Majid Daliri

370 Jay St, Brooklyn, NY
11th Floor, Desk A1119

Cell: +1-646-750-4667
daliri.majid@nyu.edu
majid-daliri.github.io

Education

New York University, New York, USA 2022 - present
Ph.D. in Computer Science (in progress)
Student in the Theoretical Computer Science group
Advised by Prof. Christopher Musco

New York University, New York, USA 2022 - 2024
M.S. in Computer Science

University of Tehran, Tehran, Iran 2017 - 2022
B.S. in Computer Engineering (Cumulative GPA: **3.97/4.0**)

Publications

(1) **Coupling without Communication and Drafter-Invariant Speculative Decoding** (Arxiv)
Majid Daliri, Christopher Musco, Ananda Theertha Suresh

(2) **Unlocking the Theory Behind Scaling 1-Bit Neural Networks** (CPAL) 2025
Majid Daliri, Zhao Song, Chiwun Yang

(3) **QJL: 1-Bit Quantized JL transform for KV Cache Quantization** (AAAI) 2025
Amir Zandieh, **Majid Daliri**, Insu Han

(4) **Matrix Product Sketching via Coordinated Sampling** (ICLR) 2025
Majid Daliri, Juliana Freire, Danrong Li, Christopher Musco

(5) **Sampling Methods for Inner Product Sketching** (VLDB) 2024
Majid Daliri, Juliana Freire, Christopher Musco, Aécio Santos, Haoxiang Zhang

(6) **Simple Analysis of Priority Sampling** (SOSA) 2024
Majid Daliri, Juliana Freire, Christopher Musco, Aécio Santos, Haoxiang Zhang

(7) **KDEformer: Accelerating Transformers via Kernel Density Estimation** (ICML) 2023
Amir Zandieh, Insu Han*, **Majid Daliri***, Amin Karbasi (* equal contribution)

(8) **Weighted Minwise Hashing Beats Linear Sketching for Inner Product Estimation** (PODS) 2023
Aline Bessa, **Majid Daliri**, Juliana Freire, Cameron Musco, Christopher Musco, Aécio Santos, Haoxiang Zhang

(9) **Efficient Approximations for Cache-conscious Data Placement** (PLDI) 2022
Ali Ahmadi, **Majid Daliri**, Amir Kafshdar Goharshady, Andreas Pavlogiannis

(10) **A 10-Approximation of the $\frac{\pi}{2}$-MST** (STACS) 2022
Ahmad Biniaz, **Majid Daliri**, AmirHossein Moradpour

Research Internship

Machine Learning Intern, Max Planck Institute, Advised by Dr. Amir Zandieh 2021-2022

- Implemented Fast Attention in Transformers, optimizing accuracy and efficiency in sequence modeling tasks.
- Designed an efficient GPU-compatible LSH method, boosting performance in attention approximation.
- Technical Stack: BigGAN, PyTorch, Transformer architectures, and advanced sequence modeling tools.

Research Intern, HKUST, Advised by Prof. Amir Goharshady 2021-2022

- Made a pioneering theoretical contribution to Cache-conscious Data Placement (CDP), addressing a longstanding challenge in optimizing cache hits.
- Implemented and tested various cache management policies and algorithms, outperforming previous heuristics.
- Our method emerged as the most effective solution, setting a new standard for cache optimization.

Research Intern, University of Salzburg, Advised by Prof. Ana Sokolova Summer 2022

- Developed algorithms for Distribution Bisimilarity, focusing on finite bisimulation up to convex hull witness.
- Extended research to probabilistic system verification and collaborated on using Quantum Annealers for program verification.

Awards and Honors

Travel Grant, ACM-SIAM Symposium on Discrete Algorithms 2024
Awarded a travel grant to present a paper.

Research Grant, University of Salzburg Summer 2022
Awarded a €5,000 grant for a research internship focusing on algorithms for distribution bisimilarity, probabilistic systems verification, and quantum annealing projects.

Hong Kong PhD Fellowship Scheme (declined to attend NYU) 2022
Awarded $185,000 Ph.D. Fellowship as one of the top 300 students selected across all fields for academic excellence and research potential.

ACM ICPC - Regional (University of Tehran) 2019
Ranked 6^{th} among more than 100 team all around the Iran.

Iranian National Olympiad in Informatics Finalist (IOI, Iran) 2016
Awarded to 50 students after a year long competition involving 10,000 students.

Service

- **Reviewer for International Conference on Learning Representations (ICLR 2025)**
- **Reviewer for Conference on Neural Information Processing Systems (NeurIPS 2024)**
- **Reviewer for ACL Rolling Review (ACL, EMNLP, NAACL 2024-2025)**
- **Reviewer for International Conference on Machine Learning (ICML 2024-2025)**
- **Reviewer for Royal Society Open**
- **External Reviewer for Canadian Conference on Computational Geometry (CCCG 2023)**

Conference Presentations

- **Simple Analysis of Priority Sampling** Presentation, (SOSA) 2024
- **Accelerating Transformers via Kernel Density Estimation** Poster, (ICML) 2023
- **Weighted MinHash for Inner Product Estimation** Poster, (PODS) 2023
- **Efficient Approximations for Cache-conscious Data Placement** Presentation, (PLDI) 2022

Teaching

- **Section Leader for NYU CSCI-UA 310 Basic Algorithms** Spring 2023
- **Teaching Assistant NYU CS-GY 6763 Algorithmic Machine Learning** Fall 2022
- **Teaching Assistant University of Tehran Design and Analysis of Algorithms** Fall 2020-2021

Work Experience

Site Reliability Engineer at Cafebazaar 2021 - 2022
- Designed, implemented, and maintained both Redis-as-a-Service/PostgreSQL-as-a-Service on a Kubernetes-based cloud.
- Achieved consistent performance benchmarks for both services with 100% uptime and a 99.9% response rate.
- Technical Stack: Kubernetes, Docker, Sentry, S3, Prometheus

Skills

Theoretical Background:
Proficient in Machine Learning Theory, Neural Networks, Linear Algebra, and Probability.

Technical Skills:
Highly skilled in C/C++, CUDA, Go, Python, Bash-Scripting, PHP, JavaScript. Experience with PyTorch, TensorFlow, Django, CSS3, HTML5, and git.