We thank reviewers for their thoughtful feedback. We are pleased to see that most reviewers found the work to be interesting and innovative. Here, we provide clarifications and conduct additional experiments to demonstrate the improvement in performance achieved by our proposed technique for certification under the ℓ_∞ norm threat model.

Reviewer 1 : 1) Significance of results. Although we only give empirical evidence for ℓ_1, ℓ_2 norm and subspace ℓ_2 norm in the manuscript, the theoretical guarantees in fact extend to ℓ_∞ and subspace ℓ_1, ℓ_∞ norms. We only focus on ℓ_1, ℓ_2 and ℓ_∞ as these are the most intensely researched / relevant threat models in the field. In response to Reviewer 3's comments, we also provide empirical evidence to show improvement for ℓ_∞ norm on the CIFAR10 dataset. As the current state-of-the-art for ℓ_∞ norm is given under Gaussian smoothing [1], our empirical result can give the new state-of-the-art for ℓ_∞ norm certification. **2) Cohen ℓ_1 radius calculation :** The ℓ_1, ℓ_∞ norm results were not explicitly stated in the original paper [2] but they can be derived by following the same analysis, which are also stated in the recent paper [1, Appendix Table A] . **3) "Higher-order" in title :** The "higher-order" in the title is in reference to the fact that the paper lays down the ground work for using higher-order information for certification. However, we see that it might be ambiguous as we only fully explore first-order smoothing. So, we plan to change the title to "first-order smoothing".

Reviewer 2 : Limited empirical evidence : We note that our experiments provide numerical evidence of our theoretical results and demonstrate that the certification performance can be greatly improved by incorporating higher-order information. We have followed standard experiment setup and conducted various experiments on CIFAR10 (Sec 5) and ImageNet (Appendix E) and compared all the current baselines the ℓ_1, ℓ_2 norm and subspace ℓ_2 norm, which is in line with other works in this field. Notably, we conduct additional CIFAR10 experiments for ℓ_∞ certification in Figure R1.

Reviewer 3 : Experiments for ℓ_∞ : In the current paper, we have only focused on giving the bounds for $\ell_1, \ell_2, \ell_\infty$ norms. For general ℓ_p norm we can use the current results to provide lower bounds on the certified radii. As for empirical results for certifying the ℓ_∞ norm radius, it requires the estimation of $\left\|y^{(1)}\right\|_1$. As mentioned in line 270 in the paper the current estimators used to calculate $\left\|y^{(1)}\right\|_1$ need a lot of samples in order to find non-vacuous high-confidence bounds. Although the certification cost is higher, **using the proposed method gives us significant ($\sim 10\%$) improvement over the bounds given by Cohen et al. for CIFAR10.** The CIFAR10 ℓ_∞ results in Figure R1 are calculated using 4M samples for certification (6 minutes/image). One of the major limitations of the current estimators is that they are biased. In the paper we have proposed a new unbiased estimator for $\left\|y^{(1)}\right\|_2$ (Table 1 in the current paper). We think a similar new estimator is needed to make ℓ_∞ certification more scalable. This is left for future work.

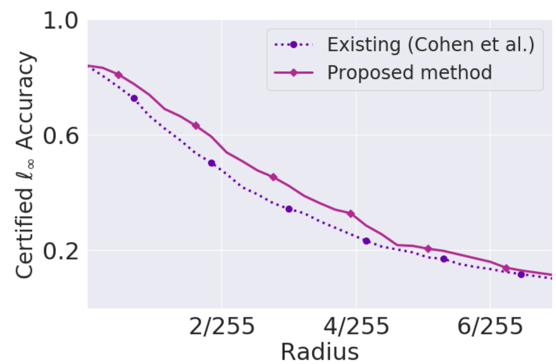

Figure R1: Comparing certified accuracy for CIFAR10 under ℓ_∞ threat models. Our results show that around 10% improvement can be obtained by using the proposed method.

Reviewer 4 : 1) Typos and clarifications : ***i)*** Sorry it is a major typo. The statement should read *"Under the proposed general framework for calculating certified radii, it is easy to see that adding more local constraints (H_i^x) in Equation (2) gives a ~~smaller~~ bigger value of $\boldsymbol{p}_x(z)$ for any x, z which makes the super-level set of $\boldsymbol{p}_x$, equivalently the certified safety region, bigger."* ***ii)*** With a slight abuse of notation, we use μ to denote both the measure and the probability density function of the measure. Here, $D_x^\alpha \mu(y-x)$ corresponds to taking the multivariate differential of the probability density function (μ) at $(y-x)$ with respect to variable x. ***iii)*** In Figure 1 of our manuscript, the direction of the gradient $y^{(1)}$ is along the negative x-axis. We plan to add an arrow to clarify this. Also in order to better motivate the idea we plan to add numerical values for the directional certified radii on the figure. ***iv)*** Given the images in the pixel space, we do change of basis to orient the basis along the gradient $y^{(1)}$ to simplify calculations. In line 515-518, $z_1, z_2, \ldots, z_d$ denotes the variables corresponding to the new basis vectors for the space after the transformation. In corollary 1 we abuse the notation and use z_1, z_2 to denote the variables involved in the system of equations we reduce our initial constraints to. The two sets of variables are not linked. We will change the variable names to avoid confusion in the future and also give a description of z_i's in the proof before using them. **2) Certified bounds vs Attack bounds :** We do agree that the experimental evidence would be great. However we are aware of attacks on randomized smoothing classifiers only for the ℓ_2 norm threat model currently. For this scenario the current bounds are near optimal as the attacks are close to the current state-of-the-art certification bounds [3] (equivalently our proposed certification bounds).

References

[1] G. Yang, T. Duan, E. Hu, H. Salman, I. Razenshteyn, and J. Li, "Randomized smoothing of all shapes and sizes," *ICML*, 2020.

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

[3] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, "Provably robust deep learning via adversarially trained smoothed classifiers," *Neural Information Processing Systems*, 2019.