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 2024 - Present
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

AWARDS AND FELLOWSHIPS

2024 OpenAI Superalignment Fast Grant ($150,000)
2023 Microsoft Accelerate Foundation Models Research Award ($25,000)
2020 Bert Kay Dissertation Award for best dissertation in CS at UT Austin
2019 Rising Star in ML by University of Maryland and in EECS by UIUC
2019 The University of Texas at Austin Graduate Dean's Prestigious Fellowship Supplement
2019 J.P. Morgan AI PhD Fellowship
2019 Simons-Berkeley Research Fellowship for Foundations of Deep Learning program
2018 The University of Texas at Austin Graduate Continuing Bruton Fellowship
2017 The University of Texas at Austin Graduate School Summer Fellowship
2015 ICIM Stay Ahead Award and Suresh Chandra Memorial Trust Award for Undergraduate Thesis
2011 Aditya Birla Scholarship & OPJEM Scholarship
2011 All India Rank 37 (*Rank 2 among all women applicants*) in IITJEE among 450,000 students
2010 Indian National Mathematics Olympiad Top 30

PUBLICATIONS

* *indicates α-β (alphabetical) ordering.*

PREPRINTS

Surbhi Goel*, Adam R. Klivans*, Konstantinos Stavropoulos*, Arsen Vasilyan*
Testing Noise Assumptions of Learning Algorithms

Natalie Collina*, **Surbhi Goel***, Varun Gupta*, Aaron Roth*
Tractable Agreement Protocols
Pluralistic Alignment Workshop, NeurIPS 2024

Maxon Rubin-Toles, Maya Gambhir, Keshav Ramji, Aaron Roth, **Surbhi Goel**
Conformal Language Model Reasoning with Coherent Factuality
Statistical Foundations of LLMs and Foundation Models Workshop, NeurIPS 2024

Abhishek Panigrahy, Bingbin Liu, Sadhika Malladi, Andrej Risteski, **Surbhi Goel**
Progressive Distillation Induces an Implicit Curriculum
Mechanistic Interpretability Workshop, ICML 2024
Theoretical Foundations of Foundation Models (T2FM) Workshop, ICML 2024
Mathematics of Modern Machine Learning (M3L) Workshop, NeurIPS 2024

Anton Xue, Avishree Khare, Rajeev Alur, **Surbhi Goel**, Eric Wong
Logicbreaks: A Framework for Understanding Subversion of Rule-based Inference
New Frontiers in Adversarial Machine Learning (AdvML-Frontiers) Workshop, NeurIPS 2024
Safe & Trustworthy Agents (SATA) Workshop, NeurIPS 2024
Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024

CONFERENCE PAPERS

Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, Eran Malach, **Surbhi Goel**
The Evolution of Statistical Induction Heads: In-Context Learning Markov Chains
NeurIPS 2024
Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024

Surbhi Goel*, Abhishek Shetty*, Konstantinos Stavropoulos*, Arsen Vasilyan*
Tolerant Algorithms for Learning with Arbitrary Covariate Shift
Spotlight presentation, NeurIPS 2024

GuanWen Qiu, Da Kuang, **Surbhi Goel**
Complexity Matters: Feature Learning in the Presence of Spurious Correlations
ICML 2024
M*athematics of Modern Machine Learning (M3L) Workshop, NeurIPS 2023*

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

Surbhi Goel*, Steve Hanneke*, Shay Moran*, Abhishek Shetty*
Adversarial Resilience in Sequential Prediction via Abstention
NeurIPS 2023

Benjamin L. Edelman*, **Surbhi Goel***, Sham M. Kakade*, Eran Malach*, Cyril Zhang*
Pareto Frontiers in Neural Feature Learning: Data, Compute, Width, and Luck
Spotlight presentation, NeurIPS 2023

Bingbin Liu, Jordan T. Ash, **Surbhi Goel**, Akshay Krishnamurthy, Cyril Zhang
Exposing Attention Glitches with Flip-Flop Language Modeling
Spotlight presentation, NeurIPS 2023
Challenges of Deploying Generative AI Workshop, ICML 2023
Knowledge and Logical Reasoning in the Era of Data-driven Learning Workshop, ICML 2023

Sitan Chen*, Zehao Dou*, **Surbhi Goel***, Adam R. Klivans*, Raghu Meka*
Learning Narrow One-Hidden-Layer ReLU Networks
COLT 2023

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

Surbhi Goel*, Sham M. Kakade*, Adam T. Kalai*, Cyril Zhang*
Recurrent Convolutional Neural Networks Learn Succinct Learning Algorithms
NeurIPS 2022

Boaz Barak*, Benjamin L. Edelman*, **Surbhi Goel***, Sham M. Kakade*, Eran Malach*, Cyril Zhang*
Hidden Progress in Deep Learning: SGD Learns Parities Near the Computational Limit
NeurIPS 2022

Benjamin L. Edelman*, **Surbhi Goel***, Sham M. Kakade*, Cyril Zhang *
Inductive Biases and Variable Creation in Self-Attention Mechanisms
ICML 2022

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

Jordan T. Ash, Cyril Zhang, **Surbhi Goel**, Akshay Krishnamurthy, Sham M. Kakade
Anti-Concentrated Confidence Bonuses For Scalable Exploration
ICLR 2022

Jordan T. Ash*, **Surbhi Goel***, Akshay Krishnamurthy*, Dipendra Misra*
Investigating the Role of Negatives in Contrastive Representation Learning
AISTATS 2022

Jordan T. Ash, **Surbhi Goel**, Akshay Krishnamurthy, Sham M. Kakade
Gone Fishing: Neural Active Learning with Fisher Embeddings
NeurIPS 2021

Naman Agarwal*, **Surbhi Goel***, Cyril Zhang*
Acceleration via Fractal Learning Rate Schedules
ICML 2021

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

Surbhi Goel*, Adam R. Klivans*, Pasin Manurangsi*, Daniel Reichman*
Tight Hardness Results for Learning One-Layer ReLU Networks
ITCS 2021

Surbhi Goel*, Adam R. Klivans*, Frederic Koehler*
From Boltzmann Machines to Neural Networks and Back Again
NeurIPS 2020

Surbhi Goel*, Aravind Gollakota*, Adam R., Klivans*
Statistical-Query Lower Bounds via Functional Gradients
NeurIPS 2020

Surbhi Goel*, Aravind Gollakota*, Zhihan Jin*, Sushrut Karmalkar*, Adam R. Klivans*
Superpolynomial Lower Bounds for Learning One-Layer Neural Networks using Gradient Descent
ICML 2020

Omar Montasser, **Surbhi Goel**, Ilias Diakonikolas, Nathan Srebro
Efficiently Learning Adversarially Robust Halfspaces with Noise
ICML 2020

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

Ilias Diakonikolas*, **Surbhi Goel***, Sushrut Karmalkar*, Adam R. Klivans*, Mahdi Soltanolkotabi*
Approximation Schemes for ReLU Regression
COLT 2020

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

Surbhi Goel*, Sushrut Karmalkar*, Adam R. Klivans*
Time/Accuracy Trade-offs for Learning a ReLU with respect to Gaussian Marginals
Spotlight presentation, NeurIPS 2019

Surbhi Goel*, Daniel Kane*, Adam R. Klivans*
Learning Ising Models with Independent Failures
COLT 2019

Surbhi Goel*, Adam R. Klivans*
Learning Neural Networks with Two Nonlinear Layers in Polynomial Time
COLT 2019
Deep Learning: Bridging Theory and Practice Workshop, NeurIPS 2017

Surbhi Goel*, Adam R. Klivans*, Raghu Meka*
Learning One Convolutional Layer with Overlapping Patches
Oral presentation, ICML 2018

Surbhi Goel*, Adam R. Klivans*
Eigenvalue Decay Implies Polynomial-Time Learnability for Neural Networks
NeurIPS 2017

Surbhi Goel*, Varun Kanade*, Adam R. Klivans*, Justin Thaler*
Reliably Learning ReLU in Polynomial Time
COLT 2017
Oral presentation, *Optimization for Machine Learning (OPT-ML) Workshop, NeurIPS 2016*

UNPUBLISHED MANUSCRIPTS

Mahdi Sabbaghi, George J. Pappas, Hamed Hassani, **Surbhi Goel**
Encoding Structural Symmetry is Key for Length Generalization in Arithmetic Tasks

Surbhi Goel*, Rina Panigrahy*
Learning Two layer Networks with Multinomial Activation and High Thresholds

Matthew Jordan, Naren Manoj, **Surbhi Goel**, Alexandros Dimakis
Quantifying Perceptual Distortion of Adversarial Examples

Simon Du*, **Surbhi Goel***
Improved Learning of One-hidden-layer Convolutional Neural Networks with Overlaps

INVITED TALKS

Beyond Worst-case Sequential Prediction: Adversarial Robustness via Abstention
Emerging Paradigms of Generalization Workshop at Simons Institute — *September 2024*
Rutgers/DIMACS Theory Seminar — *April 2024*
MINDS Seminar at JHU — *March 2024*
EnCORE Workshop at IPAM, UCLA — *March 2024*
Theory Seminar at UPenn — *November 2023*
Alg-ML Seminar at Princeton — *November 2023*
BLISS Seminar at UC Berkeley — *October 2023*
Math Machine Learning Seminar at MPI MIS + UCLA — *August 2023*
FODSI Workshop on Computational Complexity of Statistical Problems at MIT — *June 2023*

How do Large Language Models Think?
AI for Executives: Daylong Immersion at Penn Engineering — *May 2024*
Women in Data Science at UPenn — *February 2024*

Understanding Training Dynamics in Deep Learning using Simplified Models
Optimization Seminar at UPenn — *March 2024*

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*
MaD Seminar at NYU — *April 2023*

ASSET Seminar at UPenn — *April 2023*

Sparse Feature Emergence in Deep Learning
Symposium on New Directions in Theoretical Machine Learning [slides] — *September 2022*

What Functions do Self-attention Blocks Prefer to Represent?
Demystifying Attention-based Architectures in Deep Learning
Joint IFML/Data-Driven Decision Processes Workshop at Simons Institute — *October 2022*
ML Foundations Seminar at MSR Redmond — *August 2022*
Workshop on Algorithms for Learning and Economics (WALE) in Greece — *June 2022*
ML Symposium at USC — *December 2021*
ELLIS Talk Series at IST Austria — *December 2021*
Learning Theory Workshop at Google — *October 2021*

The Hidden Progress Behind Loss Curves
Workshop on Learning: Optimization and Stochastics at EPFL — *July 2022*

Principled Algorithm Design in the Era of Deep Learning
CS/CSE Colloquium at NYU Courant/Tandon — *April 2022*
CS Colloquium at UW-Madison — *March 2022*
CS Colloquium at Halicioglu Data Science Institute, UCSD — *March 2022*
CS Colloquium at UMD — *February 2022*
SCS Talk at CMU — *February 2022*
CS Colloquium at Duke — *February 2022*
CIS Colloquium at UPenn — *February 2022*
CS Colloquium at Cornell — *February 2022*
Talks at TTIC — *February 2022*

Computational Barriers For Learning Some Generalized Linear Models
Information-Computation Trade-offs Workshop at Simons Institute [video][slides] — *September 2021*

Computational Complexity of ReLU Regression
The Multifaceted Complexity of Machine Learning Workshop at IMSI [video] — *April 2021*

Computational Complexity of Learning Neural Networks over Gaussian Marginals
Statistics Seminar at Stanford University — *July 2021*
SILO Seminar at UW-Madison — *January 2021*
TOC Colloquium at MIT — *December 2020*
IDEAL Seminar at TTIC — *November 2020*
ARC Colloquium at Georgia Tech — *November 2020*
ML Theory Seminar at Harvard University [video] — *October 2020*
Algorithms Seminar at Duke University — *October 2020*
MIC Seminar at NYU — *May 2020*

Approximation Schemes for ReLU Regression
Deep Learning Program Reunion at Simons Institute — *August 2020*

Provably Efficient Algorithms for Learning Neural Networks
Microsoft Research New York — *February 2020*
Microsoft Research New England — *February 2020*
Microsoft Research Redmond — *February 2020*

Exploring Surrogate Losses for Learning Neural Networks
TTIC Young Researcher Seminar Series — *December 2019*

Efficiently Learning Simple Neural Networks
Rising Star in ML Talk at University of Maryland *September 2019*

Learning Ising Models with Independent Failures
Research Fellows Talk at Simons Institute *July 2019*

Efficiently Learning Simple Convolutional Networks
China Theory Week at Tsinghua University *September 2018*

Learning One Convolutional Layer with Overlapping Patches
Google Research Theory Reading Group *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*

OUTREACH

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

Co-organized a social at NeurIPS 2024
Co-organized the Fall 2023 Mentoring Workshop
Co-organized the Fall 2022 Mentoring Workshop in collaboration with FODSI
Co-organized the COLT 2022 Mentoring Panel
Co-organized the ALT 2022 Mentoring Workshop
Co-organized the Graduate Applications Support Program in collaboration with WiML-T
Co-organized the COLT 2021 Mentoring Workshop
Co-organized the ALT 2021 Mentoring Workshop

Mentor
Women in Machine Learning Theory (WiML-T) Mentoring Program *2021-Present*
UT Austin's Women in CS (GWC-WiCS) Mentoring Program *2018-19*

Panelist

New in ML Workshop, NeurIPS 2023 — *Decemeber 2023*
WiML Un-Workshop, ICML 2022 — *July 2022*
New Horizons in Theoretical Computer Science — *June 2022*
VMware Nirman for Women in Tech — *January 2021*

SERVICE ROLES

Steering Committee Member 2024-Present
Association for Algorithmic Learning Theory

Action Editor 2024-Present
Transactions on Machine Learning Research

Co-treasurer 2024-Present
Association for Computational Learning

Workshop Co-organizer 2024
Transformers as a Computational Model, Simons Institute's Special Year on LLMs and Transformers

Workshop Co-organizer 2023-Present
Unknown Futures of Generalization, Simons Institute's Program on Modern Paradigms in Generalization

Program Co-organizer 2023-Present
Simons Institute's Special Year on LLMs and Transformers

Office Hours Co-chair 2023-24
ICLR 2024

Workshop Reviewing Committee 2024
ICML 2024

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

Virtual Experience Co-chair 2023
COLT 2023

Online Experience Co-chair 2021
COLT 2021

Co-organized the virtual part of the hybrid conference, including the 2-day virtual-only program

Seminar Co-organizer 2020-21
One World Machine Learning Seminar Series

Treasurer 2016-17
Graduate Representative Association of Computer Sciences (GRACS) 2024

Program Committee

ALT	*2021/22/23*
COLT	*2021/22*
AISTATS (area chair)	*2023*
NeurIPS (area chair)	*2023/24*
ALT (senior program committee)	*2024*
COLT (senior program committee)	*2024*

Conference Reviewing

STOC	*2019/20/21*
NeurIPS	*2018 (top 30%)/20/21*
COLT	*2018/19/20*
ICLR	*2019/20/23*
SODA	*2020/23*
FOCS	*2020/22*
ICML	*2019 (top 5%)*

Journal Reviewing

TheoretiCS	*2024*
Journal of Machine Learning Research	*2021/22*
IEEE Transactions on Information Theory	*2020*