SURBHI GOEL

https://www.surbhigoel.com
[first name][last initial]@cis.upenn.edu

EDUCATION

The University of Texas at Austin August 2015 - June 2020
M.S. and Ph.D. in Computer Science
Advisor: Adam R. Klivans
Committee: Alex Dimakis, Raghu Meka, Eric Price
Dissertation: Towards Provably Efficient Algorithms for Learning Neural Networks
Received the Bert Kay dissertation award

Indian Institute of Technology, Delhi July 2011 - May 2015
B.Tech. in Computer Science and Engineering

APPOINTMENTS

University of Pennsylvania, Philadelphia, PA January 2023 - Present
Magerman Term Assistant Professor, Computer and Information Science

Simons Institute for Theory of Computing, Berkeley, CA August - December 2024
Visiting Scientist, Special Year on Large Language Models and Transformers
Visiting Scientist, Modern Paradigms of Generalization

Microsoft Research, New York, NY July 2020 - December 2022
Postdoctoral Researcher, Machine Learning Group

Institute for Advanced Study, Princeton, NJ January - May 2020
Visiting Graduate Student, Theoretical Machine Learning Program

Simons Institute for Theory of Computing, Berkeley, CA May - August 2019
Research Fellow, Foundations of Deep Learning Program

RESEARCH FUNDING

OpenAI Superalignment Fast Grant ($150,000) 2024 - Present
Microsoft Accelerate Foundation Models Research Award ($25,000) 2023 - 2024

AWARDS AND FELLOWSHIPS

Bert Kay Dissertation Award for best dissertation in CS at UT Austin 2020
Rising Star in ML by University of Maryland and in EECS by UIUC 2019
J.P. Morgan AI PhD Fellowship 2019
Simons-Berkeley Research Fellowship for Foundations of Deep Learning program 2019
The University of Texas at Austin Graduate Continuing Bruton Fellowship 2018
The University of Texas at Austin Graduate School Summer Fellowship 2017
ICIM Stay Ahead Award and Suresh Chandra Memorial Trust Award for Undergraduate Thesis 2015
Aditya Birla Scholarship & OPJEM Scholarship 2011
All India Rank 37 (*Rank 2 among all women applicants*) in IITJEE among 450,000 students 2011
Indian National Mathematics Olympiad Top 30 2010

PUBLICATIONS

(α-β) indicates alphabetical ordering of authors.

PREPRINTS

P5. (α-β) Surbhi Goel, Adam Klivans, Konstantinos Stavropoulos, Arsen Vasilyan. *Testing Noise Assumptions of Learning Algorithms*.

P4. (α-β) Natalie Collina, Surbhi Goel, Varun Gupta, Aaron Roth. *Tractable Agreement Protocols*.

P3. Max Rubin-Toles, Maya Gambhir, Keshav Ramji, Aaron Roth, Surbhi Goel. *Conformal Language Model Reasoning with Coherent Factuality*.

P2. Abhishek Panigrahy, Bingbin Liu, Sadhika Malladi, Andrej Risteski, Surbhi Goel. *Progressive Distillation Induces an Implicit Curriculum*.

P1. Anton Xue, Avishree Khare, Rajeev Alur, Surbhi Goel, Eric Wong. *Logicbreaks: A Framework for Understanding Subversion of Rule-based Inference*.

CONFERENCE PAPERS

C32. Ezra Edelman, Nikolaos Tsilivis, Ben Edelman, Eran Malach, Surbhi Goel. *The Evolution of Statistical Induction Heads*. NeurIPS 2024

C31. (α-β) Surbhi Goel, Abhishek Shetty, Konstantinos Stavropoulos, Arsen Vasilyan. *Tolerant Algorithms for Learning with Arbitrary Covariate Shift*. ***Spotlight***, NeurIPS 2024

C30. GuanWen Qiu, Da Kuang, Surbhi Goel. *Complexity Matters: Feature Learning in the Presence of Spurious Correlations*. ICML 2024

C29. Kan Xu, Hamsa Bastani, Surbhi Goel, Osbert Bastani. *Stochastic Bandits with ReLU Neural Networks*. ICML 2024

C28. (α-β) Surbhi Goel, Steve Hanneke, Shay Moran, Abhishek Shetty. *Adversarial Resilience in Sequential Prediction via Abstention*. NeurIPS 2023

C27. (α-β) Ben Edelman, Surbhi Goel, Sham Kakade, Eran Malach, Cyril Zhang. *Pareto Frontiers in Neural Feature Learning*. ***Spotlight***, NeurIPS 2023

C26. Bingbin Liu, Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, Cyril Zhang. *Exposing Attention Glitches with Flip-Flop Language Modeling*. ***Spotlight***, NeurIPS 2023

C25. (α-β) Sitan Chen, Zehao Dou, Surbhi Goel, Adam Klivans, Raghu Meka. *Learning Narrow One-Hidden-Layer ReLU Networks*. COLT 2023

C24. Bingbin Liu, Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, Cyril Zhang. *Transformers Learn Shortcuts to Automata*. ***Notable top-5%***, ICLR 2023

C23. (α-β) Surbhi Goel, Sham Kakade, Adam Kalai, Cyril Zhang. *Recurrent CNNs Learn Succinct Learning Algorithms*. NeurIPS 2022

C22. (α-β) Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, Cyril Zhang. *Hidden Progress in Deep Learning*. NeurIPS 2022

C21. (α-β) Ben Edelman, Surbhi Goel, Sham Kakade, Cyril Zhang. *Inductive Biases and Variable Creation in Self-Attention*. ICML 2022

C20. Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, Akshay Krishnamurthy. *Understanding Contrastive Learning Requires Incorporating Inductive Biases*. ICML 2022

C19. Jordan Ash, Cyril Zhang, Surbhi Goel, Akshay Krishnamurthy, Sham Kakade. *Anti-Concentrated Confidence Bonuses*. ICLR 2022

C18. (α-β) Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, Dipendra Misra. *Investigating the Role of Negatives in Contrastive Learning*. AISTATS 2022

C17. Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, Sham Kakade. *Gone Fishing: Neural Active Learning*. NeurIPS 2021

C16. (α-β) Naman Agarwal, Surbhi Goel, Cyril Zhang. *Acceleration via Fractal Learning Rate Schedules*. ICML 2021

C15. Vardis Kandiros, Yuval Dagan, Nishanth Dikkala, Surbhi Goel, Constantinos Daskalakis. *Statistical Estimation from Dependent Data*. ICML 2021

C14. (α-β) Surbhi Goel, Adam Klivans, Pasin Manurangsi, Daniel Reichman. *Tight Hardness Results for Learning One-Layer ReLU Networks*. ITCS 2021

C13. (α-β) Surbhi Goel, Adam Klivans, Frederic Koehler. *From Boltzmann Machines to Neural Networks*. NeurIPS 2020

C12. (α-β) Surbhi Goel, Aravind Gollakota, Adam Klivans. *Statistical-Query Lower Bounds*. NeurIPS 2020

C11. (α-β) Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, Adam Klivans. *Superpolynomial Lower Bounds for Learning Neural Networks*. ICML 2020

C10. Omar Montasser, Surbhi Goel, Ilias Diakonikolas, Nathan Srebro. *Learning Adversarially Robust Halfspaces*. ICML 2020

C9. Jessica Hoffmann, Soumya Basu, Surbhi Goel, Constantine Caramanis. *Learning Mixtures of Graphs from Epidemic Cascades*. ICML 2020

C8. (α-β) Ilias Diakonikolas, Surbhi Goel, Sushrut Karmalkar, Adam Klivans, Mahdi Soltanolkotabi. *Approximation Schemes for ReLU Regression*. COLT 2020

C7. Surbhi Goel. *Learning Ising and Potts Models with Latent Variables*. AISTATS 2020

C6. (α-β) Surbhi Goel, Sushrut Karmalkar, Adam Klivans. *Time/Accuracy Trade-offs for Learning ReLU*. ***Spotlight***, NeurIPS 2019

C5. (α-β) Surbhi Goel, Daniel Kane, Adam Klivans. *Learning Ising Models with Independent Failures*. COLT 2019

C4. (α-β) Surbhi Goel, Adam Klivans. *Learning Neural Networks with Two Nonlinear Layers*. COLT 2019

C3. (α-β) Surbhi Goel, Adam Klivans, Raghu Meka. *Learning One Convolutional Layer*. ***Oral***, ICML 2018 (***Oral*** at OPT-ML Workshop, NeurIPS 2016)

C2. (α-β) Surbhi Goel, Adam Klivans. *Eigenvalue Decay Implies Polynomial-Time Learnability*. NeurIPS 2017

C1. (α-β) Surbhi Goel, Varun Kanade, Adam Klivans, Justin Thaler. *Reliably Learning ReLU in Polynomial Time*. COLT 2017

REPORTS

R4. Mahdi Sabbaghi, George Pappas, Hamed Hassani, Surbhi Goel. *Encoding Structural Symmetry for Length Generalization*. 2024.

R3. (α-β) Surbhi Goel, Rina Panigrahy. *Learning Two Layer Networks with High Thresholds*. 2019.

R2. Matthew Jordan, Naren Manoj, Surbhi Goel, Alexandros Dimakis. *Quantifying Perceptual Distortion*. 2019.

R1. (α-β) Simon Du, Surbhi Goel. *Improved Learning of One-hidden-layer CNNs*.

INVITED TALKS

T45-44. *Synthetic Tasks as Sandboxes for Understanding Model Behavior*

ATTRIB Workshop at NeurIPS, New Orleans	December 2024
Optimization Seminar at UPenn	March 2024

T43-37. *Beyond Worst-case Sequential Prediction: Adversarial Robustness via Abstention*

Emerging Paradigms Workshop at Simons Institute	September 2024
EnCORE Workshop at IPAM, UCLA	March 2024
Seminars at UPenn, JHU, Princeton, UC Berkeley, and MPI MIS + UCLA	March - August 2023

T36-35. *How do Large Language Models Think?*

AI for Executives at Penn Engineering	May 2024
Women in Data Science at UPenn	February 2024

T35-31. *Thinking Fast with Transformers - Algorithmic Reasoning via Shortcuts*

Deep Learning Down Under Workshop, Lorne, Australia	January 2024
IFML Workshop on Generative AI at UT Austin	November 2023
Youth in High Dimensions, Trieste, Italy	May 2023
Seminars at NYU and UPenn	April 2023

T30-29. *Sparse Feature Emergence in Deep Learning*

Simons Foundation Symposium on Theoretical Machine Learning, Germany	September 2022
Workshop on Learning at EPFL, Switzerland	July 2022

T29-25. *Demystifying Attention-based Architectures in Deep Learning*

IFML Workshop at Simons Institute, Berkeley	October 2022
WALE Workshop, Greece	June 2022
ML Symposium at USC	December 2021
ELLIS Talk Series at IST Austria	December 2021
Statistics Seminar at Stanford	July 2021

T25-17. *Principled Algorithm Design in the Era of Deep Learning*

Seminars at NYU, UW-Madison, UCSD, UMD, CMU, Duke, UPenn, Cornell, and TTIC	February-April 2022

T16-1. *Computational Complexity of Learning Neural Networks*

IMSI Workshop	April 2021
Seminars at UW-Madison, MIT, TTIC, GaTech, Harvard, Duke, NYU	May 2020 - January 2021
Deep Learning Program Reunion at Simons Institute, Berkeley	August 2020
Microsoft Research (NYC, NE, Redmond)	February 2020
TTIC Young Researcher Series	December 2019
Rising Star in ML at UMD	September 2019
Research Fellows Talk at Simons Institute, Berkeley	July 2019
Theory Reading Group at Google, Mountain View	June 2018

WORK EXPERIENCE

Google, Mountain View CA — May - August 2018
Research Intern — *Supervisor: Rina Panigrahy*

Dell, Round Rock TX — June - August 2017
Research Intern

Google, New York, NY — May - August 2016
Research Intern — *Supervisor: Natalia Ponomareva*

Google, Mountain View CA — May - August 2014
Software Engineering Intern — *Supervisor: Neha Jha*

University of Michigan, Ann Arbor MI — May - July 2013
Research Scholar — *Supervisor: Atul Prakash*

TEACHING

CIS 5200: Machine Learning — Spring 2023, 2024, 2025
Co-instructor with Eric Wong — *University of Pennsylvania*

CIS 7000: Foundations of Modern ML - Theory and Empirics — Fall 2023
Instructor — *University of Pennsylvania*

ADVISING

Ezra Edelman (PhD)	2023 - Present
Max Rubin-Toles (Undergraduate)	2024 - Present
Maya Gambhir (Undergraduate)	2024 - Present
Dante Lokitiyakul (Masters)	2024 - Present
GuanWen Qiu (Masters)	2023 - 2024
Bingbin Liu (Intern at MSR, now Postdoc at Kempner Institute, Harvard)	Summer 2022
Nikunj Saunshi (Intern at MSR, now Research Scientist at Google)	Summer 2021
Ben Edelman (Intern at MSR, now Fellow at US AI Safety Institute)	Summer 2021

PROFESSIONAL SERVICE AND LEADERSHIP

Steering Committee Member, Association for Algorithmic Learning Theory — 2024 - Present

Action Editor, Transactions on Machine Learning Research — 2024 - Present

Co-treasurer, Association for Computational Learning — 2024 - Present

Program Co-organizer, Simons Institute's Special Year on LLMs and Transformers — 2023 - 2024

Workshop Co-organizer, Transformers as a Computational Model (Simons Institute) — 2024

Workshop Co-organizer, Unknown Futures of Generalization (Simons Institute) — 2023 - 2024

Workshop Co-organizer, Mathematics of Modern Machine Learning (M3L) at NeurIPS — 2023

Virtual Experience Co-chair, COLT — 2023

Online Experience Co-chair, COLT — 2021

Co-founder and Organizing Committee Member, Learning Theory Alliance (LeT-All) — 2020 - Present

- Co-organized 10 mentoring events at major conferences (NeurIPS, COLT, ALT)
- Created Graduate Applications Support Program with WiML-T

Senior Program Committee: ALT, COLT, NeurIPS, AISTATS — 2023 - 2024

Program Committee: ALT, COLT — 2021 - 2023