We thank all reviewers for their valuable feedback. Below we address each reviewer's questions and concerns.

—— Response to Reviewer 1 ——

[The theoretical results are "not complete" because only two-layer fully-connected networks are considered.] First of all, we'd like to remind the reviewer that many interesting deep learning theory papers focus on two-layer networks. There are *dozens or even hundreds* of such papers published in top conferences and journals. Just for example, [15,16,17,20,25,34,35,36,43] in our paper's references are all theoretical studies of two-layer networks. Second, it's already challenging to formally establish the two-layer results in our paper; we believe that the conceptual and technical contributions of the paper are interesting and insightful enough to meet the standard of publication, as acknowledged by all other reviewers. Third, we didn't claim we have rigorous results for multi-layer nets, but only provided empirical results and partial theory to support them, and we explicitly mentioned formally proving them as a future work direction in Section 5. For these reasons, we find it incorrect and unfair to call our theoretical results "not complete" and to recommend weak reject mainly based on this. We hope the reviewer can reconsider their decision.

["The experimental results are not sufficient. This work only conducts experiments on single data set, and the experimental results could not fully verify the theoretical results."] We do not understand this comment. We provided experimental results on synthetic data, CIFAR-10, as well as MNIST in the supplementary. These are more than a "single data set." Also, our experimental results match theoretical predictions very well, so we do not understand why they "could not fully verify the theoretical results." Note that all other reviewers find our experiments convincing.

[Related work not sufficient.] We do not understand what's the reviewer's specific concern about related work. We have tried to provide a thorough discussion of related work, and all other reviewers think our discussion is sufficient. If the reviewer can point out exactly what they think is missing in the discussion, we are happy to incorporate it in the paper.

["The part of the theoretical results is not written well, since the experimental results appear among them."] Thank you for the feedback about writing. In fact, we separated the experiments in two sections *on purpose*: Section 3.4 is to verify the two-layer results in Section 3, and Section 4.2 is for multi-layer and convolutional nets. We thought this is the clearest way to present our results, and all other reviewers think our paper is well-written (in particular, R2 finds it "a quite enjoyable read"). Also, we did separate theory and experiments in different subsections so that they are not really muddled together. Regarding the reviewer's comment that "theoretical results are not written well", we do not understand the reasoning that our arrangement of experiments affects whether the theoretical results are written well.

—— Response to Reviewer 2 ——

[What if $\alpha > \frac{1}{4}$?] If $\alpha > \frac{1}{4}$, the conditions are still satisfied with $\alpha = \frac{1}{4}$, so our result still applies, which means the network still behaves like a linear model early in training. It is indeed an interesting question to characterize what happens after the linear learning period, possibly related to some higher-order kernel as the reviewer points out.

—— Response to Reviewer 3 ——

[Is Assumption 3.1 approximately satisfied by real-world data? Can we transform a given dataset so that Assumption 3.1 holds?] We think it's possible that Assumption 3.1 is approximately satisfied by certain datasets, but not all. Yes, it's possible to transform the data to move closer to this assumption. For example, for CIFAR-10 and MNIST we applied standard pre-processing to normalize each image to have zero mean and fixed norm (lines 468-469). This already enables the linear learning behavior to hold. We believe whitening the data can make the assumption better satisfied.

[About experiments: (1) How is the linear model calculated? (2) How long does the early phase last if we use a large learning rate?] (1) For two-layer NN experiments, the linear model is exactly calculated using Eqn. (11). For multi-layer NNs, since we use erf activation in the experiments, it's easy to show that the corresponding linear model is $x \mapsto cx$ for some constant c (without bias or the norm-dependent feature). We estimate c by $c^2 \approx \frac{\lambda_{\max}(\mathrm{NTK})}{\lambda_{\max}(XX^\top/d)}$, since we expect $\mathrm{NTK} \approx c^2 XX^\top/d$. In general we can use the method sketched in Section 4.1 to compute the linear model analytically. (2) If the learning rate is 5 times larger, the number of iterations of the agreement will be 5 times smaller. Note that the correct way to think about this should be progress of learning rather than specific number of iterations. That is, the shapes of the learning curves will be the same regardless of learning rate (as long as training doesn't diverge), i.e., the agreement will last until the linear function finishes learning.

—— Response to Reviewer 4 ——

[What does "we show that these common perceptions can be completely false" mean?] We simply meant that the network mimics a linear model and doesn't use its nonlinear capacity early in training, which is a rephrasing of our main result. We will try to modify the sentence to make it clearer. Thanks for pointing out the confusion.

[Any intuition/reason why the networks eventually escape the linear behavior?] The network has the capacity to express complex nonlinear functions, so it should not be a surprise that it eventually becomes nonlinear. [Extension to deep networks?] We discussed extension to deep networks in Section 4. [In Section 3.4, what if ReLU was used instead of ERF?] We indeed provided experiment on ReLU in Figure 2. The corresponding linear model is given by Eqn. (11).