Alexander Robey

3401 Walnut Street
Philadelphia, PA 19104
✉ arobey1@upenn.edu
https://arobey1.github.io/

Education

2018–present **PhD, Electrical and Systems Engineering**, *University of Pennsylvania*
2014-2018 **Bachelor of Science, Engineering**, *Swarthmore College*
2014-2018 **Bachelor of Arts, Mathematics**, *Swarthmore College*

Work Experience

2022-2023 Student researcher, Google Cloud AI
2022 Research intern, Google Cloud AI
2017 Research intern, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory
2016, 2018 Research assistant, Department of Engineering, Swarthmore College

Fellowships

2023 *Amazon AWS Research Fellowship*
2018 *Dean's Fellowship*, Department of Electrical and Systems Engineering, University of Pennsylvania
2016 *Undergraduate Research Fellowship*, Department of Engineering, Swarthmore College

Awards

2023 *Best Paper Award*, 2nd AdvML-Frontiers Workshop at ICML 2023.
2022 *Outstanding Reviewer Award*, NeurIPS 2022.
2022 *Outstanding Reviewer Award*, ICML 2022.
2021 *Outstanding Reviewer Award*, ICLR 2021.
2021 *Outstanding Reviewer Award*, NeurIPS 2021.
2020 *Outstanding Reviewer Award*, ICML 2020.
2020 *Teaching Assistant of the Year*, Department of Electrical and Systems Engineering, University of Pennsylvania.

Refereed Conference Papers

2023 Haoze Wu*, Teruhiro Tagomori*, **Alexander Robey***, Fengjun Yang*, Nikolai Matni, George J. Pappas, Hamed Hassani, Corina Pasareanu, and Clark Barrett. Toward certified robustness against real-world distribution shifts. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 2023.

2022 Allan Zhou, Fahim Tajwar, **Alexander Robey**, Tom Knowles, George J. Pappas, Hamed Hassani, and Chelsea Finn. Do Deep Networks Transfer Invariances across Classes? In *International Conference on Learning Representations*, 2022.

2022 Anton Xue, Lars Lindemann, **Alexander Robey**, Hamed Hassani, George J. Pappas, and Rajeev Alur. Chordal Sparsity for Lipschitz Constant Estimation of Deep Neural Networks. In *2022 61st IEEE Conference on Decision and Control (CDC)*. IEEE, 2022.

2022 Stephen Tu, **Alexander Robey**, Tingnan Zhang, and Nikolai Matni. On the Sample Complexity of Stability Constrained Imitation Learning. In *Learning for Dynamics and Control*. PMLR, 2022.

2022 **Alexander Robey**, Luiz F. O. Chamon, George J. Pappas, and Hamed Hassani. Probabilistically Robust Learning: Balancing Average-and Worst-case Performance. In *International Conference on Machine Learning*. PMLR, 2022.

2022 Cian Eastwood*, **Alexander Robey***, Shashank Singh, Julius von Kügelgen, Hamed Hassani, George J. Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 2022.

2021 **Alexander Robey***, Luiz F. O. Chamon*, George J. Pappas, Hamed Hassani, and Alejandro Ribeiro. Adversarial Robustness with Semi-Infinite Constrained Learning. In *Advances in Neural Information Processing Systems*, 2021.

2021 **Alexander Robey**, George J. Pappas, and Hamed Hassani. Model-Based Domain Generalization. In *Advances in Neural Information Processing Systems*, 2021.

2021 **Alexander Robey**, Lars Lindemann, Stephen Tu, and Nikolai Matni. Learning Robust Hybrid Control Barrier Functions for Uncertain Systems. *IFAC Conference on Analysis and Design of Hybrid Systems*, 2021.

2021 **Alexander Robey**, Arman Adibi, Brent Schlotfeldt, George J. Pappas, and Hamed Hassani. Optimal Algorithms for Submodular Maximization with Distributed Constraints. In *Learning for Dynamics and Control*. PMLR, 2021.

2021 Lars Lindemann, Haimin Hu, **Alexander Robey**, Hanwen Zhang, Dimos V Dimarogonas, Stephen Tu, and Nikolai Matni. Learning Hybrid Control Barrier Functions from Data. *Conference on Robot Learning*. PMLR, 2021.

2020 **Alexander Robey***, Haimin Hu*, Lars Lindemann, Hanwen Zhang, Dimos V Dimarogonas, Stephen Tu, and Nikolai Matni. Learning Control Barrier Functions from Expert Demonstrations. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3717–3724. IEEE, 2020.

2019 Mahyar Fazlyab, **Alexander Robey**, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and Accurate Estimation of Lipschitz Constants for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, pages 11427–11438, 2019.

Journal Papers

2022 Edgar Dobriban, Hamed Hassani, David Hong, and **Alexander Robey**. Provable Tradeoffs in Adversarially Robust Classification. *IEEE Transactions on Information Theory*. IEEE, 2022.

2018 **Alexander Robey** and Vidya Ganapati. Optimal Physical Preprocessing for Example-based Super Resolution. *Optics Express*, volume 26, pages 31333–31350. Optical Society of America, 2018.

Preprints

2023 **Alexander Robey***, Fabian Latorre*, George J. Pappas, Hamed Hassani, and Volkan Cevher. Adversarial training should be cast as a non-zero-sum game. *arXiv preprint arXiv:2306.11035*, 2023.

2023 **Alexander Robey**, Eric Wong, Hamed Hassani, and George J Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

2023 Patrick Chao, **Alexander Robey**, Eric Wong, Hamed Hassani, George J. Pappas, and Edgar Dobriban. Jailbreaking llms in twenty questions. *arXiv preprint arXiv:2310.08419*, 2023.

2021 Lars Lindemann, **Alexander Robey**, Lejun Jiang, Stephen Tu, and Nikolai Matni. Learning Robust Output Control Barrier Functions from Safe Expert Demonstrations. *arXiv preprint arXiv:2111.09971*, 2021.

2020 **Alexander Robey**, Hamed Hassani, and George J. Pappas. Model-Based Robust Deep Learning. *arXiv preprint arXiv:2005.10247*, 2020.

Patents

2020 **Alexander Robey**, Hamed Hassani, and George J Pappas. Model-Based Robust Deep Learning, 2020. United States Provisional Patent 63/034,355.

Professional Activities

Organizing

2022 ECCV workshop on *Adversarial Robustness in the Real World*.

2021 ICCV workshop on *Adversarial Robustness in the Real World*.

Reviewing (conferences)

Neural Information Processing Systems (NeurIPS)

International Conference on Machine Learning (ICML)
International Conference on Learning Representations (ICLR)
The AAAI Conference on Artifical Intelligence (AAAI)
International Conference on Cyber-Physical Systems (ICCPS)
Learning for Dynamics and Control (L4DC)
Conference on Decision and Control (CDC)
American Control Conference (ACC)
International Conference on Computer Vision (ICCV)
European Conference on Computer Vision (ECCV)
International Symposium on Information Theory (ISIT)

Reviewing (journals)

Journal of Machine Learning Research (JMLR)
Transactions on Machine Learning Research (TMLR)
IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)
IEEE Transactions on Neural Networks and Learning Systems
IEEE Transactions on Knowledge and Data Engineering
IEEE Transactions on Artificial Intelligence
IEEE Robotics and Automation Letters
Transactions on Automatic Control (TAC)
SIAM Journal on Mathematics of Data Science (SIMODS)
Springer Nature Journal on Machine Learning
Springer International Journal on Computer Vision (IJCV)

Reviewing (workshops & special tracks)

2023 R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (NeurIPS)
2023 Distribution Shifts: New Frontiers with Foundation Models (NeurIPS)
2023 Adversarial Robustness in the Real World (ICCV)
2023 Out-of-Distribution Generalization in Computer Vision (ICCV)
2023 Adversarial Machine Learning Frontiers (ICML)
2023 Blog post track (ICLR)
2023 Workshop on Domain Generalization (ICLR)
2023 Safe and Robust AI special track (AAAI)
2022 Distribution Shifts (NeurIPS)
2022 Robustness in Sequence Modeling (NeurIPS)
2022 Out-Of-Distribution Generalization in Computer Vision (ECCV)
2022 Adversarial Robustness in the Real World (ECCV)
2022 Adversarial Machine Learning Frontiers (ICML)
2021 Distribution Shifts: Connecting Methods and Applications (NeurIPS)
2021 Adversarial Robustness in the Real World (ICCV)
2020 Adversarial Robustness in the Real World (ECCV)

Technical Skills

Programming languages: Python, MATLAB, JavaScript, HTML, CSS, R, C/C++, SQL (Postgres), Verilog HDL, LaTeX

Frameworks: Pytorch, TensorFlow, Jax, Django, Slurm

Teaching Experience

Spring 2022 ENGR 56: *Modeling and Optimization for Engineering*, Swarthmore College (guest lecturer)
Spring 2021 ESE 605: *Modern Convex Optimization*, University of Pennsylvania (teaching assistant)
Spring 2020 ESE 290: *Introduction to Research Methodologies*, University of Pennsylvania (teaching assistant)

Fall 2020 ESE 530: *Elements of Probability Theory*, University of Pennsylvania (teaching assistant)
Fall 2019 ESE 530: *Elements of Probability Theory*, University of Pennsylvania (teaching assistant)
Spring 2018 ENGR 019: *Numerical Methods for Engineering Applications*, Swarthmore College (teaching assistant)
Fall 2017 ENGR 011: *Electrical Circuit Analysis*, Swarthmore College (teaching assistant)
Spring 2017 ENGR 012: *Linear Physical Systems Analysis*, Swarthmore College (teaching assistant)
Fall 2016 ENGR 011: *Electrical Circuit Analysis*, Swarthmore College (teaching assistant)
Spring 2016 ENGR 006: *Engineering Mechanics*, Swarthmore College (teaching assistant)

Selected Talks

Oct. 2023 *SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks*, EnCORE Institute for Emerging CORE Methods in Data Science
July 2023 *Adversarial Training Should Be Cast as a Non-Zero-Sum Game*, ICML workshop on New Frontiers in Adversarial Machine Learning
Nov. 2022 *Learning Under Robustness Constraints: From perturbations to Distribution Shifts*, UCSD
Oct. 2022 *Learning Under Robustness Constraints*, INFORMS Annual Meeting
July 2022 *Probabilistically Robust Learning: Balancing Worst- and Average-case Performance*, ICML
Mar. 2022 *Probabilistically Robust Learning: Balancing Worst- and Average-case Performance*, UCSD
Mar. 2022 *Toward Robust, Generalizable Deep Learning*, Swarthmore College
Dec. 2021 *Model-Based Domain Generalization*, CDC Workshop on Robust Deep Learning-Based Control
Oct. 2021 *Robustness against Natural Variation: Theory and Practice*, University of Pennsylvania
Sept. 2021 *Robustness against Natural Variation: Theory and Practice*, Simons Foundation
July 2021 *Learning Robust Hybrid Control Barrier Functions for Uncertain Systems*, ADHS
Apr. 2021 *Model-Based Robust Deep Learning: Generalizing to Natural Out-of-Distribution Data*, Italian Conference on CyberSecurity (ITASEC), AI for Security and Security for AI Workshop
Mar. 2021 *Model-Based Domain Generalization*, Simons Foundation
Dec. 2020 *Learning Control Barrier Functions from Expert Demonstrations*, CDC
Nov. 2020 *Model-Based Robust Deep Learning: Generalizing to Natural Out-of-Distribution Data*, Intel
Oct. 2020 *Generalizing to Natural Out-of-Distribution Data*, C3.ai workshop on the Analytical Foundations of Deep Learning
Sept. 2020 *Model-Based Robust Deep Learning: Generalizing to Natural Out-of-Distribution Data*, Intel
Sept. 2020 *Model-Based Robust Deep Learning: Generalizing to Natural Out-of-Distribution Data*, Data Augmentation and Equivariance Workshop
Aug. 2020 *Model-Based Robust Deep Learning: Generalizing to Natural Out-of-Distribution Data*, ECCV workshop on Adversarial Robustness in the Real World
July 2020 *Model-Based Robust Deep Learning*, Stanford University
Dec. 2019 *Efficient and Accurate Estimation of Lipschitz Constants of Deep Neural Networks*, NeurIPS
May 2018 *Computationally Expediting Fourier Ptychographic Miscroscopy*, Swarthmore College

Selected Poster Presentations

Oct. 2023 *SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks*, Yale Foundations of Data Science workshop on Theory and Practice of Foundation Models
July 2023 *Adversarial Training Should Be Cast as a Non-Zero-Sum Game*, ICML workshop on New Frontiers in Adversarial Machine Learning
Feb. 2023 *Toward Certified Robustness Against Real-World Distribution Shifts*, SaTML
Dec. 2022 *Probable Domain Generalization via Quantile Risk Minimization*, NeurIPS
Oct. 2022 *Probabilistically Robust Learning: Balancing Average and Worst-case Performance*, NSF
Sept. 2022 *Probabilistically Robust Learning: Balancing Average and Worst-case Performance*, Simons Foundation
Sept. 2022 *Probabilistically Robust Learning: Balancing Average and Worst-case Performance*, The Institute for Emerging CORE Methods in Data Science
July 2022 *Probabilistically Robust Learning: Balancing Average and Worst-case Performance*, ICML

July 2022	*Toward Certified Robustness Against Real-World Distribution Shifts*, ICML workshop on Formal Verification of Machine Learning
Apr. 2022	*Do Deep Networks Transfer Invariance Across Classes?*, ICLR
Dec. 2021	*Adversarial Robustness via Semi-Infinite Constrained Learning*, NeurIPS
Dec. 2021	*Model-Based Domain Generalization*, NeurIPS
Sept. 2021	*Model-Based Domain Generalization*, Simons Foundation
July 2021	*Model-Based Robust Deep Learning: Generalizing to Natural Out-of-Distribution Data*, ICML workshop on Uncertainty and Robustness in Deep Learning
June 2021	*Optimal Algorithms for Submodular Maximization with Distributed Constraints*, L4DC
Nov. 2020	*Learning Hybrid Control Barrier Functions from Data*, CoRL
Dec. 2019	*Efficient and Accurate Estimation of Lipschitz Constants of Deep Neural Networks*, NeurIPS
Aug. 2017	*Quantifying the Impact of Factors Affecting Communication Performance*, Lawrence Livermore National Laboratory
Sept. 2016	*Bird-Window Collision Prevention*, Swarthmore College