A Proofs

In this section, we provide proofs of the main results stated in our work.

Theorem 1. *Suppose that $f : \mathbb{R}^D \to \mathbb{R}^d$ is a deterministic function with corresponding continuously differentiable smoothed function $g(x) = \mathbb{E}_{\varepsilon \sim \mathcal{N}(0,\sigma^2 I)} f(x+\varepsilon)$. Then if for all x, $\|f(x)\|_2 = 1$, then $g(x)$ is $L-$Lipschitz in l_2-norm with $L = \sqrt{\frac{2}{\pi\sigma^2}}$.*

Proof. It is known that everywhere differentiable function $g(x)$ with Jacobian matrix $J_g(x) : (J_g(x))_{ij} = \frac{\partial g_i}{\partial x_j}$ is $L-$Lipschitz in l_2-norm with $L = \sup\limits_{x \in \mathbb{R}^D} \|J_g(x)\|_2$, where $\|J_g(x)\|_2 = \sup\limits_{v : \|v\|_2 = 1} \|J_g(x) \cdot v\|_2$ is the spectral norm of $J_g(x)$.

Taking into account the fact that

$$g(x) = \mathop{\mathbb{E}}_{\varepsilon \sim \mathcal{N}(0,\sigma^2 I)} f(x+\varepsilon) = \frac{1}{(2\pi\sigma^2)^{n/2}} \int_{\mathbb{R}^D} f(x+\varepsilon) \exp\left(-\frac{\|\varepsilon\|_2^2}{2\sigma^2}\right) d\varepsilon, \tag{15}$$

we may derive its Jacobian matrix:

$$J_g = \nabla g(x) = \nabla \mathop{\mathbb{E}}_{\varepsilon \sim \mathcal{N}(0,\sigma^2 I)} f(x+\varepsilon) = \nabla \left(\frac{1}{(2\pi\sigma^2)^{n/2}} \int_{\mathbb{R}^D} f(x+\varepsilon) \exp\left(-\frac{\|\varepsilon\|_2^2}{2\sigma^2}\right) d\varepsilon \right) = \tag{16}$$

$$= \nabla \left(\frac{1}{(2\pi\sigma^2)^{n/2}} \int_{\mathbb{R}^D} f(y) \exp\left(-\frac{\|y-x\|_2^2}{2\sigma^2}\right) dy \right) = \tag{17}$$

$$= \frac{1}{M} \int_{\mathbb{R}^D} f(y)(x-y)^\top \exp\left(-\frac{\|y-x\|_2^2}{2\sigma^2}\right) dy, \tag{18}$$

where $M = (2\pi\sigma^2)^{n/2}\sigma^2$. In order to estimate the spectral norm of J_g, one can estimate the norm of dot product with normalized vector v:

$$\|J_g \cdot v\|_2 = \left\| \frac{1}{M} \int_{\mathbb{R}^D} f(y)(x-y)^\top \cdot v \exp\left(-\frac{\|y-x\|_2^2}{2\sigma^2}\right) dy \right\|_2. \tag{19}$$

Here, we apply a trick: it is possible to rotate vectors in dot product $(x-y)^\top \cdot v$ in such a way that one of the resulting vectors will have one nonzero component after rotation (without loss of generality, assume that this is the first component, e_1). Namely, given a rotation matrix $Q = Q(v)$ that is unitary ($QQ^\top = I$), the expression from Eq. 19 becomes

$$\|J_g \cdot v\|_2 = \left\| \frac{1}{M} \int_{\mathbb{R}^D} f(y)(x-y)^\top QQ^\top \cdot v \exp\left(-\frac{\|y-x\|_2^2}{2\sigma^2}\right) dy \right\|_2. \tag{20}$$

Now, since the rotation does not affect the l_2-norm, $\|Q^\top v\|_2 = \|v\|_2 = 1$ and thus $Q^\top v = e$ and $|e_1| = 1$. More than that, under the change of the variables $z^\top = (x-y)^\top Q$, the following holds:

- $\|z^\top\|_2 = \|(x-y)^\top Q\|_2 = \|(x-y)^\top\|_2$ since rotation is l_2-norm preserving operation;
- $y = x - Qz$;
- $dz = -Qdy$ for the diffentials, leading to $dy = -Q^\top dz$.

Thus, expression from Eq. 20 becomes

$$\|J_g \cdot v\|_2 = \left\| \frac{1}{M} \int_{\mathbb{R}^D} f(x - Qz) z^\top e \exp\left(-\frac{\|z\|_2^2}{2\sigma^2}\right) (-Q^\top) dz \right\|_2 . \tag{21}$$

Now, we bound the norm from Eq. 21 using Cauchy–Schwarz inequality:

$$\|J_g \cdot v\|_2 \leq \frac{1}{M} \int_{\mathbb{R}^D} \left\| f(x - Qz) z^\top e \exp\left(-\frac{\|z\|_2^2}{2\sigma^2}\right) (-Q^\top) \right\|_2 dz \leq \tag{22}$$

$$\leq \frac{1}{(2\pi\sigma^2)^{n/2}\sigma^2} \int_{\mathbb{R}^D} |z_1| \exp\left(-\frac{\|z\|_2^2}{2\sigma^2}\right) dz = \frac{1}{\sigma^2} \mathop{\mathbb{E}}_{z \sim \mathcal{N}(0,\sigma^2 I)} |z_1|. \tag{23}$$

since $\|f(x)\|_2 = 1 \; \forall x$ and $\|Q\|_2 = 1$, and $\|z^T e\|_2 = |z_1|$. Here z_1 is the first component of $z^\top e$.

The expectation from Eq. 22 is known to be equal to $\sigma \sqrt{\frac{2}{\pi}}$, and, thus

$$\|J_g \cdot v\|_2 \leq \sqrt{\frac{2}{\pi\sigma^2}} \; \forall v : \|v\|_2 = 1. \tag{24}$$

Taking a supremum over all unit vectors v, we immediately get $L \leq \sqrt{\frac{2}{\pi\sigma^2}}$ what finalizes the proof.

□

Theorem 2. *(Adversarial embedding risk) Given an input image $x \in \mathbb{R}^D$ and the embedding $g : \mathbb{R}^D \to \mathbb{R}^d$ the closest point on to decision boundary in the embedding space (see Figure 2) is located at a distance (defined as adversarial embedding risk):*

$$\gamma = \|\Delta\|_2 = \frac{\|c_2 - g(x)\|_2^2 - \|c_1 - g(x)\|_2^2}{2\|c_2 - c_1\|_2^2}, \tag{25}$$

where $c_1 \in \mathbb{R}^d$ and $c_2 \in \mathbb{R}^d$ are the two closest prototypes. The value of γ is the distance between classifying embedding and the decision boundary between classes represented by c_1 and c_2. Note that this is the minimum l_2-distortion in the embedding space required to change the prediction of g.

Proof. For the convenience, redraw Figure 2 in the Figure 6 with labeled points.

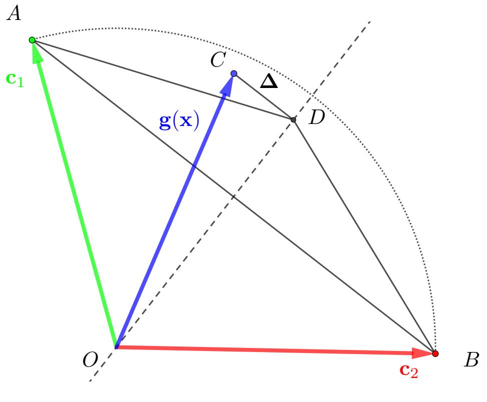


Figure 6: The direction of adversarial risk in the space of embeddings is always parallel to the vector $c_1 - c_2$. This is also true for the case of $\|c_1\|_2 \neq \|c_2\|_2$.

In the Figure 6, O is the origin, $\overrightarrow{OA} = c_1$, $\overrightarrow{OB} = c_2$, $\overrightarrow{OC} = g(x)$, $\overrightarrow{CD} = \Delta$, $\overrightarrow{AB} = c_2 - c_1$. We need to solve the following problem:

$$\min \|\overrightarrow{CD}\|_2 = \|\Delta\|_2 \tag{26}$$

$$\text{s.t. } \|\overrightarrow{BD}\|_2 \le \|\overrightarrow{AD}\|_2 \tag{27}$$

It is obvious that to satisfy minimality requirement we should consider $\|\overrightarrow{BD}\|_2 = \|\overrightarrow{AD}\|_2$. Thus we have $\overrightarrow{AB}$ is perpendicular to the ray $\overrightarrow{OD}$. Therefore, to minimize $\overrightarrow{CD} = \Delta$, we need to find distance from C to the ray $\overrightarrow{OD}$.

The closest distance is perpendicular to $\overrightarrow{OD}$.

$$\left.\begin{array}{c} \overrightarrow{AB} \perp \overrightarrow{OD} \\ \overrightarrow{CD} \perp \overrightarrow{OD} \end{array}\right\} \implies \overrightarrow{CD} \parallel \overrightarrow{AB} \tag{28}$$

$$\implies \Delta = \overrightarrow{CD} = \gamma \frac{\overrightarrow{AB}}{\|\overrightarrow{AB}\|_2} = \gamma \frac{c_2 - c_1}{\|c_2 - c_1\|_2} \tag{29}$$

$$\implies \overrightarrow{OD} = \overrightarrow{OC} + \overrightarrow{CD} = g(x) + \gamma \frac{c_2 - c_1}{\|c_2 - c_1\|_2} \tag{30}$$

$$\implies \left\{\begin{array}{l} \overrightarrow{AD} = \overrightarrow{OD} - \overrightarrow{OA} = g(x) + \gamma \frac{c_2 - c_1}{\|c_2 - c_1\|_2} - c_1 \\ \overrightarrow{BD} = \overrightarrow{OD} - \overrightarrow{OB} = g(x) + \gamma \frac{c_2 - c_1}{\|c_2 - c_1\|_2} - c_2 \end{array}\right. \tag{31}$$

Solving equation $\|\overrightarrow{BD}\|_2 = \|\overrightarrow{AD}\|_2$ implies

$$\|g(x) + \Delta - c_1\|_2^2 = \|g(x) + \Delta - c_2\|_2^2. \tag{32}$$

$$\|g(x) - c_1\|_2^2 + \|\Delta\|_2^2 + 2(g(x) - c_1)^T\Delta = \|g(x) - c_2\|_2^2 + \|\Delta\|_2^2 + 2(g(x) - c_2)^T\Delta \tag{33}$$

$$\|g(x) - c_1\|_2^2 + 2(g(x) - c_1)^T\Delta = \|g(x) - c_2\|_2^2 + 2(g(x) - c_2)^T\Delta \tag{34}$$

$$2(g(x) - c_1)^T\Delta - 2(g(x) - c_2)^T\Delta = \|g(x) - c_2\|_2^2 - \|g(x) - c_1\|_2^2 \tag{35}$$

$$2(c_2 - c_1)^T\Delta = \|g(x) - c_2\|_2^2 - \|g(x) - c_1\|_2^2 \tag{36}$$

Using the fact that $\Delta = \gamma \frac{c_2 - c_1}{\|c_2 - c_1\|_2}$ we find that

$$\gamma = \|\Delta\|_2 = \frac{\|c_2 - g(x)\|_2^2 - \|c_1 - g(x)\|_2^2}{2\|c_2 - c_1\|_2^2}. \tag{37}$$

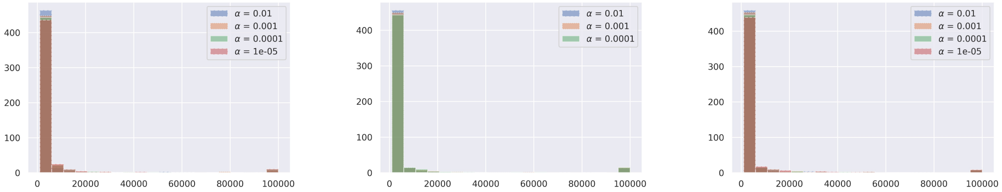
□

B Additional experiments

In this section, we provide the results of additional experiments conducted to ease the assessment of our method.

B.1 The distribution of required sample size

In figure 7 we depict the histograms showing the distribution of the smallest sample size required to reach the confidence level on each dataset. All the experiments were conducted with the following parameters: the variance of additive noise $\sigma = 1.0$, maximum number of samples $n = 100000$. Note that for all the experiments and for all values α of interest, the required number of samples is less than 10000 for most of the input images.



(a) Cub-200-2011 (b) CIFAR-FS (c) *mini*ImageNet

Figure 7: Histogram of the required sample sizes, $n \leq 100000$, $\sigma = 1.0$.

B.2 Empirical robustness and effect of data augmentation

In this section, we assess our approach against random attacks, against the adversarial attack, and evaluate the effect of the data augmentation during the baseline model training.

For each dataset, we have sampled a random subset S of 500 test images and used it as a test set.

The random attacks were conducted on the plain baseline model f. In this experiment, the input x of the model was perturbed by a random noise δ of the fixed magnitude. In this case, empirical robust accuracy is computed as follows:

$$CA(S, \varepsilon) = \frac{|(x, y) \in S : h(f(x)) = h(f(x + \delta)) = y|}{|S|}, \ \|\delta\|_2 = \varepsilon. \tag{38}$$

Here and below, $h(\cdot)$ corresponds to the classification rule from Section 5.3.

For the adversarial attack, the smoothed model g was attacked at input point x with the FGSM attack [10]. Namely, given an approximation of the smoothed model in the form $\hat{g}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x + \varepsilon_i)$, the additive perturbation is computed as follows:

$$\delta(x) = \text{sign}\left(\frac{1}{n}\sum_{i=1}^{n} \nabla_x f(x + \varepsilon_i)\right), \tag{39}$$

so the additive perturbation corresponds to the projection of the mean gradient. In this case, the empirical robust accuracy looks as follows:

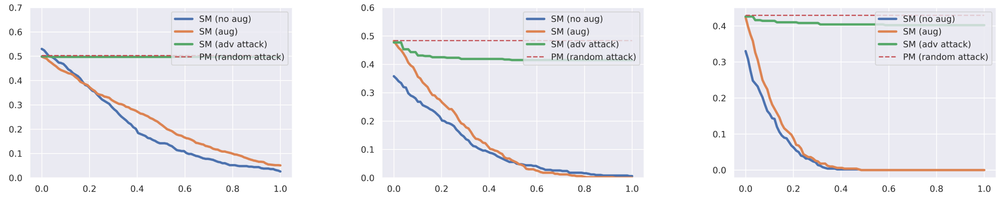
$$CA(S, \varepsilon) = \frac{|(x, y) \in S : h(g(x)) = h(g(x + \delta)) = y|}{|S|}, \ \|\delta\|_2 = \varepsilon. \tag{40}$$

Also, we have evaluated the effect of the data augmentation during the training of the baseline model. For all datasets for both $1-$shot and $5-$shot settings, we have trained new baseline models without data augmentation and evaluated associated smoothed models.

In the figures 8-9, we report results of the experiments. For all the smoothed models, we fixed the confidence level $\alpha = 10^{-4}$ and number of samples $n = 1000$. The variance σ of additive noise for Cub-200-2011 and CIFAR-FS datasets is set $\sigma = 1.0$, for *mini*ImageNet dataset $\sigma = 0.5$.

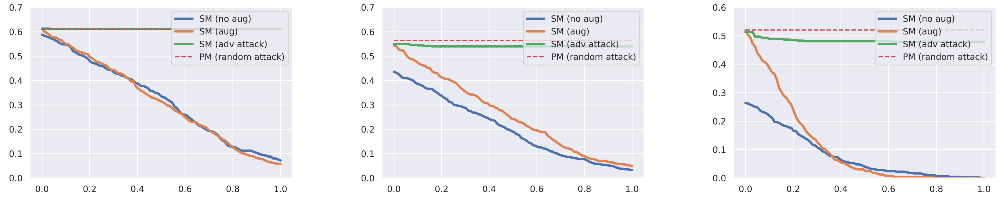
In the figures, *SM (no aug)* corresponds to the certified accuracy of the smoothed model with the baseline model trained without augmentation, *SM (aug)* corresponds to the certified accuracy of the smoothed model with the baseline model trained with augmentation, *PM (random attack)* corresponds to the empirical robust accuracy of plain baseline model under random attack and *SM (adv attack)* corresponds to the empirical robust accuracy of the smoothed model under adversarial attack.

From the pictures, we made several observations. Firstly, the augmentation of the baseline model with an additive noise increases the empirical robustness of the smoothed model, so it is natural to augment the data to have both better robustness and larger accuracy when there is no attack. Secondly, even the baseline models could not be attacked with random additive perturbation. Although the fact



(a) Cub-200-2011 (b) CIFAR-FS (c) *mini*ImageNet

Figure 8: Certified accuracy of smoothed models, empirical robust accuracy of the plain model under random attack and empirical robust accuracy of the smoothed model under adversarial attack, $1-$shot case.



(a) Cub-200-2011 (b) CIFAR-FS (c) *mini*ImageNet

Figure 9: Certified accuracy of smoothed models, empirical robust accuracy of the plain model under random attack and empirical robust accuracy of the smoothed model under adversarial attack, $5-$shot case.

that the probability of randomly generated additive noise being adversarial is low, it is interesting to show that it holds for the prototypical networks too. Finally, we observe that even adversarial attack is not effective against smoothed models.

Note that our theoretical guarantees are provided for the worst-case behavior of the smoothed models, in practice, smoothed encoders may have very strong empirical robustness.

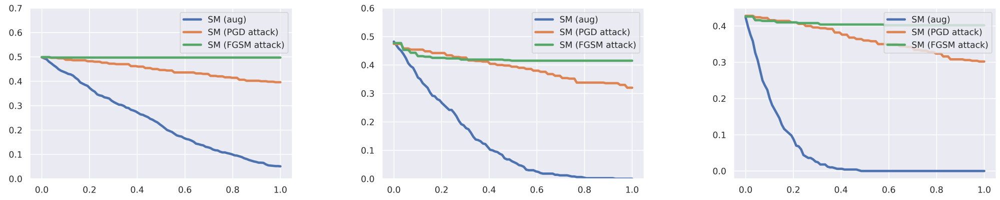
B.2.1 Adversarial attack on the smoothed model: FGSM vs PGD

It is important to mention that a one-step FGSM attack is not the best tool to assess the model's robustness. In this section, we report the results of additional experiments where we compare FGSM and its multi-step version, PGD [31] (projected gradient descent).

In all experiments, we run PGD attack for $s = 20$ iterations. The attack norm on each iteration is s times smaller than the one in the FGSM experiment; it is done to compare methods of attacks on corresponding magnitudes. For all the smoothed models, we fixed the confidence level $\alpha = 10^{-4}$ and the number of samples $n = 1000$. The variance σ of additive noise for Cub-200-2011 and CIFAR-FS datasets is set $\sigma = 1.0$, for *mini*ImageNet dataset $\sigma = 0.5$.

All the models were evaluated in $1-$shot setting. Results are reported in the figure 10.

It is notable that a multi-step attack is more effective against a smoothed model, especially on larger norms of perturbations. Still, the model performs well even against a PGD attack.



(a) Cub-200-2011

(b) CIFAR-FS

(c) *mini*ImageNet

Figure 10: Certified accuracy of smoothed models, empirical robust accuracy of smoothed models under FGSM and PGD attacks, $1-$shot case.