



MIT Open Access Articles

Tight certificates of adversarial robustness for randomly smoothed classifiers

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

| | |
|---------------------|---|
| Citation | Lee, Guang-He et al. "Tight certificates of adversarial robustness for randomly smoothed classifiers." 32nd Conference on Neural Information Processing Systems, December 2018, Montreal, Canada, Neural Information Processing Systems, 2018. © 2018 The Author(s) |
| As Published | https://papers.nips.cc/paper/2019/hash/fa2e8c4385712f9a1d24c363a2cbe5b8-Abstract.html |
| Version | Final published version |
| Citable link | https://hdl.handle.net/1721.1/129439 |
| Terms of Use | Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use. |

Tight Certificates of Adversarial Robustness for Randomly Smoothed Classifiers

Guang-He Lee¹, Yang Yuan^{1,2}, Shiyu Chang³, Tommi S. Jaakkola¹

¹MIT Computer Science and Artificial Intelligence Lab

²Institute for Interdisciplinary Information Sciences, Tsinghua University

³MIT-IBM Watson AI Lab

{guanghe, yangyuan, tommi}@csail.mit.edu, shiyu.chang@ibm.com

Abstract

Strong theoretical guarantees of robustness can be given for ensembles of classifiers generated by input randomization. Specifically, an ℓ_2 bounded adversary cannot alter the ensemble prediction generated by an additive isotropic Gaussian noise, where the radius for the adversary depends on both the variance of the distribution as well as the ensemble margin at the point of interest. We build on and considerably expand this work across broad classes of distributions. In particular, we offer adversarial robustness guarantees and associated algorithms for the discrete case where the adversary is ℓ_0 bounded. Moreover, we exemplify how the guarantees can be tightened with specific assumptions about the function class of the classifier such as a decision tree. We empirically illustrate these results with and without functional restrictions across image and molecule datasets.¹

1 Introduction

Many powerful classifiers lack robustness in the sense that a slight, potentially unnoticeable manipulation of the input features, e.g., by an adversary, can cause the classifier to change its prediction [15]. The effect is clearly undesirable in decision critical applications. Indeed, a lot of recent work has gone into analyzing such failures together with providing certificates of robustness.

Robustness can be defined with respect to a variety of metrics that bound the magnitude or the type of adversarial manipulation. The most common approach to searching for violations is by finding an adversarial example within a small neighborhood of the example in question, e.g., using gradient-based algorithms [13, 15, 26]. The downside of such approaches is that failure to discover an adversarial example does not mean that another technique could not find one. For this reason, a recent line of work has instead focused on certificates of robustness, i.e., guarantees that ensure, for specific classes of methods, that no adversarial examples exist within a certified region. Unfortunately, obtaining exact guarantees can be computationally intractable [20, 25, 36], and guarantees that scale to realistic architectures have remained somewhat conservative [7, 27, 38, 39, 42].

Ensemble classifiers have recently been shown to yield strong guarantees of robustness [6]. The ensembles, in this case, are simply induced from randomly perturbing the input to a base classifier. The guarantees state that, given an additive isotropic Gaussian noise on the input example, an adversary cannot alter the prediction of the corresponding ensemble within an ℓ_2 radius, where the radius depends on the noise variance as well as the ensemble margin at the given point [6].

In this work, we substantially extend robustness certificates for such noise-induced ensembles. We provide guarantees for alternative metrics and noise distributions (e.g., uniform), develop a stratified

¹Project page: http://people.csail.mit.edu/guanghe/randomized_smoothing.

likelihood ratio analysis that allows us to provide certificates of robustness over discrete spaces with respect to ℓ_0 distance, which are *tight* and *applicable* to any measurable classifiers. We also introduce scalable algorithms for computing the certificates. The guarantees can be further tightened by introducing additional assumptions about the family of classifiers. We illustrate this in the context of ensembles derived from decision trees. Empirically, our ensemble classifiers yield the state-of-the-art certified guarantees with respect to ℓ_0 bounded adversaries across image and molecule datasets in comparison to the previous methods adapted from continuous spaces.

2 Related Work

In a classification setting, the role of robustness certificates is to guarantee a constant classification within a local region; a certificate is always sufficient to claim robustness. When a certificate is both sufficient and necessary, it is called an exact certificate. For example, the exact ℓ_2 certificate of a linear classifier is the ℓ_2 distance between the classifier and a given point. Below we focus the discussions on the recent development of robustness guarantees for deep networks.

Most of the exact methods are derived on piecewise linear networks, defined as any network architectures with piecewise linear activation functions. Such class of networks has a mix integer-linear representation [22], which allows the usage of mix integer-linear programming [4, 9, 14, 25, 36] or satisfiability modulo theories [3, 12, 20, 33] to find the exact adversary under an ℓ_q radius. However, the exact method is in general NP-complete, and thus does not scale to large problems [36].

A certificate that only holds a sufficient condition is conservative but can be more scalable than exact methods. Such guarantees may be derived as a linear program [39, 40], a semidefinite program [30, 31], or a dual optimization problem [10, 11] through relaxation. Alternative approaches conduct layer-wise relaxations of feasible neuron values to derive the certificates [16, 27, 34, 38, 42]. Unfortunately, there is no empirical evidence of an effective certificate from the above methods in large scale problems. This does not entail that the certificates are not tight enough in practice; it might also be attributed to the fact that it is challenging to obtain a robust network in a large scale setting.

Recent works propose a new modeling scheme that ensembles a classifier by input randomization [2, 24], mostly done via an additive isotropic Gaussian noise. Lecuyer et al. [21] first propose a certificate based on differential privacy, which is improved by Li et al. [23] using Rényi divergence. Cohen et al. [6] proceed with the analysis by proving the *tight* certificate with respect to *all the measurable classifiers* based on the Neyman-Pearson Lemma [28], which yields the state-of-the-art provably robust classifier. However, the tight certificate is tailored to an isotropic Gaussian distribution and ℓ_2 metric, while we generalize the result across broad classes of distributions and metrics. In addition, we show that such tight guarantee can be tightened with assumptions about the classifier.

Our method of certification also yields the first tight and actionable ℓ_0 robustness certificates in discrete domains (cf. continuous domains where an adversary is easy to find [15]). Robustness guarantees in discrete domains are combinatorial in nature and thus challenging to obtain. Indeed, even for simple binary vectors, verifying robustness requires checking an exponential number of predictions for any black-box model.²

3 Certification Methodology

Given an input $\mathbf{x} \in \mathcal{X}$, a randomization scheme ϕ assigns a probability mass/density $\Pr(\phi(\mathbf{x}) = \mathbf{z})$ for each randomized outcome $\mathbf{z} \in \mathcal{X}$. We can define a probabilistic classifier either by specifying the associated conditional distribution $\mathbb{P}(y|\mathbf{x})$ for a class $y \in \mathcal{Y}$ or by viewing it as a random function $f(\mathbf{x})$ where the randomness in the output is independent for each \mathbf{x} . We compose the randomization scheme ϕ with a classifier f to get a randomly smoothed classifier $\mathbb{E}_\phi[\mathbb{P}(y|\phi(\mathbf{x}))]$, where the probability for outputting a class $y \in \mathcal{Y}$ is denoted as $\Pr(f(\phi(\mathbf{x})) = y)$ and abbreviated as p , whenever f, ϕ, \mathbf{x} and y are clear from the context. Under this setting, we first develop our framework for tight robustness certificates in §3.1, exemplify the framework in §3.2-3.4, and illustrate how the guarantees can be refined with further assumption in §3.5-3.6. We defer all the proofs to Appendix A.

²We are aware of two concurrent works also yielding certificates in discrete domain [18, 19].

3.1 A Framework for Tight Certificates of Robustness

In this section, we develop our framework for deriving tight certificates of robustness for randomly smoothed classifiers, which will be instantiated in the following sections.

Point-wise Certificate. Given p , we first identify a tight lower bound on the probability score $\Pr(f(\phi(\bar{\mathbf{x}})) = y)$ for another (neighboring) point $\bar{\mathbf{x}} \in \mathcal{X}$. Here we denote the set of measurable classifiers with respect to ϕ as \mathcal{F} . Without any additional assumptions on f , a lower bound can be found by the minimization problem:

$$\rho_{\mathbf{x}, \bar{\mathbf{x}}}(p) \triangleq \min_{\bar{f} \in \mathcal{F}: \Pr(\bar{f}(\phi(\mathbf{x})) = y) = p} \Pr(\bar{f}(\phi(\bar{\mathbf{x}})) = y) \leq \Pr(f(\phi(\bar{\mathbf{x}})) = y). \quad (1)$$

Note that bound is tight since f satisfies the constraint.

Regional Certificate. We can extend the point-wise certificate $\rho_{\mathbf{x}, \bar{\mathbf{x}}}(p)$ to a regional certificate by examining the worst case $\bar{\mathbf{x}}$ over the neighboring region around \mathbf{x} . Formally, given an ℓ_q metric $\|\cdot\|_q$, the neighborhood around \mathbf{x} with radius r is defined as $\mathcal{B}_{r,q}(\mathbf{x}) \triangleq \{\bar{\mathbf{x}} \in \mathcal{X} : \|\mathbf{x} - \bar{\mathbf{x}}\|_q \leq r\}$. Assuming $p = \Pr(f(\phi(\mathbf{x})) = y) > 0.5$ for a $y \in \mathcal{Y}$, a robustness certificate on the ℓ_q radius can be found by

$$R(\mathbf{x}, p, q) \triangleq \sup r, \text{ s.t. } \min_{\bar{\mathbf{x}} \in \mathcal{B}_{r,q}(\mathbf{x})} \rho_{\mathbf{x}, \bar{\mathbf{x}}}(p) > 0.5. \quad (2)$$

Essentially, the certificate $R(\mathbf{x}, p, q)$ entails the following robustness guarantee:

$$\forall \bar{\mathbf{x}} \in \mathcal{X} : \|\mathbf{x} - \bar{\mathbf{x}}\|_q < R(\mathbf{x}, p, q), \text{ we have } \Pr(f(\phi(\bar{\mathbf{x}})) = y) > 0.5. \quad (3)$$

When the maximum can be attained in Eq. (2) (which will be the case in ℓ_0 norm), the above $<$ can be replaced with \leq . Note that here we assume $\Pr(f(\phi(\mathbf{x})) = y) > 0.5$ and ignore the case that $0.5 \geq \Pr(f(\phi(\bar{\mathbf{x}})) = y) > \max_{y' \neq y} \Pr(f(\phi(\bar{\mathbf{x}})) = y')$. By definition, the certified radius $R(\mathbf{x}, p, q)$ is tight for binary classification, and provides a reasonable sufficient condition to guarantee robustness for $|\mathcal{Y}| > 2$. The tight guarantee for $|\mathcal{Y}| > 2$ will involve the maximum prediction probability over all the remaining classes (see Theorem 1 of [6]). However, when the prediction probability $p = \Pr(f(\phi(\mathbf{x})) = y)$ is intractable to compute and relies on statistical estimation for each class y (e.g., when f is a deep network), the tight guarantee is statistically challenging to obtain. The actual algorithm used by Cohen et al. [6] is also a special case of Eq. (2).

3.2 A Warm-up Example: the Uniform Distribution

To illustrate the framework, we show a simple (but new) scenario when $\mathcal{X} = \mathbb{R}^d$ and ϕ is an additive uniform noise with a parameter $\gamma \in \mathbb{R}_{>0}$:

$$\phi(\mathbf{x})_i = \mathbf{x}_i + \epsilon_i, \epsilon_i \stackrel{i.i.d.}{\sim} \text{Uniform}([- \gamma, \gamma]), \forall i \in \{1, \dots, d\}. \quad (4)$$

Given two points \mathbf{x} and $\bar{\mathbf{x}}$, as illustrated in Fig. 1, we can partition the space \mathbb{R}^d into 4 disjoint regions: $\mathcal{L}_1 = \mathcal{B}_{\gamma, \infty}(\mathbf{x}) \setminus \mathcal{B}_{\gamma, \infty}(\bar{\mathbf{x}})$, $\mathcal{L}_2 = \mathcal{B}_{\gamma, \infty}(\mathbf{x}) \cap \mathcal{B}_{\gamma, \infty}(\bar{\mathbf{x}})$, $\mathcal{L}_3 = \mathcal{B}_{\gamma, \infty}(\bar{\mathbf{x}}) \setminus \mathcal{B}_{\gamma, \infty}(\mathbf{x})$ and $\mathcal{L}_4 = \mathbb{R}^d \setminus (\mathcal{B}_{\gamma, \infty}(\bar{\mathbf{x}}) \cup \mathcal{B}_{\gamma, \infty}(\mathbf{x}))$. Accordingly, $\forall \bar{f} \in \mathcal{F}$, we can rewrite $\Pr(\bar{f}(\phi(\mathbf{x})) = y)$ and $\Pr(\bar{f}(\phi(\bar{\mathbf{x}})) = y)$ as follows:

$$\begin{aligned} \Pr(\bar{f}(\phi(\mathbf{x})) = y) &= \sum_{i=1}^4 \int_{\mathcal{L}_i} \Pr(\phi(\mathbf{x}) = \mathbf{z}) \Pr(\bar{f}(\mathbf{z}) = y) d\mathbf{z} = \sum_{i=1}^4 \pi_i \int_{\mathcal{L}_i} \Pr(\bar{f}(\mathbf{z}) = y) d\mathbf{z}, \\ \Pr(\bar{f}(\phi(\bar{\mathbf{x}})) = y) &= \sum_{i=1}^4 \int_{\mathcal{L}_i} \Pr(\phi(\bar{\mathbf{x}}) = \mathbf{z}) \Pr(\bar{f}(\mathbf{z}) = y) d\mathbf{z} = \sum_{i=1}^4 \bar{\pi}_i \int_{\mathcal{L}_i} \Pr(\bar{f}(\mathbf{z}) = y) d\mathbf{z}, \end{aligned}$$

where $\pi_{1:4} = ((2\gamma)^{-d}, (2\gamma)^{-d}, 0, 0)$, and $\bar{\pi}_{1:4} = (0, (2\gamma)^{-d}, (2\gamma)^{-d}, 0)$. With this representation, it is clear that, in order to solve Eq. (1), we only have to consider the integral behavior of \bar{f} within each region $\mathcal{L}_1, \dots, \mathcal{L}_4$. Concretely, we have:

$$\begin{aligned} \rho_{\mathbf{x}, \bar{\mathbf{x}}}(p) &= \min_{\bar{f} \in \mathcal{F}: \sum_{i=1}^4 \pi_i \int_{\mathcal{L}_i} \Pr(\bar{f}(\mathbf{z}) = y) d\mathbf{z} = p} \sum_{i=1}^4 \bar{\pi}_i \int_{\mathcal{L}_i} \Pr(\bar{f}(\mathbf{z}) = y) d\mathbf{z} \\ &= \min_{\substack{g: \{1,2,3,4\} \rightarrow [0,1], \\ \pi_1 |\mathcal{L}_1| g(1) + \pi_2 |\mathcal{L}_2| g(2) = p}} \bar{\pi}_2 |\mathcal{L}_2| g(2) + \bar{\pi}_3 |\mathcal{L}_3| g(3) = \min_{\substack{g: \{1,2,3,4\} \rightarrow [0,1], \\ \pi_1 |\mathcal{L}_1| g(1) + \pi_2 |\mathcal{L}_2| g(2) = p}} \bar{\pi}_2 |\mathcal{L}_2| g(2), \end{aligned}$$

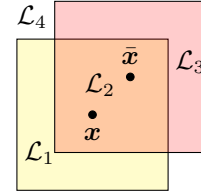


Figure 1: Uniform distributions.

where the second equality filters the components with $\pi_i = 0$ or $\bar{\pi}_i = 0$, and the last equality is due to the fact that $g(3)$ is unconstrained and minimizes the objective when $g(3) = 0$. Since $\pi_2 = \bar{\pi}_2$,

$$\begin{cases} \rho_{\mathbf{x}, \bar{\mathbf{x}}}(p) = 0, & \text{if } 0 \leq p \leq \pi_1 |\mathcal{L}_1| = \Pr(\phi(\mathbf{x}) \in \mathcal{L}_1), \\ \rho_{\mathbf{x}, \bar{\mathbf{x}}}(p) = p - \pi_1 |\mathcal{L}_1|, & \text{if } 1 \geq p > \pi_1 |\mathcal{L}_1| = \Pr(\phi(\mathbf{x}) \in \mathcal{L}_1). \end{cases}$$

To obtain the regional certificate, the minimizers of $\min_{\bar{\mathbf{x}} \in \mathcal{B}_{r,q}(\mathbf{x})} \rho_{\mathbf{x}, \bar{\mathbf{x}}}(p)$ are simply the points that maximize the volume of $\mathcal{L}_1 = \mathcal{B}_1 \setminus \mathcal{B}_2$. Accordingly,

Proposition 1. *If $\phi(\cdot)$ is defined as Eq. (4), we have $R(\mathbf{x}, p, q = 1) = 2p\gamma - \gamma$ and $R(\mathbf{x}, p, q = \infty) = 2\gamma - 2\gamma(1.5 - p)^{1/d}$.*

Discussion. Our goal here was to illustrate how certificates can be computed with the uniform distribution using our technique. However, the certificate radius itself is inadequate in this case. For example, $R(\mathbf{x}, p, q = 1) \leq \gamma$, which arises from the bounded support in the uniform distribution. The derivation nevertheless provides some insights about how one can compute the point-wise certificate $\rho_{\mathbf{x}, \bar{\mathbf{x}}}(p)$. The key step is to partition the space into regions $\mathcal{L}_1, \dots, \mathcal{L}_4$, where the likelihoods $\Pr(\phi(\mathbf{x}) = \mathbf{z})$ and $\Pr(\phi(\bar{\mathbf{x}}) = \mathbf{z})$ are both constant within each region \mathcal{L}_i . The property allows us to substantially reduce the optimization problem in Eq. (1) to finding a single probability value $g(i) \in [0, 1]$ for each region \mathcal{L}_i .

3.3 A General Lemma for Point-wise Certificate

In this section, we generalize the idea in §3.2 to find the point-wise certificate $\rho_{\mathbf{x}, \bar{\mathbf{x}}}(p)$. For each point $\mathbf{z} \in \mathcal{X}$, we define the likelihood ratio $\eta_{\mathbf{x}, \bar{\mathbf{x}}}(\mathbf{z}) \triangleq \Pr(\phi(\mathbf{x}) = \mathbf{z}) / \Pr(\phi(\bar{\mathbf{x}}) = \mathbf{z})$.³ If we can partition \mathcal{X} into n regions $\mathcal{L}_1, \dots, \mathcal{L}_n : \cup_{i=1}^n \mathcal{L}_i = \mathcal{X}$ for some $n \in \mathbb{Z}_{>0}$, such that the likelihood ratio within each region \mathcal{L}_i is a constant $\eta_i \in [0, \infty]$: $\eta_{\mathbf{x}, \bar{\mathbf{x}}}(\mathbf{z}) = \eta_i, \forall \mathbf{z} \in \mathcal{L}_i$, then we can sort the regions such that $\eta_1 \geq \eta_2 \geq \dots \geq \eta_n$. Note that \mathcal{X} can still be uncountable (see the example in §3.2).

Informally, we can always “normalize” \bar{f} so that it predicts a constant probability value $g(i) \in [0, 1]$ within each likelihood ratio region \mathcal{L}_i . This preserves the integral over \mathcal{L}_i and thus over \mathcal{X} , generalizing the scenario in §3.2. Moreover, to minimize $\Pr(f(\phi(\bar{\mathbf{x}})) = y)$ under a fixed budget $\Pr(\bar{f}(\phi(\mathbf{x})) = y)$, as in Eq. (1), it is advantageous to set $\bar{f}(\mathbf{z})$ to y in regions with high likelihood ratio. These arguments suggest a greedy algorithm for solving Eq. (1) by iteratively assigning $f(\mathbf{z}) = y, \forall \mathbf{z} \in \mathcal{L}_i$ for $i \in (1, 2, \dots)$ until the budget constraint is met. Formally,

Lemma 2. *$\forall \mathbf{x}, \bar{\mathbf{x}} \in \mathcal{X}, p \in [0, 1]$, let $H^* \triangleq \min_{H \in \{1, \dots, n\} : \sum_{i=1}^H \Pr(\phi(\mathbf{x}) \in \mathcal{L}_i) \geq p} H$, then $\eta_{H^*} > 0$, any f^* satisfying Eq. (5) is a minimizer of Eq. (1),*

$$\forall i \in \{1, 2, \dots, n\}, \forall \mathbf{z} \in \mathcal{L}_i, \Pr(f^*(\mathbf{z}) = y) = \begin{cases} 1, & \text{if } i < H^*, \\ \frac{p - \sum_{i=1}^{H^*-1} \Pr(\phi(\mathbf{x}) \in \mathcal{L}_i)}{\Pr(\phi(\mathbf{x}) \in \mathcal{L}_{H^*})}, & \text{if } i = H^*, \\ 0, & \text{if } i > H^*. \end{cases} \quad (5)$$

and $\rho_{\mathbf{x}, \bar{\mathbf{x}}}(p) = \sum_{i=1}^{H^*-1} \Pr(\phi(\bar{\mathbf{x}}) \in \mathcal{L}_i) + (p - \sum_{i=1}^{H^*-1} \Pr(\phi(\mathbf{x}) \in \mathcal{L}_i)) / \eta_{H^*}$

We remark that Eq. (1) and Lemma 2 can be interpreted as a likelihood ratio testing [28], by casting $\Pr(\phi(\mathbf{x}) = \mathbf{z})$ and $\Pr(\phi(\bar{\mathbf{x}}) = \mathbf{z})$ as likelihoods for two hypothesis with the significance level p . We refer the readers to [37] to see a similar Lemma derived under the language of hypothesis testing.

Remark 3. *$\rho_{\mathbf{x}, \bar{\mathbf{x}}}(p)$ is an increasing continuous function of p ; if $\eta_1 < \infty$, $\rho_{\mathbf{x}, \bar{\mathbf{x}}}(p)$ is a strictly increasing continuous function of p ; if $\eta_1 < \infty$ and $\eta_n > 0$, $\rho_{\mathbf{x}, \bar{\mathbf{x}}} : [0, 1] \rightarrow [0, 1]$ is a bijection.*

Remark 3 will be used in §3.4 to derive an efficient algorithm to compute robustness certificates.

Discussion. Given $\mathcal{L}_i, \Pr(\phi(\mathbf{x}) \in \mathcal{L}_i)$, and $\Pr(\phi(\bar{\mathbf{x}}) \in \mathcal{L}_i), \forall i \in [n]$, Lemma 2 provides an $O(n)$ method to compute $\rho_{\mathbf{x}, \bar{\mathbf{x}}}(p)$. For any actual randomization ϕ , the key is to find a partition $\mathcal{L}_1, \dots, \mathcal{L}_n$ such that $\Pr(\phi(\mathbf{x}) \in \mathcal{L}_i)$ and $\Pr(\phi(\bar{\mathbf{x}}) \in \mathcal{L}_i)$ are easy to compute. Having constant likelihoods in each $\mathcal{L}_i : \Pr(\phi(\mathbf{x}) = \mathbf{z}) = \Pr(\phi(\mathbf{x}) = \mathbf{z}'), \forall \mathbf{z}, \mathbf{z}' \in \mathcal{L}_i$ (cf. only having constant likelihood ratio η_i) is a way to simplify $\Pr(\phi(\mathbf{x}) \in \mathcal{L}_i) = |\mathcal{L}_i| \Pr(\phi(\mathbf{x}) = \mathbf{z})$, and similarly for $\Pr(\phi(\bar{\mathbf{x}}) \in \mathcal{L}_i)$.

³If $\Pr(\phi(\bar{\mathbf{x}}) = \mathbf{z}) = \Pr(\phi(\mathbf{x}) = \mathbf{z}) = 0$, $\eta_{\mathbf{x}, \bar{\mathbf{x}}}(\mathbf{z})$ can be defined arbitrarily in $[0, \infty]$ without affecting the solution in Lemma 2.

3.4 A Discrete Distribution for ℓ_0 Robustness

We consider ℓ_0 robustness guarantees in a discrete space $\mathcal{X} = \{0, \frac{1}{K}, \frac{2}{K}, \dots, 1\}^d$ for some $K \in \mathbb{Z}_{>0}$;⁴ we define the following discrete distribution with a parameter $\alpha \in (0, 1)$, independent and identically distributed for each dimension $i \in \{1, 2, \dots, d\}$:

$$\begin{cases} \Pr(\phi(\mathbf{x})_i = \mathbf{x}_i) = \alpha, \\ \Pr(\phi(\mathbf{x})_i = z) = (1 - \alpha)/K \triangleq \beta \in (0, 1/K), \quad \text{if } z \in \{0, \frac{1}{K}, \frac{2}{K}, \dots, 1\} \text{ and } z \neq \mathbf{x}_i. \end{cases} \quad (6)$$

Here $\phi(\cdot)$ can be regarded as a composition of a Bernoulli random variable and a uniform random variable. Due to the symmetry of the randomization with respect to all the configurations of \mathbf{x} , $\bar{\mathbf{x}} \in \mathcal{X}$ such that $\|\mathbf{x} - \bar{\mathbf{x}}\|_0 = r$ (for some $r \in \mathbb{Z}_{\geq 0}$), we have the following Lemma for the equivalence of $\rho_{\mathbf{x}, \bar{\mathbf{x}}}$:

Lemma 4. *If $\phi(\cdot)$ is defined as Eq. (6), given $r \in \mathbb{Z}_{\geq 0}$, define the canonical vectors $\mathbf{x}_C \triangleq (0, 0, \dots, 0)$ and $\bar{\mathbf{x}}_C \triangleq (1, 1, \dots, 1, 0, 0, \dots, 0)$, where $\|\bar{\mathbf{x}}_C\|_0 = r$. Let $\rho_r \triangleq \rho_{\mathbf{x}_C, \bar{\mathbf{x}}_C}$. Then for all $\mathbf{x}, \bar{\mathbf{x}}$ such that $\|\mathbf{x} - \bar{\mathbf{x}}\|_0 = r$, we have $\rho_{\mathbf{x}, \bar{\mathbf{x}}} = \rho_r$.*

Based on Lemma 4, finding $R(\mathbf{x}, p, q)$ for a given p , it suffices to find the maximum r such that $\rho_r(p) > 0.5$. Since the likelihood ratio $\eta_{\mathbf{x}, \bar{\mathbf{x}}}(z)$ is always positive and finite, the inverse ρ_r^{-1} exists (due to Remark 3), which allows us to pre-compute $\rho_r^{-1}(0.5)$ and check $p > \rho_r^{-1}(0.5)$ for each $r \in \mathbb{Z}_{\geq 0}$, instead of computing $\rho_r(p)$ for each given p and r . Then $R(\mathbf{x}, p, q)$ is simply the maximum r such that $p > \rho_r^{-1}(0.5)$. Below we discuss how to compute $\rho_r^{-1}(0.5)$ in a scalable way. Our first step is to identify a set of likelihood ratio regions $\mathcal{L}_1, \dots, \mathcal{L}_n$ such that $\Pr(\phi(\mathbf{x}) \in \mathcal{L}_i)$ and $\Pr(\phi(\bar{\mathbf{x}}) \in \mathcal{L}_i)$ as used in Lemma 2 can be computed efficiently. Note that, due to Lemma 4, it suffices to consider $\mathbf{x}_C, \bar{\mathbf{x}}_C$ such that $\|\bar{\mathbf{x}}_C\|_0 = r$ throughout the derivation.

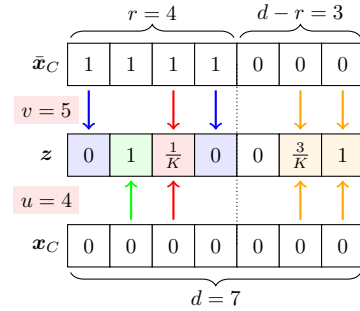


Figure 2: Illustration for Eq. (7)

For an ℓ_0 radius $r \in \mathbb{Z}_{\geq 0}$, $\forall (u, v) \in \{0, 1, \dots, d\}^2$, we construct the region

$$\mathcal{L}(u, v; r) \triangleq \{z \in \mathcal{X} : \Pr(\phi(\mathbf{x}_C) = z) = \alpha^{d-u} \beta^u, \Pr(\phi(\bar{\mathbf{x}}_C) = z) = \alpha^{d-v} \beta^v\}, \quad (7)$$

which contains points that can be obtained by “flipping” u coordinates from \mathbf{x}_C or v coordinates from $\bar{\mathbf{x}}_C$. See Figure 2 for an illustration, where different colors represent different types of coordinates: orange means both $\mathbf{x}_C, \bar{\mathbf{x}}_C$ are flipped on this coordinate and they were initially the same; red means both are flipped and were initially different; green means only \mathbf{x}_C is flipped and blue means only $\bar{\mathbf{x}}_C$ is flipped. By denoting the numbers of these coordinates as $i, j^*, u - i - j^*, v - i - j^*$, respectively, we have the following formula for computing the cardinality of each region $|\mathcal{L}(u, v; r)|$.

Lemma 5. *For any $u, v \in \{0, 1, \dots, d\}$, $u \leq v$, $r \in \mathbb{Z}_{\geq 0}$ we have $|\mathcal{L}(u, v; r)| = |\mathcal{L}(v, u; r)|$, and*

$$|\mathcal{L}(u, v; r)| = \sum_{i=\max\{0, v-r\}}^{\min(u, d-r, \lfloor \frac{u+v-r}{2} \rfloor)} \frac{(K-1)^{j^*} r!}{(u-i-j^*)!(v-i-j^*)! j^*!} \frac{K^i (d-r)!}{(d-r-i)! i!},$$

where $j^* \triangleq u + v - 2i - r$.

Therefore, for a fixed r , the complexity of computing all the cardinalities $|\mathcal{L}(u, v; r)|$ is $\Theta(d^3)$. Since each region $\mathcal{L}(u, v; r)$ has a constant likelihood ratio $\alpha^{v-u} \beta^{u-v}$ and we have $\cup_{u=0}^d \cup_{v=0}^d \mathcal{L}(u, v; r) = \mathcal{X}$, we can apply the regions to find the function $\rho_{\mathbf{x}, \bar{\mathbf{x}}} = \rho_r$ via Lemma 2. Under this representation, the number of nonempty likelihood ratio regions n is bounded by $(d+1)^2$, the perturbation probability $\Pr(\phi(\mathbf{x}) \in \mathcal{L}(u, v; r))$ used in Lemma 2 is simply $\alpha^{d-u} \beta^u |\mathcal{L}(u, v; r)|$, and similarly for the $\Pr(\phi(\bar{\mathbf{x}}) \in \mathcal{L}(u, v; r))$. Based on Lemma 2 and Lemma 5, we may use a for-loop to compute the bijection $\rho_r(\cdot)$ for the input p until $\rho_r(p) = 0.5$, and return the corresponding p as $\rho_r^{-1}(0.5)$. The procedure is illustrated in Algorithm 1.

⁴More generally, the method applies to the $\ell_0 /$ Hamming distance in a Hamming space (i.e., fixed length sequences of tokens from a discrete set, e.g., $(\spadesuit 10, \spadesuit J, \spadesuit Q, \spadesuit K, \spadesuit A) \in \{\spadesuit A, \spadesuit K, \dots, \clubsuit 2\}^5$).

Scalable implementation. In practice, Algorithm 1 can be challenging to implement; the probability values (e.g., $\alpha^{d-u}\beta^u$) can be extremely small, which is infeasible to be computationally represented using floating points. If we set α to be a rational number, both α and β can be represented in fractions, and thus all the corresponding probability values can be represented by two (large) integers; we also observe that computing the (large) cardinality $|\mathcal{L}(u, v; r)|$ is feasible in modern large integer computation frameworks in practice (e.g., python), which motivates us to adapt the computation in Algorithm 1 to large integers.

For simplicity, we assume $\alpha = \alpha'/100$ with some $\alpha' \in \mathbb{Z} : 100 \geq \alpha' \geq 0$. If we define $\tilde{\alpha} \triangleq 100K\alpha \in \mathbb{Z}, \tilde{\beta} \triangleq 100K\beta \in \mathbb{Z}$, we may implement Algorithm 1 in terms of the non-normalized, integer version $\tilde{\alpha}, \tilde{\beta}$. Specifically, we replace α, β and the constant 0.5 with $\tilde{\alpha}, \tilde{\beta}$ and $50K \times (100K)^{d-1}$, respectively. Then all the computations in Algorithm 1 can be trivially adapted except the division $(0.5 - \rho_r)/\rho'_r$. Since the division is bounded by $|\mathcal{L}(u_i, v_i; r)|$ (see the comparison between line 9 and line 11), we can implement the division by a binary search over $\{1, 2, \dots, |\mathcal{L}\{m_i, n_i\}|\}$, which will result in an upper bound with an error bounded by ρ'_r in the original space, which is in turn bounded by α^d assuming $\alpha > \beta$. Finally, to map the computed, unnormalized $\rho_r^{-1}(0.5)$, denoted as $\tilde{\rho}_r^{-1}(0.5)$, back to the original space, we find an upper bound of $\rho_r^{-1}(0.5)$ up to the precision of 10^{-c} for some $c \in \mathbb{Z}_{>0}$ (we set $c = 20$ in the experiments): we find the smallest upper bound of $\tilde{\rho}_r^{-1}(0.5) \leq \hat{\rho} \times (10K)^c (100K)^{d-c}$ over $\hat{\rho} \in \{1, 2, \dots, 10^c\}$ via binary search, and report an upper bound of $\rho_r^{-1}(0.5)$ as $\hat{\rho} \times 10^{-c}$ with an error bounded by $10^{-c} + \alpha^d$ in total. Note that an upper bound of $\rho_r^{-1}(0.5)$ is still a valid certificate.

As a side note, simply computing the probabilities in the log-domain will lead to uncontrollable approximate results due to floating point arithmetic; using large integers to ensure a verifiable approximation error in Algorithm 1 is necessary to ensure a computationally accurate certificate.

3.5 Connection Between the Discrete Distribution and an Isotropic Gaussian Distribution

When the inputs are binary vectors $\mathcal{X} = \{0, 1\}^d$, one may still apply the prior work [6] using an additive isotropic Gaussian noise ϕ to obtain an ℓ_0 certificates since there is a bijection between ℓ_0 and ℓ_2 distance in $\{0, 1\}^d$. If one uses a denoising function $\zeta(\cdot)$ that projects each randomized coordinate $\phi(\mathbf{x})_i \in \mathbb{R}$ back to the space $\{0, 1\}$ using the (likelihood ratio testing) rule

$$\zeta(\phi(\mathbf{x}))_i = \mathbb{I}\{\phi(\mathbf{x})_i > 0.5\}, \forall i \in [d],$$

then the composition $\zeta \circ \phi$ is equivalent to our discrete randomization scheme with $\alpha = \Phi(0.5; \mu = 0, \sigma^2)$, where Φ is the CDF function of the Gaussian distribution with mean μ and variance σ^2 .

If one applies a classifier upon the composition (or, equivalently, the discrete randomization scheme), then the certificates obtained via the discrete distribution is always tighter than the one via Gaussian distribution. Concretely, we denote $\mathcal{F}_\zeta \subset \mathcal{F}$ as the set of measurable functions with respect to the Gaussian distribution that can be written as the composition $\bar{f}' \circ \zeta$ for some \bar{f}' , and we have

$$\min_{\bar{f} \in \mathcal{F}_\zeta: \Pr(\bar{f}(\phi(\bar{\mathbf{x}})) = y) = p} \Pr(\bar{f}(\phi(\bar{\mathbf{x}})) = y) \geq \min_{\bar{f}' \in \mathcal{F}: \Pr(\bar{f}'(\phi(\bar{\mathbf{x}})) = y) = p} \Pr(\bar{f}'(\phi(\bar{\mathbf{x}})) = y),$$

where the LHS corresponds to the certificate derived from the discrete distribution (i.e., applying ζ to an isotropic Gaussian), and the RHS corresponds to the certificate from the Gaussian distribution.

3.6 A Certificate with Additional Assumptions

In the previous analyses, we assume nothing but the measurability of the classifier. If we further make assumptions about the functional class of the classifier, we can obtain a tighter certificate than the ones outlined in §3.1. Assuming an extra denoising step in the classifier over an additive Gaussian noise as illustrated in §3.5 is one example.

Algorithm 1 Computing $\rho_r^{-1}(0.5)$

```

1: sort  $\{(u_i, v_i)\}_{i=1}^n$  by likelihood ratio
2:  $p, \rho_r = 0, 0$ 
3: for  $i = 1, \dots, n$  do
4:    $p' = \alpha^{d-u_i} \beta^{u_i}$ 
5:    $\rho'_r = \alpha^{d-v_i} \beta^{v_i}$ 
6:    $\Delta\rho_r = \rho'_r \times |\mathcal{L}(u_i, v_i; r)|$ 
7:   if  $\rho_r + \Delta\rho_r < 0.5$  then
8:      $\rho_r = \rho_r + \Delta\rho_r$ 
9:      $p = p + p' \times |\mathcal{L}(u_i, v_i; r)|$ 
10:  else
11:     $p = p + p' \times (0.5 - \rho_r) / \rho'_r$ 
12:  return  $p$ 
13: end if
14: end for

```

Here we illustrate the idea with another example. We assume that the inputs are binary vectors $\mathcal{X} = \{0, 1\}^d$, the outputs are binary $\mathcal{Y} = \{0, 1\}$, and that the classifier is a decision tree that each input coordinate can be used at most once in the entire tree. Under the discrete randomization scheme, the prediction probability can be computed via tree recursion, since a decision tree over the discrete randomization scheme can be interpreted as assigning a probability of visiting the left child and the right child for each decision node. To elaborate, we denote $\text{idx}[i]$, $\text{left}[i]$, and $\text{right}[i]$ as the split feature index, the left child and the right child of the i^{th} node. Without loss of generality, we assume that each decision node i routes its input to the right branch if $\mathbf{x}_{\text{idx}[i]} = 1$. Then $\Pr(f(\phi(\mathbf{x})) = 1)$ can be found by the recursion

$$\text{pred}[i] = \alpha^{\mathbb{I}\{\mathbf{x}_{\text{idx}[i]}=1\}} \beta^{\mathbb{I}\{\mathbf{x}_{\text{idx}[i]}=0\}} \text{pred}[\text{right}[i]] + \alpha^{\mathbb{I}\{\mathbf{x}_{\text{idx}[i]}=0\}} \beta^{\mathbb{I}\{\mathbf{x}_{\text{idx}[i]}=1\}} \text{pred}[\text{left}[i]], \quad (8)$$

where the boundary condition is the output of the leaf nodes. Effectively, we are recursively aggregating the partial solutions found in the left subtree and the right subtree rooted at each node i , and $\text{pred}[\text{root}]$ is the final prediction probability. Note that changing one input coordinate in \mathbf{x}_k is equivalent to changing the recursion in the corresponding unique node i' (if exists) that uses feature k as the splitting index, which gives

$$\text{pred}[i'] = \alpha^{\mathbb{I}\{\mathbf{x}_{\text{idx}[i']}=0\}} \beta^{\mathbb{I}\{\mathbf{x}_{\text{idx}[i']}=1\}} \text{pred}[\text{right}[i']] + \alpha^{\mathbb{I}\{\mathbf{x}_{\text{idx}[i']}=1\}} \beta^{\mathbb{I}\{\mathbf{x}_{\text{idx}[i']}=0\}} \text{pred}[\text{left}[i']].$$

In addition, changes in the left subtree do not affect the partial solution found in the right subtree, and vice versa. Hence, we may use dynamic programming to find the *exact* adversary under each ℓ_0 radius r by aggregating the worst case changes found in the left subtree and the right subtree rooted at each node i . See Appendix B.1 for details.

4 Learning and Prediction in Practice

Since we focus on the development of certificates, here we only briefly discuss how we train the classifiers and compute the prediction probability $\Pr(f(\phi(\mathbf{x})) = y)$ in practice.

Deep networks: We follow the approach proposed by the prior work [21]: training is conducted on samples drawn from the randomization scheme via a cross entropy loss. The prediction probability $\Pr(f(\phi(\mathbf{x})) = y)$ is estimated by the lower bound of the Clopper-Pearson Bernoulli confidence interval [5] with 100K samples drawn from the distribution and the 99.9% confidence level. Since $\rho_{\mathbf{x}, \bar{\mathbf{x}}}(p)$ is an increasing function of p (Remark 3), a lower bound of p entails a valid certificate.

Decision trees: we train the decision tree greedily in a breadth-first ordering with a depth limit; for each split, we only search coordinates that are not used before to enforce the functional constraint in §3.6, and optimize a weighted gini index, which weights each training example \mathbf{x} by the probability that it is routed to the node by the discrete randomization. The details of the training algorithm is in Appendix B.2. The prediction probability is computed by Eq. (8).

5 Experiment

In this section, we validate the robustness certificates of the proposed discrete distribution (\mathcal{D}) in ℓ_0 norm. We compare to the state-of-the-art additive isotropic Gaussian noise (\mathcal{N}) [6], since an ℓ_0 certificate with radius r in $\mathcal{X} = \{0, \frac{1}{K}, \dots, 1\}^d$ can be obtained from an ℓ_2 certificate with radius \sqrt{r} . Note that the derived ℓ_0 certificate from Gaussian distribution is still tight with respect to all the measurable classifiers (see Theorem 1 in [6]). We consider the following evaluation measures:

- $\mu(R)$: the average certified ℓ_0 radius $R(\mathbf{x}, p, q)$ (with respect to the labels) across the testing set.
- $\text{ACC}@r$: the certified accuracy within a radius r (the average $\mathbb{I}\{R(\mathbf{x}, p, q) \geq r\}$ in the testing set).

5.1 Binarized MNIST

We use a 55,000/5,000/10,000 split of the MNIST dataset for training/validation/testing. For each data point \mathbf{x} in the dataset, we binarize each coordinate by setting the threshold as 0.5. Experiments are conducted on randomly smoothed CNN models and the implementation details are in Appendix C.1.

The results are shown in Table 1. For the same randomly smoothed CNN model (the 1st and 2nd rows in Table 1), our certificates are consistently better than the ones derived from the Gaussian

Table 1: Randomly smoothed CNN models on the MNIST dataset. The first two rows refer to the same model with certificates computed via different methods (see details in §3.5).

| ϕ | Certificate | $\mu(R)$ | ACC@ r | | | | | | |
|---------------|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | | $r = 1$ | $r = 2$ | $r = 3$ | $r = 4$ | $r = 5$ | $r = 6$ | $r = 7$ |
| \mathcal{D} | \mathcal{D} | 3.456 | 0.921 | 0.774 | 0.539 | 0.524 | 0.357 | 0.202 | 0.097 |
| \mathcal{D} | \mathcal{N} [6] | 1.799 | 0.830 | 0.557 | 0.272 | 0.119 | 0.021 | 0.000 | 0.000 |
| \mathcal{N} | \mathcal{N} [6] | 2.378 | 0.884 | 0.701 | 0.464 | 0.252 | 0.078 | 0.000 | 0.000 |

Table 2: The guaranteed accuracy of randomly smoothed ResNet50 models on ImageNet.

| ϕ and certificate | ACC@ r | | | | | | |
|------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | $r = 1$ | $r = 2$ | $r = 3$ | $r = 4$ | $r = 5$ | $r = 6$ | $r = 7$ |
| \mathcal{D} | 0.538 | 0.394 | 0.338 | 0.274 | 0.234 | 0.190 | 0.176 |
| \mathcal{N} [6] | 0.372 | 0.292 | 0.226 | 0.194 | 0.170 | 0.154 | 0.138 |

distribution (see §3.5). The gap between the average certified radius is about 1.7 in ℓ_0 distance, and the gap between the certified accuracy can be as large as 0.4. Compared to the models trained with Gaussian noise (the 3rd row in Table 1), our model is also consistently better in terms of the measures.

Since the above comparison between our certificates and the Gaussian-based certificates is *relative*, we conduct an exhaustive search over all the possible adversary within ℓ_0 radii 1 and 2 to study the tightness against the *exact* certificate. The resulting certified accuracies at radii 1 and 2 are 0.954 and 0.926, respectively, which suggest that our certificate is reasonably tight when $r = 1$ (0.954 vs. 0.921), but still too pessimistic when $r = 2$ (0.926 vs. 0.774). The phenomenon is expected since the certificate is based on *all the measurable functions* for the discrete distribution. A tighter certificate requires additional assumptions on the classifier such as the example in §3.6.

5.2 ImageNet

We conduct experiments on ImageNet [8], a large scale image dataset with 1,000 labels. Following common practice, we consider the input space $\mathcal{X} = \{0, 1/255, \dots, 1\}^{224 \times 224 \times 3}$ by scaling the images. We consider the same ResNet50 classifier [17] and learning procedure as Cohen et al. [6] with the only modification on the noise distribution. The details and visualizations can be found in Appendix C.2. For comparison, we report the best guaranteed accuracy of each method for each ℓ_0 radius r in Table 2. Our model outperforms the competitor by a large margin at $r = 1$ (0.538 vs. 0.372), and consistently outperforms the baseline across different radii.

Analysis. We analyze our method in ImageNet in terms of 1) the number n of nonempty likelihood ratio region $\mathcal{L}(u, v; r)$ in Algorithm 1, 2) the pre-computed $\rho_r^{-1}(0.5)$, and 3) the certified accuracy at each α . The results are in Figure 3. For reproducibility, the detailed accuracy numbers of 3) is available in Table 3 in Appendix C.2, and the pre-computed $\rho_r^{-1}(0.5)$ is available at our code repository. 1) The number n of nonempty likelihood ratio regions is much smaller than the bound $(d + 1)^2 = (3 \times 224 \times 224)^2$ for small radii. 2) The value $\rho_r^{-1}(0.5)$ approaches 1 more rapidly for a higher α value than a lower one. Note that $\rho_r^{-1}(0.5)$ only reaches 1 when $r = d$ due to Remark 3. Computing $\rho_r^{-1}(0.5)$ in large integer is time-consuming, which takes about 4 days for each α and r , but this can be trivially parallelized across different α and r .⁵ For each radius r and randomization parameter α , note that the 4-day computation only has to be done *once*, and the pre-computed $\rho_r^{-1}(0.5)$ can be applied to any ImageNet scale images and models. 3) The certified accuracy behaves nonlinearly across different radii; relatively, a high α value exhibits a high certified accuracy at small radii and low certified accuracy at large radii, and vice versa.

⁵As a side note, computing $\rho_r^{-1}(0.5)$ in MNIST takes less than 1 second for each α and r .

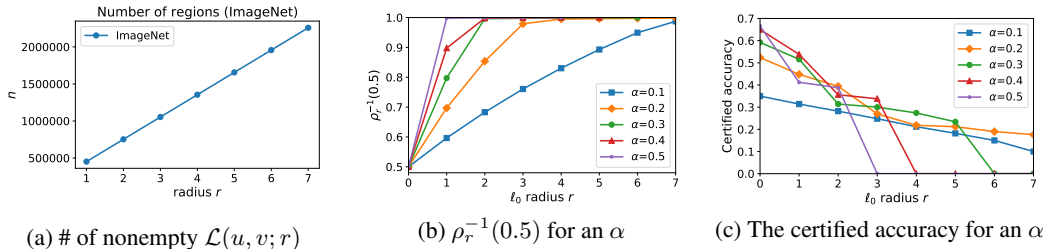


Figure 3: Analysis of the proposed method in the ImageNet dataset.

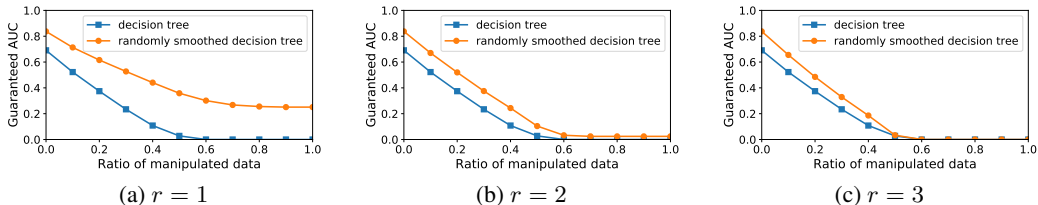


Figure 4: The guaranteed AUC in the Bace dataset across different ℓ_0 radius r and the ratio of testing data that the adversary can manipulate.

5.3 Chemical Property Prediction

The experiment is conducted on the Bace dataset [35], a binary classification dataset for biophysical property prediction on molecules. We use the Morgan fingerprints [32] to represent molecules, which are commonly used binary features [41] indicating the presence of various chemical substructures. The dimension of the features (fingerprints) is 1,024. Here we focus on an ablation study comparing the proposed randomly smoothed decision tree with a vanilla decision tree, where the adversary is found by dynamic programming in §3.6 (thus the exact worst case) and a greedy search, respectively. More details can be found in Appendix C.3.

Since the chemical property prediction is typically evaluated via AUC [41], we define a robust version of AUC that takes account of the radius of the adversary as well as the ratio of testing data that can be manipulated. Note that to maximally decrease the score of AUC via a positive (negative) example, the adversary only has to maximally decrease (increase) its prediction probability, regardless of the scores of the other examples. Hence, given an ℓ_0 radius r and a ratio of testing data, we first compute the adversary for each testing data, and then find the combination of adversaries and the clean data under the ratio constraint that leads to the worst AUC score. See details in Appendix C.4.

The results are in Figure 4. Empirically, the adversary of the decision tree at $r = 1$ always changes the prediction probability of a positive (negative) example to 0 (1). Hence, the plots of the decision tree model are constant across different ℓ_0 radii. The randomly smoothed decision tree is consistently more robust than the vanilla decision tree model. We also compare the exact certificate of the prediction probability with the one derived from Lemma 2; the average difference across the training data is 0.358 and 0.402 when r equals to 1 and 2, respectively. The phenomenon encourages the development of a classifier-aware guarantee that is tighter than the classifier-agnostic guarantee.

6 Conclusion

We present a stratified approach to certifying the robustness of randomly smoothed classifiers, where the robustness guarantees can be obtained in various resolutions and perspectives, ranging from a point-wise certificate to a regional certificate and from general results to specific examples. The hierarchical investigation opens up many avenues for future extensions at different levels.

Acknowledgments

GH and TJ were in part supported by a grant from Siemens Corporation.

References

- [1] G. W. Bemis and M. A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.
- [2] X. Cao and N. Z. Gong. Mitigating evasion attacks to deep neural networks via region-based classification. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 278–287. ACM, 2017.
- [3] N. Carlini, G. Katz, C. Barrett, and D. L. Dill. Provably minimally-distorted adversarial examples. *arXiv preprint arXiv:1709.10207*, 2017.
- [4] C.-H. Cheng, G. Nührenberg, and H. Ruess. Maximum resilience of artificial neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pages 251–268. Springer, 2017.
- [5] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.
- [6] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. In *the 36th International Conference on Machine Learning*, 2019.
- [7] F. Croce, M. Andriushchenko, and M. Hein. Provable robustness of relu networks via maximization of linear regions. In *the 22nd International Conference on Artificial Intelligence and Statistics*, 2018.
- [8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE international conference on computer vision*, pages 248–255. Ieee, 2009.
- [9] S. Dutta, S. Jha, S. Sankaranarayanan, and A. Tiwari. Output range analysis for deep feedforward neural networks. In *NASA Formal Methods Symposium*, pages 121–138. Springer, 2018.
- [10] K. Dvijotham, S. Gowal, R. Stanforth, R. Arandjelovic, B. O’Donoghue, J. Uesato, and P. Kohli. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265*, 2018.
- [11] K. Dvijotham, R. Stanforth, S. Gowal, T. Mann, and P. Kohli. A dual approach to scalable verification of deep networks. In *the 34th Annual Conference on Uncertainty in Artificial Intelligence*, 2018.
- [12] R. Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pages 269–286. Springer, 2017.
- [13] C. Finlay, A.-A. Pooladian, and A. M. Oberman. The logbarrier adversarial attack: making effective use of decision boundary information. *arXiv preprint arXiv:1903.10396*, 2019.
- [14] M. Fischetti and J. Jo. Deep neural networks and mixed integer linear optimization. *Constraints*, 23:296–309, 2018.
- [15] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.
- [16] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, T. Mann, and P. Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.
- [17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [18] P.-S. Huang, R. Stanforth, J. Welbl, C. Dyer, D. Yogatama, S. Gowal, K. Dvijotham, and P. Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. *arXiv preprint arXiv:1909.01492*, 2019.

- [19] R. Jia, A. Raghunathan, K. Göksel, and P. Liang. Certified robustness to adversarial word substitutions. *arXiv preprint arXiv:1909.00986*, 2019.
- [20] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer, 2017.
- [21] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. *IEEE Symposium on Security and Privacy (SP)*, 2019.
- [22] G.-H. Lee, D. Alvarez-Melis, and T. S. Jaakkola. Towards robust, locally linear deep networks. In *International Conference on Learning Representations*, 2019.
- [23] B. Li, C. Chen, W. Wang, and L. Carin. Second-order adversarial attack and certifiable robustness. *arXiv preprint arXiv:1809.03113*, 2018.
- [24] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–385, 2018.
- [25] A. Lomuscio and L. Maganti. An approach to reachability analysis for feed-forward relu neural networks. *arXiv preprint arXiv:1706.07351*, 2017.
- [26] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
- [27] M. Mirman, T. Gehr, and M. Vechev. Differentiable abstract interpretation for provably robust neural networks. In *the 35th International Conference on Machine Learning*, 2018.
- [28] J. Neyman and E. S. Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.
- [29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.
- [30] A. Raghunathan, J. Steinhardt, and P. Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.
- [31] A. Raghunathan, J. Steinhardt, and P. S. Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10877–10887, 2018.
- [32] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.
- [33] K. Scheibler, L. Winterer, R. Wimmer, and B. Becker. Towards verification of artificial neural networks. In *MBMV*, pages 30–40, 2015.
- [34] G. Singh, T. Gehr, M. Mirman, M. Püschel, and M. Vechev. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems*, pages 10802–10813, 2018.
- [35] G. Subramanian, B. Ramsundar, V. Pande, and R. A. Denny. Computational modeling of β -secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling*, 56(10):1936–1949, 2016.
- [36] V. Tjeng, K. Xiao, and R. Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2017.
- [37] K. Tocher. Extension of the neyman-pearson theory of tests to discontinuous variates. *Biometrika*, 37(1/2):130–144, 1950.

- [38] T.-W. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, D. Boning, I. S. Dhillon, and L. Daniel. Towards fast computation of certified robustness for relu networks. In *the 35th International Conference on Machine Learning*, 2018.
- [39] E. Wong and J. Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *the 35th International Conference on Machine Learning*, 2018.
- [40] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, pages 8400–8409, 2018.
- [41] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.
- [42] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, pages 4939–4948, 2018.