# On the Learning Dynamics of Two-layer Linear Networks with Label Noise SGD

**Tongcheng Zhang**[1*], **Zhanpeng Zhou**[1*], **Mingze Wang**[2], **Andi Han**[3,4], **Wei Huang**[4,5†], **Taiji Suzuki**[4,6], **Junchi Yan**[1†]

[1]Sch. of Computer Science and Zhiyuan College, Shanghai Jiao Tong University
[2]Peking University
[3]University of Sydney
[4]RIKEN Center for Advanced Intelligence Project
[5]The Institute of Statistical Mathematics
[6]The University of Tokyo
{a-usually, zzp1012, yanjunchi}@sjtu.edu.cn, wei.huang.vr@riken.jp

## Abstract

One crucial factor behind the success of deep learning lies in the implicit bias induced by noise inherent in gradient-based training algorithms. Motivated by empirical observations that training with noisy labels improves model generalization, we delve into the underlying mechanisms behind stochastic gradient descent (SGD) with label noise. Focusing on a two-layer over-parameterized linear network, we analyze the learning dynamics of label noise SGD, unveiling a two-phase learning behavior. In *Phase I*, the magnitudes of model weights progressively diminish, and the model escapes the lazy regime; enters the rich regime. In *Phase II*, the alignment between model weights and the ground-truth interpolator increases, and the model eventually converges. Our analysis highlights the critical role of label noise in driving the transition from the lazy to the rich regime and minimally explains its empirical success. Furthermore, we extend these insights to Sharpness-Aware Minimization (SAM), showing that the principles governing label noise SGD also apply to broader optimization algorithms. Extensive experiments, conducted under both synthetic and real-world setups, strongly support our theory.

**Code** — https://github.com/a-usually/Label-Noise-SGD

## Introduction

One central factor behind the success of modern deep learning stems from the implicit bias induced by inherent stochastic noise in gradient-based training algorithms. While clean training data is ideal, recent studies (Shallue et al. 2019; HaoChen et al. 2021; Damian, Ma, and Lee 2021) revealed that injecting label noise, or label smoothing during training can paradoxically improve the generalization of neural networks. Figure 1 demonstrates the observation where stochastic gradient descent (SGD) with label noise enhances model generalization capability and inherently favors sparser solutions. Wang and Jacot (2023) also demonstrated that training with SGD alone jumps from the local minimum solution with only a small possibility. These phenomena challenge conventional wisdom and raise a fundamental question:

*How does label noise, often undesirable in statistical learning, confer benefits in over-parameterized models?*

Existing theoretical works have tried to understand the mechanisms behind SGD with noisy labels. Blanc et al. (2020); Damian, Ma, and Lee (2021); Li, Wang, and Arora (2022) focused on the local geometry around the global minimizers selected by label noise SGD. HaoChen et al. (2021); Vivien, Reygner, and Flammarion (2022) analyzed the behavior of diagonal linear networks under label noise SGD. However, few attempts to study the learning dynamics of label noise SGD in a more realistic setting.

**Our Contributions.** In this work, we provide a theoretical analysis of the learning dynamics in a two-layer linear network trained with label noise SGD.

• *Theoretical analysis.* In this work, we rigorously characterize the learning dynamics of a two-layer linear network where both layers are trainable by label noise SGD on a regression task. In particular, we identify two phases:

- **Phase I.** The magnitudes of neuron weights progressively diminish, and the model escapes from the lazy regime (Chizat, Oyallon, and Bach 2019); enters the rich regime (Geiger et al. 2020).
- **Phase II.** The neurons increasingly align ground-truth interpolator, and the model becomes sparser.

Refer to Table 1 for a summary. In the lazy regime, the dynamics are linear and the model achieves zero training loss with parameters hardly varying (Du et al. 2019b; Li and Liang 2018; Du et al. 2019a; Allen-Zhu, Li, and Song 2019). The lazy regime is often considered undesirable in practice and fails to explain the surprising generalization of neural networks (Chizat, Oyallon, and Bach 2019). In contrast, the rich regime, or feature learning, which draws significant attention recently, captures complex non-linear dynamics and is considered beneficial for generalization. Our analysis highlights the effect of label noise SGD in shifting

[*]These authors contributed equally.
[†]Corresponding authors. Junchi Yan is also with School of AI, SJTU.
The SJTU authors were partly supported by NSFC 92370201.

dynamics from *lazy* to *rich* regime, serving as a minimalist example to explain its intriguing properties.

Notably, the combination of over-parameterization and the intricate coupling between the first and second layers makes the theoretical analysis of label noise SGD far more challenging than for simpler linear models. To the best of our knowledge, our work presents the first detailed theoretical investigation of label noise SGD in networks with two or more trainable layers.

• *Extension.* Furthermore, we explore whether the principles underlying label noise SGD apply to broader optimization algorithms. We extend our findings on label noise SGD to Sharpness-Aware Minimization (SAM) (Foret et al. 2021). We show that SAM also fosters the transition from the *lazy* to *rich* regime and promotes sparsity in neural networks.

*In summary*, our work unveils a richer set of implicit bias of label noise SGD. We theoretically analyze the dynamics of SGD with label noise and carefully characterize how it transitions from lazy to rich regime. Our results offer valuable insights into the mechanisms behind the noise inherent in stochastic learning algorithms.

## Related Work

**Lazy Regime.** Numerous theoretical studies investigated the learning dynamics of highly over-parameterized neural networks in the *lazy* (or kernel) regime (Jacot, Gabriel, and Hongler 2018; Du et al. 2019b,a; Allen-Zhu, Li, and Song 2019; Zou et al. 2020). In this regime, the model behaves as its linearized model around initialization throughout training, making it equivalent to a deterministic kernel, specifically the neural tangent kernel (NTK) (Jacot, Gabriel, and Hongler 2018). The lazy regime typically occurs in over-parameterized models with *relatively large initialization* (Chizat, Oyallon, and Bach 2019). While global exponential convergence can be established in this setting, the lazy dynamics fail to explain the generalization advantage over kernel methods—a fundamental question in understanding the success of deep learning.

**Rich Regime.** In contrast to the lazy regime, where learning dynamics remain linear, the *rich* regime[1], also known as feature learning regime, exhibits complex nonlinear dynamics (Chizat and Bach 2018; Mei, Misiakiewicz, and Montanari 2019), including the initial alignment phenomenon (Maennel, Bousquet, and Gelly 2018; Luo et al. 2021) and saddle-to-saddle dynamics (Jacot et al. 2021; Abbe, Boix-Adsera, and Misiakiewicz 2023; Pesme and Flammarion 2023). Some studies have demonstrated that the initialization scale governs the emergence of the rich regime in (S)GD, which typically occurs at *small initialization scales* (Geiger et al. 2020; Woodworth et al. 2020). In this regime, it is shown that small initialization induces simplicity biases, leading to sparse or low-rank features (Maennel, Bousquet, and Gelly 2018; Li, Luo, and Lyu 2020; Lyu et al. 2021; Boursier, Pillaud-Vivien, and Flammarion 2022; Wang and Ma 2023; Min, Mallada, and Vidal 2023). Subsequent work further revealed that the relative scale of initializations (Azulay et al. 2021; Kunin et al. 2024) and their effective rank (Liu et al. 2024) can similarly induce feature learning. Beyond initializations, factors like weight decay (Lewkowycz and Gur-Ari 2020; Jacot et al. 2022; Lyu et al. 2023) and large learning rates (Lewkowycz et al. 2020; Ba et al. 2022) have also been shown to drive the rich regimes.

[1]This term broadly refers to learning behaviors that deviate from the lazy regime.

Algorithm 1: Label Noise SGD (HaoChen et al. 2021; Damian, Ma, and Lee 2021)

**Input**: Initial parameters $\boldsymbol{\theta}(0)$, step size $\eta$, label flipping probability $\tau$, batch size $B$, steps $T$.

**for** $t = 0$ **to** $T-1$ **do**
  Sample a batch $\mathcal{B}_t \in [n]^B$ uniformly.
  **for** each $i \in \mathcal{B}_t$ **do**
    Sample $u \sim \mathrm{Uniform}(0,1)$.
    **if** $u < \tau$ **then**
      $\tilde{y}_i =$ sample from $[c] \backslash \{y_i\}$.
    **else**
      $\tilde{y}_i = y_i$.
    **end if**
    $\hat{\ell}_i(\boldsymbol{\theta}(t)) = \ell(f(\boldsymbol{\theta}(t); \boldsymbol{x}_i), \tilde{y}_i)$.
  **end for**
  $\hat{\mathcal{L}}(\boldsymbol{\theta}(t)) = \frac{1}{B} \sum_{i \in \mathcal{B}_t} \hat{\ell}_i(\boldsymbol{\theta}(t))$.
  $\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \nabla \hat{\mathcal{L}}(\boldsymbol{\theta}(t))$.
**end for**

**Label Noise SGD Theories.** Many existing theoretical works have analyzed label noise SGD from the perspective of implicit regularization. Blanc et al. (2020); Damian, Ma, and Lee (2021); Li, Wang, and Arora (2022) showed that label noise implicitly regularizes the sharpness of the minimizers. HaoChen et al. (2021); Vivien, Reygner, and Flammarion (2022) proved that training with label noise helps recover the sparse ground-truth interpolator in a diagonal linear network setup. Takakura and Suzuki (2024) analyzed the implicit regularization of label noise from a kernel perspective. In addition to implicit regularization, Huh and Rebeschini (2024) derived a generalization bound for label noise SGD. Han et al. (2025) theoretically analyzed the training dynamics with static label noise. Huang et al. (2025) analyzed the generalization capability in low SNR regimes. Varre, Sagitova, and Flammarion (2024) showed that label noise SGD tends to gradually reduce the rank of the parameter matrix.

*In comparison*, we theoretically analyze the learning dynamics of label noise SGD in an over-parameterized two-layer linear network, highlighting the transition from lazy to rich regime. Analyzing the two-layer linear network with label noise SGD requires careful treatment of the update rule of both layers, which introduces complex coupling effect between the first and the second layer parameters, thus posing significant challenges to theoretical analysis.

## Preliminaries

**Basic Notation and Setup.** Denote $[k] = \{1, 2, \ldots, k\}$. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be the training set, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the input and $y_i \in \mathbb{R}$ is the label/target of the $i$-th data point. Let $f : \mathcal{D} \times \mathbb{R}^d \to \mathbb{R}$ be the model function and let $f(\boldsymbol{x}_i; \boldsymbol{\theta})$ be the model output on the $i$-th data point, where $\boldsymbol{\theta} \in \mathbb{R}^p$ are the model parameters. The loss of the model at

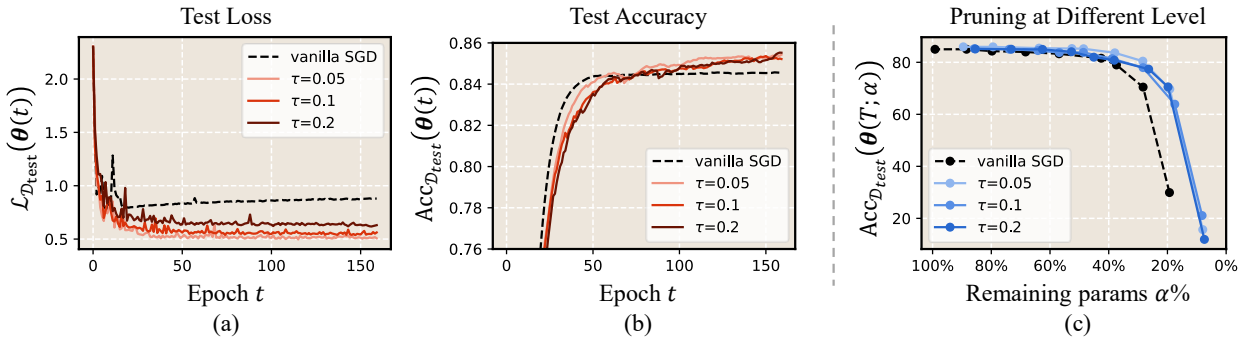


Figure 1: (Left).Label noise SGD (Algorithm 1) leads to better generalization. Test loss $\mathcal{L}(\boldsymbol{\theta}(t))$ and accuracy $\text{Acc}(\boldsymbol{\theta}(t))$ vs. training epochs $t$. (Right). Label noise SGD leads to sparser solutions. Testing accuracy of pruned mode $\text{Acc}(\boldsymbol{\theta}(T;\alpha))$ vs. the percentage of remaining parameters $\alpha$. Here, $\boldsymbol{\theta}(T;\alpha)$ represents the pruned model derived from the pretrained model $\boldsymbol{\theta}(T)$, with $\alpha\%$ of parameters remaining. The details of the model pruning algorithm are provided in Appendix D. We use both vanilla SGD and label noise SGD to train the models, with no weight decay or momentum. The learning rate is set to $0.1$, and the total number of epochs is $160$. We employ exponential moving averaging to smooth the test accuracy curves. Results are presented for ResNet-18 (He et al. 2016) trained on CIFAR-10 (Krizhevsky, Hinton et al. 2009), across different label noise probabilities, ($\tau \in \{0.05, 0.1, 0.2\}$). As shown in figure (a) and figure (b), label noise SGD consistently outperforms vanilla SGD in both test loss and accuracy across different values of the label flipping probability $\tau$, providing an around $1.5\%$ improvement in test accuracy. As shown in figure (c), models trained with label noise SGD maintain higher performance at the same sparsity level.

the $i$-th sample $(\boldsymbol{x}_i, y_i)$ is denoted as $\ell(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i)$, simplified to $\ell_i(\boldsymbol{\theta})$. The loss over the training set is then given by $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})$. Note that we consider classification tasks in our empirical observations, where $y_i \in [c]$ and $c$ are the number of classes. We also use $\text{Acc}_{\mathcal{D}}(\boldsymbol{\theta})$ to denote the classification accuracy of $f(\boldsymbol{\theta})$ on the dataset $\mathcal{D}$.

Throughout the paper, bold lowercase letters denote vectors, and bold uppercase letters represent matrices. The unbolded lowercase letters with subscripts indicate the entries of vectors or matrices, such as $x_i$ for the $i$-th entry of $\boldsymbol{x}$ and $a_{i,j}$ for the $(i,j)$-th entry of $\boldsymbol{A}$. For simplicity, we use $\|\boldsymbol{x}\|$ as $\|\boldsymbol{x}\|_2$ for a vector $\boldsymbol{x}$ and $\|\boldsymbol{X}\|$ as $\|\boldsymbol{X}\|_F$ for a matrix $\boldsymbol{X}$.

**Label Noise SGD.** We recall the algorithm of label noise SGD in Algorithm 1. We focus on a classification setting, where the label flipping probability $\tau$ governs the noise level; for simplicity, our theoretical analysis considers a regression task. Specifically, label noise SGD can be adapted to regression by replacing $\tau$ with the noise variance $\sigma^2$. In this context, the noisy label $\tilde{y}_i$ is generated by $\tilde{y}_i = y_i + \epsilon$, where $\epsilon \sim \{-\sigma, \sigma\}$. Assuming the squared loss without loss of generality, the training loss at the $i$-th data is given by:

$$\hat{\ell}_i(\boldsymbol{\theta}(t)) = \frac{1}{2}\left|f(\boldsymbol{\theta}(t);\boldsymbol{x}_i) - y_i - \epsilon\right|^2. \tag{1}$$

This setup is widely adopted in recent theoretical advances on label noise SGD (Damian, Ma, and Lee 2021; HaoChen et al. 2021; Li, Wang, and Arora 2022; Vivien, Reygner, and Flammarion 2022; Eun Huh and Rebeschini 2024).

## Theoretical Analysis: The Learning Dynamics of Label Noise SGD

This section presents a theoretical analysis of learning dynamics in a two-layer linear network, characterizing the phase transition from lazy to rich regimes under label noise SGD.

**Roadmap.** In Subsection "Setup and Overview: A Two-Layer Linear Network", we formulate the problem setup and provide an overview of our theoretical results. In Subsection "Phase I: Progressively Diminishing; From the Lazy to the Rich Regime" , we present formal theorems. In Subsection "Experiments: Synthetic and Real-World Setups", we validate our theory by extensive experiments under both synthetic and real-world setups.

### Setup and Overview: A Two-Layer Linear Network

**Problem Setup.** We consider a regression task where each data pair $(\boldsymbol{x}_i, y_i) \in \mathcal{D}$ maps input $\boldsymbol{x}_i \in \mathbb{R}^d$ to its corresponding target $y_i \in \mathbb{R}$. We solve this task using a two-layer linear network of the form:

$$\hat{y}_i = \boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{x}_i, \tag{2}$$

where $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ and $\boldsymbol{a} \in \mathbb{R}^m$. Here, $m$ represents the number of neurons and $\boldsymbol{w}_i$ denotes the $i$-th neuron of $\boldsymbol{W}$.

• *Label noise SGD.* The network's parameters $\boldsymbol{\theta} = \boldsymbol{a}^\top \boldsymbol{W}$ are optimized using label noise SGD with a squared loss function (see Equation (1)). The update rule is written as:

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \nabla_{\boldsymbol{\theta}} \hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t)), \tag{3}$$

$$\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t)) = \frac{1}{2}\left|f(\boldsymbol{\theta}(t);\boldsymbol{x}_{\xi_t}) - y_{\xi_t} - \epsilon_t\right|^2, \tag{4}$$

where $\xi_t \in [n]$ represents the index of a randomly sampled training sample at iteration $t$, and the noise $\epsilon_t \sim \{-\sigma, \sigma\}$ is controlled by the variance $\sigma^2$.

• *Initialization.* We consider label noise SGD starting from the following initializations: for $i \in [m]$ and $j \in [d]$,

$$w_{i,j}(0) \overset{\text{i.i.d.}}{\sim} \frac{1}{\sqrt{d}}\mathcal{N}(0, I) \text{ and } a_i(0) \overset{\text{i.i.d.}}{\sim} \frac{1}{\sqrt{m}}\mathcal{N}(0, I). \tag{5}$$

| | | |
|---|---|---|
| **I** | *The magnitudes of model weights progressively diminish; The model escapes the lazy regime; enters the rich regime.* | **Lemma 3 Theorem 2** |
| **II** | *The alignment between model weights and the ground-truth interpolator increases; the model eventually converges.* | **Lemma 5 Lemma 6** |

Table 1: Overview of the two-phase picture and corresponding theoretical results.

This initialization scheme is commonly referred to as the NTK initialization (Jacot, Gabriel, and Hongler 2018). Allen-Zhu, Li, and Song (2019) showed that training over-parameterized models initialized as Equation (5) with SGD stays in the lazy regime.

• *Data generation.* Without loss of generality, we assume each input $\boldsymbol{x}_i$ is drawn from $\mathcal{N}(0, \boldsymbol{I}_{d\times d})$, and that there exists at least one interpolating parameter $\boldsymbol{\theta}^\star$ that perfectly fits the training set, i.e., $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}^\star) = 0$[2].

**Overview.** We first state our main conditions.

**Condition 1.** *Suppose there exists a sufficiently large constant* [3] $C$ *such that the following holds:*

1. **(A1) Model width.** *The width of the network* $m$ *satisfies* $m = \Omega\left(\frac{1}{\sqrt{\eta}}\right)$.
2. **(A2) Learning rate.** *The learning rate satisfies* $\eta \leq \frac{1}{C^{96}}$.
3. **(A3) Dataset size.** *The training set size satisfies* $n \geq \frac{1}{\eta^2}$.
4. **(A4) Scale of the optimal parameter.** *The ground-truth interpolator satisfies* $\|\boldsymbol{\theta}^\star\| \leq m^{-1/4}$.
5. **(A5) Input magnitude.** *The maximum norm of the input samples satisfies* $\max_i \|\boldsymbol{x}_i\| \leq C_{data}$.
6. **(A6) Dimension of sample.** *The dimension of a single sample* $d$ *satisfies* $d \geq \frac{9(\ln 2)\cdot K^4}{2c}$.

**Remarks on Condition 1.** Specifically, A1 ensures over-parameterization, where $m \gg d$, which is important for enabling the progressively diminishing of weight norm in Phase I. A2 ensures the step size is small enough to allow Phase I to persist over a long range of iterations. A3 addresses the dataset size: when using gradient descent, a substantial volume of training data is typically required. A4 assumes the sparse ground-truth interpolator. For the sake of clarity in our theoretical analysis, we additionally introduce the A5 to give an upper bound of input norm and A6 to give a lower bound of input dimension. In A5, $C_{data}$ is a constant and by definition of $C$ in A1, we have $C_{data} \leq C \ll m$. In A6, both $K$ and $c$ are constant and defined in Appendix (Lemma A.5).

Under these conditions, we can state our main results: we identify a two-phase picture in the training dynamics of label noise SGD, as outlined in Table 1. In the following, we will formally present our formal theorems, with each claim supported by corresponding analysis.

[2] $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}^\star) = \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{\theta}^\star) = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}|f(\boldsymbol{\theta}^\star; \boldsymbol{x}_i), y_i|^2$

[3] $C \geq \max(e^{-\frac{(\sigma C_{data})^2}{3}}, (\frac{(1-3/(4\sqrt{\pi}))\cdot 2\sqrt{d}}{1/2-3/(4\sqrt{\pi})}\sqrt{\pi})^8, eC_{data}^2)$, where $C_{data}$ is a constant defined in Condition 1A(5)

## Phase I: Progressively Diminishing; From the Lazy to the Rich Regime

In this section, we present theoretical support for Phase I. Inspired by Du et al. (2019b,a); Allen-Zhu, Li, and Song (2019), we first introduce the definition of the lazy regime.

**Definition 1** (The lazy regime)**.** $\forall i \in [m]$, it holds that $\|\boldsymbol{w}_i(t) - \boldsymbol{w}_i(0)\| \leq \frac{1}{\sqrt{m}}$.

Definition 1 depicts a minimal variation of model weights from its initialization at time $t$. Based on Definition 1, we establish the following theorem.

**Theorem 2** (Escaping the lazy regime)**.** *Suppose Condition 1 (A1-2, 4-6) hold and consider the update rule in Equation* (3)*. With probability at least* $1 - O(\frac{1}{m})$*, all the neurons* $\boldsymbol{w}_i$ *(*$i \in [m]$*) escape from the lazy regime at time* $T_1 = \frac{384\sqrt{\log m}}{\sigma^2\eta^2\sqrt{m}}$.

**Insights from Theorem 2.** Theorem 2 indicates that in Phase I, label noise SGD facilitates the transition from the lazy to the rich regime. Indeed, such transition is induced by the *progressively diminishing* of the first-layer weights $\boldsymbol{W}$. Specifically, for each neuron $\boldsymbol{w}_i (i \in [m])$ at time $T$, we can easily derive that $\|\boldsymbol{w}_i(T)\|^2 = \|\boldsymbol{w}_i(0)\|^2 + \eta^2\sum_{j=0}^{T-1}\Delta W_i(j) - a_i(0)^2 + a_i(T)^2$ and $\Delta W_i(j) = -\nabla\hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2((\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2)$. Since $\boldsymbol{a}(0)$ is initialized small, the term $\nabla\hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2$ dominates the evolution of the weight norm. Notably, by Equation (3), we have

$$\nabla\hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 = (a_i(j+1) - a_i(j))^2.$$

Consequently, the evolution of the first-layer weight norm is primarily determined by the oscillations of the neurons in the second layer. Intuitively, label noise accelerates the oscillations in the second layer, thereby contributing to the progressive diminishing of the first-layer weights.

**Proof Sketch of Theorem 2.** The proof relies on Lemma 3.

**Lemma 3** (Progressively diminishing at each step)**.** *Suppose Condition 1 (A1-2, 4-6) hold and consider the update rule in Equation* (3)*. Assume the model is still under the lazy regime at step* $T$*, then with probability at least* $1 - O(\frac{1}{m})$*, for all the iterative steps* $j \leq T_1$ *and for every* $i \in [m]$*:*

1. $\Delta W_i(j) \leq 0$ *with probability at least* $1 - \frac{\rho}{m^{1/8}}$*. Furthermore,* $\Delta W_i(j) \leq -(\frac{\sigma}{4})^2$ *with probability at least* $\frac{1}{4}$.
2. $\Delta W_i(j) > 0$ *with probability at most* $\frac{\rho}{m^{1/8}}$*. Furthermore, we have* $\Delta W_i(j) \leq O(1)$.

*where* $\rho = \frac{2\sqrt{d}}{\sqrt{\pi}}$ *is a constant.*

Lemma 3 states that with high probability, $\Delta W_i(j)$ is negative at each step. Furthermore, with probability at least $\frac{1}{4}$,

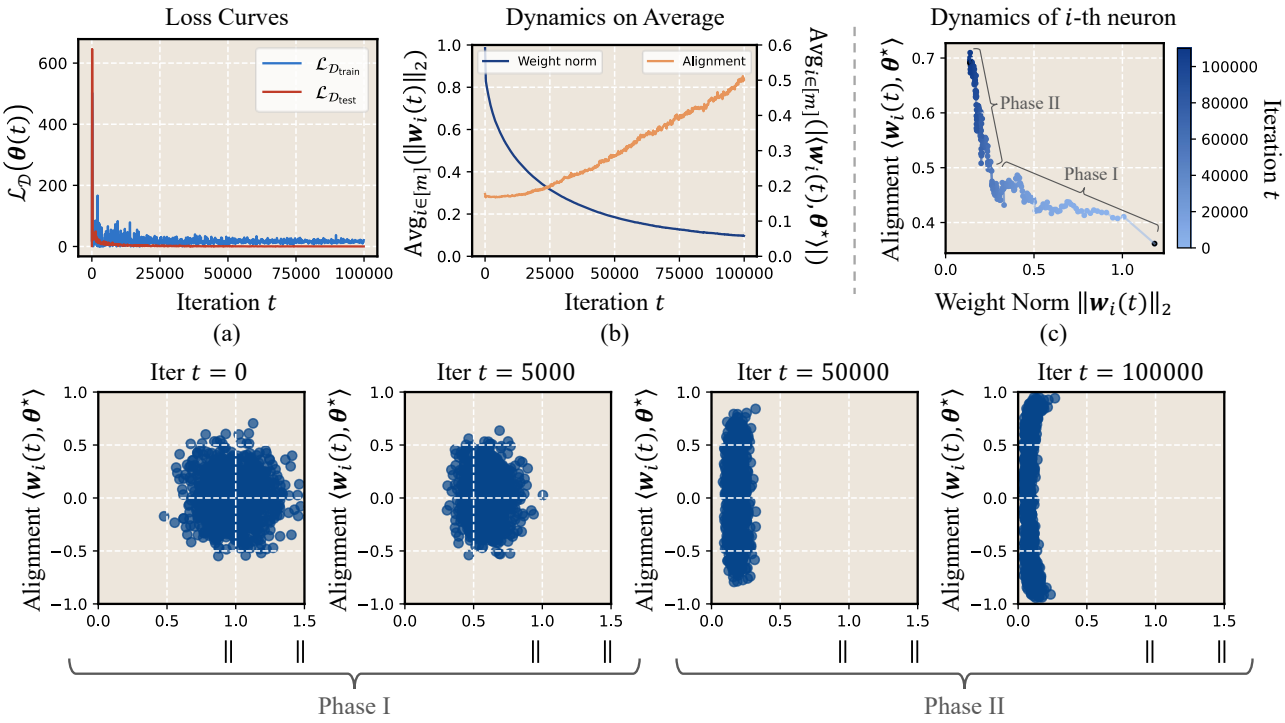


Figure 2: Two-phase dynamics of label noise SGD under synthetic setup. We replicate the synthetic problem setup from Subsection "Experiments: Synthetic and Real-World Setups". (a) Loss curves. Training $\mathcal{L}_{\mathcal{D}_{\text{train}}}(\boldsymbol{\theta}(t))$ and test loss $\mathcal{L}_{\mathcal{D}_{\text{test}}}(\boldsymbol{\theta}(t))$ vs. training iteration $t$. (b) Learning dynamics on average. The mean neuron norm $\text{Avg}_{i\in[m]}(\|\boldsymbol{w}_i(t)\|_2)$ and the mean neuron alignment $\text{Avg}_{i\in[m]}(\langle \boldsymbol{w}_i(t), \boldsymbol{\theta}^\star\rangle)$ vs. training iteration $t$. (c) Learning dynamics of $i$-th neuron. The alignment of the $i$-th neuron $\langle \boldsymbol{w}_i(t), \boldsymbol{\theta}^\star\rangle$ vs. its weight norm $\|\boldsymbol{w}_i(t)\|_2$, with darker points indicating larger $t$. (Bottom) Complete view of dynamics of each neuron. This plot is similar to (c); yet instead of focusing on a single neuron, we plot the status over iterations.

$\Delta W_i(j)$ significantly deviates from zero. In our setup, we can prove that with high probability, $|a_i(t)|$ remains small throughout training. Thus, $\Delta W_i(j)$ primarily governs the change in $\|\boldsymbol{w}_i\|^2$ at each step. With Lemma 3, we can demonstrate the progressive diminishing of the first-layer weights $\boldsymbol{W}$, thereby driving the transition from lazy to rich regime.

## Simulation Setup: Oscillation Induces Progressive Diminishing

In the previous analysis, we have shown that the oscillation of the second layer plays a central role in the progressive diminishing of the first-layer weights $\boldsymbol{W}$, and label noise SGD facilitates the oscillations, leading to the phase transition.

**Simulation Setup.** Inspired by the discovery, we propose to simulate the oscillation of the second layer via a simple three-state Markov process. To eliminate the interference of SGD noise, the first-layer weights $\boldsymbol{w}_i$ follow GD update rule:

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) - \eta \cdot a_i(t) \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}(t)) \boldsymbol{x}_i, \quad (6)$$

$$a_i(t+1) = a_i(t) + \delta_i(t), \quad (7)$$

where

$$\delta_i(t) = \begin{cases} -\eta^{0.25} & \text{if } a_i(t) = \eta^{0.25} \\ \eta^{0.25} & \text{if } a_i(t) = -\eta^{0.25} \\ \sim \{-\eta^{0.25}, \eta^{0.25}\} & \text{if } a_i(t) = 0 \end{cases}$$

The initialization of $a_i$ is set to $\eta^{0.25}$ or $-\eta^{0.25}$, each with probability $1/4$ and set to 0 with probability $1/2$. The initialization of $\boldsymbol{w}_i$ remains consistent with Equation (5). With this design, the neurons in the second layer exhibit strong oscillations within a small range. The following lemma demonstrates the progressive diminishing under this algorithm.

**Lemma 4** (Progressively diminishing under simulation setup)**.** *Suppose Condition 1 (A1-3, 5-6) holds (let $m = \frac{1}{\sqrt{\eta}}$) and consider the update rule in Equations* (6) *and* (7)*, there exists a step $t_0 \le \frac{1}{\eta^2}$ such that* [4]

$$\mathbb{E}[\frac{1}{m}\sum_{i=1}^{m} \|\boldsymbol{w}_i(t_0)\|^2] \le \sqrt{\eta}. \quad (8)$$

**Insights into Lemma 4.** Lemma 4 further confirms our key message that label noise SGD primarily contributes to the oscillation of the second layer, which induces the progressively diminishing phenomenon, ultimately leading to the transition from the lazy to the rich regime.

## Phase II: Alignment and Convergence

In this section, we present theoretical support for Phase II. When all the neurons satisfy $\|\boldsymbol{w}_i\|, |a_i| \le \sqrt{\eta}$, we say that

[4] We simply denote $\mathbb{E}_{\{x^{(t-1)}\}_{i=0}^{t-1} \in D, \{\epsilon_i\}_{i=0}^{t-1} \sim \{-\sigma, +\sigma\}}$ as $\mathbb{E}$

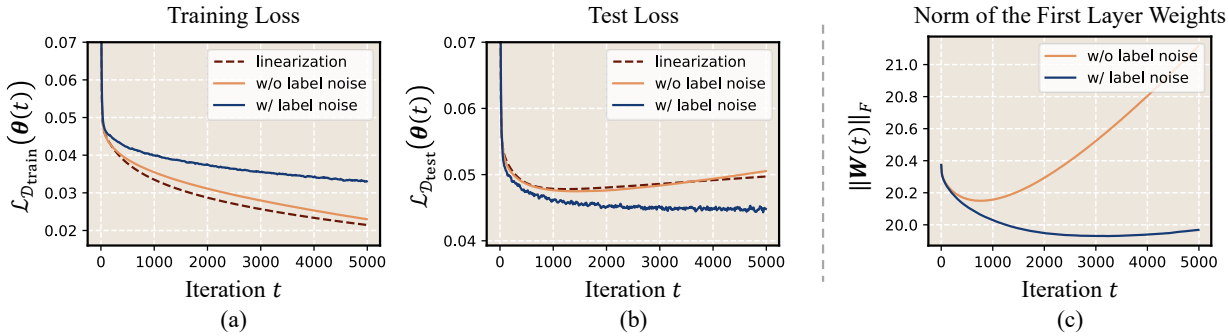


Figure 3: Label noise SGD induces the rich regime. (a, b). Training $\mathcal{L}_{\mathcal{D}_{\text{train}}}(\boldsymbol{\theta}(t))$ and test loss $\mathcal{L}_{\mathcal{D}_{\text{test}}}(\boldsymbol{\theta}(t))$ vs. training epochs $t$. Label noise SGD induces the progressively diminishing phenomenon. (c). The first-layer weight norm $\|\boldsymbol{W}(t)\|_F$ vs. training epochs $t$. We use GD to train the models with NTK parameterization (Jacot, Gabriel, and Hongler 2018), both with and without label noise. We also train a linearized model with GD as baseline. Results are presented for WideResNets trained on a random subset of 64 images from CIFAR-10 due to the $O(n^2)$ computational complexity of NTK.

Phase II begins. This situation is analogous to small initialization (Geiger et al. 2020; Woodworth et al. 2020). During this phase, the neurons in the first layer rapidly align with the ground-truth interpolator $\boldsymbol{\theta}^\star$.

Notice that we consider gradient descent in Phase II for simplicity. This simplification maintains mathematical tractability without affecting our conclusion in Phase II. The following lemmas formalize our results in Phase II.

**Lemma 5** (Alignment)**.** *Suppose Condition 1 (A1-3,5-6) holds and consider gradient descent for updates. Assume that phase II begins at time $t_1$, then at time $t_2 = t_1 + T_2$, $T_2 = \frac{1}{\|\boldsymbol{\theta}^\star\|} \cdot \ln(\frac{1}{\eta})$, for any neuron $\boldsymbol{w}_i$ it holds*

$$\frac{|\langle \boldsymbol{\theta}^\star, \boldsymbol{w}_i(t_2)\rangle|}{\|\boldsymbol{\theta}^\star\| \cdot \|\boldsymbol{w}_i(t_2)\|} \geq 1 - \left| O(\ln \frac{1}{\eta} \cdot \sqrt{\eta}) \right|. \tag{9}$$

**Lemma 6** (Convergence)**.** *Suppose Condition 1 (A1-3, 5-6) holds and consider gradient descent for updates. Assume all the neurons are perfectly aligned at step $t_2$. Let $t_3 = t_2 + \frac{1}{\|\boldsymbol{\theta}^\star\|^2} \cdot \frac{\ln(1/\eta)}{\eta}$. Using gradient descent, we have $\|\boldsymbol{\theta}(t_3) - \boldsymbol{\theta}^\star\| \leq |O(\eta \cdot \ln \frac{1}{\eta})|$. Furthermore, for any neuron $\|\boldsymbol{w}_i(t_3)\| \geq \sqrt{\eta}$ ($i \in [m]$), we have*

$$\frac{|\langle \boldsymbol{\theta}^\star, \boldsymbol{w}_i(t_3)\rangle|}{\|\boldsymbol{\theta}^\star\| \cdot \|\boldsymbol{w}_i(t_3)\|} \geq 1 - \left| O(\eta \cdot \ln \frac{1}{\eta}) \right|. \tag{10}$$

**Insights from Lemmas 5 and 6.** Lemma 5 indicates that the directions of each neuron rapidly align to a common direction, that of the ground-truth interpolator $\boldsymbol{\theta}^\star$. This alignment process is critical in Phase II, where the optimization shifts from the progressive diminishing phase to a more stable and efficient convergence towards the global minimum. Once perfect alignment is achieved, Lemma 6 guarantees that after $T_3 = O(\frac{-\ln \eta}{\eta})$ steps, $\boldsymbol{\theta}(t)$ converges to the solution $\boldsymbol{\theta}^\star$.

## Experiments: Synthetic and Real-World Setups

In this section, we provide extensive empirical evidence support for our theory.

**The Two-phase picture under synthetic setups.** The synthetic experiments precisely replicate the problem setup in Subsection "Setup and Overview: A Two-Layer Linear Network". In Figure 2 *(b)*, the averaged neuron norm $\frac{1}{m}\sum_{i=1}^m \|\boldsymbol{w}_i(t)\|$ initially drops as $t$ increases, suggesting the progressive diminishing phenomenon in Phase I. Afterwards, the averaged neuron alignment $\frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{w}_i(t), \boldsymbol{\theta}^\star\rangle$ rapidly increases, implying the convergence to the global solution in Phase II. Additionally, in Figure 2 *(c)* and *(bottom)*, we visualize the dynamics of each neuron in the training process, where a clear two-phase pattern is observed.

**The transition from the lazy to rich regime under real-world setups.** The real-world experiments are presented for WideResNets (Zagoruyko and Komodakis 2016) trained on a small subset of CIFAR-10. Specifically, we compare the loss curves of models trained with and without label noise. We also train a linearized model without label noise as a baseline. In Figure 3 *(a)* and *(b)*, the model trained without label noise behaves similarly to its linearized counterparts, indicating the lazy regime; whereas the model trained with label noise follows a distinctly different training trajectory, suggesting the rich regime. Additionally, we also plot the evolution process of the first-layer weight norm $\|\boldsymbol{W}(t)\|$. In Figure 3 *(c)*, when training with label noise, the first-layer weight norm notably decreases, especially compared to the case without label noise This result further validates our progressively diminishing phenomenon in real-world setups.

## Extension: From Label Noise SGD to SAM

We conjecture that the principles governing label noise SGD also apply to broader optimization algorithms. In this section, we extend our findings from label noise SGD to SAM.

**Sharpness-Aware Minimization (SAM).** SAM, introduced by Foret et al. (2021), has substantially improved the generalization of neural networks. The core idea of SAM is to perform an inner adversarial gradient perturbation, followed by an outer gradient update. Its update rule is written as $\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \nabla \ell_{\xi_t}\left(\boldsymbol{\theta}(t) + \rho \frac{\nabla \ell_{\xi_t}(\boldsymbol{\theta}(t))}{\|\nabla \ell_{\xi_t}(\boldsymbol{\theta}(t))\|_2}\right)$, where $\xi_t \in [n]$ is the index sampled at iteration $t$ and $\rho$ is the perturbation radius. Recent works (Monzio Compagnoni et al.

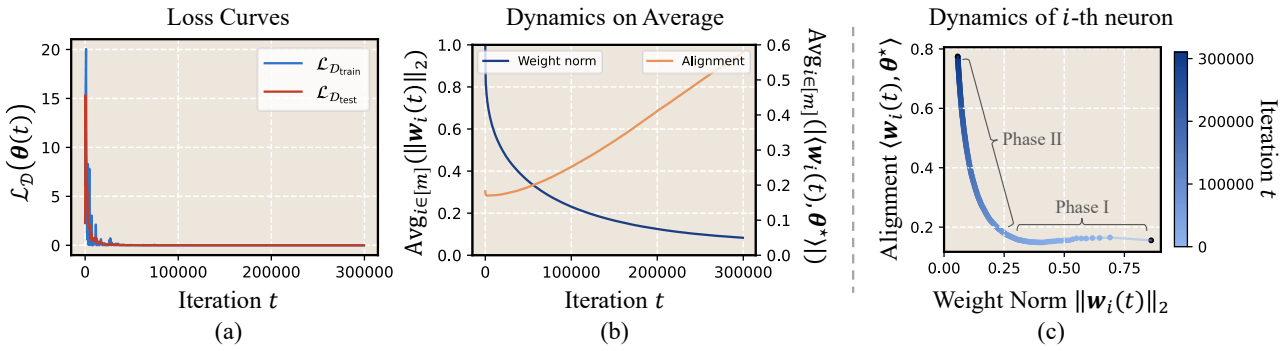


Figure 4: Two-phase dynamics of SAM under synthetic setup. We replicate the synthetic problem setup from Subsection "Experiments: Synthetic and Real-World Setups", replacing label noise SGD with SAM. (a) Loss curves. Training $\mathcal{L}_{\mathcal{D}_{\text{train}}}(\boldsymbol{\theta}(t))$ and test loss $\mathcal{L}_{\mathcal{D}_{\text{test}}}(\boldsymbol{\theta}(t))$ vs. training iteration $t$. (b) Learning dynamics on average. The averaged neuron norm $\text{Avg}_{i\in[m]}(\|\boldsymbol{w}_i(t)\|_2)$ and the averaged neuron alignment $\text{Avg}_{i\in[m]}(\langle \boldsymbol{w}_i(t), \boldsymbol{\theta}^\star\rangle)$ vs. training iteration $t$. (c) Learning dynamics of $i$-th neuron. The alignment of $i$-th neuron $\langle \boldsymbol{w}_i(t), \boldsymbol{\theta}^\star\rangle$ vs. its weight norm $\|\boldsymbol{w}_i(t)\|_2$, with darker points indicating larger iteration $t$.

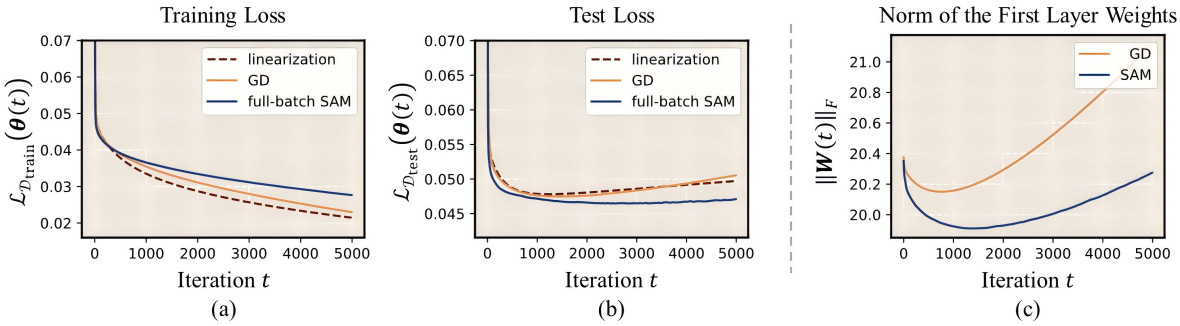


Figure 5: SAM induces the rich regime. Training $\mathcal{L}_{\mathcal{D}_{\text{train}}}(\boldsymbol{\theta}(t))$ and test loss $\mathcal{L}_{\mathcal{D}_{\text{test}}}(\boldsymbol{\theta}(t))$ vs. training epochs $t$. We use both GD and full-batch SAM to train the models with NTK parameterization (Jacot, Gabriel, and Hongler 2018). We also train a linearized model with GD as baseline. Results are presented for WideResNets trained on a random subset of 64 images from CIFAR-10 due to the $O(n^2)$ computational complexity of NTK.

2023; Zhou et al. 2025) have shown that the effectiveness of SAM stems from its ability to amplify the stochastic noise inherent in SGD. As label noise SGD also benefits from additional noise, a natural question arises:

*Can our findings on label noise SGD extend to SAM?*

**SAM exhibits the two-phase learning dynamics under synthetic setups.** First, we focus on the exact dynamics of SAM on synthetic data. We conduct the synthetic experiments similar to previous sections except for replacing label noise SGD with SAM. In Figure 4 *(b)*, the norm of the first-layer weights $\boldsymbol{W}$ initially decreases, demonstrating the progressive diminishing phenomenon; subsequently, the alignment between $\boldsymbol{W}$ and ground-truth interpolator $\boldsymbol{\theta}^\star$ increases, indicating an active feature learning process. Furthermore, in Figure 4, we visualize the learning dynamics of a single neuron, evidently confirming that the two-phase picture holds for SAM.

**SAM promotes feature learning in practice.** Second, we explore the properties of SAM in applications. As in phase I, we compare the loss curves of models trained with different strategies. In Figure 5, the model trained with GD shows loss curves that closely align with its linearized approximations, a hallmark of the lazy regime. In contrast, training with SAM results in loss curves that deviate significantly from GD trajectories, signaling entry into the rich regime.

In summary, SAM shows similar behaviors to label noise SGD across various setups, suggesting that the underlying ideology of label noise SGD is generalizable.

## Conclusion and Outlook

We present an in-depth study on the implicit regularization effect of label noise SGD from empirical observations to theoretical analysis. Notably, our theory demonstrates the surprising effect of label noise on the oscillation of the second layer, which induces the progressively diminishing phenomenon, leading to the transition from lazy to rich regime.

**Limitations.** Our theories are derived from a two-layer linear network. A future direction is to consider non-linear activation functions and study whether non-linearity influences the training dynamics. Besides, extending our theory to classification tasks remains an open challenge, which we will explore in future work.

## References


Abbe, E.; Boix-Adsera, E.; and Misiakiewicz, T. 2023. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. *arXiv preprint arXiv:2302.11055*.

Allen-Zhu, Z.; Li, Y.; and Song, Z. 2019. A Convergence Theory for Deep Learning via Over-Parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 242–252.

Azulay, S.; Moroshko, E.; Nacson, M. S.; Woodworth, B. E.; Srebro, N.; Globerson, A.; and Soudry, D. 2021. On the Implicit Bias of Initialization Shape: Beyond Infinitesimal Mirror Descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 468–477.

Ba, J.; Erdogdu, M. A.; Suzuki, T.; Wang, Z.; Wu, D.; and Yang, G. 2022. High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation. In *Advances in Neural Information Processing Systems*, volume 35.

Blanc, G.; Gupta, N.; Valiant, G.; and Valiant, P. 2020. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, 483–513.

Boursier, E.; Pillaud-Vivien, L.; and Flammarion, N. 2022. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35.

Chizat, L.; and Bach, F. 2018. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, volume 31.

Chizat, L.; Oyallon, E.; and Bach, F. 2019. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems*, volume 32.

Damian, A.; Ma, T.; and Lee, J. D. 2021. Label Noise SGD Provably Prefers Flat Global Minimizers. In *Advances in Neural Information Processing Systems*, volume 34, 27449–27461.

Du, S.; Lee, J.; Li, H.; Wang, L.; and Zhai, X. 2019a. Gradient Descent Finds Global Minima of Deep Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 1675–1685.

Du, S. S.; Zhai, X.; Poczos, B.; and Singh, A. 2019b. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. In *International Conference on Learning Representations*.

Eun Huh, J.; and Rebeschini, P. 2024. Generalization Bounds for Label Noise Stochastic Gradient Descent. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, 1360–1368.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.

Geiger, M.; Spigler, S.; Jacot, A.; and Wyart, M. 2020. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11): 113301.

Han, A.; Huang, W.; Zhou, Z.; Niu, G.; Chen, W.; Yan, J.; Takeda, A.; and Suzuki, T. 2025. On the Role of Label Noise in the Feature Learning Process. *arXiv preprint arXiv:2505.18909*.

HaoChen, J. Z.; Wei, C.; Lee, J.; and Ma, T. 2021. Shape Matters: Understanding the Implicit Bias of the Noise Covariance. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, 2315–2357.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Huang, W.; Han, A.; Song, Y.; Chen, Y.; Wu, D.; Zou, D.; and Suzuki, T. 2025. How Does Label Noise Gradient Descent Improve Generalization in the Low SNR Regime? *arXiv preprint arXiv:2510.17526*.

Huh, J. E.; and Rebeschini, P. 2024. Generalization Bounds for Label Noise Stochastic Gradient Descent. In *International Conference on Artificial Intelligence and Statistics*, 1360–1368.

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31.

Jacot, A.; Ged, F.; Simsek, B.; Hongler, C.; and Gabriel, F. 2021. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*.

Jacot, A.; Golikov, E.; Hongler, C.; and Gabriel, F. 2022. Feature Learning in $L_2$-regularized DNNs: Attraction/Repulsion and Sparsity. In *Advances in Neural Information Processing Systems*, volume 35.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. Accessed: 2025-1-19.

Kunin, D.; Raventos, A.; Domine, C. C. J.; Chen, F.; Klindt, D.; Saxe, A. M.; and Ganguli, S. 2024. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lewkowycz, A.; Bahri, Y.; Dyer, E.; Sohl-Dickstein, J.; and Gur-Ari, G. 2020. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*.

Lewkowycz, A.; and Gur-Ari, G. 2020. On the training dynamics of deep networks with L2 regularization. In *Advances in Neural Information Processing Systems*, volume 33, 4790–4799.

Li, Y.; and Liang, Y. 2018. Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data. In *Advances in Neural Information Processing Systems*, volume 31.

Li, Z.; Luo, Y.; and Lyu, K. 2020. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*.

Li, Z.; Wang, T.; and Arora, S. 2022. What Happens after SGD Reaches Zero Loss? –A Mathematical Framework. In *International Conference on Learning Representations*.

Liu, Y. H.; Baratin, A.; Cornford, J.; Mihalas, S.; SheaBrown, E. T.; and Lajoie, G. 2024. How connectivity structure shapes rich and lazy learning in neural circuits. In *The Twelfth International Conference on Learning Representations*.

Luo, T.; Xu, Z.-Q. J.; Ma, Z.; and Zhang, Y. 2021. Phase diagram for two-layer relu neural networks at infinite-width limit. *The Journal of Machine Learning Research*, 22(1): 3327–3373.

Lyu, K.; Jin, J.; Li, Z.; Du, S. S.; Lee, J. D.; and Hu, W. 2023. Dichotomy of early and late phase implicit biases can provably induce grokking. *arXiv preprint arXiv:2311.18817*.

Lyu, K.; Li, Z.; Wang, R.; and Arora, S. 2021. Gradient Descent on Two-layer Nets: Margin Maximization and Simplicity Bias. In *Advances in Neural Information Processing Systems*, volume 34.

Maennel, H.; Bousquet, O.; and Gelly, S. 2018. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*.

Mei, S.; Misiakiewicz, T.; and Montanari, A. 2019. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, 2388–2464.

Min, H.; Mallada, E.; and Vidal, R. 2023. Early neuron alignment in two-layer relu networks with small initialization. *arXiv preprint arXiv:2307.12851*.

Monzio Compagnoni, E.; Biggio, L.; Orvieto, A.; Proske, F. N.; Kersting, H.; and Lucchi, A. 2023. An SDE for Modeling SAM: Theory and Insights. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 25209–25253.

Pesme, S.; and Flammarion, N. 2023. Saddle-to-Saddle Dynamics in Diagonal Linear Networks. *arXiv preprint arXiv:2304.00488*.

Shallue, C. J.; Lee, J.; Antognini, J.; Sohl-Dickstein, J.; Frostig, R.; and Dahl, G. E. 2019. Measuring the Effects of Data Parallelism on Neural Network Training. *Journal of Machine Learning Research*, 20(112): 1–49.

Takakura, S.; and Suzuki, T. 2024. Mean-field Analysis on Two-layer Neural Networks from a Kernel Perspective. In *Forty-first International Conference on Machine Learning*.

Varre, A.; Sagitova, M.; and Flammarion, N. 2024. SGD vs GD: Rank Deficiency in Linear Networks. In *Advances in Neural Information Processing Systems*, volume 37, 60133–60161.

Vivien, L. P.; Reygner, J.; and Flammarion, N. 2022. Label noise (stochastic) gradient descent implicitly solves the Lasso for quadratic parametrisation. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, 2127–2159.

Wang, M.; and Ma, C. 2023. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of relu networks. In *Advances in Neural Information Processing Systems*, volume 36.

Wang, Z.; and Jacot, A. 2023. Implicit bias of SGD in $L_2$-regularized linear DNNs: One-way jumps from high to low rank. *arXiv preprint arXiv:2305.16038*.

Woodworth, B.; Gunasekar, S.; Lee, J. D.; Moroshko, E.; Savarese, P.; Golan, I.; Soudry, D.; and Srebro, N. 2020. Kernel and Rich Regimes in Overparametrized Models. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, 3635–3673.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhou, Z.; Wang, M.; Mao, Y.; Li, B.; and Yan, J. 2025. Sharpness-Aware Minimization Efficiently Selects Flatter Minima Late in Training. In *The Thirteenth International Conference on Learning Representations*.

Zou, D.; Cao, Y.; Zhou, D.; and Gu, Q. 2020. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine learning*, 109: 467–492.