SCALING LAW OF LABEL-NTK ALIGNMENT AND A TIGHTER CONVERGENCE BOUND IN THE NTK REGIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural tangent kernel (NTK) provides a central framework for understanding the training dynamics of wide neural networks. In this work, we empirically identify a scaling law that often holds on real-world datasets, governing the alignment of training labels and label residuals with NTK eigenspaces. Specifically, the squared eigen-components of labels and label residuals scale approximately as λ_i^2 and λ_i, respectively, where λ_i is the corresponding NTK eigenvalue. We theoretically explain this phenomenon through the Lipschitz continuity of real-world label functions. Finally, leveraging this scaling law, we derive a tighter convergence bound for gradient descent in the NTK regime. Unlike existing bounds that depend on the smallest NTK eigenvalue, which is typically extremely small and leads to pessimistic convergence rates, our bound depends on the entire spectrum and closely matches empirical training curves.

1 INTRODUCTION

Neural tangent kernel (NTK) has become a central framework for understanding the optimization of neural networks. Before the development of NTK-based theory, the optimization landscape of neural networks was generally viewed as highly nonconvex, with complex training dynamics that were difficult to characterize theoretically. NTK theory revealed that, in the infinite-width or sufficiently wide regime, neural networks evolve approximately linearly in their parameters (Lee et al., 2019; Liu et al., 2020). This linearization makes the training dynamics analytically tractable and, under appropriate conditions, yields global convergence with exponential rates (Allen-Zhu et al., 2019; Du et al., 2019a; Oymak & Soltanolkotabi, 2020; Liu et al., 2022a). Despite its limitations in capturing feature evolution and nonlinear optimization phenomena observed in practice, such as Edge of Stability (Cohen et al., 2021) and catapult dynamics (Lewkowycz et al., 2020), NTK-based theory remains one of the most foundational and tractable frameworks for analyzing and understanding neural network optimization.

Given its central role in optimization theory, substantial effort has been devoted to understanding the properties of the NTK, with particular attention to its spectrum. Several studies have revealed strong similarity between the NTK and classical kernels such as the Laplace kernel (Geifman et al., 2020; Chen & Xu, 2021). Its eigenvalues typically exhibit polynomial decay toward zero, and although the smallest eigenvalue is often positive (Du et al., 2019b), it can be extremely small in magnitude (Nguyen et al., 2021; Velikanov & Yarotsky, 2021). The eigenvectors or eigenfunctions of the NTK have also been characterized explicitly for certain simple data distributions (Murray et al., 2023). These spectral properties have important implications and play a key role in the theoretical analysis of neural network optimization.

An important but less understood aspect of the NTK spectrum is its correlation with the training labels. Rather than contributing equally across eigenspaces - or being governed by the smallest NTK eigenvalue, as in many worst-case analyses - the labels may be preferentially aligned with certain parts of the NTK spectrum. For example, Arora et al. (2019) empirically observed that the labels tend to have stronger correlations with the leading NTK eigenvectors than with those associated with other eigenvalues. A related line of work has studied the scalar quantity $\mathbf{y}^\top K\mathbf{y}$, or its normalized variants, as a measure of label-NTK alignment and investigated how this quantity evolves during training (Atanasov et al., 2022; Shan & Bordelon, 2021). However, these scalar alignment measures do not characterize how the label information is distributed across individual NTK eigenspaces.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

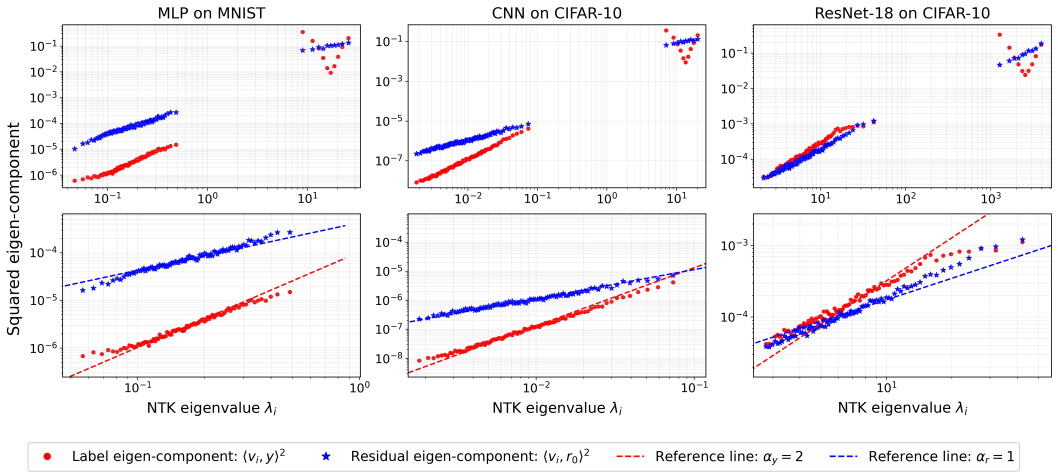


Figure 1: The scaling law of label-NTK alignment. In each plot, both axes are log-scaled, and the relation becomes approximately linear. Moreover, $\alpha_y \approx 2$ and $\alpha_r \approx 1$. **Top:** full spectrum. **Bottom:** zoomed in to show the bulk of NTK spectrum.

In this paper, we study this finer-grained relationship between the labels and the NTK spectrum. Specifically, we examine the magnitudes of the components of labels and label residuals along individual NTK eigenvectors, $\langle \mathbf{v}_i, \mathbf{y}\rangle$ and $\langle \mathbf{v}_i, \mathbf{r}\rangle$, and investigate how they depend on the corresponding NTK eigenvalues λ_i. Across a wide range of neural network architectures and real-world datasets, we consistently observe a striking scaling law:

$$\langle \mathbf{v}_i, \mathbf{y}\rangle^2 \sim \lambda_i^{\alpha_y}, \qquad \langle \mathbf{v}_i, \mathbf{r}\rangle^2 \sim \lambda_i^{\alpha_r}, \qquad \forall i \in [n], \tag{1}$$

where "$\sim$" indicates proportional scaling up to a multiplicative constant. In other words, the squared eigen-components of labels and label residuals exhibit approximately power-law dependence on the NTK eigenvalues. As illustrated in Figure 1, this relationship appears as an approximately linear trend in a log-log plot. Surprisingly, the exponents are highly consistent across architectures and datasets: we empirically find α_y to be close to 2 and α_r to be close to 1. We refer to this phenomenon as the *scaling law of label-NTK alignment*.

We further propose a theoretical explanation for why such a relationship emerges in real-world data. First, We observe empirically that the scaling law is not a generic property for arbitrary datasets: for example, it can break down when the labels are randomly assigned. This observation suggests that the scaling law is tied to an intrinsic property of labels of real-world datasets. Our analysis shows that it can be derived from a natural regularity assumption on real-world data: nearby inputs tend to have similar labels, which can be formalized through Lipschitz continuity of the underlying label function $y(\mathbf{x})$. Under this assumption, our theory predicts $\alpha_y = 2$, precisely matching the exponent observed empirically. In contrast, when labels are randomly assigned and the label function no longer satisfies the Lipschitz continuity, our theory shows that the predicted scaling law disappears, consistent with our empirical observations.

Beyond being a striking phenomenon in its own right, the scaling law has important implications for the optimization theory of neural networks. Prior NTK-based convergence analyses typically yield rates of the order $O\left(\exp(-C\eta\lambda_{\min}t)\right)$, which explicitly depends on the smallest NTK eigenvalue $\lambda_{\min}$. This dependence arises because these analyses have to consider a worst-case scenario where the training dynamics dominantly occurs in the eigenspace of $\lambda_{\min}$. Now, this scaling law shows that this worst-case scenario does not happen in real situations: the label and residual components associated with smaller eigenvalues are systematically weaker and these eigendirections have a correspondingly smaller contribution to the training dynamics. Leveraging this scaling law, we are able to avoid the overly pessimistic dependence on $\lambda_{\min}$ and obtain a much sharper characterization of the convergence of neural networks, based on existing NTK-based analysis:

Theorem 1.1 (Informal version of Theorem 5.4)**.** *Suppose the scaling law of label-NTK alignment holds. For sufficiently wide networks, gradient descent with constant step size η on the squared loss $\mathcal{L}(\mathbf{w}_t)$ satisfies*

$$\mathcal{L}(\mathbf{w}_t) \sim \mathrm{tr}[(I - \eta K)^{2t} K]. \tag{2}$$

This result, no longer relying on the smallest NTK eigenvalue, provides a substantially sharper characterization of the training dynamics than prior convergence bounds on neural networks. Empirically, we observe that this convergence bound is tight and closely matches the practical training curves (See Figure 3 for example).

Summary of Contributions.

- We identify and empirically characterize a novel scaling law describing how the label and residual components distributed across NTK eigenspaces depend on the corresponding NTK eigenvalues.
- We show that the observed scaling law can be explained by the Lipschitz continuous of label functions of real data. Our theory precisely predicts the exponent $\alpha_y = 2$, consistent with empirical observations, and explains why the scaling law breaks down for randomly assigned labels.
- Leveraging this scaling law, we developed a sharper convergence rate of gradient descent in the NTK regime. This improved rate provides is substantially tighter and closely matches empirical training curves.

1.1 Related Work

Optimization and spectral analyses in the NTK regime: Neural-network optimization has been extensively studied through the NTK and related linearization frameworks (Jacot et al., 2018; Lee et al., 2019; Liu et al., 2020; 2022b; Zhu et al., 2022). In the infinite-width limit, gradient-based training can be approximated by kernel gradient descent with a kernel that remains close to its initialization (Jacot et al., 2018; Lee et al., 2019; Chizat et al., 2019; Geiger et al., 2020). This perspective has enabled a broad class of convergence guarantees for sufficiently over-parameterized neural networks (Du et al., 2019b;a; Allen-Zhu et al., 2019; Zou & Gu, 2019; Zou et al., 2020; Oymak & Soltanolkotabi, 2020; Nguyen & Mondelli, 2020; Ji & Telgarsky, 2020). Many of these analyses characterize worst-case convergence through the smallest positive NTK eigenvalue, $\lambda_{\min}$ (Jacot et al., 2018; Du et al., 2019b;a; Lee et al., 2019), which can lead to pessimistic predictions when the empirical spectrum contains very small eigenvalues.

Several works move beyond minimum-eigenvalue analyses by exploiting the full NTK spectrum. Under idealized spherical data distributions, Basri et al. (2019); Geifman et al. (2020); Murray et al. (2023) characterized the spectral structure of neural kernels, including eigenvalue decay and spherical-harmonic eigenfunctions. This structure was used by Basri et al. (2019) to derive refined convergence rates for two-layer networks. In a related direction, Arora et al. (2019) expressed optimization and generalization quantities in terms of both NTK eigenvalues and label components along the corresponding eigendirections. More broadly, Bordelon et al. (2020); Canatar et al. (2021) emphasized that learning behavior depends jointly on the kernel spectrum and the distribution of the target across spectral modes. Our work builds on this spectral perspective, but focuses on empirically characterizing how target-related quantities themselves scale across the NTK eigenspectrum on real-world datasets.

NTK Alignment: Prior work has studied several notions of alignment between neural-network targets and kernels. Arora et al. (2019) observed that labels tend to have stronger components along leading NTK eigendirections, suggesting that these directions can dominate early optimization. Other works study aggregate alignment quantities such as $\mathbf{y}^\top K \mathbf{y}$, or normalized variants, and their evolution during training (Atanasov et al., 2022; Shan & Bordelon, 2021; Jiang et al., 2025; Geiger et al., 2020). For example, under small initialization, Atanasov et al. (2022) identified training phases associated with changes in normalized kernel-target alignment, while Shan & Bordelon (2021) provided theoretical analysis of related alignment behavior in shallow networks. More recent work Jiang et al. (2025) tracks alignment between individual target components and kernel eigendirections over training and studies how these components shift toward leading directions under larger learning rates. Our study is complementary. We investigate the eigen-components of labels onto each eigen-direction and how they depend on the corresponding NTK eigenvalues.

2 Preliminaries

Consider a neural network f, parameterized by $\mathbf{w}$, that takes an input $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ and outputs a scalar prediction $f(\mathbf{w}; \mathbf{x})$. Denote the derivative of the network output with respect to the parameters

$\mathbf{w}$ as $\nabla_{\mathbf{w}} f$. The neural tangent kernel (NTK) (Jacot et al., 2018) is defined as

$$\mathcal{K}(\mathbf{w}; \mathbf{x}_1, \mathbf{x}_2) = \langle \nabla_{\mathbf{w}} f(\mathbf{w}; \mathbf{x}_1), \nabla_{\mathbf{w}} f(\mathbf{w}; \mathbf{x}_2) \rangle , \tag{3}$$

where $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ are arbitrary inputs.

Given a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{X}$ denotes the input and $y_i \in \mathbb{R}$ its corresponding label, the NTK can be evaluated for each pair of training inputs, yielding an NTK matrix $K \in \mathbb{R}^{n \times n}$ whose entries are given by $K_{ij}(\mathbf{w}) = \mathcal{K}(\mathbf{w}; \mathbf{x}_i, \mathbf{x}_j)$. Let $J(\mathbf{w}) \in \mathbb{R}^{n \times p}$ be the Jacobian matrix, where the i-th row is $J_i(\mathbf{w}) \triangleq \nabla_{\mathbf{w}} f(\mathbf{w}; \mathbf{x}_i)^\top$. The NTK matrix $K(\mathbf{w}) \in \mathbb{R}^{n \times n}$ can be written as $K(\mathbf{w}) = J(\mathbf{w}) J(\mathbf{w})^\top$.

Since K is symmetric and positive semi-definite, it admits the eigen-decomposition

$$K = V \Lambda V^\top = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top , \tag{4}$$

where eigenvalues are ranked in descending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ and $\mathbf{v}_i$'s denote the corresponding orthonormal eigenvectors.

The spectrum of NTK matrix has been extensively studied in recent years. Du et al. (2019b) theoretically showed that all NTK eigenvalues are positive for ReLU networks provided that no pair of data inputs is parallel. Empirically, the NTK spectrum is also often observed to exhibit a small number of large outlier eigenvalues separated from a bulk of substantially smaller eigenvalues (Fan & Wang, 2020; Murray et al., 2023). Theoretically, the NTK spectrum has been shown to resemble that of a Laplace kernel (Geifman et al., 2020; Chen & Xu, 2021), with eigenvalues that decay according to a power law as the eigenvalue index increases. Consequently, the smallest NTK eigenvalue is typically extremely small (Nguyen & Mondelli, 2020; Velikanov & Yarotsky, 2021).

In this paper, we study the relationship between the NTK spectrum and the training labels. Let $\mathbf{f} \triangleq (f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n))^\top \in \mathbb{R}^n$ denote the vector of model outputs on the training dataset and $\mathbf{y} \triangleq (y_1, y_2, \ldots, y_n)^\top \in \mathbb{R}^n$ denote the label vector. We further define the *label residual* as

$$\mathbf{r} \triangleq \mathbf{f} - \mathbf{y}. \tag{5}$$

For notational simplicity, we omit the dependence of $\mathbf{f}$ and $\mathbf{r}$ on the model parameters $\mathbf{w}$ when the context is clear. Specifically, we investigate the alignments of the labels $\mathbf{y}$ and label residuals $\mathbf{r}$ with each of the eigenvectors $\mathbf{v}_i$ of NTK, which are measured by the component squares:

$$\langle \mathbf{v}_i, \mathbf{y} \rangle^2, \quad \text{and} \quad \langle \mathbf{v}_i, \mathbf{r} \rangle^2, \quad \text{respectively.} \tag{6}$$

The correlation between the label vectors and the NTK spectrum is not yet fully understood. As a result, existing NTK-based optimization theories for wide neural networks (Allen-Zhu et al., 2019; Du et al., 2019a; Oymak & Soltanolkotabi, 2020; Liu et al., 2022a) have to accommodate the least favorable case, in which $\mathbf{y}$ or $\mathbf{r}$ is aligned with the eigenvector $\mathbf{v}_{\min} \triangleq \mathbf{v}_n$ corresponding to the smallest NTK eigenvalue $\lambda_{\min} \triangleq \lambda_n$. This causes the $\lambda_{\min}$-eigenspace to dominate the optimization dynamics, resulting in two major limitations: these theories yield very slow convergence rates of $O(\exp(-C\lambda_{\min} t))$ for some constant $C > 0$ and require the network width to be as large as $\Omega(1/\lambda_{\min})$. Both limitations are particularly severe in practice because $\lambda_{\min}$ can be extremely small. Arora et al. (2019) initially observed in experiments that the label vector $\mathbf{y}$ often has large components in the eigenspaces associated with a few large eigenvalue outliers, and consequently formulated their convergence guarantee in terms of the eigen-components $\langle \mathbf{v}_i, \mathbf{y} \rangle$. However, without a characterization of these eigen-components, particularly those associated with the bulk of the NTK spectrum, the resulting convergence rate remains unclear.

3 Scaling law of label-NTK alignment

In this section, we empirically characterize the relationship between the label vectors $\mathbf{y}$ and $\mathbf{r}$ and the NTK spectrum on real-world datasets. In Section 4, we provide a theoretical explanation for it.

We measure the alignment of the label vector $\mathbf{y}$ and residual vector $\mathbf{r}$ with each NTK eigenspace through their squared eigen-components, $\langle \mathbf{v}_i, \mathbf{y} \rangle^2$ and $\langle \mathbf{v}_i, \mathbf{r} \rangle^2$, where a larger value indicates stronger alignment with the corresponding eigenspace. For a generic vector $\mathbf{v}$, there is no inherent reason for $\langle \mathbf{v}_i, \mathbf{v} \rangle^2$ to depend on λ_i; these squared eigen-components would be independent of

the NTK eigenvalues, and on average being uniformly distributed across λ_i. Furthermore, to obtain worst-case convergence bound, many NTK-based theories often need to account for the worst case that $\langle \mathbf{v}_{\min}, \mathbf{v} \rangle^2$ dominates over the others, resulting in convergence rates governed by $\lambda_{\min}$.

In fact, we empirically observe that this alignment exhibits a favorable behavior on real-world datasets: the label and label residual vectors tend to be more strongly aligned with eigenspaces associated with larger NTK eigenvalues, while their components along eigenspaces with smaller eigenvalues are systematically weaker. More surprisingly, this dependence follows a power law over the bulk of the NTK spectrum when averaged over batch randomness:

$$\langle \mathbf{v}_i, \mathbf{y} \rangle^2 \sim \lambda_i^{\alpha_y}, \quad \langle \mathbf{v}_i, \mathbf{r} \rangle^2 \sim \lambda_i^{\alpha_r}, \quad \forall i \in [n]. \tag{7}$$

We refer to this phenomenon as the *scaling law of label-NTK alignment*. Furthermore, we consistently observe that the scaling exponents are remarkably $\alpha_y \approx 2$ and $\alpha_r \approx 1$.

Figure 1 illustrates this phenomenon on various network architectures: fully connected networks, CNNs, and ResNets. We see that the scaling law holds tightly – almost linear on the log-log plots – for both labels $\mathbf{y}$ and label residuals $\mathbf{r}$ on almost all NTK eigenvalues except for a small number of large-eigenvalue outliers. In addition, their slopes are close to $\alpha_y \approx 2$ and $\alpha_r \approx 1$, respectively. The same phenomenon holds across a broad range of real-world datasets, including MNIST, CIFAR-10, SVHN, Fashion-MNIST, and Tiny ImageNet; additional results are provided in Appendix D.1.

We observe that the scaling law also extends to transformer architectures, when the NTK is evaluated only with respect to the embedding-related parameters W_V, W_O, and the parameters of the fully-connected feed-forward modules. More numerical results are provided in Appendix D.2. Moreover, the scaling law also seems to persist beyond initialization across extended training time. This is interesting, because even though the NTK evolves during training, this power law scaling relation still approximately hold. More results can be found in Appendix D.3.

Remark 3.1. Prior work by Arora et al. (2019) made initial qualitative observations that the top few NTK eigenvalue outliers often have larger label components $\langle \mathbf{v}_i, \mathbf{y} \rangle^2$ than the remainder of the spectrum. Our observation goes beyond: we identify a systematic and quantitative dependence, given by Eq. (7), between the label components and NTK eigenvalues throughout the bulk of the spectrum. As we will see later, this bulk-spectrum relationship has a greater impact on the optimization theory.

We note that the scaling law of label-NTK alignment, Eq.(7), does not generally hold for arbitrary datasets, especially those are not real. To demonstrate this, we replace the labels of real-world datasets with independently sampled random labels while keeping the inputs unchanged. As shown in Figure 2, the eigen-components become approximately independent of λ_i, and the scaling law no longer holds. Additional results are provided in Appendix D.4. This counterexample demonstrates that the observed scaling law is not a generic consequence of the NTK spectrum alone. Rather, it reflects additional structure present in labels of real-world datasets.

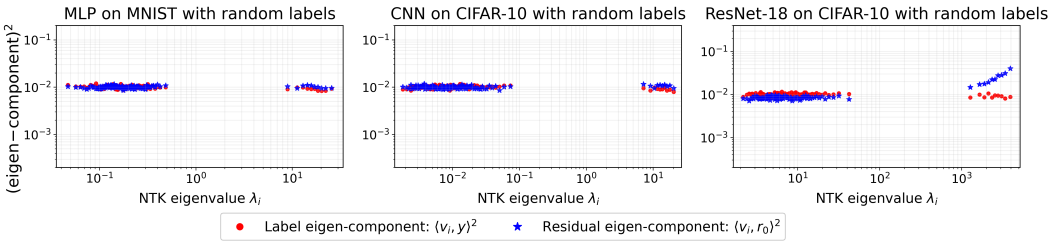


Figure 2: Datasets with random labels do not follow the scaling law. Labels of the datasets are i.i.d. randomly generated, while the inputs are kept as original.

4 A THEORETICAL EXPLANATION OF THE SCALING LAW OF LABEL-NTK ALIGNMENT

In this section, we seek to provide a theoretical explanation for the scaling law of label-NTK alignment on real-world datasets observed above.

We *hypothesize* that this scaling law of label-NTK alignment arises from "smoothness" - technically, Lipschitz continuity - of the ground-truth label function $y(\mathbf{x})$ underlying real-world datasets.

Definition 4.1 (Lipschitz continuity)**.** A function $y(\mathbf{x})$ is L_y-Lipschitz continuous, for some $L_y > 0$, if $|y(\mathbf{x}_i) - y(\mathbf{x}_j)| \leq L_y \|\mathbf{x}_i - \mathbf{x}_j\|$ hold for all $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$.

Lipschitz continuity of the ground-truth label function $y(\mathbf{x})$ is a natural assumption for many real-world prediction tasks, reflecting the intuition that *similar inputs tend to have similar labels*. For example, in image classification, a small change in an image, such as a slight change in lighting, viewpoint, or a small perturbation of its pixels, typically does not change its semantic class. More generally, if two inputs $\mathbf{x}_i$ and $\mathbf{x}_j$ are close to each other, i.e., $|\mathbf{x}_i - \mathbf{x}_j| \ll 1$, their corresponding labels are expected to be close as well, i.e., $|y(\mathbf{x}_i) - y(\mathbf{x}_j)| \ll 1$. For classification problems with discrete labels, Lipschitz continuity of the label function on the training dataset can be viewed as requiring a nonzero margin between different classes, as assumed, for example, in (Ji & Telgarsky, 2020). Note that such continuity does not generally hold for arbitrary datasets, for example when the data labels are randomly generated.

In the following, we theoretically show how the scaling law of label-NTK alignment arises from the Lipschitz continuity of label function $y(\mathbf{x})$.

Dataset with two data points. We start with a simple dataset $\mathcal{D}_2 = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\}$ that contains two data points, where the inputs $\mathbf{x}_1$ and $\mathbf{x}_2$ are very close: $0 < \|\mathbf{x}_1 - \mathbf{x}_2\| \ll 1$. We require Lipschitz continuity on the training labels, i.e., $|y_1 - y_2| \leq L_y \|\mathbf{x}_1 - \mathbf{x}_2\|$. Without loss of generality, we further assume that $\|\mathbf{x}_1\| = \|\mathbf{x}_2\| = 1$ and denote the angle between the two inputs as $\theta > 0$.

Consider the NTK of a sufficiently wide neural network at its initialization. For this dataset with two points, the NTK is a 2×2 matrix of the following form:

$$\bar{K} \triangleq \frac{1}{(L+1)^2} K = \begin{pmatrix} 1 & \cos\phi \\ \cos\phi & 1 \end{pmatrix},$$

where $\bar{K}$ is a depth-normalized version of the NTK matrix K, L is the depth of the neural network and ϕ is the angle between the two tangent vectors $\nabla f(\mathbf{x}_1)$ and $\nabla f(\mathbf{x}_2)$. The diagonal entries are 1 because the tangent vectors are shown to satisfy $\frac{1}{L+1}\|\nabla f(\mathbf{x}_i)\| = \|\mathbf{x}_i\|$ for all inputs by (Liu et al., 2025). Moreover, angles ϕ and θ are shown to have the following relation (Liu et al., 2025):

$$\cos\phi = \big(1 - \gamma\theta + o(\theta)\big)\cos\theta, \qquad \gamma := \frac{L}{2\pi}. \tag{8}$$

It is easy to see that the eigenvalues and eigenvectors of the NTK $\bar{K}$ are as follows:

$$\lambda_{\max} = 1 + \cos\phi, \;\; \mathbf{v}_{\max} = \tfrac{1}{\sqrt{2}}(1,1)^\top; \quad \lambda_{\min} = 1 - \cos\phi, \;\; \mathbf{v}_{\min} = \tfrac{1}{\sqrt{2}}(1,-1)^\top.$$

By Eq.(8), we have

$$\lambda_{\max} \geq 1, \quad \lambda_{\min} = O(\theta) \ll 1. \tag{9}$$

For the eigen-components of the label vector, we have $\langle \mathbf{v}_{\max}, \mathbf{y}\rangle^2 = \frac{(y_1+y_2)^2}{2} = O(1)$, and by the Lipschitz continuity of $y(\mathbf{x})$, we have

$$\langle \mathbf{v}_{\min}, \mathbf{y}\rangle^2 = \frac{(y_1 - y_2)^2}{2} \leq \frac{L_y^2}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 = L_y^2(1 - \cos\theta) = O(\theta^2) \ll 1. \tag{10}$$

This leads to the following theorem.

Theorem 4.2 (Scaling law of label-NTK alignment for two-point dataset)**.** *Consider the dataset $\mathcal{D}_2$ defined above with $0 < \theta \ll 1$ and a sufficiently wide fully-connected neural network of depth L. Suppose the label function $y(\mathbf{x})$ is Lipschitz continuous, then, at initialization of the neural network*

$$\langle \mathbf{v}_{\min}, \mathbf{y}\rangle^2 \sim \lambda_{\min}^2 = O(\theta^2), \;\; \langle \mathbf{v}_{\max}, \mathbf{y}\rangle^2 \sim \lambda_{\max}^2 = O(1). \tag{11}$$

The above theorem is consistent with our empirical observations made in Section 3 that the magnitude of eigen-component $|\langle \mathbf{v}_{\min}, \mathbf{y}\rangle|$ is small when the corresponding NTK eigenvalue $\lambda_{\min}$ is small. Notably, the above theory precisely predicts the scaling exponent $\alpha_y = 2$.

It is also easy to see that, when the Lipschitz continuity of $y(\mathbf{x})$ fails such as when labels are random, Eq.(10) no longer holds and the scaling law breaks down. This is also consistent with the experimental findings, as shown in Figure 2.

Dataset with multiple data points. We now consider the general case with $n > 2$ data points and show that the above result extends. We assume that there exists at least one pair of distinct data points, $\mathbf{x}_i$ and $\mathbf{x}_j$ with $i \neq j$, that are very close to each other; in this case, the NTK has eigenvalues of small magnitude. This scenario commonly arises in real-world datasets, particularly when the dataset size n is large.

Theorem 4.3. *Consider a dataset $\mathcal{D}$ of size n and a sufficiently wide fully-connected neural network of depth L at initialization. Suppose the inputs $\|\mathbf{x}_i\| = 1$ for all $i \in [n]$, $\mathbf{x}_i \nparallel \mathbf{x}_j$ for all $i \neq j \in [n]$ and there is at least a pair of data points $\mathbf{x}_i$ and $\mathbf{x}_j$, $i \neq j$ such that the angle θ between the two data points is small $\theta \ll 1$. Then, there exists eigenvalue λ and its corresponding eigenvector $\mathbf{v}$ satisfy the following:*

$$\langle \mathbf{v}, \mathbf{y} \rangle^2 \sim \lambda^2 = O(\theta^2). \tag{12}$$

The proof of the theorem is deferred to Appendix B.1.

Remark 4.4. The small eigenvalue λ and the corresponding small eigen-projection $\langle \mathbf{v}_i, \mathbf{y} \rangle^2$ arise from the small angular separation between the pair of data points $\mathbf{x}_i$ and $\mathbf{x}_j$. We expect that datasets containing multiple closely spaced pairs may exhibit multiple small-eigenvalue directions with correspondingly small label projections.

Alignment for label residual. Above, we theoretically explained the alignment between $\mathbf{y}$ and the NTK spectrum through the Lipschitz continuity of the ground-truth label function $y(\mathbf{x})$. However, this argument does not directly extend to the label residual vector $\mathbf{r} = \mathbf{f} - \mathbf{y}$, since the network prediction $\mathbf{f}$ is not necessarily Lipschitz continuous. We conjecture that a certain form of "smoothness" of the residual underlies the observed scaling law $\langle \mathbf{v}_i, \mathbf{r} \rangle^2 \sim \lambda_i^{\alpha_r}$. Developing a theoretical explanation for this $\mathbf{r}$-NTK alignment is an interesting direction for future work.

5 An Improved Convergence for Wide Neural Networks

In this section, we show that the scaling law of label-NTK alignment can be leveraged to improve the convergence rates of existing optimization theories for wide neural networks.

5.1 Limitation of Existing Optimization Theory

Consider the optimization problem of minimizing the squared loss

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}), \quad \text{with} \quad \mathcal{L}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{n} \big(f_{\mathbf{w}}(\mathbf{x}_i) - y_i\big)^2 \triangleq \frac{1}{2} \|\mathbf{f}(\mathbf{w}) - \mathbf{y}\|^2, \tag{13}$$

using gradient descent $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}_t)$, where t denotes the training timestamp and $\eta > 0$ is the learning rate.

It has been shown by Jacot et al. (2018) that, for sufficiently wide neural networks, the NTK matrix $K(\mathbf{w}_t)$ remains arbitrarily close to its value at initialization: $K(\mathbf{w}_t) \approx K(\mathbf{w}_0)$, for all t. Consequently, the learning dynamics of these neural networks is approximately linear (Lee et al., 2019; Liu et al., 2020):

$$\mathbf{r}_t = \mathbf{f}(\mathbf{w}_t) - \mathbf{y} \approx (I - \eta K)^t \big(\mathbf{f}(\mathbf{w}_0) - \mathbf{y}\big) = (I - \eta K)^t \mathbf{r}_0. \tag{14}$$

Based on these, a rich literature has developed tractable theories that guarantee convergence of GD to a global minimum.

Theorem 5.1 (Informal (Liu et al., 2022a))**.** *Consider a fully-connected neural network f with L hidden layers. If the network width satisfies $m = \tilde{\Omega}\left(\frac{nR^{6L+2}}{\lambda_{\min}^2}\right)$ with $R = \frac{8\beta\sqrt{\mathcal{L}(\mathbf{w}_0)}}{\lambda_{\min}}$ and the learning rate $\eta \leq \frac{1}{2\beta + \beta_f \sqrt{2\mathcal{L}(\mathbf{w}_0)}}$, then the square loss $\mathcal{L}(\mathbf{w})$ under gradient descent satisfies, for some constant $C > 0$:*

$$\mathcal{L}(\mathbf{w}_t) \leq (1 - C\eta\lambda_{\min})^t \mathcal{L}(\mathbf{w}_0). \tag{15}$$

The convergence rate $O(\exp(-C\eta\lambda_{\min} t))$ explicitly depends on the smallest NTK eigenvalue $\lambda_{\min}$, which can be arbitrarily small in real cases. As a result, the convergence rate becomes extremely slow and could not match the fast decay of training loss in practice, see Figure 3 for example.

Our **key observation** regarding this limitation (i.e., the dependence of convergence rate on $\lambda_{\min}$) is that such dependence is not intrinsic to the approximate constancy of the NTK or the approximately linear dynamics. Rather, it arises from the lack of knowledge about the alignment between the label/residual vectors and the NTK spectrum. To obtain a convergence guarantee, prior analyses must therefore consider in Eq. (14) the worst-case scenario, where $\mathbf{r}$ or $\mathbf{y}$ lies entirely in the eigenspace corresponding to the smallest eigenvalue $\lambda_{\min}$ leading to the $\lambda_{\min}$-dominance.

In the rest of this section, we show that the scaling law of label-NTK alignment provides precisely the missing spectral information needed to overcome this limitation and significantly improve the resulting convergence bound.

5.2 The scaling law improves the optimization theory

Condition 5.2 (Scaling law of label-NTK alignment)**.** There exist constants $c \gg \delta > 0$ such that

$$(c-\delta)\lambda_i \le \langle \mathbf{v}_i, \mathbf{r}_0 \rangle^2 \le (c+\delta)\lambda_i, \quad \forall i \in [n]. \tag{16}$$

Remark 5.3. This condition allows fluctuations around Eq.(7), to tolerate certain inexactness of the empirical scaling law. In addition, we only require that the relation holds at initialization, although we empirically observed in Section 3 that the same relation may still hold during training.

We assume the loss function $\mathcal{L}(\mathbf{w})$ is β-smooth and the network function $f(\mathbf{w})$ is β_f-smooth with $\beta, \beta_f \ge 1$. Without of loss of generality, we assume the loss at initialization $\mathcal{L}(\mathbf{w}_0) \ge 1$. Now, we are ready to present the theorem for the improved convergence rate.

Theorem 5.4. *Consider a fully-connected neural network f with L hidden layers. Suppose Condition 5.2 holds. For all $\varepsilon \in (0,1)$, if the network width satisfies $m = \tilde{\Omega}\left(\frac{nR^{6L+2}}{\lambda_{\min}^2 \varepsilon^2}\right)$ with $R = \frac{8\beta\sqrt{\mathcal{L}(\mathbf{w}_0)}}{\lambda_{\min}}$ and the learning rate $\eta \le \frac{1}{2\beta+\beta_f\sqrt{2\mathcal{L}(\mathbf{w}_0)}}$, then the square loss $\mathcal{L}(\mathbf{w})$ under gradient descent has the following convergence bounds:*

$$\frac{c-\delta}{2} tr[(I-\eta K)^{2t}K] - \frac{\varepsilon}{2} \le \mathcal{L}(\mathbf{w}_t) \le \frac{c+\delta}{2} tr[(I-\eta K)^{2t}K] + \frac{\varepsilon}{2}. \tag{17}$$

Proof. The proof is deferred to Appendix B.2. □

Remark 5.5. The analysis in (Arora et al., 2019) expresses the convergence bound in terms of the eigen-components $\langle \mathbf{v}_i, \mathbf{y} \rangle$. Plugging our scaling law of label-NTK alignment into their analysis would therefore lead to a similar improved convergence bound. However, their analysis requires a substantially larger network width, with an additional factor of $1/\kappa^2$, where κ is the initialization scale of the network and is required to be extremely small (see Theorem 4.1 of (Arora et al., 2019)). In contrast, the theorem above does not impose this additional width requirement and uses the same network-width scaling as prior optimization theories. Thus, the convergence rate established by our theorem provides a direct improvement over existing results without requiring a stronger over-parameterization condition.

The upper bound in this theorem depends on the trace of a function of the NTK K, placing greater weight on larger eigenvalues than on smaller ones. As a result, it is substantially tighter than prior bounds that depend solely on $\lambda_{\min}$. We numerically demonstrate this improvement in Figure 3 (more in Figure 10). As shown in these figures, the convergence bound in Theorem 5.4 is substantially tighter than prior bounds and matches the actual training curves much closer.

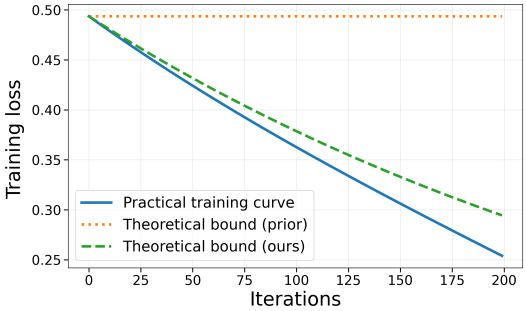


Figure 3: Comparison of theoretical convergence bounds. Setting: a five-layer fully-connected neural network trained on a subset of CIFAR-10 with 1000 samples.

Moreover, the above theorem also provides a lower bound on the convergence, which is of the same order as the upper bound, suggesting the optimality of the upper bound.

Leveraging the left inequality of Eq.(16), we have $(c-\delta)\mathtt{tr}[K] \leq 2\mathcal{L}(\mathbf{w}_0)$, and obtain the following corollary which writes out the dependence of the convergence bound on the initial loss value $\mathcal{L}(\mathbf{w}_0)$.

Corollary 5.6. *In the same setting as Theorem 5.4, the loss function during training is upper bounded by:*

$$\mathcal{L}(\mathbf{w}_t) \leq \frac{c+\delta}{c-\delta}\frac{\mathtt{tr}[(I-\eta K)^{2t}K]}{\mathtt{tr}[K]}\mathcal{L}(\mathbf{w}_0) + \frac{\varepsilon}{2}. \tag{18}$$

Note that when $t = 0$, the upper bound in Eq.(18) reduces to $\frac{c+\delta}{c-\delta}\mathcal{L}(\mathbf{w}_0) + \frac{\varepsilon}{2}$. Given that $\delta \ll c$ and $\varepsilon \ll 1$, we have this upper bound tight.

Time Complexity bound. Now, we consider the time complexity to achieve a certain loss value $\varepsilon > 0$. We additionally assume the following power law decay in NTK eigenvalues.

Assumption 5.7 (Eigenvalue power law decay)**.** NTK matrix $K \in \mathbb{R}^{n\times n}$ is positive definite, and there exist constants $C_1, C_2 > 0$ and $q > 1$ such that the following holds:

$$C_1 i^{-q} \leq \lambda_i \leq C_2 i^{-q}, \quad \forall i \in [n]. \tag{19}$$

This power law decay of NTK eigenvalues has been discovered by a large body of prior work Basri et al. (2019; 2020); Geifman et al. (2020); Velikanov & Yarotsky (2021); Murray et al. (2023) in various settings, including the special setting of uniformly distributed data in unit sphere $\mathbb{S}^d$.

Theorem 5.8. *For all $\varepsilon \in (0, 1)$, consider the same setting as Theorem 5.4. A training loss $\mathcal{L}(\mathbf{w}_T) \leq \varepsilon$ can be achieved within T_ε iterations where*

$$T_\varepsilon = \frac{1}{2\eta C_1}\left(\frac{(q-1)\varepsilon}{2(c+\delta)C_2} + n^{1-q}\right)^{\frac{q}{1-q}} \log\left(\frac{4(c+\delta)\mathcal{L}_0}{(c-\delta)\varepsilon}\right). \tag{20}$$

Proof. Proof is deferred to Appendix B.4. □

The above time complexity bound is of the following order:

$$T_\varepsilon = \min\left\{O\left(\frac{1}{\eta\varepsilon^{1+\frac{1}{q-1}}}\log\left(\frac{1}{\varepsilon}\right)\right), O\left(\frac{n^q}{\eta}\log\left(\frac{1}{\varepsilon}\right)\right)\right\}. \tag{21}$$

For comparison, the prior $\lambda_{\min}$-dependence convergence bound $O(\exp(-\eta\lambda_{\min}t))$ yields a time complexity of $O(\frac{1}{\eta\lambda_{\min}}\log\frac{1}{\varepsilon})$. As ε is typically much larger than $\lambda_{\min}$ in practice (for example, $\varepsilon = 10^{-3}$), the complexity obtained in the above theorem is tighter.

Leveraging the scaling law, we further show that prior generalization bound for neural network can be improved. See Appendix A.

6 Conclusion and Discussion

In this paper, we experimentally identified a scaling law governing the alignment of labels and label residuals with the NTK spectrum that consistently emerges in real-world datasets. We proposed a theoretical explanation of this phenomenon via Lipschitz continuity of the target function. Leveraging this scaling law, we derived a convergence bound for gradient descent in the NTK regime, which is much tighter than existing bounds and closely matches the fast-decaying training curves observed in practice.

Limitation. Our theoretical explanation of the scaling law is currently limited to the alignment between labels $\mathbf{y}$ and the NTK spectrum and does not extend to the label residuals $\mathbf{r}$. Understanding the residual alignment remains an open problem, as the residuals evolve throughout training and therefore cannot be directly characterized by the regularity of the target function alone. In addition, our improved convergence bound is limited to the NTK regime. Extending the analysis beyond the NTK regime to account for feature learning and the resulting evolution of the NTK spectrum is an important direction for future work.

AI USE STATEMENT

We used AI tools solely for language editing and polishing of the manuscript. Specifically, AI assistance was used to improve grammar, clarity, and readability. No AI tools were used to generate research ideas, conduct experiments, analyze results, develop theoretical results, or otherwise contribute to the scientific content of the paper. All final content was reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the experimental setup, including the model architectures, datasets, training settings, NTK computation, initialization, and evaluation details in Appendix C. Similarly, we provided all assumptions, theorem statements, and proofs in the main text and appendix for the theoretical results. Together, the empirical and theoretical results are reproducible from the information provided.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022.

Ronen Basri, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. *Advances in Neural Information Processing Systems*, 32, 2019.

Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In *International conference on machine learning*, pp. 685–694. PMLR, 2020.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034. PMLR, 2020.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, May 2021. doi: 10.1038/s41467-021-23103-1. URL https://www.nature.com/articles/s41467-021-23103-1.

Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. In *International Conference on Learning Representations*, 2021.

Lénaïc Chizat, Édouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2937–2947, 2019.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019a.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=S1eK3i09YQ.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33:7710–7721, 2020.

Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. On the similarity between the laplace and neural tangent kernels. *Advances in Neural Information Processing Systems*, 33:1451–1461, 2020.

Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020 (11):113301, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HygegyrYwH.

Kaiqi Jiang, Jeremy Cohen, and Yuanzhi Li. Understanding the evolution of the neural tangent kernel at the edge of stability. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=QKo4c3LAz3.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: The catapult mechanism. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Chaoyue Liu, Libin Zhu, and Misha Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. *Advances in Neural Information Processing Systems*, 33: 15954–15964, 2020.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022a.

Chaoyue Liu, Libin Zhu, and Misha Belkin. Transition to linearity of wide neural networks is an emerging property of assembling weak models. In *International Conference on Learning Representations*, 2022b.

Chaoyue Liu, Han Bi, Like Hui, and Xiao Liu. Better NTK conditioning: A free lunch from (reLU) nonlinear activation in wide neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8dTBydCMSr.

Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar. Characterizing the spectrum of the ntk via a power series expansion. In *The Eleventh International Conference on Learning Representations*, 2023.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL http://stanford.edu.

Quynh Nguyen, Marco Mondelli, and Guido F Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8119–8129. PMLR, 2021.

Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33:11961–11972, 2020.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Haozhe Shan and Blake Bordelon. A theory of neural tangent kernel alignment and its influence on training. *arXiv preprint arXiv:2105.14301*, 2021.

Maksim Velikanov and Dmitry Yarotsky. Explicit loss asymptotics in the gradient descent training of neural networks. *Advances in Neural Information Processing Systems*, 34:2570–2582, 2021.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL http://arxiv.org/abs/1708.07747.

Libin Zhu, Chaoyue Liu, and Misha Belkin. Transition to linearity of general neural networks with directed acyclic graph architecture. *Advances in neural information processing systems*, 35:5363–5375, 2022.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109:467–492, 2020. doi: 10.1007/s10994-019-05839-6.

A AN IMPROVED NTK GENERALIZATION BOUND

We show that the scaling law of label-NTK alignment also helps to improve the generalization bound for an over-parameterized neural network. Our analysis is based on the following established generalization bound:

Theorem A.1 (Informal version of Theorem 5.1 of Arora et al. (2019))**.** *Let network initialization scale $\kappa = O(\lambda_{\min}/n)$ and network width $m = \Omega(\kappa^{-2}\mathrm{poly}(\lambda_{\min}^{-1}))$. The generalization error $\mathcal{L}_{test}$ of a two-layer ReLU network trained by gradient descent is upper bounded by*

$$\mathcal{L}_{test} \leq \sqrt{2\mathbf{y}^\top K^{-1}\mathbf{y}/n}. \tag{22}$$

Theorem A.2. *Consider the same setting as in Theorem 5.1 of Arora et al. (2019). Suppose the Label-NTK alignment holds: $\langle \mathbf{v}_i^\top \mathbf{y}\rangle^2 \sim \lambda_i^2$. Then, the test loss $\mathcal{L}_{test}$ of a two-layer ReLU network trained by gradient descent is upper bounded by*

$$\mathcal{L}_{test} \leq O\left(\sqrt{\mathrm{tr}[K]/n}\right). \tag{23}$$

Proof.

$$\mathbf{y}^\top K^{-1}\mathbf{y} = \sum_{i=1}^{n} \lambda_i^{-1}\langle \mathbf{v}_i, \mathbf{y}\rangle^2 = \sum_{i=1}^{n} \lambda_i^{-1}(c\cdot\lambda_i^2) = c\sum_{i=1}^{n}\lambda_i = c\cdot\mathrm{tr}[K], \tag{24}$$

where c is a certain constant. The result directly follows. □

B OMITTED PROOFS

B.1 PROOF OF THEOREM 4.3

Without loss of generality, we assume that $i = 1$ and $j = 2$; namely, the angle θ between the pair $\mathbf{x}_1$ and $\mathbf{x}_2$ is small $\theta \ll 1$.

We consider the normalized NTK matrix $\bar{K} = \frac{1}{(L+1)^2}K$. As has been shown in Liu et al. (2025), $\frac{1}{(L+1)^2}\|\nabla f(\mathbf{x}_i)\| = \|\mathbf{x}_i\| = 1$ for all $\mathbf{x}_i \in \mathcal{D}$. Therefore, all the diagonal entries of $\bar{K}$ are 1, and each off-diagonal entry $\bar{K}_{ij}$, $i \neq j$, is $\cos\phi_{ij}$ with ϕ_{ij} being the angle between the two tangent vectors $\nabla f(\mathbf{x}_i)$ and $\nabla f(\mathbf{x}_j)$.

We write the normalized NTK matrix $\bar{K}$ in the following form:

$$\bar{K} = \begin{pmatrix} 1 & \cos\phi_{12} & \mathbf{a}^\top \\ \cos\phi_{12} & 1 & \mathbf{b}^\top \\ \mathbf{a} & \mathbf{b} & B \end{pmatrix}, \tag{25}$$

where vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{n-2}$ and matrix $B \in \mathbb{R}^{(n-2)\times(n-2)}$. By Theorem 3.1 of Du et al. (2019b), the matrix B is positive-definite and invertible.

By Corollary of Liu et al. (2025), we get

$$\cos\phi_{12} = (1 - \gamma\theta + o(\theta))\cos\theta, \text{ with } \gamma = \frac{L}{2\pi}.$$

In addition, considering components of $\mathbf{a}$ and $\mathbf{b}$, we have, for each $i \in [n-2]$,

$$|a_i - b_i| = |\cos\phi_{1(i+2)} - \cos\phi_{2(i+2)}| = O(\theta_{1(i+2)} - \theta_{2(i+2)}) = O(\theta).$$

In brief, we can write

$$\bar{K} = \begin{pmatrix} 1 & 1-\gamma\theta+o(\theta) & \mathbf{a}^\top \\ 1-\gamma\theta+o(\theta) & 1 & \mathbf{a}^\top + O(\theta) \\ \mathbf{a} & \mathbf{a}+O(\theta) & B \end{pmatrix}. \tag{26}$$

We can see that it has the following eigenvalue λ and corresponding eigenvector $\mathbf{v}$:

$$\lambda = \gamma\theta + o(\theta), \quad \mathbf{v} = \frac{1}{\sqrt{2}}(1,\ -1,\ -B^{-1}(\mathbf{a}-\mathbf{b})) + o(\theta).$$

Hence,

$$\langle \mathbf{v}, \mathbf{y} \rangle = \frac{1}{\sqrt{2}}(y_1 - y_2 - \tilde{\mathbf{y}}^\top B^{-1}(\mathbf{a} - \mathbf{b})) + o(\theta).$$

Note that $\mathbf{a} - \mathbf{b} = O(\theta)$. Additionally, by Lipschitz continuity of label function, we have $|y_1 - y_2| \leq L_y \|\mathbf{x}_1 - \mathbf{x}_2\| = O(\theta)$. Therefore, we obtain

$$\langle \mathbf{v}, \mathbf{y} \rangle^2 = O(\theta^2), \text{ and } \lambda = O(\theta). \tag{27}$$

We complete the proof.

B.2 Proof of Theorem 5.4

We start the proof with the following key lemma (Proof of lemma is deferred to Appendix B.3).

Lemma B.1. *Given any $\varepsilon \in (0, 1)$, if the model Hessian spectral norm $\|\nabla^2 f\|_{op} \leq \frac{\lambda_{\min}\varepsilon}{32\beta R\sqrt{n}\mathcal{L}(\mathbf{w}_0)}$ holds within the ball $B_R(\mathbf{w}_0)$ with $R = \frac{8\sqrt{\beta\mathcal{L}(\mathbf{w}_0)}}{\lambda_{\min}}$ and learning rate $\eta \leq \frac{1}{2\beta + \beta_f\sqrt{2\mathcal{L}(\mathbf{w}_0)}}$, then*

$$\mathbf{f}(\mathbf{w}_t) - \mathbf{y} = (I - \eta K)^t(\mathbf{f}(\mathbf{w}_0) - \mathbf{y}) + \Delta_t, \tag{28}$$

with $\|\Delta_t\| \leq \frac{\varepsilon}{4\sqrt{\mathcal{L}(\mathbf{w}_0)}}$.

Prior work, Theorem 3.2 of Liu et al. (2020), has proven that the model Hessian norm can be arbitrarily small given a large enough m: $\|\nabla^2 f\|_{op} = \tilde{O}\left(\frac{R^{3L}}{\sqrt{m}}\right)$. Therefore, the conditions of the lemma is met by the neural network set in the theorem.

Taking square on both sides of Eq. (28), we get

$$\mathcal{L}(\mathbf{w}_t) = \frac{1}{2}\|\mathbf{f}(\mathbf{w}_t) - \mathbf{y}\|^2 = \frac{1}{2}\|(I - \eta K)^t\mathbf{r}_0\|^2 + ((I - \eta K)^t\mathbf{r}_0)^\top\Delta_t + \|\Delta_t\|^2.$$

On the other hand, multiplying a factor $(1 - \eta\lambda_i)^{2t}$ onto both sizes of Eq. (16) in Assumption 5.2 and summing over i gives

$$(c - \delta)\sum_{i=1}^n (1 - \eta\lambda_i)^{2t}\lambda_i \leq \sum_{i=1}^n (1 - \eta\lambda_i)^{2t}\langle \mathbf{v}_i, \mathbf{r}_0 \rangle^2 \leq (c + \delta)\sum_{i=1}^n (1 - \eta\lambda_i)^{2t}\lambda_i$$

Note that $\|(I - \eta K)^t\mathbf{r}_0\|^2 = \sum_{i=1}^n (1 - \eta\lambda_i)^{2t}\langle \mathbf{v}_i, \mathbf{r}_0 \rangle^2$ and $\sum_{i=1}^n (1 - \eta\lambda_i)^{2t}\lambda_i = \mathtt{tr}[(I - \eta K)^{2t}K]$, then we obtain

$$(c - \delta)\mathtt{tr}[(I - \eta K)^{2t}K] \leq \|(I - \eta K)^t\mathbf{r}_0\|^2 \leq (c + \delta)\mathtt{tr}[(I - \eta K)^{2t}K].$$

In addition, we have

$$|((I - \eta K)^t\mathbf{r}_0)^\top\Delta_t| \leq \|(I - \eta K)^t\mathbf{r}_0\| \cdot \|\Delta_t\| \leq \|\mathbf{r}_0\| \cdot \|\Delta_t\| \leq \frac{\varepsilon}{4},$$

and $\|\Delta_t\|^2 \leq \frac{\varepsilon^2}{16\mathcal{L}(\mathbf{w}_0)} \leq \frac{\varepsilon}{4}$. Combining all the above, we finally get the desired result.

B.3 Proof of Lemma B.1

Using Taylor expansion with Lagrange remainder term, as well as the GD update rule $\mathbf{w}_{t+1} - \mathbf{w}_t = -\eta J(\mathbf{w}_t)^\top(\mathbf{f}(\mathbf{w}_t) - \mathbf{y})$, we have the following:

$$\begin{aligned}\mathbf{f}(\mathbf{w}_{t+1}) - \mathbf{f}(\mathbf{w}_t) &= -\eta K(\mathbf{w}_t)(\mathbf{f}(\mathbf{w}_t) - \mathbf{y}) + \boldsymbol{\zeta}_t \\ &= -\eta K(\mathbf{w}_0)(\mathbf{f}(\mathbf{w}_t) - \mathbf{y}) - \eta\delta K_t(\mathbf{f}(\mathbf{w}_t) - \mathbf{y}) + \boldsymbol{\zeta}_t\end{aligned} \tag{29}$$

where $\delta K_t = K(\mathbf{w}_t) - K(\mathbf{w}_0)$ and vector $\boldsymbol{\zeta}_t \in \mathbb{R}^n$ has elements

$$\zeta_{t,i} = \eta^2(\mathbf{f}(\mathbf{w}_t) - \mathbf{y})^\top J(\mathbf{w}_t)\nabla^2 f(\xi_t(\mathbf{x}_i); \mathbf{x}_i)J(\mathbf{w}_t)^\top(\mathbf{f}(\mathbf{w}_t) - \mathbf{y}), \ \ \forall i \in [n], \tag{30}$$

where $\xi(\mathbf{x}_i)$ is some point on the line segment between $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$ depending on $\mathbf{x}_i$.

Note that, under the setting of the model Hessian spectral norm $\|\nabla^2 f\|_{op} \leq \frac{\lambda_{\min}\varepsilon}{32\beta R\sqrt{n}\mathcal{L}(\mathbf{w}_0)}$, $R = \frac{8\sqrt{\beta\mathcal{L}(\mathbf{w}_0)}}{\lambda_{\min}}$ and learning rate $\eta \leq \frac{1}{2\beta + \beta_f\sqrt{2\mathcal{L}(\mathbf{w}_0)}}$, Liu et al. (2022a) has proven that:

- $\mathbf{w}_t$ remains in the ball $B_R(\mathbf{w}_0)$ for all $t \geq 0$.
- $\mathcal{L}(\mathbf{w}_t) = \frac{1}{2}\|\mathbf{f}(\mathbf{w}_t) - \mathbf{y}\|^2 \leq \left(1 - \eta\frac{\lambda_{\min}}{2}\right)^t \frac{1}{2}\|\mathbf{f}(\mathbf{w}_0) - \mathbf{y}\|^2$.

Now, let's find the upper bounds for the magnitudes of the two terms: $\eta\delta K_t(\mathbf{f}(\mathbf{w}_t) - \mathbf{y})$ and $\boldsymbol{\zeta}_t$.

Bounding term $\eta\delta K_t(\mathbf{f}(\mathbf{w}_t) - \mathbf{y})$. For all $i \in [n]$, we have

$$\nabla_{\mathbf{w}} f(\mathbf{w}_t; \mathbf{x}_i) = \nabla_{\mathbf{w}} f(\mathbf{w}_0; \mathbf{x}_i) + \int_0^1 \nabla^2 f(\mathbf{w}_0 + \tau(\mathbf{w}_t - \mathbf{w}_0; \mathbf{x}_i)(\mathbf{w}_t - \mathbf{w}_0) d\tau.$$

Then we have

$$\begin{aligned}
\|\nabla_{\mathbf{w}} f(\mathbf{w}_t; \mathbf{x}_i) - \nabla_{\mathbf{w}} f(\mathbf{w}_0; \mathbf{x}_i)\| &\leq \sup_{\tau \in [0,1]} \|\nabla^2 f(\mathbf{w}_0 + \tau(\mathbf{w}_t - \mathbf{w}_0; \mathbf{x}_i)\|_{op} \cdot \|\mathbf{w}_t - \mathbf{w}_0\| \\
&\leq \frac{\lambda_{\min}\varepsilon}{32\beta R\sqrt{n}\mathcal{L}(\mathbf{w}_0)} \cdot R \\
&= \frac{\lambda_{\min}\varepsilon}{32\beta\sqrt{n}\mathcal{L}(\mathbf{w}_0)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\|J(\mathbf{w}_t) - J(\mathbf{w}_0)\|_{op} &\leq \|J(\mathbf{w}_t) - J(\mathbf{w}_0)\|_F \\
&= \sqrt{\sum_{i \in [n]} \|\nabla_{\mathbf{w}} f(\mathbf{w}_t; \mathbf{x}_i) - \nabla_{\mathbf{w}} f(\mathbf{w}_0; \mathbf{x}_i)\|^2} \\
&\leq \frac{\lambda_{\min}\varepsilon}{32\beta\mathcal{L}(\mathbf{w}_0)}.
\end{aligned}$$

Now, the spectral norm of the NTK change δK_t is bounded by

$$\begin{aligned}
\|\delta K_t\|_{op} &= \|J(\mathbf{w}_t)J(\mathbf{w}_t)^\top - J(\mathbf{w}_0)J(\mathbf{w}_0)^\top\|_{op} \\
&= \|J(\mathbf{w}_t)(J(\mathbf{w}_t) - J(\mathbf{w}_0))^\top + (J(\mathbf{w}_t) - J(\mathbf{w}_0))J(\mathbf{w}_0)^\top\|_{op} \\
&\leq (\|J(\mathbf{w}_t)\|_{op} + \|J(\mathbf{w}_0)\|_{op})\|J(\mathbf{w}_t) - J(\mathbf{w}_0)\|_{op} \\
&\leq \frac{\lambda_{\min}\varepsilon}{8\mathcal{L}(\mathbf{w}_0)}.
\end{aligned}$$

In the last inequality above, we use the fact that $\|J(\mathbf{w})\|_{op}^2 = \|K(\mathbf{w})\|_{op}$.

Therefore, the term $\eta\delta K_t(\mathbf{f}(\mathbf{w}_t) - \mathbf{y})$ is bounded by

$$\|\eta\delta K_t(\mathbf{f}(\mathbf{w}_t) - \mathbf{y})\| \leq \frac{\eta\lambda_{\min}\varepsilon}{8\mathcal{L}(\mathbf{w}_0)}\|\mathbf{f}(\mathbf{w}_t) - \mathbf{y}\|.$$

Bounding term $\boldsymbol{\zeta}_t$. By Eq. (30), we have, for all $t > 0, i \in [n]$,

$$\begin{aligned}
|\zeta_{t,i}| &\leq \eta^2\|\mathbf{f}(\mathbf{w}_t) - \mathbf{y}\|^2\|K(\mathbf{w}_t)\|_{op} \cdot \|\nabla^2 f(\xi_t(\mathbf{x}_i); \mathbf{x}_i)\|_{op} \\
&\leq \eta^2 \frac{\lambda_{\min}\varepsilon}{16R\sqrt{n}\mathcal{L}(\mathbf{w}_0)}\|\mathbf{f}(\mathbf{w}_t) - \mathbf{y}\|^2.
\end{aligned}$$

Then, we have

$$\|\boldsymbol{\zeta}_t\| = \sqrt{\sum_{i \in [n]} |\zeta_{t,i}|^2} \leq \eta^2 \frac{\lambda_{\min}\varepsilon}{16R\mathcal{L}(\mathbf{w}_0)}\|\mathbf{f}(\mathbf{w}_t) - \mathbf{y}\|^2.$$

Finally, applying Eq. (29) recursively, we get

$$\begin{aligned}
\mathbf{f}(\mathbf{w}_t) - \mathbf{y} &= (I - \eta K(\mathbf{w}_0))^t(\mathbf{f}(\mathbf{w}_0) - \mathbf{y}) + \sum_{\tau=0}^{t-1}(I - \eta K(\mathbf{w}_0))^{t-1-\tau}\left(\eta\delta K_\tau(\mathbf{f}(\mathbf{w}_\tau) - \mathbf{y}) + \boldsymbol{\zeta}_\tau\right) \\
&\triangleq (I - \eta K(\mathbf{w}_0))^t(\mathbf{f}(\mathbf{w}_0) - \mathbf{y}) + \Delta_t.
\end{aligned}$$

We can see that the first term $(I - \eta K(\mathbf{w}_0))^t(\mathbf{f}(\mathbf{w}_0) - \mathbf{y})$ is equivalent to a linear dynamics. The deviation term (deviating from linear dynamics) can be bounded by

$$\begin{aligned}
\|\Delta_t\| &\le \sum_{\tau=0}^{t-1} \|I - \eta K(\mathbf{w}_0)\|_{op}^{t-1-\tau} \left(\|\eta\delta K_\tau(\mathbf{f}(\mathbf{w}_\tau) - \mathbf{y})\| + \|\boldsymbol{\zeta}_\tau\|\right) \\
&\le \sum_{\tau=0}^{t-1} (1 - \eta\lambda_{\min})^{t-1-\tau} \cdot \frac{\eta\lambda_{\min}\varepsilon}{4\mathcal{L}(\mathbf{w}_0)} \|\mathbf{f}(\mathbf{w}_t) - \mathbf{y}\| \\
&\le \frac{\eta\lambda_{\min}\varepsilon}{4\sqrt{\mathcal{L}(\mathbf{w}_0)}} \sum_{\tau=0}^{t-1} (1 - \eta\lambda_{\min})^{t-1-\tau} \\
&\le \frac{\varepsilon}{4\sqrt{\mathcal{L}(\mathbf{w}_0)}}.
\end{aligned}$$

Above we used $\|\mathbf{f}(\mathbf{w}_t) - \mathbf{y}\|^2 \le \left(1 - \eta\frac{\lambda_{\min}}{2}\right)^t \|\mathbf{f}(\mathbf{w}_0) - \mathbf{y}\|^2 \le \|\mathbf{f}(\mathbf{w}_0) - \mathbf{y}\|^2 = \mathcal{L}(\mathbf{w}_0)$ for the third inequality, and $\sum_t q^t \le 1/(1-q)$ for $q < 1$ for the fourth inequality.

B.4 Proof of Theorem 5.8

According to Theorem 5.4, we have

$$\mathcal{L}(\mathbf{w}_t) \le \frac{c+\delta}{2} \mathtt{tr}[(I - \eta K)^{2t} K] + \frac{\varepsilon}{2} = \frac{c+\delta}{2} \sum_{i=1}^{n} (1 - \eta\lambda_i)^{2t} \lambda_i + \frac{\varepsilon}{2}. \tag{31}$$

By Proposition 5.7, we have $\lambda_i \le C_2 i^{-q}$. Define

$$i_\varepsilon^* = \left(\frac{(q-1)\varepsilon}{2C_2(c+\delta)} + n^{1-q}\right)^{\frac{1}{1-q}},$$

then we have

$$\begin{aligned}
\frac{c+\delta}{2} \sum_{i=i_\varepsilon^*+1}^{n} (1 - \eta\lambda_i)^{2t} \lambda_i &\le \frac{c+\delta}{2} \sum_{i=i_\varepsilon^*+1}^{n} \lambda_i \\
&\le \frac{c+\delta}{2} \sum_{i=i_\varepsilon^*+1}^{n} C_2 i^{-q} \\
&\le \frac{c+\delta}{2} \int_{i_\varepsilon^*}^{n} C_2 z^{-q} dz \\
&= \frac{(c+\delta)C_2}{2(q-1)} \left[(i_\varepsilon^*)^{1-q} - (n)^{1-q}\right] \\
&\le \frac{\varepsilon}{4}.
\end{aligned}$$

Then Eq. (31) reduces to

$$\mathcal{L}(\mathbf{w}_t) \le \frac{c+\delta}{2} \sum_{i=1}^{i_\varepsilon^*} (1 - \eta\lambda_i)^{2t} \lambda_i + \frac{3\varepsilon}{4}. \tag{32}$$

Now consider the term $\sum_{i=1}^{i_\varepsilon^*} (1 - \eta\lambda_i)^{2t} \lambda_i$. Denote $\lambda_{i=i_\varepsilon^*}$ as λ_*. We have

$$\begin{aligned}
\sum_{i=1}^{i_\varepsilon^*} (1 - \eta\lambda_i)^{2t} \lambda_i &\le (1 - \eta\lambda_*)^{2t} \sum_{i=1}^{i_\varepsilon^*} \lambda_i \\
&\le (1 - \eta\lambda_*)^{2t} \sum_{i=1}^{n} \lambda_i \\
&\le (1 - \eta C_1 (i_\varepsilon^*)^{-q})^{2t} \cdot \frac{2\mathcal{L}_0}{c-\delta}.
\end{aligned}$$

In the last inequality above, we used $\lambda_* \geq C_1(i_\varepsilon^*)^{-q}$ (Assumption 5.7) and Condition 5.2. When

$$t \geq \frac{1}{2\eta C_1}\left(\frac{(q-1)\varepsilon}{2(c+\delta)C_2} + n^{1-q}\right)^{\frac{q}{1-q}} \log\left(\frac{4(c+\delta)\mathcal{L}_0}{(c-\delta)\varepsilon}\right),$$

we can bound $\mathcal{L}(\mathbf{w}_t) \leq \varepsilon$. Hence, we finish the proof.

C Experimental setup

Architectures. We evaluate the scaling law of label-NTK alignment across four model architectures: fully-connected neural network (a.k.s. Multi Layer Perceptrons (MLP)), Convolutional Neural Network (CNN), Residual Network (ResNet), and Transformers. All experiments are implemented in PyTorch (Paszke et al., 2019).

The MLP consists of five hidden layers, each of width 512 with ReLU activations after every hidden layer and a final linear output layer. The CNN consists of five 3×3 convolutional layers with channel sizes $32, 64, 128, 128,$ and 256, ReLU activations throughout, and max-pooling after the second, fourth, and fifth convolution blocks. This is followed by adaptive 4×4 pooling and a three-layer classifier $4096 \to 512 \to 256 \to 10$.

We additionally evaluate ResNet-18 and Vision Transformer architectures. For the ResNet experiments, we use a CIFAR-style ResNet-18 with a 3×3 convolutional stem, no initial max-pooling, and four residual stages with channel dimensions 64, 128, 256, and 512. Each stage contains two basic residual blocks, each with two 3×3 convolutions and ReLU activations. Downsampling is performed with stride 2 at the beginning of stages 2-4, with a 1×1 convolution used in the shortcut whenever the spatial resolution or channel dimension changes. The network ends with adaptive global average pooling followed by a linear classifier. Unless otherwise stated, batch normalization is disabled (`use_bn=False`). Convolutional layers use kaiming normal initialization, while the final layer is initialized from a zero-mean Guassian distribution with standard deviation of 0.01.

We use a lightweight pre-norm Vision Transformer (ViT) with convolutional patch embedding, a learnable class token, and learnable positional embeddings. Each Transformer block contains multi-head self-attention followed by a two-layer MLP with GELU activations, with LayerNorm applied before each sublayer and residual connections around both. We consider the model configuration $(6, 6, 192)$ where the entries denote the number of attention heads, network depth, and embedding dimension, respectively. The MLP hidden dimension is set as a fixed multiple of the embedding dimension. Depending on the experiment, the NTK is computed for both scenarios, using all model parameters and using selected parameter groups (W_V, W_O, and W_{MLP} parameters). For all the experiments, SGD Optimizer with momentum as 0, learning rate as 0.01 are chosen consistently.

Datasets For MLPs, we evaluate MNIST (LeCun et al., 1998), FashionMNIST (Xiao et al., 2017), SVHN (Netzer et al., 2011), and CIFAR-10 (Krizhevsky, 2009). For CNNs, we consider CIFAR-10, SVHN, TinyImageNet-10 (Deng et al., 2009). For ResNet-18, we evaluate CIFAR-10, SVHN, and TinyImageNet-10. For ViTs, we consider CIFAR-10, and TinyImageNet-10 datasets.

NTK Alignment Computation. The numerical results for Label-NTK and Residual-NTK alignments are obtained by computing the eigen-spectrum of the NTK together with target projections across eigen-modes at multiple timestamps. In particular, Label-dependent projections $\langle \mathbf{v}_i, \mathbf{y} \rangle^2$ and residual-dependent projections $\langle \mathbf{v}_i, \mathbf{r} \rangle^2$ are evaluated on batches of 100 samples across 500 batches, and then aggregated to obtain mean projections-eigenvalues correlation plots which we refer to as NTK alignment plots. To understand the linearity and slope in the small eigen-value regime, performed a separate small eigen-value analysis using a reference fit limited to the tail of the spectrum.

Convergence. For the convergence experiments, full-batch GD was run with a learning rate 0.01, 200 iterations, and a subset of 1000 samples. For each run, the initialization NTK was computed and eigen-decomposed to obtain its spectrum, which was then combined with the initialization outputs and residuals to construct the improved full-spectrum theoretical prediction. Improved convergence prediction incorporates contributions from the initialization, rescaled to match the initial empirical loss and then compared with the practical training loss.

Compute resources. All experiments were implemented in PyTorch and executed on a single $x86_64$ compute worker equipped with dual AMD EPYC 9554 64-Core processors (256 logical

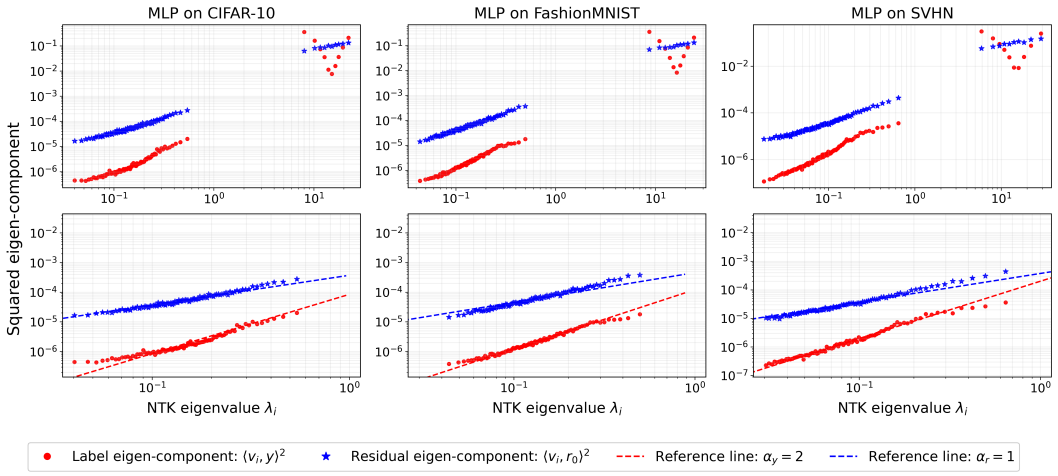


Figure 4: The scaling law of label-NTK alignment across on MLP. Both axes are log-scaled, under which the relationship becomes approximately linear. The label projections exhibit a slope close to $\alpha_y \approx 2$, while the residual projections exhibit a slope close to $\alpha_r \approx 1$. **Top:** full NTK eigenspectrum. **Bottom:** zoomed in to show the bulk of NTK spectrum.

CPUs in total), 755 GiB of system memory, and four NVIDIA H100 NVL GPUs, each with approximately 95 GiB of GPU memory. The implementation automatically used `cuda` when available and otherwise fell back to `cpu`. The empirical study includes MLP, CNN, ResNet-18, and ViT experiments across multiple datasets, training timestamps, random label controls, parameter subsets and model configurations. Computational cost varies substantially across these settings, particularly for larger ResNet-18 and ViT NTK computations, we report hardware and experimental settings rather than a single aggregate runtime. For smaller MLP and CNN, generating the saved data used for plotting typically required approximately 8 minutes per run, while the larger ResNet-18 and ViT experiments required longer depending on the dataset size, model configuration, and parameter subset used for the NTK computation.

D Additional Numerical Results

D.1 The scaling law across architectures and datasets

In addition to the results in Figure 1, Figure 4 provides more empirical evidence on the scaling law of label-NTK alignment on MLP model across CIFAR-10, FashionMNIST, and SVHN datasets. Figure 5 and 6 shows more evidence on CNN models on TinyImageNet-10 and SVHN datasets, and ResNet-18 architecture on TinyImageNet-10 and SVHN datasets, respectively. Across all these experimental settings of various neural network architecture and real-world datasets, we observe that the scaling law holds: $\log\langle \mathbf{v}_i, \mathbf{y}\rangle^2$ and $\log\langle \mathbf{v}_i, \mathbf{r}_0\rangle^2$ are both linear in $\log \lambda_i$, and their slopes are approximately matching $\alpha_y \approx 2$ and $\alpha_r \approx 1$.

D.2 The scaling law of label-NTK alignment on transformers

As mentioned above, we compute ViT NTK using selected parameter groups, in which we include the value-projection parameters W_V, output-projection parameters W_O, and MLP parameters W_{MLP}. These parameters are not involved in the attention score computation and only perform transformations in the embedding space, while the other parameters W_Q and W_K are for the inference of the attention scores between tokens. We empirically find that the scaling law of label-NTK alignment still holds well if we evaluate the NTK matrix only based on these selected parameters W_V, W_O, and W_{MLP}. Refer to Figure 7 for the results, the top panel shows the full eigenspectrum for CIFAR-10 and TinyImageNet-10 and the bottom panel shows the corresponding zoomed-in small eigenvalue spectrum. The label and residual projections exhibit the same scaling observed in MLP, CNN, ResNet-18 experiments where the trends remain closer to $\alpha_y \approx 2$ and $\alpha_r \approx 1$, respectively.

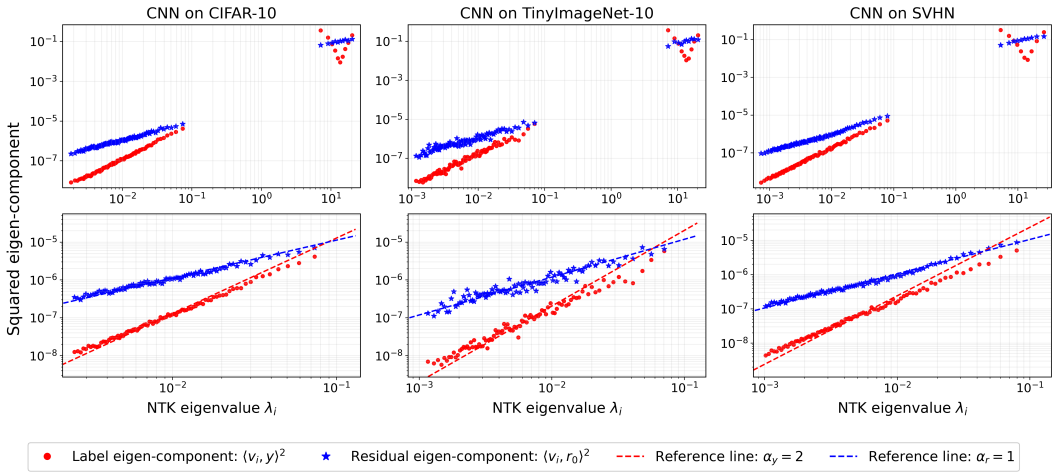


Figure 5: The scaling law of label-NTK alignment on CNN. Both axes are log-scaled, under which the relationship becomes approximately linear. The label projections exhibit a slope close to $\alpha_y \approx 2$, while the residual projections exhibit a slope close to $\alpha_r \approx 1$. **Top:** full NTK eigenspectrum. **Bottom:** zoomed in to show the bulk of NTK spectrum.

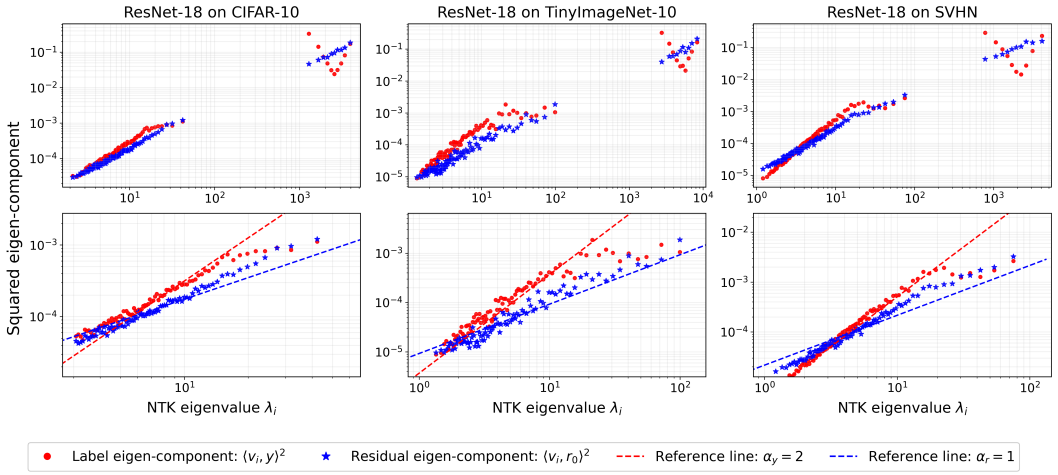


Figure 6: The scaling law of label-NTK alignment on ResNet-18. Both axes are log-scaled, under which the relationship becomes approximately linear. The label projections exhibit a slope close to $\alpha_y \approx 2$, while the residual projections exhibit a slope close to $\alpha_r \approx 1$. **Top:** full NTK eigenspectrum. **Bottom:** zoomed in to show the bulk of NTK spectrum.

D.3 The scaling law of label-NTK alignment during training

To examine whether the observed scaling is restricted to initialization, we compute the label and residual alignments at multiple stages of architecture training. At every checkpoint, NTK is recomputed using the current model parameters, and the label and residual projections are averaged across fixed-size batches.

Figure 8 illustrates the scaling law at intermediate epochs during training across different architectures and datasets. The scaling law remains broadly consistent with that observed at initialization. Throughout training, we observe that the squared eigen-components of the labels and label residuals remain systematically related to the NTK spectrum: they continue to be larger for larger NTK eigenvalues and smaller for smaller ones, although the linear relationship in the log-log plots may not hold exactly. Notably, both the NTK and the label residuals can change substantially during

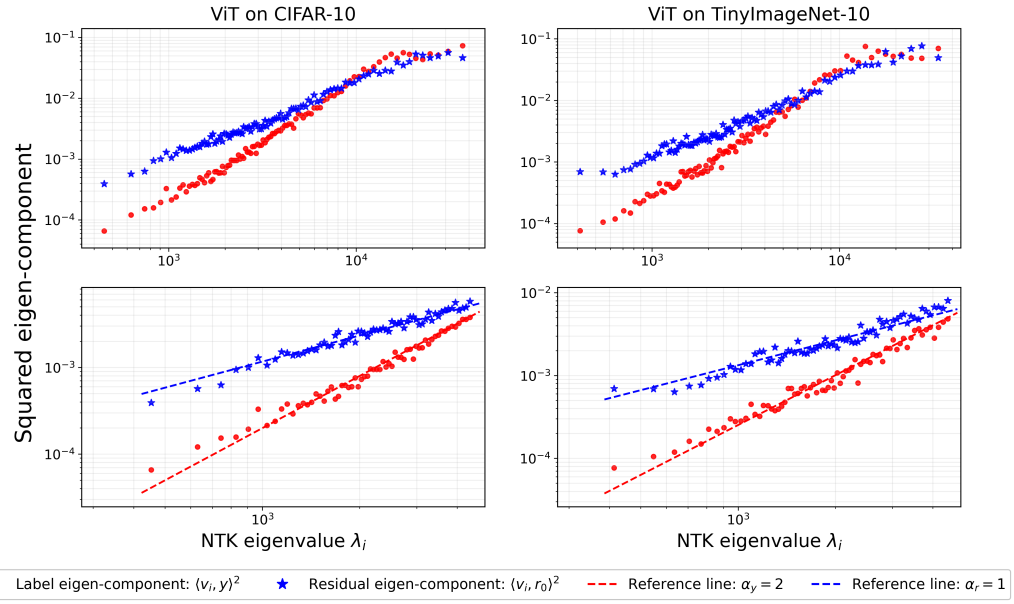


Figure 7: The scaling law of label-NTK alignment for Vision Transformer with 6 attention heads, depth 6, and embedding dimension 192, using the NTK computed on W_V, W_0, and W_{MLP} only. Both axes are log-scaled, under which the relationship becomes approximately linear. The label projections exhibit a slope close to $\alpha_y \approx 2$, while the residual projections exhibit a slope close to $\alpha_r \approx 1$. **Top:** full NTK eigenspectrum. **Bottom:** zoomed in to show the bulk of NTK spectrum.

training. The persistence of this scaling relationship despite these changes suggests that it reflects a nontrivial and robust property of this scaling law.

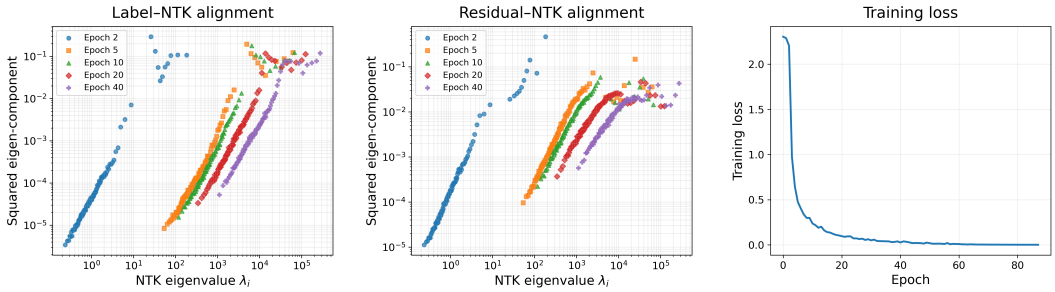
D.4 The scaling law does not hold for random labels

We provide more empirical evidence that the scaling law does not generally hold for non-real datasets. We repeat the analysis by replacing the original labels with i.i.d. random targets while keeping the input samples and model architecture fixed. These controls were performed across MLP, CNN, ResNet-18 settings. The goal of this experiment is to observe and compare the spectral structure induced by the targets when the input-label relationship is removed. Figure 9 shows the additional random label experiments across MLP, CNN, and ResNet-18 architectures.

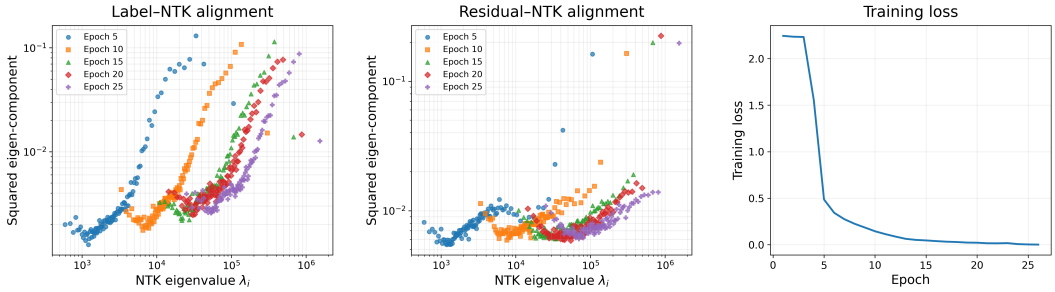
Across most settings, replacing the original targets with i.i.d. random labels weakens the scaling behavior and becomes much flatter across the NTK eigenspectrum. The effect is much clear for MLP, CNN experiments while in ResNet-18 settings, although the residual projections have some structure at larger eigenvalues, the label projections show the same flattening effect. These results indicate that the scaling law does not hold for non-real datasets.

D.5 Additional Convergence results

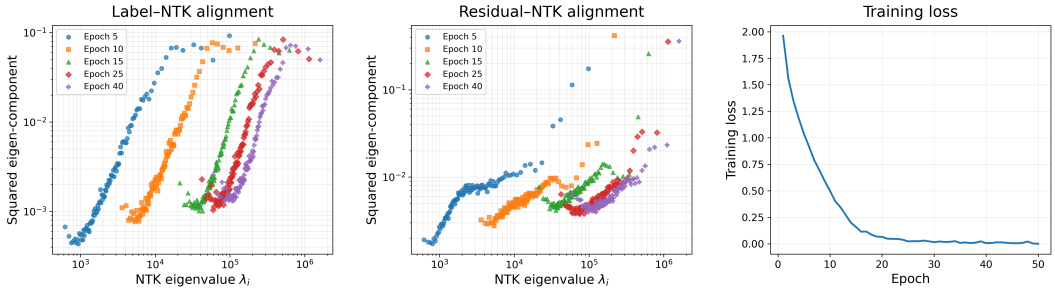
We provide additional experiments to evaluate the new convergence bound provided by Theorem 5.4. We consistently observe that the new bound, which is full-spectrum dependent, matches the practical training curves much better than prior bound across architectures and datasets. For MLPs, we report results on MNIST, Fashion-MNIST, SVHN, and CIFAR-10. For CNNs, we consider CIFAR-10 and TinyImageNet-10. The corresponding results are shown in Figure 10.



(a) MLP on MNIST



(b) CNN on SVHN



(c) ResNet-18 on CIFAR-10

Figure 8: Evolution of the scaling law of label-NTK alignment during training. **Left column:** $\langle \mathbf{v}_i, \mathbf{y} \rangle^2$ vs. λ_i, **Middle column:** $\langle \mathbf{v}_i, \mathbf{r}_t \rangle^2$ vs. λ_i, and **Right column:** Training loss over epochs.

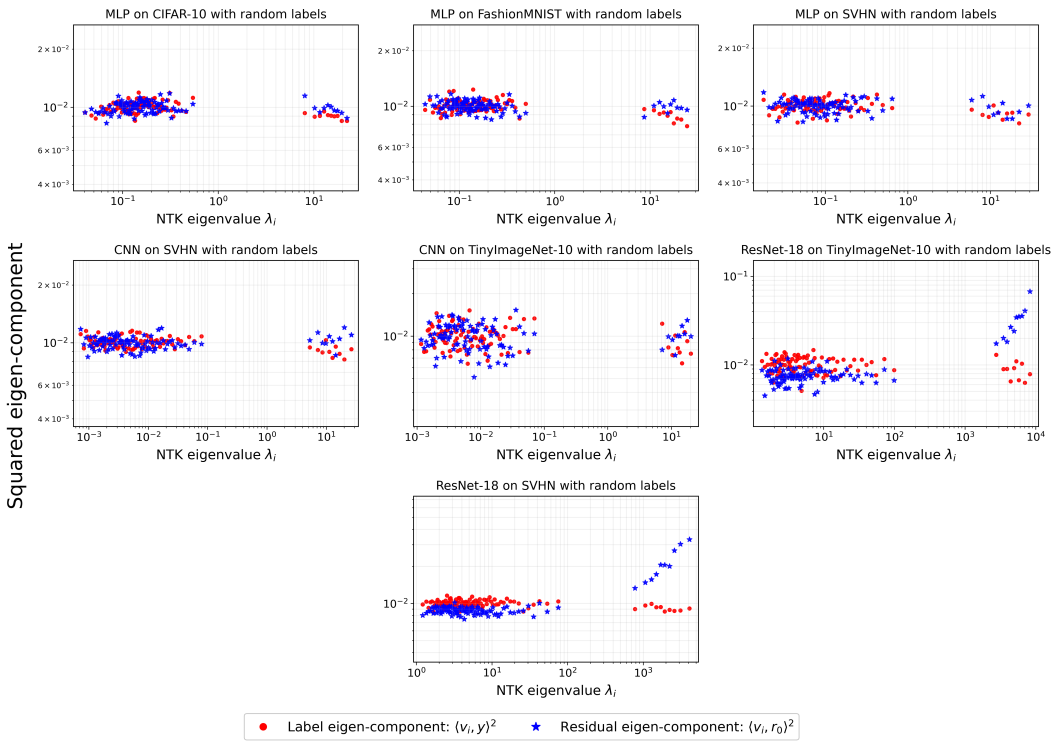
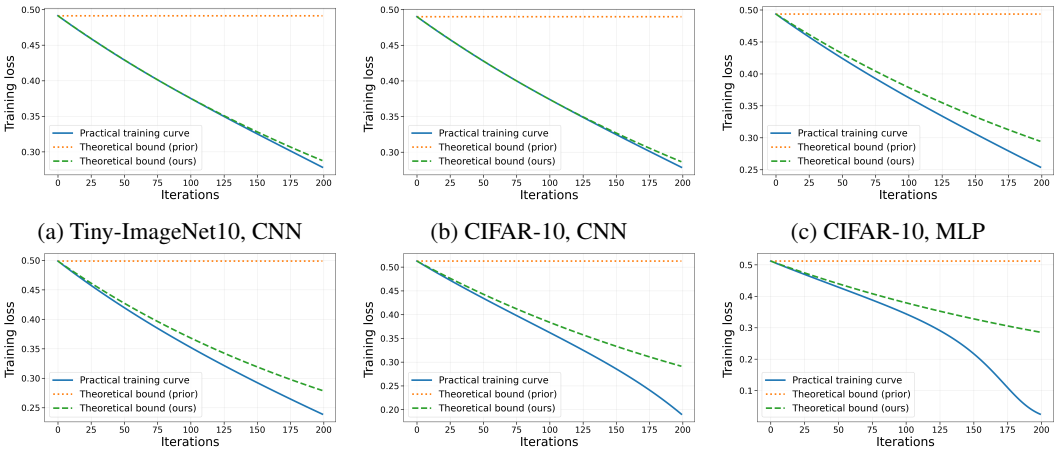


Figure 9: The scaling law of label-NTK alignment does not hold on dataset with i.i.d. random labels.



(a) Tiny-ImageNet10, CNN (b) CIFAR-10, CNN (c) CIFAR-10, MLP

(d) SVHN, MLP (e) Fashion-MNIST, MLP (f) MNIST, MLP

Figure 10: Comparison of theoretical convergence bounds across architectures and datasets. The MLP experiments use a five-layer MLP architecture on subsets of CIFAR-10, SVHN, Fashion-MNIST, and MNIST, while the CNN experiment is evaluated on Tiny-ImageNet10, and CIFAR-10. Each training subset contains 1000 samples.