Response to Reviewer $\#1$: Thank you for the careful reading and feedback. In a revision, we will address the detailed comments:

1. We will clarify the discussion around Lemma 3 to reflect that activation regions and linear regions are typically identified in prior literature but are in fact not quite the same.
2. We will add intuitive explanations of the +/- 1s and activation patterns in the definition of activation regions.
3. Around (3), we will clarify that $\text{activationregions}(\mathcal{N}, \theta)$ is the set of all non-empty activation regions for the network $\mathcal{N}$ with trainable parameters θ.
4. We will emphasize in line 39 that T is constant.
5. We will restate conditions 1,2 in terms of continuous random variables.
6. We will correct the reference to (5) in line 177.
7. We will sharpen the intuition in Section 3.2 to reflect that $|z'(x)| = O(1)$ guarantees $O(1)$ high amplitude oscillations for $z(x)$ when x varies over a fixed bounded interval.
8. We will make it clearer that this terminology is deliberately vague - the terms F_{learn} etc are only referenced in the caption to Figure 1.
9. We are not sure what work from ICML 2019 the reviewer has in mind. In case Hanin and Rolnick was meant here, we do make a point of citing it as [14].

Response to Reviewer $\#2$: Thank you for the careful reading and feedback. About point 3 in the reviewer's list of three contributions: we found the fact that networks cannot learn many activation regions to be a surprising counterpoint to the well-known ability of networks to memorize high-dimensional noise. We plan to amplify this point in the revision. In the revision, we will also address the reviewer's detailed comments:

1. We agree that more intuition can be helpful and plan to add more (see points 1,2,7 in our response to Reviewer $\#1$).
2. We will give a more thorough discussion of the constants C_{grad}, C_{bias}. Previous work [11,12] shows that C_{grad} is like d/n only at init, and hence can in principle grow through training, as the reviewer suggests. It is not clear how to rule this out *a priori*.
3. About Lemma 6, we agree that our discussion could be clarified and will write simply that "we conjecture" that the inhomogeneous scaling of biases does not strongly affect the number of regions.
4. We agree that a discussion of which architectures have large C_{grad} is warranted. We will explain that prior work [11,12,13] shows that unless C_{grad} is small, fully connected ReLU nets have unstable forward and backward passes at init. Thus, for such networks, as long as they are trainable, C_{grad} will not be too large. This is the reason we used terms like "depth-independent", and we will amplify this point.

Response to Reviewer $\#3$: Thank you for the careful reading and feedback. About the reviewer's comment that some of our experiments could be seen as illustrations rather than empirical evidence: we will emphasize in the revision that, indeed, at init, they are simply illustrations of our results. However, after init, it is not clear how C_{grad}, C_{bias} behave and hence empirical validation that our results apply is provided by these experiments. In the revision, we will also address the reviewer's detailed comments:

1A. We agree that Definitions 1-2 and Lemmas 1-4 are elementary, and their purpose is primarily for clarity in exposition. Moreover, we wanted to give a clear delineation between linear and activation regions, which have often been conflated in prior work.

2A. We agree that the potential dependence of C_{grad} on depth needs to be discussed. See points $2, 4$ in our response to Reviewer $\#2$.

About the reviewer's suggestions on how to improve:

1B. Our results show both theoretically and empirically that not only *can* the number of regions be small but that it typically *is* small both at init and throughout training. We believe this is an important point and will emphasize it in the revision.

2B. In the revision we will emphasize that although our results do not directly influence architecture selection, they make more clear the role of depth and hence suggest to practitioners the intuition that network depth is mainly useful for optimization and not for expressivity.

3B. See point 2A above and point 4 in our response to Reviewer $\#2$.