A Theoretical Analysis of Mamba's Training Dynamics: Filtering Relevant Features for Generalization in State Space Models

Anonymous authors
Paper under double-blind review

Abstract

The recent empirical success of Mamba and other selective state space models (SSMs) has renewed interest in non-attention architectures for sequence modeling, yet their theoretical foundations remain underexplored. We present a first-step analysis of generalization and learning dynamics for a simplified but representative Mamba block: a single-layer, single-head selective SSM with input-dependent gating, followed by a two-layer MLP trained via gradient descent (GD). Our study adopts a structured data model with tokens that include both class-relevant and class-irrelevant patterns under token-level noise and examines two canonical regimes: majority-voting and locality-structured data sequences. We prove that the model achieves guaranteed generalization by establishing non-asymptotic sample complexity and convergence rate bounds, which improve as the effective signal increases and the noise decreases. Furthermore, we show that the gating vector aligns with class-relevant features while ignoring irrelevant ones, thereby formalizing a feature-selection role similar to attention but realized through selective recurrence. Numerical experiments on synthetic data justify our theoretical results. Overall, our results provide principled insight into when and why Mamba-style selective SSMs learn efficiently, offering a theoretical counterpoint to Transformer-centric explanations.

1 Introduction

Transformers (Vaswani et al., 2017) have become the mainstream framework in large language models (Achiam et al., 2023; Guo et al., 2025; Brown et al., 2020; Touvron et al., 2023). However, due to the quadratic time and memory complexity introduced by the attention mechanism with respect to input length (Gu & Dao, 2023; Dao & Gu, 2024), Transformers are inefficient when handling long input sequences. Recently, State Space Models (SSMs) (Gu & Dao, 2023; Dao & Gu, 2024; Zhu et al., 2024; Wang et al., 2024a; Behrouz & Hashemi, 2024; Liu et al., 2024; Wang et al., 2024b) have shown competitive or superior performance to Transformers across domains such as language (Gu & Dao, 2023), vision (Zhu et al., 2024; Liu et al., 2024), graphs (Wang et al., 2024a; Behrouz & Hashemi, 2024), audio (Yadav & Tan, 2024), and reinforcement learning (Lu et al., 2023). SSMs has brought many advantages absent in Transformer-based models, such as linear computational complexity and hardware-friendly properties that enable efficient parallelization. Among these models, Mamba (Gu & Dao, 2023) proposes a selection mechanism, which parameterizes the SSM with the input, which allows the model to dynamically retain or discard relevant and irrelevant information. This enables the Mamba model to achieve performance comparable to Transformer-based models in long-text modeling as well as tasks such as visual classification and dense prediction (Zhu et al., 2024; Liu et al., 2024), but in a more efficient manner.

Although recent work has primarily focused on the empirical performance of Mamba and its architectural comparisons with other models, the theoretical understanding of Mamba remains less investigated. In addition, recent empirical evidence shows that Mamba's success is highly sensitive to hyperparameter tuning (Okpekpe & Orvieto, 2025). Such dependence on fragile optimization choices raises fundamental questions about why and when Mamba succeeds. These include fundamental inquiries such as:

- *Under what conditions can a Mamba be trained to achieve satisfactory generalization?*
- *How is the selection mechanism implemented through Mamba's components?*

Existing theoretical studies on Mamba or related SSMs mainly focus on the expressive power and the mechanisms of optimal parameters. Orvieto et al. (2024) and Nishikawa & Suzuki prove SSMs augmented with MLPs are universal approximators of regular functionals and can mimic token selection dynamically. Muca Cirone et al. (2024) and Huang et al. (2025) show that Mamba has stronger expressive power than its diagonal SSM predecessor, especially in approximating discontinuous functions. Li et al. (2024c) and Li et al. (2025b) respectively prove that two simplified SSMs, H3 and GLA, implicitly perform weighted preconditioned GD at the global minima of in-context learning problems when input with context examples. However, these works do not explain whether the selection mechanisms and advantages of Mamba can actually be obtained through practical training. Moreover, these studies do not analyze the generalization ability of Mamba models.

Contributions of this paper. In this work, we study a nonlinear neural model composed of a one-layer Mamba block and a two-layer perceptron, which is simplified but sufficiently representative to reflect the gating structure in Mamba. By assuming the presence of the class-relevant feature that influence the label and class-irrelevant features that do not, we respectively formulate majority-voting and locality-structured data, whose labels depend on the proportion and the spatial/temporal locality of a certain class-relevant feature in the data. To the best of our knowledge, this work provides the first theoretical analysis of Mamba's training dynamics with input-dependent gating, together with generalization guarantees under the two structured data regimes. The highlights of our technical contributions include:

First, we develop a general theoretical framework for analyzing gated architectures trained with gradient descent on structured data. Our analysis explains how the selection mechanism within Mamba interacts with data structure to enable efficient learning and guaranteed generalization, complementing prior results that focus mainly on attention-based models.

Second, we provide a theoretical characterization of the gating mechanism in Mamba. We show that the gating parameter vector is trained to amplify class-relevant features while ignoring class-irrelevant ones, thereby formalizing the intuition that the gating network dynamically allocates capacity to informative patterns.

Third, we establish the sample complexity and the required number of iterations for two canonical data types: majority-voting and locality-structured data sequences. For majority-voting data, these bounds scale with the gap between the class-relevant and confusion features; for locality-structured data, they depend on the concentration of class-relevant tokens. In both regimes, stronger signal and lower token-level noise yield faster convergence and better generalization.

1.1 Related Work

State Space Models (SSMs). Building upon the early S4 models (Gu et al., 2021; Gupta et al., 2022; Smith et al.), Mamba (Gu & Dao, 2023; Dao & Gu, 2024) introduced input-dependent gating to dynamically select relevant features, achieving remarkable performance in NLP and CV. Recent works extending SSMs beyond 1D sequences have highlighted the importance of input ordering and scanning. For example, VMamba (Liu et al., 2024) introduces SS2D, employing multiple scanning routes to bridge sequential structure with the non-sequential nature of vision inputs, while Graph Mamba (Wang et al., 2024a; Behrouz & Hashemi, 2024) adapts SSMs to non-Euclidean domains by leveraging graph connectivity. Collectively, these works show that the effectiveness of SSMs is tightly linked to input ordering and scanning strategies, a challenge that also motivates our theoretical analysis.

Theoretical Analysis of SSMs. Theoretical understanding of Mamba is still in its early stages and has so far centered primarily on approximation theory, such as connections to attention-like mechanisms (Dao & Gu, 2024; Nishikawa & Suzuki), expressive capacity (Cohen-Karlik et al., 2025; Huang et al., 2025; Muca Cirone et al., 2024; Bao et al., 2025), long-range dependency modeling (Ma & Najarian, 2025; Yu & Erichson, 2025), and the comparison with Transformers Jelassi et al. (2024). Beyond approximation theory, several recent works have begun examining optimization and generalization aspects of SSMs. Honarpisheh et al. (2025) provide a generalization-error bound based on Rademacher complexity; Slutzky et al. (2024) study implicit bias under a teacher–student setting and show that gradient flow can converge to a low-rank solution, though their model does not incorporate Mamba's input-dependent gating. These analyses provide valuable intuition about the representational strengths and weaknesses of Mamba blocks. However, such results remain largely structural: they establish only the existence of desirable representations, without explaining whether or how these capabilities arise during training, particularly under Mamba's unique mechanism. Motivated by this gap, we focus on studying how Mamba interacts with structured data, with particular emphasis on the role of its gating mechanism in shaping training dynamics and generalization.

Feature Learning Framework. Recent theoretical studies of deep learning have shifted focus from the NTK framework (Jacot et al., 2018; Allen-Zhu et al., 2019b; Arora et al., 2019; Wen & Li, 2021) to the feature-learning framework, where data is modeled as a combination of features and the central question is how neural networks align with these features. Much of the recent work has concentrated on transformers (Li et al., 2023a; 2024b; 2023b; 2025a), feedforward neural networks (Bakshi et al., 2019; Arora et al., 2019), and graph neural networks (Zhang et al., 2023; Li et al., 2024a). Due to the inherent complexity of non-convex optimization and modern architectures, prior works on feature learning have, to the best of our knowledge, focused primarily on shallow networks. In this work, we extend the structural data model to analyze the training dynamics of a shallow yet representative Mamba block, with particular emphasis on how its data-dependent gating mechanism shapes learning and generalization.

2 Preliminaries

Structured state space models (S4). For the t-th token, e.g., at time step t, let $\boldsymbol{x}_t \in \mathbb{R}^d$ be the input, $\boldsymbol{H}_t \in \mathbb{R}^{N\times d}$ denote the corresponding hidden state, and $\boldsymbol{y}_t \in \mathbb{R}^d$ denote the output. Let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ and $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^N$ be model parameters. The discrete-time SSM is given by

$$\boldsymbol{H}_t = \overline{\boldsymbol{A}}\boldsymbol{H}_{t-1} + \bar{\boldsymbol{b}}\,\boldsymbol{x}_t^\top, \qquad y_t = \boldsymbol{H}_t^\top \boldsymbol{c}, \tag{1}$$

where $\overline{\boldsymbol{A}} = \exp(\Delta\boldsymbol{A})$ and $\bar{\boldsymbol{b}} = \boldsymbol{A}^{-1}(\exp(\Delta\boldsymbol{A}) - \boldsymbol{I})\,\boldsymbol{b}$ with $\Delta > 0$ as the sampling step.

Mamba. To overcome the data-independence of S4, recent work introduced *selective state space models* (Gu & Dao, 2023), where key parameters are made input-dependent. Concretely, given input tokens $\boldsymbol{x}_t \in \mathbb{R}^d$, the recurrence parameters are defined as

$$\boldsymbol{b}_t = \boldsymbol{W}_B^\top \boldsymbol{x}_t, \qquad \Delta_t = \log\left(1 + e^{\boldsymbol{w}_\Delta^\top \boldsymbol{x}_t}\right), \qquad \boldsymbol{c}_t = \boldsymbol{W}_C^\top \boldsymbol{x}_t, \tag{2}$$

with learnable projections $\boldsymbol{W}_B, \boldsymbol{W}_C \in \mathbb{R}^{d\times N}$ and a gating vector $\boldsymbol{w}_\Delta \in \mathbb{R}^d$. The discretization then yields two input-dependent gates,

$$\bar{\boldsymbol{b}}_t = \sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_t)\,\boldsymbol{b}_t, \qquad \bar{a}_t = 1 - \sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_t), \tag{3}$$

which respectively control the input update and the carry-over of past states. With hidden state $\boldsymbol{H}_t \in \mathbb{R}^{N\times d}$, the recurrence becomes

$$\boldsymbol{H}_t = \bar{a}_t \boldsymbol{H}_{t-1} + \bar{\boldsymbol{b}}_t \boldsymbol{x}_t^\top. \tag{4}$$

Mamba output at token t is given by:

$$\begin{aligned}\boldsymbol{y}_t(\boldsymbol{X}) := \boldsymbol{H}_t^\top \boldsymbol{c}_t &= \sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_t)\,(\boldsymbol{W}_B^\top \boldsymbol{x}_t)^\top(\boldsymbol{W}_C^\top \boldsymbol{x}_t)\,\boldsymbol{x}_t + \left(1 - \sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_t)\right)\boldsymbol{H}_{t-1}^\top \boldsymbol{c}_t \\ &= \sum_{s=1}^{t}\left(\prod_{j=s+1}^{t}\left(1 - \sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_j)\right)\right)\cdot \sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_s)\,(\boldsymbol{W}_B^\top \boldsymbol{x}_s)^\top(\boldsymbol{W}_C^\top \boldsymbol{x}_t)\,\boldsymbol{x}_s.\end{aligned} \tag{5}$$

Connection and Difference with Transformer. The Mamba formulation reveals a natural analogy to attention mechanisms (Dao & Gu, 2024; Sieber et al., 2024). In particular, the input-dependent matrices $\boldsymbol{W}_B$ and $\boldsymbol{W}_C$ can be interpreted as counterparts to queries and keys in the self-attention, while the gating term $\sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_t)$ acts as a dynamic weight controlling how past information contributes to the current output (Dao & Gu, 2024). This structure yields a formulation closely related to gated linear attention (Yang et al.; Li et al., 2025b; Lu et al., 2025), thereby highlighting a connection between SSM and Transformer models. Meanwhile, Mamba departs from these architectures: its gating mechanism is defined through *multiplicative interactions*, effectively involving products of successive terms. This nonlinearity makes the analysis of Mamba substantially different and more challenging than that of gated linear attention. Unlike additive attention-style weighting, Mamba's gating introduces input-dependent multiplicative modulation in the selection mechanism. This alters how information is propagated through the model and results in training dynamics that differ from attention-based architectures.

3 Problem Formulation

Following existing works (Brutzkus & Globerson, 2021; Zhang et al., 2023; Li et al., 2023a), we consider a binary classification problem with training data $\{(\boldsymbol{X}^{(n)}, z^{(n)})\}_{n=1}^{N}$ sampled i.i.d. from an unknown distribution $\mathcal{D}$, where $z^{(n)} \in \{+1, -1\}$ is the label. The goal is to learn a model that maps $\boldsymbol{X}$ to z for any $(\boldsymbol{X}, z) \sim \mathcal{D}$. Each input takes the form $\boldsymbol{X}^{(n)} = [\boldsymbol{x}_1^{(n)}, \dots, \boldsymbol{x}_L^{(n)}] \in \mathbb{R}^{d \times L}$ with L tokens, where each token is d-dimensional. Tokens can be image patches (Dosovitskiy et al., 2021; Touvron et al., 2021) or subwords (Sennrich et al., 2016; Kudo & Richardson, 2018).

Learning is performed using a simplified Mamba block formulated by (5), followed by a two-layer MLP. Formally, the model output can be expressed as

$$F(\boldsymbol{X}) = \frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{m} v_i \phi \left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l(\boldsymbol{X}) \right), \tag{6}$$

where $\phi(\cdot)$ denotes the ReLU function, and $\boldsymbol{W}_O \in \mathbb{R}^{m \times d}$, with $\boldsymbol{W}_{O(i,\cdot)}$ being the i-th row of $\boldsymbol{W}_O$. Here, $\boldsymbol{y}_l(\boldsymbol{X})$ corresponds to the l-th token output of Mamba, as defined in (5). In addition, v_i represents the output-layer weight for the i-th hidden unit.

Model Training. Let $\boldsymbol{\Psi} = (\boldsymbol{v}, \boldsymbol{W}_O, \boldsymbol{w}_\Delta, \boldsymbol{W}_B, \boldsymbol{W}_C)$ denote the set of model parameters. The training process is to minimize the empirical risk $f_N(\boldsymbol{\Psi})$,

$$\min_{\boldsymbol{\Psi}} \; f_N(\boldsymbol{\Psi}) = \frac{1}{N} \sum_{n=1}^{N} \ell(\boldsymbol{X}^{(n)}, z^{(n)}; \boldsymbol{\Psi}), \tag{7}$$

where $\ell(\boldsymbol{X}^{(n)}, z^{(n)}; \boldsymbol{\Psi})$ is the hinge loss function, i.e.,

$$\ell(\boldsymbol{X}^{(n)}, z^{(n)}; \boldsymbol{\Psi}) = \max\{0, 1 - z^{(n)} \cdot F(\boldsymbol{X}^{(n)})\}. \tag{8}$$

The empirical risk minimization problem in (7) is solved via gradient descent (GD). For the theoretical analysis, we consider the full batch gradient update with a learning rate of η at each iteration $t = 1, 2, \dots, T$. Each entry of $\boldsymbol{W}_O \in \mathbb{R}^{m \times d}$ is independently initialized from $\mathcal{N}(0, c_0^2)$, and $\boldsymbol{w}_\Delta$ is initialized to 0. Similarly, each entry of $\boldsymbol{v} \in \mathbb{R}^m$ is independently sampled from $\{+\frac{1}{\sqrt{m}}, -\frac{1}{\sqrt{m}}\}$ with equal probability. $\boldsymbol{v}$ is fixed during training, as in other theoretical works (Allen-Zhu & Li, 2022; Arora et al., 2019; Karp et al., 2021; Allen-Zhu et al., 2019a; Li et al., 2023a; 2024b).

Generalization. The generalization error of the learned model $\boldsymbol{\Psi}$ is evaluated using the population risk $f(\boldsymbol{\Psi})$, defined as

$$f(\boldsymbol{\Psi}) = f(\boldsymbol{v}, \boldsymbol{W}_O, \boldsymbol{w}_\Delta, \boldsymbol{W}_B, \boldsymbol{W}_C) = \mathbb{E}_{(\boldsymbol{X}, z) \sim \mathcal{D}} \, \ell(\boldsymbol{X}, z). \tag{9}$$

4 Theoretical Results

Table 1: Some important notations

$\boldsymbol{y}_l$	Mamba block output at token position l	N	Number of samples in a batch
d	Embedding dimension	m	The number of neurons in $\boldsymbol{W}_O$
η	Learning rate for gradient descent	L	Length of the squence
$\Delta L^{+}_{\boldsymbol{o}_+}$	Concentration of class-relevant tokens	α_r	Average fraction of class-relevant tokens
$\Delta L^{-}_{\boldsymbol{o}_+}$	Dispersion of the confusion tokens	α_c	Average fraction of confusion tokens

4.1 Key Takeaways and Insights of the Findings

Before formally presenting our data assumptions and theoretical results, we first summarize key insights derived from our theoretical findings. We consider a data model where tokens are noisy versions of *class-relevant* patterns that determine the data label and *class-irrelevant* patterns that do not affect the label. Some important parameters are summarized in Table 1.

(T1). Convergence and sample complexity analysis of GD to achieve guaranteed generalization. We introduce a theoretical framework for analyzing gated architectures with structured data.

Compared with existing results on attention-based models, our framework captures the role of the gating mechanism inside the Mamba block and structured weight interactions, explaining how gradient descent (GD) exploits data structure to improve learning efficiency. Based on this analysis, we show that a model trained with GD achieves guaranteed generalization with high probability over the randomness of the data and the GD updates.

(T2). Theoretical characterization of the gating mechanism in Mamba. We prove that during training, the gating network learns to prioritize class-relevant features while ignoring irrelevant ones. In the majority-voting regime, the gating vector $\boldsymbol{w}_\Delta$ becomes increasingly aligned with class-relevant directions: gradients along those directions grow, while those along irrelevant features remain negligible. In the locality-structured data regime, learning emphasizes the elimination of irrelevant features. Their directions are consistently pushed downward by negative updates, while the directions of relevant features remain nearly unchanged. This occurs because class-relevant and confusion tokens appear in equal proportion, so the model cannot amplify the former and instead reduces the influence of the latter. These dynamics strengthen informative tokens and weaken uninformative ones, inducing effective sparsity in the activations and formalizing the intuition that Mamba allocates capacity to the most important patterns in the data.

(T3). Larger fraction or higher local concentration of class-relevant features accelerates learning. We show that both the number of iterations and the sample complexity required for generalization depend on the discriminative structure of the data and the token-level noise τ. For majority-voting data, these quantities scale as $(\alpha_r - \alpha_c)^{-2}$, so learning is faster when the fraction of class-relevant tokens is larger. For locality-structured data, the number of iterations scales as $\big[(\frac{1}{2})^{\Delta L^+_{\boldsymbol{o}_+}} - (\frac{1}{2})^{\Delta L^-_{\boldsymbol{o}_+}}\big]^{-1}$, while the sample complexity scales as $\big[(\frac{1}{2})^{\Delta L^+_{\boldsymbol{o}_+}} - (\frac{1}{2})^{\Delta L^-_{\boldsymbol{o}_+}}\big]^{-2}$. Here, $\Delta L^+_{\boldsymbol{o}_+}$ denotes the separation between class-relevant features $\boldsymbol{o}_+$ in positive samples (capturing their locality), and $\Delta L^-_{\boldsymbol{o}_+}$ denotes the separation between confusion features $\boldsymbol{o}_+$ in negative samples (capturing the locality of confusing patterns). Thus, when $\Delta L^+_{\boldsymbol{o}_+} \gg \Delta L^-_{\boldsymbol{o}_+}$, the locality of class-relevant features dominates, which reduces both the number of iterations and the sample complexity needed for convergence, implying faster learning when class-relevant tokens are more concentrated locally. Finally, in both regimes, smaller token-level noise τ further accelerates learning.

4.2 Data Model

Consider an arbitrary set of orthonormal vectors $\mathcal{O} = \{\boldsymbol{o}_+, \boldsymbol{o}_-, \boldsymbol{o}_3, \ldots, \boldsymbol{o}_d\}$ in $\mathbb{R}^d$, where $\boldsymbol{o}_+$ and $\boldsymbol{o}_-$ are discriminative features and the remaining vectors $\boldsymbol{o}_j, j \geq 3$, are class-irrelevant (filler) features. Depending on the class label, either $\boldsymbol{o}_+$ or $\boldsymbol{o}_-$ serves as the class-relevant pattern, while the other acts as a confusion pattern. Each token $\boldsymbol{x}_l^{(n)}$ in $\boldsymbol{X}^{(n)}$ is a noisy version of one of the input patterns (features), i.e., $\boldsymbol{x}_l^{(n)} = \boldsymbol{o} + \boldsymbol{\xi}$, where $\boldsymbol{o} \in \mathcal{O}$ and $\boldsymbol{\xi}$ is the Gaussian noise. We consider two different data types: majority-voting and locality-structured data.

Majority Voting Data. For the majority voting data type, the label is determined by a majority vote over the class-relevant patterns. Let α_r and α_c denote the average fractions of class-relevant tokens and confusion tokens over the distribution $\mathcal{D}$, respectively. In positive samples, noisy variants of $\boldsymbol{o}_+$ are class-relevant, while noisy variants of $\boldsymbol{o}_-$ act as confusion tokens. In negative samples, the roles are reversed. All other tokens correspond to class-irrelevant features.

Locality-structured Data. For the locality-structured data type, each sequence contains two $\boldsymbol{o}_+$ tokens and two $\boldsymbol{o}_-$ tokens, while all other tokens correspond to class-irrelevant features. In positive samples, the two $\boldsymbol{o}_+$ tokens are close to each other, while the two $\boldsymbol{o}_-$ tokens are far apart; formally, $\Delta L^+_{\boldsymbol{o}_+} \ll \Delta L^+_{\boldsymbol{o}_-}$, where $\Delta L^+_{\boldsymbol{o}_+}$ and $\Delta L^+_{\boldsymbol{o}_-}$ denote the distances between the two $\boldsymbol{o}_+$ and $\boldsymbol{o}_-$ tokens, respectively. In negative samples, the pattern is reversed: $\Delta L^-_{\boldsymbol{o}_-} \ll \Delta L^-_{\boldsymbol{o}_+}$.

In addition, we consider a **balanced dataset** sampled from the unknown distribution $\mathcal{D}$. Let $\mathcal{N}_+ = \{(\boldsymbol{X}^{(n)}, z^{(n)}) : z^{(n)} = +1, n \in [N]\}$ and $\mathcal{N}_- = \{(\boldsymbol{X}^{(n)}, z^{(n)}) : z^{(n)} = -1, n \in [N]\}$ denote the sets of positively and negatively labeled samples, respectively. Then the class balance satisfies $\big|\, |\mathcal{N}_+| - |\mathcal{N}_-| \,\big| = O(\sqrt{N})$.

Interpreting the Data Model in Practice. Our theoretical data models are motivated by common patterns observed in practical machine learning tasks.

On the one hand, the **majority-voting** data model captures a widely adopted assumption (Li et al., 2023a; 2024b) in theoretical analysis, whereby the label is determined by the aggregate contribution through majority vote. For example, in image classification tasks (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014; He et al., 2016), the class label is often driven by multiple discriminative patches corresponding to foreground objects (class-relevant tokens). In contrast, background patches may contain other objects or patterns that are not associated with the target class (confusing tokens), along with random patches that are entirely unrelated (class-irrelevant tokens) (Dosovitskiy et al., 2021; Touvron et al., 2021).

On the other hand, the **locality-structured** data corresponds to tasks where semantic meaning is concentrated in spatially or temporally localized clusters, while background features are more dispersed. This structure is most familiar in vision tasks such as object detection and localization (Ren et al., 2015; Carion et al., 2020; Zhou et al., 2016) and image captioning (Vinyals et al., 2016; Xu et al., 2015; Radford et al., 2021), where the decisive content is often confined to a small region of the image. For example, in an image labeled "dog in a park," the prediction relies primarily on the contiguous region containing the dog rather than on scattered background textures. A similar principle holds in audio and speech recognition (Yadav & Tan, 2024; Gulati et al., 2020), where short phonetic segments capture the information needed to recognize words, and in genomics (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015), where functional elements such as sequence motifs and regulatory regions are localized to short windows of DNA. In these settings, the local structure of nearby tokens strongly correlates with the label.

Together, the majority-voting and locality-structured models offer complementary perspectives on when selective recurrence can most effectively support learning from structured real-world data.

4.3 Formal Theoretical Results

4.3.1 Theoretical Results for Majority-Voting Data

We next present a lemma characterizing how the gating vector aligns with different features under the majority-voting data.

Lemma 4.1 (Gating Vector Alignment for Majority Voting Data)**.** *With initialization where each entry of $\boldsymbol{W}_O$ is drawn independently from $\mathcal{N}(0, \xi^2)$ and $\boldsymbol{w}_\Delta^{(0)} = 0$. With a sufficient number of training samples and iterations, we have*

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_+ \right\rangle \geq \frac{\eta T}{8L^2}\Theta((\alpha_r L - \alpha_c L)^2) \tag{10}$$

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_- \right\rangle \geq \frac{\eta T}{8L^2}\Theta((\alpha_r L - \alpha_c L)^2) \tag{11}$$

$$\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_j \rangle \leq \widetilde{\mathcal{O}}\left(1/\text{poly}(d)\right), \quad \forall j \geq 3. \tag{12}$$

Lemma 4.1 establishes that after sufficient training, the gating vector $\boldsymbol{w}_\Delta$ aligns positively with the class-relevant features $\boldsymbol{o}_+$(10) and $\boldsymbol{o}_-$ as shown in (11), while its alignment with irrelevant features remains strictly negative as shown in (12). In other words, the selection mechanism implicitly acts as a feature selector, amplifying relevant tokens and ignoring irrelevant ones. Lemma 4.1 serves as an informal version of Lemmas B.5 and B.6.

Remark 1: With majority voting data, the gating vector aligns with discriminative features, i.e., $\boldsymbol{o}_+$ and $\boldsymbol{o}_-$. As a result, the model's output focuses primarily on these features, giving more weight to tokens that carry discriminative features while reducing the influence of less important tokens. Since the number of class-relevant tokens is greater than the number of confusing ones, e.g., in a positive sample, the tokens containing $\boldsymbol{o}+$ outnumber those containing $\boldsymbol{o}-$, the model can correctly assign the label through this majority effect. Furthermore, as the difference between the counts of class-relevant features and confusing features (i.e., $\alpha_r - \alpha_L$) increases, the gating vector converges much faster. Overall, this gating mechanism allows the model to use its training samples more efficiently because it learns to emphasize the most relevant feature early on and ignore irrelevant features.

We now present the theorem establishing the generalization guarantee for Mamba under the majority-voting data.

Theorem 1 (Generalization for Majority Voting Data)**.** *Suppose the model width satisfies $m \geq d^2 \log q$ for some constant $q > 0$, and the token noise level is bounded as $\tau < \mathcal{O}\big(\frac{1}{d}\big)$. Then, with probability at least $1 - N^{-d}$, if the number of training samples N satisfies*

$$N \geq \Omega\bigg(\frac{L^2 d}{\eta^2(\alpha_r - \alpha_c)^2}\bigg), \tag{13}$$

and the number of iterations T satisfies

$$T = \Theta\bigg(\frac{L^2}{\eta(\alpha_r - \alpha_c)^2}\bigg), \tag{14}$$

the model returned by Algorithm 1 achieves guaranteed generalization, i.e.,

$$f(\boldsymbol{v}^{(0)}, \boldsymbol{W}_O^{(T)}, \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{W}_B^{(0)}, \boldsymbol{W}_C^{(0)}) = 0. \tag{15}$$

Theorem 1 establishes the sample complexity, as shown in (13), and the convergence rate, as given in (14), that are required to guarantee desirable generalization when training the model in (6) using GD for the majority-voting data type. In other words, the model achieves good generalization once a sufficient number of samples is available, as specified in (13), and training has proceeded for a sufficient number of iterations, as specified in (14).

Remark 2: With majority-voting data, the Mamba architecture can effectively capture the underlying data distribution by first identifying discriminative features through its gating mechanism and then aggregating them via a data-dependent recurrent mechanism. In this sense, Mamba behaves similarly to the Transformer (Li et al., 2023a), suggesting a close connection between the two models despite their architectural differences. According to the results of Lemma 4.1, the model further benefits from a faster convergence rate and reduced sample complexity when the gap between class-relevant and confusing features is larger.

4.3.2 Theoretical Results for Locality-Structured Data

We next present a lemma characterizing how the gating vector aligns with different features under the locality-structured data.

Lemma 4.2 (Gating Vector Alignment for Locality-structured Data)**.** *With initialization where each entry of $\boldsymbol{W}_O$ is drawn independently from $\mathcal{N}(0, \xi^2)$ and $\boldsymbol{w}_\Delta^{(0)} = 0$. With a sufficient number of training samples and iterations, we have*

$$\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_+ \rangle \geq -\widetilde{\mathcal{O}}\left(1/\mathrm{poly}(d)\right), \tag{16}$$

$$\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_- \rangle \geq -\widetilde{\mathcal{O}}\left(1/\mathrm{poly}(d)\right), \tag{17}$$

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_j \right\rangle \leq \frac{-\eta T c'^3}{16L} \left[\left(\frac{1}{2}\right)^{\Delta L_{\boldsymbol{o}_+}^{+}-2} - \left(\frac{1}{2}\right)^{\Delta L_{\boldsymbol{o}_+}^{-}-2}\right] \left[\left(\frac{1}{2}\right)^{\Delta L_{\boldsymbol{o}_+}^{+}} + \left(\frac{1}{2}\right)^{\Delta L_{\boldsymbol{o}_-}^{-}}\right]. \tag{18}$$

Lemma 4.2 establishes that after sufficient training, the gating vector $\boldsymbol{w}_\Delta$ remains close to zero for class-relevant features $\boldsymbol{o}_+$ as shown in (16) and $\boldsymbol{o}_-$ as shown in (17), however its alignment with irrelevant features remains strongly negative as shown in (18). Through this mechanism, the gating favors class-relevant features to select the most informative feature for learning. Lemma 4.2 serves as an informal version of Lemmas C.5 and C.6.

Remark 3: The gating vector behaves differently from majority voting, though the overall insights remain similar. We can no longer guarantee that $\boldsymbol{w}_\Delta$ will always grow in the direction of discriminative features, because we assume that the number of class-relevant features can be comparable to the number of confusing features. This assumption is introduced to highlight the role of data locality in shaping the gating vector, which is more challenging to analyze in isolation since majority voting can readily reinforce it; however, their combined effect better reflects real-world data. Although this direct growth no longer holds, the gating vector consistently decreases in the direction of irrelevant features. At a higher level, this can be seen as a synergistic interaction: the recurrent mechanism captures locality and suppresses irrelevant features, which pushes the gating vector to decrease along those directions, while the gating itself further amplifies this suppression. From another perspective, by making the model pay less attention to irrelevant features, the gating vector effectively shifts more attention toward discriminative features.

We now present the theorem establishing the generalization guarantee for Mamba under the locality-structured data.

Theorem 2 (Generalization for Locality-structured Data)**.** *Suppose the model width satisfies $m \geq d^2 \log q$ for some constant $q > 0$, and the token noise level is bounded as $\tau < \mathcal{O}\big(\frac{1}{d}\big)$. Then, with probability at least $1 - N^{-d}$, if the number of training samples N satisfies*

$$N \geq \Omega\bigg(\frac{L^2 d}{\eta^2 \big[\, (1/2)^{\Delta L_{o+}^{+}} - (1/2)^{\Delta L_{o+}^{-}}\, \big]^2}\bigg), \tag{19}$$

and the number of iterations T satisfies

$$T = \Theta\bigg(\frac{L^2}{\eta \big[\, (1/2)^{\Delta L_{o+}^{+}} - (1/2)^{\Delta L_{o+}^{-}}\, \big]}\bigg), \tag{20}$$

the model returned by Algorithm 1 achieves guaranteed generalization, i.e.,

$$f\big(\boldsymbol{v}^{(0)}, \boldsymbol{W}_O^{(T)}, \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{W}_B^{(0)}, \boldsymbol{W}_C^{(0)}\big) = 0. \tag{21}$$

Theorem 2 shows that good generalization on locality-structured data is guaranteed if the sample complexity meets (19) and training proceeds for at least (20) iterations.

Remark 4: We establish that Mamba can also effectively learn this type of data through its ability to exploit locality, in contrast to Transformers, where no such guarantee is provided in (Li et al., 2023a). In our analysis, ΔL_{o+}^{+} captures the distance between class-relevant tokens, reflecting the locality of class-relevant features, while ΔL_{o+}^{-} captures the locality of confusing features. The effectiveness of learning is governed by the separation between these two quantities. In particular, when $\Delta L_{o+}^{+} \gg \Delta L_{o+}^{-}$, the locality of class-relevant features dominates that of confusing features. In particular, when $\Delta L_{o+}^{+} \gg \Delta L_{o+}^{-}$, the locality of class-relevant features dominates that of confusing ones, which reduces both the sample complexity and the number of iterations required for convergence, allowing Mamba to learn more effectively and efficiently.

4.4 Technical Novelty and Challenges

Differences with Existing Works. Our work is mainly inspired by prior feature-learning analyses of (Bakshi et al., 2019; Arora et al., 2019; Brutzkus & Globerson, 2021; Li et al., 2023a; 2025a). Building on these foundations, we develop a framework specifically tailored to *gated* architectures with structured data. Unlike these existing models, Mamba introduces an input-dependent gating mechanism, absent from other network architectures, which acts as a dynamic selection operator and requires new analytical techniques to capture its learning dynamics. Moreover, while the majority-voting data model has been previously studied in the context of Transformers (Li et al., 2023a), we show that Mamba can also learn this type of data with comparable performance. Furthermore, we find that Mamba is particularly effective at capturing the inherent locality of the data, which motivates us to introduce a new *locality-structured* data model. For both regimes, we establish generalization guarantees within the framework of selective state space models, thereby advancing our understanding of this class of architectures and clarifying their distinctions from Transformers. A proof sketch can be found in Appendix A.2.

Technical Challenges. Our analysis faces several unique technical challenges stemming from the structure of selective SSMs. Unlike attention-based models, where interactions are primarily additive, Mamba's gating mechanism introduces multiplicative recurrences across tokens, with dynamics that are explicitly sensitive to token order. These multiplicative effects accumulate over time, substantially complicating the training analysis. To capture this behavior, we systematically track the gradient updates of the gating vector $\boldsymbol{w}_\Delta$, decomposing the contributions from different token positions and analyzing how token placement influences training dynamics.

Specifically, in the **majority-voting data**, the gradient decomposition of the gating includes off-diagonal terms $\beta_{s,s+1}^{(l)}$ that exhibit additional multiplicative decay due to the recursive gating structure, whereas the diagonal term $\beta_{s,s}^{(l)}$ is independent of token position. Hence, it is important to carefully consider competing token contributions to prove indeed the gating vector indeed aligns with class-relevant feature directions.

Instead, in the **locality-structured data**, the variation introduced by the number of class-relevant and confusion tokens in positive and negative samples is negligible, as our data model assumes an equal number of class-relevant and confusion tokens. Consequently, we need to rely on $\Delta L_{\boldsymbol{o}_+}^{+}$ and $\Delta L_{\boldsymbol{o}_-}^{+}$ to ensure that the lucky neuron $\boldsymbol{W}_{O(i,\cdot)}$ learns the class-relevant feature. Moreover, since the number of class-relevant and confusion tokens is balanced, updates along the class-relevant feature direction for the gating vector remain close to zero. To demonstrate how the gate filters information, we show that gradient updates along class-irrelevant features are driven strongly negative. To prove this, in addition to the terms considered in the majority-voting setting, we must also bound the positively contributing terms that hinder the gate's ability to suppress irrelevant features. Specifically, we bound these opposing terms as $\mathcal{O}\left((1-\sigma(p_2))^{\Delta L_{\boldsymbol{o}_+}^{-}}\right)+\mathcal{O}\left((1-\sigma(p_2))^{\Delta L_{\boldsymbol{o}_-}^{+}}\right)$ ensuring that their effect remains minimal. This reveals that the gate effectively suppresses irrelevant features while preserving class-relevant features for this data model.

5 Numerical Experiments

We verify our theoretical results through synthetic experiments based on the data models described in Section 4.2. Due to the space limit, we defer the experiment details to Appendix A.3

Faster convergence with larger majority-voting gap. Fig. 1 illustrates that increasing the majority-voting gap $\alpha_r - \alpha_c$ consistently reduces the number of epochs across various sizes of training samples. These findings are consistent with our theoretical results in (13) and (14).

Gating mechanism amplifies relevant features in majority-voting data. Fig. 2 shows the cosine similarity between the gating vector $\boldsymbol{w}_\Delta$ and both class-relevant and class-irrelevant features. The similarity with class-relevant features steadily increases, while that with class-irrelevant features remains essentially unchanged. This empirically confirms Lemma 4.1, demonstrating that the gate prioritizes informative features while ignoring irrelevant ones.

MLP weights selectively align with only class-relevant features. Fig. 3 tracks the average cosine similarity between each neuron $\boldsymbol{W}_O(i,:)$ and both class-relevant & class-irrelevant features. The alignment increases for class-relevant features and stays essentially unchanged for irrelevant features, which is consistent with our findings in Lemmas B.1 and B.3 in the Appendix.

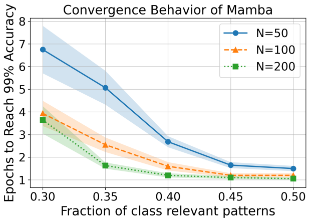

Figure 1: Convergence vs. majority-voting gap.

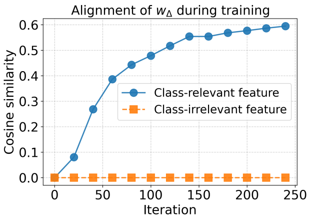

Figure 2: Alignment of $\boldsymbol{w}_\Delta$ for majority-voting data.

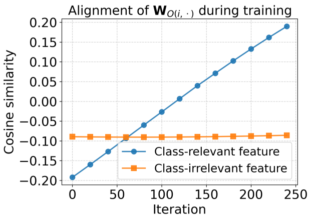

Figure 3: Average alignment of $\boldsymbol{W}_{O(i,\cdot)}$ during training.

Locality affects the learning. Fig. 4 illustrate the effect of class-relevant token separation ΔL on the convergence in the locality-structured data. Larger Δ slows convergence across different training sample sizes, which is consistent with our results in (19) and (20).

Gating mechanism suppresses irrelevant features. Fig. 5 illustrates that while the cosine similarity is negative for both types of features, it stays close to zero for class-relevant features but becomes largely negative for class-irrelevant ones. This contrast drives the gating mechanism to prioritize class-relevant features, consistent with Lemma 4.2.

Mamba outperforms Transformer and local attention on locality-structured data. Intuitively, locality-structured data favors models that exploit local biases. Global attention performs only marginally better than random guessing, whereas both local attention and Mamba learn meaningful patterns, with Mamba achieving the best performance.

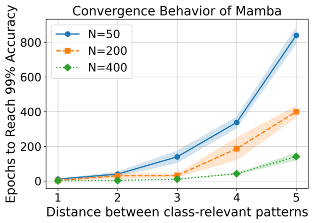

Figure 4: Convergence under locality-structured data.

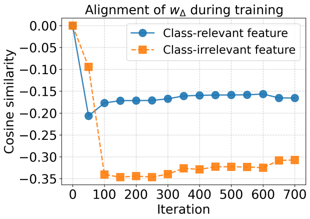

Figure 5: Alignment of $\boldsymbol{w}_\Delta$ for locality-structured data.

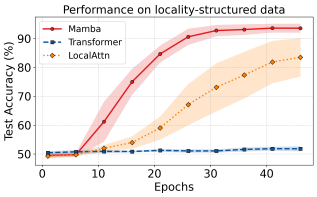

Figure 6: Mamba outperforms on locality data.

6 Conclusion

Encouraged by the emergence and successful applications of the Transformer alternative architecture Mamba, this paper provides a novel theoretical generalization analysis of Mamba by considering its unique gated selection mechanism. Focusing on a data model with class-relevant and class-irrelevant tokens, we establish the non-asymptotic sample complexity and the convergence rate required to achieve desirable test accuracy. Our analysis further shows that the gating parameter vector filters out the class-relevant features while ignoring irrelevant ones. To the best of our knowledge, this is the first theoretical analysis of Mamba's training dynamics, with its input-dependent gating mechanism, together with generalization guarantees.

Finally, we note some limitations of our work. First, our theoretical analysis focuses on a simplified Mamba setting that abstracts away practical components such as depth, multiple heads, residual connections, and layer normalization. Second, our data model, while standard in theoretical studies, also simplifies real-world sequence structures. Extending the analysis to more realistic multi-layer and multi-head Mamba architectures, richer data models, and alternative designs such as gated Transformers or hybrid Mamba–Transformer frameworks remains an important direction for future work.

LLM usage disclosure

We used large-language models (ChatGPT) to aid in polishing the writing of this paper. For numerical experiments, we employed AI-assisted coding tools (GitHub Copilot and ChatGPT) to support code development.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33 (8):831–838, 2015.

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)*, pp. 977–988. IEEE, 2022.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019a.

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

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

Ainesh Bakshi, Rajesh Jayaram, and David P Woodruff. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory*, pp. 195–268. PMLR, 2019.

Zeyu Bao, Penghao Yu, Haotian Jiang, and Qianxiao Li. The effect of depth on the expressivity of deep linear state-space models. *arXiv preprint arXiv:2506.19296*, 2025.

Ali Behrouz and Farnoosh Hashemi. Graph mamba: Towards learning on graphs with state space models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 119–130, 2024.

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.

Alon Brutzkus and Amir Globerson. An optimization and generalization analysis for max-pooling networks. In *Uncertainty in Artificial Intelligence*, pp. 1650–1660. PMLR, 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Edo Cohen-Karlik, Itamar Zimerman, Liane Galanti, Ido Atad, Amir Globerson, and Lior Wolf. On the expressivity of selective state-space layers: A multivariate polynomial approach. *arXiv preprint arXiv:2502.02209*, 2025.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech 2020*, pp. 5036–5040, 2020.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. *Advances in neural information processing systems*, 35:22982–22994, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Arya Honarpisheh, Mustafa Bozdag, Octavia Camps, and Mario Sznaier. Generalization error analysis for selective state-space models through the lens of attention. *arXiv preprint arXiv:2502.01473*, 2025.

Ningyuan Huang, Miguel Sarabia, Abhinav Moudgil, Pau Rodriguez, Luca Zappella, and Federico Danieli. Understanding input selectivity in mamba: Impact on approximation power, memorization, and associative recall capacity. *arXiv preprint arXiv:2506.11891*, 2025.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*, 2024.

Stefani Karp, Ezra Winston, Yuanzhi Li, and Aarti Singh. Local signal adaptivity: Provable feature learning in neural networks beyond kernels. *Advances in Neural Information Processing Systems*, 34:24883–24897, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *EMNLP 2018*, pp. 66, 2018.

Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *International Conference on Learning Representations (ICLR 2023)*, 2023a.

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. In *International Conference on Machine Learning*, pp. 28784–28829. PMLR, 2024a.

Hongkang Li, Meng Weng, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? In *International Conference on Machine Learning*, 2024b.

Hongkang Li, Yihua Zhang, Shuai Zhang, Pin-Yu Chen, Sijia Liu, and Meng Wang. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. In *International Conference on Learning Representations*, 2025a.

Yingcong Li, Ankit S Rawat, and Samet Oymak. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. *Advances in Neural Information Processing Systems*, 37: 138324–138364, 2024c.

Yingcong Li, Davoud Ataee Tarzanagh, Ankit Singh Rawat, Maryam Fazel, and Samet Oymak. Gating is weighting: Understanding gated linear attention through in-context learning. *arXiv preprint arXiv:2504.04308*, 2025b.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*, pp. 19689–19729. PMLR, 2023b.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

Chris Lu, Yannick Schroecker, Albert Gu, Emilio Parisotto, Jakob Foerster, Satinder Singh, and Feryal Behbahani. Structured state space models for in-context reinforcement learning. *Advances in Neural Information Processing Systems*, 36:47016–47031, 2023.

Peng Lu, Ivan Kobyzev, Mehdi Rezagholizadeh, Boxing Chen, and Philippe Langlais. Regla: Refining gated linear attention. *arXiv preprint arXiv:2502.01578*, 2025.

Cong Ma and Kayvan Najarian. Rethinking the long-range dependency in mamba/ssm and transformer models. *arXiv preprint arXiv:2509.04226*, 2025.

Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. *Advances in Neural Information Processing Systems*, 37:127226–127272, 2024.

Naoki Nishikawa and Taiji Suzuki. State space models are provably comparable to transformers in dynamic token selection. In *The Thirteenth International Conference on Learning Representations*.

Destiny Okpekpe and Antonio Orvieto. Revisiting associative recall in modern recurrent models. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*, 2025.

Antonio Orvieto, Soham De, Caglar Gulcehre, Razvan Pascanu, and Samuel L Smith. Universality of linear recurrences followed by non-linear projections: finite-width guarantees and benefits of complex eigenvalues. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 38837–38863, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, 2016.

Jerome Sieber, Carmen A Alonso, Alexandre Didier, Melanie N Zeilinger, and Antonio Orvieto. Understanding the differences in foundation models: Attention, state space models, and recurrent neural networks. *Advances in Neural Information Processing Systems*, 37:134534–134566, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Yonatan Slutzky, Yotam Alexander, Noam Razin, and Nadav Cohen. The implicit bias of structured state space models can be poisoned with clean labels. *arXiv preprint arXiv:2410.10473*, 2024.

Jimmy TH Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016.

Chloe Wang, Oleksii Tsepa, Jun Ma, and Bo Wang. Graph-mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv preprint arXiv:2402.00789*, 2024a.

Junxiong Wang, Tushaar Gangavarapu, Jing Nathan Yan, and Alexander M Rush. Mambabyte: Token-free selective state space model. *arXiv preprint arXiv:2401.13660*, 2024b.

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

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Sarthak Yadav and Zheng-Hua Tan. Audio mamba: Selective state spaces for self-supervised audio representations. *arXiv preprint arXiv:2406.02178*, 2024.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Forty-first International Conference on Machine Learning*.

Annan Yu and N Benjamin Erichson. Block-biased mamba for long-range sequence processing. *arXiv preprint arXiv:2505.09022*, 2025.

Shuai Zhang, Meng Wang, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Miao Liu. Joint edge-model sparse learning is provably efficient for graph neural networks. *The Eleventh International Conference on Learning Representations*, 2023.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62429–62442. PMLR, 2024.

A NOTATIONS, PROOF SKETCH AND ADDITIONAL EXPERIMENTS

A.1 NOTATIONS

A.1.1 LUCKY NEURON DEFINITION

Let

$$\mathcal{K}_+ = \{i \in [m] : v_i > 0\}, \quad \mathcal{K}_- = \{i \in [m] : v_i < 0\} \tag{22}$$

denote the sets of neurons with positive and negative output layer weights, respectively.

We define the sets of lucky neurons at initialization as:

$$\mathcal{W}(0) = \left\{ i \in \mathcal{K}_+ : \boldsymbol{W}_{O(i,\cdot)}(0)\boldsymbol{o}_+ > 0 \right\}, \tag{23}$$

$$\mathcal{U}(0) = \left\{ i \in \mathcal{K}_- : \boldsymbol{W}_{O(i,\cdot)}(0)\boldsymbol{o}_- > 0 \right\}, \tag{24}$$

where $\boldsymbol{o}_+$ and $\boldsymbol{o}_-$ denote the class-relevant features for the positive and negative classes, respectively.

A.1.2 LOSS FUNCTION

The loss function for the n^{th} sample is defined as

$$\begin{aligned} \ell(\boldsymbol{X}^{(n)}, z^{(n)}) &= \max\{0, 1 - z^{(n)} \cdot F(\boldsymbol{X}^{(n)})\} \\ &= \max\left\{0, 1 - z^{(n)} \cdot \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m} v_i \phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\right\}. \end{aligned} \tag{25}$$

The empirical loss is denoted by $\hat{\mathcal{L}}$ and is given by

$$\hat{\mathcal{L}} = \frac{1}{N}\sum_{n=1}^{N} \ell(\boldsymbol{X}^{(n)}, z^{(n)}). \tag{26}$$

The population loss is denoted by $\mathcal{L}$ and is defined as

$$\mathcal{L} = \mathbb{E}_{(\boldsymbol{X},z)\sim\mathcal{D}} \ell(\boldsymbol{X}, z). \tag{27}$$

With additional important notations can be found in Table 2.

A.2 PROOF SKETCH

The major idea of our proof is to analyze how GD gradually aligns both the hidden-layer weights and gating vector with class-relevant features while ignoring the irrelevant ones. A key tool in our analysis is the notion of a *lucky neuron*, i.e., a hidden layer neuron whose initialization is well aligned with a class-relevant feature. For the majority-voting data model, the signal driving this alignment is proportional to the gap between the fractions of class-relevant and confusion tokens, $\Theta(\alpha_r - \alpha_c)$, as established by Lemmas B.1–B.4. Lucky neurons move consistently toward their class-relevant feature, while the magnitude of unlucky ones remains small (upper-bounded by the inverse square root of the number of samples). For the locality-structured data model, we prove that the update in the class-relevant feature direction for the gating vector remains close to zero because an equal number of class-relevant and confusion tokens are present in the data. We then show that the gating vector consistently decreases along irrelevant feature directions, thereby enabling the gate to effectively select the class-relevant feature.

Due to these properties, the training dynamics can be simplified to show that the network output in (6) changes linearly with the iteration number t. In particular, we prove that, for a new positive sample (w.l.o.g.) during inference, the learned model's output is strictly positive. From this analysis, we derive the sample complexity and the required number of iterations for achieving zero generalization error for both data types, as shown in (13) and (14) for the majority-voting setting in Theorem 1, and similarly in (19) and (20) for the locality-structured setting in Theorem 2.

Table 2: Summary of notations

$F(\boldsymbol{X}^{(n)})$	The final model output for $\boldsymbol{X}^{(n)}$
α_r	The average fractions of class-relevant tokens
α_c	The average fractions of confusion tokens
$\Delta L_{\boldsymbol{o}_+}^{+}$	Separation between class-relevant features $\boldsymbol{o}_+$ in positive samples
$\Delta L_{\boldsymbol{o}_+}^{-}$	Separation between confusion features $\boldsymbol{o}_+$ in negative samples
$\Delta L_{\boldsymbol{o}_-}^{-}$	Separation between class-relevant features $\boldsymbol{o}_-$ in negative samples
$\Delta L_{\boldsymbol{o}_-}^{+}$	Separation between confusion features $\boldsymbol{o}_-$ in positive samples
$\mathcal{O}$	The set of class-relevant and class-irrelevant patterns
$\mathcal{K}_+$	The set of lucky neurons with respect to $\boldsymbol{W}^{(0)}$
$\mathcal{K}_-$	The set of lucky neurons with respect to $\boldsymbol{U}^{(0)}$
$\mathcal{N}$	The set of training data
$\mathcal{N}_+$	The set of training data with positive labels
$\mathcal{N}_-$	The set of training data with negative labels
$\mathcal{W}(t)$	Set of lucky neurons for the positive class at iteration t
$\mathcal{U}(t)$	Set of lucky neurons for the negative class at iteration t
$\mathcal{O}(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$	We use the standard convention: $f(x) = \mathcal{O}(g(x))$ (resp. $\Omega(g(x))$, $\Theta(g(x))$) means $f(x)$ grows at most (resp. at least, on the order of) $g(x)$.
$\widetilde{\mathcal{O}}(\cdot)$	Soft-$\mathcal{O}$ notation: hides polylog factors
$\mathrm{poly}(d)$	An unspecified polynomial in d
$\gtrsim$, $\lesssim$	$f(x) \gtrsim g(x)$ (resp. $f(x) \lesssim g(x)$) abbreviates $f(x) \geq \Omega\big(g(x)\big)$ (resp. $f(x) \leq \mathcal{O}\big(g(x)\big)$).

A.3 Additional Numerical Experiments

Experiment settings.

The data dimension and token embedding size are both set to $d = 32$, which also corresponds to the number of feature directions. Unless otherwise stated, experiments in the main text use exactly the model defined in Eq. (6) to match our theoretical setting. We also use the model without convolution, and keep $\mathbf{W}_B = \mathbf{W}_C = I$ frozen as in Eq. (15). The total number of neurons in the hidden layer $\boldsymbol{W}_O$ is set to $m = 50$. For simplicity, we fix the ratio of different features to be the same across all data. The sequence length is set to $L = 30$.

We run 100 independent trials and consider only the successful trials to compute the mean epochs for convergence for a given fraction of class-relevant patterns. An experiment is successful if the testing loss is smaller than 10^{-3}. For this experiment, we fixed the fraction of the confusion tokens at 0.10 and varied the fraction of class-relevant features.

Additional Results on MLP Weight Alignment.

Figure 7 illustrates the alignment of sampled neurons with the class-relevant feature. We observe that, with a good initialization, a subset of neurons, denoted as lucky neurons, consistently increases in the direction of the class-relevant feature, while another subset, denoted as unlucky neurons, remains almost unchanged, which supports our findings in Lemmas B.3 and C.3.

In contrast, Figure 8 shows the alignment of sampled neurons with the class-irrelevant feature. In this case, we observe that all neurons, both lucky and unlucky, remain nearly unchanged in the direction of the class-irrelevant feature, which further supports our findings in Lemmas B.4 and C.4.

Additional Experiments

To further strengthen the empirical connection between our theoretical analysis and practical Mamba architectures, we conducted additional experiments using the multi-layer, multi-head Mamba model

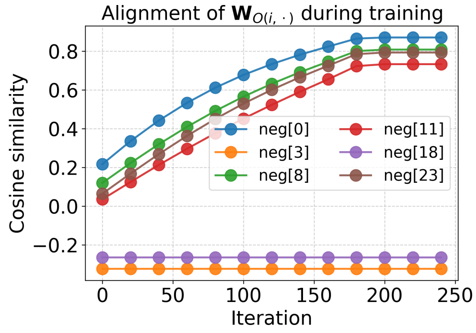

Figure 7: Alignment of $\boldsymbol{W}_{O(i,\cdot)}$ with class-relevant feature directions during training on the majority-voting data.

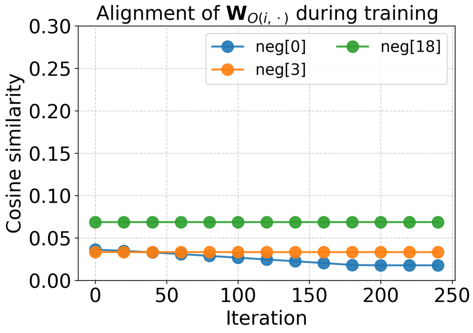

Figure 8: Alignment of $\boldsymbol{W}_{O(i,\cdot)}$ with class-irrelevant feature directions during training on the majority-voting data.

from Dao & Gu (2024), trained on synthetic datasets that follow the same structured data models as in our theory.

We first evaluated the Mamba2 block, which includes residual connections and RMSNorm. We focused on a 2-block Mamba model with 4 heads and report the cosine similarity of the learned gating vectors and MLP weights with class-relevant and class-irrelevant features in Figures 9 and 10. For a deeper 5-block Mamba model with the same configuration, we summarize the final alignment values in Table 3, which exhibit the same qualitative trends predicted by our analysis.

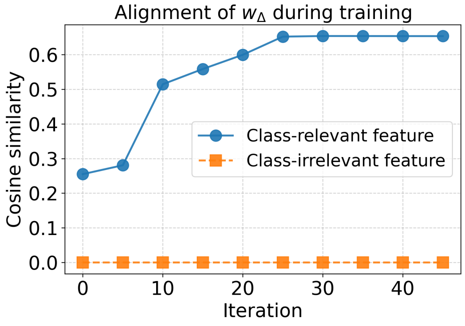

Figure 9: Alignment of the gating vector in the 2-block Mamba model.

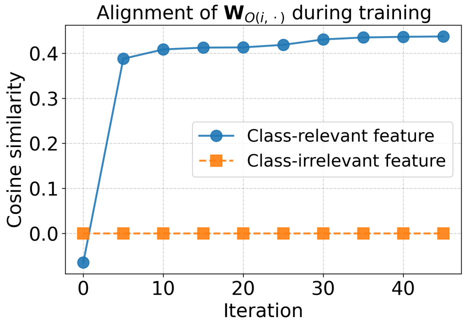

Figure 10: Alignment of the MLP weights in the 2-block Mamba model.

Table 3: Cosine similarity alignment in the 5-block Mamba model

Component	**Class-relevant**	**Class-irrelevant**
Gating vector	0.53	0.00
MLP weights	0.73	0.00

Next, we examined the effect of the gating mechanism by comparing models trained with and without gating across both structured data regimes. On the majority-voting data, the gated model consistently outperforms the ungated variant (Figure 11). On the locality-structured data, gating becomes essential: the ungated model fails to learn the task, whereas the gated model converges reliably (Figure 12).

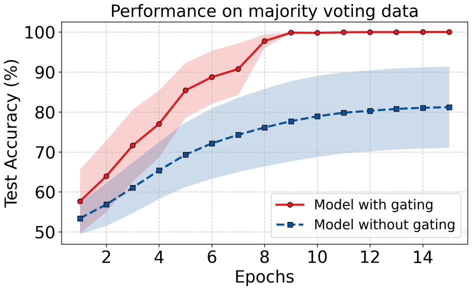

Figure 11: Test accuracy with and without gating on the majority-voting data.

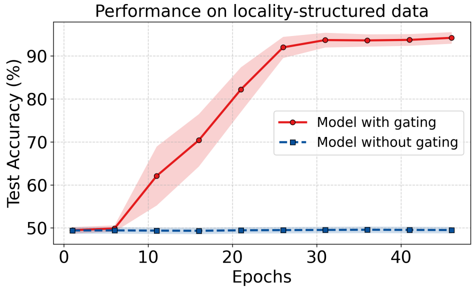

Figure 12: Test accuracy with and without gating on the locality-structured data.

We also conducted two controlled ablations. First, we varied the feature dimension $d \in \{32, 64, 128\}$ and observed that the qualitative behavior of the model remained consistent across all three settings (Figures 13–15). Second, we varied the data distribution parameter α_c, the fraction of confusion tokens in the majority-voting data. Across all three choices of α_c, the empirical results remained closely aligned with the theoretical predictions (Figures 16–18).

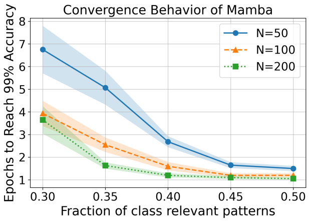

Figure 13: Ablation with feature dimension $d = 32$.

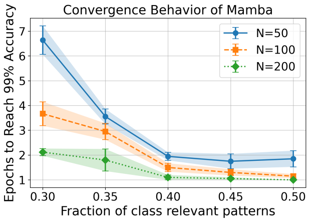

Figure 14: Ablation with feature dimension $d = 64$.

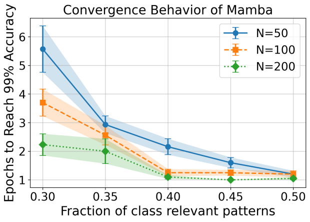

Figure 15: Ablation with feature dimension $d = 128$.

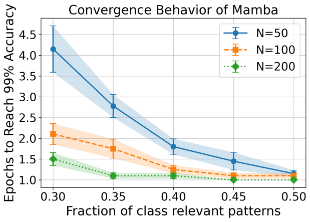

Figure 16: Ablation with confusion fraction $\alpha_c = 0.17$.

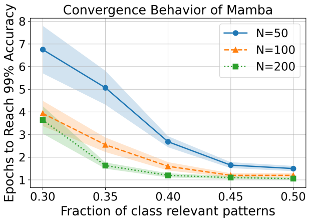

Figure 17: Ablation with confusion fraction $\alpha_c = 0.20$.

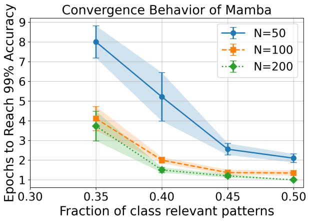

Figure 18: Ablation with confusion fraction $\alpha_c = 0.23$.

B Majority-voting data

B.1 Useful Lemmas

Lemma B.1 provides bounds on the gradient updates of lucky neurons $i \in \mathcal{W}(t)$ in the directions of both class-relevant features ($\boldsymbol{o}_+$, $\boldsymbol{o}_-$) and irrelevant features.

Lemma B.1. *Suppose $p_1 \leq \langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+ \rangle \leq q_1$ and $p_1 \leq \langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_- \rangle \leq q_1$. Then, for any lucky neuron $i \in \mathcal{W}(t)$ at iteration t, the following bounds hold:*

(L1.1) *A lower bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{1}{\sqrt{m}L} \cdot \sigma(p_1)\Theta(\alpha_r L - \alpha_c L) - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) - \mathcal{O}(\tau). \tag{28}$$

(L1.2) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \frac{1}{\sqrt{m}L} \cdot \sigma(q_1)\Theta(\alpha_r L - \alpha_c L) + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau). \tag{29}$$

(L1.3) *A lower bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \geq -\frac{1}{\sqrt{m}L} \cdot \sigma(q_1)\Theta(\alpha_r L - \alpha_c L) - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) - \mathcal{O}(\tau). \tag{30}$$

(L1.4) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau). \tag{31}$$

(L1.5) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_j$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau), \quad \textit{for } j \neq 1, 2. \tag{32}$$

Lemma B.2 shows that, for unlucky neurons associated with the positive class, the gradients in the directions of both class-relevant and irrelevant features are small.

Lemma B.2. *For any unlucky neuron $i \in \mathcal{K}_+ \setminus \mathcal{W}(t)$ at iteration t, the following bounds hold:*

(L2.1) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau). \tag{33}$$

(L2.2) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau). \tag{34}$$

(L2.3) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_j$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau), \quad \textit{for } j \neq 1, 2. \tag{35}$$

Lemmas B.3 and B.4, by symmetry, state the analogous results for lucky and unlucky neurons associated with the negative class.

Lemma B.3. *Suppose $p_1 \leq \langle \boldsymbol{w}_{\Delta}^{(t)}, \boldsymbol{o}_+ \rangle \leq q_1$ and $p_1 \leq \langle \boldsymbol{w}_{\Delta}^{(t)}, \boldsymbol{o}_- \rangle \leq q_1$. Then, for any lucky neuron $i \in \mathcal{U}(t)$ at iteration t, the following bounds hold:*

(L3.1) *A lower bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \geq \frac{1}{\sqrt{m}L} \cdot \sigma(p_1)\Theta(\alpha_r L - \alpha_c L) - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) - \mathcal{O}(\tau). \tag{36}$$

(L3.2) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \frac{1}{\sqrt{m}L} \cdot \sigma(q_1)\Theta(\alpha_r L - \alpha_c L) + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau). \tag{37}$$

(L3.3) *A lower bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \geq -\frac{1}{\sqrt{m}L} \cdot \sigma(q_1)\Theta(\alpha_r L - \alpha_c L) - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) - \mathcal{O}(\tau). \tag{38}$$

(L3.4) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau). \tag{39}$$

(L3.5) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_j$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau), \quad \textit{for } j \neq 1, 2. \tag{40}$$

Lemma B.4. *For any unlucky neuron $i \in \mathcal{K}_- \setminus \mathcal{U}(t)$ at iteration t, the following bounds hold:*

(L4.1) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau). \tag{41}$$

(L4.2) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau). \tag{42}$$

(L4.3) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_j$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau), \quad \textit{for } j \neq 1, 2. \tag{43}$$

Lemma B.5 establishes bounds for the gradient updates of $\boldsymbol{w}_{\Delta}$ in the class-relevant feature directions.

Lemma B.5. *Suppose $r_1^* \leq \langle {\boldsymbol{W}_{O(i,\cdot)}^{(t+1)}}^\top, \boldsymbol{o}_+ \rangle \leq s_1^*$. Let $|\mathcal{W}(t)| = \rho_t^+$ and $|\mathcal{U}(t)| = \rho_t^-$. Then, at iteration t, the following bounds hold:*

(L5.1) *A lower bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{w}_\Delta$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{r_1^*}{2\sqrt{m}L} \cdot \rho_t^+ \cdot \Theta(\alpha_r L) - \frac{\sqrt{m} s_1^*}{4L} \cdot \Theta(\alpha_c L) - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) - \mathcal{O}(\tau). \quad (44)$$

(L5.2) *A lower bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{w}_\Delta$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_- \right\rangle \geq \frac{r_1^*}{2\sqrt{m}L} \cdot \rho_t^- \cdot \Theta(\alpha_r L) - \frac{\sqrt{m} s_1^*}{4L} \cdot \Theta(\alpha_c L) - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) - \mathcal{O}(\tau) \quad (45)$$

Lemma B.6 establishes bounds for the gradient updates of $\boldsymbol{w}_\Delta$ in the directions of irrelevant features.

Lemma B.6. *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{w}_\Delta$ at iteration t, in the direction of $\boldsymbol{o}_j$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau), \quad \textit{for } j \neq 1, 2. \quad (46)$$

B.2 Proof of Convergence

Proof of Theorem 1. The proof starts with the base case at $t = 0$ and proceeds to analyze the training dynamics in a deductive manner, providing additional details in deriving the corresponding convergence and sample complexity bounds.

(**S1**) Warm-up (Base case): Training dynamics at the first iteration $t = 0$.

Recall that we set $\boldsymbol{w}_\Delta^{(0)} = \boldsymbol{0}$. Then, we have

$$\langle \boldsymbol{w}_\Delta^{(0)}, \boldsymbol{o}_+ \rangle = 0 \quad \text{and} \quad \langle \boldsymbol{w}_\Delta^{(0)}, \boldsymbol{o}_- \rangle = 0.$$

(**S1.1**) Training dynamics of $\boldsymbol{W}_{O(i,:)}$ at the first iteration $t = 0$.

From Lemma B.1, identify $p_1 = 0$ and $q_1 = 0$. Let α_r and α_c denote the average fraction of label-relevant tokens and confusion tokens, respectively. Then, for any lucky neuron $i \in \mathcal{W}(0)$, we obtain

$$\begin{aligned} \frac{1}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right) \leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(0)}}, \boldsymbol{o}_+ \right\rangle& \\ \leq \frac{1}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) + \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right)&. \end{aligned} \quad (47)$$

$$\text{and} \quad \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(0)}}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right) \quad \text{for } j \neq 1. \quad (48)$$

Recall that we set the number of samples in a batch $N = \text{poly}(d)$.

Recall that the initialization is

$$\boldsymbol{W}_{O(i,\cdot)}(0) = \delta_1 \boldsymbol{o}_+ + \delta_2 \boldsymbol{o}_- + \cdots + \delta_d \boldsymbol{o}_d, \quad \delta_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \xi^2) \quad j = 1, 2, \cdots, d. \quad (49)$$

Then, after one gradient descent step, we have

$$\begin{aligned}
&\delta_1 + \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
&\quad \leq \left\langle \boldsymbol{W}_{O(i,\cdot)}^{\top}{}^{(1)}, \boldsymbol{o}_+ \right\rangle \\
&\quad \leq \delta_1 + \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)
\end{aligned} \tag{50}$$

$$\text{and} \quad \left\langle \boldsymbol{W}_{O(i,\cdot)}^{\top}{}^{(1)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 1. \tag{51}$$

By applying Lemma B.3, for any lucky neuron $i \in \mathcal{U}(0)$, we obtain

$$\begin{aligned}
&\delta_2 + \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
&\quad \leq \left\langle \boldsymbol{W}_{O(i,\cdot)}^{\top}{}^{(1)}, \boldsymbol{o}_- \right\rangle \\
&\quad \leq \delta_2 + \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)
\end{aligned} \tag{52}$$

$$\text{and} \quad \left\langle \boldsymbol{W}_{O(i,\cdot)}^{\top}{}^{(1)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 2. \tag{53}$$

For any unlucky neuron $i \in \mathcal{K}_- \setminus \mathcal{U}(0)$, Lemma B.4 gives

$$\left\langle \boldsymbol{W}_{O(i,\cdot)}^{\top}{}^{(1)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } \forall j. \tag{54}$$

<u>(**S1.2**) Training dynamics of $\boldsymbol{W}_\Delta$ at the first iteration $t = 0$.</u>

Now consider the gradient update for $\boldsymbol{w}_\Delta$. Define:

$$\begin{aligned}
a &= \delta_1 + \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
b &= \delta_1 + \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)
\end{aligned}$$

Applying Lemma B.5 with $r_1^* = a$, $s_1^* = b$, and $\rho_0^+ = |\mathcal{W}(0)|$, we get

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(0)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{a}{2\sqrt{m}L} \cdot \rho_0^+ \cdot \Theta(\alpha_r L) - \frac{\sqrt{m}b}{4L} \cdot \Theta(\alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) =: \alpha \tag{55}$$

Let $\delta_1 = \frac{1}{\mathrm{poly}(d)}$. Since $a - b = \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)$ that is sufficiently small,

$$\begin{aligned}
\alpha &= \frac{1}{2L}\left[\frac{a}{\sqrt{m}} \cdot \frac{m}{2}\Theta(\alpha_r L) - \frac{\sqrt{m}b}{2}\Theta(\alpha_c L)\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
&= \frac{1}{2L}\left[\frac{\sqrt{m}a}{2}(\Theta(\alpha_r L) - \Theta(\alpha_c L))\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
&= \frac{\sqrt{m}}{4L} \cdot \frac{\eta}{2\sqrt{m}L}\Theta((\alpha_r L - \alpha_c L)^2) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
&= \frac{\eta}{8L^2}\Theta((\alpha_r L - \alpha_c L)^2) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) > 0
\end{aligned} \tag{56}$$

From Lemma B.6, we also obtain

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_{\Delta}^{(0)}}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) =: \gamma \quad \text{for } j \neq 1, 2. \tag{57}$$

(**S2**) Induction Step: Training dynamics at a general iteration t.

Suppose $\langle \boldsymbol{w}_{\Delta}^{(t)}, \boldsymbol{o}_+ \rangle = \alpha^* \geq \alpha \cdot t$, $\langle \boldsymbol{w}_{\Delta}^{(t)}, \boldsymbol{o}_- \rangle = \beta^* \geq \beta \cdot t$, and $\langle \boldsymbol{w}_{\Delta}^{(t)}, \boldsymbol{o}_j \rangle = \gamma^* \leq \gamma \cdot t$, where

$$\beta = \frac{a'}{2\sqrt{m}L} \cdot \rho_0^- \cdot \Theta(\alpha_r L) - \frac{\sqrt{m}b'}{4L} \cdot \Theta(\alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) > 0, \tag{58}$$

$$a' = \delta_2 + \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right), \tag{59}$$

$$b' = \delta_2 + \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right). \tag{60}$$

Following the same approach as in (56), we can simplify and obtain

$$\beta = \frac{\eta}{8L^2}\Theta((\alpha_r L - \alpha_c L)^2) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) > 0. \tag{61}$$

For any lucky neuron $i \in \mathcal{W}(t)$ at the $(t+1)$-th iteration, we have

$$\begin{aligned} &\frac{1}{\sqrt{m}L} \cdot \sigma(\alpha^*)\Theta(\alpha_r L - \alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\ &\quad \leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \\ &\quad \leq \frac{1}{\sqrt{m}L} \cdot \sigma(\alpha^*)\Theta(\alpha_r L - \alpha_c L) + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right), \end{aligned}$$

and

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 1 \tag{62}$$

Next, we have $\sigma(\alpha^*) > \frac{1}{2}$ since $\alpha^* > 0$ when $t = 1$. By a simple induction, this further ensures

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(0)}}, \boldsymbol{o}_+ \right\rangle \leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(1)}}, \boldsymbol{o}_+ \right\rangle \leq \cdots \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t+1)}}, \boldsymbol{o}_+ \right\rangle. \tag{63}$$

Thus, we obtain the following bound after the second gradient descent step:

$$\begin{aligned} &\delta_1 + \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L)\left[1 + 2\sigma(\alpha^*)\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) =: u \\ &\quad \leq \left\langle (\boldsymbol{W}_{O(i,\cdot)}^{(2)})^\top, \boldsymbol{o}_+ \right\rangle \\ &\quad \leq \delta_1 + \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L)\left[1 + 2\sigma(\alpha^*)\right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) =: v. \end{aligned} \tag{64}$$

Similarly, applying Lemma B.3 to any lucky neuron $i \in \mathcal{U}(1)$ at iteration 2, we get

$$\begin{aligned} &\frac{1}{\sqrt{m}L} \cdot \sigma(\beta^*)\Theta(\alpha_r L - \alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\ &\quad \leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(1)}}, \boldsymbol{o}_- \right\rangle \\ &\quad \leq \frac{1}{\sqrt{m}L} \cdot \sigma(\beta^*)\Theta(\alpha_r L - \alpha_c L) + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right), \end{aligned} \tag{65}$$

$$\text{and} \quad \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(1)}}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\operatorname{poly}(d)}\right) \quad \text{for } j \neq 2. \tag{66}$$

Applying Lemma B.5 with $r_1^* = u$, and $s_1^* = v$, we obtain

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_{\Delta}^{(1)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{u}{2\sqrt{m}L} \cdot \rho_1^+ \cdot \Theta(\alpha_r L) - \frac{\sqrt{m}v}{4L} \cdot \Theta(\alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\operatorname{poly}(d)}\right) =: \chi. \tag{67}$$

Since $u - v = \widetilde{\mathcal{O}}\left(\frac{1}{\operatorname{poly}(d)}\right)$ that is sufficiently small, we have $\chi \geq 0$.

By applying Lemma B.6, we get

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_{\Delta}^{(0)}}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\operatorname{poly}(d)}\right) \quad \text{for } j \neq 1, 2. \tag{68}$$

(**S3**) Induction conclusion: Training dynamics when the algorithm ends.

We proceed by induction on t: the base case $t = 0$ is established in (S1), and the induction step for general t is shown in (S2). For any lucky neuron $i \in \mathcal{W}(T)$, we obtain

$$\left\langle \boldsymbol{W}_{O(i,\cdot)}^{\top}{}^{(T)}, \boldsymbol{o}_+ \right\rangle \geq aT, \tag{69}$$

$$\text{and} \quad \left\langle \boldsymbol{W}_{O(i,\cdot)}^{\top}{}^{(T)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\operatorname{poly}(d)}\right) \quad \text{for } j \neq 1 \tag{70}$$

For any lucky neuron $i \in \mathcal{U}(T)$, we obtain

$$\left\langle \boldsymbol{W}_{O(i,\cdot)}^{\top}{}^{(T)}, \boldsymbol{o}_- \right\rangle \geq aT, \tag{71}$$

$$\text{and} \quad \left\langle \boldsymbol{W}_{O(i,\cdot)}^{\top}{}^{(T)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\operatorname{poly}(d)}\right) \quad \text{for } j \neq 2 \tag{72}$$

Also, we obtain

$$\left\langle \boldsymbol{w}_{\Delta}^{(T)}, \boldsymbol{o}_+ \right\rangle \geq \alpha T, \tag{73}$$

$$\left\langle \boldsymbol{w}_{\Delta}^{(T)}, \boldsymbol{o}_- \right\rangle \geq \beta T, \tag{74}$$

$$\text{and} \quad \left\langle \boldsymbol{w}_{\Delta}^{(T)}, \boldsymbol{o}_j \right\rangle \leq \gamma T. \tag{75}$$

(**S4**) Derivation for the generalization bound.

We will demonstrate that once the weights have converged at iteration T, the model accurately captures the underlying data distribution, which leads to zero generalization error, as shown in (94).

Consider $z^{(n)} = +1$ as an example. The sequence $\boldsymbol{X}^{(n)} = \begin{bmatrix} \boldsymbol{x}_1^{(n)} & \boldsymbol{x}_2^{(n)} & \cdots & \boldsymbol{x}_L^{(n)} \end{bmatrix}$ has first $\alpha_r L$ tokens correspond to the feature $\boldsymbol{o}_+$, while the following $\alpha_c L$ tokens correspond to the feature $\boldsymbol{o}_-$.

$$
\begin{aligned}
F(\boldsymbol{X}^{(n)}) &= \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m} v_i\,\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \\
&= \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}^+}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) - \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}^-}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \\
&\geq \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) - \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{U}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \\
&\quad - \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}^-\backslash\mathcal{U}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \qquad (76)
\end{aligned}
$$

The Mamba output $\boldsymbol{y}_l^{(n)}$ is defined as

$$
\boldsymbol{y}_l^{(n)} = \sum_{s=1}^{l}\left(\prod_{j=s+1}^{l}\left(1-\sigma(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_j^{(n)})\right)\right)\cdot\sigma(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_s^{(n)})\cdot(\boldsymbol{x}_s^{(n)\top}\boldsymbol{x}_l^{(n)})\boldsymbol{x}_s^{(n)}. \qquad (77)
$$

We now derive a lower bound for

$$
\sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\phi(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l).
$$

To that end, consider the aggregated projection

$$
\sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l = \sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\sum_{j=1}^{d}\langle\boldsymbol{W}_{O(i,\cdot)}^\top,\boldsymbol{o}_j\rangle\cdot\langle\boldsymbol{y}_l,\boldsymbol{o}_j\rangle. \qquad (78)
$$

For any $i\in\mathcal{W}(0)$, we know that

$$
\langle\boldsymbol{W}_{O(i,\cdot)}^\top,\boldsymbol{o}_+\rangle \geq aT. \qquad (79)
$$

Hence, let's obtain a lower bound for $\langle\boldsymbol{y}_l,\boldsymbol{o}_+\rangle$.

We only need to consider the cases where $\boldsymbol{x}_s = \boldsymbol{o}_+$ for some s in the range $1\leq s\leq l$.

After T iterations, we know

$$
\langle\boldsymbol{w}_\Delta,\boldsymbol{o}_+\rangle \geq \alpha T, \quad \langle\boldsymbol{w}_\Delta,\boldsymbol{o}_-\rangle \geq \beta T, \quad \langle\boldsymbol{w}_\Delta,\boldsymbol{o}_j\rangle \leq \gamma T \quad \text{for } j\neq 1,2. \qquad (80)
$$

Therefore, we have

$$
\langle\boldsymbol{y}_l,\boldsymbol{o}_+\rangle = \Theta(\sigma\left(\langle\boldsymbol{w}_\Delta,\boldsymbol{o}_+\rangle\right)) = \Theta(\sigma(\alpha T)), \quad \text{for } l=1,2,\ldots,\alpha_r L. \qquad (81)
$$

We now lower bound the objective

$$
\sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\phi(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l).
$$

Note that

$$
\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l = \sum_{j=1}^{d}\left\langle\boldsymbol{W}_{O(i,\cdot)}^\top,\boldsymbol{o}_j\right\rangle\left\langle\boldsymbol{y}_{L_1^+},\boldsymbol{o}_j\right\rangle,
$$

and $\boldsymbol{y}_l$ has only $\boldsymbol{o}_+$ component for $l=1,2,\ldots,\alpha_r L$.

Therefore,

$$\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l = \left\langle \boldsymbol{W}_{O(i,\cdot)}^\top, \boldsymbol{o}_+ \right\rangle \langle \boldsymbol{y}_l, \boldsymbol{o}_+ \rangle \geq aT \cdot \Theta(\sigma(\alpha T)) > 0, \quad \text{for } l = 1, 2, \ldots, \alpha_r L.$$

Applying $\phi(z) = z$ for positive z, we obtain

$$\phi(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l) \geq aT \cdot \Theta(\sigma(\alpha T)), \quad \text{for } l = 1, 2, \ldots, \alpha_r L.$$

Hence,

$$\sum_{i \in \mathcal{W}(0)} \sum_{l=1}^{L} \phi(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l) \geq \sum_{i \in \mathcal{W}(0)} aT \cdot \Theta(\sigma(\alpha T)) \cdot \alpha_r L \tag{82}$$

Next, we derive an upper bound for

$$\sum_{i \in \mathcal{U}(0)} \sum_{l=1}^{L} \phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right).$$

For any $i \in \mathcal{U}(0)$, we know that

$$0 < \langle \boldsymbol{W}_{O(i,\cdot)}^\top, \boldsymbol{o}_- \rangle \leq bT. \tag{83}$$

We now derive an upper bound for $\langle \boldsymbol{y}_l, \boldsymbol{o}_- \rangle$. We only need to consider the cases where $\boldsymbol{x}_s = \boldsymbol{o}_-$ such that $1 \leq s \leq l$.

We have,

$$\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_+ \rangle \leq WT, \qquad \langle \boldsymbol{w}_\Delta, \boldsymbol{o}_- \rangle \leq WT,$$

where

$$W = \frac{\eta}{8L^2}\Theta((\alpha_r L - \alpha_c L)^2) + \tilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right). \tag{84}$$

$$\langle \boldsymbol{y}_l, \boldsymbol{o}_- \rangle = \Theta(\sigma\left(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_- \rangle\right)) = \Theta(\sigma(WT)), \quad \text{for } l = 1, 2, \ldots, \alpha_c L. \tag{85}$$

$$\sum_{l=1}^{L} \boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l \leq bT \cdot \Theta(\sigma(WT)) \cdot \alpha_c L. \tag{86}$$

$$\sum_{i \in \mathcal{U}(0)} \sum_{l=1}^{L} \phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \leq \sum_{i \in \mathcal{U}(0)} bT \cdot \Theta(\sigma(WT)) \cdot \alpha_c L. \tag{87}$$

In addition, we have

$$\sum_{i \in \mathcal{K}^- \setminus \mathcal{U}(0)} \sum_{l=1}^{L} \phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \leq \tilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right). \tag{88}$$

By (76), we can write

$$F(\boldsymbol{X}^{(n)}) \geq \frac{1}{\sqrt{m}L}\left\{\frac{m}{2} \cdot aT \cdot \Theta(\sigma(\alpha T)) \cdot \alpha_r L - \frac{m}{2} \cdot bT \cdot \Theta(\sigma(WT)) \cdot \alpha_c L - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right)\right\}, \tag{89}$$

with

$$a = \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right), \tag{90}$$

$$\text{and} \quad b = \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L) + \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right). \tag{91}$$

$$\begin{aligned}\alpha =&\frac{\eta}{8L^2}\Theta((\alpha_r L-\alpha_c L)^2)-\widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right)\\ =&W-\widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right).\end{aligned} \tag{92}$$

Therefore, we conclude that

$$F(\boldsymbol{X}^{(n)})\geq \frac{\sqrt{m}}{2}\cdot aT\cdot\Theta(\sigma(\alpha T))\cdot(\alpha_r-\alpha_c)-\widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right) \tag{93}$$

There, for any positive sample, we can prove that

$$F(\boldsymbol{X}^{(n)})\geq C,\ \text{where } C \text{ is some positive constant.} \tag{94}$$

Similar to the previous analysis, one can show that the negtive sample $\boldsymbol{X}_n$ leads to

(**S4.1**) Derivation for the convergence rate. Let's find the number of iterations T required such that $F(\boldsymbol{X}^{(n)})\geq 1$, since the label is $+1$. We require

$$\frac{\sqrt{m}}{2}\cdot aT(\alpha_r-\alpha_c)\geq 1+\epsilon. \tag{95}$$

Substituting the value of $a\approx b=\frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L-\alpha_c L)$, the condition becomes

$$\begin{aligned}\frac{\sqrt{m}aT}{2}(\alpha_r-\alpha_c)&=\frac{\sqrt{m}}{2}\cdot\frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L-\alpha_c L)T\cdot(\alpha_r-\alpha_c)\\ &=\frac{\eta T}{4}\Theta((\alpha_r-\alpha_c)^2)\geq 1+\epsilon.\end{aligned} \tag{96}$$

Solving for T, we obtain

$$T\geq\frac{4(1+\epsilon)}{\eta\Theta((\alpha_r-\alpha_c)^2)}\geq\frac{4}{\eta\Theta((\alpha_r-\alpha_c)^2)}. \tag{97}$$

Now, we additionally require that the sigmoid activation $\sigma(\alpha T)$ be sufficiently large, i.e.,

$$\sigma(\alpha T)\geq 1-\epsilon. \tag{98}$$

When z is sufficiently large we can approximate

$$\sigma(z)=\frac{1}{1+e^{-z}}\approx 1-e^{-z}.$$

Substituting $z=\alpha T$, condition (98) becomes:

$$\begin{aligned}\sigma(\alpha T)&\approx 1-e^{-\alpha T}\geq 1-\epsilon,\\ e^{-\alpha T}&\leq\epsilon,\\ \alpha T&\geq-\ln(\epsilon)\\ T&\geq-\frac{\ln(\epsilon)}{\alpha}.\end{aligned} \tag{99}$$

Substituting $\alpha=\frac{\eta}{8L^2}\Theta((\alpha_r L-\alpha_c L)^2)$, we get:

$$T\geq-\ln(\epsilon)\cdot\frac{8L^2}{\eta\Theta((\alpha_r L-\alpha_c L)^2)}. \tag{100}$$

$$T\geq-\ln(\epsilon)\cdot\frac{8}{\eta\Theta((\alpha_r-\alpha_c)^2)}. \tag{101}$$

Hence, by combining (97) and (101), we obtain

$$T \geq \max\left\{\frac{4}{\eta\Theta((\alpha_r - \alpha_c)^2)}, -\ln(\epsilon) \cdot \frac{8}{\eta\Theta((\alpha_r - \alpha_c)^2)}\right\}. \tag{102}$$

By combining (95) and (98) with the expression for the model output $F(\boldsymbol{X}^{(n)})$ in (93), we obtain

$$\begin{aligned} F(\boldsymbol{X}^{(n)}) &\geq (1+\epsilon)\cdot(1-\epsilon) \\ &\geq 1 - \mathcal{O}(\epsilon^2) \end{aligned} \tag{103}$$

Hence, for sufficiently small $\epsilon > 0$, the model output satisfies $F(\boldsymbol{X}^{(n)}) \geq 1$.

Similarly, for a negative sample, one can show by symmetry that the model output satisfies $F(\boldsymbol{X}^{(n)}) \leq 1$.

(**S4.2**) Derivation for the sample complexity.

Now we derive a sample-complexity bound that guarantees zero generalization error.

Assuming enough samples, we can write for sufficiently small $\lambda \ll 1$

$$\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \leq \lambda \cdot \frac{\eta}{2\sqrt{m}L}\Theta(\alpha_r L - \alpha_c L). \tag{104}$$

From this, we can derive a lower bound on the required sample size,

$$\begin{aligned} N &\geq \Omega\left(\lambda^{-2}\cdot\frac{4L^2 d}{\eta^2\Theta((\alpha_r-\alpha_c)^2)}\right) \\ &\geq \Omega\left(\frac{L^2 d}{\eta^2\Theta((\alpha_r-\alpha_c)^2)}\right), \end{aligned} \tag{105}$$

which will be (13) in Theorem 1.

□

C LOCALITY-STRUCTURED DATA

C.1 USEFUL LEMMAS

Lemma C.1 provides bounds on the gradient updates of lucky neurons $i \in \mathcal{W}(t)$ in the directions of both class-relevant features ($\boldsymbol{o}_+$, $\boldsymbol{o}_-$) and irrelevant features.

Lemma C.1. *Suppose $p_1 \leq \langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle \leq q_1$, $p_1 \leq \langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_-\rangle \leq q_1$, and $p_2 \leq \langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_j\rangle \leq q_2$ for $j \neq 1, 2$. Then, for any lucky neuron $i \in \mathcal{W}(t)$ at iteration t, the following bounds hold:*

(L1.1) *A lower bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\begin{aligned} \left\langle -\frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+\right\rangle \geq & \frac{1}{\sqrt{m}L}\cdot\sigma(p_1)\cdot(1-\sigma(q_1))^2\left[(1-\sigma(q_2))^{\Delta L_{\boldsymbol{o}_+}^{+}-2} - (1-\sigma(p_2))^{\Delta L_{\boldsymbol{o}_+}^{-}-2}\right] \\ & - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \end{aligned} \tag{106}$$

(L1.2) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\begin{aligned} \left\langle -\frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+\right\rangle \leq & \frac{1}{\sqrt{m}L}\cdot\sigma(q_1)\cdot(1-\sigma(p_1))^2\left[(1-\sigma(p_2))^{\Delta L_{\boldsymbol{o}_+}^{+}-2} - (1-\sigma(q_2))^{\Delta L_{\boldsymbol{o}_+}^{-}-2}\right] \\ & + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \end{aligned} \tag{107}$$

***(L1.3)** A lower bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\begin{aligned}\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \geq &-\frac{1}{\sqrt{m}L} \cdot \sigma(q_1) \cdot (1-\sigma(p_1))^2 \left[(1-\sigma(p_2))^{\Delta L_{\boldsymbol{o}_-}^{-} -2} - (1-\sigma(q_2))^{\Delta L_{\boldsymbol{o}_-}^{+} -2}\right] \\ &- \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right).\end{aligned} \tag{108}$$

***(L1.4)** An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \tag{109}$$

***(L1.5)** An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_j$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right), \quad \textit{for } j \neq 1, 2. \tag{110}$$

Lemma C.2 shows that, for unlucky neurons associated with the positive class, the gradients in the directions of both class-relevant and irrelevant features are small.

Lemma C.2. *For any unlucky neuron $i \in \mathcal{K}_+ \setminus \mathcal{W}(t)$ at iteration t, the following bounds hold:*

***(L2.1)** An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \tag{111}$$

***(L2.2)** An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \tag{112}$$

***(L2.3)** An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_j$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right), \quad \textit{for } j \neq 1, 2. \tag{113}$$

Lemmas C.3 and C.4, by symmetry, state the analogous results for lucky and unlucky neurons associated with the negative class.

Lemma C.3. *Suppose $p_1 \leq \langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_- \rangle \leq q_1$, $p_1 \leq \langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+ \rangle \leq q_1$, and $p_2 \leq \langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_j \rangle \leq q_2$ for $j \neq 1, 2$. Then, for any lucky neuron $i \in \mathcal{U}(t)$ at iteration t, the following bounds hold:*

***(L3.1)** A lower bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\begin{aligned}\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \geq &\frac{1}{\sqrt{m}L} \cdot \sigma(p_1) \cdot (1-\sigma(q_1))^2 \left[(1-\sigma(q_2))^{\Delta L_{\boldsymbol{o}_-}^{-} -2} - (1-\sigma(p_2))^{\Delta L_{\boldsymbol{o}_-}^{+} -2}\right] \\ &- \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right).\end{aligned} \tag{114}$$

(L3.2) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \frac{1}{\sqrt{mL}} \cdot \sigma(q_1) \cdot (1-\sigma(p_1))^2 \left[(1-\sigma(p_2))^{\Delta L_{\boldsymbol{o}_-}^- -2} - (1-\sigma(q_2))^{\Delta L_{\boldsymbol{o}_-}^+ -2} \right] + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{115}$$

(L3.3) *A lower bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \geq -\frac{1}{\sqrt{mL}} \cdot \sigma(q_1) \cdot (1-\sigma(p_1))^2 \left[(1-\sigma(p_2))^{\Delta L_{\boldsymbol{o}_+}^+ -2} - (1-\sigma(q_2))^{\Delta L_{\boldsymbol{o}_+}^- -2} \right] - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{116}$$

(L3.4) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{117}$$

(L3.5) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_j$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right), \quad \text{for } j \neq 1, 2. \tag{118}$$

Lemma C.4. *For any unlucky neuron $i \in \mathcal{K}_- \setminus \mathcal{U}(t)$ at iteration t, the following bounds hold:*

(L4.1) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_-$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{119}$$

(L4.2) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_+$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{120}$$

(L4.3) *An upper bound on the gradient of $\hat{\mathcal{L}}$ with respect to $\boldsymbol{W}_{O(i,\cdot)}$ at iteration t, in the direction of $\boldsymbol{o}_j$, is given by*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right), \quad \text{for } j \neq 1, 2. \tag{121}$$

Lemma C.5 establishes bounds for the gradient updates of $\boldsymbol{w}_\Delta$ in the class-relevant feature directions.

Lemma C.5. *Suppose $p_1 \leq \langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+ \rangle \leq q_1$ and $r_1^* \leq \langle {\boldsymbol{W}_{O(i,\cdot)}^{(t+1)}}^\top, \boldsymbol{o}_+ \rangle \leq s_1^*$. Let $|\mathcal{W}(t)| = \rho_t^+$ and $|\mathcal{U}(t)| = \rho_t^-$. Then, we have:*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{\sigma(p_1)\,(1-\sigma(q_1))\, r_1^* \cdot \rho_t^+}{\sqrt{m}} - \frac{\sigma(q_1)\,(1-\sigma(p_1))\, s_1^* \cdot \sqrt{m}}{2} - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{122}$$

Suppose $p_1 \leq \langle \boldsymbol{w}_{\Delta}^{(t)}, \boldsymbol{o}_- \rangle \leq q_1$ *and* $r_1^* \leq \langle \boldsymbol{W}_{O(i,\cdot)}^{(t+1)^\top}, \boldsymbol{o}_- \rangle \leq s_1^*$. *Let* $|\mathcal{W}(t)| = \rho_t^+$ *and* $|\mathcal{U}(t)| = \rho_t^-$. *Then, we have:*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_{\Delta}^{(t)}}, \boldsymbol{o}_- \right\rangle \geq \frac{\sigma(p_1)\,(1-\sigma(q_1))\,r_1^* \cdot \rho_t^-}{\sqrt{m}} - \frac{\sigma(q_1)\,(1-\sigma(p_1))\,s_1^* \cdot \sqrt{m}}{2} - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \quad (123)$$

Lemma C.6 establishes bounds for the gradient updates of $\boldsymbol{w}_{\Delta}$ in the directions of irrelevant features.

Lemma C.6. *Suppose* $p_1 \leq \langle \boldsymbol{w}_{\Delta}^{(t)}, \boldsymbol{o}_+ \rangle \leq q_1$, $p_1 \leq \langle \boldsymbol{w}_{\Delta}^{(t)}, \boldsymbol{o}_- \rangle \leq q_1$, $\langle \boldsymbol{w}_{\Delta}^{(t)}, \boldsymbol{o}_j \rangle \leq q_2$ *for* $j \neq 1, 2$, *and* $r_1^* \leq \langle \boldsymbol{W}_{O(i,\cdot)}^{(t)^\top}, \boldsymbol{o}_+ \rangle$. *Let* $\rho_t^+ = |\mathcal{W}(t)|$ *and* $\rho_t^- = |\mathcal{U}(t)|$. *Then we have:*

$$\begin{aligned}\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_{\Delta}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq & -\frac{r_1^*}{2\sqrt{m}} \cdot \sigma(p_1)\,(1-\sigma(q_1)) \left[(1-\sigma(q_2))^{\Delta L_{\boldsymbol{o}_+}^+} \rho_t^+ + (1-\sigma(q_2))^{\Delta L_{\boldsymbol{o}_-}^-} \rho_t^- \right] \\ & + \mathcal{O}\left((1-\sigma(p_2))^{\Delta L_{\boldsymbol{o}_+}^-} \right) + \mathcal{O}\left((1-\sigma(p_2))^{\Delta L_{\boldsymbol{o}_-}^+} \right) + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right).\end{aligned} \quad (124)$$

C.2 Proof of Convergence

Proof of Theorem 2. Similar to the proof of Theorem 1, the proof starts with the base case at $t = 0$ and proceeds to analyze the training dynamics in a deductive manner, providing additional details in deriving the corresponding convergence and sample complexity bounds.

(**S1**) Warm-up (Base case): Training dynamics at the first iteration $t = 0$.

Recall that we set $\boldsymbol{w}_{\Delta}^{(0)} = \boldsymbol{0}$. Then, we have

$$\langle \boldsymbol{w}_{\Delta}^{(0)}, \boldsymbol{o}_+ \rangle = 0, \quad \langle \boldsymbol{w}_{\Delta}^{(0)}, \boldsymbol{o}_- \rangle = 0, \quad \text{and} \quad \langle \boldsymbol{w}_{\Delta}^{(0)}, \boldsymbol{o}_j \rangle = 0 \quad \forall j.$$

(**S1.1**) Training dynamics of $\boldsymbol{W}_{O(i,:)}$ at the first iteration $t = 0$.

From Lemma C.1, identify $p_1 = 0, q_1 = 0, p_2 = 0$ and $q_2 = 0$. Let $\Delta L_{\boldsymbol{o}_+}^+$ and $\Delta L_{\boldsymbol{o}_-}^+$ be the distance between two $\boldsymbol{o}_+$ and $\boldsymbol{o}_-$ features respectively in the positive sample. Similarly, in a negative sample, let the distance between the two $\boldsymbol{o}_+$ tokens as $\Delta L_{\boldsymbol{o}_+}^-$, and the distance between the two $\boldsymbol{o}_-$ tokens as $\Delta L_{\boldsymbol{o}_-}^-$. Then, for any lucky neuron $i \in \mathcal{W}(0)$, we obtain

$$\begin{aligned} & \frac{c'^2}{2\sqrt{m}L} \left[\left(\frac{1}{2}\right)^{\Delta L_{\boldsymbol{o}_+}^+ - 2} - \left(\frac{1}{2}\right)^{\Delta L_{\boldsymbol{o}_+}^- - 2} \right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\ & \leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(0)}}, \boldsymbol{o}_+ \right\rangle \\ & \leq \frac{1}{2\sqrt{m}L} \left[\left(\frac{1}{2}\right)^{\Delta L_{\boldsymbol{o}_+}^+} - \left(\frac{1}{2}\right)^{\Delta L_{\boldsymbol{o}_+}^-} \right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \end{aligned} \quad (125)$$

$$\text{and} \quad \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(0)}}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 1. \quad (126)$$

Recall that we set the number of samples in a batch $N = \mathrm{poly}(d)$.

Suppose the initialization is

$$\boldsymbol{W}_{O(i,\cdot)}(0) = \delta_1 \boldsymbol{o}_+ + \delta_2 \boldsymbol{o}_- + \cdots + \delta_d \boldsymbol{o}_d, \quad \delta_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \xi^2) \quad j = 1, 2, \cdots, d. \quad (127)$$

Then, after one gradient descent step, we have

$$\begin{aligned}
&\delta_1 + \frac{\eta c'^2}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}-2} - \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_+}-2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
&\quad\leq \left\langle {\boldsymbol{W}_{O(i,\cdot)}^{\top}}^{(1)}, \boldsymbol{o}_+ \right\rangle \\
&\quad\leq \delta_1 + \frac{\eta}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}} - \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_+}}\right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)
\end{aligned} \tag{128}$$

$$\text{and} \quad \left\langle {\boldsymbol{W}_{O(i,\cdot)}^{\top}}^{(1)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 1. \tag{129}$$

By applying Lemma C.3, for any lucky neuron $i \in \mathcal{U}(0)$, we obtain

$$\begin{aligned}
&\delta_2 + \frac{\eta c'^2}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_-}-2} - \left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_-}-2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
&\quad\leq \left\langle {\boldsymbol{W}_{O(i,\cdot)}^{\top}}^{(1)}, \boldsymbol{o}_+ \right\rangle \\
&\quad\leq \delta_2 + \frac{\eta}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_-}} - \left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_-}}\right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)
\end{aligned} \tag{130}$$

$$\text{and} \quad \left\langle {\boldsymbol{W}_{O(i,\cdot)}^{\top}}^{(1)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 2. \tag{131}$$

For any unlucky neuron $i \in \mathcal{K}_- \setminus \mathcal{U}(0)$, Lemma C.4 gives

$$\left\langle {\boldsymbol{W}_{O(i,\cdot)}^{\top}}^{(1)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } \forall j. \tag{132}$$

(**S1.2**) Training dynamics of $\boldsymbol{W}_\Delta$ at the first iteration $t = 0$.

Now consider the gradient update for $\boldsymbol{w}_\Delta$. Define:

$$a = \delta_1 + \frac{\eta c'^2}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}-2} - \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_+}-2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)$$

$$b = \delta_1 + \frac{\eta}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}} - \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_+}}\right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)$$

Applying Lemma C.5 with $p_1 = 0$, $q_1 = 0$, $r_1^* = a$, $s_1^* = b$, and $\rho_0^+ = |\mathcal{W}(0)|$, we get

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(0)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{a}{4\sqrt{m}} \cdot \rho_0^+ - \frac{b\sqrt{m}}{8} - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \tag{133}$$

We can relax this lower bound and obtain

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(0)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{c'}{4}\left[\frac{2a}{\sqrt{m}} \cdot \rho_0^+ - \sqrt{m}b\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) =: \alpha \tag{134}$$

Recall that $\delta_1 = \frac{1}{\mathrm{poly}(d)}$. Since $a - b = \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)$ that is sufficiently small,

$$\begin{aligned}\alpha &= \frac{c'}{4}\left[\frac{2a}{\sqrt{m}}\cdot\frac{m}{2} - \sqrt{m}b\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)\\ &= \frac{c'}{4}\left[\sqrt{m}a - \sqrt{m}a\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)\\ &= 0 - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)\\ &= -\widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \approx 0 \end{aligned} \tag{135}$$

From Lemma C.6, we also obtain

$$\left\langle -\frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(0)}}, \boldsymbol{o}_j \right\rangle \le \frac{-a}{8\sqrt{m}}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}}\cdot \rho_0^+ + \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_-}}\cdot\rho_0^-\right] \tag{136}$$

where we apply the lemma with the values

$$p_1 = 0, \quad q_1 = 0, \quad q_2 = 0, \quad \text{and} \quad r_1^* = a.$$

We can relax this upper bound and obtain

$$\left\langle -\frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(0)}}, \boldsymbol{o}_j \right\rangle \le \frac{-ac'}{4\sqrt{m}}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}}\cdot \rho_0^+ + \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_-}}\cdot\rho_0^-\right] =: \gamma \tag{137}$$

Taking $\rho_0^+ = \rho_0^- = \frac{m}{2} + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)$, we can simplify and write

$$\begin{aligned}\gamma &= \frac{-ac'}{4\sqrt{m}}\cdot\frac{m}{2}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}} + \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_-}}\right]\\ &= -\sqrt{m}a\cdot\frac{c'}{8}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}} + \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_-}}\right]\\ &= \frac{-\eta c'^3}{16L}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}-2} - \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_+}-2}\right]\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}} + \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_-}}\right]\\ &\qquad - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right). \end{aligned} \tag{138}$$

<u>(**S2**) Induction Step: Training dynamics at a general iteration t.</u>

Let $\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle = \alpha^* \ge \alpha\cdot t$, $\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_-\rangle = \beta^* \ge \beta\cdot t$, and $\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_j\rangle = \gamma^* \le \gamma\cdot t$, where

$$\beta = \frac{c'}{4}\left[\frac{2a'}{\sqrt{m}}\cdot\rho_0^- - \sqrt{m}b'\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) > 0 \tag{139}$$

$$a' = \delta_2 + \frac{\eta c'^2}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_-}-2} - \left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_-}-2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)$$

$$b' = \delta_2 + \frac{\eta}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_-}} - \left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_-}}\right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)$$

Following the same approach as in (135), we can simplify and obtain

$$\beta = -\widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right). \tag{140}$$

For any lucky neuron $i \in \mathcal{W}(t)$ at the $(t+1)$-th iteration, we have

$$\begin{aligned}
&\frac{c'^2}{\sqrt{mL}} \cdot \sigma(\alpha^*) \left[(1-\sigma(\gamma^*))^{\Delta L^+_{\boldsymbol{o}_+}-2} - (1-\sigma(\gamma^*))^{\Delta L^-_{\boldsymbol{o}_+}-2} \right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
&\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+ \right\rangle \\
&\leq \frac{1}{\sqrt{mL}} \cdot \sigma(\alpha^*) \cdot (1-\sigma(\alpha^*))^2 \left[(1-\sigma(\gamma^*))^{\Delta L^+_{\boldsymbol{o}_+}-2} - (1-\sigma(\gamma^*))^{\Delta L^-_{\boldsymbol{o}_+}-2} \right] \\
&\qquad + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)
\end{aligned} \tag{141}$$

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(1)}_{O(i,\cdot)}}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 1 \tag{142}$$

Note that, $\sigma(\alpha^*) > \frac{1}{2}$ and $\sigma(\gamma^*) < \frac{1}{2}$.

Thus, we obtain the following bound after the second gradient descent step.

$$\begin{aligned}
&\left\langle (\boldsymbol{W}^{(2)}_{O(i,\cdot)})^\top, \boldsymbol{o}_+ \right\rangle \\
&\geq \delta_1 + \frac{\eta c'^2}{\sqrt{mL}} \left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}-1} - \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_+}-1} \right. \\
&\quad \left. + \sigma(\alpha^*) \left((1-\sigma(\gamma^*))^{\Delta L^+_{\boldsymbol{o}_+}-2} - (1-\sigma(\gamma^*))^{\Delta L^-_{\boldsymbol{o}_+}-2} \right) \right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right).
\end{aligned} \tag{143}$$

and

$$\begin{aligned}
&\left\langle (\boldsymbol{W}^{(2)}_{O(i,\cdot)})^\top, \boldsymbol{o}_+ \right\rangle \\
&\leq \delta_1 + \frac{\eta}{2\sqrt{mL}} \left[\left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_-}} - \left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_-}} \right. \\
&\quad \left. + 2\sigma(\alpha^*) \cdot (1-\sigma(\alpha^*))^2 \left((1-\sigma(\gamma^*))^{\Delta L^+_{\boldsymbol{o}_+}-2} - (1-\sigma(\gamma^*))^{\Delta L^-_{\boldsymbol{o}_+}-2} \right) \right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right).
\end{aligned} \tag{144}$$

For the convenience of presentation, we use u to denote the lower bound in (143), and v to denote the upper bound in (144).

Similarly, applying Lemma C.3 to any lucky neuron $i \in \mathcal{U}(1)$ at iteration 2, we get

$$\begin{aligned}
\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(1)}_{O(i,\cdot)}}, \boldsymbol{o}_- \right\rangle \geq & \frac{c'^2}{\sqrt{mL}} \cdot \sigma(\beta^*) \left[(1-\sigma(\gamma^*))^{\Delta L^-_{\boldsymbol{o}_-}-2} - (1-\sigma(\gamma^*))^{\Delta L^+_{\boldsymbol{o}_-}-2} \right] \\
& - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right),
\end{aligned} \tag{145}$$

$$\begin{aligned}
\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(1)}_{O(i,\cdot)}}, \boldsymbol{o}_- \right\rangle \leq & \frac{1}{\sqrt{mL}} \cdot \sigma(\beta^*) \cdot (1-\sigma(\beta^*))^2 \left[(1-\sigma(\gamma^*))^{\Delta L^-_{\boldsymbol{o}_-}-2} - (1-\sigma(\gamma^*))^{\Delta L^+_{\boldsymbol{o}_-}-2} \right] \\
& + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right),
\end{aligned} \tag{146}$$

and

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(1)}_{O(i,\cdot)}}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 1 \tag{147}$$

Applying Lemma C.5 with $p_1 = \alpha^*$, $q_1 = \alpha^*$, $r_1^* = u$, and $s_1^* = v$, we obtain

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(1)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{\sigma(\alpha^*)c'}{2} \left[\frac{2u}{\sqrt{m}} \cdot \rho_1^+ - \sqrt{m}v \right] - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)} \right) =: \chi \tag{148}$$

Since $\rho_1^+ = \frac{m}{2}$ and $u - v = \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right)$ that is sufficiently small,, we have $\chi \approx 0$.

By applying Lemma C.6 with

$$p_1 = \alpha^*(= \beta^*), \quad q_1 = \alpha^*(= \beta^*), \quad q_2 = \gamma^*, \text{and} \quad r_1^* = u, \text{we have}$$

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_j \right\rangle \leq -\frac{c'u}{2\sqrt{m}} \sigma(\alpha^*) \left[(1 - \sigma(\gamma^*))^{\Delta L^+_{\boldsymbol{o}_+}} \rho_t^+ \ + \ (1 - \sigma(\gamma^*))^{\Delta L^-_{\boldsymbol{o}_-}} \rho_t^- \right] =: \iota \tag{149}$$

Note that here we assumed the distribution of ΔL^+ is identical to ΔL^- to have $\alpha^* = \beta^*$.

(**S3**) Induction conclusion: Training dynamics when the algorithm ends.

We proceed by induction on t: the base case $t = 0$ is established in (S1), and the induction step for general t is shown in (S2). For, any lucky neuron $i \in \mathcal{W}(T)$, we obtain

$$\left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(T)}, \boldsymbol{o}_+ \right\rangle \geq aT \tag{150}$$

$$\left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(T)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)} \right) \quad \text{for } j \neq 1 \tag{151}$$

For any lucky neuron $i \in \mathcal{U}(T)$, we obtain

$$\left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(T)}, \boldsymbol{o}_- \right\rangle \geq aT \tag{152}$$

$$\left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(T)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)} \right) \quad \text{for } j \neq 2 \tag{153}$$

Also, we obtain

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_+ \right\rangle \geq \alpha T, \tag{154}$$

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_- \right\rangle \geq \beta T, \tag{155}$$

$$\text{and} \quad \left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_j \right\rangle \leq \gamma T. \tag{156}$$

(**S4**) Derivation for the generalization bound.

We will demonstrate that once the weights have converged at iteration T, the model accurately captures the underlying data distribution, which leads to zero generalization error, as shown in (180).

Consider $z^{(n)} = +1$ as an example. The sequence $\boldsymbol{X}^{(n)} = \begin{bmatrix} \boldsymbol{x}_1^{(n)} & \boldsymbol{x}_2^{(n)} & \cdots & \boldsymbol{x}_L^{(n)} \end{bmatrix}$ contains two $\mathbf{o}_+$ at L_1^+ and L_2^+ and two $\mathbf{o}_-$ at L_1^- and L_2^-.

$$\begin{aligned}
F(\boldsymbol{X}^{(n)}) &= \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m} v_i\,\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \\
&= \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}^+}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) - \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}^-}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \\
&\geq \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) - \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{U}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \\
&\quad - \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}^-\setminus\mathcal{U}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \qquad (157)
\end{aligned}$$

The Mamba output $\boldsymbol{y}_l^{(n)}$ is defined as

$$\boldsymbol{y}_l^{(n)} = \sum_{s=1}^{l}\left(\prod_{j=s+1}^{l}\left(1-\sigma(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_j^{(n)})\right)\right)\cdot\sigma(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_s^{(n)})\cdot(\boldsymbol{x}_s^{(n)\top}\boldsymbol{x}_l^{(n)})\boldsymbol{x}_s^{(n)}. \qquad (158)$$

We now derive a lower bound for

$$\sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\phi(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l).$$

To that end, consider the aggregated projection

$$\sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l = \sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\sum_{j=1}^{d}\langle\boldsymbol{W}_{O(i,\cdot)}^\top,\boldsymbol{o}_j\rangle\cdot\langle\boldsymbol{y}_l,\boldsymbol{o}_j\rangle. \qquad (159)$$

For any $i\in\mathcal{W}(0)$, we know that

$$\langle\boldsymbol{W}_{O(i,\cdot)}^\top,\boldsymbol{o}_+\rangle \geq aT. \qquad (160)$$

Hence, let's obtain a lower bound for $\langle\boldsymbol{y}_l,\boldsymbol{o}_+\rangle$

We only need to consider the cases where $\boldsymbol{x}_s=\boldsymbol{o}_+$ for some s in the range $1\leq s\leq l$. In particular, we will focus on the following instances:

$$s=L_1^+ \text{ and } l\in\{L_1^+,L_2^+\}, \qquad s=L_2^+ \text{ and } l=L_2^+.$$

After T iterations, we know

$$\langle\boldsymbol{w}_\Delta,\boldsymbol{o}_+\rangle\geq\alpha T, \quad \langle\boldsymbol{w}_\Delta,\boldsymbol{o}_-\rangle\geq\beta T, \quad \langle\boldsymbol{w}_\Delta,\boldsymbol{o}_j\rangle\leq\gamma T \quad \text{for } j\neq 1,2. \qquad (161)$$

Therefore,

$$\left\langle\boldsymbol{y}_{L_1^+},\boldsymbol{o}_+\right\rangle = \sigma\left(\langle\boldsymbol{w}_\Delta,\boldsymbol{o}_+\rangle\right)\geq\sigma(\alpha T). \qquad (162)$$

We have,

$$\langle\boldsymbol{w}_\Delta,\boldsymbol{o}_+\rangle\leq W_1T, \qquad \langle\boldsymbol{w}_\Delta,\boldsymbol{o}_-\rangle\leq W_2T,$$

where

$$W_1=\tilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right), \qquad (163)$$

$$W_2=\tilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right). \qquad (164)$$

Then we obtain the following:

$$\begin{aligned}\left\langle \boldsymbol{y}_{L_2^+}, \boldsymbol{o}_+ \right\rangle &\geq \sigma(\alpha T) + (1-\sigma(W_1 T))\,(1-\sigma(W_2 T))\,(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L_{\boldsymbol{o}_+}^+ - 2} \cdot \sigma(\alpha T) \\ &= \sigma(\alpha T)\left[1 + (1-\sigma(W_1 T))\,(1-\sigma(W_2 T))\,(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L_{\boldsymbol{o}_+}^+ - 2}\right]. \qquad (165)\end{aligned}$$

We now lower bound the objective

$$\sum_{i\in\mathcal{W}(0)} \sum_{l=1}^{L} \phi(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l).$$

We begin with

$$\sum_{i\in\mathcal{W}(0)} \sum_{l=1}^{L} \phi(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l) \geq \sum_{i\in\mathcal{W}(0)} \left[\phi(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_{L_1^+}) + \phi(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_{L_2^+})\right].$$

Note that

$$\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_{L_1^+} = \sum_{j=1}^{d} \left\langle \boldsymbol{W}_{O(i,\cdot)}^\top, \boldsymbol{o}_j \right\rangle \left\langle \boldsymbol{y}_{L_1^+}, \boldsymbol{o}_j \right\rangle,$$

and $\boldsymbol{y}_{L_1^+}$ has only $\boldsymbol{o}_+$ component.

Therefore,

$$\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_{L_1^+} = \left\langle \boldsymbol{W}_{O(i,\cdot)}^\top, \boldsymbol{o}_+ \right\rangle \left\langle \boldsymbol{y}_{L_1^+}, \boldsymbol{o}_+ \right\rangle \geq aT \cdot \sigma(\alpha T) > 0.$$

Similarly, we can write

$$\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_{L_2^+} \geq aT \cdot \sigma(\alpha T)\left[1 + (1-\sigma(W_1 T))\,(1-\sigma(W_2 T))\,(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L_{\boldsymbol{o}_+}^+ - 2}\right] > 0.$$

Applying $\phi(z) = z$ for positive z, we obtain

$$\phi(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_{L_1^+}) \geq aT \cdot \sigma(\alpha T),$$

$$\phi(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_{L_2^+}) \geq aT \cdot \sigma(\alpha T)\left[1 + (1-\sigma(W_1 T))\,(1-\sigma(W_2 T))\,(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L_{\boldsymbol{o}_+}^+ - 2}\right].$$

Hence,

$$\begin{aligned}\sum_{i\in\mathcal{W}(0)} \sum_{l=1}^{L} \phi(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l) \geq \sum_{i\in\mathcal{W}(0)} & aT \cdot \sigma(\alpha T) \cdot \\ & \left[2 + (1-\sigma(W_1 T))\,(1-\sigma(W_2 T))\,(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L_{\boldsymbol{o}_+}^+ - 2}\right]. \qquad (166)\end{aligned}$$

Next, we derive an upper bound for

$$\sum_{i\in\mathcal{U}(0)} \sum_{l=1}^{L} \phi\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right).$$

For any $i \in \mathcal{U}(0)$, we know that

$$0 < \langle \boldsymbol{W}_{O(i,\cdot)}^\top, \boldsymbol{o}_- \rangle \leq bT. \qquad (167)$$

We now derive an upper bound for $\langle \boldsymbol{y}_l, \boldsymbol{o}_- \rangle$. We need to focus on the following instances:

$$s = L_1^- \text{ and } l \in \{L_1^-, L_2^-\}, \qquad s = L_2^- \text{ and } l = L_2^-.$$

$$\left\langle \boldsymbol{y}_{L_1^-}, \boldsymbol{o}_- \right\rangle = \sigma\left(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_- \rangle\right) \leq \sigma(W_2 T). \qquad (168)$$

$$\begin{aligned}\left\langle \boldsymbol{y}_{L_2^-}, \boldsymbol{o}_- \right\rangle &\leq \sigma(W_2T) + (1-\sigma(\alpha T))\,(1-\sigma(\beta T))\,(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L^+_{\boldsymbol{o}_-}-2} \cdot \sigma(W_2T) \\ &= \sigma(W_2T)\left[1 + (1-\sigma(\alpha T))\,(1-\sigma(\beta T))\,(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L^+_{\boldsymbol{o}_-}-2}\right]. \end{aligned} \tag{169}$$

Hence,

$$\begin{aligned}\sum_{i\in\mathcal{U}(0)} \sum_{l=1}^{L} \phi(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l) \leq \sum_{i\in\mathcal{U}(0)} bT\cdot\sigma(W_2T)\cdot& \\ \left[2 + (1-\sigma(\alpha T))\,(1-\sigma(\beta T))\,(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L^+_{\boldsymbol{o}_-}-2}\right]&.\end{aligned}$$

In addition, we have

$$\sum_{i\in\mathcal{K}^-\setminus\mathcal{U}(0)} \sum_{l=1}^{L} \phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \leq \tilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right). \tag{170}$$

By (157), we can write

$$\begin{aligned}F(\boldsymbol{X}^{(n)}) \geq \frac{1}{\sqrt{m}L}\Bigg\{&\frac{m}{2}\cdot aT\cdot\sigma(\alpha T)\left[2 + (1-\sigma(W_1T))(1-\sigma(W_2T))(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L^+_{\boldsymbol{o}_+}-2}\right] \\ &-\frac{m}{2}\cdot bT\cdot\sigma(W_2T)\left[2 + (1-\sigma(\alpha T))\,(1-\sigma(\beta T))\,(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L^+_{\boldsymbol{o}_-}-2}\right] \\ &-\widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)\Bigg\},\end{aligned} \tag{171}$$

with

$$a = \frac{\eta c'^2}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}-2} - \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_+}-2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right), \tag{172}$$

$$\text{and}\quad b = \frac{\eta}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^+_{\boldsymbol{o}_+}} - \left(\frac{1}{2}\right)^{\Delta L^-_{\boldsymbol{o}_+}}\right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right). \tag{173}$$

$$\begin{aligned}\alpha &= \frac{c'}{4}\left[\frac{2a}{\sqrt{m}}\cdot\rho_0^+ - \sqrt{m}b\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\ &= -\widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)\end{aligned} \tag{174}$$

Therefore, we conclude that

$$\begin{aligned}F(\boldsymbol{X}^{(n)}) \geq \frac{1}{\sqrt{m}L}\Bigg\{&\frac{m}{2}\cdot aT\cdot\sigma(\alpha T)(1-\sigma(W_1T))(1-\sigma(W_2T)) \\ &\left[(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L^+_{\boldsymbol{o}_+}-2} - (1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L^+_{\boldsymbol{o}_-}-2}\right]\Bigg\} - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)\end{aligned} \tag{175}$$

If we can show $\left[(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L^+_{\boldsymbol{o}_+}-2} - (1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j\rangle))^{\Delta L^+_{\boldsymbol{o}_-}-2}\right] > 0$, then we can prove $F(\boldsymbol{X}^{(n)}) \geq C$ for some positive constant C.

First define a random variable $\psi_1 = \langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j \rangle$. Then, we have from the definition of our locality-structured data type

$$\mathbb{E}_n \left[(1-\sigma(\psi_1))^{\Delta L^+_{o_+} - 2} - (1-\sigma(\psi_1))^{\Delta L^+_{o_-} - 2} \right] = k' > 0 \tag{176}$$

for some positive constant k'.

The random variable $\psi_2 = (1-\sigma(\psi_1))^{\Delta L^+_{o_+} - 2} - (1-\sigma(\psi_1))^{\Delta L^+_{o_-} - 2}$ is bounded above by 1.

Applying Hoeffding's bound, for any $q > 0$,

$$\mathbb{P}\left(|\psi_2 - \mathbb{E}\psi_2| \gtrsim \sqrt{\frac{q \log N}{N}} \right) \leq N^{-q}. \tag{177}$$

From this we can conclude that,

$$\psi_2 = \left[(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j \rangle))^{\Delta L^+_{o_+} - 2} - (1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j \rangle))^{\Delta L^+_{o_-} - 2} \right] \geq k' - \mathcal{O}\left(\sqrt{\frac{q \log N}{N}} \right), \tag{178}$$

with probability at most N^{-q}.

Hence, for sufficiently large N, we have from (176)

$$\left[(1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j \rangle))^{\Delta L^+_{o_+} - 2} - (1-\sigma(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j \rangle))^{\Delta L^+_{o_-} - 2} \right] > 0 \tag{179}$$

Therefore,

$$F(\boldsymbol{X}^{(n)}) \geq C, \text{ where } C \text{ is some positive constant.} \tag{180}$$

Similarly, for a negative sample, one can show by symmetry that the model output satisfies $F(\boldsymbol{X}^{(n)}) \leq 1$.

(**S4.1**) Derivation for the convergence rate. Let's find the number of iterations T required such that $F(\boldsymbol{X}^{(n)}) \geq 1$, since the label is $+1$. We require

$$\frac{1}{\sqrt{mL}} \cdot \frac{m}{2} \cdot aT \cdot \sigma(\alpha T) \geq 1 + \epsilon. \tag{181}$$

Substituting the value of $a = \frac{\eta}{2\sqrt{mL}} \left[\left(\frac{1}{2}\right)^{\Delta L^+_{o_+}} - \left(\frac{1}{2}\right)^{\Delta L^-_{o_+}} \right]$ and $\sigma(\alpha T) \approx \frac{1}{2}$ since $\alpha \approx 0$, the condition becomes

$$\begin{aligned} \frac{\sqrt{m} aT}{4L} &= \frac{\sqrt{m}}{4L} \cdot \frac{\eta}{2\sqrt{mL}} \left[\left(\frac{1}{2}\right)^{\Delta L^+_{o_+}} - \left(\frac{1}{2}\right)^{\Delta L^-_{o_+}} \right] T \\ &= \frac{\eta T}{8L^2} \left[\left(\frac{1}{2}\right)^{\Delta L^+_{o_+}} - \left(\frac{1}{2}\right)^{\Delta L^-_{o_+}} \right] \geq 1 + \epsilon. \end{aligned} \tag{182}$$

Solving for T, we obtain

$$T \geq \frac{8L^2(1+\epsilon)}{\eta \left[\left(\frac{1}{2}\right)^{\Delta L^+_{o_+}} - \left(\frac{1}{2}\right)^{\Delta L^-_{o_+}} \right]} \geq \frac{8L^2}{\eta \left[\left(\frac{1}{2}\right)^{\Delta L^+_{o_+}} - \left(\frac{1}{2}\right)^{\Delta L^-_{o_+}} \right]}. \tag{183}$$

By combining (181) with the expression for the model output $F(\boldsymbol{X}^{(n)})$ in (175), we obtain

$$F(\boldsymbol{X}^{(n)}) \geq (1+\epsilon) \tag{184}$$

Hence, the model output satisfies $F(\boldsymbol{X}^{(n)}) \geq 1$.

(**S4.2**) Derivation for the sample complexity. Now we derive a sample-complexity bound that guarantees zero generalization error.

Assuming enough samples, we can write for sufficiently small $\lambda \ll 1$

$$\mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) \leq \lambda \cdot \frac{\eta}{2\sqrt{m}L}\left[\left(\frac{1}{2}\right)^{\Delta L^{+}_{\boldsymbol{o}_+}} - \left(\frac{1}{2}\right)^{\Delta L^{-}_{\boldsymbol{o}_+}}\right]. \tag{185}$$

From this we can derive a lower bound on the required sample size,

$$\begin{aligned} N &\geq \Omega\left(\lambda^{-2} \cdot \frac{4L^2 d}{\eta^2\left[\left(\frac{1}{2}\right)^{\Delta L^{+}_{\boldsymbol{o}_+}} - \left(\frac{1}{2}\right)^{\Delta L^{-}_{\boldsymbol{o}_+}}\right]^2}\right) \\ &\geq \Omega\left(\frac{L^2 d}{\eta^2\left[\left(\frac{1}{2}\right)^{\Delta L^{+}_{\boldsymbol{o}_+}} - \left(\frac{1}{2}\right)^{\Delta L^{-}_{\boldsymbol{o}_+}}\right]^2}\right), \end{aligned} \tag{186}$$

which will be (19) in Theorem 2.

□

D Proof of Lemmas in Appendix B

Please refer to the supplementary material for this section. We defer all proofs to the supplementary material, as the high-level ideas underlying the lemmas overlap with those presented in Appendix C for locality data. However, the case of locality-structured data presents additional challenges. Appendix E provides the complete proofs for the locality-structured data, which contain the main technical ideas.

E Proof of Lemmas in Appendix C

E.1 Proof of Lemma C.1

Proof. We know that the gradient of the loss function for the n^{th} sample is

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{W}_{O(i,\cdot)}} &= \frac{\partial \ell}{\partial F(\boldsymbol{X}^{(n)})} \cdot \frac{\partial F(\boldsymbol{X}^{(n)})}{\partial \boldsymbol{W}_{O(i,\cdot)}} \\ &= -\frac{z^{(n)}}{L}\sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}. \end{aligned} \tag{187}$$

If we consider the gradient for the population loss,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}} = -\mathbb{E}\left[\frac{z^{(n)}}{L}\sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right] \tag{188}$$

$$\begin{aligned} &= -\mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right] \\ &\quad + \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right]. \end{aligned} \tag{189}$$

We are given that

$$p_1 \leq \left\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+ \right\rangle \leq q_1, \quad p_2 \leq \left\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_- \right\rangle \leq q_2, \quad \text{and} \quad p_3 \leq \left\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_j \right\rangle \leq q_3 \quad \text{for } j \neq 1, 2. \tag{190}$$

The Mamba output can be written as

$$\boldsymbol{y}_l(t) = \sum_{s=1}^{l} \left(\prod_{j=s+1}^{l} \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{x}_j) \right) \right) \cdot \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{x}_s) \cdot (\boldsymbol{x}_s^\top \boldsymbol{x}_l) \boldsymbol{x}_s \tag{191}$$

We have to consider FOUR cases.

Case I: $l = s = L_1^+$

$$\boldsymbol{x}_s = \boldsymbol{x}_l = \boldsymbol{o}_+ \tag{192}$$

$$\left\langle \mathbb{E}\boldsymbol{y}_{L_1^+}, \boldsymbol{o}_+ \right\rangle = \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_+) \geq \sigma(p_1) \pm \mathcal{O}(\tau) = \frac{1}{1 + e^{-p_1}} \pm \mathcal{O}(\tau). \tag{193}$$

Case II: $l = s = L_2^+$

$$\left\langle \mathbb{E}\boldsymbol{y}_{L_2^+, L_2^+}, \boldsymbol{o}_+ \right\rangle = \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_+) \pm \mathcal{O}(\tau). \tag{194}$$

Case III: $l = L_2^+$, $s = L_1^+$

$$\begin{aligned} \left\langle \mathbb{E}\boldsymbol{y}_{L_2^+, L_1^+}, \boldsymbol{o}_+ \right\rangle = & \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_+) \right) \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_-) \right) \\ & \cdot \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_j) \right)^{\Delta L_{\boldsymbol{o}_+}^+ - 2} \cdot \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_+) \pm \mathcal{O}(\tau). \end{aligned} \tag{195}$$

Combining (194) and (195), we obtain

$$\begin{aligned} \left\langle \mathbb{E}\boldsymbol{y}_{L_2^+}, \boldsymbol{o}_+ \right\rangle = & \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_+) \\ & + \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_+) \right) \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_-) \right) \\ & \cdot \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_j) \right)^{\Delta L_{\boldsymbol{o}_+}^+ - 2} \cdot \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_+) \pm \mathcal{O}(\tau). \end{aligned} \tag{196}$$

Case IV: Others

For the other token positions, $\boldsymbol{x}_l \neq \boldsymbol{o}_+$. Since we assume orthogonality among the features, $\boldsymbol{y}_l = 0$.

From our initialization, for the lucky neuron $i \in \mathcal{W}(0)$, $v_i = +\frac{1}{\sqrt{m}}$. For $i \in \mathcal{W}(0)$, and $z^{(n)} = +1$, we have

$$\begin{aligned} & \left\langle \mathbb{E}_{z=+1} \left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)} \right) \cdot \boldsymbol{y}_l^{(n)} \right], \boldsymbol{o}_+ \right\rangle \\ & = \frac{1}{\sqrt{m}L} \cdot \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_+) \left[2 + \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_+) \right) \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_-) \right) \right. \\ & \quad \left. \cdot \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{o}_j) \right)^{\Delta L_{\boldsymbol{o}_+}^+ - 2} \right] \pm \mathcal{O}(\tau). \end{aligned} \tag{197}$$

Similarly for $z = -1$, we can obtain

$$\begin{aligned}
&\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\
&= \frac{1}{\sqrt{m}L} \cdot \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_+) \bigg[2 + \Big(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_+)\Big)\Big(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_-)\Big) \\
&\quad \cdot \Big(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_j)\Big)^{\Delta L^-_{\boldsymbol{o}_+}-2}\bigg] \pm \mathcal{O}(\tau).
\end{aligned} \tag{198}$$

Therfore, combining (197) and (198),

$$\begin{aligned}
\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\
&\quad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\
&= \frac{1}{\sqrt{m}L} \cdot \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_+) \cdot \Big(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_+)\Big) \cdot \Big(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_-)\Big) \cdot \\
&\quad \bigg[\Big(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_j)\Big)^{\Delta L^+_{\boldsymbol{o}_+}-2} - \Big(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_j)\Big)^{\Delta L^-_{\boldsymbol{o}_+}-2}\bigg] \pm \mathcal{O}(\tau).
\end{aligned} \tag{199}$$

We aim to bound the deviation between the gradient of the population loss and that of the empirical loss. Specifically, $\left\|\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}} - \frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}\right\|_2 = \left\|\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\gamma}_n\right\|_2$, where

$$\boldsymbol{\gamma}_n = \frac{z^{(n)}}{L} \sum_{l=1}^{L} v_i \, \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \boldsymbol{y}_l^{(n)}. \tag{200}$$

Consider a fixed vector $\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\|_2 = 1$. We will show that $\boldsymbol{\alpha}^\top \boldsymbol{\gamma}_n$ is a sub-Gaussian random variable.

$$\left|\boldsymbol{\alpha}^\top \boldsymbol{\gamma}_n\right| \le \|\boldsymbol{\alpha}\|_2 \cdot \|\boldsymbol{\gamma}_n\|_2 = \|\boldsymbol{\gamma}_n\|_2. \tag{201}$$

By the problem setup, we know that

$$|v_i| = \frac{1}{\sqrt{m}}, \quad |z^{(n)}| = 1, \quad \left|\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\right| \le 1. \tag{202}$$

Recall the Mamba output,

$$\boldsymbol{y}_l^{(n)}(t) = \sum_{s=1}^{l} \left(\prod_{j=s+1}^{l} \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{x}_j)\right)\right) \cdot \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{x}_s) \cdot (\boldsymbol{x}_s^\top \boldsymbol{x}_l)\boldsymbol{x}_s. \tag{203}$$

Since $\|\boldsymbol{x}_s\|_2 = 1$, we get

$$\begin{aligned}
\left\|\boldsymbol{y}_l^{(n)}\right\|_2 &\le \sum_{s=1}^{l} \left|a^{l-s+1} \cdot \left(\boldsymbol{x}_s^\top \boldsymbol{x}_l\right)\right| \cdot \|\boldsymbol{x}_s\|_2 \\
&\le \sum_{s=1}^{l} \frac{a}{1-a} \cdot 1 \cdot 1 = a' \quad \text{(where } a' \text{ denotes a constant).}
\end{aligned} \tag{204}$$

Therefore, the norm of $\boldsymbol{\gamma}_n$ satisfies

$$\begin{aligned}\|\boldsymbol{\gamma}_n\|_2 &\leq \frac{1}{L}\sum_{l=1}^{L}|v_i|\cdot\left|\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\right|\cdot\left\|\boldsymbol{y}_l^{(n)}\right\|_2\\ &\leq \frac{1}{L}\cdot\frac{1}{\sqrt{m}}\sum_{l=1}^{L}\left\|\boldsymbol{y}_l^{(n)}\right\|_2\\ &\leq \frac{1}{L}\cdot\frac{1}{\sqrt{m}}\cdot\sum_{l=1}^{L}a' = \frac{a'}{\sqrt{m}}.\end{aligned} \tag{205}$$

Hence,

$$\left|\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n\right| \leq \frac{a'}{\sqrt{m}} \quad \text{(bounded)}. \tag{206}$$

This implies that $\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n$ is sub-Gaussian with variance proxy

$$\sigma^2 = \mathcal{O}\left(\frac{1}{m}\right). \tag{207}$$

Now consider the independent sub-Gaussian variables $\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_N$, each bounded as

$$-\frac{1}{\sqrt{m}} \leq \boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n \leq \frac{1}{\sqrt{m}}. \tag{208}$$

Applying Hoeffding's inequality, for any $q>0$,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n\right| \gtrsim \sqrt{\frac{q\log N}{mN}}\right) \leq N^{-q}. \tag{209}$$

Observe that this can be written as

$$\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n = \boldsymbol{\alpha}^\top\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\gamma}_n\right) := \boldsymbol{\alpha}^\top\boldsymbol{\zeta}. \tag{210}$$

Therefore, by Hoeffding's inequality (cf. (209)),

$$\mathbb{P}\left(\left|\boldsymbol{\alpha}^\top\boldsymbol{\zeta}\right| \gtrsim \sqrt{\frac{q\log N}{mN}}\right) \leq N^{-q}. \tag{211}$$

To bound $\|\boldsymbol{\zeta}\|_2$, we use the dual norm identity

$$\|\boldsymbol{\zeta}\|_2 = \sup_{\|\boldsymbol{\alpha}\|_2=1}\boldsymbol{\alpha}^\top\boldsymbol{\zeta}. \tag{212}$$

We apply an ε-cover argument to obtain

$$\begin{aligned}\sup_{\|\boldsymbol{\alpha}\|_2=1}\boldsymbol{\alpha}^\top\boldsymbol{\zeta} &\leq \frac{1}{1-\varepsilon}\max_{\boldsymbol{\alpha}\in\mathcal{C}_\varepsilon}\boldsymbol{\alpha}^\top\boldsymbol{\zeta}\\ &\leq 2\max_{\boldsymbol{\alpha}\in\mathcal{C}_{1/2}}\boldsymbol{\alpha}^\top\boldsymbol{\zeta}.\end{aligned} \tag{213}$$

We have shown that for any fixed $\boldsymbol{\alpha}$,

$$\mathbb{P}\left(\left|\boldsymbol{\alpha}^\top\boldsymbol{\zeta}\right| \gtrsim \sqrt{\frac{q\log N}{mN}}\right) \leq N^{-q}. \tag{214}$$

Therefore, for all fixed $\boldsymbol{\alpha} \in \mathcal{C}_{1/2}$,

$$\left|\boldsymbol{\alpha}^\top \boldsymbol{\zeta}\right| \gtrsim \sqrt{\frac{q \log N}{mN}} \quad \text{with probability at most } N^{-q}. \tag{215}$$

Then,

$$\max_{\boldsymbol{\alpha} \in \mathcal{C}_{1/2}} \left|\boldsymbol{\alpha}^\top \boldsymbol{\zeta}\right| \gtrsim \sqrt{\frac{q \log N}{mN}} \quad \text{with probability at most } |\mathcal{C}_{1/2}| N^{-q}. \tag{216}$$

Recall that the covering number satisfies

$$|\mathcal{C}_\varepsilon| \leq \left(\frac{3B}{\varepsilon}\right)^d. \tag{217}$$

For $B = 1$ and $\varepsilon = \frac{1}{2}$, we have

$$|\mathcal{C}_{1/2}| \leq 6^d. \tag{218}$$

We can therefore write

$$\mathbb{P}\left(\|\boldsymbol{\zeta}\|_2 \gtrsim \sqrt{\frac{q \log N}{mN}}\right) \leq 6^d \cdot N^{-q}. \tag{219}$$

We want this probability to be sufficiently small. Set $q = d$, so that

$$\mathbb{P}\left(\|\boldsymbol{\zeta}\|_2 \gtrsim 2\sqrt{\frac{d \log N}{mN}}\right) \leq \left(\frac{N}{6}\right)^{-d}. \tag{220}$$

Hence, the deviation is bounded with high probability:

$$\|\boldsymbol{\zeta}\|_2 > \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) \quad \text{with probability at most } \mathcal{O}(N^{-d}). \tag{221}$$

Or equivalently, with probability at most $\mathcal{O}(N^{-d})$,

$$\left\|\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\gamma}_n\right\|_2 > \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \tag{222}$$

That is, with high probability $1 - \mathcal{O}(N^{-d})$, we have

$$\left\|\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\gamma}_n\right\|_2 \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \tag{223}$$

Using the identities

$$-\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\gamma}_n, \quad -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} = \mathbb{E}\boldsymbol{\gamma}_n, \tag{224}$$

we conclude that, with high probability,

$$\left\|\left(-\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}\right) - \left(-\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}\right)\right\|_2 = \left\|\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} - \frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}\right\|_2 \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \tag{225}$$

Using the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\left|\left\langle \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} - \frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+ \right\rangle\right| &\leq \left\| \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} - \frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} \right\|_2 \cdot \|\boldsymbol{o}_+\|_2 \\ &= \left\| \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} - \frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} \right\|_2 \qquad (\text{since } \|\boldsymbol{o}_+\|_2 = 1) \\ &\leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right).\end{aligned} \tag{226}$$

Therefore, we obtain

$$\begin{aligned}\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+ \right\rangle - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) &\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+ \right\rangle \\ &\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+ \right\rangle + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right).\end{aligned} \tag{227}$$

By pairing (199) with the given the conditions on $\boldsymbol{w}_\triangle$ in (190), we can write

$$\begin{aligned}&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+ \right\rangle \\ \geq & \frac{1}{\sqrt{m}L} \cdot \sigma(p_1) \cdot (1-\sigma(q_1)) \cdot (1-\sigma(q_2)) \left[(1-\sigma(q_3))^{\Delta L^{+}_{\boldsymbol{o}_+}-2} - (1-\sigma(p_3))^{\Delta L^{-}_{\boldsymbol{o}_+}-2} \right] - \mathcal{O}(\tau)\end{aligned} \tag{228}$$

and

$$\begin{aligned}&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+ \right\rangle \\ \leq & \frac{1}{\sqrt{m}L} \cdot \sigma(q_1) \cdot (1-\sigma(p_1)) \cdot (1-\sigma(p_2)) \left[(1-\sigma(p_3))^{\Delta L^{+}_{\boldsymbol{o}_+}-2} - (1-\sigma(q_3))^{\Delta L^{-}_{\boldsymbol{o}_+}-2} \right] + \mathcal{O}(\tau)\end{aligned} \tag{229}$$

Therefore, we can obtain the lower bound and the upper bound of $\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+ \right\rangle$ as

$$\begin{aligned}&\frac{1}{\sqrt{m}L} \cdot \sigma(p_1) \cdot (1-\sigma(q_1)) \cdot (1-\sigma(q_2)) \Big[(1-\sigma(q_3))^{\Delta L^{+}_{\boldsymbol{o}_+}-2} \\ &\qquad\qquad - (1-\sigma(p_3))^{\Delta L^{-}_{\boldsymbol{o}_+}-2} \Big] - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) - \mathcal{O}(\tau) \\ &\qquad\qquad \leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+ \right\rangle\end{aligned} \tag{230}$$

and

$$\begin{aligned}&\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+ \right\rangle \\ \leq & \frac{1}{\sqrt{m}L} \cdot \sigma(q_1) \cdot (1-\sigma(p_1)) \cdot (1-\sigma(p_2)) \left[(1-\sigma(p_3))^{\Delta L^{+}_{\boldsymbol{o}_+}-2} - (1-\sigma(q_3))^{\Delta L^{-}_{\boldsymbol{o}_+}-2} \right] \\ & + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau).\end{aligned} \tag{231}$$

This concludes the proof of (106) and (107) in Lemma C.1.

To obtain $\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_-\right\rangle$, we have to consider $\mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right]$.

If $\boldsymbol{W}^{(t)}_{O(i,\cdot)}\boldsymbol{o}_- > 0$,

$$\begin{aligned}
&\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_-\right\rangle \\
&\quad = \frac{1}{\sqrt{m}L} \cdot \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_-)\left[2 + \Big(1-\sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_-)\Big)\Big(1-\sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_+)\Big)\right. \\
&\qquad \left. \cdot \Big(1-\sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_j)\Big)^{\Delta L^-_{\boldsymbol{o}_-}-2}\right] \pm \mathcal{O}(\tau).
\end{aligned} \tag{232}$$

If $\boldsymbol{W}^{(t)}_{O(i,\cdot)}\boldsymbol{o}_- \leq 0$,

$$\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_-\right\rangle = 0 \pm \mathcal{O}(\tau). \tag{233}$$

From (189), We know that

$$\begin{aligned}
\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_-\right\rangle = &\left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_-\right\rangle \\
&- \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_-\right\rangle.
\end{aligned} \tag{234}$$

Hence, combining both cases, we conclude

$$\begin{aligned}
&- \frac{1}{\sqrt{m}L} \cdot \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_-)\left[2 + \Big(1-\sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_-)\Big)\Big(1-\sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_+)\Big)\right. \\
&\qquad \left. \cdot \Big(1-\sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_j)\Big)^{\Delta L^-_{\boldsymbol{o}_-}-2}\right] - \mathcal{O}(\tau) \\
&\quad \leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_-\right\rangle \leq \mathcal{O}(\tau).
\end{aligned} \tag{235}$$

From (225), similar to (227), we can write

$$\begin{aligned}
&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_-\right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\
&\quad \leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_-\right\rangle \\
&\quad \leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_-\right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).
\end{aligned} \tag{236}$$

Hence, we have

$$\begin{aligned}
&-\frac{1}{\sqrt{mL}}\cdot\sigma(\boldsymbol{w}_{\Delta}^{(t)^{\top}}\boldsymbol{o}_-)\left[2+\left(1-\sigma(\boldsymbol{w}_{\Delta}^{(t)^{\top}}\boldsymbol{o}_-)\right)\left(1-\sigma(\boldsymbol{w}_{\Delta}^{(t)^{\top}}\boldsymbol{o}_+)\right)\right.\\
&\qquad\left.\cdot\left(1-\sigma(\boldsymbol{w}_{\Delta}^{(t)^{\top}}\boldsymbol{o}_j)\right)^{\Delta L^-_{\boldsymbol{o}_-}-2}\right]-\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right)-\mathcal{O}(\tau)\\
&\le\left\langle-\frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{W}^{(t)}_{O(i,\cdot)}},\boldsymbol{o}_-\right\rangle\le\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right)+\mathcal{O}(\tau).
\end{aligned}\tag{237}$$

This concludes the proof of (108) and (109) in Lemma C.1.

Now consider $\left\langle-\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}^{(t)}_{O(i,\cdot)}},\boldsymbol{o}_j\right\rangle$ for $j\neq 1,2$.

$$\begin{aligned}
\left\langle-\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}^{(t)}_{O(i,\cdot)}},\boldsymbol{o}_j\right\rangle&=\left\langle\mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i\cdot\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot\boldsymbol{y}_l^{(n)}\right],\boldsymbol{o}_j\right\rangle\\
&\qquad-\left\langle\mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i\cdot\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot\boldsymbol{y}_l^{(n)}\right],\boldsymbol{o}_j\right\rangle\\
&:=\langle I_1,\boldsymbol{o}_j\rangle-\langle I_2,\boldsymbol{o}_j\rangle.
\end{aligned}\tag{238}$$

Because $\boldsymbol{o}_j$ for $j\neq 1,2$ is identical in both I_1 and I_2, $\langle I_1,\boldsymbol{o}_j\rangle-\langle I_2,\boldsymbol{o}_j\rangle=0\pm\mathcal{O}(\tau)$. Hence, $\left\langle-\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}^{(t)}_{O(i,\cdot)}},\boldsymbol{o}_j\right\rangle=0\pm\mathcal{O}(\tau)$. From (225), similar to (227), we can write

$$\begin{aligned}
&\left\langle-\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}^{(t)}_{O(i,\cdot)}},\boldsymbol{o}_j\right\rangle-\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right)\\
&\le\left\langle-\frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{W}^{(t)}_{O(i,\cdot)}},\boldsymbol{o}_j\right\rangle\\
&\le\left\langle-\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}^{(t)}_{O(i,\cdot)}},\boldsymbol{o}_j\right\rangle+\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right)+\mathcal{O}(\tau).
\end{aligned}\tag{239}$$

Therefore,

$$\left\langle-\frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{W}^{(t)}_{O(i,\cdot)}},\boldsymbol{o}_j\right\rangle\le\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right)\ \text{for } j\neq 1,2.\tag{240}$$

This concludes the proof of (110) in Lemma C.1.

□

E.2 Proof of Lemma C.2

Proof. By definition, for any unlucky neuron $i\in\mathcal{K}_+\setminus\mathcal{W}(0)$, we have

$$\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{o}_+\le 0.\tag{241}$$

We first consider the alignment with $\boldsymbol{o}_+$. That is,

$$\left\langle-\frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{W}^{(t)}_{O(i,\cdot)}},\boldsymbol{o}_+\right\rangle.\tag{242}$$

The gradient is given in (188). We only need to consider the cases where $\left\langle\boldsymbol{y}_l^{(n)},\boldsymbol{o}_+\right\rangle>0$. However, since $\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{o}_+\le 0$, we have

$$\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)=0.\tag{243}$$

$$\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle = \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle$$
$$- \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle$$
$$= 0 \pm \mathcal{O}(\tau). \tag{244}$$

We know by (227),

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \tag{245}$$

Hence,

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) + \mathcal{O}(\tau). \tag{246}$$

We now analyze the alignment with $\boldsymbol{o}_-$. To obtain the bound on $\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle$, we consider the expectation $\mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right]$.

If $\boldsymbol{W}_{O(i,\cdot)}^{(t)} \boldsymbol{o}_- > 0$, the inner product satisfies

$$\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle$$
$$= \frac{1}{\sqrt{m}L} \cdot \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{o}_-) \Big[2 + \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{o}_-)\right) \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{o}_+)\right)$$
$$\cdot \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{o}_j)\right)^{\Delta L_{\boldsymbol{o}_-}^- - 2} \Big] \pm \mathcal{O}(\tau). \tag{247}$$

If $\boldsymbol{W}_{O(i,\cdot)}^{(t)} \boldsymbol{o}_- \leq 0$, then

$$\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle = 0 \pm \mathcal{O}(\tau). \tag{248}$$

From (189), We know that

$$\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle = \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle$$
$$- \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle. \tag{249}$$

Hence, combining both cases, we conclude

$$- \frac{1}{\sqrt{m}L} \cdot \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{o}_-) \Big[2 + \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{o}_-)\right) \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{o}_+)\right)$$
$$\cdot \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{o}_j)\right)^{\Delta L_{\boldsymbol{o}_-}^- - 2} \Big] - \mathcal{O}(\tau).$$
$$\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}(\tau). \tag{250}$$

From (225), similar to (227), we can write

$$\begin{aligned}
&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\
&\quad \le \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \\
&\quad \le \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).
\end{aligned} \tag{251}$$

Hence,

$$\begin{aligned}
&-\frac{1}{\sqrt{m}L}\cdot \sigma(\boldsymbol{w}_\Delta^{(t)\top}\boldsymbol{o}_-)\left[2+\left(1-\sigma(\boldsymbol{w}_\Delta^{(t)\top}\boldsymbol{o}_-)\right)\left(1-\sigma(\boldsymbol{w}_\Delta^{(t)\top}\boldsymbol{o}_+)\right)\right. \\
&\qquad \left.\cdot\left(1-\sigma(\boldsymbol{w}_\Delta^{(t)\top}\boldsymbol{o}_j)\right)^{\Delta L_{\boldsymbol{o}_-}^- - 2}\right] - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) - \mathcal{O}(\tau) \\
&\quad \le \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \le \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) + \mathcal{O}(\tau).
\end{aligned} \tag{252}$$

Now consider $\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle$ for $j \neq 1, 2$.

$$\begin{aligned}
\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i\cdot\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot\boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_j \right\rangle \\
&\qquad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i\cdot\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot\boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_j \right\rangle \\
&:= \langle I_1, \boldsymbol{o}_j\rangle - \langle I_2, \boldsymbol{o}_j\rangle.
\end{aligned} \tag{253}$$

Because $\boldsymbol{o}_j$ for $j \neq 1, 2$ is identical in both I_1 and I_2, $\langle I_1, \boldsymbol{o}_j\rangle - \langle I_2, \boldsymbol{o}_j\rangle = 0 \pm \mathcal{O}(\tau)$. Hence, $\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle = 0 \pm \mathcal{O}(\tau)$. From (225), similar to (227), we can write

$$\begin{aligned}
&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\
&\quad \le \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \\
&\quad \le \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) + \mathcal{O}(\tau).
\end{aligned} \tag{254}$$

Therefore,

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \le \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \text{ for } j \neq 1, 2. \tag{255}$$

□

E.3 PROOF OF LEMMA C.3

By symmetry, the proof is analogous to that of Lemma C.1; Please see Appendix E.1.

E.4 Proof of Lemma C.4

By symmetry, the proof is analogous to that of Lemma C.2; Please see Appendix E.2.

E.5 Proof of Lemma C.5

Proof. The gradient of the loss with respect to $\boldsymbol{w}_\Delta$ for the n^{th} sample is given by

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{w}_\Delta} = &-\frac{z^{(n)}}{L}\cdot\sum_{i=1}^{m}\sum_{l=1}^{L} v_i\,\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot\sum_{s=1}^{l}\left(\boldsymbol{W}_B^\top\boldsymbol{x}_s^{(n)}\right)^\top\left(\boldsymbol{W}_C^\top\boldsymbol{x}_l^{(n)}\right)\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{x}_s^{(n)}\right)\\ &\cdot\sigma\left(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_s^{(n)}\right)\cdot\prod_{r=s+1}^{l}\left(1-\sigma\left(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_r^{(n)}\right)\right)\\ &\cdot\left[\left(1-\sigma\left(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_s^{(n)}\right)\right)\boldsymbol{x}_s^{(n)}-\sum_{j=s+1}^{l}\left(1-\sigma\left(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_j^{(n)}\right)\right)\boldsymbol{x}_j^{(n)}\right]\\ :=&-\frac{z^{(n)}}{L}\cdot\sum_{i=1}^{m}\sum_{l=1}^{L} v_i\,\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot\sum_{s=1}^{l}\boldsymbol{I}_{l,s}^{(n)}.\end{aligned}\tag{256}$$

We define the gradient summand $\boldsymbol{I}_{l,s}^{(n)}$ as

$$\boldsymbol{I}_{l,s}^{(n)} = \beta_{s,s}\cdot\boldsymbol{x}_s^{(n)} - \sum_{j=s+1}^{l}\beta_{s,j}\boldsymbol{x}_j^{(n)},\tag{257}$$

where the coefficients $\beta_{s,s}$ and $\beta_{s,j}$ are given by

$$\begin{aligned}\beta_{s,s} = &(\boldsymbol{W}_B^\top\boldsymbol{x}_s^{(n)})^\top(\boldsymbol{W}_C^\top\boldsymbol{x}_l^{(n)})(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{x}_s^{(n)})\sigma(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_s^{(n)})\\ &\times\left[\prod_{r=s+1}^{l}\left(1-\sigma(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_r^{(n)})\right)\right](1-\sigma(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_s^{(n)})).\end{aligned}\tag{258}$$

and

$$\begin{aligned}\beta_{s,j} = &(\boldsymbol{W}_B^\top\boldsymbol{x}_s^{(n)})^\top(\boldsymbol{W}_C^\top\boldsymbol{x}_l^{(n)})(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{x}_s^{(n)})\sigma(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_s^{(n)})\\ &\times\left[\prod_{r=s+1}^{l}\left(1-\sigma(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_r^{(n)})\right)\right](1-\sigma(\boldsymbol{w}_\Delta^\top\boldsymbol{x}_j^{(n)})).\end{aligned}\tag{259}$$

If we consider the gradient of the empirical loss,

$$\frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{w}_\Delta} = -\frac{1}{N}\sum_{n=1}^{N}\frac{z^{(n)}}{L}\cdot\sum_{i=1}^{m}\sum_{l=1}^{L}v_i\,\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot\sum_{s=1}^{l}\boldsymbol{I}_{l,s}^{(n)}.\tag{260}$$

We are given that

$$p_1 \le \langle\boldsymbol{w}_\Delta^{(t)},\boldsymbol{o}_+\rangle \le q_1,\quad\text{and}\quad r_1^* \le \langle{\boldsymbol{W}_{O(i,\cdot)}^{(t+1)}}^\top,\boldsymbol{o}_+\rangle \le s_1^*.\tag{261}$$

From our initialization, for all $i\in\mathcal{K}^+$, we have $v_i = \frac{1}{\sqrt{m}}$. This gives

$$\left\langle -\frac{\partial\ell}{\partial\boldsymbol{w}_\Delta},\boldsymbol{o}_+\right\rangle = \frac{z^{(n)}}{L}\sum_{i=1}^{m}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\cdot\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle\boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_+\right\rangle.\tag{262}$$

Averaging over the training samples, the inner product of the empirical gradient becomes

$$\begin{aligned}\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta}, \boldsymbol{o}_+ \right\rangle &= \frac{1}{N}\sum_{n=1}^{N} \frac{z^{(n)}}{L} \cdot \sum_{i=1}^{m}\sum_{l=1}^{L} v_i \phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}) \cdot \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \\ &= \frac{1}{N}\sum_{n:z^{(n)}=+1} \frac{1}{L}\left[\sum_{i\in\mathcal{K}_+}\sum_{l=1}^{L} \frac{1}{\sqrt{m}} \phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \right. \\ &\quad \left. + \sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L} \left(-\frac{1}{\sqrt{m}}\right) \phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \right] \\ &\quad + \frac{1}{N}\sum_{n:z^{(n)}=-1} \frac{-1}{L}\left[\sum_{i\in\mathcal{K}_+}\sum_{l=1}^{L} \frac{1}{\sqrt{m}} \phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \right. \\ &\quad \left. + \sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L} \left(-\frac{1}{\sqrt{m}}\right) \phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \right]. \end{aligned} \tag{263}$$

First, we focus on the contribution from the samples where $z^{(n)} = +1$, for which we seek a lower bound. We analyze the inner terms by considering four cases.

Case I: $l = L_1^+,\ s = L_1^+$

Since $l = s$ and $\boldsymbol{x}_s = \boldsymbol{o}_+$, it follows from (257) that

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = \beta_{s,s}. \tag{264}$$

Using (258), with $\boldsymbol{W}_B = \boldsymbol{W}_C = I$ and $\boldsymbol{x}_l = \boldsymbol{x}_s = \boldsymbol{o}_+$, we obtain

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = \beta_{s,s} = \langle \boldsymbol{W}_{O(i,\cdot)}^{(t+1)\top}, \boldsymbol{o}_+ \rangle \cdot \sigma(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+ \rangle) \cdot \left(1 - \sigma(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+ \rangle)\right). \tag{265}$$

Given the conditions in (261), we can write

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \geq (r_1^* - \mathcal{O}(\tau)) \cdot \sigma(p_1 - \mathcal{O}(\tau)) \cdot (1 - \sigma(q_1 + \mathcal{O}(\tau))). \tag{266}$$

We can approximate $\sigma(p_1 - \mathcal{O}(\tau)) \approx \sigma(p_1) - \mathcal{O}(\tau)$ and $1 - \sigma(q_1 + \mathcal{O}(\tau)) \approx 1 - \sigma(q_1) - \mathcal{O}(\tau)$, since $\mathcal{O}(\tau) < \mathcal{O}(\frac{1}{d})$.

Therefore, we obtain

$$\begin{aligned}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle &\geq (r_1^* - \mathcal{O}(\tau)) \cdot (\sigma(p_1) - \mathcal{O}(\tau)) \cdot (1 - \sigma(q_1) - \mathcal{O}(\tau)) \\ &\geq r_1^* \cdot \sigma(p_1) \cdot (1 - \sigma(q_1)) - \mathcal{O}(\tau).\end{aligned} \tag{267}$$

Case II: $l = L_2^+,\ s = L_2^+$

This configuration yields the same result as in Case I. We again obtain

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \geq r_1^* \cdot \sigma(p_1) \cdot (1 - \sigma(q_1)) - \mathcal{O}(\tau). \tag{268}$$

Case III: $l = L_2^+,\ s = L_1^+$ Comparing (258) with (259), we see that the two expressions differ only in their last term. In this setting, $\boldsymbol{x}_j$ equals $\boldsymbol{o}_+$ only when $j = L_2^+$. Consequently, $\boldsymbol{x}_s = \boldsymbol{x}_j = \boldsymbol{o}_+$, which implies $\beta_{s,s} = \beta_{s,j}$. Hence,

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = \beta_{s,s} - \beta_{s,j} = 0 \pm \mathcal{O}(\tau). \tag{269}$$

Case IV: Others

For the other token positions, $\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = 0$ due to orthogonality among the features.

Combining the above, the total contribution becomes

$$\sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \geq 2r_1^* \cdot \sigma(p_1) \cdot (1 - \sigma(q_1)) - \mathcal{O}(\tau). \tag{270}$$

We now bound the entire sum over all tokens:

$$\frac{1}{L} \sum_{l=1}^{L} \frac{1}{\sqrt{m}} \phi' \left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l \right) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \geq \frac{1}{L} \sum_{l=1}^{L} \frac{1}{\sqrt{m}} \cdot 1 \cdot 2r_1^* \cdot \sigma(p_1) \cdot (1 - \sigma(q_1)) - \mathcal{O}(\tau). \tag{271}$$

Let $\rho_t^+ = |\mathcal{W}(t)|$ be the number of contributing neurons. Then the total contribution from the active neurons is lower bounded as

$$\frac{1}{L} \sum_{i \in \mathcal{K}_+} \sum_{l=1}^{L} v_i \phi' \left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l \right) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \geq \frac{2r_1^* \cdot \sigma(p_1) \cdot (1 - \sigma(q_1))}{\sqrt{m}} \cdot \rho_t^+ - \mathcal{O}(\tau). \tag{272}$$

Next, we consider $z^{(n)} = -1$ for $i \in \mathcal{K}_+$. For $z^{(n)} = -1$, the negative sample also contains two $\boldsymbol{o}_+$ features.

Similar to the above, we have to consider 4 cases.

Case I: $l = L_1^+,\ s = L_1^+$

Since $l = s$, it follows from (257) that

$$\boldsymbol{I}_{l,s}^{(n)} = \beta_{s,s} \cdot \boldsymbol{x}_l. \tag{273}$$

Since $\boldsymbol{x}_l = \boldsymbol{o}_+$, we have

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = \beta_{s,s}. \tag{274}$$

We now seek an upper bound for this contribution. From the initial conditions in (261), we know

$$\langle {\boldsymbol{W}_{O(i,\cdot)}^{(t+1)}}^{\top}, \boldsymbol{o}_+ \rangle \leq s_1^* + \mathcal{O}(\tau). \tag{275}$$

Hence, we obtain

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \leq (s_1^* + \mathcal{O}(\tau)) \cdot \sigma(q_1 + \mathcal{O}(\tau)) \cdot (1 - \sigma(p_1 - \mathcal{O}(\tau))). \tag{276}$$

We can approximate $\sigma(q_1 + \mathcal{O}(\tau)) \approx \sigma(q_1) + \mathcal{O}(\tau)$ and $1 - \sigma(p_1 - \mathcal{O}(\tau)) \approx 1 - \sigma(p_1) + \mathcal{O}(\tau)$, since $\mathcal{O}(\tau) < \mathcal{O}(\frac{1}{d})$.

Therefore, we obtain

$$\begin{aligned} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle &\leq (s_1^* + \mathcal{O}(\tau)) \cdot (\sigma(q_1) + \mathcal{O}(\tau)) \cdot (1 - \sigma(p_1) + \mathcal{O}(\tau)) \\ &\leq s_1^* \cdot \sigma(q_1) \cdot (1 - \sigma(p_1)) + \mathcal{O}(\tau). \end{aligned} \tag{277}$$

Case II: $l = L_2^+,\ s = L_2^+$

This configuration yields the same result as in Case I. We again obtain

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \leq s_1^* \cdot \sigma(q_1) \cdot (1 - \sigma(p_1)) + \mathcal{O}(\tau). \tag{278}$$

Case III: $l = L_2^+$, $s = L_1^+$

In this case, the contribution vanishes:

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = 0 \pm \mathcal{O}(\tau). \tag{279}$$

Case IV: Others

For the other token positions, $\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = 0$ due to orthogonality among the features.

The maximum number of such contributing neurons is $\frac{m}{2}$. Therefore, the total contribution is bounded above by

$$\begin{aligned} \frac{1}{L}\sum_{i\in\mathcal{K}_+}\sum_{l=1}^{L} v_i\,\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l\right)\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \leq & \frac{2s_1^* \cdot \sigma(q_1)\cdot(1-\sigma(p_1))}{\sqrt{m}} \cdot \frac{m}{2} + \mathcal{O}(\tau) \\ = & \sqrt{m}\cdot s_1^* \cdot \sigma(q_1)\cdot(1-\sigma(p_1)) + \mathcal{O}(\tau). \end{aligned} \tag{280}$$

Thirdly, let us consider the contribution for $z^{(n)} = +1$ from $i \in \mathcal{K}_-$. From our initialization, for $i \in \mathcal{K}^-$, $v_i = -\frac{1}{\sqrt{m}}$. For $z^{(n)} = +1$, we seek an upper bound on the contribution from such neurons.

Let $z^{(n)} = +1$. To maximize the term $\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{x}_s^{(n)}$ in (258), we can consider the token locations which contain $\boldsymbol{o}_-$ features since $\boldsymbol{W}_{O(i,\cdot)}$ has a large component in the $\boldsymbol{o}_-$ direction. Then $\boldsymbol{x}_l = \boldsymbol{o}_- \Rightarrow \boldsymbol{y}_l$ contains the $\boldsymbol{o}_-$ feature.

However, in this case, $\boldsymbol{x}_s = \boldsymbol{o}_- = \boldsymbol{x}_l$, and due to orthogonality,

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = 0. \tag{281}$$

Hence, we only need to consider time steps $l = L_1^+, L_2^+$, where $\boldsymbol{o}_+$ features appear.

Recall that

$$\left\langle -\frac{\partial \ell}{\partial \boldsymbol{w}_\Delta}, \boldsymbol{o}_+ \right\rangle = \frac{1}{L}\sum_{i=1}^{m}\sum_{l=1}^{L} -\frac{1}{\sqrt{m}}\cdot\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l)\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle. \tag{282}$$

We analyze the inner contributions case by case.

Case I: $l = L_1^+$, $s = L_1^+$

Given that

$$\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{o}_+ \leq \delta_1 + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) =: c, \tag{283}$$

we obtain

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \leq c\cdot\sigma(q_1)\cdot(1-\sigma(p_1)) + \mathcal{O}(\tau). \tag{284}$$

Case II: $l = L_2^+$, $s = L_2^+$

This configuration yields the same bound:

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \leq c\cdot\sigma(q_1)\cdot(1-\sigma(p_1)) + \mathcal{O}(\tau). \tag{285}$$

Case III: $l = L_2^+$, $s = L_1^+$

In this case, the contribution vanishes:

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = 0 \pm \mathcal{O}(\tau). \tag{286}$$

Case IV: Others

For the other token positions, $\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = 0$ due to orthogonality among the features.

Thus, the total contribution from each $i \in \mathcal{K}^-$ satisfies

$$\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \leq 2c\cdot\sigma(q_1)\cdot(1-\sigma(p_1)) + \mathcal{O}(\tau). \tag{287}$$

The maximum number of such contributing neurons is $\frac{m}{2}$, so the full contribution is bounded by

$$\begin{aligned}\frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L}\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l\right)\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_+\right\rangle &\le \frac{2c\cdot\sigma(q_1)\cdot(1-\sigma(p_1))}{\sqrt{m}}\cdot\frac{m}{2}+\mathcal{O}(\tau)\\ &= \sqrt{m}c\cdot\sigma(q_1)\cdot(1-\sigma(p_1))+\mathcal{O}(\tau).\end{aligned} \tag{288}$$

Therefore, the overall contribution is

$$-\frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L}\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l)\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_+\right\rangle \ge -\sqrt{m}c\cdot\sigma(q_1)\cdot(1-\sigma(p_1))-\mathcal{O}(\tau). \tag{289}$$

Finally, we consider $z^{(n)}=-1$ for $i\in\mathcal{K}_-$. For $z^{(n)}=-1$, we want a lower bound since $v_i=-\frac{1}{\sqrt{m}}$.

We could consider $l=L^+\Rightarrow \boldsymbol{x}_l=\boldsymbol{o}_+$, and write

$$\left\langle \boldsymbol{W}_{O(i,\cdot)},\boldsymbol{o}_+\right\rangle \ge \delta_1-\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{290}$$

However, the minimum number of such contributing neurons is not tractable. Thus, if we consider the worst case where $\boldsymbol{W}_{O(i,\cdot)}$ for $i\in\mathcal{K}^-$ does not learn the $\boldsymbol{o}_+$ feature, the obvious lower bound is zero:

$$\frac{1}{L}\sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l\right)\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_+\right\rangle \ge 0. \tag{291}$$

We now combine the bounds for the four terms identified in (263), corresponding to the contributions from: (i) $\mathcal{K}_+$ with $z^{(n)}=+1$ as shown in (272), (ii) $\mathcal{K}_+$ with $z^{(n)}=-1$ as shown in (280), (iii) $\mathcal{K}_-$ with $z^{(n)}=+1$ as shown in (289), and (iv) $\mathcal{K}_-$ with $z^{(n)}=-1$ as shown in (291). We assume the batch is balanced, so the number of positive and negative samples is equal, with each class contributing $\frac{N}{2}$ samples. Then we have

$$\begin{aligned}\left\langle -\frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{w}_\Delta},\boldsymbol{o}_+\right\rangle &\ge \frac{1}{2}\Bigg[\frac{2r_1^*\cdot\sigma(p_1)\left(1-\sigma(q_1)\right)}{\sqrt{m}}\cdot\rho_t^+-\sqrt{m}\cdot c\cdot\sigma(q_1)\left(1-\sigma(p_1)\right)\\ &\qquad -\sqrt{m}\cdot s_1^*\cdot\sigma(q_1)\cdot(1-\sigma(p_1))+0\Bigg]-\mathcal{O}(\tau)\\ &=\frac{\sigma(p_1)\left(1-\sigma(q_1)\right)r_1^*\cdot\rho_t^+}{\sqrt{m}}-\frac{\sigma(q_1)\left(1-\sigma(p_1)\right)s_1^*\cdot\sqrt{m}}{2} && (292)\\ &\quad-\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right)-\mathcal{O}(\tau). && (293)\end{aligned}$$

where we have used the fact $\frac{\sqrt{m}}{2}\cdot\sigma(q_1)\left(1-\sigma(p_1)\right)\cdot c=\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right)$ since $c=\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right)$.

□

E.6 Proof of Lemma C.6

Proof. The gradient is given in (256).

Let's consider the alignment with $\boldsymbol{o}_k$ for $k\neq 1,2$.

$$\left\langle -\frac{\partial\ell}{\partial\boldsymbol{w}_\Delta},\boldsymbol{o}_k\right\rangle=\frac{z^{(n)}}{L}\sum_{i=1}^{m}\sum_{l=1}^{L}v_i\cdot\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\right\rangle \tag{294}$$

From our initialization, for all $i \in \mathcal{K}^+$, we have $v_i = \frac{1}{\sqrt{m}}$.

We first consider the case $z^{(n)} = +1$ for $i \in \mathcal{K}^+$. Since $\boldsymbol{W}_{O(i,\cdot)}$, for $i \in \mathcal{K}^+$ has a large $\boldsymbol{o}_+$ component, we have to consider the token features with $\boldsymbol{o}_+$. For $z^{(n)} = +1$, only when $l = L_2^+,\ s = L_1^+$ we have $\boldsymbol{x}_l = \boldsymbol{x}_s = \boldsymbol{o}_+$. Therefore, $\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{x}_s$ is significant. Hence, we have

$$\begin{aligned}
&\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k \right\rangle\\
=& - \sum_{j=s+1}^{l} \beta_{s,j}\langle \boldsymbol{x}_j^{(n)}, \boldsymbol{o}_k\rangle\\
\le& -\beta_{s,s+1} \qquad\qquad (\text{Assuming W.L.O.G. } \boldsymbol{x}_{s+1}^{(n)} = \boldsymbol{o}_k)\\
\le& -\langle \boldsymbol{W}_{O(i,\cdot)}^{(t+1)\top}, \boldsymbol{o}_+\rangle \cdot \sigma(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle) \cdot \left(1 - \sigma(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle)\right) \cdot \left(1 - \sigma(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_k\rangle)\right)^{\Delta L^+_{\boldsymbol{o}_+}}.
\end{aligned} \tag{295}$$

Using the the conditions in (261), we can write

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k \right\rangle \le (-r_1^* + \mathcal{O}(\tau)) \cdot \sigma(p_1 + \mathcal{O}(\tau)) \cdot (1 - \sigma(q_1 - \mathcal{O}(\tau))) \cdot (1 - \sigma(q_2 - \mathcal{O}(\tau)))^{\Delta L^+_{\boldsymbol{o}_+}}. \tag{296}$$

We can approximate $\sigma(p_1 + \mathcal{O}(\tau)) \approx \sigma(p_1) + \mathcal{O}(\tau)$, $1 - \sigma(q_1 - \mathcal{O}(\tau)) \approx 1 - \sigma(q_1) + \mathcal{O}(\tau)$ and $1 - \sigma(q_2 - \mathcal{O}(\tau)) \approx 1 - \sigma(q_2) + \mathcal{O}(\tau)$, since $\mathcal{O}(\tau) < \mathcal{O}(\frac{1}{d})$.

Hence, we obtain

$$\begin{aligned}
&\frac{1}{L}\sum_{l=1}^{L} \frac{1}{\sqrt{m}} \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k \right\rangle\\
\le& \frac{1}{L}\sum_{l=1}^{L} \frac{1}{\sqrt{m}} \cdot 1 \cdot \left[-r_1^* \cdot \sigma(p_1) \cdot (1 - \sigma(q_1)) \cdot (1 - \sigma(q_2))^{\Delta L^+_{\boldsymbol{o}_+}}\right] + \mathcal{O}(\tau)).
\end{aligned} \tag{297}$$

Let $\rho_t^+ = |\mathcal{W}(t)|$ be the number of contributing neurons. Then the total contribution from $\mathcal{K}_+$ neurons is bounded as

$$\begin{aligned}
\frac{1}{L}\sum_{i \in \mathcal{K}_+}\sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k \right\rangle \le& -\frac{r_1^*}{\sqrt{m}} \cdot \sigma(p_1)\left(1 - \sigma(q_1)\right)\\
&\cdot (1 - \sigma(q_2))^{\Delta L^+_{\boldsymbol{o}_+}} \cdot \rho_t^+ + \mathcal{O}(\tau)).
\end{aligned} \tag{298}$$

Next, we consider $z^{(n)} = -1$ for $i \in \mathcal{K}_+$. Since $\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k \right\rangle < 0$, we require a lower bound for this.

$$\begin{aligned}
&\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k \right\rangle\\
=& - \sum_{j=s+1}^{l} \beta_{s,j}\langle \boldsymbol{x}_j^{(n)}, \boldsymbol{o}_k\rangle\\
\gtrsim& -\langle \boldsymbol{W}_{O(i,\cdot)}^{(t+1)\top}, \boldsymbol{o}_+\rangle \cdot \sigma(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle) \cdot \left(1 - \sigma(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle)\right) \cdot \left(1 - \sigma(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_k\rangle)\right)^{\Delta L^-_{\boldsymbol{o}_+}}.
\end{aligned} \tag{299}$$

Using the the conditions in (261), we can write

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k \right\rangle \gtrsim -s_1^* \cdot \sigma(q_1) \cdot (1 - \sigma(p_1)) \cdot (1 - \sigma(p_2))^{\Delta L^-_{\boldsymbol{o}_+}}. \tag{300}$$

$$\begin{aligned}
&\frac{1}{L}\sum_{i \in \mathcal{K}_+}\sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k \right\rangle\\
\gtrsim& -\frac{s_1^*}{\sqrt{m}} \cdot \sigma(q_1)\left(1 - \sigma(p_1)\right)\left(1 - \sigma(p_2)\right)^{\Delta L^-_{\boldsymbol{o}_+}} \cdot \rho_t^+.
\end{aligned} \tag{301}$$

Since $\Delta L^-_{\boldsymbol{o}_+} \gg \Delta L^+_{\boldsymbol{o}_+}$, this term is negligible which leads to

$$\frac{1}{L}\sum_{i\in\mathcal{K}_+}\sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k\right\rangle \geq -\mathcal{O}\left((1-\sigma(p_2))^{\Delta L^-_{\boldsymbol{o}_+}}\right) \approx 0. \tag{302}$$

Thirdly, we consider the case $i \in \mathcal{K}^-$, for $z^{(n)} = -1$. Similar to (295) and(296), when $l = L_2^-,\ s = L_1^-$ the contribution is significant.

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k\right\rangle \leq -r_1^* \cdot \sigma(p_1)\cdot(1-\sigma(q_1))\cdot(1-\sigma(q_2))^{\Delta L^-_{\boldsymbol{o}_-}} + \mathcal{O}(\tau)). \tag{303}$$

Hence, we obtain

$$\begin{aligned}&\frac{1}{L}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\cdot\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k\right\rangle\\ \leq&\frac{1}{L}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\cdot 1\cdot\left[-r_1^*\cdot\sigma(p_1)\cdot(1-\sigma(q_1))\cdot(1-\sigma(q_2))^{\Delta L^-_{\boldsymbol{o}_-}}\right]+\mathcal{O}(\tau))\end{aligned} \tag{304}$$

Let $\rho_t^- = |\mathcal{U}(t)|$ be the number of contributing neurons. Then the total contribution from $\mathcal{K}_-$ neurons is bounded as

$$\begin{aligned}\frac{1}{L}\sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L} v_i\,\phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big)\sum_{s=1}^{l}\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k\rangle \leq -\frac{r_1^*}{\sqrt{m}}\,\sigma(p_1)\,\big(1-\sigma(q_1)\big)\,\big(1-\sigma(q_2)\big)^{\Delta L^-_{\boldsymbol{o}_-}}\,\rho_t^-\\ +\mathcal{O}(\tau)).\end{aligned} \tag{305}$$

Finally, we consider $i \in \mathcal{K}_-$ for $z^{(n)} = +1$. Following the same approach as in (299) to (301), we can write

$$\begin{aligned}&\frac{1}{L}\sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k\right\rangle\\ \gtrsim&-\frac{s_1^*}{\sqrt{m}}\cdot\sigma(q_1)\,(1-\sigma(p_1))\,(1-\sigma(p_2))^{\Delta L^+_{\boldsymbol{o}_-}}\cdot\rho_t^-.\end{aligned} \tag{306}$$

Since $\Delta L^+_{\boldsymbol{o}_-} \gg \Delta L^-_{\boldsymbol{o}_-}$, this term is negligible which leads to

$$\frac{1}{L}\sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k\right\rangle \geq -\mathcal{O}\left((1-\sigma(p_2))^{\Delta L^+_{\boldsymbol{o}_-}}\right) \approx 0. \tag{307}$$

Putting it together, We know

$$\begin{aligned}\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta}, \boldsymbol{o}_k \right\rangle &= \frac{1}{N}\sum_{n=1}^{N}\frac{z^{(n)}}{L}\cdot\sum_{i=1}^{m}\sum_{l=1}^{L} v_i \phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\cdot\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\right\rangle \\ &= \frac{1}{N}\sum_{n:z^{(n)}=+1}\frac{1}{L}\Bigg[\sum_{i\in\mathcal{K}_+}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\right\rangle \\ &\quad + \sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L}\left(-\frac{1}{\sqrt{m}}\right)\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\right\rangle\Bigg] \\ &\quad + \frac{1}{N}\sum_{n:z^{(n)}=-1}\frac{-1}{L}\Bigg[\sum_{i\in\mathcal{K}_+}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\right\rangle \\ &\quad + \sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L}\left(-\frac{1}{\sqrt{m}}\right)\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\right\rangle\Bigg]. \qquad (308)\end{aligned}$$

We now combine the bounds for the two terms identified in equation (308), corresponding to the contributions from: (i) $\mathcal{K}_+$ with $z^{(n)} = +1$ (298), (ii) $\mathcal{K}_+$ with $z^{(n)} = -1$ (306), (iii) $\mathcal{K}_-$ with $z^{(n)} = +1$ (301), and (iv) $\mathcal{K}_-$ with $z^{(n)} = -1$ (305). We assume the batch is balanced, so the number of positive and negative samples is equal, with each class contributing $\frac{N}{2}$ samples. Then we have

$$\begin{aligned}&\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_k \right\rangle \\ &\leq -\frac{r_1^*}{2\sqrt{m}}\sigma(p_1)\left(1-\sigma(q_1)\right)\left[\left(1-\sigma(q_2)\right)^{\Delta L_{\boldsymbol{o}_+}^+}\rho_t^+ + \left(1-\sigma(q_2)\right)^{\Delta L_{\boldsymbol{o}_-}^-}\rho_t^-\right] \\ &\quad + \frac{s_1^*}{\sqrt{m}}\cdot\sigma(q_1)\left(1-\sigma(p_1)\right)\left[\mathcal{O}\left(\left(1-\sigma(p_2)\right)^{\Delta L_{\boldsymbol{o}_-}^+}\right)\cdot\rho_t^- + \mathcal{O}\left(\left(1-\sigma(p_2)\right)^{\Delta L_{\boldsymbol{o}_+}^-}\right)\cdot\rho_t^+\right] \qquad (309) \\ &\quad + \mathcal{O}(\tau)) \qquad (310)\end{aligned}$$

$$\begin{aligned}&\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_k \right\rangle \\ &\leq -\frac{r_1^*}{2\sqrt{m}}\sigma(p_1)\left(1-\sigma(q_1)\right)\left[\left(1-\sigma(q_2)\right)^{\Delta L_{\boldsymbol{o}_+}^+}\rho_t^+ + \left(1-\sigma(q_2)\right)^{\Delta L_{\boldsymbol{o}_-}^-}\rho_t^-\right] \\ &\quad + \mathcal{O}\left(\left(1-\sigma(p_2)\right)^{\Delta L_{\boldsymbol{o}_+}^-}\right) + \mathcal{O}\left(\left(1-\sigma(p_2)\right)^{\Delta L_{\boldsymbol{o}_-}^+}\right) + \mathcal{O}(\tau)) \qquad (311)\end{aligned}$$

From (302) and (307), we can conclude

$$\begin{aligned}\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_k \right\rangle \leq& -\frac{r_1^*}{2\sqrt{m}}\sigma(p_1)\left(1-\sigma(q_1)\right)\left[\left(1-\sigma(q_2)\right)^{\Delta L_{\boldsymbol{o}_+}^+}\rho_t^+ + \left(1-\sigma(q_2)\right)^{\Delta L_{\boldsymbol{o}_-}^-}\rho_t^-\right] \\ &+ \mathcal{O}(\tau)). \qquad (312)\end{aligned}$$

□

F Extension to Multi-Class Classification

Consider the classification problem with four classes, where each example is assigned a label $\boldsymbol{z} = (z_1, z_2) \in \{+1, -1\}^2$ representing four distinct classes. Similarly to the binary setting, there exist four orthogonal discriminative patterns. In the output layer, the scalar coefficient v_i associated with hidden neuron i is replaced by a two-dimensional vector $\boldsymbol{v}_i \in \mathbb{R}^2$.

Hence, we define the model output as

$$\boldsymbol{F}(\boldsymbol{X}) = \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m} \boldsymbol{v}_i\, \phi\big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l(\boldsymbol{X})\big)\,. \tag{313}$$

$$F_1(\boldsymbol{X}^{(n)}) = \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m}(\boldsymbol{v}_i)_1\, \phi\big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l(\boldsymbol{X}^{(n)})\big), \tag{314}$$

$$F_2(\boldsymbol{X}^{(n)}) = \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m}(\boldsymbol{v}_i)_2\, \phi\big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l(\boldsymbol{X}^{(n)})\big). \tag{315}$$

The dataset can be divided into four groups as

$$\begin{aligned}
\mathcal{D}_1 &= \{(\boldsymbol{X}^{(n)}, \boldsymbol{z}^{(n)}) \mid \boldsymbol{z}^{(n)} = (1,1)\},\\
\mathcal{D}_2 &= \{(\boldsymbol{X}^{(n)}, \boldsymbol{z}^{(n)}) \mid \boldsymbol{z}^{(n)} = (1,-1)\},\\
\mathcal{D}_3 &= \{(\boldsymbol{X}^{(n)}, \boldsymbol{z}^{(n)}) \mid \boldsymbol{z}^{(n)} = (-1,1)\},\\
\mathcal{D}_4 &= \{(\boldsymbol{X}^{(n)}, \boldsymbol{z}^{(n)}) \mid \boldsymbol{z}^{(n)} = (-1,-1)\}.
\end{aligned} \tag{316}$$

The loss function for data $(\boldsymbol{X}^{(n)}, \boldsymbol{z}^{(n)})$ is

$$\mathrm{Loss}(\boldsymbol{X}^{(n)}, \boldsymbol{z}^{(n)}) = \max\Big\{1 - \boldsymbol{z}^{(n)\top}\boldsymbol{F}(\boldsymbol{X}^{(n)}),\, 0\Big\}\,. \tag{317}$$

Since $\boldsymbol{v}_i \in \{\pm\frac{1}{\sqrt{m}}\}^2$, we divide neurons into four groups:

$$\begin{aligned}
\mathcal{W}_1 &= \big\{i : \boldsymbol{v}_i = \tfrac{1}{\sqrt{m}}(1,1)\big\},\\
\mathcal{W}_2 &= \big\{i : \boldsymbol{v}_i = \tfrac{1}{\sqrt{m}}(1,-1)\big\},\\
\mathcal{W}_3 &= \big\{i : \boldsymbol{v}_i = \tfrac{1}{\sqrt{m}}(-1,1)\big\},\\
\mathcal{W}_4 &= \big\{i : \boldsymbol{v}_i = \tfrac{1}{\sqrt{m}}(-1,-1)\big\}.
\end{aligned} \tag{318}$$

For neuron i, the gradient decomposes as

$$\frac{\partial \mathrm{Loss}}{\partial \boldsymbol{W}_{O(i,\cdot)}} = -z_1^{(n)}\,\frac{\partial F_1(\boldsymbol{X}^{(n)})}{\partial \boldsymbol{W}_{O(i,\cdot)}} - z_2^{(n)}\,\frac{\partial F_2(\boldsymbol{X}^{(n)})}{\partial \boldsymbol{W}_{O(i,\cdot)}}. \tag{319}$$

Let $\boldsymbol{o}_1, \boldsymbol{o}_2, \boldsymbol{o}_3, \boldsymbol{o}_4$ denote the four discriminative directions. Consider $i \in \mathcal{W}_2$, i.e. $\boldsymbol{v}_i = \frac{1}{\sqrt{m}}(1,-1)$. Projecting the gradient onto $\boldsymbol{o}_2$, for any $(\boldsymbol{X}^{(n)}, \boldsymbol{z}^{(n)}) \in \mathcal{D}_2$ we obtain

$$-\Big\langle \frac{\partial \mathrm{Loss}}{\partial \boldsymbol{W}_{O(i,\cdot)}}, \boldsymbol{o}_2 \Big\rangle \approx \frac{2}{\sqrt{m}}\|\boldsymbol{o}_2\|^2 \ > \ 0, \tag{320}$$

showing GD moves $\boldsymbol{W}_{O(i,\cdot)}$ toward $\boldsymbol{o}_2$.

For samples from the other classes:

$$\begin{aligned}
(\boldsymbol{X}^{(n)}, \boldsymbol{z}^{(n)}) \in \mathcal{D}_1 :&\quad -\Big\langle \frac{\partial \mathrm{Loss}}{\partial \boldsymbol{W}_{O(i,\cdot)}}, \boldsymbol{o}_1 \Big\rangle \approx 0,\\
(\boldsymbol{X}^{(n)}, \boldsymbol{z}^{(n)}) \in \mathcal{D}_3 :&\quad -\Big\langle \frac{\partial \mathrm{Loss}}{\partial \boldsymbol{W}_{O(i,\cdot)}}, \boldsymbol{o}_3 \Big\rangle \approx -\frac{2}{\sqrt{m}}\|\boldsymbol{o}_3\|^2,\\
(\boldsymbol{X}^{(n)}, \boldsymbol{z}^{(n)}) \in \mathcal{D}_4 :&\quad -\Big\langle \frac{\partial \mathrm{Loss}}{\partial \boldsymbol{W}_{O(i,\cdot)}}, \boldsymbol{o}_4 \Big\rangle \approx 0.
\end{aligned} \tag{321}$$

Thus, for $i \in \mathcal{W}_2$, the update direction aligns with $\boldsymbol{o}_2$, and similarly neurons in $\mathcal{W}_1, \mathcal{W}_3, \mathcal{W}_4$ align with $\boldsymbol{o}_1, \boldsymbol{o}_3, \boldsymbol{o}_4$ respectively. Similarly, we can analyze the gradient dynamics of the gating vector $\boldsymbol{w}_\triangle$.