We thank the reviewers for their comments. Before we address individual concerns, we make some general comments.

1) We agree that our conceptual contribution can be viewed as "taking the right perspective on the problem". In hindsight this perspective was clearly the right thing to do. However we would like to point out that both previous papers (Lecuyer et al. and Cohen et al.) made the same mistake, indicating that this "right perspective" was perhaps not obvious a priori (and indeed, as we explain in Section 2.2, the previously used objective (4) also has a natural interpretation, albeit not the correct one for the problem at hand). In fact, driving this point home, Reviewer 1 cites yet another work (Athalye et al.) which uses the suboptimal objective (4) instead of our proposed objective (S).

2) We would like to emphasize a key point from reviewer 3: "the authors had to address a number of technical challenge [. . .] which they did in a very careful, systematic and clear way". Indeed, writing down the correct objective is only the first step in creating an efficient and state of the art deep learning system. Our double digit improvements over the previous state of the art numbers from Cohen et al. speak for themselves: this can only happen by getting all parts of the system right (in the present case this includes for example using a *biased* estimator of the gradient, see (6) and the discussion after).

3) Finally several reviewers ask about our theoretical insights. As reviewer 4 points out, the Stein's lemma observation sounds promising, and we leave it as an open avenue for future works. We also obtained another new theoretical result, which we left out from the submission as we wanted to focus more on the practical implications of our work (we we believe is the SOTA certified accuracy). This new insight is an alternative, and much simpler, proof of Cohen et al.'s key theorem, which relies on rephrasing Cohen et al.'s statement as a nonlinear Lipschitz bound on the smoothed classifier. This also opens up another avenue for future work, namely by finding better nonlinear Lipschitz guarantees. We would be happy to include this derivation if the reviewers think it would improve the paper.

Reviewer 1: We will change the scale of accuracy plotted in Figure 3, to better illustrate the difference between the empirical accuracy against SmoothAdv and the vanilla PGD.

As mentioned above, the Expectation over Transformation attack of Athalye et al. has the opposite order of log and expectation compared to our SmoothAdv object (Eqn. (S)), and identifying this correct order is a key contribution of our paper (see Sec 2.2). We will add this citation and a discussion of this interesting connection.

Reviewer 2: *Possible confusion concerning the abstention rate.* The reported certifiably robust accuracies in our paper follows the formula $\frac{\text{certified robust}}{\text{total samples}} = \frac{\text{certified robust}}{\text{certified robust+not certified robust+abstained}}$, where abstained samples are counted toward the denominator. Thus, high abstention rate leads to lower certified accuracies. Given that our certified accuracies are higher than Cohen et al.'s (who reports abstention rate of 1% with the parameters described below; see their appendix D2), our abstention rate is not likely to be lower, and it should not be the major component of our improvement over their results. Nevertheless, we will add this experiment to the final version of the paper.

SmoothAdv vs PGD Training. We emphasize that SmoothAdv is first and foremost an objective function that we argue is the correct one for attacking smoothed classifiers (see Eqn.(S)). It is thus somewhat orthogonal to "PGD training" as the objective SmoothAdv can be empirically optimized by PGD, DDN, or other attacks (see Pseudocode 1), and it can be used in both adversarial attack and adversarial training. If the reviewer is referring to adversarial training using SmoothAdv objective (with PGD optimizer), this is different from PGD training, as it carefully combines Gaussian data augmentation and adversarial training (Pseudocode 1).

Our main contribution is not theoretical and we will update the draft to make this more clear.

We will summarize the experiments referred to in Lines 246-247.

Reviewer 3: We thank Reviewer 3 for recognizing our careful experimental work and our technical contributions to the robust training of smoothed classifiers. We will include the standard accuracy in Table 1. The "representative" models in Figures 1 and 2 were picked at random from the set of models we train.

Reviewer 4: *Suggested Improvement 1, new theoretical perspective.* As mentioned in general comment (3) above, we would like to include an alternate proof of the tight certified bound of smoothed classifiers (Theorem 1 in Cohen et al.) purely by bounding the Lipschitz constant of $\Phi^{-1}(p_c)$, where p_c is the probability of predicting class c by the smoothed classifier and Φ^{-1} is the inverse of the standard Gaussian CDF. This greatly simplifies the proof and provide new intuitions on how or why smoothing tends to improve the robustness under L2 perturbations: it reduces the Lipschitz constant of the classifier.

Suggested Improvement 2, improved results that leverage theoretical contributions. It is promising to design stronger attacks to smoothed classifiers based on our alternative derivation of SmoothAdv in Appendix B. This is on-going work.