We thank the reviewers for their thorough and thoughtful reviews. We greatly appreciate the positive comments and address the questions below.

To Reviewer #1: We thank the reviewer for the comments.

(1) Although there has been no theoretical guarantees before, the convergence of adversarial training to zero loss is well-observed in practice. There are papers, e.g. Madry et al. [24], showing that as the capacity of network increases, adversarial training will converge to nearly zero loss. Moreover, we also conducted experiments showing the convergence of adversarial training for different architectures. For the 3x-wide and 10x-wide Resnet-32 (solid red and green lines), the training accuracy is close to 100%.

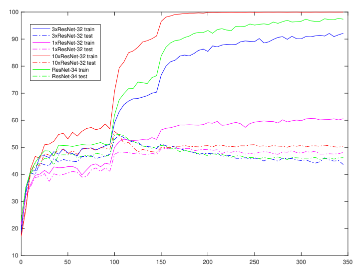

Figure 1: Adversarial Training with Different Architectures. y-axis is accuracy and x-axis is epochs.

(2) Generalization in the robustness literature is an important problem that is not addressed in this paper. We will add a discussion of the work on robust generalization which is complementary to this paper. In future work, we plan on investigating how our adversarial training results can be combined with robust generalization to yield end-to-end guarantees on the robust test loss.

(3) Thank you for pointing out the two papers related to the capacity argument; we will cite these in the next version and discuss the relationship.

To Reviewer #2: We thank reviewer 2 for giving insightful suggestions on both theory and writing.

We wholeheartedly agree that we should talk more about the limitations of the current theory and point out the future directions more clearly. Some of this discussion is in Section 7, especially the possibility of reducing the exponential dependence of the depth into polynomial dependence (we believe using similar techniques in reference [1], reducing to polynomial dependence is indeed possible without changing the structure of the arguments in this paper and potentially even only logarithmic depth dependence via using a ResNet architecture). We will discuss in more detail in the revision, including the need for more fine-grained analysis on the role of depth, architecture, and input data. Of course even for natural training, many of these questions remain open. We will expand upon Section 7 the limitations of the current analysis and routes to improve the analysis, and finally testable hypotheses (stronger attack algorithm leads to stronger adversarial training loss guarantee and adversarial training requires additional capacity even to minimize the training loss).

To Reviewer #3: We thank reviewer 3 for the positive comments, and for giving insightful suggestions on both writing and future directions.

We will follow the reviewer's suggestions in the revision. In particular we will mention early on that the success of adversarial training is dependent upon the ability of the kernel method's expressivity. We will also try to reword the abstract to remove the ambiguity caused by exact vs heuristic inner maximization solving.

In addition, we are currently working on removing the projection in the gradient descent algorithm. For example, we can prove that for the two-layer case the projection step is not needed as remarked under Theorem 4.1; we will include the proof of this in the next version.