

Mehrdad Moghimi

mehrdad.m7496@gmail.com | mehrdadmoghimi.github.io

linkedin.com/in/mehrdad-moghimi | github.com/MehrdadMoghimi | Google Scholar

SUMMARY

Machine learning researcher and final-year PhD candidate with five years of **reinforcement learning** research, spanning theoretical foundations and scalable algorithms for safe online and offline RL, with first-author publications at **NeurIPS (×2)** and **ICML**. Currently working on **LLM post-training**, with a focus on **knowledge distillation** and supervised fine-tuning for improving reasoning in smaller language models. PhD expected December 2026.

RESEARCH EXPERIENCE

Machine Learning Research Intern, RBC Borealis, Toronto, ON, Canada

May 2026 – Present

- SMC-KD: Knowledge Distillation Using Teacher-Guided Sequential Monte Carlo | Under review at ICLR 2027
 - * Conducting research on **LLM post-training** via **knowledge distillation**, transferring mathematical reasoning from large teachers to smaller students using **supervised fine-tuning (SFT)**, off-policy, and on-policy distillation.
 - * Proposed SMC-KD, which **outperforms on-policy distillation** on **MATH500, GSM8K, and AMC23** across Qwen, Llama, and Gemma teacher–student pairs.
 - * Built an end-to-end post-training pipeline using Best-of-N and SMC-based data generation, fine-tuning with **Hugging Face TRL**, and automated evaluation.

Research Collaboration with Bernardo Ávila Pires (Google DeepMind)

Sep 2025 – Present

- Utility-Constrained Policy Optimization | *NeurIPS 2026*
 - * Developed a practical framework for utility-constrained policy optimization using stock-augmented distributional RL.
 - * Designed a safe Lagrangian actor–critic algorithm with safety budgets adjustable at test time.
 - * Implemented and evaluated the method on **22 Safety Gymnasium benchmarks**, achieving **state-of-the-art performance**.

Research Assistant, York University, Toronto, ON, Canada

Sep 2021 – Present

- Decoupling Time and Risk: Risk-Sensitive Reinforcement Learning with General Discounting | *NeurIPS 2026*
 - * Studied how general discount functions and risk-sensitive objectives jointly influence sequential decision-making.
 - * Demonstrated that existing frameworks for hyperbolic discounting can induce time-inconsistent behavior, leading to suboptimal policies.
 - * Developed a time-consistent formulation, achieving a **40% average improvement** in episodic return over standard hyperbolic baselines across 50 Atari games.
- Beyond CVaR: Leveraging Static Spectral Risk Measures for Enhanced Decision-Making in Distributional Reinforcement Learning | *ICML 2025*
 - * Studied time inconsistency in risk-sensitive reinforcement learning and developed a novel theory to characterize the behavior of risk-sensitive policies.
 - * Designed a risk-sensitive RL algorithm that maintains a consistent objective across decisions, with convergence guarantees.
 - * Demonstrated the effectiveness of time-consistent decision-making through empirical evaluation on Trading and Box2D environments.
- Risk-Sensitive Actor–Critic with Static Spectral Risk Measures for Online and Offline Reinforcement Learning | *Expert Systems with Applications*
 - * Expanded the risk-sensitive framework with spectral risk measures to high-dimensional problems by formulating actor–critic algorithms suitable for both online and offline learning.
 - * Established convergence guarantees for the policy gradient updates in the tabular setting.
 - * Extended a JAX codebase with risk-sensitive objectives and evaluated the algorithms on online and offline healthcare, finance, and MuJoCo benchmarks.

Research Assistant, RiskLab, Tehran, Iran

Sep 2019 – Aug 2021

- Developed “Encoded Value-at-Risk,” a novel machine learning framework utilizing Variational Auto-Encoders (VAEs) to model non-linear dependencies in high-dimensional financial data.

Research Assistant, Sharif Image Processing Lab, Tehran, Iran

Jan 2018 – Aug 2018

- Bachelor’s thesis: designed a computer vision pipeline for dynamic advertisement replacement in broadcast soccer footage, using planar homography estimation to localize field banners under camera perspective.

PUBLICATIONS

- **M. Moghimi**, A. Rafiey, A. Ma, X. Liu, L. Pishdad, L. El Asri, “SMC-KD: Knowledge Distillation Using Teacher-Guided Sequential Monte Carlo”, Under review at ICLR 2027
- **M. Moghimi**, B. Ávila Pires, “Utility-Constrained Policy Optimization”, *NeurIPS 2026* ([Paper](#)) ([Code](#))
- **M. Moghimi**, A. Coache, H. Ku, “Decoupling Time and Risk: Risk-Sensitive Reinforcement Learning with General Discounting”, *NeurIPS 2026* ([Paper](#)) ([Code](#))
- **M. Moghimi**, H. Ku, “Risk-Sensitive Actor–Critic with Static Spectral Risk Measures for Online and Offline Reinforcement Learning”, *Expert Systems with Applications*, 2026 ([Paper](#)) ([Code](#))
- **M. Moghimi**, H. Ku, “Beyond CVaR: Leveraging Static Spectral Risk Measures for Enhanced Decision-Making in Distributional Reinforcement Learning”, *ICML 2025* ([Paper](#)) ([Code](#))
- H. Arian, **M. Moghimi**, E. Tabatabaei, S. Zamani, “Encoded Value-at-Risk: A Machine Learning Approach for Portfolio Risk Measurement”, *Mathematics and Computers in Simulation*, 2022 ([Paper](#))

EDUCATION

Ph.D. in Applied Mathematics, York University, Toronto, Canada Sep 2021 – Dec 2026 (Expected)

- Supervisor: Prof. Hyejin Ku | Research focus: Safe and Risk-Sensitive Reinforcement Learning

MBA in Finance, Sharif University of Technology, Tehran, Iran Sep 2018 – Sep 2021

- Supervisor: Prof. Hamid Arian | Research focus: Machine Learning for Finance

BSc in Computer Science, Sharif University of Technology, Tehran, Iran Sep 2014 – Sep 2018

- GPA: 18.71/20 (**Ranked 1st in the class of 2018**)

TECHNICAL SKILLS

- **ML Frameworks:** PyTorch, JAX, Hugging Face Transformers, TRL, vLLM
- **Tools:** SLURM, Git, Linux
- **Programming:** Python (proficient); R, MATLAB (intermediate)

HONOURS AND AWARDS

- **York University Graduate Scholarship**, Fall 2021
- **Exceptional Talents Scholarship**, direct admission to graduate studies (waived National Entrance Exam) due to top academic performance, Fall 2018