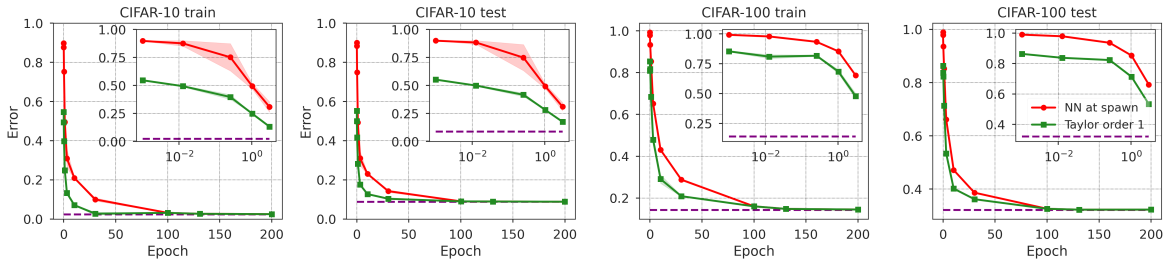

Taylorized training with errorbars. Insets zoom in on early period of training.

Dear reviewers, thank you for your comments! We are happy to see that our empirical findings provide rich insight and suggestions for further developing the theory of loss landscape and NTK (R2), that our metrics are numerous (R1), our empirical analysis extensive (R6), and our results interesting and likely to contribute to the discussion in this field (R6).

R3 : To produce uncertainty estimates, we have now re-run Taylorized training experiments multiple times and **added standard deviation error bars, which turn out to be very small**. An example is shown in the Figure attached.

R1 : We **added training epochs until convergence at epoch 200**. NTK (Taylor 1) outperforms full network training in terms of training error, when both the full network and the NTK are trained up to 200 epochs. However, we found that this lower training error for the NTK comes at the expense of overfitting, resulting in higher test error for the NTK. Therefore, in our new analysis up to 200 epochs, we terminated NTK training based on the optimal **early stopping time as determined by the test error of NTK**. Under this reasonable condition, we found that both NTK training and test performance (green lines) almost matches that of the full network's test error at 200 epochs (dashed purple line), using data-dependent NTK kernels that were created from the full network at 30 epochs for CIFAR-10 and 100 epochs for CIFAR-100 (see Figure - compare green curves to dashed purple line). Notably, the NTK that is used at initialization, and all fractional epochs less than 1, cannot be used to obtain test errors anywhere close to that of the full network at 200 epochs. However, within 3 epochs, we obtain an NTK that does significantly better than the one at initialization.

R1 : In terms of conclusions, these results provide important information to the field, namely that: (1) **NTK kernels at initialization perform quite poorly**, but (2) **such kernels created after very few epochs of training perform substantially better than at initialization**, and finally, **within 30-90 epochs of training these kernels have enough information about the data in them to closely match the test error of the full network** at 200 epochs of full training. We believe this first analysis of the dynamics of a data-dependent NTK kernel and its relation to landscape geometry will be very useful for understanding relations between deep learning, NTK kernels, and geometry. Especially important is the rapid improvement of a data-dependent kernel relative to initialization with less than 2 epochs of training (compare green points at 2 epochs to green points at 0 epochs in the insets).

R3 : To falsify the hypothesis that children spawned late are in the same linearly connected mode because they don't travel as far after being spawned, we measured the distance traveled in both cases. In Figure 5 of our paper, the L2 distance travelled from parent to trained child is **58 for spawn at init** and **43 for spawn at e=30**. While $58 > 43$, this difference is not large enough to account for the effect. Therefore a **small distance travelled is likely not the explanation** for the effect we see. Moreover, if we train children for a full T epochs after spawn time t_s, rather than $T - t_s$, we see similar effects. Therefore we conclude that it is the effect of the spawning time, not the length of the subsequent optimization, that keeps late spawned children in the same basin.

R6 : *It sounds highly non-trivial that the green point always out-performs red one even at epochs very close to 100*: The final accuracy of the Taylorized regime training depends on a) at what epoch you spawn it = develop the Taylorized approximation, b) how well the Taylorized training works, and c) the existence of good optima within the Taylorized approximation. While it is not a priori obvious, we found the green curve (error of trained Taylorized model) to be lower than the red curve (original NN) at 100. However, at (newly added) epoch 200, this was still the case for train but not test – training error went down, but overfitted and caused test error to go above the red line. This might be due to the difficulty of training the Taylorized regime and our sub-optimal choice of hyperparameters. The Taylorized model spawned at epoch e gets additional training compared to the NN there, which is why the green curve can even *in principle* be lower than the red curve.

R1 : We focused on image datasets as they are still a very important and prominent part of ML research. While we show results for CIFAR-10 and CIFAR-100, we also ran experiments on MNIST, Fashion MNIST and SVHN, with equivalent results. Due to the computational requirements of our experimental sweeps and Taylorized training in general, we couldn't extend our analysis to ImageNet scale. However, we did use powerful models such as WideResNet and SOTA training schedules and HP choices to make sure we can get as close to real settings as possible. We train to comparable train and test error (see He et al.'15 for ResNet20 and Zagoruyko et al. '16 for WRN).