Proceedings of Machine Learning Research vol 144:1–16, 2022

Certified Adversarial Robustness via Locally Biased Randomized Smoothing

Brendon G. Anderson BGANDERSON@BERKELEY.EDU
Somayeh Sojoudi SOJOUDI@BERKELEY.EDU
University of California, Berkeley

Abstract

The successful incorporation of machine learning models into safety-critical control systems requires rigorous robustness guarantees. Randomized smoothing remains the state-of-the-art method for robustification with theoretical guarantees. We show that using uniform and unbiased smoothing measures, as is standard in the literature, relies on the underlying assumption that smooth decision boundaries yield good robustness, which manifests into a robustness-accuracy tradeoff. We generalize the smoothing framework to rid this assumption and learn a locally optimal robustification of the decision boundary based on training data, a method we term *locally biased randomized smoothing*. We prove nontrivial closed-form certified robust radii for the resulting model, avoiding Monte Carlo certifications as used by other smoothing methods. Experiments on synthetic, MNIST, and CIFAR-10 data show increased certified radii and accuracy over conventional smoothing.

Keywords: Adversarial robustness, randomized smoothing

1. Introduction

In light of their impressive representation capabilities and computational efficiency, machine learning models are rapidly being adopted in a variety of control tasks, ranging from autonomous driving (Bojarski et al., 2016; Wu et al., 2017) to reinforcement learning for uncertain systems (Levine et al., 2016; Sutton and Barto, 2018). Nevertheless, these models (and in particular, neural networks) can be extremely sensitive to small perturbations in their inputs (Biggio et al., 2013; Szegedy et al., 2014; Nguyen et al., 2015), a property directly at odds with the robustness and stability guarantees cherished by the control community (Recht, 2019). Recent works have tried to address this gap in the forms of adversarial training (Goodfellow et al., 2015; Madry et al., 2018; Shafahi et al., 2019) and robustness certification (Wong and Kolter, 2018; Weng et al., 2018a; Raghunathan et al., 2018; Fazlyab et al., 2020; Anderson et al., 2020). However, the challenge of developing nontrivial robustness guarantees that scale to practically-sized settings remains an open problem.

Randomized smoothing, introduced and popularized by Lecuyer et al. (2019); Li et al. (2019); Cohen et al. (2019), is commonly accepted as the state-of-the-art method for efficiently robustifying large-scale models with rigorous robustness guarantees. Instead of relying on the model's baseline prediction, randomized smoothing assigns the most probable prediction when considering random perturbations of the input. Intuitively, this ensemble approach averages out any outlier inputs that may have drastically changed the prediction—such inputs are termed *adversarial inputs* or *adversarial attacks*. This averaging operation smooths the level sets of the model's input-output map. By considering randomized smoothing with specific probability distributions, e.g., normal or Laplacian, researchers have proven the non-existence of adversarial inputs within balls corresponding to

© 2022 B. G. Anderson & S. Sojoudi.

some norm or metric, e.g., ℓ_2- or ℓ_1-norm, or Wasserstein metrics (Cohen et al., 2019; Teng et al., 2020; Levine and Feizi, 2020). The radius of such a ball is called a *certified radius* or *robust radius*.

Despite the popularity of randomized smoothing, the method still presents a handful of limitations and open questions, many of which have only recently been considered or remain under investigation. For example, Salman et al. (2019) blends randomized smoothing with adversarial training and significantly improved the resulting model's certified robustness. The paper Yang et al. (2020a) determines the geometry of optimal smoothing distributions for ℓ_1-, ℓ_2-, and ℓ_∞-norm bounded attacks. Contrarily, Zhang et al. (2020) considers optimizing the base classifier to maximize the robust radius for a fixed distribution. The work Dvijotham et al. (2020) develops a measure-theoretic approach for robustness certification of models smoothed using arbitrary distributions. Many negative results have also been shown, e.g., Mohapatra et al. (2021) shows that smoothed classifiers suffer from a "shrinking phenomenon": decision regions shrink and eventually vanish as the variance of the smoothing distribution increases. Many works have also identified a robustness-accuracy tradeoff in relation to the smoothness of models (Tsipras et al., 2019; Krishnan et al., 2020; Yang et al., 2020b; Gao et al., 2020), a limitation we discuss in Section 2.3 and address in our proposed approach. Finally, very recent works have considered more general formulations of randomized smoothing in an attempt to increase certified radii—we discuss these works in-depth in Section 2.4.

Randomized smoothing is usually considered in a static classification setting, and this is the setting we study. Nonetheless, such works are actively being incorporated into dynamic settings with more general outputs, e.g., smoothing of neural network policies in reinforcement learning (Kumar et al., 2021; Wu et al., 2021). Consequently, the results of this paper are of interest in more general dynamic learning problems than the static classification setting that we present.

1.1. Contributions

We show that standard randomized smoothing methods possess the informal assumption that making models smoother is a good surrogate for making them more robust. This manifests into a robustness-accuracy tradeoff, and we show that to eradicate the assumption it is necessary to generalize to biased and input-dependent distributions. Accordingly, we propose *locally biased randomized smoothing*, which uses training data to directly learn model robustification without relying on the assumption that smoothness yields robustness. We obtain a closed-form smoothed model with closed-form certified radii for arbitrary norms, overcoming the Monte Carlo estimations relied on by many current smoothing methods. Our experiments demonstrate increased accuracy both on clean and adversarially attacked data, as well as increased certified radii.

2. Randomized Smoothing: Review, Limitations, and Generalizations

2.1. Preliminaries

We denote by $\mathcal{P}(\mathbb{R}^d)$ the set of probability measures on $\mathbb{R}^d$ equipped with the Borel σ-algebra. If $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $g\colon \mathbb{R}^d \to \mathbb{R}^n$ has μ-integrable components $g_i\colon \mathbb{R}^d \to \mathbb{R}$, $i \in \{1, 2, \dots, n\}$, we define the expectation $E_{x\sim\mu} g(x) := \int_{\mathbb{R}^d} g(x) d\mu(x) = \big(\int_{\mathbb{R}^d} g_1(x) d\mu(x), \dots, \int_{\mathbb{R}^d} g_n(x) d\mu(x)\big)$. We assume μ-integrability whenever we write $E_{x\sim\mu} g(x)$ or $\int_{\mathbb{R}^d} g(x) d\mu(x)$. The dual norm of a norm $\|\cdot\|\colon \mathbb{R}^d \to [0, \infty)$ is denoted by $\|\cdot\|_*$, and is given by $\|y\|_* = \sup\{x^\top y : \|x\| \leq 1\}$ for all $y \in \mathbb{R}^d$. Throughout, we allow $\|\cdot\|$ to denote an arbitrary norm, the domain of which will be clear

from context. We let $\rho\colon \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty)$ denote the metric defined by $\rho(x,y) = \|x-y\|_2$. For ease of exposition, we assume that all $\arg\max$ and $\arg\min$ yield singleton sets.[1]

Consider a classifier $f\colon \mathbb{R}^d \to \{1,2,\ldots,n\}$, $f(x) \in \arg\max_{i\in\{1,2,\ldots,n\}} g_i(x)$, where $g\colon \mathbb{R}^d \to \mathbb{R}^n$. In this paper we consider robustifying f using the framework of randomized smoothing.

2.2. Review of Randomized Smoothing

The idea of randomized smoothing is this: instead of assigning the class $f(x)$ to an input $x \in \mathbb{R}^d$, assign the expected class under f of random perturbations of x. This amounts to choosing a smoothing measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ and replacing f with the smoothed classifier $f^\mu\colon \mathbb{R}^d \to \{1,2,\ldots,n\}$ defined by $f^\mu(x) \in \arg\max_{i\in\{1,2,\ldots,n\}} g_i^\mu(x)$, with $g^\mu\colon \mathbb{R}^d \to \mathbb{R}^n$ given by $g^\mu(x) = E_{\epsilon\sim\mu} g(x+\epsilon)$.

Some works consider directly manipulating the hard classifier f without regard to the soft classifier g (Cohen et al., 2019; Teng et al., 2020). In contrast, we smooth the soft classifier g before the $\arg\max$ is taken, as is done in many other works (Salman et al., 2019; Zhai et al., 2020; Levine and Feizi, 2020; Kumar et al., 2020). Smoothing g, which generalizes smoothing f (Salman et al., 2019), takes into account the confidence of the base classifier, whereas hard smoothing does not (Kumar et al., 2020). Consequently, we concern ourselves only with soft smoothing.

Intuitively, randomized smoothing flattens jagged regions of the decision boundary, where adversarial inputs are conjectured to exist (Fawzi et al., 2018). This intuition can be formalized in the framework of convolution. If μ has density $\phi\colon \mathbb{R}^d \to [0,\infty)$ (with respect to Lebesgue measure) that is even symmetric (i.e., $\phi(-x) = \phi(x)$), then randomized smoothing is the convolution

$$g^\mu(x) = \int_{\mathbb{R}^d} \phi(\epsilon) g(x-\epsilon)\, d\epsilon =: \phi * g(x).$$

In general, the convolution $g^\mu = \phi * g$ is smoother than the functions ϕ and g being convolved (Folland, 1999). From the control and signal processing perspective, this convolutional representation shows that randomized smoothing acts like a low-pass filter on g. Upon attenuating the high-frequency behavior in g via smoothing, the radius of robustness around clean inputs has been found to increase, with certified robust radii given for special cases of the smoothing measure μ.

The most popular form of randomized smoothing, introduced in Cohen et al. (2019), takes the smoothing measure μ to be that of the normal distribution $N(0,\sigma^2 I)$. We refer to this scheme as *normal smoothing*. In this case, g^μ becomes the Weierstrass transform of g, which is well known to attenuate high-frequency components in g. Since evaluating $g^\mu(x)$ in this case requires computing an integral that has no closed form in general, implementing normal smoothing typically requires Monte Carlo estimation. The authors of Cohen et al. (2019) proved a certified robust ℓ_2-radius for normal smoothing, which must also be estimated via Monte Carlo methods. We recall the result below in terms of soft classifier smoothing—see Zhai et al. (2020) for this generalization.

Theorem 1 (Cohen et al. (2019); Zhai et al. (2020)) *Assume that $g\colon \mathbb{R}^d \to [0,1]^n$. Let $\sigma^2 > 0$, and let μ be the probability measure of the normal distribution $N(0,\sigma^2 I)$. Denote the distribution function of $N(0,I)$ by Φ. Consider a point $x \in \mathbb{R}^d$ and let $y = f^\mu(x) \in \arg\max_{i\in\{1,2,\ldots,n\}} g_i^\mu(x)$ and $y' \in \arg\max_{i\in\{1,2,\ldots,n\}\setminus\{y\}} g_i^\mu(x)$. Then $f^\mu(x+\delta) = y$ for all $\delta \in \mathbb{R}^d$ such that*

$$\|\delta\|_2 \le r^\mu_{nrml}(x) := \frac{\sigma}{2}\left(\Phi^{-1}(g_y^\mu(x)) - \Phi^{-1}(g_{y'}^\mu(x))\right).$$

1. This assumption is violated in some cases, e.g., when considering inputs on decision boundaries. In practice, however, we need not worry about these cases, as they correspond to sets of zero Lebesgue measure.

2.3. Limitations of Randomized Smoothing

We remark two important restrictions on the measure μ that are common to most randomized smoothing methods in the literature: 1) μ is uniform with respect to the input x, and 2) μ is centered at $0 \in \mathbb{R}^d$. In this section, we formalize these restrictions and show why they should be relaxed.

We begin with Proposition 3 below, which, as a direct consequence of the uniform smoothing measure, shows that g^μ is necessarily "more constant" than g. This forces classification to remain constant over larger regions of the input space, but when these regions become too large (i.e., when L becomes very small), the accuracy of the predictions degrades (Krishnan et al., 2020; Yang et al., 2020b). The proposition has an obvious generalization to the case of local Lipschitzness.

Definition 2 *Let $L \in \mathbb{R}$, and let $\|\cdot\|$ and $\|\cdot\|'$ be norms on $\mathbb{R}^d$ and $\mathbb{R}^n$, respectively. A function $h\colon \mathbb{R}^d \to \mathbb{R}^n$ is called* L-Lipschitz in norms $(\|\cdot\|, \|\cdot\|')$ *if for all $x, x' \in \mathbb{R}^d$ it holds that $\|h(x) - h(x')\|' \leq L\|x - x'\|$.*

Proposition 3 *If g is L-Lipschitz in norms $(\|\cdot\|, \|\cdot\|')$, then g^μ is L-Lipschitz in norms $(\|\cdot\|, \|\cdot\|')$.*

Proof Suppose that g is L-Lipschitz in norms $(\|\cdot\|, \|\cdot\|')$ and let $x, x' \in \mathbb{R}^d$. Then $\|g^\mu(x) - g^\mu(x')\|' \leq \int_{\mathbb{R}^d} \|g(x+\epsilon) - g(x'+\epsilon)\|' d\mu(\epsilon) \leq \int_{\mathbb{R}^d} L\|x - x'\| d\mu(\epsilon) = L\|x - x'\|$. ■

We next show that smoothing measures centered at the origin $0 \in \mathbb{R}^d$ cannot change a linear decision boundary, even if they are allowed to depend on the input x (which we denote by μ_x). This is true even when doing so would increase robustness with respect to the data distribution at hand. Thus, unbiased smoothing distributions cannot robustify linear classifiers.

Definition 4 *A measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is called* unbiased *if $E_{\epsilon\sim\mu}\epsilon = 0$. The measure μ is called* biased *if it is not unbiased.*

Proposition 5 *Suppose that g is affine, so that $g(x) = Ax + b$ for some $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, and consider g^μ with input-dependent smoothing measure μ_x, so that $g^\mu(x) = E_{\epsilon\sim\mu_x} g(x+\epsilon)$. If μ_x is unbiased for all $x \in \mathbb{R}^d$, then $f^\mu = f$.*

Proof Suppose that μ_x is unbiased for all $x \in \mathbb{R}^d$. Then $g^\mu(x) = E_{\epsilon\sim\mu_x}(A(x+\epsilon) + b) = Ax + b + AE_{\epsilon\sim\mu_x}\epsilon = g(x)$. Hence, $g^\mu = g$, and consequently $f^\mu = f$. ■

When the smoothing measure is unbiased and uniform with respect to the input, we refer to the scheme as *standard smoothing*. Together, the two limitations in Propositions 3 and 5 point to a fundamental informal assumption that underlies standard smoothing: *making classifiers smoother, as characterized by their Lipschitz constant or the linearity of their decision boundaries, is a good surrogate for increasing robustness*. Although standard smoothing has been shown to work well in many settings, this assumption that smoothness yields robustness is fundamentally flawed, since minimizing the Lipschitz constant degrades accuracy (Krishnan et al., 2020; Yang et al., 2020b). If the assumption were to hold, then a constant classifier, obtained, e.g., by letting $\sigma^2 \to \infty$ in Theorem 1, would be the most robust classifier, which is nonsensical when we take accuracy into account. The work Madry et al. (2018) corroborates this conclusion, arguing that simultaneous accuracy and robustness often requires a complicated decision boundary. Thus, our goal should be to directly increase robustness with respect to the data distribution, without resorting to surrogate notions such as smoothness. To do so, Propositions 3 and 5 show that we must generalize the smoothing framework to allow for input-dependent and biased smoothing measures.

2.4. Generalizing Randomized Smoothing and Related Works

Going forward we consider $\mu = \{\mu_x \in \mathcal{P}(\mathbb{R}^d) : x \in \mathbb{R}^d\}$ with all μ_x possibly biased, and define

$$g^{\mu}(x) = E_{\epsilon\sim\mu_x} g(x+\epsilon). \tag{1}$$

We refer to this scheme as *generalized smoothing*.

A handful of recent works have considered generalized smoothing (although most of them in a blind attempt to increase robust radii, not because they have recognized and seek to overcome the flawed informal assumption previously discussed). For example, Wang et al. (2021) uses normal distributions $N(0, \sigma_i^2 I)$ to maximize ℓ_2-robust regions around every training point x_i. If a test input x is not contained in any such certified region, they optimize a new variance $\sigma^2(x)$ to allocate a certified region around x, which is then used for future classification. Not only is this restricted to ℓ_2-adversaries and computationally heavy due to two-stage training, but also the resulting classifier depends on the order of incoming inputs, introducing new performance and robustness concerns.

The works Alfarra et al. (2020) and Chen et al. (2021) use μ_x being the measure associated with $N(0, \sigma^2(x)I)$, where the variance is everywhere chosen to maximize the certified ℓ_2-radius of normal smoothing; $\sigma^2(x) \in \arg\max_{\sigma^2>0} r^{\mu}_{\text{nrml}}(x)$. The authors of Eiras et al. (2021) use the same idea, albeit they extend the approach to specific anisotropic distributions. The work Súkeník et al. (2021) shows that, in addition to suffering from the curse of dimensionality, the robustness certificates issued by these three works are actually invalid in practice. To see this, consider a fixed input $x \in \mathbb{R}^d$ and its chosen smoothing measure $\mu_x \in \mathcal{P}(\mathbb{R}^d)$. These works certify that, for $\delta \in \mathbb{R}^d$ within a specified robust radius, $x+\delta$ is classified the same as x under the smoothed classifier using μ_x. However, the classifier uses the measure $\mu_{x+\delta} \neq \mu_x$ when classifying $x+\delta$ (since the measure is optimized per-input), and therefore, the robustness certificate doesn't apply to the actual classifier used at test time. To overcome this, Súkeník et al. (2021) proposes a specific parameterization of $\sigma^2(x)$ for generalized smoothing with $N(0, \sigma^2(x))$ that leads to valid robust ℓ_2-radii. However, they find that this $\sigma^2(x)$ does not notably increase the certified radii over normal smoothing in practice.

These works, all still preprints, showcase the importance and timeliness of generalized smoothing, and highlight its difficulties in deriving robust radii. In the sequel, we use generalized smoothing to learn a closed-form manipulation of the decision boundary from data. Our approach culminates into robust radii for arbitrary norms that are mathematically rigorous and practically valid.

3. Robustifying Binary Linear Classifiers

Since standard smoothing is unable to robustify linear classifiers, we start from the basics: we assume a binary linear setting, with $g\colon \mathbb{R}^d \to \mathbb{R}$, $g(x) = a^\top x + b$, and $f(x) = \text{sign}(g(x))$.

3.1. Optimal Robustification Under the Direction Oracle

Consider $x \in \mathbb{R}^d$. We start by assuming that we know that the true class of x is 1. Formally, we assume that there exists an oracle function $y\colon \mathbb{R}^d \to \{-1, 1\}$ that gives the true class of x, and for this x it holds that $y(x) = 1$. With this in mind, we remark that

$$g^{\mu}(x) = E_{\epsilon\sim\mu_x}(a^\top(x+\epsilon)+b) = g(x) + a^\top E_{\epsilon\sim\mu_x}\epsilon. \tag{2}$$

Since x has true class 1, robustification at x is equivalent to $g^{\mu}(x) > g(x)$, so that the neighborhood around x classified into class 1 increases in size. Hence, our goal amounts to maximizing $a^\top E_{\epsilon\sim\mu_x}\epsilon$.

Without constraints on μ_x, this optimization would be unbounded. Therefore, we consider measures with bounded expectation $E_{\epsilon\sim\mu_x}\epsilon$, and we find that the optimal μ_x is one attaining[2] $a^\top E_{\epsilon\sim\mu_x}\epsilon = \sup\left\{a^\top E_{\epsilon\sim\nu_x}\epsilon : \|E_{\epsilon\sim\nu_x}\epsilon\| \leq \alpha,\ \nu_x \in \mathcal{P}(\mathbb{R}^d)\right\} = \alpha\|a\|_*$.

If, on the other hand, the true class of x is -1, then the optimal μ_x is one attaining $a^\top E_{\epsilon\sim\mu_x}\epsilon = -\alpha\|a\|_*$. Therefore, for general $x \in \mathbb{R}^d$, we find that the optimal smoothed classifier is given by

$$g^\mu(x) = g(x) + \alpha y(x)\|a\|_*, \tag{3}$$

where $y(x) \in \{-1, 1\}$ is the oracle class assigned to x. We call y the *direction oracle*, since its value at x determines which direction to push the decision boundary (either in the a or $-a$ direction).

3.2. Approximating the Direction Oracle

Suppose that we have a subset of training data $\{(x_1, y_1), \ldots, (x_N, y_N)\} \subseteq \mathbb{R}^d \times \{-1, 1\}$. We have the true classes $y(x_i) = y_i$ for these data points, so the optimal robustified classification is given by $g^\mu(x_i) = g(x_i) + \alpha y_i\|a\|_*$. However, for general x we do not have access to $y(x)$ (if we did, we wouldn't need to learn anything). So, for $x \notin \{x_1, x_2, \ldots, x_N\}$, we propose to approximate $y(x)$ based on the given data. This approximation of the direction oracle will be denoted by $\hat{y}\colon \mathbb{R}^d \to \{-1, 1\}$, and will be used in place of y in our smoothed classifier (3).

It is insightful to note that g^μ does not use the oracle value $y(x)$ to directly classify x. Rather, it is used to encode which direction to push the decision boundary for robustification. Thus, a "good" approximation of the direction oracle is one that encodes a "good" manipulation of the decision boundary to achieve robustification, not necessarily one that accurately predicts the true label.

Going forward, we operate under the following informal assumption: *a good manipulation of the decision boundary at x_i remains good locally around x_i*. Since robustness is a local property and most classifiers are continuous, this is a natural assumption to make. This assumption vastly differs from the informal assumption underlying standard smoothing, namely, that smoothness of decision boundaries is a good surrogate for robustness. We will soon see that our assumption may result in the creation of nonlinearity, contradicting the informal assumption of standard smoothing.

We choose the approximate direction oracle to be the 1-nearest neighbor: $\hat{y}(x) = y_{i^*(x)}$, where $i^*\colon \mathbb{R}^d \to \{1, 2, \ldots, N\}$ is defined[3] by $i^*(x) \in \arg\min_{i\in\{1,2,\ldots,N\}} \rho(x, x_i)$. This choice is natural based on our informal assumption. Note that $\hat{y}(x_i)$ recovers y_i for the data x_i. In Theorems 6 and 8, we will see that this choice yields closed-form certified robust radii. Using other approximate direction oracles could present an interesting direction for future research (for example, k-nearest neighbors or learning a neural network to output labels that optimize the induced robustness).

3.3. Locally Biased Randomized Smoothing

With our smoothing scheme now finalized, the classifier becomes

$$g^\mu(x) = g(x) + \alpha y_{i^*(x)}\|a\|_*. \tag{4}$$

We remark the two underlying features that distinguish our scheme from standard smoothing: the direction oracle encodes an informed manipulation of the decision boundary that is determined *locally*

2. This optimization is always attained by an appropriately chosen Dirac measure.
3. This is well-defined under our assumption that the $\arg\min$ yields a singleton set.

based on data, and this manipulation optimized for robustness using *biased* smoothing measures. For this reason, we term our framework *locally biased randomized smoothing*.

In contrast to standard smoothing, g^μ may be nonlinear when the data informs us that nonlinearity is required to increase robustness. We will continue to refer to f^μ (and g^μ) as the smoothed classifier, despite the fact that it may be less smooth than the base classifier. Unlike normal smoothing, our classifier requires no Monte Carlo estimation, since the smoothing distribution has a closed-form expectation. As $\alpha \to \infty$, the classifier f^μ converges pointwise to the 1-nearest neighbor classifier. On the other hand, normal smoothing converges pointwise to a constant function as $\sigma^2 \to \infty$. Thus, we may view both methods as interpolating between the base classifier, typically optimized for clean accuracy, and a limiting classifier. With this perspective, a good limiting classifier is one that is optimized for robust accuracy, and we posit that our data-informed 1-nearest neighbor better serves this purpose than the constant function. Indeed, it has been shown that 1-nearest neighbor classifiers are accurate and certifiably robust when the data follows mild separation properties (Wang et al., 2018).

We provide closed-form certified robust radii in the following theorem.

Theorem 6 *Consider $x \in \mathbb{R}^d$ and fix $i = i^*(x)$. Then $f^\mu(x+\delta) = f^\mu(x)$ for all $\delta \in \mathbb{R}^d$ such that*

$$\|\delta\| < r^\mu_{lin}(x) \coloneqq \min\left\{\frac{|g^\mu(x)|}{\|a\|_*}, \min\left\{\frac{\rho(x,x_j)^2 - \rho(x,x_i)^2}{2\|x_i - x_j\|_*} : y_j \neq y_i,\ j \in \{1,2,\ldots,N\}\right\}\right\}.$$

Proof See Appendix A. ■

Note that $r^\mu_{\text{lin}}(x) \geq 0$ for all $x \in \mathbb{R}^d$. Furthermore, the term $\frac{|g^\mu(x)|}{\|a\|_*}$ is the distance (in norm $\|\cdot\|$) from x to the hyperplane $\{x' \in \mathbb{R}^d : g(x') + \alpha y_{i^*(x)}\|a\|_* = 0\}$ (Mangasarian, 1999), which is the decision boundary of f offset by $\alpha y_{i^*(x)}\|a\|_*$. Thus, when $y_{i^*(x)} = f(x)$, meaning the base classifier and the approximate direction oracle agree at x, then $\frac{|g^\mu(x)|}{\|a\|_*} = \frac{|g(x)|}{\|a\|_*} + \alpha$, so this term of the certified radius is strictly larger than the robust radius $r(x) \coloneqq \frac{|g(x)|}{\|a\|_*}$ under f. The term $\frac{\rho(x,x_j)^2 - \rho(x,x_i)^2}{2\|x_i - x_j\|_*}$ with $y_j \neq y_i$ quantifies how close x is to a data point of class different from that assigned by the approximate direction oracle. If x is sufficiently far from such data points and $y_{i^*(x)} = f(x)$, then $r^\mu_{\text{lin}}(x) = r(x) + \alpha$, so the robust radius increases by α. Of course, this need not happen in general. If x is relatively close to a data point of class $y_j \neq y_i$ or if $y_{i^*(x)} \neq f(x)$, then $r^\mu_{\text{lin}}(x)$ may be less than $r(x)$. This is expected, since in these cases the nearby data is informing us that x may not belong to class $f(x)$ predicted by the base classifier. These are the sacrificial points that may move closer to the resulting decision boundary in the name of robustifying where the data says to. Such points must exist since it is not possible to robustify everywhere simultaneously.

4. Extension to Nonlinear Classifiers

We now extend the approach to binary nonlinear classifiers. Assume that $g\colon \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and possibly nonlinear. For $x, \epsilon \in \mathbb{R}^d$, the mean value theorem gives that $g(x+\epsilon) = g(x) + \nabla g(x')^\top \epsilon$ for some x' on the line segment between x and $x+\epsilon$. By continuity of ∇g, we have that $\lim_{x' \to x} \nabla g(x') = \nabla g(x)$, so, informally, $g(x+\epsilon) \approx g(x) + \nabla g(x)^\top \epsilon$ for ϵ with small norm. Therefore, instead of using the expectation of $g(x+\epsilon)$ to define the smoothed classifier, we propose to use the expectation of $g(x) + \nabla g(x)^\top \epsilon$. In doing so, we define g^μ by $g^\mu(x) = g(x) + \nabla g(x)^\top E_{\epsilon \sim \mu_x} \epsilon$. Unlike the linear case, $g^\mu(x)$ may not equal $E_{\epsilon \sim \mu_x} g(x+\epsilon)$. This

slight modification enables us to prove certified radii while maintaining notable increases in robust accuracy in practice. When the base classifier is linear, g^μ reduces to the prior formulation (2).

Performing the same analysis as in the linear case, the smoothed classifier becomes

$$g^\mu(x) = g(x) + \alpha y_{i^*(x)} \|\nabla g(x)\|_*. \tag{5}$$

Interestingly, when $\|\cdot\| = \|\cdot\|_\infty$, the value $g^\mu(x)$ approximates the soft classification (under g) of the adversarial attack $x_{\text{fgsm}} := x + \alpha\,\text{sign}(\nabla \ell(x))$ generated by the well-known fast gradient sign method with loss $\ell(\cdot) = y_{i^*(\cdot)} g(\cdot)$ to be maximized (Goodfellow et al., 2015). High values of this particular loss are actually beneficial with respect to the given data, and therefore an alternative interpretation of our proposed method is as a preemptive "anti-attack" everywhere in the input space.

Theorem 8 below generalizes the certified radii of Theorem 6 to nonlinear base classifiers. The result uses a global Lipschitz constant of the gradient, which is easily modified to use local constants if desired (e.g., the local Lipschitz constant over a $\|\cdot\|$-norm ball at x of radius $r^\mu_{\text{nghbr}}(x)$). In general, local constants give stronger bounds but are difficult to compute. See the number of works on estimating and upper-bounding Lipschitz constants, e.g., Weng et al. (2018b); Fazlyab et al. (2019).

Assumption 7 *The gradient $\nabla g\colon \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz in norms $(\|\cdot\|, \|\cdot\|_*)$ for some $L > 0$.*

Theorem 8 *Suppose that Assumption 7 holds. Consider $x \in \mathbb{R}^d$ and fix $i = i^*(x)$. Then $f^\mu(x+\delta) = f^\mu(x)$ for all $\delta \in \mathbb{R}^d$ such that $\|\delta\| < r^\mu(x) := \min\left\{r^\mu_{bndry}(x), r^\mu_{nghbr}(x)\right\}$, where*

$$r^\mu_{bndry}(x) = \frac{\sqrt{(\alpha L + \|\nabla g(x)\|_*)^2 + 4L|g^\mu(x)|} - (\alpha L + \|\nabla g(x)\|_*)}{2L},$$
$$r^\mu_{nghbr}(x) = \min\left\{\frac{\rho(x,x_j)^2 - \rho(x,x_i)^2}{2\|x_i - x_j\|_*} : y_j \neq y_i,\ j \in \{1,2,\ldots,N\}\right\}.$$

Proof See Appendix A. ■

Similar to the linear case, the certified radius depends on two terms that, informally, quantify the distance from x to the shifted linearized decision boundary and to the nearest neighboring data point of a differing class, respectively. The certified radius is nonnegative; $r^\mu(x) \geq 0$. When g is linear, then ∇g is constant and is therefore 0-Lipschitz. In this case, Theorem 8 holds for all $L > 0$, so $\lim_{L \downarrow 0} r^\mu_{\text{bndry}}(x) = \frac{|g^\mu(x)|}{\|\nabla g(x)\|_*}$ implies that $r^\mu(x) = r^\mu_{\text{lin}}(x)$, i.e., the certified radius recovers that of Theorem 6, despite the proof for the nonlinear case involving more bounding steps.

5. Numerical Experiments

5.1. Illustrative Example

Consider the spiral dataset with test data shown in Figure 1 and a support vector machine (SVM) learned on isolated training data. Using the SVM as the base classifier, we apply locally biased randomized smoothing (with an unused subset of training data) with $\alpha \in [0, 10]$, denoted α-LBRS. The certified radius from Theorem 6 is computed at every test point using both $\|\cdot\| = \|\cdot\|_2$ and $\|\cdot\| = \|\cdot\|_\infty$. These radii are denoted by ℓ_2-avg$(r^\mu(x))$ and ℓ_∞-avg$(r^\mu(x))$, respectively. Also computed are the average true certified radii, ℓ_2- and ℓ_∞-avg$_{\text{true}}(r^\mu(x))$, which are found by setting

the certified radius to zero for test points that are classified incorrectly by f^μ. From the decision region plots in Figure 1, we see for $\alpha > 0$ that α-LBRS learns to increase the nonlinearity of the base classifier in order to enhance robustness. In contrast, standard smoothing leaves the base SVM classifier unchanged, failing to increase robustness. The average certified radii, the average true certified radii, and the clean accuracy all simultaneously increase upon applying α-LBRS (see Figure 2). We see that the α-LBRS converges pointswise to the 1-nearest neighbor (1-NN) as $\alpha \to \infty$. We will see in the next section that this is beneficial even on larger non-synthetic examples.

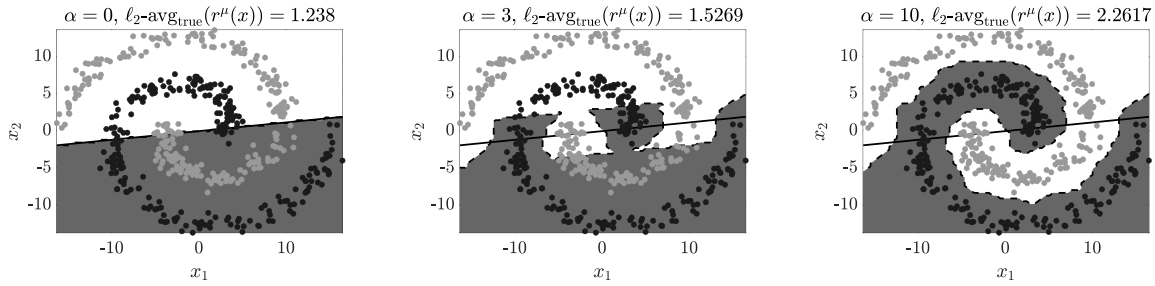

Figure 1: Test data, SVM decision boundary (bold line), and f^μ decision regions (shaded).

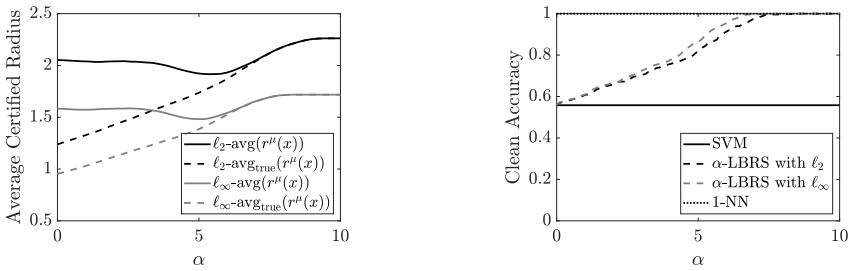

Figure 2: Average certified radius and clean accuracy for α-LBRS versus α.

5.2. Evaluating Clean and Robust Accuracy

The MNIST dataset (LeCun, 1998) is considered in a binary setting, where images with digit eight are labeled 1 and the rest are labeled -1. The training and testing data are randomly selected so that the number of data points in class -1 equals the number in class 1. Of the training data, $N = 10$ points are reserved for locally biased randomized smoothing. We train a convolutional neural network (CNN) containing three convolutional layers with ReLU activations and one fully connected layer. Using the CNN as the base classifier, we apply normal smoothing (Cohen et al., 2019) with $\sigma \in [0, 0.5]$, denoted σ-NS, and locally biased randomized smoothing with $\alpha \in [0, 1000]$, denoted α-LBRS. We also consider the 1-nearest neighbor (1-NN) using the N reserved training data points.

The accuracy of each model is computed on the test set as well as on an adversarially attacked version of the test set using a 10-step ℓ_2-PGD attack (Madry et al., 2018) with attack radius $\epsilon \in \{0.5, 1\}$, and the results are shown in Figure 3. We see that, although σ-NS achieves good robustification for small σ, the accuracy rapidly degrades to that of a constant function (0.5 for this binary problem) as σ increases. On the other hand, α-LBRS converges to the accuracy of the 1-NN

as $\alpha \to \infty$. The 1-NN is seen to be robust against the attacks (which may in part be due to the fact that the attacks are designed for the base CNN classifier—a benefit to the defender from using smoothing at test time), and therefore α-LBRS inherits this robustness for sufficiently large α.

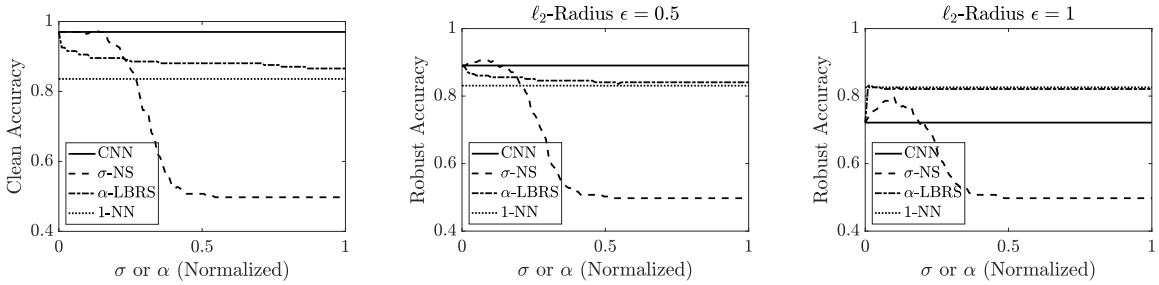

Figure 3: Clean and robust accuracy versus smoothing parameter σ or α.

Next, we fix the parameters $\sigma = 0.05$ and $\alpha = 10$ near the "corners" in Figure 3 (recall that the abscissa was normalized) that yield both good clean accuracy and good robust accuracy for $\epsilon \in \{0.5, 1\}$. We attack these models with the wider range of ℓ_2-radii $\epsilon \in [0, 3]$ and find that α-LBRS maintains its resistance to larger attacks for longer than σ-NS does—the accuracy of σ-NS degrades at a faster rate—see Figure 4. We also demonstrate the generality of our method by considering the same experiment using $\|\cdot\| = \|\cdot\|_\infty$ along with ℓ_∞-PGD attacks. Normal smoothing is not catered towards ℓ_∞-attacks, which explains the performance increase of α-LBRS over σ-NS relative to the CNN for this attack when compared to the ℓ_2-attack.

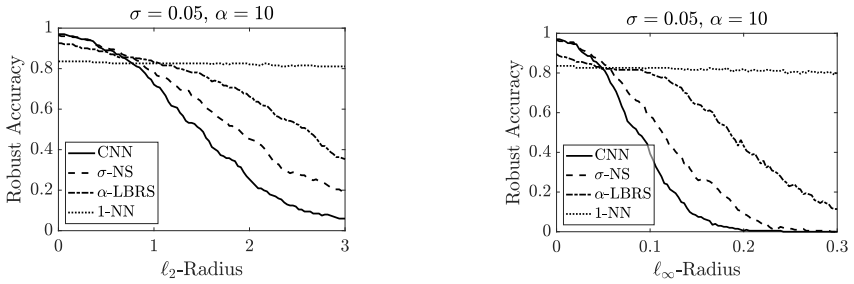

Figure 4: Robust accuracy versus attack radius.

We run the same experiments on CIFAR-10 (Krizhevsky et al., 2009) and arrive at the same conclusions, albeit with generally lower accuracies and higher sensitivities to attacks.

6. Conclusions

Locally biased randomized smoothing is introduced to learn locally optimal robustification of a classifier's decision boundary based on data. The method directly induces robustness without relying on the surrogate notion of smoothness, in contrast to existing methods. Certified robust radii are proved for the binary setting, and experiments show increased certified, clean, and robust accuracy over conventional randomized smoothing. Directions for future research include extending the approach to the multiclass setting, and studying alternate direction oracles.

Appendix A. Proofs

Theorem 6 *Consider $x \in \mathbb{R}^d$ and fix $i = i^*(x)$. Then $f^\mu(x+\delta) = f^\mu(x)$ for all $\delta \in \mathbb{R}^d$ such that*

$$\|\delta\| < r^\mu_{lin}(x) \coloneqq \min\left\{\frac{|g^\mu(x)|}{\|a\|_*}, \min\left\{\frac{\rho(x,x_j)^2 - \rho(x,x_i)^2}{2\|x_i - x_j\|_*} : y_j \neq y_i,\ j \in \{1,2,\dots,N\}\right\}\right\}.$$

Proof Let $\delta \in \mathbb{R}^d$ be such that $\|\delta\| < r^\mu_{\text{lin}}(x)$. Since $\|\delta\| \leq \frac{\rho(x,x_j)^2 - \rho(x,x_i)^2}{2\|x_i - x_j\|_*}$ for all $j \in \{1,2,\dots,N\}$ such that $y_j \neq y_i$, it holds that

$$\begin{aligned}
\rho(x+\delta,x_j)^2 - \rho(x+\delta,x_i)^2 &= (x+\delta-x_j)^\top(x+\delta-x_j) - (x+\delta-x_i)^\top(x+\delta-x_i) \\
&= 2(x_i-x_j)^\top(x+\delta) + \|x_j\|_2^2 - \|x_i\|_2^2 \\
&= 2x_i^\top x - 2x_j^\top x + \|x_j\|_2^2 - \|x_i\|_2^2 + 2(x_i-x_j)^\top\delta \\
&= (x-x_j)^\top(x-x_j) - (x-x_i)^\top(x-x_i) + 2(x_i-x_j)^\top\delta \\
&= \rho(x,x_j)^2 - \rho(x,x_i)^2 + 2(x_i-x_j)^\top\delta \\
&\geq \rho(x,x_j)^2 - \rho(x,x_i)^2 - 2|(x_i-x_j)^\top\delta| \\
&\geq \rho(x,x_j)^2 - \rho(x,x_i)^2 - 2\|x_i-x_j\|_*\|\delta\| \\
&\geq 0
\end{aligned}$$

for all such j. Hence, $\rho(x+\delta,x_i) \leq \rho(x+\delta,x_j)$ for all j such that $y_j \neq y_i$. Therefore, it must be that $i^*(x+\delta) \in \{j \in \{1,2,\dots,N\} : y_j = y_i\}$, and hence $y_{i^*(x+\delta)} = y_i$. Therefore,

$$\begin{aligned}
g^\mu(x+\delta) &= g(x+\delta) + \alpha y_{i^*(x+\delta)}\|a\|_* \\
&= a^\top(x+\delta) + b + \alpha y_i\|a\|_* \\
&= g(x) + \alpha y_i\|a\|_* + a^\top\delta \\
&= g^\mu(x) + a^\top\delta.
\end{aligned}$$

This gives that

$$|g^\mu(x+\delta) - g^\mu(x)| = |a^\top\delta| \leq \|a\|_*\|\delta\| < |g^\mu(x)|,$$

where the last inequality follows from the fact that $\|\delta\| < \frac{|g^\mu(x)|}{\|a\|_*}$. Therefore,

$$-|g^\mu(x)| < g^\mu(x+\delta) - g^\mu(x) < |g^\mu(x)|, \tag{6}$$

so

$$g^\mu(x) - |g^\mu(x)| < g^\mu(x+\delta) < g^\mu(x) + |g^\mu(x)|.$$

If $g^\mu(x) \geq 0$, then the left-hand inequality gives that $0 = g^\mu(x) - |g^\mu(x)| < g^\mu(x+\delta)$, whereas if $g^\mu(x) < 0$, then the right-hand inequality gives that $g^\mu(x+\delta) < g^\mu(x) + |g^\mu(x)| = 0$. In both cases, $\operatorname{sign}(g^\mu(x+\delta)) = \operatorname{sign}(g^\mu(x))$, which proves that $f^\mu(x+\delta) = f^\mu(x)$, as desired. ■

Assumption 7 *The gradient $\nabla g\colon \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz in norms $(\|\cdot\|, \|\cdot\|_*)$ for some $L > 0$.*

Theorem 8 *Suppose that Assumption 7 holds. Consider $x \in \mathbb{R}^d$ and fix $i = i^*(x)$. Then $f^\mu(x+\delta) = f^\mu(x)$ for all $\delta \in \mathbb{R}^d$ such that $\|\delta\| < r^\mu(x) \coloneqq \min\left\{r^\mu_{bndry}(x), r^\mu_{nghbr}(x)\right\}$, where*

$$r^\mu_{bndry}(x) = \frac{\sqrt{(\alpha L + \|\nabla g(x)\|_*)^2 + 4L|g^\mu(x)|} - (\alpha L + \|\nabla g(x)\|_*)}{2L},$$
$$r^\mu_{nghbr}(x) = \min\left\{\frac{\rho(x,x_j)^2 - \rho(x,x_i)^2}{2\|x_i - x_j\|_*} : y_j \neq y_i,\ j \in \{1,2,\dots,N\}\right\}.$$

Proof Let $\delta \in \mathbb{R}^d$ be such that $\|\delta\| < r^\mu(x)$. Since $\|\delta\| \leq \frac{\rho(x,x_i)^2 - \rho(x,x_j)^2}{2\|x_i - x_j\|_*}$ for all $j \in \{1,2,\dots,N\}$ such that $y_j \neq y_i$, it holds that

$$\begin{aligned}
\rho(x+\delta,x_j)^2 - \rho(x+\delta,x_i)^2 &= (x+\delta-x_j)^\top(x+\delta-x_j) - (x+\delta-x_i)^\top(x+\delta-x_i)\\
&= 2(x_i-x_j)^\top(x+\delta) + \|x_j\|_2^2 - \|x_i\|_2^2\\
&= 2x_i^\top x - 2x_j^\top x + \|x_j\|_2^2 - \|x_i\|_2^2 + 2(x_i-x_j)^\top\delta\\
&= (x-x_j)^\top(x-x_j) - (x-x_i)^\top(x-x_i) + 2(x_i-x_j)^\top\delta\\
&= \rho(x,x_j)^2 - \rho(x,x_i)^2 + 2(x_i-x_j)^\top\delta\\
&\geq \rho(x,x_j)^2 - \rho(x,x_i)^2 - 2|(x_i-x_j)^\top\delta|\\
&\geq \rho(x,x_j)^2 - \rho(x,x_i)^2 - 2\|x_i-x_j\|_*\|\delta\|\\
&\geq 0
\end{aligned}$$

for all such j. Hence, $\rho(x+\delta,x_i) \leq \rho(x+\delta,x_j)$ for all j such that $y_j \neq y_i$. Therefore, it must be that $i^*(x+\delta) \in \{j \in \{1,2,\dots,N\} : y_j = y_i\}$, and hence $y_{i^*(x+\delta)} = y_i$. Therefore,

$$\begin{aligned}
g^\mu(x+\delta) &= g(x+\delta) + \alpha y_{i^*(x+\delta)}\|\nabla g(x+\delta)\|_*\\
&= g(x) + \nabla g(x')^\top\delta + \alpha y_i\|\nabla g(x+\delta)\|_*
\end{aligned}$$

for some x' on the line segment between x and $x+\delta$. This gives that

$$\begin{aligned}
|g^\mu(x+\delta) - g^\mu(x)| &= |g^\mu(x+\delta) - g(x) - \alpha y_i\|\nabla g(x)\|_*|\\
&= \left|\nabla g(x')^\top\delta + \alpha y_i\|\nabla g(x+\delta)\|_* - \alpha y_i\|\nabla g(x)\|_*\right|\\
&= \left|\nabla g(x)^\top\delta + \left(\nabla g(x') - \nabla g(x)\right)^\top\delta + \alpha y_i\left(\|\nabla g(x+\delta)\|_* - \|\nabla g(x)\|_*\right)\right|\\
&\leq \left|\nabla g(x)^\top\delta\right| + \left|\left(\nabla g(x') - \nabla g(x)\right)^\top\delta\right|\\
&\quad + \alpha\left|\|\nabla g(x+\delta)\|_* - \|\nabla g(x)\|_*\right|\\
&\leq \|\nabla g(x)\|_*\|\delta\| + \|\nabla g(x') - \nabla g(x)\|_*\|\delta\|\\
&\quad + \alpha\left|\|\nabla g(x+\delta)\|_* - \|\nabla g(x)\|_*\right|\\
&\leq \|\nabla g(x)\|_*\|\delta\| + L\|x'-x\|\|\delta\|\\
&\quad + \alpha\left|\|\nabla g(x+\delta)\|_* - \|\nabla g(x)\|_*\right|\\
&\leq \|\nabla g(x)\|_*\|\delta\| + L\|\delta\|^2\\
&\quad + \alpha\left|\|\nabla g(x+\delta)\|_* - \|\nabla g(x)\|_*\right|.
\end{aligned}$$

To bound the last term, note that $-\|\nabla g(x+\delta) - \nabla g(x)\|_* \leq \|\nabla g(x+\delta)\|_* - \|\nabla g(x)\|_* \leq \|\nabla g(x+\delta) - \nabla g(x)\|_*$, so $|\|\nabla g(x+\delta)\|_* - \|\nabla g(x)\|_*| \leq \|\nabla g(x+\delta) - \nabla g(x)\|_* \leq L\|\delta\|$. Thus,

$$|g^\mu(x+\delta) - g^\mu(x)| \leq (\alpha L + \|\nabla g(x)\|_*)\,\|\delta\| + L\|\delta\|^2.$$

Since $\|\delta\| < r^\mu_{\text{bndry}}(x)$, this gives that

$$|g^\mu(x+\delta) - g^\mu(x)| \leq |g^\mu(x)|.$$

The remainder of the proof is identical to that of Theorem 6 from (6) onwards. ■

References

Motasem Alfarra, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Data dependent randomized smoothing. *arXiv preprint arXiv:2012.04351*, 2020.

Brendon G. Anderson, Ziye Ma, Jingqi Li, and Somayeh Sojoudi. Tightened convex relaxations for neural network robustness certification. In *Proceedings of the 59th IEEE Conference on Decision and Control*, 2020.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402. Springer, 2013.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Chen Chen, Kezhi Kong, Peihong Yu, Juan Luque, Tom Goldstein, and Furong Huang. Instars: Instance-wise randomized smoothing for improved robustness and accuracy. *arXiv preprint arXiv:2103.04436*, 2021.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020.

Francisco Eiras, Motasem Alfarra, M Pawan Kumar, Philip HS Torr, Puneet K Dokania, Bernard Ghanem, and Adel Bibi. Ancer: Anisotropic certification via sample-wise volume maximization. *arXiv preprint arXiv:2107.04570*, 2021.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2018.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 11427–11438, 2019.

Mahyar Fazlyab, Manfred Morari, and George J Pappas. Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. *IEEE Transactions on Automatic Control*, 2020.

Gerald B Folland. *Real Analysis: Modern Techniques and Their Applications*, volume 40. John Wiley & Sons, 1999.

Yue Gao, Harrison Rosenberg, Kassem Fawaz, Somesh Jha, and Justin Hsu. Analyzing accuracy loss in randomized smoothing defenses. *arXiv preprint arXiv:2003.01595*, 2020.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Vishaal Krishnan, Al Makdah, Abed AlRahman, and Fabio Pasqualetti. Lipschitz bounds and provably robust training by laplacian smoothing. In *Advances in Neural Information Processing Systems*, volume 33, pages 10924–10935, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical report*, 2009.

Aounon Kumar, Alexander Levine, Soheil Feizi, and Tom Goldstein. Certifying confidence via randomized smoothing. In *Advances in Neural Information Processing Systems*, volume 33, pages 5165–5177, 2020.

Aounon Kumar, Alexander Levine, and Soheil Feizi. Policy smoothing for provably robust reinforcement learning. *arXiv preprint arXiv:2106.11420*, 2021.

Yann LeCun. The mnist database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy*, pages 656–672. IEEE, 2019.

Alexander Levine and Soheil Feizi. Wasserstein smoothing: Certified robustness against Wasserstein adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 3938–3947. PMLR, 2020.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Olvi L Mangasarian. Arbitrary-norm separating plane. *Operations Research Letters*, 24(1-2):15–23, 1999.

Jeet Mohapatra, Ching-Yun Ko, Lily Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Hidden cost of randomized smoothing. In *International Conference on Artificial Intelligence and Statistics*, pages 4033–4041. PMLR, 2021.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10877–10887, 2018.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Peter Súkeník, Aleksei Kuvshinov, and Stephan Günnemann. Intriguing properties of input-dependent randomized smoothing. *arXiv preprint arXiv:2110.05365*, 2021.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Jiaye Teng, Guang-He Lee, and Yang Yuan. ℓ_1 adversarial robustness certificates: a randomized smoothing approach. *Preprint*, 2020. URL https://openreview.net/forum?id=H1lQIgrFDS.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Lei Wang, Runtian Zhai, Di He, Liwei Wang, and Li Jian. Pretrain-to-finetune adversarial training via sample-wise randomized smoothing. *Preprint*, 2021. URL https://openreview.net/pdf?id=Te1aZ2myPIu.

Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. In *International Conference on Machine Learning*, pages 5133–5142. PMLR, 2018.

Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for ReLU networks. In *International Conference on Machine Learning*, pages 5276–5285. PMLR, 2018a.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations*, 2018b.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

Bichen Wu, Forrest Iandola, Peter H Jin, and Kurt Keutzer. SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 129–137, 2017.

Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. Crop: Certifying robust policies for reinforcement learning through functional smoothing. *arXiv preprint arXiv:2106.09292*, 2021.

Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR, 2020a.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In *Advances in Neural Information Processing Systems*, volume 33, pages 8588–8601, 2020b.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020.

Dinghuai Zhang, Mao Ye, Chengyue Gong, Zhanxing Zhu, and Qiang Liu. Black-box certification with randomized smoothing: A functional optimization based framework. In *Advances in Neural Information Processing Systems*, volume 33, pages 2316–2326, 2020.