# Kayhan Behdin

Sunnyvale, CA 94085
behdin1675@gmail.com • http://www.mit.edu/∼behdink/

## EDUCATION

**Massachusetts Institute of Technology**, Cambridge, MA

- Ph.D. in Operations Research Sep 2019 – May 2024
  - Supervisor: Prof. Rahul Mazumder
  - GPA: 5.0/5.0
  - Thesis: Statistical Learning with Discrete Structures: Statistical and Computational Perspectives

**Sharif University of Technology**, Tehran, Iran

- B.Sc. in Electrical Engineering (Communications) Sep 2014 – Jul 2019
  - GPA: 19.34/20
  - Supervisor: Prof. Babak Khalaj
  - Thesis: Remote feedback stabilization of LTI systems under fading channel constraint

## RESEARCH EXPERIENCE

**LinkedIn**, Sunnyvale, CA

- Senior Software Engineer - Machine Learning
  - I lead the model pruning and compression efforts at LinkedIn. Specifically, my work focuses on creating Small Language Models (SLMs) that can be deployed online under strict latency and throughput requirements. Through the use of optimization-based model pruning techniques, we have been able to ramp SLMs online for a plethora of tasks related to recommendation systems, such as semantic search and ranking. An open-source implementation of my work at LinkedIn can be found at https://github.com/linkedin/fmchisel. Jul 2024 – Now

**Operations Research Center**, MIT

- Graduate Research Assistant Sep 2019 – May 2024

**LinkedIn**, Sunnyvale, CA

- Artificial Intelligence/Machine Learning Engineering Intern May 2023 –Aug 2023

**LinkedIn**, Sunnyvale, CA

- Artificial Intelligence/Machine Learning Engineering Intern Jun 2022 –Aug 2022

**Institute of Theoretical Computer Science and Communications**, The Chinese University of Hong Kong

- Summer Research Program Jul 2018 – Aug 2018

**Electrical Engineering Department**, Sharif University of Technology

- Undergraduate Research Student 2016 – 2019

## SELECTED PAPERS

### PREPRINTS

[1] **K. Behdin**, R. Benbaki, P. Radchenko and R. Mazumder "Modeling with Categorical Features via Exact Fusion and Sparsity Regularisation", Major Revision, The Journal of the Royal Statistical Society, Series B, 2024.

[2] **K. Behdin**, W. Chen and R. Mazumder "Sparse Gaussian Graphical Models with Discrete Optimization: Computational and Statistical Perspectives", Minor Revision, Operations Research, 2023.

[3] **K. Behdin**$^*$, G. Loewinger$^*$, K. T. Kishida, G. Parmigiani and R. Mazumder "Multi-Task Learning for Sparsity Pattern Heterogeneity: Statistical and Computational Perspectives", Minor Revision, The Journal of the Royal Statistical Society, Series B, 2022. ($^*$: Equal Contribution)

### ACCEPTED

[1] **K. Behdin** et al. "Scaling Down, Serving Fast: Compressing and Deploying Efficient LLMs for Recommendation Systems", EMNLP 2025 Industry Track (to appear).

[2] **K. Behdin** and R. Mazumder "Sparse PCA: A New Scalable Estimator Based On Integer Programming", Annals of Statistics (to appear).

[3] P. Prastakos, **K. Behdin** and R. Mazumder “Differentially Private High-dimensional Variable Selection via Integer Programming”, NeurIPS, 2025 (to appear).

[4] R. Lucas, **K. Behdin**, Z. Wang, Q. Song, S. Tang and R. Mazumder “Reasoning Models Can be Accurately Pruned Via Chain-of-Thought Reconstruction”, NeurIPS 2025 Workshop on Efficient Reasoning (to appear).

[5] M. Makni, **K. Behdin**, G. Afriat, Z. Xu, S. Vassilvitskii, N. Ponomareva, R. Mazumder and H. Hazimeh “SPARTA: An Optimization Framework for Differentially Private Sparse Fine-Tuning”, KDD, 2025.

[6] **K. Behdin** et al. “Efficient Algorithms for Leveraging LLMs for Generative and Predictive Recommender Systems”, WWW, 2025 (Tutorial).

[7] M. Makni, **K. Behdin**, Z. Xu, N. Ponomareva and R. Mazumder “A Unified Framework for Sparse Plus Low-Rank Matrix Decomposition for LLMs”, CPAL, 2025.

[8] X. Meng, **K. Behdin**, H. Wang and R. Mazumder “ALPS: Improved Optimization for Highly Sparse One-Shot Pruning for Large Language Models”, NeurIPS, 2024.

[9] X. Meng, S. Ibrahim, **K. Behdin**, H. Hazimeh, N. Ponomareva and R. Mazumder “OSSCAR: One-Shot Structured Pruning in Vision and Language Models with Combinatorial Optimization”, ICML 2024.

[10] **K. Behdin** and R. Mazumder “Sparse NMF with Archetypal Regularization: Computational and Robustness Properties”, Journal of Machine Learning Research, 2024.

[11] S. Ibrahim, **K. Behdin** and R. Mazumder “End-to-end Feature Selection Approach for Learning Skinny Trees”, AISTATS, 2024.

[12] S. Ibrahim, G. Afriat, **K. Behdin** and R. Mazumder, “GRAND-SLAMIN’ Interpretable Additive Modeling with Structural Constraints ”, NeurIPS, 2023.

[13] **K. Behdin**, Q. Song, A. Gupta, D. Durfee, A. Acharya, S. Keerthi and R. Mazumder “Improved Deep Neural Network Generalization Using m-Sharpness-Aware Minimization ”, NeurIPS OPT Workshop, 2022.

[14] A. Esmaeili, **K. Behdin**, M. A. Fakharian and F. Marvasti “Transductive Multi-label Learning From Missing Data Using Smoothed Rank Function”, Pattern Analysis and Applications, 2020.

[15] M. Azghani, A. Esmaeili, **K. Behdin** and F. Marvasti “Missing Low-Rank and Sparse Decomposition Based on Smoothed Nuclear Norm”, IEEE Transactions on Circuits and Systems for Video Technology, 2019.

## TALKS AND PRESENTATIONS

- Sparse PCA: A New Scalable Estimator Based On Integer Programming
  - IPCO 2021 (poster), MIP Workshop 2021 (poster), JSM 2021 (contributed), INFORMS Annual Meeting 2021 (invited), INFORMS IOS 2022
- Gaussian Graphical Models: A Scalable Framework Based on Combinatorial Optimization
  - MIP Workshop 2022 (poster), INFORMS Annual Meeting 2022 (invited)
- On Statistical Properties of Sharpness-Aware Minimization
  - INFORMS Annual Meeting 2023 (invited)
- Modeling with Categorical Features via Exact Fusion and Sparsity Regularization
  - INFORMS Annual Meeting 2024 (invited)
- Efficient Algorithms for Leveraging LLMs for Generative and Predictive Recommender Systems
  - WWW 2025 (tutorial)

## AWARDS & HONORS

- Outstanding Student Paper Highlight, AISTATS 2024 — May 2024
  End-to-end Feature Selection Approach for Learning Skinny Trees
- Most Popular Poster, MIP Workshop 2021 — May 2021
  Scalable Sparse PCA: Computation-friendly MIP formulations under Statistical Assumptions
- Silver Medal in 8th International Olympiad on Astronomy and Astrophysics (IOAA) — Aug 2014
  Suceava, Romania
- Gold Medal in 9th National Olympiad on Astronomy and Astrophysics — Sep 2013
  Tehran, Iran

## TEACHING EXPERIENCE

**Sloan School of Management**, MIT

- Teaching Assistant
  - Fundamentals of Probability - Fall 2021 - Instructor: Prof. Gamarnik (MSc/PhD level)
- Teaching Assistant
  - Optimization Methods in Business Analytics - Spring 2022 - Instructors: Profs. Orlin and Magnanti (BSc level)
- Teaching Assistant
  - Advanced Analytics Edge - Fall 2022 - Instructor: Prof. Van Parys (MSc level)
- Teaching Assistant
  - Advanced Analytics Edge - Fall 2023 - Instructor: Prof. Van Parys (MSc level)
- Teaching Assistant
  - Statistical Thinking and Data Analysis - Spring 2024 - Instructor: Prof. Mazumder (BSc level)

## EXECUTIVE EXPERIENCE

Conference Session Chair

- INFORMS Annual Meeting (2022, 2023, 2024)
- INFORMS Optimization Society Conference 2022

Reviewer

- Mathematics of Operations Research, Operations Research, Journal of Machine Learning Research, Biometrika, IEEE Transactions on Pattern Analysis and Machine Intelligence, Annals of Statistics
- ICML, KDD, NeurIPS, ICLR, AISTATS