We would like to thank the reviewers for their valuable comments. Below, we address the concerns raised.

Reviewer 1.

Q. The assumption in Line 207 seems strong compared to the results in [Shamir, 2018].

Please note that [S18] considers only one-block ResNets, and as mentioned in Lines 236–239, the condition in Line 207 is always satisfied when $L = 1$. In other words, our paper subsumes and extends the results of [S18].

Q. The result in Section 3.2 seems like a trivial application of [Shamir, 2018].

The example in Section 3.2 is not a result of [S18]. Rather, the example is meant to show that direct application of [S18] is **not possible** for deeper ResNets. As discussed in Lines 178–182, an application of [S18] only proves $\mathrm{err}(H_2, Y) \leq \mathrm{err}(H_1, Y)$, so a comparison of $\mathrm{err}(H_2, Y)$ and $\mathrm{err}(X, Y)$ does not follow from [S18]. Further, our example shows that even $\mathrm{err}(H_1, Y) > \mathrm{err}(X, Y)$ is possible, showing that theoretically proving $\mathrm{err}(H_2, Y) \leq \mathrm{err}(X, Y)$ is challenging even for $L = 2$. (We hope that the reviewer's confusion regarding Lines 179–181 is clarified by this answer.)

Q. Section 3 feels separated from the following results in the paper.

The examples make the following points: (1) the advantage of ResNets; and (2) the difficulty of analyzing *deep* ResNets. In particular, the second example highlights that the results in the subsequent sections are not trivial extensions of [S18].

We will also reflect the other comments in the revision.

Reviewer 2.

Q. The paper argues that the geometric conditions [...] sufficient conditions when it should hold.

Thanks for pointing this out. In fact, we do provide a sufficient architectural condition; please refer to Lines 226–235.

Reviewer 3.

Q. My main concern regarding this paper is that proofs [...] future improvement based on them.

Our proof is not based on a recursive analysis of each residual block. By recursive analysis, we believe that the reviewer means some induction-like argument, which is not the case here. We analyze each block of the network separately using partial derivatives and carefully designed perturbations, and collect the results from each block to prove the theorem. We should mention that our proof technique already reveals the importance of identity skip connection in ResNets: If the identity terms in residual blocks are replaced with tunable parameter matrices $\boldsymbol{W}_l$, it is impossible to apply the argument in Lines 405–409 unless strong assumptions on $\boldsymbol{W}_l$ are made. This suggests that identity skip connections are indeed the key to the benign loss landscape.

Q. Moreover, the second assumption made in theorem 2 [...] a rather strong assumption on the architecture.

We agree that if the condition in Corollary 3 is not satisfied, the second condition may not necessarily hold. We believe that weakening/removing the condition is a highly nontrivial future challenge. But as noted in Lines 240–246, we'd like to stress that not only does Theorem 2 subsume previous results on shallow ResNets, but it also shows that a **chain of multiple** skip connections, *as opposed to direct skip connections to the output*, has beneficial effects on the loss landscape. It has been argued that deep ResNets have benign loss landscapes compared to fully-connected networks, but a rigorous theoretical understanding is still missing; we believe that our results take a step forward in that direction.

Q. The examples provided make understanding and reading the paper easier, [...] standalone results.

We agree; the examples are just to note the advantage of ResNets and highlight the difficulty of analyzing *deep* ResNets.

Q. Additionally, section 5 strongly relies on the near-identity assumption [...] incremental improvements.

As summarized in the paper, previous theoretical results have shown that near-identity regions are expressive and have certain benign optimization properties. In addition, a recent paper *"Fixup Initialization: Residual Learning Without Normalization"* showed that initializing ResNets in near-identity regions also leads to good empirical performance. The finding of that paper is rather surprising. For each residual part $\Phi^l_{\boldsymbol{\theta}}(\cdot)$ (according to our notation), Fixup initializes the last layer of $\Phi^l_{\boldsymbol{\theta}}(\cdot)$ at zero, and initializes the other layers by using standard random schemes; and then it multiplies a factor inversely proportional to depth L. This means that each $\Phi^l_{\boldsymbol{\theta}}(\cdot)$ at initialization is zero, hence the network *does start in the near-identity region*. Using this initialization scheme, their experiments demonstrate that ResNets can be stably trained **without batch normalization**, and trained networks match the **generalization** performance of the state-of-the-art models. The Fixup paper thus suggests that understanding optimization and generalization of ResNets in near-identity regions is a meaningful and important problem, thus further motivating our analysis in Section 5.

Q. Lastly, in line 119 it is claimed that augmenting one dimension [...] Can you please elaborate on this trick?

We admit that we didn't provide enough details about this trick. The trick requires some additional structures, e.g., the residual part $\Phi^l_{\boldsymbol{\theta}}$ must have an additional output dimension that always outputs 0. Due to space limits, we refer the reviewer to Remark 1 of [Shamir, 2018]; we will add more details as we revise our paper.

Q. Definition of $\mathrm{col}(\cdot)$?

The notation $\mathrm{col}(A)$ means the column space of a matrix A. We will add the definition in the next revision.