Response to Reviews on "Towards Understanding the Importance of Shortcut Connections in Residual Networks"

We appreciate reviewers' valuable comments. We will correct typos and reply to comments in the following.

To Reviewer #1:

Our analysis can be extended to more general cases. We will add more discussions in the revision.

• Extension to SGD: Our analysis can be extended to mini-batch SGD when the batch size is large. We can show its convergence by applying tools from the super-martingale theory, however, the analysis is more involved.

• Extension to deep networks: As mentioned in Lines 71-77, the weights in well-trained deep ResNet have small magnitudes. Thus, we expect that the shortcut prior assumption generally holds true for deep ResNets, and consequently the shortcut connection can analogously ease the optimization as in our two-layer ResNet model. Moreover, the partial dissipativity condition (PD for short, Definition 3,) provides a potential outlet to analyze deep ResNet. We will provide an empirical verification of the partial dissipativity condition for deep ResNets in the revision.

• Initialization and step size: For the first layer w, simply initializing w at 0 has been observed working well in deep ResNets (as mentioned in the previous paragraph, w tends to have a small magnitude). Thus, we believe that initializing w at origin can still work well in deep networks. For the second layer a, we use $O(1/\sqrt{k})$ type initialization. This coincides with common initialization techniques for deep networks (Glorot and Bengio, 2010; He et al., 2015; LeCun et al., 1998). Although deep networks undoubtedly need a more complex step size scheme, our analysis provides useful insights: For example, our choice of step sizes is consistent with the step size warmup scheme for deep ResNet (Goyal et al. 2017) as mentioned in Section 5.

To Reviewer #2:

Comment 1: Lemma 5 does not require $w_0 = 0$. We will remove it in the revision.

Comment 2: We briefly discuss the conditions for GD to be trapped in the spurious local optimum. Define the basins of attraction of the global optimum and spurious optimum as $\mathcal{R}_1 = \{(w,a)|a^\top a^* > 0, \phi = \angle(1/\sqrt{p}+w, 1/\sqrt{p}+w^*) < \pi/2\}$ and $\mathcal{R}_2 = \{(w,a)|a^\top a^* < 0, \phi = \angle(1/\sqrt{p}+w, 1/\sqrt{p}+w^*) > \pi/2\}$, respectively.

• Larger step size with $w = 0$: There exists a small constant $\epsilon > 0$, such that when $a_0^\top a^* < -\epsilon$, GD will be trapped in the spurious optimum. This is because $a_t^\top a^*$ takes a large amount of iterations to increase from negative to 0. Consequently, with a large step size, w can move far away from w^* before $a_t^\top a^*$ becomes nonnegative. This implies there exists T such that $(w_T, a_T) \in \mathcal{R}_2$, i.e., GD will be trapped in the spurious local optimum. On the other hand, if $a_0^\top a^* > -\epsilon$, $a_t^\top a^*$ becomes positive in a few iterations, and $\phi < \pi/2$ still holds. GD stays in $\mathcal{R}_1$, and converges to the global optimum. Note that the constant ϵ depends on w^* and is hard to characterize. We leave it for future investigation.

• Different initialization: If (w_0, a_0) falls in $\mathcal{R}_2$, GD will get trapped in the spurious optimum.

Comment 3: We appreciate reviewer's suggestion. Training the network in a layer-wise manner (setting learning rate of w to 0) is actually a special case considered in our analysis. We will add a discussion in the revision.

Comment 4: Our analysis applies to multiple spurious optima, as long as the partial dissipativity (PD) condition holds for the unique global optimum. For problems with multiple global optima, our analysis can still be applied if the following condition holds: there exists one global optimum such that the PD condition holds globally with respect to this optimum. In fact, we can empirically verify that PD condition holds globally with respect to the well trained model for some deep networks where multiple global optima exist. Thus, our analysis provides useful insights towards understanding multiple global optima cases. The theoretical analysis for multiple global optima cases needs a more detailed characterization of the landscape, which is technically difficult. We leave it for future investigation.

Comment 5: When we randomly initialize a in the ball, we have $a_0^\top a^* = O_p(1/k)$, where k is the dimension of a. This means a_0 has a larger chance to fall in the region where $a_0^\top a^* > -\epsilon$. As mentioned in the first item of our response to comment 2, GD initialized in this region will converge to the global optima. That is why we observe the increase of the success rate in our experiment. We further empirically observe that the success rate does not increase to 1 exponentially fast.

Comment 6: We show that using a small learning rate helps GD avoid the spurious optimum. In the two-layer ResNet, the spurious optimum yields bad generalization. We will make a clarification in the revision.

Comment 7: We thank the reviewer for this suggestion. We will revise our claim and make a clarification.

To Reviewer #3:

Our understanding of two-layer ResNet can provide useful insight towards understanding more general networks as we discussed in our response to Reviewer 1. The assumption $\|v\| = 1$ is used to stabilize the training, and ReLU is one of the most commonly used activation function. These assumptions eases our analysis but is not necessary. The key, partial dissipativity condition, is possible to hold for other networks. We remark that, even for this simple network, the analysis is already quite challenging.

We must emphasize that we are the FIRST to analyze the convergence of GD for two-layer ResNet, and there exists no other bound for comparison. We show that GD converge to the global optimum in polynomial time, but the degree of the polynomial may not be tight. In fact, the lower bound of GD is unknown and hard to characterize since GD can be trapped in the spurious optimum.