General responses: We thank all the reviewers for helpful insights, comments, and suggestions.

- (R1 & R2) Our ℓ_0 certificate and baseline: providing robustness certificates in discrete spaces is highly non-trivial. Brute-force solution would require checking $\binom{d}{r}(K+1)^r$ prediction values. To our best knowledge, this work is the first one to establish an actionable ℓ_0 robustness certificate besides adapting certificates in other norms to ℓ_0 norm. In addition, our certificate is *tight* and *applicable* to *any* measurable classifiers, and we theoretically and empirically demonstrate its feasibility to large problems. Under this circumstance, we believe the most appropriate (strongest) baseline is the adaptation of the state-of-the-art certificate in other norms (ℓ_2 via the Gaussian perturbations).
- (R1 & R3) Extra assumptions / decision trees: the significance is in the elaboration that adding an extra assumption *always* leads to a *stronger* certificate (§4.3-4), which was not observed in prior work. An extra assumption is not difficult to find in deep networks, e.g., overall L-Lipschitz continuity or a matrix characterization per layer. The challenge is how to convert it to an actionable certificate, where we hope our finding would enable and solicit further investigations. The decision tree example also has moderate significance due to its unique role in interpretability.
- (R1 & R2) The warmup example: it is only used for warming up how to use the theoretical foundation (§3.1-2). Despite the tightness, the distribution is restrictive (supp$(\phi) \neq \mathcal{X}$), and thus the certificate is limited even if $p = 1$.

Review 1: We thank the reviewer for helpful comments, but hope the reviewer reconsiders the significance of the paper.

- (Contributions-1) The construction of the smoothing distribution: similarly to Cohen et al., our contribution is not in the smoothing distribution but rather in deriving the certificate and associated algorithms.
- (Contributions-1) Suitable comparison: besides the general clarification, here we use Bafna et al. as an example to clarify. They consider a *specific setting* where the clean input $\boldsymbol{x}$ is approximately k-sparse in the Fourier domain. Under this assumption, they used compressed sensing technique to get $\boldsymbol{x}'$ from some corrupted input $\boldsymbol{x} + \boldsymbol{\delta}$ such that $\|\boldsymbol{x} - \boldsymbol{x}'\|_2$ is small. Note that this is not a *robustness guarantee*, which further requires $f(\boldsymbol{x}) = f(\boldsymbol{x}')$. Empirically, they only *show* successful defense against *some* adversarial attacks w.r.t. the base classifier (instead of w.r.t. the recovery algorithm + base classifier), while we *guarantee* the robustness against *all* adversaries up to some radius.
- (Contributions-1) Speed of certification: for a perturbation ϕ on $\mathcal{X}$, it only requires *pre-computing* $\rho_r^{-1}(0.5)$ *once* for each r, and then the certification for any base classifiers f and data $\boldsymbol{x} \in \mathcal{X}$ is done by checking if $p > \rho_r^{-1}(0.5)$.
- Contributions-3: generalization beyond Cohen et al. in uncountable $\mathcal{X}$ and randomized predictions is indeed not the main contribution of our paper. (cf. general response #1 & 2)
- Accuracy: both Table 1 (MNIST) and Figure 4 (Bace) show the certified results for $\alpha = 0.8$ (tuned by validation). The certified accuracy for different α and radii are available for ImageNet in Figure 3c. The raw accuracy / AUC can be found in Figure 3c (ImageNet) and Figure 4 (Bace) when the x-axis is 0. The raw accuracy for MNIST is 0.983.
- Tuning α: for MNIST, ImageNet, and Bace, we consider α values in $\{0.72, 0.76, \ldots, 0.96\}$, $\{0.1, 0.2, \ldots, 0.5\}$, and $\{0.75, 0.8, \ldots, 0.9\}$, respectively. The information is available in Appendix C.1-3.
- How α scales with d: unfortunately the interaction between α and d has no known closed form.

Review 2: We will try to make the paper more understandable. The responses are ordered according to the questions.

1. Lemma 1 solves the minimization problem in Eq. 1. It shows how much the probability of the correct class at $\boldsymbol{x}$ can decrease if we move to $\bar{\boldsymbol{x}}$ for any (worst case) smoothed predictor provided that we know the likelihood ratio regions for the perturbation distribution. The worst case classifier assigns one class for regions with high likelihood ratio, and another class for the remaining regions. This is why they are ordered. The classification in the borderline regions may need to be randomized (the middle case in Eq. 2).
2. We study ℓ_0 robustness in discrete domains (ℓ_0 distance in $\mathbb{R}^d$ would not work for any guarantees).
3. Bijection holds in $\{0,1\}^d$, not in $\mathbb{R}^d$. We can map any Gaussian perturbation result into a discrete domain by embedding $\{0,1\}^d$ as points in $\mathbb{R}^d$, obtaining the continuous certified radius, and mapping back to an ℓ_0 radius over the binary points. Or we can map the Gaussian perturbation first to a discrete perturbation to obtain the guarantee. In the latter case, the guarantee derived from the discrete perturbation is tighter than the one from the Gaussian perturbation, since the more precise assumption yields a tighter result.
4. The 2nd row in Table 1 is about the thresholding discussion. The 3rd row in Table 1 and $\mathcal{N}$ in Table 2 adapts the ℓ_2 certificate of radius r using the Gaussian perturbations to an ℓ_0 certificate of radius $\lfloor r^2 \rfloor$.
5. $K = 1$ (binary) in binarized MNIST and Bace, and $K = 255$ (pixel) in ImageNet. The certificate is always tight regardless of K, and the complexity (pre-computing $\rho_r^{-1}(0.5)$) is $\Theta(d^3)$ (line 166), which does not depend on K.
6. Yes, we can provide additional comparisons in terms of AUC in the final version. However, such comparison will only indirectly show the tightness of the bounds (the purpose of this experiment).

Review 3: We thank the reviewer for accurate explanations and helpful suggestions. We only clarify a few things.

- We will publicly release $\rho_r^{-1}(0.5)$, with the implementation, for all the α and r used for MNIST and ImageNet.
- Despite the restrictive assumption, the smoothed tree yields large improvements even in the raw AUC (x=0 in Fig. 4).
- The first 2 rows in Table 1 refer to the same smoothed classifier; the discrete perturbation in $\{0,1\}^d$ can be interpreted as a Gaussian perturbation with denoising ζ, so it can be certified in both ways. (also see the 3rd response to R2)