Thanks to all the reviewers for their thorough feedback and valuable suggestions.

Reviewer 1

We will add more intuitions and pictures to make the proofs in our appendix clearer.

Reviewer 2

Typos: We will fix the typos in Definitions 1 (i.e. $h_j/2$ for layer j) and 3 (i.e. largest number instead of smallest).

Justifying our stability assumptions: For neural networks trained with standard algorithms, noise stability was previously observed in [Marcos et al. 2018, Arora et al. 2018]. While there is no formal proof showing any optimization algorithm must find a noise stable solution, there is evidence that the solutions found are indeed noise stable. Our own experiments also show that the paths constructed using these properties can indeed connect two different solutions.

Purpose of θ_2 through θ_{d-1}: Our path construction for Theorem 1 passes through each θ_i, not just θ_1. For example, in Figure 1 Step (2) uses θ_2 and Step (4) uses θ_1. We describe how to connect each θ_i to θ_{i-1} in our proof of Lemma 1. We will add some helpful figures in the revised version that should hopefully clarify to our path construction.

Properties in Definition 2: It is better to have larger μ. This definition is exactly the same as in [Arora et al. 2018], due to space limitations we chose to focus on properties that are slightly different.

Definition 4, question about activation contraction c: Main confusion here is that the noise stability quantities are defined *specifically* for ReLU networks. For ReLU activations, the constant c is mostly related to the fraction of neurons that are active. The definitions would need to be different for other activations. While there are a lot of terms in Definition 4, the overall message is that larger layer cushion and larger interlayer cushion leads to better noise stability (i.e. smaller ϵ). We also note that the theory is asymptotic, and that the numbers computed in a real network might be loose; however our experimental results suggest that the theory does indeed lead to paths connecting different solutions.

Explaining interlayer smoothness, "cirular" argument: At a high level, interlayer smoothness assumes the network is close to its *linear* approximation, but what we need to prove is that the network output is nearly *constant*. A large class of functions have good linear approximations, but not all linear functions are constants. Only in combination with our other noise stability assumptions can we show that the dropped out network has similar loss as the original network.

Experiments on CNNs: As explained in Remark 1, our dropout-based path constructions naturally extend to convolutional nets since we can think of each channel as one hidden unit. All our noise stability properties also apply to convolutional nets (as in [Arora et al. 2018]). We chose to show experiments with convolutional nets because these architectures are widespread and of practical interest.

Figure 2: Figure 2 is meant to construct a path based on Theorem 3 (instead of Theorem 1 which requires dropping-out 1/2 the units). In Theorem 3 we show that if there exists a small network with low loss, then one only needs to drop out a smaller fraction of the units in each hidden layer. The left plot shows the error under different levels of dropout; the middle plot shows that connecting a network with its dropout version with $p = 0.2$ has almost constant loss; the right plot shows that there exists a neural network whose size is 0.2 times the original network that has low loss. Combining these three plots with Theorem 3, we know that there exists a path with almost constant loss. We also note that data in Figure 2 are from networks NOT trained with dropout. We will clarify these points in the final version.

Comparison to interpolation: Interpolating results in substantially higher loss than our path construction (loss/accuracy 1.61/67.2% on MNIST and 2.34/10% on CIFAR). We will add a plot in the revision. Thank you for the suggestion!

Reviewer 3

Typos: We will fix the typos in Definition 1 (same one pointed out by Reviewer 2), Lemma 2 and the construction of our counterexample. Thanks for pointing these out!

What we get from interlayer smoothness: Interlayer smoothness does not in and of itself ensure that the output of the network is stable as we discuss in our response to Reviewer 2. Requiring the property to hold for all t is crucial for us to be able to directly interpolate between the original network and the fully dropped out version using a single linear segment. Note however that even if we only assume the property holds at the endpoint of this path we would still be able to connect the original network to its dropped out version, but doing so would come at the cost of needing more segments to construct the path as we did in Lemma 1.

Figure 3: The leftmost plot in Figure 3 is simply meant to give readers a sense for what the values for each of the different components that comprise our ultimate definition of noise stability tend to look like in practice: the quantities appearing in the denominator of noise stability are reasonable constants bounded away from zero, and conversely those appearing in the numerator are not too large. It is true that the bounds are asymptotic and directly computing the bound may not give a good number on real data, but our experiments show that the paths constructed are reasonable (although they are not the same as the paths constructed in [Garipov et al. 2018, Draxler et al. 2018]).