We sincerely thank all the reviewers for their thoughtful comments, efforts, and time. We are delighted that every reviewers has left a positive impression on our research, particularly appreciating the simplicity and effectiveness of our method, with thorough and strong experimental results. We respond to each comment one-by-one in what follows.

(R1) Novelty. We use the framework in [4] since it is a natural way to extend the previous natural-robust error decomposition [5] for smoothed classifiers. We believe the novelty of our method is in its simplicity, and in how to come up with this simplest form from the common principle of accuracy-robustness trade-off under the framework.

(R2) Variance over multiple runs. We observe ACR of a training method is fairly robust to network initialization, e.g., as given in the right table: each value reports the mean and standard deviation across 5 seeds. As another support, we point out all the baselines considered in this paper [1, 3, 2, 4] also report single-run results in their papers, possibly based on observations similar to ours. Finally, we plan to publicly release our code and models for better reproducibility.

ACR (MNIST)	$\sigma = 0.25$	$\sigma = 0.50$	$\sigma = 1.00$
Gaussian [1]	0.9108±0.0003	1.5581±0.0016	1.6184±0.0021
+ Consistency	**0.9300**±0.0004	**1.6653**±0.0007	**1.7486**±0.0025
SmoothAdv [3]	0.9315±0.0001	1.6830±0.0006	1.7706±0.0019
+ Consistency	**0.9323**±0.0004	**1.6905**±0.0002	**1.8087**±0.0022
Stability [2]	0.9152±0.0007	1.5719±0.0028	1.6341±0.0018
MACER [4]	0.9201±0.0006	1.5899±0.0069	1.5950±0.0051

(R2) Clean accuracy is often worse than [4]. Our method can effectively explore the accuracy-robustness trade-off [5] with λ. We expect the demands for higher clean accuracy could be compensated by using lower λ, e.g., our method achieves better clean accuracy than MACER in Table 1 when $\lambda = 10$ on $\sigma = 0.50$, even with better ACR. Nevertheless, it is remarkable that MACER sometimes achieve higher clean accuracy even than Gaussian, e.g., at $\sigma = 0.25$ of Table 1, and we agree with R2 that improving clean accuracy of smoothed classifier is also an important future direction.

(R2) Logit margin vs. input margin. Regarding Figure 1, it is important to notice that our focus is NOT the robustness of the base classifiers f tested, but the robustness of its *smoothed* counterparts $\hat{f}$. A key benefit of smoothed classifier is that it elegantly transforms a matter of *input* margin on f into that of *output* margin on $\hat{f}$: the robustness guarantee of Cohen et al. [1] in Eq. 3 implies that one is enough to minimize $\mathbb{P}_\delta(f(x+\delta) \neq y)$ to improve the robustness of $\hat{f}$ at x, which corresponds to the shaded areas in Figure 1. This can be also viewed in terms of the Lipschitzness: Salman et al. [3] have shown that any Gaussian-smoothed classifier $\hat{f}$ has an explicit Lipschitz constant, leading to a simpler proof of Eq. 3 [1]. We will incorporate the respective discussion in the final draft.

(R2) "Sufficient condition"? The last $\mathbb{E}$ in Eq. 6? Why $m > 1$? (*i*) The condition we refer is at L107: "$F(x+\delta)$ returns a constant output over δ". This directly implies "Eq. 6 $\to 0$", and consequently "the robust error of Eq. 5 $\to 0$", which is why we refer this as *sufficient* condition. (*ii*) Regarding Eq. 6, we remark the outer $\mathbb{E}$ is not over δ, but over $(x, y) \sim \mathcal{D}$ (as given in Eq. 5), thereby the last $\mathbb{E}$ of Eq. 6 cannot be discarded. (*iii*) Finally, our regularization requires $m > 1$ to work, as the term would vanish if $m = 1$: with only a single sample, say δ_1, $F(x+\delta_1)$ would be the best estimation of $\hat{F}(x)$ in Eq. 7, and $L^{\texttt{con}} = 0$ in this case. We will make all these points more clearer in the final draft.

(R3) Eq. 6 $\to$ Eq. 7? For a fixed x, the cross-entropy loss in Eq. 7 is a natural surrogate loss of the 0-1 risk $\mathbb{E}_\delta[\mathbf{1}_{f(x+\delta)\neq\hat{f}(x)}] = \mathbb{P}_\delta(f(x+\delta) \neq \hat{f}(x))$, and this 0-1 risk minimizes the last upper bound in Eq. 6 when minimized across $(x, y) \sim \mathcal{D}$. We also remark that this surrogate loss is *calibrated* [6, 5], i.e., minimizers of Eq. 7 are also minimizers of the 0-1 risk (as mentioned in L104-110). We will clarify this point in the final draft.

(R3) Suggestions for better clarity. (*i*) We use $\mathbf{1}_A$ to denote the *indicator* random variable, formally defined by $\mathbf{1}_A(\omega) = 1$ if $\omega \in A$, and 0 otherwise. We will specify this in the final draft. (*ii*) As suggested by R3, we will update the legends in Figure 1 to better indicate our method. (*iii*) Also, we thank R3 for a detailed assessment of the Broader Impact statements. The final draft will include more discussions regarding the points made by R3.

(R3) SmoothAdv + Consistency on ImageNet? We conduct ImageNet experiments primarily on Gaussian training, and report SmoothAdv results for a comparison. Conducting the suggested experiments with SmoothAdv would be interesting, but currently we found it incurs too much costs to execute during the rebuttal period, e.g., ∼600 GPU hours per single run. Nevertheless, we are willing to incorporate them in the final draft for thoroughness of our experiments.

(R4) Other architectures. In our experiments, all the architectures per dataset is exactly from the prior works [1, 3, 4] for a fair comparison. Nevertheless, we agree with R4 that the effect of architectures on smoothed classifiers is an important question to explore, and we will include more results on other architectures, e.g., DenseNet, in the final draft.

[1] J. Cohen et al. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019.
[2] B. Li et al. Certified adversarial robustness with additive noise. In *NeurIPS*, 2019.
[3] H. Salman et al. Provably robust deep learning via adversarially trained smoothed classifiers. In *NeurIPS*, 2019.
[4] R. Zhai et al. MACER: Attack-free and scalable robust training via maximizing certified radius. In *ICLR*, 2020.
[5] H. Zhang et al. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.
[6] B. Ávila Pires and C. Szepesvári. Multiclass classification calibration functions, 2016.