We appreciate the valuable comments from all reviewers. We first respond to common issues brought by the reviewers, and then respond to individual comments.

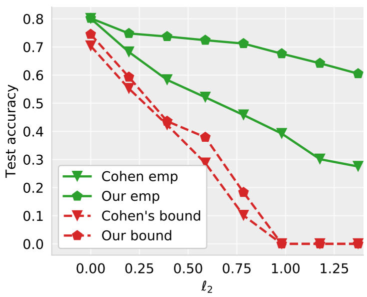

Figure 1: Bounds & empirical robust accuracy comparisons on CIFAR10 with ours and Cohen et al.

Comparison to Cohen et al, [2019]: Both Reviewers 2 and 3 asked for a comparison to Cohen et al. We did not include such a comparison because their work is actually a follow-up of our work. As a work released earlier (posted on arXiv), we thought it was not necessary to include all follow-up works in comparisons. Although they do indeed improve our bound by a small margin, our work has novel contributions in several ways: 1) We propose stability training to improve the bound and robustness, while they only use Gaussian augmentation. As pointed out by Reviewer 3, stability training works better than Gaussian augmentation. The improvement from stability training is more significant than the improvement of the bound. As a result, our bound plus stability training could yield a higher certified bound and empirical robustness accuracy than Cohen et al. We show the improvement of the bound and robust accuracy on CIFAR10 in Figure 1. The gap in robust accuracy is more obvious as their tighter bound does not help improve robustness against real attacks. Thus, stability training is an important and unique contribution to the literature. 2) We conduct empirical evaluation against real attacks and compare to the state-of-the-art defense method to show our approach is competitive. Cohen et al. only discusses the certified bound and does not provide evaluation against real attacks. 3) The analysis from Cohen et al. is difficult to be extended to other norms, while ours can be easily achieved. Both Reviewer 1 and 2 asked about extension to other norms and we now explain. For example, noticing the Renyi divergence between two Laplacian distribution $\Lambda(x,\lambda)$ and $\Lambda(x',\lambda)$ is $\frac{\|x-x'\|_1}{\alpha-1}\log\left(\frac{\alpha}{2\alpha-1}e^{\frac{\alpha-1}{\lambda}}+\frac{\alpha-1}{2\alpha-1}e^{-\frac{\alpha}{\lambda}}\right)$, one can obtain a certified ℓ_1 bound by adding Laplacian noise. The derivation is similar to the proof of Theorem 1 by replacing Gaussian with Laplacian. In general, our framework extends to any norm as long as we find a corresponding distribution whose Renyi divergence of two is a function of the distance of their means. Thus, our analysis provides better flexibility. We will discuss the differences more thoroughly in the revision.

Caption of Figure 1: We apologize for the confusion in the caption. We will revise it so the colours match the plots.

Reviewer 1: 1) The fundamental difference from TRADES is our approach provides certified robustness, *i.e.*, our model is theoretically guaranteed to be robust as long as the norm of the perturbation is smaller than the bound. Although TRADES is empirically robust, no theoretical guarantee is provided. The fact that our approach is empirically competitive to TRADES should be considered as a great bonus. 2) The certified bound from Gowal et al. is for ℓ_∞ norm which is not comparable to ℓ_2 norm. The state-of-the-art ℓ_2 certified robustness is achieved at Wong et al. [2018] in Table 4, to which we compare in our paper. 3) Our approach is much cheaper than TRADES. Training TRADES on Wide-Resnet for CIFAR10 takes more than 5 days on a T4 GPU; whereas our approach does not require constructing adversarial examples and only takes 8 hours to train in the same setting. Thus our method is about 15 times cheaper than TRADES. In fact, due to the computational constraint, TRADES is not scalable to ImageNet, while our approach is, as shown in Appendix D. 4) We argue the comment that we care more about the smaller perturbations than the large ones is not precise. The goal of adversaries is to reduce the classification as much as the perturbation is not perceptible. For example, if any perturbation smaller than 2.0 is not perceptible, the adversaries should always use 2.0 instead of any number smaller, because larger perturbations strictly reduce more accuracy. In this sense, larger perturbations are more important than smaller ones. 5) We use two commonly used gradient-free attacks (transfer attack and boundary attack) to evaluate the robustness of our method. We believe it is sufficient to show there is no gradient masking. We will add the results for PGD-1000 with 20 random initializations in the revised version. 6) In our evaluation, we use 20 steps PGD with step size $\alpha = \epsilon/10$. These parameters can be found in our code. We will make it more clear in the revised version.

Reviewer 2: 1) Our bound is strictly tighter than the one from Lecuyer et al. which is shown in Appendix A with simulation. We also evaluate PixelDP and our bound without STN on MNIST and CIFAR10 to show the empirical gap in Figure 1 in our paper (orange and green). 2) It is exactly because the two probabilities are not correlated that we can multiply them together. The probability that two independent events (bound is correctly estimated and the example is correctly classified) happen simultaneously is the product of their individual probabilities. 3) Yes, the radius is chosen to maintain consistency with Wong et al. 5) We will include results for ImageNet in the main text in the revised version.

Reviwer 3: 1) In Table 1, we are comparing to the best results (bold numbers) reported in Wong et al [2018]. For the single model on CIFAR10, the robust accuracy should be 52% instead 53% as they reported the robust error being 48%. Similarly, the natural accuracy for Cascade model on CIFAR10 should be 58.8% instead of 68.8%. We apologize for the mistakes, but in both cases the actual numbers are worse than what is reported. 2) The orange curve corresponds to a vanilla trained model attacked by PGD. We will make it more clear and fix the typos in the revised version.