We thank all reviewers for the insightful and encouraging comments. Below we provide a point by point response to Reviewers 1,2,3 (R1, R2, R3).

R1 + R2: [Difference from "lazy training"]. One key contribution of our work is that we identify constancy of tangent kernel, first observed in Jacot et al. (2018), as related to the scaling of the Hessian norm. Notice that the constancy of the tangent kernel cannot be explained from the point of view of "lazy training": when the last layer is non-linear, the change of parameter from the initialization is of the same order as for the linear case, but the tangent kernel is no longer constant along the optimization path, as the Hessian norm is no longer small (see Section 4).

R1: *Theorem 3.2 and results in Appendix G has been proved previously (e.g., [1]). ... I tend to think of this paper as a summarization of the previous line of NTK papers. ([1] Sanjeev Arora, et al. On Exact Computation...*)

Our Hessian analysis results, including Theorem 3.2 and Theorem 3.1, are new. Note that previous works, including [1], only analyze the tangent kernel matrix, which is first order. In contrast, we analyze the Hessian, a second order differential operator. We note that in some related works, including [1], the notation H stands for the tangent kernel, while we use it to denote the Hessian matrix. This can perhaps cause confusion. Our novel contributions include identifying the underlying reasons for constancy of NTK (small Hessian norm, as opposed to "lazy training"), and the finding that NTK is **not constant** when the last layer is non-linear, even in the infinite width limit (Section 4).

R2: *The paper mostly focuses on squared loss while widely applied NNs use softmax-cross entropy loss. I would encourage putting discussion on ... optimization of those networks in the context of this paper's result.*

Thank you for the suggestion. The main focus in this submission is to uncover the underlying reasons for the constancy of NTK (which depends only on the model, rather than the loss function). Still, it is an important issue, we will add a discussion.

R2: *It (supplementary B) does have significant overlap with current submission ... may be subject to dual submission ... or just cite a separate paper distinguishing contribution.*

There is no dual submission issue as the supplementary B has not been submitted to NeurIPS or any other conference/journal. Given the space constraints, it does not seem feasible to have a full discussion of the optimization-related issues. For the final version, we are planning to cite the optimization results as a separate document.

R2: *What is the important point of emphasizing Euclidean norm change is O(1)? ... Should I understand the point to be while literature casually talks about "small weight change", one should be aware that in Euclidean norm it could be O(1) due to large dimensions?*

Yes. The measurement of the change of parameters from initialization depends on the norm. In dimension m, the Euclidean norm and infinity-norm can be different by a factor of $\sqrt{m}$. When dimension increases, the infinity norm of the difference from the initialization to the solution converges to zero. However, the Euclidean norm of the difference is always $O(1)$. Importantly, the remainder term of the Taylor expansion (and hence the constancy of TK) is controlled by the Euclidean norm of the difference, not the infinity norm. In contrast, "lazy training" suggests that the optimization path stays close to the initialization point. We will clarify this in the paper.

R2: *Is the condition (b) in Theorem 3.1 violated for these non-linearities (softmax, maxout) and does not have a constancy guarantee? Do authors believe these networks would not have constant tangent kernels*

This is an open question so far. These non-linearities do not fit in our current analysis. If softmax is in the output layer, our analysis in Section 4 shows that the model does not have a constant tangent kernel.

R2: *L251: logically violation of conditions in Theorem 3.1 does not necessarily lead to breakdown of linearity since Theorem 3.1 is not if and only if statement, correct?*

Yes. The conditions in Thm 3.1 are sufficient but not necessary. But note that the neural networks that are shown to have constant NTK satisfy these conditions. Hence, Thm 3.1 is enough to explain the phenomenon of constant NTK.

R3: Thank you for the positive comments and the helpful suggestions.

R3: *Could the authors comment on the applicability of their results assuming a hinge loss? It seems that the hinge loss ought to satisfy conditions in Eq(13) and Eq(14), given that its second derivative is zero almost everywhere.*

First, note that the Hessian is defined for the model, not the loss function. For hinge-loss-like activation functions, the Hessian is zero at most locations, but it is infinite at the hinge point. We note that the Hessian affects the linearity of the model (i.e., constancy of tangent kernel) through Taylor expansion, see Proposition 2.1 and its proof. Hence, the tangent kernel, which is the first-derivative, depends on the integral of the Hessian, which is second-derivative. This infinite Hessian for hinge-like activation functions is likely to have a non-trivial contribution to the tangent kernel after integration, implying non-constant tangent kernel.