We thank all the reviewers for their thorough reviews and insightful comments. Reviewers' comments are in blue.

Reviewer 1: *The authors should explain more how this is actually done, and why it doesn't represent a computational bottleneck for the proposal.* The detailed algorithms for finding robust error regions under ℓ_∞ and ℓ_2 perturbations are provided in Section C in our supplementary materials. For ℓ_∞, we construct the systems of hyperrectangles by first precomputing an approximate k-NN distance estimate using Ball Trees for each data point, and then clustering the top-q densest data points into T partitions using the k-means algorithm, where we binary search for the optimal parameter q. The time complexity of precomputing and sorting the nearest neighbor distance estimates is approximately $O(nd\log(n))$, where n is the total number of data points in $\mathbb{R}^d$. In addition, the time complexity of k-means algorithm is $O(ndTI)$, where I is the averaged number of iterations for k-means algorithm to converge. Therefore, the total computational complexity of our algorithm for ℓ_∞ is $O(nd\log(n) + ndTI\log(1/\delta))$, where δ is the stopping threshold for binary search. In our experiments, we applied our algorithms to medium-sized datasets including CIFAR-10 and SVHN, and they finished reasonably quickly. We will include a runtime analysis of algorithms in the final version.

Reviewer 2: *It is worth including a discussion on https://arxiv.org/abs/1805.12152, which many researchers point to as refuting the concentration of measure hypothesis.* We will include the discussions. In a nutshell, that work uses a definition that coincides with the definition of adversarial examples that we use for the interesting range of tampering parameters (in which the ground truth is robust). But, if the tampering goes up and can change the ground truth, even learning a concept *exactly* might leave room for adversarial examples under the definition used in that work (but not under ours). *I would also be a bit more specific in the introduction that l_p perturbations are a toy threat model that is not intended to model how actual adversaries choose to break systems.* We will make sure to add comments about shortcomings of l_p norm in capturing the whole picture. *Any reason the authors defined an adversarial example as the nearest input which is classified differently, and not the nearest error?* We agree with you that nearest error point is a natural definition of adversarial examples and we indeed use this definition in the paper. Specifically, in Definition 2.1, we compare the true label of x' with the predicted label of x', which means x' should be an error point to be counted as an adversarial example. We will make this point more clear in the statement of our definition. *It's worth noting that MNIST may be a degenerate case with respect to the l_∞ metric. In particular, a trivial defense is to first threshold the inputs about .5 and classify the resulting binary image. Because of this, I would not expect any meaningful bounds to hold for this dataset and metric.* We agree with the reviewer that thresholding MNIST (and any other dataset) will make the transformed distribution not concentrated under l_∞. However, the original distribution (before transformation) might still be concentrated. In particular, one might be able to add a small perturbation to the image before thresholding the features and make the binary transformation of the perturbed image different from that of the original image. For the case of MNIST, it seems that binarizing images should not change the distribution much, as the original images have close to binary form. Our experiments support this intuition and show that regions in MNIST dataset could have a very small expansion (it only grows from $\sim 1\%$ to $\sim 10\%$ when allowing $\epsilon = 0.4$ perturbations). *It would be very interesting if the authors could strengthen their bounds by making additional assumptions on the shape of the error set. Additionally, one could strengthen the bounds by approximating the content-preserving threat model.* Thank you for pointing out these interesting future directions, particularly for the content-preserving mode. Interestingly, part of our theoretical results do already prove such results for restricted forms of error sets, and this does set the stage on how we choose the sets for our experiments. The *proof* of Theorem 3.5 first proves such result for limited shapes. In particular, we obtain such result when VC dimension of the sets *and* their expansion are bounded (e.g., union of hyperrectangles).

Reviewer 3: *Q1- the theoretical innovations in this paper are not practically relevant to the study of adversarial vulnerability. Q2- to disprove the "adversarial examples are inevitable" theory, you only need to show an *upper bound* on the concentration function, i.e. a finding that there exists some set with measure alpha whose epsilon-expansion has measure at most Y. Given a sample from the data distribution, here is a simple way to do that: split the sample into a "training set" and a "test set".* The two questions/comments are relevant. We start with Q2, then will address Q1 as well. Yes, indeed to show that a distribution does not concentrate beyond a parameter, one can aim to show the existence of *some* set (found based on "training set") and test its expansion using the "test set". However, the question is how to design algorithms that come up with such sets. Our theory tells us that by looking at specific types of sets (e.g., collection of hyperrectangles), we can get "generalization" bounds for estimating expansion. Note that we tried different collections of subsets (e.g., subsets decided by neural networks) that were not supported by our theory and we observed huge generalization error that made the experiment meaningless. Therefore, in our experiments we use exactly the subset collections that theory suggests and the results of experiments verify our theory. Our theory is also important for *future* work. If one wants to find the concentration of measure under another metric probability space, they can use our theory to come up with suitable subset collections with generalization guarantees. *As a sidenote, while the authors interpret this to mean that there is room to develop better robust classifiers, it could also mean that robustness is impossible for reasons other than concentration of measure.* Thank you for pointing this out. We tried to be cautious in interpreting our results and consider the "robust classification is impossible for other reasons" hypothesis. However, after reading through the paper we found an occasion in our discussions (line 284-285) that we did not consider this hypothesis. We will make sure to clarify this in the next version of our paper.