First of all, we would like to thank all reviewers for the insightful comments and suggestions! Reviewers have also raised many inspiring questions on asymmetric valleys (AVs), most of which we have addressed in this rebuttal. But for some of them (like what network structure or loss function tend to cause AVs, and what other new theoretical results could be obtained based on AVs), we may NOT have satisfying answers yet. However, this is perhaps one of the most valuable contributions of the paper – spawning new research problems and inspiring future research.

General Response

Significance and Novelty. Optimization landscape analysis is an important research topic in deep learning. To the best of our knowledge, this work for the first time introduces and formally defines AV. This goes beyond simply characterizing a local minimum by sharp or flat, which are popular terminology in the literature. The concept of AV leads to new results which may NOT be possible to derive based on existing terminology (see our next response).

What can be explained by AVs but not symmetric valleys (SVs). Here we give two examples (more details can be found in Sec 6.1 and 6.2): (1) Recent work [25,5,51] found that stochastic weight averaging (SWA) over iterations leads to HIGHER TRAINING LOSS but lower test error. If local minimum are SVs, then by simple concentration arguments, SWA should lead to LOWER TRAINING LOSS! In contrast, AVs gives a nice intuitive explanation for those interesting observations, and we have provided rigid theoretical analysis. (2) Recent work [12,43] observed that the local minimum of deep networks are well connected, meaning that a wide minimum and a sharp minimum could be in fact from the SAME basin. This seemingly contradictory observation can be well explained by AVs, but not SVs.

Are AVs prevalent? Yes. In our experiments, we can always find asymmetric directions at every local minimum that SGD finds, for all networks and datasets. To be conservative, we used the word "decent probability" in our paper.

Do AVs only appear in deep nets? What about 2-D loss surfaces? Apart from the SOTA networks stated in our paper, we also conduct experiments on a simple MLP in Appendix. Following Reviewer 2's suggestion, we also tried a 2D-MLP (1 single neuron with its weight, bias and sigmoid activation) experiment: data is drown from two 1D Gaussian distributions, forming a binary classification problem. It turns out that in such a simple case we can also find AVs (shown in Figure 1).

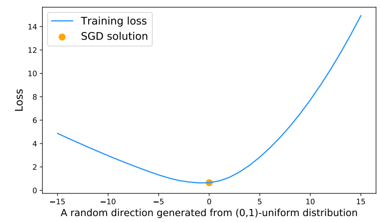

Figure 1: An AV in 2D-MLP

To Reviewer #1

Thanks a lot for your inspiring questions. As AV is a novel concept, we are not able to study *all* its specific properties in this work, and we do to have answers to several of your questions yet. But we believe that our work provides a new perspective to understand the loss landscape of deep networks, and may inspire many interesting future research topics.

SGD automatically avoids the training problem? In fact, SGD does not automatically lead to desired solutions for AVs, but averaged SGD does (Theorem 2).

AV and objective function. We believe that the structure of AV depends on the objective function. The whole story is quite complicated. But adding a L-2 penalty (as you suggested) seems to have little effect on AVs: we could also find AVs, and averaged SGD still gives better performance. Studying the relation between AV and objective function (and network architectures) is an interesting future research direction.

What leads to AVs? The reason of why AV exist is not fully understand yet, but we believe batch normalization is one of the important reasons (please refer to Appendix H). We leave it as the future work.

To Reviewer #2

Motivation and contribution? Our result is important because the notion of AVs can be used for explaining many interesting observations (e.g., [25,5,51]) which can not be well explained by existing concepts. Please refer to the "General Response" above.

Flatness includes AVs? This is true if we still simply characterize both asymmetric and symmetric valleys by flatness, without differentiating SVs and AVs. However, without introducing AVs, we will not be able to obtain the theoretical results on bias and generalization, and those observations made by [25,5,51] cannot be well explained.

To Reviewer #3

Bias and SVs? In fact, bias is good for AVs, but not for SVs. Fortunately, SGD averaging automatically generate bias for AVs (Theorem 2) and less bias for SVs (following a simple concentration argument). As our paper is on AVs, we dismissed the discussion on SVs. **Learning rate.** We follow all the hyperparameter configurations used in [25].

Are SGD and SWA end up in the same neighborhood? As SWA is the averaged solution of SGD iterates, they are located in the same neighborhood under mild conditions. Therefore, our theoretical generalization guarantee can be ensured. In our experiment, we further run SGD from SWA because we want to find a solution that clearly has lower training loss than SWA, but has a higher test loss, thus validating our theoretical results. Also notice that empirically SWA still generalizes better than SGD even when they are not in the same local basin (see e.g. [25]).