Center Smoothing: Provable Robustness for Functions with Metric-Space Outputs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Randomized smoothing has been successfully applied to classification tasks on high-dimensional inputs, such as images, to obtain models that are provably robust against adversarial perturbations of the input. We extend this technique to produce provable robustness for functions that map inputs into an arbitrary metric space rather than discrete classes. Such functions are used in many machine learning problems like image reconstruction, dimensionality reduction, facial recognition, etc. Our robustness certificates guarantee that the change in the output of the smoothed model as measured by the distance metric remains small for any norm-bounded perturbation of the input. We can certify robustness under a variety of different output metrics, such as total variation distance, Jaccard distance, norm-based metrics, etc. In our experiments, we apply our procedure to create certifiably robust models with disparate output spaces – from sets to images – and show that it yields meaningful certificates without significantly degrading the performance of the base model.

1 Introduction

The study of adversarial robustness in machine learning has gained a lot of attention ever since deep neural networks (DNNs) have been demonstrated to be vulnerable to adversarial attacks. They are tiny perturbations of the input that can completely alter a model's predictions [46, 36, 16, 25]. These maliciously chosen perturbations can significantly degrade the performance of a model, like an image classifier, and make it output almost any class that the attacker wants. However, these attacks are not just limited to classification problems. Recently, they have also been shown to exist for DNN-based models with many different kinds of outputs like images, probability distributions, sets, etc. For instance, facial recognition systems can be deceived to evade detection, impersonate authorized individuals and even render them completely ineffective [48, 45, 13]. Image reconstruction models have been targeted to introduce unwanted artefacts or miss important details, such as tumors in MRI scans, through adversarial inputs [1, 40, 5, 6]. Similarly, super-resolution systems can be made to generate distorted images that can in turn deteriorate the performance of subsequent tasks that rely on the high-resolution outputs [8, 52]. Deep neural network based policies in reinforcement learning problems also have been shown to succumb to imperceptible perturbations in the state observations [14, 21, 2, 38]. Such widespread presence of adversarial attacks is concerning as it threatens the use of deep neural networks in critical systems, such as facial recognition, self-driving vehicles, medical diagnosis, etc., where safety, security and reliability are of utmost importance.

Adversarial defenses have mostly focused on classification tasks [24, 3, 19, 11, 34, 18, 15]. Provable defenses based on convex-relaxation [50, 39, 43, 7, 44], interval-bound propagation [17, 20, 12, 37] and randomized smoothing [9, 26, 32, 41] that guarantee that the predicted class will remain the same in a certified region around the input point have also been studied. Among these approaches

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

randomized smoothing scales up to high-dimensional inputs, such as images, and does not need access to or make assumptions about the underlying model. The robustness certificates produced are probabilistic, meaning that they hold with high probability. First studied by Cohen et al. in [9], smoothing methods sample a set of points in a Gaussian cloud around an input, and aggregate the predictions of the classifier on these points to generate the final output.

While accuracy is the standard quality measure for classification, more complex tasks may require other quality metrics like total variation for images, intersection over union for object localization, earth-mover distance for distributions, etc. In general, networks can be cast as functions of the type $f : \mathbb{R}^k \to (M, d)$ which map a k dimensional real-valued space into a metric space M with distance function $d : M \times M \to \mathbb{R}_{\geq 0}$. In this work, we extend randomized smoothing to obtain provable robustness for functions that map into arbitrary metric spaces. We generate a robust version $\bar{f}$ such that the change in its output, as measured by d, is small for a small change in its input. More formally, given an input x and an ℓ_2-perturbation size ϵ_1, we produce a value ϵ_2 with the guarantee that, with high probability,

$$\forall x' \text{ s.t. } \|x - x'\|_2 \leq \epsilon_1,\ d(\bar{f}(x), \bar{f}(x')) \leq \epsilon_2.$$

Our contributions: We develop *center smoothing*, a technique to make functions like f provably robust against adversarial attacks. For a given input x, center smoothing samples a collection of points in the neighborhood of x using a Gaussian smoothing distribution, computes the function f on each of these points and returns the center of the smallest ball enclosing at least half the points in the output space (see figure 1). Computing the minimum enclosing ball in the output space is equivalent to solving the 1-center problem with outliers (hence the name of our procedure), which is an NP-complete problem for a general metric [42]. We approximate it by computing the point that has the smallest median distance to all the other points in the sample. We show that the output of the smoothed function is robust to input perturbations of bounded ℓ_2-size. Although we defined the output space as a metric, our proofs only require the symmetry property and triangle inequality to hold. Thus, center smoothing can also be applied to pseudometric distances that need not satisfy the identity of indiscernibles. Many distances defined for images, such as total variation, cosine distance, perceptual distances, etc., fall under this category. Center smoothing steps outside the world of ℓ_p metrics, and certifies robustness in metrics like IoU/Jaccard distance for object localization, and total-variation, which is a good measure of perceptual similarity for images. In our experiments, we show that this method can produce meaningful certificates for a wide variety of output metrics without significantly compromising the quality of the base model.

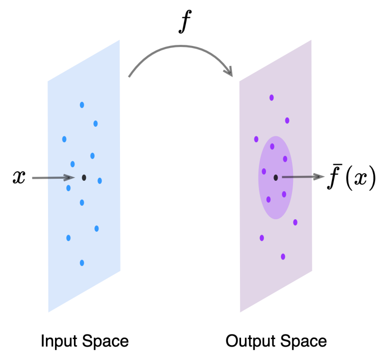


Figure 1: Center smoothing.

Related Work: Randomized smoothing has been extensively studied for classification problems to obtain provably robust models against many different ℓ_p [9, 26, 41, 47, 33, 31, 27, 30] and non-ℓ_p [28, 29] threat models. Beyond classification tasks, it has also been used for certifying the median output of regression models [51] and the expected softmax scores of neural networks [23]. Smoothing a vector-valued function by taking the mean of the output vectors has been shown to have a bounded Lipschitz constant when both input and output spaces are ℓ_2-metrics [49]. However, existing methods do not generate the type of certificates described above for general distance metrics. Center smoothing takes the distance function of the output space into account for generating the robust output and thus results in a more natural smoothing procedure for the specific distance metric.

2 Preliminaries and Notations

Given a function $f : \mathbb{R}^k \to (M, d)$ and a distribution $\mathcal{D}$ over the input space $\mathbb{R}^k$, let $f(\mathcal{D})$ denote the probability distribution of the output of f in M when the input is drawn from $\mathcal{D}$. For a point $x \in \mathbb{R}^k$, let $x + \mathcal{P}$ denote the probability distribution of the points $x + \delta$ where δ is a smoothing noise drawn from a distribution $\mathcal{P}$ over $\mathbb{R}^k$ and let X be the random variable for $x + \mathcal{P}$. For elements in M, define $\mathcal{B}(z, r) = \{z' \mid d(z, z') \leq r\}$ as a ball of radius r centered at z. Define a smoothed version of f under $\mathcal{P}$ as the center of the ball with the smallest radius in M that encloses at least half of the probability mass of $f(x + \mathcal{P})$, i.e.,

$$\bar{f}_{\mathcal{P}}(x) = \underset{z}{\text{argmin}}\ r \text{ s.t. } \mathbb{P}[f(X) \in \mathcal{B}(z, r)] \geq \frac{1}{2}.$$

If there are multiple balls with the smallest radius satisfying the above condition, return one of the centers arbitrarily. Let $r^*_{\mathcal{P}}(x)$ be the value of the minimum radius. Hereafter, we ignore the subscripts and superscripts in the above definitions whenever they are obvious from context. In this work, we sample the noise vector δ from an i.i.d Gaussian distribution of variance σ^2 in each dimension, i.e., $\delta \sim \mathcal{N}(0, \sigma^2 I)$.

2.1 Gaussian Smoothing

Cohen et al. in 2019 showed that a classifier $h : \mathbb{R}^k \to \mathcal{Y}$ smoothed with a Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ as,

$$\bar{h}(x) = \underset{c \in \mathcal{Y}}{\text{argmax}} \; \mathbb{P}\left[h(x + \delta) = c\right],$$

where $\mathcal{Y}$ is a set of classes, is certifiably robust to small perturbations in the input. Their certificate relied on the fact that, if the probability of sampling from the top class at x under the smoothing distribution is p, then for an ℓ_2 perturbation of size at most ϵ, the probability of the top class is guaranteed to be at least

$$p_\epsilon = \Phi(\Phi^{-1}(p) - \epsilon/\sigma), \tag{1}$$

where Φ is the CDF of the standard normal distribution $\mathcal{N}(0, 1)$. This bound applies to any $\{0, 1\}$-function over the input space $\mathbb{R}^k$, i.e., if $\mathbb{P}[h(x) = 1] = p$, then for any ϵ-size perturbation x', $\mathbb{P}[h(x') = 1] \geq p_\epsilon$.

We use this bound to generate robustness certificates for center smoothing. We identify a ball $\mathcal{B}(\bar{f}(x), R)$ of radius R enclosing a very high probability mass of the output distribution. One can define a function that outputs one if f maps a point to inside $\mathcal{B}(\bar{f}(x), R)$ and zero otherwise. The bound in (1) gives us a region in the input space such that for any point inside it, at least half of the mass of the output distribution is enclosed in $\mathcal{B}(\bar{f}(x), R)$. We show in section 3 that the output of the smoothed function for a perturbed input is guaranteed to be within a constant factor of R from the output of the original input.

3 Center Smoothing

As defined in section 2, the output of $\bar{f}$ is the center of the smallest ball in the output space that encloses at least half the probability mass of the $f(x + \mathcal{P})$. Thus, in order to significantly change the output, an adversary has to find a perturbation such that a majority of the neighboring points map far away from $\bar{f}(x)$. However, for a function that is roughly accurate on most points around x, a small perturbation in the input cannot change the output of the smoothed function by much, thereby making it robust.

For an ℓ_2 perturbation size of ϵ_1 of an input point x, let R be the radius of a ball around $\bar{f}(x)$ that encloses more than half the probability mass of $f(x' + \mathcal{P})$ for all x' satisfying $\|x - x'\|_2 \leq \epsilon_1$, i.e.,

$$\forall x' \text{ s.t. } \|x - x'\|_2 \leq \epsilon_1, \; \mathbb{P}[f(X') \in \mathcal{B}(\bar{f}(x), R)] > \frac{1}{2}, \tag{2}$$

where $X' \sim x' + \mathcal{P}$. Basically, R is the radius of a ball around $\bar{f}(x)$ that contains at least half the probability mass of $f(x' + \mathcal{P})$ for any ϵ_1-size perturbation x' of x. Then, we have the following robustness guarantee on $\bar{f}$:

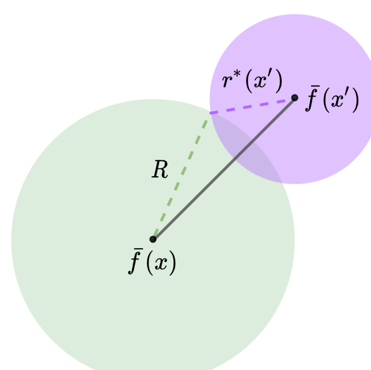


Figure 2: Robustness guarantee.

Theorem 1. *For all x' such that $\|x - x'\|_2 \leq \epsilon_1$,*

$$d(\bar{f}(x), \bar{f}(x')) \leq 2R.$$

Proof. Consider the balls $\mathcal{B}(\bar{f}(x'), r^*(x'))$ and $\mathcal{B}(\bar{f}(x), R)$ (see figure 2). From the definition of $r^*(x')$ and R, we know that the sum of the probability masses of $f(x' + \mathcal{P})$ enclosed by the two balls must be strictly greater than one. Thus, they must have an element y in common. Since d satisfies the triangle inequality, we have:

$$\begin{aligned} d(\bar{f}(x), \bar{f}(x')) &\leq d(\bar{f}(x), y) + d(y, \bar{f}(x')) \\ &\leq R + r^*(x'). \end{aligned}$$

Since, the ball $\mathcal{B}(\bar{f}(x), R)$ encloses more than half of the probability mass of $f(x+\mathcal{P})$, the minimum ball with at least half the probability mass cannot have a radius greater than R, i.e., $r^*(x') \leq R$. Therefore, $d(\bar{f}(x), \bar{f}(x')) \leq 2R$. $\square$

The above result, in theory, gives us a smoothed version of f with a provable guarantee of robustness. However, in practice, it may not be feasible to obtain $\bar{f}$ just from samples of $f(x+\mathcal{P})$. Instead, we will use some procedure that approximates the smoothed output with high probability. For some $\Delta \in [0, 1/2]$, let $\hat{r}(x, \Delta)$ be the radius of the smallest ball that encloses at least $1/2 + \Delta$ probability mass of $f(x+\mathcal{P})$, i.e.,

$$\hat{r}(x, \Delta) = \min_{z'} r \text{ s.t. } \mathbb{P}[f(X) \in \mathcal{B}(z', r)] \geq \frac{1}{2} + \Delta.$$

Now define a probabilistic approximation $\hat{f}(x)$ of the smoothed function $\bar{f}$ to be a point $z \in M$, which with probability at least $1 - \alpha_1$ (for $\alpha_1 \in [0, 1]$), encloses at least $1/2 - \Delta$ probability mass of $f(x+\mathcal{P})$ within a ball of radius $\hat{r}(x, \Delta)$. Formally, $\hat{f}(x)$ is a point $z \in M$, such that, with at least $1 - \alpha_1$ probability,

$$\mathbb{P}\left[f(X) \in \mathcal{B}(z, \hat{r}(x, \Delta))\right] \geq \frac{1}{2} - \Delta.$$

Defining $\hat{R}$ to be the radius of a ball centered at $\hat{f}(x)$ that satisfies:

$$\forall x' \text{ s.t. } \|x - x'\|_2 \leq \epsilon_1,\ \mathbb{P}[f(X') \in \mathcal{B}(\hat{f}(x), \hat{R})] > \frac{1}{2} + \Delta, \tag{3}$$

we can write a probabilistic version of theorem 1,

Theorem 2. *With probability at least* $1 - \alpha_1$*,*

$$\forall x' \textit{ s.t. } \|x - x'\|_2 \leq \epsilon_1,\ d(\hat{f}(x), \hat{f}(x')) \leq 2\hat{R},$$

The proof of this theorem is in the appendix, and logically parallels the proof of theorem 1.

3.1 Computing $\hat{f}$

For an input x and a given value of Δ, sample n points independently from a Gaussian cloud $x + \mathcal{N}(0, \sigma^2 I)$ around the point x and compute the function f on each of these points. Let $Z = \{z_1, z_2, \ldots, z_n\}$ be the set of n samples of $f(x+\mathcal{N}(0, \sigma^2 I))$ produced in the output space. Compute the minimum enclosing ball $\mathcal{B}(z, r)$ that contains at least half of the points in Z. The following lemma bounds the radius r of this ball by the radius of the smallest ball enclosing at least $1/2 + \Delta_1$ probability mass of the output distribution (proof in appendix).

Lemma 1. *With probability at least* $1 - e^{-2n\Delta_1^2}$*,*

$$r \leq \hat{r}(x, \Delta_1).$$

Now, sample a fresh batch of n random points and compute the $1 - e^{-2n\Delta_1^2}$ probability Hoeffding lower-bound p_{Δ_1} of the probability mass enclosed inside $\mathcal{B}(z, r)$ by conting the number of points that fall inside the ball, i.e., calculate the p_{Δ_1} for which, with probability at least $1 - e^{-2n\Delta_1^2}$,

$$\mathbb{P}\left[f(X) \in \mathcal{B}(z, r)\right] \geq p_{\Delta_1}.$$

Let $\Delta_2 = 1/2 - p_{\Delta_1}$. If $\max(\Delta_1, \Delta_2) \leq \Delta$, the point z satisfies the conditions in the definition of $\hat{f}$, with at least $1 - 2e^{-2n\Delta_1^2}$ probability. If $\max(\Delta_1, \Delta_2) > \Delta$, discard the computed center z and abstain. In our experiments, we select Δ_1, n and α_1 appropriately so that the above process succeeds easily.

Computing the minimum enclosing ball $\mathcal{B}(z, r)$ exactly can be computationally challenging, as for certain norms, it is known to be NP-complete [42]. Instead, we approximate it by computing a ball β-MEB$(Z, 1/2)$ that contains at least half the points in Z, but has a radius that is within βr units of the optimal radius, for a constant β. We modify theorem 1 to account for this approximation (see appendix for proof).

Algorithm 1 Smooth

Input: $x \in \mathbb{R}^k, \sigma, \Delta, \alpha_1$.
Output: $z \in M$.
Set $Z = \{z_i\}_{i=1}^m$ s.t. $z_i \sim f(x + \mathcal{N}(0, \sigma^2 I))$.
Set $\Delta_1 = \sqrt{\ln(2/\alpha_1)/2n}$.
Compute $z = \beta$-MEB$(Z, 1/2)$.
Re-sample Z.
Compute p_{Δ_1}.
Set $\Delta_2 = 1/2 - p_{\Delta_1}$.
If $\Delta < \max(\Delta_1, \Delta_2)$, discard z and abstain.

Algorithm 2 Certify

Input: $x \in \mathbb{R}^k, \epsilon_1, \sigma, \Delta, \alpha_1, \alpha_2$.
Output: $\epsilon_2 \in \mathbb{R}$.
Compute $\hat{f}(x)$ using algorithm 1.
Set $Z = \{z_i\}_{i=1}^m$ s.t. $z_i \sim f(x + \mathcal{N}(0, \sigma^2 I))$.
Compute $\tilde{\mathcal{R}} = \{d(\hat{f}(x), f(z_i)) \mid z_i \in Z\}$.
Set $p = \Phi(\Phi^{-1}(1/2 + \Delta) + \epsilon_1/\sigma)$.
Set $q = p + \sqrt{\ln(1/\alpha_2)/2m}$.
Set $\hat{R} = q$th-quantile of $\tilde{\mathcal{R}}$.
Set $\epsilon_2 = (1+\beta)\hat{R}$.

Theorem 3. *With probability at least* $1 - \alpha_1$,

$$\forall x' \text{ s.t. } \|x - x'\|_2 \le \epsilon_1,\ d(\hat{f}(x), \hat{f}(x')) \le (1+\beta)\hat{R}$$

where $\alpha_1 = 2e^{-2n\Delta_1^2}$.

We use a simple approximation that works for all metrics and achieves an approximation factor of two, producing a certified radius of $3\hat{R}$. It computes a point from the set Z, instead of a general point in M, that has the minimum median distance from all the points in the set (including itself). This can be achieved in $O(n^2)$ steps. To see how the factor 2-approximation is achieved, consider the optimal ball with radius r. Each pair of points is at most $2r$ distance from each other. Thus, a ball with radius $2r$, centered at one of these points will cover every other point in the optimal ball. Better approximations can be obtained for specific norms, e.g., there exists a $(1+\epsilon)$-approximation algorithm for the ℓ_2 norm [4]. For graph distances, the optimal radius can be computed exactly using the above algorithm. The smoothing procedure is outlined in algorithm 1.

3.2 Certifying $\hat{f}$

Given an input x, compute $\hat{f}(x)$ as described above. Now, we need to compute a radius $\hat{R}$ that satisfies condition 3. As per bound 1, in order to maintain a probability mass of at least $1/2 + \Delta$ for any ϵ_1-size perturbation of x, the ball $\mathcal{B}(\hat{f}(x), \hat{R})$ must enclose at least

$$p = \Phi\left(\Phi^{-1}\left(\frac{1}{2} + \Delta\right) + \frac{\epsilon_1}{\sigma}\right) \tag{4}$$

probability mass of $f(x + \mathcal{P})$. Again, just as in the case of estimating $\bar{f}$, we may only compute $\hat{R}$ from a finite number of samples m of the distribution $f(x + \mathcal{P})$. For each sample $z_i \sim x + \mathcal{P}$, we compute the distance $d(\hat{f}(x), f(z_i))$ and set $\hat{R}$ to be the qth-quantile $\tilde{R}_q$ of these distances for a q that is slightly greater than p (see equation 5 below). The qth-quantile $\tilde{R}_q$ is a value larger than at least q fraction of the samples. We set q as,

$$q = p + \sqrt{\frac{\ln(1/\alpha_2)}{2m}}, \tag{5}$$

for some small $\alpha_2 \in [0, 1]$. This guarantees that, with high probability, the ball $\mathcal{B}(\hat{f}(x), \tilde{R}_q)$ encloses at least p fraction of the probability mass of $f(x + \mathcal{P})$. We prove the following lemma by bounding the cumulative distribution function of the distances of $f(z_i)$s from $\hat{f}(x)$ using the Dvoretzky–Kiefer–Wolfowitz inequality.

Lemma 2. *With probability* $1 - \alpha_2$,

$$\mathbb{P}\left[f(X) \in \mathcal{B}(\hat{f}(x), \tilde{R}_q)\right] > p$$

Combining with theorem 3, we have the final certificate:

$$\forall x' \text{ s.t. } \|x - x'\|_2 \le \epsilon_1,\ d(\hat{f}(x), \hat{f}(x')) \le (1+\beta)\hat{R},$$

with probability at least $1 - \alpha$, for $\alpha = \alpha_1 + \alpha_2$. In our experiments, we set $\alpha_1 = \alpha_2 = 0.005$ to achieve an overall success probability of $1 - \alpha = 0.99$, and calculate the required Δ and q values accordingly. We use a $\beta = 2$-approximation for computing the minimum enclosing ball in the smoothing step. Algorithm 2 provides the pseudocode for the certification procedure.

4 Relaxing Metric Requirements

Although we defined our procedure for metric outputs, our analysis does not critically use all the properties of a metric. For instance, we do not require $d(z_1, z_2)$ to be strictly greater than zero for $z_1 \neq z_2$. An example of such a distance measure is the total variation distance that returns zero for two vectors that differ by a constant amount on each coordinate. Our proofs do implicitly use the symmetry property, but asymmetric distances can be converted to symmetric ones by taking the sum or the max of the distances in either directions. Perhaps the most important property of metrics that we use is the triangle inequality as it is critical for the robustness guarantee of the smoothed function. However, even this constraint may be partially relaxed. It is sufficient for the distance function d to satisfy the triangle inequality approximately, i.e., $d(a, c) \leq \gamma(d(a, b) + d(b, c))$, for some constant γ. The theorems and lemmas can be adjusted to account for this approximation, e.g., the bound in theorem 1 will become $2\gamma R$. A commonly used distance measure for comparing images and documents is the cosine distance defined as the inner-product of two vectors after normalization. This distance can be show to be proportional to the squared Euclidean distance between the normalized vectors which satisfies the relaxed version of triangle inequality for $\gamma = 2$.

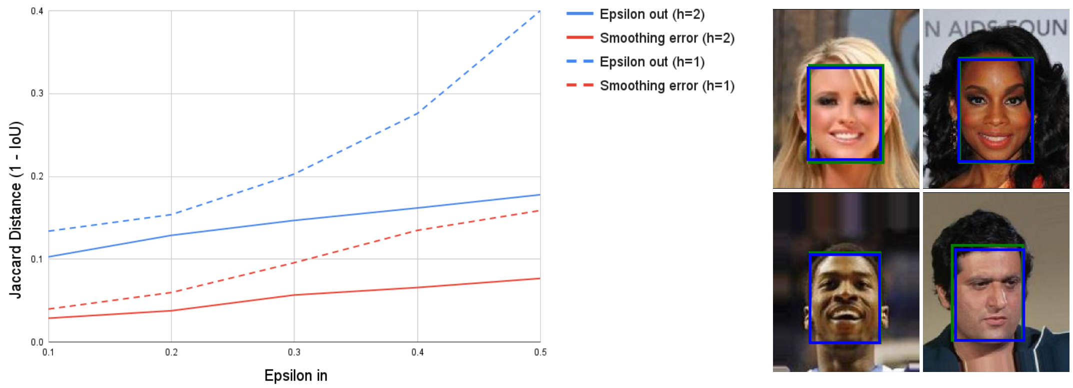
These relaxations extend the scope of center smoothing to many commonly used distance measures that need not necessarily satisfy all the metric properties. For instance, perceptual distances measure the distance between two images in some feature space rather than image space. Such distances align well with human judgements when the features are extracted from a deep neural network [54] and are considered more natural measures for image similarity. For two images I_1 and I_2, let $\phi(I_1)$ and $\phi(I_2)$ be their feature representations. Then, for a distance function d in the feature space that satisfies the relaxed triangle inequality, we can define a distance function $d_\phi(I_1, I_2) = d(\phi(I_1), \phi(I_2))$ in the image space, which also satisfies the relaxed triangle inequality. For any image I_3,

$$\begin{aligned} d_\phi(I_1, I_2) &= d(\phi(I_1), \phi(I_2)) \\ &\leq \gamma\left(d(\phi(I_1), \phi(I_3)) + d(\phi(I_3), \phi(I_2))\right) \\ &= \gamma\left(d_\phi(I_1, I_3) + d_\phi(I_3, I_2)\right). \end{aligned}$$

5 Experiments

We apply center smoothing to certify a wide range of output metrics: Jaccard distance based on intersection over union (IoU) of sets, total variation distances for images, and angular distance. We certify the bounding box generated by a face detector – a key component of most facial recognition systems – by guaranteeing the minimum overlap (measured using IoU) it must have with the output under an adversarial perturbation of the input. For instance, if $\epsilon_1 = 0.2$, the Jaccard distance (1-IoU) is guaranteed to be bounded by 0.2, which implies that the bounding box of a perturbed image must have at least 80% overlap with that of the clean image. We use a pre-trained face detection model for this experiment. For total variation and angular distance, we use simple, easy-to-train convolutional neural network based dimensionality reduction (autoencoder) and image reconstruction models. Our goal is to demonstrate the effectiveness of our method for a wide range of applications and so, we place less emphasis on the performance of the underlying models being smoothed. In each case, we show that our method is capable of generating certified guarantees without significantly degrading the performance of the underlying model. We provide additional experiments for other metrics and parameter settings in the appendix.

As is common in the randomized smoothing literature, we train our base models (except for the pre-trained ones) on noisy data with different noise levels $\sigma = 0.1, 0.2, \ldots, 0.5$ to make them more robust to input perturbations. We use $n = 10^4$ samples to estimate the smoothed function and $m = 10^6$ samples to generate certificates, unless stated otherwise. We set $\Delta = 0.05, \alpha_1 = 0.005$ and $\alpha_2 = 0.005$ as discussed in previous sections. We grow the smoothing noise σ linearly with the input perturbation ϵ_1. Specifically, we maintain $\epsilon_1 = h\sigma$ for different values of $h = 2, 1, 1.5$ in our experiments. We plot the median certified output radius ϵ_2 and the median smoothing loss, defined as the distance between the outputs of the base model and the smoothed model $d(f(x), \hat{f}(x))$, of fifty random test examples for different values of ϵ_1. In all our experiments, we observe that both these quantities increase as the input radius ϵ_1 increases, but the smoothing error remains significantly below the certified output radius. Also, increasing the value of h improves the quality of the certificates (lower ϵ_2). This could be due to the fact that for a higher h, the smoothing noise



(a) Certifying Jaccard Distance (1 - IoU). (b) Smoothed Output.

Figure 3: Face Detection on CelebA using MTCNN detector: Part (a) plots the certified output radius ϵ_2 and the smoothing error for $h = 1$ and 2. Part (b) compares the smoothed output (blue box) to the output of the base classifier (green box, mostly hidden behind the blue box) showing a significant overlap.

σ is lower (keeping ϵ_1 constant), which means that the radius of the minimum enclosing ball in the output space is smaller leading to a tighter certificate. We ran all our experiments on a single NVIDIA GeForce RTX 2080 Ti in an internal cluster. Each of the fifty examples we certify took somewhere between 1-3 minutes depending on the underlying model.

5.1 Jaccard distance

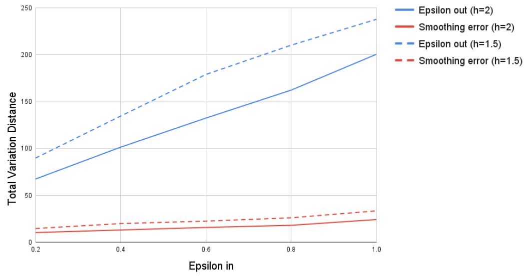
It is known that facial recognition systems can be deceived to evade detection, impersonate authorized individuals and even render completely ineffective [48, 45, 13]. Most facial recognition systems first detect a region that contains a persons face, e.g. a bounding box, and then uses facial features to identify the individual in the image. To evade detection, an attacker may seek to degrade the quality of the bounding boxes produced by the detector and can even cause it to detect no box at all. Bounding boxes are often interpreted as sets and the their quality is measured as the amount of overlap with the desired output. When no box is output, we say the overlap is zero. The overlap between two sets is defined as the ratio of the size of the intersection between them to the size of their union (IoU). Thus, to certify the robustness of the output of a face detector, it makes sense to bound the worst-case IoU of the output of an adversarial input to that of a clean input. The corresponding distance function, known as Jaccard distance, is defined as $1 - IoU$ which defines a metric over the universe of sets.

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad d_J(A, B) = 1 - IoU(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}.$$

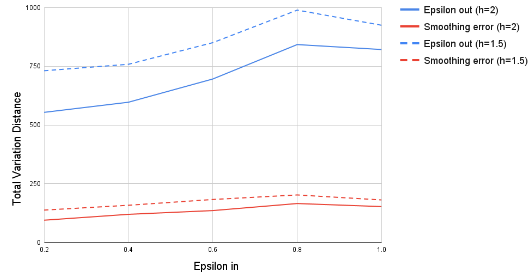
In this experiment, we certify the output of a pre-trained face detection model MTCNN [53] on the CelebA face dataset [35]. We set $n = 5000$ and $m = 10000$, and use default values for other parameters discussed above. Figure 3a plots the certified output radius ϵ_2 and the smoothing error for $h = \epsilon_1/\sigma = 1$ and 2 for $\epsilon_1 = 0.1, 0.2, \ldots, 0.5$. Certifying the Jaccard distance allows us to certify IoU as well, e.g., for $h = 2$, ϵ_2 is consistently below 0.2 which means that even the worst bounding box under adversarial perturbation of the input has an overlap of at least 80% with the box for the clean input. The low smoothing error shows that the performance of the base model does not drop significantly as the actual output of the smoothed model has a large overlap with that of the base model. Figure 3b compares the outputs of the smoothed model (blue box) and the base model (green box). For most of the images, the blue box overlaps with the green one almost perfectly.

5.2 Total Variation Distance

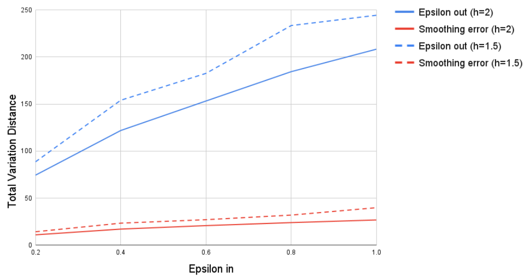
The total variation norm of a vector x is defined as the sum of the magnitude of the difference between pairs of coordinates defined by a *neighborhood* set N. For a 1-dimensional array x with k elements,



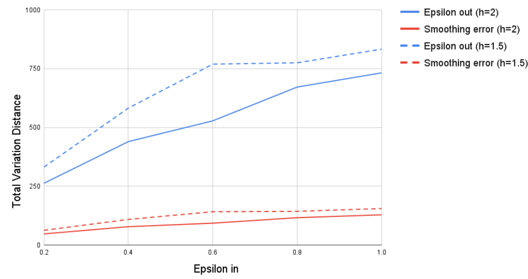
(a) Dimensionality Reduction on MNIST



(b) Dimensionality Reduction on CIFAR-10



(c) Image Reconstruction on MNIST



(d) Image Reconstruction on CIFAR-10

Figure 4: Certifying Total Variation Distance

one can define the neighborhood as the set of consecutive elements.

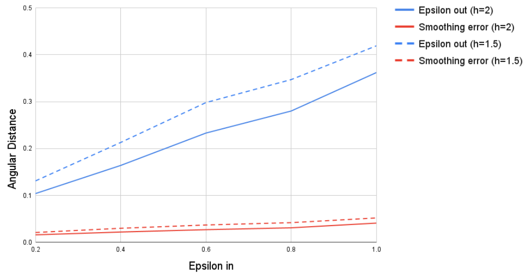
$$TV(x) = \sum_{(i,j)\in N} |x_i - x_j|, \quad TV_{1D}(x) = \sum_{i=1}^{k-1} |x_i - x_{i+1}|.$$

Similarly, for a grayscale image represented by a $h \times w$ 2-dimensional array x, the neighborhood can be defined as the next element (pixel) in the row/column. In case of an RGB image, the difference between the neighboring pixels is a vector, whose magnitude can be computed using an ℓ_p-norm. For, our experiments we use the ℓ_1-norm.

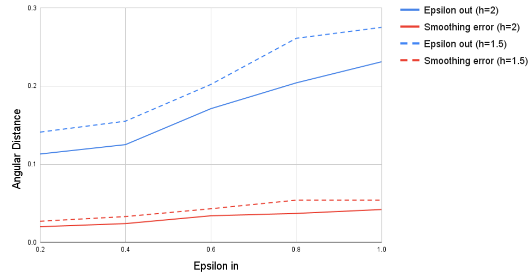
$$TV_{RGB}(x) = \sum_{i=1}^{h-1} \sum_{j=1}^{w-1} \|x_{i,j} - x_{i+1,j}\|_1 + \|x_{i,j} - x_{i,j+1}\|_1$$

The total variation distance between two images I_1 and I_2 can be defined as the total variation norm of the difference $I_1 - I_2$, i.e., $TVD(I_1, I_2) = TV(I_1 - I_2)$. The above distance defines a pseudometric over the space of images as it satisfies the symmetry property and the triangle inequality, but may violate the identity of indiscernibles as an image obtained by adding the same value to all the pixel intensities has a distance of zero from the original image. However, as noted in section 4, our certificates hold even for this setting.

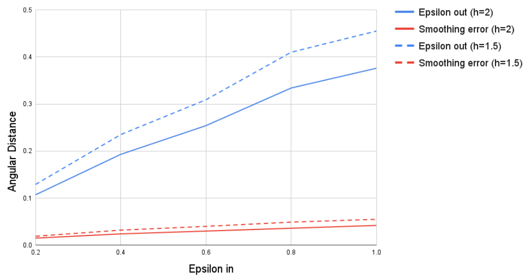
We certify total variation distance for the problems of dimensionality reduction and image reconstruction on MNIST [10] and CIFAR-10 [22]. The base-model for dimensionality reduction is an autoencoder that uses convolutional layers in its encoder module to map an image down to a small number of latent variables. The decoder applies a set of de-convolutional operations to reconstruct the same image. We insert batch-norm layers in between these operations to improve performance. For image reconstruction, the goal is to recover an image from small number of measurements of the original image. We apply a transformation defined by Gaussian matrix A on each image to obtain the measurements. The base model tries to reconstruct the original image from the measurements. The attacker, in this case, is assumed to add a perturbation in the measurement space instead of the image space (as in dimensionality reduction). The model first reverts the measurement vector to a vector in the image space by simply applying the pseudo-inverse of A and then passes it through a similar autoencoder model as for dimensionality reduction. We present results for $\epsilon_1 = 0.2, 0.4, \ldots, 1.0$ and $h = 2, 1.5$ and use 256 latent dimensions and measurements for these experiments in figure 4. To put these plots in perspective, the maximum TVD between two CIFAR-10 images could be $6 \times 31 \times 31 = 5766$ and between MNIST images could be $2 \times 27 \times 27 = 1458$ (pixel values between 0 and 1).



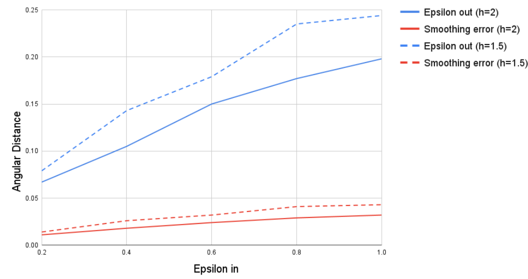
(a) Dimensionality Reduction on MNIST



(b) Dimensionality Reduction on CIFAR-10



(c) Image Reconstruction on MNIST



(d) Image Reconstruction on CIFAR-10

Figure 5: Certifying Angular Distance

5.3 Angular Distance

A common measure for similarity of two vectors A and B is the cosine similarity between them, defined as below:

$$\cos(A, B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2} = \frac{\sum_i A_i B_i}{\sqrt{\sum_j A_j^2}\sqrt{\sum_k B_k^2}}.$$

In order to convert it into a distance, we can compute the angle between the two vectors by taking the cosine inverse of the above similarity measure, which is known as angular distance:

$$AD(A, B) = \cos^{-1}(\cos(A, B))/\pi.$$

Angular distance always remains between 0 and 1, and similar to the total variation distance, angular distance also defines a pseudometric on the output space. We repeat the same experiments with the same models and hyper-parameter settings as in the previous subsection for total variation distance (figure 5). The results are similar in trend in all the experiments conducted, showing that center smoothing can be reliably applied to a vast range of output metrics to obtain similar robustness guarantees.

6 Conclusion

Randomized smoothing can be extended beyond classification tasks to obtain provably robust models for problems where the quality of the output is measured using a distance metric. We design a procedure that can make any model of this kind provably robust against norm bounded adversarial perturbations of the input. In our experiments, we demonstrate that it can generate meaningful certificates under a wide variety of distance metrics without significantly compromising the quality of the base model. We also note that the metric requirements on the distance measure can be partially relaxed in exchange for weaker certificates.

In this work, we focus on ℓ_2-norm bounded adversaries and the Gaussian smoothing distribution. An important direction for future investigation could be whether this method can be generalised beyond ℓ_p-adversaries to more natural threat models, e.g., adversaries bounded by total variation distance, perceptual distance, cosine distance, etc. Center smoothing does not critically rely on the shape of the smoothing distribution or the threat model. Thus, improvements in these directions could potentially be coupled with our method to broaden the scope of provable robustness in machine learning.

References

[1] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. On instabilities of deep learning in image reconstruction - does AI come at a cost? *CoRR*, abs/1902.05300, 2019. URL `http://arxiv.org/abs/1902.05300`.

[2] Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition - 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings*, volume 10358 of *Lecture Notes in Computer Science*, pages 262–275. Springer, 2017. doi: 10.1007/978-3-319-62416-7_19. URL `https://doi.org/10.1007/978-3-319-62416-7_19`.

[3] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian J. Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[4] Mihai Bundefineddoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC '02, page 250–257, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134959. doi: 10.1145/509907.509947. URL `https://doi.org/10.1145/509907.509947`.

[5] Francesco Calivá, Kaiyang Cheng, Rutwik Shah, and Valentina Pedoia. Adversarial robust training in mri reconstruction. *arXiv preprint arXiv:2011.00070*, 2020.

[6] Kaiyang Cheng, Francesco Calivá, Rutwik Shah, Misung Han, Sharmila Majumdar, and Valentina Pedoia. Addressing the false negative problem of deep learning mri reconstruction models by adversarial attacks and robust training. In Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Herve Lombaert, and Christopher Pal, editors, *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 121–135, Montreal, QC, Canada, 06–08 Jul 2020. PMLR. URL `http://proceedings.mlr.press/v121/cheng20a.html`.

[7] Ping-yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studer, and Tom Goldstein. Certified defenses for adversarial patches. In *8th International Conference on Learning Representations*, 2020.

[8] Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, and Jong-Seok Lee. Evaluating robustness of deep image super-resolution against adversarial attacks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 303–311. IEEE, 2019. doi: 10.1109/ICCV.2019.00039. URL `https://doi.org/10.1109/ICCV.2019.00039`.

[9] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[10] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

[11] Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[12] Krishnamurthy Dvijotham, Sven Gowal, Robert Stanforth, Relja Arandjelovic, Brendan O'Donoghue, Jonathan Uesato, and Pushmeet Kohli. Training verified learners with learned verifiers, 2018.

[13] Morgan Frearson and Kien Nguyen. Adversarial attack on facial recognition using visible light. *arXiv preprint arXiv:2011.12680*, 2020.

[14] Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=HJgEMpVFwB`.

[15] Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and clean data are not twins. *CoRR*, abs/1704.04960, 2017.

[16] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[17] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models, 2018.

[18] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick D. McDaniel. On the (statistical) detection of adversarial examples. *CoRR*, abs/1702.06280, 2017.

[19] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[20] Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4081–4091, 2019. doi: 10.18653/v1/D19-1419. URL `https://doi.org/10.18653/v1/D19-1419`.

[21] Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=ryvlRyBKl`.

[22] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

[23] Aounon Kumar, Alexander Levine, Soheil Feizi, and Tom Goldstein. Certifying confidence via randomized smoothing. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/37aa5dfc44dddd0d19d4311e2c7a0240-Abstract.html`.

[24] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL `https://openreview.net/forum?id=BJm4T4Kgx`.

[25] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 10408–10418, 2019. URL `http://papers.nips.cc/paper/9228-functional-adversarial-attacks`.

[26] Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 656–672, 2019.

[27] Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi S. Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 4911–4922, 2019.

[28] Alexander Levine and Soheil Feizi. Wasserstein smoothing: Certified robustness against wasserstein adversarial attacks, 2019.

[29] Alexander Levine and Soheil Feizi. (de)randomized smoothing for certifiable defense against patch attacks. *CoRR*, abs/2002.10733, 2020. URL https://arxiv.org/abs/2002.10733.

[30] Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 4585–4593. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/5888.

[31] Alexander Levine, Aounon Kumar, Thomas Goldstein, and Soheil Feizi. Tight second-order certificates for randomized smoothing, 2020.

[32] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 9459–9469, 2019.

[33] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Second-order adversarial attack and certifiable robustness, 2019. URL https://openreview.net/forum?id=SyxaYsAqY7.

[34] Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5775–5783, 2017.

[35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[37] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3578–3586. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/mirman18b.html.

[38] Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. In Elisabeth André, Sven Koenig, Mehdi Dastani, and Gita Sukthankar, editors, *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, pages 2040–2042. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018. URL http://dl.acm.org/citation.cfm?id=3238064.

[39] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 10900–10910, Red Hook, NY, USA, 2018. Curran Associates Inc.

[40] Ankit Raj, Yoram Bresler, and Bo Li. Improving robustness of deep-learning-based image reconstruction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7932–7942. PMLR, 2020. URL http://proceedings.mlr.press/v119/raj20a.html.

[41] Hadi Salman, Jerry Li, Ilya P. Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 11289–11300, 2019.

[42] Vladimir Shenmaier. Complexity and approximation of the smallest k-enclosing ball problem. *European Journal of Combinatorics*, 48:81 – 87, 2015. ISSN 0195-6698. doi: https://doi.org/10.1016/j.ejc.2015.02.011. URL http://www.sciencedirect.com/science/article/pii/S0195669815000335.

[43] Sahil Singla and Soheil Feizi. Robustness certificates against adversarial examples for relu networks. *CoRR*, abs/1902.01235, 2019.

[44] Sahil Singla and Soheil Feizi. Second-order provable defenses against adversarial attacks, 2020.

[45] Qing Song, Yingqi Wu, and Lu Yang. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. *CoRR*, abs/1811.12026, 2018. URL http://arxiv.org/abs/1811.12026.

[46] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

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

[48] Fatemeh Vakhshiteh, Raghavendra Ramachandra, and Ahmad Nickabadi. Threat of adversarial attacks on face recognition: A comprehensive survey. *arXiv preprint arXiv:2007.11709*, 2020.

[49] Adva Wolf. Making medical image reconstruction adversarially robust. 2019. URL http://cs229.stanford.edu/proj2019spr/report/97.pdf.

[50] Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5283–5292, 2018.

[51] Ping yeh Chiang, Michael J. Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection by median smoothing, 2020.

[52] Minghao Yin, Yongbing Zhang, Xiu Li, and Shiqi Wang. When deep fool meets deep prior: Adversarial attack on super-resolution network. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 1930–1938, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356657. doi: 10.1145/3240508.3240603. URL https://doi.org/10.1145/3240508.3240603.

[53] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878, 2016. URL http://arxiv.org/abs/1604.02878.

[54] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In the conclusion (section 6), we discuss that more work is needed to generalize our method to threat models other than ℓ_2.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work aims to help make machine learning models more robust and reliable for real-world applications.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Complete proofs of all theorems and lemmas can be found in the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code is included in the supplemental.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In the experiments section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In the experiments section.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A] We are using datasets that are available in the public domain with custom license terms that allow non-commercial use, like MNIST, CIFAR-10 and CelebA.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Our code is included in the supplemental.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We are using datasets that are available in the public domain with custom license terms that allow non-commercial use, like MNIST, CIFAR-10 and CelebA.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]