# Amartya Sanyal

*Assistant Professor of Machine Learning*

amartya18x@gmail.com | amartya18x.github.io | github.com/amartya18x | +4531216790

## PROFESSIONAL EXPERIENCE

**Tenure Track Assistant Professor in Machine Learning** Aug 2024–
Department of Computer Science, University of Copenhagen Copenhagen

**Adjunct Assistant Professor** Nov 2024-
Department of Computer Science, Indian Institute of Technology, Kanpur Kanpur

**Postdoctoral Fellow, Max Planck Institute for Intelligent Systems** 2023 - 2024
With Prof. Bernhard Schölkopf Tübingen

**Postdoctoral Fellow, ETH AI Center** 2021 - 2023
With Prof. Fanny Yang Zürich

**Part Time Researcher, Facebook AI Research** 2020 - 2021
with Dr. Edward Grefenstette London

## EDUCATION

**D.Phil in Computer Science (Advisor: Dr. Varun Kanade and Dr. Philip Torr)** 11 Sep, 2021
University of Oxford, Thesis: Identifying and Exploiting Structures for Reliable Deep Learning

**B.Tech. in Computer Science and Engineering (Minor in Linguistics Theory)** 15 Jun, 2017
Indian Institute of Technology, Kanpur

## AWARDS AND MEMBERSHIPS

| | |
|---|---|
| ’25 | Villum Young Investigator Award, **Copenhagen** |
| ’24 | Rising Star, **Workshop in Applied Algorithms for ML, Paris** |
| ’23 | Rising Star in AI, **KAUST** |
| ’21-’23 | ETH AI Center Postdoctoral Fellowships Award, **ETH AI Center, Zürich, Switzerland** |
| ’17-’21 | Turing Doctoral Studentship Award, **The Alan Turing Institute, London, UK** |
| ’14 & ’16 | Academic Excellence Award, **Awarded to top 10% student, IIT Kanpur** |

## GRANTS AND MANAGEMENT

| | | |
|---|---|---|
| 2025-2027 | P1 Program on Data Privacy for ML 50,000 DKK, **Co-Director** | University of Copenhagen |
| 2025-2029 | Villum Young Investigator Grant 7,000,000 DKK (≈ 985k USD), **PI** | University of Copenhagen |
| 2024-2028 | Novo Nordisk Foundation (NNF) Startup Grant 4,000,000 DKK (≈ 563k USD), **PI** | University of Copenhagen |
| 2025 | NNF Conference Grant 300,000 DKK (≈ 42k USD), **General Chair of SaTML, 2025** | University of Copenhagen |
| 2025 | DDSA Large Event Grant 100,000 DKK (≈ 14k USD), **General Chair of SaTML, 2025** | University of Copenhagen |
| 2022-23 | Hasler Stiftung Grant 50,000 CHF, **Privacy and Fairness in Machine Learning, PI** | ETH Zürich |

## ACADEMIC SERVICE

| | | |
|---|---|---|
| ’25 | General Chair for IEEE Conference on Secure and Trustworthy Machine Learning, | Copenhagen, Denmark |
| ’25 | ENCORE Workshop on Defining Holistic Private Data Science for Practice, | San Diego, USA |
| ’23 | ICLR Workshop on Pitfalls of limited data and computation for Trustworthy ML, | Kigali, Rwanda |

## REVIEWING

| | |
|---|---|
| JMLR, JPC, TPAMI, TMLR, IJCV, | Journal |
| AISTATS, NeurIPS, | Area Chair |
| NeurIPS, ICML, ICLR, COLT, AISTATS, UAI, SODA, ECCV, CVPR, | Reviewing |
| ERC StG, DfG, Schmidt sciences, | Grant Reviewing |

## RECENT TUTORIALS AND TALKS

| | | |
|---|---|---|
| Feb ’25 | Building trustworthy ML: The role of label quality and availability, **AAAI Tutorial** | Philadelphia, USA |
| March ’25 | Machine Unlearning: Its promises and failures, **Keynote Speaker** | Oulu, Finland |
| Feb ’25 | ELSA Workshop on Privacy-Preserving Machine Learning, **Invited Speaker** | Bertinoro, Italy |
| June ’24 | Applied Algorithms for ML Workshop,Rice Global University, **Invited Speaker** | Paris, France |
| Nov ’23 | CISPA, Helmholtz Institute, **Invited Speaker** | Saarbrücken, Germany |
| Oct ’23 | Data Science Seminar, University of Michigan, **Invited Speaker** | Michigan, USA |
| Oct ’23 | META AI, Facebook, **Invited Speaker** | New York, USA |
| Sep ’23 | Department of Computer Science, University of Helsinki, **Invited Speaker** | Helsinki, Finland |

## ADVISED PHD STUDENTS

| | |
|---|---|
| ’25- | Giorgio Racca, **University of Copenhagen** |
| ’25- | Luka Radic, **University of Copenhagen** |
| ’24- | Carolin Heinzler, **University of Copenhagen** |
| ’24- | Johanna Duëngler, **University of Copenhagen** |
| ’23- | Omri Ben Dov , **Max Planck Institute for Intelligent Systems** |
| ’23- | Anmol Goel, **ELLIS PhD, TUDarmstadt** |
| ’23- | Yaxi Hu, **Max-Planck Institute for Intelligent Systems** |

## TEACHING

| | |
|---|---|
| ’23-’25 | Privacy in Machine Learning, **University of Copenhagen** |
| ’24 | Machine Learning B, **University of Copenhagen** |
| ’22 | Guarantees in Machine Learning, **ETH Zürich** |
| ’22 | Projects in Machine Learning Research, **ETH Zürich** |

| | |
|---|---|
| ’21 | Tutor, Computational Learning Theory, **University of Oxford** |
| ’19’, ’20 | Tutor, Theory of Optimization, **University of Oxford** |
| ’18-’20 | Tutor, Machine Learning, **Wadham College, Worcester College, Somerville College, University of Oxford** |
| ’14 | Tutor, Linear Algebra, Real Analysis and ODEs, **IIT Kanpur** |

## CONFERENCE PUBLICATIONS

Alex, Neel, Shoaib Ahmed Siddiqui, Amartya Sanyal, and David Krueger (2025). “Protecting against simultaneous data poisoning attacks”. In: *The Thirteenth International Conference on Learning Representations (ICLR)*.

Goel, Anmol, Yaxi Hu, Iryna Gurevych, and Amartya Sanyal (2025). “Differentially Private Steering for Large Language Model Alignment”. In: *The Thirteenth International Conference on Learning Representations (ICLR)*.

Sanyal, Amartya, Yaxi Hu, Yaodong Yu, Yian Ma, Yixin Wang, and Bernhard Schölkopf (2025). “Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation”. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Wei, Stanley, Sadhika Malladi, Sanjeev Arora, and Amartya Sanyal (2025). “Provable unlearning in topic modeling and downstream tasks”. In: *The Thirteenth International Conference on Learning Representations (ICLR)*.

Ben-Dov, Omri, Jake Fawkes, Samira Samadi, and Amartya Sanyal (2024). “The Role of Learning Algorithms in Collective Action”. In: *International Conference on Machine Learning (ICML)*.

Dmitriev, Daniil, Rares-Darius Buhai, Stefan Tiegel, Alexander Wolters, Gleb Novikov, Amartya Sanyal, David Steurer, and Fanny Yang (2024). “Robust Mixture Learning when Outliers Overwhelm Small Groups”. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Dmitriev, Daniil, Kristóf Szabó, and Amartya Sanyal (2024). “On the Growth of Mistakes in Differentially Private Online Learning: A Lower Bound Perspective”. In: *Conference on Learning Theory (COLT) 2024*
*Theory and Practice of Differential Privacy (TPDP)*.

Goel, Shashwat, Ameya Prabhu, Philip Torr, Ponnurangam Kumaraguru, and Amartya Sanyal (2024). “Corrective Machine Unlearning”. In: *Transactions in Machine Learning Research*.

Hu, Yaxi, Amartya Sanyal, and Bernhard Schölkopf (2024). “Provable Privacy with Non-Private Pre-Processing”. In: *International Conference on Machine Learning (ICML)*
*Theory and Practice of Differential Privacy (TPDP) 2024*.

Jain, Samyak, Ekdeep Singh Lubana, Kemal Oksuz, Tom Joy, Philip Torr, Amartya Sanyal, and Puneet K. Dokania (2024). “What Makes and Breaks Safety Fine-tuning? A Mechanistic Study”. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Donhauser, Konstantin, Johan Lokna, Amartya Sanyal, March Boedihardjo, Robert Hönig, and Fanny Yang (2023). “Sample-efficient private data release for Lipschitz functions under sparsity assumptions”. In: *International Conference on Artificial Intelligence and Statistics (AISTATS) 2024*
*Theory and Practice of Differential Privacy (TPDP)*.

Paleka, Daniel and Amartya Sanyal (2023). “A law of adversarial risk, interpolation, and label noise”. In: *International Conference on Learning Representations (ICLR)*.

Petrov, Aleksander, Francisco Eiras, Amartya Sanyal, Philip H.S. Torr, and Adel Bibi (2023). “Certifying Ensembles: A General Certification Theory with $\mathcal{S}$-Lipschitzness”. In: *International Conference on Machine Learning (ICML)*.

Pinto, Francesco, Yaxi Hu, Fanny Yang, and Amartya Sanyal (2023). “PILLAR: How to make Semi-private learning more effective”. In: *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) 2024*
*Theory and Practice of Differential Privacy (TPDP)*.

Shi, Yuge, Imant Daunhawer, Julia E. Vogt, Philip H.S. Torr, and Amartya Sanyal (2023). “How robust are pre-trained models to distribution shift?” In: *International Conference on Learning Representations (ICLR)*.

Yüce, Gizem, Alenxandru Tifrea, Amartya Sanyal, and Fanny Yang (2023). “Can semi-supervised learning use all the data effectively? A lower bound perspective”. In: *Neural Information Processing Systems (NeurIPS)* ***Spotlight Paper***.

Jorge, Pau de, Amartya Sanyal, Adel Bibi, Ricardo Volpi, Gregory Rogez, Puneet K. Dokania, and Philip H. S. Torr (2022). “Make Some Noise: Reliable and Efficient Single-Step Adversarial Training”. In: *Advanced in Neural Information Processing Systems (NeurIPS)*.

Sanyal, Amartya, Yaxi Hu, and Fanny Yang (2022). “How unfair is private learning?” In: *Conference on Uncertainty in Artificial Intelligence (UAI)* ***Oral Paper***.

Sanyal, Amartya and Giorgia Ramponi (2022). “Open Problem: Do you pay for Privacy in Online learning?” In: *Conference on Learning Theory (COLT), Open Problem*.

Jorge, Pau de, Amartya Sanyal, Harkirat S. Behl, Philip H. S. Torr, Gregory Rogez, and Puneet K. Dokania (2021). “Progressive Skeletonization: Trimming more fat from a network at initialization”. In: *International Conference on Learning Representations (ICLR),*

Sanyal, Amartya, Varun Kanade, Philip H.S. Torr, and Puneet K. Dokania (2021). “ How Benign is Benign Overfitting ?” In: *International Conference on Learning Representations (ICLR),* ***Spotlight Paper***.

Mukhoti, Jishnu, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania (2020). “The Intriguing Effects of Focal Loss on the Calibration of Deep Neural Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Sanyal, Amartya, Philip H.S. Torr, and Puneet K. Dokania (2020). “ Stable Rank Normalization for Improved Generalization in Neural Networks and GANs”. In: *International Conference on Learning Representations (ICLR),* ***Spotlight Paper***.

Sanyal, Amartya, Matt Kusner, Adria Gascon, and Varun Kanade (2018). “TAPAS: Tricks to Accelerate (encrypted) Prediction As a Service”. In: *International Conference on Machine Learning (ICML)*.

## PREPRINTS

Barez, Fazl, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O’Gara, Robert Kirk, Ben Bucknall, Tim Fist, et al. (2025). “Open problems in machine unlearning for ai safety”. In: *arXiv preprint arXiv:2501.04952*.

Ben-Dov, Omri, Samira Samadi, Amartya Sanyal, and Alexandru Ţifrea (2025). “Fairness for the People, by the People: Minority Collective Action”. In: *arXiv preprint arXiv:2508.15374*.

Düngler, Johanna and Amartya Sanyal (2025). “An Iterative Algorithm for Differentially Private $k$-PCA with Adaptive Noise”. In: *arXiv preprint arXiv:2508.10879*.

Hu, Yaxi, Bernhard Schölkopf, and Amartya Sanyal (2025). “Online Learning and Unlearning”. In: *arXiv:2505.08557*.

Li, Wenjie, Jiawei Li, Christian Schroeder de Witt, Ameya Prabhu, and Amartya Sanyal (2024). “Delta-Influence: Unlearning Poisons via Influence Functions”. In: *arXiv preprint arXiv:2411.13731*.

Hu, Yaxi, Francesco Pinto, Amartya Sanyal, and Fanny Yang (2023). “Semi-private learning via low dimensional structures”. In: *Third Workshop on Seeking Low-Dimensionality in Deep Neural Networks*.

Bartolomeis, Piersilvio De, Jacob Clarysse, Fanny Yang, and Amartya Sanyal (2022). “How robust accuracy suffers from certified training with convex relaxations”. In: *NeurIPS 2022: Workshop on Understanding Deep Learning Through Empirical Falsification* ***Contributed Talk***,

Ortiz-Jimenez, Guillermo, Pau de Jorge, Amartya Sanyal, Adel Bibi, Puneet Dokania, Pascal Frossard, Gregory Rogez, and Philip H. S. Torr (2022). “Catastrophic Overfitting is a bug but also a feature”. In: *ICML 2022: Workshop on New Frontiers In Adversarial Machine Learning*.

Sanyal, Amartya, Varun Kanade, Philip H.S. Torr, and Puneet Dokania (2018). “Robustness via Deep Low Rank Representations”. In: *ICML 2018: Workshop on Theory and Application of Deep Generative Models*.

Merriënboer, Bart van, Amartya Sanyal, Hugo Larochelle, and Yoshua Bengio (2017). “Multiscale sequence modeling with a learned dictionary”. In: *ICML 2017: Workshop on Machine Learning in Speech and Language Processing*.