Transformers as Multi-Task Feature Selectors: Generalization Analysis of In-Context Learning

Hongkang Li[*]
Rensselaer Polytechnic Institute

Meng Wang
Rensselaer Polytechnic Institute

Songtao Lu
IBM Research

Hui Wan
Google

Xiaodong Cui
IBM Research

Pin-Yu Chen
IBM Research

Abstract

Transformer-based large language models have displayed impressive capabilities in the domain of in-context learning, wherein they use multiple input-output pairs to make predictions on unlabeled test data. To lay the theoretical groundwork for in-context learning, we delve into optimization and generalization of a single-head, one-layer Transformer in the context of multi-task learning for classification. Our investigation uncovers that lower sample complexity is associated with increased training-relevant features and reduced noise in prompts, resulting in improved learning performance. The trained model exhibits the mechanism to first attend to demonstrations of training-relevant features and then decode the corresponding label embedding. Furthermore, we delineate the necessary conditions for successful out-of-domain generalization for in-context learning, specifically regarding the relationship between training and testing prompts.

1 Introduction

Transformers now serve as the backbone architecture for a wide range of modern, large-scale foundation models, including prominent language models like GPT-3 [5], PaLM [9], LLaMa [32], as well as versatile visual and multi-modal models such as CLIP [27], DALL-E [28], and GPT-4 [25]. One intriguing capability exhibited by certain large language models (LLMs) is known as [5] "in-context learning (**ICL**)." In other words, these models can accurately predict outcomes for new tasks without fine-tuning their internal parameters. This is achieved simply by providing a small number of testing examples and necessary instructions for the testing query as a prompt.

While Transformer-based LLMs have found diverse applications [21, 22, 23, 35, 34], there is relatively less exploration into the generalization of ICL across multiple tasks using these models. A recent work [12] proposed a framework for studying ICL on linear regression under a supervised learning setup, where the inputs consist of queries augmented with input-output pairs as prompts. This learning process yields a model capable of implementing ICL, serving as a foundation for further investigation. Several theoretical studies have followed this framework. For instance, [1] and [33] have demonstrated that Transformers implement gradient descent during the forward pass. [36] interprets ICL as implicit Bayesian inference and establishes generalization guarantees when the pre-training distribution follows a mixture of Hidden Markov Models (HMMs). Additionally, [19] studies the generalization and stability of ICL by treating the Transformer as an algorithm. Notably, [37] is the only work to simultaneously explore the optimization and generalization of Transformers in the context of ICL, especially with distribution shifts during inference. However, their Transformer architecture employs linear self-attention and linear Multilayer Perceptrons (MLPs), omitting the nonlinear components commonly applied in practical applications.

[*]Work done in an IBM internship

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

To the best of our knowledge, our work represents the first comprehensive theoretical analysis of the optimization dynamics and generalization aspects of ICL in the context of multi-task classification using a nonlinear Transformer. Our approach involves training a simplified one-layer Transformer using data from multiple tasks and subsequently quantifying in- and out-of-domain generalizations on testing data originating from distinct distributions or tasks. We present several technical contributions:

First, this study introduces a unique analytical framework for ICL using shallow Transformers within a multi-task classification setup. Unlike prior works such as [14, 16, 20, 31, 30], which typically focus on single tasks, and ICL research [12, 1, 19, 37] on linear regression, we explore ICL on classification tasks under the multi-task learning setup. We delve into how the quantity of training-relevant features impacts sample complexity, prompt length requirements, and the number of iterations necessary to achieve desired in- and out-of-domain generalization performance.

Second, we conduct an in-depth analysis of how various components of Transformers contribute to multi-task learning. Our analysis uncovers a two-step mechanism within Transformers: first, they enhance the salience of training-relevant features through self-attention, and subsequently, they decode the resulting label embeddings into predictions via the MLP layer. This mechanism extends existing theoretical insights [16, 20, 31, 30], which demonstrate that trained Transformers on single tasks attend to key features through self-attention.

Thirdly, we provide a theoretical characterization of the scenarios in which the trained Transformer performs well with out-of-domain data from previously unseen tasks. Our analysis focuses on a generalized inference setting where we evaluate the model on unseen classification tasks using features that may not have been encountered during the training phase. We outline the sufficient conditions for achieving a desired generalization based on our data model. Furthermore, we show that few-shot generalization becomes attainable when the testing prompt is thoughtfully chosen to encompass a significant portion of the testing-relevant features.

Notation: Let $\boldsymbol{A}_{r_1:r_2,c_1:c_2}$ be the submatrix of a matrix $\boldsymbol{A}$ from rows r_1 to r_2 and columns c_1 to c_2.

2 Problem Formulation

In this work, we study a set of binary classification problems. Consider there are N data samples, each consisting of l input-label pairs, referred to as demonstrations, and one query. Let $\tilde{\boldsymbol{x}}_i^n, i \in [l]$ denote the input of the i-th demonstration of the n-th data. $\tilde{\boldsymbol{x}}_{l+1}^n$ denotes the query of the n-th data. Label $y^n \in \{+1, -1\}$ is a scalar for $n \in [N]$ and $f^{(n)}(\cdot) : \mathbb{R}^{d_\mathcal{X}} \mapsto \mathbb{R}$ represents a task that maps $\tilde{\boldsymbol{x}}_i^n$ to $\{+1, -1\}$. Here, $f^{(n)}$ can be different tasks for different $n \in [N]$. Subsequently, a raw training dataset is $\{\tilde{\boldsymbol{P}}^n, y^n\}_{n=1}^N$ where $\tilde{\boldsymbol{P}}^n = (\tilde{\boldsymbol{x}}_1^n, f^{(n)}(\tilde{\boldsymbol{x}}_1^n), \tilde{\boldsymbol{x}}_2^n, f^{(n)}(\tilde{\boldsymbol{x}}_2^n), \cdots, \tilde{\boldsymbol{x}}_l^n, f^{(n)}(\tilde{\boldsymbol{x}}_l^n), \tilde{\boldsymbol{x}}_{l+1}^n)$. Following [37, 4], we consider the input $\tilde{\boldsymbol{P}}^n$ encoded as

$$\boldsymbol{P}^n = \begin{pmatrix} \boldsymbol{x}_1^n & \boldsymbol{x}_2^n & \cdots & \boldsymbol{x}_l^n & \boldsymbol{x}_{l+1}^n \\ \boldsymbol{y}_1^n & \boldsymbol{y}_2^n & \cdots & \boldsymbol{y}_l^n & \boldsymbol{0} \end{pmatrix} := (\boldsymbol{p}_1^n, \boldsymbol{p}_2^n, \cdots, \boldsymbol{p}_{l+1}^n) \in \mathbb{R}^{(d_\mathcal{X}+d_\mathcal{Y})\times(l+1)} \tag{1}$$

where $\boldsymbol{x}_i^n \in \mathbb{R}^{d_\mathcal{X}}$ and $\boldsymbol{y}_i^n \in \mathbb{R}^{d_\mathcal{Y}}$ for $n \in [N]$ and $i \in [l]$. We use a single-head, one-layer Transformer with a self-attention layer and a two-layer perceptron as the learning network. Mathematically, it can be written as

$$F(\Psi; \boldsymbol{P}^n) = \boldsymbol{a}^\top \text{Relu}(\boldsymbol{W}_O \cdot \textbf{sa}(\Psi, \boldsymbol{p}^n)),\ \textbf{sa}(\Psi, \boldsymbol{p}^n) = \sum_{i=1}^{l} \boldsymbol{W}_V \boldsymbol{p}_i^n \text{softmax}((\boldsymbol{W}_K \boldsymbol{p}_i^n)^\top \boldsymbol{W}_Q \boldsymbol{p}_{l+1}^n), \tag{2}$$

where $\Psi := \boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{m_a\times(d_\mathcal{X}+d_\mathcal{Y})}$, $\boldsymbol{W}_V \in \mathbb{R}^{m_b\times(d_\mathcal{X}+d_\mathcal{Y})}$, $\boldsymbol{W}_O \in \mathbb{R}^{m\times m_b}$, $\boldsymbol{a} \in \mathbb{R}^m$ denotes the model parameters of the Transformer. Typically, $m_a,\ m_b > d_\mathcal{X} + d_\mathcal{Y}$. The training problem minimizes the empirical risk loss $R_N(\Psi)$, which is $\min_\Psi R_N(\Psi) := \frac{1}{N}\sum_{n=1}^N \ell(\Psi; \boldsymbol{P}^n, y^n)$, The loss function is a Hinge loss, i.e., $\ell(\Psi; \boldsymbol{P}^n, y^n) = \max\{0, 1 - y^n \cdot F(\Psi; \boldsymbol{P}^n)\}$.

Training Algorithm The model is trained on a set $\mathcal{T}$ of tasks using mini-batch stochastic gradient descent with step size η under a supervised learning setup. $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$ and $\boldsymbol{W}_V$ are initialized as (non-square) diagonal matrices, where all diagonal entries of $\boldsymbol{W}_V^{(0)}$, and the first $d_\mathcal{X}$ entries of $\boldsymbol{W}_Q^{(0)}$ and $\boldsymbol{W}_K^{(0)}$ are $\delta \in (0, 0.1)$. Each entry of $\boldsymbol{W}_O$ is generated from $\mathcal{N}(0, \xi^2)$ and each entry of $\boldsymbol{a}$ is uniformly sampled from $\{1/m, -1/m\}$. Besides, $\boldsymbol{a}$ does not update during training.

3 Theoretical Results

3.1 Main Theoretical Insights

Before formally presenting the theoretical results, we summarize the main insights as follows.

P1. Sample Complexity for Zero In-Domain Generalization Error. Our findings reveal that, with a sufficiently large model, the sample complexity required to achieve zero in-domain generalization error is proportional to the following key factors, including λ_*^{-1} (where λ_* is the minimum fraction of training-relevant pattern in any training demonstration), $(1-\alpha^{-1})^{-1}$ (where α is the average fraction of training-relevant patterns in prompts), and $(1-\tau M_1)^{-1/2}$ (where τ is the noise level).

P2. Mechanism of Transformers in In-Context Learning. We elucidate the mechanism where Transformers learn multiple tasks in context. Transformers first promote the magnitude of multiple training-relevant features through self-attention to select gold demonstrations. Subsequently, they decode the resulting embeddings using the MLP layer, mainly based on the label part, to make predictions. Such a mechanism differs from existing works [16, 26, 31] on single tasks.

P3. Out-of-domain generalization. Based on our formulated data model, we show that zero generalization error can be achieved under some conditions, even if the testing data follows a different distribution from the training data. We consider any task formed by any two testing-relevant patterns. When the testing-relevant patterns in testing prompts and queries are positive linear combinations of training-relevant features, and the label embeddings match those in the training data, our trained model achieves zero generalization error if the testing prompt is long enough to adequately cover demonstrations with the same testing-relevant features as the testing query.

3.2 Training Data Modeling

To be more specific, let M_1 $(2 \leq M_1 \leq m_a, m_b)$ denote the number of training-relevant patterns represented by $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_{M_1}\}$ and M_2 $(2 \leq M_2 \leq m_a, m_b)$ as the number of training-irrelevant features represented by $\{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \cdots, \boldsymbol{\nu}_{M_2}\}$ in $\mathbb{R}^{d_\mathcal{X}}$. Here, $\boldsymbol{\mu}_i \perp \boldsymbol{\mu}_j$ for $1 \leq i \neq j \leq M_1$, $\boldsymbol{\nu}_i \perp \boldsymbol{\nu}_j$ for $1 \leq i \neq j \leq M_2$, and $\boldsymbol{\mu}_i \perp \boldsymbol{\nu}_j$ for $1 \leq i \leq M_1$ and $1 \leq j \leq M_2$. Also, $\|\boldsymbol{\mu}_i\| = \|\boldsymbol{\nu}_j\| = \beta = \Theta(\log\log M_1)$ for $i \in [M_1],\ j \in [M_2]$. Let $M = M_1 + M_2 \geq M_1^2$. Then, each input embedding $\boldsymbol{x}_i^n$ satisfies that for a certain $j \in [M_1]$ and $k \in [M_2]$,

$$\boldsymbol{x}_i^n = \lambda_i^n \boldsymbol{\mu}_j + \kappa_i^n \boldsymbol{\nu}_k + \boldsymbol{n}_i^n \tag{3}$$

where $\lambda_i^n > 0$ and $|\kappa_i^n| \leq 1$, $i \in [l+1]$, $n \in [N]$, and $\boldsymbol{n}_i^n$ is a bounded noise with $\|\boldsymbol{n}_i^n\| \leq \tau$. Let

$$\lambda_* = \min\{\lambda_i^n, n \in [N], i \in [l+1]\} > 0 \tag{4}$$

We define that $\boldsymbol{y}_i^n \in \{\boldsymbol{q}, -\boldsymbol{q}\}$ for $i \in [l+1]$ and $n \in [N]$, where $\boldsymbol{q}, -\boldsymbol{q}$ represent the label embeddings for labels $+1$ and -1, respectively. $\|\boldsymbol{q}\| = \beta$.

Each task is a binary classification that decides the label based on two training-relevant patterns in input embeddings. Specifically, for a certain task that respectively maps inputs with $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$, $1 \leq a \neq b \leq M_1$, to $+1$ and -1, we have $f^{(n)}(\tilde{\boldsymbol{x}}_i^n)$ (including y^n) is $+1$ (or -1) if $j = a$ (or $j = b$) in (3). If the training-relevant pattern in $\boldsymbol{x}_i^n$ is neither $\boldsymbol{\mu}_a$ nor $\boldsymbol{\mu}_b$, the label of $\boldsymbol{x}_i^n$ is randomly chosen from $\{+1, -1\}$ with equal probability.

The demonstrations of the training prompts are randomly selected following a categorical distribution. For the task dependent on $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ introduced above as an example, the demonstration inputs with $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ are selected with probability $\alpha/2$ where $\alpha = \Theta(1) \in (0, 1)$, while demonstration inputs with other $\boldsymbol{\mu}_j, j \in [M_1]$ are selected with probability $(1-\alpha)/(M_1 - 2)$. Denote the set of demonstrations with the same training-relevant patterns as the query $\boldsymbol{p}_{l+1}^n$ as $\mathcal{N}_*^n$.

3.3 In-Domain Generalization with Sample Complexity Analysis

In-domain generalization means the testing data follows the same distribution as the training data. To avoid the bias in multi-task learning, we need the training tasks to uniformly cover all training-relevant patterns for simplicity, i.e., the number of times where every $\boldsymbol{\mu}_i$ represents labels $+1$ and -1 are equal and are at least 1 among all tasks. Therefore, we have the following lemma on the required number of training tasks.

Lemma 1. *The number of training tasks* $|\mathcal{T}| \geq M_1$.

The following theorem is built on the training and testing on the M_1 required training tasks.

Theorem 1. *(In-Domain Generalization) As long as* $m \geq \epsilon_m^{-2} M_1^2 \log N$ *for* $\epsilon_m \in (0, 1/2)$*, the mini-batch* $B > \Omega(M_1 \log M_1)$*, the length of training prompts satisfies*

$$l_{tr} \geq \Omega(2 \log M/\alpha), \tag{5}$$

then after

$$T = \Theta\big(\sqrt{M}\lambda_*^{-1}\eta^{-1}(1 - \epsilon_m - \tau M_1)^{-1/2} \cdot (C - \alpha^{-1})^{-1}\big) \tag{6}$$

iterations for some $\tau \leq O(1/M_1)$*,* $C > \Omega(1)$ *with* $N = BT$ *samples, with a high probability, the returned model achieves zero generalization error on all training tasks.*

Theorem 1 characterizes the condition on the iterations and sample complexity such that the trained model achieves zero in-domain generalization error. The next section will investigate the mechanism of in-context learning by a one-layer Transformer.

3.4 How Does the Trained Transformer Learn in Context?

We summarize Propositions 1 and 2 to illustrate what the trained self-attention layer and the MLP layer contribute to the prediction.

Proposition 1. *The trained model satisfy that, after* T *iterations,*

$$\|{\boldsymbol{W}_Q^{(T)}}_{1:m_a,1:d_\mathcal{X}}\boldsymbol{\mu}_j\|,\ \|{\boldsymbol{W}_K^{(T)}}_{1:m_a,1:d_\mathcal{X}}\boldsymbol{\mu}_j\| \geq \Theta(\sqrt{\log M}) \text{ for } j \in [M_1]. \tag{7}$$

$$\|{\boldsymbol{W}_Q^{(T)}}_{1:m_a,1:d_\mathcal{X}}\boldsymbol{\nu}_l\|,\ \|{\boldsymbol{W}_K^{(T)}}_{1:m_a,1:d_\mathcal{X}}\boldsymbol{\nu}_l\| \leq \Theta(1) \text{ for } l \in [M_2]. \tag{8}$$

For any training data $\boldsymbol{P}^n$ *and* $C > 1$*, at a sublinear rate of* $O(1/T)$*,*

$$\sum_{s \in \mathcal{N}_*^n} \text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(T)}}^\top \boldsymbol{W}_Q^{(T)} \boldsymbol{p}_{l+1}^n) \rightarrow 1 - \Theta(1/M^C). \tag{9}$$

Proposition 2. *For a constant fraction of* $i \in [m]$*, we have*

$$ {\boldsymbol{W}_O^{(T)}}_{i,d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}} \textbf{sa}(\Psi^{(T)}, \boldsymbol{P}^n)_{d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}} > {\boldsymbol{W}_O^{(T)}}_{i,1:d_\mathcal{X}} \textbf{sa}(\Psi^{(T)}, \boldsymbol{P}^n)_{1:d_\mathcal{X}}. \tag{10}$$

For other i*,* $\|{\boldsymbol{W}_O^{(T)}}_{i,1:m_b} \textbf{sa}(\Psi^{(T)}, \boldsymbol{P}^n)\| \leq O(\xi)$*.*

Proposition 1 indicates that the returned self-attention layer promotes the magnitude of the training-relevant patterns from $\Theta(\log\log M)$ to $\Theta(\sqrt{\log M})$ and maintains the training-irrelevant features close to the initialization. Proposition 2 states that the MLP layer decodes the obtained feature by the self-attention layer with a high weight on the embedding of the label part.

Such a mechanism is discovered for multi-task learning with Transformers for the first time. Figure 3.4 verifies these two propositions with a one-layer Transformer defined in (2).

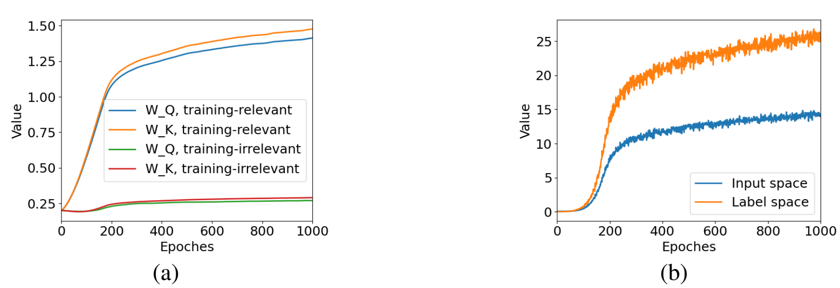

Figure 1: (a) The average value of $\|{\boldsymbol{W}_Q^{(T)}}_{1:m_a,1:d_\mathcal{X}}\boldsymbol{\mu}_j\|$, $\|{\boldsymbol{W}_K^{(T)}}_{1:m_a,1:d_\mathcal{X}}\boldsymbol{\mu}_j\|$, $\|{\boldsymbol{W}_Q^{(T)}}_{1:m_a,1:d_\mathcal{X}}\boldsymbol{\nu}_l\|$, $\|{\boldsymbol{W}_K^{(T)}}_{1:m_a,1:d_\mathcal{X}}\boldsymbol{\nu}_l\|$, for $j \in [M_1]$ and $l \in [M_2]$. (b) The growth of the MLP layer output before ReLU. The blue curve means the output contribution from the feature embeddings. The orange curve refers to the output contribution from the label embeddings.

3.5 Can the Model Generalize to Out-of-Domain Data and Unseen Tasks?

Similar to the data assumptions we make for the training data, we define that $\{\boldsymbol{\mu}'_1, \boldsymbol{\mu}'_2, \cdots, \boldsymbol{\mu}'_{M'_1}, \boldsymbol{\nu}'_1, \boldsymbol{\nu}'_2, \cdots, \boldsymbol{\nu}'_{M_2}\}$ form another orthonormal basis, where $\boldsymbol{\mu}'_j$ and $\boldsymbol{\nu}'_j$ are testing-relevant and testing-irrelevant patterns, respectively. Each input embedding of the test demonstration $\boldsymbol{x}_i^n$ such that

$$\boldsymbol{x}_i^n = \lambda_i^n \boldsymbol{\mu}'_j + \kappa_i^n \boldsymbol{\nu}'_k + \boldsymbol{o}_i^n \tag{11}$$

where $\|\boldsymbol{o}_i^n\| \leq \tau$. All testing tasks are binary classification problems dependent on two certain testing-relevant patterns. The formulation of a testing prompt mirrors that of a training prompt. In this setup, the inputs involving testing-relevant patterns produce labels of either $+1$ or -1 depending on the particular task. Conversely, the inputs of testing-irrelevant patterns result in labels randomly selected from $\{+1, -1\}$ by equal probability. We maintain the label embedding for $\boldsymbol{y}_i^n$ as either $\boldsymbol{q}$ or $-\boldsymbol{q}$ throughout the testing tasks. Each testing data $\boldsymbol{P}^n = (\boldsymbol{p}_1^n, \cdots, \boldsymbol{p}_l^n, \boldsymbol{p}_{l+1}^n)$ is defined as training data in (1) given testing demonstration and label embeddings described above. The demonstrations for testing are randomly selected, following a categorical distribution with a parameter α' on the inputs of the testing-relevant patterns, where the set of demonstrations with the same testing-relevant patterns as the query $\boldsymbol{p}_{l+1}^n$ is $\mathcal{N}_*^n$. Then, we have the following result.

Theorem 2. *(Out-of-Domain Generalization) As long as any $\boldsymbol{\mu}'_j \in \{\sum_{i=1}^{M_1} k_i \boldsymbol{\mu}_i | k_i \geq 0\}$ with $M'_1 \leq M_1$, $\boldsymbol{\nu}'_j \in \mathcal{R}^{d_{\mathcal{X}}} \backslash span\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_{M_1}\}$, and the length of the testing prompt satisfies $l_{ts} \geq 2/\alpha'$, then with high probability, the model learned with training data achieves zero generalization error.*

Corollary 1. *For any testing data $\boldsymbol{P}^n$,*

$$\sum_{s \in \mathcal{N}_*^n} softmax({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(T)}}^\top \boldsymbol{W}_Q^{(T)} \boldsymbol{p}_{l+1}^n) \geq 1 - \Theta(1/M) \tag{12}$$

Remark 1. *Theorem 2 indicates that a one-layer Transformer can generalize well even in the presence of distribution shifts between the training and testing data on the unseen binary classification tasks. The conditions for this favorable generalization encompass the following: (1) the testing-relevant patterns are linear combinations of training-irrelevant patterns with non-negative coefficients; (2) The label embeddings of testing and training prompts are the same, i.e., either $\boldsymbol{q}$ or $-\boldsymbol{q}$; (3) the testing prompt is long enough to include demonstrations involving testing-relevant patterns. With these conditions, Corollary 1 indicates that, despite distribution shift, the attention weights of testing data also concentrate on tokens of testing-relevant patterns as training data does in Proposition 1*

The success of out-of-domain generalization can be understood at a high level by considering the properties of the trained model. The trained self-attention layer can perform demonstration selection based on training-relevant patterns. Hence, the learned parameters enable similarity measurement between out-of-domain testing queries and demonstrations, given that testing-relevant patterns can be represented by training-relevant patterns. Consequently, when provided with the same label embedding, the model can still make accurate predictions.

4 Conclusion and Future Works

This paper studies both optimization and generalization of a one-layer Transformer implementing ICL for multi-task classification. We theoretically analyze the impact of the prompt length, the number of iterations, and sample complexity on the performance of the Transformer for ICL. Additionally, we investigate the conditions essential for successful out-of-domain generalization. Future research directions include exploring generation tasks using more practical Transformer architectures and conducting comparative studies on variants of ICL.

References

[1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pages 6155–6166, 2019.

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

[4] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

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

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

[7] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10836–10846, 2019.

[8] Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. A generalized neural tangent kernel analysis for two-layer neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[10] Amit Daniely and Eran Malach. Learning parities with neural networks. *Advances in Neural Information Processing Systems*, 33:20356–20365, 2020.

[11] Haoyu Fu, Yuejie Chi, and Yingbin Liang. Guaranteed recovery of one-hidden-layer neural networks via cross entropy. *IEEE Transactions on Signal Processing*, 68:3225–3235, 2020.

[12] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

[13] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

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

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

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

[17] Hongkang Li, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. Generalization guarantee of training graph convolutional networks with graph topology sampling. In *International Conference on Machine Learning*, pages 13014–13051. PMLR, 2022.

[18] Hongkang Li, Shuai Zhang, and Meng Wang. Learning and generalization of one-hidden-layer neural networks, going beyond standard gaussian data. In *2022 56th Annual Conference on Information Sciences and Systems (CISS)*, pages 37–42. IEEE, 2022.

[19] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, 2023.

[20] Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. *arXiv preprint arXiv:2303.04245*, 2023.

[21] Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, 2022.

[22] Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, 2022.

[23] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, 2022.

[24] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[25] OpenAI. Gpt-4 technical report. *OpenAI*, 2023.

[26] Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. *arXiv preprint arXiv:2306.03435*, 2023.

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

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[29] Zhenmei Shi, Junyi Wei, and Yingyu Liang. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. In *International Conference on Learning Representations*, 2021.

[30] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *CoRR*, 2023.

[31] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *arXiv preprint arXiv:2305.16380*, 2023.

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

[33] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR, 2023.

[34] Hui Wan, Hongkang Li, Songtao Lu, Xiaodong Cui, and Marina Danilevsky. How can context help? exploring joint retrieval of passage and personalized context. *arXiv preprint arXiv:2308.13760*, 2023.

[35] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. *ACL*, 2023.

[36] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2021.

[37] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

[38] Shuai Zhang, Hongkang Li, Meng Wang, Miao Liu, Pin-Yu Chen, Songtao Lu, Sijia Liu, Keerthiram Murugesan, and Subhajit Chaudhury. On the convergence and sample complexity analysis of deep q-networks with ϵ-greedy exploration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[39] 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. In *The Eleventh International Conference on Learning Representations*, 2023.

[40] Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen, and Jinjun Xiong. Fast learning of graph neural networks with guaranteed generalizability: One-hidden-layer case. In *International Conference on Machine Learning*, pages 11268–11277. PMLR, 2020.

[41] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4140–4149, 2017.

[42] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

A Proof of the main theorems

We first provide several key lemms for the proof of the main theorems.

Lemma 2. *(Multiplicative Chernoff bounds, Theorem D.4 of [24]) Let $X_1, \cdots, \boldsymbol{X}_m$ be independent random variables drawn according to some distribution $\mathcal{D}$ with mean p and support included in $[0,1]$. Then, for any $\gamma \in [0, \frac{1}{p}-1]$, the following inequality holds for $\hat{p} = \frac{1}{m}\sum_{i=1}^m X_i$:*

$$\Pr(\hat{p} \geq (1+\gamma)p) \leq e^{-\frac{mp\gamma^2}{3}}. \tag{13}$$

$$\Pr(\hat{p} \leq (1-\gamma)p) \leq e^{-\frac{mp\gamma^2}{2}}. \tag{14}$$

Lemma 3. *When $t \geq \Omega(1)$, we have that for $i \in \cup_{l=1}^{M_1}\mathcal{W}_l(t) \cup \mathcal{U}_l(t)$,*

$$\Big\|\eta\frac{1}{B}\sum_{b=0}^{t-1}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}[d_\mathcal{X}+1 : d_\mathcal{X}+d_\mathcal{Y}]\Big\| = \Theta(\delta\beta^2/a), \tag{15}$$

while

$$\Big\|\eta\frac{1}{B}\sum_{b=0}^{t-1}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}[1 : d_\mathcal{X}]\Big\| = \Theta(\delta\beta^2\lambda_*/a). \tag{16}$$

For $i \notin \cup_{l=1}^{M_1}\mathcal{W}_l(t) \cup \mathcal{U}_l(t)$, we can obtain

$$\Big\|\eta\frac{1}{B}\sum_{b=0}^{t-1}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}[d_\mathcal{X}+d_\mathcal{Y} : m_b]\Big\| \lesssim \eta\sqrt{\frac{1}{B}}\cdot\frac{\beta^2}{a}. \tag{17}$$

Lemma 4. *For any $j, l \in [M_1]$, $l \neq j$*

$$\begin{aligned}&(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{\mu}_j^\top, \boldsymbol{q}^\top)^\top\\ \gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^4,\end{aligned} \tag{18}$$

$$\begin{aligned}&\Big|(\boldsymbol{\mu}_l^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)^\top\Big|\\ \lesssim&\eta\frac{1}{BM_1}\sum_{b=0}^{t_0-1}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\frac{\zeta_i\delta(1-\gamma_b)\gamma_b(1+\tau)^2\beta^4}{M_1},\end{aligned} \tag{19}$$

$$\begin{aligned}&\Big\|\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{\mu}_j^\top, \boldsymbol{q}^\top)^\top\Big\|\\ \gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^2,\end{aligned} \tag{20}$$

$$\begin{aligned}&(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{\mu}_j^\top, \boldsymbol{q}^\top)^\top\\ \gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^4,\end{aligned} \tag{21}$$

$$\begin{aligned}&\Big|(\boldsymbol{\mu}_l^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_K}\Big|_{t=0}(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)^\top\Big|\\ \lesssim&\eta\frac{1}{BM_1}\sum_{b=0}^{t_0-1}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\frac{\zeta_i\delta(1-\gamma_b)\gamma_b(1+\tau)^2\beta^4}{M_1}.\end{aligned} \tag{22}$$

$$\begin{aligned}&\Big\|\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{\mu}_j^\top,\boldsymbol{q}^\top)^\top\Big\|\\&\gtrsim\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^2,\end{aligned}\tag{23}$$

For any $j\in[M_2]$,

$$\Big|(\boldsymbol{\nu}_l^\top,\boldsymbol{q}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0+1}(\boldsymbol{\nu}_j^\top,\boldsymbol{0}^\top)^\top\Big|\lesssim\eta t_0\frac{1}{BM}\zeta_i\delta\beta^4\tag{24}$$

$$\Big\|\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0+1}(\boldsymbol{\nu}_j^\top,\boldsymbol{0}^\top)^\top\Big\|\lesssim\eta t_0\frac{1}{BM}\zeta_i\delta\beta^2\tag{25}$$

Lemma 5. *For $\boldsymbol{p}_j^n$ that corresponds to $\boldsymbol{\mu}_j$*

$$\begin{aligned}&\eta\frac{1}{B}\sum_{b=0}^{t_0}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_V^{(t)}}\boldsymbol{p}_j^n\\&=\eta\sum_{b=0}^{t_0}\Big(\sum_{i\in\cup_{l=1}^{M_1}\mathcal{W}_l(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\cup_{l=1}^{M_1}\mathcal{U}_l(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\cup_{l=1}^{M_1}\mathcal{W}_l(b)\cup\mathcal{U}_l(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\Big),\end{aligned}\tag{26}$$

where

$$V_i(b)\lesssim\sqrt{\frac{\log B}{B}}\cdot\frac{\beta^2}{a},\quad i\notin\cup_{l=1}^{M_1}\mathcal{W}_l(b)\cup\mathcal{U}_l(b),\tag{27}$$

and if $\boldsymbol{p}_j^n$ corresponds to $\boldsymbol{q}$,

$$V_i(b)\gtrsim\lambda_*^2(1-2\gamma_t)\beta^2/a,\quad i\in\cup_{l=1}^{M_1}\mathcal{W}_l(b),\tag{28}$$

$$V_i(b)\leq 0,\quad i\in\cup_{l=1}^{M_1}\mathcal{U}_l(b),\tag{29}$$

otherwise,

$$V_i(b)\gtrsim\lambda_*^2(1-2\gamma_t)\beta^2/a,\quad i\in\cup_{l=1}^{M_1}\mathcal{U}_l(b),\tag{30}$$

$$V_i(b)\leq 0,\quad i\in\cup_{l=1}^{M_1}\mathcal{W}_l(b).\tag{31}$$

Lemma 6. *For $i\in\cup_{l=1}^{M_1}\mathcal{W}_l(t)\cup\mathcal{U}_l(t)$,*

$$\begin{aligned}&\eta\frac{1}{B}\sum_{b=0}^{t_0}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{p}_j^{n\top},\boldsymbol{0})^\top\\&\gtrsim\frac{\eta}{Ba}\sum_{b=0}^{t_0}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\delta\beta^2\lambda_*^2+\eta\sum_{c=0}^{b}m(1-\epsilon_m-\tau M_1)\lambda_*^2(1-2\gamma_c)\frac{\beta^2}{a}\cdot\beta)(1-2\gamma_c))\end{aligned}\tag{32}$$

$$\begin{aligned}&\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)}\|\\&\gtrsim\frac{\eta}{Ba}\sum_{b=0}^{t_0}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\delta\beta\lambda_*+\eta\sum_{c=0}^{b}m(1-\epsilon_m-\tau M_1)\lambda_*^2(1-2\gamma_c)\frac{\beta^2}{a}(1-2\gamma_c))\end{aligned}\tag{33}$$

For $i\notin\cup_{l=1}^{M_1}(\mathcal{W}_l(t)\cup\mathcal{U}_l(t))$, we have

$$\eta\frac{1}{B}\sum_{b=0}^{t_0}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{p}_j^{n\top},\boldsymbol{0})^\top\leq\eta\sqrt{\frac{\log Bt_0}{Bt_0}}\frac{\beta^2}{a}\tag{34}$$

Lemma 7. *If the number of neurons m is larger enough such that*

$$m\geq\epsilon_m^{-2}M_1^2\log N,\tag{35}$$

the number of lucky neurons at the initialization $|\mathcal{W}(0)|$, $|\mathcal{U}(0)|$ satisfies

$$|\mathcal{W}(0)|,\ |\mathcal{U}(0)|\geq\frac{m}{16}(1-\epsilon_m-\tau M_1)\tag{36}$$

A.1 Proof of Theorem 1

We first look at the required length of the prompt. Define m_i as the corresponding task-relevant features in the i-th demonstration. Consider the multinomial distribution where the probabilities of selecting $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ are $\alpha/2$ respectively. By the Chernoff bound of Bernoulli distribution in Lemma 2, we can obtain

$$\Pr(\frac{1}{l}\sum_{i=1}^{l}\mathbb{1}[m_i=\boldsymbol{\mu}_a]\leq(1-c)\frac{\alpha}{2})\leq e^{-lc^2\frac{\alpha}{2}}=M_1^{-C}, \tag{37}$$

for some $c\in(0,1)$ and $C>0$. Hence, with a high probability,

$$l\geq\frac{2\log M_1}{\alpha}, \tag{38}$$

By the solution to the Coupon collector's problem, we know that

$$B\geq M_1\log M_1, \tag{39}$$

For $y^n=+1$, we have that for i such that $a_i>0$ but $i\notin\cup_{l\in[M_1]}\mathcal{W}_l(t)$,

$$a_i\text{Relu}(\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n)\text{softmax}((\boldsymbol{W}_K^{(T)}\boldsymbol{p}_s^n)^\top\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{l+1}^n))\geq 0. \tag{40}$$

Furthermore, we have that for $i\in\mathcal{W}_l(t)$ where $l\in[M_1]$,

$$\begin{aligned}
&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n\\
=&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}(\delta(\boldsymbol{p}_s^{n\top},\boldsymbol{0}^\top)^\top+\sum_{b=0}^{t-1}\eta(\sum_{i\in\mathcal{W}(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\mathcal{U}(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\\
&+\sum_{i\notin\mathcal{W}(b)\cup\mathcal{U}(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top)\\
\gtrsim&\delta\cdot\frac{\eta}{Ba}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\delta\beta^2\lambda_*^2+\eta\sum_{c=0}^{b}m(1-\epsilon_m-\tau M_1)\lambda_*^2(1-2\gamma_c)\frac{\beta^2}{a}\cdot\beta(1-2\gamma_c))\\
&+\sum_{b=0}^{T-1}\eta\frac{\eta}{Ba}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\delta\beta\lambda_*+\eta\sum_{c=0}^{b}m(1-\epsilon_m-\tau M_1)\lambda_*^2(1-2\gamma_c)\frac{\beta^2}{a}(1-2\gamma_c))\\
&\cdot\lambda_*^2(1-\gamma_T)\frac{\beta^2}{a}.
\end{aligned} \tag{41}$$

$$\begin{aligned}
&\sum_{i\in\mathcal{W}(t)}a_i\text{Relu}(\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n)\text{softmax}((\boldsymbol{W}_K^{(T)}\boldsymbol{p}_s^n)^\top\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{l+1}^n))\\
\gtrsim&\frac{m}{a}(1-\epsilon_m-\tau M_1)(1-\gamma_T)\cdot\Big(\delta\cdot\frac{\eta}{Ba}\sum_{b=0}^{T}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\lambda_*^2\delta\beta^2+\eta\sum_{c=0}^{b}m(1-\epsilon_m-\tau M_1)\\
&\cdot\lambda_*^2(1-2\gamma_c)\frac{\beta^2}{a}\cdot\beta(1-2\gamma_c))+\sum_{b=0}^{T-1}\eta\frac{\eta}{Ba}\sum_{c=0}^{T-1}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\delta\beta\lambda_*\\
&+\eta\sum_{c=0}^{b}m(1-\epsilon_m-\tau M_1)\lambda_*^2(1-2\gamma_c)\frac{\beta^2}{a}(1-2\gamma_c))\cdot\lambda_*^2(1-\gamma_T)\frac{\beta^2}{a}\Big)
\end{aligned} \tag{42}$$

We next give a bound for γ_T. Note that

$$1-\gamma_T=\sum_{s\in\mathcal{N}_{n_1}^n}\text{softmax}((\boldsymbol{W}_K^{(T)}\boldsymbol{p}_s^n)^\top\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{l+1}^n). \tag{43}$$

When $T=\Theta(M^\omega)$, we have

$$\begin{aligned}&(\boldsymbol{W}_K^{(T)}\boldsymbol{p}_s^n)^\top \boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{l+1}^n\\ \gtrsim&(\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{T-1}\frac{m}{a}(1-\epsilon_m-\tau M)\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^2)^2\\ \gtrsim&\frac{\eta^2}{M_1^2}(\sum_{b=0}^{T-1}\frac{\eta^2 b^2}{a}\gamma_b)^2,\end{aligned}\tag{44}$$

where in the last step, we only consider the term related to T and γ_b. Then,

$$\begin{aligned}&\sum_{s\in\mathcal{N}_{n_1}^n}\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\ \geq&\frac{\sum_{s\in\mathcal{N}_{n_1}^n}e^{\Theta(\delta^2\beta^2)+\frac{\eta^2}{M_1^2}(\sum_{b=0}^{T-1}\frac{\eta^2b^2}{a}\gamma_b)^2}}{\sum_{s\in\mathcal{N}_{n_1}^n}e^{\Theta(\delta^2\beta^2)+\frac{\eta^2}{M_1^2}(\sum_{b=0}^{T-1}\frac{\eta^2b^2}{a}\gamma_b)^2}+\sum_{s\in[l]-\mathcal{N}_{n_1}^n}e^{\delta^2\beta^2(\tau+\kappa)+\frac{\eta^2}{M_1^4}(\sum_{b=0}^{T-1}\frac{\eta^2b^2}{a}\gamma_b)^2}}\\ \geq&1-\frac{1-\alpha}{\alpha}e^{-\frac{\eta^2}{M_1^2}(\sum_{b=0}^{T-1}\frac{\eta^2b^2}{a}\gamma_b)^2}.\end{aligned}\tag{45}$$

Combining with (43), we can derive

$$\gamma_T\leq\frac{1-\alpha}{\alpha}e^{-\frac{\eta^2}{M_1^2}(\sum_{b=0}^{T-1}\frac{\eta^2b^2}{a}\gamma_b)^2}=\frac{1-\alpha}{\alpha}e^{-\frac{\eta^2}{M_1^2}(\sum_{b=0}^{T-2}\frac{\eta^2b^2}{a}\gamma_b)^2}\cdot e^{-\frac{\eta^2}{M_1^2}\gamma_{T-1}\frac{\eta^2(T-1)^2}{a}\cdot 2\sum_{b=0}^{T-1}\frac{\eta^2b^2}{a}\gamma_b}.\tag{46}$$

When T is large, γ_T is approaching zero. Hence, the equality of (84) is close to being achieved, in which case,

$$\gamma_T\approx\frac{1-\alpha}{\alpha}\gamma_{T-1}\cdot e^{-\frac{\eta^2}{M_1^2}\gamma_{T-1}\frac{\eta^2(T-1)^2}{a}\cdot 2\sum_{b=0}^{T-1}\frac{\eta^2b^2}{a}\gamma_b}.\tag{47}$$

We can observe that when $\sum_{b=0}^{t_0-1}\eta^2b^2\gamma_b/a\geq(1-\alpha)/\alpha\cdot\eta^{-1}M_1\sqrt{\log M}$, γ_{t_0} reaches $\Theta(1/M)$. Similarly, when $\sum_{b=0}^{t_0'-1}\eta^2b^2\gamma_b/a\leq(1-\alpha)/\alpha\cdot\eta^{-1}M_1\sqrt{\log C}$ for some $C>1$, $\gamma_{t_0'}$ is still $\Theta(1)$, which indicates $t_0'\leq C\eta^{-1}((1-\alpha)/\alpha\cdot MM_1\sqrt{\log C})^{\frac{1}{3}}$. Since we require $M\geq M_1^2$, we have $T\geq\Theta(\eta^{-1}((1-\alpha)/\alpha\cdot MM_1\sqrt{\log C})^{\frac{1}{3}})$. Therefore, we can conclude that $\gamma_T=\Theta(1/M)$. Then, for some large $C>1$,

$$\begin{aligned}&\sum_{i\in\mathcal{W}(t)}a_i\text{Relu}(\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n)\text{softmax}((\boldsymbol{W}_K^{(T)}\boldsymbol{p}_s^n)^\top\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{l+1}^n))\\ \gtrsim&\frac{m}{a}(1-\epsilon_m-\tau M_1)(1-\frac{1}{M})\cdot\Big(\frac{\eta T}{a}(1-\frac{1}{C}\alpha^{-1})+\frac{\eta^2T^2}{a}(1-\frac{1}{C}\alpha^{-1})^2\lambda_*^2\\ &\cdot(1-\epsilon_m-\tau M_1)+\frac{\eta^2T^2}{a}((1-\frac{1}{C}\alpha^{-1})^2+(1-\frac{1}{C}\alpha^{-1})^3\eta T(1-\epsilon_m-\tau M_1)\lambda_*^2)\lambda_*^2\frac{1}{a}\Big).\end{aligned}\tag{48}$$

We next look at i where $a_i<0$. If $i\in\mathcal{U}_l(t)$ where $l\in[M_1]$, we have that for s such that the y-embedding of $\boldsymbol{p}_s^n$ is $\boldsymbol{q}$, the summation of corresponding softmax value is $1-\gamma_T$. Furthermore,

$$\begin{aligned}&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n\\ \lesssim&-\sum_{b=0}^{T-1}\eta\frac{\eta}{Ba}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\delta\beta\lambda_*^2\\ &-\eta m(1-\epsilon_m-\tau M_1)\lambda_*^2(1-\gamma_b)\frac{\beta^2}{a})(1-2\gamma_b))\cdot\lambda_*^2(1-\gamma_T)\frac{\beta^2}{a}.\end{aligned}\tag{49}$$

Hence,

$$\text{Relu}(\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n)\text{softmax}((\boldsymbol{W}_K^{(T)}\boldsymbol{p}_s^n)^\top\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{l+1}^n))=0.\tag{50}$$

If $i \notin \mathcal{W}(T) \cup \mathcal{U}(T)$, we have,

$$\begin{aligned}
&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n\\
\lesssim&\eta\sqrt{\frac{\log BT}{BT}}\frac{\beta}{a}\cdot\eta T\lambda_*^2(1-\gamma_t)\frac{\beta^2}{a}\|\boldsymbol{W}_{O_{(j,\cdot)}}^{(T)}\|\\
\lesssim&\eta\cdot\frac{1}{\sqrt{M}},
\end{aligned}\tag{51}$$

where the last step holds when $\eta T=\Theta(\sqrt{M})$ and $\|\boldsymbol{W}_{O_{(j,\cdot)}}^{(T)}\|=\Theta(1)$ by its lower bound. Since the further computation of $F(\boldsymbol{p}_{l+1}^n)$ is by subtraction between terms related to the lower bound and upper bound of $\|\boldsymbol{W}_{O_{(j,\cdot)}}^{(T)}\|$, the final lower bound of $F(\boldsymbol{p}_{l+1}^n)$ is based on the lower bound of $\|\boldsymbol{W}_{O_{(j,\cdot)}}^{(T)}\|$. Then, combining (40), (48), (50), and (51), we can derive

$$\begin{aligned}
&F(\boldsymbol{p}_{l+1}^n)\\
\gtrsim&\frac{m}{a}(1-\epsilon_m-\tau M_1)(1-\frac{1}{M})\cdot\frac{\eta^2T^2}{a}(1-\frac{1}{C}\alpha^{-1})^2\lambda_*^2\\
\geq&1.
\end{aligned}\tag{52}$$

Therefore, as long as

$$T=\Theta(\frac{\sqrt{M}\lambda_*^{-1}\eta^{-1}}{\sqrt{(1-\epsilon_m-\tau M_1)}}\cdot\frac{C}{C-\alpha^{-1}}),\tag{53}$$

for some large $C>1$, we can obtain

$$F(\boldsymbol{p}_{l+1}^n)>1.\tag{54}$$

Hence, $\omega=1/2$. Combining (112), we have

$$BT\gtrsim(\frac{M_1\cdot\frac{1}{M}}{\eta^{-1}M_1\sqrt{\log M}})^2.\tag{55}$$

We can conclude that $B\gtrsim\Theta(M_1\log M_1)$.
Similarly, we can derive that for $y^n=-1$,

$$F(\boldsymbol{p}_{l+1}^n)<-1.\tag{56}$$

Hence, for all $n\in[N]$,

$$\text{Loss}(\tilde{\boldsymbol{P}}^n,y^n)=0.\tag{57}$$

We also have

$$\mathbb{E}_{(\tilde{\boldsymbol{P}}^n,y^n)\sim\mathcal{D}}[\text{Loss}(\tilde{\boldsymbol{P}}^n,y^n)]=0.\tag{58}$$

with the conditions of sample complexity and the number of iterations.

A.2 Proof of Proposition 1

This is a corollary of Lemma 4. We can derive that

$$\begin{aligned}
&\left\|\boldsymbol{W}_Q^{(T)}[:,0:d_{\mathcal{X}}]\boldsymbol{\mu}_j\right\|\\
\geq&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{T}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^4\\
\geq&\frac{\eta}{M_1}\sum_{b=0}^{T-1}\frac{\eta^2b^2}{a}\gamma_b\\
\geq&\frac{\eta}{M_1}\cdot\frac{1-\alpha}{\alpha}\cdot\eta^{-1}M_1\sqrt{\log M_1}\\
\gtrsim&\sqrt{\log M_1},
\end{aligned}\tag{59}$$

where the second to last step follows the derivation of γ_T in proving Theorem 1. Similarly,

$$\left\|\boldsymbol{W}_K^{(T)}[:,0:d_\mathcal{X}]\boldsymbol{\mu}_j\right\| \gtrsim \sqrt{\log M_1}, \tag{60}$$

Meanwhile,

$$\begin{aligned}&\left\|\boldsymbol{W}_Q^{(T)}[:,0:d_\mathcal{X}]\boldsymbol{\nu}_j\right\|\\ \lesssim&\eta T\frac{1}{BM}\zeta_i\delta\beta^2+\delta^2\beta^2\\ \lesssim&\frac{1}{M_1^2\log M_1}+\delta^2\beta^2\\ \lesssim&\Theta(1),\end{aligned} \tag{61}$$

$$\left\|\boldsymbol{W}_K^{(T)}[:,0:d_\mathcal{X}]\boldsymbol{\nu}_j\right\| \lesssim \Theta(1), \tag{62}$$

A.3 Proof of Proposition 2

By (136), we know that the contribution of the label space embedding is more than that of the feature space embedding in the MLP layer for each $\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s$. Since that $\gamma_T \leq 1/M_1$, we have that there exists a constant $C>1$, such that

$$\begin{aligned}&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}[d_\mathcal{X}:d_{\mathcal{X}+d_\mathcal{Y}}]\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)[d_\mathcal{X}:d_{\mathcal{X}+d_\mathcal{Y}}]\\ \geq&C\cdot\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}[0:d_\mathcal{X}]\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)[0:d_\mathcal{X}]\end{aligned} \tag{63}$$

A.4 Proof of Theorem 2

Note that we need at least one demonstration of the same $\boldsymbol{\mu}_l'$ as the query in the testing prompt. Hence, with high probability,

$$l\geq\frac{2}{\alpha'}. \tag{64}$$

Consider ${\boldsymbol{p}_{l+1}^n}'$ such that the label is $+1$. Let $\boldsymbol{\mu}_j'=\sum_{j=1}^{M_1}c_j\boldsymbol{\mu}_j$ where $\sum_{j=1}^{M_1}c_j^2=1$. By Lemma 4, we have that for $s\in\mathcal{N}^n$,

$$\begin{aligned}&(\boldsymbol{W}_K^{(T)}{\boldsymbol{p}_s^n}')^\top\boldsymbol{W}_Q^{(T)}{\boldsymbol{p}_{l+1}^n}'\\ \gtrsim&\sum_{j=1}^{M_1}c_j^2\cdot(\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{T-1}\frac{m}{a}(1-\epsilon_m-\tau M)\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^2)^2\\ &\cdot(1-\frac{1}{\sqrt{M_1}}\cdot\frac{1}{\sqrt{M_1}})\\ \gtrsim&\frac{\eta^2}{M_1^2}(\sum_{b=0}^{T-1}\frac{\eta^2b^2}{a}\gamma_b)^2\\ \gtrsim&(\alpha^{-1}-1)^2\log M.\end{aligned} \tag{65}$$

Therefore,

$$\sum_{s\in\mathcal{N}_{n_1}^n}\text{softmax}((\boldsymbol{W}_K^{(T)}{\boldsymbol{p}_s^n}')^\top(\boldsymbol{W}_Q^{(T)}{\boldsymbol{p}_{l+1}^n}'))\geq1-\Theta(\frac{1}{M}). \tag{66}$$

Meanwhile, we have that for a certain $i \in \mathcal{W}_l(t)$ where $l \in [M_1]$ and $\boldsymbol{p}_s^n$ where the corresponding y-space embedding is $\boldsymbol{q}$,

$$\begin{aligned}
&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^{n\prime}\\
\gtrsim&\Big(\delta\frac{\eta}{Ba}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\delta\beta^2\lambda_*^2+\eta\sum_{c=0}^{b}m(1-\epsilon_m-\tau M_1)\lambda_*^2(1-2\gamma_c)\frac{\beta^2}{a}\cdot\beta(1-2\gamma_c))\\
&+\sum_{b=0}^{T-1}\eta\frac{\eta}{Ba}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\delta\beta\lambda_*+\eta\sum_{c=0}^{b}m(1-\epsilon_m-\tau M_1)\lambda_*^2(1-2\gamma_c)\frac{\beta^2}{a}(1-2\gamma_c))\\
&\cdot\lambda_*^2(1-\gamma_T)\frac{\beta^2}{a}\Big)\cdot(1-\frac{1}{M_1})\\
\gtrsim&(\frac{\sqrt{M}}{M}+1+\frac{M}{M^2}+\frac{M^{\frac{3}{2}}}{M^2})\cdot(1-\frac{1}{M_1})\\
\geq&1-\frac{1}{M_1},
\end{aligned}\tag{67}$$

$$\begin{aligned}
&\sum_{i\in\mathcal{W}_l(t)}a_i\text{Relu}(\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\sum_{s=1}^{l+1}\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^{n\prime}\text{softmax}((\boldsymbol{W}_K^{(T)}\boldsymbol{p}_s^{n\prime})^\top\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{l+1}^{n}{}'))\\
\gtrsim&\frac{m}{a}(1-\epsilon_m-\tau M_1)(1-\gamma_T)(1-\frac{1}{M_1})\cdot\Theta(1)\\
>&1-\frac{1}{M_1}.
\end{aligned}\tag{68}$$

We can similarly derive

$$F(\boldsymbol{p}_{l+1}^{n}{}')>(1-\frac{1}{M_1})\tag{69}$$

by bounding the components where $a_i<0$ following the proof of Theorem 1.
Likewise, for $\boldsymbol{p}_{l+1}^{n}{}'$ such that the label is -1, we can obtain

$$F(\boldsymbol{p}_{l+1}^{n}{}')<-(1-\frac{1}{M_1}).\tag{70}$$

Therefore, as long as $M_1\geq\epsilon^{-1}$,

$$\text{Loss}(\tilde{\boldsymbol{P}}^n,y^n)\leq\epsilon.\tag{71}$$

B Partial proof of key lemmas

B.1 Proof of Lemma 3

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial\boldsymbol{W}_{O_{(i,\cdot)}}}\\
=&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial F(\boldsymbol{p}_{l+1}^n)}\frac{\partial F(\boldsymbol{p}_{l+1}^n)}{\partial\boldsymbol{W}_{O_{(i,\cdot)}}}\\
=&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}(-y^n)a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}(\boldsymbol{p}_s^{n\top}\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\geq0]\\
&\cdot\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}(\boldsymbol{p}_s^{n\top}\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n).
\end{aligned}\tag{72}$$

We have that for $s\in\mathcal{N}_{n_1}^n$, $i\in\mathcal{W}(t)$ and $y^n=1$,

$$\begin{aligned}&\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\ \geq & \frac{e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}}{\sum_{s\in\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n} e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}+\sum_{s\in\mathcal{N}_{n_2}^n-\mathcal{N}_{n_u}^n} e^{\delta^2\beta^2(\tau+\kappa)}},\end{aligned}\tag{73}$$

and for $s\notin\mathcal{N}_{n_1}^n$,

$$\begin{aligned}&\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\ \leq & \frac{e^{\delta^2\beta^2(\tau+\kappa)}}{\sum_{s\in\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n} e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}+\sum_{s\in\mathcal{N}_{n_2}^n-\mathcal{N}_{n_u}^n} e^{\delta^2\beta^2(\tau+\kappa)}}.\end{aligned}\tag{74}$$

Hence, we can obtain that

$$\begin{aligned}&\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\ \geq&(|\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n|\zeta_i\delta e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}-|[l]-\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n|\zeta_i\delta e^{\delta^2\beta^2(\tau+\kappa)}\\ &+|\mathcal{N}_{n_1}^n\cap\mathcal{N}_{n_u}^n|\zeta_i\delta e^{\delta^2\beta^2((1-\lambda_{l+1}^n)(1-\lambda_s^n)+\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}-|([l]-\mathcal{N}_{n_1}^n)\cap\mathcal{N}_{n_u}^n|\\ &\zeta_i\delta e^{\delta^2\beta^2((1-\lambda_{l+1}^n)(1-\lambda_s^n)+\tau+\kappa)})\cdot(|\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n|e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}\\ &+|[l]-\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n|e^{\delta^2\beta^2(\tau+\kappa)})^{-1}\\ >&0,\end{aligned}\tag{75}$$

where $\zeta_i=\|\boldsymbol{W}_{O_{(i,\cdot)}}[d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}]\|$ with high probability. Hence, for $i\in\mathcal{W}_l(t)\cup\mathcal{U}_l(t)$,

$$\eta\frac{1}{B}\sum_{b=0}^{t-1}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{0}^\top,\boldsymbol{q}^\top,\boldsymbol{0}^\top)^\top\gtrsim\eta\sum_{b=0}^{t-1}\delta\beta^2(1-2\gamma_b)/a,\tag{76}$$

$$\eta\frac{1}{B}\sum_{b=0}^{t-1}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_l,\boldsymbol{0}^\top)^\top\gtrsim\eta\sum_{b=0}^{t-1}\delta\beta^2\lambda_*(1-2\gamma_b)/a.\tag{77}$$

For $i\notin\mathcal{W}(t)\cup\mathcal{U}(t)$, we have

$$\eta\frac{1}{B}\sum_{b=0}^{t-1}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}({\boldsymbol{p}_j^n}^\top,\boldsymbol{0}^\top)^\top\lesssim\eta\sqrt{\frac{\log Bt}{Bt}}\frac{\beta^2}{a}.\tag{78}$$

Therefore, when $t\geq\Omega(\eta^{-1})$, we have that for $i\in\mathcal{W}_l(t)\cup\mathcal{U}_l(t)$,

$$\Big\|\eta\frac{1}{B}\sum_{b=0}^{t-1}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}[d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}]\Big\|=\Theta(\delta\beta^2/a),\tag{79}$$

while

$$\Big\|\eta\frac{1}{B}\sum_{b=0}^{t-1}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}[1:d_\mathcal{X}]\Big\|=\Theta(\delta\beta^2\lambda_*/a),\tag{80}$$

For $i\notin\mathcal{W}(t)\cup\mathcal{U}(t)$, we can obtain

$$\Big\|\eta\frac{1}{B}\sum_{b=0}^{t-1}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}[d_\mathcal{X}+d_\mathcal{Y}:m_b]\Big\|\lesssim\eta\sqrt{\frac{1}{B}}\cdot\frac{\beta^2}{a}.\tag{81}$$

B.2 Proof of Lemma 4

We first study the gradient of $\boldsymbol{W}_Q^{(t+1)}$ in part (a) and the gradient of $\boldsymbol{W}_K^{(t+1)}$ in part (b). The proof is derived with a framework of induction combined with Lemma 5 and 6.
(a) From the training loss function, we can obtain

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n;\Psi)}{\partial \boldsymbol{W}_Q}\\
=&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n;\Psi)}{\partial F(\boldsymbol{p}_{l+1}^n)}\frac{\partial F(\boldsymbol{p}_{l+1}^n)}{\partial \boldsymbol{W}_Q}\\
=&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}(-y^n)\sum_{i=1}^m a_i \mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top \boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top \boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\\
&\cdot\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top \boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K(\boldsymbol{p}_s^n-\boldsymbol{p}_r^n){\boldsymbol{p}_{l+1}^n}^\top\Big)\\
=&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-y^n)\sum_{i=1}^m a_i \mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top \boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top \boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top \boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K\boldsymbol{p}_r^n){\boldsymbol{p}_{l+1}}^\top\Big).
\end{aligned}\tag{82}$$

If $t=0$, we have that

$$(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n)^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n=\delta^2{\boldsymbol{p}_s^n}^\top\boldsymbol{p}_{l+1}^n.\tag{83}$$

When $y^n=+1$, let $\boldsymbol{p}_{l+1}^n$ be a noisy version of $\lambda_{l+1}^n\boldsymbol{\mu}_{n_1}+(1-\lambda_{l+1}^n)\boldsymbol{\mu}_{n_u}$ where $n_1\in\{1,2,\cdots,M_1\}$ and $n_u\in\{M_1+1,M_1+2,\cdots,M_1+M_2\}$ and $\lambda_{l+1}^n\in(0,1)$. Let $i\in\mathcal{W}(t)$, $s\in\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n$, then

$$\begin{aligned}
&\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\
\geq& e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}\cdot\Big(\sum_{s\in\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n}e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}\\
&+\sum_{s\in\mathcal{N}_{n_1}^n\cap\mathcal{N}_{n_u}^n}e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n+(1-\lambda_{l+1}^n)(1-\lambda_s^n)-\tau-\kappa)}\\
&+\sum_{s\in\mathcal{N}_{n_2}^n-\mathcal{N}_{n_u}^n}e^{\delta^2\beta^2(\tau+\kappa)}+\sum_{s\in\mathcal{N}_{n_2}^n\cap\mathcal{N}_{n_u}^n}e^{\delta^2\beta^2((1-\lambda_{l+1}^n)(1-\lambda_s^n)+\tau+\kappa)}\Big)^{-1}\\
\gtrsim&\frac{e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}}{\sum_{s\in\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n}e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}+\sum_{s\in\mathcal{N}_{n_2}^n-\mathcal{N}_{n_u}^n}e^{\delta^2\beta^2(\tau+\kappa)}},,
\end{aligned}\tag{84}$$

where the second step is by $\log M_2\geq\delta^2\beta^2$. Similarly, for $s\in\mathcal{N}_{n_1}^n\cap\mathcal{N}_{n_u}^n$,

$$\begin{aligned}
&\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\
\gtrsim&\frac{e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n+(1-\lambda_{l+1}^n)(1-\lambda_s^n)-\tau-\kappa)}}{\sum_{s\in\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n}e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}+\sum_{s\in\mathcal{N}_{n_2}^n-\mathcal{N}_{n_u}^n}e^{\delta^2\beta^2(\tau+\kappa)}}.
\end{aligned}\tag{85}$$

For $s \in \mathcal{N}_{n_2}^n - \mathcal{N}_{n_u}^n$,

$$\begin{aligned}&\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\\ \lesssim&\frac{e^{\delta^2\beta^2(\tau+\kappa)}}{|\mathcal{N}_{n_1}^n - \mathcal{N}_{n_u}^n|e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}+|[l]-\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n|e^{\delta^2\beta^2(\tau+\kappa)}}.\end{aligned} \tag{86}$$

For $s \in \mathcal{N}_{n_2}^n \cap \mathcal{N}_{n_u}^n$,

$$\begin{aligned}&\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\\ \lesssim&\frac{e^{\delta^2\beta^2((1-\lambda_{l+1}^n)(1-\lambda_s^n)+\tau+\kappa)}}{|\mathcal{N}_{n_1}^n - \mathcal{N}_{n_u}^n|e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}+|[l]-\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n|e^{\delta^2\beta^2(\tau+\kappa)}}.\end{aligned} \tag{87}$$

Therefore, since $\log M_1 \leq \delta^2\beta^2$, we can obtain that

$$\begin{aligned}&\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\\ \geq&(|\mathcal{N}_{n_1}^n - \mathcal{N}_{n_u}^n|\zeta_i\delta e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}-|[l]-\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n|\zeta_i\delta e^{\delta^2\beta^2(\tau+\kappa)}\\ &+|\mathcal{N}_{n_1}^n\cap\mathcal{N}_{n_u}^n|\zeta_i\delta e^{\delta^2\beta^2((1-\lambda_{l+1}^n)(1-\lambda_s^n)+\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}-|([l]-\mathcal{N}_{n_1}^n)\cap\mathcal{N}_{n_u}^n|\\ &\zeta_i\delta e^{\delta^2\beta^2((1-\lambda_{l+1}^n)(1-\lambda_s^n)+\tau+\kappa)})\cdot(|\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n|e^{\delta^2\beta^2(\lambda_{l+1}^n\lambda_s^n-\tau-\kappa)}\\ &+|[l]-\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n|e^{\delta^2\beta^2(\tau+\kappa)})^{-1}\\ >&0,\end{aligned} \tag{88}$$

where $\zeta_i = \|\boldsymbol{W}_{O_{(i,\cdot)}}[d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}]\|$. Then we derive

$$\begin{aligned}&\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n\\ =&\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)+\boldsymbol{n}\\ =&\Big(\sum_{r\in\mathcal{N}_{n_1}^n-\mathcal{N}_{n_u}^n}+\sum_{r\in\mathcal{N}_{n_1}^n\cap\mathcal{N}_{n_u}^n}+\sum_{r\in\mathcal{N}_{n_2}^n-\mathcal{N}_{n_u}^n}+\sum_{r\in\mathcal{N}_{n_2}^n-\mathcal{N}_{n_u}^n}\Big)\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n\\ &)(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)+\boldsymbol{n},\end{aligned} \tag{89}$$

for some $\|\boldsymbol{n}\| \leq \tau$. One can observe that

$$\begin{aligned}
&\sum_{s\in\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n - \sum_{r=1}^{l+1}\text{softmax}(\\
&{\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\\
=&\sum_{s\in\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n - (\sum_{r\in\mathcal{N}_{n_1}^n} + \sum_{r\notin\mathcal{N}_{n_1}^n})\text{softmax}(\\
&{\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\\
=&\sum_{r\notin\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\cdot \sum_{s\notin\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n\\
&)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n - \sum_{s\notin\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\\
&\cdot\sum_{r\notin\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n\\
=&\sum_{s\notin\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\cdot \sum_{r\notin\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n\\
&)\boldsymbol{W}_K^{(t)}(\boldsymbol{p}_s^n - \boldsymbol{p}_r^n).
\end{aligned} \tag{90}$$

Hence, denote

$$\sum_{r\in[l]-\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n) = \gamma_t < 1. \tag{91}$$

Since that the $\boldsymbol{x}_\cdot^n$-space latent features of $({\boldsymbol{p}_r^n}^\top, \boldsymbol{0}^\top)^\top$ are orthogonal to $\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n$ for $r\in[l]-\mathcal{N}_{n_1}^n$, we have that for $s\in\mathcal{N}_{n_1}^n$,

$$\begin{aligned}
&\Big|({\boldsymbol{x}_r^n}^\top, \boldsymbol{0}^\top)\sum_{r\in[l]-\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n - \boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\Big| \leq \frac{\gamma_t(1+\tau)\delta\beta^2}{M_1},
\end{aligned} \tag{92}$$

$$\begin{aligned}
&({\boldsymbol{x}_s^n}^\top, \boldsymbol{0}^\top)\sum_{r\in[l]-\mathcal{N}_{n_1}^n} \text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n - \boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n) \geq \gamma_t\lambda_*^2\delta\beta^2(1-\tau).
\end{aligned} \tag{93}$$

Therefore,

$$\begin{aligned}
&(\boldsymbol{x}_{l+1}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s\in\mathcal{N}^n}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n - \sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{l+1}}^\top(\boldsymbol{x}_{l+1}^\top, \boldsymbol{0}^\top)^\top\\
\geq&\zeta_i\delta\lambda_*^4\beta^4(1-\gamma_t)\gamma_t(1-\tau)^2,
\end{aligned} \tag{94}$$

and for $j\in[l]-\mathcal{N}^n$,

$$\begin{aligned}
&(\boldsymbol{x}_j^\top, \boldsymbol{0}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s\in\mathcal{N}^n}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n - \sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{l+1}}^\top(\boldsymbol{x}_{l+1}^\top, \boldsymbol{0}^\top)^\top\\
\leq&\frac{\zeta_i\delta(1-\gamma_t)\gamma_t\lambda_*^4\beta^4(1+\tau)^2}{M_1},
\end{aligned} \tag{95}$$

To deal with $s \in [l] - \mathcal{N}_{n_1}^n$, we compare (84) and (86). We can then derive that for $s \in \mathcal{N}_{n_1}^n$,

$$\begin{aligned}
&(\boldsymbol{x}_{l+1}^\top, \mathbf{0}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{l+1}}^\top(\boldsymbol{x}_{l+1}^\top,\mathbf{0}^\top)^\top\\
\geq&\zeta_i\delta(1-\gamma_t)\gamma_t(1-\tau)^2(1-e^{-\delta^2\beta^2(\lambda_{l+1}^n-2\tau-2\kappa)})\lambda_*^4\beta^4,
\end{aligned} \tag{96}$$

Note that here for the computation of $\boldsymbol{y}_s^n$ space, we consider the majority voting, which enables us to only focus on $y_s^n = y^n$ for $s \in \mathcal{N}^n$.
Similarly, for $j \in [l] - \mathcal{N}_{n_1}^n$,

$$\begin{aligned}
&(\boldsymbol{x}_j^\top, \mathbf{0}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{l+1}}^\top(\boldsymbol{x}_{l+1}^\top,\mathbf{0}^\top)^\top\\
\leq&\frac{\zeta_i\delta^4\beta^4(1-\gamma_t)\gamma_t(1+\tau)^2}{M_1}(1-e^{-\delta^2\beta^2(\lambda_{l+1}^n-2\tau-2\kappa)}).
\end{aligned} \tag{97}$$

If $i \in \mathcal{U}(t)$, since that $y^n = 1$, by the majority voting, the resulting gradient update does not exceed that of $i \in \mathcal{W}(t)$ by magnitude. If $i \notin \mathcal{W}(t) \cup \mathcal{U}(t)$, by the uniform distribution of a_i, we have that,

$$\begin{aligned}
&({\boldsymbol{x}_{l+1}^n}^\top, \mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-y^n)\sum_{i\notin\mathcal{W}_t\cup\mathcal{U}(t)}a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\\
&\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{l+1}}^\top\Big)(\boldsymbol{x}_{l+1}^\top,\mathbf{0}^\top)^\top\\
\leq&\eta\sqrt{\frac{\log B}{B}}\frac{m}{aM_1}\xi\delta\beta,
\end{aligned} \tag{98}$$

and for $j \in [l] - \mathcal{N}_{n_1}^n$,

$$\begin{aligned}
&({\boldsymbol{x}_j^n}^\top, \mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-y^n)\sum_{i\notin\mathcal{W}_t\cup\mathcal{U}(t)}a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\\
&\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{l+1}}^\top\Big)(\boldsymbol{x}_{l+1}^\top,\mathbf{0}^\top)^\top\\
\leq&(1+\tau)\eta\sqrt{\frac{\log B}{B}}\frac{m}{aM_1}\xi\delta\beta,
\end{aligned} \tag{99}$$

Therefore,

$$\begin{aligned}
&({\boldsymbol{x}_{l+1}^n}^\top, \mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-y^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\\
&\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\boldsymbol{p}_{l+1}{}^\top\Big)(\boldsymbol{x}_{l+1}^\top,\mathbf{0}^\top)\\
\geq&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\zeta_i\delta(1-\gamma_t)\gamma_t(1-\tau)^2(1-e^{-\delta^2\beta^2(\lambda_{l+1}^n-2\tau-2\kappa)})\lambda_*^4\beta^4\\
&-\eta\sqrt{\frac{\log B}{B}}\frac{m}{a}\xi\\
\gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\zeta_i\delta(1-\gamma_t)\gamma_t(1-\tau)^2\lambda_*^4\beta^4(1-e^{-\delta^2\beta^2(\lambda_{l+1}^n-2\tau-2\kappa)}),
\end{aligned}\tag{100}$$

as long as

$$B\geq\left(\frac{\xi M_1}{\zeta_i\delta(1-\gamma_t)\gamma_t(1-\tau)^2(1-e^{-\delta^2\beta^2(\lambda_{l+1}^n-2\tau-2\kappa)})}\right)^2.\tag{101}$$

Meanwhile, for $j\in[l]-\mathcal{N}_{n_1}^n$,

$$\begin{aligned}
&\Big|({\boldsymbol{x}_j^n}^\top, \mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-y^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\\
&\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\boldsymbol{p}_{l+1}{}^\top\Big)(\boldsymbol{x}_{l+1}^\top,\mathbf{0}^\top)^\top\Big|\\
\lesssim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\frac{\zeta_i\delta(1-\gamma_t)\gamma_t(1+\tau)^2\lambda_*^4\beta^4}{M_1}(1-e^{-\delta^2\beta^3(\lambda_{l+1}^n-2\tau-2\kappa)}).
\end{aligned}\tag{102}$$

Then, by combining (100) and (102), we have

$$\begin{aligned}
&(\boldsymbol{\mu}_j^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial\boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{\mu}_j^\top,\mathbf{0}^\top)^\top\\
\gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_i\delta(1-\gamma_t)\gamma_t(1-\tau)^2(1-e^{-\delta^2\beta^2(1-2\tau-2\kappa)})\lambda_*\beta^4\\
\gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\zeta_i(1-2\tau)\delta\lambda_*\beta^4(1-\gamma_t)\gamma_t(1-\tau)^2(1-\frac{1}{M_1^{(1-2\tau-2\kappa)}})\\
\gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\zeta_i(1-2\tau)\delta\lambda_*\beta^4(1-\gamma_t)\gamma_t(1-\tau)^2,
\end{aligned}\tag{103}$$

$$\begin{aligned} &\Big|(\boldsymbol{\mu}_l^\top, \mathbf{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{\mu}_j^\top, \mathbf{0}^\top)^\top\Big|\\ \lesssim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\frac{\zeta_i\delta(1-\gamma_t)\gamma_t(1+\tau)^2\beta^3}{M_1}, \end{aligned} \tag{104}$$

where $l\neq j$. Similarly, since that with probability $1/M_2=\Theta(1/M)$ one demonstration contains $\boldsymbol{\nu}_j$, then by Chernoff bounds in Lemma 2,

$$\Pr(\sum_{i=1}^{Bl}\mathbb{1}[\boldsymbol{x}_i^n \text{ contains } \boldsymbol{\nu}_j]\geq(1+\frac{M_2}{Bl}-1)\frac{Bl}{M_2})\leq e^{-Bl\cdot(\frac{M_2}{Bl})^2\cdot\frac{1}{M_2}}=e^{-\frac{M_2}{Bl}}\lesssim e^{-\frac{M}{Bl}} \tag{105}$$

If $Bl\lesssim M_1\log^2 M$, we have that at most one demonstration contains $\boldsymbol{v}_j$ in the whole batch $\mathcal{B}_b$ for any $j\in[M_2]$. Therefore,

$$\Big|(\boldsymbol{\nu}_j^\top, \mathbf{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{\nu}_j^\top, \mathbf{0}^\top)^\top\Big|\lesssim\eta\frac{1}{BM}\zeta_i\delta\beta^4 \tag{106}$$

$$\Big|(\boldsymbol{\nu}_l^\top, \mathbf{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{\nu}_j^\top, \mathbf{0}^\top)^\top\Big|\lesssim\eta\frac{1}{BM}\zeta_i\delta\beta^4 \tag{107}$$

For the $\boldsymbol{y}_{(\cdot)}^n$-space feature, we have

$$\begin{aligned} &\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=0}[:, d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}]\\ =&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-y^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\\ &\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\ &\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\ &\cdot\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\boldsymbol{p}_{l+1}{}^\top\Big)[:, d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}].\\ =&\mathbf{0} \end{aligned} \tag{108}$$

Since that $\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n[d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}]=\pm\boldsymbol{q}$, i.e., the data with the same $\boldsymbol{x}_\cdot^n$-space feature have the opposite labels, we have

$$\begin{aligned} &\Big|\boldsymbol{q}^\top\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-y^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\\ &\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\ &\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\Big)[d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}]\Big|\\ \leq&\eta\sqrt{\frac{\log B}{B}}\frac{m}{a}\zeta_t\delta^3(\frac{1}{M}+\xi), \end{aligned} \tag{109}$$

where the last step comes from the equal probability of two signs and the upper bound of the inner product.
We need

$$B\geq \frac{(\epsilon_y^{-1}\xi M)^2}{\zeta_i^2} \tag{110}$$

to make (109) upper bounded by $\epsilon_y\in(0,1/2)$.
Hence, the conclusion holds when $t=1$. Suppose that the statement also holds when $t=t_0$. When $t=t_0+1$, the gradient update is the same as in (100) and (102). The only difference is the changes in ζ_t and γ_t. Thus, we can obtain

$$\begin{aligned}
&(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0+1}(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)^\top\\
\gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-2\tau)\delta(1-\gamma_b)\gamma_b(1-\tau)^2\lambda_*\beta^4\\
&-\eta\sqrt{\frac{\log Bt_0}{Bt_0}}\frac{m}{aM_1}\gamma_t\zeta_i\delta\\
\gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-2\tau)\delta(1-\gamma_b)\gamma_b(1-\tau)^2\lambda_*\beta^4,
\end{aligned}\tag{111}$$

where the last step holds as long as

$$Bt_0\gtrsim (\frac{\gamma_{t_0}M_1}{\sum_{b=0}^{t_0-1}\gamma_b\frac{\eta^2b^2}{a}\lambda_*\beta^4})^2, \tag{112}$$

$$\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0+1}[:,d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}]=\boldsymbol{0}. \tag{113}$$

We know that $\boldsymbol{W}_Q$ is used for the computation with the $l+1$-th input. Then we have

$$\begin{aligned}
&(\boldsymbol{\mu}_j^\top,\boldsymbol{q}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0+1}(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)^\top\\
\gtrsim&\eta\frac{1}{BM}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-2\tau)\delta(1-\gamma_b)\gamma_b(1-\tau)^2(1-\epsilon_y)\lambda_*\beta^4\\
\gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^4,
\end{aligned}\tag{114}$$

$$\begin{aligned}
&\Big\|\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0+1}(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)^\top\Big\|\\
\gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^2,
\end{aligned}\tag{115}$$

where the last step comes from the basic mathematical computation.
Similarly, for $j\neq l\in[M_1]$,

$$\begin{aligned}
&\Big|(\boldsymbol{\mu}_l^\top,\boldsymbol{q}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)^\top\Big|\\
\lesssim&\eta\frac{1}{BM_1}\sum_{b=0}^{t_0}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\frac{\zeta_i\delta(1-\gamma_b)\gamma_b(1+\tau)^2\beta^4}{M_1},
\end{aligned}\tag{116}$$

$$\begin{aligned}&\Big\|\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)^\top\Big\|\\ \lesssim&\eta\frac{1}{BM_1}\sum_{b=0}^{t_0}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\frac{\zeta_i\delta(1-\gamma_b)\gamma_b(1+\tau)^2\beta^2}{M_1},\end{aligned}\tag{117}$$

Meanwhile, for $j\neq l\in[M_2]$,

$$\Big|(\boldsymbol{\nu}_j^\top,\boldsymbol{q}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0+1}(\boldsymbol{\nu}_j^\top,\boldsymbol{0}^\top)^\top\Big|\lesssim \eta t_0\frac{1}{BM}\zeta_i\delta\beta^4 \tag{118}$$

$$\Big|(\boldsymbol{\nu}_l^\top,\boldsymbol{q}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0+1}(\boldsymbol{\nu}_j^\top,\boldsymbol{0}^\top)^\top\Big|\lesssim \eta t_0\frac{1}{BM}\zeta_i\delta\beta^4 \tag{119}$$

$$\Big\|\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0+1}(\boldsymbol{\nu}_j^\top,\boldsymbol{0}^\top)^\top\Big\|\lesssim \eta t_0\frac{1}{BM}\zeta_i\delta\beta^2 \tag{120}$$

(b) Then we study the updates of $\boldsymbol{W}_K$. We can compute the gradient as

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n,\Psi)}{\partial \boldsymbol{W}_K}\\ =&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-y^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\\ &\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\geq 0]\\ &\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1})\boldsymbol{W}_Q^\top\boldsymbol{p}_{l+1}^n\\ &\cdot(\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\boldsymbol{p}_r^n)^\top\Big).\end{aligned}\tag{121}$$

If we investigate $\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n$, we can tell that the output is a weighed summation of multiple $\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n$. Similarly, the output of $\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n$ is a weighed summation of multiple $\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s$. Given the initialization $\boldsymbol{W}_Q^{(0)}$ and $\boldsymbol{W}_K^{(0)}$, the update of $\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n$ and $\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n$ only contains the contribution from the feature space embeddings at the initialization. Therefore, along further iterations, only feature space embeddings matter.
Following the steps in Part (a), we can obtain

$$\begin{aligned}&(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_K}\Big|_{t=t_0+1}(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)^\top\\ \gtrsim&\eta\frac{1}{BM}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-2\tau)\delta(1-\gamma_b)\gamma_b(1-\tau)^2\lambda_*\beta^4,\end{aligned}\tag{122}$$

$$\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_K}\Big|_{t=t_0+1}[d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y},:]=\boldsymbol{0}, \tag{123}$$

and for $j\neq l\in[M_1]$,

$$\begin{aligned}&(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_K}\Big|_{t=t_0+1}(\boldsymbol{\mu}_j^\top,\boldsymbol{q}^\top)^\top\\ \gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^4.\end{aligned}\tag{124}$$

$$
\begin{aligned}
&\Big\|\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_K}\Big|_{t=t_0+1}(\boldsymbol{\mu}_j^\top,\boldsymbol{q}^\top)^\top\Big\|\\
\gtrsim&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{m}{a}(1-\epsilon_m-\frac{\tau M}{\pi})\zeta_b(1-4\tau-\epsilon_y)\delta(1-\gamma_b)\gamma_b\lambda_*\beta^2.
\end{aligned}\tag{125}
$$

$$
\begin{aligned}
&\Big|(\boldsymbol{\mu}_l^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_K}\Big|_{t=t_0}(\boldsymbol{\mu}_j^\top,\boldsymbol{q}^\top)^\top\Big|\\
\lesssim&\eta\frac{1}{BM_1}\sum_{b=0}^{t_0}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\frac{\zeta_i\delta(1-\gamma_b)\gamma_b(1+\tau)^2\beta^4}{M_1},
\end{aligned}\tag{126}
$$

$$
\begin{aligned}
&\Big\|\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_K}\Big|_{t=t_0}(\boldsymbol{\mu}_j^\top,\boldsymbol{q}^\top)^\top\Big\|\\
\lesssim&\eta\frac{1}{BM_1}\sum_{b=0}^{t_0}\sum_{n\in\mathcal{B}_b}\frac{m}{a}(1-\epsilon_m-\frac{\tau M_1}{\pi})\frac{\zeta_i\delta(1-\gamma_b)\gamma_b(1+\tau)^2\beta^2}{M_1},
\end{aligned}\tag{127}
$$

Meanwhile, for $j\neq l\in[M_2]$,

$$
\Big|(\boldsymbol{\nu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_K}\Big|_{t=t_0+1}(\boldsymbol{\nu}_j^\top,\boldsymbol{q}^\top)^\top\Big|\lesssim\eta t_0\frac{1}{BM}\zeta_i\delta\beta^4\tag{128}
$$

$$
\Big|(\boldsymbol{\nu}_l^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_K}\Big|_{t=t_0+1}(\boldsymbol{\nu}_j^\top,\boldsymbol{q}^\top)^\top\Big|\lesssim\eta t_0\frac{1}{BM}\zeta_i\delta\beta^4\tag{129}
$$

$$
\Big\|\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_K}\Big|_{t=t_0+1}(\boldsymbol{\nu}_j^\top,\boldsymbol{q}^\top)^\top\Big\|\lesssim\eta t_0\frac{1}{BM}\zeta_i\delta\beta^2\tag{130}
$$

B.3 Proof of Lemma 5

$$
\begin{aligned}
&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_V}\\
=&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial F(\boldsymbol{p}_{l+1}^n)}\frac{\partial F(\boldsymbol{p}_{l+1}^n)}{\partial \boldsymbol{W}_V}\\
=&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}(-y^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\geq 0]\\
&\cdot\boldsymbol{W}_{O_{(i,\cdot)}}^\top\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n){\boldsymbol{p}_s^n}^\top.
\end{aligned}\tag{131}
$$

For $\boldsymbol{p}_{l+1}^n$ which corresponds to the task-relevant feature $\boldsymbol{\mu}_a$,

$$
\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n){\boldsymbol{p}_s^n}^\top({\boldsymbol{x}_a^n}^\top,\boldsymbol{q}^\top,\boldsymbol{0}^\top)^\top\gtrsim\lambda_*^2(1-\gamma_t)\cdot 2\beta^2,\tag{132}
$$

$$
\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n){\boldsymbol{p}_s^n}^\top({\boldsymbol{x}_b^n}^\top,\boldsymbol{q}^\top,\boldsymbol{0}^\top)^\top\lesssim\beta^2\gamma_t,\tag{133}
$$

for $\boldsymbol{x}_b^n$ and $\boldsymbol{x}_a^n$ correspond to different task-relevant features. When $t=0$, for all $i\in\mathcal{W}_a(0)$, we have that for $\boldsymbol{p}_{l+1}^n$ that corresponds to $\boldsymbol{\mu}_a$,

$$
\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)>0\tag{134}
$$

Since that $\mathcal{W}_a(t) \subset \mathcal{W}_a(0)$, such conclusion holds when $t \geq 1$. Note that for $i \in \mathcal{W}_e(t)$ where $e \neq a$ and

$$\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}[d_{\mathcal{X}} : d_{\mathcal{X}+d_{\mathcal{Y}}}]\| \geq C \cdot \|\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}[0 : d_{\mathcal{X}}]\|, \tag{135}$$

where $C > 1$, we also have (134) holds. By Gaussian initialization, with high probability, (135) attains equality with some constant C. Hence, by the summation of $\boldsymbol{W}_{O_{(i,\cdot)}}$ in (131), we can obtain that with high probability, when $t \geq \Theta(1)$, for $i \in \cup_{l=1}^{M_1} \mathcal{W}_l(t) = \mathcal{W}(t)$,

$$\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)[d_{\mathcal{X}} : d_{\mathcal{X}+d_{\mathcal{Y}}}]\| \geq C' \cdot \|\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)[0 : d_{\mathcal{X}}]\| \tag{136}$$

for some $C' > 1$. This indicates that as long as t is large enough such that γ_t is trivial (such condition is achievable finally), we have

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s=1}^{l+1} (\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{l+1}^n) > 0 \tag{137}$$

During our analysis, for simplicity, we directly say (137) for $i \in \mathcal{W}(t)$ holds when t is large without characterizing the lower bound of t. This shall only hold in a subset of $\mathcal{W}(t)$, but it does not affect the conclusion of this lemma.
Therefore, for any $\boldsymbol{p}_j^n = ({\boldsymbol{x}_j^n}^\top, {\boldsymbol{y}_j^n}^\top, \boldsymbol{0}^\top)^\top$ where $f^{(n)}(\tilde{\boldsymbol{x}}_j^n) = +1$,

$$\|\boldsymbol{x}_j^n - \lambda_j^n \boldsymbol{\mu}_a - (1 - \lambda_j^n)\boldsymbol{\nu}_b\| \leq \tau, \tag{138}$$

we have

$$\begin{aligned}
&\sum_{i \in \mathcal{W}(t)} {\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top \eta \frac{1}{B} \sum_{l \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_V^{(t)}} \boldsymbol{p}_j^n \\
\gtrsim & \lambda_*^2 (1 - 2\gamma_t)\beta^2 \cdot \frac{\eta}{a} {\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top (\sum_{j \in \mathcal{W}(t)} \boldsymbol{W}_{O_{(j,\cdot)}}^{(t)}) \\
\gtrsim & \lambda_*^2 (1 - 2\gamma_t)\beta^2 \cdot \frac{\eta}{a} {\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top (\sum_{j \in \mathcal{W}(t)} \boldsymbol{W}_{O_{(j,\cdot)}}^{(t)}),
\end{aligned} \tag{139}$$

$$\begin{aligned}
&\eta \frac{1}{B} \sum_{l \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_V^{(t)}} \boldsymbol{p}_j^n \\
=&\eta (\sum_{i \in \mathcal{W}(t)} V_i(t) \boldsymbol{W}_{O_{(i,\cdot)}}^{(t)} + \sum_{i \in \mathcal{U}(t)} V_i(t) \boldsymbol{W}_{O_{(i,\cdot)}}^{(t)} + \sum_{i \notin \mathcal{W}(t) \cup \mathcal{U}(t)} V_i(t) \boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}),
\end{aligned} \tag{140}$$

where

$$V_i(t) \gtrsim \lambda_*^2 (1 - 2\gamma_t)\beta^2 / a, \quad i \in \mathcal{W}(t), \tag{141}$$

$$V_i(t) \lesssim (1 - 1)\beta^2 / a \leq 0, \quad i \in \mathcal{U}(t) \tag{142}$$

$$V_i(t) \lesssim \sqrt{\frac{\log B}{B}} \cdot \frac{\beta^2}{a}, \quad i \notin \mathcal{W}(t) \cup \mathcal{U}(t). \tag{143}$$

We require that

$$\eta t (1 - 2\gamma_t)(\lambda_*^2 - \kappa^2 - \tau)\beta^2 - > 0 \tag{144}$$

Similarly, for any $\boldsymbol{p}_j^n = ({\boldsymbol{x}_j^n}^\top, {\boldsymbol{y}_j^n}^\top, \boldsymbol{0}^\top)^\top$ where $f^{(n)}(\tilde{\boldsymbol{x}}_j^n) = -1$,

$$\|\boldsymbol{x}_j^n - \lambda_j^n \boldsymbol{\mu}_a - (1 - \lambda_j^n)\boldsymbol{\nu}_b\| \leq \tau, \tag{145}$$

we have

$$\begin{aligned}
&\sum_{i \in \mathcal{U}(t)} \boldsymbol{W}_{O_{(i,\cdot)}}^\top \eta \frac{1}{B} \sum_{l \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n; \Psi)}{\partial \boldsymbol{W}_V^{(t)}} \boldsymbol{p}_j^n \\
\gtrsim & \lambda_*^2 (1 - 2\gamma_t)\beta^2 \cdot \frac{\eta}{a} (\sum_{i \in \mathcal{W}(t)} \boldsymbol{W}_{O_{(i,\cdot)}})^2 \\
\gtrsim & \lambda_*^2 (1 - 2\gamma_t)\beta^2 \cdot \frac{\eta}{a} (\sum_{i \in \mathcal{U}(t)} \boldsymbol{W}_{O_{(i,\cdot)}})^2,
\end{aligned} \tag{146}$$

$$\begin{aligned}&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_V^{(t)}}\boldsymbol{p}_j^n\\=&\eta(\sum_{i\in\mathcal{W}(t)}V_i(t)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}+\sum_{i\in\mathcal{U}(t)}V_i(t)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}+\sum_{i\notin\mathcal{W}(t)\cup\mathcal{U}(t)}V_i(t)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}),\end{aligned}\tag{147}$$

where

$$V_i(t)\gtrsim \lambda_*^2(1-2\gamma_t)\beta^2/a,\quad i\in\mathcal{U}(t),\tag{148}$$

$$V_i(t)\lesssim (1-1)\beta^2/a\leq 0,\quad i\in\mathcal{W}(t),\tag{149}$$

$$V_i(t)\lesssim \sqrt{\frac{\log B}{B}}\cdot\frac{\beta^2}{a},\quad i\notin\mathcal{W}(t)\cup\mathcal{U}(t).\tag{150}$$

B.4 Proof of Lemma 6

$$\begin{aligned}&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}\\=&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial F(\boldsymbol{p}_{l+1}^n)}\frac{\partial F(\boldsymbol{p}_{l+1}^n)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}\\=&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}(-y^n)a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)\geq 0]\\&\cdot\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n).\end{aligned}\tag{151}$$

We have that

$$\begin{aligned}&\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n\\=&\delta({\boldsymbol{p}_s^n}^\top,\boldsymbol{0}^\top)^\top+\sum_{b=0}^{t-1}\eta(\sum_{i\in\mathcal{W}(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\mathcal{U}(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}(b)\cup\mathcal{U}(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top.\end{aligned}\tag{152}$$

Consider a certain $\boldsymbol{p}_s^n=({\boldsymbol{x}_s^n}^\top,{\boldsymbol{y}_s^n}^\top,\boldsymbol{0}^\top)^\top$ where $f^{(n)}(\tilde{\boldsymbol{x}}_s^n)=+1$, and

$$\|\boldsymbol{x}_s^n-\lambda_s^n\boldsymbol{\mu}_a-\kappa_s^n\boldsymbol{\nu}_b\|\leq\tau.\tag{153}$$

When $t=1$, we can obtain that for $i\in\mathcal{W}_a(t)$,

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}({\boldsymbol{p}_s^n}^\top,\boldsymbol{0}^\top)^\top\gtrsim\beta.\tag{154}$$

$$\begin{aligned}&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}({\boldsymbol{p}_j^n}^\top,\boldsymbol{0})^\top\\=&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{1}{a}\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{l+1}^n)(\delta{\boldsymbol{p}_s^n}^\top\boldsymbol{p}_j^n+\sum_{b=0}^{t-1}\eta(\sum_{i\in\mathcal{W}(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\\&+\sum_{i\in\mathcal{U}(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}(b)\cup\mathcal{U}(b)}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top({\boldsymbol{p}_j^n}^\top,\boldsymbol{0}^\top)^\top)\\\geq&\frac{\eta}{Ba}\sum_{n\in\mathcal{B}_b}(((1-\gamma_t)2\delta\beta^2-\gamma_t\delta\beta^2)\lambda_*^2+\eta m(1-\epsilon_m-\tau M_1)(\lambda_*^2(1-2\gamma_t)-\sqrt{\frac{\log B}{B}})\frac{\beta^2}{a}\cdot\beta\\&\cdot(1-2\gamma_t))\\\gtrsim&\frac{\eta}{Ba}\sum_{n\in\mathcal{B}_b}((2-3\gamma_t)\delta\beta^2\lambda_*^2+\eta m(1-\epsilon_m-\tau M_1)\lambda_*^2(1-2\gamma_t)\frac{\beta^2}{a}\cdot\beta(1-2\gamma_t)),\end{aligned}\tag{155}$$

as long as $\log M_1 \geq \delta^2\beta^2$ and $B \geq \lambda_*^{-4}$. For $i \in \mathcal{U}_a(t)$, we also have

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{p}_j^{n\top}, \boldsymbol{0})^\top \\
\gtrsim & \frac{\eta}{Ba}\sum_{n\in\mathcal{B}_b}((2-3\gamma_t)\delta\beta^2\lambda_*^2 + \eta m(1-\epsilon_m - \tau M_1)\lambda_*^2(1-2\gamma_t)\frac{\beta^2}{a}\cdot\beta(1-2\gamma_t)).
\end{aligned} \tag{156}$$

if $\boldsymbol{p}_j^n$ corresponds to label -1 in this task. For $i \notin \mathcal{W}_a(t)\cup\mathcal{U}_a(t)$, we have

$$\eta\frac{1}{B}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{p}_j^{n\top}, \boldsymbol{0})^\top \leq \eta\sqrt{\frac{\log B}{B}}\frac{\beta^2}{a}. \tag{157}$$

Suppose that the conclusion holds when $t \leq t_0$. Then when $t = t_0+1$, we have that for $i \in \mathcal{W}_a(t)$,

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{p}_j^{n\top}, \boldsymbol{0})^\top \\
\gtrsim & \frac{\eta}{Ba}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\delta\beta^2\lambda_*^2 + \eta\sum_{c=0}^{b} m(1-\epsilon_m - \tau M_1)\lambda_*^2(1-2\gamma_c)\frac{\beta^2}{a}\cdot\beta)(1-2\gamma_c)),
\end{aligned} \tag{158}$$

$$\begin{aligned}
&\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\| \\
\gtrsim & \frac{\eta}{Ba}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}((2-3\gamma_b)\delta\beta\lambda_* + \eta\sum_{c=0}^{b} m(1-\epsilon_m - \tau M_1)\lambda_*^2(1-2\gamma_c)\frac{\beta^2}{a}(1-2\gamma_c)).
\end{aligned} \tag{159}$$

For $i \notin \mathcal{W}_a(t)\cup\mathcal{U}_a(t)$, we have

$$\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{l\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, y^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{p}_j^{n\top}, \boldsymbol{0})^\top \leq \eta\sqrt{\frac{\log B(t_0+1)}{B(t_0+1)}}\frac{\beta^2}{a}. \tag{160}$$

By the derivation of (137), we have that (158), (159), and (160) holds for $i \in \mathcal{W}(t)$.

C Related works

Theoretical analysis of learning and generalization of neural networks. Some works [41, 11, 40, 18, 38] study the generalization performance following the model recovery framework by probing the local convexity around a ground truth parameter. The neural-tangent-kernel (NTK) analysis [13, 2, 3, 7, 42, 8, 17] considers strongly overparameterized networks to linearize the neural network around the initialization. The generalization performance is independent of the feature distribution. [10, 29, 15, 6, 39, 16] investigate the generalization of neural networks assuming a data model consisting of discriminative patterns and background patterns.

Theoretical study on in-context learning. Existing theoretical works on in-context learning include the expressive power of the introduced parameter [4, 1], the optimization process [33], and the generalization analysis [36, 37, 19]. Most studies concentrate on linear regression tasks on in-context learning.