PROSAC: Provably Safe Certification for Machine Learning Models under Adversarial Attacks

Chen Feng[1], Ziquan Liu[2], Zhuo Zhi[1], Ilija Bogunovic[1], Carsten Gerner-Beuerle[3], Miguel Rodrigues[4]

[1]Department of Electronic and Electrical Engineering, University College London
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London
[3]Faculty of Laws, University College London
[4]AI Centre, Department of Electronic and Electrical Engineering, University College London
chen.feng@ucl.ac.uk, ziquan.liu@qmul.ac.uk, {zhuo.zhi.21, i.bogunovic, c.gerner, m.rodrigues}@ucl.ac.uk

Abstract

It is widely known that state-of-the-art machine learning models, including vision and language models, can be seriously compromised by adversarial perturbations. It is therefore increasingly relevant to develop capabilities to certify their performance in the presence of the most effective adversarial attacks. Our paper offers a new approach to certify the performance of machine learning models in the presence of adversarial attacks with population level risk guarantees. In particular, we introduce the notion of (α, ζ)-safe machine learning model. We propose a hypothesis testing procedure, based on the availability of a calibration set, to derive statistical guarantees providing that the probability of declaring that the adversarial (population) risk of a machine learning model is less than α (i.e. the model is safe), while the model is in fact unsafe (i.e. the model adversarial population risk is higher than α), is less than ζ. We also propose Bayesian optimization algorithms to determine efficiently whether a machine learning model is (α, ζ)-safe in the presence of an adversarial attack, along with statistical guarantees. We apply our framework to a range of machine learning models - including various sizes of vision Transformer (ViT) and ResNet models - impaired by a variety of adversarial attacks, such as PGDAttack, MomentumAttack, GenAttack and BanditAttack, to illustrate the operation of our approach. Importantly, we show that ViT's are generally more robust to adversarial attacks than ResNets, and large models are generally more robust than smaller models. Our approach goes beyond existing empirical adversarial risk-based certification guarantees. It formulates rigorous (and provable) performance guarantees that can be used to satisfy regulatory requirements mandating the use of state-of-the-art technical tools.

Introduction

With the development of increasingly capable autonomous machine learning systems and their use in a range of domains from healthcare to banking and finance, education, and e-commerce, to name just a few, policy makers across the world are in the process of formulating detailed regulatory requirements that will apply to developers and operators of AI systems. The EU is at the forefront of the drive to regulate AI systems. Proposals for an EU AI Act, an AI Liability Directive and an extension of the EU Product Liability Directive to AI systems and AI-enabled goods are at advanced stages of the legislative process. Other jurisdictions, too, pursue a variety of regulatory initiatives, and standard setters such as the National Institute of Standards and Technology in the United States and the Supreme Audit Institutions of Germany, the UK, and other countries have started work on more precise standards, including standards concerning the robustness of machine learning systems in the presence of adversarial attacks.

Regulatory frameworks adopted so far are mostly high-level, but those that establish more detailed requirements for AI systems to be put in service or for ongoing compliance, such as the EU AI Act, require an assessment of the performance of AI systems based on precise metrics. These metrics must include, among other things, an evaluation of the accuracy and resilience of a system in case of perturbations or unauthorised use.

It is thus important for those who deploy an AI system to have technical capabilities that allow a precise measurement of performance. However, developing certification procedures is not trivial due to the fact that state-of-the-art machine learning models are black-boxes that are poorly understood; furthermore, the standard train/validate/test paradigm often lacks rigorous quantifiable statistical guarantees and is therefore a poor certification instrument. Therefore, recent years have witnessed the introduction of various promising procedures, building on recent advances in statistics, that can be used to endow black-box / complex state-of-the-art machine learning models with statistical guarantees (Bates et al. 2021; Angelopoulos et al. 2021; Laufer-Goldshtein et al. 2023). For example, Bates et al. (2021) have proposed a framework to offer rigorous distribution-free error control of machine learning models for a variety of tasks. Angelopoulos et al. (2021) have proposed a procedure, the Learn-then-Test framework, that leverages multiple hypothesis testing techniques to calibrate machine learning models so that their predictions satisfy explicit, finite-sample statistical guarantees. Building on the Learn-then-Test framework, Laufer-Goldshtein et al. (2023) introduce a procedure to identify machine learning model risk-controlling configurations that also satisfy a variety of other objectives. Additionally, conformal prediction techniques have been proposed to quantify

the reliability of the predictions of machine learning models, e.g. Angelopoulos and Bates (2023).

Our paper builds on this line of research to offer an approach – Provably Safe Certification (PROSAC) – to certify the robustness of a machine learning model under adversarial attacks (Bruna et al. 2014; Chakraborty et al. 2018) with population-level guarantees, thereby differing from existing approaches that are limited to the certification of the empirical risk such as Cohen, Rosenfeld, and Kolter (2019); Wong and Kolter (2018) (see Section 2). In particular, we build on hypothesis testing techniques akin to those in Angelopoulos et al. (2021); Laufer-Goldshtein et al. (2023) to determine whether a model is robust against a specific adversarial attack. However, our approach differs from those in Angelopoulos et al. (2021); Laufer-Goldshtein et al. (2023) because we aim to guarantee that a machine learning model is safe for any attacker hyper-parameter configuration, rather than for at least one such hyper-parameter configuration. PROSAC is then used to benchmark a wide variety of state-of-the-art machine learning models, such as vision Transformers (ViT) and ResNet models, against a number of adversarial attacks, such as PGDAttack (Madry et al. 2018), MomentumAttack (Dong et al. 2018), GenAttack (Alzantot et al. 2019) and BanditAttack (Ilyas, Engstrom, and Madry 2018) in vision tasks.

Contributions: Our main contributions are as follows:

- We propose PROSAC, a new framework to certify whether a machine learning model is robust against a specific adversarial attack. Specifically, we propose a hypothesis testing procedure based on a notion of (α, ζ) machine learning model safety, entailing (loosely) that the adversarial risk of a model is less than a (pre-specified) threshold α with a (pre-specified) probability higher than ζ.
- We propose a Bayesian optimization algorithm — concretely, the (Improved) GP-UCB algorithm — to approximate the p-values associated with the underlying hypothesis testing problems, with a number of queries that scale much slower than the number of hyper-parameter configurations available to the attacker.
- We also demonstrate that – under a slightly more stringent testing procedure – the proposed Bayesian optimization algorithm allows us to rigorously certify (α, ζ)-safety of a specific machine learning model in the presence of a specific adversarial attack.
- Finally, we offer a series of experiments elaborating on (α, ζ)-safety of different machine learning models in the presence of different adversarial attacks. Notably, our framework reveals that ViTs appear to be more robust to adversarial perturbations than ResNets, and that large models appears to be more robust to adversarial perturbations than smaller models.

Organization: Our paper is organized as follows: The following section briefly reviews related work. Section 3 presents the problem statement, including the notion of (α, ζ) machine learning model safety under adversarial attacks. Section 4 presents our procedure to certify (α, ζ) machine learning model safety. It describes the algorithm to certify (α, ζ) machine learning model safety and presents associated guarantees. Section 5 offers experimental results to benchmark (α, ζ)-safety of various machine learning models under various attacks. Finally, we offer concluding remarks in Section 6. The proofs of the main technical results are relegated to the Supplementary Material.

Related Works

Adversarial Robustness Certification Different approaches have been proposed to certify the adversarial robustness of machine learning models (Li, Xie, and Li 2023). For example, a) set propagation methods (Wong and Kolter 2018; Wong et al. 2018; Gowal et al. 2018, 2019; Zhang et al. 2019); b) Lipschitz constant controlling methods (Hein and Andriushchenko 2017; Tsuzuku, Sato, and Sugiyama 2018; Trockman and Kolter 2020; Leino, Wang, and Fredrikson 2021; Zhang et al. 2021; Xu, Li, and Li 2022); and c) randomized smoothing techniques (Cohen, Rosenfeld, and Kolter 2019; Lecuyer et al. 2019; Salman et al. 2019; Carlini et al. 2023). Set propagation approaches need access to the model architecture and parameters so that an input polytope can be propagated from the input layer to the output layer to produce an upper bound for the worse-case input perturbation. This approach however requires the model architecture to be able to propagate sets, e.g. (Wong and Kolter 2018) relies on ReLU activation functions. Lipschitz constant controlling approaches produce adversarial robustness certification by bounding local Lipschitz constants; however, these approaches are limited to certain model architectures such as LipConvnet (Singla and Feizi 2021). In contrast, randomized smoothing (RS) represents a versatile certification methodology free from model architectural constraints or model parameter access.

Other Certification Approaches There are various other recent approaches to certify (audit) machine learning models in relation to issues including fairness or bias (Black, Yeom, and Fredrikson 2020; Xue, Yurochkin, and Sun 2020; Si et al. 2021; Taskesen et al. 2021; Chugg et al. 2023). For example, Black, Yeom, and Fredrikson (2020), Xue, Yurochkin, and Sun (2020), Taskesen et al. (2021) and Si et al. (2021) leverage hypothesis testing techniques – coupled with optimal transport approaches – to test whether a model discriminates against different demographic groups; Chugg et al. (2023) leverages recent advances in (sequential) hypothesis testing techniques – the "testing by betting" framework – to continuously test (monitor) whether a model is fair. Our certification framework also leverages hypothesis testing techniques, but the focus is on certifying the model adversarial robustness rather than model fairness.

Distribution-free Uncertainty Quantification Our certification framework builds on recent work on distribution-free risk quantification (Bates et al. 2021; Angelopoulos et al. 2021). In particular, Bates et al. (2021); Angelopoulos et al. (2021) seek to identify model hyper-parameter configurations that offer a pre-specified level of risk control (under a variety of risk functions). See also similar follow-up work by Laufer-Goldshtein et al. (2023) and Quach et al. (2023).

Our proposed PROSAC framework departs from these existing approaches in that it seeks to offer risk guarantees for a machine learning model in the presence of an adversarial attack. Via the use of a GP-UCB algorithm, it seeks to ascertain the risk of a machine learning model in the presence of the worst-case attacker hyper-parameter configuration.

Problem Statement

Adversarial Attack

We consider how to certify the robustness of a (classification) machine learning model against specific adversarial attacks. We assume that we have access to a machine learning model $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ that maps features $X \in \mathcal{X}$ onto a (categorical) target $Y \in \mathcal{Y}$ where (X, Y) are drawn from an unknown distribution $\mathcal{D}_{X,Y}$. We also assume that this machine learning model has already been optimized (trained) *a priori* to solve a specific multi-class classification task using a given training set (hence, $\mathcal{Y} = \{1, 2, \ldots, K\}$). We denote the corresponding positive loss function as $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_0^+$.

We consider that the machine learning model $\mathcal{M}$ is attacked by an adversarial attack

$$\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X},$$

that given a pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ – (ideally) converts the original model input $X \in \mathcal{X}$ onto an adversarial one $\tilde{X} \in \mathcal{X}$ as follows:

$$\tilde{X} = \mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q}(X, Y) = X + \arg\max_{\delta \in \mathcal{B}_\epsilon^q} \mathcal{L}(\mathcal{M}(X + \delta), Y), \tag{1}$$

with the intent of maximizing the loss for the given sample $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{B}_\epsilon^q$ is an ℓ_q-norm bounded ball with radius ϵ (where ϵ measures the capability of the attacker, i.e., the attack budget).

However, practically, it is nontrivial to directly obtain the optimal adversarial sample – calculating $\arg\max_{\delta \in \mathcal{B}_\epsilon^q} \mathcal{L}(\mathcal{M}(X + \delta), Y)$ analytically with Eq. (1) is often inaccessible. Most attackers often need to iterate and update based on the original sample, which requires manually setting the number of iterations and iteration step size, etc. Specifically, according to the accessibility of the model information, common attackers are divided into two categories: *white-box* attacks, where the attacker has full access to the machine learning model, including its architecture/parameters/gradients, and *black-box* attacks, where the attacker does not have full access to the machine learning model. In both conditions, we assume that the attacker draws its hyper-parameter configuration (for example, iteration steps) λ from a (finite) set of hyper-parameter configurations Λ, where each hyper-parameter configuration is d-dimensional i.e. $\lambda \in \mathbb{R}^d$.

Moreover, in general, the various attacks are still stochastic. Given fixed attack hyper-parameters λ, the white-box and black-box attacks do not deliver a deterministic perturbation $\tilde{\delta}$ given fixed sample (X, Y) but rather random perturbations, because the attacks depend on other random variables. For example, the white-box PGDAttack (Madry et al. 2018) depends on the random initialization. Denoted as Z such remaining randomness in the attack, therefore, the adversarial attacks[1] can be represented as

$$\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q,\lambda} : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{Z} \to \mathcal{X},$$

and the corresponding adversarial sample as:

$$\tilde{X} = \mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q,\lambda}(X, Y, Z), \tag{2}$$

Model Safety

Given an adversarial attack $\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q,\lambda}$, we can consequently characterize the safety of a machine learning model using two quantities: the *adversarial risk* and the *max adversarial risk*. We define the adversarial (population) risk induced by an attack $\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q,\lambda}$ on a model $\mathcal{M}$ as follows:

$$\mathcal{R}_{\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q,\lambda}}(\mathcal{M}) := \mathbb{E}_{(X,Y,Z)\sim\mathcal{D}_{X,Y}\times\mathcal{D}_Z}\{R_{\mathcal{M}}\}, \tag{3}$$

where

$$R_{\mathcal{M}} = \mathbb{1}[\mathcal{M}(\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q,\lambda}(X, Y, Z)) \neq Y] \cdot \mathbb{1}[\mathcal{M}(X) = Y],$$

and we define the max adversarial (population) risk induced by an attack $\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q,\lambda}$ on a model $\mathcal{M}$ independently of how the attacker chooses its hyper-parameters as follows:

$$\mathcal{R}^\star_{\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q}}(\mathcal{M}) = \max_\lambda \mathcal{R}_{\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q,\lambda}}(\mathcal{M}), \tag{4}$$

where we use the 0-1 loss to measure the per-sample loss[2]. Note that the adversarial (population) risk characterizes the performance of the machine learning model for a specific attack with a given budget / norm and a fixed hyper-parameter configuration, whereas the max adversarial (population) risk characterizes the performance of the machine learning model for an attack with a given budget / norm, independently of how the attacker chooses its hyper-parameter configuration.

Our main goal is to determine whether a machine learning model is safe by establishing whether the max (adversarial) population risk is below some threshold with high probability.

Definition 1 *((α, ζ)-Model Safety) Fix $0 \leq \alpha \leq 1$, $0 \leq \zeta \leq 1$. Then, we say that a machine learning model $\mathcal{M}$ is (α, ζ)-safe under an adversarial attack $\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q,\lambda}$ with fixed budget ϵ and ℓ_q-norm, and for all attack hyper-parameters $\lambda \in \Lambda$, provided that*

$$\mathbb{P}\left(\text{reject } \mathcal{R}^* > \alpha \middle| \, \mathcal{R}^* > \alpha \text{ is true}\right) \leq \zeta. \tag{5}$$

Here, $\mathcal{R}^* \triangleq \mathcal{R}^*_{\mathcal{A}_{\mathcal{M},\mathcal{B}_\epsilon^q,\lambda}}(\mathcal{M})$.

We will see in the following that this entails formulating a hypothesis testing problem where the null hypothesis is associated with a max adversarial risk higher than α. Therefore, (α, ζ)-model safety means that the probability of declaring that the max adversarial risk of a model is less than α when it is in fact higher than α is smaller than ζ, or, more loosely speaking, a model max adversarial risk is less than α with a probability higher than $1 - \zeta$.

[1]Note that, we do not consider the attack budget, ϵ, to be a hyper-parameter since it would not be possible to control the risk where the adversary has the ability to choose any attack budget $\epsilon \in (0, \infty)$. We also do not consider the attack norm to be a hyper-parameter.

[2]This work concentrates primarily on classification problems with 0-1 loss. However, our work readily extends to other losses subject to some modifications.

Certification Procedure

We now describe our proposed certification approach allowing us to establish (α, ζ)- safety of a machine learning model in the presence of an adversarial attack. We will omit the dependency of adversarial risk on the model, the attack, and the attack parameters in order to simplify notation. We will also omit the fact that the attack depends on the model, its budget / norm, and the hyper-parameters.

Procedure

Our procedure is related to, but also departs from, a recent line of research concerning risk control in machine learning models, pursued by Bates et al. (2021); Angelopoulos et al. (2021) and Laufer-Goldshtein et al. (2023) (see also references therein). In particular, Bates et al. (2021); Angelopoulos et al. (2021) and Laufer-Goldshtein et al. (2023) offer a methodology to identify a set of model hyper-parameter configurations that control the (statistical) risk of a machine learning model. However, we are not interested in determining a set of attacker hyper-parameters guaranteeing risk control, but rather in guaranteeing risk control independently of how an attacker chooses the hyper-parameters (since a user cannot control the choice of hyper-parameters).

We fix the machine learning model $\mathcal{M}$, the adversarial attack $\mathcal{A}$, the adversarial attack budget ϵ, and the adversarial attack ℓ_q-norm. We leverage – in line with Bates et al. (2021); Angelopoulos et al. (2021); Laufer-Goldshtein et al. (2023) – access to a calibration set $\mathcal{S} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ (independent of any training set) where the samples (X_i, Y_i) are drawn i.i.d. from the distribution $\mathcal{D}_{X,Y}$ to construct our certification procedure.

Our certification procedure then involves the following sequence of steps:

- First, we set up a hypothesis testing problem where the null hypothesis is $\mathcal{H}_0 : \mathcal{R}^* > \alpha$ or, equivalently, $\mathcal{H}_0 : \exists\, \lambda \in \Lambda, \mathcal{R}_\lambda > \alpha$, where $\mathcal{R}^*$ represents the max adversarial risk in Eq. (4) and $\mathcal{R}_\lambda$ represents the adversarial risk in Eq. (3) that depends on the attacker hyper-parameters $\lambda \in \Lambda$.
- Second, we leverage the calibration set (plus another set with a number of instances / objects characterizing the randomness of the attack) to determine a finite-sample p-value p^* that can be used to reject the null hypothesis $\mathcal{H}_0 : \mathcal{R}^* > \alpha$ or, equivalently, $\mathcal{H}_0 : \exists\, \lambda \in \Lambda, \mathcal{R}_\lambda > \alpha$.
- Finally, we reject the null hypothesis $\mathcal{H}_0 : \mathcal{R}^* > \alpha$ or, equivalently, $\mathcal{H}_0 : \exists\, \lambda \in \Lambda, \mathcal{R}_\lambda > \alpha$ provided that the p-value p^* is less than ζ.

This procedure allows us to establish (α, ζ)- safety of the machine learning model $\mathcal{M}$ in the presence of an adversarial attack $\mathcal{A}$, in accordance with Definition 1.

Proposition 1 *Let p^* be a p-value associated with the hypothesis testing problem where the null hypothesis is $\mathcal{H}_0 : \mathcal{R}^* > \alpha$ or, equivalently, $\mathcal{H}_0 : \exists\, \lambda \in \Lambda, \mathcal{R}_\lambda > \alpha$. It follows immediately that the machine learning model is (α, ζ)- safe, i.e.*

$$\mathbb{P}\left(\text{reject } \mathcal{R}^* > \alpha \;\middle|\; \mathcal{R}^* > \alpha \text{ is true}\right) \leq \zeta, \tag{6}$$

provided that the null hypothesis is rejected if and only if $p^* \leq \zeta$.

We next show how to derive a p-value for our hypothesis testing problem where $\mathcal{H}_0 : \exists\, \lambda \in \Lambda, \mathcal{R}_\lambda > \alpha$ from the p-values for the hypothesis testing problems where $\mathcal{H}_0 : \mathcal{R}_\lambda > \alpha$, for all $\lambda \in \Lambda$ (see also Laufer-Goldshtein et al. 2023). [3]

Theorem 2 *If $p(\lambda)$ is a p-value associated with the null $\mathcal{H}_0 : \mathcal{R}_\lambda > \alpha$ then $p^* = \max_{\lambda \in \Lambda} p(\lambda)$ is a p-value associated with the null hypothesis $\mathcal{H}_0 : \exists\, \lambda \in \Lambda, \mathcal{R}_\lambda > \alpha$.*

Therefore, building on Theorem 2, we can immediately determine a p-value for our hypothesis testing problem.

Theorem 3 *A (super-uniform) p-value associated with the null hypothesis $\mathcal{H}_0 : \exists\, \lambda \in \Lambda, \mathcal{R}_\lambda > \alpha$ is given by (Bates et al. 2021; Angelopoulos et al. 2021):*

$$p^* = \max_{\lambda \in \Lambda} \min\Big\{ \exp\Big(-n \cdot h_1(\hat{\mathcal{R}}(\lambda); \alpha)\Big), \; e \cdot \mathbb{P}\left(\mathsf{Bin}(n, \alpha) \leq \left\lceil n \cdot \hat{\mathcal{R}}(\lambda) \right\rceil\right) \Big\}, \tag{7}$$

where $\hat{\mathcal{R}}(\lambda)$ represents the adversarial empirical risk induced by the attack $\mathcal{A}_\lambda$ on model $\mathcal{M}$ given a specific hyper-parameter configuration $\lambda \in \Lambda$ i.e.

$$\hat{\mathcal{R}}(\lambda, \mathcal{S}, \mathcal{Z}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\mathcal{M}(\mathcal{A}_\lambda(X_i, Y_i, Z_i)) \neq Y_i] \cdot \mathbb{1}[\mathcal{M}(X_i) = Y_i], \tag{8}$$

where $\mathcal{S} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is the set containing the calibration data, $\mathcal{Z} = \{Z_1, \ldots, Z_n\}$ is a set containing a series of random objects that capture the randomness of the attack, and $h_1(a, b) = a \cdot \log(a/b) + (1 - a) \cdot \log((1 - a)/(1 - b))$.

Algorithm and Associated Guarantees

Our procedure to establish (α, ζ)-safety of a machine learning model $\mathcal{M}$ in the presence of an adversarial attack $\mathcal{A}$, in accordance with Definition 1, relies on the ability to approximate the p-value associated with the null hypothesis $\mathcal{H}_0 : \exists\, \lambda \in \Lambda, \mathcal{R}_\lambda > \alpha$ as per Theorem 3. However, this involves solving a complex optimization problem concerning the maximization of a function (a Hoeffding-Bentkus p-value (Bates et al. 2021; Angelopoulos et al. 2021)) over the set of attacker hyper-parameter configurations. We therefore propose to adopt a Bayesian optimization (BO) procedure, based on the established Gaussian Process Upper Confidence Bound (GP-UCB) algorithm (Srinivas et al. 2010), which can be used to search effectively over the set of hyper-parameter configurations of the attack in order to identify the configuration leading to the highest p-value[4].

[3] Note the difference between the hypothesis testing problems. The problem with the null $\mathcal{H}_0 : \exists\, \lambda \in \Lambda, \mathcal{R}_\lambda > \alpha$ tests whether the max adversarial risk is above α independently of the choice of hyper-parameters associated with the attack, whereas the hypothesis testing problem with the null $\mathcal{H}_0 : \mathcal{R}_\lambda > \alpha$ tests whether the risk is above α for a particular choice of hyper-parameters associated with the attack.

[4] We require a sample-efficient optimization method since the evaluation of the p-value involves computation of the empirical risk

Algorithm 1: GP-UCB for hyperparameter optimization

Input: Hyper-parameter configuration grid Λ. Gaussian Process prior mean $\mu_0 = 0$; Gaussian Process prior covariance $\sigma_0 = k$ where k corresponds to the kernel function.
for t = 1, 2, 3 ... T **do**
  Compute $\lambda_t = \arg\max_{\lambda \in \Lambda} \mu_{t-1}(\lambda) + \beta_t \sigma_{t-1}(\lambda)$.
  Compute $\hat{p}_t = p(\lambda_t) + \nu_t$
  Perform Bayesian update to obtain new GP mean μ_t and covariance σ_t using the sampled points $(\lambda_t, \hat{p}_t)$.
end for
return $\hat{p}_T = 1/T \sum_{t=1}^{T} \hat{p}_t$

Algorithm 1 summarizes the GP-UCB procedure used to search for the attack hyperparameters that solve Eq. (7). The algorithm first ingests the attacker hyper-parameter grid configuration, the Gaussian process mean function, and the Gaussian process prior covariance (kernel) function. We select the kernel to be Matern kernel (Genton 2001). At round t, the algorithm determines the hyper-parameter configuration $\lambda_t \in \Lambda$ that maximizes the upper confidence bound. The algorithm then determines a p-value $\hat{p}_t$ corresponding to the sum of $p(\lambda_t)$ plus some i.i.d. zero-mean Gaussian noise ν_t (where $p(\lambda_t)$ is derived from Eq. (7), and the algorithm performs Bayesian updating to obtain a new GP mean function μ_t and covariance function σ_t. The algorithm finally delivers the p-value estimate after T rounds. We choose β_t to be 0.1 with hyper-parameter search from β={0.01,0.1,1.0}.

The following theorem shows that we can establish (α, ζ)-safety of the machine learning model $\mathcal{M}$ in the presence of an adversarial attack $\mathcal{A}$ (in accordance with Definition 1) by relying on Algorithm 1. In particular, in view of the fact that the GP-UCB procedure in Algorithm 1 delivers a p-value estimate that is close to the true p-value with probability $(1 - \delta)$, where $0 < \delta < 1$ (see guarantees in (Srinivas et al. 2010)), the hypothesis testing procedure underlying Theorem 4 compares the GP-UCB p-value estimate $\hat{p}_T$ to a more conservative threshold $\zeta' < \zeta$, rather than ζ, where

$$\zeta' = \zeta - \mathcal{O}\left(B\sqrt{\gamma_T/T} + \sqrt{\gamma_T\left(\gamma_T + \log(1/\delta)\right)/T}\right) - \delta, \tag{9}$$

where the value B bounds the smoothness of the p-value function , γ_T corresponds to the maximum information gain at round T , and T is the number of GP-UCB rounds. According to Eq. (9), with probability nearly 1, we can expect $\zeta' \to \zeta$ with $T \to \infty$. The Supplementary Material demonstrates that this more conservative testing procedure is sufficient to retain the (α, ζ) safety guarantees in Definition 1.

Theorem 4 ((α, ζ)-Safe Model with GP-UCB) *Fix $0 \leq \alpha \leq 1$, $0 \leq \zeta \leq 1$, $0 \leq \delta \leq 1$ (with $\delta < \zeta$), the machine learning model $\mathcal{M}$, the adversarial attack $\mathcal{A}$ (its budget ϵ and ℓ_q-norm). Assume that one rejects the null hypothesis provided that GP-UCB p-value estimate $\hat{p}_T$ is below ζ' in Eq.* (9)*. Then, one can guarantee that the machine learning model $\mathcal{M}$ is (α, ζ)-safe under an adversarial attack $\mathcal{A}$ for all attack hyper-parameters, i.e.,*

$$\mathbb{P}\left(\text{reject } \mathcal{R}^* > \alpha \;\middle|\; \mathcal{R}^* > \alpha \text{ is true}\right) \leq \zeta. \tag{10}$$

of the model subject to the attack for each individual attack hyperparameter configuration; this is very time-consuming for complex models used in our experiments.

Experiments

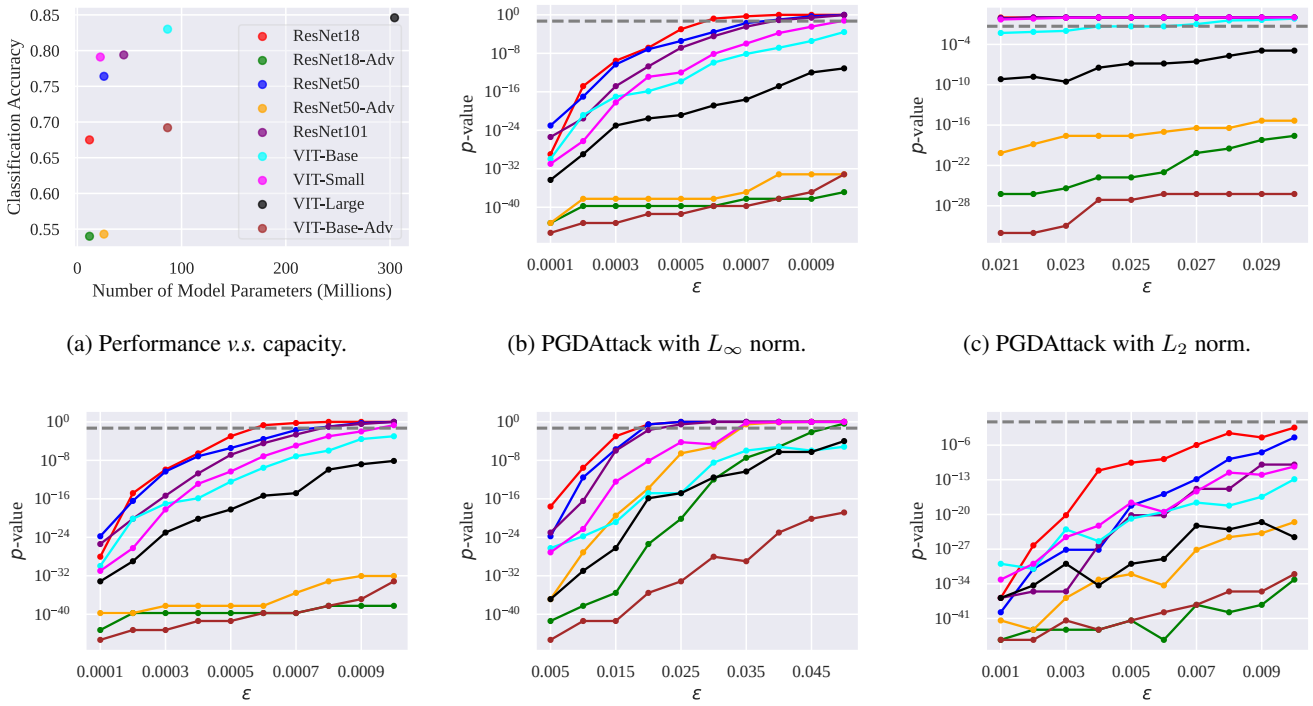
In this section, we conduct extensive experiments with PROSAC to certify the performance of various state-of-the-art vision models in the presence of various adversarial attacks; how the framework recovers existing trends relating to the robustness of different models against different adversarial attacks; and how the framework also suggests new trends relating to state-of-the-art model robustness against attacks.

Experimental Settings

Datasets We will consider primarily classification tasks on the ImageNet-1k dataset (Deng et al. 2009). We follow the common experimental setting in black-box adversarial attacks, using 1,000 images from ImageNet-1k (Andriushchenko et al. 2020; Ilyas et al. 2018) to apply our proposed certification procedure. In particular, we take our calibration set to correspond to this dataset.

Models We use two representative state-of-the-art models in computer vision in our experiments, vision transformer (ViT) (Dosovitskiy et al. 2020) and ResNet (He et al. 2016). We first consider supervised pre-trained models on ImageNet-1k: We use small, base and large models for both ResNet and ViT. Specifically, we test *ViT-Small*, *ViT-Base* and *ViT-Large* for ViT, and *ResNet-18*, *ResNet-50* and *ResNet-101* for ResNet. To certify, we also consider adversarially pre-trained (Adv) models on ImageNet-1k. We adopt the adversarial training models provided by RobustBench, in particular, *ResNet18-Adv* (Salman et al. 2020), *ResNet50-Adv* (Wong, Rice, and Kolter 2020) and *ViT-Base-Adv* (Mo et al. 2022). In summary, we test 9 pre-trained models in our experiments, also see Fig. 1(a) for their classification performance and number of parameters.

Attackers We consider both white-box and black-box attackers in our experiments - PGDAttack (Madry et al. 2018) and MomentumAttack (Dong et al. 2018) for white-box, and BanditAttack (Ilyas, Engstrom, and Madry 2018) and GenAttack (Alzantot et al. 2019) for black-box. We default to L_∞ perturbation. For comparison, we also consider PGDAttack with L_2 perturbation. The hyperparameters of each attacker were carefully selected to explore a wide range of configurations. Specifically, detailed range/values of hyperparameters for each attackers are shown in APPENDIX A. *As our focus is not to investigate and compare different adversarial attackers, for better clarity, we leave it to the interested readers more details and mechanisms about each attacker.* We follow the implementation of different attackers in `advertorch` (Ding, Wang, and Jin 2019) in our experiments. All hyperparameters can be corresponded back to. We set $\alpha = 0.10$ and $\zeta = 0.05$ in our safety certification procedure, per Definition 1.



(a) Performance *v.s.* capacity.

(b) PGDAttack with L_∞ norm.

(c) PGDAttack with L_2 norm.

(d) MomentumAttack with L_∞ norm.

(e) GenAttack with L_∞ norm.

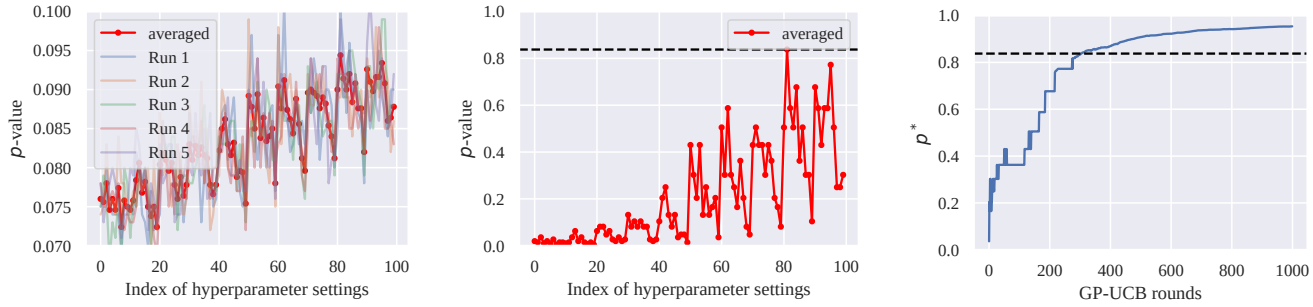
(f) BanditAttack with L_∞ norm.

Figure 1: Model certification *w.r.t* different attacking budgets ϵ.

Model Safety Certification under Variable Attack Budget

In this section, we certify the safety of nine different models against adversarial attacks, focusing on five different adversarial attack strategies with the hyperparameter range defined earlier, as illustrated in Fig. 1. As previously discussed, the attack budget ϵ controls the degree of perturbation applied to the input. We can generally observe that as the attack budget ϵ increases, the safety of various models decreases (indicated by a rise in the p-value). This further explains why we do not consider attack budget ϵ as a interested hyperparameter - since it has a clear linear relationship with model safety. Additionally, we highlight the following observations:

- ***Large Models vs. Small Models***: It is commonly agreed that although larger models are capable of capturing complex patterns and generalizing more effectively, they are often more susceptible to overfitting, particularly when trained with limited datasets. This leads to a natural hypothesis that larger models, with more parameters, might be more vulnerable to adversarial attacks, as they could be easily deceived by perturbations that exploit these overfitted details. However, contrary to this hypothesis, we find that smaller models tend to exhibit lower safety against adversarial attacks across both ResNet and ViT model families (i.e., $p^*_{\text{ResNet18}} > p^*_{\text{ResNet50}} > p^*_{\text{ResNet101}}$, $p^*_{\text{ViT-Small}} > p^*_{\text{ViT-Base}} > p^*_{\text{ViT-Large}}$). This counterintuitive result suggests that the relationship between model size and adversarial robustness is more complex than initially assumed, and it highlights the need for further exploration and analysis in this area.
- ***ResNet vs. ViT***: In addition to the intriguing relationship between model size and model security, we observe a notable contrast between ResNet-family models and ViT-family models - that the ViT-family consistently demonstrates superior adversarial safety compared to the ResNet-family. While a detailed investigation of this phenomenon is beyond the scope of this paper, one key difference lies in their architectural approaches. Unlike ResNet, which relies on local convolutions for feature extraction, Vision Transformers (ViT) utilize a self-attention mechanism. This mechanism allows ViTs to capture long-range dependencies and global context more effectively. By focusing on global interactions between patches of the input image, ViTs may potentially counteract adversarial perturbations that exploit local features.
- ***Adversarial Training***: We also evaluate the performance of existing adversarial training methods. As expected, models incorporating adversarial training techniques (ResNet18-Adv, ResNet50-Adv and ViT-Base-Adv) generally exhibit improved adversarial safety. However, it is important to note that these models also experience a significant impact on their classification performance, see Fig. 1(a). The trade-off between enhanced adversarial resilience and reduced classification accuracy must



(a) Grid search of $\hat{R}$. (b) Estimated p-value with averaged $\hat{R}$. (c) GP-UCB estimation.

Figure 2: Performance of GP-UCB estimation as the rounds change.

be carefully considered when implementing adversarial training.

- ***Different Attackers***: Finally, we conducted a comparison of different attackers. Note that, since the attack budget ϵ heavily influences the attacker's capability, we only provide a rough summary based on significant observations available. It is noted that, compared to black-box attacks, white-box attacks are more harmful even with lower attack budgets. We conjecture that, in white-box attacks, the attacker has complete access to the model's internal parameters and gradient information, enabling the design of more effective adversarial samples.

Approximation Performance of GP-UCB

Before the primary model validation process, we first empirically validated the performance of the GP-UCB method in estimating the p-value. Specifically, this section focuses on evaluating the robustness of the ResNet18 model when subjected to the GenAttack. For comparison, in addition to the proposed GP-UCB method for estimating the p-value, a straightforward grid search approach was also employed. To ensure computational efficiency, we only considered two variable parameters , resulting in a total of 100 hyperparameter combinations.

As shown in Fig. 2(a), we present the *attack success rate* $\hat{R}$ corresponding to these 100 parameter combinations. The final attack success rate ('averaged') was obtained by averaging the results from five random repetitions ('Run 1' - 'Run 5'). Based on Eq. (7) and Eq. (8), we calculated the p-values corresponding to each hyperparameter combination, shown in Fig. 2(b). Notably, we marked the maximum value (p^*) with a horizontal line. For comparison, Fig. 2(c) illustrates the p^* values estimated by the GP-UCB algorithm at different rounds. We observed that, compared to grid search, after a certain number of iterations, the GP-UCB method consistently provides a more conservative p-value ($p^*_{gpu-cb} > p^*_{grid-search}$), which helps further reduce the false positive rate in the model robustness assessment.

Conclusions

We have proposed PROSAC, a new approach to certify the performance of a machine learning model in the presence of an adversarial attack, with population level adversarial risk guarantees. PROSAC builds on recent work on distribution-free risk quantification approaches, offering an instrument to ascertain whether a model is likely to be safe in the presence of an adversarial attack, independently of how the attacker chooses the attack hyperparameters. We show via experiments that PROSAC is able to certify various state-of-the-art models, leading to results that are in line with existing results in the literature.

The technical framework developed here is likely to be of high relevance to AI regulation, such as the EU's AI Act, which requires providers of certain AI systems to ensure that their systems are resilient to adversarial attacks. Beyond its utility as a certification instrument, our framework also suggests that a number of areas in adversarial robustness may merit further attention. First, PROSAC has shown that large ViT models appear to be more adversarially robust than smaller models, pointing to new directions for research on the relationship between the capacity of a ViT and its adversarial robustness. Second, CLIP-ViT models appear to be more vulnerable to adversarial attacks than supervised trained ViTs, raising questions about how to use self-supervised models better in improving adversarial robustness of downstream tasks.

We highlight the importance of measuring the resilience of ML models against adversarial use and misuse, not only in order to comply with the EU AI Act, but also because general duties to act with due care and not negligently presuppose that providers of AI systems have clarity about the safety of their models. With our approach to certifying robustness, we seek to provide a tool to reduce these societal risks. In addition, in future work, we aim to further investigate the security challenges of models in multimodal learning, particularly in the presence of noise and missing modalities (Feng, Tzimiropoulos, and Patras 2022, 2024; Zhi et al. 2024a,b).

Acknowledgements

We acknowledge support from Leverhulme Trust via research grant RPG-2022-198.

References

Alzantot, M.; Sharma, Y.; Chakraborty, S.; Zhang, H.; Hsieh, C.-J.; and Srivastava, M. B. 2019. Genattack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the genetic and evolutionary computation conference*, 1111–1119.

Andriushchenko, M.; Croce, F.; Flammarion, N.; and Hein, M. 2020. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, 484–501. Springer.

Angelopoulos, A.; and Bates, S. 2023. Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning*, 31.

Angelopoulos, A. N.; Bates, S.; Candès, E. J.; Jordan, M. I.; and Lei, L. 2021. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Bates, S.; Angelopoulos, A.; Lei, L.; Malik, J.; and Jordan, M. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6): 1–34.

Black, E.; Yeom, S.; and Fredrikson, M. 2020. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 111–121.

Bruna, J.; Szegedy, C.; Sutskever, I.; Goodfellow, I.; Zaremba, W.; Fergus, R.; and Erhan, D. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Carlini, N.; Tramer, F.; Dvijotham, K. D.; Rice, L.; Sun, M.; and Kolter, J. Z. 2023. (Certified!!) Adversarial Robustness for Free! In *The Eleventh International Conference on Learning Representations*.

Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; and Mukhopadhyay, D. 2018. Adversarial Attacks and Defences: A Survey. *arXiv:1810.00069*.

Chugg, B.; Cortes-Gomez, S.; Wilder, B.; and Ramdas, A. 2023. Auditing Fairness by Betting. *arXiv preprint arXiv:2305.17570*.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, 1310–1320. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Ding, G. W.; Wang, L.; and Jin, X. 2019. AdverTorch v0.1: An Adversarial Robustness Toolbox based on PyTorch. *arXiv preprint arXiv:1902.07623*.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Feng, C.; Tzimiropoulos, G.; and Patras, I. 2022. SSR: An Efficient and Robust Framework for Learning with Unknown Label Noise. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press.

Feng, C.; Tzimiropoulos, G.; and Patras, I. 2024. CLIP-Cleaner: Cleaning Noisy Labels with CLIP. In *The 32nd ACM International Conference on Multimedia*.

Genton, M. G. 2001. Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, 2(Dec): 299–312.

Gowal, S.; Dvijotham, K.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Arandjelovic, R.; Mann, T.; and Kohli, P. 2018. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*.

Gowal, S.; Dvijotham, K. D.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Arandjelovic, R.; Mann, T.; and Kohli, P. 2019. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4842–4851.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hein, M.; and Andriushchenko, M. 2017. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in neural information processing systems*, 30.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, 2137–2146. PMLR.

Ilyas, A.; Engstrom, L.; and Madry, A. 2018. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*.

Laufer-Goldshtein, B.; Fisch, A.; Barzilay, R.; and Jaakkola, T. 2023. Efficiently controlling multiple risks with Pareto testing. In *International Conference on Learning Representations*.

Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; and Jana, S. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, 656–672. IEEE.

Leino, K.; Wang, Z.; and Fredrikson, M. 2021. Globally-robust neural networks. In *International Conference on Machine Learning*, 6212–6222. PMLR.

Li, L.; Xie, T.; and Li, B. 2023. Sok: Certified robustness for deep neural networks. In *2023 IEEE Symposium on Security and Privacy (SP)*, 1289–1310. IEEE.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Mo, Y.; Wu, D.; Wang, Y.; Guo, Y.; and Wang, Y. 2022. When adversarial training meets vision transformers: Recipes from training to architecture. *Advances in Neural Information Processing Systems*, 35: 18599–18611.

Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J. H.; Jaakkola, T. S.; and Barzilay, R. 2023. Conformal Language Modeling. *arXiv preprint arXiv:2306.10193*.

Salman, H.; Ilyas, A.; Engstrom, L.; Kapoor, A.; and Madry, A. 2020. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33: 3533–3545.

Salman, H.; Li, J.; Razenshteyn, I.; Zhang, P.; Zhang, H.; Bubeck, S.; and Yang, G. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32.

Si, N.; Murthy, K.; Blanchet, J.; and Nguyen, V. A. 2021. Testing group fairness via optimal transport projections. In *International Conference on Machine Learning*, 9649–9659. PMLR.

Singla, S.; and Feizi, S. 2021. Skew orthogonal convolutions. In *International Conference on Machine Learning*, 9756–9766. PMLR.

Srinivas, N.; Krause, A.; Kakade, S.; and Seeger, M. 2010. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 1015–1022.

Taskesen, B.; Blanchet, J.; Kuhn, D.; and Nguyen, V. A. 2021. A statistical test for probabilistic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 648–665.

Trockman, A.; and Kolter, J. Z. 2020. Orthogonalizing Convolutional Layers with the Cayley Transform. In *International Conference on Learning Representations*.

Tsuzuku, Y.; Sato, I.; and Sugiyama, M. 2018. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, 31.

Wong, E.; and Kolter, Z. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, 5286–5295. PMLR.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*.

Wong, E.; Schmidt, F.; Metzen, J. H.; and Kolter, J. Z. 2018. Scaling provable adversarial defenses. *Advances in Neural Information Processing Systems*, 31.

Xu, X.; Li, L.; and Li, B. 2022. Lot: Layer-wise orthogonal training on improving l2 certified robustness. *Advances in Neural Information Processing Systems*, 35: 18904–18915.

Xue, S.; Yurochkin, M.; and Sun, Y. 2020. Auditing ml models for individual bias and unfairness. In *International Conference on Artificial Intelligence and Statistics*, 4552–4562. PMLR.

Zhang, B.; Jiang, D.; He, D.; and Wang, L. 2021. Boosting the Certified Robustness of L-infinity Distance Nets. In *International Conference on Learning Representations*.

Zhang, H.; Chen, H.; Xiao, C.; Gowal, S.; Stanforth, R.; Li, B.; Boning, D.; and Hsieh, C.-J. 2019. Towards Stable and Efficient Training of Verifiably Robust Neural Networks. In *International Conference on Learning Representations*.

Zhi, Z.; Liu, Z.; Elbadawi, M.; Daneshmend, A.; Orlu, M.; Basit, A.; Demosthenous, A.; and Rodrigues, M. 2024a. Borrowing Treasures from Neighbors: In-Context Learning for Multimodal Learning with Missing Modalities and Data Scarcity. *arXiv preprint arXiv:2403.09428*.

Zhi, Z.; Liu, Z.; Wu, Q.; and Rodrigues, M. R. D. 2024b. Wasserstein Modality Alignment Makes Your Multimodal Transformer More Robust. In *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*.