Aounon Kumar

Harvard University, Cambridge, MA.
aounon.github.io
aokumar@hbs.edu

Experience

2023 - Present **Harvard University**, Research Associate (Postdoc)
Working in Trustworthy Machine Learning focusing on AI Safety and Certifiable Adversarial Robustness.

Summer 2022 **Amazon Go**, Applied Scientist Intern
Worked on uncertainty estimation and out-of-distribution detection for action recognition. Used temporal information in video data to improve out-of-distribution detection.

Summer 2019 **Nokia Bell Labs**, Research Intern
Used machine learning techniques to improve network security. The objective was to build a firewall that could adapt to changing threat patterns and block suspicious IP addresses.

Summer 2018 **Nokia Bell Labs**, Research Intern
Worked on a theoretical analysis of single-layer autoencoders. The goal was to understand the class of problems that can be solved using feed-forward networks with ReLU activation.

2021-2023 **University of Maryland**, Graduate Research Assistant

Education

2017 – 2023 **Ph.D. in Computer Science**, *University of Maryland – College Park*, GPA 3.84/4.0
Thesis: Extending the Scope of Provable Adversarial Robustness in Machine Learning
Advisors: Soheil Feizi and Tom Goldstein

2015 – 2017 **Master of Science (Research) in Computer Science and Engineering**, *Indian Institute of Technology Delhi*, GPA 9.69/10
Thesis: The Capacitated k-Center Problem and its Variant with Vertex Weights
Advisors: Naveen Garg and Amit Kumar

2011 – 2015 **B-Tech in Computer Science and Engineering**, *Indian Institute of Technology Mandi*, GPA 8.58/10
Project Title: The Steiner Tree problem

Research Interests

Machine Learning, AI Safety, Certifiable Adversarial Robustness, Distributional Robustness, Language Models, and Reinforcement Learning.

Research Statement (Summary)

Full version available here

Deep neural networks and other machine learning models are known to malfunction under minor changes in the input. My research aims to design methods that have provable guarantees of robustness, also known as **robustness certificates**, against input corruptions. I research certified robustness techniques for a wide range of learning settings, including tasks with **structured outputs** such as semantic segmentation and image generation, as well as the dynamic and adaptive settings of **reinforcement learning** and **distribution shifts**. The goal of my research is to extend provable robustness to real-world applications in different learning paradigms.

Previously, I have also worked in theoretical computer science studying NP-hard **combinatorial optimization problems** such as the k-center clustering problem. I worked on designing approximation algorithms with fairness constraints for this problem and also studied its computational hardness. In the future, I plan to continue my research in machine learning robustness and the broader field of trustworthy AI. Given my background in theoretical computer science and machine learning, I am also interested in working in the overlap of these two areas.

Publications

Preprint 2023 — Certifying LLM Safety against Adversarial Prompting [PDF]
Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, Hima Lakkaraju
`https://arxiv.org/abs/2309.02705`
Media Coverage: Science News Magazine, D^3 Institute at Harvard.

Preprint 2023 — Can AI-Generated Text be Reliably Detected? [PDF]
Vinu Sankar Sadasivan, **Aounon Kumar**, Sriram Balasubramanian, Wenxiao Wang, Soheil Feizi
`https://arxiv.org/abs/2303.11156`
Media Coverage: The Washington Post, Bloomberg, Wired, New Scientist, The Register, TechSpot.

ICML 2020 — Curse of Dimensionality on Randomized Smoothing for Certifiable Robustness [PDF]
Aounon Kumar, Alexander Levine, Tom Goldstein, Soheil Feizi
`https://arxiv.org/abs/2002.03239`

ICLR 2022 — Policy Smoothing for Provably Robust Reinforcement Learning [PDF]
Aounon Kumar, Alexander Levine, Soheil Feizi
`arxiv.org/abs/2106.11420`

ICLR 2023 — Provable Robustness against Wasserstein Distribution Shifts via Input Randomization [PDF]
Aounon Kumar, Alexander Levine, Tom Goldstein, Soheil Feizi
`arxiv.org/abs/2201.12440`

NeurIPS 2021 — Center Smoothing: Provable Robustness for Networks with Structured Outputs [PDF]
Aounon Kumar, Tom Goldstein
`arxiv.org/abs/2102.09701`

NeurIPS 2020 — Certifying Confidence via Randomized Smoothing [PDF]
Aounon Kumar, Alexander Levine, Soheil Feizi, Tom Goldstein
`arxiv.org/abs/2009.08061`

NeurIPS 2020 Detection as Regression: Certified Object Detection by Median Smoothing [PDF]
Ping-yeh Chiang, Michael J. Curry, Ahmed Abdelkader, **Aounon Kumar**, John Dickerson, Tom Goldstein
`arxiv.org/abs/2007.03730`

Preprint 2020 Tight Second-Order Certificates for Randomized Smoothing [PDF]
Alexander Levine, **Aounon Kumar**, Thomas Goldstein, Soheil Feizi
`arxiv.org/abs/2010.10549`

APPROX 2019 On the cost of essentially fair clusterings [PDF]
Ioana O. Bercea, Martin Groß, Samir Khuller, **Aounon Kumar**, Clemens Rösner, Daniel R. Schmidt and Melanie Schmidt
`arxiv.org/abs/1811.10319`

FSTTCS 2016 Capacitated k-Center Problem with Vertex Weights [PDF]
Aounon Kumar

Media Coverage

My recent works on LLM safety and reliability have been featured in popular tech magazines and academic news outlets:

1. Science News Magazine, D^3 Institute at Harvard. Work featured: Certifying LLM Safety against Adversarial Prompting.
2. The Washington Post, Bloomberg, Wired, New Scientist, The Register, TechSpot. Work featured: Can AI-Generated Text be Reliably Detected?.

Select Projects

1. **Provable Robustness against Wasserstein Shifts:** I developed a robustness certificate for the performance of machine learning models under shifts in the input distribution such as RGB shifts, hue shifts, and brightness/saturation changes. It is an efficient technique that can certify neural networks that are several layers deep.
2. **Certified Reinforcement Learning:** In this work, presented at ICLR 2022, I proved robustness guarantees for an RL agent that randomizes its observations of the environment before passing them through the policy network. The robustness certificate guarantees that the total reward obtained by the agent under an adversarial attack remains above a certain threshold.
3. **Certified Robustness beyond Classification:** One of the objectives of my research is to extend provable robustness beyond classifier outputs to more complex outputs like images, segmentation masks, and abstract latent representations. In NeurIPS 2021, I presented a procedure for certifying such structured outputs under several commonly used distance metrics such as LPIPS, cosine distance, and intersection-over-union. In another work, presented at NeurIPS 2020, I develop a procedure for certifying the *confidence score* produced by conventional neural networks which is often used to estimate the uncertainty in their predictions.

4. **Curse of Dimensionality:** In this work, presented at ICML 2020, I studied the limitations of a popular certified robustness technique called randomized smoothing that obtains good certificates against ℓ_1 and ℓ_2-norm bounded adversaries. My work shows that it suffers from the curse of dimensionality for higher norms such as the ℓ_∞-norm. The theoretical results prove that the best possible ℓ_∞-certificate decays at the rate of $O(1/\sqrt{d})$ with the input dimensionality d, regardless of the choice of the smoothing distribution.

Academic Service

I have served as a reviewer for prominent machine learning conferences such as NeurIPS (2022, 2021), and ICLR (2024).

Relevant Courses

Ph. D. Introduction to Quantum Information Processing, Scientific Computing, Advanced Numerical Optimization

M.S. (R) Advanced Algorithms, Theory of Computation and Complexity Theory, Cryptography and Computer Security, Machine Learning

B-Tech Advanced Algorithms, Modern Techniques in Theory of Computation, Advanced Theory of Computation, Advanced Complexity Theory, Mathematical Concepts in Computer Science, Algorithm Design and Analysis, Advanced Data Structures and Algorithms, Formal Languages and Automata Theory, Artificial Intelligence, Pattern Recognition, Machine Learning

Teaching Experience

2017-2020 **University of Maryland**, Teaching Assistant
- CMSC250: Discrete structures
- CMSC351: Algorithms
- CMSC451: Design and analysis of computer algorithms

2015-2017 **Indian Institute of Technology Delhi**, Teaching Assistant
- Discrete mathematics
- Introduction to Automata and Theory of Computation
- Analysis and Design of Algorithms

Programming Languages

Python, MATLAB, C++

Deep Learning Frameworks: PyTorch, Torchvision, TensorFlow, Keras.

Other Tools: Numpy, Scipy, Matplotlib, Linux, LaTeX