We thank all the reviewers for their constructive comments. The following are our point-by-point replies.

Rev. 1.

1) *this is only for 1HL networks but the Theorems do not specify this*: The presentation in the submission might cause confusion. While Theorems 5, 9 and 10 consider the three-layer networks for simplicity, their extension to L-layer networks is easy by replacing the derivatives of the loss with the back-propagated delta. We briefly mentioned this at the beginning of Sec. 3.2. But, we will make a clearer argument in an update.

2) *No definition of saddle and local minimum points*: We will include their standard definitions in Sec. 2. See also 13) for semi-flat minima.

3) *Strict or non-strict saddles and what order?*: It is easy to derive the index of the saddle point based on the Hessian in Lemma 4, but the details were skipped by space limitation. If G is positive (or negative) definite and the off-diagonal F is of full rank, the index is easily given by Eq.(37) in Supplements. In the other cases, the index may not be explicit, depending on the eigenspaces of $\tilde{F}$ and $\tilde{G}$. We will add a brief comment and show the details in Supplements.

4) *Implications to learning with GD*: We totally agree that this is an important topic, and are currently working on it.

5) *Saddle by inactive units*: Theorems 2 and 4 in [4] consider a special parameter, which corresponds to $\boldsymbol{v}_{H_0} = 0$ in the current paper. For general $\boldsymbol{v}_{H_0}$, they are not critical points, and this fact is mentioned in lines 111-112 with full description in Supplements. If $\boldsymbol{v}_{H_0} = 0$, we can show similar results to [4]. In an update, we will clarify this at the paragraph in line 111.

6) *Embedding of a saddle point*: In the unit replication, Lemma 4 tells that the Hessian of the narrower network is recovered in the wider one. So, embedding of a saddle point with positive and negative eigenvalues gives saddle points. For the embedding by inactive units and propagation, a critical point is not in general embedded into a critical point.

7) *Refer Section 2 in Introduction*: We will reflect this.

Rev. 2.

8) *Embeddings are artificial*: We do not think the three methods are artificial for the reason described in the paragraph, line 88-93. Overparameterization is an important issue in this field as we describe in Sec. 1, and it refers to the situation where a network has more sizes than the one necessary to realize a function. We rigorously formulate and discuss this situation in the paper. It is also important to note that, as discussed in lines 88-93, the classical results [10,14] already proved that the three methods of embedding are the ONLY ways of realizing a network function by a wider network in the case of three-layer models. These embeddings are thus the essential ones in discussing overparameterization.

9) *Embedding methods are not clear in Theorems*: Theorem 5 does specify the embedding by $\theta_\lambda^{(H)}$, which is defined as a symbol of the Unit Replication in Eq.(4). Additionally, Sec. 3.2 starts with a statement that we consider this type. Theorem 9 clearly states that the embedded point is "defined by Eq(10)", which is the definition of Inactive Units (line 173). In an update, we will explicitly place the name of embedding at the theorems to avoid confusion.

10) *Difference of considerations for the smooth and ReLU activation*: For smooth activation, the inactive units and propagation do not embed a critical point to a critical one in general (lines 111-112), so we discuss only the unit replication for local minima. For ReLU, note that the definition of inactive units is different from that of smooth activation, and discussing an embedding with inactive units is meaningful. Inactive propagation is the same as the smooth case, and a critical point does not give a critical point in general. We will clarify this in an update.

11) *Different choices of P and Q in PAC-Bayes*: Using the same P and Q does not necessarily give a fair comparison, but they should be chosen so that the bounds be tight. The distributions of the parameters that give similar loss values are different, and the choice of the posterior Q must reflect this difference for meaningful bounds. The prior P is arbitrary, as long as it does not depend on the training data. The choices in the paper reflect these conditions.

12) *Inconsistency between experiment and theory*: Sec. 5 highlights the difference between the smooth and ReLU models, and the form Eq.(11) holds also in 1-dimensional output (see [4]), which means the difference exits also. While the theory in Sec.5 are based on Theorem 5 or Lemma 4, the experiments use 1-dimensional network for simplicity of realizing zero error. We will make clearer explanations on this point in the paragraph on the experiments.

Rev. 3.

13) *Definition of semi-flatness*: In Sec. 1, we say "semi-flat minima, at which a lower dimensional affine subset in the parameter space gives a constant value of error", and use it as a definition. We will make it clearer. Flat-minima often refer to points at which all the directions are flat. "Semi-flat" allows the case that only part of the directions are flat.

14) *PAC-Bayes analysis already exits*: Our motivation in Sec. 5 is to compare the smooth activation and ReLU in overparaemterized cases. We will clarify this point better and cite the reference.

15) *How does the number of surplus units affect the landscape?*: Thank you for pointing out this important question. We must admit that the detailed discussions were just left in Supplements by space limitation. For smooth networks, some of the surplus parameters make sharp directions so the bound is linearly increasing to the surplus number of units, while in ReLU there are no sharp directions so that no such major increasing terms exist in the bounds. This is illustrated in the experimental results (Fig.2(b)). We will emphasize this in an update.