Research Statement

Aounon Kumar (aounon@umd.edu)

My primary research interest is in machine learning robustness with provable guarantees. Deep neural networks and other machine learning models are known to malfunction under minor changes in the input. This weakness poses a serious risk to the deployment of such models in real-world scenarios, particularly for safety-critical applications such as autonomous driving and medical diagnosis. Unstable behavior can lead to a loss of trust in machine learning models, hindering their adoption in society. While several empirical methods have been developed to defend models against input corruptions, they often break down against unseen perturbations, making it difficult to determine the true robustness of a model. My research aims to design methods that have provable guarantees of robustness, also known as **robustness certificates**. Unlike empirical defenses, certified methods can produce a mathematical description of a set of perturbations, both seen and unseen, for which a model is guaranteed to be robust.

I conduct research in certified robustness for a wide range of learning setups, 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 overarching goal of my research is to extend provable robustness to real-world applications in different machine learning paradigms. Existing techniques often lack the practical feasibility needed for real-world applications. For instance, randomized smoothing evaluates a model over millions of noisy input samples just to make a single inference. Methods based on Lipschitz continuity and interval-bound propagation struggle to scale up to deeper neural networks and higher input dimensions. Most existing approaches focus only on classification tasks and are difficult to adapt to other modalities. My research aims to address these limitations by designing methods that require fewer model evaluations, are viable for large neural networks, and can adapt to multiple tasks. The key principles guiding my research are **efficiency**, **scalability** and **flexibility**. In the following sections, I will discuss some of my thesis research and highlight how these guiding principles shape my work. I am grateful to have had the opportunity to work with a diverse group of collaborators whose expertise and dedication have been instrumental in driving my research forward.

My Main Contributions

In the first paper of my thesis research, presented at ICML 2020, I study the limitations of a popular certified robustness technique called randomized smoothing [1]. This is the first technique that could scale up to high-dimensional problems such as ImageNet classification and provide good robustness certificates against ℓ_1 and ℓ_2-norm bounded adversaries [2, 3]. My work shows that it suffers from the **curse of dimensionality for higher norms** such as the ℓ_∞-norm. The ℓ_∞-norm produces a more interpretable threat model than ℓ_1 and ℓ_2-norms, as it puts independent perturbation bounds on each input element, e.g., each pixel's intensity in an image being perturbed by at most $8/255$. My work mathematically proves 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. This suggests that, despite its impressive performance for lower norms, randomized smoothing may not be suitable against ℓ_∞-adversaries for high-dimensional inputs like images. While this is a setback towards making provable robustness more flexible and scalable, the focus of the upcoming papers will be on achieving these qualities in other settings such as reinforcement learning and distribution shifts.

The literature on provable robustness focuses mainly on static, supervised learning tasks like image classification. However, deep neural networks are extensively used for dynamic settings like reinforcement learning (RL) and streaming tasks, making such systems vulnerable to adversarial attacks as well. My ICLR 2022 paper designs a **robustness certificate for RL** that can certify the expected total reward obtained by an agent [4]. It presents an efficient method called Policy Smoothing that simply randomizes the agent's observations before passing them through the policy network. It only requires one sample per time-step, keeping the computational complexity of the original policy unaffected. The certificate takes the adaptive nature of an RL adversary into account and works well even for long episodes. It can certify for challenging RL problems such as Atari game environments like Pong and Freeway.

While conventional robustness certificates focus on adversaries with a fixed attack budget for every sample in the input distribution, a real-world adversary may choose to allocate a different budget for each sample depending on its adversarial vulnerability. Natural perturbations, such as RGB shifts, blur and noise, also vary in size depending on factors like time, location and operating conditions. One of my recent papers presents a method that can certify a model's accuracy under **Wasserstein shifts** of the data distribution [5]. It allows the datum-specific perturbation size to vary across different points in the input distribution and can certify neural networks that are several layers deep. By randomizing the input within a certain transformation space this method can produce distributional certificates for a host of visual corruptions like color shifts, hue shifts, and brightness/saturation (SV) changes (see Figure 1).

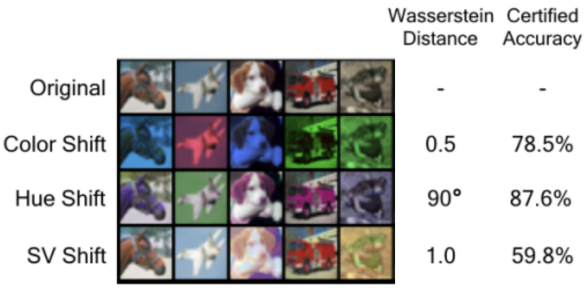

Figure 1: Distributional certificates against natural perturbations.

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. Unlike classification certificates which guarantee that the predicted class remains unchanged, the robustness of such structured outputs can be measured using an appropriate distance function in the output or latent space. My NeurIPS 2021 paper presents a procedure called Center Smoothing that can **certify neural networks with structured outputs** [6]. It can produce robustness guarantees for 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 [7]. It certifies the expected value of the confidence score by observing its distribution around an input point.

Before beginning my thesis research, I worked in theoretical computer science studying NP-hard **combinatorial optimization problems** such as the k-center clustering problem. I worked on approximation algorithms with fairness constraints for this problem [8] and also studied its computational hardness [9]. Conventional clustering algorithms used in several machine learning and data science applications can lead to clusters that are biased in favor of a particular group. My work studied algorithms with explicit fairness constraints on the composition of the clusters. In the hardness paper, I showed that the k-center problem with vertex weights and capacities cannot be approximated within a constant factor of the optimal solution, unless P = NP.

Future Work

I plan to continue my research in machine learning robustness and the broader field of trustworthy AI. More specifically, I am interested in studying robustness in dynamic and adaptive settings like reinforcement learning and streaming applications. Proving robustness guarantees in these settings is often more challenging because a worst-case adversary could strengthen itself by adapting to the defense strategy used by the victim model. I am also interested in studying non-ℓ_p threat models that can capture changes in the semantics of the input, which are difficult to analyse using existing mathematical tools. I am also curious about how certified defenses for abstract vector representations, such as the Center Smoothing procedure above, can produce robustness guarantees in unsupervised and semi-supervised learning settings.

Given my background in theoretical computer science and machine learning, I am also interested in working in the overlap of these two areas. For instance, I am interested in studying how reinforcement learning techniques could be applied to solve discrete optimization problems for which efficient algorithms may not be easy to design. Reinforcement learning can automate the search for new algorithms with minimal human intervention [10]. However, current methods are designed with the specific problem in mind, and model architectures need to be adapted every time the problem description changes. If we could build a unified framework that decouples the model architecture from the problem description, it could have a profound impact in several industries ranging from logistics and transportation to chip design and telecommunications.

References

[1] **Aounon Kumar**, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*.

[2] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy, SP 2019*.

[3] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[4] **Aounon Kumar**, Alexander Levine, and Soheil Feizi. Policy smoothing for provably robust reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022*.

[5] **Aounon Kumar**, Alexander Levine, Tom Goldstein, and Soheil Feizi. Certifying model accuracy under distribution shifts. *CoRR*, abs/2201.12440, 2022.

[6] **Aounon Kumar** and Tom Goldstein. Center smoothing: Certified robustness for networks with structured outputs. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*.

[7] **Aounon Kumar**, Alexander Levine, Soheil Feizi, and Tom Goldstein. Certifying confidence via randomized smoothing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*.

[8] Ioana Oriana Bercea, Martin Groß, Samir Khuller, **Aounon Kumar**, Clemens Rösner, Daniel R. Schmidt, and Melanie Schmidt. On the cost of essentially fair clusterings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2019, September 20-22, 2019, Massachusetts Institute of Technology, Cambridge, MA, USA*.

[9] **Aounon Kumar**. Capacitated k-center problem with vertex weights. In *36th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2016*.

[10] Nina Mazyavkina, Sergey Sviridov, Sergei Ivanov, and Evgeny Burnaev. Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, 134:105400, 2021.