We thank the reviewers for their insightful feedback, and we appreciate the opportunity to improve our paper. We will address typos and notational inconsistencies in the updated version.

Response to Reviewer 1:

We would like to emphasize that Theorem 1 is the most important contribution of our paper due to its generality. By considering the set of all possible classifiers, it provides lower bounds on adversarial robustness for *any pair of class-conditional distributions*. As we show in our experimental results in Section 6, we are able to obtain lower bounds for arbitrary real-world datasets by constructing the empirical distribution for these. In our estimation, these results serve to provide theoretical validation for adversarial training for low perturbation budgets as well as to highlight the gap to optimality for higher budgets.

Our focus on the Gaussian case, as a concrete application of the general theorem, is due to the attention this setting has received in relevant previous work such as Schmidt et al. and our results provide a conclusive characterization of the behavior of the optimal loss under different adversarial constraints. We show that the common assumption of the optimality of a linear classifier even in the presence of an adversary is justified through a primal-dual equivalence.

In the Gaussian case, our sample complexity result follows directly from the expression for the optimal loss. In the updated version, we will add experiments with synthetic data which validates this result empirically using standard learning algorithms.

Response to Reviewer 2: We thank the reviewer for pointing us to Dohmatob's "Generalized No Free Lunch Theorem for Adversarial Robustness" from ICML 2019. There are several key differences between the results as well as methods in the two papers. We require very mild assumptions on the example space, distribution, and adversarial constraints while the assumptions in Dohmatob's paper are more restrictive. Further, ours explicitly concern the adversarial risk of the optimal classifier, while Dohmatob's relate adversarial and ordinary risks of a classifier. Thus, our bounds on adversarial risk can still be nontrivial even when there is a classifier with an ordinary risk of zero, which is exactly the case in our MNIST experiments. Finally, while Dohmatob's bounds become non-trivial only when the adversarial budget exceeds a critical threshold depending on the properties of the space, ours apply for any adversarial budget.

We will add the explicit but mild conditions required on the example spaces, neighborhood relations, and potential functions throughout Section 3. X_1 is a random example from class 1 and X_{-1} is an example from class -1. We will add back the explanation of this notation which we accidentally removed. $\tilde{P}_{X_1}$ should have been $P_{\tilde{X}_1}$ everywhere. We will add a more explicit description of $\tilde{X}_1$ for the translate adversarial strategy. This makes the dependence on z explicit on line 196. We will also add a clearer description of the "translate and pair in place" coupling. Finally, in B.2., going from (2) to (3) is a standard calculation for the total variation distance between Gaussians with the same covariance, which we will add in the updated version.

There are at least a few interesting examples of adversaries that produce examples in a different space than the clean examples, e.g. by erasing pixels in a image. Allowing symmetric nearness relations does not complicate the proofs, it only requires us to keep track of the difference between $N(x)$ and $N^{-1}(\tilde{x})$.

While we were unable to find the same calculation for the upper bound on classification accuracy for the Gaussian case in Tsipras et al. [74], we did find it in concurrent work from the same group (Ilyas et al., Arxiv: 1905.02175). We will add a citation and comparison in the updated version.

Response to Reviewer 3: While other papers such as Sinha et al. [68] and Dohmatob use ideas from optimal transport, we are the first to identify the precise distributional distance metric that provides tight lower bounds on adversarial robustness. Comparisons with Sinha et al. are in Section 7 and we compare to Dohmatob above. We would like to emphasize that our identification of this metric has allowed us to apply our theoretical results directly to practical datasets of interest.

We are currently investigating the extension of our results to the multi-class case. There is a close connection between our framework and targeted adversarial examples in the multi-class setting. In this case, the transportation distances between all pairs of classes characterize the performance of an optimal classifier. Since the number of distances required for this characterization scales as the square of the number of classes, we are attempting to understand how much information is contained in the one-vs-rest distances. Exact characterization of classification accuracy with untargeted adversarial examples seems to require higher order interactions between class distributions. However, usable bounds using only pairwise distances are available, which we will demonstrate in follow-up work.

As the reviewer correctly notes, the robust classifier loss on CIFAR-10 is high even for small budgets. Nevertheless, we will add these in the updated version of the paper. We also ran experiments with a $4\times$ larger model for MNIST, per the reviewer's suggestion, and observed some mitigation of the robust training issues up to an L_2 budget of 4.6. We thank the reviewer for pointing this out and we will update our plots with this model.