We thank all the reviewers for their helpful comments. We address specific questions below.

Reply to Reviewer 2

Question: New work by Diakonikolas et al.

We thank the reviewer for their thorough review and for alerting us to the recent work by Diakonikolas et al. [1]. We will be sure to provide a detailed comparison with this paper in the camera-ready. [1] showed that learning the single ReLU neuron up to $O(\mathsf{OPT}) + \varepsilon$ risk for log-concave and isotropic distributions is possible if one uses gradient descent on a convex surrogate risk for the squared loss; it was previously known that learning up to exactly $\mathsf{OPT} + \varepsilon$ is impossible in polynomial time [3]. The updates by gradient descent on this surrogate correspond to the GLMTron updates of [2]. By contrast, our bounds for strictly increasing and Lipschitz activations cover *any* distribution over x with bounded marginals, with dimension-independent sample complexity, by minimizing the (nonconvex) empirical risk with vanilla G.D., although we achieve a weaker guarantee of $O(\mathsf{OPT}^{1/2})$. Although our risk guarantee is weaker, we believe a complete characterization of what distributions can be agnostic PAC learned using neural networks trained by gradient descent on the empirical risk is a fundamental research question. Our work provides, to our knowledge, the first positive result on this question for the single neuron in the agnostic PAC learning setting.

Question: Lower bounds

Thank you for your suggestion for studying lower bounds for this problem. There are two types of lower bounds that we are interested in: (1) whether there exist distributions for which no algorithm can achieve population risk $\mathsf{OPT} + \varepsilon$ for the single neuron; (2) whether there exist distributions for which gradient descent on the empirical risk cannot achieve population risk $\mathsf{OPT} + \varepsilon$ (or even $O(\mathsf{OPT}^{1/2}) + \varepsilon$). For ReLU, [3] addresses (1), and [4] addresses (2), but to our knowledge no such results are known for nontrivial strictly increasing and Lipschitz activations. We hope to explore these questions in future work.

Reply to Reviewer 3 and Reviewer 4

Question: Significance of agnostic learning of a single neuron

We thank the reviewers for their comments. We agree with the reviewers that the agnostic PAC learning of neural networks with multiple neurons and layers is an important problem. Unfortunately, there are very few works that have been able to show any results in this direction, as we describe in lines 41–53 and 75–92 of our submission. Even in the single neuron setting, the question of what distributions can be PAC learned has only recently begun to be understood [1,3,4].

We believe that the agnostic PAC learning of the single neuron using gradient descent is a fundamental problem for the understanding of neural networks. Without a full characterization of what can be learned for the simplest possible neural network—the single neuron—it seems unlikely that we will find satisfying explanations for why complicated, deep neural networks trained by gradient descent are so successful. As our work is the first result for agnostic PAC learning for a single neuron using gradient descent on the empirical risk, we think we have made significant progress on this problem.

[1] I. Diakonikolas, S. Goel, S. Karmalkar, A. R. Klivans, and M. Soltanolkotabi. Approximation schemes for relu regression. In *Conference on Learning Theory (COLT)*, 2020.

[2] S. M. Kakade, A. Kalai, V. Kanade, and O. Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2011.

[3] S. Goel, S. Karmalkar, and A. R. Klivans. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[4] G. Yehudai and O. Shamir. Learning a single neuron with gradient methods. In *Conference on Learning Theory (COLT)*, 2020.