, C. Park, D. Abel, M. Jin. *ICLR*. 2025 & *ICLR 2025 Advances in Financial AI Workshop* (Oral, Top 1.2%). [**pdf**]

[**C3**] Pausing Policy Learning in Non-stationary Reinforcement Learning.
H. Lee, M. Jin, J. Lavaei, and S. Sojoudi. *ICML*. 2024. (Oral, Top 1.2%) [**pdf/codes/talk**]

[**J3**] Policy-based Primal-Dual Methods for Concave CMDP with Variance Reduction.
D. Ying, M. Guo, **H. Lee**, Y. Ding, J. Lavaei, and Z. Shen. *JAIR*. 2025. [**pdf /codes**]

[**C2**] Tempo Adaptation in Non-stationary Reinforcement Learning.
H. Lee, Y. Ding, J. Lee, M. Jin, J. Lavaei, and S. Sojoudi. *NeurIPS*. 2023 [**pdf/codes/slides**]

[**J2**] Beyond Exact Gradients: Convergence of Stochastic Soft-Max Policy Gradient Methods with Entropy Regularization.
Y. Ding, J. Zhang, **H. Lee**, and J. Lavaei. *IEEE TAC*. 2025 [**pdf**]

[**C1**] Initial State Interventions for Deconfounded Imitation Learning.
S. Pfrommer, Y. Bai, **H. Lee**, and S. Sojoudi. *IEEE CDC*. 2023. [**pdf**]

[**J1**] Explainable Deep Learning Model for EMG Based Finger Angle Estimation Using Attention.
H. Lee, D. Kim, and Y. Park. *IEEE TNSRE*. vol. 30, pp. 1877-1886 2022. [**pdf/codes**]

Work Experience

Meta — May 2025 – Oct 2025
Research Scientist Intern — *Ranking AI Research Team*

- Researched on transformer-based **multi-modal recommendation model**, enhancing ranking performance in sparse data setting. (Intern manager: Yong Zhang, Xue Feng)
- Proposed a novel cross-attention mechanism—orthogonal alignment—achieving 1.5x improvement in **scaling law efficiency**, validated on both Meta ranking model and public baselines. [**p1**]

OpenAI — Mar 2025 – Aug 2025
Research Associate — *Human Data (Research) Team*

- Engaged with the safety and preparedness team to evaluate an AI agent's ability to replicate ML research and to create a comprehensive rubric that defines objective success criteria for accurately reproducing given ML papers. (Hired by Benjamin Kinsella)

University of California, Berkeley Aug 2022 – Present
Graduate Student Researcher *Advisor: Prof. Somayeh Sojoudi*

- Investigating **non-stationary reinforcement learning** and optimization under distributional shift. **[p1, p2]**
- Investigating **Human-AI alignment** focusing on ensuring **safe AI systems** in the presence of unpredictable or non-standard human behaviour. **[p1, p2, p3]**

OUTTA **[Homepage]** **[LinkedIn]** Aug 2021 – Present
Co-Founder *Seoul, South Korea*

- Deliver online AI courses to **800+ students** annually across South Korea.

Knowledge AI Jul 2021 – Jul 2022
Machine Learning Engineer *Boston, MA*

- Implemented a **Bayesian inference algorithm** in Python to quantify students' mastery of mathematics topics.
- Developed a question-recommendation deep-learning system for an online math-learning platform, boosting student performance.

Seoul National University Mar 2021 – Nov 2021
Undergraduate Research Intern *Soft Robotics & Bionics Lab*

- Designed an **attention-based sequential decision-making algorithm** in Python to predict finger-joint angles from forearm EMG signals, improving accuracy by more than 10%.

Seoul National University Sep. 2020 – Jun. 2021
Undergraduate Research Intern *Robot Learning Lab*

- Develop **deep generative Q learning algorithm** to reconstruct a reward kernel using Python **[pdf]** **[video]**

Academic Activitiy

Reviewer ICLR 2024-2025, ICML 2024-2025, NeurIPS 2024-2025, RLC 2025, AISTATS 2025
Program Chair Committee AAAI 2025

Teaching Experience

Graduate Student Instructor, Statistics and Data Science for Engineers Fall, Spring 2025

Graduate Student Instructor, Dynamic Systems and Feedback Fall 2024

Teaching Tutor, Math and Physics for Freshman Spring 2019, Spring 2020

Teaching Assistant, Dynamic Fall 2019

Teaching Assistant, Mechanical Product Design Fall 2020

Grants and Honors

Berkeley Summer Research Fellowship | *Mechanical Engineering Department* Summer 2024

NeurIPS scholar award | *Conference on Neural Information Processing Systems* Dec. 2023

Kwanjeong Abroad Scholarship | *Kwanjeong Educational Foundation* Fall 2022 – Present

Berkeley Fellowship for Graduate Study | *Graduate Division* Fall 2022 – Spring 2023

National Science & Technology Scholarship | *Korea Student Aid Foundation* Spring 2017, Fall 2019

Spring 2020, Fall 2020

Certificate of Appreciation (OUTTA) | *Dean, college of Engineering, Seoul National University* Jun. 2021

Scholarship to Academic Excellence | *Seoul National University* Spring 2015, Fall 2015

Spring 2016, Fall 2016

Graduate courses

Specialization: Non-convex Optimization & Reinforcement Learning

Theoretical statistics I, II, **Probability Theory I, II**
Convex Optimization (convex optimziation, robust optimization)
Mathematical Programming II (Advanced optimization theory, non-convex optimization)
Advanced control system I (canonical state-space representation forms, Lyapunov stability, LQR control)
Experiential advanced control design I, II (model predictive control, kalman filter)
Linear System, Nonlinear System

Technical Skills

Languages: Python (Advanced), MatLab (Advanced), C++
Software library, platform : Pytorch (Advanced), Tensorflow. Gurobi (Advanced), CPLEX (Advanced)