How Does Label Noise Gradient Descent Improve Generalization in the Low SNR Regime?

Wei Huang[1,2]*, **Andi Han**[3,1]*, **Yujin Song**[4,1], **Yilan Chen**[5], **Denny Wu**[6,7], **Difan Zou**[8], **Taiji Suzuki**[4,1]
[1] RIKEN AIP [2] The Institute of Statistical Mathematics
[3] The University of Sydney [4] The University of Tokyo [5] University of California San Diego
[6] New York University [7] Flatiron Institute [8] The University of Hong Kong
wei.huang.vr@riken.jp; andi.han@sydney.edu.au;
y.song.research@gmail.com; yic031@ucsd.edu; dennywu@nyu.edu;
dzou@hku.hk; taiji@mist.i.u-tokyo.ac.jp

Abstract

The capacity of deep learning models is often large enough to both learn the underlying statistical signal and overfit to noise in the training set. This noise memorization can be harmful especially for data with a low signal-to-noise ratio (SNR), leading to poor generalization. Inspired by prior observations that label noise provides implicit regularization that improves generalization, in this work, we investigate whether introducing label noise to the gradient updates can enhance the test performance of neural network (NN) in the low SNR regime. Specifically, we consider training a two-layer NN with a simple label noise gradient descent (GD) algorithm, in an idealized signal-noise data setting. We prove that adding label noise during training suppresses noise memorization, preventing it from dominating the learning process; consequently, label noise GD enjoys rapid signal growth while the overfitting remains controlled, thereby achieving good generalization despite the low SNR. In contrast, we also show that NN trained with standard GD tends to overfit to noise in the same low SNR setting and establish a non-vanishing lower bound on its test error, thus demonstrating the benefit of introducing label noise in gradient-based training.

1 Introduction

The success of deep learning across various domains [34, 48, 7] is often attributed to their ability to extract features [20, 15] via gradient-based training [14, 3]. One desirable property of gradient-based feature learning is the algorithmic regularization that prioritizes learning of the underlying signal instead of overfitting to noise: real-world data contains noise due to mislabeling, data corruption, or inherent ambiguity, yet despite having the capacity to memorize noise, neural networks (NNs) trained by gradient descent (GD) tend to identify informative features and "low-complexity" solutions that generalize [57, 43].

To understand this behavior, recent theoretical works considered data models that partition the features into signal and noise components [19, 5, 54], and studied the performance of gradient-based training in different signal-noise conditions. Among existing theoretical settings, the signal-noise model proposed in [2, 8] has been extensively studied in the feature learning theory literature. In this model, input features are constructed by combining a label-dependent *signal* with label-independent *noise*. The signal represents meaningful patterns relevant to the predictive task while the noise

[*] Equal Contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

component captures background features unrelated to the learning task. This idealized setting has shed light on how various algorithms, neural network architectures, and other factors influence optimization and generalization of neural networks, depending on the signal-to-noise ratio (SNR) [18, 58, 29, 25, 56, 9, 27, 22, 26, 23, 35].

In such model, it is known that the SNR dictates a transition from *benign overfitting* to *harmful overfitting*. In the high SNR regime, gradient-based feature learning prioritizes signal learning over noise memorization; hence upon convergence, the trained NN recovers the signal and generalizes to unseen data despite some degree of noise memorization, a phenomenon known as benign overfitting [4, 51, 40, 44, 46, 31]. In contrast, when the SNR is low, noise memorization dominates the training dynamics, and the network fails to identify useful features before the training loss becomes small, leading to harmful overfitting [8, 33].

Given these challenges, recent works have explored algorithmic modifications that either enhance signal learning or suppress noise memorization, to improve generalization in the challenging low SNR regime. [25] showed that the smoothing effect of graph convolution in graph neural networks mitigates overfitting to noise; however, this approach requires the graph to be sufficiently dense and exhibits high homophily. [10] found that the sharpness-aware minimization (SAM) method [17] prevents noise memorization in early stages of training, thereby promoting effective feature learning; this being said, SAM has higher computational cost than standard GD due to the two forward and backward passes per step, and it involves more complex hyperparameter tuning. The goal of this work is to address the following question:

Is there a simple modification of GD with no computational overhead that achieves small generalization error in low SNR settings where standard GD fails to generalize?

1.1 Our contributions

We provide an affirmative answer to the question above by introducing **random label noise** to the training dynamics as a form of regularization, inspired by label noise (stochastic) gradient descent (GD) [6, 45, 49]. Specifically, we analyze the classification extension of label noise GD considered in [24, 13], where random label flipping is introduced to prevent overfitting.

Empirically, we first present findings in a controlled classification setting, where we train a VGG-16 model on (a subset of) the CIFAR-10 dataset. To modulate the SNR, we follow [19] and add varying levels of noise to the high-frequency Fourier components of the images – higher noise strength corresponds to lower SNR and vice versa. The results, shown in Figure 1, demonstrate that as the SNR decreases, the performance gap between label noise GD and standard GD becomes more significant, hence suggesting that label noise GD improves generalization in the low SNR regime. The goal of this work is to rigorously establish this separation *in an idealized theoretical model.*

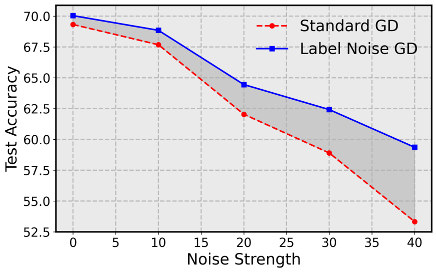


Figure 1: **Test accuracy of VGG-16 on CIFAR-10 with varying SNR**. Label noise GD consistently outperforms standard GD, and the gap increases with the noise strength.

Theoretically, we characterize the properties of Label Noise GD in low SNR regimes, by considering the learning of a two-layer convolutional neural network in a binary classification problem studied in [8], and show that by randomly flipping the labels of a small proportion of training samples at each iteration, noise memorization can be suppressed despite the low SNR, whereas signal learning experiences a period of fast growth. As a result, neural network trained by label noise GD attains good generalization performance in regimes where standard GD fails, as summarized in the following informal theorem:

Theorem 1.1 (Informal)**.** *Given n training samples drawn from the distribution in Definition 2.1 in the low SNR regime where $n^{-1}\mathrm{SNR}^{-2} = \tilde{\Omega}(1)$. Then for any $\epsilon > 0$, after a polynomial number of training steps t (depending on ϵ), with high probability we have: (i)* ***Standard GD*** *minimizes the logistic training loss to $L_S^{(t)} \leq \epsilon$, but the generalization error (0-1 loss) remains large, i.e., $L_{\mathcal{D}}^{(t)} = \Omega(1)$. (ii)* ***Label noise GD*** *cannot reduce the logistic training loss to a small value $L_S^{(t)} = \Omega(1)$, but achieves small generalization error (0-1 loss), i.e., $L_{\mathcal{D}}^{(t)} = o(1)$.*

We make the following remarks on our main results.

- **Improved Generalization due to Label Noise.** The theorem provides an upper bound on the test error of label noise GD and lower bound on the error of standard GD. This demonstrates that incorporating label noise into the gradient descent updates improves generalization in the low SNR regime. We note that our conditions on label noise GD learnability are weaker than those required for SAM as specified in [10], even though our studied algorithm is arguably simpler and more computationally efficient – see Section 3 for more comparisons.
- **Analysis of Feature Learning Dynamics.** We establish the main theorem via a refined characterization of the training dynamics of label noise GD on a two-layer convolutional NN with squared ReLU activation. A key observation in our analysis is that label noise introduces regularization to the noise memorization process, preventing it from growing beyond a constant level; meanwhile, signal learning continues to exhibit a rapid growth rate, allowing the model to identify the informative features and avoid harmful overfitting in low SNR regimes.

2 Problem Setup

In this section, we describe the signal-noise data model, the neural network architecture used for training, and the label noise gradient descent algorithm considered in this work.

Data generating process. We consider the signal-noise data model from [8, 10, 25]. Let $\boldsymbol{\mu} \in \mathbb{R}^d$ be a fixed signal vector, and for each data point $(\mathbf{x}, y)$, the feature $\mathbf{x}$ is composed of two patches, denoted as $\mathbf{x} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\} \in \mathbb{R}^{2d}$. The target variable y is a binary label, taking values in $\{\pm 1\}$. Then the data is generated according to the following process.

Definition 2.1. We consider the following generating process for $(\mathbf{x}, y)$:

1. The true label y is drawn from a Rademacher distribution, i.e., $\mathbb{P}[y = 1] = \mathbb{P}[y = -1] = 1/2$.
2. One of the patches, $\mathbf{x}^{(1)}$ or $\mathbf{x}^{(2)}$ is randomly selected to be $y\boldsymbol{\mu}$ (the signal), while the other is set to be $\boldsymbol{\xi}_i \sim \mathcal{N}(0, \sigma_p^2(\mathbf{I}_d - \boldsymbol{\mu}\boldsymbol{\mu}^\top \|\boldsymbol{\mu}\|_2^{-2}))$ (the noise). Here, σ_p^2 denotes the strength of the noise vector.

We make the following remarks on the data distribution.

- The data model simulates a setting where the input features are composed of both signal and noise components. Specifically, each data point is divided into two patches, and one of these patches contains meaningful information (signal) related to the classification label, while the other patch only contains random noise independent of the label. The noise covariance $\sigma_p^2(\mathbf{I}_d - \boldsymbol{\mu}\boldsymbol{\mu}^\top \|\boldsymbol{\mu}\|_2^{-2})$ is set to ensure that the noise vector is orthogonal to the signal vector for simplicity.
- This setup is designed to reflect real-world scenarios where data contains a mix of relevant and irrelevant features (see Appendix A in [2] for discussions). Note that in high dimensions ($n \ll d$), the NN can achieve small training loss just by overfitting to the noise component. Therefore, the challenge for the learning algorithm in the low SNR regime is to identify and learn the signal patch while ignoring the noisy patch.
- We use the minimum number of patches in the multi-patch model for concise presentation. Our results can be extended to more general cases where the number of patches is greater than 2; see [2, 47] for such extension.

Neural network and loss function. Following [8], we consider a two-layer convolutional neural network with squared ReLU activation and shared filters applied separately to each patch. The network is defined as $f(\mathbf{W}, \mathbf{x}) = F_{+1}(\mathbf{W}_{+1}, \mathbf{x}) - F_{-1}(\mathbf{W}_{-1}, \mathbf{x})$, where

$$F_j(\mathbf{W}_j, \mathbf{x}) = \frac{1}{m}\sum_{r=1}^{m}\sum_{p=1}^{2}\sigma\big(\langle \mathbf{w}_{j,r}, \mathbf{x}^{(p)}\rangle\big) = \frac{1}{m}\sum_{r=1}^{m}\Big(\sigma\big(\langle \mathbf{w}_{j,r}, y\boldsymbol{\mu}\rangle\big) + \sigma\big(\langle \mathbf{w}_{j,r}, \boldsymbol{\xi}_i\rangle\big)\Big),$$

in which m denotes the size of the hidden layer, and $\sigma(z) = (\max\{0, z\})^2$. Note that $j \in \{-1, +1\}$ corresponds to the fixed second-layer. The symbol $\mathbf{W}_j$ represents the collection of weight vectors in

Algorithm 1 Label noise gradient descent

```
Initialize $\mathbf{W}_0$, step size $\eta$, flipping probability $p \in [0,1]$
for $t = 0, ..., T-1$ do
    Sample $\epsilon_i^{(t)} \sim \mathrm{Rademacher}(1-p, p), \forall i \in [n]$.
    $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \nabla_{\mathbf{W}} L_S^{\epsilon}(\mathbf{W}^{(t)})$, where $L_S^{\epsilon}(\mathbf{W}^{(t)}) = \frac{1}{n}\sum_{i\in[n]} \ell\big(\epsilon_i^{(t)} y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big)$.
end for
```

the first layer, i.e., $\mathbf{W}_j = [\mathbf{w}_{j,1}, \mathbf{w}_{j,2}, \ldots, \mathbf{w}_{j,m}] \in \mathbb{R}^{d\times m}$, where $\mathbf{w}_{j,r} \in \mathbb{R}^d$ is the weight vector of the r-th neuron. Here, $j \in \{-1, +1\}$ indicates the fixed value in the second layer. The initial weights $\mathbf{W}_{\pm 1}$ has entries sampled from $\mathcal{N}(0, \sigma_0^2)$.

Remark 2.2. Since we do not optimize the 2nd-layer parameters, we expect the 2-homogeneous squared ReLU activation to mimic the behavior of training both layers simultaneously in a ReLU network; such higher-order homogeneity amplifies feature learning (e.g., see [12, 21]) and creates a significant gap between signal learning and noise memorization. Similar effect can be achieved by smoothed ReLU with local polynomial growth as in [2, 47].

We use the logistic loss computed over n training samples, denoted as $S = \{(\mathbf{x}_i, y_i)\}_{i\in[n]}$:

$$L_S(\mathbf{W}) = \frac{1}{n}\sum_{i\in[n]} \ell(y_i f(\mathbf{W}, \mathbf{x}_i)),$$

where $\ell(z) = \log(1+\exp(-z))$. To evaluate the generalization performance of the trained network, we measure its expected 0-1 loss on unseen data, defined as

$$L_{\mathcal{D}}^{0-1}(\mathbf{W}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}(y \neq \mathrm{sign}(f(\mathbf{W}, \mathbf{x}))], \tag{1}$$

where $\mathcal{D}$ denotes the data distribution specified in Definition 2.1, and $\mathbb{1}(\cdot)$ is the indicator function.

Label noise GD for binary classification. We train the above neural network by gradient descent on either (i) the original loss function (standard GD), or (ii) the loss function with label-flipping noise defined as

$$L_S^{\epsilon}(\mathbf{W}^{(t)}) \triangleq \frac{1}{n}\sum_{i\in[n]} \ell\big(\epsilon_i^{(t)} y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i)\big).$$

Here, $\epsilon_i^{(t)}$ is a random variable equal to 1 with probability $1-p$ and -1 with p, i.e., $\epsilon_i^{(t)} \sim \mathrm{Rademacher}(1-p, p)$. In other words, labels flip with probability p independently at each step.

Remark 2.3. We remark that label smoothing [45, 49] and label flipping are equivalent in expectation. This connection has also been discussed in [38]. However, note that this equivalence in expectation does not imply closeness in training dynamics due to the stochasticity introduced by the label-flipping.

The label noise GD update is then given as follows:

$$\mathbf{w}_{j,r}^{(t+1)} = \mathbf{w}_{j,r}^{(t)} - \frac{\eta}{nm}\sum_{i=1}^{n} \widetilde{\ell}_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \boldsymbol{\mu}\rangle)\epsilon_i^{(t)} j \boldsymbol{\mu} - \frac{\eta}{nm}\sum_{i=1}^{n} \widetilde{\ell}_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\epsilon_i^{(t)} y_i j \boldsymbol{\xi}_i, \tag{2}$$

where η is the learning rate, and we defined $\widetilde{\ell}_i'^{(t)} = \ell'(\epsilon_i^{(t)} y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i))$ as the derivative of the loss function. This *label noise GD* training procedure is outlined in Algorithm 1. Observe that the proposed algorithm is *computationally efficient*, as the introduced label noise does not modify the original gradient descent framework. Hence this method is simple to implement, does not add significant computational overhead, and requires no complex hyperparameter tuning.

3 Main results

In this section, we quantify the benefits of label noise gradient descent by comparing its generalization performance against standard gradient descent (GD) training without label noise. We begin by outlining the assumptions that apply to both label noise GD and standard GD.

Assumption 3.1. Define $\mathrm{SNR} = \frac{\|\boldsymbol{\mu}\|_2}{\sigma_p \sqrt{d}}$. We consider the following setting for both algorithms:

(i) data dimension $d = \tilde{\Omega}(\max\{n^2, n\|\boldsymbol{\mu}\|_2^2/\sigma_p^2\})$; signal-to-noise ratio $\mathrm{SNR} = \tilde{O}(1/\sqrt{n})$.

(ii) network width $m = \tilde{\Omega}(1)$; number of training samples $n = \tilde{\Omega}(1)$.

(iii) learning rate $\eta \leq \tilde{O}(\sigma_p^{-2} d^{-1})$.

(iv) initialization variance $\tilde{O}(n\sigma_p^{-1} d^{-3/4}) \leq \sigma_0 \leq \tilde{O}(\min\{\|\boldsymbol{\mu}\|_2^{-1} d^{-5/8}, \sigma_p^{-1} d^{-1/2}\})$.

(v) flipping rate of label noise p lies in the interval $p \in (\frac{C \log d}{\sqrt{mn}}, \frac{1}{C})$, where C is a sufficient large constant.

We make the following remarks on the above assumption.

- The high-dimensional assumption (i) is standard in the benign overfitting analysis of NNs (e.g., see [8, 18]). The low SNR condition is derived from the comparison between the magnitude of signal learning and noise memorization – see Section 4.1; similar conditions has been established in [8, 33] for different activations.
- The requirements on the hidden layer size m and the sample size n being at least polylogarithmic in the dimension d ensure that certain statistical properties regarding weight initialization and the training data hold with high probability at least $1 - 1/d$.
- The upper bound on the learning rate η ensures that the iterates in (4-6) remain bounded, which is required for standard GD to reach low training loss; see Proposition 4.1.
- The upper bound on initialization scale σ_0 is used to ensure convergence of GD, and the lower bound is used for anti-concentration at initialization. Similar requirements can be found in [8, Condition 4.2].
- The lower bound ensures that the number of flipped samples concentrates around its expectation so that our theoretical analysis remains valid, while the upper bound on label flipping rate p prevents the label noise from dominating the true signal.

We first state the negative result for standard gradient descent (GD) without label noise.

Theorem 3.2 (GD fails to generalize under low SNR)**.** *Under Assumption 3.1, for any* $\epsilon > 0$*, there exists* $t = \Theta(\frac{nm \log(1/(\sigma_0 \sigma_p \sqrt{d}))}{\eta \sigma_p^2 d} + \frac{m^3 n}{\eta \epsilon \sigma_p^2 d})$*, such that with probability at least* $1 - d^{-1/4}$*, it holds that*

- *The training error converges, i.e.,* $L_S(\mathbf{W}^{(t)}) \leq \epsilon$*.*
- *The test error is large, i.e.,* $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \geq 0.24$*.*

Theorem 3.2 indicates that even though standard GD can minimize the training error to an arbitrarily small value, the generalization performance remains poor. This is mainly because the neural network overfits to the noise components in the input data instead of learning the useful features. Next, we present the positive result for label noise GD.

Theorem 3.3 (Label Noise GD generalizes under low SNR)**.** *Under Assumption 3.1, there exists* $t = \Theta(\frac{nm \log(1/(\sigma_0 \sigma_p \sqrt{d}))}{\eta \sigma_p^2 d} + \frac{m \log(6/(\sigma_0 \|\boldsymbol{\mu}\|_2))}{\eta \|\boldsymbol{\mu}\|_2^2})$ *and constants* $C > 0$*, such that with probability at least* $1 - d^{-1/4}$*, it holds that*

- *The training error is at constant order, i.e.,* $L_S(\mathbf{W}^{(t)}) = \Theta(1)$*.*
- *The test error is small, i.e.,* $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \leq 2 \exp\left(-\frac{Cd}{n^2}\right)$*.*

Theorem 3.3 shows that label noise GD achieves vanishing generalization error when the input dimensionality is large (i.e., $d = \Omega(n^2)$) despite the low SNR.

Remark 3.4. Theorems 3.3 and 3.2 present contrasting outcomes for standard GD and label noise GD in the low SNR regime. In particular,

- Standard GD minimizes the training error effectively but does so by primarily overfitting to noise in the training data. This significant noise memorization leads to harmful overfitting.

- In contrast, label noise GD introduces a regularization effect through label noise, which prevents the network from fully memorizing the noise components. This allows the network to focus on learning the true signal, resulting in a phase of accelerated signal learning. Consequently, the model generalizes even though the training loss does not vanish (due to noise injection).

Comparison with sharpness-aware minimization [10]. We briefly discuss the differences between our findings and those in [10] for the sharpness-aware minimization (SAM) method, where the authors established conditions on the SNR under which SAM can generalize better than stochastic gradient descent (SGD). However, their analysis requires the additional condition that the signal norm satisfies $\|\boldsymbol{\mu}\|_2 \geq \tilde{\Omega}(1)$, indicating the necessity of a sufficiently strong signal. In contrast, we show that label noise GD enjoys good generalization without this strong signal condition. This highlights the robustness of label noise GD in low SNR regimes (even when the signal strength is considerably weaker compared to the noise).

Comparison with stopping times across theorems. We compare the stopping times in Theorems 3.2 and 3.3. The stopping times for standard GD and label noise GD are not directly comparable, as they correspond to different evaluation criteria. Specifically, the stopping time for standard GD is the number of iterations required for the training loss to converge below a threshold ϵ, whereas the stopping time for label noise GD is defined as the number of iterations needed to achieve sufficiently low 0-1 test loss. To enable a meaningful comparison, we derive the ratio between the two stopping times under Assumption 3.1. By setting $m^2 = \log\left(\frac{6}{\sigma_0\|\boldsymbol{\mu}\|_2}\right)/\epsilon$, we obtain$\frac{t_{\text{Standard GD}}}{t_{\text{label noise GD}}} = \Theta\left(\frac{n\|\boldsymbol{\mu}\|_2^2}{\sigma_p^2 d}\right) = \Theta(n\mathrm{SNR}^2)$. According to Assumption 3.1, we assume $n\mathrm{SNR}^2 \ll 1$, which implies that label noise GD requires more iterations to achieve good test performance compared to the time required for the training loss of standard GD to converge.

4 Proof Sketch

In this section, we give an overview of of our analysis of the optimization dynamics of standard GD and label noise GD . Our key technical contributions are summarized as follows: (i) **Boundary characterization in low SNR regimes.** Unlike previous studies [8, 33, 10] that focus on the higher polynomial or standard ReLU activation, we analyze the 2-homogeneous squared ReLU activation, leading to a different boundary characterization of the low SNR regime for standard GD – see Section 4.2. (ii) **Upper bound via supermartingale.** We apply supermartingale arguments with Azuma's inequality to bound noise memorization in label noise GD. This yields high-probability guarantees on training dynamics, previously unestablished in this context.

4.1 Signal-noise decomposition

To analyze the training dynamics, we adopt a parameter decomposition technique from [8, 33]: there exist $\{\gamma_{j,r}^{(t)}\}$ and $\{\rho_{j,r,i}^{(t)}\}$ such that

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j\gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\boldsymbol{\mu} + \sum_{i=1}^{n}\rho_{j,r,i}^{(t)}\|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i. \tag{3}$$

This decomposition originates from the observation that the gradient descent update always evolves in the direction of $\boldsymbol{\mu}$ and $\mathbf{x}_i$ for $i \in [n]$. In particular, $\gamma_{j,r}^{(t)} \approx \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu}\rangle$ serves as the *signal learning* coefficient, whereas $\rho_{j,r,i}^{(t)} \approx \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle$ characterizes the *noise memorization* during training. Next we let $\overline{\rho}_{j,r,i}^{(t)} = \rho_{j,r,i}^{(t)}\mathbb{1}(y_i = j)$ and $\underline{\rho}_{j,r,i}^{(t)} = \rho_{j,r,i}^{(t)}\mathbb{1}(y_i = -j)$. Combined with the gradient descent update given by Equation (2), we obtain the iteration rules for these coefficients:

$$\gamma_{j,r}^{(t+1)} = \gamma_{j,r}^{(t)} - \frac{\eta}{nm}\sum_{i=1}^{n}\tilde{\ell}_i'^{(t)}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i\boldsymbol{\mu}\rangle)\|\boldsymbol{\mu}\|_2^2\epsilon_i^{(t)}, \tag{4}$$

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm}\tilde{\ell}_i'^{(t)}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2\epsilon_i^{(t)}\mathbb{1}(y_i = j), \tag{5}$$

$$\underline{\rho}_{j,r,i}^{(t+1)} = \underline{\rho}_{j,r,i}^{(t)} + \frac{\eta}{nm}\tilde{\ell}_i'^{(t)}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2\epsilon_i^{(t)}\mathbb{1}(y_i = -j). \tag{6}$$

where the initial values of the coefficients are given by $\gamma_{j,r}^{(0)} = 0$ and $\rho_{j,r,i}^{(0)} = 0$ for all $i \in [n]$, $j \in \{-1, 1\}$ and $r \in [m]$.

To analyze the optimization trajectory, we track the dynamics of signal learning coefficients ($\gamma_{j,r}^{(t)}$) and noise memorization coefficients ($\rho_{j,r,i}^{(t)}$) using the iteration rules in Equations (4-6). To facilitate a detailed analysis, we first provide upper bounds on the absolute value of both the signal learning and noise memorization coefficients throughout the entire training process.

Proposition 4.1. *Given Assumption 3.1 and* $\epsilon > 0$. *Let* $\beta = 2\max_{j,r,i}\{|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|, |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle|\}$ *and* $\alpha = 4\log(T^*)$. *For* $0 \le t \le T^*$, *where* $T^* = \eta^{-1}\mathrm{poly}(n, m, d, \|\boldsymbol{\mu}\|_2^{-1}, (\sigma_p^2 d)^{-1}, \sigma_0^{-1}, \epsilon^{-1})$, *for all* $i \in [n]$, $r \in [m]$ *and* $j \in \{-1, 1\}$, *it holds that*

$$0 \le \gamma_{j,r}^{(t)} \le \alpha, \quad 0 \le \overline{\rho}_{j,r,i}^{(t)} \le \alpha, \tag{7}$$

$$0 \ge \underline{\rho}_{j,r,i}^{(t)} \ge -\beta - 16\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \ge -\alpha. \tag{8}$$

The proof is in Appendix C. Proposition 4.1 shows that throughout training, the absolute values of signal learning and noise memorization coefficients have a logarithmic upper bound. This result is key for a stage-wise characterization of training dynamics. Notably, this bound holds for both standard GD and label noise GD.

4.2 Proof Sketch for Theorem 3.2

We first establish the negative result for standard GD based on a two-stage analysis. As previously mentioned, we consider the 2-homogeneous $\sigma(z) = \mathrm{ReLU}^2(z)$ which differs from [8, 33, 10]. This leads to a key difference in the boundary characterization of the low SNR regime.

First stage. Notice that starting from small initialization, the loss derivative remains close to a constant. Based on this observation, we establish the difference in magnitude between the coefficients of signal learning and noise memorization.

According to the update rule for the signal learning coefficient given by Equation (4) and by setting $\epsilon_i^{(t)} = 1$ for all t and $i \in [n]$ (i.e., no label flipping), the upper bound of signal learning can be achieved as $\gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \le \exp\big(\frac{2\eta\|\boldsymbol{\mu}\|_2^2}{m}t\big)|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|$. Meanwhile, the bounds for the noise memorization coefficients can be derived from the update rules (5) and (6). The results are given as $\max_{j,r} |\underline{\rho}_{j,r,i}^{(t)}| \le \frac{3\eta\sigma_p^2 td}{nm}\sqrt{\log(8mn/\delta)}\sigma_0\sigma_p\sqrt{d}$, and $\max_{j,r} \overline{\rho}_{j,r,i}^{(t)} \ge \exp\big(\frac{\eta C_1\sigma_p^2 d}{2nm}t\big)\sigma_0\sigma_p\sqrt{d}/4 - 0.6\overline{\beta}$, for all $i \in [n]$, where we define $\bar{\beta} = \min_{i\in[n]}\max_{r\in[m]}\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle$, and use $|\tilde{\ell}_i'^{(t)}| \ge C_1$. In the low SNR setting, where $\sigma_p\sqrt{d}$ is much larger than $\|\boldsymbol{\mu}\|_2$, we observe that noise memorization dominates the feature learning process during the first stage, as shown in the following lemma.

Lemma 4.2. *Under the same condition as Theorem 3.2, and let* $T_1 = \Theta(\frac{nm\log(1/(\sigma_0\sigma_p\sqrt{d}))}{\eta\sigma_p^2 d})$, *the following results hold with high probability at least* $1 - d^{-1}$: *(i)* $\max_{j,r} \overline{\rho}_{j,r,i}^{(T_1)} \ge 1$, *for all* $i \in [n]$; *(ii)* $\max_{j,r,i} |\underline{\rho}_{j,r}^{(t)}| \le \tilde{O}(\sigma_0\sigma_p\sqrt{d})$, *for all* $t \in [T_1]$; *(iii)* $\max_{j,r} \gamma_{j,r}^{(t)} \le \tilde{O}(\sigma_0\|\boldsymbol{\mu}\|_2)$, *for all* $t \in [T_1]$.

Lemma 4.2 indicates that when the SNR is sufficiently low, i.e., $\mathrm{SNR} = \tilde{O}(1/\sqrt{n})$, noise memorization dominates the training dynamics during the early phase of standard GD optimization. We highlight that this "low-SNR" condition differs from that of [8, 33] due to the choice of activation function. In particular, [8] assumed $\sigma(z) = (\max\{0, z\})^q$ with $q > 2$ and established a low-SNR boundary $n^{-1}\mathrm{SNR}^{-q} = \tilde{\Omega}(1)$, whereas [33] considered the ReLU activation and derived the condition $n\frac{\|\boldsymbol{\mu}\|_2^4}{\sigma_p^4 d} \le O(1)$.

Second stage. After the first stage, the loss derivative is no longer bounded by a constant value. To prove convergence of the training loss $L(t) \le \epsilon$, we build upon the analysis from the first stage and define $\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^{(0)} + 2m\log(2/\epsilon)\sum_{i=1}^n \|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i$. We show that, as gradient descent progresses, the distance between $\mathbf{W}^{(t)}$ and $\mathbf{W}^*$ decreases until $L(t) \leq \epsilon$: $\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2 \leq \eta L_S(t) - \eta\epsilon$. Moreover, we show that the difference between signal learning and noise memorization still holds in the second stage, as summarized below.

Lemma 4.3. *Let $T_2 = \eta^{-1}\sigma_p^{-2}d^{-1}nm\log(1/(\sigma_0\sigma_p d)) + \eta^{-1}\epsilon^{-1}m^3n\sigma_p^{-2}d^{-1}$. Under the same assumptions as Theorem 3.2, for training step $t \in [T_1, T_2]$, it holds that $\gamma_{j,r}^{(t)} \leq \tilde{O}(\sigma_0\|\boldsymbol{\mu}\|_2)$, $|\underline{\rho}_{j,r,i}^{(t)}| \leq \tilde{O}(\sigma_0\sigma_p\sqrt{d})$, and $\overline{\rho}_{j,r,i}^{(t)} \geq 1$. Besides, there exists a step $t \in [T_1, T_2]$, such that $L_S(t) < \epsilon$.*

Lemma 4.3 shows that standard GD achieves low training error after polynomially many steps, and noise memorization dominates the entire training process, which results in harmful overfitting.

4.3 Proof Sketch for Theorem 3.3

We also divide the training dynamics of label noise GD into two phases. In the first phase, both signal learning and noise memorization increase exponentially despite the presence of random label noise. In the second phase, label noise suppresses the growth of noise memorization, causing it to oscillate within a constant range; meanwhile, signal learning continues to grow exponentially until stabilizing at constant value, which leads to beneficial feature learning and low generalization error.

First stage. Leveraging the fact that the derivative of the loss function remains within a constant range due to small initialization, we demonstrate that both signal learning and noise memorization exhibit exponential growth rates, even in the presence of label noise. According to the iterative update of the signal learning coefficient in Equation (4), the upper and lower bounds are given as $\gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \leq \exp\big(\frac{2\eta\|\boldsymbol{\mu}\|_2^2}{m}t\big)|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|$, and $\max_{r\in[m]}\{\gamma_{j,r}^{(t)} + j\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle\} \geq \exp(\frac{C_0\eta\|\boldsymbol{\mu}\|_2^2}{8m})\big(\max_{r\in[m]}\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle\big)$, respectively. Here C_0 is the lower bound for the absolute loss derivative. These bounds indicate that signal learning grows exponentially with the number of training iterations. On the other hand, from the update equation (5), we characterize the behavior of noise memorization. Despite the injected label noise, we can show a lower bound on the noise memorization rate: $\max_{j,r}\{\overline{\rho}_{j,r,i}^{(t)} + 0.6|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle|\} \geq \exp(\frac{\eta C_0\sigma_p^2 d}{2nm})|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle|$. The main results for the first stage are summarized in the following lemma.

Lemma 4.4. *Under the same condition as Theorem 3.3, and let $T_1 = \Theta(\frac{nm\log((1/\sigma_0\sigma_p d))}{\eta\sigma_p^2 d})$. Then the following holds with probability at least $1 - d^{-1}$: (i) $\max_{j,r}\overline{\rho}_{j,r,i}^{(T_1)} \geq 0.1$, for all $i \in [n]$; (ii) $\max_{j,r,i}|\underline{\rho}_{j,r}^{(t)}| \leq \tilde{O}(\sigma_0\sigma_p\sqrt{d})$, for all $t \in [T_1]$. (iii) $\max_{j,r}\gamma_{j,r}^{(t)} \geq \tilde{O}(\sigma_0\|\boldsymbol{\mu}\|_2)$, for all $t \in [T_1]$.*

Lemma 4.4 states that both signal learning and noise memorization grow exponentially during the first stage.

In order to guarantee that the number of flipped labels remains within its expected range with high probability, we require $p = \tilde{\Omega}(1/\sqrt{t})$, where t is the number of training steps. We set $t = \tilde{\Theta}(n/(\eta\sigma_p^2 d))$ according to Lemma 4.4. Together with the upper bound on η from Assumption 3.1, we obtain $p = \tilde{\Omega}(1/\sqrt{mn})$.

For the analysis of label noise GD, one additional technical challenge is the instability of training dynamics caused by the injected noise, which we address as follows. For signal learning, we make use of the small label flipping rate p and aggregate information across all samples via concentration. Whereas for noise memorization (which is tied to individual samples), we leverage the broad range of time steps in the first stage to establish the overall increment rate.

Second stage. As shown in Lemma 4.4, at the end of the first phase, noise memorization has reached a significant level, dominating the model's output. However, label noise introduces randomness in the labels, which affects the updates of noise memorization coefficients. We track the evolution of $\overline{\rho}_{j,r,i}^{(t)}$ via the following approximation. Define $\iota_i^{(t)} \triangleq \frac{1}{m}\sum_{r=1}^m(\overline{\rho}_{y_i,r,i}^{(t)})^2$. The evolution of noise memorization under label noise GD can be approximated as

$$\iota_i^{(t+1)} \approx \begin{cases} (1 + \frac{\eta\sigma_p^2 d}{(1+\exp((\iota_i^{(t)})^2))nm})^2\iota_i^{(t)}, & \text{with prob } 1 - p. \\ (1 - \frac{\eta\sigma_p^2 d}{(1+\exp(-(\iota_i^{(t)})^2))nm})^2\iota_i^{(t)}, & \text{with prob } p. \end{cases}$$

Unlike conventional approaches such as [8, 33], we analyze this process using a supermartingale argument and apply Azuma's inequality with a union bound over the second-stage training period. Via a martingale argument, we show that noise memorization remains at a constant level with high probability. While noise memorization stabilizes, signal learning continues to grow exponentially. This discrepancy enables signal learning to eventually dominate the generalization. The analysis of the second stage is summarized by the following lemma.

Lemma 4.5. *Under the same condition as Theorem 3.3, during* $t \in [T_1, T_2]$ *with* $T_2 = T_1 + \log(6/(\sigma_0\|\boldsymbol{\mu}\|_2))4m(1+\exp(c_2))\eta^{-1}\|\boldsymbol{\mu}\|_2^{-2}$*, there exist a sufficient large positive constant* C_ι *and a constant* ι_i^* *depending on sample index* i *such that the following results hold with probability at least* $1 - 1/d$*: (i)* $|\iota_i^{(t)} - \iota_i^*| \le C_\iota$*; (ii)* $\gamma_{j,r}^{(t)} \le 0.1$ *for all* $j \in \{-1, 1\}$ *and* $r \in [m]$ *(iii)*$\frac{1}{2m}(\sum_{r=1}^m \overline{\rho}_{y_i,r,i}^{(t)})^2 \le f_i^{(t)} \le \frac{2}{m}(\sum_{r=1}^m \overline{\rho}_{y_i,r,i}^{(t)})^2$ *and (iv)* $\max_{r\in[m]}(\gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|) \ge \exp\big(\frac{\eta\|\boldsymbol{\mu}\|_2^2}{16m}(t - T_1)\big)\max_{r\in[m]}|\gamma_{j,r}^{(T_1)} + \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|$.

Lemma 4.5 demonstrates that label noise introduces a regularizing effect preventing the noise memorization coefficients from growing unchecked, while simultaneously allowing signal learning to grow to a sufficiently large value. Building on this result, we show that both signal learning and noise memorization reach a constant order of magnitude. Consequently, the population loss can be bounded by $L_{\mathcal{D}}(\mathbf{W}^{(t)}) \le 2\exp\big(-\frac{Cd}{n^2}\big)$, corresponding to the second bullet point of Theorem 3.3.

5 Synthetic Experiments

We conduct experiments using synthetic data to validate our theoretical results. The samples are generated according to Definition 2.1. The train and test sample size is $n = 200$ and $n_{\text{test}} = 2000$, and the input dimension is set to $d = 2000$. The label noise flip rate is $p = 0.1$. We train the two-layer network with squared ReLU activation using standard GD and label noise GD for $t = 2000$ steps. The network width is $m = 20$ and the learning rate is $\eta = 0.5$. The signal vector is defined as $\boldsymbol{\mu} = [2, 0, 0, \ldots, 0] \in \mathbb{R}^d$ and the noise variance is set to $\sigma_p^2 = 0.25$.

Dynamics of signal and noise coefficients. In Figure 2, we present the feature learning coefficients defined in Section 4.1, the training loss and test accuracy for both algorithms. We observe that GD successfully minimizes the training loss to a near-zero value; however, noise memorization (ρ) significantly exceeds signal learning (γ), leading to poor test performance. In contrast, label noise GD does not fully minimize the training loss, as it oscillates around 0.5; consistent with our theoretical analysis, this behavior causes noise memorization to remain constant in the second stage, while signal learning continues to grow rapidly. Hence the test accuracy of label noise GD steadily improves.

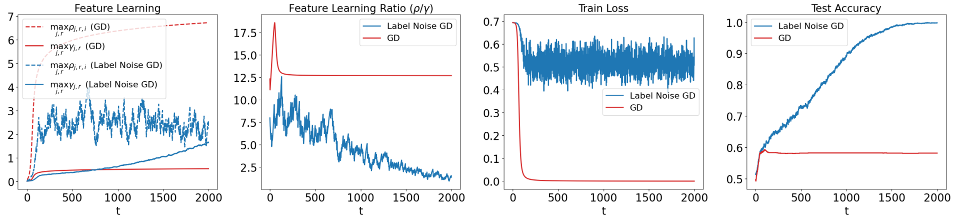
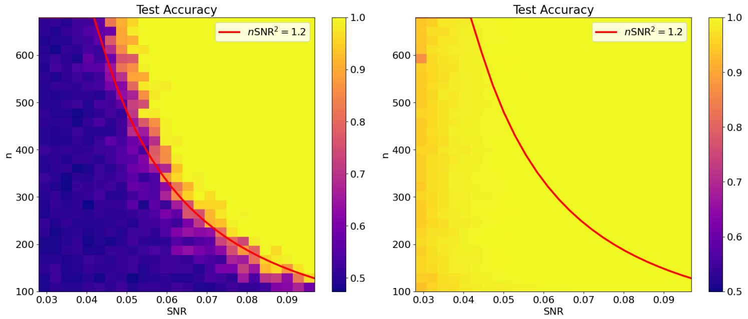


Figure 2: Ratio of noise memorization over signal learning, training loss, and test accuracy, of standard GD and label noise GD. See Section 4.1 for definitions of signal learning (γ) and noise memorization (ρ).

Heatmap of generalization error. Next we explore a range of SNR values from 0.03 to 0.10 and sample sizes n ranging from 100 to 700. For each combination of SNR and sample size n, we train the NN for 1000 steps with $\eta = 1.0$ using standard GD or label noise GD. The resulting test error is visualized in Figure 3. Observe that standard GD (left) fails to generalize when $\mathrm{SNR} = O(n^{-1/2})$, which is consistent with our theoretical prediction in Theorem 3.2. On the other hand, label noise GD (right) achieves perfect test accuracy across a broader range of SNR, which agrees with Theorem 3.3.

Additional Experiments and Extended Analysis. In addition to our primary experiments, we extend our analysis in Appendix G by evaluating Label Noise GD on deeper neural networks, modified MNIST and CIFAR datasets, different types of label noise (e.g., Gaussian), and higher-order ReLU activation functions, demonstrating its robustness across various settings.



(a) Standard GD

(b) Label Noise GD

Figure 3: Test accuracy heatmap of Standard GD (left) and Label Noise GD (right) after training.

6 Conclusion and limitation

We presented a theoretical analysis of gradient-based feature learning in the challenging low SNR regime. Our main contribution is to demonstrate that label noise gradient descent (GD) can effectively enhance signal learning while suppressing noise memorization; this implicit regularization mechanism enables label noise GD to generalize in low SNR settings where standard GD suffers from harmful overfitting. Our theoretical findings are supported by experiments on synthetic data.

Limitations and Broader Impacts. Our current theoretical analysis is limited to a specific choice of activation function (squared ReLU) and network architecture (two-layer convolutional neural network). Extending this theoretical framework to more complex architectures, such as deeper or residual networks, would be a promising direction for future research. Additionally, investigating label noise GD under other optimization algorithms, including stochastic gradient descent (SGD) and adaptive methods like Adam, could provide further insight into its implicit regularization effects in practical settings. This work aims to advance the theoretical understanding of generalization in neural networks. We are not aware of any immediate negative societal impacts resulting from this research.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. WH was supported by JSPS KAKENHI (24K20848) and JST BOOST (JPMJBY24G6). TS was partially supported by JSPS KAKENHI (24K02905) and JST CREST (JPMJCR2015). DZ is supported by NSFC 62306252 and central fund from HKU IDS.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 977–988. IEEE, 2022.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *The Eleventh International Conference on Learning Representations*, 2023.

[3] 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. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

[4] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[5] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35:25349–25362, 2022.

[6] 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.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in Neural Information Processing Systems*, 35:25237–25250, 2022.

[9] Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding mixture of experts in deep learning. *Advances in Neural Information Processing Systems*, 2022.

[10] Zixiang Chen, Junkai Zhang, Yiwen Kou, Xiangning Chen, Cho-Jui Hsieh, and Quanquan Gu. Why does sharpness-aware minimization generalize better than SGD? *Advances in neural information processing systems*, 2023.

[11] Muthu Chidambaram, Xiang Wang, Chenwei Wu, and Rong Ge. Provably learning diverse features in multi-view data with midpoint mixup. In *International Conference on Machine Learning*, pages 5563–5599. PMLR, 2023.

[12] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on learning theory*, pages 1305–1338. PMLR, 2020.

[13] Alex Damian, Tengyu Ma, and Jason D Lee. Label noise SGD provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

[14] Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[16] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

[17] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *International Conference on Learning Representations*, 2021.

[18] Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pages 2668–2703. PMLR, 2022.

[19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33:14820–14830, 2020.

[20] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[21] Margalit Glasgow. SGD finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the XOR problem. *The Twelfth International Conference on Learning Representations*, 2024.

[22] Andi Han, Wei Huang, Yuan Cao, and Difan Zou. On the feature learning in diffusion models. *The Thirteenth International Conference on Learning Representations*, 2025.

[23] Andi Han, Wei Huang, Zhanpeng Zhou, Gang Niu, Wuyang Chen, Junchi Yan, Akiko Takeda, and Taiji Suzuki. On the role of label noise in the feature learning process. In *Forty-second International Conference on Machine Learning*, 2025.

[24] Jeff Z HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357. PMLR, 2021.

[25] Wei Huang, Yuan Cao, Haonan Wang, Xin Cao, and Taiji Suzuki. Quantifying the optimization and generalization advantages of graph neural networks over multilayer perceptrons. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[26] Wei Huang, Andi Han, Yongqiang Chen, Yuan Cao, Zhiqiang Xu, and Taiji Suzuki. On the comparison between multi-modal and single-modal contrastive learning. *Advances in Neural Information Processing Systems*, 37:81549–81605, 2024.

[27] Wei Huang, Ye Shi, Zhongyi Cai, and Taiji Suzuki. Understanding convergence and generalization in federated learning through feature learning theory. In *The Twelfth International Conference on Learning Representations*, 2023.

[28] 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.

[29] Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pages 9965–10040. PMLR, 2022.

[30] Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

[31] Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization. *Advances in Neural Information Processing Systems*, 37:135464–135625, 2024.

[32] Yiwen Kou, Zixiang Chen, Yuan Cao, and Quanquan Gu. How does semi-supervised learing with pseudo-labelers work? a case study. In *International Conference on Learning Representations*, 2023.

[33] Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer relu convolutional neural networks. In *International Conference on Machine Learning*, pages 17615–17659. PMLR, 2023.

[34] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[35] Bingrui Li, Wei Huang, Andi Han, Zhanpeng Zhou, Taiji Suzuki, Jun Zhu, and Jianfei Chen. On the optimization and generalization of two-layer transformers with sign gradient descent. In *The Thirteenth International Conference on Learning Representations*, 2025.

[36] Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. *The Eleventh International Conference on Learning Representations*, 2023.

[37] Hongkang Li, Meng Wang, Tengfei Ma, Sijia Liu, Zaixi Zhang, and Pin-Yu Chen. What improves the generalization of graph transformers? a theoretical dive into the self-attention and positional encoding. *The Forty-First International Conference on Machine Learning*, 2024.

[38] Weizhi Li, Gautam Dasarathy, and Visar Berisha. Regularization via structural label smoothing. In *International Conference on Artificial Intelligence and Statistics*, pages 1453–1463. PMLR, 2020.

[39] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*, 2021.

[40] Zhu Li, Weijie J Su, and Dino Sejdinovic. Benign overfitting and noisy features. *Journal of the American Statistical Association*, 118(544):2876–2888, 2023.

[41] Miao Lu, Beining Wu, Xiaodong Yang, and Difan Zou. Benign oscillation of stochastic gradient descent with large learning rates. *The Twelfth International Conference on Learning Representations*, 2024.

[42] Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, pages 26724–26768. PMLR, 2023.

[43] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

[44] Amartya Sanyal, Puneet K Dokania, Varun Kanade, and Philip HS Torr. How benign is benign overfitting? *International Conference on Learning Representations*, 2021.

[45] 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.

[46] Ohad Shamir. The implicit bias of benign overfitting. *Journal of Machine Learning Research*, 24(113):1–40, 2023.

[47] Ruoqi Shen, Sébastien Bubeck, and Suriya Gunasekar. Data augmentation as feature manipulation. In *International conference on machine learning*, pages 19773–19808. PMLR, 2022.

[48] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[49] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

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

[51] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

[52] Roman Vershynin. *High-Dimensional Probability: An Introduction With Applications in Data Science*. Cambridge, UK: Cambridge Univ. Press, 2018.

[53] Loucas Pillaud Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In *Conference on Learning Theory*, pages 2127–2159. PMLR, 2022.

[54] Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4891–4957. PMLR, 2024.

[55] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021.

[56] Zhiwei Xu, Yutong Wang, Spencer Frei, Gal Vardi, and Wei Hu. Benign overfitting and grokking in reLU networks for XOR cluster data. *The Twelfth International Conference on Learning Representations*, 2024.

[57] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[58] Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. The benefits of mixup for feature learning. In *International Conference on Machine Learning*, pages 43423–43479. PMLR, 2023.

[59] Difan Zou, Yuan Cao, Yuanzhi Li, and Quanquan Gu. Understanding the generalization of Adam in learning neural networks with proper regularization. *The Eleventh International Conference on Learning Representations*, 2023.

Appendix

Contents

A Additional related Works **16**

B Preliminary Lemmas **16**

C Proof of Proposition 4.1 **17**

D Standard GD Fails to Generalize with low SNR **19**

D.1 Proof of Lemma 4.2 19

D.2 Proof of Lemma 4.3 21

D.3 Proof of Theorem 3.2 22

E label noise GD Successfully Generalizes with Low SNR **25**

E.1 Proof of Lemma 4.4 25

E.2 Proof of Lemma 4.5 29

E.3 Proof of Theorem 3.3 33

F Experimental Details for Figure 1 **35**

F.1 Dataset and Noise Injection 35

F.2 Model and Training Setup 35

F.3 Results Analysis 36

F.4 Reproducibility 36

G Additional Experiments **36**

G.1 Deeper Neural Network 36

G.2 Real World Dataset 36

G.3 Different Type of Label Noise 38

G.4 Higher Order Polynomial ReLU 39

A Additional related Works

Label Noise SGD. Recent works have empirically shown that label noise stochastic gradient descent (SGD) exhibits favorable generalization properties due to the regularization effect of the injected noise [24, 13]. From a theoretical standpoint, label noise SGD has been primarily explored in the context of linear regression or shallow neural networks, particularly in regression settings [6, 13, 24, 28, 39, 53, 50]; these studies have highlighted the implicit regularization benefits of label noise in SGD. For instance, [50] illustrated the implicit regularization of label noise in mean-field neural networks, while [39, 13] proved that label noise introduces bias towards flat minima. In contrast to these existing literature, our work focuses on the binary classification setting specified by the signal-noise model, providing a quantitative analysis of the training dynamics and the generalization benefits of label noise GD in the low SNR regime.

Signal-Noise Data Models. Recent theoretical works have studied the signal-noise model in various contexts, including (i) *optimization algorithms*, such as Adam [59], momentum [29], sharpness-aware minimization [10], large learning rates [41]; (ii) *learning paradigms*, such as ensembling and knowledge distillation [2], semi-and self-supervised learning [32, 55], Mixup [58, 11], adversarial training [1], and prompt tuning [42]; and (iii) *neural network structures*, such as convolutional neural network [8, 33], vision transformer [30, 36], graph neural network [25, 37]. Our work is in line with [9, 25], with the goal of showing that a simple algorithmic modification (label noise GD) facilitates feature learning in the challenging low SNR regime.

B Preliminary Lemmas

Lemma B.1 ([8]). *Suppose that $\delta > 0$ and $d = \Omega(\log(4n/\delta))$). Then with probability $1 - \delta$,*

$$\sigma_p^2 d/2 \le \|\boldsymbol{\xi}_i\|_2^2 \le 3\sigma_p^2 d/2,$$
$$|\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'} \rangle| \le 2\sigma_p^2 \sqrt{d \log(4n^2/\delta)},$$

for all $i, i' \neq i \in [n]$.

Lemma B.2 ([8]). *Suppose that $d \geq \Omega(\log(mn/\delta))$, $m = \Omega(\log(1/\delta))$. Then with probability at least $1 - \delta$, it satisfies that for all $r \in [m], j \in \{\pm 1\}, i \in [n]$,*

$$|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle| \le \sqrt{2\log(8m/\delta)}\sigma_0 \|\boldsymbol{\mu}\|_2$$
$$|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle| \le 2\sqrt{\log(8mn/\delta)}\sigma_0\sigma_p\sqrt{d}$$

and for all $j \in \{\pm 1\}, i \in [n]$

$$\sigma_0\|\boldsymbol{\mu}\|_2/2 \le \max_{r\in[m]} j\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle \le \sqrt{2\log(8m/\delta)}\sigma_0\|\boldsymbol{\mu}\|_2,$$
$$\sigma_0\sigma_p\sqrt{d}/4 \le \max_{r\in[m]} j\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle \le 2\sqrt{\log(8mn/\delta)}\sigma_0\sigma_p\sqrt{d}.$$

Lemma B.3. *Let $\mathcal{S}_{\pm}^{(t)} = \{i : \epsilon_i^{(t)} = \pm 1\}$ and $\mathcal{S}_j = \{i : y_i = j\}$. Then $\forall t \geq 0$, we have following with probability at least $1 - \delta$,*

1. $||\mathcal{S}_+^{(t)}| - n(1-p)| \le \sqrt{\frac{n}{2}\log\left(\frac{4T^*}{\delta}\right)}$*, and* $||\mathcal{S}_-^{(t)}| - np| \le \sqrt{\frac{n}{2}\log\left(\frac{4T^*}{\delta}\right)}$.
2. *The size of set follows,* $\forall j \in \{\pm 1\}$

$$\left||\mathcal{S}_+^{(t)} \cap \mathcal{S}_j| - \frac{(1-p)n}{2}\right| \le \sqrt{\frac{n}{2}\log\left(\frac{8T^*}{\delta}\right)}, \quad \left||\mathcal{S}_-^{(t)} \cap \mathcal{S}_j| - \frac{pn}{2}\right| \le \sqrt{\frac{n}{2}\log\left(\frac{8T^*}{\delta}\right)}.$$

Suppose $n \geq \frac{8\log(8T^*/\delta)}{p^2} \geq \frac{8\log(8T^*/\delta)}{(1-p)^2}$, we have

$$|\mathcal{S}_+^{(t)} \cap \mathcal{S}_j| \in \left[\frac{(2-3p)n}{4}, \frac{(2-p)n}{4}\right], \quad |\mathcal{S}_-^{(t)} \cap \mathcal{S}_j| \in \left[\frac{pn}{4}, \frac{3pn}{4}\right].$$

Proof of Lemma B.3. By independence, we have $\mathbb{E}|\mathcal{S}_+^{(t)}| = (1-p)n$ and $\mathbb{E}|\mathcal{S}_-^{(t)}| = pn$. By Hoeffding's inequality, we have for arbitrary $\tau > 0$,

$$\mathbb{P}\big(||\mathcal{S}_+^{(t)}| - (1-p)n| \geq \tau\big) \leq 2\exp\big(-\frac{2\tau^2}{n}\big), \quad \mathbb{P}\big(||\mathcal{S}_-^{(t)}| - pn| \geq \tau\big) \leq 2\exp\big(-\frac{2\tau^2}{n}\big).$$

Setting $\tau = \sqrt{(n/2)\log(4/\delta)}$ and taking the union bound over $[T^*]$ gives

$$||\mathcal{S}_+^{(t)}| - (1-p)n| \leq \sqrt{\frac{n}{2}\log\Big(\frac{4T^*}{\delta}\Big)}, \quad ||\mathcal{S}_-^{(t)}| - pn| \leq \sqrt{\frac{n}{2}\log\Big(\frac{4T^*}{\delta}\Big)},$$

which holds with probability at least $1-\delta$.

Similarly, by the same argument, we can show the result for $|\mathcal{S}_+^{(t)} \cap \mathcal{S}_j|$ and $|\mathcal{S}_-^{(t)} \cap \mathcal{S}_j|$.

Suppose $n \geq \frac{8\log(8T^*/\delta)}{p^2} \geq \frac{8\log(8T^*/\delta)}{(1-p)^2}$, then we have with probability at least $1-\delta$, we have $|\mathcal{S}_+^{(t)} \cap \mathcal{S}_j| \in \Big[\frac{(2-3p)n}{4}, \frac{(2-p)n}{4}\Big]$, $\quad |\mathcal{S}_-^{(t)} \cap \mathcal{S}_j| \in \big[\frac{pn}{4}, \frac{3pn}{4}\big]$. □

Lemma B.4. *Let* $\mathcal{S}_{i,\pm}^{(t)} := \{s \leq t : \epsilon_i^{(s)} = \pm 1\}$. *Then for any* $i \in [n]$, $t > 0$, *with probability at least* $1-\delta$,

1. $||\mathcal{S}_{i,+}^{(t)}| - (1-p)t| \leq \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}$ *and* $||\mathcal{S}_{i,-}^{(t)}| - pt| \leq \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}$.
2. *In addition, suppose* $t \geq \frac{2\log(4n/\delta)}{p^2}$, *we have* $|\mathcal{S}_{i,+}^{(t)}| \in [\frac{(2-3p)t}{2}, \frac{(2-p)t}{2}]$, $|\mathcal{S}_{i,-}^{(t)}| \in [\frac{pt}{2}, \frac{3pt}{2}]$.

Proof of Lemma B.4. By independence, we have $\mathbb{E}|\mathcal{S}_{i,+}^{(t)}| = (1-p)t$ and $\mathbb{E}|\mathcal{S}_{i,-}^{(t)}| = pt$. By Hoeffding's inequality, we have for arbitrary $\tau > 0$,

$$\mathbb{P}\big(||\mathcal{S}_{i,+}^{(t)}| - (1-p)t| \geq \tau\big) \leq 2\exp\big(-\frac{2\tau^2}{t}\big), \quad \mathbb{P}\big(||\mathcal{S}_{i,-}^{(t)}| - pt| \geq \tau\big) \leq 2\exp\big(-\frac{2\tau^2}{t}\big).$$

Setting $\tau = \sqrt{(t/2)\log(4/\delta)}$ and taking the union bound gives

$$||\mathcal{S}_{i,+}^{(t)}| - (1-p)t| \leq \sqrt{\frac{t}{2}\log\Big(\frac{4n}{\delta}\Big)}, \quad ||\mathcal{S}_{i,-}^{(t)}| - pt| \leq \sqrt{\frac{t}{2}\log\Big(\frac{4n}{\delta}\Big)},$$

which holds with probability at least $1-\delta$.

Suppose $t \geq \frac{2\log(4n/\delta)}{p^2} \geq \frac{2\log(4n/\delta)}{(1-p)^2}$, then we have with probability at least $1-\delta$, we have $|\mathcal{S}_{i,+}^{(t)}| \in [\frac{(2-3p)t}{2}, \frac{(2-p)t}{2}]$, $|\mathcal{S}_{i,-}^{(t)}| \in [\frac{pt}{2}, \frac{3pt}{2}]$. □

C Proof of Proposition 4.1

In this section, we provide a proof for Proposition 4.1, which establishes upper bounds for the absolute values of the signal learning and noise memorization coefficients throughout the entire training stage. Additionally, we present some preliminary lemmas that will be used in the proof of Proposition 4.1 as well as in other results in the subsequent sections.

Lemma C.1. *Suppose that inequalities (7) and (8) hold for all* $r \in [m]$, $j \in \{-1, 1\}$, $i \in [n]$ *and* $t \in [0, T^*]$. *For any* $\delta > 0$, *with probability at least* $1-\delta$, *it holds that*

$$|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - \rho_{j,r,i}^{(t)}| \leq 8\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha,$$
$$|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, j\boldsymbol{\mu}\rangle - \gamma_{j,r}^{(t)}| = 0.$$

Proof of Lemma C.1. From the signal-noise decomposition of $\mathbf{w}_{j,r}^{(t)}$, we have

$$\begin{aligned}|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - \rho_{j,r,i}^{(t)}| &\overset{(a)}{=} |j\gamma_{j,r}^{(t)}\langle \boldsymbol{\mu}, \boldsymbol{\xi}_i\rangle \|\boldsymbol{\mu}\|_2^{-2} + \sum_{i' \neq i} \rho_{j,r,i}^{(t)}\langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i\rangle \|\boldsymbol{\xi}_{i'}\|_2^{-2}| \\ &\overset{(b)}{\leq} 8\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha,\end{aligned}$$

where (a) follows from the weight decomposition, and inequality (b) is due to Lemma B.1 and the upper bound of $\rho_{j,r,i}^{(t)}$ based on inequalities (7) and (8).

Next, for the projection of the weight difference onto the signal vector, we have:

$$|\langle \mathbf{w}_{j,r}^{(t)} - \mathbf{w}_{j,r}^{(0)}, j\boldsymbol{\mu}\rangle - \gamma_{j,r}^{(t)}| = |\sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \|\boldsymbol{\xi}_i\|_2^{-2} \langle \boldsymbol{\xi}_i, \boldsymbol{\mu}\rangle| = 0,$$

where the equality holds because $\langle \boldsymbol{\xi}_i, \boldsymbol{\mu}\rangle = 0$ for $i \in [n]$ due to the covariance property of the noise vector distribution. □

With Lemma C.1 in place, we are now prepared to prove Proposition 4.1. The general proof strategy follows the approach outlined in [8]. However, we present a complete proof here for the sake of clarity and to provide a unified analysis for both gradient descent and label noise GD.

Proof of Proposition 4.1. The proof uses induction and covers both gradient descent and label noise gradient descent.

At $t = 0$, it is straightforward that the results hold for all coefficients, as they are initialized to zero. Now, assume that there exists a time step $\hat{T}$ such that for $t \in [1, \hat{T}]$ the following inequalities hold:

$$0 \le \gamma_{j,r}^{(t)} \le \alpha, \quad 0 \le \overline{\rho}_{j,r,i}^{(t)} \le \alpha,$$
$$0 \ge \underline{\rho}_{j,r,i}^{(t)} \ge -\beta - 16\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha \ge -\alpha.$$

To complete the induction, we need to show that the above inequalities hold for $t = \hat{T} + 1$. First, we examine $\underline{\rho}_{j,r,i}^{(\hat{T}+1)}$ for $j = -y_i$, since $\underline{\rho}_{j,r,i}^{(\hat{T}+1)} = 0$ when $j = y_i$ by definition. Using Lemma C.1, if $\underline{\rho}_{j,r,i}^{(\hat{T})} \le -0.5\beta - 8\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha$, we have

$$\langle \mathbf{w}_{j,r}^{(\hat{T})}, \boldsymbol{\xi}_i\rangle \le \underline{\rho}_{j,r,i}^{(\hat{T})} + 8\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha + \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle \le 0.$$

Thus,

$$\begin{aligned}\underline{\rho}_{j,r,i}^{(\hat{T}+1)} &= \underline{\rho}_{j,r,i}^{(\hat{T})} + \frac{\eta}{nm} \ell_i^{'(\hat{T})} \sigma'(\langle \mathbf{w}_{j,r}^{(\hat{T})}, \boldsymbol{\xi}_i\rangle) \|\boldsymbol{\xi}_i\|_2^2 \epsilon_i^{(\hat{T})} \\ &= \underline{\rho}_{j,r,i}^{(\hat{T})} \ge -\beta - 16\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha,\end{aligned}$$

where we have used $\sigma'(\langle \mathbf{w}_{j,r}^{(\hat{T})}, \boldsymbol{\xi}_i\rangle) = 0$. On the other hand, if $\underline{\rho}_{j,r,i}^{(\hat{T})} \ge -0.5\beta - 8\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha$, the update function implies:

$$\begin{aligned}\underline{\rho}_{j,r,i}^{(\hat{T}+1)} &\overset{(a)}{\ge} \underline{\rho}_{j,r,i}^{(\hat{T})} + \frac{\eta}{nm} \ell_i^{'(\hat{T})} \langle \mathbf{w}_{j,r}^{(\hat{T})}, \boldsymbol{\xi}_i\rangle \|\boldsymbol{\xi}_i\|_2^2 \\ &\overset{(b)}{\ge} -0.5\beta - 8\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha - \frac{3\eta\sigma_p^2 d}{2nm}(0.5\beta + 8\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha) \\ &\overset{(c)}{\ge} -\beta - 16\sqrt{\frac{\log(4n^2/\delta)}{d}} n\alpha,\end{aligned}$$

where (a) is due to choosing $\epsilon_i^{(\hat{T})} = 1$ and $\langle \mathbf{w}_{j,r}^{(\hat{T})}, \boldsymbol{\xi}_i\rangle > 0$, follows from Lemma B.1, and (c) holds when $\eta \le \frac{2nm}{3\sigma_p^2 d}$.

Next, consider $\overline{\rho}_{j,r,i}^{(\hat{T}+1)}$ for $j = y_i$. Let $\hat{T}_1$ to be the last time that $\overline{\rho}_{j,r,i}^{(t)} \leq 0.5\alpha$. By propagation, we have:

$$
\begin{aligned}
\overline{\rho}_{j,r,i}^{(\hat{T}+1)} &= \overline{\rho}_{j,r,i}^{(\hat{T}_1)} - \frac{\eta}{nm}\ell_i^{'(\hat{T}_1)}\sigma'(\langle \mathbf{w}_{j,r}^{(\hat{T}_1)}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2\epsilon_i^{(\hat{T}_1)} - \sum_{\hat{T}_1 < t \leq \hat{T}} \frac{\eta}{nm}\ell_i^{'(t)}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2\epsilon_i^{(t)} \\
&\overset{(a)}{\leq} 0.5\alpha + \frac{\eta}{nm}\langle \mathbf{w}_{j,r}^{(\hat{T}_1)}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|_2^2 + \sum_{\hat{T}_1 < t \leq \hat{T}} \frac{\eta}{nm}\ell_i^{'(t)}\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|_2^2 \\
&\overset{(b)}{\leq} 0.5\alpha + \frac{3\eta\sigma_p^2 d}{2nm}(0.5\alpha + \beta + 16\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha) \\
&\quad + \sum_{\hat{T}_1 < t \leq \hat{T}} \exp(-4\alpha^2 + 1)\frac{3\eta\sigma_p^2 d}{2nm}(\alpha + \beta + 16\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha) \\
&\overset{(c)}{\leq} 0.5\alpha + 0.25\alpha + 0.25\alpha = \alpha,
\end{aligned}
$$

where (a) holds since $\ell_i^{'(\hat{T}_1)} \geq -1$ and $\epsilon_i^{(t)} \leq 1$ for all $t \in [\hat{T}_1, \hat{T}]$, (b) is by Lemma B.1, Lemma C.1, and $-\tilde{\ell}_i^{'(t)} \leq \exp(-F_{y_i} + 1) \leq \exp(-4\alpha^2 + 1)$. Here we have used that $\beta + 16\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha \leq 2\alpha$ with the condition that $d = \tilde{\Omega}(n^2)$ and $\sigma_0 \leq \tilde{O}(1)\min\{\|\boldsymbol{\mu}\|_2^{-1}, \sigma_p^{-1}d^{-1/2}\}$. The final inequality (c) holds because $\eta = O(\frac{nm}{\sigma_p^2 d})$ and $\exp(-4\alpha^2 + 1)\alpha < 1$ with $\alpha = 4\log(T^*)$.

Similarly, we can prove that $\gamma_{j,r}^{(\hat{T}+1)} \leq \alpha$ using $\eta = O(\frac{nm}{\|\boldsymbol{\mu}\|_2^2})$, which completes the induction proof. □

D Standard GD Fails to Generalize with low SNR

D.1 Proof of Lemma 4.2

In this section, we provide a proof for the result obtained in the first stage of gradient descent training. Several preliminary lemmas are established to facilitate the analysis.

Lemma D.1 (Upper bound on $\gamma_{j,r}^{(t)}$**).** *Under Assumption 3.1, in the first stage, where* $0 \leq t \leq T_1 = \frac{nm\log(1/(\sigma_0\sigma_p\sqrt{d}))}{\eta\sigma_p^2 d}$*, there exists an upper bound for* $\gamma_{j,r}^{(t)}$*, for all* $j \in \{-1, 1\}, r \in [m]$*:*

$$
\gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \leq \exp\big(\frac{2\eta\|\boldsymbol{\mu}\|_2^2}{m}t\big)|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|.
$$

Proof of Lemma D.1. By the iterative update rule of signal learning, we have:

$$
\begin{aligned}
\gamma_{j,r}^{(t+1)} &\overset{(a)}{\leq} \gamma_{j,r}^{(t)} + \frac{\eta}{nm}\sum_{i=1}^{n}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i\boldsymbol{\mu}\rangle)\|\boldsymbol{\mu}\|_2^2 \\
&\overset{(b)}{=} \gamma_{j,r}^{(t)} + \frac{\eta}{nm}\sum_{i=1}^{n}\sigma'(y_i\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle + jy_i\gamma_{j,r}^{(t)})\|\boldsymbol{\mu}\|_2^2 \\
&\overset{(c)}{\leq} \gamma_{j,r}^{(t)} + \frac{2\eta}{m}(\gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|)\|\boldsymbol{\mu}\|_2^2.
\end{aligned}
$$

where (a) follows from $|\ell_i^{'(t)}| \leq 1$, (b) is derived using Lemma C.1, and (c) is due to the properties of the squared ReLU activation function.

Define $A^{(t)} := \gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|$. Then, we have:

$$
A^{(t+1)} \leq \big(1 + \frac{2\eta\|\boldsymbol{\mu}\|_2^2}{m}\big)A^{(t)} \leq \big(1 + \frac{2\eta\|\boldsymbol{\mu}\|_2^2}{m}\big)^{(t)}A^{(0)} \leq \exp\big(\frac{2\eta\|\boldsymbol{\mu}\|_2^2}{m}t\big)A^{(0)},
$$

where we use $1 + x \leq \exp(x)$. This suggests:

$$\gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \leq \exp\big(\frac{2\eta\|\boldsymbol{\mu}\|_2^2}{m}t\big)|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|.$$

□

Lemma D.2 (Upper bound on $\underline{\rho}_{j,r,i}^{(t)}$)**.** *Under Assumption 3.1, in the first stage, where* $0 \leq t \leq T_1 = \frac{nm\log(1/(\sigma_0\sigma_p\sqrt{d}))}{\eta\sigma_p^2 d}$*, there exists an upper bound for* $|\underline{\rho}_{j,r,i}^{(t)}|$*, for all* j, r, i*:*

$$|\underline{\rho}_{j,r,i}^{(t)}| = \tilde{O}(\sigma_0\sigma_p\sqrt{d}).$$

Proof of Lemma D.2. The proof uses the induction method. By the iterative update rule for noise memorization, we have:

$$\begin{aligned}
|\underline{\rho}_{j,r,i}^{(t+1)}| &\overset{(a)}{\leq} |\underline{\rho}_{j,r,i}^{(t)}| + \frac{\eta}{nm}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2 \\
&\overset{(b)}{\leq} |\underline{\rho}_{j,r,i}^{(t)}| + \frac{3\eta\sigma_p^2 d}{2nm}\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle + 16\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha + \underline{\rho}_{j,r,i}^{(t)}) \\
&\overset{(c)}{\leq} |\underline{\rho}_{j,r,i}^{(t)}| + \frac{3\eta\sigma_p^2 d}{nm}\sqrt{\log(8mn/\delta)}\sigma_0\sigma_p\sqrt{d},
\end{aligned}$$

where the inequality (a) is by the upper bound on $|\ell_i^{'(t)}| \leq 1$; Inequality (b) is derived using Proposition 4.1, Lemma B.1, and Lemma C.1. Finally, the inequality (c) uses the fact that $\underline{\rho}_{j,r,i}^{(t)} < 0$ and Lemma B.2.

Taking a telescoping sum over t form 0 to T_1, we obtain:

$$|\underline{\rho}_{j,r,i}^{(T_1)}| \leq \frac{3\eta\sigma_p^2 d T_1}{nm}\sqrt{\log(8mn/\delta)}\sigma_0\sigma_p\sqrt{d} = \tilde{O}(\sigma_0\sigma_p\sqrt{d}),$$

where we substituted $T_1 = \Theta(\frac{nm\log(1/(\sigma_0\sigma_p\sqrt{d}))}{\eta\sigma_p^2 d})$, thereby completing the proof. □

Lemma D.3. *Let* $\bar{\beta} = \min_{i\in[n]}\max_{r\in[m]}\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle$*. Suppose that* $\sigma_0 \geq 160n\sqrt{\frac{\log(4n^2/\delta)}{d}}(\sigma_p\sqrt{d})^{-1}\alpha$*. Then it holds that* $\bar{\beta} \geq 40n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha$*.*

Proof of Lemma D.3. The proof follows directly from Lemma B.2. With high probability, we have: $\overline{\beta} \geq \sigma_0\sigma_p\sqrt{d}/4$. Substituting the condition on σ_0, we obtain:

$$\overline{\beta} \geq 40n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha.$$

□

Lemma D.4 (Lower bound on $\overline{\rho}_{j,r,i}^{(t)}$)**.** *Under Assumption 3.1, in the first stage, where* $0 \leq t \leq T_1 = \frac{nm\log(1/(\sigma_0\sigma_p\sqrt{d}))}{\eta\sigma_p^2 d}$*, there exists a lower bound for* $\max_{j,r}\overline{\rho}_{j,r,i}^{(t)}$*, for all* $i \in [n]$*:*

$$\max_{j,r}\overline{\rho}_{j,r,i}^{(t)} + \overline{\beta} \geq \exp\big(\frac{\eta C_1\sigma_p^2 d}{2nm}t\big)\sigma_0\sigma_p\sqrt{d}/4.$$

Proof of Lemma D.4. By the iterative update rule for noise memorization, we have:

$$\begin{aligned}
\max_{j,r}\overline{\rho}_{j,r,i}^{(t+1)} &\overset{(a)}{\geq} \max_{j,r}\overline{\rho}_{j,r,i}^{(t)} + \max_{j,r}\frac{\eta C_1}{nm}\sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)\|\boldsymbol{\xi}_i\|_2^2 \\
&\overset{(b)}{\geq} \max_{j,r}\overline{\rho}_{j,r,i}^{(t)} + \max_{j,r}\frac{\eta\sigma_p^2 d C_1}{2nm}\sigma'(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - 16\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha + \overline{\rho}_{j,r,i}^{(t)}) \\
&\overset{(c)}{\geq} \max_{j,r}\overline{\rho}_{j,r,i}^{(t)} + \frac{\eta\sigma_p^2 d C_1}{nm}(\max_{j,r}\overline{\rho}_{j,r,i}^{(t)} + \frac{2}{5}\overline{\beta}),
\end{aligned}$$

where the inequality (a) is by the lower bound on $|\ell_i^{'(t)}| \geq C_1$ in the first stage; Inequality (b) is by Lemma B.1 and Lemma C.1. Finally, the inequality (c) is by Lemma D.3.

Define $B_i^{(t)} := \max_{j,r} \overline{\rho}_{j,r,i}^{(t)} + 0.6\overline{\beta}$. Then

$$B_i^{(t+1)} \geq \big(1 + \frac{\eta C_1 \sigma_p^2 d}{nm}\big) B_i^{(t)} \geq \big(1 + \frac{\eta C_1 \sigma_p^2 d}{nm}\big)^{(t)} B_i^{(0)} \geq \exp\big(\frac{\eta C_1 \sigma_p^2 d}{2nm} t\big) B_i^{(0)},$$

where we used $1 + x \geq \exp(x/2)$ for $x \leq 2$. □

With the above lemmas in place, we are now ready to prove Lemma 4.2.

Proof of Lemma 4.2. We choose the end of stage 1 as $T_1 = \frac{4nm}{\eta\sigma_p^2 d}\log(1/(\sigma_0\sigma_p\sqrt{d}))$. Then by Lemma D.4, we conclude that $\max_{j,r} \overline{\rho}_{j,r,i}^{(T_1)} \geq 1$, for all $i \in [n]$. Besides, by Lemma D.2, we directly obtain the result that

$$|\underline{\rho}_{j,r,i}^{(T_1)}| \leq \frac{3\eta\sigma_p^2 d T_1}{nm}\sqrt{\log(8mn/\delta)}\sigma_0\sigma_p\sqrt{d} = \tilde{O}(\sigma_0\sigma_p\sqrt{d}).$$

Finally, Lemma D.1 yields

$$\gamma_{j,r}^{(T_1)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \leq \exp\big(\frac{2\eta\|\boldsymbol{\mu}\|_2^2}{m}\frac{4nm}{\eta\sigma_p^2 d}\log(1/(\sigma_0\sigma_p d))\big)|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \leq 2|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|,$$

where we have used the condition of low SNR, namely $n\mathrm{SNR}^2 \leq 1/\log(\sigma_0\sigma_p d)$. By Lemma B.2, we conclude the proof for $\max_{j,r}\gamma_{j,r}^{(T_1)} = \tilde{O}(\sigma_0\|\boldsymbol{\mu}\|_2)$. □

D.2 Proof of Lemma 4.3

In this section, we provide a complete proof for Lemma 4.3 based on Lemma 4.2 and an iterative analysis of the training dynamics. We introduce several necessary preliminary lemmas that will be used in the proof for $t \in [T_1, T_2]$ with $T_2 = \eta^{-1}\sigma_p^{-2}d^{-1}nm\log(1/(\sigma_0\sigma_p\sqrt{d})) + \eta^{-1}\epsilon^{-1}m^3 n\sigma_p^{-2}d^{-1}$.

Lemma D.5 ([8]). *Under the same condition as Theorem 3.2, for all* $t \in [T_1, T_2]$ *and* $i \in [n]$*, the following properties hold:*

$$\begin{aligned}
&\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2 = O(\sigma_p^2 d) L_S(\mathbf{W}^{(t)}),\\
&\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F = \tilde{O}(m^{3/2}n^{1/2}\sigma_p^{-1}d^{-1/2}),\\
&y_i\langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^*\rangle \geq 2\log(2/\epsilon).
\end{aligned}$$

With the above lemmas at hand, we are now ready to provide the complete proof for Lemma 4.3.

Proof of Lemma 4.3. We start by showing the convergence of gradient descent. The key idea is to construct a reference weight matrix $\mathbf{W}^*$ defined as $\mathbf{w}_{j,r}^* = \mathbf{w}_{j,r}^{(0)} + 2m\log(2/\epsilon)\sum_{i=1}^n \|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i$.

Summing the above inequality from $\mathbf{W}^{(t)}$ and $\mathbf{W}^*$:

$$\begin{aligned}
&\|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t+1)} - \mathbf{W}^*\|_F^2\\
&= \|\mathbf{W}^{(t)} - \mathbf{W}^*\|_F^2 - \|\mathbf{W}^{(t)} - \eta\nabla L_S(\mathbf{W}^{(t)}) - \mathbf{W}^*\|_F^2\\
&= 2\eta\langle \nabla L_S(\mathbf{W}^{(t)}), \mathbf{W}^{(t)} - \mathbf{W}^*\rangle - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2\\
&\overset{(a)}{=} \frac{2\eta}{n}\sum_{i=1}^n \ell_i^{'(t)}[2y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) - \langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^*\rangle] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2\\
&\overset{(b)}{\geq} \frac{2\eta}{n}\sum_{i=1}^n \ell_i^{'(t)}[2y_i f(\mathbf{W}^{(t)}, \mathbf{x}_i) - 2\log(2/\epsilon)] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2\\
&\overset{(c)}{\geq} \frac{4\eta}{n}\sum_{i=1}^n [\ell_i^{(t)} - \epsilon/2] - \eta^2\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2\\
&\overset{(d)}{\geq} 2\eta(L_S(\mathbf{W}^{(t)}) - \epsilon),
\end{aligned}$$

where in equation (a), we have applied the homogeneity property of the squared ReLU activation. The inequality (b) is by $\langle \nabla f(\mathbf{W}^{(t)}, \mathbf{x}_i), \mathbf{W}^* \rangle \geq 2\log(2/\epsilon)$ as stated in Lemma D.5, and the inequality (c) is due to the convexity of the logistic function. Finally, the inequality (d) is by Lemma D.5 and the condition on the learning rate.

Taking a summation over the above inequality from T_1 to T_2, we have

$$\begin{aligned}\sum_{t=T_1}^{T_2} L_S(\mathbf{W}^{(t)}) &\leq \frac{\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2 + \eta\epsilon(T_2 - T_1 + 1)}{2\eta} \\ &\leq \frac{\|\mathbf{W}^{(T_1)} - \mathbf{W}^*\|_F^2}{\eta} \\ &\leq \tilde{O}(\eta^{-1} m^3 n \sigma_p^{-2} d^{-1}), \qquad (9)\end{aligned}$$

where in the second inequality, we have applied Lemma D.5. Finally, plugging in the $T_2 = \eta^{-1}\epsilon^{-1} m^3 n \sigma_p^{-2} d^{-1} + \eta^{-1}\sigma_p^{-2} d^{-1} nm \log(1/(\sigma_0 \sigma_p \sqrt{d}))$, we achieve $L_S(\mathbf{W}^{(t)}) \leq \epsilon$.

Next, we provide the lower bound for the noise memorization coefficient $\overline{\rho}_{j,r,i}^{(t)}$ and the upper bound for the signal learning coefficient $\gamma_{j,r}^{(t)}$ in the second stage. For the noise memorization coefficient, using its update equation:

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{\eta}{nm} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \|\boldsymbol{\xi}_i\|_2^2 \geq \overline{\rho}_{j,r,i}^{(t)}.$$

Here, we have used $\ell_i'^{(t)} \geq 0$ and property of the squared ReLU activation. This implies that $\overline{\rho}_{j,r,i}^{(t)}$ never decreases during training. Therefore, we have $\max_{j,r} \overline{\rho}_{j,r,i}^{(t)} \geq 1$, for all $i \in [n]$ and $t \in [T_1, T_2]$. For the signal learning coefficient, we use the induction method. From Lemma 4.2, we know that $\max_{j,r} \gamma_{j,r}^{(T_1)} = \tilde{O}(\sigma_0 \|\boldsymbol{\mu}\|_2) \triangleq \hat{\beta}$. Suppose that there exists $T \in [T_1, T_2]$ such that $\max_{j,r} \gamma_{j,r}^{(t)} \leq 2\hat{\beta}$ for all $t \in [T_1, T]$. Then we analyze:

$$\begin{aligned}\gamma_{j,r}^{(T+1)} &= \gamma_{j,r}^{(T_1)} - \frac{\eta}{nm} \sum_{t=T_1}^{T} \sum_{i=1}^{n} \ell_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}, \boldsymbol{\mu} \rangle) \|\boldsymbol{\mu}\|_2^2 \\ &\overset{(a)}{\leq} \gamma_{j,r}^{(T_1)} + \frac{2\eta\hat{\beta}}{nm} \|\boldsymbol{\mu}\|_2^2 \sum_{t=T_1}^{T} \sum_{i=1}^{n} |\ell_i'^{(t)}| \\ &\overset{(b)}{\leq} \gamma_{j,r}^{(T_1)} + \frac{2\eta\hat{\beta}}{nm} \|\boldsymbol{\mu}\|_2^2 \sum_{t=T_1}^{T} L_S(\mathbf{W}^{(t)}) \\ &\overset{(c)}{\leq} \gamma_{j,r}^{(T_1)} + \frac{2\eta\hat{\beta}}{nm} \|\boldsymbol{\mu}\|_2^2 \tilde{O}(\eta^{-1} m^3 n \sigma_p^{-2} d^{-1}) \\ &\leq \gamma_{j,r}^{(T_1)} + \tilde{O}(n\mathrm{SNR}^2) \overset{(d)}{\leq} 2\hat{\beta}.\end{aligned}$$

where the inequality (a) is due to Lemma C.1, the inequality (b) is by $|\ell_i'| \leq \ell_i$ for $i \in [n]$, and the inequality (c) is due to the inequality (9). Finally, the inequity (d) is by the condition that $n^{-1}\mathrm{SNR}^{-2} = \tilde{\Omega}(1)$. Similarly, with the induction method, we can show that $|\underline{\rho}_{j,r,i}^{(t)}| \leq \tilde{O}(\sigma_0 \sigma_p \sqrt{d})$. □

D.3 Proof of Theorem 3.2

To complete the proof of Theorem 3.2, we provide a proof for the generalization result.

Lemma D.6. *Define $g(\boldsymbol{\xi}_i) = \frac{1}{m} j \sum_{j,r} \sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle)$. Under Assumption 3.1, there exists a fixed vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 \leq 0.02\sigma_p$ such that*

$$\sum_{j \in \{\pm 1\}} [g(j\boldsymbol{\xi}_i + \mathbf{v}) - g(\boldsymbol{\xi}_i)] \geq 4\tilde{\Omega}(\sigma_0^2 \|\boldsymbol{\mu}\|_2^2).$$

Proof of Lemma D.6. To proceed with the proof, we construct the vector $\mathbf{v} \triangleq \lambda \sum_{i:y_i=1} \boldsymbol{\xi}_i$, where $\lambda = 0.01/\sqrt{nd}$. Then we show that

$$\begin{aligned}
\|\mathbf{v}\|_2^2 = \|\lambda \sum_{i:y_i=1} \boldsymbol{\xi}_i\|_2^2 &= \lambda^2 \langle \sum_{i:y_i=1} \boldsymbol{\xi}_i, \sum_{i:y_i=1} \boldsymbol{\xi}_i \rangle \\
&= \lambda^2 \sum_{i:y_i=1} \|\boldsymbol{\xi}_i\|_2^2 + 2\lambda^2 \sum_i \sum_{j \neq i} \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_j \rangle \\
&\leq \lambda^2 n \sigma_p^2 d + 4n^2 \lambda^2 \sigma_p^2 \sqrt{2d \log(4n^2/\delta)} \\
&\leq 4\lambda^2 n \sigma_p^2 d = 0.02^2 \sigma_p^2,
\end{aligned}$$

where the first inequity is by Lemma B.1, the second inequality is by $d \geq \tilde{\Omega}(n^2)$, and the final equality is by $\lambda = 0.01/\sqrt{nd}$, which confirms that $\|\mathbf{v}\|_2 \leq 0.02\sigma_p$.

By the convexity property of the squared ReLU function, we have that

$$\begin{aligned}
\sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i + \mathbf{v} \rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i \rangle) &\geq \sigma'(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle, \\
\sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}_i + \mathbf{v} \rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}_i \rangle) &\geq \sigma'(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}_i \rangle) \langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle.
\end{aligned}$$

With the above inequalities, we have that almost surely for all $\boldsymbol{\xi}_i$:

$$\begin{aligned}
&\sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i + \mathbf{v} \rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i \rangle) + \sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}_i + \mathbf{v} \rangle) - \sigma(\langle \mathbf{w}_{1,r}^{(t)}, -\boldsymbol{\xi}_i \rangle) \\
&\geq 4|\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i \rangle| |\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle|.
\end{aligned}$$

On the other hand, using the properties of the squared ReLU function and the triangle inequality, we have:

$$\begin{aligned}
&\sigma(\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}_i + \mathbf{v} \rangle) - \sigma(\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}_i \rangle) + \sigma(\langle \mathbf{w}_{-1,r}^{(t)}, -\boldsymbol{\xi}_i + \mathbf{v} \rangle) - \sigma(\langle \mathbf{w}_{-1,r}^{(t)}, -\boldsymbol{\xi}_i \rangle) \\
&\leq (\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}_i \rangle + |\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v} \rangle|)^2 + (-\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}_i \rangle + |\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v} \rangle|)^2 - \langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}_i \rangle^2 \\
&\leq |\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}_i \rangle|^2 + 2|\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v} \rangle|^2.
\end{aligned}$$

Next, we compare $|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle|$ and $|\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v} \rangle|$ with $|\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i \rangle|$ and $|\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}_i \rangle|$. We show that

$$\begin{aligned}
|\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v} \rangle| &= \lambda |(\sum_{i:y_i=1} \underline{\rho}_{-1,r,i}^{(t)} + \langle \mathbf{w}_{-1,r}^{(0)}, \sum_{i:y_i=1} \boldsymbol{\xi}_i \rangle)| \\
&\leq \lambda(n\sqrt{\log(12mn/\delta)})\sigma_0 \sigma_p \sqrt{d}) \leq \lambda n/4,
\end{aligned}$$

where the first inequality is by Lemma B.2 and Lemma 4.3, and the second inequality is by the condition on σ_0 from Assumption 3.1. Besides,

$$\begin{aligned}
|\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle| &= \lambda |(\sum_{i:y_i=1} \overline{\rho}_{1,r,i}^{(t)} + \langle \mathbf{w}_{1,r}^{(0)}, \sum_{i:y_i=1} \boldsymbol{\xi}_i \rangle)| \\
&\geq \lambda(n - n\sqrt{\log(12mn/\delta)})\sigma_0 \sigma_p \sqrt{d}) \geq \lambda n/2,
\end{aligned}$$

where the first inequality is by Lemma B.2 and Lemma 4.3; and the second inequality is by the condition on σ_0 from Assumption 3.1.

Finally, by Lemma B.2, Proposition 4.1, and Lemma B.1 it holds that

$$\begin{aligned}
|\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i \rangle| &= |\langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\xi}_i \rangle + \sum_{i'=1}^{n} \rho_{j,r,i'}^{(t)} \|\boldsymbol{\xi}_{i'}\|_2^{-2} \langle \boldsymbol{\xi}_{i'}, \boldsymbol{\xi}_i \rangle| \\
&\leq \sqrt{\log(12mn/\delta)})\sigma_0 \sigma_p \sqrt{d} + 8\sqrt{\frac{\log(4n^2/\delta)}{d}} \sqrt{n}\alpha.
\end{aligned}$$

On the other hand, it is observed that $\langle \mathbf{w}_{1,r}^{(t)} - \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\xi}_i \rangle \sim \mathcal{N}(0, \sigma_w^2)$, where the variance σ_w follows

$$
\begin{aligned}
\sigma_w^2 &= \sigma_p^2 \sum_{k=1}^{d} (\sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \|\boldsymbol{\xi}_i\|_2^{-2} \xi_{i,k})^2 \\
&\overset{(a)}{\geq} \frac{1}{2}\sigma_p^2 \sum_{k=1}^{d} \sum_{i=1}^{n} (\rho_{j,r,i}^{(t)})^2 \|\boldsymbol{\xi}_i\|_2^{-4} \xi_{i,k}^2 \\
&= \frac{1}{2}\sigma_p^2 \sum_{i=1}^{n} (\rho_{j,r,i}^{(t)})^2 \|\boldsymbol{\xi}_i\|_2^{-2} \\
&\geq \frac{1}{3d} \sum_{i=1}^{n} (\rho_{j,r,i}^{(t)})^2 \geq \frac{n}{6d},
\end{aligned}
$$

where (a) is by Lemma B.1 and condition on d from Assumption 3.1, (b) is due to Lemma B.1, and (c) is by Lemma 4.3.

By the anti-concentration inequality of Gaussian variance, we have

$$
\begin{aligned}
\mathbb{P}(|\langle \mathbf{w}_{1,r}^{(t)} - \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\xi}_i \rangle| \leq \tau) &\leq 2\mathrm{erf}(\frac{\tau}{\sqrt{2}\sigma_w}) \leq 2\mathrm{erf}(\frac{\tau\sqrt{6d}}{\sqrt{2n}}) \\
&\leq 2\sqrt{1 - \exp(-\frac{12d\tau^2}{\pi n})}.
\end{aligned}
$$

Then with probability at least $1 - \delta$, it holds that

$$
|\langle \mathbf{w}_{1,r}^{(t)} - \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\xi}_i \rangle| \geq \sqrt{\frac{\pi n}{12d} \log(\frac{1}{1 - (\delta/2)^2})} \geq \sqrt{\frac{\pi n \delta^2}{96d}},
$$

where we have used $\log(1+x) \geq \frac{x}{1+x}$ for $x > -1$ and $\delta^2 \leq 1/8$.

Together, we conclude that

$$
\begin{aligned}
\sum_{j \in \{\pm 1\}} [g(j\boldsymbol{\xi}_i + \mathbf{v}) - g(\boldsymbol{\xi}_i)] &\geq 4|\langle \mathbf{w}_{1,r}^{(t)}, \boldsymbol{\xi}_i \rangle||\langle \mathbf{w}_{1,r}^{(t)}, \mathbf{v} \rangle| + |\langle \mathbf{w}_{-1,r}^{(t)}, \boldsymbol{\xi}_i \rangle|^2 + 2|\langle \mathbf{w}_{-1,r}^{(t)}, \mathbf{v} \rangle|^2 \\
&\geq 4(\lambda/2)\sqrt{\frac{\pi n \delta^2}{96d}} \geq 4\tilde{\Omega}(\sigma_0^2 \|\boldsymbol{\mu}\|_2^2),
\end{aligned}
$$

where the final inequality holds by $\sigma_0^2 \leq \tilde{O}(\frac{1}{d^{5/4}\|\boldsymbol{\mu}\|_2^2})$ with δ chosen as $d^{-1/4}$, thus completing the proof. □

Proof of Theorem 3.2. For the population loss, we expand the expression

$$
\begin{aligned}
L_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbb{1}(y \neq \mathrm{sign}(f(\mathbf{W}, \mathbf{x}))] = \mathbb{P}(yf(\mathbf{W}^{(t)}, \mathbf{x}) < 0) \\
&= \mathbb{P}\Big(\frac{1}{m}\sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}_i \rangle) - \frac{1}{m}\sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i \rangle) \geq \\
&\qquad \frac{1}{m}\sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu} \rangle) - \frac{1}{m}\sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu} \rangle)\Big).
\end{aligned}
$$

Recall the weight decomposition:

$$
\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j\gamma_{j,r}^{(t)} \|\boldsymbol{\mu}\|_2^{-2} \boldsymbol{\mu} + \sum_{i=1}^{n} \overline{\rho}_{j,r,i}^{(t)} \|\boldsymbol{\xi}_i\|_2^{-2} \boldsymbol{\xi}_i + \sum_{i=1}^{n} \underline{\rho}_{j,r,i}^{(t)} \|\boldsymbol{\xi}_i\|_2^{-2} \boldsymbol{\xi}_i.
$$

Then we conclude that:

$$
\begin{aligned}
\langle \mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu} \rangle &= \langle \mathbf{w}_{-y,r}^{(0)}, y\boldsymbol{\mu} \rangle - \gamma_{-y,r}^{(t)}, \\
\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu} \rangle &= \langle \mathbf{w}_{y,r}^{(0)}, y\boldsymbol{\mu} \rangle + \gamma_{y,r}^{(t)}.
\end{aligned}
$$

First, we provide the bound for the signal learning part:

$$
\begin{aligned}
&\frac{1}{m}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}\rangle) - \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu}\rangle) \\
&\leq \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}\rangle) = \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{y,r}^{(0)}, y\boldsymbol{\mu}\rangle + \gamma_{y,r}^{(t)}) \\
&\leq (\langle \mathbf{w}_{y,r}^{(0)}, y\boldsymbol{\mu}\rangle + \gamma_{y,r}^{(t)})^2 \\
&\leq \tilde{O}(\sigma_0^2\|\boldsymbol{\mu}\|_2^2),
\end{aligned}
$$

where the first and second inequalities follow from the properties of the squared ReLU function, and the last inequality is by Lemma B.2 and Lemma 4.3.

Denote that $g(\boldsymbol{\xi}_i) = \frac{1}{m} j \sum_{j,r}\sigma(\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle)$. It follows that:

$$
\begin{aligned}
&\mathbb{P}(yf(\mathbf{W}^{(t)}, \mathbf{x}) < 0) \\
&= \mathbb{P}\Big(\frac{1}{m}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}_i\rangle) - \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}_i\rangle) \geq \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}\rangle) - \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle \mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu}\rangle)\Big) \\
&\geq 0.5\mathbb{P}\Big(|g(\boldsymbol{\xi}_i)| \geq \tilde{\Omega}(\sigma_0^2\|\boldsymbol{\mu}\|_2^2)\Big).
\end{aligned}
$$

Define the set $\mathcal{A} = \{\boldsymbol{\xi}_i : |g(\boldsymbol{\xi}_i)| \geq \tilde{\Omega}(\sigma_0^2\|\boldsymbol{\mu}\|_2^2)\}$. By Lemma D.6, we have:

$$
\sum_{j\in\{\pm 1\}}[g(j\boldsymbol{\xi}_i + \mathbf{v}) - g(\boldsymbol{\xi}_i)] \geq 4\tilde{\Omega}(\sigma_0^2\|\boldsymbol{\mu}\|_2^2).
$$

Thus, there must exist at least one of $\boldsymbol{\xi}_i$, $\boldsymbol{\xi}_i + \mathbf{v}$, $-\boldsymbol{\xi}_i$ and $-\boldsymbol{\xi}_i + \mathbf{v}$ that belongs to $\mathcal{A}$ and the probability is larger than 0.25. Furthermore, we have:

$$
\begin{aligned}
|\mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{A} - \mathbf{v})| &= |\mathbb{P}_{\boldsymbol{\xi}_i\sim\mathcal{N}(\mathbf{0},\sigma_p^2\mathbf{I})}(\boldsymbol{\xi}_i \in \mathcal{A}) - \mathbb{P}_{\boldsymbol{\xi}_i\sim\mathcal{N}(\mathbf{v},\sigma_p^2\mathbf{I})}(\boldsymbol{\xi}_i \in \mathcal{A})| \\
&\leq \frac{\|\mathbf{v}\|_2}{2\sigma_p} \leq 0.02,
\end{aligned}
$$

where the first inequality is by Proposition 2.1 in [16] and the second inequality is by $\|\mathbf{v}\|_2 \leq 0.01\sigma_p$ according to Lemma D.6. Combined with that $\mathbb{P}(\mathcal{A}) = \mathbb{P}(-\mathcal{A})$, we finally achieve that $\mathbb{P}(\mathcal{A}) \geq 0.24$, corresponding to the second bullet result. Combined with Lemma 4.3, which establishes the first bullet point, this completes the proof of 3.2

□

E label noise GD Successfully Generalizes with Low SNR

E.1 Proof of Lemma 4.4

Lemma E.1 (Lower bound on $\gamma_{j,r}^{(t)}$). *Under Assumption 3.1, during the first stage, where $0 \leq t \leq T_1 = \frac{nm\log(1/(\sigma_0\sigma_p\sqrt{d}))}{\eta\sigma_p^2 d}$, there exists an lower bound for $\gamma_{j,r}^{(t)}$, for all j:*

$$
\max_{r\in[m]}\gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \geq \exp\big(\frac{C_0\eta\|\boldsymbol{\mu}\|_2^2}{8m}t\big)\max_{r\in[m]}|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|.
$$

where C_0 is the lower bound on $|\tilde{\ell}'^{(t)}| \geq C_0$ is the first stage.

Proof of Lemma E.1. If $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle \geq 0$, then

$$
\begin{aligned}
\gamma_{j,r}^{(t+1)} &= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \sum_{i=1}^{n} \tilde{\ell}_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \boldsymbol{\mu} \rangle) \|\boldsymbol{\mu}\|_2^2 \epsilon_i^{(t)} \\
&= \gamma_{j,r}^{(t)} - \frac{\eta}{nm} \Big[\sum_{i \in \mathcal{S}_+^{(t)}} \tilde{\ell}_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \boldsymbol{\mu} \rangle) - \sum_{i \in \mathcal{S}_-^{(t)}} \tilde{\ell}_i'^{(t)} \sigma'(\langle \mathbf{w}_{j,r}^{(t)}, y_i \boldsymbol{\mu} \rangle) \Big] \|\boldsymbol{\mu}\|_2^2 \\
&= \gamma_{j,r}^{(t)} - \frac{2\eta}{nm} \Big[\sum_{i \in \mathcal{S}_+^{(t)} \cap \mathcal{S}_1} \tilde{\ell}_i'^{(t)} - \sum_{i \in \mathcal{S}_-^{(t)} \cap \mathcal{S}_1} \tilde{\ell}_i'^{(t)} \Big] \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle \|\boldsymbol{\mu}\|_2^2 \\
&\geq \gamma_{j,r}^{(t)} + \frac{2\eta}{nm} \big(C_0 |\mathcal{S}_+^{(t)} \cap \mathcal{S}_1| - |\mathcal{S}_-^{(t)} \cap \mathcal{S}_1| \big) \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle \|\boldsymbol{\mu}\|_2^2.
\end{aligned}
$$

Note that we have defined $\mathcal{S}_{\pm}^{(t)} = \{i : \epsilon_i^{(t)} = \pm 1\}$ and $\mathcal{S}_j = \{i : y_i = j\}$ in Lemma B.3.

On the other hand, when $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle < 0$,

$$
\begin{aligned}
\gamma_{j,r}^{(t+1)} &= \gamma_{j,r}^{(t)} - \frac{2\eta}{nm} \Big[\sum_{i \in \mathcal{S}_+^{(t)} \cap \mathcal{S}_{-1}} \tilde{\ell}_i'^{(t)} - \sum_{i \in \mathcal{S}_-^{(t)} \cap \mathcal{S}_{-1}} \tilde{\ell}_i'^{(t)} \Big] \langle -\mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle \|\boldsymbol{\mu}\|_2^2 \\
&\geq \gamma_{j,r}^{(t)} + \frac{2\eta}{nm} \big(C_0 |\mathcal{S}_+^{(t)} \cap \mathcal{S}_{-1}| - |\mathcal{S}_-^{(t)} \cap \mathcal{S}_{-1}| \big) \langle -\mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle \|\boldsymbol{\mu}\|_2^2.
\end{aligned}
$$

By Lemma B.3, we have

$$
\begin{aligned}
\frac{|\mathcal{S}_+^{(t)} \cap \mathcal{S}_1|}{|\mathcal{S}_-^{(t)} \cap \mathcal{S}_1|}, \frac{|\mathcal{S}_+^{(t)} \cap \mathcal{S}_{-1}|}{|\mathcal{S}_-^{(t)} \cap \mathcal{S}_{-1}|} &\geq \frac{(1-p)n - \sqrt{2n \log(8T^*/\delta)}}{pn + \sqrt{2n \log(8T^*/\delta)}}, \\
|\mathcal{S}_+^{(t)} \cap \mathcal{S}_1|, |\mathcal{S}_+^{(t)} \cap \mathcal{S}_{-1}| &\geq (1-p)n - \sqrt{2n \log(8T^*/\delta)}.
\end{aligned}
$$

These hold with probability at least $1-\delta$. This suggests that when $p < C_0/6, n \geq 72 C_0^{-2} \log(8T^*/\delta)$, we have:

$$
\begin{aligned}
&|\mathcal{S}_+^{(t)} \cap \mathcal{S}_1| \geq \frac{2}{C_0} |\mathcal{S}_-^{(t)} \cap \mathcal{S}_1|, \quad |\mathcal{S}_+^{(t)} \cap \mathcal{S}_{-1}| \geq \frac{2}{C_0} |\mathcal{S}_-^{(t)} \cap \mathcal{S}_{-1}|, \\
&|\mathcal{S}_+^{(t)} \cap \mathcal{S}_1|, |\mathcal{S}_+^{(t)} \cap \mathcal{S}_{-1}| \geq \frac{n}{4}.
\end{aligned}
$$

Hence, we have:

$$
\begin{aligned}
\gamma_{j,r}^{(t+1)} &\geq \gamma_{j,r}^{(t)} + \frac{C_0 \eta \|\boldsymbol{\mu}\|_2^2}{4m} \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle = \gamma_{j,r}^{(t)} + \frac{C_0 \eta \|\boldsymbol{\mu}\|_2^2}{4m} \big(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle + j \gamma_{j,r}^{(t)} \big), \quad \text{if } \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle \geq 0 \\
\gamma_{j,r}^{(t+1)} &\geq \gamma_{j,r}^{(t)} - \frac{C_0 \eta \|\boldsymbol{\mu}\|_2^2}{4m} \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle = \gamma_{j,r}^{(t)} - \frac{C_0 \eta \|\boldsymbol{\mu}\|_2^2}{4m} \big(\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle + j \gamma_{j,r}^{(t)} \big), \quad \text{if } \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\mu} \rangle < 0.
\end{aligned}
$$

When $j = 1$, due to the increase of $\gamma_{j,r}^{(t)}$, we have

$$
\gamma_{1,r}^{(t+1)} \geq \gamma_{1,r}^{(t)} + \frac{C_0 \eta \|\boldsymbol{\mu}\|_2^2}{4m} \big(\langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\mu} \rangle + \gamma_{1,r}^{(t)} \big).
$$

Let $B_j^{(t)} = \max_{r \in [m]} \{ \gamma_{j,r}^{(t)} + j \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle \}$, then we have

$$
\begin{aligned}
B_1^{(t+1)} &\geq \big(1 + \frac{C_0 \eta \|\boldsymbol{\mu}\|_2^2}{4m} \big) B_1^{(t)} \geq \big(1 + \frac{C_0 \eta \|\boldsymbol{\mu}\|_2^2}{4m} \big)^t B_1^{(0)} \\
&\geq \exp \big(\frac{C_0 \eta \|\boldsymbol{\mu}\|_2^2}{8m} t \big) \max_r \langle \mathbf{w}_{1,r}^{(0)}, \boldsymbol{\mu} \rangle \\
&\geq \exp \big(\frac{C_0 \eta \|\boldsymbol{\mu}\|_2^2}{8m} t \big) \frac{\sigma_0 \|\boldsymbol{\mu}\|_2}{2},
\end{aligned}
$$

where we use the fact that $1 + x \geq \exp(x/2)$ for $x \leq 2$.

Similarly when $j = -1$, we have $\gamma_{-1,r}^{(t+1)} \geq \gamma_{-1,r}^{(t)} - \frac{C_0\eta\|\boldsymbol{\mu}\|_2^2}{4m}(\langle \mathbf{w}_{-1,r}^{(0)}, \boldsymbol{\mu}\rangle - \gamma_{-1,r}^{(t)})$ and

$$\begin{aligned}
B_{-1}^{(t+1)} &\geq \big(1 + \frac{C_0\eta\|\boldsymbol{\mu}\|_2^2}{4m}\big)B_{-1}^{(t)} \geq \big(1 + \frac{C_0\eta\|\boldsymbol{\mu}\|_2^2}{4m}\big)^t B_{-1}^{(0)} \\
&\geq \exp\big(\frac{C_0\eta\|\boldsymbol{\mu}\|_2^2}{8m}t\big)\max_r \langle -\mathbf{w}_{-1,r}^{(0)}, \boldsymbol{\mu}\rangle \\
&\geq \exp\big(\frac{C_0\eta\|\boldsymbol{\mu}\|_2^2}{8m}t\big)\frac{\sigma_0\|\boldsymbol{\mu}\|_2}{2}.
\end{aligned}$$

Thus, we obtain $B_j^{(t)} \geq \exp\big(\frac{C_0\eta\|\boldsymbol{\mu}\|_2^2}{8m}t\big)\frac{\sigma_0\|\boldsymbol{\mu}\|_2}{2}, \forall j \in \{\pm 1\}$. □

Lemma E.2. *Let* $\bar{\beta} = \min_{i\in[n]}\max_{r\in[m]}\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle$. *Suppose that* $\sigma_0 \geq 160n\sqrt{\frac{\log(4n^2/\delta)}{d}}(\sigma_p\sqrt{d})^{-1}\alpha d^{1/4}$ *. Then we have that* $\bar{\beta}/d^{1/4} \geq 40n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha$.

Proof of Lemma E.2. The proof follows from Lemma B.2. It is known that, with high probability, we have $\overline{\beta} \geq \sigma_0\sigma_p\sqrt{d}/4$. By substituting the condition for σ_0, we obtain

$$\overline{\beta}/d^{1/4} \geq 40n\sqrt{\frac{\log(4n^2/\delta)}{d}}\alpha.$$

□

Lemma E.3 (Lower bound on $\overline{\rho}_{j,r,i}^{(t)}$)**.** *Let* $\bar{\beta} = \min_{i\in[n]}\max_{r\in[m]}\langle \mathbf{w}_{y_i,r}^{(0)}, \boldsymbol{\xi}_i\rangle$ *and* $A_{y_i,r,i}^{(t)} := \overline{\rho}_{j,r,i}^{t} + \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - 0.4\bar{\beta}/d^{1/4}$. *Under Assumption 3.1, if* $\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle \geq \bar{\beta}$, *then at time step* $T_1 = \frac{nm\log(1/(\sigma_0\sigma_p\sqrt{d}))}{\eta\sigma_p^2 d}$, *with high probability, it holds that*

$$A_{y_i,r,i}^{(T_1)} \geq (1 + \frac{\eta C_0\sigma_p^2 d}{2nm})^{T_1}A_{y_i,r,i}^{(0)}.$$

Proof of Lemma E.3. First, consider $y_i = j$ as the case of $\overline{\rho}_{j,r,i}^{(t)}$. By Lemma C.1 and Lemma D.3, when $y_i = j$,

$$|\langle \mathbf{w}_{j,r,}^{(t)}, \boldsymbol{\xi}_i\rangle - \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - \overline{\rho}_{j,r,i}^{(t)}| \leq 16n\sqrt{\frac{\log(4n^2/\delta)}{d}} \leq 0.4\overline{\beta}/d^{1/4}. \tag{10}$$

From the update of $\overline{\rho}_{j,r,i}^{(t)}$, when $\epsilon_i^{(t)} = 1$ and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle > 0$,

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{2\eta}{nm}\ell_i'^{(t)}\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|_2^2\epsilon_i^{(t)} \geq \overline{\rho}_{j,r,i}^{(t)} + \frac{\eta C_0\sigma_p^2 d}{nm}\Big(\overline{\rho}_{j,r,i}^{(t)} + \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - 0.4\overline{\beta}/d^{1/4}\Big),$$

On the other hand, when $\epsilon_i^{(t)} = -1$ and $\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle > 0$,

$$\overline{\rho}_{j,r,i}^{(t+1)} = \overline{\rho}_{j,r,i}^{(t)} - \frac{2\eta}{nm}\ell_i'^{(t)}\langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i\rangle\|\boldsymbol{\xi}_i\|_2^2\epsilon_i^{(t)} \geq \overline{\rho}_{j,r,i}^{(t)} - \frac{3\eta\sigma_p^2 d}{nm}\Big(\overline{\rho}_{j,r,i}^{(t)} + \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle + 0.4\overline{\beta}/d^{1/4}\Big).$$

For simplification of notations, denote $\zeta = 0.8\overline{\beta}/d^{1/4}$. Let $A_{y_i,r,i}^{(t)} := \overline{\rho}_{j,r,i}^{t} + \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle - 0.4\overline{\beta}/d^{1/4}$. Then when $\epsilon_i^{(t)} = 1$, we have

$$A_{y_i,r,i}^{(t+1)} \geq (1 + \frac{\eta C_0\sigma_p^2 d}{nm})A_{y_i,r,i}^{(t)},$$

and when $\epsilon_i^{(t)} = -1$, we have

$$A_{y_i,r,i}^{(t+1)} \geq (1 - \frac{3\eta\sigma_p^2 d}{nm})A_{y_i,r,i}^{(t)} - \frac{3\eta\sigma_p^2 d\zeta}{nm}.$$

Here we prove when $\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i\rangle \geq \overline{\beta}$, $A_{y_i,r,i}^{(t)} > \zeta$. The proof is by the induction method.

First it is clear that $A^{(0)}_{y_i,r,i} = \langle \mathbf{w}^{(0)}_{j,r}, \boldsymbol{\xi}_i \rangle - 0.5\zeta > \zeta$ because $d \gg \Theta(1)$. Then we consider when $t \le \frac{2\log(4n/\delta)}{p^2}$ (where the condition for Lemma B.4 does not hold). In this case, $|\mathcal{S}^{(t)}_+| \ge (1-p)t - \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}, |\mathcal{S}^{(t)}_-| \le pt + \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}$. In addition, the worst case lower bound is achieved by the case where all the $\mathcal{S}^{(t)}_-$ events happen at the first few iterations. This gives

$$\begin{aligned}
A^{(t)}_{y_i,r,i} &\ge (1 + \frac{\eta C_0 \sigma_p^2 d}{nm})^{(1-p)t - \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}} (1 - \frac{3\eta\sigma_p^2 d}{nm})^{pt + \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}} A^{(0)}_{y_i,r,i} \\
&\qquad - (1 + \frac{\eta C_0 \sigma_p^2 d}{nm})^{(1-p)t - \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}} \Big[\sum_{s=0}^{pt + \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}} (1 - \frac{3\eta\sigma_p^2 d}{nm})^s \Big] \frac{\zeta\eta\sigma_p^2 d}{3nm} \\
&\ge (1 + \frac{\eta C_0 \sigma_p^2 d}{nm})^{(1-p)t - \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}} \big((1 - \frac{3\eta\sigma_p^2 d}{nm})^{pt + \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}} A^{(0)}_{y_i,r,i} - \zeta \big) \\
&\ge (1 + \frac{\eta C_0 \sigma_p^2 d}{nm})^{(1-p)t - \sqrt{\frac{t}{2}\log(\frac{4}{\delta})}} \zeta \ge \zeta,
\end{aligned}$$

where the last inequality follows from the fact that $d \gg \Theta(1)$. To see this, suppose there exists a $t \le \frac{2\log(4n/\delta)}{p^2}$ such that

$$(1 - \frac{3\eta\sigma_p^2 d}{nm})^{pt + \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}} A^{(0)}_{y_i,r,i} \le 2\zeta,$$

then we have

$$pt + \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})} \ge \frac{\log(d^{1/4}/2)}{\log\Big(\frac{1}{1 - \frac{3\eta\sigma_p^2 d}{nm}}\Big)},$$

while $t \le \frac{2\log(4/\delta)}{p^2}$ raises a contradiction by the choice of d. This proves for all $t \le \frac{2\log(4/\delta)}{p^2}$, we have $(1 - \frac{3\eta\sigma_p^2 d}{nm})^{pt + \sqrt{\frac{t}{2}\log(\frac{4n}{\delta})}} A^{(0)}_{y_i,r,i} \ge 2\zeta$ and thus $A^{(t)}_{y_i,r,i} \ge \zeta$.

Then we consider the case when $t \ge \frac{2\log(4/\delta)}{p^2}$ where the condition for Lemma B.4 holds. Now suppose for all $s \le t-1$, we have $A^{(s)}_{y_i,r,i} \ge \zeta$, which clearly holds for $t = \frac{2\log(4n/\delta)}{p^2}$. For all $s \le t-1$, we have $A^{(s)}_{y_i,r,i} \ge (1 - \frac{3\eta\sigma_p^2 d\zeta}{nm}) A^{(s)}_{y_i,r,i}$ when $\epsilon^{(s)}_i = -1$. This leads to the following lower bound for $A^{(t)}_{y_i,r,i}$ as

$$\begin{aligned}
A^{(t)}_{y_i,r,i} &\ge (1 + \frac{\eta C_0 \sigma_p^2 d}{nm})^{(1-1.5p)t} (1 - \frac{3\eta\sigma_p^2 d}{nm})^{1.5pt} A^{(0)}_{y_i,r,i} \\
&\ge (1 + \frac{\eta C_0 \sigma_p^2 d}{2nm})^t A^{(0)}_{y_i,r,i} \ge \zeta,
\end{aligned}$$

where the second last inequality follows from the choice of

$$p \le \frac{2}{3} \frac{\log(1 + \frac{\eta C_0 \sigma_p^2 d}{nm}) - \log(1 + \frac{\eta C_0 \sigma_p^2 d}{2nm})}{\log(1 + \frac{\eta C_0 \sigma_p^2 d}{nm}) - \log(1 - \frac{3\eta\sigma_p^2 d}{nm})}.$$

We can verify that $p = \frac{C_0}{24}$ satisfies the above inequality. This concludes the proof that, for all t, we have $A^{(t)}_{y_i,r,i} \ge \zeta$ and thus for all t. Finally, we conclude that

$$\begin{aligned}
A^{(t)}_{y_i,r,i} &\ge (1 + \frac{\eta C_0 \sigma_p^2 d}{nm})^{(1-1.5p)t} (1 - \frac{2\eta\sigma_p^2 d}{3nm})^{1.5pt} A^{(0)}_{y_i,r,i} \\
&\ge (1 + \frac{\eta C_0 \sigma_p^2 d}{2nm})^t A^{(0)}_{y_i,r,i}.
\end{aligned}$$

□

With the above lemmas at hand, we are ready to prove Lemma 4.4:

Proof of Lemma 4.4. By Lemma E.3, at $t = T_1$, taking the maximum over r yields

$$\begin{aligned}\max_r A_{y_i,r,i}^{(t)} &\geq (1 + \frac{\eta C_0 \sigma_p^2 d}{2nm})^t 0.6\bar{\beta} \\ &\geq (1 + \frac{\eta C_0 \sigma_p^2 d}{2nm})^t 0.15\sigma_0\sigma_p\sqrt{d} \\ &\geq \exp\big(\frac{\eta C_0 \sigma_p^2 d}{4nm} t\big) 0.15\sigma_0\sigma_p\sqrt{d},\end{aligned}$$

where the first inequality is by $\max_r \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle \geq \bar{\beta}$ and $0.4\bar{\beta}d^{-1/4} \leq 0.4\bar{\beta}$. In the last inequality, we use $(1+z) \geq \exp(z/2)$ for $z \leq 2$.

Then we see $\max_r A_{y_i,r,i}^{(t)} \geq 1$ in at least $T_1 = \frac{\log(20/(\sigma_0\sigma_p\sqrt{d}))4nm}{\eta C_0 \sigma_p^2 d}$ and because $\max_{j,r} \overline{\rho}_{j,r,i}^{T_1} \geq A_{y_i,r,i}^{T_1} - \max_{j,r} |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\xi}_i \rangle| + 0.4\bar{\beta} \geq 1$.

Besides, by Lemma D.2, we directly obtain the result that

$$|\underline{\rho}_{j,r,i}^{(T_1)}| \leq \frac{3\eta\sigma_p^2 d T_1}{nm} \sqrt{\log(8mn/\delta)}\sigma_0\sigma_p\sqrt{d} = \tilde{O}(\sigma_0\sigma_p\sqrt{d}).$$

Furthermore, Lemma D.1 yields

$$\gamma_{j,r}^{(T_1)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle| \leq \exp\big(\frac{2\eta\|\boldsymbol{\mu}\|_2^2}{m} \frac{4nm}{\eta\sigma_p^2 d} \log(1/(\sigma_0\sigma_p\sqrt{d}))\big)|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle| \leq 2|\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle|,$$

where we have used the condition of low SNR, namely $n\mathrm{SNR}^2 \leq 1/\log(20/(\sigma_0\sigma_p\sqrt{d}))$. By Lemma B.2, we conclude the proof for $\max_{j,r} \gamma_{j,r}^{(T_1)} = \tilde{O}(\sigma_0\|\boldsymbol{\mu}\|_2)$.

Lastly, according to Lemma E.1, at the end of stage1, we have the lower bound on signal learning coefficient

$$\begin{aligned}\max_{r\in[m]} \gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle| &\geq \exp\big(\frac{C_0\eta\|\boldsymbol{\mu}\|_2^2}{8m} t\big) \max_{r\in[m]} |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle| \\ &= \exp\big(\frac{C_0\eta\|\boldsymbol{\mu}\|_2^2}{8m} \frac{\log(20/(\sigma_0\sigma_p\sqrt{d}))4nm}{\eta C_0 \sigma_p^2 d}\big) \max_{r\in[m]} |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu} \rangle| \\ &\geq \exp(n\mathrm{SNR}^2 \log(20/(\sigma_0\sigma_p\sqrt{d})))\sigma_0\|\boldsymbol{\mu}\|_2 \geq \sigma_0\|\boldsymbol{\mu}\|_2.\end{aligned}$$

□

E.2 Proof of Lemma 4.5

The key idea is to show $\overline{\rho}_{j,r,i}^{(t)}$ oscillates during the second stage, where the growth tends to offset the drop over a given time frame. This would suggest the $f(\mathbf{W}^{(t)}, \mathbf{x})$ is both upper and lower bounded by a constant, which is crucial to ensuring that $\gamma_{j,r}^{(t)}$ increases exponentially during the second stage.

Without loss of generality, for each i with $\langle \mathbf{w}_{j,r,i}^{(t)}, \boldsymbol{\xi}_i \rangle > 0$ and $j = y_i = 1$, the evolution of $\overline{\rho}_{j,r,i}^{t+1}$ is written as

$$\begin{aligned}\overline{\rho}_{j,r,i}^{t+1} &= \overline{\rho}_{j,r,i}^{(t)} - \frac{2\eta}{nm} \tilde{\ell}_i^{\prime(t)} \langle \mathbf{w}_{j,r}^{(t)}, \boldsymbol{\xi}_i \rangle \|\boldsymbol{\xi}_i\|^2 \epsilon_i^{(t)} \\ &\approx \begin{cases} (1 + \frac{2\eta\|\boldsymbol{\xi}_i\|^2}{nm(1+\exp(f_i^{(t)}))})\overline{\rho}_{j,r,i}^{(t)}, & \text{if } \epsilon_i^{(t)} = 1 \\ (1 - \frac{2\eta\|\boldsymbol{\xi}_i\|^2}{nm(1+\exp(-f_i^{(t)}))})\overline{\rho}_{j,r,i}^{(t)} & \text{if } \epsilon_i^{(t)} = -1 \end{cases}\end{aligned}$$

where we denote $f_i^{(t)} = f(\mathbf{W}^{(t)}, \mathbf{x}_i)$. Note that $f_i^{(t)} \approx \frac{1}{m}\sum_{r=1}^m (\overline{\rho}_{+1,r,i}^{(t)})^2$ when $\gamma_{j,r}^{(t)} \ll 1$.

To simplify the notation, we define that $\iota_i^{(t)} \triangleq \frac{1}{m}\sum_{r=1}^m(\overline{\rho}_{+1,r,i}^{(t)})^2$. Then the dynamics can be approximated to

$$\iota_i^{(t+1)} \approx \begin{cases}(1+\frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{nm(1+\exp((\iota_i^{(t)})^2))})^2\iota_i^{(t)} & \text{with prob } 1-p\\ (1-\frac{2\eta\|\mathbf{x}_i\|_2^2}{nm(1+\exp(-(\iota_i^{(t)})^2))})^2\iota_i^{(t)} & \text{with prob } p\end{cases}$$

Lemma E.4 (Restatement of Lemma 4.5)**.** *Under the same condition as Theorem 3.3, during $t \in [T_1, T_2]$ with $T_2 = T_1 + \log(6/(\sigma_0\|\boldsymbol{\mu}\|_2))4m(1+\exp(c_2))\eta^{-1}\|\boldsymbol{\mu}\|_2^{-2}$, there exist a sufficient large positive constant C_ι and a constant ι_i^* depending on sample index i such that the following results hold with high probability at least $1-1/d$:*

- $|\iota_i^{(t)} - \iota_i^*| \le C_\iota$
- $\gamma_{j,r}^{(t)} \le 0.1$ *for all $j \in \{-1, 1\}$ and $r \in [m]$*
- $\frac{1}{2m}\sum_{r=1}^m(\overline{\rho}_{y_i,r,i}^{(t)})^2 \le f_i^{(t)} \le \frac{2}{m}\sum_{r=1}^m(\overline{\rho}_{y_i,r,i}^{(t)})^2$
- $\max_{r\in[m]}(\gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|) \ge \exp\big(\frac{\eta\|\boldsymbol{\mu}\|_2^2}{16m}(t-T_1)\big)\max_{r\in[m]}|\gamma_{j,r}^{(T_1)} + \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|$.

Proof of Lemma E.4. The proof is based on the method of induction. Without loss of generality, we consider all i with $y_i = 1$. We first check that at time step $t = T_1$, by Lemma 4.4, there exists a constant C such that

$$|\frac{1}{m}\sum_{r=1}^m \overline{\rho}_{+1,r,i}^{(T_1)} - \iota_i^*| \le C.$$

Besides, by Lemma 3.3, it is straightforward to check that $\gamma_{j,r}^{(T_1)} \le 1$ for all $j \in \{-1, 1\}$ and $r \in [m]$, and $\max_j \gamma_{j,r}^{(T_1)} \ge 0$. Next, we can show the following result at time $t = T_1$:

$$\begin{aligned}
f_i^{(T_1)} &= F_{+1}(\mathbf{W}_{+1}^{(T_1)}, \mathbf{x}_i) - F_{-1}(\mathbf{W}_{-1}^{(T_1)}, \mathbf{x}_i)\\
&= \frac{1}{m}\sum_{r=1}^m \sigma\big(\langle \mathbf{w}_{+1,r}^{(0)}, \boldsymbol{\mu}\rangle + \gamma_{+1,r}^{(T_1)}\big) + \frac{1}{m}\sum_{r=1}^m \sigma\Big(\langle \mathbf{w}_{+1,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \overline{\rho}_{+1,r,i}^{(T_1)} + \sum_{i'\ne i}\frac{\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2}\rho_{+1,r,i'}^{(T_1)}\Big)\\
&\quad - \frac{1}{m}\sum_{r=1}^m \sigma\big(\langle \mathbf{w}_{-1,r}^{(0)}, \boldsymbol{\mu}\rangle - \gamma_{-1,r}^{(T_1)}\big) - \frac{1}{m}\sum_{r=1}^m \sigma\Big(\langle \mathbf{w}_{-1,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \underline{\rho}_{-1,r,i}^{(T_1)} + \sum_{i'\ne i}\frac{\langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2}\rho_{-1,r,i'}^{(T_1)}\Big)\\
&\ge -\tilde{\Omega}(\sigma_0^2\|\boldsymbol{\mu}\|_2^2) - \tilde{\Omega}(\sigma_0\sigma_p\sqrt{d}) + \frac{1}{m}\sum_{r=1}^m(\overline{\rho}_{+1,r,i}^{T_1)} - \beta - 16\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha)^2\\
&\ge \frac{1}{2m}\sum_{r=1}^m(\overline{\rho}_{+1,r,i}^{(T_1)})^2,
\end{aligned}$$

where the first inequality is by Lemma 4.5, Proposition 4.1, and Lemma B.1, The second inequality follows from the condition on σ_0 and d in Assumption 3.1. Similarly, we have

$$\begin{aligned}
f_i^{(T_1)} &= F_{+1}(\mathbf{W}_{+1}^{(T_1)}, \mathbf{x}_i) - F_{-1}(\mathbf{W}_{-1}^{(T_1)}, \mathbf{x}_i)\\
&\le \tilde{O}(\sigma_0^2\|\boldsymbol{\mu}\|_2^2) + \tilde{O}(\sigma_0\sigma_p\sqrt{d}) + \frac{1}{m}\sum_{r=1}^m(\overline{\rho}_{+1,r,i}^{(T_1)} + \beta + 16\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha)^2\\
&\le \frac{2}{m}\sum_{r=1}^m(\overline{\rho}_{+1,r,i}^{(T_1)})^2.
\end{aligned}$$

Next, we assume that all the results hold for $T_1 < t \le T$. By the induction hypothesis, we can bound $c_1 \le f_i^{(T)} \le c_2$ for all $i \in [n]$. Then we can show that $\gamma_{j,r}^{(T+1)}$ continues to exhibit exponential growth:

$$\begin{aligned}
\gamma_{j,r}^{(T+1)} &= \gamma_{j,r}^{(T)} - \frac{2\eta}{nm}\Big(\sum_{i\in\mathcal{S}_+^{(T)}\cap\mathcal{S}_1} \tilde{\ell}_i'^{(t)} - \sum_{i\in\mathcal{S}_-^{(T)}\cap\mathcal{S}_1} \tilde{\ell}_i'^{(t)}\Big)\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\mu}\rangle \|\boldsymbol{\mu}\|_2^2 \\
&\geq \gamma_{j,r}^{(T)} + \frac{2\eta}{nm}\big(|\mathcal{S}_+^{(T)}|\frac{1}{1+\exp(c_2)} - |\mathcal{S}_-^{(T)}|\frac{1}{1+\exp(-c_2)}\big)\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\mu}\rangle \|\boldsymbol{\mu}\|_2^2 \\
&\geq \gamma_{j,r}^{(T)} + \frac{2\eta}{m}\big(\frac{2-3p}{4}\frac{1}{1+\exp(c_2)} - \frac{3p}{4}\frac{1}{1+\exp(-c_2)}\big)\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\mu}\rangle \|\boldsymbol{\mu}\|_2^2 \\
&= \gamma_{j,r}^{(T)} + \frac{\eta}{m}\big(\frac{1}{1+\exp(c_2)} - \frac{3p}{2}\big)(\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\mu}\rangle + j\gamma_{j,r}^{(T)}) \|\boldsymbol{\mu}\|_2^2 \\
&\geq \gamma_{j,r}^{(T)} + \frac{\eta\|\boldsymbol{\mu}\|_2^2}{2m(1+\exp(c_2))}(\langle \mathbf{w}_{j,r}^{(T)}, \boldsymbol{\mu}\rangle + j\gamma_{j,r}^{(T)}),
\end{aligned}$$

where the last inequality is by $\frac{3}{2}p \leq \frac{1}{2}\frac{1}{1+\exp(c_2)}$. Next, define $B^{(t)} = \max_{r\in[m]}(\gamma_{j,r}^{(t)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle|)$, we have:

$$\begin{aligned}
B^{(T+1)} &\geq B^{(T)}(1 + \frac{\eta\|\boldsymbol{\mu}\|_2^2}{2m(1+\exp(c_2))}) \\
&\geq \exp(\frac{\eta\|\boldsymbol{\mu}\|_2^2}{4m(1+\exp(c_2))}(t-T_1))B^{(T_1)} \\
&\geq \exp\big(\frac{\eta\|\boldsymbol{\mu}\|_2^2}{16m}(t-T_1)\big)B^{(T_1)}.
\end{aligned}$$

At the same time, there exists an upper bound on the signal learning:

$$\gamma_{j,r}^{(T)} + |\langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \leq \exp\big(\frac{2\eta\|\boldsymbol{\mu}\|_2^2}{m}(T-T_1)\big)|\gamma_{j,r}^{(T_1)} + \langle \mathbf{w}_{j,r}^{(0)}, \boldsymbol{\mu}\rangle| \leq 0.01,$$

where we used the condition that $T < T_2$.

To show that $\iota_i^{(T+1)}$ remains within a constant range, we define $M_i^{(t)} \triangleq (\iota_i^{(t)} - \iota_i^*)^2$ where ι_i^* is a sufficiently large constant depending on i. Using the relation $\frac{1}{2m}\sum_{r=1}^m (\overline{\rho}_{y_i,r,i}^{(T)})^2 \leq f_i^{(T)} \leq \frac{2}{m}\sum_{r=1}^m (\overline{\rho}_{y_i,r,i}^{(T)})^2$ we have:

$$\begin{aligned}
\mathbb{E}[\iota_i^{(T+1)}|\iota_i^{(T)}] \geq{}& (1-p)\Big(1 + \frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{(1+2\exp((\iota_i^{(T)})^2))nm}\Big)^2\iota_i^{(T)} \\
&+ p\Big(1 - \frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{(1+1/2\exp(-(\iota_i^{(T)})^2))nm}\Big)^2\iota_i^{(T)}.
\end{aligned}$$

At the same time,

$$\begin{aligned}
\mathbb{E}[(\iota_i^{(T+1)})^2|\iota_i^{(T)}] \leq{}& (1-p)\Big(1 + \frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{(1+1/2\exp((\iota_i^{(T)})^2))nm}\Big)^4(\iota_i^{(T)})^2 \\
&+ p\Big(1 - \frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{(1+2\exp(-(\iota_i^{(T)})^2))nm}\Big)^4(\iota_i^{(T)})^2.
\end{aligned}$$

Then we show that

$$\begin{aligned}
\mathbb{E}[M_i^{(T+1)}|\iota_i^{(T)}] &= \mathbb{E}[(\iota_i^{(T+1)})^2|\iota_i^{(T)}] - 2\iota^*\mathbb{E}[\iota_i^{(T+1)}|\iota_i^{(T)}] + (\iota^*)^2 \\
&\leq (1-p)\Big(1 + \frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{(1+1/2\exp((\iota_i^{(T)})^2))nm}\Big)^4(\iota_i^{(T)})^2 \\
&\quad + p\Big(1 - \frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{(1+2\exp(-(\iota_i^{(T)})^2))nm}\Big)^4(\iota_i^{(T)})^2 \\
&\quad - 2\iota^*\Bigg((1-p)\Big(1 + \frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{(1+2\exp((\iota_i^{(T)})^2))nm}\Big)^2\iota_i^{(T)} \\
&\quad + p\Big(1 - \frac{2\eta\|\boldsymbol{\xi}_i\|_2^2}{(1+1/2\exp(-(\iota_i^{(T)})^2))nm}\Big)^2\iota_i^{(T)}\Bigg) + (\iota^*)^2.
\end{aligned}$$

Subtracting $M_i^{(T)}$ yields

$$
\begin{aligned}
&\mathbb{E}[M_i^{(T+1)}|\iota_i^{(T)}] - M_i^{(T)} \\
&\leq (1-p)\left[\frac{8\eta\|\boldsymbol{\xi}_i\|_2^2}{(1+1/2\exp((\iota_i^{(T)})^2))nm} + O\left((\frac{\eta\|\boldsymbol{\xi}_i\|_2}{nm})^2\right)\right](\iota_i^{(T)})^2 \\
&\quad + p\left[-\frac{8\eta\|\boldsymbol{\xi}_i\|_2^2}{(1+2\exp(-(\iota_i^{(T)})^2))nm} + O\left((\frac{\eta\|\boldsymbol{\xi}_i\|_2}{nm})^2\right)\right](\iota_i^{(T)})^2 \\
&\quad - \frac{8\eta\|\boldsymbol{\xi}_i\|^2}{nm}\Big(\frac{1}{1+2\exp((\iota_i^{(T)})^2)} - p\Big)\iota_i^{(T)}\iota^* + O\left((\frac{\eta\|\boldsymbol{\xi}_i\|_2}{nm})^2\right) \\
&= \frac{8\eta\|\boldsymbol{\xi}_i\|_2^2\iota_i^{(T)}}{nm}\left[\frac{1-p(1+1/2\exp((\iota_i^{(T)})^2))}{1+1/2\exp((\iota_i^{(T)})^2)}\iota_i^{(T)} - \frac{1-p(1+2\exp((\iota_i^{(T)})^2))}{1+2\exp((\iota_i^{(T)})^2)}\iota^*\right] + O\left((\frac{\eta\|\boldsymbol{\xi}_i\|_2}{nm})^2\right) \\
&\leq 0,
\end{aligned}
$$

where the final inequality is by $\iota_i^{(T)} \leq 4\iota^*$ and $p < 1/(1+2\exp((\iota_i^{(T)})^2))$ and condition the learning rate from Assumption 3.1, which confirms that $\{M_i^{(t)}\}_{t\in[T_1,T]}$ is a super martingale. By one-sided Azuma inequality, with probability at least $1-\delta$, for any $\tau > 0$, it holds that

$$
P(M_i^{(T)} - M_i^{(T_1)} \geq \tau) \leq \exp\left(-\frac{\tau^2}{\sum_{k=T_1}^{T} c_k^2}\right),
$$

where,

$$
\begin{aligned}
c_k &= |M_i^{(k)} - M_i^{(k-1)}| = |(\iota_i^{(k)} - \iota^*)^2 - (\iota_i^{(k-1)} - \iota^*)^2| \\
&= |(\iota_i^{(k)} - \iota_i^{(k-1)})(\iota_i^{(k)} + \iota_i^{(k-1)} - 2\iota_i^*)| \leq \eta C_2.
\end{aligned}
$$

Taking the upper bound of $c_k \leq \eta C_2$ yields

$$
P((\iota_i^{(T+1)} - \iota_i^*)^2 - C_0^2 \geq \tau) \leq \exp\left(-\frac{\tau^2}{(T+1-T_1)\eta^2 C_2^2}\right),
$$

where we define $C_0^2 \triangleq (\iota_i^{(T)} - \iota_i^*)^2 > 0$. Therefore, we conclude with probability at least $1-\delta$,

$$
|\iota_i^{(T+1)} - \iota_i^*| \leq \sqrt{C_0^2 + \sqrt{\eta^2 t C_2^2 \log(1/\delta)}} \leq C_\iota,
$$

where the last inequality is by $\eta \leq \tilde{O}(\sigma_p^{-2}d^{-1})$ and $T < T_2$.

Finally, we check that

$$
\begin{aligned}
f_i^{(T+1)} &= F_{+1}(\mathbf{W}_{+1}^{(T+1)}, \mathbf{x}_i) - F_{-1}(\mathbf{W}_{-1}^{(T+1)}, \mathbf{x}_i) \\
&= \frac{1}{m}\sum_{r=1}^{m}\sigma\big(\langle\mathbf{w}_{+1,r}^{(0)}, \boldsymbol{\mu}\rangle + \gamma_{+1,r}^{(T+1)}\big) + \frac{1}{m}\sum_{r=1}^{m}\sigma\Big(\langle\mathbf{w}_{+1,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \overline{\rho}_{+1,r,i}^{(T+1)} + \sum_{i'\neq i}\frac{\langle\boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2}\rho_{+1,r,i'}^{(T+1)}\Big) \\
&\quad - \frac{1}{m}\sum_{r=1}^{m}\sigma\big(\langle\mathbf{w}_{-1,r}^{(0)}, \boldsymbol{\mu}\rangle - \gamma_{-1,r}^{(T+1)}\big) - \frac{1}{m}\sum_{r=1}^{m}\sigma\Big(\langle\mathbf{w}_{-1,r}^{(0)}, \boldsymbol{\xi}_i\rangle + \underline{\rho}_{-1,r,i}^{(T+1)} + \sum_{i'\neq i}\frac{\langle\boldsymbol{\xi}_i, \boldsymbol{\xi}_{i'}\rangle}{\|\boldsymbol{\xi}_{i'}\|_2^2}\rho_{-1,r,i'}^{(T+1)}\Big) \\
&\geq -\tilde{\Omega}(\sigma_0^2\|\boldsymbol{\mu}\|_2^2) - 0.01 + \frac{1}{m}\sum_{r=1}^{m}(\overline{\rho}_{+1,r,i}^{(T+1)} - \beta - 16\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha)^2 \\
&\geq \frac{1}{2m}\sum_{r=1}^{m}(\overline{\rho}_{+1,r,i}^{(T+1)})^2,
\end{aligned}
$$

where the first inequality is by Lemma 4.4 and the induction claim, and the second inequality is by condition on d from Assumption 3.1. Similarly, by the same argument, we conclude that:

$$\begin{aligned}
f_i^{(T+1)} &= F_{+1}(\mathbf{W}_{+1}^{(t)}, \mathbf{x}_i) - F_{-1}(\mathbf{W}_{-1}^{(t)}, \mathbf{x}_i) \\
&\le \tilde{O}(\sigma_0^2\|\boldsymbol{\mu}\|_2^2) + 0.01 + \tilde{O}(\sigma_0\sigma_p\sqrt{d}) + \frac{1}{m}\sum_{r=1}^{m}(\overline{\rho}_{+1,r,i}^{(t)} + \beta + 16\sqrt{\frac{\log(4n^2/\delta)}{d}}n\alpha)^2 \\
&\le 2\frac{1}{m}\sum_{r=1}^{m}(\overline{\rho}_{+1,r,i}^{(t)})^2.
\end{aligned}$$

Let $T_2 = T_1 + \log(6/(\sigma_0\|\boldsymbol{\mu}\|_2))4m(1+\exp(c_2))\eta^{-1}\|\boldsymbol{\mu}\|_2^{-2}$, then by lemma 4.4 we can show that

$$\begin{aligned}
\gamma_{j,r}^{(T_2)} &\ge \exp(\frac{\eta\|\boldsymbol{\mu}\|_2^2}{4m(1+\exp(c_2))}t)\gamma_{j,r}^{(T_1)} \\
&= \exp(\frac{\eta\|\boldsymbol{\mu}\|_2^2}{4m(1+\exp(c_2))}\log(6/(\sigma_0\|\boldsymbol{\mu}\|_2))4m(1+\exp(c_2))\eta^{-1}\|\boldsymbol{\mu}\|_2^{-2})\gamma_{j,r}^{(T_1)} \\
&= C_0/(\sigma_0\|\boldsymbol{\mu}\|_2)\gamma_{j,r}^{(T_1)} \\
&\ge 0.01.
\end{aligned}$$

□

E.3 Proof of Theorem 3.3

Proof of Theorem 3.3. For the population loss, we expand the expression as follows:

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y \ne f(\mathbf{W}^{(t)}, \mathbf{x}))] = \mathbb{P}(yf(\mathbf{W}^{(t)}, \mathbf{x}) < 0) \\
&= \mathbb{P}\Big(\frac{1}{m}\sum_{r=1}^{m}\sigma(\langle\mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\rangle) - \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle\mathbf{w}_{y,r}^{(t)}, \boldsymbol{\xi}\rangle) \ge \\
&\qquad \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle\mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}\rangle) - \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle\mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu}\rangle)\Big).
\end{aligned}$$

Recall the weight decomposing

$$\mathbf{w}_{j,r}^{(t)} = \mathbf{w}_{j,r}^{(0)} + j\gamma_{j,r}^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\boldsymbol{\mu} + \sum_{i=1}^{n}\overline{\rho}_{j,r,i}^{(t)}\|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i + \sum_{i=1}^{n}\underline{\rho}_{j,r,i}^{(t)}\|\boldsymbol{\xi}_i\|_2^{-2}\boldsymbol{\xi}_i.$$

From this, we obtain:

$$\begin{aligned}
\langle\mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu}\rangle &= \langle\mathbf{w}_{-y,r}^{(0)}, y\boldsymbol{\mu}\rangle - \gamma_{-y,r}^{(t)}, \\
\langle\mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}\rangle &= \langle\mathbf{w}_{y,r}^{(0)}, y\boldsymbol{\mu}\rangle + \gamma_{y,r}^{(t)}.
\end{aligned}$$

By Lemma 4.5, we conclude that

$$\langle\mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}\rangle = \Theta(1), \quad \langle\mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu}\rangle = -\Theta(\gamma_{y,r}^{(t)}) < 0.$$

Therefore, it holds that

$$\begin{aligned}
&\frac{1}{m}\sum_{r=1}^{m}\sigma(\langle\mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}\rangle) - \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle\mathbf{w}_{-y,r}^{(t)}, y\boldsymbol{\mu}\rangle) \\
&= \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle\mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu}\rangle) \\
&= \frac{1}{m}\sum_{r=1}^{m}\sigma(\langle\mathbf{w}_{y,r}^{(0)}, y\boldsymbol{\mu}\rangle + \gamma_{y,r}^{(t)}) \\
&= \Theta(1),
\end{aligned}$$

where the last inequity is by Lemma 4.5.

Next, we provide the bound for the noise memorization part. Define that $g(\boldsymbol{\xi}) = \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\rangle)$. By Theorem 5.2.2 in [52], for any $\tau > 0$, it holds

$$\mathbb{P}(g(\boldsymbol{\xi}) - \mathbb{E}[g(\boldsymbol{\xi}] \geq \tau) \leq \exp(-\frac{c\tau^2}{\sigma_p^2 \|g\|_{\mathrm{Lip}}^2}),$$

where c is a constant and $\|g\|_{\mathrm{Lip}}$ is the Lipschitz norm of function $g(\boldsymbol{\xi})$, which can be calculated as follows:

$$\begin{aligned}
|g(\boldsymbol{\xi}_i) - g(\boldsymbol{\xi}')| &= |\sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\rangle) - \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}'\rangle)| \\
&\leq \sum_{r=1}^{m} |\sigma(\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\rangle) - \sigma(\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}'\rangle)| \\
&\leq 2\sum_{r=1}^{m} |\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\rangle| \cdot |\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi} - \boldsymbol{\xi}'\rangle| \\
&\leq 2\sum_{r=1}^{m} \|\mathbf{w}_{-y,r}^{(t)}\|_2^2 \cdot \|\boldsymbol{\xi}\|_2 \cdot \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2 \\
&\leq 3\sum_{r=1}^{m} \|\mathbf{w}_{-y,r}^{(t)}\|_2^2 \sigma_p \sqrt{d} \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2,
\end{aligned}$$

where the first inequality is by the triangle inequality, the second inequality follows from the the convexity of the activation function, the third inequality is by the Cauchy-Schwarz inequality, and the last inequality follows from B.1. Therefore we conclude that

$$\|g\|_{\mathrm{Lip}} \leq 3\sum_{r=1}^{m} \|\mathbf{w}_{-y,r}^{(t)}\|_2^2 \sigma_p \sqrt{d}.$$

Furthermore, given that $\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}\rangle \sim \mathcal{N}(0, \sigma_p^2 \|\mathbf{w}_{-y,r}^{(t)}\|_2^2)$ we have:

$$\mathbb{E}[g(\boldsymbol{\xi})] = \sum_{r=1}^{m} \mathbb{E}[\sigma(\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi}'\rangle)] = \sum_{r=1}^{m} \sigma_p^2/2 \|\mathbf{w}_{-y,r}^{(t)}\|_2^2.$$

To obtain the the upper bound of $g(\boldsymbol{\xi})$, we show that:

$$\begin{aligned}
\|\mathbf{w}_{-y,r}^{(t)}\|_2^2 &= \left\| \sum_{i=1}^{n} \rho_{j,r,i}^{(t)} \|\boldsymbol{\xi}_i\|_2^{-2} \boldsymbol{\xi}_i \right\|_2^2 \\
&= \sum_{i=1}^{n} (\rho_{j,r,i}^{(t)})^2 \|\boldsymbol{\xi}_i\|_2^{-2} \boldsymbol{\xi}_i + 2\sum_{i=1}^{n} \sum_{j \neq i} \rho_{j,r,i}^{(t)} \rho_{j,r,j}^{(t)} \|\boldsymbol{\xi}_i\|_2^{-2} \|\boldsymbol{\xi}_j\|_2^{-2} \langle \boldsymbol{\xi}_i, \boldsymbol{\xi}_j\rangle \\
&\leq 3nC(\sigma_p^2 d)^{-1} + 2n^2(\sigma_p^2 d)^{-2} \sigma_p^2 \sqrt{d \log(4n^2/\delta)} \\
&\leq 4nC(\sigma_p^2 d)^{-1},
\end{aligned}$$

where the first inequality is by Lemma B.1, and the second inequality is by the condition on d in Assumption 3.1. With the results above, we conclude that

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}^{0-1}(\mathbf{W}^{(t)}) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[y \neq f(\mathbf{W}^{(t)}, \mathbf{x}))] = \mathbb{P}(yf(\mathbf{W}^{(t)}, \mathbf{x}) < 0) \\
&\leq \mathbb{P}(\sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{-y,r}^{(t)}, \boldsymbol{\xi} \rangle) \geq \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu} \rangle)) \\
&= \mathbb{P}(g(\boldsymbol{\xi}) - \mathbb{E}[g(\boldsymbol{\xi})] \geq \sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu} \rangle) - \sum_{r=1}^{m} \sigma_p^2/2 \|\mathbf{w}_{-y,r}^{(t)}\|_2^2) \\
&\leq \exp\left(-\frac{c(\sum_{r=1}^{m} \sigma(\langle \mathbf{w}_{y,r}^{(t)}, y\boldsymbol{\mu} \rangle) - \sum_{r=1}^{m} \sigma_p^2/2 \|\mathbf{w}_{-y,r}^{(t)}\|_2^2)^2}{\sigma_p^2 (3\sum_{r=1}^{m} \|\mathbf{w}_{-y,r}^{(t)}\|_2^2 \sigma_p \sqrt{d})^2} \right) \\
&\leq \exp\left(-\left(\frac{C_1 - \sigma_p^2/2 \cdot 4nC(\sigma_p^2 d)^{-1}}{3\sigma_p^2 \sqrt{d} 4nC(\sigma_p^2 d)^{-1}} \right)^2 \right) \\
&\leq \exp(\frac{1}{36d}) \exp(-\frac{C_1^2 d}{12^2 n^2 C^2}) \\
&\leq 2\exp\left(-\frac{C_1^2 d}{12^2 n^2 C^2} \right),
\end{aligned}$$

which corresponds to the second bullet point of Theorem 3.3. Combined with Lemma 4.5, which establishes the first bullet point, this completes the proof of Theorem 3.3. □

F Experimental Details for Figure 1

In this section, we provide a detailed description of the experimental setup used to generate the results shown in Figure 1, which compares the performance of Label Noise GD and Standard GD on the CIFAR-10 dataset under varying SNR conditions.

F.1 Dataset and Noise Injection

We used the CIFAR-10 dataset, selecting 1,000 images in total, with 100 images per class, to perform both Standard GD and Label Noise GD training. The random seed used for selecting training samples was fixed to ensure a fair comparison across different hyperparameters.

To simulate varying SNR conditions, inspired by [19], we introduced noise to the high-frequency Fourier components of the images using the following procedure:

- Each image was transformed into the frequency domain using a 2D Fourier transform.
- Gaussian noise was added to the high-frequency components, excluding the low-frequency region near the center of the Fourier spectrum. The intensity of the noise was controlled by a *noise level* parameter, where higher values correspond to noisier data and lower SNR.
- Finally, the image was transformed back into the spatial domain using an inverse Fourier transform.

The noise level was adjusted to control the SNR factor, which is represented on the x-axis in Figure 1.

F.2 Model and Training Setup

The experiments were conducted using a VGG-16 model trained from scratch on the CIFAR-10 dataset. The final fully connected layer of the model was modified to output predictions for the 10 classes in CIFAR-10. Both Standard GD and Label Noise GD were trained using cross-entropy loss and gradient descent (GD) with a learning rate of 0.05. Training was performed with a full-batch setup over 5,000 epochs. For Label Noise GD, labels were flipped randomly with a probability of 20% at each iteration to simulate label noise.

F.3 Results Analysis

The results shown in Figure 1 demonstrate that Label Noise GD consistently achieves higher test accuracy than Standard GD across all SNR levels. The performance gap is most evident under low SNR conditions, where Standard GD suffers significant accuracy degradation due to noise memorization, while Label Noise GD effectively suppresses noise and promotes robust feature learning.

F.4 Reproducibility

To ensure reproducibility, all experiments were implemented in PyTorch. The codebase, including dataset preprocessing, model training, and evaluation, is provided in the supplementary material.

G Additional Experiments

In this section, we provide additional experiments to further support our theoretical findings.

G.1 Deeper Neural Network

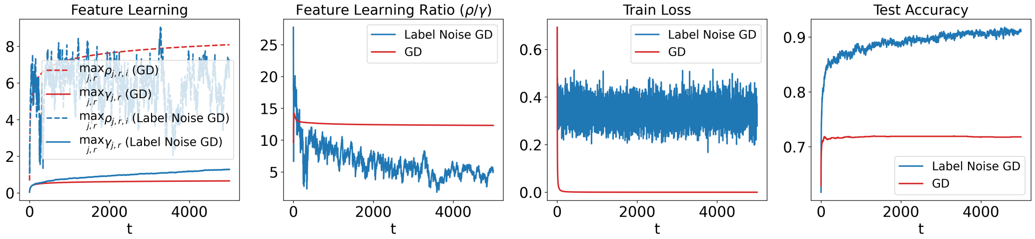


Figure 4: Performance of a 3-layer ReLU neural network: The ratio of noise memorization to signal learning, along with training loss and test accuracy, for standard GD and label noise GD.

We have conducted additional experiments using a 3-layer neural network with ReLU activation. The network is defined as $f(\mathbf{W}, \mathbf{x}) = F_{+1}(\mathbf{W}_{+1}, \mathbf{W}, \mathbf{x}) - F_{-1}(\mathbf{W}_{-1}, \mathbf{W}, \mathbf{x})$, where

$$F_j(\mathbf{W}_j, \mathbf{W}, \mathbf{x}) = \frac{1}{m} \sum_{r=1}^{m} \sum_{p=1}^{2} \sigma\big(\langle \mathbf{w}_{j,r}, \mathbf{z}^{(p)} \rangle\big), \quad \mathbf{z}^{(p)} = \sigma(\mathbf{W}^\top \mathbf{x}^{(p)}),$$

in which $\sigma(\cdot)$ is the ReLU activation, $\mathbf{W} \in \mathbb{R}^{d\times m}$ denotes the weight in the first layer, and $\mathbf{W}_{\pm 1} \in \mathbb{R}^{m\times m}$ are weights in the second layer. The last layer is fixed.

Specifically, we train the first two layers. The number of training samples is $n = 200$, and the number of test samples is $n_{\text{test}} = 2000$. The input dimension was set to $d = 2000$. We set the width to $m = 20$, the learning rate to $\eta = 0.5$, and the noise flip rate to $p = 0.1$. The data model follows our theoretical setting, where $\boldsymbol{\mu} = [1, 0, 0, \cdots, 0]$ and the noise strength is $\sigma_p = 1$. The experimental results, shown in Figure 4, are consistent with our original findings: compared to standard gradient descent, label noise GD boosts signal learning (as shown in the first plot) and achieves better generalization (as shown in the last plot).

G.2 Real World Dataset

We conducted an experiment using the MNIST dataset, in which Gaussian noise was added to the borders of the images while retaining the digits in the middle. The noise level was set to $\sigma_p = 5$. Moreover, the original pixel values of the digits ranged from 0 to 255, and we chose a normalization factor of 80. In this setup, the added noise formed a “noise patch" and the digits formed a “signal patch". We focused on the digits ‘0’ and ‘1’, using $n = 100$ samples for training and 200 samples for testing. The learning rate was set to $\eta = 0.001$, and the width was set to $m = 20$, with a label noise level of $p = 0.15$. The results, shown in Figure 5, were consistent with our theoretical conclusions, reinforcing the insights derived from our analysis.

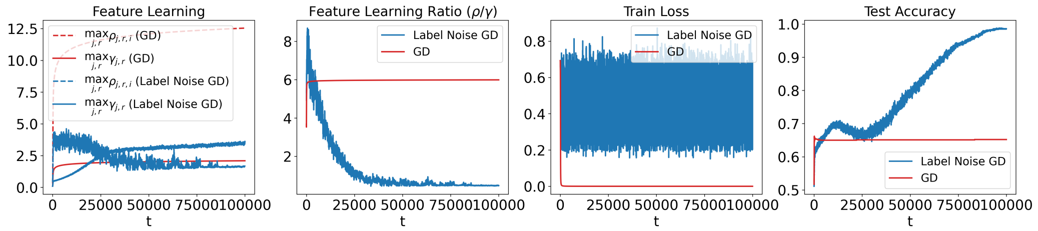
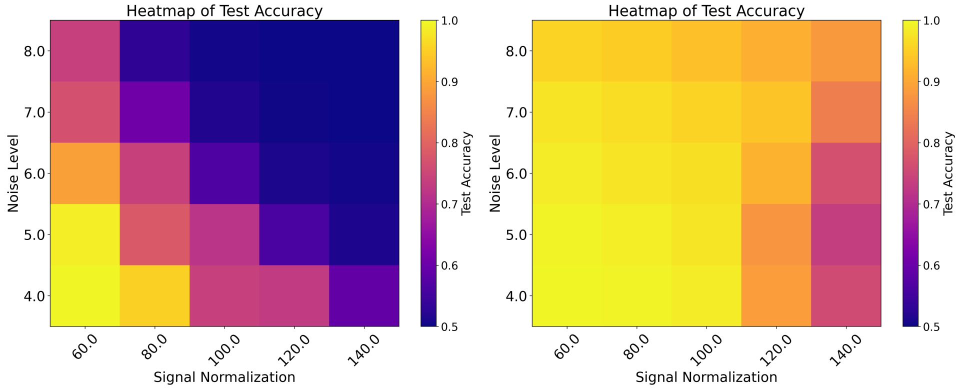


Figure 5: Performance on the modified MNIST dataset: The ratio of noise memorization to signal learning, along with training loss and test accuracy, for standard GD and label noise GD.



(a) Performance of standard GD

(b) Performance of Label Noise GD

Figure 6: Test accuracy heatmap of standard GD (left) and Label Noise GD (right) after training on modified MNIST dataset.

To assess the sensitivity of the methods to the choice of noise parameters and signal normalization, we conducted additional experiments on a modified MNIST dataset. The signal normalization values were varied from 60 to 140, while the noise levels ranged from 4 to 8. For each combination of noise level and signal normalization, we trained the neural network for 200,000 steps with a learning rate $\eta = 0.001$, using either standard gradient descent (GD) or label noise GD.

The resulting test errors are visualized in Figure 6. Notably, label noise GD (right) consistently achieves higher test accuracy than standard GD (left) across all configurations. This demonstrates the robustness of label noise GD to variations in noise and signal normalization parameters.

The motivation behind using MNIST was its clearer signal, which allows us to more directly observe the effects of label noise without other confounding factors. However, we also conducted experiments on a subset of CIFAR-10, using two classes: *airplane* and *automobile*. Gaussian noise was added to a portion of the images, following a similar setup to MNIST. For these experiments, we set $q = 2$, the number of neurons $m = 20$, the learning rate $\eta = 0.001$, the signal norm signal_norm $= 64$, the noise level noise_level $= 5$, the number of samples $n = 100$, the label noise probability $p = 0.15$, and the input dimension $d = 6144$.

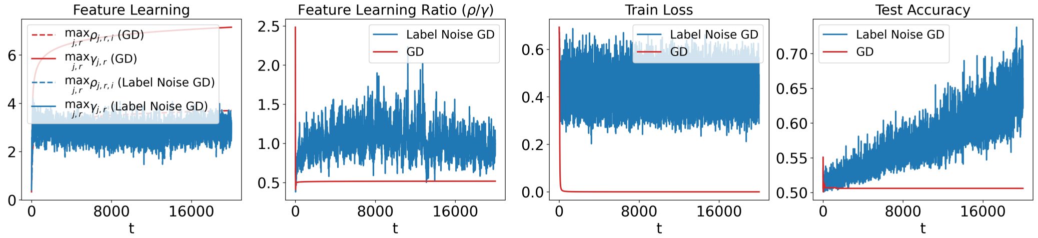


Figure 7: Performance on the modified CIFAR-10 dataset: The ratio of noise memorization to signal learning, along with training loss and test accuracy, for standard GD and label noise GD.

The results shown in Figure 7 indicate that label noise GD continues to provide benefits in terms of generalization compared to standard GD. We believe these extended experiments help establish a broader applicability of our findings to more complex benchmarks.

G.3 Different Type of Label Noise

To validate the robustness of label noise GD under different noise forms, we varied p across different values. For example, we show the results for $p = 0.3$ in Figure 8 and $p = 0.4$ in Figure 9. The results consistently indicate that label noise helps reduce overfitting and boost generalization, especially in low SNR settings.

In addition, we extended our empirical analysis to include Gaussian noise and uniform distribution noise added to the labels. For Gaussian noise, we used two examples, namely $\epsilon_i^{(t)} \sim \mathcal{N}(1, 1)$ and $\epsilon_i^{(t)} \sim \mathcal{N}(1, 1)$, with the results shown in Figures 10 and 11, respectively. Furthermore, for the uniform distribution, we simulated the noise with $\epsilon_i^{(t)} \sim \text{unif}[-1, 2]$ and $\epsilon_i^{(t)} \sim \text{unif}[-2, 3]$. The results are shown in Figures 12 and 13, respectively.

Our results indicate that label noise GD still performs effectively, achieving better generalization compared to standard GD, providing further evidence of the robustness of label noise GD under different noise forms.

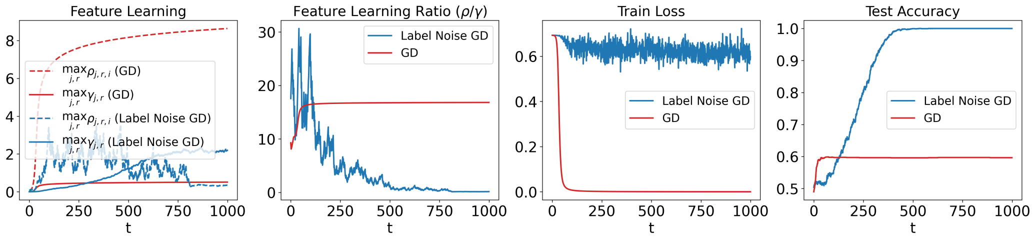


Figure 8: Performance with flip noise $p = 0.3$: The ratio of noise memorization to signal learning, training loss, and test accuracy of standard GD and label noise GD.

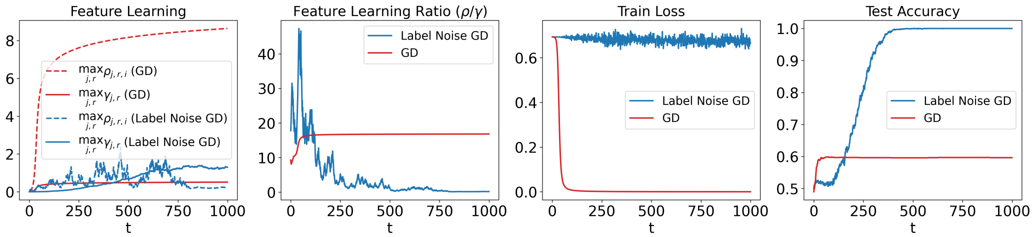


Figure 9: Performance with flip noise $p = 0.4$: The ratio of noise memorization to signal learning, training loss, and test accuracy of standard GD and label noise GD.

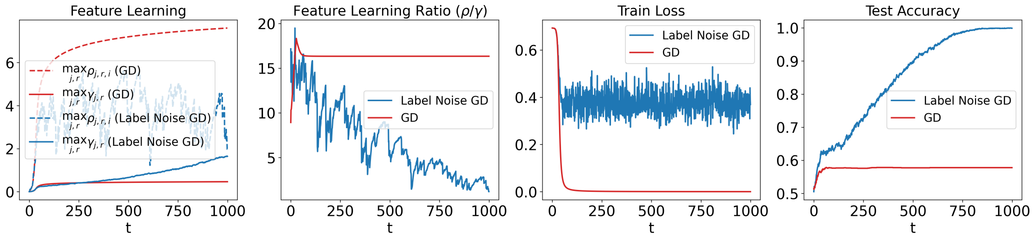


Figure 10: Performance with Gaussian noise $\mathcal{N}(1, 1)$: The ratio of noise memorization to signal learning, training loss, and test accuracy of standard GD and label noise GD.

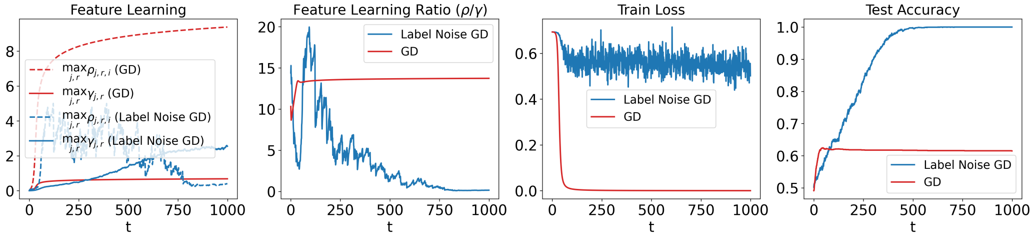


Figure 11: Performance with Gaussian noise $\mathcal{N}(0.6, 1)$: The ratio of noise memorization to signal learning, training loss, and test accuracy of standard GD and label noise GD.

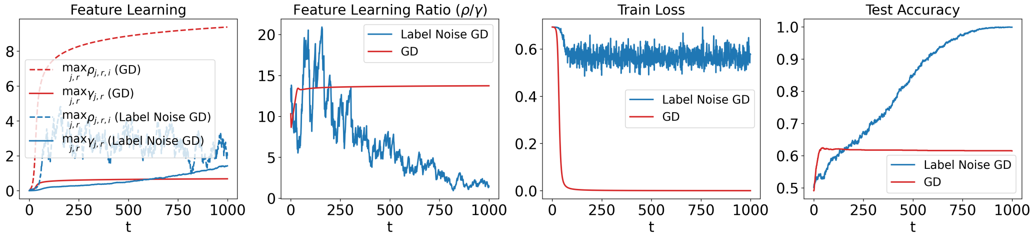


Figure 12: Performance with uniform distribution noise $\mathrm{unif}[-1, 2]$: The ratio of noise memorization to signal learning, training loss, and test accuracy of standard GD and label noise GD.

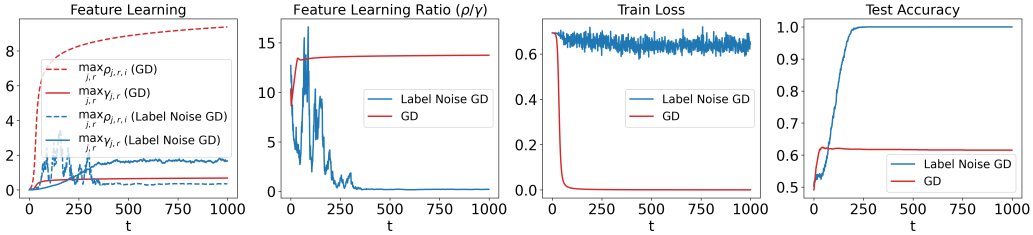


Figure 13: Performance with uniform distribution noise $\mathrm{unif}[-2, 3]$: The ratio of noise memorization to signal learning, training loss, and test accuracy of standard GD and label noise GD.

G.4 Higher Order Polynomial ReLU

In this work, we set the activation function as squared ReLU. This choice makes $q = 2$ a particularly interesting and challenging case to analyze, as it allows us to study the interaction between signal and noise in a setting that closely resembles practical two-layer ReLU networks.

For higher values of q, we also conducted experiments with $q = 3$ and $q = 4$. For $q = 3$, we set the learning rate $\eta = 0.5$, the number of neurons $m = 20$, the number of samples $n = 200$, the signal mean $\boldsymbol{\mu} = [2, 0, 0, \cdots, 0]$, and the noise strength $\sigma_p = 0.5$. The results are shown in Figure 14. For $q = 4$, the parameters were set as $\eta = 0.1$, $m = 20$, $n = 50$, $\boldsymbol{\mu} = [5, 0, 0, \cdots, 0]$, and $\sigma_p = 0.5$. The results are shown in Figure 15.

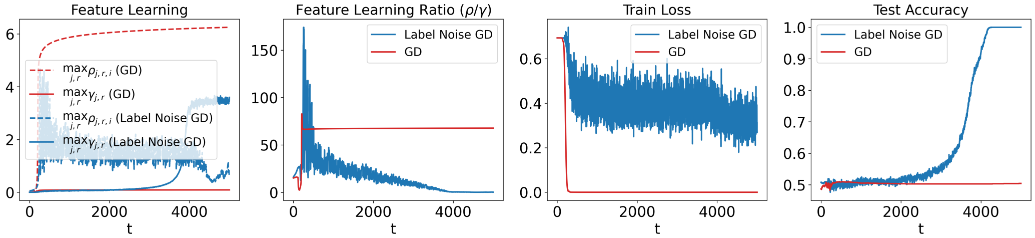


Figure 14: Performance with $q = 3$ for polynomial ReLU: The ratio of noise memorization to signal learning, training loss, and test accuracy of standard GD and label noise GD.

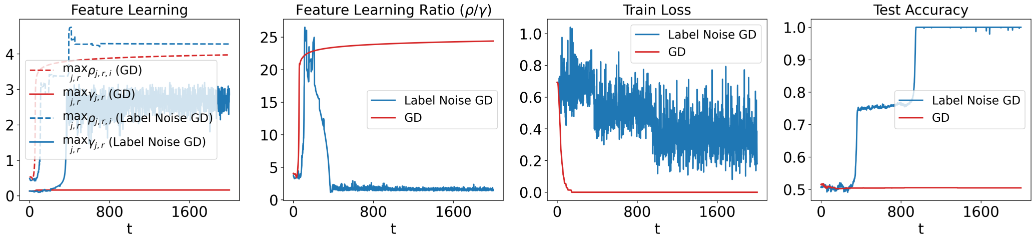


Figure 15: Performance with $q = 4$ for polynomial ReLU: The ratio of noise memorization to signal learning, training loss, and test accuracy of standard GD and label noise GD.

In all these cases, the experimental results consistently show that using a higher polynomial ReLU activation helps label noise GD suppress noise memorization while enhancing signal learning. This ultimately leads to improved test accuracy compared to standard GD.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction precisely match the contributions made in the paper, which include both the theoretical analysis and empirical validation of label noise GD for improving generalization under low SNR.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the "Conclusion and limitation" section, where we clearly state the current theoretical analysis is limited to specific architectures and activation functions, and outline directions for future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions for the theoretical results are clearly stated (e.g., Assumption 3.1), and complete proofs are provided in the appendix, with proof sketches included in the main text (Sections 3 and 4, Appendix B–E).

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All necessary experimental details, including dataset construction, noise injection, model architecture, training protocol, and evaluation metrics, are provided in Section 5 and Appendix F and G.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code necessary to reproduce the main experimental results is included in the supplementary material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 and Appendix F and G provide all relevant training and test details, including data splits, hyperparameters, optimizer choices, and noise settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The primary findings were robust and stable across all tested settings, so statistical significance analysis was not included. If requested, we can provide additional runs and error bar analysis during the rebuttal period.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 5 and Appendix F specify the compute environment (e.g., GPU type) and approximate runtime for main experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research fully conforms to the NeurIPS Code of Ethics. No human or sensitive data is used.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The broader impact is discussed in the "Conclusion and limitation" section, where we state there are no immediate negative societal impacts, and the work aims to advance theoretical understanding.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release any high-risk data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and code used (e.g., CIFAR-10, VGG-16) are cited in the references and used in accordance with their licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are introduced in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs are not used as an important, original, or non-standard component of the core methods in this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.