We thank all the reviewers for their valuable comments. We would be happy to revise our paper using their suggestions.

Reviewer 1: Lipschitz constant and adversarial robustness. While we agree that different approaches can be used to improve adversarial robustness, optimizing the Lipschitz constant of a model is shown to tightly bound the *worst-case* error due to bounded input perturbation (See Fig 1(e) below from our experiments; our SI and *Fazlyab et al., NeurIPS 2019*). The Lipschitz constraint is particularly effective at targeting precisely the regions of peak sensitivity of the model, and our approach thereby provides provable guarantees for the trained model against adversarial perturbations.

We respectfully disagree with the statement that a "model can have large Lipschitz constant but can still be robust". In the provided example, if g achieves a higher performance than f, then the model g lacks robustness. In fact, a very small perturbation (of the order of $1/1000$) can degrade the performance of g at all testing points close to the decision boundaries (which are, in the absence of perturbation, correctly classified due to the high-frequency component). If g achieves the same performance as f, then g contains an unnecessary high-frequency term. In fact, because our loss function is strictly convex, it admits a unique minimizer, and high-frequency components are present in our models only if they improve the overall performance (our Laplacian approach favors smooth solutions). This analysis is compatible with the result that the Lipschitz constant provides a worst-case bound on the robustness of a model to perturbations.

Finally, despite its title, the reference cited by Reviewer 1 does not claim that bounding the Lipschitz constant is not a good approach to obtain a robust model. Rather, it argues in favor of our work: it claims that the existing methods to compute the Lipschitz constant had limitations, and that overcoming them would lead to adversarially robust models.

Reviewers 1 and 4: Empirical validations, figures, and reproducibility. We thank the reviewers for pointing out these weaknesses. We recreated our figures (below; which now align with Reviewer 1's intuition). We can now provide more MNIST implementation details, upload our code to GitHub, and further demonstrate our robust training scheme.

Reviewer 2: Confidence vs accuracy. Confidence and accuracy are directly related (Figure 1(d)). Confidence is a smoother metric for optimization and encoding the Lipschitz constraint, and is readily mapped onto one-hot vectors.

Reviewer 2: Scalability. We are able to successfully demonstrate our approach using the MNIST dataset. While additional studies are needed to investigate scalability, our numerical studies show that our performance increases rapidly with the number of vertices (model complexity), with an accuracy of 96% with only 10000 vertices (Figure 1(c)). For comparison, our study shows that neural networks with the same number of parameters achieve only 68% accuracy.

Reviewer 2: Dataset. We choose the checkerboard dataset precisely because it requires a Lipschitz constant that tends to infinity, so as to compare the performance over a broad range. We use the MNIST dataset for a more realistic study.

Reviewer 3: Conclusion and novelty. To the best of our knowledge, our results are the first to prove that a fundamental tradeoff exists between Lipschitz constant and accuracy. Prior works (cited) provide empirical or less general results. Our paper contributes both new results to the literature to address the immediate questions on tradeoffs, and a novel way of looking at the learning problem (connections to PDE) that has broader scope and will pave the way for future works.

Reviewer 3: Connections between problems and motivation. The first problem generates models with minimal loss and desired robustness. The second problem fixes the performance (loss) and improves the robustness of a model. While the first problem yields a training scheme, the second is used to prove a fundamental tradeoff between performance and robustness. We have drawn motivation for our work from prior works on adversarial attacks that have demonstrated a glaring problem of robustness of machine learning models and a need for adversarially robust training.

Reviewer 4: Discretized models and neural networks. The discretized versions of the problems offer a way to construct an implementable training scheme with robustness guarantees. Implementable models are obtained by various discretizations of the original infinite-dimensional loss minimization problem. We have considered a graph discretization that directly inherits the convexity of the infinite-dimensional problem. As the reviewer points out, other function space parametrizations, such as neural networks, result in non-convex finite-dimensional optimization problems. Our infinite-dimensional analysis holds regardless of parametrization and characterizes the solution that particular models approximate, while our graph-based design offers a novel and alternative scheme to construct provably-robust models.

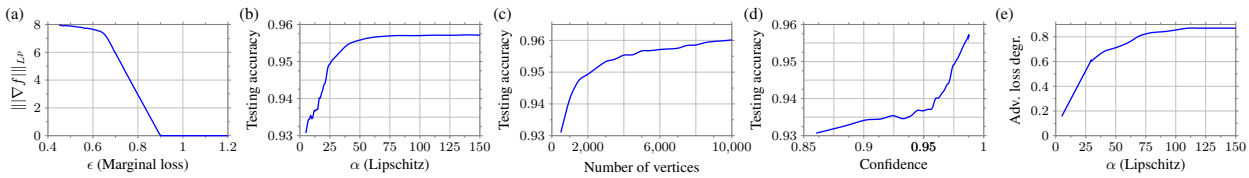

Figure 1: Figure (a) updates Fig. 2c in the paper. Figures (b)-(e) complement our study of the MNIST dataset by looking at the relations between accuracy and Lipschitz, accuracy and complexity, accuracy and confidence, adversarial robustness and Lipschitz.