Input-Specific Robustness Certification for Randomized Smoothing

Ruoxin Chen[1], **Jie Li**[1 *], **Junchi Yan**[1], **Ping Li**[2], **Bin Sheng**[1]

[1] Shanghai Jiao Tong University
[2] The Hong Kong Polytechnic University
{chenruoxin, lijiecs, yanjunchi}@sjtu.edu.cn, p.li@polyu.edu.hk, shengbin@cs.sjtu.edu.cn

Abstract

Although randomized smoothing has demonstrated high certified robustness and superior scalability to other certified defenses, the high computational overhead of the robustness certification bottlenecks the practical applicability, as it depends heavily on the large sample approximation for estimating the confidence interval. In existing works, the sample size for the confidence interval is universally set and agnostic to the input for prediction. This Input-Agnostic Sampling (IAS) scheme may yield a poor Average Certified Radius (ACR)-runtime trade-off which calls for improvement. In this paper, we propose Input-Specific Sampling (ISS) acceleration to achieve the cost-effectiveness for robustness certification, in an adaptive way of reducing the sampling size based on the input characteristic. Furthermore, our method universally controls the certified radius decline from the ISS sample size reduction. The empirical results on CIFAR-10 and ImageNet show that ISS can speed up the certification by more than three times at a limited cost of 0.05 certified radius. Meanwhile, ISS surpasses IAS on the average certified radius across the extensive hyperparameter settings. Specifically, ISS achieves ACR=0.958 on ImageNet in 250 minutes, compared to ACR=0.917 by IAS under the same condition. We release our code in https://github.com/roy-ch/Input-Specific-Certification.

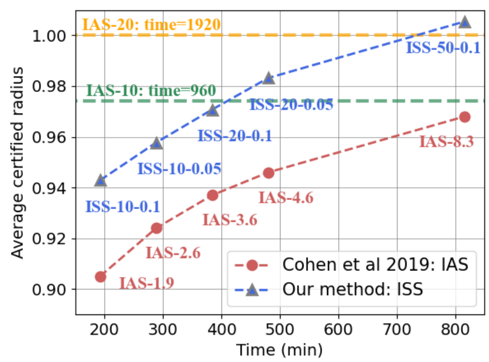


Figure 1: ISS achieves a better ACR-time trade-off than IAS. $\mathrm{ISS}-\mathrm{c}_1-\mathrm{c}_2$ denotes that ISS accelerates $\mathrm{IAS}-\mathrm{c}_1$ at the controllable certified radius decline ($\leq c_2$). $\mathrm{IAS}-\mathrm{c}_1$ denotes that IAS accelerates $\mathrm{IAS}-10, \mathrm{IAS}-20$ by reducing the sample size to $c_1 \times 10,000$. ISS always surpasses IAS on ACR in the same time cost. The results are evaluated on the ImageNet ($\sigma = 1.0$) model trained by (Jeong and Shin 2020).

Introduction

Neural networks are known susceptible to adversarial attacks (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015). A line of empirical defenses (Buckman et al. 2018; Song et al. 2018) have been proposed to defend adversarial attacks, but are often broken by the newly devised stronger attacks (Athalye, Carlini, and Wagner 2018). Existing certified defenses (Wong et al. 2018; Raghunathan, Steinhardt, and Liang 2018; Cohen, Rosenfeld, and Kolter 2019) provide the theoretical guarantees for their robustness. In particular, *Randomized smoothing* (Cohen, Rosenfeld, and Kolter 2019) is one of the few certified defenses that can scale to ImageNet-scale classification task, showing its great potential for wide application. Moreover, randomized smoothing has shown high robustness against various types of adversarial attacks, including norm-constrained perturbations (e.g. $\ell_0, \ell_2, \ell_\infty$ norms) and image transformations (e.g. rotations and image shift).

Despite these advances, randomized smoothing suffers the costly robustness certification. Specifically, computing a certified radius close to the exact value needs a relatively tight lower bound of the top-1 label probability, which requires running forward passes on a large number of samples (Salman et al. 2019; Jeong and Shin 2020). Such expensive overheads make them less applicable to the real-world scenarios. Some works (Jia et al. 2020; Feng et al. 2020) proposed to leverage the runner-up label probability in the certification, but their performances may suffer from the inevitable loss in the simultaneous confidence intervals. Traditionally, the robustness certification is accelerated by reducing the sample size used for estimating the lower bound (Cohen, Rosenfeld, and Kolter 2019; Jia et al. 2020), but the vanilla sample size reduction will lead to a poor ACR-runtime trade-off. It is critical to develop a cost-effective certification method.

In this paper, we propose Input-Specific Sampling (ISS) to speed up the certification for randomized smoothing, without hurting too much on the certification performance. The idea behind ISS is to *minimize the sample size for the given input at the bounded cost of the certified radius decline,* in-

*Jie Li is the corresponding author. Email: lijiecs@sjtu.edu.cn stead of directly applying the same sample size to all inputs. The idea is realized by precomputing a mapping from the input characteristics to the sample size. Consequently, ISS can accelerate the certification at a controllable cost. Empirical results validate that ISS consistently outperforms IAS (Cohen, Rosenfeld, and Kolter 2019) on ACR. As shown in Fig. 1, $\mathrm{ISS}-10-0.05$ (ACR=0.958) accelerates the standard certification $\mathrm{IAS}-10$, shortening the certification time $962 \rightarrow 250$ mins at the controllable decline (≤ 0.05). Furthermore, ISS is compatible with all the randomized smoothing works that need confidence interval, since ISS has no additional constraint on the base classifier or the smoothing scheme.

Our contributions can be summarized as follows;

1. We propose Input-Specific Sampling (ISS) to adaptively reduce the sample size for each input. The proposed input-specific sampling, for the first time to our best knowledge, can significantly reduce the cost for accelerating the robustness certification of randomized smoothing.
2. ISS can universally control the difference between the certified radii before and after the acceleration. In particular, the sample size computed by ISS is theoretically tight for bounding the radius decline.
3. The results on CIFAR-10 and ImageNet demonstrate that: 1) ISS significantly accelerates the certification at a controllable decline in the certified radii. 2) ISS consistently achieves a higher average certified radius when compared to the mainstream acceleration IAS.

Related Works

Certified defenses. Neural networks are vulnerable to adversarial attacks (Athalye, Carlini, and Wagner 2018; Eykholt et al. 2018; Kurakin, Goodfellow, and Bengio 2017; Eykholt et al. 2018; Jia and Gong 2018). Compared to empirical defenses (Goodfellow, Shlens, and Szegedy 2015; Svoboda et al. 2019; Buckman et al. 2018; Ma et al. 2018; Guo et al. 2018; Dhillon et al. 2018; Xie et al. 2017; Song et al. 2018), certified defenses can provide provable robustness guarantees for their predictions. Recently, a line of certified defenses have been proposed, including dual network (Wong et al. 2018), convex polytope (Wong and Kolter 2018), CROWN-IBP (Zhang et al. 2019), Lipschitz bounding (Cisse et al. 2017). However, those certified defenses suffer from either the low scalability or the hard constraints on the neural network architecture.

Randomized smoothing. In the seminal work (Cohen, Rosenfeld, and Kolter 2019), the authors for the first time propose randomized smoothing to defend the ℓ_2-norm perturbations, which significantly outperforms other certified defenses. Recently, series of works further extend randomized smoothing to defend various attacks, including $\ell_0, \ell_1, \ell_2, \ell_\infty$-norm perturbations and geometric transformations. For instance, (Levine and Feizi 2020) introduce the random ablation against ℓ_0-norm adversarial attacks. (Yang et al. 2020) propose Wulff Crystal uniform distribution against ℓ_1-norm perturbations. (Awasthi et al. 2020) introduce $\infty \rightarrow 2$ matrix operator for Gaussian smoothing to defend ℓ_∞-norm perturbations. (Fischer, Baader, and Vechev 2020; Li et al. 2020) exploit randomized smoothing to defend adversarial translations. Remarkably, almost all the randomized smoothing works (Salman et al. 2019; Cohen, Rosenfeld, and Kolter 2019; Zhai et al. 2020; Jeong and Shin 2020; Yang et al. 2020; Jia et al. 2020) have achieved superior certified robustness to other certified defenses in their respective fields.

Robustness certification in randomized smoothing. Despite its sound performance, the certification of randomized smoothing is seriously costly. Unfortunately, accelerating the certification is a fairly under-explored field. The mainstream acceleration method (Jia et al. 2020; Feng et al. 2020), which we call IAS, is to apply a smaller sample size for certifying the radius. However, IAS accelerates the certification at a seriously sacrifice ACR and the certified radii of specific inputs. Therefore, it calls for approaches to achieve a better time-cost trade-off, which is the main purpose of this paper.

Preliminaries

Randomized smoothing The basic idea of randomized smoothing (Cohen, Rosenfeld, and Kolter 2019) is to generate a smoothed version of the base classifier f. Given an arbitrary base classifier $f(x) : \mathbb{R}^d \rightarrow \mathcal{Y}$ where $\mathcal{Y} = \{1, \ldots, k\}$ is the output space, the smoothed classifier $g(\cdot)$ is defined as:

$$g(x) := \arg\max_{c \in \mathcal{Y}} Pr[f(x') = c],\ x' \sim \mathcal{N}(x, \sigma^2 I^d) \quad (1)$$

$g(x)$ returns the most likely predicted label of $f(\cdot)$ when input the data with Gaussian augmentation $x' \sim \mathcal{N}(x, \sigma^2 I^d)$. The tight lower bound of ℓ_2-norm certified radius (Cohen, Rosenfeld, and Kolter 2019) for the prediction $c_A = g(x)$ is:

$$\begin{aligned} &\sigma\Phi^{-1}(p_A) \\ &\text{where } p_A := Pr[f(x') = c_A],\ x' \sim \mathcal{N}(x, \sigma^2 I^d) \end{aligned} \quad (2)$$

where Φ^{-1} is the inverse of the standard Gaussian CDF. We emphasize that computing the deterministic value of $g(x)$ is impossible because $g(\cdot)$ is built over the random distribution $\mathcal{N}(x, \sigma^2 I^d)$. Therefore, we use Clopper-Pearson method (Clopper and Pearson 1934) to guarantee $Pr[f(x') = c_A] > Pr[f(x') = c], \forall c \neq c_A$ with the confidence level $1 - \alpha$, and then we have $g(x) = c_A$ with the confidence level $1 - \alpha$.

Robustness certification In practice, the main challenge in computing the radius $\sigma\Phi^{-1}(p_A)$ is that p_A is inaccessible because iterating all possible $f(x') : x' \in \mathbb{R}^d$ is impossible. Therefore, we estimate $\underline{p_A}$, the standard one-sided Clopper-Pearson confidence lower bound of p_A instead of p_A and certify a lower bound $\sigma\Phi^{-1}(\underline{p_A})$. Estimating a tight $\underline{p_A}$ needs a large size of samples for $f(x') : x' \sim \mathcal{N}(x, \sigma^2 I^d)$. Generally, the estimated $\underline{p_A}$ increases with the sample size[1].

[1]The seminal work (Cohen, Rosenfeld, and Kolter 2019) derives the certified radius: $\frac{\sigma}{2}[\Phi^{-1}(p_A) - \Phi^{-1}(p_B)]$ where p_B is the runner-up label probability. Currently, most works (Cohen, Rosenfeld, and Kolter 2019; Zhai et al. 2020; Jeong and Shin 2020; Jia et al. 2020) compute the certified radius by Eq. (2), which substitutes p_B with $1 - p_A$, to avoid doing interval estimation twice.

Standard certification and vanilla acceleration: IAS The standard certification algorithm (Cohen, Rosenfeld, and Kolter 2019) can be summarized in two steps:

1. **Sampling:** Given the input x, sample k (e.g. $k = 100,000$) iid samples $\{x'_i : i = 1, \dots, k\} \sim \mathcal{N}(x, \sigma^2 I^d)$ and run k times forward passes $\{f(x'_i) : i = 1, \dots, k\}$.
2. **Interval estimation:** Count $k_A = \sum_{i=1}^{k} \mathbb{I}\{f(x'_i) = c_A\}$ ($\mathbb{I}$ denotes the indicator function) where c_A is the label with top-1 label counts. Compute the confidence lower bound $\underline{p_A}$ with the confidence level $1 - \alpha$. Return the certified radius $\sigma\Phi^{-1}(\underline{p_A})$.

The high computation is mainly due to the k times forward passes in **Sampling**. The certification is accelerated by the vanilla sample size reduction, which we call input-agnostic sample size reduction (IAS). This acceleration is at the cost of unpredictable radius declines, which yields a poor ACR-runtime trade-off since it reduces the sample size equally for each input, without considering the input characteristics.

Methodology

We first introduce the notions of Absolute Decline and Relative Decline. Then we propose Input-Specific Sampling (ISS), which aims to use the minimum sample size with the constraint that the radius decline is less than the given bound.

Overview and Main Idea

The key idea of ISS is to appropriately reduce the sample size for each input, instead of applying the same sample size to the certifications for all inputs. Since the sample size reduction will inevitably cause the decline in the certified radius, thus we aim to quantify the radius decline and bound the decline to be less than the pre-specified value. First we define the radius decline as follows:

Definition 1 (**Absolute Decline** $\mathrm{AD}(k; \overline{k}, p)$)**.** *Given the input x and the pre-specified desired sample size $\overline{k}$ (e.g. $\overline{k} = 100,000$), suppose we know p_A of x, Absolute Decline $\mathrm{AD}(k; \overline{k}, p)$ is the gap between the radius certified at the sample size $\overline{k}$ and the radius certified at $k : k \leq \overline{k}$:*

$$\mathrm{AD}(k; \overline{k}, p_A) := \underbrace{\sigma\Phi^{-1}(\underline{p_1})}_{\textit{Desired radius}} - \underbrace{\sigma\Phi^{-1}(\underline{p_2})}_{\textit{Estimated radius}}$$
$$\text{where } \underline{p_1} = \mathbf{B}(\alpha;\ p_A\overline{k},\ \overline{k} - p_A\overline{k} + 1),$$
$$\underline{p_2} = \mathbf{B}(\alpha;\ p_A k,\ k - p_A k + 1) \tag{3}$$

where $\mathbf{B}(\alpha; k_A, k - k_A + 1)$ denotes the one-sided Clopper-Pearson lower bound (Clopper and Pearson 1934) with the confidence level $1-\alpha$, which is equal to the αth quantile from a Beta distribution with shape parameters $k_A, k - k_A + 1$.

Definition 2 (**Relative Decline** $\mathrm{RD}(k; \overline{k}, p_A)$)**.** *Similar to absolute decline, Relative Decline $\mathrm{RD}(k; \overline{k}, p_A)$ is*

$$\mathrm{RD}(k; \overline{k}, p_A) := \frac{\sigma\Phi^{-1}(\underline{p_1}) - \sigma\Phi^{-1}(\underline{p_2})}{\sigma\Phi^{-1}(\underline{p_1})}$$
$$\text{where } \underline{p_1} = \mathbf{B}(\alpha;\ p_A\overline{k},\ \overline{k} - p_A\overline{k} + 1),$$
$$\underline{p_2} = \mathbf{B}(\alpha;\ p_A k,\ k - p_A k + 1) \tag{4}$$

Algorithm 1: Compute ISS mapping $\psi_{ISS}(\cdot)$

Input: The maximum decline U, the decline type, the desired sample size $\overline{k}$, the noise level σ, the confidence level $1 - \alpha$, the length of the subinterval δ
Output: the ISS mapping $\psi_{ISS}(\cdot)$

for $N = 0, 1, 2, \dots, \frac{1}{\delta}$ **do**
  $p \leftarrow N \cdot \delta$;
  Compute $\overline{r} = \sigma \cdot \Phi^{-1}\left(\mathrm{B}(\alpha;\ \mathrm{p}\overline{\mathrm{k}},\ \overline{\mathrm{k}})\right)$;
  Compute the minimum required certified radius:
  If the decline type is AD: $\tilde{r} \leftarrow \overline{r} - \mathrm{U_{AD}}$ or
  If the decline type is RD: $\tilde{r} \leftarrow (1 - \mathrm{U_{RD}})\overline{\mathrm{r}}$;
  if $\tilde{r} \leq 0$ **then**
    $\psi_{ISS}(p) \leftarrow 0$;
  else
    $\psi_{ISS}(p) \leftarrow \arg\min_k \sigma \cdot \Phi^{-1}\left(\mathrm{B}(\alpha;\ \mathrm{pk},\ \mathrm{k})\right) \geq \tilde{r}$;
  end if
end for
Return $\psi_{ISS}(p) : p = \delta, 2\delta, \dots, 1$;

Remark 1. *The absolute (or relative) decline is the expected gap between the radius certified at the sample size $\overline{k}$ and k when fixing $k_A/k \equiv p_A$ where $k_A := \sum_{i=1}^{k} \mathbb{I}\{f(x'_i) = c_A\}$. It connects the expected radius decline to the sample size when given p_A. In particular, when $\overline{k} = \infty$, the absolute (or relative) decline measures the gap between the optimal certified radius that randomized smoothing can provide and the radius certified at the sample size k.*

Formulate our key idea Given the input x and the pre-specified upper bound of the decline $\mathrm{U_{AD}} \in \mathbb{R}^+$ (or $\mathrm{U_{RD}} \in \mathbb{R}^+$), our idea for AD (or RD) is formulated as follows:

1. find $\min k$ with the constraint $\mathrm{AD}(k; \overline{k}, p) \leq \mathrm{U_{AD}}$.
2. find $\min k$ with the constraint $\mathrm{RD}(k; \overline{k}, p) \leq \mathrm{U_{RD}}$.

In practice, solving the above two problems is non-trival because p_A of x is inaccessible. Simply treating the estimated k_A/k as p_A is obviously unreasonable. We propose ISS, a practical solution to the above two problems.

Certification with Input-Specific Sampling

Fig. 2 shows an overview. Given the input x, we first estimate a relatively loose two-sided Clopper-Pearson confidence interval $p_A \in [p_{\mathrm{low}}, p_{\mathrm{up}}]$ by k_0 samples where $k_0 < \overline{k}$ is a relatively small sample size. Given $\overline{k}, \mathrm{U_{AD}}$ (or $\mathrm{U_{AD}}$), ISS assigns the sample size $\hat{k}$ for certifying $g(x)$ where $\hat{k}$ is:

$$\hat{k} = \max\left(\psi(p_{\mathrm{low}}), \psi(p_{\mathrm{up}})\right) \tag{5}$$

$$\text{For Absolute Decline} : \psi(p) := \arg\min_k\ \mathrm{AD}(\mathrm{k}; \overline{\mathrm{k}}, \mathrm{p}) \leq \mathrm{U_{AD}}$$

$$\text{For Relative Decline} : \psi(p) := \arg\min_k\ \mathrm{RD}(\mathrm{k}; \overline{\mathrm{k}}, \mathrm{p}) \leq \mathrm{U_{RD}}$$

Formally, we present the following two propositions to theoretically prove that $\hat{k}$ (AD) computed from Eq. (5) is optimal. Prop. 1 guarantees that the sample size $\hat{k}$ computed from Eq. (5) must satisfy the constraint $\mathrm{AD}(\hat{k}; \overline{k}, p_A) \leq$

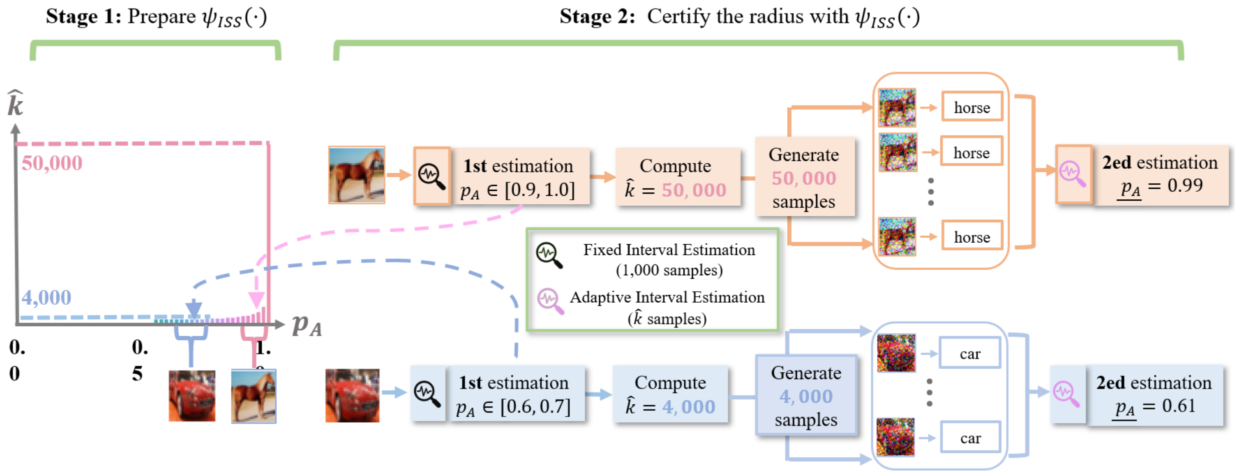


Figure 2: Overview of the robustness certification with ISS. In Stage 1, we compute $\psi_{ISS}(\cdot)$, a mapping from p_A to $\hat{k}$. In Stage 2, given the image x, we first loosely estimate the confidence interval for p_A and determine the sample size for x with $\psi_{ISS}(\cdot)$.

Algorithm 2: Certification with input-specific sampling (ISS)

Input: The input x, the base classifier f, the maximum sample size $\overline{k}$, the sample size $k_0 : k_0 \le \overline{k}$, the confidence level α, the ISS mapping $\psi_{ISS}(\cdot)$
Output: Prediction *pred*, radius r

Sample k_0 noisy samples $x'_1, \dots, x'_k \sim \mathcal{N}(x, \sigma^2 I^d)$;
Compute the prediction:
  $pred \leftarrow \arg\max_{y \in \mathcal{Y}} \sum_{i=1}^{k_0} \mathbb{I}\{f(x'_i) = y\}$;
Count $k_A^0 \leftarrow \max_{y \in \mathcal{Y}} \sum_{i=1}^{k_0} \mathbb{I}\{f(x'_i) = y\}$;
Compute the two-sided confidence interval:
  $p_{\text{low}} \leftarrow \mathbf{B}(\alpha/2;\ k_A^0,\ k_0 - k_A^0 + 1)$
  $p_{\text{up}} \leftarrow \mathbf{B}(1 - \alpha/2;\ k_A^0 + 1,\ k_0 - k_A^0)$;
Compute $\hat{k} \leftarrow \max(\psi_{ISS}(p_{\text{low}}), \psi_{ISS}(p_{\text{up}}))$;
Sample $\max(\hat{k} - k_0, 0)$ noisy samples:
  $x'_{k_0+1}, \dots, x'_{\hat{k}} \sim \mathcal{N}(x, \sigma^2 I^d)$;
Count $k_A \leftarrow \max_{y \in \mathcal{Y}} \sum_{i=1}^{\hat{k}} \mathbb{I}\{f(x'_i) = y\}$;
Compute the one-sided confidence lower bound:
  $\underline{p_A} \leftarrow \mathbf{B}(\alpha;\ k_A,\ \hat{k} - k_A + 1)$;
if $\underline{p_A} < \frac{1}{2}$ **then**
  $pred \leftarrow$ ABSTAIN, $r \leftarrow 0$;
else
  Compute the radius $r \leftarrow \sigma\Phi^{-1}(\underline{p_A})$;
end if
Return *pred* and r;

$\mathrm{U_{AD}}$. Prop. 2 guarantees that $\hat{k}$ is the minimize sample size that can guarantee $\mathrm{AD}(\hat{k}; \overline{k}, p_A) \le \mathrm{U_{AD}}$.

Proposition 1. *[Bounded absolute radius decline] Suppose $p_A \in [p_{\text{low}}, p_{\text{up}}]$ with $1 - \alpha$ confidence level, then we guarantee that there is at least $1 - \alpha$ probability that $\hat{k}$ computed from Eq. (5) satisfies* $\mathrm{AD}(\mathrm{k}; \hat{\mathrm{k}}, \mathrm{p_A}) \le \mathrm{U_{AD}}$.

Proposition 2. *[Tightness for $\hat{k}$] Suppose $p_A \in [p_{\text{low}}, p_{\text{up}}]$ and $\hat{k}$ is computed from Eq. (5), then for an arbitrary sample size $k : k < \hat{k}$, there exists $p_A \in [p_{\text{low}}, p_{\text{up}}]$ that breaks the constraint* $\mathrm{AD}(\mathrm{k}; \hat{\mathrm{k}}, \mathrm{p_A}) \le \mathrm{U_{AD}}$.

Implementation

In the practical algorithm of ISS, we substitute $\psi(\cdot)$ in Eq. (5) with a piecewise constant function approximation $\psi_{ISS}(\cdot)$. The advantage of $\psi_{ISS}(\cdot)$ over $\psi(\cdot)$ is that we can compute $\psi_{ISS}(p) : p \in [0, 1]$ previously before the certification to save the cost in computing $\psi(p_{\text{low}}), \psi(p_{\text{up}})$ in Eq. (5) when certifying the radius for the testing data. Constructing $\psi_{ISS}(\cdot)$ is feasible because that the value of $\psi(p)$ only depends on p when fixing $\overline{k}$, regardless of the testing set or the base classifier architecture. Specifically, $\psi_{ISS}(p)$ is

$$\psi_{ISS}(p) = \begin{cases} \psi(p) & p/\delta \in \mathbb{N} \\ \max(\psi(N_1\delta), \psi(N_2\delta)) & p/\delta \in (N_1, N_2) \end{cases} \tag{6}$$

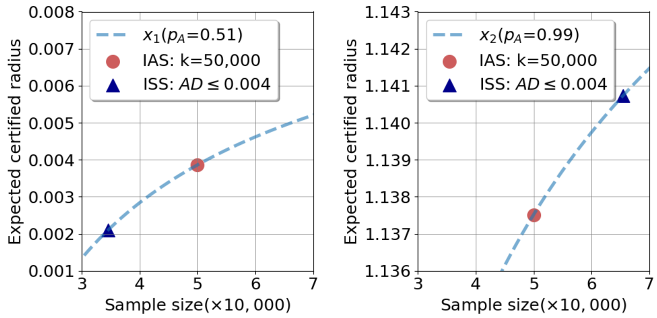
where $N_1, N_2 \in \mathbb{N}$. Obviously, $\forall p \in [0, 1], \psi_{ISS}(p) \ge \psi(p)$, thus Prop. 1 still holds for the substitution $\psi_{ISS}(\cdot)$. Prop. 2 holds for $\psi_{ISS}(\cdot)$ when $p_{\text{low}}/\delta \in \mathbb{N}, p_{\text{up}}/\delta \in \mathbb{N}$.

The practical algorithm can summarized into two stages:

Stage 1: prepare $\psi_{ISS}(\cdot)$. Given $\overline{k}$ and the decline upper bound $\mathrm{U_{AD}}$ (or $\mathrm{U_{RD}}$), compute $\psi_{ISS}(p)$ by Eq. (5) and Eq. (6). The detailed algorithm is shown in Alg. 1.

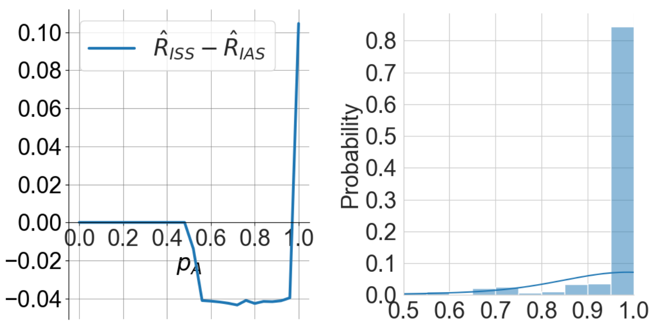
Stage 2: certify the radius with $\psi_{ISS}(\cdot)$. Given x, we first estimate a loose confidence interval $p_A \in [p_{\text{low}}, p_{\text{up}}]$ by k_0 samples. With $[p_{\text{low}}, p_{\text{up}}]$ and $\psi_{ISS}(\cdot)$, we compute $\hat{k}$. Then we estimate the certified radius by sampling $\hat{k}$ noisy samples. The algorithm is shown in Alg. 2.

Compare ISS(AD) to IAS We compare ISS to IAS in Fig. 3a, Fig. 3b where $\hat{R}(k, p_A; \sigma) := \sigma\Phi^{-1}(\mathbf{B}(\alpha; p_A k, k - p_A k + 1))$. As presented, IAS assigns 50,000 for both x_1, x_2, while ISS assigns 35,000 and 65,000 for x_1, x_2 respectively. The sample size of ISS are computed by solving $\mathrm{AD}(\mathrm{k}; 100,000, \mathrm{p_A}) \le 0.0042$. For each certified radius, the



(a) $\hat{R}$-k curve of x_1 ($p_A = 0.51$). (b) $\hat{R}$-k curve of x_2 ($p_A = 0.99$).

Figure 3: ISS certifies higher ACR on x_1, x_2 than IAS.



(a) $\hat{R}_{ISS} - \hat{R}_{IAS}$ w.r.t. p_A (b) p_A distribution (ImageNet).

Figure 4: ISS fits the practical smoothed classifiers.

decline in x_2 certified radius is up to 0.0075 due to the sample size reduction $100,000 \to 50,000$, which is $1.78\times$ $\mathrm{U_{AD}}$ of ISS. For the average certified radius, ISS trades 0.002 radius of x_1 for 0.003 radius of x_2, thus ACR of ISS improves IAS 0.0005 under the same average sample size. The improvement is because ISS tends to assign larger sample sizes to the high-p_A inputs, which meets the property of $\hat{R}(k, p_A; \sigma)$. Namely, $\hat{R}(k + \Delta k, p_A; \sigma) - \hat{R}(k, p_A; \sigma)$ increases with p_A, meaning that assigning larger sample sizes to the high-p_A inputs is more efficient than input-agnostic sampling.

ISS fits the well-trained smoothed classifiers Fig. 4a reports $\hat{R}_{ISS} - \hat{R}_{IAS}$ where $\hat{R}_{ISS}$ denotes the radius certified by ISS of $\overline{k} = 100,000, \mathrm{U_{AD}} = 0.05$ and $\hat{R}_{IAS}$ denotes the radius certified by IAS at $k = 30,000$[2]. We observe that ISS certifies higher certified radius when $p_A > 0.94$. Fig. 4b reports the p_A distribution of the test set [3] on the real ImageNet base classifier ($\sigma = 0.5$) trained by Consistency (Jeong and Shin 2020). We found that the probability mass of p_A distribution is concentrated around $p_A = 1.0$, which is the interval where $\hat{R}_{ISS} - \hat{R}_{IAS} > 0$. Furthermore, ISS is expected to outperform IAS on the smoothed classifiers trained by other algorithms, since their p_A distributions have the similar property (see appendix).

[2]Here we choose to compare ISS to IAS ($k = 30,000$) is because that the average sample size of ISS ($\overline{k} = 100,000, \mathrm{U_{AD}} = 0.05$) on the ImageNet model trained by Consistency (Jeong and Shin 2020) ($\sigma = 0.5$) is roughly $30,000$.

[3]We sample $k = 1000,000$ Monte Carlo samples and approximately regard k_A/k as the exact value of p_A.

Experiments

We evaluate our proposed method ISS on two benchmark datasets: CIFAR-10 (Krizhevsky 2009) and ImageNet (Russakovsky et al. 2015). All the experiments are conducted on CPU (16 Intel(R) Xeon(R) Gold 5222 CPU @ 3.80GHz) and GPU (one NVIDIA RTX 2080 Ti). We observe that the certification runtime is roughly proportional to the average sample size when fixing the model architecture. The hyper-parameters are listed in Table 2. For clarity, $\mathrm{ISS - c_1 - c_2}$ denotes ISS at $\overline{k} = c_1 \cdot 10,000, \mathrm{U_{AD}} = \mathrm{c_2}$, and $\mathrm{IAS - c_1}$ denotes IAS at $k = c_1 \cdot 10,000$. The overhead of computing ψ_{ISS} is reported in Table 3.

Evaluation Metrics

Our evaluation metrics include average sample size, runtime, MAD, ACR and certified accuracy, where MAD denotes the maximum absolute decline between the radius certified before and after the acceleration among all the testing data[4]. ACR and certified accuracy $\mathrm{CA(r)}$ at the radius r are computed as follows:

$$\mathrm{ACR} := \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{(\mathrm{x},\mathrm{y}) \in \mathcal{D}_{\mathrm{test}}} \mathrm{R(x;g)} \cdot \mathbb{I}(\mathrm{g(x) = y}) \tag{7}$$

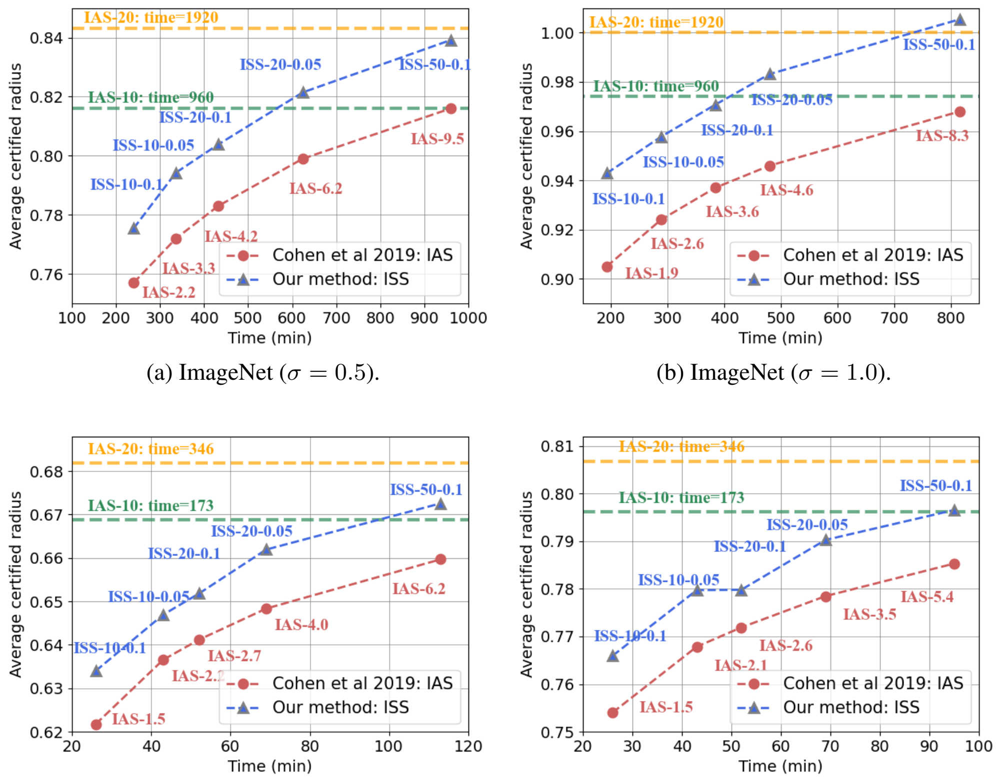
$$\mathrm{CA(r)} := \frac{1}{|\mathcal{D}_{\mathrm{test}}|} \sum_{(\mathrm{x},\mathrm{y}) \in \mathcal{D}_{\mathrm{test}}} \mathbb{I}(\mathrm{R(x;g) > r}) \cdot \mathbb{I}(\mathrm{g(x) = y}) \tag{8}$$

where $\mathrm{R(x;g)}$ denotes the estimated certified radius of $g(x)$.

Overall Analysis of ACR and Runtime

Fig. 5c, Fig. 5d, Fig. 5a, Fig. 5b present the overall empirical results of ISS and IAS on CIFAR-10 and ImageNet. As presented, ISS significantly accelerates the certification for randomized smoothing. Specifically, on ImageNet ($\sigma = 0.5, 1.0$), $\mathrm{ISS{-}10{-}0.05}$, $\mathrm{ISS{-}10{-}0.1}$ reduce the original time cost 962 minutes (the green dotted lines) to roughly $300, 200$ respectively at $\mathrm{U_{AD}} = 0.05, 0.1$ respectively. Overall, the speedups of ISS are even higher on CIFAR-10. We also compare ISS to IAS on two datasets. We found that ISS always achieves higher ACR than IAS in the similar time cost. For ImageNet ($\sigma = 1.0$), $\mathrm{ISS - 20 - 0.05}$ even further improves $\mathrm{IAS - 20}$ by a moderate margin, while the time cost of $\mathrm{ISS - 20 - 0.05}$ is only $0.56\times$ of $\mathrm{IAS - 20}$. The full results are reported in the supplemental material.

[4]Note the speedup of ISS deterministically depends on the p_A distribution of the testing set. Since the smoothed classifiers trained by different training algorithms, including SmoothAdv (Salman et al. 2019), MACER (Zhai et al. 2020) and Consistency (Jeong and Shin 2020), report the similar p_A distributions, ISS will perform similarly on the models trained by other algorithms.



(c) CIFAR-10 ($\sigma = 0.5$). (d) CIFAR-10 ($\sigma = 1.0$).

Figure 5: Overall analysis on ImageNet and CIFAR-10.

σ	Method	Avg	Time (min)	MAD	ACR	0.00	0.50	1.00	1.50	2.00	2.5	3.0	3.5	4.0
0.50	$\text{ISS}_{\text{AD}-10-0.05}$	**32992**	**317**	**0.05**	**0.794**	54.6	49.8	42.4	**33.2**	0.0	0.0	0.0	0.0	0.0
	IAS − 3.3	33000	**317**	0.14	0.77	**54.8**	**50.2**	**43.4**	32.8	0.0	0.0	0.0	0.0	0.0
	$\text{ISS}_{\text{AD}-10-0.10}$	**22144**	**213**	**0.10**	**0.775**	54.6	49.8	42.0	**33.0**	0.0	0.0	0.0	0.0	0.0
	IAS − 2.2	22200	**213**	0.19	0.752	**54.8**	**50.2**	**43.4**	32.6	0.0	0.0	0.0	0.0	0.0
	$\text{ISS}_{\text{AD}-50-0.05}$	**144220**	**1385**	**0.05**	**0.856**	**54.8**	**50.2**	42.8	**34.2**	**29.8**	0.0	0.0	0.0	0.0
	IAS − 14.4	144400	1386	0.15	0.831	**54.8**	**50.2**	**43.4**	33.4	0.0	0.0	0.0	0.0	0.0
	$\text{ISS}_{\text{AD}-50-0.10}$	**95381**	**916**	**0.10**	**0.839**	**54.8**	**50.2**	42.8	**34.2**	0.0	0.0	0.0	0.0	0.0
	IAS − 9.5	95400	**916**	0.20	0.815	**54.8**	**50.2**	**43.4**	33.2	0.0	0.0	0.0	0.0	0.0
1.00	$\text{ISS}_{\text{AD}-10-0.05}$	**25987**	**250**	**0.05**	**0.958**	40.6	36.8	31.8	**28.0**	**24.0**	**20.2**	**17.4**	**13.4**	0.0
	IAS − 2.6	26000	**250**	0.35	0.917	**41.0**	**37.2**	**32.6**	**28.0**	**24.0**	20.0	16.6	0.0	0.0
	$\text{ISS}_{\text{AD}-10-0.10}$	**19209**	**185**	**0.10**	**0.943**	40.2	36.8	31.8	27.4	**24.0**	**20.0**	**17.4**	**13.4**	0.0
	IAS − 1.9	19400	186	0.43	0.903	**41.0**	**37.0**	**32.2**	**28.0**	**24.0**	**20.0**	16.2	0.0	0.0
	$\text{ISS}_{\text{AD}-50-0.05}$	**104037**	**999**	**0.05**	**1.015**	40.6	36.8	32.2	**28.0**	**24.0**	**20.4**	**17.4**	**13.4**	**13.4**
	IAS − 10.4	104200	1000	0.37	0.976	**41.4**	**37.4**	**32.6**	**28.0**	**24.0**	**20.4**	**17.4**	**13.4**	0.0
	$\text{ISS}_{\text{AD}-50-0.10}$	**82899**	**796**	**0.10**	**1.005**	40.6	36.8	32.2	**28.0**	**24.0**	20.0	**17.4**	**13.4**	**13.4**
	IAS − 8.3	83000	797	0.43	0.967	**41.4**	**37.4**	**32.6**	**28.0**	**24.0**	**20.4**	**17.4**	**13.4**	0.0

Table 1: ImageNet: compare ISS to IAS on average sample size (Avg), certification runtime, maximum absolute decline (MAD), verage certified radii (ACR) and certified accuracies (%) on the models trained by Consistency (Jeong and Shin 2020). Results are evaluated on 500 trials by $id = 0, 100, \ldots, 49800, 49900$. The bold denotes better performance under the similar setting.

Results of ISS (AD) on ImageNet

Table 1 reports the results of ISS[5]. Remarkably, ISS reduces the average sample size to roughly $\frac{3}{10}\times, \frac{1}{5}\times$ at the cost of $\text{U}_{\text{AD}} = 0.05, 0.10$ respectively, meaning the speedups are roughly $\frac{10}{3}\times, 5\times$. We found that the MADs of IAS are higher

[5]Here we only report the results at $\sigma = 0.5$, and $\sigma = 1.0$ because the work (Jeong and Shin 2020) only releases the training hyperparameters at $\sigma = 0.5, 1.0$ for consistency training algorithm.

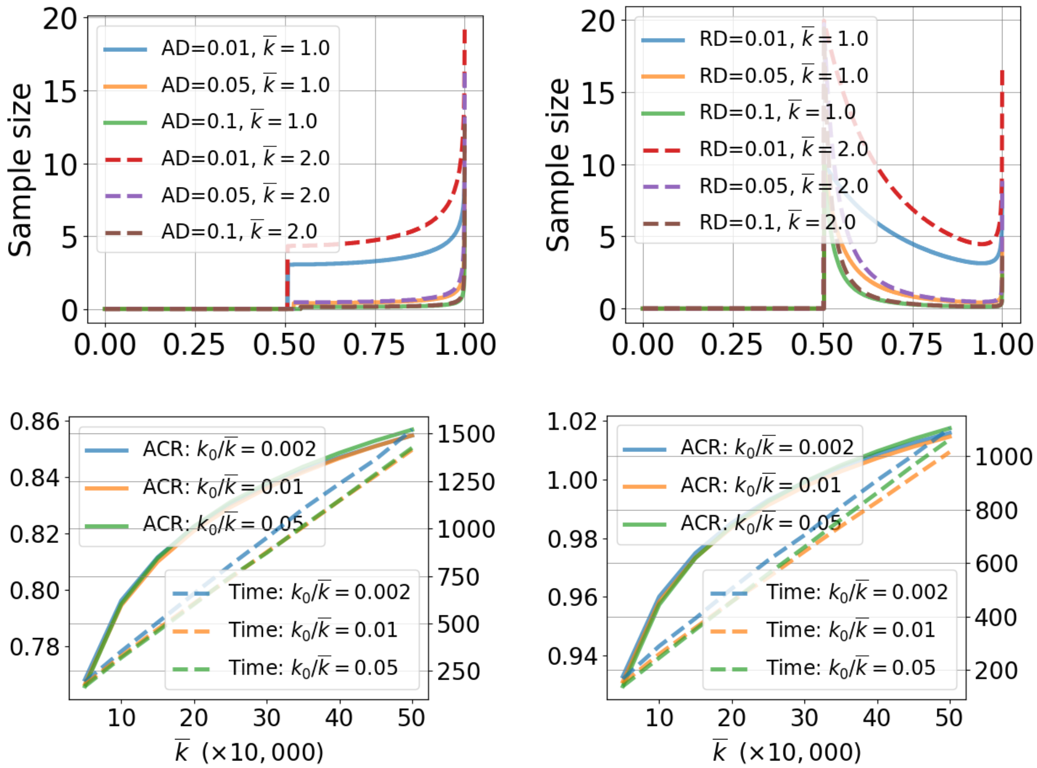


Figure 6: Ablation studies. Upper: k-p_A curves w.r.t. AD, $\overline{\mathrm{k}}$ (Upper Left) and RD, $\overline{\mathrm{k}}$ (Upper Right). Lower: ACR-$\overline{k}$ curves and Time-$\overline{k}$ curves w.r.t. $k_0/\overline{k}$ on ImageNet $\sigma = 0.5$ (Lower Left) and ImageNet $\sigma = 1.0$ (Lower Right).

Dataset	CIFAR-10	ImageNet
Model	ResNe-110	ResNet-50
Training by	MACER	Consistency
$\overline{k}$	100,000, 500,000	
k_0	$0.01\overline{k}$	
σ	0.25, 0.5, 1.0	0.5, 1.0

Table 2: Experiment setting.

AD	Time (s)	RD	Time (s)
0.01	0.70	0.01	39.47
0.05	0.65	0.05	13.52
0.10	0.57	0.10	7.50

Table 3: Runtime for computing ψ_{ISS}.

than ISS, meaning that IAS will cause a large radius decline on the specific inputs. Namely, the MAD of IAS $-$ 10.4 is more than $7\times$ $\mathrm{ISS}_{\mathrm{AD}-50-0.05}$. ISS consistently surpasses IAS on ACR. $\mathrm{ISS}_{\mathrm{AD}-50-0.10}(\sigma = 1.0)$ achieves ACR $=$ 1.005 in 796 minutes while IAS only achieves ACR $= 0.976$ in $1,000$ minutes. We also observe that ISS slightly lower than IAS on the low-radius certified accuracies. It is because ISS tends to assign the small sample sizes to those inputs with low p_A, which inevitably sacrifices the certified radii of low-p_A inputs. Meanwhile, ISS significantly improves the high-radius certified accuracies and ACR in return.

Results of ISS (RD) on ImageNet

Table 4 reports the results of ISS (RD) on ImageNet at $\mathrm{U}_{\mathrm{RD}} = 0.05, 0.10$. ISS reduces the average sample size to roughly $\frac{7}{10}\times$, $\frac{7}{20}\times$ at controllable cost of RD $= 1\%, 5\%$ respectively. Compared to IAS, ISS (RD) also improves ACR.

Results of ISS (AD) on CIFAR-10

Table 5 reports the results of ISS (OF AD) on CIFAR-10. ISS reduces the average sample size to roughly $\frac{1}{5}\times$, $\frac{1}{10}\times$ at $\mathrm{U}_{\mathrm{AD}} = 0.05, 0.10$. Remarkably, ISS still improves ACRs and MADs, high-radius certified accuraries by a moderate margin on CIFAR-10. These empirical comparisons suggest that ISS is a better acceleration.

Ablation Study

Choice on AD or RD As shown in Fig. 6, when p_A : $p_A \in [0.5, 1.0]$ increases, the sample size of ISS (AD) monotonically increases, while the sample size of ISS (RD) first decreases and then increases around $p_A = 1.0$. ISS (AD) can greatly improve ACR, but tends to sacrifice the certified radii of low-p_A inputs a relatively larger proportion. ISS (RD) sacrifices all inputs the same proportion of radius.

σ	Method	Avg	Time	ACR	0.00	0.50	1.00	1.50	2.00	2.5	3.0	3.5
0.50	$\mathrm{ISS}_{\mathrm{RD}-10-0.01}$	**70919**	**682**	**0.809**	**54.8**	**50.2**	**43.4**	**33.2**	0.0	0.0	0.0	0.0
	$\mathrm{IAS}-7.1$	71000	**682**	0.803	**54.8**	**50.2**	**43.4**	**33.2**	0.0	0.0	0.0	0.0
	$\mathrm{ISS}_{\mathrm{RD}-10-0.05}$	**34591**	333	**0.781**	**54.8**	**50.2**	42.2	**33.0**	0.0	0.0	0.0	0.0
	$\mathrm{IAS}-3.5$	34600	**332**	0.772	**54.8**	**50.2**	**43.4**	32.8	0.0	0.0	0.0	0.0
1.00	$\mathrm{ISS}_{\mathrm{RD}-10-0.01}$	**67207**	**646**	**0.966**	**41.4**	**37.4**	**32.6**	**28.0**	**24.0**	**20.4**	**17.4**	**13.4**
	$\mathrm{IAS}-6.7$	67400	647	0.959	41.2	**37.4**	**32.6**	**28.0**	**24.0**	**20.4**	**17.4**	**13.4**
	$\mathrm{ISS}_{\mathrm{RD}-10-0.05}$	**32515**	**313**	**0.935**	**41.2**	37.0	31.8	27.6	23.8	**20.0**	**17.2**	**13.4**
	$\mathrm{IAS}-3.3$	32600	**313**	0.927	41.0	**37.2**	**32.6**	**28.0**	**24.0**	**20.0**	16.8	**13.4**

Table 4: ImageNet: comparison on average sample size (Avg), certification runtime (in minutes), average certified radii (ACR) and certified accuracies (%) on models trained by Consistency (Jeong and Shin 2020).

σ	Method	Avg	Time	MAD	ACR	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.0
0.25	$\mathrm{ISS}_{\mathrm{AD}-10-0.05}$	**22237**	**39**	**0.05**	**0.492**	76.8	68.0	49.4	**38.8**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathrm{IAS}-2.2$	22400	**39**	0.10	0.483	**77.8**	**68.6**	**52.0**	37.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathrm{ISS}_{\mathrm{AD}-10-0.10}$	**10945**	**19**	**0.10**	**0.473**	76.8	68.0	48.8	**37.6**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathrm{IAS}-1.1$	11000	**19**	0.15	0.462	**77.4**	**68.4**	**51.6**	36.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathrm{ISS}_{\mathrm{AD}-50-0.05}$	**98984**	172	**0.05**	**0.529**	77.4	68.4	51.6	**39.8**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathrm{IAS}-9.9$	99000	**171**	0.10	0.518	**77.8**	**69.0**	**52.2**	39.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathrm{ISS}_{\mathrm{AD}-50-0.10}$	**46509**	**81**	**0.10**	**0.512**	77.4	68.4	51.6	**39.4**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathrm{IAS}-4.7$	46600	**81**	0.14	0.501	**77.8**	**68.8**	**52.0**	38.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.50	$\mathrm{ISS}_{\mathrm{AD}-10-0.05}$	**21836**	**38**	**0.05**	**0.647**	60.6	53.0	46.8	39.8	32.4	**26.0**	**19.8**	**13.0**	0.0	0.0	0.0	0.0	0.0
	$\mathrm{IAS}-2.2$	22000	**38**	0.20	0.633	**61.8**	**54.0**	**47.8**	**40.2**	**32.8**	**26.0**	19.4	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathrm{ISS}_{\mathrm{AD}-10-0.10}$	**14620**	**26**	**0.10**	**0.634**	60.6	53.0	46.8	39.4	31.0	**26.0**	**19.6**	**11.2**	0.0	0.0	0.0	0.0	0.0
	$\mathrm{IAS}-1.5$	14800	**26**	0.25	0.621	**61.8**	**54.0**	**47.8**	**40.2**	**32.6**	**26.0**	19.0	0.0	0.0	0.0	0.0	0.0	0.0
	$\mathrm{ISS}_{\mathrm{AD}-50-0.05}$	**91293**	**158**	**0.05**	**0.68**	61.8	54.0	47.6	40.0	32.6	26.0	**20.2**	**14.2**	**10.2**	0.0	0.0	0.0	0.0
	$\mathrm{IAS}-9.1$	91400	**158**	0.20	0.667	**62.2**	**54.4**	**48.0**	**40.2**	**33.0**	**26.6**	19.8	13.4	0.0	0.0	0.0	0.0	0.0
	$\mathrm{ISS}_{\mathrm{AD}-50-0.10}$	**61567**	**107**	**0.10**	**0.673**	61.8	54.0	47.6	40.0	32.6	25.2	**20.0**	**13.8**	0.0	0.0	0.0	0.0	0.0
	$\mathrm{IAS}-6.2$	61600	**107**	0.25	0.659	**62.2**	**54.4**	**47.8**	**40.2**	**33.0**	**26.4**	19.8	12.4	0.0	0.0	0.0	0.0	0.0
1.00	$\mathrm{ISS}_{\mathrm{AD}-10-0.05}$	**21153**	**37**	**0.05**	**0.78**	42.6	40.2	37.2	33.4	30.4	27.0	24.4	21.2	**18.4**	**14.6**	**13.4**	**10.4**	**8.8**
	$\mathrm{IAS}-2.1$	21200	**37**	0.40	0.763	**42.8**	**40.6**	**37.4**	**34.0**	**31.0**	**27.4**	**24.8**	**21.4**	**18.4**	**14.6**	12.8	9.8	8.0
	$\mathrm{ISS}_{\mathrm{AD}-10-0.10}$	**14751**	**26**	**0.10**	**0.766**	42.4	39.6	36.8	32.8	30.0	26.4	24.0	20.6	18.2	**14.4**	**12.8**	**10.2**	**8.8**
	$\mathrm{IAS}-1.5$	14800	**26**	0.50	0.754	**42.8**	**40.4**	**37.4**	**33.8**	**30.8**	**27.4**	**24.8**	**21.4**	**18.4**	**14.4**	**12.8**	9.6	7.2
	$\mathrm{ISS}_{\mathrm{AD}-50-0.05}$	**70204**	**123**	**0.05**	**0.803**	**42.8**	40.4	**37.4**	33.8	30.4	27.4	24.6	**21.4**	**18.4**	14.6	13.4	**10.4**	**9.2**
	$\mathrm{IAS}-7.0$	70400	**123**	0.47	0.79	**42.8**	**40.6**	**37.4**	**34.4**	**31.0**	**27.6**	**25.0**	**21.4**	**18.4**	**14.8**	**13.6**	10.2	8.8
	$\mathrm{ISS}_{\mathrm{AD}-50-0.10}$	**54102**	**94**	**0.10**	**0.797**	**42.8**	40.4	**37.4**	33.8	30.4	27.4	24.6	21.2	**18.4**	14.4	13.0	**10.0**	**9.2**
	$\mathrm{IAS}-5.4$	54200	**94**	0.53	0.785	**42.8**	**40.6**	**37.4**	**34.4**	**31.0**	**27.6**	**25.0**	**21.4**	**18.4**	**14.8**	**13.4**	**10.0**	8.6

Table 5: CIFAR-10: comparison on average sample size (Avg), certification runtime (in minutes), maximum absolute decline (MAD), average certified radii (ACR) and certified accuracies (%) on the models trained by MACER (Zhai et al. 2020). Results are evaluated on 500 testing data of $id = 0, 20, \ldots, 9960, 9980$. The bold denotes better performance.

Impact of p_A and $\overline{k}$ We investigate the impact of p_A and $\overline{k}$ in Fig. 6 (**Upper**). For both AD and RD, the sample size is 0 when $p_A \leq 0.5$. It is because the certified radius is 0 when $p_A \leq 0.5$. As expected, the sample size monotonically increases with $\overline{k}$ and decreases with AD (or RD).

Impact of $k_0/\overline{k}$ We investigate the impact $k_0/\overline{k}$ on the runtime and ACR in Fig. 6 (**Lower**). Too small $k_0/\overline{k}$ results in a loose confidence interval $[p_{\mathrm{low}}, p_{\mathrm{up}}]$, which can cause the ISS sample size $\hat{k}$ to be much larger than required. Too large $k_0/\overline{k}$ may waste too much computation in estimating $[p_{\mathrm{low}}, p_{\mathrm{up}}]$. Our choice $k_0/\overline{k} = 0.01$ performs well across various noise levels on CIFAR-10 and ImageNet.

Conclusion

Randomized smoothing has been suffering from the long certification runtime, but the current acceleration methods are low-efficiency. Therefore, we propose input-specific sampling, which adaptively assigns the sample size. Our work establishes an initial step towards a better performance-time trade-off for the certification of randomized smoothing. Specifically. Our strong empirical results suggest that ISS is a promising acceleratio. Specifically, ISS speeds up the certification by more than $4\times$ only at the controllable cost of 0.10 certified radius on ImageNet. An interesting direction for future work is to make the confidence interval estimation method adapt to the input.

Acknowledgements

This work was supported by the National Key Research and Development Program of China No. 2020YFB1806700, Shanghai Municipal Science and Technology Major Project Grant 2021SHZDZX0102, NSFC Grant 61932014, NSFC Grant 61872241, Project BE2020026, the Key R&D Program of Jiangsu, China. This work is also partially supported by The Hong Kong Polytechnic University under Grant P0030419, P0030929, and P0035358.

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In *ICML*.

Awasthi, P.; Jain, H.; Rawat, A. S.; and Vijayaraghavan, A. 2020. Adversarial robustness via robust low rank representations. In *NeurIPS*.

Buckman, J.; Roy, A.; Raffel, C.; and Goodfellow, I. 2018. Thermometer encoding: One hot way to resist adversarial examples. In *ICLR*.

Cisse, M.; Bojanowski, P.; Grave, E.; Dauphin, Y.; and Usunier, N. 2017. Parseval networks: Improving robustness to adversarial examples. In *ICML*.

Clopper, C. J.; and Pearson, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*.

Cohen, J. M.; Rosenfeld, E.; and Kolter, J. Z. 2019. Certified Adversarial Robustness via Randomized Smoothing. In Chaudhuri, K.; and Salakhutdinov, R., eds., *ICML*.

Dhillon, G. S.; Azizzadenesheli, K.; Lipton, Z. C.; Bernstein, J.; Kossaifi, J.; Khanna, A.; and Anandkumar, A. 2018. Stochastic activation pruning for robust adversarial defense. In *ICLR*.

Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Tramer, F.; Prakash, A.; Kohno, T.; and Song, D. 2018. Physical adversarial examples for object detectors. *USENIX Workshop*.

Feng, H.; Wu, C.; Chen, G.; Zhang, W.; and Ning, Y. 2020. Regularized training and tight certification for randomized smoothed classifier with provable robustness. In *AAAI*.

Fischer, M.; Baader, M.; and Vechev, M. 2020. Certified Defense to Image Transformations via Randomized Smoothing. In *NeurIPS*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *ICLR*.

Guo, C.; Rana, M.; Cisse, M.; and van der Maaten, L. 2018. Countering Adversarial Images using Input Transformations. In *ICLR*.

Jeong, J.; and Shin, J. 2020. Consistency regularization for certified robustness of smoothed classifiers. In *NeurIPS*.

Jia, J.; Cao, X.; Wang, B.; and Gong, N. Z. 2020. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *ICLR*.

Jia, J.; and Gong, N. 2018. AttriGuard: A practical defense against attribute inference attacks via adversarial machine learning. In *USENIX Security Symposium*.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017. Adversarial machine learning at scale. In *ICLR*.

Levine, A.; and Feizi, S. 2020. Robustness Certificates for sparse adversarial attacks by randomized ablation. In *AAAI*.

Li, L.; Weber, M.; Xu, X.; Rimanic, L.; Xie, T.; Zhang, C.; and Li, B. 2020. Provable robust learning based on transformation-specific smoothing. *arXiv preprint arXiv:2002.12398*.

Ma, X.; Li, B.; Wang, Y.; Erfani, S. M.; Wijewickrema, S.; Schoenebeck, G.; Song, D.; Houle, M. E.; and Bailey, J. 2018. Characterizing adversarial subspaces using local intrinsic dimensionality. In *ICLR*.

Raghunathan, A.; Steinhardt, J.; and Liang, P. S. 2018. Semidefinite relaxations for certifying robustness to adversarial examples. In *NeurIPS*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet large scale visual recognition challenge. *IJCV*.

Salman, H.; Li, J.; Razenshteyn, I.; Zhang, P.; Zhang, H.; Bubeck, S.; and Yang, G. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. In *NeurIPS*.

Song, Y.; Kim, T.; Nowozin, S.; Ermon, S.; and Kushman, N. 2018. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *ICLR*.

Svoboda, J.; Masci, J.; Monti, F.; Bronstein, M. M.; and Guibas, L. 2019. Peernets: Exploiting peer wisdom against adversarial attacks. In *ICLR*.

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

Wong, E.; and Kolter, J. Z. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*.

Wong, E.; Schmidt, F.; Metzen, J. H.; and Kolter, J. Z. 2018. Scaling provable adversarial defenses. In *NeurIPS*.

Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; and Yuille, A. 2017. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*.

Yang, G.; Duan, T.; Hu, E.; Salman, H.; Razenshteyn, I. P.; and Li, J. 2020. Randomized smoothing of all shapes and sizes. In *ICML*.

Zhai, R.; Dan, C.; He, D.; Zhang, H.; Gong, B.; Ravikumar, P.; Hsieh, C.; and Wang, L. 2020. MACER: attack-free and scalable robust training via maximizing certified Radius. In *ICLR*.

Zhang, H.; Chen, H.; Xiao, C.; Li, B.; Boning, D.; and Hsieh, C.-J. 2019. Towards Stable and Efficient Training of Verifiably Robust Neural Networks. *arXiv preprint arXiv:1906.06316*.