Our gratitude to the reviewers for their constructive comments and the useful references they pointed out. We will revise the paper accordingly.

Takeaway message from Theorem 1. Following Rev.1's comments on the significance and interpretation of Theorem 1, we will clarify the message in the final version of the paper. Theorem 1 states that noise injection ensures robustness (in the sense of Def 2). The degree of robustness, ϵ in Theorem 1, is in $O(||\theta||)$, where θ is the parameter of a given Exponential distribution E_F. Since θ is decreasing w.r.t. $std(E_F)$ in general, the larger the amount of the added noise is, closer to each other are the output distributions of the randomized classifier. The control of the added noise and its impact on the accuracy gap is the subject of Theorem 2 (see discussion at lines 219-228).

About the notion of certified accuracy. Reviewers discussed our use of the term "certified accuracy". What we named "certified accuracy" was indeed a probabilistic guarantee (this will be precised in the paper), and thus is different from the notion mentioned in papers reviewers referred to. Giving this kind of "certified accuracy" bounds was not the main focus of our work, nevertheless, we are able to devise certificates in the spirit of these papers by leveraging our concepts. Notably, our work covers the certificate obtained in [LAG$^+$18] as presented below.

Theorem *Let $x \in \mathcal{X}$ be some input vector, and M be a probabilistic mapping such that for any $y \sim \mathrm{M}(x)$, $y = (y_i)_{i\in[K]}$ is a probability vector of size K. If M is d_λ-(α, ϵ) robust, and if there is some k^*, and some $0 < \delta^* < 1$ for which $\mathbb{E}_{y\sim\mathrm{M}(x)}[y_{k^*}] > e^{2\epsilon'} \max_{i\neq k^*} \mathbb{E}_{y\sim\mathrm{M}(x)}[y_i] + (1 + e^{\epsilon'})\delta^*$, with $\epsilon' = \epsilon + \frac{\log(1/\delta^*)}{\lambda-1}$. Then, for the classifier $f(.) = \mathrm{argmax}_k \mathbb{E}_{y\sim\mathrm{M}(.)}[y_k]$ there is no perturbation $\tau \in \mathrm{B}(\alpha)$ such that $f(x) \neq f(x + \tau)$.*

Proof *Let us consider some $x \in \mathcal{X}$. If M is d_λ-(α, ϵ) robust, then with a proof similar to [Mir17] Proposition 3, one easily gets that $\mathbb{E}_{y\sim\mathrm{M}(x)}[y] \leq e^{\epsilon + \frac{\log(1/\delta)}{\lambda-1}} \mathbb{E}_{y\sim\mathrm{M}(x+\tau)}[y] + \delta$ (element wise), for any $\delta \in (0, 1)$. Then one can use [LAG$^+$18] Proposition 1 to get the expected result.*

On the probabilistic extensions of the theorems. Rev.3 pointed out that the results we show in Theorem 1 only hold for $\gamma = 0$. It is possible to extend the result for $\gamma > 0$. To do so, one possible workaround is to replace the robustness guarantee in Theorem 1 as follows: adding noise drawn from $E_F(\theta, t, k)$ ensures that the randomized network is $d_{R,\lambda}$-$(\alpha, \epsilon, \gamma)$ robust with $\epsilon = ||\theta||_2 \omega_t^{B,2}(L) + \omega_k^{B,1}(L)$ and $\gamma = \mathbb{P}_{x\sim\mathcal{D}_x}(\exists\tau \in B_A(\alpha), ||\phi(x + \tau) - \phi(x)||_B > L)$. The same guarantee can be derived for the Gaussian case. Theorem 2 can also be extended: if M is $d_{R,\lambda}$-$(\alpha, \epsilon, \gamma)$ robust for some $\lambda \geq 1$ then: $|\mathrm{Risk}_\alpha(\mathrm{M}) - \mathrm{Risk}(\mathrm{M})| \leq 1 - (1 - \gamma)e^{-\epsilon}\mathbb{E}_x\left[e^{-H(\mathrm{M}(x))}\right]$. However, in practice, it is intractable to compute γ as the data distribution is unknown. A discussion will be added in the paper.

On the convergence of the entropy estimator. Rev.3 raised a question about the convergence of the entropy estimator. We used the MLE estimator from [Pan03] which is endowed with convergences guarantees. By integrating the correcting bias (Miller-Madow estimator, Section 3), we can derive bounds for the entropy estimator. We will clarify this point in the final version of the paper.

Accuracy guarantee for the case with no adversary. Studying the impact of corrupted noise on generalization is another line of research, complementary to the scope of this paper. Outside some specific cases (GLM, Gaussian noise, etc.), proving generalization bounds w.r.t. noise injection in general settings is still an open question to the best of our knowledge.

Other remarks.

- *Misnomer of "generalization":* We agree with Rev.1. In the final version we will use "risk gap" instead.
- *On Table 1:* The accuracy without attack is indeed evaluated empirically. We will clarify that in the final version.
- *On Table 2:* We share the same intuition with Rev.2: the noise makes the attack problem harder for both attacks.
- *PGD iterations:* For the sake of comparison we used the same number of iterations as in [MMS$^+$18] (Table 2). We are eager to use experiments with a large number of iterations and report the results in the final version of the paper.

References

[LAG$^+$18] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 727–743, 2018.

[Mir17] Ilya Mironov. Renyi differential privacy. In *IEEE Computer Security Foundations Symposium*, 2017.

[MMS$^+$18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[Pan03] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 2003.