We thank all reviewers for their positive reception of our paper and for their constructive feedback.

Reviewer #2

- **On dual norms and prior work.** Thank you for pointing us to the relevant prior work of Demontis et al. and Xu et al. which we apparently missed.

  Our results about the tradeoff between ℓ_1 and ℓ_∞ robustness do indeed share similarities with the dual-norm behavior studied by Demontis et al. for linear classifiers. The main difference is that for our Gaussian dataset, we show that the robustness tradeoff is inherent to *any* classifier (i.e., not only linear ones). Another difference is that we extend our results to tradeoffs between ℓ_∞ and spatial perturbations.

  For arbitrary dual norms ℓ_p and ℓ_q, we can also exhibit a tradeoff as follows (somewhat informally):

  - Flipping the features $x_1, \ldots, x_n$ requires a ℓ_p-perturbation of magnitude $O(\mu \cdot n^{1/p}) = O(n^{1/p-1/2})$.
  - Flipping the feature x_0 requires a ℓ_q-perturbation of magnitude $O(1)$.
  - So if a model is robust to perturbations of size $O(n^{1/p-1/2})$ in the ℓ_p-norm, it cannot also be robust to perturbations of size $O(1)$ in the dual norm.

  We will discuss these connections between our work and the prior work of Demontis et al. and Xu et al. in the final version of our paper. We thank the reviewer for suggesting this dual-norm view which nicely generalizes one of our results.

- **On structure and readability.** We agree that our paper contains many contributions (the formal analysis, a new ℓ_1-attack, an experimental evaluation) that are somewhat heterogenous. We will make an effort to clarify our main contributions and to better structure our paper to improve its readability.

Reviewer #3

- **On MNIST artifacts.** The gradient masking effect we discover and explain is indeed specific to MNIST (for multiple ℓ_p norms), and we do not claim otherwise. In fact, this gradient masking effect seems inherently due to the mostly binary nature of the MNIST images, which leads to thresholding being a viable defense against ℓ_∞-perturbations.

  Nevertheless, as MNIST is the only vision dataset for which we've been able to train models to high levels of robustness (for individual attack models), we believe it is worthwhile to observe that extending this robustness to multiple ℓ_p attacks may be particularly challenging for this dataset. In this sense, even a dataset as simple as MNIST is clearly not solved from an adversarial robustness perspective.

  We do believe that it should be possible to train MNIST models to a robustness tradeoff similar to that we found on CIFAR10. But this will require new techniques that somehow circumvent gradient masking as a spurious solution. We think this is an interesting open problem for the community to consider.

  On CIFAR10, the robustness tradeoff is indeed smaller but still quite noticeable. It is worth noting that our experiments on CIFAR10 only ever consider the combination of two attack types. It is not clear what would happen if we were to try to train models to be robust to 4 or 5 attacks at a time for instance. For these types of experiments we are mainly limited by the poor scalability of adversarial training (e.g., for two attack types, adversarially training a wide ResNet takes about two GPU weeks). There is some promising recent research on speeding up adversarial training, so these types of experiments might become tractable for future work.

- **On black-box attacks.** The Ensemble Adversarial Training technique of Tramer et al. was proposed to increase a model's robustness to black-box attacks, but it was found to have no noticeable effect on the model's robustness to stronger white-box attacks. As our evaluation focuses on the white-box robustness of the trained models, we have not incorporated black-box attack examples at training time.

  We also considered using black-box attacks at evaluation time (e.g., as a test against gradient-masking), but found decision-based attacks to be stronger and more reliable for this purpose.