We thank the reviewers for their very positive and encouraging feedback.

Regarding the main comments raised by the reviewers:

- The strong distributional assumptions: we believe similar techniques can be used for showing similar results under more general families of distributions, a problem that we leave for future work. Specifically, the key property that we used for showing our positive result is the fact that the gradient-descent draws the weights of the irrelevant bits to zero, a behavior that can be observed in other distributions as well. We focused on this choice of distribution to make the analysis simpler, and since this setting suffices to demonstrate the separation between neural networks and linear classes, which is the main goal of the paper.
- The choice of a parities over odd number of bits is important since we assume that, in half of the distribution, the parity bits are either all 1 or all -1. We believe this assumption can be removed when considering different distributional assumptions.
- We introduced the regularization term to simplify the theoretical analysis, but we believe similar results can be shown in different settings as well.

We will fix and address the additional comments and suggestions raised by the reviewers in the final version of the paper.