Theoretical Analysis of the Selection Mechanism in Mamba: Training Dynamics and Generalization

Mugunthan Shandirasegaran
New Jersey Institute of Technology
ms3537@njit.edu

Yating Zhou
Rensselaer Polytechnic Institute
zhouy26@rpi.edu

Songyang Zhang
University of Louisiana at Lafayette
songyang.zhang@louisiana.edu

Shuai Zhang
New Jersey Institute of Technology
sz457@njit.edu

Abstract

While Transformers rely on a distinctive attention mechanism, the recent emergence of Mamba and other selective state space models (SSMs) offers a strong alternative. These models incorporate attention-like mechanisms with hardware-aware efficiency and a unique selection strategy, yet their theoretical properties remain poorly understood. In this work, we present a first-step theoretical analysis of the *selection mechanism in Mamba*. We study a simplified single-layer Mamba block trained with gradient descent on structured data containing both label-relevant and irrelevant tokens. Our results show that the gating vector dynamically aligns with label-relevant features while negating irrelevant ones, formalizing its role as an implicit feature selector. Moreover, we prove that training achieves guaranteed generalization, with explicit bounds on sample size and convergence rate. These findings offer principled insight into when and why Mamba's selection mechanism enables efficient learning, offering a theoretical counterpoint to Transformer-centric explanations of generalization.

1 Introduction

The selective SSM [4, 3] and its variants [30, 24, 2, 12, 25], or commonly referred to as Mamba, have shown strong performance across language, vision, graphs, audio [26], medicine [27, 13, 16, 22], and genomics [15, 29, 27], revitalizing interest in non-attention architectures for sequence modeling. Unlike Transformers [23], which rely on distinctive attention weights to capture token interactions, Mamba employs a *selection mechanism* based on input-dependent gating. This design has demonstrated strong empirical performance across language and vision tasks, while offering hardware efficiency and autoregressive modeling capabilities. Yet, despite this progress, the theoretical principles underlying Mamba's selection mechanism remain largely unexplored. In particular, it is unclear how the gating vector evolves during training, and under what conditions it enables efficient learning and generalization. Prior theoretical work has mostly focused on Transformers, analyzing how attention aligns with label-relevant features [7, 11, 14, 1, 19–21]. Comparable results for selective SSMs [5, 6, 17] are missing, leaving open fundamental questions about their learnability.

Scope of this work. We present a theoretical study of Mamba's selection mechanism in a simplified but representative setting: a single-layer, single-head Mamba block followed by a two-layer MLP trained via gradient descent (GD). Our analysis is conducted on a simplified structured data model with both label-relevant and label-irrelevant features under token-level noise, chosen to enable a

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: What Can('t) Transformers Do?.

tractable theoretical study of the gating mechanism, which constitutes a distinctive property of Mamba in contrast to Transformer-based architectures.

Connection to Transformers. Prior work has noted that Mamba can be viewed as a *gated linear attention mechanism*, where the gating term dynamically modulates how past information contributes to the current output [3]. From this perspective, our framework complements existing theoretical analyses of Transformer attention by showing how an alternative gating-based mechanism can also align with relevant features while ignoring irrelevant ones.

Contributions. Our results provide the first theoretical characterization of the training dynamics of Mamba's gating mechanism. We prove that: (i) training with GD achieves guaranteed generalization once the sample size and number of iterations scale according to our derived sample complexity and convergence rate bounds; (ii) the gating vector aligns with class-relevant features while negating irrelevant ones, thereby formalizing its role as a feature selector (See Lemmas 5 and 6 in Appendix D); (iii) **the distance between class-relevant features**, captured by ΔL, directly impacts learning speed, highlighting the role of token order and scanning strategies in Mamba. Overall, this work establishes foundational insights into when and why Mamba's selection mechanism enables efficient learning, offering a theoretical counterpart to Transformer-centric analyses.

2 Problem Setup and Main Theoretical Result

We study a binary classification task with training data $\{(\boldsymbol{X}^{(n)}, z^{(n)})\}_{n=1}^{N}$, where each sequence $\boldsymbol{X}^{(n)} = [\boldsymbol{x}_1^{(n)}, \ldots, \boldsymbol{x}_L^{(n)}] \in \mathbb{R}^{d\times L}$ contains label-relevant and irrelevant tokens. Labels are determined by a majority vote over the class-relevant tokens. For example, a positive sample may contain two occurrences of the positive token $\boldsymbol{o}+$ and only one occurrence of the negative token $\boldsymbol{o}-$, whereas a negative sample would contain the opposite. We use a simplified *Mamba block* with input-dependent gating, followed by a two-layer MLP trained with SGD on hinge loss. Given a sequence $\boldsymbol{X}$, the model output is

$$F(\boldsymbol{X}) = \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m} v_i\,\phi\left(\boldsymbol{W}_{O(i,\cdot)}\sum_{s=1}^{t}\left(\prod_{j=s+1}^{t}\left(1-\sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_j)\right)\right)\cdot\sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_s)\,(\boldsymbol{W}_B^\top \boldsymbol{x}_s)^\top(\boldsymbol{W}_C^\top \boldsymbol{x}_t)\,\boldsymbol{x}_s\right), \quad (1)$$

where $\phi(\cdot)$ is the ReLU activation.

Our main result, Theorem 1, gives conditions under which the model generalizes.

Theorem 1 (Generalization of Mamba)**.** *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 high probability, if the number of training samples N satisfies $N \geq \Omega\left(\frac{4L^2 d}{\eta^2\left[1+(1/2)^{\Delta L}\right]^2}\right)$, and the number of iterations T satisfies $T = \Theta\left(\frac{L^2}{\eta\left[1+(1/2)^{\Delta L}\right]}\right)$, the returned model achieves guaranteed generalization.*

3 Numerical Experiments

We complement our theory with simple synthetic experiments under the majority-voting data model. The results confirm our prediction that larger ΔL slows convergence (Table 1). We also measured the cosine similarity between $\boldsymbol{w}_\Delta$ and feature directions, finding 0.18 for $\boldsymbol{o}+$, 0.22 for $\boldsymbol{o}-$, and -0.08 for irrelevant features, confirming that the gating filters in relevant tokens while ignoring irrelevant ones.

Table 1: Average epochs for convergence under varying distances between class-relevant features ΔL.

N || ΔL=1	2	4	6	10	25
100 || 1.90	4.10	18.55	22.25	32.15	47.65
200 || 1.50	2.50	10.60	19.50	26.65	36.50

4 Conclusion and Future Work

This paper provides a first-step theoretical analysis of the Mamba architecture by examining its gated selection mechanism under the majority-voting data model. We establish sample complexity

and convergence guarantees for training with GD, and show that the gating vector aligns with class-relevant features while ignoring irrelevant ones. To the best of our knowledge, this is the first generalization result for Mamba, highlighting its role as an implicit feature selector. Our results also reveal that token order has a significant impact on the performance, and that different scanning strategies in Mamba can lead to distinct behaviors. Future research directions include extending the analysis to more general data distributions, deeper or multi-head Mamba variants, and hybrid architectures that combine Mamba with attention.

Acknowledgments

This work was supported by the National Science Foundation (NSF) #2349879 and #2349878. We also thank all anonymous reviewers for their constructive comments.

References

[1] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *Advances in neural information processing systems*, 36: 48314–48362, 2023.

[2] 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*, pages 119–130, 2024.

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

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

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

[6] 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.

[7] Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. *arXiv preprint arXiv:2302.06015*, 2023.

[8] Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? *arXiv preprint arXiv:2402.15607*, 2024.

[9] 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*, pages 28784–28829. PMLR, 2024.

[10] Hongkang Li, Yihua Zhang, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. *arXiv preprint arXiv:2504.10957*, 2025.

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

[12] 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.

[13] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

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

[15] Cong Qi, Hanzhang Fang, Tianxing Hu, Siqi Jiang, and Wei Zhi. Bidirectional mamba for single-cell data: Efficient context learning with biological fidelity. *arXiv preprint arXiv:2504.16956*, 2025.

[16] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

[17] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

[18] Jiawei Sun, Shuai Zhang, Hongkang Li, and Meng Wang. Theoretical guarantees and training dynamics of contrastive learning: How misaligned data influence feature purity. In *High-dimensional Learning Dynamics 2025*, 2025. URL `https://openreview.net/forum?id=rO9riCUHD9`.

[19] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

[20] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in neural information processing systems*, 36:71911–71947, 2023.

[21] Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*, 2023.

[22] Ting Yu Tsai, Li Lin, Shu Hu, Ming-Ching Chang, Hongtu Zhu, and Xin Wang. Uu-mamba: Uncertainty-aware u-mamba for cardiac image segmentation. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 267–273, 2024. doi: 10.1109/MIPR62202.2024.00050.

[23] 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.

[24] 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*, 2024.

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

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

[27] Yubiao Yue and Zhenzhang Li. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024.

[28] 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.

[29] Qi Zhao, Ze Li, Qian Mao, Tingwei Chen, Yiran Zhang, Bingle Li, Zheng Zhao, and Xiaoya Fan. Mambacpg: an accurate model for single-cell dna methylation status imputation using mamba. *Briefings in Bioinformatics*, 26(4):bbaf360, 2025.

[30] 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*, pages 62429–62442. PMLR, 2024.

A Data Model

Consider an arbitrary set of orthonormal vectors $\{\boldsymbol{o}_+, \boldsymbol{o}_-, \boldsymbol{o}_3, \ldots, \boldsymbol{o}_d\}$ in $\mathbb{R}^d$, where $\boldsymbol{o}_+$ and $\boldsymbol{o}_-$ represent discriminative features, and the remaining $\boldsymbol{o}_j$ are non-discriminative (filler) features ($\boldsymbol{o}_+ \equiv \boldsymbol{o}_1$, $\boldsymbol{o}_- \equiv \boldsymbol{o}_2$). We add token-level noise to all input tokens; each token $\boldsymbol{x}_l^{(n)}$ in $\boldsymbol{X}^{(n)}$ is a noisy version of one of the input patterns. The label is determined by a majority vote over the class-relevant features in the input sequence. In our data model, a positive sample contains two instances of $\boldsymbol{o}_+$ and one instance of $\boldsymbol{o}_-$. Let L_1^+ and L_2^+ denote the positions of the two $\boldsymbol{o}_+$ tokens (with $L_1^+ < L_2^+$), and let L^- denote the position of the $\boldsymbol{o}_-$ token. We assume that L_1^+, L_2^+, and L^- are drawn independently from the same distribution over sequence positions.

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

B Related Work

State Space Models (SSMs). State space models (SSMs) have recently gained attention as a competitive alternative to attention-based architectures for sequence modeling. Early works such as S4[5, 6, 17] demonstrated that carefully parameterized state space recurrences can capture long-range dependencies with strong hardware efficiency, achieving competitive results on long-sequence language and audio benchmarks. More recent variants, including Mamba [4, 3], extend this line of research by introducing input-dependent gating, which enables dynamic selection of relevant features and further improves performance across natural language and vision tasks. Beyond 1D sequences, SSMs have also been adapted to two-dimensional data. For instance, VMamba [12] proposes SS2D, which leverages multiple scanning routes to align the ordered structure of selective 1D scans with the non-sequential nature of vision inputs, thereby enhancing contextual aggregation and boosting recognition performance. Other extensions, such as Graph Mamba [24, 2], demonstrate the adaptability of selective SSMs to non-Euclidean data domains by incorporating structured graph information. These empirical advances underscore the importance of input ordering and scanning strategies in determining the effectiveness of SSMs, a phenomenon that resonates with the structured data models considered in our theoretical analysis.

Theoretical Analysis of SSMs. While empirical studies have demonstrated the effectiveness of SSMs across language, vision, and graph domains, theoretical investigations remain relatively scarce. Existing analyses primarily focus on drawing connections between SSMs and attention-like mechanisms, such as interpreting the recurrence as a form of linear attention or highlighting similarities in feature selection[3]. Although these works provide valuable intuition, they do not address fundamental questions regarding how selective mechanisms in SSMs affect learning dynamics, sample complexity, or generalization under different data distributions. This gap motivates us to study how selective mechanisms in SSMs interact with structured data, a perspective inspired by recent theoretical works on structured models such as vision transformers and graph neural networks(GNNs) [7, 8, 28]. In particular, our work develops a theoretical framework that directly characterizes the role of the gating mechanism in Mamba under structured data models.

Feature Learning Framework. The feature learning framework provides a systematic perspective for analyzing how neural networks gradually emphasize informative features while suppressing uninformative ones during training. Several recent works have adopted this framework to study the optimization and generalization of Transformers [7, 11, 8, 10, 18], showing how attention mechanisms evolve to highlight class-relevant tokens and ignore irrelevant noise. Similar analyses have also been conducted for GNNs [28, 9], where message passing is shown to amplify discriminative structures in the graph progressively.

Despite the recent empirical success of Mamba, its feature learning properties in selective SSMs remain theoretically unexplored. To the best of our knowledge, we provide the first such characterization. In particular, we analyze how the gating mechanism in Mamba evolves under stochastic gradient descent, thereby extending the feature learning perspective to this new family of architectures.

C Main Theorems

In this section, we present the formal statements of our main theorems.

Let $\mathbf{\Psi} = (\boldsymbol{v}, \boldsymbol{W}_O, \boldsymbol{w}_\Delta, \boldsymbol{W}_B, \boldsymbol{W}_C)$ denote the set of parameters to train. The generalization performance of the learned model $\mathbf{\Psi}$ is evaluated using the population risk $f(\mathbf{\Psi})$, defined as

$$f(\mathbf{\Psi}) = f(\boldsymbol{v}, \boldsymbol{W}_O, \boldsymbol{w}_\Delta, \boldsymbol{W}_B, \boldsymbol{W}_C) = \mathbb{E}_{(\boldsymbol{X},z)\sim\mathcal{D}}\, \ell(\boldsymbol{X}, z), \tag{2}$$

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

Theorem 1 establishes the sample complexity (3) and convergence rate (4) required to guarantee generalization when training with SGD for the majority-voting data model. In other words, the model generalizes once enough samples are available (3) and training has run for the required number of iterations (4).

Theorem 1 (Generalization of Mamba)**.** *Let the learning rate $\eta > 0$ be a positive constant. 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

$$N \geq \Omega\left(\frac{4L^2 d}{\eta^2 \left[1 + \left(\frac{1}{2}\right)^{\Delta L}\right]^2}\right), \tag{3}$$

and the number of iterations T satisfies

$$T = \Theta\left(\frac{L^2}{\eta \left[1 + \left(\frac{1}{2}\right)^{\Delta L}\right]}\right), \tag{4}$$

the model obtained after T iterations of SGD achieves guaranteed generalization, i.e.,

$$f\Big(\boldsymbol{v}^{(0)}, \boldsymbol{W}_O^{(T)}, \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{W}_B^{(0)}, \boldsymbol{W}_C^{(0)}\Big) = 0. \tag{5}$$

Theorem 2 establishes that after sufficient training, the gating vector $\boldsymbol{w}_\Delta$ aligns positively with the class-relevant features $\boldsymbol{o}_+$(6) and $\boldsymbol{o}_-$(7), while its alignment with irrelevant features remains strictly negative (8). In other words, the selection mechanism implicitly acts as a feature selector, amplifying relevant tokens and suppressing irrelevant ones.

Theorem 2 (Gating Vector Alignment)**.** *Suppose training is performed under the same conditions as in Theorem 1, with initialization where each entry of $\boldsymbol{W}_O$ is drawn independently from $\mathcal{N}(0, \xi^2)$ and $\boldsymbol{w}_\Delta^{(0)} = 0$. If the number of training samples N satisfies $N \geq \Omega\left(\frac{4L^2 d}{\eta^2\left[1+(1/2)^{\Delta L}\right]^2}\right)$, and the number of iterations T satisfies $T = \Theta\left(\frac{L^2}{\eta\left[1+(1/2)^{\Delta L}\right]}\right)$,*

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_+ \right\rangle \geq \frac{\eta T c'^3}{16L}\left[\frac{1}{c'^2} + \left(\frac{1}{2}\right)^{\Delta L^+ - 2}\right] \tag{6}$$

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_- \right\rangle \geq \frac{\eta T c'^3}{16L}\left[\frac{1}{c'^2} + \left(\frac{1}{2}\right)^{\Delta L^- - 2}\right] \tag{7}$$

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_j \right\rangle \leq \frac{-\eta T c'^3}{16L}\left[\frac{1}{c'^2} + \left(\frac{1}{2}\right)^{\Delta L^+ - 2}\right]\left[\left(\frac{1}{2}\right)^{\Delta L^+} + \left(\frac{1}{2}\right)^{\Delta L^-}\right], \tag{8}$$

for any irrelevant feature $\boldsymbol{o}_j$.

D Useful Lemmas

Lemma 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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{1}{\sqrt{m}L} \cdot \sigma(p_1) \left[1 + (1-\sigma(q_1))^2 (1-\sigma(q_2))^{L_2^+ - L_1^+ - 2} \right] - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{9}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \frac{1}{\sqrt{m}L} \cdot \sigma(q_1) \left[1 + (1-\sigma(p_1))^2 (1-\sigma(p_2))^{L_2^+ - L_1^+ - 2} \right] + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{10}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \geq -\frac{1}{\sqrt{m}L} \cdot \sigma(p_1) \left[2 + (1-\sigma(q_1))^2 (1-\sigma(q_2))^{L_2^- - L_1^- - 2} \right] - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{11}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{12}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right), \quad \text{for } j \neq 1, 2. \tag{13}$$

Lemma 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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{14}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right). \tag{15}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}} \right), \quad \text{for } j \neq 1, 2. \tag{16}$$

Lemma 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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \geq \frac{1}{\sqrt{m}L}\cdot\sigma(p_1)\left[1+(1-\sigma(q_1))^2(1-\sigma(q_2))^{L_2^- - L_1^- - 2}\right] - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{17}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \frac{1}{\sqrt{m}L}\cdot\sigma(q_1)\left[1+(1-\sigma(p_1))^2(1-\sigma(p_2))^{L_2^- - L_1^- - 2}\right] + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{18}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \geq -\frac{1}{\sqrt{m}L}\cdot\sigma(p_1)\left[2+(1-\sigma(q_1))^2(1-\sigma(q_2))^{L_2^+ - L_1^+ - 2}\right] - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{19}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{20}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right), \quad \textit{for } j \neq 1, 2. \tag{21}$$

Lemma 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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{22}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{23}$$

(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*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right), \quad \textit{for } j \neq 1, 2. \tag{24}$$

Lemma 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:*

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

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{\sigma(p_1)\left(1-\sigma(q_1)\right)}{2}\left[\frac{2r_1^*\rho_t^+}{\sqrt{m}} - \frac{\sqrt{m}}{2}s_1^*\right] - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{25}$$

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:*

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

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_- \right\rangle \geq \frac{\sigma(p_1)\left(1-\sigma(q_1)\right)}{2}\left[\frac{2r_1^*\rho_t^-}{\sqrt{m}} - \frac{\sqrt{m}}{2}s_1^*\right] - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{26}$$

Lemma 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:*

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_j \right\rangle \leq -\frac{r_1^*}{2\sqrt{m}} \cdot \sigma(p_1)\left(1-\sigma(q_1)\right)\left[\left(1-\sigma(q_2)\right)^{L_2^+ - L_1^+}\rho_t^+ + \left(1-\sigma(q_2)\right)^{L_2^- - L_1^-}\rho_t^-\right]. \tag{27}$$

E Proof of Convergence

Proof. Suppose $\boldsymbol{w}_\Delta^{(0)} = \boldsymbol{0}$. Then, we have

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

From Lemma 1, identify $p_1 = 0$, $q_1 = 0$, $p_2 = 0$ and $q_2 = 0$. Let L_1^+ and L_2^+ denote the positions of the two $\boldsymbol{o}_+$ tokens in the positive sample, with $L_1^+ < L_2^+$. Define the gap between them as

$$\Delta L^+ := L_2^+ - L_1^+. \tag{28}$$

Similarly, for the negative sample, let L_1^- and L_2^- be the positions of the two $\boldsymbol{o}_-$ tokens, with $L_1^- < L_2^-$, and define

$$\Delta L^- := L_2^- - L_1^-. \tag{29}$$

Then, for any lucky neuron $i \in \mathcal{W}(0)$, we obtain

$$\begin{aligned} \frac{1}{2\sqrt{m}L}\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) &\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(0)}}, \boldsymbol{o}_+ \right\rangle \\ &\leq \frac{1}{2\sqrt{m}L}\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+}\right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right). \end{aligned} \tag{30}$$

We can relax the lower bound and obtain

$$\begin{aligned} \frac{c'^2}{2\sqrt{m}L}\left[\frac{1}{c'^2} + \left(\frac{1}{2}\right)^{\Delta L^+ - 2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) &\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(0)}}, \boldsymbol{o}_+ \right\rangle \\ &\leq \frac{1}{2\sqrt{m}L}\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+}\right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \end{aligned} \tag{31}$$

$$\text{and} \quad \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(0)}}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 1. \tag{32}$$

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

Suppose the initialization is

$$\boldsymbol{W}_{O(i,\cdot)}(0) = \delta_1 \boldsymbol{o}_+ + \delta_2 \boldsymbol{o}_- + \cdots + \delta_d \boldsymbol{o}_d, \quad \delta_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \xi^2) \quad j = 1, 2, \cdots, d. \tag{33}$$

Then, after one gradient descent step, we have

$$\begin{aligned} \delta_1 + \frac{\eta c'^2}{2\sqrt{m}L}\left[\frac{1}{c'^2} + \left(\frac{1}{2}\right)^{\Delta L^+ - 2}\right] & \\ -\widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \le \left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(1)}, \boldsymbol{o}_+ \right\rangle & \\ \le \delta_1 + \frac{\eta}{2\sqrt{m}L}\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+}\right] & \\ + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right). & \end{aligned} \tag{34}$$

$$\text{and} \quad \left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(1)}, \boldsymbol{o}_j \right\rangle \le \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 1. \tag{35}$$

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

$$\begin{aligned} \delta_2 + \frac{\eta c'^2}{2\sqrt{m}L}\left[\frac{1}{c'^2} + \left(\frac{1}{2}\right)^{\Delta L^- - 2}\right] & \\ - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \le \left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(1)}, \boldsymbol{o}_- \right\rangle & \\ \le \delta_2 + \frac{\eta}{2\sqrt{m}L}\left[1 + \left(\frac{1}{2}\right)^{\Delta L^-}\right] & \\ + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right). & \end{aligned} \tag{36}$$

$$\text{and} \quad \left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(1)}, \boldsymbol{o}_j \right\rangle \le \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 2. \tag{37}$$

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

$$\left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(1)}, \boldsymbol{o}_j \right\rangle \le \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } \forall j. \tag{38}$$

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

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

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

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

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(0)}}, \boldsymbol{o}_+ \right\rangle \ge \frac{1}{8}\left[\frac{2a}{\sqrt{m}} \cdot \rho_0^+ - \frac{\sqrt{m}}{2} b\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \tag{39}$$

We can relax this lower bound and obtain

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(0)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{c'}{4}\left[\frac{2a}{\sqrt{m}} \cdot \rho_0^+ - \frac{\sqrt{m}}{2} b\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right) =: \alpha \tag{40}$$

Since $\rho_0^+ \approx \frac{m}{2}$ under random initialization, when m is sufficiently large, the above gradient update is positive.

Let $\delta_1 = \frac{1}{\text{poly}(d)}$. Since $a \approx b$,

$$\begin{aligned}
\alpha &= \frac{c'}{4}\left[\frac{2a}{\sqrt{m}} \cdot \frac{m}{2} - \frac{\sqrt{m} b}{2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right) \\
&= \frac{c'}{4}\left[\sqrt{m} a - \frac{\sqrt{m} a}{2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right) \\
&= \frac{c'}{4} \cdot \frac{\eta c'^2}{4L}\left[\frac{1}{c'^2} + \left(\frac{1}{2}\right)^{\Delta L^+ - 2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right) \\
&= \frac{\eta c'^3}{16L}\left[\frac{1}{c'^2} + \left(\frac{1}{2}\right)^{\Delta L^+ - 2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right) > 0
\end{aligned} \tag{41}$$

From Lemma 6, we also obtain

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(0)}}, \boldsymbol{o}_j \right\rangle \leq \frac{-a}{8\sqrt{m}}\left[\left(\frac{1}{2}\right)^{\Delta L^+} \cdot \rho_0^+ + \left(\frac{1}{2}\right)^{\Delta L^-} \cdot \rho_0^-\right] \tag{42}$$

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

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(0)}}, \boldsymbol{o}_j \right\rangle \leq \frac{-ac'}{4\sqrt{m}}\left[\left(\frac{1}{2}\right)^{\Delta L^+} \cdot \rho_0^+ + \left(\frac{1}{2}\right)^{\Delta L^-} \cdot \rho_0^-\right] =: \gamma \tag{43}$$

Taking $\rho_0^+ = \rho_0^- \approx \frac{m}{2}$, we can simplify and write

$$\begin{aligned}
\gamma &= \frac{-ac'}{4\sqrt{m}} \cdot \frac{m}{2}\left[\left(\frac{1}{2}\right)^{\Delta L^+} + \left(\frac{1}{2}\right)^{\Delta L^-}\right] \\
&= -\sqrt{m} a \cdot \frac{c'}{8}\left[\left(\frac{1}{2}\right)^{\Delta L^+} + \left(\frac{1}{2}\right)^{\Delta L^-}\right] \\
&= \frac{-\eta c'^3}{16L}\left[\frac{1}{c'^2} + \left(\frac{1}{2}\right)^{\Delta L^+ - 2}\right]\left[\left(\frac{1}{2}\right)^{\Delta L^+} + \left(\frac{1}{2}\right)^{\Delta L^-}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right).
\end{aligned} \tag{44}$$

Let $\langle \boldsymbol{w}_\Delta^{(1)}, \boldsymbol{o}_+ \rangle = \alpha^* \geq \alpha$, $\langle \boldsymbol{w}_\Delta^{(1)}, \boldsymbol{o}_- \rangle = \beta^* \geq \beta$, and $\langle \boldsymbol{w}_\Delta^{(1)}, \boldsymbol{o}_j \rangle = \gamma^* \leq \gamma$, where

$$\beta = \frac{c'}{4}\left[\frac{2a'}{\sqrt{m}} \cdot \rho_0^- - \frac{\sqrt{m}}{2} b'\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right) > 0 \tag{45}$$

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

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

Following the same approach as in (41), we can simplify and obtain

$$\beta = \frac{\eta c'^3}{16L}\left[\frac{1}{c'^2} + \left(\frac{1}{2}\right)^{\Delta L^- - 2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) > 0 \tag{46}$$

For any lucky neuron $i \in \mathcal{W}(1)$ at iteration 2, we have

$$\begin{aligned}
&\frac{c'^2}{\sqrt{m}L}\cdot\sigma(\alpha^*)\left[\frac{1}{c'^2} + (1-\sigma(\gamma^*))^{\Delta L^+ - 2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
&\quad\le \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(1)}}, \boldsymbol{o}_+ \right\rangle \\
&\quad\le \frac{1}{\sqrt{m}L}\cdot\sigma(\alpha^*)\left[1 + (1-\sigma(\alpha^*))^2(1-\sigma(\gamma^*))^{\Delta L^+ - 2}\right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)
\end{aligned} \tag{47}$$

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

Note that, $\sigma(\alpha^*) > \frac{1}{2}$ and $\sigma(\gamma^*) < \frac{1}{2}$. This ensures $\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(0)}}, \boldsymbol{o}_+ \right\rangle \le \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(1)}}, \boldsymbol{o}_+ \right\rangle$. Thus, we obtain the following bound after the second gradient descent step.

$$\begin{aligned}
&\delta_1 + \frac{\eta c'^2}{\sqrt{m}L}\left[\frac{1}{2c'^2} + \left(\frac{1}{2}\right)^{\Delta L^+ - 1} + \sigma(\alpha^*)\left(1/c'^2 + (1-\sigma(\gamma^*))^{\Delta L^+ - 2}\right)\right] \\
&- \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) =: u \\
&\le \left\langle (\boldsymbol{W}_{O(i,\cdot)}^{(2)})^\top, \boldsymbol{o}_+ \right\rangle \\
&\le \delta_1 + \frac{\eta}{2\sqrt{m}L}\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+} + 2\sigma(\alpha^*)\left(1 + (1-\sigma(\alpha^*))^2(1-\sigma(\gamma^*))^{\Delta L^+ - 2}\right)\right] \\
&+ \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) =: v
\end{aligned} \tag{49}$$

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

$$\begin{aligned}
&\frac{c'^2}{\sqrt{m}L}\cdot\sigma(\beta^*)\left[\frac{1}{c'^2} + (1-\sigma(\gamma^*))^{\Delta L^- - 2}\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \\
&\quad\le \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(1)}}, \boldsymbol{o}_- \right\rangle \\
&\quad\le \frac{1}{\sqrt{m}L}\cdot\sigma(\beta^*)\left[1 + (1-\sigma(\beta^*))^2(1-\sigma(\gamma^*))^{\Delta L^- - 2}\right] + \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)
\end{aligned} \tag{50}$$

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(1)}}, \boldsymbol{o}_j \right\rangle \le \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 2 \tag{51}$$

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

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(1)}}, \boldsymbol{o}_+ \right\rangle \geq \frac{\sigma(\alpha^*)c'}{2}\left[\frac{2u}{\sqrt{m}}\cdot\rho_1^+ - \frac{\sqrt{m}}{2}v\right] - \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) =: \chi \tag{52}$$

Since $u \approx v$, we have $\chi \geq 0$.

By applying Lemma 6 with

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

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta^{(t)}}, \boldsymbol{o}_j \right\rangle \leq -\frac{c'u}{2\sqrt{m}}\sigma(\alpha^*)\left[(1-\sigma(\gamma^*))^{\Delta L^+}\,\rho_t^+ \;+\; (1-\sigma(\gamma^*))^{\Delta L^-}\,\rho_t^-\right] =: \iota \tag{53}$$

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

Finally, for any lucky neuron $i \in \mathcal{W}(T)$, we obtain

$$\left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(T)}, \boldsymbol{o}_+ \right\rangle \geq aT \tag{54}$$

$$\left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(T)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 1 \tag{55}$$

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

$$\left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(T)}, \boldsymbol{o}_- \right\rangle \geq aT \tag{56}$$

$$\left\langle {\boldsymbol{W}_{O(i,\cdot)}^\top}^{(T)}, \boldsymbol{o}_j \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right) \quad \text{for } j \neq 2 \tag{57}$$

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_+ \right\rangle \geq \alpha T \tag{58}$$

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_- \right\rangle \geq \beta T \tag{59}$$

$$\left\langle \boldsymbol{w}_\Delta^{(T)}, \boldsymbol{o}_j \right\rangle \leq \gamma T \tag{60}$$

$$\boldsymbol{X}^{(n)} = \left[\boldsymbol{x}_1^{(n)} \quad \boldsymbol{x}_2^{(n)} \quad \cdots \quad \boldsymbol{x}_L^{(n)}\right]$$

Consider $z^{(n)} = +1$ as an example. The sequence $\boldsymbol{X}^{(n)}$ has two $\boldsymbol{o}_+$ and one $\boldsymbol{o}_-$ at locations L_1^+, L_2^+, and L^-.

$$\begin{aligned}
F(\boldsymbol{X}^{(n)}) &= \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m} v_i\,\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \\
&= \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}^+}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) - \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}^-}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \\
&\geq \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) - \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{U}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \\
&\quad - \frac{1}{\sqrt{m}L}\sum_{i\in\mathcal{K}^-\setminus\mathcal{U}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)
\end{aligned} \tag{61}$$

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

$$\boldsymbol{y}_l^{(n)} = \sum_{\tau=0}^{l-1} \left(\prod_{r=\tau+1}^{l-1} \left(1 - \sigma\left(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_{l-r}^{(n)}\right)\right) \right) \cdot \sigma\left(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_{l-\tau}^{(n)}\right) \cdot (\boldsymbol{x}_{l-\tau}^{(n)\top} \boldsymbol{x}_l^{(n)}) \boldsymbol{x}_{l-\tau}^{(n)}. \tag{62}$$

Equivalently, letting $s = l - \tau$, we write

$$\boldsymbol{y}_l^{(n)} = \sum_{s=1}^{l} \left(\prod_{j=s+1}^{l} \left(1 - \sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_j^{(n)})\right) \right) \cdot \sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_s^{(n)}) \cdot (\boldsymbol{x}_s^{(n)\top} \boldsymbol{x}_l^{(n)}) \boldsymbol{x}_s^{(n)}. \tag{63}$$

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

$$\sum_{i \in \mathcal{W}(0)} \sum_{l=1}^{L} \boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l = \sum_{i \in \mathcal{W}(0)} \sum_{l=1}^{L} \sum_{j=1}^{d} \langle \boldsymbol{W}_{O(i,\cdot)}^\top, \boldsymbol{o}_j \rangle \cdot \langle \boldsymbol{y}_l, \boldsymbol{o}_j \rangle. \tag{64}$$

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

$$\langle \boldsymbol{W}_{O(i,\cdot)}^\top, \boldsymbol{o}_+ \rangle \geq aT. \tag{65}$$

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

$$\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_+ \rangle \geq \alpha T, \quad \langle \boldsymbol{w}_\Delta, \boldsymbol{o}_- \rangle \geq \beta T, \quad \langle \boldsymbol{w}_\Delta, \boldsymbol{o}_j \rangle \leq \gamma T \quad \text{for } j \neq 1, 2. \tag{66}$$

Therefore,

$$\left\langle \boldsymbol{y}_{L_1^+}, \boldsymbol{o}_+ \right\rangle = \sigma\left(\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_+ \rangle\right) \geq \sigma(\alpha T). \tag{67}$$

We have,

$$\langle \boldsymbol{w}_\Delta, \boldsymbol{o}_+ \rangle \leq W_1 T, \qquad \langle \boldsymbol{w}_\Delta, \boldsymbol{o}_- \rangle \leq W_2 T,$$

where

$$W_1 = \frac{\eta}{16L} \left[\frac{1}{2} + \left(\frac{1}{2}\right)^{\Delta L^+ + 1} \right] + \tilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right), \tag{68}$$

$$W_2 = \frac{\eta}{16L} \left[\frac{1}{2} + \left(\frac{1}{2}\right)^{\Delta L^- + 1} \right] + \tilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right). \tag{69}$$

Then we obtain the following:

$$\begin{aligned} \left\langle \boldsymbol{y}_{L_2^+}, \boldsymbol{o}_+ \right\rangle &\geq \sigma(\alpha T) + \left(1 - \sigma(W_1 T)\right) \left(1 - \sigma(W_2 T)\right) \left(1 - \sigma(\gamma T)\right)^{\Delta L^+ - 2} \cdot \sigma(\alpha T) \\ &= \sigma(\alpha T) \left[1 + \left(1 - \sigma(W_1 T)\right) \left(1 - \sigma(W_2 T)\right) \left(1 - \sigma(\gamma T)\right)^{\Delta L^+ - 2} \right]. \end{aligned} \tag{70}$$

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

$$\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_{L_2^+}\geq aT\cdot\sigma(\alpha T)\left[1+\left(1-\sigma(W_1T)\right)\left(1-\sigma(W_2T)\right)\left(1-\sigma(\gamma T)\right)^{\Delta L^+-2}\right]>0.$$

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

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

$$\phi(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_{L_2^+})\geq aT\cdot\sigma(\alpha T)\left[1+\left(1-\sigma(W_1T)\right)\left(1-\sigma(W_2T)\right)\left(1-\sigma(\gamma T)\right)^{\Delta L^+-2}\right].$$

Hence,

$$\begin{aligned}\sum_{i\in\mathcal{W}(0)}\sum_{l=1}^{L}\phi(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l)\geq&\sum_{i\in\mathcal{W}(0)}aT\cdot\sigma(\alpha T)\\&\times\left[2+\left(1-\sigma(W_1T)\right)\left(1-\sigma(W_2T)\right)\left(1-\sigma(\gamma T)\right)^{\Delta L^+-2}\right].\end{aligned}\tag{71}$$

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

$$0<\langle\boldsymbol{W}_{O(i,\cdot)}^{\top},\boldsymbol{o}_-\rangle\leq bT.\tag{72}$$

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 $1\leq s\leq l$.

$$s=L^-\text{ and }l=L^-.$$

$$\langle\boldsymbol{y}_{L^-},\boldsymbol{o}_-\rangle=\sigma\left(\langle\boldsymbol{w}_\Delta,\boldsymbol{o}_-\rangle\right)\leq\sigma(W_2T).\tag{73}$$

$$\sum_{l=1}^{L}\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l\leq bT\cdot\sigma(W_2T).\tag{74}$$

$$\sum_{i\in\mathcal{U}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\leq\sum_{i\in\mathcal{U}(0)}bT\cdot\sigma(W_2T).\tag{75}$$

In addition, we have

$$\sum_{i\in\mathcal{K}^-\setminus\mathcal{U}(0)}\sum_{l=1}^{L}\phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\leq\tilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right).\tag{76}$$

By (61), we can write

$$F(\boldsymbol{X}^{(n)}) \geq \frac{1}{\sqrt{m}L}\Bigg\{\frac{m}{2}\cdot aT\cdot\sigma(\alpha T)\left[2+(1-\sigma(W_1T))(1-\sigma(W_2T))(1-\sigma(\gamma T))^{\Delta L^+-2}\right] \\ -\frac{m}{2}\cdot bT\cdot\sigma(W_2T)-\widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right)\Bigg\}, \tag{77}$$

with

$$a=\frac{\eta}{\sqrt{m}L}\left[\frac{1}{2}+c'^2\left(\frac{1}{2}\right)^{\Delta L^+-1}\right]-\widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right), \tag{78}$$

$$\text{and}\quad b=\frac{\eta}{\sqrt{m}L}\left[\frac{1}{2}+\left(\frac{1}{2}\right)^{\Delta L^+ +1}\right]+\widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right). \tag{79}$$

$$a\approx b. \tag{80}$$

$$\begin{aligned}\alpha&=\frac{\eta c'^3}{16L}\left[\frac{1}{c'^2}+\left(\frac{1}{2}\right)^{\Delta L^+-2}\right]-\widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right)\\&=\frac{\eta}{16L}\left[c'+c'^3\left(\frac{1}{2}\right)^{\Delta L^+-2}\right]-\widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right)\\&\approx W_1\approx W_2.\end{aligned} \tag{81}$$

The last step assumes the distribution of ΔL^+ is identical to ΔL^- .

Therefore, we conclude that

$$F(\boldsymbol{X}^{(n)}) \geq \frac{1}{\sqrt{m}L}\left\{\frac{m}{2}\cdot aT\cdot\sigma(\alpha T)\left[1+(1-\sigma(W_1T))(1-\sigma(W_2T))(1-\sigma(\gamma T))^{\Delta L^+-2}\right]\right\} \\ -\widetilde{\mathcal{O}}\left(\frac{1}{\text{poly}(d)}\right) \tag{82}$$

$$F(\boldsymbol{X}^{(n)})\geq C,\ \text{where C is some positive constant.} \tag{83}$$

E.1 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

$$\frac{1}{\sqrt{m}L}\cdot\frac{m}{2}\cdot aT\geq 1+\epsilon. \tag{84}$$

Substituting the value of $a\approx b=\frac{\eta}{\sqrt{m}L}\left[\frac{1}{2}+\left(\frac{1}{2}\right)^{\Delta L^+ +1}\right]$, the condition becomes

$$\begin{aligned}\frac{\sqrt{m}aT}{2L}&=\frac{\sqrt{m}}{2L}\cdot\frac{\eta}{\sqrt{m}L}\left[\frac{1}{2}+\left(\frac{1}{2}\right)^{\Delta L^+ +1}\right]T\\&=\frac{\eta T}{2L^2}\left[\frac{1}{2}+\left(\frac{1}{2}\right)^{\Delta L^+ +1}\right]\geq 1+\epsilon.\end{aligned} \tag{85}$$

Solving for T, we obtain

$$T\geq\frac{2L^2(1+\epsilon)}{\eta\left[\frac{1}{2}+\left(\frac{1}{2}\right)^{\Delta L^+ +1}\right]}=\frac{4L^2(1+\epsilon)}{\eta\left[1+\left(\frac{1}{2}\right)^{\Delta L^+}\right]}\approx\frac{4L^2}{\eta\left[1+\left(\frac{1}{2}\right)^{\Delta L^+}\right]}. \tag{86}$$

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

$$\sigma(\alpha T) \geq 1 - \epsilon. \tag{87}$$

When z is sufficiently large we can approximate

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

Substituting $z = \alpha T$, condition (87) becomes:

$$\begin{aligned} \sigma(\alpha T) &\approx 1 - e^{-\alpha T} \geq 1 - \epsilon, \\ e^{-\alpha T} &\leq \epsilon, \\ \alpha T &\geq -\ln(\epsilon) \\ T &\geq -\frac{\ln(\epsilon)}{\alpha}. \end{aligned} \tag{88}$$

Substituting $\alpha = \frac{\eta}{16L}\left[\frac{1}{2} + \left(\frac{1}{2}\right)^{\Delta L^+ + 1}\right]$, we get:

$$T \geq -\ln(\epsilon) \cdot \frac{16L}{\eta\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+ + 1}\right]}. \tag{89}$$

Multiplying both the numerator and the denominator of the second term by 2, we obtain the final bound:

$$T \geq -\ln(\epsilon) \cdot \frac{32L}{\eta\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+}\right]}. \tag{90}$$

Hence, by combining (86) and (90), we obtain

$$T \geq \max\left\{\frac{4L^2}{\eta\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+}\right]}, -\ln(\epsilon) \cdot \frac{32L}{\eta\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+}\right]}\right\}. \tag{91}$$

By combining (84) and (87) with the expression for the model output $F(\boldsymbol{X}^{(n)})$ in (82), we obtain

$$\begin{aligned} F(\boldsymbol{X}^{(n)}) &\geq (1+\epsilon)\cdot(1-\epsilon)\cdot(1+\epsilon) \\ &\geq (1+\epsilon)^2 \cdot (1-\epsilon) \\ &\geq (1 + 2\epsilon + \epsilon^2)\cdot(1-\epsilon) \\ &\geq 1 + \epsilon - \mathcal{O}(\epsilon^2) \end{aligned} \tag{92}$$

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

E.2 Sample Complexity

Previously, we used $N = \mathrm{poly}(d)$ to obtain the asymptotic term $\widetilde{\mathcal{O}}\left(\frac{1}{\mathrm{poly}(d)}\right)$. We now derive a sample-complexity bound that guarantees zero generalization error.

Assume that, for sufficiently small $\lambda \ll 1$,

$$\mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \leq \lambda \cdot \frac{\eta}{2\sqrt{m}L}\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+}\right]. \tag{93}$$

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

$$\begin{aligned} N &\geq \Omega\left(\lambda^{-2} \cdot \frac{4L^2 d}{\eta^2\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+}\right]^2}\right) \\ &\geq \Omega\left(\frac{4L^2 d}{\eta^2\left[1 + \left(\frac{1}{2}\right)^{\Delta L^+}\right]^2}\right). \end{aligned} \tag{94}$$

□

F Proof of Lemmas

F.1 Proof of Lemma 1

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

$$\begin{aligned}\ell(\boldsymbol{X}^{(n)}, z^{(n)}) &= \max\{0, 1 - z^{(n)} \cdot F(\boldsymbol{X}^{(n)})\} \\ &= \max\left\{0, 1 - z^{(n)} \cdot \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m} v_i \phi\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\right\}. \end{aligned} \tag{95}$$

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

$$\hat{\mathcal{L}} = \frac{1}{N}\sum_{n=1}^{N} \ell(\boldsymbol{X}^{(n)}, z^{(n)}). \tag{96}$$

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

$$\mathcal{L} = \mathbb{E}_{(\boldsymbol{X},z)\sim\mathcal{D}}\ell(\boldsymbol{X}, z). \tag{97}$$

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

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{W}_{O(i,\cdot)}} &= \frac{\partial \ell}{\partial F(\boldsymbol{X}^{(n)})} \cdot \frac{\partial F(\boldsymbol{X}^{(n)})}{\partial \boldsymbol{W}_{O(i,\cdot)}} \\ &= -\frac{z^{(n)}}{L}\sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}. \end{aligned} \tag{98}$$

If we consider the gradient for the population loss,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}} = -\mathbb{E}\left[\frac{z^{(n)}}{L}\sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right] \tag{99}$$

$$\begin{aligned} &= -\mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right] \\ &\quad + \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right]. \end{aligned} \tag{100}$$

We are given that

$$p_1 \le \left\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+ \right\rangle \le q_1, \quad p_2 \le \left\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_- \right\rangle \le q_2, \quad \text{and} \quad p_3 \le \left\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_j \right\rangle \le q_3 \quad \text{for } j \ne 1, 2. \tag{101}$$

The Mamba output can be written as

$$\boldsymbol{y}_l(t) = \sum_{s=1}^{l}\left(\prod_{j=s+1}^{l}\left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{x}_j)\right)\right) \cdot \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{x}_s) \cdot (\boldsymbol{x}_s^\top \boldsymbol{x}_l)\boldsymbol{x}_s \tag{102}$$

We have to consider 4 cases.

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

$$\boldsymbol{x}_s = \boldsymbol{x}_l = \boldsymbol{o}_+ \tag{103}$$

$$\left\langle \mathbb{E}\boldsymbol{y}_{L_1^+}, \boldsymbol{o}_+ \right\rangle = \sigma(\boldsymbol{w}_\Delta^{(t)^\top}\boldsymbol{o}_+) \ge \sigma(p_1) = \frac{1}{1 + e^{-p_1}} \tag{104}$$

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

$$\left\langle \mathbb{E}\boldsymbol{y}_{L_2^+, L_2^+}, \boldsymbol{o}_+ \right\rangle = \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+) \tag{105}$$

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

$$\begin{aligned}\left\langle \mathbb{E}\boldsymbol{y}_{L_2^+, L_1^+}, \boldsymbol{o}_+ \right\rangle = &\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right) \\ &\cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^+ - L_1^+ - 2} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\end{aligned} \tag{106}$$

Combining (105) and (106), we obtain

$$\begin{aligned}\left\langle \mathbb{E}\boldsymbol{y}_{L_2^+}, \boldsymbol{o}_+ \right\rangle = &\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+) \\ &+ \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right) \\ &\cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^+ - L_1^+ - 2} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+).\end{aligned} \tag{107}$$

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

$$\begin{aligned}&\left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\ &= \frac{1}{\sqrt{m}L} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\left[2 + \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\right. \\ &\left. \cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^+ - L_1^+ - 2}\right].\end{aligned} \tag{108}$$

For $z = -1$, when $l = s = L^+$, $\boldsymbol{x}_s = \boldsymbol{x}_l = \boldsymbol{o}_+$. Therefore, we obtain

$$\langle \mathbb{E}\boldsymbol{y}_{L^+}, \boldsymbol{o}_+ \rangle = \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+) \geq \sigma(p_1) \tag{109}$$

$$\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle = \frac{1}{\sqrt{m}L} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+) \tag{110}$$

Therfore, combining (108) and (110),

$$\begin{aligned}\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle = &\left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\ &- \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\ = &\frac{1}{\sqrt{m}L} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\left[1 + \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\right. \\ &\left. \cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^+ - L_1^+ - 2}\right].\end{aligned} \tag{111}$$

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

$$\boldsymbol{\gamma}_n = \frac{z^{(n)}}{L} \sum_{l=1}^{L} v_i \, \phi' \left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)} \right) \boldsymbol{y}_l^{(n)}. \tag{112}$$

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.

$$\left| \boldsymbol{\alpha}^\top \boldsymbol{\gamma}_n \right| \leq \|\boldsymbol{\alpha}\|_2 \cdot \|\boldsymbol{\gamma}_n\|_2 = \|\boldsymbol{\gamma}_n\|_2. \tag{113}$$

By the problem setup, we know that

$$|v_i| = \frac{1}{\sqrt{m}}, \quad |z^{(n)}| = 1, \quad \left| \phi' \left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)} \right) \right| \leq 1. \tag{114}$$

Recall the Mamba output,

$$\boldsymbol{y}_l^{(n)}(t) = \sum_{s=1}^{l} \left(\prod_{j=s+1}^{l} \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{x}_j) \right) \right) \cdot \sigma(\boldsymbol{w}_\Delta^{(t)\top} \boldsymbol{x}_s) \cdot (\boldsymbol{x}_s^\top \boldsymbol{x}_l) \boldsymbol{x}_s. \tag{115}$$

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

$$\begin{aligned} \left\| \boldsymbol{y}_l^{(n)} \right\|_2 &\leq \sum_{s=1}^{l} \left| a^{l-s+1} \cdot \left(\boldsymbol{x}_s^\top \boldsymbol{x}_l \right) \right| \cdot \|\boldsymbol{x}_s\|_2 \\ &\leq \sum_{s=1}^{l} \frac{a}{1-a} \cdot 1 \cdot 1 = a' \quad \text{(where } a' \text{ denotes a constant).} \end{aligned} \tag{116}$$

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

$$\begin{aligned} \|\boldsymbol{\gamma}_n\|_2 &\leq \frac{1}{L} \sum_{l=1}^{L} |v_i| \cdot \left| \phi' \left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)} \right) \right| \cdot \left\| \boldsymbol{y}_l^{(n)} \right\|_2 \\ &\leq \frac{1}{L} \cdot \frac{1}{\sqrt{m}} \sum_{l=1}^{L} \left\| \boldsymbol{y}_l^{(n)} \right\|_2 \\ &\leq \frac{1}{L} \cdot \frac{1}{\sqrt{m}} \cdot \sum_{l=1}^{L} a' = \frac{a'}{\sqrt{m}}. \end{aligned} \tag{117}$$

Hence,

$$\left| \boldsymbol{\alpha}^\top \boldsymbol{\gamma}_n \right| \leq \frac{a'}{\sqrt{m}} \quad \text{(bounded).} \tag{118}$$

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

$$\sigma^2 = \mathcal{O}\left(\frac{1}{m} \right). \tag{119}$$

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

$$-\frac{1}{\sqrt{m}} \leq \boldsymbol{\alpha}^\top \boldsymbol{\gamma}_n \leq \frac{1}{\sqrt{m}}. \tag{120}$$

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

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n\right| \gtrsim \sqrt{\frac{q\log N}{mN}}\right) \leq N^{-q}. \tag{121}$$

Observe that this can be written as

$$\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\alpha}^\top\boldsymbol{\gamma}_n = \boldsymbol{\alpha}^\top\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\gamma}_n\right) := \boldsymbol{\alpha}^\top\boldsymbol{\zeta}. \tag{122}$$

Therefore, by Hoeffding's inequality (cf. (121)),

$$\mathbb{P}\left(\left|\boldsymbol{\alpha}^\top\boldsymbol{\zeta}\right| \gtrsim \sqrt{\frac{q\log N}{mN}}\right) \leq N^{-q}. \tag{123}$$

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

$$\|\boldsymbol{\zeta}\|_2 = \sup_{\|\boldsymbol{\alpha}\|_2=1}\boldsymbol{\alpha}^\top\boldsymbol{\zeta}. \tag{124}$$

We apply an ε-cover argument to obtain

$$\begin{aligned}\sup_{\|\boldsymbol{\alpha}\|_2=1}\boldsymbol{\alpha}^\top\boldsymbol{\zeta} &\leq \frac{1}{1-\varepsilon}\max_{\boldsymbol{\alpha}\in\mathcal{C}_\varepsilon}\boldsymbol{\alpha}^\top\boldsymbol{\zeta}\\ &\leq 2\max_{\boldsymbol{\alpha}\in\mathcal{C}_{1/2}}\boldsymbol{\alpha}^\top\boldsymbol{\zeta}.\end{aligned} \tag{125}$$

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

$$\mathbb{P}\left(\left|\boldsymbol{\alpha}^\top\boldsymbol{\zeta}\right| \gtrsim \sqrt{\frac{q\log N}{mN}}\right) \leq N^{-q}. \tag{126}$$

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

$$\left|\boldsymbol{\alpha}^\top\boldsymbol{\zeta}\right| \gtrsim \sqrt{\frac{q\log N}{mN}} \quad \text{with probability at most } N^{-q}. \tag{127}$$

Then,

$$\max_{\boldsymbol{\alpha}\in\mathcal{C}_{1/2}}\left|\boldsymbol{\alpha}^\top\boldsymbol{\zeta}\right| \gtrsim \sqrt{\frac{q\log N}{mN}} \quad \text{with probability at most } |\mathcal{C}_{1/2}|N^{-q}. \tag{128}$$

Recall that the covering number satisfies

$$|\mathcal{C}_\varepsilon| \leq \left(\frac{3B}{\varepsilon}\right)^d. \tag{129}$$

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

$$|\mathcal{C}_{1/2}| \leq 6^d. \tag{130}$$

We can therefore write

$$\mathbb{P}\left(\|\boldsymbol{\zeta}\|_2 \gtrsim \sqrt{\frac{q\log N}{mN}}\right) \leq 6^d \cdot N^{-q}. \tag{131}$$

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

$$\mathbb{P}\left(\|\boldsymbol{\zeta}\|_2 \gtrsim 2\sqrt{\frac{d\log N}{mN}}\right) \leq \left(\frac{N}{6}\right)^{-d}. \tag{132}$$

Hence, the deviation is bounded with high probability:

$$\|\boldsymbol{\zeta}\|_2 > \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \quad \text{with probability at most } \mathcal{O}(N^{-d}). \tag{133}$$

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

$$\left\|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\gamma}_n\right\|_2 > \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{134}$$

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

$$\left\|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\gamma}_n\right\|_2 \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{135}$$

Using the identities

$$-\frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\gamma}_n, \quad -\frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} = \mathbb{E}\boldsymbol{\gamma}_n, \tag{136}$$

we conclude that, with high probability,

$$\left\|\left(-\frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}\right) - \left(-\frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}\right)\right\|_2 = \left\|\frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} - \frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}\right\|_2 \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{137}$$

Using the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\left|\left\langle \frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} - \frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+\right\rangle\right| &\leq \left\|\frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} - \frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}\right\|_2 \cdot \|\boldsymbol{o}_+\|_2 \\
&= \left\|\frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}} - \frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}\right\|_2 \quad (\text{since } \|\boldsymbol{o}_+\|_2 = 1) \\
&\leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).
\end{aligned} \tag{138}$$

Therefore, we obtain

$$\begin{aligned}
\left\langle -\frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+\right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) &\leq \left\langle -\frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+\right\rangle \\
&\leq \left\langle -\frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+\right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).
\end{aligned} \tag{139}$$

By pairing (111) with the given the conditions on $\boldsymbol{w}_\Delta$ in (101), we can write

$$\frac{1}{\sqrt{m}L}\cdot\sigma(p_1)\left[1+(1-\sigma(q_1))\cdot(1-\sigma(q_2))\cdot(1-\sigma(q_3))^{L_2^+-L_1^+-2}\right] \leq \left\langle -\frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+\right\rangle \tag{140}$$

and

$$\left\langle -\frac{\partial\mathcal{L}}{\partial \boldsymbol{W}^{(t)}_{O(i,\cdot)}}, \boldsymbol{o}_+\right\rangle \leq \frac{1}{\sqrt{m}L}\cdot\sigma(q_1)\left[1+(1-\sigma(p_1))\cdot(1-\sigma(p_2))\cdot(1-\sigma(p_3))^{L_2^+-L_1^+-2}\right]. \tag{141}$$

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

$$\frac{1}{\sqrt{m}L} \cdot \sigma(p_1) \left[1 + (1-\sigma(q_1)) \cdot (1-\sigma(q_2)) \cdot (1-\sigma(q_3))^{L_2^+ - L_1^+ - 2}\right] - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) \leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \tag{142}$$

$$\text{and} \quad \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \frac{1}{\sqrt{m}L} \cdot \sigma(q_1) \left[1 + (1-\sigma(p_1)) \cdot (1-\sigma(p_2)) \cdot (1-\sigma(p_3))^{L_2^+ - L_1^+ - 2}\right] + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \tag{143}$$

This concludes the proof of (9) and (10) in Lemma 1.

To obtain $\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle$, 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 $\boldsymbol{W}_{O(i,\cdot)}^{(t)} \boldsymbol{o}_- > 0$,

$$\begin{aligned}
&\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle \\
&= \frac{1}{\sqrt{m}L} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-) \left[2 + \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right) \right. \\
&\left. \cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^- - L_1^- - 2}\right].
\end{aligned} \tag{144}$$

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

$$\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle = 0. \tag{145}$$

From (100), We know that

$$\begin{aligned}
\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle \\
&\quad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle.
\end{aligned} \tag{146}$$

Hence, combining both cases, we conclude

$$\begin{aligned}
&-\frac{1}{\sqrt{m}L} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-) \left[2 + \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right) \right. \\
&\left. \cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^- - L_1^- - 2}\right] \\
&\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq 0.
\end{aligned} \tag{147}$$

From (137), similar to (139), we can write

$$\begin{aligned}
&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\
&\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \\
&\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \qquad (148)
\end{aligned}$$

Hence,

$$\begin{aligned}
&-\frac{1}{\sqrt{m}L}\cdot\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\left[2+\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\cdot\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\right. \\
&\quad\left.\cdot\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^- - L_1^- - 2}\right] - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\
&\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \qquad (149)
\end{aligned}$$

This concludes the proof of (11) and (12) in Lemma 1.

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

$$\begin{aligned}
\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i\cdot\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot\boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_j \right\rangle \\
&\quad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i\cdot\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot\boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_j \right\rangle \\
&:= \langle I_1, \boldsymbol{o}_j\rangle - \langle I_2, \boldsymbol{o}_j\rangle. \qquad (150)
\end{aligned}$$

Because $\boldsymbol{o}_j$ for $j \neq 1, 2$ is identical in both I_1 and I_2, $\langle I_1, \boldsymbol{o}_j\rangle - \langle I_2, \boldsymbol{o}_j\rangle = 0$. Hence, $\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle = 0$. From (137), similar to (139), we can write

$$\begin{aligned}
&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\
&\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \\
&\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \qquad (151)
\end{aligned}$$

Therefore,

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \text{ for } j \neq 1, 2. \qquad (152)$$

This concludes the proof of (13) in Lemma 1.

□

F.2 Proof of Lemma 2

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

$$\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{o}_+ \leq 0. \tag{153}$$

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

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle. \tag{154}$$

The gradient is given in (99). We only need to consider the cases where $\left\langle \boldsymbol{y}_l^{(n)}, \boldsymbol{o}_+ \right\rangle > 0$. However, since $\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{o}_+ \leq 0$, we have

$$\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) = 0. \tag{155}$$

$$\begin{aligned}
\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\
&\quad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\
&= 0.
\end{aligned} \tag{156}$$

We know by (139),

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{157}$$

Hence,

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{158}$$

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

$$\begin{aligned}
&\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle \\
&= \frac{1}{\sqrt{m}L}\cdot \sigma\left({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-\right)\left[2 + \left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\right. \\
&\quad \left.\cdot \left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^- - L_1^- - 2}\right].
\end{aligned} \tag{159}$$

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

$$\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle = 0. \tag{160}$$

From (100), We know that

$$\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle = \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle. \tag{161}$$

Hence, combining both cases, we conclude

$$\begin{aligned} &-\frac{1}{\sqrt{m}L}\cdot\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\left[2+\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\right. \\ &\quad\left.\cdot\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^- - L_1^- - 2}\right] \\ &\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq 0. \end{aligned} \tag{162}$$

From (137), similar to (139), we can write

$$\begin{aligned} &\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\ &\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \\ &\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \end{aligned} \tag{163}$$

Hence,

$$\begin{aligned} &-\frac{1}{\sqrt{m}L}\cdot\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\left[2+\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\cdot\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\right. \\ &\quad\left.\cdot\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^- - L_1^- - 2}\right] - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\ &\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \end{aligned} \tag{164}$$

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

$$\begin{aligned} \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_j \right\rangle \\ &\quad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_j \right\rangle \\ &:= \langle I_1, \boldsymbol{o}_j\rangle - \langle I_2, \boldsymbol{o}_j\rangle. \end{aligned} \tag{165}$$

Because $\boldsymbol{o}_j$ for $j \neq 1, 2$ is identical in both I_1 and I_2, $\langle I_1, \boldsymbol{o}_j \rangle - \langle I_2, \boldsymbol{o}_j \rangle = 0$. Hence, $\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle = 0$. From (137), similar to (139), we can write

$$\begin{aligned}
&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) \\
&\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \\
&\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \qquad (166)
\end{aligned}$$

Therefore,

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) \text{ for } j \neq 1, 2. \qquad (167)$$

□

F.3 Proof of Lemma 3

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

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{W}_{O(i,\cdot)}} &= \frac{\partial \ell}{\partial F(\boldsymbol{X}^{(n)})} \cdot \frac{\partial F(\boldsymbol{X}^{(n)})}{\partial \boldsymbol{W}_{O(i,\cdot)}} \\
&= -\frac{z^{(n)}}{L} \sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}. \qquad (168)
\end{aligned}$$

If we consider the gradient for the population loss,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}} &= -\mathbb{E}\left[\frac{z^{(n)}}{L} \sum_{l=1}^{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right] \qquad (169) \\
&= -\mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right] \\
&\quad + \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right]. \qquad (170)
\end{aligned}$$

We are given that

$$p_1 \leq \left\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_- \right\rangle \leq q_1, \quad p_2 \leq \left\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+ \right\rangle \leq q_2, \quad \text{and} \quad p_3 \leq \left\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_j \right\rangle \leq q_3 \quad \text{for } j \neq 1, 2. \qquad (171)$$

The Mamba output can be written as

$$\boldsymbol{y}_l(t) = \sum_{s=1}^{l} \left(\prod_{j=s+1}^{l} \left(1 - \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{x}_j)\right) \right) \cdot \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{x}_s) \cdot (\boldsymbol{x}_s^\top \boldsymbol{x}_l) \boldsymbol{x}_s \qquad (172)$$

We have to consider 4 cases.

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

$$\boldsymbol{x}_s = \boldsymbol{x}_l = \boldsymbol{o}_- \qquad (173)$$

$$\left\langle \mathbb{E}\boldsymbol{y}_{L_1^+}, \boldsymbol{o}_- \right\rangle = \sigma(\boldsymbol{w}_\Delta^{(t)^\top} \boldsymbol{o}_-) \geq \sigma(p_1) = \frac{1}{1 + e^{-p_1}} \qquad (174)$$

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

$$\left\langle \mathbb{E}\boldsymbol{y}_{L_2^-, L_2^-}, \boldsymbol{o}_- \right\rangle = \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-) \tag{175}$$

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

$$\begin{aligned}\left\langle \mathbb{E}\boldsymbol{y}_{L_2^-, L_1^-}, \boldsymbol{o}_- \right\rangle = &\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right) \\ &\cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^- - L_1^- - 2} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\end{aligned} \tag{176}$$

Combining (175) and (176), we obtain

$$\begin{aligned}\left\langle \mathbb{E}\boldsymbol{y}_{L_2^-}, \boldsymbol{o}_- \right\rangle = &\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-) \qquad (177)\\ &+ \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right) \\ &\cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^- - L_1^- - 2} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-). \qquad (178)\end{aligned}$$

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 $i \in \mathcal{U}(0)$, $v_i = -\frac{1}{\sqrt{m}}$. For $i \in \mathcal{U}(0)$, and $z^{(n)} = -1$, we have

$$\begin{aligned}&\left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle \\ &= \frac{1}{\sqrt{m}L} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\left[2 + \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\right. \\ &\left. \cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^- - L_1^- - 2}\right].\end{aligned} \tag{179}$$

For $z = +1$, when $l = s = L^-$, $\boldsymbol{x}_s = \boldsymbol{x}_l = \boldsymbol{o}_-$. Therefore, we obtain

$$\langle \mathbb{E}\boldsymbol{y}_{L^-}, \boldsymbol{o}_- \rangle = \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-) \geq \sigma(p_1) \tag{180}$$

$$\left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle = \frac{1}{\sqrt{m}L} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-) \tag{181}$$

Therfore, combining (179) and (181),

$$\begin{aligned}\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i\cdot\phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big)\cdot\boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle \\ &\quad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i\cdot\phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big)\cdot\boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle \\ &= \frac{1}{\sqrt{m}L}\cdot\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top\boldsymbol{o}_-)\left[1+\Big(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top\boldsymbol{o}_-)\Big)\Big(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top\boldsymbol{o}_+)\Big)\right. \\ &\quad \left.\cdot\Big(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top\boldsymbol{o}_j)\Big)^{L_2^- - L_1^- - 2}\right].\end{aligned} \tag{182}$$

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

$$\boldsymbol{\gamma}_n = \frac{z^{(n)}}{L}\sum_{l=1}^{L} v_i\,\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\boldsymbol{y}_l^{(n)}. \tag{183}$$

That is, w.h.p. $1-\mathcal{O}(N^{-d})$, we have $\left\|\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\gamma}_n - \mathbb{E}\boldsymbol{\gamma}_n\right\|_2 \le \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right)$. Hence, we can write

$$\begin{aligned}&\left|\left\langle\left(-\frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{W}_{O(i,\cdot)}^{(t)}}\right) - \left(-\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}_{O(i,\cdot)}^{(t)}}\right), \boldsymbol{o}_-\right\rangle\right| \\ &\quad = \left|\left\langle\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}_{O(i,\cdot)}^{(t)}} - \frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_-\right\rangle\right| \\ &\quad \le \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).\end{aligned} \tag{184}$$

Therefore, we obtain

$$\begin{aligned}\left\langle -\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_-\right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) &\le \left\langle -\frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_-\right\rangle \\ &\le \left\langle -\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_-\right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).\end{aligned} \tag{185}$$

By pairing (182) with the given the conditions on $\boldsymbol{w}_\Delta$ in (171), we can write

$$\frac{1}{\sqrt{m}L}\cdot\sigma(p_1)\left[1+(1-\sigma(q_1))\cdot(1-\sigma(q_2))\cdot(1-\sigma(q_3))^{L_2^- - L_1^- - 2}\right] \le \left\langle -\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_-\right\rangle \tag{186}$$

and $$\left\langle -\frac{\partial\mathcal{L}}{\partial\boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_-\right\rangle \le \frac{1}{\sqrt{m}L}\cdot\sigma(q_1)\left[1+(1-\sigma(p_1))\cdot(1-\sigma(p_2))\cdot(1-\sigma(p_3))^{L_2^- - L_1^- - 2}\right]. \tag{187}$$

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

$$\frac{1}{\sqrt{m}L} \cdot \sigma(p_1) \left[1 + (1-\sigma(q_1)) \cdot (1-\sigma(q_2)) \cdot (1-\sigma(q_3))^{L_2^- - L_1^- - 2}\right] - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) \leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \tag{188}$$

and $$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \frac{1}{\sqrt{m}L} \cdot \sigma(q_1) \left[1 + (1-\sigma(p_1)) \cdot (1-\sigma(p_2)) \cdot (1-\sigma(p_3))^{L_2^- - L_1^- - 2}\right] + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \tag{189}$$

This concludes the proof of (17) and (18) in Lemma 3.

To obtain $\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle$, 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 $\boldsymbol{W}_{O(i,\cdot)}^{(t)} \boldsymbol{o}_+ > 0$,

$$\begin{aligned}
&\left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\Big(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\Big) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\
&= \frac{1}{\sqrt{m}L} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+) \left[2 + \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right) \right. \\
&\quad \left. \cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^+ - L_1^+ - 2}\right].
\end{aligned} \tag{190}$$

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

$$\left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle = 0. \tag{191}$$

From (170), We know that

$$\begin{aligned}
\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\
&\quad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle.
\end{aligned} \tag{192}$$

Hence, combining both cases, we conclude

$$\begin{aligned}
&- \frac{1}{\sqrt{m}L} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+) \left[2 + \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right) \right. \\
&\quad \left. \cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^+ - L_1^+ - 2}\right] \\
&\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq 0.
\end{aligned} \tag{193}$$

From (137), similar to (139), we can write

$$\begin{aligned}
&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\
&\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \\
&\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).
\end{aligned} \tag{194}$$

Hence,

$$\begin{aligned}
-\frac{1}{\sqrt{m}L} \cdot \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+) \Bigg[2 + \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right) \\
\cdot \left(1 - \sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^+ - L_1^+ - 2}\Bigg] - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\
\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).
\end{aligned} \tag{195}$$

This concludes the proof of (19) and (20) in Lemma 3.

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

$$\begin{aligned}
\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_j \right\rangle \\
&\quad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L} \frac{1}{L} v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l^{(n)}\right) \cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_j \right\rangle \\
&:= \langle I_1, \boldsymbol{o}_j \rangle - \langle I_2, \boldsymbol{o}_j \rangle .
\end{aligned} \tag{196}$$

Because $\boldsymbol{o}_j$ for $j \neq 1, 2$ is identical in both I_1 and I_2, $\langle I_1, \boldsymbol{o}_j \rangle - \langle I_2, \boldsymbol{o}_j \rangle = 0$. Hence, $\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle = 0$. From (137), similar to (139), we can write

$$\begin{aligned}
&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\
&\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \\
&\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).
\end{aligned} \tag{197}$$

Therefore,

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \text{ for } j \neq 1, 2. \tag{198}$$

This concludes the proof of (21) in Lemma 3. □

F.4 Proof of Lemma 4

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

$$\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{o}_- \leq 0. \tag{199}$$

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

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle. \tag{200}$$

The gradient is given in (169). We only need to consider the cases where $\left\langle \boldsymbol{y}_l^{(n)}, \boldsymbol{o}_- \right\rangle > 0$. However, since $\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{o}_- \leq 0$, we have

$$\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right) = 0. \tag{201}$$

$$\begin{aligned}
\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle \\
&\quad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_- \right\rangle \\
&= 0.
\end{aligned} \tag{202}$$

We know by (139),

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{203}$$

Hence,

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_- \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \tag{204}$$

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

$$\begin{aligned}
&\left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle \\
&= \frac{1}{\sqrt{m}L}\cdot \sigma\left({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+\right)\left[2 + \left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\right. \\
&\quad \left.\cdot \left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^+ - L_1^+ - 2}\right].
\end{aligned} \tag{205}$$

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

$$\left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle = 0. \tag{206}$$

From (100), We know that

$$\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle = \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_+ \right\rangle. \tag{207}$$

Hence, combining both cases, we conclude

$$\begin{aligned} &-\frac{1}{\sqrt{m}L}\cdot\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\left[2+\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\right. \\ &\quad\left.\cdot\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^+-L_1^+-2}\right] \\ &\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq 0. \end{aligned} \tag{208}$$

From (137), similar to (139), we can write

$$\begin{aligned} &\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\ &\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \\ &\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \end{aligned} \tag{209}$$

Hence,

$$\begin{aligned} &-\frac{1}{\sqrt{m}L}\cdot\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\left[2+\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_+)\right)\cdot\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_-)\right)\right. \\ &\quad\left.\cdot\left(1-\sigma({\boldsymbol{w}_\Delta^{(t)}}^\top \boldsymbol{o}_j)\right)^{L_2^+-L_1^+-2}\right] - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\ &\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_+ \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right). \end{aligned} \tag{210}$$

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

$$\begin{aligned} \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle &= \left\langle \mathbb{E}_{z=+1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_j \right\rangle \\ &\quad - \left\langle \mathbb{E}_{z=-1}\left[\sum_{l=1}^{L}\frac{1}{L}v_i \cdot \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \boldsymbol{y}_l^{(n)}\right], \boldsymbol{o}_j \right\rangle \\ &:= \langle I_1, \boldsymbol{o}_j\rangle - \langle I_2, \boldsymbol{o}_j\rangle. \end{aligned} \tag{211}$$

Because $\boldsymbol{o}_j$ for $j \neq 1,2$ is identical in both I_1 and I_2, $\langle I_1, \boldsymbol{o}_j\rangle - \langle I_2, \boldsymbol{o}_j\rangle = 0$. Hence, $\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle = 0$. From (137), similar to (139), we can write

$$\begin{aligned}
&\left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \\
&\leq \left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \\
&\leq \left\langle -\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle + \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).
\end{aligned} \tag{212}$$

Therefore,

$$\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{W}_{O(i,\cdot)}^{(t)}}, \boldsymbol{o}_j \right\rangle \leq \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right) \text{ for } j \neq 1,2. \tag{213}$$

□

F.5 Proof of Lemma 5

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

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{w}_\Delta} =& -\frac{z^{(n)}}{L}\cdot \sum_{i=1}^{m}\sum_{l=1}^{L} v_i\, \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \sum_{s=1}^{l}\left(\boldsymbol{W}_B^\top \boldsymbol{x}_s^{(n)}\right)^\top \left(\boldsymbol{W}_C^\top \boldsymbol{x}_l^{(n)}\right)\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{x}_s^{(n)}\right) \\
&\cdot \sigma\left(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_s^{(n)}\right)\cdot \prod_{r=s+1}^{l}\left(1-\sigma\left(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_r^{(n)}\right)\right) \\
&\cdot \left[\left(1-\sigma\left(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_s^{(n)}\right)\right)\boldsymbol{x}_s^{(n)} - \sum_{j=s+1}^{l}\left(1-\sigma\left(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_j^{(n)}\right)\right)\boldsymbol{x}_j^{(n)}\right] \\
:=& -\frac{z^{(n)}}{L}\cdot \sum_{i=1}^{m}\sum_{l=1}^{L} v_i\, \phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot \sum_{s=1}^{l}\boldsymbol{I}_{l,s}^{(n)}.
\end{aligned} \tag{214}$$

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

$$\boldsymbol{I}_{l,s}^{(n)} = \beta_{s,s}\cdot \boldsymbol{x}_s^{(n)} - \sum_{j=s+1}^{l}\beta_{s,j}\boldsymbol{x}_j^{(n)}, \tag{215}$$

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

$$\begin{aligned}
\beta_{s,s} =& (\boldsymbol{W}_B^\top \boldsymbol{x}_s^{(n)})^\top (\boldsymbol{W}_C^\top \boldsymbol{x}_l^{(n)})(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{x}_s^{(n)})\sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_s^{(n)}) \\
&\times \left[\prod_{r=s+1}^{l}\left(1-\sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_r^{(n)})\right)\right](1-\sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_s^{(n)})).
\end{aligned} \tag{216}$$

and

$$\begin{aligned}
\beta_{s,j} =& (\boldsymbol{W}_B^\top \boldsymbol{x}_s^{(n)})^\top (\boldsymbol{W}_C^\top \boldsymbol{x}_l^{(n)})(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{x}_s^{(n)})\sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_s^{(n)}) \\
&\times \left[\prod_{r=s+1}^{l}\left(1-\sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_r^{(n)})\right)\right](1-\sigma(\boldsymbol{w}_\Delta^\top \boldsymbol{x}_j^{(n)})).
\end{aligned} \tag{217}$$

If we consider the gradient of the empirical loss,

$$\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta} = -\frac{1}{N}\sum_{n=1}^{N}\frac{z^{(n)}}{L}\cdot\sum_{i=1}^{m}\sum_{l=1}^{L} v_i\,\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\right)\cdot\sum_{s=1}^{l}\boldsymbol{I}_{l,s}^{(n)}. \tag{218}$$

We are given that

$$p_1 \le \langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle \le q_1, \quad \text{and} \quad r_1^* \le \langle {\boldsymbol{W}_{O(i,\cdot)}^{(t+1)}}^\top, \boldsymbol{o}_+\rangle \le s_1^*. \tag{219}$$

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

$$\left\langle -\frac{\partial \ell}{\partial \boldsymbol{w}_\Delta}, \boldsymbol{o}_+ \right\rangle = \frac{z^{(n)}}{L}\sum_{i=1}^{m}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\cdot\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle. \tag{220}$$

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

$$\begin{aligned}
\left\langle -\frac{\partial \hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta}, \boldsymbol{o}_+ \right\rangle &= \frac{1}{N}\sum_{n=1}^{N}\frac{z^{(n)}}{L}\cdot\sum_{i=1}^{m}\sum_{l=1}^{L} v_i\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\cdot\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle \\
&= \frac{1}{N}\sum_{n:z^{(n)}=+1}\frac{1}{L}\left[\sum_{i\in\mathcal{K}_+}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle\right. \\
&\quad \left. + \sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L}\left(-\frac{1}{\sqrt{m}}\right)\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle\right] \\
&\quad + \frac{1}{N}\sum_{n:z^{(n)}=-1}\frac{-1}{L}\left[\sum_{i\in\mathcal{K}_+}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle\right. \\
&\quad \left. + \sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L}\left(-\frac{1}{\sqrt{m}}\right)\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle\right].
\end{aligned} \tag{221}$$

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 (215) that

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle = \beta_{s,s}. \tag{222}$$

Using (216), with $\boldsymbol{W}_B = \boldsymbol{W}_C = I$ and $\boldsymbol{x}_l = \boldsymbol{x}_s = \boldsymbol{o}_+$, we obtain

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle = \beta_{s,s} = \langle {\boldsymbol{W}_{O(i,\cdot)}^{(t+1)}}^\top, \boldsymbol{o}_+\rangle\cdot\sigma(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle)\cdot\left(1-\sigma(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle)\right). \tag{223}$$

Given the conditions in (219), we can write

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle \ge r_1^*\cdot\sigma(p_1)\cdot(1-\sigma(q_1)). \tag{224}$$

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

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

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle \ge r_1^*\cdot\sigma(p_1)\cdot(1-\sigma(q_1)). \tag{225}$$

Case III: $l = L_2^+,\ s = L_1^+$ Comparing (216) with (217), 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,

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = \beta_{s,s} - \beta_{s,j} = 0. \tag{226}$$

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

$$\sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \geq 2r_1^* \cdot \sigma(p_1) \cdot (1 - \sigma(q_1)). \tag{227}$$

We now bound the entire sum over all tokens:

$$\frac{1}{L} \sum_{l=1}^{L} \frac{1}{\sqrt{m}} \phi' \left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l \right) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \geq \frac{1}{L} \sum_{l=1}^{L} \frac{1}{\sqrt{m}} \cdot 1 \cdot 2r_1^* \cdot \sigma(p_1) \cdot (1 - \sigma(q_1)). \tag{228}$$

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

$$\frac{1}{L} \sum_{i \in \mathcal{K}_+} \sum_{l=1}^{L} v_i \phi' \left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l \right) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \geq \frac{2r_1^* \cdot \sigma(p_1) \cdot (1 - \sigma(q_1))}{\sqrt{m}} \cdot \rho_t^+. \tag{229}$$

Next, we consider $z^{(n)} = -1$ for $i \in \mathcal{K}_+$. For $z^{(n)} = -1$, the negative sample contains two $\boldsymbol{o}_-$ and one $\boldsymbol{o}_+$ at positions L_1^-, L_2^-, L^+, respectively.

Let $l = L^+$. Then $\boldsymbol{x}_l = \boldsymbol{o}_+$, and $\boldsymbol{y}_l$ contains a component in the direction of $\boldsymbol{o}_+$. Since $\boldsymbol{W}_{O(i,\cdot)}$ has an $\boldsymbol{o}_+$ component for $i \in \mathcal{K}_+$, this contributes to the gradient.

Let $s = l$, so that

$$\boldsymbol{I}_{l,s}^{(n)} = \beta_{s,s} \cdot \boldsymbol{x}_l. \tag{230}$$

We now seek an upper bound for this contribution. From the initial conditions in (219), we know

$$\langle \boldsymbol{W}_{O(i,\cdot)}^{(t+1)^\top}, \boldsymbol{o}_+ \rangle \leq s_1^*. \tag{231}$$

Hence, we obtain

$$\langle \boldsymbol{I}_{l,s}, \boldsymbol{o}_+ \rangle \leq s_1^* \cdot \sigma(q_1) \cdot (1 - \sigma(p_1)). \tag{232}$$

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

$$\begin{aligned} \frac{1}{L} \sum_{i \in \mathcal{K}_+} \sum_{l=1}^{L} v_i \, \phi' \left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l \right) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle &\leq \frac{s_1^* \cdot \sigma(q_1) \cdot (1 - \sigma(p_1))}{\sqrt{m}} \cdot \frac{m}{2} \\ &= \frac{\sqrt{m} \cdot s_1^* \cdot \sigma(q_1) \cdot (1 - \sigma(p_1))}{2}. \end{aligned} \tag{233}$$

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 (216), we consider $l = L^-$ 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,

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = 0. \tag{234}$$

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

Recall that

$$\left\langle -\frac{\partial \ell}{\partial \boldsymbol{w}_\Delta}, \boldsymbol{o}_+ \right\rangle = \frac{1}{L} \sum_{i=1}^{m} \sum_{l=1}^{L} -\frac{1}{\sqrt{m}} \cdot \phi'(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle. \tag{235}$$

We analyze the inner contributions case by case.

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

Given that

$$\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{o}_+ \leq \delta_1 + \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right) =: c, \tag{236}$$

we obtain

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \leq c \cdot \sigma(q_1) \cdot (1 - \sigma(p_1)). \tag{237}$$

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

This configuration yields the same bound:

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \leq c \cdot \sigma(q_1) \cdot (1 - \sigma(p_1)). \tag{238}$$

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

In this case, the contribution vanishes:

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle = 0. \tag{239}$$

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

$$\sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \leq 2c \cdot \sigma(q_1) \cdot (1 - \sigma(p_1)). \tag{240}$$

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

$$\begin{aligned} \frac{1}{\sqrt{m}L} \sum_{i \in \mathcal{K}_-} \sum_{l=1}^{L} \phi'\left(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l\right) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle &\leq \frac{2c \cdot \sigma(q_1) \cdot (1 - \sigma(p_1))}{\sqrt{m}} \cdot \frac{m}{2} \\ &= \sqrt{m} c \cdot \sigma(q_1) \cdot (1 - \sigma(p_1)). \end{aligned} \tag{241}$$

Therefore, the overall contribution is

$$-\frac{1}{\sqrt{m}L} \sum_{i \in \mathcal{K}_-} \sum_{l=1}^{L} \phi'(\boldsymbol{W}_{O(i,\cdot)} \boldsymbol{y}_l) \sum_{s=1}^{l} \left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+ \right\rangle \geq -\sqrt{m} c \cdot \sigma(q_1) \cdot (1 - \sigma(p_1)). \tag{242}$$

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

$$\left\langle \boldsymbol{W}_{O(i,\cdot)}, \boldsymbol{o}_+ \right\rangle \geq \delta_1 - \mathcal{O}\left(\sqrt{\frac{d \log N}{mN}}\right). \tag{243}$$

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:

$$\frac{1}{L}\sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\phi'\left(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l\right)\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_+\right\rangle \geq 0. \tag{244}$$

We now combine the bounds for the four terms identified in equation (221), corresponding to the contributions from: (i) $\mathcal{K}_+$ with $z^{(n)} = +1$ (229), (ii) $\mathcal{K}_+$ with $z^{(n)} = -1$ (233), (iii) $\mathcal{K}_-$ with $z^{(n)} = +1$ (242), and (iv) $\mathcal{K}_-$ with $z^{(n)} = -1$ (244). 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

$$\begin{aligned}\left\langle -\frac{\partial\hat{\mathcal{L}}}{\partial \boldsymbol{w}_\Delta}, \boldsymbol{o}_+\right\rangle &\geq \frac{1}{2}\left[\frac{2r_1^*\cdot\sigma(p_1)\left(1-\sigma(q_1)\right)}{\sqrt{m}}\cdot\rho_t^+ - \sqrt{m}\cdot c\cdot\sigma(q_1)\left(1-\sigma(p_1)\right)\right.\\ &\qquad\left. -\frac{\sqrt{m}\cdot s_1^*\cdot\sigma(q_1)\left(1-\sigma(p_1)\right)}{2}+0\right]\\ &= \frac{\sigma(p_1)\left(1-\sigma(q_1)\right)r_1^*\cdot\rho_t^+}{\sqrt{m}} - \frac{\sigma(q_1)\left(1-\sigma(p_1)\right)s_1^*\cdot\sqrt{m}}{2} - \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right).\end{aligned} \tag{245}$$

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 $c = \mathcal{O}\left(\sqrt{\frac{d\log N}{mN}}\right)$.

□

F.6 Proof of Lemma 6

Proof. The gradient is given in (214).

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

$$\left\langle -\frac{\partial\ell}{\partial \boldsymbol{w}_\Delta}, \boldsymbol{o}_k\right\rangle = \frac{z^{(n)}}{L}\sum_{i=1}^{m}\sum_{l=1}^{L} v_i\cdot\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k\right\rangle \tag{246}$$

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

$$\begin{aligned}\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k\right\rangle &= -\sum_{j=s+1}^{l}\beta_{s,j}\langle \boldsymbol{x}_j^{(n)}, \boldsymbol{o}_k\rangle\\ &\leq -\beta_{s,s+1} \quad \text{(Assuming W.L.O.G. } \boldsymbol{x}_{s+1}^{(n)} = \boldsymbol{o}_k)\\ &\leq -\langle \boldsymbol{W}_{O(i,\cdot)}^{(t+1)^\top}, \boldsymbol{o}_+\rangle\cdot\sigma\Big(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle\Big)\cdot\Big(1-\sigma\Big(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_+\rangle\Big)\Big)\\ &\quad\cdot\Big(1-\sigma\Big(\langle \boldsymbol{w}_\Delta^{(t)}, \boldsymbol{o}_k\rangle\Big)\Big)^{L_2^+-L_1^+}.\end{aligned} \tag{247}$$

Using the the conditions in (219), we can write

$$\left\langle \boldsymbol{I}_{l,s}^{(n)}, \boldsymbol{o}_k\right\rangle \leq -r_1^*\cdot\sigma(p_1)\cdot\left(1-\sigma(q_1)\right)\cdot\left(1-\sigma(q_2)\right)^{L_2^+-L_1^+}. \tag{248}$$

Hence, we obtain

$$\frac{1}{L}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\cdot\phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big)\sum_{s=1}^{l}\Big\langle\boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\Big\rangle\leq\frac{1}{L}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\cdot 1 \cdot\Big[-r_1^*\cdot\sigma(p_1)\cdot(1-\sigma(q_1))\cdot(1-\sigma(q_2))^{L_2^+-L_1^+}\Big]. \tag{249}$$

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

$$\frac{1}{L}\sum_{i\in\mathcal{K}_+}\sum_{l=1}^{L}v_i\cdot\phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big)\sum_{s=1}^{l}\Big\langle\boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\Big\rangle\leq-\frac{r_1^*}{\sqrt{m}}\cdot\sigma(p_1)\,(1-\sigma(q_1))\cdot(1-\sigma(q_2))^{L_2^+-L_1^+}\cdot\rho_t^+. \tag{250}$$

Similarly for $i\in\mathcal{K}^-$, for $z^{(n)}=-1$, when $l=L_2^-,\ s=L_1^-$ the contribution is significant.

$$\Big\langle\boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\Big\rangle\leq-r_1^*\cdot\sigma(r_2)\cdot(1-\sigma(s_2))\cdot(1-\sigma(q_2))^{L_2^--L_1^-}. \tag{251}$$

Hence, we obtain

$$\frac{1}{L}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\cdot\phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big)\sum_{s=1}^{l}\Big\langle\boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\Big\rangle\leq\frac{1}{L}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\cdot 1 \cdot\Big[-r_1^*\cdot\sigma(r_2)\cdot(1-\sigma(s_2))\cdot(1-\sigma(q_2))^{L_2^--L_1^-}\Big]. \tag{252}$$

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

$$\frac{1}{L}\sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L}v_i\cdot\phi'\Big(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)}\Big)\sum_{s=1}^{l}\Big\langle\boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\Big\rangle\leq-\frac{r_1^*}{\sqrt{m}}\cdot\sigma(r_2)\,(1-\sigma(s_2))\cdot(1-\sigma(q_2))^{L_2^--L_1^-}\cdot\rho_t^-. \tag{253}$$

Putting it together, We know

$$\begin{aligned}\Big\langle-\frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{w}_\Delta},\boldsymbol{o}_k\Big\rangle&=\frac{1}{N}\sum_{n=1}^{N}\frac{z^{(n)}}{L}\cdot\sum_{i=1}^{m}\sum_{l=1}^{L}v_i\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\cdot\sum_{s=1}^{l}\Big\langle\boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\Big\rangle\\&\leq\frac{1}{N}\sum_{n:z^{(n)}=+1}\frac{1}{L}\left[\sum_{i\in\mathcal{K}_+}\sum_{l=1}^{L}\frac{1}{\sqrt{m}}\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\Big\langle\boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\Big\rangle\right]\\&\quad-\frac{1}{N}\sum_{n:z^{(n)}=-1}\frac{1}{L}\left[\sum_{i\in\mathcal{K}_-}\sum_{l=1}^{L}\frac{-1}{\sqrt{m}}\phi'(\boldsymbol{W}_{O(i,\cdot)}\boldsymbol{y}_l^{(n)})\sum_{s=1}^{l}\Big\langle\boldsymbol{I}_{l,s}^{(n)},\boldsymbol{o}_k\Big\rangle\right]\end{aligned} \tag{254}$$

We now combine the bounds for the two terms identified in equation (254), corresponding to the contributions from: (i) $\mathcal{K}_+$ with $z^{(n)}=+1$ (250), and (ii) $\mathcal{K}_-$ with $z^{(n)}=-1$ (253). 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

$$\Big\langle-\frac{\partial\hat{\mathcal{L}}}{\partial\boldsymbol{w}_\Delta^{(t)}},\boldsymbol{o}_k\Big\rangle\leq-\frac{r_1^*}{2\sqrt{m}}\Big[\sigma(p_1)\,(1-\sigma(q_1))\,(1-\sigma(q_2))^{L_2^+-L_1^+}\,\rho_t^+ + \sigma(r_2)\,(1-\sigma(s_2))\,(1-\sigma(q_2))^{L_2^--L_1^-}\,\rho_t^-\Big] \tag{255}$$

□