Extensions and limitations of randomized smoothing for robustness guarantees

Jamie Hayes
University College London
j.hayes@cs.ucl.ac.uk

Abstract

Randomized smoothing, a method to certify a classifier’s decision on an input is invariant under adversarial noise, offers attractive advantages over other certification methods. It operates in a black-box and so certification is not constrained by the size of the classifier’s architecture. Here, we extend the work of Li et al. [26], studying how the choice of divergence between smoothing measures affects the final robustness guarantee, and how the choice of smoothing measure itself can lead to guarantees in differing threat models. To this end, we develop a method to certify robustness against any $\ell_p$ ($p \in \mathbb{N}_{>0}$) minimised adversarial perturbation. We then demonstrate a negative result, that randomized smoothing suffers from the curse of dimensionality; as $p$ increases, the effective radius around an input one can certify vanishes.

1. Introduction

Image classification is vulnerable to adversarial examples. Given an image classifier $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that the decision function $F = \arg \max_i f_i$, classifies an input, $x$, correctly as $F(x) = y$, an adversarial example is an input, $x + \delta$, such that $F(x + \delta) \neq y$ where $x$ and $x + \delta$ are assigned the same label by an oracle classifier, $\mathcal{O}$, which is usually taken to be the human vision system. To preserve oracle classification, it is common to minimize the perturbation, $\delta$, with respect to an $\ell_p$ norm. Constructing a perturbation such that $\|\delta\|_p \ll \|x\|_p$, will result in an input such that $\|x + \delta\|_p \approx \|x\|_p$. With high likelihood $x$ and $x + \delta$ will be visually similar and $\mathcal{O}$ will classify both correctly.

The vulnerability to adversarial examples requires a suitable defense. Many empirical defenses have been proposed and subsequently shown to be broken, implying more theoretically grounded techniques to measure robustness are required [1, 6, 7, 16, 34]. Recently, methods from verification literature have been used to provide guarantees of an inputs robustness to adversarial perturbations. These methods seek the minimum or a lower bound on the amount of noise required to cause a misclassification. These verification methods are most often tailored to a single $\ell_p$ norm for which the defense guarantees robustness. A number of defenses certify a neural network is robust to adversarial examples by propagating upper and lower input bounds throughout the network or by bounding the Lipschitz value of the network [4, 12, 17, 18, 27, 29, 33, 37].

Recently, randomized smoothing has been proposed to certify image classifiers to $\ell_0$, $\ell_1$, and $\ell_2$ perturbations [10, 24, 25, 26]. By constructing a classifier that outputs a label based on a majority vote under repeated addition of Laplacian or Gaussian noise, Lecuyer et al. [24] found lower bounds to the amount of noise required for misclassification of an input in the $\ell_1$ or $\ell_2$ norm, respectively. Following this, Li et al. [26] and Cohen et al. [10] provided improved bounds in the $\ell_2$ norm. As explained by Cohen et al. [10], randomized smoothing has attractive advantages over other certification methods: it is scalable to large classifiers and makes no assumption about the architecture. In this work, we extend the general framework for randomized smoothing as proposed by Li et al. [26]. Firstly, we study how the choice of divergence between inputs smoothed with noise affects the final certificate, and secondly, we study how the choice of smoothing measure itself can lead to guarantees for differing threat models. Concretely, we show how the choice of smoothing measure allows us to extend randomized smoothing to any $\ell_p$ norm ($p \in \mathbb{N}_{>0}$), showing we can certify inputs with non-vacuous bounds over a range of $\ell_p$ norms with small $p$ values. We then show that randomized smoothing fails to certify meaningfully large radii around inputs as $p$ increases.

2. Certified defenses

In this section, we discuss related work on certified defenses to adversarial examples, introduce extensions to randomized smoothing approaches to certified defenses, and provide a method to compute a certified robust area around an input under any $\ell_p$ norm attack, where $p \in \mathbb{N}_{>0}$.

2.1. Background on certified defenses

The vulnerability of empirical defenses to adversarial examples has driven the need for formal guarantees of robustness. We define certified robustness as a guarantee that the
decision of a classifier is preserved within an $\epsilon$-ball around an input, and we refer to size of this $\epsilon$-ball as the certified radius. Formal methods can be separated into complete and incomplete methods. Complete methods such as Satisfiability Modulo Theory (SMT) [8, 15, 20] or Mixed-Integer Programming (MIP) [5, 9, 35] provide exact robustness bounds but are expensive to implement. Incomplete methods solve a convex relaxation of the verification problem. The bounds given by incomplete methods can be loose but are quicker to find than exact bounds [4, 12, 17, 18, 27, 29, 37].

Lecuyer et al. [24] developed the certification technique, referred to as randomized smoothing, by noticing a connection between differential privacy [14] and robustness, and show that robustness can be proven under concentration measures of classification under noise. This work was expanded upon by Lee et al. [25], Li et al. [26], and Cohen et al. [10], who found improved robustness guarantees in the $\ell_0$, $\ell_1$, and $\ell_2$ norms, respectively. Similarly to this work, Dvijotham et al. [13] developed a general framework for randomized smoothing that can handle arbitrary smoothing measures and so find robustness guarantees in any $\ell_p$ norm. In concurrent work, Blum et al. [3], Kumar et al. [23], and Yang et al. [36] also show that randomized smoothing may be unable to find robustness guarantees in the $\ell_{\infty}$ norm. Most related to this work are the findings of Kumar et al. [23], who also use a generalized Gaussian distribution for smoothing and show that the certified radius in an $\ell_p$ norm decreases as $O(1/\epsilon^2)$, where $\epsilon$ is the dimensionality of the data.

### 2.2. Certification via randomized smoothing

Here, we expand on how robustness guarantees can be found through randomized smoothing.

#### Problem statement

Given an input $x \in \mathcal{X}$ such that $\arg \max_i f_i(x) = y$, find the maximum $\epsilon$ such that $\forall x' \in \mathcal{X}$, $d(x, x') < \epsilon \implies \arg \max_i f_i(x') = y$, given a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. This can be cast as an optimization problem, given by

$$
\max_{x' \in \mathcal{X}} d(x, x')
\text{subject to } \arg \max_i f_i(x') = y \tag{1}
$$

In general, solving the above formulation is difficult, however randomized smoothing, introduced by Lecuyer et al. [24], can be used to solve a relaxed version of this problem. Namely, the aim is to solve

$$
\max_{x' \in \mathcal{X}} d(x + \theta, x' + \theta)
\text{subject to } E[\arg \max_i f_i(x' + \theta)] = y, \tag{2}
$$

where $\theta$ is a sample from a smoothing measure, $\mu$, and $d$ is now taken to be a suitable divergence or distance measure between random variables. For example, Li et al. [26] take $\mu$ to be the centered Gaussian, $\mathcal{N}(0, \sigma^2)$. Since Gaussians belong to the location-scale family of distributions, we can treat $x$ and $x'$ as constants and so, $x + \theta$ and $x' + \theta$ can be treated as random variables from distributions $\mathcal{N}(x, \sigma^2)$ and $\mathcal{N}(x', \sigma^2)$, respectively. We can use well known properties of divergences of Gaussians to represent $d(x + \theta, x' + \theta)$ in terms of the $\ell_2$ norm difference of their means. Specifically, $d(x + \theta, x' + \theta)$ can be represented as a function of $\|x - x'\|_2$ and $\sigma$, for common divergences such as the Rényi and KL divergences. However, we must still solve the problem of ensuring $E[\arg \max_i f_i(x' + \theta)] = y$. Given a chosen divergence, Li et al. [26] approach this problem by finding a lower bound between two multinomial distributions, $P$ and $Q$, in terms of the two largest probabilities of $P$, when $\arg \max_i P_i \neq \arg \max_i Q_i$. This shows that any distribution, $Q$, for which $P$ and $Q$ agree on the index of the top probability, the divergence between $P$ and $Q$ must be smaller than this lower bound. We denote this lower bound by $h(p_1, p_2)$, where $p_1, p_2$ represent the top two probabilities from $f$. Given this lower bound Li et al. [26], solve the following problem

$$
\max_{x' \in \mathcal{X}} d(f(x + \theta), f(x' + \theta))
\text{subject to } d(f(x + \theta), f(x' + \theta)) \leq h(p_1, p_2) \tag{3}
$$

This can be efficiently solved by finding an upper bound to the Lagrangian relaxed problem

$$
\max_{\lambda \leq 0, x' \in \mathcal{X}} d(f(x + \theta), f(x' + \theta))
+ \lambda(h(p_1, p_2) - d(f(x + \theta), f(x' + \theta))) \tag{4}
$$

$$
= \max_{\lambda \leq 0, x' \in \mathcal{X}} (1 - \lambda) d(f(x + \theta), f(x' + \theta)) + \lambda h(p_1, p_2) \tag{5}
$$

$$
= \max_{\lambda \geq 0, x' \in \mathcal{X}} (1 + \lambda) d(f(x + \theta), f(x' + \theta)) - \lambda h(p_1, p_2) \tag{6}
$$

$$
\leq \max_{\lambda \geq 0, x' \in \mathcal{X}} (1 + \lambda) g(\|x - x'\|_2, \sigma) - \lambda h(p_1, p_2), \tag{7}
$$

where in eq. (7), we use the data processing inequality property of divergences, and in eq. (8), we use the fact that for many common divergences, we can represent the divergence between two Gaussians as a function of the $\ell_2$ norm of their means and their standard deviation, which we denote by $g(\|x - x'\|_2, \sigma)$.

By choosing $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ to be the Rényi divergence,
Table 1: $\ell_2$ certified radius when using different divergences.

<table>
<thead>
<tr>
<th>Distance</th>
<th>$d(Q, P)$ (when $\arg \max q_i \neq \arg \max p_i$)</th>
<th>$d(\mathcal{N}(x, \sigma^2), \mathcal{N}(x', \sigma^2))$</th>
<th>Certified radius (for $|x - x'|_2 &lt; \epsilon$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{KL}(Q, P) = \sum_{i=1}^{k} q_i \log \frac{q_i}{p_i}$</td>
<td>$- \log(2\sqrt{p_1p_2} + 1 - p_1 - p_2)$</td>
<td>$\frac{1}{\sigma} |x - x'|_2^2$</td>
<td>$\sqrt{-\sigma^2 \log(2\sqrt{p_1p_2} + 1 - p_1 - p_2)}$</td>
</tr>
<tr>
<td>$d_{H2}(Q, P) = \frac{1}{2} \sum_{i=1}^{k} (\sqrt{q_i} - \sqrt{p_i})^2$</td>
<td>$1 - \sqrt{1 - \frac{(\sqrt{p_1} - \sqrt{p_2})^2}{2}}$</td>
<td>$1 - e^{-\frac{1}{\sqrt{2}}</td>
<td>x - x'</td>
</tr>
<tr>
<td>$d_{\chi^2}(Q, P) = \sum_{i=1}^{k} \left( \frac{p_i - q_i}{p_i + q_i} \right)^2$</td>
<td>$e^{-\frac{1}{\sigma^2}</td>
<td>x - x'</td>
<td>^2} - 1$</td>
</tr>
<tr>
<td>$d_B(Q, P) = - \log(\sum_{i=1}^{k} \sqrt{q_i}p_i)$</td>
<td>$- \log(\sqrt{p_1} + \sqrt{p_2} + 2(1 - p_1 - p_2))$</td>
<td>$\frac{1}{\sigma} |x - x'|_2^2$</td>
<td>$\sqrt{-8\sigma^2 \log(1 - \frac{1}{\sqrt{2}}</td>
</tr>
<tr>
<td>$d_{TV}(Q, P) = \frac{1}{2} \sum_{i=1}^{k}</td>
<td>q_i - p_i</td>
<td>$</td>
<td>$\frac{</td>
</tr>
</tbody>
</table>

we recover the results of Li et al. [26] with

\[ g(\|x - x'\|_2, \sigma) = \frac{\alpha \|x - x'\|_2^2}{2\sigma^2} \] (9)

\[ h(p_1, p_2) = - \log \left( 1 - p_1 - p_2 + 2\left(\frac{1}{2}(p_1^{1-\alpha} + p_2^{1-\alpha})\right) \right)^{\frac{1}{\alpha}} \] (10)

Thus, for any $x' \in \mathcal{X}$ with $\|x - x'\|_2 < \epsilon$ we can guarantee the classifier, $f$, will not change its decision for any $\epsilon$ smaller than

\[ \max_{\lambda > 0} \left( \sup_{\alpha > 1} \left( - \frac{\lambda \sigma^2}{(1 + \lambda)\alpha} \log \left( 1 - p_1 - p_2 + \left(\frac{2}{\lambda} \left(\frac{1}{2}(p_1^{1-\alpha} + p_2^{1-\alpha})\right)\right) \right) \right)^{\frac{1}{2}} \] (11)

\[ \sup_{\alpha > 1} \left( - \frac{2\sigma^2}{\alpha} \log \left( 1 - p_1 - p_2 + \left(\frac{2}{\alpha} \left(\frac{1}{2}(p_1^{1-\alpha} + p_2^{1-\alpha})\right)\right) \right) \right)^{\frac{1}{2}} \] (12)

Clearly, this framework for certifying inputs is general and extends to different choices of divergence. In the next section, we explore divergences beyond Rényi divergence and show this choice affects the certified radius, given a Gaussian smoothing measure.

2.3. Certification guarantees against $\ell_2$ perturbations for common divergences

Li et al. [26] show that, given two distributions, $P$ and $Q$, with different indexes for the top probability, a lower bound of the Rényi divergence (denoted by $d_\alpha$) is given by eq. (10). We extend this line of reasoning to find lower bounds for the KL divergence ($d_{KL}$), Hellinger distance ($d_{H2}$), Neyman chi-squared distance ($d_{\chi^2}$), Bhattacharyya distance ($d_B$), and total variation distance ($d_{TV}$). Proofs of these lower bounds are given in appendix A. To find a certified radius of a classifier’s decision around an input, we find the distances between Gaussian measures with respect to each of these divergences. These are both represented in table 1 along with the certification guarantee in the $\ell_2$ norm. We visualize the trade-off in certified radius around an input in fig. 1 for a hypothetical binary classification task as a function of the classifier’s top output probability, $p_1$. As well as including the certified radii derived from the aforementioned divergences, we include the certified radii for the $\ell_2$ norm found by Lecuyer et al. [24] and Cohen et al. [10] approaches. Lecuyer et al. [24] find a certified radius against $\ell_2$ perturbations given by $\sup_{\beta < \epsilon} \log(\frac{\beta}{1 - \exp(-\beta)}) \geq 2\log(\frac{1 + \exp(\beta)}{p_1 + \exp(\beta)})$.

Clearly, all choices of distance metrics dominate the certificates found using the Lecuyer et al. [24] method, and for values of $p_1$ close to $1/2$, $d_{TV}$ is approximately equal to the tight Cohen et al. [10] guarantee. However, the certified radius found using $d_{TV}$ is linear with respect to the top predicted probability, and so becomes a weaker guarantee for larger probabilities. Robustness guarantees provided by Rényi and chi-squared divergences are approximately equal; a finer-grained visualization of the difference between these two divergences is given in appendix B.

We formalize the trade-offs between different choices of divergences with the following proposition.

**Proposition 1.** Let $c_{d_{KL}}, c_{d_{\chi^2}}, c_{d_{H2}}, c_{d_B}, c_{d_\alpha},$ and $c_{[24]}$, denote the certificates found using $d_{KL}, d_{\chi^2}, d_{H2}, d_B, d_\alpha,$ and
the Lecuyer et al. [24] approach, respectively. Then, the following holds

1. \( \forall p_1 \in (\frac{1}{2}, 1), \epsilon_{d_a} > \epsilon_{d_{x_2}}. \)
2. \( \forall p_1 \in (\frac{1}{2}, 1), \epsilon_{d_{x_2}} > \epsilon_{d_{KL}}. \)
3. \( \forall p_1 \in (\frac{1}{2}, 1), \epsilon_{d_{x_2}} > \epsilon_{d_{H^2}}. \)
4. \( \forall p_1 \in (\frac{1}{2}, 1], \epsilon_{d_B} = \epsilon_{d_{H^2}}. \)
5. \( \forall p_1 \in (\frac{1}{2}, 1), \epsilon_{d_{KL}} > \epsilon_{d_{KL}}. \)
6. \( \forall p_1 \in (\frac{1}{2}, 1), \epsilon_{d_{KL}} > \epsilon_{d_{S_2}}. \)

Proof. See appendix C.

Proposition 1 defines a strict hierarchy, and so informs us of the best divergence one can use to certify an input against \( \ell_2 \) perturbations using the Li et al. [26] approach.

### 2.4. Certification guarantees beyond the \( \ell_2 \) based perturbations via different smoothing measures

The Gaussian distribution is a natural choice for the smoothing measure because it naturally leads to robustness guarantees in the \( \ell_2 \) norm. However, it is also a convenient choice of smoothing measure because it is a member of the location-scale family of distributions. This means that, fixing \( x \in \mathcal{X} \), sampling from \( x + \mathcal{N}(0, \sigma^2) \) is equivalent to sampling from \( \mathcal{N}(x, \sigma^2) \). Importantly, addition of a constant, \( x \), does not change the family of the smoothing measure, and so we can use well known formula for the distances between two Gaussian distributions to derive robustness guarantees.

Unfortunately, not all distributions belong to the location-scale family, and so, in our formulation, we are not free to choose any distribution for smoothing. Another convenient choice of a location-scale distribution is the generalized Gaussian distribution [30], denoted \( \mathcal{G}N(\mu, \sigma, s) \), whose density function is given by

\[
p(x) = \frac{s}{2\sigma\Gamma(\frac{1}{s})} e^{-\frac{|x - \mu|^s}{2\sigma}}
\]

where \( \mu \) is the mean, \( \sigma \) denotes a scaling factor and \( s \) denotes a shaping factor. The Laplacian distribution is recovered when \( s = 1 \), the Gaussian \( \mathcal{N}(\mu, \sigma^2) \) when \( s = 2 \), and the uniform distribution on \( (\mu - \sigma, \mu + \sigma) \) as \( s \to \infty \). We will show that by using this smoothing measure we can find robustness guarantees to \( \ell_p \) perturbations, where \( p \in \mathbb{N}_{>0} \).

We show in appendix D that given inputs \( x \) and \( x' \) the Kullback–Leibler (KL) divergence of \( \mathcal{G}N(x, \sigma, s) \) and \( \mathcal{G}N(x', \sigma, s) \) is given by

\[
\sum_{k=1}^{s} \binom{s}{k} \frac{(s)(1 + (-1)^{s-k})\Gamma\left(\frac{s-k+1}{s}\right)}{(2\sigma^k\Gamma(\frac{1}{s}))} \|x - x'\|^k_k
\]

We also show in appendix A that the KL divergence of two multinomial distributions \( P \) and \( Q \) (that disagree on the index of the top probability) is lower bounded by

\[
d_{KL}(Q, P) \geq -\log(2\sqrt{p_1p_2} + 1 - p_1 - p_2)
\]

Then we use the data processing inequality to prove robustness up to \( \|x - x'\|_p < \epsilon \) if the following holds

\[
d_{KL}(f(x + \mathcal{G}N(0, \sigma, p)), f(x' + \mathcal{G}N(0, \sigma, p))) \leq d_{KL}(x + \mathcal{G}N(0, \sigma, p), x' + \mathcal{G}N(0, \sigma, p))
\]

\[
\leq \epsilon^p + \sum_{k=1}^{p-1} \binom{p}{k} \frac{(1 + (-1)^{p-k})\Gamma\left(\frac{p-k+1}{p}\right)}{(2\sigma^k\Gamma(\frac{1}{p}))} \|x - x'\|^k_k
\]

\[
\leq -\log(2\sqrt{p_1p_2} + 1 - p_1 - p_2)
\]

Table 2 gives examples of the KL-divergence of the generalized Gaussian distribution for small \( \ell_p \) norms. For \( \ell_p \) norms with \( p = 1 \) or \( p = 2 \), the upper bound to which an input is certifiably robust is given by

\[
(-\sigma^p \log(2\sqrt{p_1p_2} + 1 - p_1 - p_2))^{\frac{1}{p}}
\]

For \( \ell_p \) norms with \( p > 2 \), \( p \in \mathbb{N} \), the upper bound to which an input is certifiably robust is given by \( \epsilon \) satisfying
Figure 2: Certified accuracy against perturbations targeting the $\ell_1$ and $\ell_2$ norms. Given as a function of the certified radius, the radius around which an input is robust.

Table 2: Examples of the KL divergence between $\mathcal{G}\mathcal{N}(\mu_1, \sigma, s)$ and $\mathcal{G}\mathcal{N}(\mu_2, \sigma, s)$ for small $s$.

<table>
<thead>
<tr>
<th>$s$</th>
<th>$\ell_s$</th>
<th>$d_{KL}(p_1, p_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\ell_1$</td>
<td>$\frac{1}{\sigma}</td>
</tr>
<tr>
<td>2</td>
<td>$\ell_2$</td>
<td>$\frac{1}{\sigma}</td>
</tr>
<tr>
<td>3</td>
<td>$\ell_3$</td>
<td>$\frac{1}{\sigma}</td>
</tr>
<tr>
<td>4</td>
<td>$\ell_4$</td>
<td>$\frac{1}{\sigma^2}</td>
</tr>
</tbody>
</table>

$$
\frac{e^p}{\sigma^p} \prod_{k=1}^{p-1} \left( p \choose k \right) \left( 1 + (-1)^{p-k} \Gamma \left( \frac{p-k+1}{p} \right) \frac{d^{1-\frac{1}{p}} \epsilon^k}{2\sigma^k \Gamma \left( \frac{1}{p} \right)} \right) 
\leq - \log \left( 2\sqrt{p_1 p_2} + 1 - p_1 - p_2 \right) 
$$

The bound given by eq. (21) is found by noting that $||x - x'||_p \leq d^{\frac{1}{p}}||x - x'||_p$, where $d$ is the dimensionality of the data. We can improve upon this naive bound to prove robustness for all norms smaller than $p$ in parallel. Without loss of generality, assume $p$ is even $^1$, then we can prove robustness for every $0 < k \leq p$, where $k$ is even, up to $||x - x'||_k < \epsilon_k$ by solving the constrained problem

$$
\max_{\epsilon_2, \epsilon_4, \ldots, \epsilon_p} \epsilon_2, \epsilon_4, \ldots, \epsilon_p
$$

subject to

$$
\sum_{k=1}^{p} p \choose k \left( 1 + (-1)^{p-k} \Gamma \left( \frac{p-k+1}{p} \right) \frac{d^{1-\frac{1}{p}} \epsilon^k}{2\sigma^k \Gamma \left( \frac{1}{p} \right)} \right) 
\leq - \log \left( 2\sqrt{p_1 p_2} + 1 - p_1 - p_2 \right)
\epsilon_{i+2} \leq \epsilon_i \leq d^{\frac{1}{p} - \frac{1}{i+2}} \epsilon_{i+2}
\epsilon_i > 0, \ 2 \leq i \leq p - 2, \ i \equiv 0 \pmod{2}
$$

Note that the certified radius of robustness around an input is probabilistic because we can only estimate $p_1$ and $p_2$, however, we can bound the probability of error to be arbitrarily small. In practice we follow the methods in [10, 24, 26] for estimating $p_1$ and $p_2$. Prediction error is bounded by collecting $n$ samples of $f(x + \theta)$, where $\theta$ is sampled from a generalized Gaussian distribution, and using the Clopper-Pearson Bernoulli confidence interval to obtain a lower bound estimate of $p_1$ and an upper bound estimate of $p_2$, that holds with probability $1 - \gamma$ over the $n$ samples, where $\gamma \ll 1$. Alternatively, we can use the Hoeffding inequality which gives a lower bound of prediction error of $1 - e^{-2n\epsilon^2}$, where $c$ is the number of classes $|P|$, $n$ is the number of samples and $\epsilon$ is the perturbation size. Clearly the error becomes arbitrarily small as we increase the number of samples.

3. Discussion & experiments

We experimentally validated the certification procedure on the CIFAR-10 [22] and ImageNet [11] datasets. The base classifier is ResNet-50 on ImageNet and ResNet-110 on CIFAR-10 [19]. Given an input $x$ and a classifier $f$ the certification procedure is as follows:

1. Collect $n_0$ Monte Carlo samples of $f(x + \theta_j)$ to estimate the true class $y_j$, where $\theta_j \sim \mathcal{G}\mathcal{N}(0, \sigma, s)$ and $j \in \{1, ..., n_0\}$, with confidence $1 - \gamma_0$.

2. Use $n_1$ Monte Carlo samples to estimate $\hat{p}_1$, a lower bound of the probability of the most-likely class with confidence $1 - \gamma_1$. We follow Cohen et al. [10] for estimating $p_2$, an upper bound of the probability of the second most-likely class, who noticed nearly all probability mass on other classes is placed on the second most-likely class and so use $\hat{p}_2 = 1 - \hat{p}_1$.

3. Use $\hat{p}_1$, $\hat{p}_2$ and eq. (20) or eq. (21) to find a certified radius around $x$.

For all experiments we use $n_0 = 100, n_1 = 100, 000, \gamma_{(0,1)} = 0.001, \sigma = 0.25$ and certify 400 test set examples for both CIFAR-10 and ImageNet datasets.

$^1$We perform experiments measuring the effect that various $\sigma$ have on the certified radius in appendix E.
Certified radius trade-off between weaker than Lecuyer et al. We compare with Lecuyer et al. weaker than related work in either norm. In fig. 2a and fig. 2b, we find a certified radius against perturbations targeting the $\ell_1$ norm. In general, the largest certified regions come against perturbations targeting the $\ell_1$ norm. In appendix F, we show qualitative examples of inputs smoothed with generalized Gaussian noise and the corresponding robustness guarantees in the $\ell_1$, $\ell_2$, and $\ell_3$ norms.

While the primary boon of our certification procedure is its ability to certify inputs to adversarial perturbations beyond $\ell_1$ and $\ell_2$ norms, the method is not substantially weaker than related work in either norm. In fig. 2a and fig. 2b, we compare with Lecuyer et al. [24] and Li et al. [26] for $\ell_1$ norm certificates. Given estimates $\tilde{p}_1$ and $\tilde{p}_2$, Lecuyer et al. [24] find a certified radius against $\ell_1$ perturbations given by $\frac{\sigma}{2} \log(\tilde{p}_1/\tilde{p}_2)$, while Li et al. [26] find a certified radius against $\ell_1$ perturbations given by $\sigma \log(1 - \tilde{p}_1 + \tilde{p}_2)$. Li et al. [26] and Teng et al. [31] show that this robustness guarantee is tight for the $\ell_1$ norm. Our $\ell_1$ certificates are slightly weaker than Lecuyer et al. [24], and both are dominated by Li et al. [26] who obtain the tightest possible certificates.

In fig. 2c and fig. 2d, we compare with Lecuyer et al. [24], Li et al. [26], and Cohen et al. [10] for $\ell_2$ norm certificates. Our $\ell_2$ certificates strictly dominate Lecuyer et al. [24], and are approximately equivalent to Li et al. [26]. This equivalence is to be expected since our certificates are closely related to Li et al. [26] certificates, which are based on the Rényi divergence between two Gaussians, while ours are based on KL divergence. Clearly, we could improve upon this $\ell_2$ guarantee if we used the chi-squared distance instead of KL divergence and a standard Gaussian smoothing measure, as proved by Proposition 1. However, our aim is to show the general capacity of the generalized Gaussian as a smoothing measure for certification.

### 3.2. Robustness trade-offs between different $\ell_p$ norms.

As described by eq. (21), to obtain robustness guarantees in $\ell_{p>2}$ norms we must factor in required robustness guarantees in smaller $\ell_p$ norms. For example, to prove robustness up to $\|x - x'\|_3 < \epsilon_3$ and $\|x - x'\|_1 < \epsilon_1$ we find $\epsilon_1$ and $\epsilon_3$ satisfying

\[
\frac{1}{\sigma^3} \epsilon_3^3 + \frac{3}{\sigma \Gamma(\frac{3}{3})} \epsilon_1^3 \leq -\log(2\sqrt{\tilde{p}_1\tilde{p}_2} + 1 - \tilde{p}_1 - \tilde{p}_2) \tag{26}
\]

\[
\wedge
\]

\[
0 < \epsilon_3 \leq \epsilon_1 \leq d^2 \epsilon_3,
\]
and to prove robustness up to \( \|x - x'\|_4 < \epsilon_4 \) and \( \|x - x'\|_2 < \epsilon_2 \) we find \( \epsilon_2 \) and \( \epsilon_4 \) satisfying

\[
\frac{1}{\sigma^4} \epsilon_4^4 + \frac{6\Gamma\left(\frac{3}{4}\right)}{\sigma^2 \Gamma\left(\frac{1}{4}\right)} \epsilon_2^2 \leq -\log(2\sqrt{\hat{p}_1 \hat{p}_2} + 1 - \hat{p}_1 - \hat{p}_2) \\
\wedge \quad 0 < \epsilon_4 \leq \epsilon_2 \leq d_4^4 \epsilon_4,
\]

(27)

We visualize this trade-off in fig. 3 for \( \ell_3 \) and \( \ell_4 \) norms. That is, the trade-off in certified robustness between those norms and certified robustness in \( \ell_1 \) and \( \ell_2 \), respectively. We visualize the trade-off as we vary the noise scale \( \sigma \), assuming a robust classifier that classifies inputs correctly with \( \hat{p}_1 = 0.99 \) and \( \hat{p}_2 = 0.01 \). We can smoothly exchange robustness in one norm for robustness in another norm. For example, given \( \sigma = 1 \) and a CIFAR-10 input, we can reduce the guaranteed robustness in the \( \ell_3 \) norm from an approximate certified radius of 0.86 to approximately 0, and increase the guaranteed robustness in the \( \ell_1 \) norm from a certified radius of 0.86 to 1.44. In fig. 4, we show certified accuracy as a function of certified radius in the \( \ell_3 \), \( \ell_4 \), and \( \ell_5 \) norms on the CIFAR-10 and ImageNet datasets. To find the maximum \( \epsilon_3 \) we solve eq. (26) such that \( \epsilon_3 = \epsilon_1 \). Similarly for \( \epsilon_4 \) we solve eq. (27) such that \( \epsilon_4 = \epsilon_2 \), and extend this line of reasoning to find \( \epsilon_5 = \epsilon_3 = \epsilon_1 \) for the \( \ell_5 \) norm. Clearly, we can find non-negligible certified radii in norms outside of \( \ell_1 \) and \( \ell_2 \).

3.3. Robustness guarantees as \( \ell_p \to \infty \).

An immediate question arises when observing our certification procedure, can we find non-vacuous robustness guarantees for arbitrarily large \( \ell_p \) norms, where \( p \) is even? \(^1\) \(^2\)

Given eq. (23), note that \((\frac{1}{\Gamma(\frac{1}{p})})^{\frac{1}{p}+\frac{1}{2}}(\frac{(\frac{1}{p}+1)^{\frac{1}{p}}}{2\Gamma(\frac{1}{p})}) \geq 1, \forall 1 \leq k \leq p, \) where \( k \) is even, and as \( p \to \infty \), \( \exists k \) such that \((\frac{1}{\Gamma(\frac{1}{p})})^{\frac{1}{p}+\frac{1}{2}}(\frac{(\frac{1}{p}+1)^{\frac{1}{p}}}{2\Gamma(\frac{1}{p})}) \to \infty \). We must therefore solve the problem given in eq. (22)-eq. (25), where eq. (23) is given by

\[
\frac{c_2 \epsilon_2^2}{\sigma^2} + \frac{c_4 \epsilon_4^4}{\sigma^4} + ... + \frac{c_p \epsilon_p^p}{\sigma^p} \leq -\log(2\sqrt{\hat{p}_1 \hat{p}_2} + 1 - \hat{p}_1 - \hat{p}_2)
\]

(28)

where \( c_k \in \mathbb{R}_{>1}, 1 \leq k \leq p, k \equiv 0 \pmod{2} \)

(29)

To satisfy eq. (24), we can find \( \epsilon_2, \epsilon_4, ..., \epsilon_p \) such that \( \epsilon_2 = \epsilon_4 = ... = \epsilon_p \); we refer to this value as \( \epsilon \), and eq. (28) becomes

\[
\frac{c_2 \epsilon^2}{\sigma^2} + \frac{c_4 \epsilon^4}{\sigma^4} + ... + \frac{c_p \epsilon^p}{\sigma^p} \leq -\log(2\sqrt{p_1 p_2} + 1 - p_1 - p_2)
\]

(30)

where \( c_k \in \mathbb{R}_{>1}, 1 \leq k \leq p, k \equiv 0 \pmod{2} \)

(31)

For a fixed \( p_1, p_2, \sigma \), since \( \forall k, c_k \geq 1 \), and \( \exists k \) such that \( c_k \to \infty \) when \( p \to \infty \), to satisfy the inequality in eq. (30), we must have \( \epsilon \to 0 \). If we do not fix \( \sigma \) then we require \((\frac{\sigma}{\epsilon})^k \to 0 \) as \( c_k \to \infty \), and so to certify a non-negligible radius, \( \epsilon \), we require \( \sigma \to \infty \). However, as \( \sigma \to \infty \), the randomized smoothing will cause the input to become too noisy for any classifier to achieve low prediction error.

Clearly, as \( p \) grows the largest possible certified radius becomes smaller, because our bound requires this robustness

\(^1\)Equivalent results for this section can be found when \( p \) is not even.

\(^2\)The subject of simultaneous robustness over every \( \ell_p \) norm is expanded upon in appendix G.
guarantee holds for every norm smaller than $p$. One may wonder if we can find an $\ell_p$ norm in which we can certify a non-vacuous radius that approximates the $\ell_{\infty}$ norm arbitrarily well. The difference in volume between a unit ball in the $\ell_p$ norm and $\ell_{\infty}$ norm is given by $\Gamma(1+1/p)^d/\Gamma(1+d/p)$, where $d$ is the data dimensionality. Unfortunately, the error in the approximation is dependent on the data dimensionality. For example, for an ImageNet input where $d = 3 \times 224 \times 224$, if we require the ratio of volumes between an $\ell_p$ unit ball and $\ell_{\infty}$ unit ball to be larger than 0.99, we must take $p = 9 \times 3 \times 224 \times 224$.

3.4. How tight is the bound?

The difference between the certified area and the size of an adversarial perturbation gives a tightness estimate. If the certified radius is close to the size of an adversarial perturbation this implies the bound is close to optimal. To check how tight our bound is we ran the PGD attack [28] minimizing perturbations in the $\ell_2$ norm. Because the certification procedure requires the addition of generalized Gaussian noise to the input, the gradient is highly stochastic, leading to extremely slow convergence of the PGD attack. We circumvent this stochasticity by optimizing using the Expectation Over Transformation [2] – we use 1000 Monte Carlo samples to estimate the gradient of an input during the attack. Figure 5 gives attack results on CIFAR-10 along with the certified radius of 400 inputs. We find adversarial examples with norms within $2 - 2.5 \times$ the certified radius. Unfortunately, this does not inform us if our bound is loose or if the attack is sub-optimal. We leave a more rigorous investigation of assessing the tightness of our bound for future work.

4. Conclusion

Randomized smoothing has offered a promising approach to scaling robustness guarantees to large architectures. By extending the framework developed by Li et al. [26], we showed how different choices of divergences affects the certified radius of robustness around an input. We verified that Rényi divergence is superior to other common $f$-divergences in this framework, for certifying an input against $\ell_2$ perturbations. We then showed that a generalized Gaussian smoothing measure leads to robustness guarantees against any $\ell_p$ ($p \in \mathbb{N}_{>0}$) minimized adversarial perturbation, however, non-negligible certified radii are only available for small $\ell_p$ norms.

Acknowledgements

Jamie Hayes is funded by a Google PhD Fellowship in Machine Learning.

Figure 5: The certified radius and size of adversarial perturbations for 400 CIFAR-10 test inputs using a PGD attack optimizing the $\ell_2$ norm. As a guide to assess how close the certified radius is to adversarial perturbation size, we also display $2 \times$ the certified radius of an input.

References


