We thank the reviewers for their time and constructive feedback on the submission, which we will incorporate to improve our manuscript.

R1

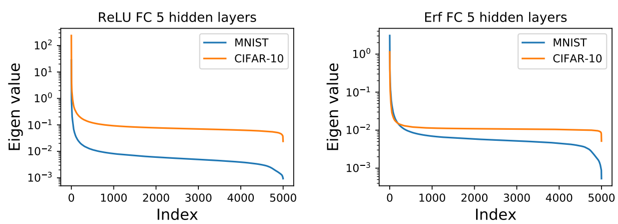
- We will define $\hat{\Theta}^{(n)}$ for clarity. Indeed it denotes an empirical tangent kernel for width n network.

- Line 118: equation reference should point to Eq. 8-11 instead of Eq. 2-3.

- Figure 4: The primary reason why the dashed lines are hard to see is that the linearized network's training dynamics match those of the nonlinear model so closely. Having said this, we agree that we should improve the clarity of the plots and will include modified versions upon revision.

R2

- Rank of tangent kernel. We observe that Θ is full rank. See line 515 in the Supplementary Material and Proposition 2 in (Jacot et al. 2018): under mild assumptions (all inputs have the same norm with non-polynomial activation functions), the kernel is positive definite. To confirm that the kernel is full-rank in practice, we can generate kernels and look at their spectrum. The following plots show the spectrum of the NTK of two five-layer fully-connected models on CIFAR-10 and MNIST. We find that they are positive-definite as expected.



- Derivation of (S84): Thank you for bringing the lack of clarity here to our attention. One source of confusion may stem from typos in (S81) which should read $\lambda_{min}(A) \geq \sqrt{N} - \sqrt{n} - t, \lambda_{max}(A) \leq \sqrt{N} + \sqrt{n} + t$. With this, we now describe how (S84) is obtained. For simplicity, let us assume $\sigma_w = 1$ and that $2 \leq l \leq L$ (as arguments for $l = 1$ and $l = L + 1$ are similar). For $2 \leq l \leq L$, when $\theta = \theta_0$ (i.e. at random initialization), W_l are $n \times n$ random Gaussian matrices, so with high probability (Thm G3), $\|W_l\|_{op} \leq (2 + 0.5)$. For any $\theta \in B(\theta_0, C/\sqrt{n})$, by the triangle inequality, the operator norm of W_l is bounded above by $(2 + 0.5 + C/\sqrt{n}) < 3$ with high probability. We have applied the fact that the operator norm of ΔW_l is bounded by its Frobenius norm, which is at most $C/\sqrt{n}$.

- We appreciate your identification of typos, which we will fix upon revision.

R4

- Relation to "Lazy Training": Thank you for bringing this oversight to our attention. *"A Note on Lazy Training in Supervised Differentiable Programming"* by Chizat and Bach is an important contribution and we will absolutely include a discussion of it in relation to our own work upon revision. While there is overlap with our own submission we would like to emphasize that the work of Chizat and Bach was concurrent with our own paper (similar preprint release dates). Additionally, the initial versions (arXiv V1, V2) of that work only performed experiments on one-hidden-layer networks and some of their results (e.g. Sec 2.2 in V1, V2) are restricted to single-hidden-layer networks.

- Applicability to modern networks: As the reviewer points out, some layers of modern networks may be operating far from the linearized regime. However, in many situations increasing the size of networks can lead to improved performance (e.g. EfficientNet, XLNet). If this trend continues to be monotonic in width, the infinite-width limit might indeed be relevant for well-performing architectures. In Figure 1 of (Novak & Xiao et al. 2019; `https://arxiv.org/abs/1810.05148`), it is shown that the comparison of performance between finite- and infinite-width networks is highly architecture-dependent. In particular, it was found that infinite-width networks perform as well as or better than their finite-width counterparts for many fully-connected or locally-connected architectures. Similarly, in Table 1 of (Arora et al, 2019; `https://arxiv.org/pdf/1904.11955.pdf`) NTK kernels outperform the corresponding finite width CNNs in 5 out of 10 configurations (though the best overall performance is achieved by a finite width CNN). It is still an open research question to determine what are the main factors that determine these performance gaps. In any case, we believe that examining the behavior of infinitely wide networks provides a strong basis from which to build up a systematic understanding of finite-width networks (and/or networks trained with large learning rates).

- Implications of "gradient descent doesn't generate samples from a probabilistic model": Very briefly: Infinitely-wide neural networks open up ways to study deep neural networks both under fully Bayesian training through the Gaussian Process correspondence, and under Gradient Descent (GD) training through the NTK or the linearization perspective. The resulting distributions over functions are inconsistent (the distribution resulting from GD training does not generally correspond to a Bayesian posterior). We believe understanding the biases over learned functions induced by different training schemes and architectures is a fascinating avenue for future work. We will expand discussion around this.