Completing the Picture: Randomized Smoothing Suffers from the Curse of Dimensionality for a Large Family of Distributions

Supplementary Material

Yihan Wu **Aleksandar Bojchevski** **Aleksei Kuvshinov** **Stephan Günnemann**
Technical University of Munich, Germany

1 MISSING PROOFS

Lemma 1. *For an arbitrary hyperspherical sector S in $B(R,x)$,*

$$\frac{|S\cap\partial B(r,x)|}{|\partial B(r,x)|}=\frac{|S\cap\partial B(R,x)|}{|\partial B(R,x)|}=\frac{V_S(R)}{V_B(R)}=\frac{1}{2}I_{sin^2(\phi)}(\frac{d-1}{2},\frac{1}{2}), \tag{1}$$

where $|A|$ is the volume of a set $A\in\mathbb{R}^d$, $\partial B(r,x)$ represents the surface of $B(r,x)$.

Lemma 2. *If $q\sim\mathcal{N}(0,\sigma^2 I_d)$, $\Psi(R;q)=Gamma(\frac{R^2}{2\sigma^2};\frac{d}{2},1)$,*

Proof.

$$\begin{aligned}\Psi(R;q)&=\int_{||z||_2<R}q(z)dz\\&=\int_{||z||_2<R}(2\pi\sigma^2)^{-d/2}e^{-\frac{z^Tz}{2\sigma^2}}dz\\&=\int_0^{R/\sigma}(2\pi)^{-d/2}e^{-\frac{r^2}{2}}dr\\&=\int_0^{R/\sigma}\frac{r^{d-1}}{2^{d/2-1}\Gamma(\frac{d}{2})}e^{-\frac{r^2}{2}}dr\\&=\int_0^{\frac{R^2}{2\sigma^2}}\frac{u^{\frac{d}{2}-1}e^{-u}}{\Gamma(\frac{d}{2})}du\\&=\text{Gamma}(\frac{R^2}{2\sigma^2};\frac{d}{2},1),\end{aligned}$$

□

Lemma 3. *For an arbitrary continuous distribution q and angle ϕ, there exists a sector S^* on the ball $B(R,x)$ with its colatitude angle equal to ϕ, which satisfies*

$$\frac{q(S^*)}{q(B(R,x))}=\frac{V_{S^*}(R)}{V_B(R)}=\frac{1}{2}I_{sin^2(\phi)}(\frac{d-1}{2},\frac{1}{2}).$$

Proof. First of all, a sector $S(a)$ is fixed by its central point $a\in\partial B(R,x)$ and the colatitude angle ϕ. We can check all possible sectors by going though all central points on $\partial B(R,x)$. Consider the integral of $\frac{q(S(a))}{q(B(R,x))}$ with all possible sectors

$$\begin{aligned}&\int_{a\in\partial B(R,x)}\frac{q(S(a))}{q(B(R,x))}da\\=&\frac{1}{q(B(R,x))}\int_{a\in\partial B(R,x)}\int_0^R q(S(a))\cap\partial B(r,x))drda\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{q(B(R,x))}\int_{a\in\partial B(R,x))}\int_0^R\int_{z\in S(a)\cap\partial B(r,x)} q(z)dzdrda\\
=&\frac{1}{q(B(R,x))}\int_0^R\int_{a\in\partial B(R,x)}\int_{z\in S(a)\cap\partial B(r,x)} q(z)dzdadr\\
=&\frac{1}{q(B(R,x))}\int_0^R\int_{z\in\partial B(r,x)}\int_{a\in S(R\frac{z}{||z||_2})\cap\partial B(R,x)} q(z)dadzdr\\
=&\frac{1}{q(B(R,x))}\int_0^R\int_{x\in\partial B(r,x)} |S(R\frac{z}{||z||_2})\cap\partial B(R,x)|q(z)dzdr
\end{aligned}
$$

Notice $|S(R\frac{z}{||z||_2})\cap\partial B(R,x)| = \frac{1}{2}I_{sin^2(\phi)}(\frac{d-1}{2},\frac{1}{2})|\partial B(R,x)|$ is the volume of the hyperspherical sector $S(R\frac{z}{||z||_2})$ on the surface of the ball, which is only depend on the colatitude angle ϕ and unrelated to z and r, therefore we have

$$
\begin{aligned}
&\int_{a\in\partial B(R,x)}\frac{q(S(a))}{q(B(R,x))}da\\
=&\frac{1}{q(B(R,x))}\int_0^R\int_{z\in\partial B(r,x)} |S(R\frac{z}{||z||_2})\cap\partial B(R,x)|q(z)dzdr\\
=&\frac{1}{q(B(R,x))}*\frac{1}{2}I_{sin^2(\phi)}(\frac{d-1}{2},\frac{1}{2})|\partial B(R,x)|\int_0^R\int_{z\in\partial B(r,x)} p(x)dzdr\\
=&\frac{1}{2}I_{sin^2(\phi)}(\frac{d-1}{2},\frac{1}{2})|\partial B(R,x)|
\end{aligned}
$$

Thus the average of $\frac{q(S(a))}{q(B(R,x))}$ on $\partial B(R,x)$ is

$$
\frac{1}{|\partial B(R,x)|}\int_{a\in\partial B(R,x)}\frac{q(S(a))}{q(B(R,x))}da = \frac{1}{2}I_{sin^2(\phi)}(\frac{d-1}{2},\frac{1}{2}),
$$

which indicates there must exist a sector S^* with colatitude angle ϕ such that $\frac{q(S^*)}{q(B(R,x))} = \frac{V_{S^*}(R)}{V_B(R)}$ (because $q(S(a))$ is continuous with a). Notice the volume of the sector is only related to the $\updownarrow_2$ norm of δ. □

1.1 Proof of proposition 2

Proposition 2. *When $R > \Psi^{-1}(g(x);q)$, there exists an perturbation δ, which fixes a hyperspherical sector S_1 in $B(R,x) := \{z|\,||z-x||_2 < R\}$ (Equation 1.1 left) and a classifier $h(x) = c_x 1_{(x\in S_1')}$ such that*

$$
p(\{z|h(x+z) = c_x\}) = q(S_1') = g(x), \tag{2}
$$

and

$$
\frac{q(S_1)}{q(B(R,x))} \le \frac{V_{S_1}(R)}{V_B(R)}. \tag{3}
$$

For brevity we set $q(S) := \mathbb{P}_{z\sim q}(z\in S)$ for a set $S\subset X$.

Proof. Since $R > \Psi^{-1}(g(x);q)$, we have $\Psi(R;q) > g(x)$. First we can select the colatitude angle ϕ such that $\frac{1}{2}I_{sin^2(\phi)}(\frac{d-1}{2},\frac{1}{2}) = 1-\frac{g(x)}{\Psi(R;q)}$, then according to Lemma 3, we can select a sector S_1 with this colatitude angle, which satisfies $\frac{q(S_1)}{q(B(R,x))} = \frac{V_{S_1}(R)}{V_B(R)}$. In this case

$$
1-\frac{g(x)}{\Psi(R;q)} = \frac{1}{2}I_{sin^2(\phi)}(\frac{d-1}{2},\frac{1}{2}) = \frac{V_{S_1}(R)}{V_B(R)} = \frac{q(S_1)}{q(B(R,x))}.
$$

Notice $q(B(R,x)) = \Psi(R;q)$, we have $q(S_1) = \Psi(R;q) - g(x)$ and $g(x) = q(S_1')$. Obviously the sector S_1 satisfies Equation 2 and Equation 3 in preposition 2. Therefore we can select the corresponding perturbation δ and classifier h. □

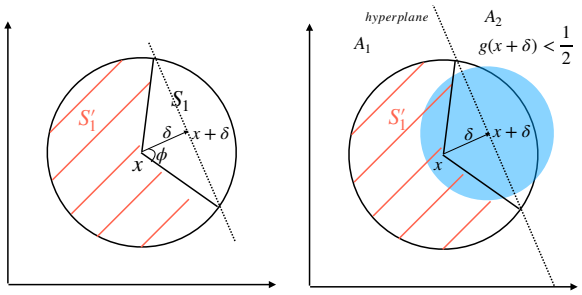


Figure 1: The classifier h and disturbance δ we choose.

1.2 Proof of proposition 4

Proposition 4. *When* $d \in [10^3, 10^7]$ *and* $\epsilon \in [10^{-6}, 1]$,

$$\sqrt{1 - I_\epsilon^{-1}(\frac{d-1}{2}, \frac{1}{2})} < \frac{5}{\sqrt{d}} \tag{4}$$

Proof. $\sqrt{1 - I_\epsilon^{-1}(\frac{d-1}{2}, \frac{1}{2})}$ is too complicate to be analyzed theoretically, luckily we have numerical approaches to evaluate it and compare it with $\frac{5}{\sqrt{d}}$. As $\sqrt{1 - I_\epsilon^{-1}(\frac{d-1}{2}, \frac{1}{2})}$ decreases with ϵ, we only need to show $\sqrt{1 - I_{10^{-6}}^{-1}(\frac{d-1}{2}, \frac{1}{2})} < \frac{5}{\sqrt{d}}$. On Equation 1.2 we plot $log_{10}(\frac{5}{\sqrt{d}} - \sqrt{1 - I_{10^{-6}}^{-1}(\frac{d-1}{2}, \frac{1}{2})})$ with respect to $log_{10}(d)$, and the plot shows that the difference between $\frac{5}{\sqrt{d}}$ and $\sqrt{1 - I_{10^{-6}}^{-1}(\frac{d-1}{2}, \frac{1}{2})}$ is always larger than 0. □

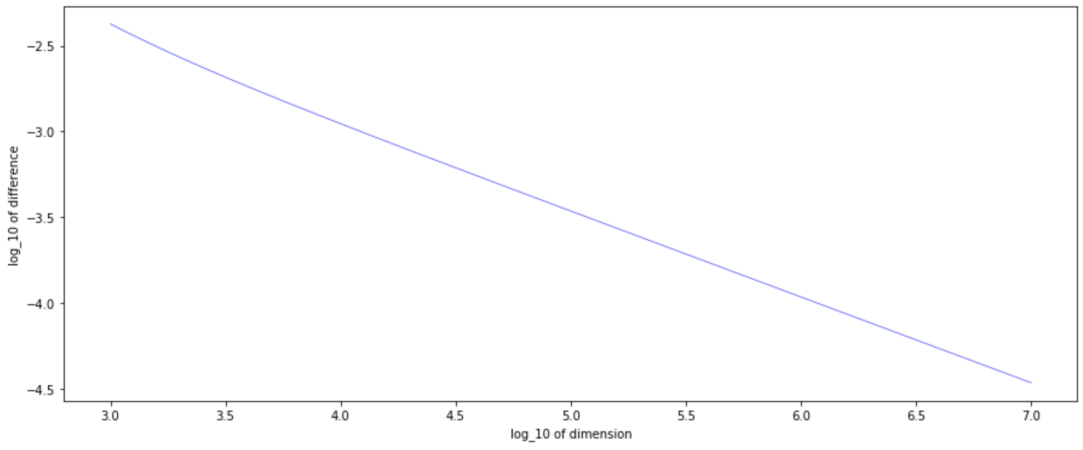


Figure 2: Numerical evaluation of $\frac{5}{\sqrt{d}} - \sqrt{1 - I_{10^{-6}}^{-1}(\frac{d-1}{2}, \frac{1}{2})}$.

1.3 Proof of proposition 5

Proposition 5. *For any distribution $q \in \mathcal{Q} := \{q|q(z) = q(-z)\}$, there exist an perturbation δ which fix a hyperspherical sector S_1 in $B(R_x, x)$ (Figure 1.3 left) and a classifier $h(x) = c_x 1_{(x\in S_1')\vee(x\in S_2')}$ such that*

$$\begin{aligned} &q(S_1') = g_{\leq R_x}(x), \\ &q(S_2') > \frac{1}{2}(1 - \Psi(R_x; q)) \geq g_{>R_x}(x), \\ &q(\{z|h(x+z) = c_x\}) = q(S_1') + q(S_2') \geq g(x), \end{aligned}$$

and

$$\frac{q(S_1)}{q(B(R_x, x))} \leq \frac{V_{S_1}(R)}{V_B(R)}$$

Proof. Since $g_{\leq R_x}(x) := \mathbb{P}_{z\sim q}(\{z|h'(x+z) = c_x\} \cap \{||z||_2 \leq R_x\}) \leq \Psi(R_x; q)$, we can choose the colatitude angle ϕ such that $\frac{1}{2} I_{sin^2(\phi)}(\frac{d-1}{2}, \frac{1}{2}) = 1 - \frac{g_{\leq R_x}(x)}{\Psi(R_x;q)}$. Analogously to the proof of proposition 2, we can select a sector S_1 such that $g(x) = q(S_1')$ and $\frac{q(S_1)}{q(B(R_x,x))} \leq \frac{V_{S_1}(R)}{V_B(R)}$.
Based on the definition of R_x, $g_{>R_x}(x) < \frac{1}{2}(1 - \Psi(R_x; q))$. Notice S_2' in figure 1.3 covers more than half of the space outside $B(x, R_x)$ and q satisfies $q(z) = q(-z)\, \forall z \in R^d$,

$$q(S_2') = \mathbb{P}(x + z \in S_2') > \frac{1}{2}\mathbb{P}(x + z \in \mathbb{R}^d / B(R_x, x)) = \frac{1}{2}(1 - \Psi(R_x; q))$$

Thus $q(S_2') > \frac{1}{2}(1-\Psi(R_x; q)) \geq g_{>R_x}(x)$. Therefore we can select the corresponding perturbation δ and classifier h. □

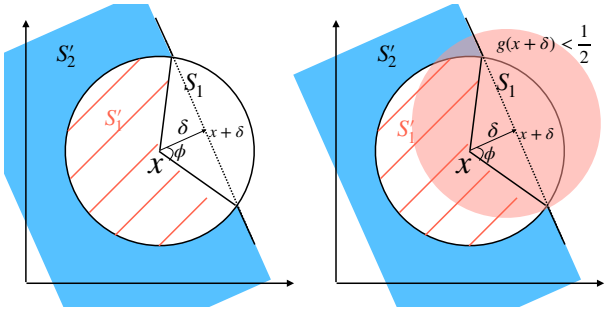
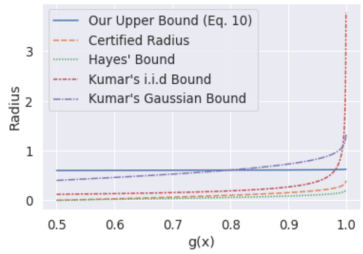


Figure 3: The classifier h and perturbation δ we choose. The classifier h we chose is $1_{(x\in S_1')\vee(x\in S_2')}$
.

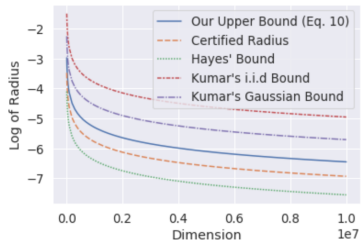
1.4 Proof of corollary 3

Corollary 3. **($\updownarrow_p$ smoothing with spherical symmetric distribution)** *If q is a spherical distribution i.e. $q(z) = q(||z||_2)$, Proposition 2 and Proposition 5 hold for any perturbation δ, we can choose $\delta = (\frac{b}{\sqrt{d}}, \frac{b}{\sqrt{d}}, \frac{b}{\sqrt{d}}, ..., \frac{b}{\sqrt{d}})$ and classifier $h = 1_{x\in S_1'}$ or $h = 1_{x\in(S_1'\cup S_2')}$ such that the $\updownarrow_p$ certified radius*

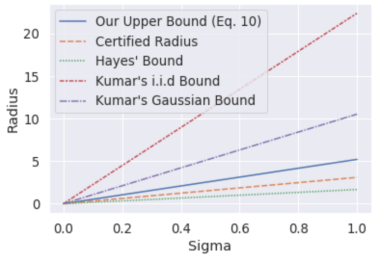
$$\begin{aligned} r_g(x) &< \frac{5}{d^{1-\frac{1}{p}}} \Psi^{-1}(\frac{g(x)}{1 - 5 * 10^{-7}}; q) \\ or \quad r_g(x) &< R_x \frac{5}{d^{1-\frac{1}{p}}} \end{aligned}$$



(a) $\sigma = 1$, $d = 3 \cdot 32 \cdot 32$, $g(x)$ varies



(b) $\sigma = 1/\sqrt{d}$, $g(x) = 0.999$, d varies



(c) $g(x) = 0.999$, $d = 3 \cdot 32 \cdot 32$, σ varies

Figure 4: Our first upper bound (blue), Kumar et al. (2020)'s i.i.d. bound (red),Kumar et al. (2020)'s generalized Gaussian bound (purple), Hayes (2020)'s bound (green), and the certified radius by Cohen et al. (2019) (orange). We choose $g(x) = 0.999$ for the second and the third case because in practice, most $g(x)$ values are close to 1. Our upper bound is better in most cases. Hayes (2020)'s upper bound is below the certified radius since it bounds the divergence-based radius which is not tight.

Proof. We only provide the proof under the condition of Proposition 2, the proof with Proposition 5 is analogously.

Firstly, if q is a spherical distribution, for an arbitrary hyperspherical sector S_1 with colatitude angle ϕ,

$$\begin{aligned}
\frac{q(S_1)}{q(B(R,x))} &= \frac{\int_0^R q(S_1 \cap \partial B(r,x))dr}{\int_0^R q(\partial B(r,x))dr} \\
&= \frac{\int_0^R |S_1 \cap \partial B(r,x)|q(r)dr}{\int_0^R |\partial B(r,x)|q(r)dr} \\
&= \frac{\int_0^R \frac{1}{2} I_{sin^2(\phi)}(\frac{d-1}{2}, \frac{1}{2})|\partial B(r,x)|q(r)dr}{\int_0^R |\partial B(r,x)|q(r)dr} \\
&= \frac{1}{2} I_{sin^2(\phi)}(\frac{d-1}{2}, \frac{1}{2}) = \frac{V_{S_1}(R)}{V_B(R)}.
\end{aligned}$$

Therefore we can select the sector S_1 with the perturbation $\delta = (\frac{b}{\sqrt{d}}, \frac{b}{\sqrt{d}}, \frac{b}{\sqrt{d}}, ..., \frac{b}{\sqrt{d}})$, where $q(S_1') = g(x)$. So

$$r_g < ||\delta||_p = \frac{b}{d^{\frac{1}{2}-\frac{1}{p}}} = \frac{||\delta||_2}{d^{\frac{1}{2}-\frac{1}{p}}}.$$

According to **Corollary 1**, $||\delta||_2$ is upper bounded by $\frac{5}{\sqrt{d}}\Psi^{-1}(\frac{g(x)}{1-5*10^{-7}}; q)$, Hence

$$r_g(x) < \frac{5}{d^{1-\frac{1}{p}}}\Psi^{-1}(\frac{g(x)}{1-5*10^{-7}}; q)$$

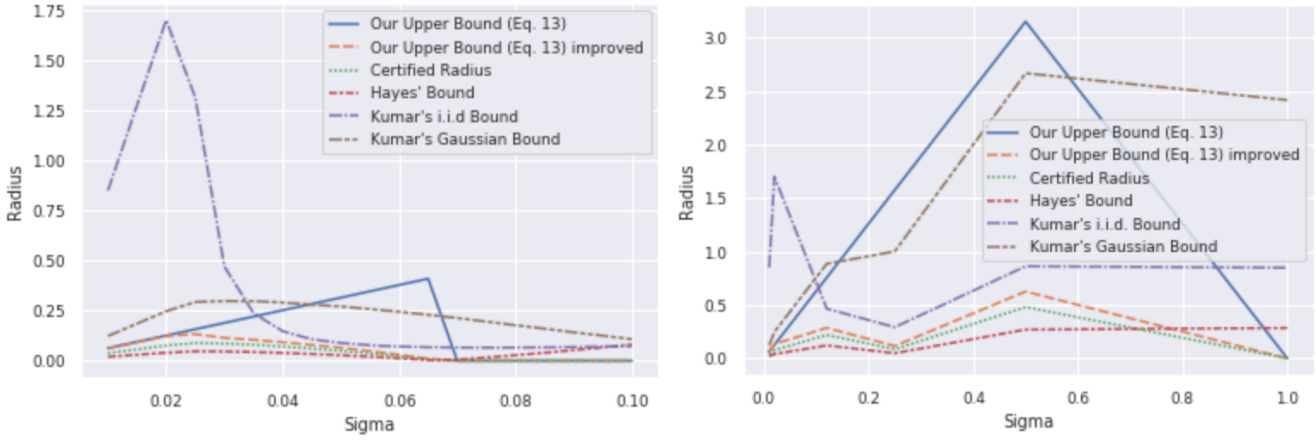
□

2 EXPERIMENTAL SETTINGS

We train the ResNet110 on CIFAR-10 with 50000 training samples and 10000 test samples. We apply SGD optimizer with momentum as 0.9 and the learning rate as 0.1. We also use a learning rate scheduler, which reduce the learning rate to one tenth of its current value every 30 epochs. The total training steps are 90 epochs. For the robust training with noise augmentation, we follow Cohen et al. (2019)'s method, which add a random noise from $\mathcal{N}(0, \sigma^2 I_d)$ to each training sample. The rest settings are the same as normal training.

3 ADDITIONAL EXPERIMENTS

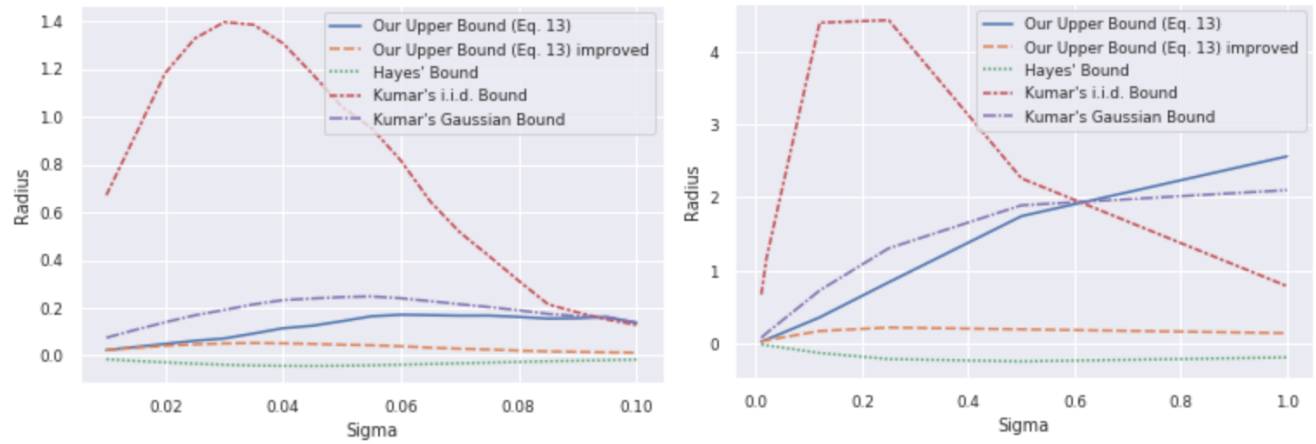
In additional experiments we will repeat the three experiments Figure 4, Figure 5 and Figure 6 in our work with Kumar et al. (2020)'s i.i.d. bound and generalized Gaussian bound separately.



(a) ResNet110 without robust training on CIFAR10

(b) ResNet110 with robust training on CIFAR10

Figure 5: Comparison of different upper bounds and the certified radius for a random test sample from CIFAR10 with respect to σ. In (a) the original classifier is a ResNet110 without robust training and σ is from a linear search space of $[0.01, 0.1]$. In (b) the original classifier is a ResNet110 with robust training and $\sigma \in \{0.01, 0.02, 0.12, 0.25, 0.5, 1.0\}$.



(a) ResNet110 without robust training on CIFAR10

(b) ResNet110 with robust training on CIFAR10

Figure 6: Average of difference between upper bounds and certified radius over 100 random samples from CIFAR10 with respect to σ. In (a) the original classifier is a ResNet110 without robust training and σ is from a linear search space of $[0.01, 0.1]$. In (b) the original classifier is a ResNet110 with robust training and $\sigma \in \{0.01, 0.02, 0.12, 0.25, 0.5, 1.0\}$.

References

J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. *ICML*, 2019.

J. Hayes. Extensions and limitations of randomized smoothing for robustness guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 786–787, 2020.

A. Kumar, A. Levine, T. Goldstein, and S. Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. *ICML*, 2020.