The final accuracy of the Taylorized regime training depends on (a) at what epoch you spawn it, (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.

Our 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).