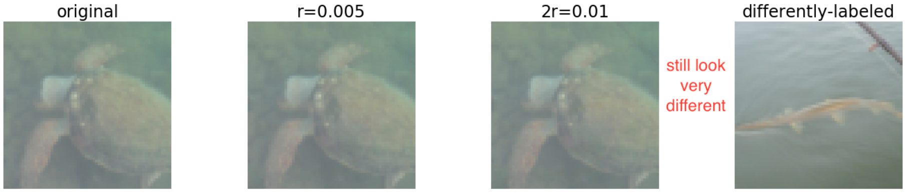
We thank all reviewers for their comments. We are glad everyone found out paper well written. Our main contribution is conceptual, showing that it is possible to achieve both robustness and accuracy in principle (contrary to previous work). We also identify that the robustness-accuracy gap seems to be caused by generalization shortcomings, which highlights the importance of addressing this with better algorithms in the future. Below we address specific comments.

R1. *"There exist no real world data where the classes are well-separated... images suffer from different lighting and lens conditions... objects do not occur in isolation..."*
It seems like there is some confusion between classes being *well-separated* and the notion we use, *r-separation*. The property of r-separation merely requires two images with different classes to have distance at least r. If the lighting changes, then this may increase the in-class variation, but it is **not** going to make the image look like it is from another class (in pixel space). To be concrete, here is a turtle and a fish from Restricted ImageNet. A perturbation distance $r = 0.005$ is used for this dataset (the separation, in terms of r-separation, is much larger). Even with a $2r$-perturbation (for a dim image), the classes do not overlap. Moving $2r$ from the turtle to the fish **still** looks much more like a turtle.



In Section 3, we show that four datasets (MNIST, CIFAR-10, SVHN, ResImageNet) are all r-separated (but, of course, they are not highly clustered). Even in these r-separated datasets, we see a robustness-accuracy tradeoff, which implies that **the problem is with the algorithms**. If robustness is needed for more overlapping datasets, then one option is to make them more separated (at the cost of some accuracy) using Adversarial Pruning from Yang et al., 2020.

R1. *"The classifier proposed in the existence proof essentially computes the distance of a test point to every point in the training data. ... In practice, this would be infeasible, or contrary to policy."*
We **agree** that there are limitations on the classifier used for the existence proof (robustness and accuracy in practice is a huge open problem). However, our theory result is not intended to be used in practice. Instead, it rigorously establishes that the robustness-accuracy trade-off is *not intrinsic* for r-separated data.

R2. *"...empirically demonstrated that robust training methods covered in the paper appear to hurt generalization. More importantly, however, the discussion on generalization is insufficient and lacks theoretical evidence at a NeurIPS level."*
The gap between training and testing accuracies are larger for the robust methods. A theory of generalization for neural networks that can explain adversarial examples adequately is a **long-standing open problem**. Some recent[1] works[2] claim to do so, but they are still in their infancy do not explain many aspects of the problem.

R2. *"It only tested against the PGD-10 attack which cannot represent the true robustness."*
Actually, in Section E.1 of our submission we test against the **multi-targeted attack** for a total of 200 iterations for a dataset with 10 classes (this is essentially SOTA for ℓ_∞) and we see that the overall trend between adversarial test accuracy and local Lipschitzness remains the same (even though the adv. accuracy goes down across the board).

R3. *" ..., it is somehow unfortunately that the authors did not propose a framework on how to attain this target of encouraging both an accuracy increase as well as robustness ... "*
Two long-standing open problems in this field are: (i) understand theoretically how networks generalize adversarially, (ii) design a framework that achieves robustness and accuracy empirically. In our work, we do not claim to design such framework. Instead, one of the main contributions of our work is to show evidence that **such a framework should exist**, especially for standard image datasets (this was not clear before). The result encourages future work to focus on finding such framework instead of believing that robustness and accuracy are at odds with each other.

R3. *The paper claimed and showed the robustness and accuracy are achievable at the same time for the real image datasets. ... So the conclusion or the claim is not very new at least to the reviewer.*
The reviewer seems to be conflating two results that we believe are quite different.We are showing that the robustness and accuracy should be *nearly perfect* for standard datasets. This is **not evidenced** by the recent works showing that robustness can help accuracy a little bit. For example, on CIFAR-10 or ResImageNet, many papers report trade-offs, and there is little evidence of any current solutions that get close to what we can show in theory. Therefore, we think that our theoretical results actually tell quite a new story about **why** the trade-off is not intrinsic, and about why generalization seems to be the key barrier.

[1] Belkin, Mikhail, Daniel J. Hsu, and Partha Mitra. "Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate.", 2018.

[2] Wei, Colin, and Tengyu Ma. "Improved sample complexities for deep networks and robust classification via an all-layer margin.", 2019.