

# UNDERSTANDING GENERALIZATION IN TRANSFORMERS: ERROR BOUNDS AND TRAINING DYNAMICS UNDER BENIGN AND HARMFUL OVERFITTING

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Transformers serve as the foundational architecture for many successful large-scale models, demonstrating the ability to overfit the training data while maintaining strong generalization on unseen data, a phenomenon known as benign overfitting. However, existing research has not sufficiently explored generalization and training dynamics of transformers under benign overfitting. This paper addresses this gap by analyzing a two-layer transformer’s training dynamics, convergence, and generalization under labeled noise. Specifically, we present generalization error bounds for benign and harmful overfitting under varying signal-to-noise ratios (SNR), where the training dynamics are categorized into three distinct stages, each with its corresponding error bounds. Additionally, we conduct extensive experiments to identify key factors in transformers that influence test losses. Our experimental results align closely with the theoretical predictions, validating our findings.

## 1 INTRODUCTION

In recent years, benign overfitting has reshaped our understanding of overparameterization in deep neural networks. Traditional viewpoints hold that models with more parameters than training samples tend to overfit, resulting in poor generalization performance on new data. However, modern deep neural networks challenge this viewpoint by demonstrating remarkable generalization capabilities. Despite having sufficient parameters to perfectly fit training data, they still maintain low test loss Zhang et al. (2017); Neyshabur et al. (2018). This phenomenon, known as benign overfitting, has attracted significant attention across both statistical and machine learning communities Belkin et al. (2018; 2019; 2020); Neyshabur et al. (2018); Hastie et al. (2022).

Researchers have investigated benign overfitting from conventional perspective, while these works are related to linear models Chatterjee & Long (2022); Zou et al. (2021), kernel methods or random feature models Montanari & Zhong (2022); Adlam & Pennington (2020); Zhu Li (2021). Researchers have expanded these theoretical analyses to study benign overfitting in neural networks Adlam & Pennington (2020); Zhu Li (2021). They are still limited to the neural tangent kernel regime (NTK) Jacot et al. (2018) because the neural network learning problem is equivalent to kernel regression. Several works further study benign overfitting and generalization in transformers. These analyses typically focus on simplified settings, such as linear transformers Frei & Vardi (2024). Yet, due to the self-attention mechanism and softmax activation function, the transformer exhibits nonlinear learning in the real world, rendering the above simplifying assumption unreasonable.

Recent theoretical works have studied the benign overfitting and generalization of transformers with nonlinear self-attention Jiang et al. (2024); Magen et al. (2024), and some even have extended to context learning tasks Li et al. (2024b). Our analysis of benign overfitting and generalization in transformers is compared to existing research, as summarized in Table 1. However, several studies Frei & Vardi (2024); Li et al. (2024a) only considered generalization in a single type of overfitting (either benign or harmful). Others Jiang et al. (2024); Li et al. (2024a) analyzed the generalization of transformers under the assumption of clean data labels, which is unreasonable in real world. Therefore, an important open question remains:

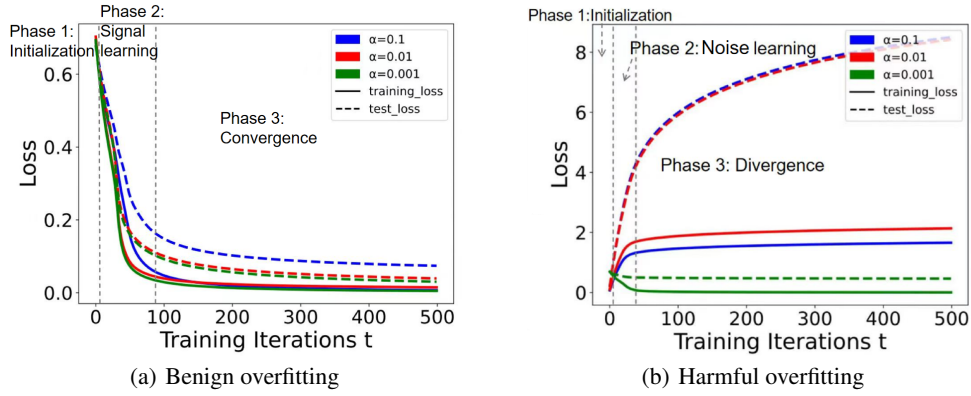


Figure 1: Test losses of benign overfitting and harmful overfitting under label noise (parameterized by  $\alpha$ ).

*How do transformers generalize under labeled noise while considering both benign overfitting and harmful overfitting?*

Our work aims to settle down the above question through feature learning framework by analyzing a two-layer transformer’s training dynamics, convergence, and generalization under labeled noise. Specifically, we consider two tokens including signal and noise, and a two-layer nonlinear transformer with softmax activation function. We explore the training dynamics of transformers in both benign overfitting and harmful overfitting, and provide corresponding error bounds. The theoretical bounds illustrate three distinct stages for benign overfitting and harmful overfitting, respectively. We then conduct extensive experiments to validate our theoretical finding. As shown in Figure 1, the test losses for benign overfitting and harmful overfitting divide into three distinct stages and the empirical loss is upper bounded by the theoretical bound (in Figure 2).

Theoretical Works	Nonlinear	Labeled Noise	Benign Overfitting	Harmful Overfitting	Stage-wise Error Bounds
Li et al. (2024a)	✓	×	×	✓	×
Sakamoto & Sato (2024)	✓	✓	✓	✓	×
Jiang et al. (2024)	✓	×	✓	✓	×
Frei & Vardi (2024)	×	✓	✓	×	×
Magen et al. (2024)	✓	✓	✓	✓	×
This work	✓	✓	✓	✓	✓

Table 1: Theoretical Comparison with existing works on benign overfitting and generalization.

Our contributions are summarized as follows:

- **Theoretical Contribution I :** We consider a nonlinear transformer with softmax activation function. Additionally, we relax the assumption of clean data labels and incorporate labeled noise to more accurately reflect real-world conditions.
- **Theoretical Contribution II :** We examine the training dynamics of transformers under labeled noise in both benign overfitting and harmful overfitting. The training dynamics associated with benign overfitting can be characterized by three distinct phases: **initialization**, **signal learning**, and **convergence**. In contrast, harmful overfitting is characterized by **initialization**, **noise learning**, and **divergence**. In Theorem 1 and Theorem 2, we provide specific stage-wise error bounds for each phase.
- **Experimental Contribution :** We investigate the transition between benign overfitting and harmful overfitting. Additionally, to further enhance the model’s generalization perfor-

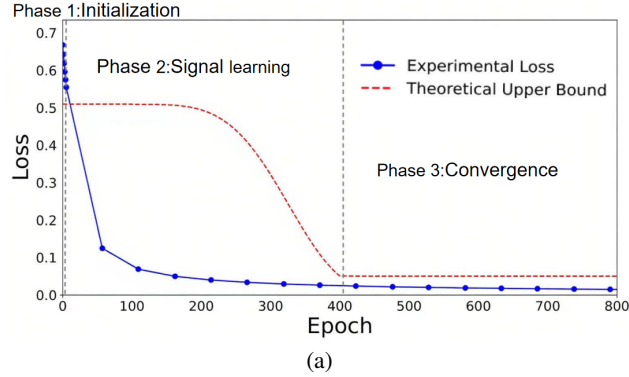


Figure 2: Numerical comparison between the theoretical upper bound and the experimental loss for benign overfitting.

mance, we analyze several key factors relevant to generalization during benign overfitting. These experimental results validate our theoretical analysis.

## 2 RELATED WORK

### 2.1 BENIGN OVERFITTING IN TRADITIONAL MODELS.

Several works explored benign overfitting in traditional models, including linear models Bartlett et al. (2020); Zou et al. (2021); Cao et al. (2021); Mo Zhou (2023), kernel methods, and random feature architectures. Zou et al. (2021) derived excess risk bounds for stochastic gradient descent with constant step sizes. Liao et al. (2021) expanded the analysis to random Fourier feature regression, focusing on fixed asymptotic ratios of sample size, data dimensionality, and feature count. As shown in Liang & Rakhlin (2020); Adlam & Pennington (2020); Zhu Li (2021); Montanari & Zhong (2022); Chatterjee & Long (2022); Spencer Frei (2022), several studies have broadened conventional perspectives to investigate benign overfitting in neural networks based on traditional models. The authors in Adlam et al. (2021) explored a precise analysis of generalization under nuclear regression, while Tsigler & Bartlett (2022) demonstrated that overparameterized ridge regression models can achieve benign overfitting even when fitting noisy data, and extended this to ridge regression conditions. Mallinar et al. (2024) discovered that interrupting training prematurely in neural networks leads to benign overfitting, while deep neural networks trained to full interpolation do not exhibit this phenomenon. Unlike these research, our work focuses on benign overfitting in transformers, which is more challenging than neural networks.

### 2.2 BENIGN OVERFITTING IN TRANSFORMER.

Towards understanding the benign overfitting and generalization in transformers, Frei & Vardi (2024) investigated the behavior of linear transformers trained on random linear classification tasks and quantifies how many examples transformers need in context learning to generalize well. Building on this, Magen et al. (2024) investigated benign overfitting in single-head attention models, revealing that this phenomenon only occurs when the signal-to-noise ratio reaches a sufficiently high level, and Sakamoto & Sato (2024) further explored benign overfitting in the token selection mechanism of the attention. The work in Li et al. (2024a) investigated the training dynamics of harmful overfitting when optimizing two-layer transformers using symbolic gradient descent. Most relevant to our work is Jiang et al. (2024), as they also study the benign overfitting and generalization of transformer with a similar data model. However, they do not take into account the effect of labeled noise, which is more common and realistic in real-world. In this paper, we bridge this gap by analyzing the generalization of transformers in benign overfitting and harmful overfitting under labeled noise condition.

### 3 PROBLEM SETUP

In this section, we denote the data generation model, two-layer transformer model, and the gradient descent training algorithm.

**Notions.** We define two sequences  $\{a_n\}$  and  $\{b_n\}$ , which have the following relationship. We define  $a_n = O(b_n)$  and  $b_n = \Omega(a_n)$  if there exist  $|a_n| \leq c_1|b_n|$  for some positive constant  $c_1$ . At the same time, we define  $a_n = \Theta(b_n)$  if  $a_n = O(b_n)$  and  $a_n = \Omega(b_n)$  hold.

**Definition 1.** Let  $\mu_+, \mu_- \in \mathbb{R}^d$  be fixed vectors which represent the signals contained in each data point  $(\mathbf{X}, y)$ , where  $\|\mu_+\|_2 = \|\mu_-\|_2 = \|\mu\|_2$  and  $\langle \mu_+, \mu_- \rangle = 0$ . Then we define each data point  $(\mathbf{X}, y)$  with the input features  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{d \times 2}$ , and  $y \in \{\pm 1\}$  is generated from the model:

- True labels  $\hat{y} \in \{\pm 1\}$  are Rademacher random variables with  $\mathbb{P}[\hat{y} = 1] = \mathbb{P}[\hat{y} = -1] = 1/2$ . Observed labels  $y$  are generated by flipping  $\hat{y}$  with probability  $\alpha$ , i.e.,  $\mathbb{P}[y = \hat{y}] = 1 - \alpha$  and  $\mathbb{P}[y = -\hat{y}] = \alpha$ .
- The signal vector  $\mathbf{x}_1$  is denoted  $\mu_+$  if  $\hat{y} = 1$ , and  $\mu_-$  if  $\hat{y} = -1$ .
- The noise vector  $\mathbf{x}_2 = \xi$  is sampled from  $\xi \sim \mathcal{N}(0, \sigma_p^2 \mathbf{I}_d)$ .

We consider each data point as a vector of two tokens,  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2)^T \in \mathbb{R}^{2 \times d}$ . The token  $\mathbf{x}_1$ , represents the signal that is inherently linked to the data's true class label, such as  $\mu_+$  and  $\mu_-$ , while  $\mathbf{x}_2$ , serves as noise and is irrelevant to the label. Building on Definition 3.1 from Jiang et al. (2024), we further refine the data distribution to enhance its practical applicability. Specifically, we introduce label-flipping noise to the true label  $\hat{y}$ .

**Signal-to-Noise Ratio (SNR).** From Cao et al. (2022), when the dimension  $d$  is large, the norm of the noise vector satisfies  $\|\xi\|_2 \approx \sigma_p \sqrt{d}$  based on standard concentration bounds. Therefore, the signal-to-noise ratio (SNR) can be expressed as  $\text{SNR} \approx \|\mu\|_2 / \sigma_p \sqrt{d}$ , which is approximately equal to  $\|\mu\|_2 / \|\xi\|_2$ . Hence, we use the expression  $\text{SNR} \approx \|\mu\|_2 / \sigma_p \sqrt{d}$  to represent the signal-to-noise ratio.

**Two-layer Transformer.** We define the model as a two-layer transformer, consisting of an attention layer with softmax activation function and a fixed linear layer. Let  $\mathbf{S}$  represent the softmax function. we categorize the output of the softmax function into four types of vectors, corresponding to the softmax outputs of the pairwise inner products involving the query signal, query noise, key signal, and key noise.

Specifically, the signal-to-signal output  $S_{11}$ , signal-to-noise output  $S_{12}$ , noise-to-signal output  $S_{21}$ , and noise-to-noise output  $S_{22}$  have been defined in the supplementary material. For example, the signal-to-signal output  $S_{11}$  can be written as:

$$\mathbf{S}_{11} = \text{Softmax}(\langle q_{\pm}^{(t)}, k_{\pm}^{(t)} \rangle) = \begin{cases} \frac{\exp(\langle q_+, k_+ \rangle)}{\exp(\langle q_+, k_+ \rangle) + \exp(\langle q_+, k_{\xi,i} \rangle)} & \text{for } i \in [\mathbf{S}_+], \\ \frac{\exp(\langle q_-, k_- \rangle)}{\exp(\langle q_-, k_- \rangle) + \exp(\langle q_-, k_{\xi,i} \rangle)} & \text{for } i \in [\mathbf{S}_-]. \end{cases}$$

Let  $\mathbf{S}_+$  be the set of indices  $i$  in  $[N]$  where  $y_i = 1$ , and let  $\mathbf{S}_-$  be the set of indices  $i$  in  $[N]$  where  $y_i = -1$ . Note that  $q_+, k_+, q_-, k_-$ , and  $k_{\xi,i}$  are related to the query with +1 label, the key with +1 label, the query with -1 label, the key with -1 label, and the key with noise, respectively. The output result can be given as:  $f(\mathbf{X}, v) = f_{+1}(\mathbf{X}, v) - f_{-1}(\mathbf{X}, v)$ , where  $f_j(\mathbf{X}, v)$  for  $j \in \{\pm 1\}$  is defined as:

$$f_j(\mathbf{X}, v) = \sum_{l=1}^2 v^\top \mathbf{W}_{V,j}^\top \mathbf{X} \mathbf{S}(\mathbf{X}^\top \mathbf{W}_Q \mathbf{W}_K^\top \mathbf{x}_l) = \sum_{l=1}^2 v^\top \left( \sum_{r=1}^{d_V} \mathbf{W}_{V,j,r}^\top \mathbf{X} \right) \mathbf{S}(\mathbf{X}^\top \mathbf{W}_Q \mathbf{W}_K^\top \mathbf{x}_l).$$

The parameter of the linear layer is denoted as  $v \in \mathbb{R}^{d_V}$ . The parameters of the attention layer are defined as  $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_{V,j}$ , where  $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d \times d_K}$  and  $\mathbf{W}_{V,j} \in \mathbb{R}^{d \times d_V}$ , representing the query matrix, the key matrix, and the value matrix respectively. We use  $\theta$  to represent all the parameters of the attention model, which is defined as  $\theta = (\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_{V,j})$ . We rewrite the model in a specific form for  $j \in \{\pm 1\}$ :

$$f_j(\theta, \mathbf{X}, v) = \sum_{r \in [d_V]} \left( v^\top \langle \mathbf{W}_{V,j,r}, \mathbf{x}_1 \rangle (\mathbf{S}_{11} + \mathbf{S}_{21}) + v^\top \langle \mathbf{W}_{V,j,r}, \mathbf{x}_2 \rangle (\mathbf{S}_{12} + \mathbf{S}_{22}) \right).$$

**Training Algorithm.** We use a training dataset  $S = \{(\mathbf{X}_i, y_i)\}_{i=1}^N$  generated from the distribution  $D$  defined in Definition 1. Our transformer model is trained by minimizing the logistic loss function:  $L_S(\theta) = \frac{1}{N} \sum_{i=1}^N \ell(y_i f(\theta, \mathbf{X}_i, v))$ , where  $\ell(z) = \log(1 + \exp(-z))$ . We employ gradient descent to minimize the training loss  $L_S(\theta)$ , and focus on characterizing the test error (i.e., true error), defined by:  $L_D^{0-1}(\theta) = \mathbb{P}_{(\mathbf{x}, y) \sim D} [y \neq \text{sign}(f(\theta, \mathbf{X}, v))]$ . For the sake of simplification, we consider gradient descent optimization, and we have  $\mathbf{W}_V^{(t+1)} = \mathbf{W}_V^{(t)} - \eta (\nabla_{\mathbf{W}_V} L_S(\mathbf{W}^{(t)}))$ ,  $\mathbf{W}_Q^{(t+1)} = \mathbf{W}_Q^{(t)} - \eta (\nabla_{\mathbf{W}_Q} L_S(\mathbf{W}^{(t)}))$ , and  $\mathbf{W}_K^{(t+1)} = \mathbf{W}_K^{(t)} - \eta (\nabla_{\mathbf{W}_K} L_S(\mathbf{W}^{(t)}))$ .

## 4 MAIN RESULTS

In this section, we present our main theoretical findings. These findings are based on several key conditions as follows:

**Assumptions .** Given a sufficiently small failure probability  $\delta > 0$ , a large constant  $c_1$ , and a target training loss  $\epsilon > 0$ , suppose the following conditions hold:

- (1) The dimension  $d_K$  satisfies:  $d_K \geq \begin{cases} \text{SNR}^4 N^4 \epsilon^{-4}, & \text{if } \|\mu\| \geq \sigma_p \sqrt{d}, \\ \text{SNR}^{-4} N^4 \epsilon^{-4}, & \text{if } \|\mu\| < \sigma_p \sqrt{d}. \end{cases}$
- (2) The dimension  $d$  satisfies:  $d \geq \text{poly}(d_K)$ .
- (3) The training sample size  $N$  satisfies:  $N \geq c_1 \cdot \text{polylog}(d)$ .
- (4) The label-flipping probability  $\alpha$  satisfies:  $\alpha \in [0, 1/2)$ .
- (5) The linear layer weight satisfies:  $\|v\|_2 = \Theta(1)$ .
- (6) The learning rate  $\eta$  satisfies:  $\eta \leq O(\min\{\sigma_p^2 d, \|\mu\|_2^2\} N^2 \epsilon^{-2})$ .
- (7) The parameters  $\mathbf{W}_Q$  and  $\mathbf{W}_K$  are initialized from a Gaussian distributions  $\mathcal{N}(0, \sigma_K^2)$  and the variance satisfies:  $\sigma_K^2 \leq O\left(\max\{(\sigma_p^2 d)^{-1}, \|\mu\|_2^{-2}\} N^{-1} \epsilon \log \frac{24N^2}{\delta}\right)^{-3/2}$ , while  $\mathbf{W}_V$  is initialized from  $\mathcal{N}(0, \sigma_V^2)$  where

$$\sigma_V \leq \begin{cases} O\left(\frac{\sqrt{\epsilon}}{\sqrt{dN}\|v\|\sigma_p}\right), & \text{if } \|\mu\| \geq \sigma_p \sqrt{d}, \\ O\left(\frac{\sqrt{\epsilon}}{\sqrt{N}\|v\|\|\mu\|}\right), & \text{if } \|\mu\| < \sigma_p \sqrt{d}. \end{cases}$$

Assumptions (1)–(3) ensure that the transformer operates in an over-parameterized setting. Similar assumptions have been made in neural networks Cao et al. (2022); Kou et al. (2023). Noise in training data is common in real-world environments. To address this gap, we relax the assumption of clean data labels and incorporate labeled noise  $\alpha$  to more accurately reflect real-world conditions. As a result, the generalization error bound derived under this assumption is more meaningful in practice. Assumption (4) ensures that we do not incorporate excessive noise, which could significantly impair the transformer’s generalization. This assumption is frequently used in theoretical analyses Kou et al. (2023); Frei & Vardi (2024); Sakamoto & Sato (2024). Assumption (5) is realistic in practice, as it controls the range of weights through appropriate training strategies. Assumptions (6)–(7) ensure that gradient descent can effectively minimize the training loss. Similar assumptions have been widely used in feature learning theories Cao et al. (2022); Jiang et al. (2024).

**Theorem 1 (Benign overfitting in transformers).** When  $N \cdot \text{SNR}^2 + h(\alpha) = \Omega(1)$ , where  $h(\alpha)$  is a function related to  $\alpha$ , for any  $\epsilon > 0$ , under the assumptions above, with probability at least  $1 - \delta$ :

- **(Phase 1: Initialization)** There exists  $T_1 = O\left(\frac{1}{\eta d_K^{\frac{1}{4}} \|\mu\|_2^2 \|v\|_2^2}\right)$ , and for  $t \in (0, T_1]$ , the test loss is upper bounded by:

$$L_D^{0-1}(\theta(t)) \leq \frac{1}{2} + \alpha + \mathcal{O}(1).$$

- **(Phase 2: Signal learning)** There exists  $T_2 = \Theta\left(\frac{1}{\eta\|\mu\|_2^2\|v\|_2^2}\right)$ , for  $t \in (T_1, T_2]$ , the test loss is upper bounded by:

$$L_{\mathcal{D}}^{0-1}(\theta(t)) \lesssim \alpha + \exp\left(-\eta^4\|\mu\|_2^8(t-T_1)^4\text{SNR}^2\right).$$

- **(Phase 3: Convergence)** There exists  $t > T_2$  such that:

- The training loss converges to  $\epsilon$ :  $L_S(\theta(t)) \leq \epsilon$ .
- The test loss is upper bounded by:

$$L_{\mathcal{D}}^{0-1}(\theta(t)) \lesssim \alpha + \exp\left(-\frac{\eta^4(t-T_2)^4\|\mu\|_2^6 \cdot \text{SNR}^2}{\sigma_V^2}\right).$$

Theorem 1 illustrates the generalization behavior of transformers under benign overfitting when  $N \cdot \text{SNR}^2 + h(\alpha) = \Omega(1)$ . Under this condition, the error bounds of transformers can be divided into three distinct phases:

- **Initialization phase:** Initially, the transformer parameters have not been adequately trained, leading to a test loss that remains at a significant constant level of  $\Omega(1)$ . This phase is primarily influenced by  $\alpha$  and the random initialization parameters ( $\sigma_V$  and  $\sigma_K^2$ ).
- **Signal learning phase:** During this phase, the model focuses more on learning the signals rather than the noises, which results in an increase in test loss. The test loss is governed by an upper bound that is directly proportional to time  $t$ , the labeled noise  $\alpha$ , the learning rate  $\eta$ , the signal strength  $\|\mu\|$ , and the square of the signal-to-noise ratio  $\text{SNR}^2$ .
- **Convergence phase:** When  $t > T_2$ , the training loss converges to a low level  $\epsilon$ . In this phase, the upper bound of the test loss is influenced by several key factors, including time  $t$ , the labeled noise  $\alpha$ , the learning rate  $\eta$ , the signal strength  $\|\mu\|$ , and  $\text{SNR}^2$ . Notably, it is inversely proportional to the initialization variance  $\sigma_V^2$ . By carefully tuning these factors under benign overfitting condition, we can achieve a lower test loss, which is the primary objective of our work in this paper.

**Theorem 2 (Harmful overfitting in transformers).** When  $N^{-1} \cdot \text{SNR}^{-2} + h(\alpha) = \Omega(1)$ , where  $h(\alpha)$  is a function related to  $\alpha$ , for any  $\epsilon > 0$ , under the assumptions, with probability at least  $1 - \delta$ :

- **(Phase 1: Initialization)** There exists  $T_1 = O\left(\frac{N}{\eta d_K^{\frac{1}{2}}\|\mu\|_2^2\|v\|_2^2}\right)$ , for  $t \in (0, T_1]$ , such that the test loss is upper bounded by:  $L_{\mathcal{D}}^{0-1}(\theta(t)) \leq \frac{1}{2} + \alpha + O(1)$ .
- **(Phase 2: Noise learning)** There exists  $T_2 = \Theta\left(\frac{N}{\eta\sigma_p^2 d\|v\|_2^2 \log(24N^2/\delta)}\right)$ . For  $t \in (T_1, T_2]$ , the test loss is bounded by:

$$L_{\mathcal{D}}^{0-1}(\theta(t)) \leq \frac{1}{2} + \alpha + O\left(\frac{1}{\|\mu\|_2^2\|v\|_2^2} + \frac{1}{\|\mu\|_2^4\|v\|_2^4}\right)$$

$$L_{\mathcal{D}}^{0-1}(\theta(t)) \geq \frac{1}{2} - O\left(\frac{1}{\|\mu\|_2^2\|v\|_2^4}\right).$$

- **(Phase 3: Divergence)** There exists  $t > T_2$  such that:

- The training loss is higher than  $\epsilon$ :  $L_S(\theta(t)) \geq \epsilon$ .
- The test loss is high:  $L_{\mathcal{D}}^{0-1}(\theta(t)) \geq \frac{1}{2}$ .

Theorem 2 characterizes the generalization behavior of the transformer in harmful overfitting when  $N^{-1} \cdot \text{SNR}^{-2} + h(\alpha) = \Omega(1)$ . The error bounds can be divided into three distinct phases:

- **Initialization phase:** Initially, the transformer parameters have not been sufficiently trained, resulting in the test loss remaining at a large constant value  $\Omega(1)$ . This is primarily influenced by  $\alpha$  and random initialization ( $\sigma_V$  and  $\sigma_K^2$ ). This indicates that the model has not yet effectively learned the signals or the noises.

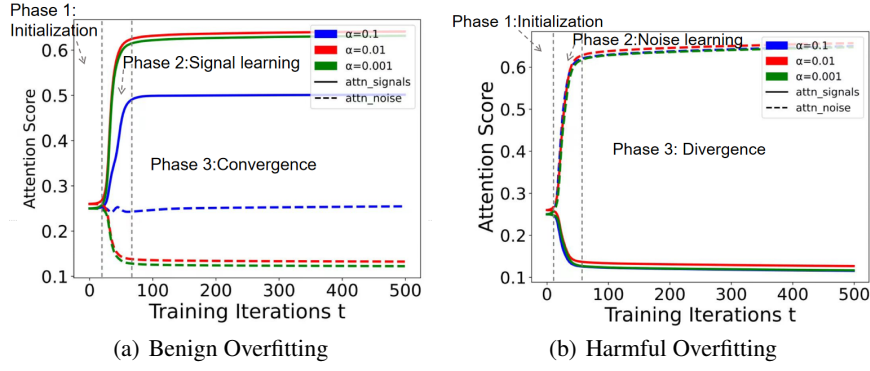


Figure 3: Attention score analysis of benign overfitting and harmful overfitting under various labeled noise  $\alpha \in \{0.2, 0.1, 0.01, 0.001\}$ . We denote `attn_signals` as the strength of the signals learned by attention, while `attn_noises` represents the strength of the noise learned by attention.

- **Noise learning phase:** During this phase, the model increasingly focuses on the noises rather than the signals, leading to an increase in test loss. The test loss is upper bounded by a function directly related to the label-flipping noise  $\alpha$ , the signal strength  $\|\mu\|$ , and the norm of the linear layer weight  $\|v\|_2$ . In contrast, the lower bound of the test loss is influenced solely by the signal strength  $\|\mu\|$  and the norm of the linear layer weight  $\|v\|_2$ .
- **Divergence phase:** When  $t > T_2$ , the model fully learns the noises. The test loss increases significantly and begins to diverge, ultimately exceeding  $1/2$ . This is higher than what would be expected from a random guess.

**Remark 1.** In summary, the model mainly learns the signals when benign overfitting occurs, resulting in lower loss values and better generalization. In contrast, when harmful overfitting occurs, the model mainly focuses on the noises, leading to poor generalization.

## 5 EXPERIMENTS

We present simulations using synthetic data to support our theoretical analysis. In this section, we demonstrate that the training dynamics can be clearly divided into three distinct phases based on varying  $\alpha$  values across both overfitting scenarios. Furthermore, we confirm the existence of benign overfitting and investigate the conditions under which it occurs. Finally, we investigated methods to further enhance the model’s generalization performance when benign overfitting occurs.

**Synthetic data setting:** We generate the training and test datasets according to Definition 1. Each data point consists of two components: signal and noise. The signal is composed of two orthogonal vectors,  $\|\mu\|_2 \cdot e_1$  and  $\|\mu\|_2 \cdot e_2$ , which are generated with equal probability.  $e_1$  and  $e_2$  are defined as  $[1, 0, \dots, 0]^T$  and  $[0, 1, \dots, 0]^T$  respectively. The noise is sampled from a Gaussian distribution  $\mathcal{N}(0, \sigma_p^2 \mathbf{I})$ . In our experiments, the sample size  $N$  is variable. Specifically, in the training dynamics and learning rate  $\eta$  experiments, we vary  $N$  from 2 to 20. In other experiments, we set  $N$  to 100 to ensure the model learns the data sufficiently. Furthermore, to investigate the effect of signal-to-noise ratio (SNR) on benign overfitting, we adjust the signal strength  $\mu$  from 1 to 100, while keeping the noise standard deviation  $\sigma_p$  constant at 4. This allows us to explore how varying SNR impacts the test loss.

### 5.1 TRAINING DYNAMICS OF BENIGN OVERFITTING AND HARMFUL OVERFITTING

We primarily illustrate the training dynamics by examining the attention scores and the values of the  $W_V$  matrix under various label-flipping noise conditions, encompassing both benign and harmful overfitting. Figure 3 (a) and (b) demonstrate that the training dynamics of attention can be characterized by three distinct phases. During the initialization phase, attention treats signals and noises equally, as it cannot distinguish between them. In the signal learning phase, attention increasingly focuses on the signals rather than the noises, and in the convergence phase, attention is entirely

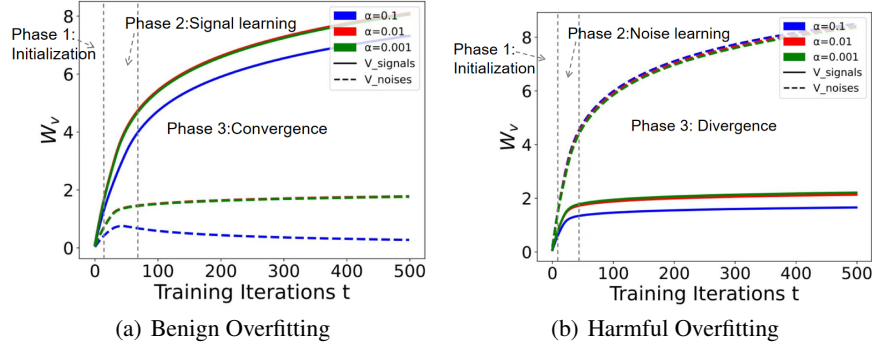


Figure 4: Analysis of the  $W_V$  matrix of benign overfitting and harmful overfitting under various labeled noise  $\alpha \in \{0.2, 0.1, 0.01, 0.001\}$ . We denote  $V\_signals$  as the strength of the signals learned by the  $W_V$  matrix, while  $V\_noises$  represents the strength of the noises learned by  $W_V$ .

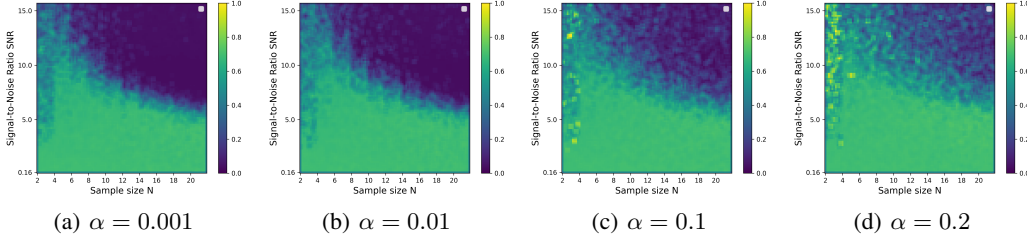


Figure 5: The heatmap of test loss on synthetic data across various  $SNR$ ,  $N$  and label-flipping probability  $\alpha$ .

directed towards the signals. In contrast, during the noise learning phase in harmful overfitting, attention increasingly concentrates on the noises rather than the signals. Eventually, attention primarily focuses on the noises, causing the model to learn irrelevant information. We also observe that as the label-flipping noise  $\alpha$  increases, a larger portion of the attention mechanism is directed towards the noises, leading the model to memorize more irrelevant information.

Figure 4 (a) and (b) demonstrate that the update of  $W_V$  matrix can be characterized by three distinct phases. According to Assumption (7), the  $W_V$  matrix starts with relatively small values due to random initialization. As training progresses, the model prefers to learn the signals rather than memorize the noises in benign overfitting, which is referred to as the signal learning phase. After a certain period,  $W_V$  stops learning noises and  $V\_noises$  converges to a constant, while  $W_V$  continues to learn signals. In contrast, as training progresses,  $W_V$  prefers to learn noises in harmful overfitting and  $W_V$  completely memorizes noises ultimately. Furthermore, we observe that the  $W_V$  matrix memorize more noises as the labeled noise  $\alpha$  increases when benign overfitting occurs.

We further conduct experiments on two types of overfitting test errors as shown in Figure 1, providing empirical verification for our theoretical results in Theorem 1 and Theorem 2. When benign overfitting occurs, the initialization stage is brief, and the test loss remains at a significantly high value. The model gradually learns the signals, with the test loss decreasing rapidly until it reaches the  $\epsilon$  level. During the convergence phase, the test loss stabilizes at the  $\epsilon$  level. In contrast, when harmful overfitting occurs, the model prefers to learn noises, with the test loss increasing rapidly and eventually diverging.

## 5.2 TRANSITION BETWEEN BENIGN OVERFITTING AND HARMFUL OVERFITTING

As illustrated in Figure 5, there is a clear distinction between benign overfitting and harmful overfitting under varying labeled noise  $\alpha$  and  $SNR$ . The test loss shows a decreasing trend as both  $N$  and  $SNR$  increase. To further explore this transition, we apply additional processing based on Figure 5. Figure 6 shows that the boundary does not undergo any significant spatial deformation as  $\alpha$



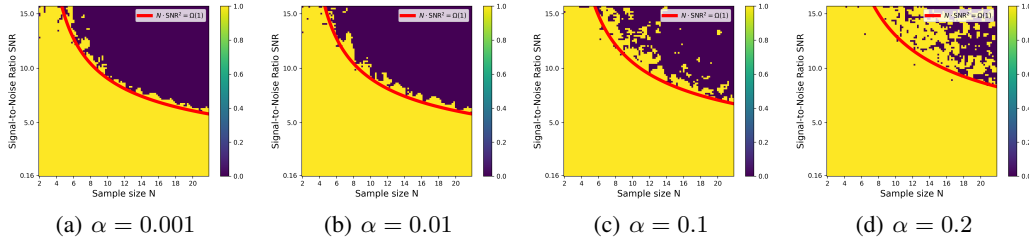
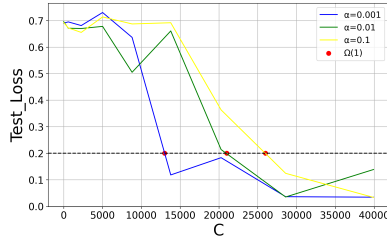
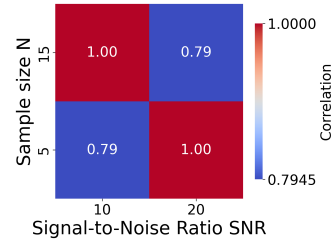


Figure 6: Under varying labeled noise  $\alpha$ , benign overfitting is depicted in yellow, while harmful overfitting is shown in purple. The transition between two types of overfitting is illustrated by a red curve.



(a) The test loss w.r.t. the critical line  $N \cdot \text{SNR}^2 + h(\alpha) = \Omega(1)$ .



(b) Similarity analysis of  $\alpha = 0.1$  and  $\alpha = 0.001$ .

Figure 7: (a) shows the variation of  $\Omega(1) - h(\alpha)$  with  $\alpha$ , while (b) shows the similarity between the two conditions  $\alpha = 0.1$  and  $\alpha = 0.001$ , with higher scores indicating higher similarity.

increases. Instead, it simply shifts spatially. This observation aligns with our theory, which indicates that the transition between benign and harmful overfitting is primarily governed by SNR and  $N$ , while  $\alpha$  only influences the translation  $h(\alpha)$ .

### 5.2.1 THE IMPACT OF CRITICAL LINE

The boundary  $N \cdot \text{SNR}^2 + h(\alpha) = \Omega(1)$  represents the minimum condition under which benign overfitting occurs. As illustrated in Figure 7 (a), we show how  $N \cdot \text{SNR}^2$  varies with changes in  $\alpha$ . The figure clearly demonstrates that the likelihood of convergence toward  $\Omega(1)$  shifts with changes in  $\alpha$ . Specifically, as  $\alpha$  increases, the term  $\Omega(1) - h(\alpha)$  rises, indicating a corresponding decrease in  $h(\alpha)$ . Furthermore, Figure 7 (b) reveals that the shape of the curve remains consistent, suggesting that  $\alpha$  affects only the spatial displacement  $h(\alpha)$  of the curve, without altering its overall form. The curves for  $\alpha = 0.1$  and  $\alpha = 0.001$  demonstrate a striking similarity, which verifies our theory: varying  $\alpha$  does not impact the distribution of the data; instead, it only influences the spatial offset  $h(\alpha)$  of the boundary line.

## 6 CONCLUSION AND FUTURE WORK

This paper studies the training dynamics, convergence, and generalization of a two-layer transformer with labeled noise. Firstly, we present generalization error bounds for both benign and harmful overfitting under varying signal-to-noise ratios (SNR). Secondly, we categorize the training dynamics into three stages and provide corresponding stage-wise error bounds. One limitation of our study is that the transformer model we analyze consists of only two layers. The more complex softmax and multi-layer attention mechanisms in deeper transformers create significant challenges in separating signal from noise, complicating the analysis of their training dynamics and generalization. An important direction for future work is to extend our analysis to deeper architectures.

## REFERENCES

- Ben Adlam, Jake Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities for deep learning, 2021.
- Brendan Adlam and James Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pp. 74–84, 2020.
- 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. ISSN 1091-6490. doi: 10.1073/pnas.1907378117. URL <http://dx.doi.org/10.1073/pnas.1907378117>.
- Mikhail Belkin, Siyuan Ma, and Saurabh Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 540–548, 2018.
- Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. ISSN 1091-6490. doi: 10.1073/pnas.1903070116. URL <http://dx.doi.org/10.1073/pnas.1903070116>.
- Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020. ISSN 2577-0187. doi: 10.1137/20m1336072. URL <http://dx.doi.org/10.1137/20M1336072>.
- Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.
- Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks, 2022.
- N. S. Chatterjee and P. M. Long. Deep linear networks can benignly overfit when shallow ones do, 2022.
- Spencer Frei and Gal Vardi. Trained transformer classifiers generalize and exhibit benign overfitting in-context, 2024.
- Trevor Hastie, Alessandro Montanari, Saharon Rosset, and Robert Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50:949–986, 2022.
- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31:8571–8580, 2018.
- Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. Unveil benign overfitting for transformer in vision: Training dynamics, convergence, and generalization, 2024.
- Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting for two-layer relu convolutional neural networks, 2023.
- 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, 2024a.
- Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning?, 2024b.
- Tengyu Liang and Alexander Rakhlin. Just interpolate: Kernel “ridgeless” regression can generalize. *The Annals of Statistics*, 48:1329–1347, 2020.
- Zhenyu Liao, Romain Couillet, and Michael W Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent\*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124006, 2021. ISSN 1742-5468. doi: 10.1088/1742-5468/ac3a77. URL <http://dx.doi.org/10.1088/1742-5468/ac3a77>.

- Roey Magen, Shuning Shang, Zhiwei Xu, Spencer Frei, Wei Hu, and Gal Vardi. Benign overfitting in single-head attention, 2024.
- Neil Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting, 2024.
- Rong Ge Mo Zhou. Implicit regularization leads to benign overfitting for sparse linear regression. *In International Conference on Machine Learning*, pp. 42543–42573, 2023.
- Alessandro Montanari and Yuchen Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50:2816–2847, 2022.
- Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks, 2018.
- Keitaro Sakamoto and Issei Sato. Benign overfitting in token selection of attention mechanism, 2024.
- Peter L. Bartlett Spencer Frei, Niladri S. Chatterji. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. *In Conference on Learning Theory*, pp. 2668–2703, 2022.
- A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression, 2022.
- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017.
- Arthur Gretton Zhu Li, Zhi-Hua Zhou. Towards an understanding of benign overfitting in neural networks, 2021.
- Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Benign overfitting of constant-stepsizes sgd for linear regression, 2021.

## A MORE EXPERIMENTS

### A.1 EXPERIMENTAL SETUP

In this section, we present simulations of synthetic data and true data to support our theoretical analysis in the previous section.

- **Synthetic data setting:** We generate the training and test datasets according to Definition 1. Each data point consists of two components: signal and noise. The signal is composed of two orthogonal vectors,  $\|\mu\|_2 \cdot \mathbf{e}_1$  and  $\|\mu\|_2 \cdot \mathbf{e}_2$ , which are generated with equal probability.  $\mathbf{e}_1$  and  $\mathbf{e}_2$  are defined as  $[1, 0, \dots, 0]^\top$  and  $[0, 1, \dots, 0]^\top$  respectively. The noise is sampled from a Gaussian distribution  $\mathcal{N}(0, \sigma_p^2 \mathbf{I})$ . In our experiments, the sample size  $N$  is variable. Specifically, in the training dynamics and learning rate  $\eta$  experiments, we vary  $N$  from 2 to 20. In other experiments, we set  $N$  to 100 to ensure the model learns the data sufficiently. Furthermore, to investigate the effect of signal-to-noise ratio (SNR) on benign overfitting, we adjust the signal strength  $\mu$  from 1 to 100, while keeping the noise standard deviation  $\sigma_p$  constant at 4. This allows us to explore how varying SNR impacts the test loss.
- **Model:** We define the model as a simple two-layer transformer used to explore the relevant factors that affect benign overfitting, consisting of an attention layer and multiple layers of perceptrons. The dimensions of the weight matrices are all 512. Parameters are initialized using PyTorch’s default initialization method. We set the target loss to 0.01 and conducted experiments using the full batch descent method.
- All experimental results were the average of 20 repeated experiments and all experiments were conducted on NVIDIA A100 GPU.

### A.2 CRITICAL FACTORS INFLUENCING THE GENERALIZATION OF BENIGN OVERFITTING

We explore the critical factors influencing the generalization of benign overfitting, including the learning rate  $\eta$ , sample size  $N$ , labeled noise  $\alpha$ , Gaussian initialization  $\sigma_V$  and signal-to-noise ratio (SNR). The following observations are consistent with our theoretical results.

#### A.2.1 LEARNING RATE $\eta$ AND SAMPLE SIZE $N$

As illustrated in Figure 8, during benign overfitting, both increasing the learning rate  $\eta$  and the sample size  $N$  contribute to improved generalization. Furthermore, Figure 8 (b) demonstrates that the test loss decreases significantly only when the signal-to-noise ratio (SNR) exceeds a specific threshold. This finding aligns with our initial assumption regarding the conditions for benign overfitting in our experimental setting.

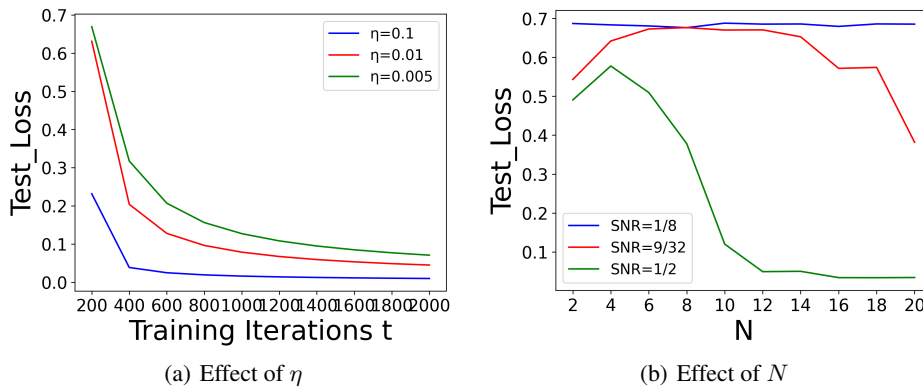


Figure 8: The effects of varying learning rate  $\eta$  and sample size  $N$  on test loss.

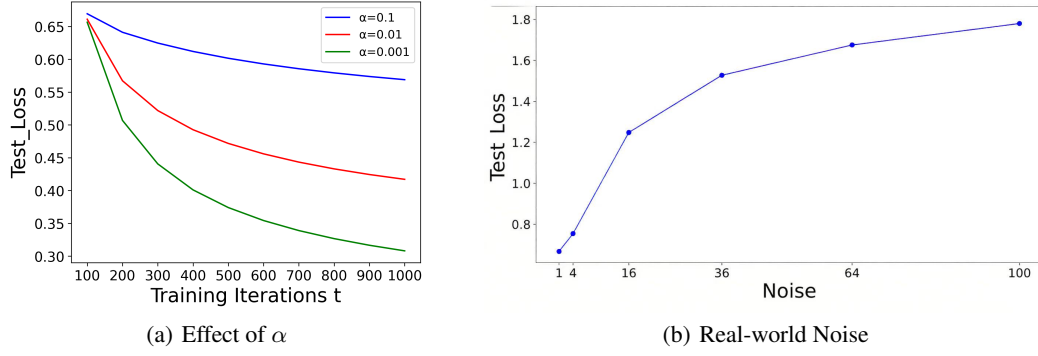


Figure 9: (a) represents the effect of varying label-flipping probability  $\alpha$  on test loss, while (b) investigates the impact of labeled noise on test loss in real-world dataset by increasing the proportion of noises within image patches.

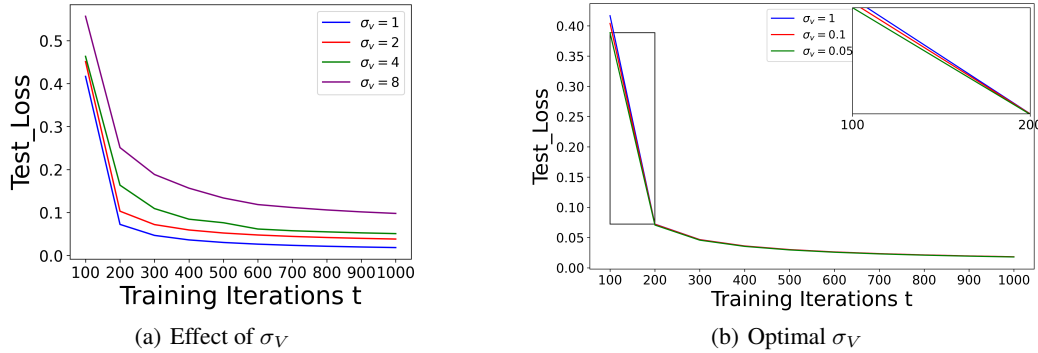


Figure 10: (a) examines the effect of varying Initialization  $\sigma_V$  on test loss, and (b) further investigates the optimal value for the initialization  $\sigma_V$ .

#### A.2.2 LABELED NOISE $\alpha$

We represent the effect of varying label-flipping probability  $\alpha$  on test loss. The synthetic experiment reveals that as  $\alpha$  increases, the test loss gradually rises as shown in Figure 9 (a). To validate this finding, we conduct experiments on the MNIST dataset, presented in Figure 9 (b), by increasing the proportion of noises within the image patches. We observe a consistent positive correlation between noise level and test loss.

#### A.2.3 GAUSSIAN INITIALIZATION $\sigma_V$

We investigate the effect of initialization  $\sigma_V$  on test loss. Specifically, we examine two aspects: Figure 10 (a) presents the test loss under various Gaussian initializations  $\sigma_V$ , while Figure 10 (b) reveals the optimal initialization values within a lower standard deviation range. Our findings indicate that as the Gaussian initialization  $\sigma_V$  decreases, the test loss tends to decline, as shown in Figure 10 (a). Additionally, Figure 10 (b) illustrates that when the standard deviation stabilizes below 1, the optimal initialization values closely align with these curves.

#### A.2.4 SIGNAL-TO-NOISE RATIO (SNR)

We also investigate the effect of varying signal-to-noise ratio (SNR) on test loss. SNR is determined by the signal norm  $\mu$  and noise standard deviation  $\sigma_p$ . We conduct experiments by varying  $\mu$  and  $\sigma_p$  while keeping other variables constant. Figure 11 demonstrates that test loss decreases with increased signal strength and decreased noise standard deviation.

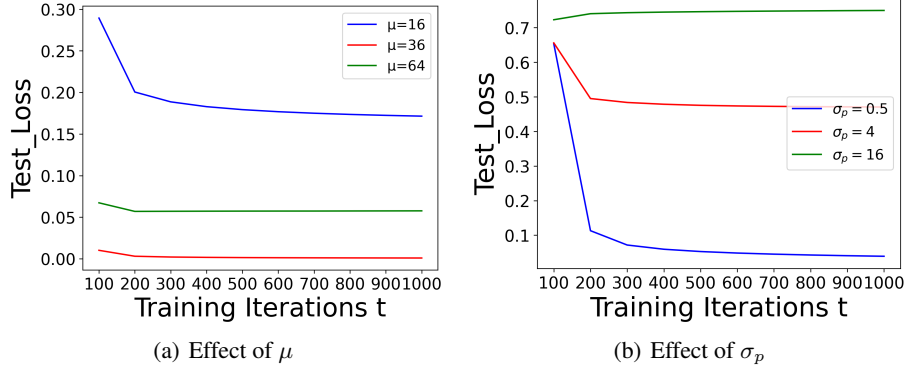


Figure 11: The effects of varying  $\mu$  and  $\sigma_p$  on test loss, where  $\mu$  represents the signal strength and  $\sigma_p$  represents the noise standard deviation.

## B SYMBOL NOTIONS

Symbols	Notions
$q_+^{(t)}, q_-^{(t)}, q_{\xi,i}^{(t)}$	vectorized Q, defined as $q_+^{(t)} = \mu_+^\top W_Q^{(t)}, q_-^{(t)} = \mu_-^\top W_Q^{(t)}, q_{\xi,i}^{(t)} = \xi_i^\top W_Q^{(t)}$
$k_+^{(t)}, k_-^{(t)}, k_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)}$	vectorized K, defined as $k_+^{(t)} = \mu_+^\top W_K^{(t)}, k_-^{(t)} = \mu_-^\top W_K^{(t)}, k_{\xi,i}^{(t)} = \xi_i^\top W_K^{(t)}, k_{\xi,i'}^{(t)} = \xi_{i'}^\top W_K^{(t)}$
$V_+^{(t)}, V_-^{(t)}, V_{\xi,i}^{(t)}$	scalarized V, defined as $V_+^{(t)} = \mu_+^\top W_V^{(t)} v, V_-^{(t)} := \mu_-^\top W_V^{(t)} v, V_{\xi,i}^{(t)} := \xi_i^\top W_V^{(t)} v$
$\Lambda_{\xi,\pm,i}^{(t)}, \Lambda_{\xi,i,\pm,i'}^{(t)}$	$\Lambda_{\xi,\pm,i}^{(t)} := \langle q_\pm^{(t)}, k_\pm^{(t)} \rangle - \langle q_\pm^{(t)}, k_{\xi,i}^{(t)} \rangle,$ $\Lambda_{\xi,i,\pm,i'}^{(t)} := \langle q_{\xi,i}^{(t)}, k_\pm^{(t)} \rangle - \langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle$

Table 2: Notions related to Query (Q), Key (K), and Value (V).

Symbols	Notions
$S_{11}$	a general reference to $\frac{\exp(\langle q_+^{(t)}, k_+^{(t)} \rangle)}{\exp(\langle q_+^{(t)}, k_+^{(t)} \rangle) + \exp(\langle q_+^{(t)}, k_{\xi,i}^{(t)} \rangle)}$ for $i \in S_+$ , and $\frac{\exp(\langle q_-^{(t)}, k_+^{(t)} \rangle)}{\exp(\langle q_-^{(t)}, k_+^{(t)} \rangle) + \exp(\langle q_-^{(t)}, k_{\xi,i}^{(t)} \rangle)}$ for $i \in S_-$
$S_{21}$	a general reference to $\frac{\exp(\langle q_{\xi,i}^{(t)}, k_+^{(t)} \rangle)}{\exp(\langle q_{\xi,i}^{(t)}, k_+^{(t)} \rangle) + \exp(\langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle)}$ for $i, i' \in S_+$ , and $\frac{\exp(\langle q_{\xi,i}^{(t)}, k_-^{(t)} \rangle)}{\exp(\langle q_{\xi,i}^{(t)}, k_-^{(t)} \rangle) + \exp(\langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle)}$ for $i, i' \in S_-$
$S_{12}$	a general reference to $\frac{\exp(\langle q_+^{(t)}, k_{\xi,i}^{(t)} \rangle)}{\exp(\langle q_+^{(t)}, k_+^{(t)} \rangle) + \exp(\langle q_+^{(t)}, k_{\xi,i}^{(t)} \rangle)}$ for $i \in S_+$ , and $\frac{\exp(\langle q_-^{(t)}, k_{\xi,i}^{(t)} \rangle)}{\exp(\langle q_-^{(t)}, k_-^{(t)} \rangle) + \exp(\langle q_-^{(t)}, k_{\xi,i}^{(t)} \rangle)}$ for $i \in S_-$
$S_{22}$	a general reference to $\frac{\exp(\langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle)}{\exp(\langle q_{\xi,i}^{(t)}, k_+^{(t)} \rangle) + \exp(\langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle)}$ for $i, i' \in S_+$ , and $\frac{\exp(\langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle)}{\exp(\langle q_{\xi,i}^{(t)}, k_-^{(t)} \rangle) + \exp(\langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle)}$ for $i, i' \in S_-$

Table 3: Notions related to softmax.

## C PROOF TECHNIQUES

In this section, we present the main proof techniques for studying the training dynamics and the specific test loss in both benign and harmful overfitting. The complete proofs are provided in the Appendix D.

### C.1 KEY TECHNIQUE 1: UPPER BOUND AND LOWER BOUND ANALYSIS OF TEST ERROR

We recognize that, due to the presence of label-flipping noise, the Bayes optimal test error is at least  $\alpha$ . Consequently, the discrepancy between the test error and the training error is no less than  $\alpha$ . This situation prevents us from employing the commonly utilized logistic loss to minimize the empirical risk. In our work, we conduct a different test error analysis. Initially, we can express the 0-1 test error as follows:

$$\begin{aligned} L_{\mathcal{D}}^{0-1}(\theta) &= \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}} [y \neq \text{sign}(f(\theta, \mathbf{X}, v))] \\ &= \alpha + (1 - 2\alpha) \mathbb{P}(\hat{y}f(\theta, \mathbf{X}, v) \leq 0). \end{aligned}$$

**Lemma 3.** *If  $\alpha \in [0, 1/C)$ ,  $f$  is the model output function, then the upper bound of the test loss function satisfies the following inequalities:*

$$\mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}} [y \neq \text{sign}(f(\theta, \mathbf{X}, v))] \leq \alpha + \mathbb{P}(\hat{y}f(\theta, \mathbf{X}, v) \leq 0). \quad (1)$$

**Lemma 4.** *If  $f(\theta, \mathbf{X}, v) \sim \mathcal{N}(E, \sigma_f^2)$ , and  $\Phi(-X)$  is the cumulative distribution function of the standard normal distribution. The test loss can be bounded as follows:*

- *Upper Bound of Joint Probability Density Function:*

$$\begin{aligned} P(\hat{y}f(\theta, \mathbf{X}, v) \leq 0) &\approx \Phi\left(-\frac{\mathbb{E}}{\sigma_f}\right) \\ &\leq \frac{1}{2} - \frac{\mathbb{E}}{\sigma_f \sqrt{2\pi}} + O\left(\frac{\mathbb{E}^2}{\sigma_f^2}\right) \end{aligned}$$

- *Lower Bound of Joint Probability Density Function:  $P(\hat{y}f(\theta, \mathbf{X}, v) \leq 0) \geq \frac{1}{2} - \frac{E}{\sigma_f \sqrt{2\pi}}$*

### C.2 KEY TECHNIQUE 2: SIMPLIFY THE JOINT PROBABILITY DENSITY FUNCTION

With Key Technique 1, the analysis of the test error can be simplified to estimating the probability of incorrect predictions  $\mathbb{P}(\hat{y}f(\theta, \mathbf{X}, v) \leq 0)$ .

**Lemma 5.** *Because it involves label flipping, the model output value can be divided into real label  $\hat{y}$  and flipped label  $-\hat{y}$ . We divide  $V$ :  $V_{+\hat{y}}^{(t)} = \sum_r a_r V_+^{(t)}$ ,  $V_{+(-\hat{y})}^{(t)} = \sum_r b_r V_+^{(t)}$ ,  $V_+^{(t)} = V_{+\hat{y}}^{(t)} + V_{+(-\hat{y})}^{(t)}$ . Similarly, we split  $V_-^{(t)}$  and  $V_{\xi}^{(t)}$  for  $j = \pm\hat{y}$ .  $\hat{y}f(\theta, \mathbf{X}, v)$  can be composed as:*

$$\hat{y}f(\theta, \mathbf{x}, v) = \sum_{j,r} [(S_{11} + S_{21})(V_+^{(t)} + V_-^{(t)}) + (S_{12} + S_{22})V_{\xi}^{(t)}].$$

**Lemma 6** (The test loss of benign overfitting). *When analyzing the second phase of benign overfitting, we have  $P(\hat{y}f(\theta, \mathbf{x}, v) \leq 0) \leq P\left(\sum_r \frac{S_{11}+S_{21}}{S_{12}+S_{22}} \left(V_{+\hat{y}}^{(t)} + V_{-(-\hat{y})}^{(t)}\right) \leq V_{\xi, (-\hat{y})}^{(t)} + o(1)\right)$ .*

### C.3 KEY TECHNIQUE 3: JOINT PROBABILITY DENSITY FUNCTION ESTABLISHES A RELATIONSHIP WITH WEIGHT UPDATE

**Lemma 7.** *Under Assumptions (1)–(7), in the theoretical analysis of test error in the second and third stages of benign overfitting, we define  $g(\xi)$  as  $V_{\xi, (-\hat{y})}^{(t)} = \sum_r \langle v W_{-\hat{y}, r}^{(t)}, \xi \rangle$ , where  $W_{-\hat{y}, r}^{(t)}$  refers to the row vector in parameter matrix  $W$  with label  $-\hat{y}$  and index  $r$  at the  $t$ -th training round. Then, we know that for any  $x \geq 0$ , if  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  is a Lipschitz function and  $c$  is a constant, we have*

$$\mathbb{P}(g(\xi) - \mathbb{E}g(\xi) \geq x) \leq \exp\left(-\frac{cx^2}{\sigma_p^2 \sum_{r=1}^{d_V} \|W_{-\hat{y}, r}^{(t)} v\|_2^2}\right).$$

## D DETAILED PROOF OF THE LEMMA

**Lemma 8** (Upper Bound of Joint Probability Density Function). *If  $\alpha \in [0, 1/C)$ ,  $f$  is the model output function, then the upper bound of the test loss function satisfies the following inequalities:*

$$\mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}} [y \neq \text{sign}(f(\theta, \mathbf{X}, v))] \leq \alpha + \mathbb{P}(\hat{y}f(\theta, \mathbf{X}, v) \leq 0). \quad (2)$$

*Proof of Lemma 3.* We can write out the test error as

$$\begin{aligned} L_{\mathcal{D}}^{0-1}(\theta(t)) &= P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (y \neq \text{sign}(f(\theta, \mathbf{X}, v))) \\ &= P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (yf(\theta, \mathbf{X}, v) \leq 0) \\ &= P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (yf(\theta, \mathbf{X}, v) \leq 0, y \neq \hat{y}) + P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (yf(\theta, \mathbf{X}, v) \leq 0, y = \hat{y}) \\ &= \alpha \cdot P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (\hat{y}f(\theta, \mathbf{X}, v) \geq 0) + (1 - \alpha) \cdot P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (\hat{y}f(\theta, \mathbf{X}, v) \leq 0) \\ &\leq \alpha + P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (\hat{y}f(\theta, \mathbf{X}, v) \leq 0), \end{aligned}$$

In the second and third equation, we used the definition of  $\mathcal{D}$  in Definition 1.

**Lemma 9** (Upper Bound and Lower Bound of Joint Probability Density Function). *If  $f(\theta, \mathbf{X}, v) \sim \mathcal{N}(E, \sigma_f^2)$ , and  $\Phi(-X)$  is the cumulative distribution function of the standard normal distribution. Upper bound and lower bound of test loss function satisfy the following inequalities:*

- *Upper Bound of Joint Probability Density Function:*

$$\begin{aligned} P(\hat{y}f(\theta, \mathbf{X}, v) \leq 0) &\approx \Phi\left(-\frac{\mathbb{E}}{\sigma_f}\right) \\ &\leq \frac{1}{2} - \frac{\mathbb{E}}{\sigma_f \sqrt{2\pi}} + O\left(\frac{\mathbb{E}^2}{\sigma_f^2}\right) \end{aligned}$$

- *Lower Bound of Joint Probability Density Function:*  $P(\hat{y}f(\theta, \mathbf{X}, v) \leq 0) \geq \frac{1}{2} - \frac{E}{\sigma_f \sqrt{2\pi}}$

*Proof of Lemma 4.* we have

$$\hat{y}f \sim \mathcal{N}(E, \sigma_f^2)$$

By applying the Taylor expansion to correct the probability, and the cumulative distribution function of the Gaussian distribution is denoted by  $\Phi$ . Using  $\Phi(-x) \approx \frac{1}{2} - \frac{x}{\sqrt{2\pi}} + O(x^2)$ , we can rewrite:

$$\begin{aligned} P(\hat{y}f \leq 0) &\approx \Phi\left(-\frac{\mathbb{E}}{\sigma_f}\right) \\ &\leq \frac{1}{2} - \frac{\mathbb{E}}{\sigma_f \sqrt{2\pi}} + O\left(\frac{\mathbb{E}^2}{\sigma_f^2}\right) \\ P(\hat{y}f(\theta, \mathbf{X}, v) \leq 0) &\geq \frac{1}{2} - \frac{E}{\sigma_f \sqrt{2\pi}} \end{aligned}$$

**Lemma 10** ( $f(\theta, \mathbf{X}, v)$  establishes a relationship with the signal and noise).

$$\hat{y}f(\theta, \mathbf{x}, v) = \sum_{j,r} [(S_{11} + S_{21})(V_+^{(t)} + V_-^{(t)}) + (S_{12} + S_{22})V_{\xi}^{(t)}].$$

*Proof of Lemma 5.*

$$\begin{aligned} &\hat{y}f(\theta, \mathbf{x}, v) \\ &= \sum_{r \in [d_V]} (v^T x_1 (S_{11} + S_{21}) W_{V,r}^{(t)} + v^T x_2 (S_{12} + S_{22}) W_{V,r}^{(t)}) \\ &= \sum_j \sum_{r \in [d_V]} [(S_{11} + S_{21}) \langle W_{V,j,r}^{(t)}, x_1 \rangle v + (S_{12} + S_{22}) \langle W_{V,j,r}^{(t)}, x_2 \rangle v] \\ &= \sum_{j,r} [(S_{11} + S_{21}) (\mu_+^T W_{v,j,r}^{(t)} v + \mu_-^T W_{V,j,r}^{(t)} v) + (S_{12} + S_{22}) \xi^T W_{V,j}^{(t)} v] \\ &= \sum_{j,r} [(S_{11} + S_{21})(V_+^{(t)} + V_-^{(t)}) + (S_{12} + S_{22})V_{\xi}^{(t)}]. \end{aligned}$$



**Lemma 11** (The test loss of benign overfitting). *When analyzing the second phase of benign overfitting, by substituting the segmented  $V$  vector and Lemma 5 into  $\mathbb{P}(\hat{y}f(\theta, \mathbf{X}, v) \leq 0)$ , we have*

$$\begin{aligned} & P(\hat{y}f(\theta, \mathbf{x}, v) \leq 0) \\ &= P\left(\sum_r (S_{11} + S_{21})(V_{+, \hat{y}}^{(t)} + V_{-, \hat{y}}^{(t)}) + (S_{12} + S_{22})V_{\xi, \hat{y}}^{(t)}\right. \\ &\leq \left.\sum_r ((S_{11} + S_{21})(V_+^{(t)} + V_-^{(t)}) + (S_{12} + S_{22})V_{\xi, (-\hat{y})}^{(t)})\right) \\ &\leq P\left(\sum_r \frac{S_{11} + S_{21}}{S_{12} + S_{22}} (V_{+, \hat{y}}^{(t)} + V_{-, \hat{y}}^{(t)}) \leq V_{\xi, (-\hat{y})}^{(t)} + o(1)\right). \end{aligned}$$

*Proof of Lemma.6.* The first equality arises from the conversion relationship among  $V_+^{(t)}$ ,  $V_{+, \hat{y}}^{(t)}$ , and  $V_{+, (-\hat{y})}^{(t)}$ . The second inequality holds because we are in a benign overfitting phase, where the signal memory exceeds the noise memory. Here, the left side of the inequality is predominantly influenced by the signal memory, while the right side satisfies Equation 14 and Equation 17.

**Lemma 12** (Joint Probability Density Function Establishes a Relationship with Weight Update). *Under Assumptions (1)–(7), in the theoretical analysis of test error in the second and third stages of benign overfitting, we define  $g(\xi)$  as  $V_{\xi, (-\hat{y})}^{(t)} = \sum_r \langle v W_{-\hat{y}, r}^{(t)}, \xi \rangle$ , where  $W_{-\hat{y}, r}^{(t)}$  refers to the row vector in parameter matrix  $W$  with label  $-\hat{y}$  and index  $r$  at the  $t$ -th training round. Then, we know that for any  $x \geq 0$ , if  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  is a Lipschitz function and  $c$  is a constant, the following inequality holds for the test loss.*

$$\mathbb{P}(g(\xi) - \mathbb{E}g(\xi) \geq x) \leq \exp\left(-\frac{cx^2}{\sigma_p^2 \sum_{r=1}^{d_V} \|W_{-\hat{y}, r}^{(t)} v\|_2^2}\right).$$

*Proof of Lemma.7.* According to Theorem 5.2.2 in Vershynin (2018), we know that for any  $x \geq 0$ , if  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  is a Lipschitz function, it holds that

$$\mathbb{P}(g(\xi) - \mathbb{E}g(\xi) \geq x) \leq \exp\left(-\frac{cx^2}{\sigma_p^2 \|g\|_{\text{Lip}}^2}\right), \quad (3)$$

Since  $g(\xi)$  is defined as  $V_{\xi, (-\hat{y})}^{(t)} = \sum_r \langle v W_{-\hat{y}, r}^{(t)}, \xi \rangle$  and  $\langle W_{-\hat{y}, r}^{(t)} v, \xi \rangle \sim \mathcal{N}(0, \|W_{-\hat{y}, r}^{(t)}\|_2^2 \|v\|_2^2 \sigma_p^2)$ , we have

$$\begin{aligned} |g(\xi) - g(\xi')| &= \left| \sum_{r=1}^{d_V} \langle v W_{-\hat{y}, r}^{(t)}, \xi \rangle - \sum_{r=1}^{d_V} \langle v W_{-\hat{y}, r}^{(t)}, \xi' \rangle \right| \\ &\leq \sum_{r=1}^{d_V} \left| \langle v W_{-\hat{y}, r}^{(t)}, \xi \rangle - \langle v W_{-\hat{y}, r}^{(t)}, \xi' \rangle \right| \\ &= \sum_{r=1}^{d_V} \left| \langle v W_{-\hat{y}, r}^{(t)}, \xi - \xi' \rangle \right| \\ &\leq \sum_{r=1}^{d_V} \|W_{-\hat{y}, r}^{(t)}\|_2 \|v\|_2 \|\xi - \xi'\|_2 \end{aligned}$$

So, we can get

$$\|g\|_{\text{Lip}} \leq \sum_{r=1}^{d_V} \|W_{-\hat{y}, r}^{(t)} v\|_2, \quad (4)$$

By plugging Equation 4 into Equation 3, we get:

$$\mathbb{P}(g(\xi) - \mathbb{E}g(\xi) \geq x) \leq \exp\left(-\frac{cx^2}{\sigma_p^2 \sum_{r=1}^{d_V} \|W_{-\hat{y}, r}^{(t)} v\|_2^2}\right).$$

**Lemma 13** (Relationship of constants). *To ensure the continuity of the test error function, it is necessary to verify that at  $t = T_1$ , the test error of the first stage matches the initial test error of the second stage. Similarly, at  $t = T_2$ , the test error of the second stage should equal the initial test error of the third stage. It is evident that this condition holds for  $t = T_1$ . We will focus on verifying this condition for  $t = T_2$ .*

*Proof.*

$$\begin{aligned} \alpha + \exp \left[ \frac{c_4}{2\pi} - c_{10}\eta^4 \|\mu\|^8 (T_2 - T_1)^2 (T_2 - T_1 - 1)^2 d \text{SNR}^2 \dot{d}_K \right] &= \alpha + \exp \left( \frac{c_{12}}{2\pi} \right) \\ \frac{c_4 - c_{12}}{2\pi \cdot c_{10}} &= \eta^4 \|\mu\|^8 (T_2 - T_1)^2 (T_2 - T_1 - 1)^2 d \text{SNR}^2 \dot{d}_K \end{aligned}$$

*It is sufficient to ensure that the above relations are satisfied among the three constants.*

## E BASIC INEQUALITY

### E.1 BENIGN OVERFITTING

According to Jiang et al. (2024), for  $t \in (T_1, T_2]$ , the following inequalities hold during the second phase of benign overfitting,

$$\frac{\exp(\langle q_+^{(t)}, k_{\xi,i}^{(t)} \rangle)}{\exp(\langle q_+^{(t)}, k_+^{(t)} \rangle) + \exp(\langle q_+^{(t)}, k_{\xi,i}^{(t)} \rangle)} \leq \frac{\exp(\langle q_+^{(t)}, k_{\xi,i}^{(t)} \rangle)}{\exp(\langle q_+^{(t)}, k_+^{(t)} \rangle)} = \frac{1}{c_5 \exp(\Lambda_{\xi,\pm,i}^{(t)})} \quad (5)$$

$$\frac{\exp(\langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle)}{\exp(\langle q_{\xi,i}^{(t)}, k_+^{(t)} \rangle) + \exp(\langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle)} \leq \frac{1}{c_7 \exp(\Lambda_{\xi,i,\pm,i'}^{(t)})} \quad (6)$$

$$V_+^{(t)} \geq \eta c_1 \|\mu\|_2^2 \|v\|_2^2 (t - T_1) \quad (7)$$

$$|V_{\pm}^{(t)}| \leq O(d^{-\frac{1}{4}}) + \eta c_4 \|\mu\|_2^2 \|v\|_2^2 (t - T_1) \quad (8)$$

$$\Lambda_{\xi,\pm,i}^{(t+1)} \geq \log \left( \exp(\Lambda_{\xi,\pm,i}^{(T_1)}) + \frac{\eta^2 c_8 \|\mu\|_2^4 \|v\|_2^2 d_K^{\frac{1}{2}}}{N(\log(24N^2/\delta))^2} \cdot (t - T_1)(t - T_1 + 1) \right) \quad (9)$$

$$\Lambda_{\xi,i,\pm,i'}^{(t+1)} \geq \log \left( \exp(\Lambda_{\xi,i,\pm,i'}^{(T_1)}) + \frac{\eta^2 c_8 \sigma_p^2 d \|\mu\|_2^2 \|v\|_2^2 d_K^{\frac{1}{2}}}{N(\log(24N^2/\delta))^2} \cdot (t - T_1)(t - T_1 + 1) \right) \quad (10)$$

$$\|\mathbf{W}_V^{(t+1)} v - \mathbf{W}_V^{(t)} v\|_2 = O \left( \eta \cdot \max \left\{ \|\mu\|_2, \sigma_p \sqrt{d} \right\} \cdot \|v\|_2 \right) \quad (11)$$

$$V_+^{(T_2)} \geq 6 \cdot |V_{\xi,i}^{(T_2)}|, \quad (12)$$

$$V_-^{(T_2)} \leq -6 \cdot |V_{\xi,i}^{(T_2)}|. \quad (13)$$

$$|V_+^{(T_2)}|, |V_-^{(T_2)}|, |V_{\xi,i}^{(T_2)}| = o(1), \quad (14)$$

When benign overfitting occurs during the third stage, the following inequality is satisfied:

$$\log \left( \exp(V_+^{(T_2)}) + \eta c_{11} \|\mu\|_2^2 \|v\|_2^2 (t - T_2) \right) \leq V_+^{(t)} \leq 2 \log \left( O \left( \frac{1}{\epsilon} \right) \right), \quad (15)$$

$$-2 \log \left( O \left( \frac{1}{\epsilon} \right) \right) \leq V_-^{(t)} \leq -\log \left( \exp(-V_-^{(T_2)}) + \eta c_{11} \|\mu\|_2^2 \|v\|_2^2 (t - T_2) \right) \quad (16)$$

$$|V_+^{(t)}|, |V_-^{(t)}|, |V_{\xi,i}^{(t)}| = o(1), \quad (17)$$

## E.2 HARMFUL OVERFITTING

According to Jiang et al. (2024), for  $t \in (T_1, T_2]$ , the following inequalities hold during the second phase of harmful overfitting

$$V_{\xi,i}^{(t)} \geq \frac{\eta c_{13} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N} \quad (18)$$

$$V_{\xi,i}^{(t)} \leq -\frac{\eta c_{14} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N} \quad (19)$$

$$|V_{\pm}^{(t)}| \leq O(d^{-\frac{1}{4}}) + \frac{\eta c_{15} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N} \quad (20)$$

$$|V_{\xi,i}^{(t)}| \leq O(d^{-\frac{1}{4}}) + \frac{\eta c_{16} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N} \quad (21)$$

$$\text{Softmax}(\langle \mathbf{q}_{\pm}^{(t)}, \mathbf{k}_{\pm}^{(t)} \rangle) = O\left(\frac{\sigma_p^2 d (\log(24N^2/\delta))^3}{\|\boldsymbol{\mu}\|_2^2 \|v\|_2^2 d^{\frac{1}{2}}}\right) \quad (22)$$

$$\text{Softmax}(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle) = 1 - O\left(\frac{(\log(24N^2/\delta))^3}{\|v\|_2^2 d^{\frac{1}{2}}}\right) \quad (23)$$

$$|V_+^{(T_2)}|, |V_-^{(T_2)}|, |V_{\xi,i}^{(T_2)}| = o(1), \quad (24)$$

When harmful overfitting occurs during the third stage, the following inequality is satisfied:

$$V_{\pm}^{(t)} = o(1) \quad (25)$$

$$|V_+^{(t)}|, |V_-^{(t)}|, |V_{\xi,i}^{(t)}| = o(1), \quad (26)$$

$$\begin{aligned} \hat{y}(f(\theta, \mathbf{X}, v)) &= \sum_{r=1}^{d_v} \left( v^T x_1 (S_{11} + S_{21}) W_{V_r}^{(t)} + v^T x_2 (S_{12} + S_{22}) W_{V_r}^{(t)} \right) \\ &\geq \log(1 + e^{-1/2}), \text{ with probability at least } \frac{1}{2}. \end{aligned} \quad (27)$$

## F UPPER BOUND OF BENIGN OVERFITTING

### F.1 STAGE I TEST LOSS

**Theorem 14** (First part of Theorem 1). *Under the same conditions as Theorem 1, When  $N \cdot \text{SNR}^2 + h(\alpha) = \Omega(1)$ , where  $h(\alpha)$  is a function related to  $\alpha$ , for any  $\epsilon > 0$ , under the assumptions above, with probability at least  $1 - \delta$ :*

$$L_{\mathcal{D}}^{0-1}((\theta(t))) \leq \frac{1}{2} + \alpha + \mathcal{O}(1) \approx \Theta(1)$$

Since  $W_Q$  and  $W_K$  are initialized as Gaussian matrices, each element of the attention score matrix  $\mathbf{X}^\top \mathbf{W}_Q \mathbf{W}_K^\top \mathbf{x}_l$  follows a zero-mean Gaussian distribution. After applying Softmax normalization, the attention weights per row approximate a uniform distribution:

$$S(\mathbf{X}^\top \mathbf{W}_Q \mathbf{W}_K^\top \mathbf{x}_l) \approx \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

For a sequence length  $M = 2$ , the attention weight per position is  $\frac{1}{2}$ .

Substituting uniform attention weights into the model:

$$\begin{aligned}
f(X, \theta^{(0)}) &\approx \sum_{l=1}^M v^\top \mathbf{W}_{V,j}^\top \mathbf{X} S (\mathbf{X}^\top \mathbf{W}_Q \mathbf{W}_K^\top \mathbf{x}_l) \\
&= \sum_{l=1}^M v^\top \mathbf{W}_{V,j}^\top \mathbf{X} \cdot \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\
&= \frac{1}{2} v^\top \sum_{l=1}^2 x_l \mathbf{W}_{V,j}^\top \\
&= \frac{1}{2} v^\top (\mu_+ W_{V,j} + \mu_- W_{V,j} + \xi W_{V,j})
\end{aligned}$$

Since  $\mu_+^\top W_{V,j}^{(0)} \sim \mathcal{N}(0, \sigma_V^2 \|\mu\|_2^2)$ ,  $\mu_-^\top W_{V,j}^{(0)} \sim \mathcal{N}(0, \sigma_V^2 \|\mu\|_2^2)$  and  $\xi^\top W_{V,j}^{(0)} \sim \mathcal{N}(0, \sigma_V^2 \sigma_p^2 d)$ , we can get the distribution of the model's initial output:

$$f(X, \theta^{(0)}) \sim \mathcal{N}\left(0, \frac{\sigma_V^2 (\|\mu\|_2^2 + \sigma_p^2 d)}{4}\right)$$

Recall from Equation 1:  $L_{\mathcal{D}}^{0-1}((\theta(0)) \leq \alpha + P(\hat{y}f(X, \theta^{(0)}) \leq 0)$ .

Since  $f(X, \theta^{(0)})$  is a symmetric distribution, hence the equation can be bounded as  $P(\hat{y}f(X, \theta^{(0)}) \leq 0) = \frac{1}{2}$ . Finally, we can obtain the test error in phase 1:

$$L_{\mathcal{D}}^{0-1}((\theta(0)) \leq \frac{1}{2} + \alpha + \mathcal{O}(1) \approx \Theta(1)$$

## F.2 STAGE II TEST LOSS

**Theorem 15** (Second part of Theorem 1). *There exists  $T_2 = \Theta\left(\frac{1}{\eta \|\mu\|_2^2 \|v\|_2^2}\right)$ , for  $t \in (T_1, T_2]$ , the test loss is upper bounded by:*

$$L_{\mathcal{D}}^{0-1}((\theta(t)) \lesssim \alpha + \exp(-\eta^4 \|\mu\|_2^8 (t - T_1)^4 \text{SNR}^2)$$

*Proof.* For the sake of convenience, we use  $(X, \hat{y}, y) \sim \mathcal{D}$  to denote the following: data point  $(X, y)$  follows distribution  $\mathcal{D}$  defined in Assumptions (1)–(7), and  $\hat{y}$  is its true label. We can write out the test error as Equation 1:

$$L_{\mathcal{D}}^{0-1}((\theta(t)) \leq \alpha + P_{(X, \hat{y}, y) \sim \mathcal{D}}(\hat{y}f(\theta, \mathbf{X}, v) \leq 0)$$

Therefore we need to calculate an upper bound for  $P_{(X, \hat{y}, y) \sim \mathcal{D}}(\hat{y}f(\theta, \mathbf{X}, v) \leq 0)$ .

To achieve this and account for the presence of label flipping, we express  $X$  as  $(\mu_+, \mu_-, \xi)$ . We can derive the following inequality from Lemma 6.

$$P(\hat{y}f(\theta, \mathbf{X}, v) \leq 0) \leq P\left(\sum_r \frac{S_{11} + S_{21}}{S_{12} + S_{22}} \left(V_{+r, \hat{y}}^{(t)} + V_{-r, \hat{y}}^{(t)}\right) \leq V_{\xi, (-r\hat{y})}^{(t)} + o(1)\right).$$

Denote  $g(\xi)$  as  $V_{\xi, (-\hat{y})}^{(t)} = \sum_r \langle v W_{-\hat{y}, r}^{(t)}, \xi \rangle$ . The initial equality is derived from equation Equation 11, while the second equality arises from the initialization of the  $\mathbf{V}$  vector.

$$\begin{aligned}
\|W_V^{(t)} v\|_2 &\leq \|W_V^{(0)} v\|_2 + \sum_{t'=0}^{t-1} \|W_V^{(t'+1)} v - W_V^{(t')} v\|_2 \\
&= \|W_V^{(0)} v\|_2 + tO\left(\eta \cdot \max\{\|\mu\|_2, \sigma_p \sqrt{d}\} \cdot \|v\|^2\right) \\
&= O\left(\sigma_V \|v\|_2 \sqrt{d} + t\eta \|v\|^2 \max\{\|\mu\|_2, \sigma_p \sqrt{d}\}\right)
\end{aligned} \tag{28}$$

Given that  $g(\xi)$  is defined as  $V_{\xi,(-\hat{y})}^{(t)} = \sum_r \langle v W_{-\hat{y},r}^{(t)}, \xi \rangle$ , and considering that  $\langle W_{-\hat{y},r}^{(t)} v, \xi \rangle \sim \mathcal{N}(0, \|W_{-\hat{y},r}^{(t)}\|_2^2 \|v\|_2^2 \sigma_p^2)$ , in the benign overfitting phase, the signal strength exceeds that of the noise. Consequently, we can deduce:

$$\begin{aligned} \mathbb{E}g(\xi) &= \sum_{r=1}^{d_V} \mathbb{E} \langle v W_{-\hat{y},r}^{(t)}, \xi \rangle \\ &= \sum_{r=1}^{d_V} \frac{\|W_{-\hat{y},r}^{(t)}\|_2 \sigma_p v}{\sqrt{2\pi}} \\ &= \frac{\sigma_p}{\sqrt{2\pi}} \sum_{r=1}^{d_V} \|W_{-\hat{y},r}^{(t)} v\|_2 \\ &\leq O\left(\sigma_V \sigma_p \|v\|_2 \sqrt{\frac{d}{2\pi}} + t\eta \sigma_p \|v\|^2 \frac{\|\mu\|}{\sqrt{2\pi}}\right) \end{aligned} \quad (29)$$

$$\begin{aligned} P(\hat{y}(f(\theta, \mathbf{X}, v)) \leq 0) &\leq P\left(V_{\xi,ir,(-\hat{y})}^{(t)} \geq \sum_r \frac{S_{11} + S_{21}}{S_{12} + S_{22}} (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)})\right) \\ &= P\left(g(\xi) - \mathbb{E}g(\xi) \geq \sum_r \frac{S_{11} + S_{21}}{S_{12} + S_{22}} (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)}) - \frac{\sigma_p}{\sqrt{2\pi}} \sum_{r=1}^{d_V} \|W_{-\hat{y},r}^{(t)} v\|_2\right) \\ &\leq \exp\left[-\frac{c_2 \left(\sum_r \frac{S_{11} + S_{21}}{S_{12} + S_{22}} (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)}) - \frac{\sigma_p \sum_{r=1}^{d_V} \|W_{-\hat{y},r}^{(t)} v\|_2}{\sqrt{2\pi}}\right)^2}{\sigma_p^2 \left(\sum_{r=1}^{d_V} \|W_{-\hat{y},r}^{(t)} v\|_2\right)^2}\right] \\ &= \exp\left[-c_3 \left(\frac{\sum_r \frac{S_{11} + S_{21}}{S_{12} + S_{22}} (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)})}{\sigma_p \sum_{r=1}^{d_V} \|W_{-\hat{y},r}^{(t)} v\|_2} - \frac{1}{\sqrt{2\pi}}\right)^2\right] \\ &\leq \exp(c_4/2\pi) \cdot \exp\left[-\frac{c_5}{2} \left(\frac{\sum_r \frac{S_{11} + S_{21}}{S_{12} + S_{22}} (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)})}{\sigma_p \sum_{r=1}^{d_V} \|W_{-\hat{y},r}^{(t)} v\|_2}\right)^2\right] \end{aligned} \quad (30)$$

The first inequalities are derived from Lemma 6, and we drop the constant  $o(1)$  to simplify the analysis. The first equality is due to Lemma 7 and Equation 29. The last inequality is due to the fact that  $(s - t)^2 \geq \frac{s^2}{2} - t^2$  for all  $s, t \geq 0$ .

When benign overfitting occurs, we are aware of the bounds of the following terms during the second stage:

- $1 - \text{Softmax}(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle)$
- $1 - \text{Softmax}(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_+^{(t)} \rangle)$
- $\text{Softmax}(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_{\xi,i}^{(t)} \rangle)$
- $\text{Softmax}(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)$

It is noted that the following inequalities hold, so we only need to calculate the last two terms:

$$\begin{aligned} 1 - \text{Softmax}(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle) &= \sum_j \text{Softmax}(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_{\xi,i}^{(t)} \rangle) \\ 1 - \text{Softmax}(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_+^{(t)} \rangle) &= \sum_j \text{Softmax}(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle) \end{aligned}$$

By substituting Equation 9 into Equation 5, and Equation 10 into Equation 6, we obtain the following inequality:

$$\begin{aligned}
& \frac{\exp(\langle q_+^{(t)}, k_{\xi,i}^{(t)} \rangle)}{\exp(\langle q_+^{(t)}, k_+^{(t)} \rangle) + \exp(\langle q_+^{(t)}, k_{\xi,i}^{(t)} \rangle)} \\
& \leq \frac{1}{c_5 \exp(\Lambda_{\xi,\pm,i}^{(t)})} \\
& \leq \frac{1}{c_5 \exp(\Lambda_{\xi,\pm,i}^{(T_1)} + \frac{\eta^2 c_8 \|\mu\|_2^4 \|v\|_2^2 d_K^{\frac{1}{2}}}{N(\log(24N^2/\delta))^2} \cdot (t - T_1)(t - T_1 - 1))} \\
& \leq \frac{1}{c_6 + \frac{\eta^2 c_{13} \|\mu\|_2^4 \|v\|_2^2 d_K^{\frac{1}{2}}}{N(\log(24N^2/\delta))^2} \cdot (t - T_1)(t - T_1 - 1)}. \tag{31}
\end{aligned}$$

$$\begin{aligned}
& \frac{\exp(\langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle)}{\exp(\langle q_{\xi,i}^{(t)}, k_+^{(t)} \rangle) + \exp(\langle q_{\xi,i}^{(t)}, k_{\xi,i'}^{(t)} \rangle)} \\
& \leq \frac{1}{c_7 \exp(\Lambda_{\xi,i,\pm,i'}^{(t)})} \\
& \leq \frac{1}{c_8 + \frac{\eta^2 C_{13} \sigma_p^2 d \|\mu\|_2^4 \|v\|_2^2 d_K^{\frac{1}{2}}}{N(\log(24N^2/\delta))^2} \cdot (t - T_1)(t - T_1 - 1)}. \tag{32}
\end{aligned}$$

By adding Equation 31 and Equation 32, we obtain the following result:

$$\begin{aligned}
\frac{S_{11} + S_{21}}{S_{12} + S_{22}} &= \frac{2}{(S_{12} + S_{22})} - 1 \\
&\geq \frac{1}{2} \left( \frac{1}{S_{22}} + \frac{1}{S_{12}} \right) - 1 \\
&\geq \frac{1}{2S_{12}} - 1 \\
&= \Theta \left( \frac{\eta^2 c_9 \|\mu\|_2^4 \|v\|_2^2 d_K^{\frac{1}{2}}}{N(\log(24N^2/\delta))^2} (t - T_1)(t - T_1 - 1) \right) \tag{33}
\end{aligned}$$

$$\begin{aligned}
\frac{\sum_r \frac{S_{11} + S_{21}}{S_{12} + S_{22}} (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)})}{\sigma_p \sum_{r=1}^{d_V} \|W_{-\hat{y},r}^{(t)} v\|_2} &\geq \frac{\frac{\eta^2 c_9 \|\mu\|_2^4 \|v\|_2^2 d_K^{\frac{1}{2}}}{N(\log(24N^2/\delta))^2} (t - T_1)(t - T_1 - 1) (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)})}{O(\sigma_p \sigma_V \|v\| \sqrt{d} + (t - T_1) \eta \sigma_p \|v\|^2 \|\mu\|)} \\
&\geq \frac{\frac{\eta^3 c_9 \|\mu\|_2^6 \|v\|_2^4 d_K^{\frac{1}{2}} (t - T_1)^2 (t - T_1 - 1)}{N(\log(24N^2/\delta))^2} - O(d^{-\frac{1}{4}}) \cdot \frac{\eta^2 c \|\mu\|_2^2 \|v\|_2^2 d_K^{\frac{1}{2}} (t - T_1)(t - T_1 - 1)}{N(\log(24N^2/\delta))^2}}{O(\sigma_p \sigma_V \|v\| \sqrt{d} + (t - T_1) \eta \sigma_p \|v\|^2 \|\mu\|)} \\
&\geq \frac{\eta^3 c_9 \|\mu\|_2^6 \|v\|_2^4 d_K^{\frac{1}{2}} (t - T_1)^2 (t - T_1 - 1)}{N(\log(24N^2/\delta))^2 \cdot O(\sigma_p \sigma_V \|v\| \sqrt{d} + (t - T_1) \eta \sigma_p \|v\|^2 \|\mu\|)} \\
&\approx \frac{\eta^3 c_9 \|\mu\|_2^6 \|v\|_2^4 d_K^{\frac{1}{2}} (t - T_1)^2 (t - T_1 - 1)}{O(N(\log(24N^2/\delta))^2 \cdot \sigma_p \eta (t - T_1) \|v\|^2 \|\mu\|)} \\
&\approx \Theta \left( \frac{\eta^2 c_9 \|\mu\|_2^5 \|v\|_2^2 d_K^{\frac{1}{2}} (t - T_1)(t - T_1 - 1)}{\sigma_p} \right) \tag{34}
\end{aligned}$$

In the subsequent equations, we substitute Equation 7, Equation 8, Equation 32, and Equation 29 to derive the initial and subsequent inequalities.

By leveraging the concept of scaling in the final steps, we then incorporate equation Equation 34 into equation Equation 30.

$$\begin{aligned}
P(\hat{y}f(\theta, \mathbf{X}, v) \leq 0) &\leq \exp(c_4/2\pi) \exp \left[ -\frac{c_5}{2} \left( \frac{\sum_r \frac{S_{11}+S_{21}}{S_{12}+S_{22}} (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)})}{\sigma_p \sum_{r=1}^{d_V} \|W_{-\hat{y},r}^{(t)} v\|} \right)^2 \right] \\
&\leq \exp\left(\frac{c_4}{2\pi}\right) \exp \left[ -\frac{c_5}{2} \Theta \left( \frac{\eta^4 c_9 \|\mu\|^{10} \|v\|^4 d_K (t-T_1)^2 (t-T_1-1)^2}{\sigma_p^2} \right) \right] \\
&\leq \exp \left[ \frac{c_4}{2\pi} - c_{10} \eta^4 \|\mu\|^8 \|v\|^4 (t-T_1)^2 (t-T_1-1)^2 d_{\text{SNR}}^2 \dot{d}_K \right] \\
&\leq \exp \left[ \frac{c_4}{2\pi} - c_{10} \eta^4 \|\mu\|^8 (t-T_1)^2 (t-T_1-1)^2 d_{\text{SNR}}^2 \dot{d}_K \right] \\
&\lesssim \alpha + \exp(-\eta^4 \|\mu\|_2^8 (t-T_1)^4 \text{SNR}^2)
\end{aligned}$$

### F.3 STAGE III TEST LOSS

**Theorem 16** (Third part of Theorem 1). *Under the same conditions as Theorem 1, there exists  $t > T_2$  such that:*

$$L_{\mathcal{D}}^{0-1}(\theta(t)) \lesssim \alpha + \exp \left( -\frac{\eta^4 (t-T_2)^4 \|\mu\|_2^6 \cdot \text{SNR}^2}{\sigma_V^2} \right).$$

*Proof.* we can get the following inequality from Lemma 6.

$$P(\hat{y}f(\theta, \mathbf{X}, v) \leq 0) \leq P \left( \sum_r \frac{S_{11}+S_{21}}{S_{12}+S_{22}} (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)}) \leq V_{\xi,ir,(-\hat{y})}^{(t)} + o(1) \right) \quad (35)$$

In the subsequent formulas, the first inequality arises because, during the benign overfitting phase, the signal memory significantly exceeds the noise memory, leading to  $S_{11} + S_{21} > S_{12} + S_{22}$ . Subsequently, by substituting Equation 14, Equation 15, Equation 16, and Equation 29, we derive the second and third inequalities. For the final inequality, we employ the Taylor series expansion of  $\log(1+x)$ . When  $x > 1$ , the  $x^2$  term dominates the expansion.

$$\begin{aligned}
\frac{\sum_r \frac{S_{11}+S_{21}}{S_{12}+S_{22}} (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)})}{\sigma_p \sum_{r=1}^{d_V} \|W_{-\hat{y},r}^{(t)} v\|_2} &\geq \frac{\sum_r (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)})}{O(\sigma_p \sigma_V \sqrt{d} \|v\|_2 + \frac{\sigma_p}{\varepsilon \|\mu\|_2})} \\
&\geq \frac{O[\log(\exp(V_+^{(T_2)}) + \eta c_{10} \cdot \|\mu\|_2^2 \|v\|_2^2 (t-T_2)) - 2 \log(O(1/\varepsilon))]}{O(\sigma_p \sigma_V \sqrt{d} \|v\|_2 + \frac{\sigma_p}{\varepsilon \|\mu\|_2})} \\
&\geq \frac{O(\log(c_{11} + 1 + \eta c_{10} \cdot \|\mu\|_2^2 \|v\|_2^2 (t-T_2)))}{O(\sigma_p \sigma_V \sqrt{d} \|v\|_2 + \frac{\sigma_p}{\varepsilon \|\mu\|_2})} \\
&\geq \frac{O[\eta c_{10} \|\mu\|_2^2 \|v\|_2^2 (t-T_2) - \frac{1}{2} \eta^2 (c_1)^2 \|\mu\|_2^4 \|v\|_2^4 (t-T_2)^2]}{O(\sigma_p \sigma_V \sqrt{d} \|v\|_2 + \frac{\sigma_p}{\varepsilon \|\mu\|_2})} \\
&\approx \frac{O(\eta^2 \|\mu\|_2^4 \|v\|_2^4 (t-T_2)^2)}{O(\sigma_p \sigma_V \sqrt{d} \|v\|_2)} \\
&\approx O \left( \frac{\eta^2 \|v\|_2^3 (t-T_2)^2 \|\mu\|_2^3}{\sigma_V} \cdot \text{SNR} \right) \quad (36)
\end{aligned}$$

By plugging Equation 36 into Equation 35, we can get :

$$\begin{aligned}
P(\hat{y}(f(\theta, \mathbf{X}, v) \leq 0) &\leq \exp(c_{12}/2\pi) \exp \left[ -\frac{c_{13}}{2} \left( \frac{\sum_r \frac{s_{11}+s_{21}}{s_{12}+s_{22}} (V_{+r,\hat{y}}^{(t)} + V_{-r,\hat{y}}^{(t)})}{\sigma_p \sum_{r=1}^{d_V} \|W_{-\hat{y},r}^{(t)} v\|} \right)^2 \right] \\
&\leq \exp(c_{12}/2\pi) \exp \left[ -\frac{c_{13}}{2} O \left( \frac{\eta^4 \|v\|^6 (t - T_2)^4 \|\mu\|^6 \cdot \text{SNR}^2}{\sigma_v^2} \right) \right] \\
&\leq \exp \left( \frac{c_{12}}{2\pi} - \frac{c_{14} \eta^4 (t - T_2)^4 \|\mu\|^6 \cdot \text{SNR}^2}{2\sigma_v^2} \right) \\
&\lesssim \alpha + \exp \left( -\frac{\eta^4 (t - T_2)^4 \|\mu\|_2^6 \cdot \text{SNR}^2}{\sigma_v^2} \right).
\end{aligned}$$

## G TEST LOSS OF HARMFUL OVERFITTING

### G.1 STAGE I TEST LOSS

**Theorem 17** (First part of Theorem 2). *When  $N^{-1} \cdot \text{SNR}^{-2} + h(\alpha) = \Omega(1)$ , where  $h(\alpha)$  is a function related to  $\alpha$ , for any  $\epsilon > 0$ , under the assumptions above, with probability at least  $1 - \delta$ , there exists  $T_1 = O \left( \frac{N}{\eta^{\frac{1}{d_K}} \|\mu\|_2^2 \|v\|_2^2} \right)$ , for  $t \in (0, T_1]$ , such that the test loss is upper bounded by:*

$$L_{\mathcal{D}}^{0-1}(\theta(t)) \leq \frac{1}{2} + \alpha + \mathcal{O}(1) \quad (37)$$

*Proof.* The proof is the same as in the first stage of benign overfitting (Theorem 14).

### G.2 STAGE II TEST LOSS UPPER BOUND

**Theorem 18** (Second part of Theorem 2). *Under the same conditions as Theorem 2, there exists  $T_2 = \Theta \left( \frac{N}{\eta \sigma_p^2 d \|v\|_2^2 \log(24N^2/\delta)} \right)$ . For  $t \in (T_1, T_2]$ , the test loss is bounded by:*

$$L_{\mathcal{D}}^{0-1}(\theta(t)) \leq \frac{1}{2} + \alpha + O \left( \frac{1}{\|\mu\|_2^2 \|v\|_2^2} + \frac{1}{\|\mu\|_2^4 \|v\|_2^4} \right).$$



*Proof.* We derive  $f(\theta, \mathbf{X}, v)$  as follows:

$$\begin{aligned}
f(\theta, \mathbf{X}, v) &= \sum_{r \in [d_V]} (v^T \mathbf{x}_1(S_{11} + S_{21}) \mathbf{W}_{Vj,r} + v^T \mathbf{x}_2(S_{12} + S_{22}) \mathbf{W}_{Vj,r}) \\
&\approx \frac{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle)}{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle) + \exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_{\xi,i}^{(t)} \rangle)} \cdot V_+^{(t)} \\
&+ \frac{\exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_+^{(t)} \rangle)}{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle) + \exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)} \cdot V_+^{(t)} \\
&+ \frac{\exp(\langle \mathbf{q}_-^{(t)}, \mathbf{k}_-^{(t)} \rangle)}{\exp(\langle \mathbf{q}_-^{(t)}, \mathbf{k}_-^{(t)} \rangle) + \exp(\langle \mathbf{q}_-^{(t)}, \mathbf{k}_{\xi,i}^{(t)} \rangle)} \cdot V_-^{(t)} \\
&+ \frac{\exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_+^{(t)} \rangle)}{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle) + \exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)} \cdot V_-^{(t)} \\
&+ \frac{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_{\xi,i}^{(t)} \rangle)}{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle) + \exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)} \cdot V_{\xi,i}^{(t)} \\
&+ \frac{\exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)}{\exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_+^{(t)} \rangle) + \exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)} \cdot V_{\xi,i}^{(t)}.
\end{aligned}$$

To simplify the analysis, let's define:

$$\begin{aligned}
A &= \frac{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle)}{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle) + \exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_{\xi,i}^{(t)} \rangle)} \cdot V_+^{(t)} \\
B &= \frac{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle)}{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle) + \exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)} \cdot V_+^{(t)} \\
C &= \frac{\exp(\langle \mathbf{q}_-^{(t)}, \mathbf{k}_-^{(t)} \rangle)}{\exp(\langle \mathbf{q}_-^{(t)}, \mathbf{k}_-^{(t)} \rangle) + \exp(\langle \mathbf{q}_-^{(t)}, \mathbf{k}_{\xi,i}^{(t)} \rangle)} \cdot V_-^{(t)} \\
D &= \frac{\exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_+^{(t)} \rangle)}{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle) + \exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)} \cdot V_-^{(t)} \\
E &= \frac{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_{\xi,i}^{(t)} \rangle)}{\exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_+^{(t)} \rangle) + \exp(\langle \mathbf{q}_+^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)} \cdot V_{\xi,i}^{(t)} \\
F &= \frac{\exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)}{\exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_+^{(t)} \rangle) + \exp(\langle \mathbf{q}_{\xi,i}^{(t)}, \mathbf{k}_{\xi,i'}^{(t)} \rangle)} \cdot V_{\xi,i}^{(t)} \\
f(\theta, \mathbf{X}, v) &= A + B + C + D + E + F
\end{aligned}$$

By plugging Equation 19–Equation 25 into the above definition, the following inequality holds:

$$\begin{aligned}
A &\leq O\left(\frac{\sigma_p^2 d (\log(24N^2/\delta))^3}{\|\mu\|_2^2 \|v\|_2^2 d^{\frac{1}{2}}}\right) \left(O(d^{-\frac{1}{4}}) + \frac{\eta c_{15} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N}\right) \\
&= O\left(\frac{\sigma_p^2 d^{\frac{1}{4}} (\log(24N^2/\delta))^3}{\|\mu\|_2^2 \|v\|_2^2}\right) + O\left(\frac{\eta c_{15} \sigma_p^4 d^{\frac{3}{2}} (\log(24N^2/\delta))^3 (t - T_1)}{\|\mu\|_2^2 N \|v\|_2^2}\right) \\
B &\leq \left(1 - O\left(\frac{\sigma_p^2 d (\log(24N^2/\delta))^3}{\|\mu\|_2^2 \|v\|_2^2 d^{\frac{1}{2}}}\right)\right) \left(O(d^{-\frac{1}{4}}) + \frac{\eta c_{15} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N}\right) \\
&\approx O(d^{-\frac{1}{4}}) + \frac{\eta c_{15} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N} \\
C &\leq O\left(\frac{\sigma_p^2 d^{\frac{1}{4}} (\log(24N^2/\delta))^3}{\|\mu\|_2^2 \|v\|_2^2}\right) + O\left(\frac{\eta c_{15} \sigma_p^4 d^{\frac{3}{2}} (\log(24N^2/\delta))^3 (t - T_1)}{\|\mu\|_2^2 N \|v\|_2^2}\right) \\
D &\leq O(d^{-\frac{1}{4}}) + \frac{\eta c_{15} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N} \\
E &\leq O\left(\frac{(\log(24N^2/\delta))^3}{\|v\|_2^2 d^{\frac{1}{2}}}\right) \frac{\eta c_{14} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N} \\
&= O\left(\frac{\eta c_{14} \sigma_p^2 d^{\frac{1}{2}} (\log(24N^2/\delta))^3 (t - T_1)}{N}\right) \\
F &\leq \left(1 - O\left(\frac{(\log(24N^2/\delta))^3}{\|v\|_2^2 d^{\frac{1}{2}}}\right)\right) \frac{\eta c_{14} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N} \\
&\approx \frac{\eta c_{14} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N}.
\end{aligned}$$

Calculate the sum of the absolute values presented above, we can get:

$$\begin{aligned}
&f(\theta, \mathbf{X}, v) \\
&\leq O\left(\frac{\sigma_p^2 d^{\frac{1}{4}} (\log(24N^2/\delta))^3}{\|\mu\|_2^2 \|v\|_2^2}\right) + O\left(\frac{\eta c_{15} \sigma_p^4 d^{\frac{3}{2}} (\log(24N^2/\delta))^3 (t - T_1)}{\|\mu\|_2^2 N \|v\|_2^2}\right) \\
&+ O\left(d^{-\frac{1}{4}} + 2\frac{\eta c_{15} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N}\right) + O\left(\frac{\eta c_{14} \sigma_p^2 d^{\frac{1}{2}} (\log(24N^2/\delta))^3 (t - T_1)}{N} + \frac{\eta c_{14} \sigma_p^2 d \|v\|_2^2 (t - T_1)}{N}\right) \\
&\leq O\left(\frac{d^{\frac{1}{4}} (\log(N^2/\delta))^3}{\|\mu\|_2^2 \|v\|_2^2}\right) + O\left(\eta d^{\frac{1}{4}} (\log(N^2/\delta))^3 (t - T_1) \left(\frac{1}{\|\mu\|_2^2 N \|v\|_2^2} + \frac{1}{N}\right)\right) + O\left(\frac{\eta d \|v\|_2^2 (t - T_1)}{N}\right) \\
&\leq O\left(\frac{d^{\frac{1}{4}} (\log(N^2/\delta))^3}{\|\mu\|_2^2 \|v\|_2^2}\right) + O\left(\frac{\eta d^{\frac{1}{4}} (\log(N^2/\delta))^3 (t - T_1)}{\|\mu\|_2^2 N \|v\|_2^2}\right) + \\
&O\left(\frac{\eta d^{\frac{1}{4}} (\log(N^2/\delta))^3 (t - T_1)}{N}\right) + O\left(\frac{\eta d \|v\|_2^2 (t - T_1)}{N}\right) \\
&\approx O\left(\frac{\eta d^{\frac{1}{4}} (\log(N^2/\delta))^3 (t - T_1)}{\|\mu\|_2^2 N \|v\|_2^2}\right) + O\left(\frac{\eta d^{\frac{1}{4}} (\log(N^2/\delta))^3 (t - T_1)}{N}\right). \tag{38}
\end{aligned}$$

Equation Equation 38 can be decomposed into two items, which are called the signal residual and the noise-dominated terms.

The term  $O\left(\frac{\eta d^{1/4} (\log(N^2/\delta))^3 (t - T_1)}{\|\mu\|_2^2 N \|v\|_2^2}\right)$  represents the signal residual, which reflects the model's residual ability to capture signal features during the harmful overfitting phase, suppressed by noise dominance. The term  $O\left(\frac{\eta d^{1/4} (\log(N^2/\delta))^3 (t - T_1)}{N}\right)$  is the noise-dominated term (gradient coupling),

which reflects the model's failure to distinguish signals from noise, leading the attention weights  $S_{12}$  and  $S_{22}$  to favor noise features in low-SNR regimes, and  $(t - T_1)$  reflects the linear accumulation of noise overfitting with training steps.

Since the noise-dominated term has a mean of zero and its variance is related to the noise magnitude  $\sigma_p\sqrt{d}$ , and the signal residual term has a mean of zero (because the signal is suppressed by noise), and its magnitude is suppressed by  $\|\mu\|_2^2$ , the distribution of  $f$  is approximately a Gaussian distribution with mean zero and variance  $\sigma_f^2$ :

$$\begin{aligned}\sigma_f^2 &= \Theta \left( \frac{\eta^2 d^{1/2} (\log(N^2/\delta))^6 (t - T_1)^2}{\|\mu\|_2^4 N^2 \|v\|_2^4} + \frac{\eta^2 d^{1/2} (\log(N^2/\delta))^6 (t - T_1)^2}{N^2} \right) \\ &= \Theta \left( \frac{\eta^2 d^{1/2} (\log(N^2/\delta))^6 (t - T_1)^2}{N^2} \left( \frac{1}{\|\mu\|_2^4 \|v\|_2^4} + 1 \right) \right).\end{aligned}\quad (39)$$

Due to the symmetry of  $\hat{y}f$ , if  $f$  is completely dominated by noise, we can get

$$P(\hat{y}f \geq 0) = P(\hat{y}f \leq 0) = \frac{1}{2}$$

But in practice, the signal residual term introduces a small bias during the second phase.

Let the mean of the signal residual be

$$E[f_s] = O \left( \frac{\eta d^{\frac{1}{4}} (\log(N^2/\delta))^3 (t - T_1)}{\|\mu\|_2^2 N \|v\|_2^2} \right) \quad (40)$$

we have

$$\hat{y}f \sim \mathcal{N}(E[f_s], \sigma_f^2) \quad (41)$$

By applying the Taylor expansion to correct the probability, and the cumulative distribution function of the Gaussian distribution is denoted by  $\Phi$ . Using this, we can rewrite:

$$\begin{aligned}P(\hat{y}f \leq 0) &\approx \Phi \left( -\frac{\mathbb{E}[f_s]}{\sigma_f} \right) \\ &\approx \frac{1}{2} - \frac{\mathbb{E}[f_s]}{\sigma_f \sqrt{2\pi}} + O \left( \frac{\mathbb{E}[f_s]^2}{\sigma_f^2} \right)\end{aligned}\quad (42)$$

We plug Equation 41 and Equation 42 into Equation 44 by using the standard Gaussian distribution function and correcting the probability term expansion by a second-order Taylor approximation. And we can plug Equation 39 and Equation 40 into Equation 45 that

$$L_D^{0-1}((\theta(t))) \quad (43)$$

$$\begin{aligned}&= \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}} [y \neq \text{sign}(f(\theta, \mathbf{X}, v))] \\ &= \alpha \cdot P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (\hat{y}f(\theta, \mathbf{X}, v) \geq 0) + (1 - \alpha) \cdot P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (\hat{y}f(\theta, \mathbf{X}, v) \leq 0)\end{aligned}\quad (44)$$

$$\begin{aligned}&= \alpha \left( 1 - \left[ \frac{1}{2} - \frac{E[f_s]}{\sigma_f \sqrt{2\pi}} + O \left( \frac{E[f_s]^2}{\sigma_f^2} \right) \right] \right) + (1 - \alpha) \left( \frac{1}{2} - \frac{E[f_s]}{\sigma_f \sqrt{2\pi}} + O \left( \frac{E[f_s]^2}{\sigma_f^2} \right) \right) \\ &\leq \frac{1}{2} + \alpha + \frac{\mathbb{E}[f_s]}{\sigma_f \sqrt{2\pi}} + O \left( \frac{\mathbb{E}[f_s]^2}{\sigma_f^2} \right)\end{aligned}\quad (45)$$

$$\leq \frac{1}{2} + \alpha + O \left( \frac{1}{\|\mu\|_2^2 \|v\|_2^2} \right) + O \left( \frac{1}{\|\mu\|_2^4 \|v\|_2^4} \right)$$

### G.3 STAGE II TEST LOSS LOWER BOUND

**Theorem 19** (Second part of Theorem 2). *Under the same conditions as Theorem 2, when  $N^{-1} \cdot \text{SNR}^{-2} = \Omega(1)$  and  $\Omega(1)$  is related to  $\alpha$ , there exists  $T_2 = \Theta \left( \frac{N}{\eta \sigma_p^2 d \|v\|_2^2 \log(24N^2/\delta)} \right)$ . For  $t \in (T_1, T_2]$ , the test error is:*

$$L_D^{0-1}((\theta(t))) \geq \frac{1}{2} - O \left( \frac{1}{\|\mu\|_2^2 \|v\|_2^4} \right)$$

*Proof.*

$$\begin{aligned}
f_j(\theta, \mathbf{X}, v) &= \sum_{r \in [d_V]} \left( v^\top \langle \mathbf{W}_{Vj,r}, \mathbf{x}_1 \rangle (S_{11} + S_{21}) + v^\top \langle \mathbf{W}_{Vj,r}, \mathbf{x}_2 \rangle (S_{12} + S_{22}) \right) \\
&= \sum_j [(S_{11} + S_{21})(V_+^{(t)} + V_-^{(t)}) + (S_{12} + S_{22})V_\xi^{(t)}] \\
&\geq \sum_j S_{11}(V_+^{(t)} + V_-^{(t)}) + \sum_j S_{22}V_\xi^{(t)}
\end{aligned} \tag{46}$$

The first summation of Equation 46 is the residual signal, and the second summation is the noise term  $S_{22}V_\xi^{(t)} \sim \mathcal{N}(0, \sigma_f^2)$  and the second summation is the dominant term.

By plugging Equation 18, Equation 20, Equation 22, Equation 23 into Equation 46, we have:

$$E = \sum_r S_{11}V_{\pm,r}^{(t)} = O\left(\frac{\sigma_p^2 d^{1/4} (\log(N^2/\delta))^3}{\|\mu\|_2^2 \|v\|_2^2}\right) \tag{47}$$

$$\sigma_f^2 = \Theta\left(\frac{\eta^2 \sigma_p^4 d^2 \|v\|_2^4 (t - T_1)^2}{N^2}\right) \tag{48}$$

We can get from  $f(\theta, \mathbf{X}, v) \sim \mathcal{N}(E, \sigma_f^2)$  that

$$\begin{aligned}
L_{\mathcal{D}}^{0-1}((\theta(t))) &= \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}} [y \neq \text{sign}(f(\theta, \mathbf{X}, v))] \\
&= \alpha \cdot P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (\hat{y}f(\theta, \mathbf{X}, v) \geq 0) + (1 - \alpha) \cdot P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}} (\hat{y}f(\theta, \mathbf{X}, v) \leq 0) \\
&= \alpha \cdot P(\mathcal{N}(E, \sigma_f^2) \leq 0) + (1 - \alpha) \cdot P(\mathcal{N}(-E, \sigma_f^2) \geq 0) \\
&\geq \alpha \cdot \Phi\left(-\frac{E}{\sigma_f}\right) + (1 - \alpha) \cdot \Phi\left(-\frac{E}{\sigma_f}\right) \\
&\geq \alpha \left(\frac{1}{2} - \frac{E}{\sigma_f \sqrt{2\pi}}\right) + (1 - \alpha) \left(\frac{1}{2} - \frac{E}{\sigma_f \sqrt{2\pi}}\right) \\
&= \frac{1}{2} - \frac{E}{\sigma_f \sqrt{2\pi}}
\end{aligned} \tag{49}$$

Where  $\Phi(-X)$  is the cumulative distribution function of the standard normal distribution ; the third inequality holds if  $\Phi(-x) \geq \frac{1}{2} - \frac{x}{\sqrt{2\pi}}$ .

By plugging Equation 47 and Equation 48 into Equation 49, we obtain that

$$\begin{aligned}
L_{\mathcal{D}}^{0-1}((\theta(t))) &\geq \frac{1}{2} - O\left(\frac{(\log N^2/\delta)^3}{\eta \|\mu\|_2^2 \|v\|_2^4 d^{3/4} (t - T_1)}\right) \\
&\approx \frac{1}{2} - O\left(\frac{1}{\|\mu\|_2^2 \|v\|_2^4}\right)
\end{aligned}$$

#### G.4 STAGE III TEST LOSS

**Theorem 20** (Third part of Theorem 2). *Under the same conditions as Theorem 2, when  $N^{-1} \cdot \text{SNR}^{-2} = \Omega(1)$  and  $\Omega(1)$  is related to  $\alpha$ , the test loss is that:*

$$L_{\mathcal{D}}^{0-1}((\theta(t))) \geq \frac{1}{2}$$

*Proof.* We have

$$\begin{aligned}
& L_{\mathcal{D}}^{0-1}((\theta(t)) \\
&= \mathbb{P}_{(\mathbf{x}, y) \sim \mathcal{D}} [y \neq \text{sign}(f(\theta, \mathbf{X}, v))] \\
&= \alpha \cdot P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}}(\hat{y}f(\theta, \mathbf{X}, v) \geq 0) + (1 - \alpha) \cdot P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}}(\hat{y}f(\theta, \mathbf{X}, v) \leq 0) \\
&\geq \alpha \cdot P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}}(\hat{y}f(\theta, \mathbf{X}, v) \geq \log(\frac{1 + 2e^{\frac{1}{2}}}{1 + e^{\frac{1}{2}}})) + (1 - \alpha) \cdot P_{(\mathbf{x}, \hat{y}, y) \sim \mathcal{D}}(\hat{y}f(\theta, \mathbf{X}, v) \leq \log(\frac{1 + 2e^{\frac{1}{2}}}{1 + e^{\frac{1}{2}}})) \\
&\geq \alpha \cdot \frac{1}{2} + (1 - \alpha) \cdot \frac{1}{2} \\
&\geq \frac{1}{2},
\end{aligned}$$

where the second equation is derived from the total probability theorem and the first and second inequalities are derived by Equation (27).