We thank all the reviewers for their positive and useful feedbacks, which we will use to improve the paper. We first address the common comments from reviewers:

To Reviewer #1

@ The γ-seperatability requirement is odd since it also requires samples from the same class to be separated: Good point! Actually, γ-seperatability can be replaced by separation only among different classes. With minor modifications, our results can be applied to classification losses, like cross-entropy. For such losses, we can define δ= minimum distance between different classes and thus $1/\gamma$ not being large becomes even more practically relevant, as an assumption (remember that $\gamma = \delta(\delta - 2\rho)$). We will add more discussions in the next version.

@ Can something be said about the robustness against the most intelligent adversary for a network trained using a polynomial time adversary?: This is certainly an interesting future direction. This might require characterization of what polynomial time adversary means, since training with a trivial adversarial which only generates the original data doesn't lead to robustness against the most intelligent adversary.

Note that our results apply to any adversary (including the most intelligent one) used in training.

To Reviewer #2

We appreciate the suggestions and will update the related-work section accordingly. Regarding experiments, we note previous works (e.g. Figure 4 in Madry et al.) have already shown empirically that reasonably wide networks achieve small robust training error. Our work serves as a potential theoretical explanation for this phenomenon.

To Reviewer #3

@ "Wang et al.": We are unable to find any related works with the reference "Wang et al". Given the comments, we believe the reviewer meant to refer to Gao et al.

@ the paper lacks novelty, and the contribution is small because Gao et al. only requires poly(d) width with a seldom used activation function where d is the input dimension: This claim made by the reviewer is incorrect. In Gao et al., in order to achieve small robust training loss, Theorem 5.2 and Corollary 5.1 require the width to be polynomial in the constant $R_{D,B,\epsilon}$, and Gao et al did not to show $R_{D,B,\epsilon}$ is poly(d) with *any* activation function. In fact, they only managed to upper bound $R_{D,B,\epsilon}$ by $(1/\epsilon)^d$ with the quadratic ReLU activation. One of our main contribution is to bound $R_{D,B,\epsilon}$ by poly(d) with the ReLU activation using a novel analysis.

This misunderstanding is possibly due to a claim made in Gao et al.'s intro that "we show that projected gradient descent converges to a network where the surrogate loss with respect to the attack $\mathcal{A}$ is within ϵ of the optimal robust loss. The required width is polynomial in the depth and the input dimension." We note that although this claim is correct, the optimal robust loss in their setting may not be necessarily small. The only concrete case where they prove it is small is for quadratic ReLU networks with $(1/\epsilon)^d$ width (Theorem C.1 in Gao et al.).

@ it is misleading that the authors claim quadratic ReLU is not used in practice: We are not aware of the use of quadratic ReLU in any practical setting but are happy to change that phrase if the reviewer could give us some references. But it is important to note that even for the quadratic ReLU, the upper bound in Gao et al. is exponential in d. That is the main point in that line.

@ appendix C.2 instead of C.1 : We thank the reviewer for pointing out this typo. We meant to write theorem C.1. We will fix it in the next version.

To Reviewer #4

@ the paper could benefit from small toy examples experimentally demonstrating the authors' claim. E.g., plot the dependence of robust training loss vs depth: By "loss vs depth", did the reviewer mean loss vs width? If so, many previous works (e.g. Figure 4 in Madry et al.) have already showcased the suggested experiments. They show larger width leads to lower robust training loss. Our work serves as a theoretical explanation for such experiments.