We thank the reviewers for their insightful and constructive feedback. We start with addressing two common issues (**C#1**, **C#2**), and then the comments from the individual reviewers (**R1**, **R2**, **R3**, **R4**).

C#1: About the connection with [A]. We will definitely draw more comprehensive discussions on [A] as suggested by **R3**. We believe the two works have different contributions and were developed concurrently. [A] derives the optimal shapes of level sets for ℓ_p attacks based on the Wulff Crystal theory, while our work, based on our functional-optimization framework and accuracy-robustness decomposition (Eq.9), proposes to use distribution that is more concentrated toward the center. Besides, we also consider a novel distribution using mixed ℓ_2 and ℓ_∞ norm for ℓ_∞ adversary, which hasn't been studied before and improve the empirical results.

C#2: About the significance of empirical results. As we simply use the pre-trained model from open source of the baselines and thus our algorithm is almost deterministic (as the variance of Monte Carlo is very small given that we use 100K samples). Our result are statistically significant, for example for the 49% vs. 50% accuracy for ImageNet on l_2 adversarial, two-proportions z-test gives p value 0.001565.

R1 On ℓ_∞ results. We achieve better or comparable performance as IBP. We will add the comparison in revision. Previously, we do not list IBP results because it relies on a special model structure while our method can be applied to any model structure. [31] uses trades to train better model, while we focus on better certification accuracy with a given model. Applying our method to [31]'s checkpoint (image size = 32, ϵ = 0.435, β=6, standard deviation = 0.12), we achieve 64.7% accuracy at 2/255 radius outperforming 62.6% in [31]. We will add this baseline in the next version.

R1 On the optimality of Gaussian. Yes, [30-32] gives the (rate) optimality of Gaussian for ℓ_p adversarial. But this theoretical results only analyzes the rate of the error while ignoring the constant. However, in practice, the constant can be important or even dominating (e.g. many non-convex optimization algorithms have same rate while can be quite different in practice due to the difference in constant factor).

R1 On the connection between the dual form and proposed distribution. The proposed distribution family is the one that gives empirical improvement. The theoretical framework serves as the motivation (decomposition of the bound in sec 3.3) for proposing the new distribution family. It is not valid to do ablation study on the theoretical framework (i.e. when we choose previous smoothing distribution, i.e., Gaussian in [C], our theoretical framework gives same result, see Corollary 1,2).

R2 [B] and our work are two different ways for generalizing the original random smoothing bounds: [B] relaxes the constrain of smoothing samples to a ball in f-divergence while ours relaxes the classifier to be in a function space. We will draw an in-depth discussion in the revision. Regarding the choice of k, we refer readers to Appendix B.3 for details. We will move the reject sampling details to main text.

R3 On the 'weakness'. Thanks for the information on [A]. Regarding work [A], please see C#1. For the questions on ℓ_2 adversarial, our improvement does not come from the optimization as we simply use the pretrain model from open source repository of the baselines. We suspect the empirical different comes from the different choice of hyper-parameters of the distribution. We are not sure what you mean by saying 'analytic form gives different findings'. Could you explain it with more detail so that we can address it in the next version? For the question on ℓ_1 adversarial, yes, [A] achieves better empirical results. But our results support our theoretical findings and is of independent values to the community. We will definitely add comprehensive discussion on [A] in the next version.

R3 On the minor remark about 'correctness'. Thanks for pointing this out. We agree using $\sup/\inf$ is more precise and will change to it. And yes, $\mathcal{F}_{[0,1]}$ isn't compact. We will fix this. We will also add proof for theorem 1 (III).

R4 On the trade-off. In line 143-147, we explain our main insights gained from Eq.9, from where we carefully design the proposed distribution to achieve a better trade-off (see Fig.1). The λ comes from the dual derivation, and its optimal value only depends on the specific value of the accuracy term and the norm ball in Eq.9. Even though the optimal λ is generally hard to derive analytically, the two components in Eq.9 clearly show the trade-off between accuracy and robustness. Similarly, performing rigorous scaling analysis on ℓ_∞ is rather challenging technically. We leave this analysis for future work.

R4 On empirical result. Please see C#2 on the statistical significance. The tightness of our bound is proved in theorem 1 (III). Empirically, the exact certified accuracy relies on $f^{\#}_{\pi_0}$, which depends on whether a good smoothing distribution that fits the data and base classifier are selected (which is also the main gain of this paper).

R4 On adversarial training. Thanks for pointing out this. We agree it's doable while it would require an unaffordable amount of computational resource for adv. training on ImageNet. Thus we leave this for future work.

Reference [A] Randomized Smoothing of All Shapes and Sizes. [B] A Framework for Robustness Certification of Smoothed Classifiers Using f-Divergences. [C] Certified Adversarial Robustness via Randomized Smoothing.