A Additional Related Work

Nonlinear Random Matrix Theory and Random Feature Regression. The limiting spectrum of CK with random input dataset has been investigated by [73, 13]; whereas [57, 30] studied the spectrum of CK with more general input data. This spectrum is actually a deformed Marčenko–Pastur distribution [30], which becomes a deformed semicircular distribution [81] when $h \gg n$. The largest eigenvalue of the CK matrix has been studied in [14], and the spectrum of the NTK was analyzed in [1, 30]. As an application, random feature ridge regression was fully determined by the limiting spectra of CK or NTK: [57, 33, 53, 61, 1]. All of these results belong to LWR.

Global Convergence of GD for Ultra Wide NNs. A recent line of work has shown the global convergences of the learning dynamics of gradient-based methods in a certain overparameterized regime, e.g. [28, 27, 69, 70, 66, 54, 75, 18]. We refer to Table 1 of [75] as a summary of these recent results. Most of the theorems in the literature require $h \gg n$, which implies that the NTK is almost static during training, while [70, 66] can consider LWR under some specific assumptions. Recently, [18] established a new criterion for the convergence of GD which results in the global convergence of general NNs with finite width h and $d \geq n$.

Beyond NTK Regime. Under the proportional limit, the initial kernel regression can only learn a linear component of the target [35]. Thus, it is reasonable to consider the cases beyond the NTK regime. To this end, [29, 39] considered the dynamics of NTK throughout training while [2, 7] have shown a second-order approximation of NTK, outperforming the initial kernel. In addition, there are many theoretical works analyzing when a NN outperforms the initial kernels in some specific settings: [51] proved a two-layer ReLU NN that is shown to beat any kernel method; [42] verified a two-layer CNN with some simple dataset can outperform the initial NTK for image classifications; [6] showed a NN can escape the kernel regime by only taking one specific large gradient step; [24] showed a specific gradient-based training can even learn polynomials with low-dimensional latent representation.

Evolution of NTK and Alignment in NNs. The feature learning can be characterized by the evolution of the kernel during training [31, 68, 55, 5, 56]. Specifically, [55] studied the hard-margin SVM for "after kernels" which are the CK and NTK matrices of trained NNs. One of the effective ways of depicting how the kernels evolve during training is to capture the evolution of kernel alignment [10, 78, 5, 56]. Kernel alignments between kernels and training labels essentially reveal how the NN accelerates training [78]. Also, several papers showed that the top eigenfunctions of the kernel align with the target function learned by the NN [44, 67, 68]. This becomes an efficient way of analyzing how NNs learn features through a particular gradient-based optimization.

Large Learning Rate Regime. As mentioned earlier, the large learning rate may contribute to feature learning. The benefits of large-learning-rate training have been studied from different aspects [52, 64, 15, 3]. Specifically, [46] observed that training dynamics with large learning rates differ from the small learning rate regime, where the latter regime exhibits monotone and fast convergence of training loss but may not generalize well on test data. At the early phase of training, [41] showed using lower learning rates may result in finding a region of the loss surface with worse conditioning of kernel and Hessian matrices. In [55], the after kernels of NNs trained with larger learning rates generalize better and stay more stable. [49] raised a "catapult mechanism", where gradient descent dynamics converge to flatter minima for extremely large learning rates. There is a transition as a function of the learning rate, from lazy training to the catapult regime. Section 4.2 illustrates a similar transition in our situations.

Heavy-tailed Phenomenon. The heavy-tailed phenomenon has appeared in many places in deep learning theory; [58, 59] observed that many state-of-the-art pre-trained models obtain heavy-tailed weight spectra. More precisely, these spectra have a "5+1" phase transition which relates to different degrees of regularization of the NN. With this heavy-tailed self-regularization theory, [60] further showed how to distinguish well-trained and poorly trained models by a power-law-based approximation. [63] classified trained weight spectra into three types: Marčenko–Pastur law, bulk with (few) outliers, and heavy-tailed spectra. We extend this classification to both weight and kernel matrices in Figure 1. Additionally, similarly to the discussion in 4.3, [63] showed that the difficulty of the classification problem is related to the emergence of heavy-tailed spectra in weight matrices. This

heavy-tailed phenomenon can be used to construct metrics for evaluating the generalization of NNs [60, 86], and early stopping of NNs to avoid over-fitting [63].

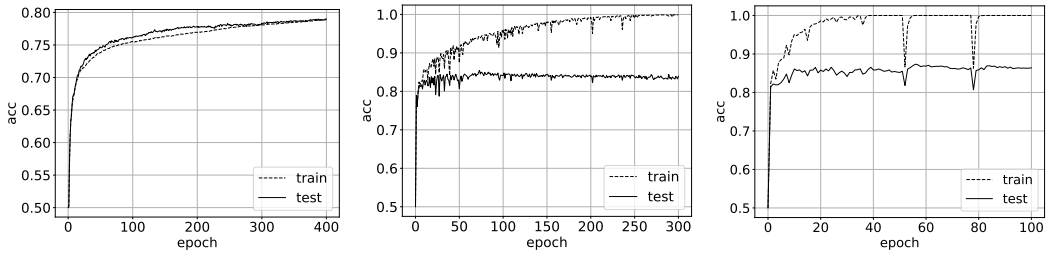
B Additional Empirical Results

There are different parameterizations for NNs at initialization. The orders of the output of NN are distinct in different cases [21, 25, 85]. This affects the size of stable and non-trivial gradient steps. The distance of trainable parameters from initialization determines whether the NN learns any features from the training data [6, Figure 2]. The performance of networks with different initializations indicates whether the NN belongs to the kernel regime or not [85]. Unlike the NTK parameterization, the *mean-field* parameterization [62, 20] and *maximal update* parameterization [85] tend to be feature learning.

For all NNs in our experiments, we apply a normalized and centered nonlinear activation function such that Assumption 3.2 holds ($\mathbb{E}[\sigma(z)] = 0$ for $z \sim \mathcal{N}(0, 1)$) because we can exclude a large but trivial spike in the initial spectra of kernel matrices. In all architectures of NNs we considered, we remove the bias term of each layer and apply the NTK parameterization (1) with standard Gaussian initialization. Specifically, all entries of $\boldsymbol{W}$ and $\boldsymbol{v}$ in (7) at initialization are i.i.d. standard Gaussian random variables. For all experiments on synthetic datasets, we use standard Gaussian random matrices to generate the training data $\boldsymbol{X}$. In addition, we consider the training label noise defined in Assumption 3.3 as $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}_n)$.

B.1 Further Discussions on Real-world Data Experiments

For the first experiment in Section 5, we fix the small-CNN architecture, which is similar to the VGG model, and CIFAR-2 dataset, and vary the methods of optimization of training. Corresponding to Figure 4, the training and test accuracy histories for three cases are shown in Figure 6. Here, in Figure 6(a), we used GD with learning rate 8×10^{-3}; we used SGD with learning rate 10^{-3}, batch size 32 and momentum 0.2 in Figure 6(b); Figure 6(c) employs the same learning rate and batch size as 6(b) but employs Adam optimization.



(a) GD (b) SGD (c) Adam

Figure 6: Training/test accuracy v.s. epochs for small-CNN model on CIFAR-2 with different optimizers.

In the experiment of the transformer language model in Section 5, we fine-tuned the BERT model with SGD for a binary classification on the Sentiment140 dataset. We apply this transformer and fine-tuning to sentiment analysis for social media data. For fine-tuning, the learning rate is 0.003, the batch size is 64 and the momentum is 0.8. The purpose of this experiment is to extract the spectral properties of pre-trained models and the evolution of the CK spectra over fine-tuning. Combining Figure 5, the following Figure 7 exhibits the evolution of the CK spectrum during fine-tuning. Similarly with Case 4 in Table 1, the CK spectrum of this pre-trained model (red histogram in Figure 7) possesses a heavy-tailed distribution, which suggests this transformer has received adequate training. From Figure 7(a) to 7(c), we observe the bulk distribution first shrinks then extends during fine-tuning. This is similar to the evolution of the first eigenvalue of CK in Figure 5(a). Accompanied by this spectra evolution, there is a rapid transformation of the features through fine-tuning, linking the features in the pre-trained model with features in the new dataset. We expect that further spectral analysis will elucidate the feature learning in this kind of transformer [82].

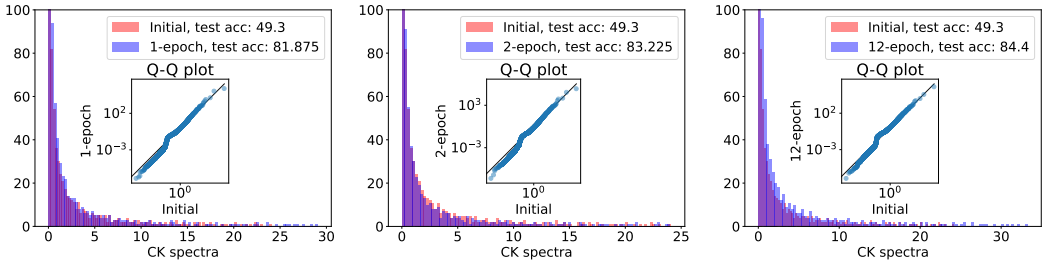


Figure 7: The spectra of CK of the BERT model on Sentiment140 dataset at epoch 1, 2 and 12.

B.2 Additional Results for Cases in Table 1

To complement the findings in Figure 1 and Section 3, we now present additional results on synthetic data and two-layer NNs. In this section, we will always use the same architecture and dataset as the typical examples in Table 1.

Norms of the Change. Based on Figure 1, the trajectories of the weight and kernel matrices are quite different among all cases in Table 1. Hence, for all cases in Table 1, we record the changes in the weight and NTK matrices in both Frobenius norm and operator norm in Figure 8:

$$\frac{1}{\sqrt{d}}\left\|\boldsymbol{W}_0-\boldsymbol{W}_t\right\|_F,\ \frac{1}{\sqrt{d}}\left\|\boldsymbol{W}_0-\boldsymbol{W}_t\right\|,\ \left\|\boldsymbol{K}_0^{\mathrm{NTK}}-\boldsymbol{K}_t^{\mathrm{NTK}}\right\|_F,\ \text{and}\ \left\|\boldsymbol{K}_0^{\mathrm{NTK}}-\boldsymbol{K}_t^{\mathrm{NTK}}\right\|,$$

at every epoch t through training. The changes in Figures 8(a) and (b) are much smaller than in the last case. Figure 8(c) has significant changes in both norms after training, which is consistent with the heavy-tailed phenomenon in Figure 1(c). The global optima of the last case is far from the initialization.

Following the settings of Theorem 4.2, in Figure 9, we compute the differences between initial $\boldsymbol{W}_0$ and final $\boldsymbol{W}_s$ in Frobenius norm, operator norm and $2,\infty$-norm. Empirically, Figure 9 shows

$$\frac{1}{\sqrt{d}}\left\|\boldsymbol{W}_0-\boldsymbol{W}_s\right\|_F,\ \frac{1}{\sqrt{d}}\left\|\boldsymbol{W}_0-\boldsymbol{W}_s\right\|,\ \frac{1}{\sqrt{d}}\left\|\boldsymbol{W}_0-\boldsymbol{W}_s\right\|_{2,\infty}=\Theta(1) \tag{16}$$

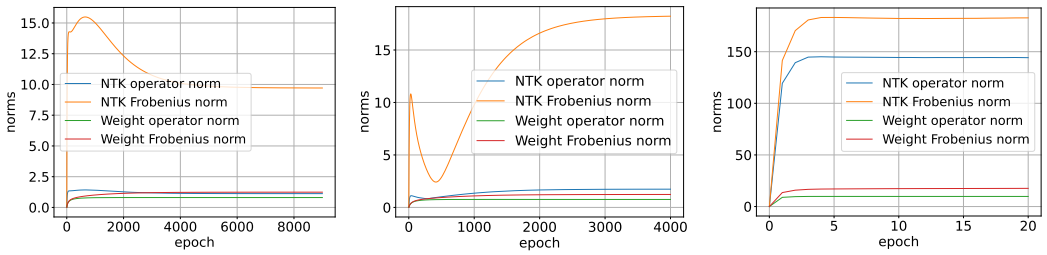
as $n\to\infty$ with $n/d\to\gamma_1$ and $N/d\to\gamma_2$, where s is the final time for GD. Here, the entry-wise 2-∞ matrix norm is defined as

$$\|\boldsymbol{M}\|_{2,\infty}:=\max_{1\le i\le N}\|\mathbf{m}_i\|,$$

for any matrix $\boldsymbol{M}\in\mathbb{R}^{N\times d}$ with the i-th row $\mathbf{m}_i\in\mathbb{R}^d$ and $1\le i\le N$. Notice that

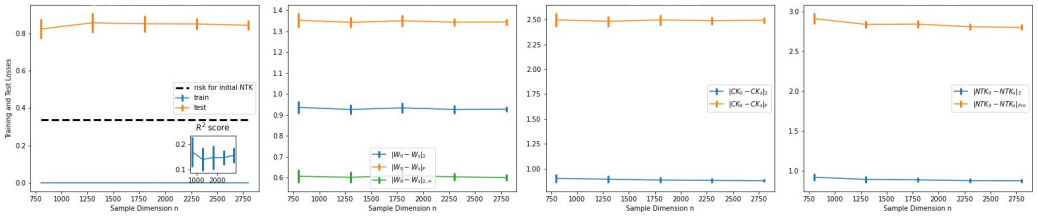
$$\|\boldsymbol{M}\|_{2,\infty}\le\|\boldsymbol{M}\|\le\|\boldsymbol{M}\|_F. \tag{17}$$

Similar observations for CK and NTK in both Frobenius norm and operator norm are also apparent in Figure 9, which empirically verifies the invariance of the spectra after training. Here, we fix the aspect ratios and let n grow to keep the NNs residing in LWR. For different n's, we repeat the experiments 10 times for average. In each experiment, we train the NN until it converges. As shown in Figure 9(a), the test losses are almost the same for different n's. Figures 9(b)-(d) empirically validate Corollary 4.3. Moreover, the observation that $\frac{1}{\sqrt{d}}\left\|\boldsymbol{W}_0-\boldsymbol{W}_s\right\|_F$ and $\frac{1}{\sqrt{d}}\left\|\boldsymbol{W}_0-\boldsymbol{W}_s\right\|$ are $\Theta(1)$ may suggest that $\frac{1}{\sqrt{d}}\left(\boldsymbol{W}_0-\boldsymbol{W}_s\right)$ is a *low-rank* perturbation. That is, training in LWR may be transferring some low-rank structures to the weight spectrum. This low-rank perturbation can help us better understand the spectral evolution during training. Notice that these norms of the change are different from ultra-wide NN [27, 28]. Similar phenomena can be also observed in Figures 10(a) and (b). Analogous result with different σ and σ^* is exhibited in Figure 30 in Appendix C. In addition, Figure 10(c) further investigates the cases when NNs can outperform lazy training as defined by (6). In these experiments, we compare the performances of GD, and SGD with small or large learning rates, and lazy training as $n\to\infty$. Each time, we take 10 trials to average. We observe that SGD with a large learning rate (green line) can asymptotically outperform lazy training.



(a) Case 1 (b) Case 2 (c) Case 4

Figure 8: The evolution of the changes in operator/Frobenius norms of the weight/CK/NTK matrices through different training processes. Each case corresponds to the case in Table 1. Case 3 is exhibited in Figure 14(c) below.



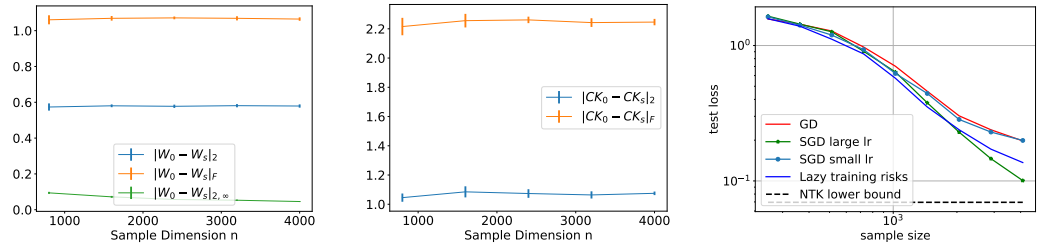
(a) Train/test losses (b) Change in weight (c) Change in CK (d) Change in NTK

Figure 9: Performances of NNs and changes in different norms for weight and kernels, when $d/n = 0.6$ and $N/n = 1.2$ are fixed as n is growing. The activation is normalized $\tanh$ and the teacher model is $f^*(\boldsymbol{x}) = \sigma^*(\boldsymbol{\beta}^\top \boldsymbol{x})$ where σ^* is a normalized *softplus*. We average over 10 trials in each case. All these curves are almost flat, which indicates these values are not growing with γ_1 and γ_2. Here, in the second figure from the left, we normalized all weights $\boldsymbol{W}$ with $\frac{1}{\sqrt{d}}$ to observe (16).

Case 1. Comparing with Figure 15, Figure 11 shows no alignments with training data in GD training. This corresponds to the performances in Table 1. The performance of Case 1 is not as good as the prediction risks in Figure 15, since Figure 11 suggests that no feature learning appears after GD training. Gradient descent requires the weights to converge to some global minima close to initialization, thereby offering no guarantees for lower generalization errors. Next, Figure 12 further presents more results on GD training and indicates more evidence of kernel regime in Case 1. This shows that, from a spectral point of view, the NTK is invariant/static through training. Based on Figures 1(a) and 12, we can empirically verify Corollary C.3 stated in Appendix C. Globally, the spectra of $\boldsymbol{W}$, $\boldsymbol{K}^{\mathrm{CK}}$ and $\boldsymbol{K}^{\mathrm{NTK}}$ are not changing over training as $n/d \to \gamma_1$ and $N/d \to \gamma_2$. The initial spectrum of weight $\boldsymbol{W}_0$ converges to Marčenko–Pastur law; the initial spectrum of NTK under proportional limit has been studied by [1, 30]. Figure 12(c) demonstrates the global convergence for GD under the proportional regime, as proved in Theorem 4.2. We can observe this global convergence even for SGD, Case 2 in Table 1, although we do not have proof for it.

Case 2. As a complement, Figure 13 exhibits the spectra of $\boldsymbol{W}$, $\boldsymbol{K}^{\mathrm{CK}}$ and $\boldsymbol{K}^{\mathrm{NTK}}$ for Case 2 in Table 1. The phenomena are similar to Case 1. This observation provides evidence that all results and conjectures in Section 4.1 can be extended to SGD training with sufficiently small learning rates, which is subject to future work. Analogously to Theorem 4.2, we conjecture that the global convergence when training both layers of NN with SGD still holds in this proportional limit. The proof strategy for global convergence, in this case, can again follow [69, 70]. Once we have the invariant global spectra in Corollary 4.3, we can apply the nonlinear RMT [73, 57, 13, 30] to characterize the limiting spectra under LWR.

Case 3. Next, in Figures 14 and 15, we present spectral properties for Case 3 in Table 1, where a spike detaches from the bulk after large-step-size training. Notice that Figures 1(b) and 14(a) imply that the bulk spectra for weight and CK remain unchanged over training despite the emergence of spikes. This is not true for NTK by observing Figures 14(b) and (c). The Frobenius norm of NTK changes significantly during training and is not $O(1)$ anymore; the spectra of the first component



(a) Change in weight. (b) Change in CK. (c) Prediction risks vs. γ_1.

Figure 10: (a) Change between initial $\boldsymbol{W}_0$ and final step $\boldsymbol{W}_s$ in operator norm, Frobenius norm and $(2,\infty)$-norm when $d/n = 0.5, N/n = 0.8$ are fixed as $n \to \infty$. We train NNs by SGD with $\eta = 2.5$ for 15 trials to average. (b) Change of $\boldsymbol{K}^{\text{CK}}$ in operator norm and Frobenius norm. (c) Prediction risks for lazy training defined by (6), GD with $\eta = \Theta(1)$ (red), SGD with $\eta = \Theta(1)$ (blue dot) and $\eta \propto \gamma_1$ (green), as $\gamma_1 \to \infty$ and $\gamma_2 = 2.5$. The black dashed line stands for the kernel lower bound given by the nonlinear part of the teacher model.

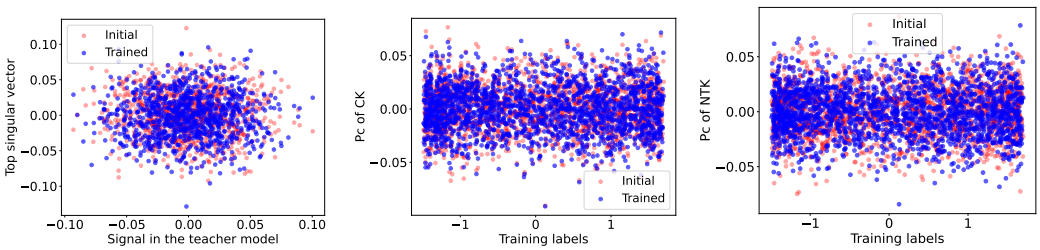


Figure 11: Alignment with leading PCs of trained weight, CK and NTK matrices in Case 1 of Table 1.

of the NTK shrinks after training (Figure 14(b)), which indicates NNs converge to flatter minima. This resembles the catapult phase in [49] for extremely large learning rates. Figure 15(a) shows the convergence rate for SGD in Case 3. Empirically, we observe that the training loss will not monotonically decrease when using a larger learning rate than Case 3, which may be analogous to catapult phases from [49].

Case 4. Additional results for Case 4 in Table 1 are shown in Figure 16. Unlike the strong alignments in Case 3 (Figure 15), Figures 16 (e)-(f) do not exhibit strong alignments for the leading singular vectors or eigenvectors. This may be due to the heavy tails present after training, with the other large spikes detaching from the bulks are also important for generalization. A similar phenomenon can be also seen in Figures 23 and 25 in Appendix B.6, where we have comparable performances to Case 4 in Table 1.

Heavy tails are essentially power laws. To measure how "heavy" the spectrum is, [58, 59, 60] provide estimates on the power law of $\boldsymbol{W}$. Consider the empirical spectrum of $\boldsymbol{W}$ as $\rho(x) \sim x^{-\alpha}$ for large x and some positive constant α. The spectrum with a heavier tail has a smaller value of α. Figure 17(a) shows how α evolves through training in Case 4 of Table 1. As α decreases, a heavy tail in the spectrum of the weight matrix emerges in Figure 1(c). In Figure 17(a) we introduce two more metrics to show this evolution: Weighted Alpha $\hat{\alpha} := \alpha\lambda_1$ and Log α-norm $\log\left(\sum_{i=1}^{N} \lambda_i^{\alpha}\right)$ where $\lambda_N \leq \ldots \leq \lambda_1$ are the eigenvalues of $\boldsymbol{W}\boldsymbol{W}^{\top}$. Remarkably, Figure 17(a) indicates the spectra change dramatically at the early stage of training, which matches the observation from fine-tuning via BERT on real-world data in Figure 5 in Section 5. These metrics are applied to measure the tails in pre-trained models [60].

B.3 Additional Results for the Emergence of A Spike

As a complement to section 4.2, in Figure 18, we show the training dynamics for SGD training with a larger learning rate in the example of Figure 2(b-d). Here we consider $\eta = 24$ which belongs to the orange region in Figure 2(b-d), where the spike and eigenvector alignment emerge. Figure 18 presents the details of the training dynamics of the NN in this case: the largest eigenvalues of CK and

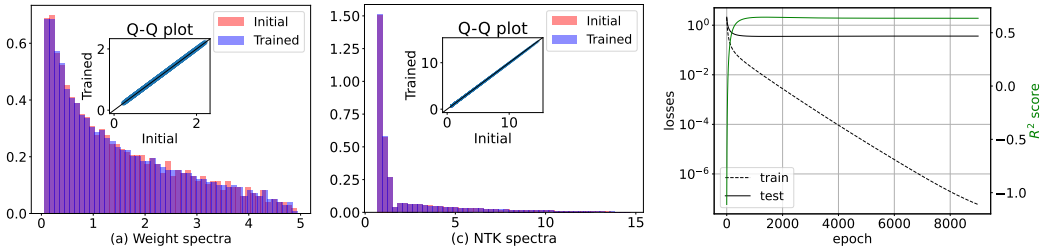


Figure 12: Performances of Case 1 in Table 1: (a) The initial and trained spectra of the first-hidden layer $\boldsymbol{W}$. (c) The initial and trained spectra of empirical NTK matrix defined by (5). Q-Q subplot shows these two spectra are almost the same. Training and test losses vs. epochs for GD (right).

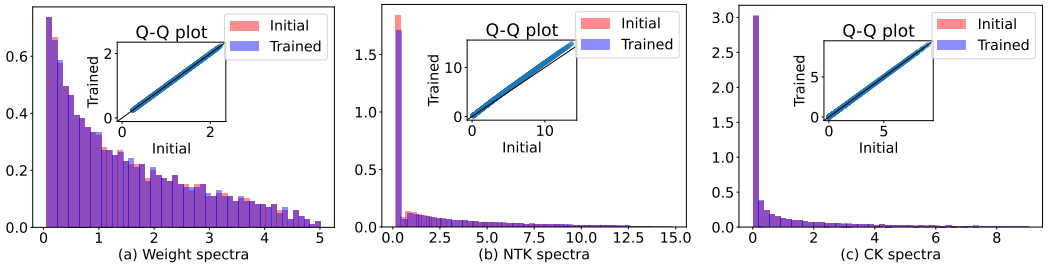
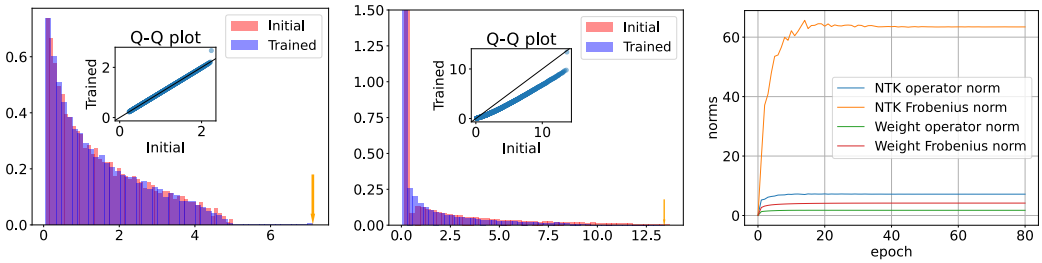


Figure 13: Spectral properties for Case 2 in Table 1: (a) The initial and trained spectra of the first-hidden layer $\boldsymbol{W}$. (b) The initial and trained spectra of empirical NTK are defined by (5). (c) The initial and trained spectra of empirical CK defined by (4).

NTK both increase and the losses first increase and then drop. In Figure 19, we empirically justify that the phase transitions we presented in section 4.2 for SGD can be also extended to full-batch GD cases. We can also observe phase transitions for test losses and R^2 scores when we are gradually increasing the learning rates. Parallel to these, a spike also appears outside the bulk distribution, which corresponds to feature alignments in Figure 19(c)&(f).

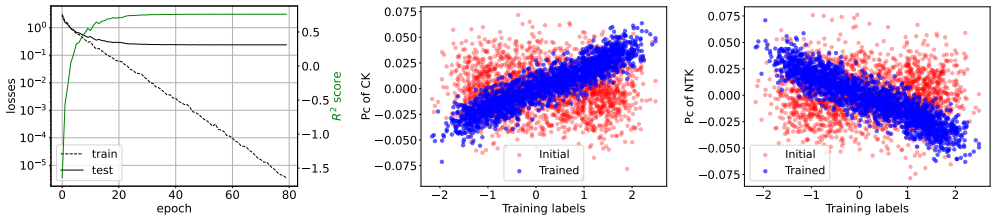
B.4 Multiple-index Examples for Heavy-Tailed Spectra in Section 4.3

Figures 17(b) and (c) are additional results for Figure 3(b) in Section 4.3. In this experiment, we consider $\sigma = \text{ReLU}$, $n = 5000$, $h = 2500$ and $d = 1000$ for NN (1). Comparing with the teacher model (9) used in Table 1, we employ the multiple-index teacher model (8) with $k = 5$ and $\sigma^* = \sigma$. We trained this student-teacher model using GD ($\eta = 15$), SGD ($\eta = 7.25$ and batch size 8), and Adam ($\eta = 0.007$ and batch size 16) for training this NN, respectively. Similarly with Figure 1, correspondingly, we observe invariant spectrum, bulk with one spike, and heavy tails after training respectively. Heuristically, to learn this f^*, the weight $\boldsymbol{W}$ of NN should gradually align with the feature space U spanned by $\boldsymbol{\beta}_i$'s. Hence, to study feature learning, we can apply principle angles to measure the alignment between $\boldsymbol{W}$ and U. Consider the eigen-decomposition of $\boldsymbol{W}_t^\top \boldsymbol{W}_t = \sum_{i=1}^d \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$. Figure 3(b) shows the heavy-tailed part (the eigenspace $E := \text{span}\{\boldsymbol{v}_i\}_{i=1}^{100}$) is aligned with U after training, which shows how features are learned in the heavy-tailed spectra. Remarkably, the test errors for training processes with SGD and Adam are even smaller than $\|\mathrm{P}_{>1} f^*\|^2$ and $\|\mathrm{P}_{>2} f^*\|^2$, where $\mathrm{P}_{>1}$ denotes the orthogonal projection onto the nonlinear part of the function w.r.t. Gaussian measure. Thus, we experimentally showed that NNs with heavy-tailed spectra can obtain feature learning and generalize better than the other two cases. Another example is exhibited in Figure 20. In this case, $k = 5$ and there are five leading outlier eigenvalues in the spectrum of the trained weight matrix, along with a heavy-tailed bulk. Interestingly, Figure 20 justifies that the eigenspace of these five leading outliers is strongly aligned with features $\boldsymbol{\beta}_i$ for $1 \leq i \leq 5$. This indicates that heavy-tailed spectra with large spikes may have a correlation with feature learning and good generalizations.



(a) Spectra of weights. (b) Spectra of the first part in NTKs. (c) Norms of change vs. epochs.

Figure 14: Additional performance for Case 3 in Table 1: (a) The initial and trained weight spectra. Notice that there is one outlier after training, while the bulk remains invariant. This is analogous to the behavior of CK spectra in Figure 1(b). (b) The spectra of the first part in (5) at initialization and after training. The orange arrow points out the outlier of the spectrum. (c) The changes $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|$, $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$, $\left\|\boldsymbol{K}_t^{\text{NTK}} - \boldsymbol{K}_0^{\text{NTK}}\right\|$ and $\left\|\boldsymbol{K}_t^{\text{NTK}} - \boldsymbol{K}_0^{\text{NTK}}\right\|_F$ at each epoch t throughout the training process.



(a) Losses vs. epochs. (b) $\boldsymbol{y}$ vs. $\boldsymbol{v}_1$ of CK. (b) $\boldsymbol{y}$ vs. $\boldsymbol{v}_1$ of NTK.

Figure 15: (a) Training/test losses and R^2 scores at each epoch of training in Case 3. (b) Alignment between training labels $\boldsymbol{y}$ and first PC of trained/initial CK. (c) Alignment between training labels $\boldsymbol{y}$ and the first PC of trained/initial NTK. We use the same setting in Case 3 of Table 1. There are strong alignments between kernels and training labels as stated in Section 4.2.

B.5 Training Only the First Hidden Layer

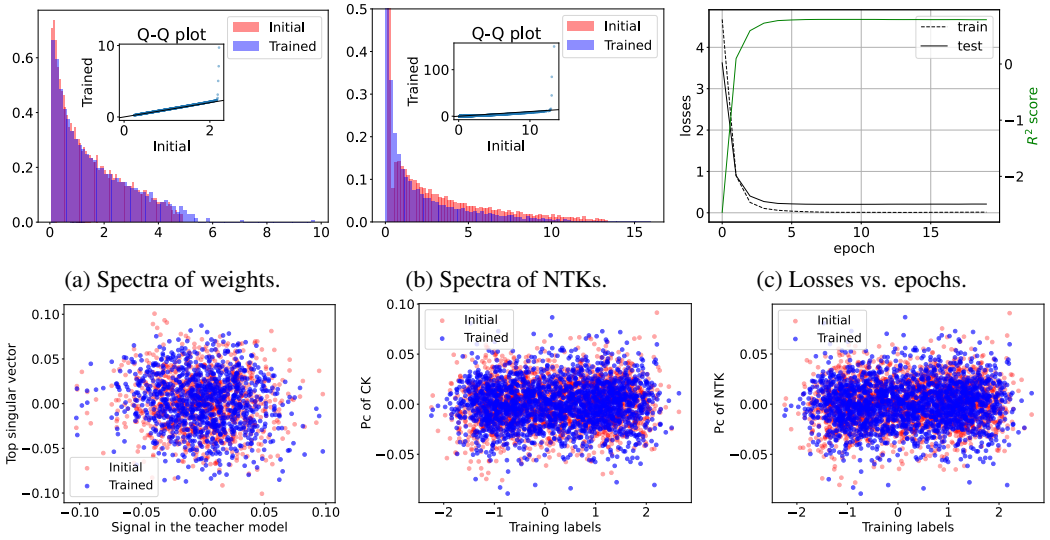
We now present additional results when only training the first layer of NN with Adam. This result resembles Figure 29 in the next section. In section 4.1, Corollary 4.3 shows that under LWR with sufficiently large width h, the limiting spectra of $\frac{1}{h}\boldsymbol{W}_t^\top \boldsymbol{W}_t$, $\boldsymbol{K}_t^{\text{CK}}$ and $\boldsymbol{K}_t^{\text{NTK}}$ are essentially the same as those of the corresponding initial matrices if we train only the first layer with GD (see (10)). Figure 21 further investigates these phenomena when training only the first layer with Adam. In particular, the Q-Q plots show the invariant spectra of weight, CK, and NTK matrices even when the training loss is approaching zero. Here, in Figure 21(c), we only consider the first component of NTK since the gradient is only taken with respect to $\boldsymbol{W}_t$. Another observation is that the smallest eigenvalues of the initial and trained NTK are both bounded away from zero. This is crucial for the proof of the global convergence as shown in Appendix C.

B.6 Adaptive Gradients

Inspired by Case 4 in Table 1, we show the spectral performances of adaptive gradient (AdaGrad) in Figures 22 and 23. The performance of this method matches Case 4 in Table 1, where we can also easily observe heavy tails and detaching spikes after training, especially in Q-Q subplots. This suggests that adaptive optimization is more likely to yield heavy-tailed distributions in trained NNs. Besides, analogously to Figure 16, there is no strong alignment in the single leading PC of the weight or kernel matrices after training in Figure 23.

B.7 Different Global Minima and Alignments

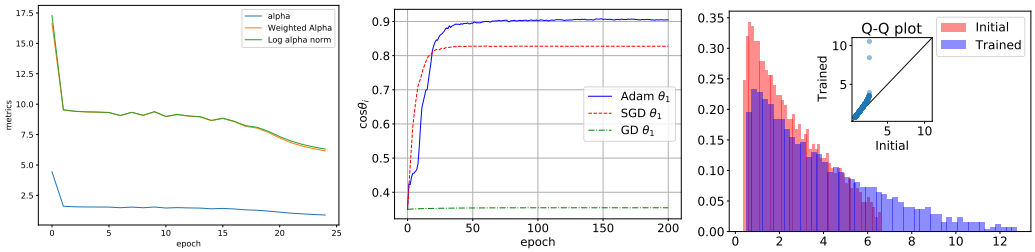
In this section, to distinguish the different alignments in Case 3&4 of Table 1, we introduce the following two simulations with slightly different optimizers to get quite different spectra and alignments among leading PCs after training. In the first experiment, Figures 24 and 25, we first take Adam with large stepsizes for a few steps and then use small-stepsize SGD for convergence. In this scenario, we



(a) Spectra of weights. (b) Spectra of NTKs. (c) Losses vs. epochs.

(d) First PC in trained & initial weights. (e) First PC in trained & initial CKs. (f) First PC in trained & initial NTKs.

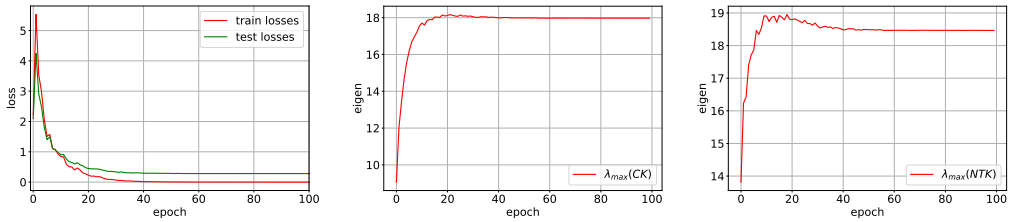
Figure 16: Additional performance for Case 4 in Table 1: (a) The initial and trained weight spectra. Notice that there are several outliers after training, while the bulk has a heavier tail. (b) The spectra of the NTK (5) at initialization and after training. (c) The test/training losses and R^2 score (green line) at each epoch t throughout training process. (d) Alignment between the leading PC of the weight matrix and the signal $\boldsymbol{\beta}$ in the teacher model before (red) and after (blue) training. (e) Alignment between the leading PC of the CK matrix and the training labels $\boldsymbol{y}$ before/after training. (f) Alignment between the leading PC of the NTK matrix and $\boldsymbol{y}$ before/after training.



(a) Evolution of metrics. (b) Evolution of PC angles. (c) Weight spectra.

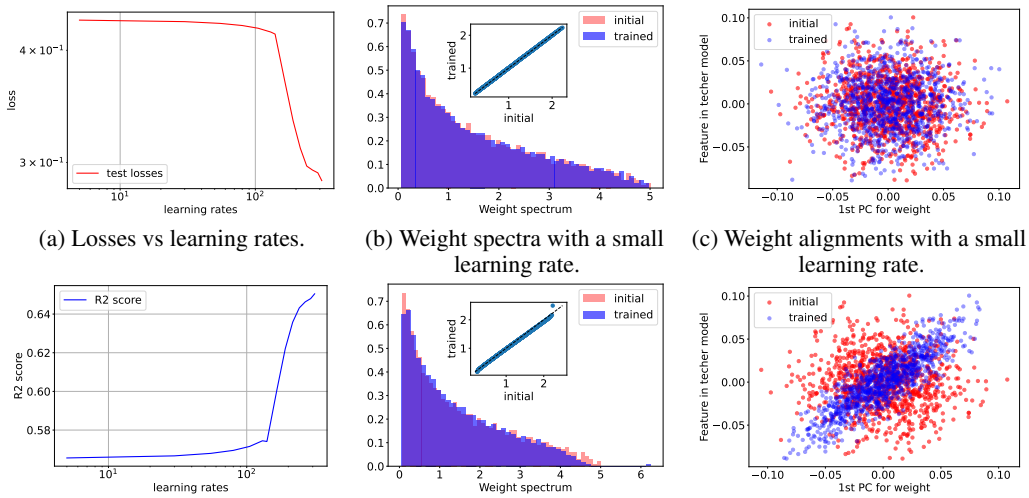
Figure 17: (a) The evolution of power α, weighted Alpha and Log α-norm (several metrics of power law tails; see [60]) during the training process in Case 4 of Table 1. (b) Evolutions of the first PC angle θ_1 between *feature subspace* $U = \text{span}\{\boldsymbol{\beta}_i\}_{i=1}^k$ of the multiple-index model (8) and the *eigenspace* spanned by top 100 of eigenvectors of $\boldsymbol{W}_t^\top \boldsymbol{W}_t$ during training with Adam (blue solid line), SGD (red dashed line) and GD (green dash-dot). The final test error is 0.33865 and the R^2 score is -0.71065 for GD. The test error is 0.10814 and the R^2 score is 0.45373 for SGD, where one spike emerges in the weight spectrum after training. The test error is 0.08672 and the R^2 score is 0.56195 for Adam. (c) Initial and trained spectra for weight matrices when training with Adam (blue solid line in (b)). Heavy tail emerges in this case.

can get heavy-tailed distributions after training, and the phenomena are essentially the same as Case 4 of Table 1. There is no strong alignment for the first leading eigenvector, while useful features may be learned by a few top eigenvectors in the heavy-tailed spectra after training. In the second experiment, we directly apply Adam with a small initial learning rate for training. In contrast, the results of this case, presented in Figures 26 and 27, are similar to Case 3 in Table 1. The test loss and R^2 score are close to previous examples, though slightly worse. As explained in the previous section, because there is only one spike appearing outside the bulk after training in the second case, the leading PC is highly aligned with the training dataset structure after training and the feature learning mainly stems from the outliers in this situation. This interprets the strongly anisotropic structures in trained spectra of NNs in the second case [67, 68, 78]. These two different spectral properties reveal significant differences between the global minima of these two training processes and different evolutions of the



(a) Losses. (b) $\lambda_{\max}(\mathrm{CK})$. (c) $\lambda_{\max}(\mathrm{NTK})$

Figure 18: The training dynamic when training neural networks with SGD and learning rate 24 in the example of Figure 2(b-c). The learning rate we chose here is above the threshold we showed in Figure 2(b-c). We use the same architecture, dataset, and teacher model as in Section 3 of our paper. The batch size is 32. (a) The evolution of the training and test errors during training. (b) The evolution of the largest eigenvalue of the CK matrix. (c) The evolution of the largest eigenvalue of the NTK matrix. This regime corresponds to the catapult phenomenon [49].



(a) Losses vs learning rates. (b) Weight spectra with a small learning rate. (c) Weight alignments with a small learning rate.

(d) R^2 vs learning rates. (e) Weight spectra with a large learning rate. (f) Weight alignments with a large learning rate.

Figure 19: Grid search for different learning rates when training NNs with full-batch GD in the same setting as Case 1 in Table 1. (a) Final test losses when varying learning rates. (d) R^2 scores when varying learning rates. For all these learning rates, we did not observe heavy-tailed spectra. (b-c) present the spectral behaviors for the smallest learning rate we used in (a)&(d). (e-f) present the spectral behaviors for the largest learning rate we can use which still ensures the convergence of GD. In this case, analogously to the SGD case in Section 4.2, we observe an outlier in the trained weight matrix and strong alignment with the spike.

spectra in NNs. Remarkably, based on the different spectral behaviors in trained weight and kernel matrices, these two experiments exhibit disparate features learned by distinct training procedures. Hence, analyzing the spectral properties in trained kernel matrices is beneficial for clarifying what features our NNs have learned during the training processes.

C Proofs of Results in Section 4.1

C.1 GD Analysis at Early Phase

From (10), the GD process with learning rate $\eta > 0$ can be written by

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \eta \cdot \boldsymbol{G}_t, \text{ where} \tag{18}$$

$$\boldsymbol{G}_t = \frac{1}{n\sqrt{dh}} \left[\left(\boldsymbol{v} \left(\boldsymbol{y} - \frac{1}{\sqrt{h}} \boldsymbol{v}^\top \sigma(\boldsymbol{W}_t \boldsymbol{X}/\sqrt{d}) \right) \right) \odot \sigma'(\boldsymbol{W}_t \boldsymbol{X}/\sqrt{d}) \right] \boldsymbol{X}^\top, \tag{19}$$

for $t \in \mathbb{N}$, where $\boldsymbol{y} \in \mathbb{R}^{1\times n}$. Following [6, Appendix B], in this section we prove the control for gradient step $\boldsymbol{G}_t$. For simplicity, denote $f_t(\boldsymbol{X}) := f_{\boldsymbol{\theta}_t}(\boldsymbol{X}) = \frac{1}{\sqrt{h}} \boldsymbol{v}^\top \sigma(\boldsymbol{W}_t \boldsymbol{X}/\sqrt{d})$ for $t \in \mathbb{N}$.

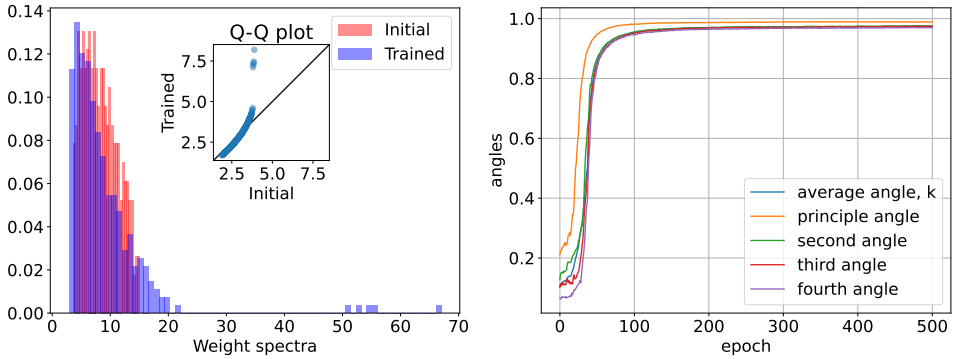


Figure 20: (Left) Initial and trained spectra for weight matrices when training with Adam. Five leading spikes emerge in this case. (Right) Evolutions of the angles between the first four PCs of $\boldsymbol{W}_t^\top \boldsymbol{W}_t$ and *feature subspace* $U = \text{span}\{\boldsymbol{\beta}_i\}_{i=1}^k$ of the multiple-index model (8) during training with Adam. Here $k = 5$. The final test error is 0.33865 and R^2 score is -0.71065 for GD. The test error is 0.01681 and R^2 score is 0.9154 for Adam.

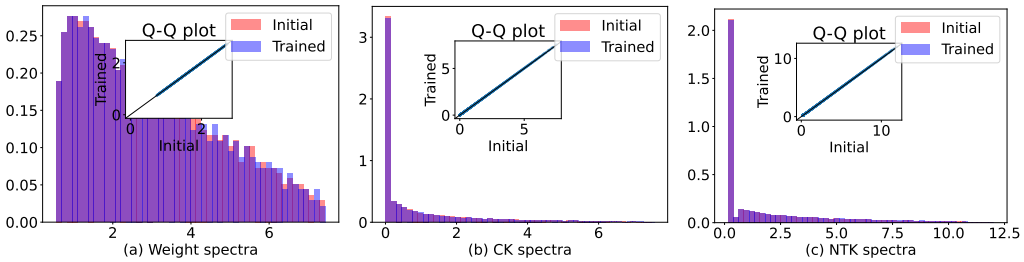


Figure 21: Additional spectral performance when training the NN by Adam with small learning rate $\eta = 0.001$, where $n = 2000, h = 3000, d = 1000, \sigma_\varepsilon = 0.3$ and batch size is 100. In this simulation, we only train the first hidden layer $\boldsymbol{W}_t$. The activation function σ is a normalized softplus and the target function is a normalized $\tanh$. The final test loss is 0.36219 and R^2 score is around 63.70%.

Lemma C.1. *Under the same assumptions as in Lemma 4.1, we have*

$$\mathbb{P}\left(\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\| \geq C\sqrt{n}\right) \leq 2e^{-cn},$$
$$\mathbb{P}\left(\|\boldsymbol{y}\| \geq C\sqrt{n}\right) \leq 2e^{-cn},$$

for some constants $C, c > 0$ only depending on σ_ε, λ_σ, γ_1, and γ_2.

Proof. Due to [30, Lemma D.4.], we can directly obtain that

$$\mathbb{P}\left(\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\| \geq C'(\sqrt{n}+\sqrt{h})\sqrt{\frac{h}{d}}\right) \leq 2e^{-cn}.$$

Here we use the fact that both $\boldsymbol{W}_0$ and $\boldsymbol{X}$ are i.i.d. Gaussian random matrices. Then by Assumption 3.1, we conclude that we control $\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})$. Recall that Assumption 3.3 implies that $\boldsymbol{y} = f^*(\boldsymbol{X}) + \boldsymbol{\varepsilon}$. Hence, by Lipschitz Gaussian concentration inequality [80, Theorem 5.2.2], each entry of $f^*(\boldsymbol{X})$ has independent sub-Gaussian coordinates, whence we can get $\|f^*(\boldsymbol{X})\| \leq C\sqrt{n}$ with probability at least $1 - 2ne^{-cn}$ for some constants $c, C > 0$. On the other hand, $[\boldsymbol{\varepsilon}]_i = \varepsilon_i$ are i.i.d. centered sub-Gaussian noises with variance σ_ε^2. By [80, Theorem 3.1.1], we have

$$\mathbb{P}\left(\|\boldsymbol{\varepsilon}\| \leq 2\sigma_\varepsilon\sqrt{n}\right) \geq 1 - 2\exp\left(-\frac{cn}{K^4}\right),$$

where the constant K is the sub-Gaussian norm defined by $K = \max_i \|\varepsilon_i\|_{\psi_2}$. Hence, combining all things together, we obtain the second inequality of this lemma.

□

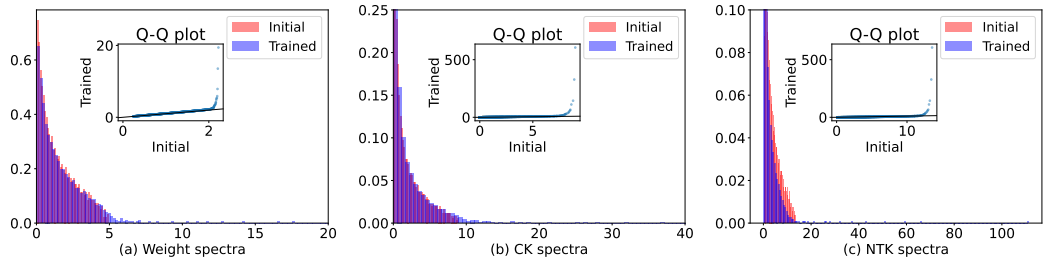
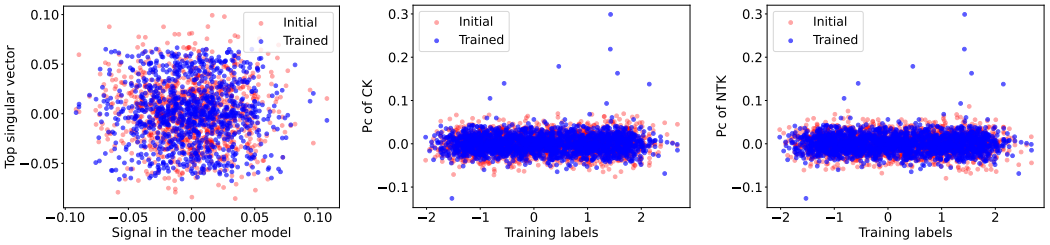


Figure 22: Additional performance for AdaGrad with learning rate $\eta = 0.5$, where $n = 2000, h = 1500, d = 1000$, $\sigma_\varepsilon = 0.3$ and small batch size is 8. Activation σ is normalized softplus and target is normalized $\tanh$. The final test loss is 0.23555 and R^2 score is around 0.76249. The black lines in the Q-Q subplots are the line of $y = x$.



(a) First PC in trained & initial weights.

(b) First PC in trained & initial CKs.

(C) First PC in trained & initial NTKs.

Figure 23: Alignment between the leading PC of the weight/kernel matrices and the signal $\boldsymbol{\beta}$ or the training labels $\boldsymbol{y}$ before/after training for experiment in Figure 22. Analogously to Case 4 in Appendix B.2, there is no strong alignment in the leading component of weight/kernel matrices.

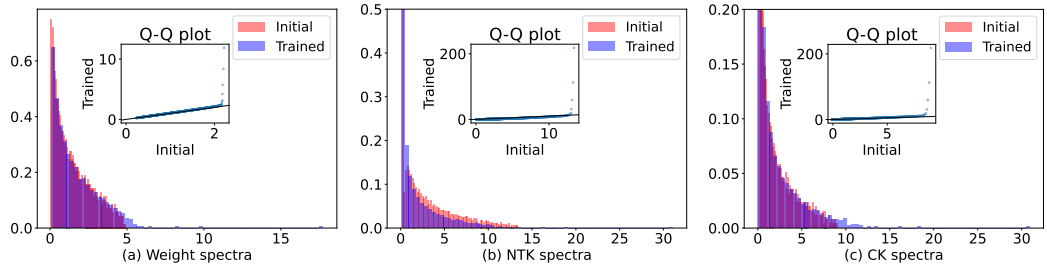


Figure 24: Additional performance for Adam with learning rate $\eta = 0.09$ and 4 epochs, then SGD with learning rate $\eta = 5 \times 10^{-4}$ and 100 epochs, where $n = 2000, h = 1500, d = 1000$, $\sigma_\varepsilon = 0.3$ and the batch size is 32. We train the NN until the training loss is less than 10^{-10}. The activation σ is normalized softplus and the target is normalized $\tanh$. The final test loss is 0.22511 and R^2 score is around 0.77462.

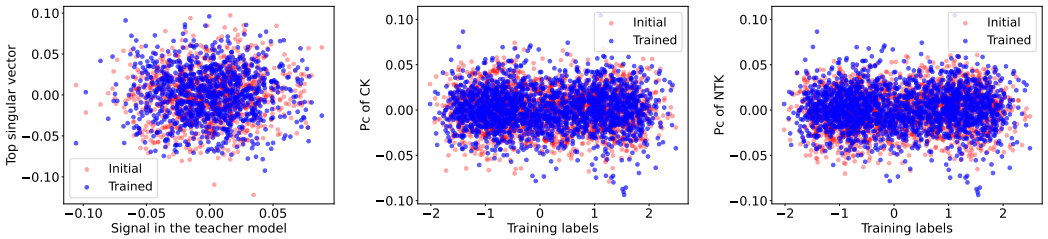
Lemma C.2. *Under the assumptions of Lemma 4.1, given any fixed $t \in \mathbb{N}$ and learning rate $\eta = \Theta(1)$, the weight matrix after t gradient steps $\boldsymbol{W}_t$ defined in* (18) *satisfies*

$$\mathbb{P}\left(\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F \geq \frac{C}{\sqrt{n}}\right) \leq \exp\left(-cn\right), \tag{20}$$

for some positive constants $c, C > 0$ only depending on $t, \eta, \sigma_\varepsilon, \lambda_\sigma, \gamma_1$ and γ_2.

Proof. Denote $\sigma_\perp(x) = \sigma(x) - \mu_1 x$ which is the nonlinear part of σ and $\mu_1 = \mathbb{E}[z\sigma(z)]$. Thus, $\mathbb{E}[\sigma_\perp(z)z] = 0$ for $z \sim \mathcal{N}(0,1)$. Based on this, we can further decompose the gradient $\boldsymbol{G}_t$ into

$$\boldsymbol{G}_t = \underbrace{\frac{\mu_1}{n\sqrt{dh}}\boldsymbol{v}\left(\boldsymbol{y} - f_t(\boldsymbol{X})\right)\boldsymbol{X}^\top}_{\boldsymbol{A}^t} + \underbrace{\frac{1}{n\sqrt{dh}}\left(\boldsymbol{v}\left(\boldsymbol{y} - f_t(\boldsymbol{X})\right) \odot \sigma_\perp'(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})\right)\boldsymbol{X}^\top}_{\boldsymbol{B}^t}. \tag{21}$$



(a) First PC in trained weight. (b) First PC in trained CK. (c) First PC in trained NTK.

Figure 25: Alignment between the leading PC of the weight/kernel matrices and the signal $\boldsymbol{\beta}$ or the training labels $\boldsymbol{y}$ before/after training for the experiment in Figure 24. This is analogous to Case 4 in Appendix B.2.

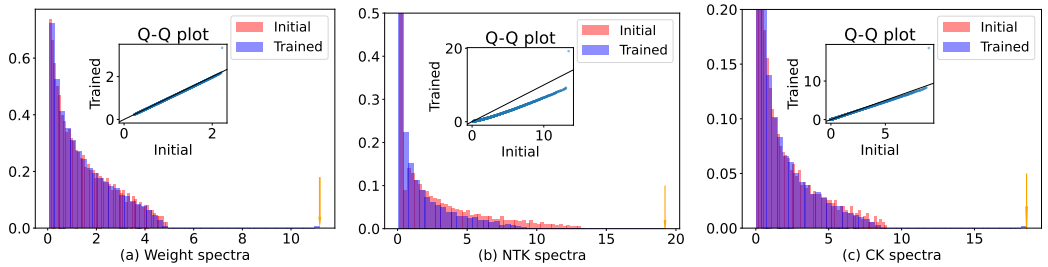


Figure 26: Additional performance for Adam with learning rate $\eta = 0.002$ and 700 epochs, where $n = 2000, h = 1500, d = 1000$, $\sigma_\varepsilon = 0.3$ and batch size is 64. The activation σ is normalized softplus and the target is normalized $\tanh$. The final test loss is 0.23954 and R^2 score is around 0.76027. The orange arrows show the positions of the outliers. Spectra behaviors in this case differ from Figure 24.

At first, consider $t = 0$ and bound the spectral norm of $\boldsymbol{W}_1$. By assumption, we know $\|\boldsymbol{v}\| \leq \sqrt{h}$. Due to Corollary 7.3.3 in [80], we have

$$\mathbb{P}\left(\frac{1}{\sqrt{d}}\|\boldsymbol{X}\| \geq 2\left(1+\sqrt{\frac{n}{d}}\right)\right) \leq 2\exp\left(-cn\right). \tag{22}$$

Therefore, by (21), we can control $\boldsymbol{A}^0$ and $\boldsymbol{B}^0$ separately. Notice that, as a rank-one matrix,

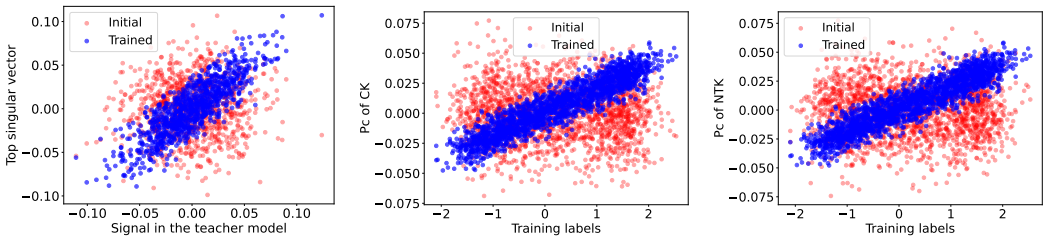
$$\begin{aligned}\left\|\boldsymbol{A}^0\right\| = \left\|\boldsymbol{A}^0\right\|_F &\leq \frac{\mu_1}{\sqrt{n}}\frac{\|\boldsymbol{X}\|}{\sqrt{d}}\frac{1}{\sqrt{n}}(\|\boldsymbol{y}\| + \|f_0(\boldsymbol{X})\|)\frac{\|\boldsymbol{v}\|}{\sqrt{h}} \\ &\leq \frac{\mu_1}{\sqrt{n}}\frac{\|\boldsymbol{X}\|}{\sqrt{d}}\frac{\|\boldsymbol{v}\|}{\sqrt{h}}\frac{1}{\sqrt{n}}\left(\|\boldsymbol{y}\| + \frac{\|\boldsymbol{v}\|}{\sqrt{h}}\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|\right).\end{aligned}$$

Hence, by Lemma C.1 and (22), one can easily claim that $\|\boldsymbol{A}^0\| \leq C/\sqrt{n}$ with probability at least $1 - e^{-cn}$ for some constants $c, C > 0$. On the other hand, since $\boldsymbol{v}\left(\boldsymbol{y} - f_t(\boldsymbol{X})\right)$ is rank-one and $\sigma'_\perp = \sigma' - \mu_1$ with $|\sigma'(x)| \leq \lambda_\sigma$, we can similarly obtain

$$\begin{aligned}\left\|\boldsymbol{B}^0\right\|_F &\leq \frac{1}{n\sqrt{dh}}\left\|\boldsymbol{v}\left(\boldsymbol{y} - f_t(\boldsymbol{X})\right) \odot \sigma'_\perp(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})\right\|_F \|\boldsymbol{X}\| \\ &\leq \frac{1}{n\sqrt{hd}}\|\boldsymbol{X}\|\left(\|\boldsymbol{y}\| + \|f_0(\boldsymbol{X})\|\right)\|\boldsymbol{v}\|\max_{i,j}\left|\sigma'_\perp(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right|_{i,j} \\ &\leq \frac{\mu_1 + \lambda_\sigma}{\sqrt{n}}\frac{\|\boldsymbol{X}\|}{\sqrt{d}}\frac{\|\boldsymbol{v}\|}{\sqrt{h}}\frac{1}{\sqrt{n}}\left(\|\boldsymbol{y}\| + \frac{\|\boldsymbol{v}\|}{\sqrt{h}}\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|\right).\end{aligned}$$

As $\boldsymbol{A}^0$, we can apply Lemma C.1 and (22) again to conclude (20) for $t = 1$.

For general t, we apply induction. We assume that after the t-th gradient step with $\eta = \Theta(1)$, Eq. (20) holds for some constants $C, c > 0$. Following [6, Lemma 16], we now show that the similar high-probability statement also holds for $\boldsymbol{W}_{t+1}$ (for some different constants c', C'). Firstly,



(a) First PC in trained & initial weights.

(b) First PC in trained & initial CKs.

(c) First PC in trained & initial NTKs.

Figure 27: Alignment between the leading PC of the weight/kernel matrices and the signal $\boldsymbol{\beta}$ or the training labels $\boldsymbol{y}$ before/after training for the experiment in Figure 26. We can observe strong alignments, in this case, comparing with Figure 25 because of outliers in above Figure 26. These kernel alignments induce anisotropic structures in the kernel matrices during training [78].

following the same argument as [71, Setion 6.6.1], we know that

$$\begin{aligned}\|f_t(\boldsymbol{X})\| &\leq \|f_0(\boldsymbol{X})\| + \|f_t(\boldsymbol{X}) - f_0(\boldsymbol{X})\| \\ &\leq \|f_0(\boldsymbol{X})\| + \frac{\lambda_\sigma}{\sqrt{h}} \|\boldsymbol{v}\| \frac{\|\boldsymbol{X}\|}{\sqrt{d}} \|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F . \end{aligned} \tag{23}$$

Note that $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F = O(1/\sqrt{n})$ with high probability by the induction hypothesis. Hence, by Lemma C.1 and (22), we have $\|f_t(\boldsymbol{X})\| \leq C\sqrt{n}$ with high probability. Indeed, the difference between $f_t(\boldsymbol{X})$ and $f_0(\boldsymbol{X})$ is significantly negligible comparing with the initial value $f_0(\boldsymbol{X})$. Similarly with $\boldsymbol{A}_0$, $\boldsymbol{A}^t$ satisfies

$$\left\|\boldsymbol{A}^t\right\| = \left\|\boldsymbol{A}^t\right\|_F \leq \frac{\mu_1}{\sqrt{n}} \frac{\|\boldsymbol{X}\|}{\sqrt{d}} \frac{1}{\sqrt{n}} (\|\boldsymbol{y}\| + \|f_t(\boldsymbol{X})\|) \frac{\|\boldsymbol{v}\|}{\sqrt{h}}.$$

Analogously for $\boldsymbol{B}^t$, we have

$$\left\|\boldsymbol{B}^t\right\|_F \leq \frac{\mu_1 + \lambda_\sigma}{\sqrt{n}} \frac{\|\boldsymbol{X}\|}{\sqrt{d}} \frac{\|\boldsymbol{v}\|}{\sqrt{h}} \frac{1}{\sqrt{n}} (\|\boldsymbol{y}\| + \|f_t(\boldsymbol{X})\|).$$

Thus, Lemma C.1, (22), and (23) ensure that

$$\mathbb{P}\left(\left\|\boldsymbol{A}^t\right\|_F \geq \frac{C'}{\sqrt{n}}\right) \leq \exp\left(-c'n\right), \ \mathbb{P}\left(\left\|\boldsymbol{B}^t\right\|_F \geq \frac{C'}{\sqrt{n}}\right) \leq \exp\left(-c'n\right),$$

for constants $c', C' > 0$. Since $\|\boldsymbol{W}_{t+1} - \boldsymbol{W}_0\|_F \leq \|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F + \eta \left\|\boldsymbol{A}^t\right\|_F + \eta \left\|\boldsymbol{B}^t\right\|_F$, by induction hypothesis, we can conclude that (20) holds for the $(t+1)$-th step with some constants $C, c > 0$, which are different from the constants at the t-th step. □

As a corollary, by (17), we can also deduce the following norm bounds:

$$\mathbb{P}\left(\|\boldsymbol{W}_t - \boldsymbol{W}_0\| \geq \frac{C}{\sqrt{n}}\right) \leq \exp\left(-cn\right), \quad \mathbb{P}\left(\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_{2,\infty} \geq \frac{C}{\sqrt{n}}\right) \leq \exp\left(-cn\right).$$

Lemma C.2 and the above bounds are empirically verified by Figure 28(a) for $t = 3$. Not only upper bounds, this simulation also shows that at early phase $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|$, $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$, and $\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_{2,\infty}$ are all of the same $\Theta(1/\sqrt{n})$ order.

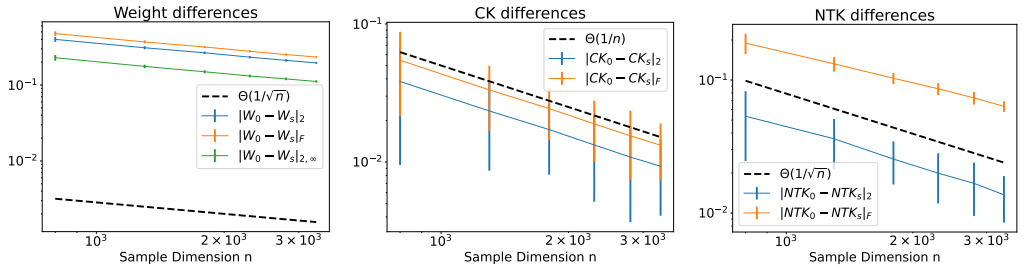
As a remark, from the bound of the second term of (23), we can deduce that the change of the output of the NN satisfies

$$|f_t(\boldsymbol{x}) - f_0(\boldsymbol{x})| \leq \frac{C}{\sqrt{n}},$$

for some t-dependent constant $C > 0$, any $\boldsymbol{x} \sim \mathcal{N}(0, \mathbf{I})$ and any finite time t. In other words, when $\eta = \Theta(1)$, the change of the output of the NN at the early phase (i.e. $t = \Theta(1)$) is negligible and its order is $O(\frac{1}{\sqrt{n}})$.

C.2 Proof of Lemma 4.1

In this section, we complete the proof of Lemma 4.1. We first mention the empirical validation of Lemma 4.1 in Figure 28. Notice that the changes in Frobenius norm for $\boldsymbol{W}$ and $\boldsymbol{K}^{\mathrm{CK}}$ are exactly $\Theta(1/\sqrt{n})$ and $\Theta(1/n)$, respectively. The operator norm of $\boldsymbol{K}^{\mathrm{NTK}}$ matches with Lemma 4.1, while the Frobenius norm of the change decays slower than the rate $\Theta(1/n)$. Additionally, in the simulation, we use $\boldsymbol{v} \sim \mathcal{N}(0, \mathbf{I})$, which indicates that our assumption for $\boldsymbol{v}$ in Lemma 4.1 can be weakened.



(a) Early phase for $\boldsymbol{W}_t$. (b) Early phase for $\boldsymbol{K}^{\mathrm{CK}}$. (c) Early phase for $\boldsymbol{K}^{\mathrm{NTK}}$.

Figure 28: Empirical validations for Lemma 4.1 and Lemma C.2 at $t = 3$. Here $\sigma_\varepsilon = 0.2$, activation σ is a normalized ReLU and the target function σ^* is normalized $\tanh$. Fix $d/n = 0.6$ and $N/n = 1.2$ as n is increasing. At each dimension, we take 25 trials to average. (a) Norms of the changes for $\boldsymbol{W}_3 - \boldsymbol{W}_0$. (b) Norms of the changes for $\boldsymbol{K}_3^{\mathrm{CK}} - \boldsymbol{K}_0^{\mathrm{CK}}$. (c) Norms of the changes for $\boldsymbol{K}_3^{\mathrm{NTK}} - \boldsymbol{K}_0^{\mathrm{NTK}}$.

Proof of Lemma 4.1. Lemma C.2 directly validates the control of $\frac{1}{\sqrt{d}} \|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$. By virtue of this result, we now present estimates for CK and NTK. Based on [71, Section 6.6.1], we have

$$\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})\right\| \leq \frac{\lambda_\sigma}{\sqrt{d}} \|\boldsymbol{X}\| \|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F. \tag{24}$$

We apply the mean value theorem to obtain this inequality. Recall the operator norm bound for Gaussian random matrix $\boldsymbol{X}$ in (22). We know $\|\boldsymbol{X}/\sqrt{d}\| \lesssim 1 + \sqrt{\gamma_1}$ with high probability as $n/d \to \gamma_1$. Hence, with the help of Lemma C.2, we can claim

$$\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})\right\| \leq C\lambda_\sigma(1 + \sqrt{\gamma_1})/\sqrt{n},$$

with probability at least $1 - \exp(-cn)$, for any fixed finite $t \in [n]$. Similarly, we can control the change in the Frobenius norm as follows:

$$\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})\right\|_F^2 \leq \frac{\lambda_\sigma^2}{d} \|\boldsymbol{X}\|^2 \|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F^2 \leq C\lambda_\sigma^2(1 + \sqrt{\gamma_1})^2/n, \tag{25}$$

with probability at least $1 - \exp(-cn)$. Therefore, we can control the change in the CK matrix in the Frobenius norm by the following inequalities:

$$\begin{aligned}
&\left\|\boldsymbol{K}_t^{\mathrm{CK}} - \boldsymbol{K}_0^{CK}\right\|_F \\
\leq\ & \frac{1}{h}\left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d})^\top\left(\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right)\right\|_F + \frac{1}{h}\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})^\top\left(\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right)\right\|_F \\
\leq\ & \frac{1}{h}\left(\left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\| + \left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d}\right\|\right) \cdot \left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|_F \\
&+ \frac{1}{h}\left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\| \cdot \left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|_F.
\end{aligned}$$

Therefore, by (24), (25) and Lemma C.1, we can claim that there exist constants $c, C > 0$ such that with probability at least $1 - \exp(-cn)$, $\left\|\boldsymbol{K}_t^{\mathrm{CK}} - \boldsymbol{K}_0^{CK}\right\|_F$ is upper bounded by C/n in the LWR.

Now we consider the change in the NTK matrix during training. Since the empirical NTK can be decomposed into two parts, one of which is exactly the CK, it suffices to consider the change of the first part of the empirical NTK. Recall that

$$\boldsymbol{K}_t := \frac{1}{d}\boldsymbol{X}^\top\boldsymbol{X} \odot \frac{1}{h}\sigma'\left(\frac{1}{\sqrt{d}}\boldsymbol{W}_t\boldsymbol{X}\right)^\top \mathrm{diag}(\boldsymbol{v}_t)^2\sigma'\left(\frac{1}{\sqrt{d}}\boldsymbol{W}_t\boldsymbol{X}\right).$$

Following the notation in [71], we denote $\mathcal{J}(\boldsymbol{W}_t) := [\mathcal{J}(\boldsymbol{w}_1^t), \ldots, \mathcal{J}(\boldsymbol{w}_N^t)] \in \mathbb{R}^{n \times hd}$ with $\mathcal{J}(\boldsymbol{w}_i) := \frac{v_i}{\sqrt{h}} \text{diag}(\sigma'(\boldsymbol{X}^\top \boldsymbol{w}_i/\sqrt{d}))\boldsymbol{X}^\top/\sqrt{d} \in \mathbb{R}^{n \times d}$. Hence, $\boldsymbol{K}_t = \mathcal{J}(\boldsymbol{W}_t)\mathcal{J}(\boldsymbol{W}_t)^\top$ and

$$\begin{aligned} \|\boldsymbol{K}_t - \boldsymbol{K}_0\| = & \left\|\mathcal{J}(\boldsymbol{W}_t)\mathcal{J}(\boldsymbol{W}_t)^\top - \mathcal{J}(\boldsymbol{W}_0)\mathcal{J}(\boldsymbol{W}_0)^\top\right\| \\ \leq & \, 2\left\|\mathcal{J}(\boldsymbol{W}_0)\right\| \left\|\mathcal{J}(\boldsymbol{W}_t) - \mathcal{J}(\boldsymbol{W}_0)\right\| + \left\|\mathcal{J}(\boldsymbol{W}_t) - \mathcal{J}(\boldsymbol{W}_0)\right\|^2. \end{aligned} \tag{26}$$

By [71, Lemma 6.6], we know $\left\|\mathcal{J}(\boldsymbol{W}_0)\right\|^2 = \left\|\boldsymbol{K}_0^{\text{NTK}}\right\|$ is upper bounded by some constant $C > 0$ with high probability. Then, we apply the inequalities from Lemma 6.5 of [71] to obtain

$$\begin{aligned} & \left\|\mathcal{J}(\boldsymbol{W}_t) - \mathcal{J}(\boldsymbol{W}_0)\right\|^2 \\ = & \left\| \left(\left(\sigma'(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma'(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right)^\top \frac{\text{diag}(\boldsymbol{v})^2}{h} \left(\sigma'(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma'(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right) \right) \odot \left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{d} \right) \right\| \\ \leq & \left\| \left(\sigma'(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma'(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right)^\top \frac{\text{diag}(\boldsymbol{v})}{\sqrt{h}} \right\|^2 \left(\max_{i\in[n]} \left\|\boldsymbol{x}_i/\sqrt{d}\right\|^2 \right) \\ \leq & \frac{1}{h}\|\boldsymbol{v}\|_\infty^2 \left\|\sigma'(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma'(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|^2 \left(\max_{i\in[n]} \left\|\boldsymbol{x}_i/\sqrt{d}\right\|^2 \right) \\ \leq & \frac{\lambda_\sigma^2}{h} \frac{\|\boldsymbol{X}\|^2}{d} \|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F^2 \left(\max_{i\in[n]} \left\|\boldsymbol{x}_i/\sqrt{d}\right\|^2 \right), \end{aligned} \tag{27}$$

where the last inequality is due to the mean value theorem, the uniform bound on σ'', and the assumption on the second layer $\boldsymbol{v}$. Notice that Gaussian random vectors satisfy

$$\mathbb{P}\left(\max_{i\in[n]} \frac{1}{d}\|\boldsymbol{x}_i\|^2 \geq 2 \right) \leq 2ne^{-cn}, \tag{28}$$

as $n/d \to \gamma_1$ and $h/d \to \gamma_2$. Thus, with (22) and Lemma C.2, we obtain

$$\mathbb{P}\left(\|\mathcal{J}(\boldsymbol{W}_t) - \mathcal{J}(\boldsymbol{W}_0)\| \geq \frac{C\lambda_\sigma(1+\gamma_1)}{n} \right) \leq 4ne^{-cn},$$

where constant C relies on the number of steps t. Hence, by (26), we can finally bound in norm the difference between the initial and the trained NTK matrices at the early phase (t is finite).

□

Corollary C.3. *For any fixed $t \in \mathbb{N}$, $i \in [d]$ and $k \in [n]$, denote λ_i^t, ν_k^t and μ_k^t the i-th, and k-th eigenvalues of $\frac{1}{h}\boldsymbol{W}_t^\top\boldsymbol{W}_t$, $\boldsymbol{K}_t^{CK}$ and $\boldsymbol{K}_t^{NTK}$, respectively. Then, under the assumptions of Lemma 4.1, we have*

$$|\lambda_i^t - \lambda_i^0|, \ |\nu_k^t - \nu_k^0|, \ |\mu_k^t - \mu_k^0| \to 0,$$

almost surely in LWR. Consequently, the eigenvalues of $\frac{1}{h}\boldsymbol{W}_t^\top\boldsymbol{W}_t$, $\boldsymbol{K}_t^{CK}$ and $\boldsymbol{K}_t^{NTK}$ are the same as corresponding the eigenvalues of initial $\frac{1}{h}\boldsymbol{W}_0^\top\boldsymbol{W}_0$, $\boldsymbol{K}_0^{CK}$ and $\boldsymbol{K}_0^{NTK}$, respectively.

This corollary is a direct outcome of Weyl's inequality from Theorem A.46 in [8]. Consequently, this corollary concludes that for any fixed $t \geq 0$, almost surely, the limiting spectra of $\frac{1}{h}\boldsymbol{W}_t^\top\boldsymbol{W}_t$, $\boldsymbol{K}_t^{\text{CK}}$ and $\boldsymbol{K}_t^{\text{NTK}}$ are the same as those of $\frac{1}{h}\boldsymbol{W}_0^\top\boldsymbol{W}_0$, $\boldsymbol{K}_0^{\text{CK}}$ and $\boldsymbol{K}_0^{\text{NTK}}$ in LWR. This corollary claims that not only does the bulk of distributions stay identical to the initialization, but also that any eigenvalues stay the same as at the initialization. This shows that the smallest eigenvalue of $\boldsymbol{K}_t^{\text{NTK}}$ has the same lower bound as $\boldsymbol{K}_0^{\text{NTK}}$ in the early phase of training.

C.3 Global Convergence for GD Under LWR

In this section, we study the final stage of (10) as training loss is approaching zero and prove Theorem 4.2. Figure 29 shows that the spectra are unchanged globally, even after training in this case. In Corollary 4.3, we confirm this observation for the weight, CK, and NTK matrices via Frobenius norm control. In the simulation, the second layer is initialized as $\boldsymbol{v} \sim \mathcal{N}(0, \mathbf{I})$, which is more general than our assumption on $\boldsymbol{v}$ in Theorem 4.2.

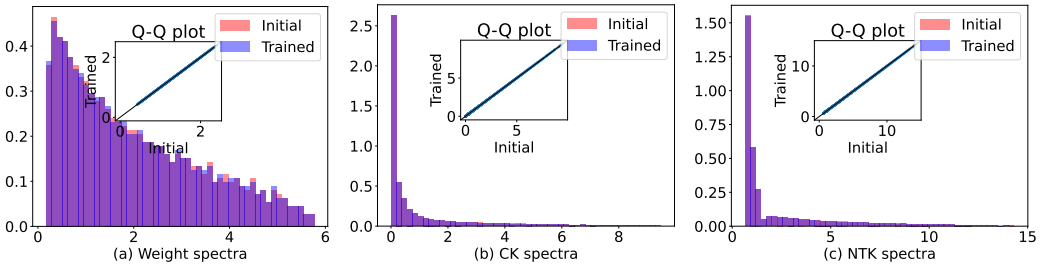


Figure 29: The initial and trained spectra (until training loss is less than 10^{-5}) when using GD only for the first layer ($h = 3000, n = 2000, d = 1000$): (a) Weight spectra. (b) $\boldsymbol{K}^{\text{CK}}$ spectra. (c) $\boldsymbol{K}^{\text{NTK}}$ spectra. The final R^2 score is 0.55964 and the test loss is 0.44724. The activation is a normalized ReLU, and the target is Sigmoid.

Proof of Theorem 4.2. Recall the Jacobian matrix $\mathcal{J}(\boldsymbol{W})$ defined in the proof of Lemma 4.1, and the definition of α based on (11) in Section 3. Denote the event

$$\mathcal{A} := \left\{ \|\boldsymbol{X}\| \leq 2\left(1+\sqrt{\gamma_1}\right)\sqrt{d},\ \max_{i\in[n]} \|\boldsymbol{x}_i\|^2 \leq 2d, \sigma_{\min}(\mathcal{J}(\boldsymbol{W}_0)) \geq 2\alpha \right\}.$$

By (22), (28) and Theorem 2.9 of [81], we have $\mathbb{P}(\mathcal{A}) \geq 1 - 2e^{-cn} - 2ne^{-cn} - n^{-7/3}$ for some constant $c > 0$ and all large n in LWR. In the following, conditionally on event $\mathcal{A}$, we will apply Theorem 6.10 of [71] to obtain the global convergence. Conditionally on $\mathcal{A}$, Lemma 6.6 of [71] implies

$$\|\mathcal{J}(\boldsymbol{W})\| \leq \lambda_\sigma \|\boldsymbol{v}\|_\infty \left\|\boldsymbol{X}/\sqrt{d}\right\| \leq 2\lambda_\sigma(1+\sqrt{\gamma_1}), \tag{29}$$

for any $\boldsymbol{W}$. Define $\beta = 2\lambda_\sigma(1+\sqrt{\gamma_1})$. Moreover, in terms of (27), we can verify the Lipschitz property for the Jacobian matrix as follows: conditionally on $\mathcal{A}$,

$$\left\|\mathcal{J}(\tilde{\boldsymbol{W}}) - \mathcal{J}(\boldsymbol{W})\right\| \leq \frac{2\beta}{\sqrt{h}} \left\|\tilde{\boldsymbol{W}} - \boldsymbol{W}\right\|_F, \tag{30}$$

for any $\tilde{\boldsymbol{W}}, \boldsymbol{W} \in \mathbb{R}^{h\times d}$. Therefore, conditionally on $\mathcal{A}$, $\mathcal{J}(\boldsymbol{W})$ is a L-Lipschitz function with respect to $\boldsymbol{W}$ where $L := \frac{2\beta}{\sqrt{h}}$. To complete the proof, it suffices to investigate the smallest singular value of $\mathcal{J}(\boldsymbol{W})$ when $\boldsymbol{W}$ is in the vicinity of $\boldsymbol{W}_0$. Recall $\ell(\boldsymbol{W}) = \|\boldsymbol{y} - f_{\boldsymbol{W}}(\boldsymbol{X})\|$. Notice that for any unit vector $\boldsymbol{u} \in \mathbb{R}^n$, we have $\boldsymbol{u}^\top f_{\boldsymbol{W}_0}(\boldsymbol{X}) = \frac{1}{\sqrt{h}} \sum_{i=1}^h v_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{X}/\sqrt{d})\boldsymbol{u}$, where $\boldsymbol{w}_i^\top$ is the i-th row of $\boldsymbol{W}_0$ for $i \in [N]$. Consider event $\mathcal{B} := \left\{ \left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\| \leq C\sqrt{n} \right\}$ for some universal constant $C > 0$. Lemma C.1 proves $\mathbb{P}(\mathcal{B}) \geq 1 - 2e^{-cn}$. By the assumption of $\boldsymbol{v}$, we know each entry v_i is a sub-Gaussian random variable with a sub-Gaussian norm at most 1. Then, according to Hoeffding's inequality, conditionally on the event $\mathcal{B}$, we have

$$\mathbb{P}\left(\left| \frac{1}{\sqrt{h}} \sum_{i=1}^h v_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{X}/\sqrt{d})\boldsymbol{u} \right| \geq t \right) \leq 2\exp\left(-ct^2\right),$$

for every $t \geq 0$ and some constant $c > 0$. Let $t = 2\sqrt{n}$. Considering an $\frac{1}{4}$-net $\mathcal{N}$ of the unit sphere $\mathbb{S}^{n-1}$, we can get

$$\mathbb{P}\left(\|f_{\boldsymbol{W}_0}(\boldsymbol{X})\| \geq \sqrt{n}\right) \leq \mathbb{P}\left(2\max_{\boldsymbol{u}\in\mathcal{N}} \left|\boldsymbol{u}^\top f_{\boldsymbol{W}_0}(\boldsymbol{X})\right| \geq \sqrt{n}\right) \leq 9^n 2\exp\left(-cn\right) \leq 2e^{-c'n}, \tag{31}$$

for some constant $c' > 0$. Hence, based on Lemma C.1 and (31), we can obtain $\ell(\boldsymbol{W}_0) \leq C_0\sqrt{n}$ with high probability for some universal constant $C_0 > 0$. Let us denote this event as $\mathcal{C} : \{\ell(\boldsymbol{W}_0) \leq C_0\sqrt{n}\}$. Define $R := 4\ell(\boldsymbol{W}_0)/\alpha$. For any $\boldsymbol{W}$ in a ball of radius R centered at $\boldsymbol{W}_0$, we have $\|\boldsymbol{W}_0 - \boldsymbol{W}\|_F \leq R$ and $\|\mathcal{J}(\boldsymbol{W}) - \mathcal{J}(\boldsymbol{W}_0)\| \leq LR$, conditionally on event $\mathcal{A}$. Thus, by (30), on event $\mathcal{A}\cap\mathcal{C}$, the smallest singular value $\sigma_{\min}(\mathcal{J}(\boldsymbol{W}))$ of the Jacobian matrix $\mathcal{J}(\boldsymbol{W})$ can be bounded by

$$\begin{aligned}
\sigma_{\min}(\mathcal{J}(\boldsymbol{W})) \geq\ & \sigma_{\min}(\mathcal{J}(\boldsymbol{W}_0)) - \|\mathcal{J}(\boldsymbol{W}) - \mathcal{J}(\boldsymbol{W}_0)\| \\
\geq\ & 2\alpha - LR \geq 2\alpha - \frac{8\beta}{\alpha}\frac{\ell(\boldsymbol{W}_0)}{\sqrt{h}} \geq 2\alpha - \frac{8C\beta}{\alpha}\sqrt{\frac{\gamma_1}{\gamma_2}},
\end{aligned}$$

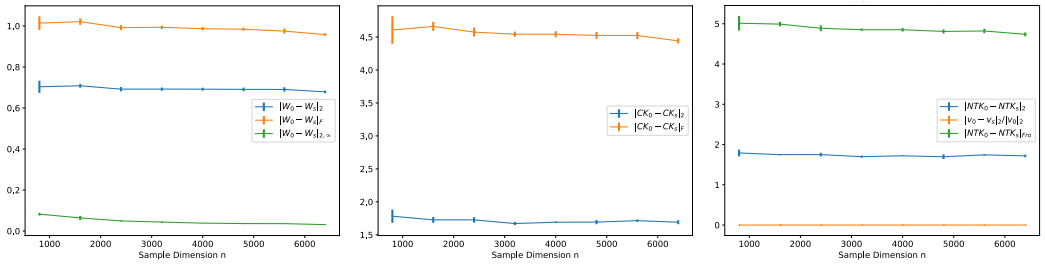
for some universal constant $C > 0$ and sufficiently large n, d, h. Notice that here constants C, β, and α do not rely on γ_2. Therefore, there exists a sufficiently large $\gamma^* > 0$ such that for all $\gamma_2 \geq \gamma^*$, we have $2\alpha - \frac{8C\beta}{\alpha}\sqrt{\frac{\gamma_1}{\gamma_2}} \geq \alpha$. In other words, when h is sufficiently large but still in the same order as n and d, for all $\|\boldsymbol{W} - \boldsymbol{W}_0\|_F \leq R$, we have $\sigma_{\min}(\mathcal{J}(\boldsymbol{W})) \geq \alpha$ conditionally on $\mathcal{C} \cap \mathcal{A}$. Combining with (29) and (30), conditionally on $\mathcal{C} \cap \mathcal{A}$, all the assumptions of Theorem 6.10 by [71] are satisfied when $\|\boldsymbol{W} - \boldsymbol{W}_0\|_F \leq R$. Therefore, when the learning rate $\frac{\eta}{n} \leq \frac{1}{\beta^2}\min\left\{1, \frac{4\alpha}{LR}\right\}$, we can get (12)-(14) for all $t \in \mathbb{N}$, conditionally on $\mathcal{C} \cap \mathcal{A}$. Both events $\mathcal{A}$ and $\mathcal{C}$ occur with high probability and only depend on initialization $\boldsymbol{W}_0$, $\boldsymbol{X}$ and $\boldsymbol{y}$. Hence we complete the proof of this theorem. Notice that since $\gamma_2 \geq \gamma^*$ is sufficiently large, $\frac{4\alpha}{LR} \geq \frac{\alpha^2}{2C\beta}\sqrt{\frac{\gamma_2}{\gamma_1}} > 1$. Therefore, it suffices to require $\eta \leq n/\beta^2$ to conclude that (12), (13) and (14) hold with high probability. This completes the proof. Moreover, (14) further shows that for all $t \in \mathbb{N}$,

$$\|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F \leq R \leq C\sqrt{n} + o_{d,\mathbb{P}}(1), \tag{32}$$

where we again apply Lemma C.1 in the following way:

$$\ell(\boldsymbol{W}_0) \leq C\sqrt{n} + o_{d,\mathbb{P}}(1),$$

for some constant $C > 0$ only depending on $\gamma_1, \gamma_2, \sigma_\varepsilon, \sigma$ and σ^*. □



(a) Changes for $\frac{1}{\sqrt{d}}\boldsymbol{W}$. (b) Changes for CK. (c) Changes for NTK.

Figure 30: Measuring the change for the weight, CK, and NTK matrices when training NN with (10). We fix $d/n = 1.2$ and $h/n = 0.6$ when n is increasing. Here, σ is normalized ReLU and the target is normalized $\tanh$. The largest $n = 6400$ and the learning rate $\eta = 5.0$ for all training processes. We train each neural network until the training losses approach zero. Each experiment repeats 4 times. In (a), we consider the changes $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|$,$\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$, and $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_{2,\infty}$.

As a corollary, (14) controls the deviation of the final step weight from the initial weight. This is empirically shown in Figure 30(a), which shows that $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|$,$\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$, and $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_{2,\infty}$ are $\Theta(1)$ when trainable parameters are convergent. This implies that the final $\boldsymbol{W}_t$ is still close to the initial weight $\boldsymbol{W}_0$, even after training. Consequently, with this observation, we can prove Corollary 4.3 in the following.

Proof of Corollary 4.3. Based on (32), we know $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_o - \boldsymbol{W}_t\|_F \leq C_0$ holds with high probability for some universal constant $C_0 > 0$. Conditionally on this event, we can then estimate changes in CK and NTK after training. The method is analogous to Lemma 4.1. For CK, we employ Lemma C.1 and (25) to get

$$\begin{aligned}
&\left\|\boldsymbol{K}_t^{\mathrm{CK}} - \boldsymbol{K}_0^{CK}\right\|_F \\
\leq\ & \frac{2}{h}\left(\left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\| + \left\|\sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d}\right\|\right) \cdot \left\|\sigma(\boldsymbol{W}_t\boldsymbol{X}/\sqrt{d}) - \sigma(\boldsymbol{W}_0\boldsymbol{X}/\sqrt{d})\right\|_F \\
\lesssim\ & \frac{2\lambda_\sigma^2(1+\sqrt{\gamma_1})^2}{h}\|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F^2 + \frac{2C\sqrt{n}\lambda_\sigma(1+\sqrt{\gamma_1})}{h}\|\boldsymbol{W}_0 - \boldsymbol{W}_t\|_F \\
\lesssim\ & \frac{2\lambda_\sigma(1+\sqrt{\gamma_1})C_0}{\gamma_2^2}\left(\lambda_\sigma(1+\sqrt{\gamma_1})C_0 + C\sqrt{\gamma_1}\right) = O_{d,\mathbb{P}}(1).
\end{aligned}$$

Hence, this shows control of the change for the CK matrix after training, compared with the initial CK.

Let us denote $\boldsymbol{w}_i^t \in \mathbb{R}^{1\times d}$ as the i-th row of $\boldsymbol{W}_t$, and $\boldsymbol{x}_j$ as the j-th column of $\boldsymbol{X}$. Additionally, by Assumption 3.2, we know that

$$|\sigma'(x) - \sigma'(y)| \leq \lambda_\sigma |x - y|, \tag{33}$$

for any $x, y \in \mathbb{R}$. For NTK, by modifying (27), one can deduce that

$$\begin{aligned}
\|\mathcal{J}(\boldsymbol{W}_t) - \mathcal{J}(\boldsymbol{W}_0)\|_F^2 &= \sum_{i=1}^h \left\|\mathcal{J}(\boldsymbol{w}_i^t) - \mathcal{J}(\boldsymbol{w}_i^0)\right\|_F^2 \\
&\overset{(i)}{\leq} \frac{1}{h}\sum_{i=1}^h \left\|\text{diag}\left(\sigma'(\boldsymbol{w}_i^t\boldsymbol{X}/\sqrt{d} - \sigma'(\boldsymbol{w}_i^0\boldsymbol{X}/\sqrt{d})\right)\right\|_F^2 \left\|\frac{\boldsymbol{X}}{\sqrt{d}}\right\|^2 \\
&\overset{(ii)}{\leq} \frac{(1+\sqrt{\gamma_1})^2}{h}\sum_{i=1}^h \left\|\text{diag}\left(\sigma'(\boldsymbol{w}_i^t\boldsymbol{X}/\sqrt{d} - \sigma'(\boldsymbol{w}_i^0\boldsymbol{X}/\sqrt{d})\right)\right\|_F^2 + o_{d,\mathbb{P}}(1) \\
&\overset{(iii)}{\leq} \frac{\lambda_\sigma^2(1+\sqrt{\gamma_1})^2}{h}\sum_{i=1}^h\sum_{j=1}^n \left(\frac{1}{\sqrt{d}}(\boldsymbol{w}_i^t - \boldsymbol{w}_i^0)\boldsymbol{x}_j\right) + o_{d,\mathbb{P}}(1) \\
&\overset{(iv)}{\leq} \frac{\lambda_\sigma^2(1+\sqrt{\gamma_1})^4}{h}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F^2 + o_{d,\mathbb{P}}(1) \leq \frac{\lambda_\sigma^2(1+\sqrt{\gamma_1})^4 C_0^2}{\gamma_2} + o_{d,\mathbb{P}}(1),
\end{aligned}$$

where (i) is because of [80, Exercise 6.3.3] and the assumption on $\boldsymbol{v}$, (ii) is due to (22), (iii) is due to the definition of Frobenius norm and (33), and (iv) is due to [80, Exercise 6.3.3] and (22). As a result, from (26), we can finally conclude that $\left\|\boldsymbol{K}_t^{\text{CK}} - \boldsymbol{K}_0^{CK}\right\|_F = O_{d,\mathbb{P}}(1)$ as $n/d \to \gamma_1$ and $h/d \to \gamma_2$.

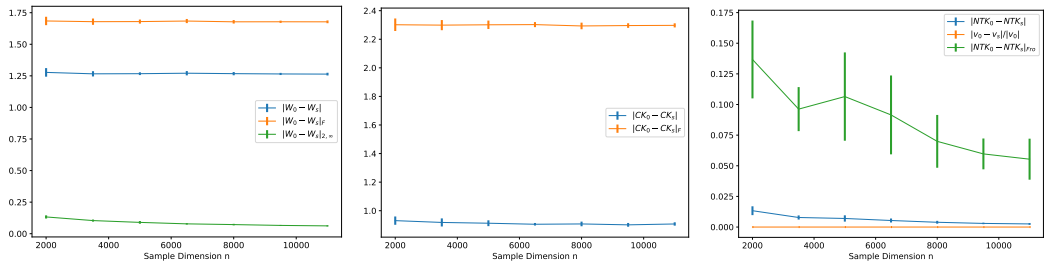
As for the limiting spectra of weight and kernel matrices, since we know that

$$\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F, \ \left\|\boldsymbol{K}_t^{\text{CK}} - \boldsymbol{K}_0^{\text{CK}}\right\|_F, \ \left\|\boldsymbol{K}_t^{\text{NTK}} - \boldsymbol{K}_0^{\text{NTK}}\right\|_F = O_{d,\mathbb{P}}(1),$$

we can automatically apply Corollary A.41 of [8]. This directly implies that the limiting empirical spectra of $\frac{1}{h}\boldsymbol{W}_t^\top\boldsymbol{W}_t$, $\boldsymbol{K}_t^{\text{CK}}$ and $\boldsymbol{K}_t^{\text{NTK}}$ are the same as the limiting spectra of $\frac{1}{h}\boldsymbol{W}_0^\top\boldsymbol{W}_0$, $\boldsymbol{K}_0^{\text{CK}}$ and $\boldsymbol{K}_0^{\text{NTK}}$, respectively, as $n/d \to \gamma_1$ and $h/d \to \gamma_2$ (see Figure 29). □

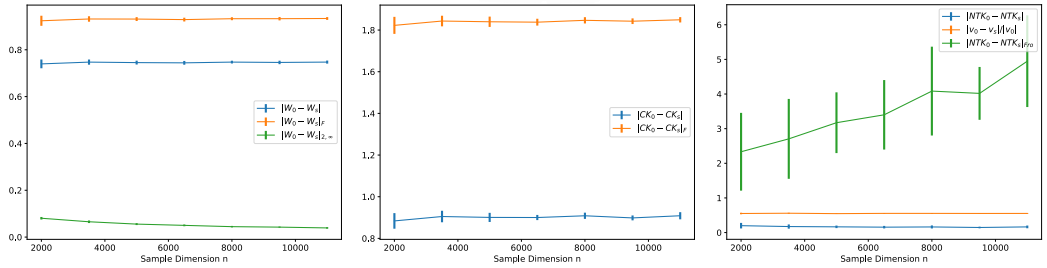
Further Simulations for Changes in Norms. The simulation of Figure 30 empirically coincides with the norm bounds in Theorem 4.2 for different norms. Because of (17), it suffices to only consider the Frobenius norm of the change for each matrix. As a remark, Theorem 4.2 requires γ_2 to be larger than some threshold γ^* to ensure norms of the change throughout training. However, Figure 30 indicates Theorem 4.2 still holds even when γ^* is small i.e. the width h is not very large. Here in Figure 30, $\gamma_2 = \frac{1}{2}\gamma_1 = 0.6$. Figure 31 suggests that similar results to Theorem 4.2 and Corollary 4.3 hold for SGD when training the first layer. This is also akin to Figure 21. Moreover, we further conjecture that similar results to Theorem 4.2 and Corollary 4.3 will hold even when training both layers (3). Denote the second layer at step t by $\boldsymbol{v}_t$. Then, indicated by Figure 32, $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|, \frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F$, $\frac{1}{\sqrt{d}}\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_{2,\infty}$, $\left\|\boldsymbol{K}_t^{\text{CK}} - \boldsymbol{K}_0^{\text{CK}}\right\|$, $\left\|\boldsymbol{K}_t^{\text{CK}} - \boldsymbol{K}_0^{\text{CK}}\right\|_F$, $\left\|\boldsymbol{K}_t^{\text{NTK}} - \boldsymbol{K}_0^{\text{NTK}}\right\|$, and $\|\boldsymbol{v}_t - \boldsymbol{v}_0\| / \|\boldsymbol{v}_0\|$ are all $\Theta(1)$ in LWR. Since $\|\boldsymbol{v}_0\| = O(\sqrt{h})$, we observe that $\|\boldsymbol{v}_t - \boldsymbol{v}_0\| = \Theta(\sqrt{h})$ in this case. Meanwhile, unlike Corollary 4.3, Figure 32(c) cannot verify that $\left\|\boldsymbol{K}_t^{\text{NTK}} - \boldsymbol{K}_0^{\text{NTK}}\right\|_F$ is still upper bounded by a constant. One possible explanation is that in this case the change in the second layer $\boldsymbol{v}_t$ is much more significant than in the first-layer weight $\boldsymbol{W}_t$, hence the NTK matrix may change a lot in this general case.

Similarly, Figure 33 shows norms of the change when training the NN with SGD for both layers. Here we use the same batch size 128 and learning rate $\eta = 1.0$ for all experiments when varying the dimension n but fixing the aspect ratios γ_1 and γ_2. All the observations are similar to the results of Corollary 4.3 except the Frobenius norm of the change for NTK. Figure 33(c) indicates that $\left\|\boldsymbol{K}_t^{\text{NTK}} - \boldsymbol{K}_0^{\text{NTK}}\right\|_F$ will not be $\Theta(1)$ anymore, which fluctuates more in this case.



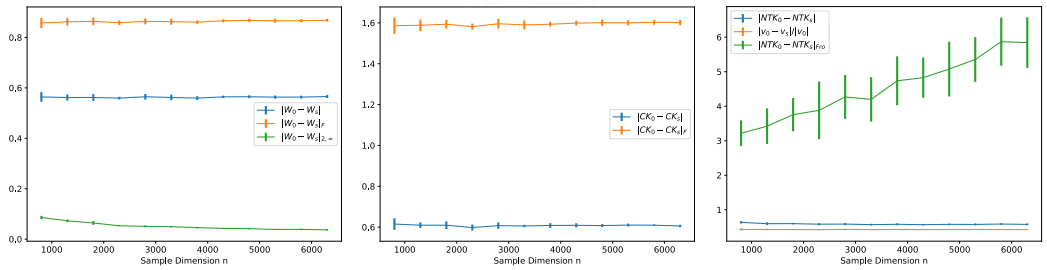
(a) Changes for $\frac{1}{\sqrt{d}}\boldsymbol{W}$. (b) Changes for CK. (c) Changes for NTK.

Figure 31: Measuring the change for weight $\boldsymbol{W}_t$, $\boldsymbol{K}^{\text{CK}}$, and $\boldsymbol{K}^{\text{NTK}}$ matrices when training NN with (10) for the first layer $\boldsymbol{W}_t$ using SGD with batch size 200. We fix $d/n = 0.6$ and $h/n = 1.2$ when n is increasing. Here σ is normalized softplus and the target is normalized $\tanh$. The largest n is 11000, $\sigma_\varepsilon = 0.3$ and the learning rate is $\eta = 3.6$ for all training processes. We train each neural network until the training losses approach zero. Each experiment repeats 15 times.



(a) Changes for $\frac{1}{\sqrt{d}}\boldsymbol{W}$. (b) Changes for CK. (c) Changes for NTK.

Figure 32: Measuring the change for $\boldsymbol{W}_t$, $\boldsymbol{K}^{\text{CK}}$, the second layer $\boldsymbol{v}_t$ and $\boldsymbol{K}^{\text{NTK}}$ when training NN with (3) for both layer $\boldsymbol{W}_t$ and $\boldsymbol{v}_t$ using GD. We fix $d/n = 0.5$ and $h/n = 0.8$ when n is increasing. Here, σ is normalized softplus and the target is normalized $\tanh$. The largest n is 11000, $\sigma_\varepsilon = 0.3$ and the learning rate is $\eta = 3.6$ for all training processes. We train each neural network until the training losses approach zero. Each experiment repeats 15 times.



(a) Changes for $\frac{1}{\sqrt{d}}\boldsymbol{W}$. (b) Changes for CK. (c) Changes for NTK.

Figure 33: Measuring the change for $\boldsymbol{W}_t$, $\boldsymbol{K}^{\text{CK}}$, the second layer $\boldsymbol{v}_t$ and $\boldsymbol{K}^{\text{NTK}}$ when training NN with SGD for both layers $\boldsymbol{W}_t$ and $\boldsymbol{v}_t$. We fix $d/n = 0.6$ and $h/n = 1.2$ when n is increasing. Here, σ is normalized ReLU and the target is normalized $\tanh$. $\sigma_\varepsilon = 0.1$ and the learning rate is $\eta = 1.0$ for all training processes. We train each neural network until the training losses approach zero. Each experiment repeats 12 times.