We thank all the reviewers for their time and effort in providing feedback. We are encouraged by the universally positive scores (7 7 7 6) and that all the reviewers appreciated the paper for the following: (i) theoretical contributions (**R1**,**R2**,**R3**,**R5**), (ii) advancing our understanding of the LTH (**R2**,**R3**), (iii) novel connection to SubSetSum(**R3**,**R5**), (iv) clear exposition (**R1**,**R2**,**R3**,**R5**), and (v) relation with prior work (**R1**,**R2**,**R3**,**R5**). Moreover, **R3** thinks this connection with SubsetSum can be applicable in other areas.

For clarity, we would like to reiterate the goal and motivation of the paper. We provide theoretical justifications for the striking empirical observations in Ramanujan et al. [1]: Do good subnetworks *provably* exist with a small factor of overparameterization? Our main theoretical contribution in this paper is to *characterize* the required overparameterization (up to logarithmic factors) for fully connected networks, offering **an exponential improvement** in the oveparameterization bounds by Malach et al.

We address the individual concerns below. We thank **R3** for pointing out the typo.

SubsetSum is NP-Hard (R2): We note that the connection to the SubsetSum problem is made so we can establish an **existential** rather than an algorithmic result: in this work we prove the existence of a subnetwork that performs as well as any target network, but do not claim to offer a "good algorithm" to do so. In general, we do not have access to the target network, but only to the labeled training data. Thus, an important–yet orthogonal to our goal in this paper–question is whether there exist poly-time pruning algorithms with provable performance. As optimizing ReLU neural network is itself NP-Hard in general, we expect all algorithms to be inefficient in the worst case. However, similar to the success of backpropagation on standard tasks, the findings of [1] suggest that a pruning-based algorithm is potentially practical. Understanding the effectiveness of pruning given certain data/model assumptions is an important future direction that we are currently working on.

Experimental comparison of target, subsetsum, and edge-popup pruned models (R2): We believe that a significant level of network isomorphism between the two pruned networks (obtained from SubsetSum and edge-popup) would be extremely unlikely, because the pruning algorithms are quite different. Because the networks differ in both weight distribution and structure, we cannot hope for a significant level network isomorphism (even upto the ordering of neurons in a layer), but we can still compare using some other metrics like sparsity (Figure 2 in our paper). For a two-layer fully-connected network, our SubsetSum approximation utilized ~ 3.5 million out of ~ 8.3 million coefficients in total. Thus, the approximated network achieved 97.17% test set accuracy with $\sim 44.69\%$ sparsity. On the other hand, one of our networks resulting from edge-popup achieved a 97.53% test set accuracy by retaining ~ 0.5 million of ~ 2 million parameters, thus giving $\sim 25\%$ sparsity, but on a much smaller network.

Lower bound for deeper networks (R3): We would like to emphasize that our lower bound controls the number of parameters and thus is valid for any depth and architecture. In particular, for constant depth greater than 2, we directly obtain $d\sqrt{\log 1/\epsilon}$ lower bound on width. We agree that obtaining a tighter lower bound is an important question, and we believe it should be possible with regards to the way that the depth and width appears in the logarithm. Answering this question requires understanding the role of depth in the representation power of neural networks. This is a notoriously difficult problem and, even after years of research, remains elusive [2].

Other activation functions (R3): Although our paper (and prior work [3]) focuses on ReLU activations, the proof strategy works for a general class of Lipschitz activations with a slight modification in the architecture. Specifically, following the structure of Section 3.1, suppose that every alternate layer is linear. The proof then goes through *exactly* as before by using the Lipschitz property in Eq. (13) (Line 468) in Appendix. We will add a comment regarding this in the final version of the paper.

Experiments on other datasets (R5): We agree that experiments with datasets beyond MNIST would be beneficial to further validate our findings. Currently, our "diamond" based architecture inspired by Lueker's random SubsetSum theorem is defined only for fully connected layers. Data sets like CIFAR10 and ImageNet require deeper convolutional networks that contain very few fully-connected layers. A good test set accuracy on MNIST is obtainable with both shallow fully connected networks and LeNet5, which has more fully connected layers than convolutional layers. Thus, although somewhat limited, we believe that MNIST is an appropriate choice for experiments in this case. However, deep convolutional networks such as ResNet and VGG generally contain up to three fully connected layers that our structure shown in Figure 1c could be applied to, so we think that investigating the performance of our structure on deep convolutional networks is valid and worth further investigation. We were not able to complete the design and execution of these experiments by the rebuttal deadline, but are eager to do so for the camera ready version.

We would again like to thank the reviewers for the positive reviews.

References

[1] V. Ramanujan, M. Wortsman, A. Kembhavi, A. Farhadi, and M. Rastegari. What's Hidden in a Randomly Weighted Neural Network? *arXiv:1911.13299*.

[2] G. Vardi and O. Shamir. Neural Networks with Small Weights and Depth-Separation Barriers. *arXiv:2006.00625*.

[3] E. Malach, G. Yehudai, S. Shalev-Shwartz, and O. Shamir. Proving the Lottery Ticket Hypothesis: Pruning is All You Need. *arXiv:2002.00585*.