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

Tongcheng Zhang[⋆] A-USUALLY@SJTU.EDU.CN
Shanghai Jiao Tong University, Shanghai, China

Zhanpeng Zhou[⋆†] ZZP1012@SJTU.EDU.CN
Shanghai Jiao Tong University, Shanghai, China

Mingze Wang MINGZEWANG@STU.PKU.EDU.CN
Peking University, Beijing, China

Andi Han ANDI.HAN@SYDNEY.EDU.AU
University of Sydney, Sydney, Australia

Wei Huang[†] WEI.HUANG.VR@RIKEN.JP
RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

Taiji Suzuki TAIJI@MIST.I.U-TOKYO.AC.JP
The University of Tokyo, Tokyo, Japan

Junchi Yan[†] YANJUNCHI@SJTU.EDU.CN
Shanghai Jiao Tong University, Shanghai, China

Abstract

[1] 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. Extensive experiments, conducted under both synthetic and real-world setups, strongly support our theory.

1. 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 [9, 14, 32] revealed that injecting label noise, or label smoothing during training can paradoxically improve the generalization of neural networks. Wang and Jacot [37] also demonstrated that training with SGD alone jumps from the local minimum solution with only a small possibility. This phenomenon challenges conventional wisdom and raises a fundamental question:

How does label noise confer benefits in over-parameterized models?

1. ⋆ denotes equal contribution; † indicates corresponding authors.

Existing Label Noise SGD Theories. Existing theoretical works have tried to understand the mechanisms behind stochastic gradient descent (SGD) with noisy labels. Blanc et al. [5], Damian et al. [9], Li et al. [23] showed that label noise implicitly regularizes the sharpness of the minimizers. HaoChen et al. [14], Vivien et al. [35] proved that training with label noise helps recover the sparse ground-truth interpolator in a diagonal linear network setup. Takakura and Suzuki [33] analyzed the implicit regularization of label noise from a kernel perspective. In addition to implicit regularization, Huh and Rebeschini [15] derived a generalization bound for label noise SGD. Varre et al. [34] showed that label noise SGD tends to gradually reduce the rank of the parameter matrix. However, few attempts to study the learning dynamics of label noise SGD in a more realistic setting.

Our contributions. 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 [8]; enters the rich regime [13].
- **Phase II.** The neurons increasingly align ground-truth interpolator, and the model becomes sparser.

Our analysis highlights the effect of label noise SGD in shifting 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.

In summary, our work unveils a richer set of implicit biases 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.

2. 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 $f : \mathcal{D} \times \mathbb{R}^p \to \mathbb{R}$ 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 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 $\mathrm{Acc}_{\mathcal{D}}(\boldsymbol{\theta})$ to denote the classification accuracy of $f(\boldsymbol{\theta})$ on the dataset $\mathcal{D}$.

Label Noise SGD. For simplicity, our theoretical analysis in Section 3 considers the Label Noise SGD in a regression task. Specifically, label noise SGD can be adapted to regression by introducing the noise variance σ^2. In this context, the noisy label $\tilde{y}_i$ is generated:

$$\tilde{y}_i = y_i + \epsilon, \text{where } \epsilon \sim \{-\sigma, \sigma\}. \tag{1}$$

Assuming the squared loss, the training loss at the i-th data is given by:

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

This setup has been widely adopted in recent theoretical advances on label noise SGD [9, 12, 14, 23, 35].

3. Theoretical Analysis: The Learning Dynamics of Label Noise SGD

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

3.1. 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 $\hat{y}_i = \boldsymbol{a}^\top \boldsymbol{W} \boldsymbol{x}_i$, where $\boldsymbol{W} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{a} \in \mathbb{R}^m$. Here, m represents the number of hidden neurons, and we denote the i-th neuron of $\boldsymbol{W}$ as $\boldsymbol{w}_i$. The network's parameters $\boldsymbol{\theta} = \boldsymbol{a}^\top \boldsymbol{W}$ are optimized using label noise SGD with a squared loss function (see Equation (2)). 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 σ^2.

We consider label noise SGD starting from the following initializations: for $i \in [m]$ and $j \in [d]$, $w_{i,j}(0) \overset{\text{i.i.d.}}{\sim} \frac{1}{\sqrt{d}} \mathcal{N}(0, I)$ and $a_i(0) \overset{\text{i.i.d.}}{\sim} \frac{1}{\sqrt{m}} \mathcal{N}(0, I)$. This initialization scheme is commonly referred to as the NTK initialization [16]. Allen-Zhu et al. [2] showed training over-parameterized models initialized as Section 3.1 with SGD stays in the lazy regime.

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., $\mathcal{L}(\boldsymbol{\theta}^\star) = 0$[2].

Main Conditions. Before proceeding to our main results, we first state our main conditions.

Condition 3.1. *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* $m = C \cdot \Omega\left(\max\left((\ln \frac{1}{\eta})^6, \frac{1}{\sqrt{\eta}}\right)\right)$.
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) Sparse interpolator.** *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* $d \geq \frac{9(\ln 2) \cdot K^4}{2c}$, *where* K *and* c *are defined in Lemma B.5.*

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 3.1A(5)

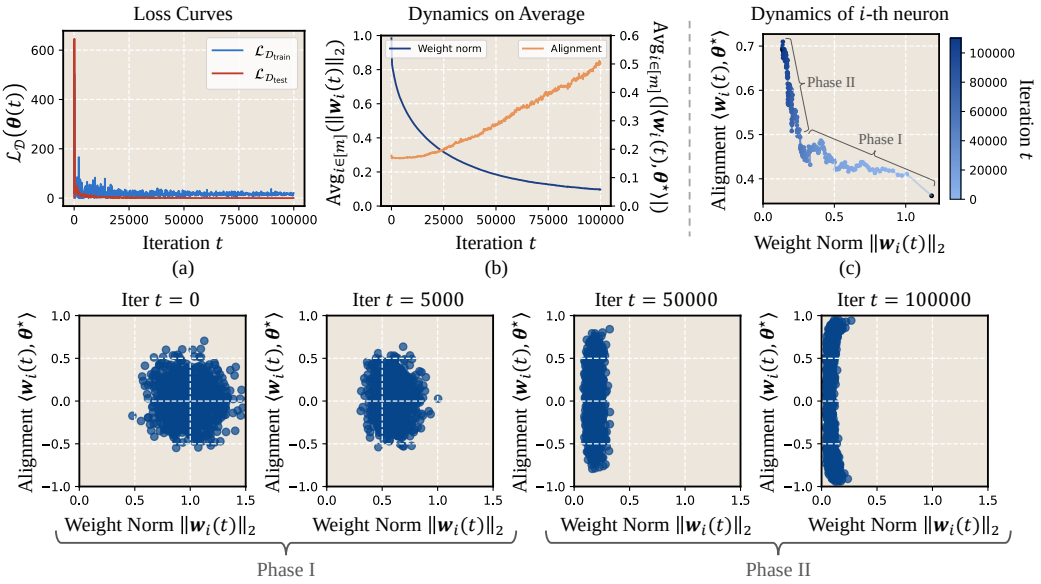

Figure 1: **Two-phase dynamics of label noise SGD under synthetic setup.** We replicate the synthetic problem setup from Section 3.1. **(a) Loss curves.** Training $\mathcal{L}_{\mathcal{D}_{\text{train}}}(\boldsymbol{\theta}(t))$ and test loss $\mathcal{L}_{\mathcal{D}_{\text{test}}}(\boldsymbol{\theta}(t))$ 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. **(Bottom) Complete view of dynamics of each neuron.** This plot is similar to (c); however, instead of focusing on a single neuron, we plot the status of each neuron at different iterations t.

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

Inspired by Allen-Zhu et al. [2], Du et al. [10, 11], we first introduce the definition of the lazy regime.

Definition 3.1 (The lazy regime)**.** $\forall i \in [m]$, it holds that $\|\boldsymbol{w}_i(t) - \boldsymbol{w}_i(0)\| \leq \frac{1}{\sqrt{m}}$.

Definition 3.1 depicts a minimal variation of model weights from its initialization at time t. Based on Definition 3.1, we establish the following theorem.

Theorem 3.2 (Escaping the lazy regime)**.** *Suppose Condition 3.1 (A1-2, 4-6) holds 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 3.2. Theorem 3.2 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,

$$\|\boldsymbol{w}_i(T)\|^2 = \|\boldsymbol{w}_i(0)\|^2 + \eta^2 \sum_{j=0}^{T-1} \Delta W_i(j) - a_i(0)^2 + a_i(T)^2,$$
$$\Delta W_i(j) = -\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2((\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2).$$

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.

3.2.1. 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 transition from the lazy to the rich regime.

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

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) - \eta \cdot a_i(t) \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}(t)) \boldsymbol{x}_i, \tag{5}$$
$$a_i(t+1) = a_i(t) + \delta_i(t), \tag{6}$$

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 Section 3.1.

With this design, the neurons in the second layer exhibit strong oscillations within a small range. The following Lemma 3.3 demonstrates the progressive diminishing of under this algorithm.

Lemma 3.3 (Progressively diminishing under simulation setup)**.** *Suppose Condition 3.1 (A1-3, 5-6) holds (let $m = \frac{1}{\sqrt{\eta}}$) and consider the update rule in Equations (5) and (6), there exists a step $t_0 \leq \frac{1}{\eta^2}$ such that* [4]

$$\mathbb{E}[\frac{1}{m} \sum_{i=1}^{m} \|\boldsymbol{w}_i(t_0)\|^2] \leq \sqrt{\eta}. \tag{7}$$

Insights into Lemma 3.3. Lemma 3.3 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.

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 $\mathbb{E}$

3.3. Phase II: Feature Alignment and Convergence

When all the neurons satisfy $\|\boldsymbol{w}_i\|, |a_i| \leq \sqrt{\eta}$, we say that Phase II begins. This situation is analogous to small initialization [13, 38]. 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 3.4 (Alignment)**.** *Suppose Condition 3.1 (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*

$$|\langle \boldsymbol{\theta}^\star, \boldsymbol{w}_i(t_2)\rangle| \geq 1 - \left| O(\ln \frac{1}{\eta} \cdot \sqrt{\eta}) \right|. \tag{8}$$

Lemma 3.5 (Convergence)**.** *Suppose Condition 3.1 (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*

$$|\langle \boldsymbol{\theta}^*, \boldsymbol{w}_i(t_3)\rangle| \geq 1 - \left| O(\eta \cdot \ln \frac{1}{\eta}) \right|. \tag{9}$$

Insights from Lemmas 3.4 and 3.5. Lemma 3.4 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 in Phase I to a more stable and efficient convergence towards the global minimum. Once perfect alignment is achieved, Lemma 3.5 guarantees that after $T_3 = O(\frac{-\ln \eta}{\eta})$ steps, $\boldsymbol{\theta}(t)$ converges to the solution $\boldsymbol{\theta}^\star$.

3.4. Experiments: Synthetic and Real-World Setups

The Two-phase picture under synthetic setups. The synthetic experiments precisely replicate the problem setup in Section 3.1. In Figure 1 *(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 1 *(b)* 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 [39] 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 2 *(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.

4. Conclusion and Outlook

We have presented 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, ultimately leading to the transition from the lazy to the rich regime.

References

[1] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. *arXiv preprint arXiv:2302.11055*, 2023.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019.

[3] Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 468–477. PMLR, 18–24 Jul 2021.

[4] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[5] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

[6] Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems*, 2022.

[7] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

[8] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[9] Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27449–27461. Curran Associates, Inc., 2021.

[10] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 09–15 Jun 2019.

[11] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[12] Jung Eun Huh and Patrick Rebeschini. Generalization bounds for label noise stochastic gradient descent. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1360–1368. PMLR, 02–04 May 2024.

[13] 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.

[14] Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2315–2357. PMLR, 15–19 Aug 2021.

[15] Jung Eun Huh and Patrick Rebeschini. Generalization bounds for label noise stochastic gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 1360–1368. PMLR, 2024.

[16] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[17] Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

[18] Arthur Jacot, Eugene Golikov, Clément Hongler, and Franck Gabriel. Feature learning in l_2-regularized DNNs: Attraction/repulsion and sparsity. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[19] Daniel Kunin, Allan Raventos, Clémentine Carla Juliette Dominé, Feng Chen, David Klindt, Andrew M Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[20] Aitor Lewkowycz and Guy Gur-Ari. On the training dynamics of deep networks with l_2 regularization. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4790–4799. Curran Associates, Inc., 2020.

[21] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

[22] Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*, 2020.

[23] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022.

[24] Yuhan Helena Liu, Aristide Baratin, Jonathan Cornford, Stefan Mihalas, Eric Todd SheaBrown, and Guillaume Lajoie. How connectivity structure shapes rich and lazy learning in neural circuits. In *The Twelfth International Conference on Learning Representations*, 2024.

[25] Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase diagram for two-layer relu neural networks at infinite-width limit. *The Journal of Machine Learning Research*, 22(1):3327–3373, 2021.

[26] Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34, 2021.

[27] Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S Du, Jason D Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. *arXiv preprint arXiv:2311.18817*, 2023.

[28] Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.

[29] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

[30] Hancheng Min, Enrique Mallada, and Rene Vidal. Early neuron alignment in two-layer relu networks with small initialization. *arXiv preprint arXiv:2307.12851*, 2023.

[31] Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. *arXiv preprint arXiv:2304.00488*, 2023.

[32] Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the effects of data parallelism on neural network training. *Journal of Machine Learning Research*, 20(112):1–49, 2019.

[33] Shokichi Takakura and Taiji Suzuki. Mean-field analysis on two-layer neural networks from a kernel perspective. In *Forty-first International Conference on Machine Learning*, 2024.

[34] Aditya Varre, Margarita Sagitova, and Nicolas Flammarion. Sgd vs gd: Rank deficiency in linear networks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 60133–60161. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6ec81faa568317949b0ff3be4d87cced-Paper-Conference.pdf.

[35] Loucas Pillaud Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2127–2159. PMLR, 02–05 Jul 2022.

[36] Mingze Wang and Chao Ma. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of relu networks. *Advances in Neural Information Processing Systems*, 36, 2023.

[37] Zihan Wang and Arthur Jacot. Implicit bias of sgd in l2-regularized linear dnns: One-way jumps from high to low rank. *ArXiv*, abs/2305.16038, 2023. URL https://api.semanticscholar.org/CorpusID:258887617.

[38] Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020.

[39] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[40] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine learning*, 109:467–492, 2020.

Appendix A. Related Work

Lazy Regime. Numerous theoretical studies investigated the learning dynamics of highly over-parameterized neural networks in the *lazy* (or kernel) regime [2, 10, 11, 16, 40]. 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) [16]. The lazy regime typically occurs in over-parameterized models with *relatively large initialization* [8]. While global exponential convergence can be established in this setting, the lazy dynamics fail to explain the generalization advantage of neural networks 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[5], also known as feature learning regime, exhibits complex nonlinear dynamics [7, 29], including the initial alignment phenomenon [25, 28] and saddle-to-saddle dynamics [1, 17, 31]. 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* [13, 38]. In this regime, it is shown that small initialization induces simplicity biases, leading to sparse or low-rank features [6, 22, 26, 28, 30, 36]. Subsequent work further revealed that the relative scale of initializations [3, 19] and their effective rank [24] can similarly induce feature learning. Beyond initializations, factors like weight decay [18, 20, 27] and large learning rates [4, 21] have also been shown to drive the rich regimes.

Label Noise SGD Theories. Many existing theoretical works have analyzed label noise SGD from the perspective of implicit regularization. Blanc et al. [5], Damian et al. [9], Li et al. [23] showed that label noise implicitly regularizes the sharpness of the minimizers. HaoChen et al. [14], Vivien et al. [35] proved that training with label noise helps recover the sparse ground-truth interpolator in a diagonal linear network setup. Takakura and Suzuki [33] analyzed the implicit regularization of label noise from a kernel perspective. In addition to implicit regularization, Huh and Rebeschini [15] derived a generalization bound for label noise SGD.

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 exhibits highly non-convex dynamics and introduce complex coupling effect between first- layer and second- layer parameters, thus posing significant challenges to theoretical analysis. Extensive experiments in both synthetic and real-world setups firmly support our theoretical analysis.

Appendix B. Preliminaries

B.1. Additional notations

Complementary Event. Let A be an event. We use $\overline{A}$ to denote the complementary event of A. We have $\Pr[A] + \Pr[\overline{A}] = 1$.

5. This term broadly refers to learning behaviors that deviate from the lazy regime.

Definition B.1. (sub-exponential) A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-exponential if there are non-negative parameters (ν, b) such that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\nu^2\lambda^2}{2}} \quad \text{for all} \quad |\lambda| < \frac{1}{b}.$$

We denote $X \in SE(\nu, b)$.

B.2. Preliminary lemmas

Lemma B.2. *Let $X = \sum_{i=1}^n X_i^2$ where $X_i \sim N(0,1)$ and i.i.d. Then $X \in SE(2\sqrt{n}, 4)$.*

Corollary B.3. *Let $Z = \sum_{i=1}^n X_i \cdot Y_i$ where $X_i, Y_i \sim N(0,1)$ and i.i.d. Then $X \in SE(2\sqrt{n}, 4)$.*

Proof of Corollary B.3. For every $i \in [n]$. we have

$$\begin{aligned}\mathbb{E}[e^{\lambda X_i^2 - 1}] &= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} e^{-\lambda(x^2-1)\cdot e^{-x^2/2}} dx \\ &= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}\end{aligned}$$

And we have

$$\begin{aligned}\mathbb{E}[e^{\lambda X_i Y_i}] &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{\lambda xy} \cdot \frac{1}{2\pi} e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}}\, dx\, dy \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{2\pi} e^{-\frac{(x-\lambda y)^2}{2}} e^{-\frac{(1-\lambda^2)y^2}{2}}\, dx\, dy \\ &= \frac{1}{\sqrt{1-\lambda^2}} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{1}{2\pi} e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}}\, dx\, dy \\ &= \frac{1}{\sqrt{1-\lambda^2}} \leq \mathbb{E}[e^{\lambda X^2 - 1}]\end{aligned}$$

So we have

$$\mathbb{E}[e^{\lambda Z}] = \mathbb{E}[e^{\lambda \sum_{i=1}^n X_i \cdot Y_i}] = \mathbb{E}[\prod_{i=1}^n e^{\lambda X_i \cdot Y_i}]$$

Since $X_i \cdot Y_i$ are independent, we have

$$\begin{aligned}\mathbb{E}[\prod_{i=1}^n e^{\lambda X_i \cdot Y_i}] &= \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i \cdot Y_i}] \\ &\leq \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i^2 - 1}] = \mathbb{E}[e^{\lambda X}]\end{aligned}$$

By lemma B.2, we have $X = \sum_{i=1}^n X_i^2 \in SE(2\sqrt{n}, 4)$. Therefore, we retain $Z = \sum_{i=1}^n X_i \cdot Y_i \in SE(2\sqrt{n}, 4)$. $\square$.

Lemma B.4. ***(Sub-exponential tail bound)*** *Suppose that X is sub-exponential with parameters (ν, b). Then*

$$\Pr[X \geq \mu + t] \leq \exp(-\frac{t^2}{2\nu^2}) \quad \textit{if } 0 \leq t \leq \frac{\nu^2}{b} \tag{10}$$

Lemma B.5. ***(Concentration of the norm)*** *Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent, sub-gaussian coordinates X_i that satisfy $\mathbb{E}X_i^2 = 1$. Then*

$$\Pr\left\{\left|\|X\| - \sqrt{n}\right| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{K^4}\right) \quad \textit{for all } t \geq 0 \tag{11}$$

where $K = \max_i \|X_i\|_{\psi_2}$ and c is an absolute constant.

Lemma B.6. ***(Chernoff's inequality)*** *Let X_i be independent Bernoulli random variables with parameters p_i. Consider their sum $S_N = \sum_{i=1}^{N} X_i$ and denote its mean by $\mu = \mathbb{E}[S_N]$. Then, for any $t > \mu$, we have*

$$\mathbb{P}\{S_N \leq t\} \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t. \tag{12}$$

Appendix C. Phase I: Progressively Diminishing and Escaping the Lazy Regime

C.1. Step 1: Bounding a_i

Let $\nabla\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t)) = f(\boldsymbol{\theta}(t); \boldsymbol{x}_{\xi_t}) - y_{\xi_t} - \epsilon_t$.

Using label noise SGD, for any $i \in [m]$, the gradient at time step t is

$$\frac{\partial\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))}{\partial\boldsymbol{w}_i^\top(t)} = a_i(t)(f(\boldsymbol{\theta}(t); \boldsymbol{x}_{\xi_t}) - y_{\xi_t} - \epsilon_t) \cdot \boldsymbol{x}_{\xi_t}^\top \tag{13}$$

$$\frac{\partial\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))}{\partial a_i(t)} = (f(\boldsymbol{\theta}(t); \boldsymbol{x}_{\xi_t}) - y_{\xi_t} - \epsilon_t) \cdot \boldsymbol{x}_{\xi_t}^\top \cdot \boldsymbol{w}_i(t) \tag{14}$$

Then we retain

$$\boldsymbol{W}(t+1) = \boldsymbol{W}(t) - \eta \cdot \boldsymbol{a}(t) \cdot (f(\boldsymbol{\theta}(t); \boldsymbol{x}_{\xi_t}) - y_{\xi_t} - \epsilon_t) \cdot \boldsymbol{x}_{\xi_t}^\top \tag{15}$$

$$\boldsymbol{a}(t+1) = \boldsymbol{a}(t) - \eta \cdot (f(\boldsymbol{\theta}(t); \boldsymbol{x}_{\xi_t}) - y_{\xi_t} - \epsilon_t) \cdot \boldsymbol{W}(t) \cdot \boldsymbol{x}_{\xi_t} \tag{16}$$

The following Lemma C.1 provides a bound of the initialization of $\|\boldsymbol{w}_i\|$ at step 0.

Lemma C.1. *Let event $B_0 = \{\|\boldsymbol{w_i(0)}\| \leq 1/4 \cdot m^{1/12}$ for all $i \in [m]\}$. Suppose Condition 3.1 (A1) holds. Under NTK initialization as in Section 3.1, we have $\Pr[B_0] \geq 1 - O(\frac{1}{m})$.*

Proof. Notice that $\boldsymbol{w_i(0)} \sim \frac{1}{\sqrt{d}}N(0, I)$, Let $Z = \sqrt{d} \cdot \boldsymbol{w_i(0)}$ and $Z \sim N(0, I)$. By Lemma B.5, we have

$$\Pr[\Big|\|Z\| - \sqrt{d}\Big| \geq \frac{1}{8} \cdot m^{1/12}] \leq 2\exp(-\frac{c \cdot m^{1/6}}{64K_w^4}) \tag{17}$$

where K_w and c both are positive constant.

Thus

$$\Pr[|\|\boldsymbol{w_i}(0)\| - 1| \geq \frac{m^{1/12}}{8\sqrt{d}}] \leq 2\exp(-\frac{c \cdot m^{1/6}}{64K_w^4}) \tag{18}$$

Since $1 \ll m$ and $d \geq 1$, we have

$$\Pr[\|\boldsymbol{w_i}(0)\| \geq \frac{1}{4} \cdot m^{1/12}] \leq 2\exp(-\frac{c \cdot m^{1/6}}{64K_w^4}) \tag{19}$$

Using union bound, we have

$$\Pr[\overline{B_0}] \leq m \cdot 2\exp(-\frac{c \cdot m^{1/6}}{64K_w^4}) \tag{20}$$

Thus

$$\Pr[B_0] \geq 1 - m \cdot 2\exp(-\frac{c \cdot m^{1/6}}{64K_w^4}) \geq 1 - \frac{1}{m} \tag{21}$$

The last inequality holds under Condition 3.1(A1). □

Lemma C.2. *Suppose Condition 3.1 (A1-2,4-6) holds and consider the update rule in Equation (3). Given the model is still in the lazy regime within $T_0 = O(\frac{\log m}{\eta^2 \cdot m^{1/2}})$, with probability at least $1 - O(\frac{1}{m})$, we have the following two propositions hold:*

*(i). (**Bound of Loss**) For every $t \leq T_0$, $\Big|\nabla \hat{\ell}_{\xi_0}(\boldsymbol{\theta}(0))\Big| = O(m^{1/4})$.*

*(ii). (**Bound of** a_i) For every $t \leq T_0$, $a_i(t) \leq m^{-1/4}$.*

Proof. We prove by induction.

When step $t = 0$, we first prove the (i) holds. When $t = 0$, for any $i \in [m]$, let $\boldsymbol{w}_i(0) = \frac{1}{\sqrt{d}}(X_{i1}, X_{i2}, \cdots, X_{id})$, $X_{ij} \sim \mathbf{N}(0,1)$ and $\boldsymbol{a}(0) = \frac{1}{\sqrt{m}}(Y_1, Y_2, \cdots, Y_m)$, $Y_i \sim \mathbf{N}(0,1)$. Let $Z_j = \sum_{i=1}^m X_{ij} \cdot Y_i$ where $j \in [d]$.

For all $i \in [m], j \in [d]$, X_{ij}, Y_i are independent, so we have

$$\mathbb{E}[Z_j] = \mathbb{E}[\sum_{i=1}^m X_{ij}Y_i] = \sum_{i=1}^m \mathbb{E}[X_{ij}Y_i] = \sum_{i=1}^m \mathbb{E}[X_{ij}]\mathbb{E}[Y_i] = 0 \tag{22}$$

and

$$\mathrm{Var}[Z_j] = \mathrm{Var}[\sum_{i=1}^m X_{ij}Y_i] = \sum_{i=1}^m \mathrm{Var}[X_{ij}Y_i] = \sum_{i=1}^m \mathrm{Var}[X_{ij}]\,\mathrm{Var}[Y_i] = m \tag{23}$$

By Corollary B.3, we have for any $j \in [d]$, $Z_j = \sum_{i=1}^m X_{ij} \cdot Y_i \in SE[2\sqrt{m}, 4]$. By Lemma B.4, let $t = \sqrt{(8 \log m)/m}$ and we retain

$$\Pr\left(\left|Z_j = \sum_{i=1}^m X_{ij} \cdot Y_i\right| \geq \sqrt{8 \log m} \cdot \sqrt{m}\right) \leq 2 \exp\left(-\frac{(\sqrt{8 \log m} \cdot \sqrt{m})^2}{2 \cdot (2\sqrt{m})^2}\right) = \frac{2}{m}$$

Using union bound, with probability at most $\frac{2d}{m}$, there exists $j \in [d]$ such that $|\sum_{i=1}^m X_{ij} \cdot Y_i| \geq \sqrt{8 \log m} \cdot \sqrt{m}$

So with probability $1 - \frac{2d}{m}$,

$$\left|\sum_{i=1}^m X_{ij} \cdot Y_i\right| \leq \sqrt{8 \log m} \cdot \sqrt{m} \quad \text{for all } j \in [d] \tag{24}$$

Then we have

$$\begin{aligned}
\|\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*\| &\leq \|\boldsymbol{\theta}(0)\| + \|\boldsymbol{\theta}^*\| \\
&= \|\sum_{i=1}^m a_i(0) \boldsymbol{w}_i(0)^\top\| + \|\boldsymbol{\theta}^*\| \\
&= \frac{1}{\sqrt{d}} \cdot \frac{1}{\sqrt{m}} \sqrt{\sum_{j=1}^d (\sum_{i=1}^m X_{ij} Y_i)^2} + \|\boldsymbol{\theta}^*\| \\
&\leq \frac{1}{\sqrt{d}} \cdot \frac{1}{\sqrt{m}} \sqrt{8 \cdot d \cdot m \cdot \log m} + m^{-1/4} \quad \text{(by Inequality (24) and Condition 3.1(A4))} \\
&\leq 3\sqrt{\log m}
\end{aligned}$$

So we have

$$\begin{aligned}
\left|\nabla \hat{\ell}_{\xi_0}(\boldsymbol{\theta}(0))\right| &= |\boldsymbol{a}(0) \boldsymbol{W}(0) \boldsymbol{x}_{\xi_0} - y_{\xi_0} - \epsilon_0| \\
&\leq |\boldsymbol{a}(0) \boldsymbol{W}(0) \boldsymbol{x}_{\xi_0} - y_{\xi_0}| + |\epsilon_0| \\
&\leq \|(\boldsymbol{\theta}(0) - \boldsymbol{\theta}^*)\| \cdot \|\boldsymbol{x}_{\xi_0}\| + \sigma \\
&\leq 3\sqrt{\log m} \cdot C_{data} + \sigma
\end{aligned}$$

Therefore, with probability at least $1 - \frac{2}{m}$, we have $\left|\nabla \hat{\ell}_{\xi_0}(\boldsymbol{\theta}(0))\right| = O(m^{1/4})$, so (i) holds when $t = 0$.

Then we prove (ii) holds at step 0.

Let event $C_t = \{|a_i(t_1)| \leq m^{-1/4}$ for all $t_1 \leq t$ and for all $i \in [m]\}$. Since $a_i(0) \sim \frac{1}{\sqrt{m}} N(0, 1)$, we have

$$\Pr[|a_i(0)| \geq m^{-1/8}] \leq 2 \exp(-\frac{c \cdot m^{3/4}}{4K_a^4}) \tag{25}$$

by Lemma B.5, where K_a and c are both positive constant.

Using union bound, we have

$$\Pr[\overline{C_0}] \leq m \cdot 2\exp(-\frac{c \cdot m^{3/4}}{4K_a^4}) \cdot (1 - \frac{3d}{m}) \tag{26}$$

Thus

$$\begin{aligned}\Pr[C_0] = 1 - \Pr[\overline{C_0}] &\geq 1 - m \cdot 2\exp(-\frac{c \cdot m^{3/4}}{4K_a^4}) \cdot (1 - \frac{3d}{m}) \\ &\geq (1 - \frac{1}{m}) \cdot (1 - \frac{3d}{m}) \\ &\geq (1 - \frac{4d}{m})\end{aligned}$$

Since d is a constant, we have $\Pr[C_0] \geq 1 - O(\frac{1}{m})$, which implies (ii) holds when $t = 0$.

Assuming the lemma holds at step t, we proceed to step $t + 1$ where we first establish property (ii) by constructing a super-martingale, and subsequently demonstrate that property (i) also holds.

By definition of online label noise SGD algorithm, x_{ξ_t}, ϵ_t are independent with $\boldsymbol{\theta}(t)$, $\boldsymbol{w}_i(t)$ and has memorylessness property. Let $\mathcal{F}_0 = \sigma(\boldsymbol{w}_i(0), a_i(0)$ for i $\in [m])$. For $t \geq 1$, let $\mathcal{F}_t = \sigma(\mathcal{F}_0, \epsilon^{(t)}, \boldsymbol{x}^{(t)})$. Here $\epsilon^{(t)}$ denotes the set $\{\epsilon_0, \epsilon_1, \ldots, \epsilon_{t-1}\}$ and $\boldsymbol{x}^{(t)}$ denotes the set $\{\boldsymbol{x}_{\xi_0}, \boldsymbol{x}_{\xi_1}, \ldots, \boldsymbol{x}_{\xi_{t-1}}\}$.

Obviously, we have $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_t$. Notice that $\mathbb{E}[\boldsymbol{x}_{\xi_t} \cdot \boldsymbol{x}_{\xi_t}^\top | \mathcal{F}_t] = I$ and $\mathbb{E}[\epsilon_t | \mathcal{F}_t] = 0$. Additionally, in the lazy regime, for any $i \in [d]$ at step t, every neuron holds $\|\boldsymbol{w}_i(t) - \boldsymbol{w}_i(0)\| \leq \frac{1}{\sqrt{m}}$. So we have

$$\begin{aligned}\mathbb{E}[|a_i(t+1)|\,|\mathcal{F}_t] &= |a_i(t) - \eta \cdot \mathbb{E}[(\boldsymbol{\theta}(t) - \boldsymbol{\theta}^\star) \cdot \boldsymbol{w_i}(t) | \mathcal{F}_t]| \\ &= \left|a_i(t) - \eta \cdot \mathbb{E}[(\boldsymbol{a}(t)^\top \boldsymbol{W}(t) - \boldsymbol{\theta}^\star) \cdot \boldsymbol{w_i}(t) | \mathcal{F}_t]\right| \\ &= \left|a_i(t) - \eta \cdot \mathbb{E}[(\sum_{j=1}^m \boldsymbol{a}_j(t) \boldsymbol{w}_j(t)^\top - \boldsymbol{\theta}^\star) \cdot \boldsymbol{w_i}(t) | \mathcal{F}_t]\right| \\ &= \left|a_i(t) - \eta \cdot (\sum_{j=1}^m \boldsymbol{a}_j(t) \mathbb{E}[\boldsymbol{w}_j(0)^\top \cdot \boldsymbol{w_i}(0)]) + \eta \cdot \boldsymbol{\theta}^\star \boldsymbol{w_i}(0) + o(\eta/\sqrt{m})\right| \\ &= \left|(1 - \eta) \cdot a_i(t) + \eta \cdot \boldsymbol{\theta}^\star \boldsymbol{w_i}(0) + o(\eta/\sqrt{m})\right|\end{aligned}$$

Let $Y_i(t) = \left|a_i(t) - \boldsymbol{\theta}^* \boldsymbol{w}_i(0) - o(\frac{1}{\sqrt{m}})\right|$. Since

$$\mathbb{E}[\left|a_i(t+1) - \boldsymbol{\theta}^* \boldsymbol{w}_i(0) - o(\frac{1}{\sqrt{m}})\right| | \mathcal{F}_t] = \left|(1 - \eta) \cdot (a_i(t) - \boldsymbol{\theta}^\star \boldsymbol{w}_i(0) - o(\frac{1}{\sqrt{m}}))\right| \tag{27}$$

We have

$$\mathbb{E}[Y_i(t+1) | \mathcal{F}_t] = (1 - \eta) \cdot Y_i(t) \leq Y_i(t) \tag{28}$$

Therefore, $Y_i(0), Y_i(1), \ldots, Y_i(t)$ are super-martingale.

By Lemma C.1, with probability at least $1 - O(\frac{1}{m})$, $\|\boldsymbol{w}_i(0)\| \leq m^{1/12}$ for all $i \in [m]$. Conditioned on $\|\boldsymbol{w}_i(0)\| \leq m^{1/12}$ and we have

$$\begin{aligned}
|Y_i(t+1) - Y_i(t)| &= \left| |a_i(t+1) - \boldsymbol{\theta}^* \cdot \boldsymbol{w}_i(0) - o(\frac{1}{\sqrt{m}})| - |a_i(t) - \boldsymbol{\theta}^* \cdot \boldsymbol{w}_i(0) - o(\frac{1}{\sqrt{m}})| \right| \\
&\leq |a_i(t+1) - a_i(t)| + o(\frac{1}{\sqrt{m}}) \quad \text{(triangle inequality)} \\
&= |\eta \cdot (\nabla \hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))) \cdot \boldsymbol{x}_{\xi_t}^\top \cdot \boldsymbol{w}_i(t)| + o(\frac{1}{\sqrt{m}}) \\
&\leq \eta \cdot |\nabla \hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))| \cdot \|\boldsymbol{x}_{\boldsymbol{\xi_t}}\| \cdot \|\boldsymbol{w}_i(t)\| + o(\frac{1}{\sqrt{m}}) \\
&\leq \eta \cdot |\nabla \hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))| \cdot \|\boldsymbol{x}_{\boldsymbol{\xi_t}}\| \cdot (\|\boldsymbol{w}_i(0)\| + \frac{1}{\sqrt{m}}) + o(\frac{1}{\sqrt{m}}) \quad \text{(in the lazy regime)} \\
&\leq 2\eta \cdot |\nabla \hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))| \cdot C_{data} \cdot m^{1/12}
\end{aligned}$$

With one-side Azuma's inequality, for any $\lambda > 0$, we have

$$\Pr[Y_i(t+1) - Y_i(0) > \lambda] \leq \exp(-\frac{\lambda^2}{(t+1) \cdot 4\eta^2 \cdot (C_{data} \cdot |\nabla \hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))| \cdot m^{1/12})^2}) \tag{29}$$

Thus

$$\Pr[|a_i(t+1)| \geq 2 \cdot |\boldsymbol{\theta}^* \cdot \boldsymbol{w}_i(0)| + |a_i(0)| + \lambda] \leq \exp(-\frac{\lambda^2}{(t+1) \cdot 4\eta^2 \cdot |\nabla \hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))| \cdot (C_{data} \cdot m^{1/12})^2}) \tag{30}$$

By Condition 3.1(A4), $\|\boldsymbol{\theta}^*\| \leq \frac{1}{m^{1/3}}$. Then we have $|\boldsymbol{\theta}^* \cdot \boldsymbol{w}_i(0)| \leq \frac{1}{4m^{1/4}}$. Notice that $\Pr[|a_i(0)| \geq m^{-1/4}] \leq 2\exp(-\frac{c \cdot m^{1/4}}{4K_a^4})$. Let $\lambda = \frac{1}{2m^{1/4}}$, so with probability at least $(1 - 2\exp(-\frac{c \cdot m^{1/4}}{4K_a^4})) \cdot (1 - \frac{3d}{m})$, we have

$$2 \cdot |\boldsymbol{\theta}^* \cdot w_i(0)| + |w_i(0)| + \lambda] \leq \frac{1}{2m^{1/4}} + \frac{1}{4m^{1/4}} + \frac{1}{4m^{1/4}} \leq \frac{1}{m^{1/4}}$$

by Condition 3.1(A1).

So we retain

$$\Pr[|a_i(t+1)| \geq m^{-1/4}] \leq \exp(-\frac{m^{1/6}}{(t+1) \cdot 8\eta^2 \cdot (C_{data} \cdot |\nabla \hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))|)^2}) \cdot (1 - \frac{3d}{m}) \tag{31}$$

By induction hypothesis, we have $|\nabla\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))| = O(m^{1/4})$. Using union bound and by Condition 3.1(A2), we have

$$\begin{aligned}
\Pr[\overline{C_t}] &\leq \sum_{j=0}^{t-1} \exp(-\frac{m^{1/12}}{j\cdot 8\eta^2\cdot(C_{data}\cdot O(m^{1/4}))^2})\cdot(1-\frac{3d}{m}) \\
&\leq t\cdot\exp(-\frac{m^{1/12}}{t\cdot 8\eta^2\cdot(C_{data}\cdot O(m^{1/4}))^2})\cdot(1-\frac{3d}{m}) \\
&\leq T_0\cdot\exp(-\frac{\sqrt{\eta}^{-1/12}}{T_0\cdot 8\eta^2\cdot(C_{data}\cdot O(m^{1/4}))^2})\cdot(1-\frac{3d}{m}) \\
&\leq T_0\cdot\exp(-\frac{\sqrt{\eta}^{-1/12}}{T_0\cdot 8\eta^2\cdot(C_{data}\cdot O(m^{1/4}))^2})\cdot(1-\frac{3d}{m}) \\
&= O(\frac{-\ln\eta}{\eta^2})\cdot\exp(-\frac{\sqrt{\eta}^{-1/12}}{O(\ln\frac{1}{\eta})\cdot C_{data}^2}\cdot O(1))\cdot(1-\frac{3d}{m}) \\
&\leq O(\eta)\cdot(1-\frac{3d}{m}) = O(\frac{1}{m})
\end{aligned}$$

The last inequality is due to Condition 3.1(A2). Therefore, we have $\Pr[C_t] \geq 1 - O(\frac{1}{m})$, which implies (ii) holds.

Then we prove (i) holds. By Definition 3.1, we have $\|\boldsymbol{w}_i(t+1) - \boldsymbol{w}_i(t)\| \leq m^{-1/2}$. Therefore, with the bound of $a_i(t)$ holds, we retain

$$\|\boldsymbol{\theta}(t+1)\| \leq \|\boldsymbol{a}(t)^\top\cdot\boldsymbol{W}(t)\| \leq \|\boldsymbol{a}(0)^\top\cdot\boldsymbol{W}(0)\| + m^{-1/2}\cdot m^{-1/4}\cdot m = \|\boldsymbol{\theta}(0)\| + m^{1/4} \tag{32}$$

Therefore,

$$\begin{aligned}
\left|\nabla\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t+1))\right| &= \left|\boldsymbol{a}(t+1)\boldsymbol{W}(t+1)\boldsymbol{x}_{\xi_{t+1}} - y_{\xi_{t+1}} - \epsilon_{t+1}\right| \\
&\leq \left|\boldsymbol{a}(t+1)\boldsymbol{W}(t+1)\boldsymbol{x}_{\xi_{t+1}} - y_{\xi_{t+1}}\right| + |\epsilon_{t+1}| \\
&\leq \|(\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}^*)\|\cdot\|\boldsymbol{x}_{\xi_{t+1}}\| + \sigma \\
&\leq 2(\|\boldsymbol{\theta}(0)\| + m^{1/4})\cdot C_{data} + \sigma \\
&= O(m^{1/4})
\end{aligned}$$

which implies (i) also holds at step $t+1$. So it follows by induction that the lemma holds. □

C.2. Step 2: Estimating ΔW_i

Let $\nabla\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t)) = f(\boldsymbol{\theta}(t);\boldsymbol{x}_{\xi_t}) - y_{\xi_t} - \epsilon_t$, by equ (13) and equ (14), we have

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) - \eta\cdot a_i(t)\cdot\nabla\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))\cdot\boldsymbol{x}_{\xi_t} \tag{33}$$

$$a_i(t+1) = a_i(t) - \eta\cdot\nabla\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))\cdot\boldsymbol{x}_{\xi_t}^\top\cdot\boldsymbol{w}_i(t) \tag{34}$$

According to equ (33), after taking the norm on both sides and then square them, we have:

$$\|\boldsymbol{w}_i(t+1)\|^2 = \|\boldsymbol{w}_i(t)\|^2 - 2\eta \cdot a_i(t)\nabla\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))\boldsymbol{x}_{\xi_t}^\top \boldsymbol{w}_i(t) + \|\eta \cdot a_i(t)\nabla\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))\boldsymbol{x}_{\xi_t}\|^2 \tag{35}$$

According to equ (34), we have

$$\eta \cdot \nabla\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t)) \cdot \boldsymbol{x}_{\xi_t}^\top \cdot \boldsymbol{w}_i(t) = a_i(t) - a_i(t+1) \tag{36}$$

Substitute equ (36) into the equ (35) and we retain:

$$\|\boldsymbol{w}_i(t+1)\|^2 = \|\boldsymbol{w}_i(t)\|^2 - 2 \cdot (a_i(t) - a_i(t+1)) \cdot a_i(t) + \|\eta \cdot a_i(t)\nabla\hat{\ell}_{\xi_t}(\boldsymbol{\theta}(t))x_{\xi_t}\|^2 \tag{37}$$

For any time $T \in \mathbb{N}^+$, summing up from 0 to T and we have:

$$\|\boldsymbol{w}_i(T)\|^2 = \|\boldsymbol{w}_i(0)\|^2 - 2\cdot\sum_{j=0}^{T-1}(a_i(j)-a_i(j+1))\cdot a_i(j) + \sum_{j=0}^{T-1}\|\eta\cdot a_i(j)\cdot\nabla\hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))\cdot\boldsymbol{x}_{\xi_j}\|^2 \tag{38}$$

Notice that

$$2 \cdot \sum_{j=0}^{T-1}(a_i(j) - a_i(j+1)) \cdot a_i(j) = \sum_{j=0}^{T-1}(a_i(j)^2 - 2 \cdot a_i(j)a_i(j+1) + a_i(j+1)^2) + a_i(0)^2 - a_i(T)^2 \tag{39}$$

$$= \sum_{j=0}^{T-1}(a_i(j) - a_i(j))^2 + a_i(0)^2 - a_i(T)^2 \tag{40}$$

Thus we have

$$\begin{aligned}
\|\boldsymbol{w}_i(T)\|^2 &= \|\boldsymbol{w}_i(0)\|^2 - \sum_{j=0}^{T-1}(a_i(j) - a_i(j+1))^2 - a_i(0)^2 + a_i(T)^2 + \sum_{j=0}^{T-1}\|\eta \cdot \boldsymbol{w}_i(j) \cdot \nabla\hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j)) \cdot x_{\xi_j}\|^2 \\
&= \|\boldsymbol{w}_i(0)\|^2 - \sum_{j=0}^{T-1}(\|a_i(j) - a_i(j+1)\|^2 - \|\boldsymbol{w}_i(j) - \boldsymbol{w}_i(j+1)\|^2) - a_i(0)^2 + a_i(T)^2 \\
&= \|\boldsymbol{w}_i(0)\|^2 - \sum_{j=0}^{T-1}\eta^2 \cdot \nabla\hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2\{(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2\} - a_i(0)^2 + a_i(T)^2
\end{aligned}$$

Let $\Delta W_i(j) = -\nabla\hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2 \cdot \{(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2\}$. Since $a_i(j)$ is small with high probability, $\Delta W_i(j)$ almost dominates the change of $\|\boldsymbol{w}_i\|^2$ at every step. We have

$$\|\boldsymbol{w}_i(T)\|^2 = \|\boldsymbol{w}_i(0)\|^2 + \eta^2 \cdot \sum_{j=0}^{T-1}\Delta W_i(j) - a_i(0)^2 + a_i(T)^2 \tag{41}$$

Lemma C.3 (Progressively diminishing at each step)**.** *Suppose Condition 3.1 (A1-2,4-6) holds and consider the update rule in Equation (3). Given 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 $j \le T_1$ and for every $i \in [m]$:*

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

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

Proof.

By Lemma C.2 (ii), with probability at least $1 - O(\frac{1}{m})$, for all $i \in [m]$ and all step $t \leq T_1$, $|a_i(t)| \leq \frac{1}{m^{1/4}}$, i.e. the event C_j happens. All the "Pr" in this lemma conditioned on C_j.

For each $j < T$, we have

$$\Delta W_i(j) > 0 \iff (\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 < a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2 \iff (\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j))^2 < a_i(j)^2$$

i.e.

$$\Big|\mathbf{x}^\top \cdot \boldsymbol{w}_i(j)| < |a_i(j)\Big| \tag{42}$$

where $\mathbf{x} = \frac{\boldsymbol{x}_{\xi_j}}{\|\boldsymbol{x}_{\xi_j}\|}$ follows a uniform distribution on the n-dimensional unit sphere.

Let x_1 denotes the element in the first dimension of $\mathbf{x}$. By symmetry, we have

$$\Pr[\Big|\mathbf{x}^\top \cdot \boldsymbol{w}_i(j)| < |a_i(j)\Big|] = \Pr[-\frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|} < x_1 < \frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|}] \tag{43}$$

The density of x_1 is $f(x_1) = \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\,\Gamma(\frac{d-1}{2})} \left(1 - x_1^2\right)^{\frac{d-3}{2}}$, where Γ denotes the gamma function.

Then we have

$$\begin{aligned}
\Pr[-\frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|} < x_1 < \frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|}] &= \int_{x_1=-\frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|}}^{\frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|}} f(x_1)dx_1 \\
&= \int_{x_1=-\frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|}}^{\frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|}} \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)} \left(1 - x_1^2\right)^{\frac{d-3}{2}} dx_1 \\
&\leq \int_{x_1=-\frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|}}^{\frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|}} \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)} dx_1 \\
&= 2 \cdot \frac{\|a_i(j)\|}{\|\boldsymbol{w}_i(j)\|} \cdot \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)} \\
&\leq \frac{2\sqrt{d}}{\sqrt{\pi}} \cdot \frac{\|a_j(j)\|}{\|\boldsymbol{w}_i(j)\|} \\
&\leq \frac{\rho}{m^{1/8}}
\end{aligned}$$

The second-to-last inequality is because $\frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})} \leq \sqrt{d}$. So we retain $\Delta W_i(j) \leq 0$ with probability at least $1 - \frac{\rho}{m^{1/8}}$ and $\Delta W_i(j) > 0$ with probability at most $\frac{\rho}{m^{1/8}}$.

By Lemma C.2 (i) , $\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j)) \leq O(m^{1/4})$. So with Condition 3.1A(5), $\|\boldsymbol{x}_{\xi_j}\| \leq C_{data}$ we have

$$\begin{aligned}
\Delta W_i(j) &= -\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2 \cdot (\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 + \nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2 \cdot (a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2) \\
&\leq \nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2 \cdot (a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2) \\
&\leq \frac{O(m^{1/4})^2 \cdot C_{data}^2}{m^{1/2}} = O(1)
\end{aligned}$$

Finally, we prove $\Pr[\Delta W_i(j) \leq -(\frac{\sigma}{4})^2] \geq \frac{1}{4}$. We have

$$\Pr[\Delta W_i(j) \leq -(\frac{\sigma}{4})^2] = \Pr[\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2\{(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2\} \geq (\frac{\sigma}{4})^2]$$

$$= \Pr[\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2\{(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2\} \geq (\frac{\sigma}{4})^2 |\, \boldsymbol{x}_{\xi_j} \;\; s.t. \;\; |\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)| < |a_i(j)|] \cdot \Pr[|\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j)| < |a_i(j)|]$$

$$+ \Pr[\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2\{(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2 \geq (\frac{\sigma}{4})^2 |\, \boldsymbol{x}_{\xi_j} \;\; s.t. \;\; |\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)| \geq |a_i(j)|] \cdot \Pr[|\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j)| \geq |a_i(j)|]$$

$$\geq \Pr[\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2\{(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2\} \geq (\frac{\sigma}{4})^2 |\, \boldsymbol{x}_{\xi_j} \;\; s.t. \;\; |\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)| \geq |a_i(j)|] \cdot \Pr[|\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j)| \geq |a_i(j)|]$$

$$\geq (1\text{- } \frac{\rho}{m^{1/8}}) \cdot \Pr[\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2\{(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2\} \geq (\frac{\sigma}{4})^2 |\, \boldsymbol{x}_{\xi_j} \;\; s.t. \;\; |\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)| \geq |a_i(j)|]$$

Since $\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2 = ((\boldsymbol{\theta}(j) - \boldsymbol{\theta}^*)\boldsymbol{x}_{\xi_j} - \epsilon_j)^2$ and ϵ_j is chosen uniformly,the probability that $(\boldsymbol{\theta}(j) - \boldsymbol{\theta}^*)\boldsymbol{x}_{\xi_j}$ and ϵ_j have the same sign is at least $\frac{1}{2}$. If the two elements have the same sign, then $\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2 \geq \sigma^2$. So we have

$$\Pr[\nabla \hat{\ell}_{\xi_j}(\boldsymbol{\theta}(j))^2\{(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2\} \geq (\frac{\sigma}{4})^2 |\, \boldsymbol{x}_{\xi_j} \;\; s.t. \;\; |\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)| \geq |a_i(j)|]$$

$$\geq \frac{1}{2} \cdot \Pr[\{(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - a_i(j)^2 \cdot \|\boldsymbol{x}_{\xi_j}\|^2\} \geq (\frac{1}{4})^2 |\, \boldsymbol{x}_{\xi_j} \;\; s.t. \;\; |\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)| \geq |a_i(j)|]$$

$$\geq \frac{1}{2} \cdot \Pr[\{(\boldsymbol{x}_{\xi_j}^\top \cdot \boldsymbol{w}_i(j))^2 - \frac{1}{m^{1/4}} \cdot \|\boldsymbol{x}_{\xi_j}\|^2\} \geq (\frac{1}{4})^2 |\, \boldsymbol{x}_{\xi_j} \;\; s.t. \;\; |\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)| \geq |a_i(j)|] \quad (a_i(j)^2 \leq m^{1/4})$$

$$= \frac{1}{2} \cdot \Pr[\left|\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)\right| \geq \sqrt{(\frac{1}{4 \cdot \|\boldsymbol{x}_{\xi_j}\|})^2 + \frac{1}{m^{1/4}}} |\, \boldsymbol{x}_{\xi_j} \;\; s.t. \;\; |\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)| \geq |a_i(j)|]$$

$$= \frac{1}{2} \cdot \Pr[\left|\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)\right| \geq \sqrt{(\frac{1}{4 \cdot \|\boldsymbol{x}_{\xi_j}\|})^2 + \frac{1}{m^{1/4}}}\,] \qquad (|a_i(j)| \leq \frac{1}{m^{1/8}} \leq \sqrt{\frac{1}{\|\boldsymbol{x}_{\xi_j}\|^2} + \frac{1}{m^{1/4}}}\,)$$

$$\geq \frac{1}{2} \cdot \Pr[\left|\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|} \cdot \boldsymbol{w}_i(j)\right| \geq \sqrt{(\frac{1}{4 \cdot \|\boldsymbol{x}_{\xi_j}\|})^2 \cdot \frac{3}{2}}\,]$$

By Condition 3.1A(6), we have

$$\Pr[\Big|\|\boldsymbol{x}_i\| - \sqrt{d}\Big| \geq \frac{\sqrt{d}}{3}] \leq 2\exp(-\frac{4\cdot c\cdot d}{9K^4}) \leq \frac{1}{2} \tag{44}$$

which implies $\Pr[\|\boldsymbol{x}_{\xi_j}\| \geq \frac{2}{3\sqrt{d}}] \geq \frac{2}{3}$. So we have

$$\begin{aligned}
\Pr[\left|\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|}\cdot \boldsymbol{w}_i(j)\right| \geq \sqrt{(\frac{1}{4\cdot\|\boldsymbol{x}_{\xi_j}\|})^2\cdot\frac{3}{2}}\,] &\geq \Pr[\left|\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|}\cdot \boldsymbol{w}_i(j)\right| \geq \frac{3}{2\sqrt{d}}\cdot\frac{1}{4\cdot\|\boldsymbol{x}_{\xi_j}\|}] \\
&= 1 - \Pr[\left|\frac{\boldsymbol{x}_{\xi_j}^\top}{\|\boldsymbol{x}_{\xi_j}\|}\cdot \boldsymbol{w}_i(j)\right| \leq \frac{3}{8\sqrt{d}}] \\
&\geq 1 - \rho\cdot\frac{3}{8\sqrt{d}} \\
&= 1 - \frac{2\sqrt{d}}{\sqrt{\pi}}\cdot\frac{3}{8\sqrt{d}} = 1 - \frac{3}{4\sqrt{\pi}}
\end{aligned}$$

By Condition 3.1(A1), $m \geq ((1-\frac{3}{4\sqrt{\pi}})\cdot\rho/(\frac{1}{2}-\frac{3}{4\sqrt{\pi}}))^8$. Therefore, we have

$$\begin{aligned}
\Pr[\Delta W_i(j) \leq -(\frac{\sigma}{4})^2] &\geq (1-\frac{\rho}{m^{1/8}})\cdot\frac{1}{2}\cdot(1-\frac{3}{4\sqrt{\pi}}) \\
&= \frac{1}{4} + \frac{1}{2}\cdot((\frac{1}{2}-\frac{3}{4\sqrt{\pi}}) - (1-\frac{3}{4\sqrt{\pi}})\cdot\frac{\rho}{m^{1/8}}) \\
&\geq \frac{1}{4}
\end{aligned}$$

which completes the proof. □

Theorem C.4 (Escaping the lazy regime)**.** *Suppose Condition 3.1 (A1-2,4-6) holds 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 $T_1 = \frac{384\sqrt{\log m}}{\sigma^2\eta^2\sqrt{m}}$.*

Proof. Let event $C_t = \{|a_i(t_1)| \leq m^{-1/4}$ for all $t_1 \leq t$ and for all $i\in[m]\}$. By Lemma C.2 (ii), with probability at least $1-O(\frac{1}{m})$, the event C_{T_1} happens. We assume C_{T_1} happens and all the "Pr" in this theorem conditioned on C_j.

In the following, we will provide a proof by contradiction. Assume with probability at least $O(\frac{1}{m})$, there exists some neurons $\boldsymbol{w}_i$ ($i\in[m]$) s.t. $\|\boldsymbol{w}_i(t)-\boldsymbol{w}_i(0)\| \leq \frac{1}{\sqrt{m}}$ holds for all T_1 steps. By Lemma B.5, we have

$$\Pr\left\{\Big|\|\boldsymbol{w}_i(0)\| - \sqrt{2\log m}\geq t\Big|\right\} \leq 2\exp\left(-\frac{c\cdot\log m}{k^4}\right) = O(\frac{1}{m}) \tag{45}$$

where k and c is an absolute constant.

Using union bound, with probability at least $1-O(\frac{1}{m})$, for these neurons stuck in the lazy regime, $\|\boldsymbol{w}_i(0)\| \leq \sqrt{2\log m}$.

We view $\Delta W_i(j)$ as a random variable. We use Ω to denote the whole sample space. Let event

$$\Omega_i = \{\omega \in \Omega \,|\Delta W_i(j)(\omega) \leq -(\frac{\sigma}{4})^2\}$$

For any $\omega \in \Omega$, by Lemma C.3, we have $\Pr[\Omega_i] \geq \frac{1}{4}$. Since the event C_{T_1} happens and $\boldsymbol{w}_i$ stays in the lazy regime, the estimation of the lower bound of $\Delta W_i(j)$ only depends on the randomness of $\boldsymbol{x}_{\xi_j}$. Thus, there exists a subset of Ω_i we denote Θ_i such that

$$\Theta_i = \{\omega \in \Omega_i | \Delta W_i(\omega) \leq -(\sigma/4)^2, \|\boldsymbol{w}_i(j) - \boldsymbol{w}_i(0)\| \leq \frac{1}{\sqrt{m}}, |a_i(j)| \leq 1/m^{1/8}\}$$

and $\Pr[\Theta_i] = \frac{1}{4}$. For Θ_i ($i \in [T_1 - 1]$), the random randomness only depends on $\boldsymbol{x}_{\xi_t}$. Therefore, $\Theta_0, \Theta_1, \cdots, \Theta_{T_1-1}$ are mutual independent events.

Then we define indicator random variable $X_0, X_1, \cdots, X_{T_1-1}$ as follow: for every $\omega \in \Omega$, for any $j \in [T_1 - 1]$,

$$X_j(\omega) = \begin{cases} 1 & \text{if } \omega \in \Theta_i. \\ 0 & \text{otherwise.} \end{cases}$$

So we have $X_0, X_1, \cdots, X_{T_1-1}$ are independent and $\Pr[X_j(\omega) = 1] = \frac{1}{4}$. And we have

$$\mathbb{E}[\sum_{j=0}^{T_1-1} X_j] = \sum_{j=0}^{T_1-1} \mathbb{E}[X_j] = \frac{T_1}{4} \tag{46}$$

Then we have

$$\begin{aligned}
&\Pr[\eta^2 \cdot \sum_{j=0}^{T_1-1} \Delta W_i(j) \leq \frac{-2\sqrt{2\log m}}{\sqrt{m}}] \\
&= \Pr[\eta^2 \cdot (\sum_{j=0}^{T_1-1} \Delta W_i(j) \cdot \mathbb{I}[\Delta W_i(j) > 0] + \sum_{j=0}^{T_1-1} \Delta W_i(j) \cdot \mathbb{I}[\Delta W_i(j) \leq 0]) \leq \frac{-2\sqrt{2\log m}}{\sqrt{m}}] \\
&\geq \Pr[\eta^2 \cdot (\sum_{j=0}^{T_1-1} \Delta W_i(j) \cdot O(\frac{1}{m^{1/8}}) + \sum_{j=0}^{T_1-1} \Delta W_i(j) \cdot \mathbb{I}[\Delta W_i(j) \leq -(\frac{\sigma}{4})^2]) \leq \frac{-2\sqrt{2\log m}}{\sqrt{m}}] \\
&\geq \Pr[\eta^2 \cdot T_1 \cdot O(1) \cdot O(\frac{1}{m^{1/8}}) - \eta^2 \cdot (\frac{\sigma}{4})^2 \sum_{j=0}^{T-1} X_j \leq \frac{-2\sqrt{2\log m}}{\sqrt{m}}] \\
&= \Pr[-\eta^2 \cdot (\frac{\sigma}{4})^2 \cdot \sum_{j=0}^{T_1-1} X_j \leq \frac{-2\sqrt{2\log m}}{\sqrt{m}} - o(\frac{1}{\sqrt{m}})] \quad (T_1 \cdot \eta^2 = O(\frac{1}{\sqrt{m}})) \\
&\geq \Pr[-\eta^2 \cdot (\frac{\sigma}{4})^2 \cdot \sum_{j=0}^{T_1-1} X_j \leq -\frac{3\sqrt{\log m}}{\sqrt{m}}] \\
&= \Pr[\sum_{j=0}^{T_1-1} X_i \geq \frac{48\sqrt{\log m}}{\sigma^2\eta^2\sqrt{m}}] = \Pr[\sum_{j=0}^{T_1-1} X_i \geq \frac{T_1}{8}] \\
&\geq 1 - \exp(-\frac{T_1}{4}) \cdot (\frac{e \cdot T_1/4}{T_1/8})^{T_1/8} = 1 - (\frac{2}{e})^{T_1/8}
\end{aligned}$$

The last inequality is by Lemma B.6.

Since $T_1 = \frac{1}{\eta^2 \cdot \sqrt{m}} \gg m$ and $a_i(T)^2 \leq \frac{1}{\sqrt{m}}$, with probability at least $(1-O(\frac{1}{m}))\cdot(1-(\frac{2}{e})^{T_1/8}) = 1 - O(\frac{1}{m})$, we have

$$\begin{aligned}
\|\boldsymbol{w}_i(T)\|^2 - \|\boldsymbol{w}_i(0)\|^2 &= \eta^2 \cdot \sum_{j=0}^{T-1} \Delta W_i(j) - a_i(0)^2 + a_i(T)^2 \\
&\leq \eta^2 \cdot \sum_{j=0}^{T-1} \Delta W_i(j) + \frac{1}{\sqrt{m}} \\
&\leq -2\frac{\sqrt{2\log m}}{\sqrt{m}} + \frac{1}{\sqrt{m}} \leq -\frac{2\sqrt{\log m}}{\sqrt{m}}
\end{aligned}$$

Thus

$$\begin{aligned}
\|\boldsymbol{w}_i(T)\| - \|\boldsymbol{w}_i(0)\| &\leq -\frac{2\sqrt{\log m}}{\sqrt{m}} \cdot \frac{1}{\|\boldsymbol{w}_i(T)\| + \|\boldsymbol{w}_i(0)\|} \\
&\leq -\frac{2\sqrt{\log m}}{\sqrt{m}} \cdot \frac{1}{\frac{1}{\sqrt{m}} + \sqrt{2\log m}} \\
&< -\frac{2\sqrt{\log m}}{\sqrt{m}} \cdot \frac{1}{2\sqrt{\log m}} = -\frac{1}{\sqrt{m}}
\end{aligned}$$

Which is a contradiction to the definition of lazy regime! Therefore, we complete the proof. □

Appendix D. Phase II: Feature Learning and Convergence

D.1. Rotation to alignment

We use gradient descent in phase II. For every neuron $\boldsymbol{w}_i$ and a_i ($i \in [m]$) at step t:

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}(t))}{\partial \boldsymbol{w}_i(t)^\top} = a_i(t) \cdot \frac{1}{n} \sum_{j=1}^{m} (\boldsymbol{a}(t)^\top \boldsymbol{W}(t)\boldsymbol{x}_j - \boldsymbol{y}_j) \cdot \boldsymbol{x}_j^\top \tag{47}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}(t))}{\partial a_i(t)} = \frac{1}{n} \sum_{j=1}^{m} (\boldsymbol{a}(t)^\top \boldsymbol{W}(t)\boldsymbol{x}_j - \boldsymbol{y}_j) \cdot \boldsymbol{x}_j^\top \cdot \boldsymbol{w}_i(t)^\top \tag{48}$$

The update rule of gradient descent is

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) - \eta \cdot a_i(t) \cdot \frac{1}{n} \sum_{j=1}^{m} (\boldsymbol{a}(t)^\top \boldsymbol{W}(t)\boldsymbol{x}_j - \boldsymbol{y}_j)\boldsymbol{x}_j \tag{49}$$

$$a_i(t+1) = a_i(t) - \eta \cdot \frac{1}{n} \sum_{j=1}^{m} (\boldsymbol{a}(t)^\top \boldsymbol{W}(t)\boldsymbol{x}_j - \boldsymbol{y}_j)\boldsymbol{x}_j^\top \cdot \boldsymbol{w}_i(t) \tag{50}$$

By law of large numbers, we have

$$\frac{1}{n} \sum_{j=1}^{n} \boldsymbol{x}_j \cdot \boldsymbol{x}_j^\top = \mathbb{E}[\boldsymbol{x}_j \cdot \boldsymbol{x}_j^\top] + O(\frac{1}{\sqrt{n}}) = I + O(\frac{1}{\sqrt{n}}) \tag{51}$$

Then we have

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) - \eta \cdot a_i(t) \cdot \frac{1}{n}\sum_{j=1}^{m} \boldsymbol{x}_j \cdot \boldsymbol{x}_j^\top \cdot (\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*) \tag{52}$$

$$= \boldsymbol{w}_i(t) - \eta \cdot a_i(t) \cdot (\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*) + O(\frac{\eta}{\sqrt{n}}) \tag{53}$$

And

$$a_i(t+1) = a_i(t) - \eta \cdot (\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*)\frac{1}{n}\sum_{j=1}^{m} \boldsymbol{x}_j \cdot \boldsymbol{x}_j^\top \cdot \boldsymbol{w}_i(t)$$

$$= a_i(t) - \eta \cdot (\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*) \cdot \boldsymbol{w}_i(t) + O(\frac{\eta}{\sqrt{n}})$$

Since $a_i(t), \|\boldsymbol{w}_i(t)\| \leq n^{0.5}$ when phase II begins, we have

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) - \eta \cdot a_i(t) \cdot (\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*) + O(\frac{\eta}{\sqrt{n}})$$

$$= \boldsymbol{w}_i(t) + \eta \cdot (a_i(t) \cdot \boldsymbol{\theta}^*) + O(\eta^{1.5})$$

Similarly, we have:

$$a_i(t+1) = a_i(t) + \eta \cdot \boldsymbol{\theta}^* \cdot \boldsymbol{w}_i(t) + O(\eta^{1.5})$$

Combine both we have:

$$\begin{bmatrix} \boldsymbol{w}_i(t+1) \\ a_i(t+1) \end{bmatrix} = (\boldsymbol{I} + \eta \boldsymbol{M}) \cdot \begin{bmatrix} \boldsymbol{w}_i(t) \\ a_i(t) \end{bmatrix} + O(\eta^{1.5}) \tag{54}$$

where $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{\theta}^{*\top} \\ \boldsymbol{\theta}^* & 0 \end{bmatrix}$. The top eigenvalue of $\boldsymbol{M}$ is $\lambda_1 = \|\boldsymbol{\theta}^*\|$ and the lowest eigenvalue of $\boldsymbol{M}$ is $\lambda_{n+1} = -\|\boldsymbol{\theta}^*\|$. All the other eigenvalues of $\boldsymbol{M}$ are equal to 0. Since $\boldsymbol{M}$ is symmetry matrix, there exists orthogonal matrix $\boldsymbol{Q}_{\boldsymbol{M}}$ such that

$$\boldsymbol{M} = \boldsymbol{Q}_M^\top \cdot \text{diag}(\|\boldsymbol{\theta}^*\|, 0, \ldots, -\|\boldsymbol{\theta}^*\|) \cdot \boldsymbol{Q}_M \tag{55}$$

Lemma D.1 (Alignment)**.** *Suppose Condition 3.1 (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*

$$|\langle \boldsymbol{\theta}^\star, \boldsymbol{w}_i(t_2) \rangle| \geq 1 - \left| O(\ln \frac{1}{\eta} \cdot \sqrt{\eta}) \right| \tag{56}$$

Proof. By equ (54), we have

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{w}_i(t_1+T_2) \\ a_i(t_1+T_2) \end{bmatrix} &= (\boldsymbol{I}+\eta\cdot\boldsymbol{M})^{T_2}\cdot\begin{bmatrix} \boldsymbol{w}_i(t_1) \\ a_i(t_1) \end{bmatrix} + O(T_2\cdot\eta^{1.5}) \\
&= \sum_{k=0}^{T_2}\binom{T_2}{k}(\eta\boldsymbol{M})^k\cdot\begin{bmatrix} \boldsymbol{w}_i(t_1) \\ a_i(t_1) \end{bmatrix} + O(T_2\cdot\eta^{1.5}) \\
&= (I+T_2\eta\boldsymbol{M}+\frac{T_2(T_2-1)}{2}(\eta\boldsymbol{M})^2+\cdots+(\eta M)^T)\cdot\begin{bmatrix} \boldsymbol{w}_i(t_1) \\ a_i(t_1) \end{bmatrix} + O(T_2\cdot\eta^{1.5}) \\
&= (\sum_{i=0}^{T_2}\frac{(T_2\eta\boldsymbol{M})^i}{i!}+O(T_2\cdot\eta^2))\cdot\begin{bmatrix} \boldsymbol{w}_i(t_1) \\ a_i(t_1) \end{bmatrix} + O(T_2\cdot\eta^{1.5}) \\
&= (\exp(T_2\eta\cdot\boldsymbol{M})+O(\frac{(T_2\eta\cdot\boldsymbol{M})^{T+1}}{(T+1)!}))\cdot\begin{bmatrix} \boldsymbol{w}_i(t_1) \\ a_i(t_1) \end{bmatrix} + O(T_2\cdot\eta^{1.5}) \quad \text{(taylor expansion)} \\
&= \frac{e\cdot\boldsymbol{M}}{\|\boldsymbol{\theta}^*\|}\begin{bmatrix} \boldsymbol{w}_i(t_1) \\ a_i(t_1) \end{bmatrix} + O(T_2\cdot\eta^{1.5}) \quad (\eta^{-\ln\eta}\ll(-\ln\eta)^{-\ln\eta})
\end{aligned}$$

So we have

$$\boldsymbol{w}_i(t_1+T_2) = \frac{e\cdot a_i(t_1)}{\|\boldsymbol{\theta}^*\|_2}\cdot\boldsymbol{\theta}^{*\top} + O(\ln\frac{1}{\eta}\cdot\eta^{1.5}) \tag{57}$$

Then we have,

$$\begin{aligned}
|\langle\boldsymbol{\theta}^*,\boldsymbol{w}_i(t_2)\rangle| &= \left|\langle\boldsymbol{\theta}^*,\frac{e\cdot a_i(t_1)}{\|\boldsymbol{\theta}^*\|_2}\cdot\boldsymbol{\theta}^{*\top}+O(\ln\frac{1}{\eta}\cdot\eta^{1.5})\rangle\right| \\
&= \frac{\left|\|\boldsymbol{\theta}^*\|\cdot e\cdot a_i(t_1)+\boldsymbol{\theta}^*\cdot O(\ln\frac{1}{\eta}\cdot\eta^{1.5})\right|}{\|\boldsymbol{\theta}^*\|\cdot\|\boldsymbol{w}_i(t_2)\|} \\
&= \frac{\left|e\cdot a_i(t_1)+\frac{\boldsymbol{\theta}^*}{\|\boldsymbol{\theta}^*\|}\cdot O(\ln\frac{1}{\eta}\cdot\eta^{1.5})\right|}{\left\|\frac{e\cdot a_i(t_1)}{\|\boldsymbol{\theta}^*\|}\cdot\boldsymbol{\theta}^{*\top}+O(\ln\frac{1}{\eta}\cdot\eta^{1.5})\right\|} \\
&\geq \frac{e\cdot a_i(t_1)-\left|O(\ln\frac{1}{\eta}\cdot\eta^{1.5})\right|}{e\cdot a_i(t_1)+\left|O(\ln\frac{1}{\eta}\cdot\eta^{1.5}))\right|} \quad \text{(triangle inequality)} \\
&= 1-\frac{\left|O(\ln\frac{1}{\eta}\cdot\eta^{1.5})\right|}{e\cdot a_i(t_1)+\left|O(\ln\frac{1}{\eta}\cdot\eta^{1.5})\right|} \geq 1-\left|O(\ln\frac{1}{\eta}\cdot\sqrt{\eta})\right|
\end{aligned}$$

Therefore, we complete the proof. □

D.2. Convergence to sparse solution

We assume that all the neurons are perfect aligned, i.e. for every neuron $\boldsymbol{w}_i$ ($i \in [m]$) at step t, there exists coefficient $\gamma_i(t)$ such that $\boldsymbol{w}_i(t) = \gamma_i(t) \cdot \boldsymbol{\theta}^*$.

Lemma D.2 (Convergence)**.** *Suppose Condition 3.1 (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*

$$|\langle \boldsymbol{w}(t_3), \boldsymbol{\theta}^* \rangle| \geq 1 - \left| O(\eta \cdot \ln \frac{1}{\eta}) \right| \tag{58}$$

By equ (49), when $t = t_2$, we have

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) - \eta \cdot a_i(t) \cdot \frac{1}{n} \sum_{j=1}^{m} (\boldsymbol{a}(t)^\top \boldsymbol{W}(t) \boldsymbol{x}_j - \boldsymbol{y}_j) \boldsymbol{x}_j \tag{59}$$

$$= \gamma_i(t) \cdot \boldsymbol{\theta}^* - \eta \cdot a_i(t) \cdot (\frac{1}{n} \sum_{j=1}^{m} \boldsymbol{x}_j \cdot \boldsymbol{x}_j^\top) \cdot (\sum_{k=1}^{m} a_k(t) \cdot \gamma_k(t) \cdot \boldsymbol{\theta}^* - \boldsymbol{\theta}^*) \tag{60}$$

$$= \left\{ \gamma_i(t) - \eta \cdot a_i(t) \cdot (\sum_{k=1}^{m} a_k(t) \cdot \gamma_k(t) - 1) \right\} \cdot \boldsymbol{\theta}^* + O(\frac{\eta}{\sqrt{n}}) \tag{61}$$

The last equation is due to $\frac{1}{n} \sum_{j=1}^{m} \boldsymbol{x}_j \cdot \boldsymbol{x}_j^\top = I + O(\frac{1}{\sqrt{n}}) = I + O(\eta^2)$ by large number law and Condition 3.1(A3). Notice that

$$\boldsymbol{\theta}(t) = \boldsymbol{a}(t)^\top \cdot \boldsymbol{W}(t) = \sum_{i=1}^{m} a_i(t) \cdot \boldsymbol{w}_i(t) = (\sum_{i=1}^{m} a_i(t) \cdot \gamma_i(t)) \cdot \boldsymbol{\theta}^* \tag{62}$$

Then we have

$$\begin{aligned}
\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) &- \eta \cdot \boldsymbol{a}(t)^\top \boldsymbol{a}(t) \cdot \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{a}(t)^\top \boldsymbol{W}(t) \boldsymbol{x}_i - y_i) \cdot \boldsymbol{x}_i^\top \\
&- \eta \cdot \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{a}^\top(t) \boldsymbol{W}(t) \boldsymbol{x}_i - y_i) \cdot \boldsymbol{x}_i^\top \boldsymbol{W}(t)^\top \boldsymbol{W}(t) + O(\eta^2) \\
&= \boldsymbol{\theta}(t) - \eta \cdot \sum_{i=1}^{m} a_i(t)^2 \cdot (\sum_{j=1}^{m} a_i(t) \cdot \gamma_i(t) - 1) \cdot \boldsymbol{\theta}^* - \eta \cdot (\sum_{j=1}^{m} a_i(t) \cdot \gamma_i(t) - 1) \cdot \|\boldsymbol{\theta}^*\|^2 \cdot \boldsymbol{\theta}^* + O(\eta^2) \\
&= \boldsymbol{\theta}(t) - \eta \cdot (\sum_{j=1}^{m} a_i(t) \cdot \gamma_i(t) - 1) \cdot (\sum_{i=1}^{m} a_i(t)^2 + \|\boldsymbol{\theta}^*\|^2) \cdot \boldsymbol{\theta}^* + O(\eta^2)
\end{aligned}$$

We consider the change of $\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*$. Subtracting θ^* on both side and we retain

$$\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}^* = \boldsymbol{\theta}(t) - \boldsymbol{\theta}^* - \eta \cdot (\sum_{j=1}^{m} a_i(t) \cdot \gamma_i(t) - 1) \cdot (\sum_{i=1}^{m} a_i(t)^2 + \|\boldsymbol{\theta}^*\|^2) \cdot \boldsymbol{\theta}^* + O(\eta^2) \tag{63}$$

$$= (\sum_{k=1}^{m} a_k(t) \cdot \gamma_k(t) - 1) \cdot (1 - \eta \cdot (\sum_{i=1}^{m} a_i(t)^2 + \|\boldsymbol{\theta}^*\|^2)) \cdot \boldsymbol{\theta}^* + O(\eta^2) \tag{64}$$

$$= (\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*) \cdot (1 - \eta \cdot (\sum_{i=1}^{m} a_i(t)^2 + \|\boldsymbol{\theta}^*\|^2)) + O(\eta^2) \tag{65}$$

In the beginning, $|a_k(t_2)|, \|\boldsymbol{w}_k(t_2)\| \leq \sqrt{\eta}$, so $\gamma_i(t_2) \leq \frac{\sqrt{\eta}}{\|\boldsymbol{\theta}^*\|}$ and we have

$$\left| \sum_{k=1}^{m} a_k(t_2) \cdot \gamma_k(t_2) \right| \leq m \cdot \frac{\eta}{\|\boldsymbol{\theta}^*\|} \tag{66}$$

by equ (65), for any $t_2 \leq t \leq t_3$ it holds

$$\|\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}^*\| = \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\| \cdot \left| (1 - \eta \cdot (\sum_{i=1}^{m} a_i(t)^2 + \|\boldsymbol{\theta}^*\|^2)) \right| + O(\eta^2) \tag{67}$$

$$\leq \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^*\| \cdot (1 - \eta \cdot \|\boldsymbol{\theta}^*\|^2) + O(\eta^2) \tag{68}$$

Let $T = \frac{-\ln \eta}{\|\boldsymbol{\theta}^*\|^2 \cdot \eta}$, we retain

$$\|\boldsymbol{\theta}(t_3) - \boldsymbol{\theta}^*\| \leq \|\boldsymbol{\theta}(t_2) - \boldsymbol{\theta}^*\| \cdot (1 - \eta \cdot \|\boldsymbol{\theta}^*\|^2)^T + O(T \cdot \eta^2) \tag{69}$$

Since $(1-x)^T \leq \exp(-x \cdot T)$, we have

$$\begin{aligned}
\|\boldsymbol{\theta}(t_3) - \boldsymbol{\theta}^*\| &\leq \|\boldsymbol{\theta}(t_2) - \boldsymbol{\theta}^*\| \cdot (1 - \eta \cdot \|\boldsymbol{\theta}^*\|^2)^T + O(T \cdot \eta^2) \\
&\leq \|\boldsymbol{\theta}(t_2) - \boldsymbol{\theta}^*\| \cdot \exp(-\eta \cdot \|\boldsymbol{\theta}^*\|^2 \cdot T) + O(\eta \cdot \ln \frac{1}{\eta}) \\
&= \|\boldsymbol{\theta}(t_2) - \boldsymbol{\theta}^*\| \cdot \eta + O(\eta \cdot \ln \frac{1}{\eta}) \\
&= \left| O(\eta \cdot \ln \frac{1}{\eta}) \right|
\end{aligned}$$

So we reatin

$$\left| 1 - \sum_{k=1}^{m} a_k(t_3) \cdot \gamma_k(t_3) \right| \leq \left| O(\eta \cdot \ln \frac{1}{\eta}) \right| \tag{70}$$

Since the summation of $\boldsymbol{w}_i(i)$ approach to $\boldsymbol{\theta}^*$, which shows that there exists some neurons with "big" norm. For any neuron $\|\boldsymbol{w}_i(t_3)\| \geq \sqrt{\eta}$, we have $\gamma_i(t_3) \geq \frac{\sqrt{\eta}}{\|\boldsymbol{\theta}^*\|}$. For these neuron, we retain

$$
\begin{aligned}
|\langle \boldsymbol{w}(t_3), \boldsymbol{\theta}^* \rangle| &= \left| \langle (\sum_{k=1}^{m} a_k(t_3) \cdot \gamma_k(t_3)) \cdot \boldsymbol{\theta}^* + O(T \cdot \eta^2), \boldsymbol{\theta}^* \rangle \right| \\
&= \left| \frac{(\sum_{k=1}^{m} a_k(t_3) \cdot \gamma_k(t_3)) \cdot \|\boldsymbol{\theta}^*\|^2 + O(\eta \cdot \ln \frac{1}{\eta})}{\gamma_i(t_3) \cdot \|\boldsymbol{\theta}^*\|^2 + O(\eta \cdot \ln \frac{1}{\eta})} \right| \\
&\geq \frac{\left|\gamma_i(t_3) \cdot \|\boldsymbol{\theta}^*\|^2\right| - \left| O(\eta \cdot \ln \frac{1}{\eta}) \right|}{|\gamma_i(t_3) \cdot \|\boldsymbol{\theta}^*\|^2| + \left| O(\eta \cdot \ln \frac{1}{\eta}) \right|} \quad \text{(by triangle inequality and equ (70))} \\
&= 1 - \left| O(\eta \cdot \ln \frac{1}{\eta}) \right|
\end{aligned}
$$

Therefore, we complete the proof. □

Appendix E. Additional Experiment Result

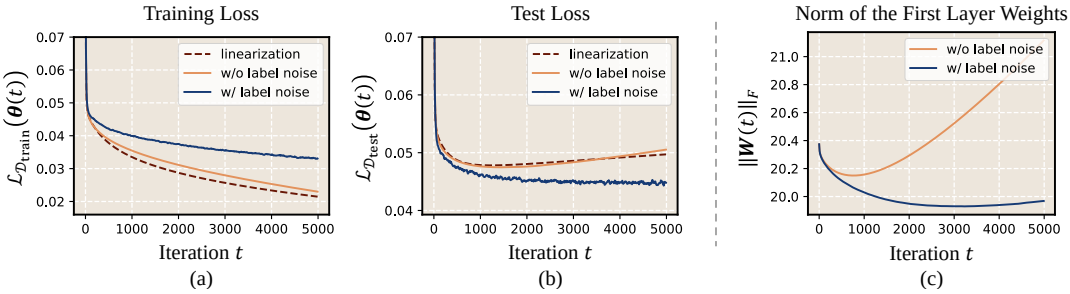

Figure 2: **Label noise SGD induces the rich regime. (a, b).** Training $\mathcal{L}_{\mathcal{D}_{\text{train}}}(\boldsymbol{\theta}(t))$ and test loss $\mathcal{L}_{\mathcal{D}_{\text{test}}}(\boldsymbol{\theta}(t))$ vs. training epochs t. **Label noise SGD induces the progressively diminishing phenomenon. (c).** The first-layer weight norm $\|\boldsymbol{W}(t)\|_F$ vs. training epochs t. We use GD to train the models with NTK parameterization [16], both with and without label noise. We also train a linearized model with GD as baseline. Results are presented for WideResNets trained on a small subset of CIFAR-10.