How Do Nonlinear Transformers Learn and Generalize in In-Context Learning?

Hongkang Li [1] **Meng Wang** [1] **Songtao Lu** [2] **Xiaodong Cui** [2] **Pin-Yu Chen** [2]

Abstract

Transformer-based large language models have displayed impressive in-context learning capabilities, where a pre-trained model can handle new tasks without fine-tuning by simply augmenting the query with some input-output examples from that task. Despite the empirical success, the mechanics of how to train a Transformer to achieve ICL and the corresponding ICL capacity is mostly elusive due to the technical challenges of analyzing the nonconvex training problems resulting from the nonlinear self-attention and nonlinear activation in Transformers. To the best of our knowledge, this paper provides the first theoretical analysis of the training dynamics of Transformers with nonlinear self-attention and nonlinear MLP, together with the ICL generalization capability of the resulting model. Focusing on a group of binary classification tasks, we train Transformers using data from a subset of these tasks and quantify the impact of various factors on the ICL generalization performance on the remaining unseen tasks with and without data distribution shifts. We also analyze how different components in the learned Transformers contribute to the ICL performance. Furthermore, we provide the first theoretical analysis of how model pruning affects ICL performance and prove that proper magnitude-based pruning can have a minimal impact on ICL while reducing inference costs. These theoretical findings are justified through numerical experiments.

[1]Department of Electrical, Computer, and System Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA [2]IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. Correspondence to: Hongkang Li <lih35@rpi.edu>, Meng Wang <wangm7@rpi.edu>, Songtao Lu <songtao@ibm.com>, Xiaodong Cui <cuix@us.ibm.com>, Pin-Yu Chen <pin-yu.chen@ibm.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

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 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), LLaMa (Touvron et al., 2023), as well as versatile visual and multi-modal models such as CLIP (Radford et al., 2021), DALL-E (Ramesh et al., 2021), and GPT-4 (OpenAI, 2023). One intriguing capability exhibited by certain large language models (LLMs) is known as "**in-context learning**" (**ICL**) (Brown et al., 2020). Given a pre-trained model $F(\Psi)$, parameterized by weights Ψ, the conventional approach fine-tunes Ψ separately for each downstream task using data from that task. In contrast, ICL allows $F(\Psi)$ to handle multiple unseen tasks simultaneously without any fine-tuning. Garg et al. (2022) is the first paper to mathematically formulate ICL. Briefly speaking, to predict $f(\boldsymbol{x}_{\text{query}})$ of a query input $\boldsymbol{x}_{\text{query}}$ for a new task represented by the label function f, ICL augments $\boldsymbol{x}_{\text{query}}$ by l example input-output pairs $(\boldsymbol{x}_i, f(\boldsymbol{x}_i))_{i=1}^l$. The resulting so-called *prompt* is sent to the model $F(\Psi)$, and, surprisingly, the model can output a prediction close to $f(x_{\text{query}})$. Thus, ICL is an efficient alternative to the resource-consuming fine-tuning methods. ICL has shown outstanding performance in multiple tasks in practice, including question answering (Liu et al., 2022b; Wu et al., 2023b), natural language inference (Liu et al., 2022a; Wu et al., 2023b), text generation (Brown et al., 2020; Lucy & Bamman, 2021), etc.

In parallel, model pruning (Han et al., 2015; Wen et al., 2016) can reduce the inference cost by removing some weights after training. It has been extensively evaluated in various applications. Among various pruning techniques, such as gradient methods (Molchanov et al., 2016) and reconstruction error minimization (Luo et al., 2017), magnitude-based pruning (Wen et al., 2016) is the most popular approach due to its simplicity and demonstrated promising empirical results. A few recent works (Frantar & Alistarh, 2023; Ma et al., 2023; Sun et al., 2023; Liu et al., 2023) also explore the pruning of LLMs to preserve their ICL capacity while accelerating the inference.

Despite the empirical success of ICL, one fundamental and theoretical question is less investigated, which is:

How can a Transformer be trained to perform ICL and *generalize in and out of domain successfully and efficiently?*

Some recent works attempt to answer this question for linear regression tasks (Li et al., 2023c; Zhang et al., 2023a). Specifically, Li et al. (2023c) investigate the generalization gap and stability of ICL. Zhang et al. (2023a) explore the training and generalization of ICL with Transformers, especially with distribution shifts during inference. Wu et al. (2023a) studies the required number of pre-training tasks for a desirable ICL property. Huang et al. (2023) characterizes the training dynamics using Transformers with softmax attention and linear MLP. However, these results are either built upon simplified Transformer models by ignoring nonlinear self-attention (Zhang et al., 2023a; Wu et al., 2023a) or nonlinear activation in the multilayer perceptron (MLP) (Huang et al., 2023; Zhang et al., 2023a; Wu et al., 2023a) or cannot characterize how to train a model to achieve the desirable ICL capability with distribution-shifted data (Huang et al., 2023; Li et al., 2023c; Wu et al., 2023a) .

1.1. Major Contributions of This Work

To the best of our knowledge, our work is the first theoretical analysis of the training dynamics of Transformers with nonlinear self-attention and nonlinear MLP, together with the ICL generalization capability of the resulting model. Moreover, our paper provides the first theoretical analysis of the impact of model pruning on ICL performance. Focusing on a group of binary classification tasks, we show that training a Transformer using prompts from a subset of these tasks can return a model with the ICL capability to generalize to the rest of these tasks. We provide a quantitative analysis of the required number of training data, iterations, the length of prompts, and the resulting ICL performance. Although our analysis is centered on a simplified single-head and one-layer Transformer with softmax self-attention and ReLU MLP, our theoretical insights shed light on practical architectures. Our major contributions include:

1. **A theoretical characterization of how to train Transformers to enhance their ICL capability.** We consider a data model where input data include both relevant patterns that determine the labels and irrelevant patterns that do not affect the labels. We quantify how the training and the resulting ICL generalization performance are affected by various factors, such as the magnitude of relevant features and the fraction of context examples that contain the same relevant pattern as the new query. In addition to proving the ICL capability of the learned Transformer to generalize to new binary tasks based on the relevant patterns that appear in the training data, we also prove the ICL capability to generalize to tasks based on patterns that are linear combinations of the relevant patterns and are unseen in the training data.

2. **Expand the theoretical understanding of the mechanism of the ICL capability of Transformers.** We prove that when sending a prompt to a properly trained Transformer, the attention weights are concentrated on contexts that share the same relevant pattern as the query. Then, the ReLU MLP layer promotes the label embedding of these examples, thus making the correct prediction for the query. Similar insights have appeared in (Huang et al., 2023). We expand the analysis to Transformers with nonlinear MLP layers and new tasks with a data distribution shift.

3. **Theoretical justification of magnitude-based pruning in preserving ICL.** Based on the characterization of the trained Transformer, our paper also provides the first theoretical analysis of the ICL inference performance when the trained model is pruned by removing neurons in the MLP layer. We show that pruning a set of neurons with a small magnitude has little effect on the generalization while pruning the remaining neurons leads to a large generalization error growing with the pruning rate. To the best of our knowledge, no theoretical analysis exists on how model pruning affects ICL.

1.2. Related Work

Expressive power of ICL Some existing works study the expressive power of Transformers to implement algorithms via ICL. Akyürek et al. (2023); Von Oswald et al. (2023) demonstrate that Transformers conduct gradient descent during the forward pass of Transformers with prompts as inputs. Ahn et al. (2023); Cheng et al. (2023) extend the conclusion to preconditioned and functional gradient descent via ICL. Garg et al. (2022); Bai et al. (2023); Guo et al. (2023) show the existence of Transformers that can implement a broad class of machine learning algorithms in context.

The optimization and generalization of Transformers Beyond in-context learning, there are several other works about the optimization and generalization analysis of fine-tuning or prompt tuning on Transformers. Jelassi et al. (2022); Li et al. (2023a;b); Luo et al. (2024) study the generalization of one-layer Transformer by assuming spatial association or the majority voting of tokens. Li et al. (2023d) delve into how one-layer Transformers learn semantic structure. Oymak et al. (2023) depict the trajectory of prompt tuning of attention networks. Tarzanagh et al. (2023b;a) characterize that the gradient updates of the prompt or weights converge to a max-margin SVM solution. Tian et al. (2023a;b) probe the training dynamics of Transformers for the next token prediction problem given infinitely long sequences.

Theoretical generalization analysis of pruning A few recent works consider analyzing the generalizations performance of model pruning theoretically. For example, Zhang et al. (2021) study the sample complexity of training a pruned network with a given sparse ground truth weight. Yang & Wang (2023) investigate the neural tangent kernel of the pruned model. Zhang et al. (2023c); Yang et al. (2023)

Theoretical Works	Nonlinear Attention	Nonlinear MLP	Training Analysis	Distribution-Shifted Data	Tasks
Li et al. (2023c)	✓	✓			linear regression
Zhang et al. (2023a)			✓	✓	linear regression
Huang et al. (2023)	✓		✓		linear regression
Wu et al. (2023a)			✓		linear regression
Ours	✓	✓	✓	✓	classification

Table 1. Comparison with existing works about training analysis and generalization guarantee of in-context learning

consider the generalization using magnitude pruning under a feature learning framework. However, these works are built on convolutional neural networks, and no theoretical works are for LLM or Transformer-based models.

2. Problem Formulation

This work studies the optimization and generalization of binary classification problems for in-context learning. Consider a query $\boldsymbol{x}_{query}$ and its label z. Define a set of binary classification tasks $\mathcal{T}$, consisting of multiple task functions. The label $z \in \{+1,-1\}$ is mapped from $\boldsymbol{x}_{query} \in \mathbb{R}^{d_{\mathcal{X}}}$ through a task f that is randomly chosen from $\mathcal{T}$, i.e., $z = f(\boldsymbol{x}_{query}) \in \{+1,-1\}, f \in \mathcal{T}$.

2.1. Training to Enhance ICL Capability

Following the framework of training for ICL in (Garg et al., 2022; Akyürek et al., 2023; Bai et al., 2023), we consider the problem of training such that the model has the ICL capability to generalize to new tasks using prompts. The idea is to update the model during the training process using pairs of the constructed prompt, embedded as $\boldsymbol{P}$ for the query $\boldsymbol{x}_{query}$, and its label $f(\boldsymbol{x}_{query})$. We start by formulating $\boldsymbol{P}$ and then introduce the learning model in this section.

Following (Von Oswald et al., 2023; Zhang et al., 2023a; Huang et al., 2023), the prompt embedding $\boldsymbol{P}$ of query $\boldsymbol{x}_{query}$ is formulated as:

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

where the last column of $\boldsymbol{P}$, denoted by $\boldsymbol{p}_{query}$, includes the query $\boldsymbol{x}_{query}$ with padding zeros, and the first l columns are the contexts for $\boldsymbol{x}_{query}$. We respectively call $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$, $i \in [l]$ *context* inputs and outputs, where l is also known as the prompt length. Let Embd$(\cdot)$ be the embedding function of each context output. $\boldsymbol{y}_i \in \mathbb{R}^{d_{\mathcal{Y}}}$ in (1) is defined as $\boldsymbol{y}_i = \text{Embd}(f(\boldsymbol{x}_i))$. Hence, $\boldsymbol{P}$ is a function of f. The first $d_{\mathcal{X}}$ dimensions of $\boldsymbol{p}_i$ are referred to as the feature embedding, while the last $d_{\mathcal{Y}}$ dimensions are called the label embedding.

The **learning model** is a single-head, one-layer Transformer with one self-attention layer and one two-layer perceptron. Mathematically, it can be written as

$$\begin{aligned} F(\Psi; \boldsymbol{P}) &= \boldsymbol{a}^\top \text{Relu}(\boldsymbol{W}_O \sum_{i=1}^{l} \boldsymbol{W}_V \boldsymbol{p}_i \cdot \text{attn}(\Psi; \boldsymbol{P}, i)), \\ \text{attn}(\Psi; \boldsymbol{P}, i) &= \text{softmax}((\boldsymbol{W}_K \boldsymbol{p}_i)^\top \boldsymbol{W}_Q \boldsymbol{p}_{query}), \end{aligned} \tag{2}$$

where $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{m_a\times(d_{\mathcal{X}}+d_{\mathcal{Y}})}$, $\boldsymbol{W}_V \in \mathbb{R}^{m_b\times(d_{\mathcal{X}}+d_{\mathcal{Y}})}$ are the embedding matrices for queries, keys, and values, respectively, and $\boldsymbol{W}_O \in \mathbb{R}^{m\times m_b}$ and $\boldsymbol{a} \in \mathbb{R}^m$ are parameters in the MLP layer. $\Psi := \{\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_O, \boldsymbol{a}\}$ denotes the set of all model weights. Typically, $\min(m_a, m_b) > d_{\mathcal{X}} + d_{\mathcal{Y}}$.

The **training problem** to enhance the ICL capability solves the empirical risk minimization problem,

$$\min_{\Psi} R_N(\Psi) := \frac{1}{N}\sum_{n=1}^{N} \ell(\Psi; \boldsymbol{P}^n, z^n), \tag{3}$$

using N pairs of prompt embedding and label pairs $\{\boldsymbol{P}^n, z^n\}_{n=1}^N$. For the n-th pair, $\boldsymbol{x}_{query}^n$ and the context input $\boldsymbol{x}_i^n$ are all sampled from an unknown distribution $\mathcal{D}$, the task f^n is sampled from $\mathcal{T}$, and $\boldsymbol{P}^n$ is constructed following (1). The loss function is a Hinge loss, i.e., $\ell(\Psi; \boldsymbol{P}^n, z^n) = \max\{0, 1 - z^n \cdot F(\Psi; \boldsymbol{P}^n)\}$, where $F(\Psi; \boldsymbol{P}^n)$ is defined in (2). Let $\mathcal{T}_{tr} = \bigcup_{n=1}^N f^n$ denote the set of tasks that appear in the training samples. Note that $\mathcal{T}_{tr} \subset \mathcal{T}$, and (3) is a *multi-task learning* problem when $|\mathcal{T}_{tr}| > 1$.

2.2. Generalization Evaluation

We define two quantities to evaluate the ICL generalization performance to new tasks as follows.

In-Domain Generalization: If the testing queries are also drawn from $\mathcal{D}$ and all the testing tasks are drawn from $\mathcal{T}$, we call it *in-domain* inference, and the in-domain generalization error is defined as[1]

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{query}\sim\mathcal{D}, f\in\mathcal{T}} [\ell(\Psi; \boldsymbol{P}, z)], \tag{4}$$

where $\boldsymbol{P}$ is defined in (1). Note that the in-domain performance includes the testing performance on *unseen* tasks in

[1]In terms of evaluating generalization on unseen tasks, (4) is almost equivalent to replacing $f \in \mathcal{T}$ with $f \in \mathcal{T}\backslash\mathcal{T}_{tr}$ in the subscript. This is because we later prove that all of our analysis can hold when training on a small fraction of tasks (Condition 3.2). Therefore, an $\mathcal{O}(\epsilon)$ generalization error on $f \in \mathcal{T}$ can indeed reflect an $\mathcal{O}(\epsilon)$ generalization error on $f \in \mathcal{T}\backslash\mathcal{T}_{tr}$

$\mathcal{T}\backslash\mathcal{T}_{tr}$ that do not appear in the training samples.

Out-of-Domain Generalization: Suppose the testing queries $\boldsymbol{x}_{query}$ follow the distribution $\mathcal{D}'$ ($\mathcal{D}' \neq \mathcal{D}$), and the binary classification tasks that map the testing queries to the labels are drawn a set $\mathcal{T}'$ ($\mathcal{T}' \neq \mathcal{T}$). Then, the *out-of-domain* generalization error can be defined as

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{query}\sim\mathcal{D}', f\in\mathcal{T}'}[\ell(\Psi;\boldsymbol{P},z)]. \tag{5}$$

2.3. Training Algorithm

The model is trained using stochastic gradient descent (SGD) with step size η with batch size B, summarized in Algorithm 1 in Appendix C. $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$ and $\boldsymbol{W}_V$ are initialized such that all diagonal entries of $\boldsymbol{W}_V^{(0)}$, and the first $d_{\mathcal{X}}$ diagonal entries of $\boldsymbol{W}_Q^{(0)}$ and $\boldsymbol{W}_K^{(0)}$ are set as δ with $\delta \in (0, 0.2]$, and all other entries are 0. Each entry of $\boldsymbol{W}_O^{(0)}$ is generated from $\mathcal{N}(0,\xi^2)$, $\xi = 1/\sqrt{m}$ and each entry of $\boldsymbol{a}$ is uniformly sampled from $\{1/m, -1/m\}$. Besides, $\boldsymbol{a}$ does not update during training.

2.4. Model Pruning

We also consider the case that the learned model Ψ is pruned to reduce the inference computation. Let $\mathcal{S}\subset[m]$ denote the index set of neurons in the output layer. Pruning neurons in $\mathcal{S}$ correspond to removing the corresponding rows in $\boldsymbol{W}_O$, resulting in the reduced matrix size of $(m-|\mathcal{S}|)\cdot m_b$.

3. Theoretical Results

We first summarize the main insights in Section 3.1. Section 3.2 formally presents our analysis model. Section 3.3 presents the formal theoretical results on the learning performance and the resulting ICL generalization. Section 3.4 provides the theoretical result that magnitude-based pruning on the out layer does not hurt ICL performance.

3.1. Main Theoretical Insights

We consider a class of binary classification tasks where the binary labels in each task are determined by two out of M_1 *in-domain-relevant patterns*. The training data include pairs of prompt embedding and labels from a small subset of these tasks. In-domain generalization evaluates the ICL capability of the learned model on tasks using all possible combinations of these M_1 patterns. Out-of-domain generalization further evaluates the binary classification tasks that are determined by pairs of *out-of-domain-relevant patterns*, which are some linear combinations of these M_1 patterns.

P1. Quantitative Learning Analysis With Guaranteed In- and Out-of-Domain Generalization. We quantitatively prove the learned model achieves desirable generalization in both in-domain and out-of-domain tasks. The required number of training data and iterations are polynomial in β^{-1} and α^{-1}, where β represents the norm of relevant patterns, and α denotes the fraction of context inputs with the same in-domain-relevant pattern as the query. A higher α implies that the context examples offer more information about the query, consequently reducing the sample requirements and expediting the learning process.

P2. Mechanism of Transformers in Implementing ICL. We elucidate the mechanism where the learned Transformers make predictions in- and out-of-domain in context. We quantitatively show that the self-attention layer attends to context examples with relevant patterns of the query task and promotes learning of these relevant patterns. Then, the two-layer perceptron promotes the label embeddings that correspond to these examples so as to predict the label of the query accurately.

P3. Magnitude-Based Pruning Preserves ICL. We quantify the ICL generalization if neurons with the smallest magnitude after training in the MLP layer are removed and prove that the generalization is almost unaffected even when a constant fraction of neurons are removed. In contrast, the generalization error is proved to be at least $\Omega(R)$ when R fraction of neurons with large magnitude are removed.

3.2. The Modeling of Training Data and Tasks

In-Domain Data and Tasks. Consider M_1 *in-domain-relevant* (*IDR*) patterns $\{\boldsymbol{\mu}_j\}_{j=1}^{M_1}$ and M_2 $(=\mathcal{O}(M_1))$ *in-domain-irrelevant* (*IDI*) patterns $\{\boldsymbol{\nu}_k\}_{k=1}^{M_2}$ $(M_1+M_2=d_{\mathcal{X}})$ in $\mathbb{R}^{d_{\mathcal{X}}}$, where these M_1+M_2 patterns are pairwise orthogonal, and $\|\boldsymbol{\mu}_j\| = \|\boldsymbol{\nu}_k\| = \beta \geq 1$ (β is a constant) for $j\in[M_1]$,$k\in[M_2]$. Each in-domain data $\boldsymbol{x}$ drawn from $\mathcal{D}$ is generated by

$$\boldsymbol{x} = \boldsymbol{\mu}_j + \kappa\boldsymbol{\nu}_k, \tag{6}$$

where $j\in[M_1]$ and $k\in[M_2]$ are arbitrarily selected. κ follows a uniform distribution $U(-K,K)$, $K\leq 1/2$. Denote IDR($\boldsymbol{x}$):$=\boldsymbol{\mu}_j$ as the IDR pattern in data $\boldsymbol{x}$. Our data assumption originates from recent feature learning works on deep learning (Allen-Zhu & Li, 2023; Li et al., 2023a; Oymak et al., 2023) for language and vision data. To the best of our knowledge, only (Huang et al., 2023) theoretically analyzes the performance of ICL with softmax attention, assuming all $\boldsymbol{x}$ are orthogonal to each other. Our assumption in (6) is more general than that in (Huang et al., 2023).

Each in-domain task is defined as a binary classification function that decides the label based on two IDR patterns in the query. Specifically,

Definition 3.1. (Definition of in-domain tasks) The in-domain task set $\mathcal{T}$ includes $M_1(M_1-1)$ tasks such that each task $f\in\mathcal{T}$ is defined as

$$f(\boldsymbol{x}) = \begin{cases} +1, & \text{IDR}(\boldsymbol{x}) = \boldsymbol{\mu}_a, \\ -1, & \text{IDR}(\boldsymbol{x}) = \boldsymbol{\mu}_b, \\ \text{random from}\{+1,-1\}, & \text{otherwise}, \end{cases} \tag{7}$$

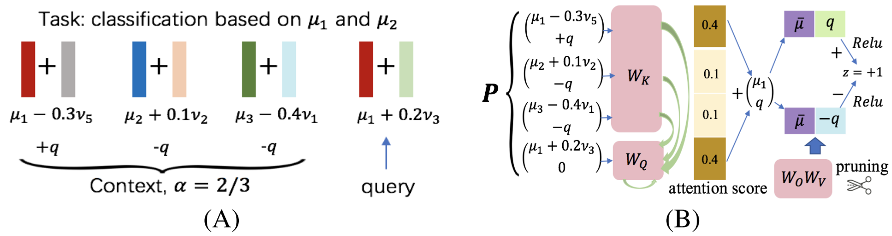


Figure 1. (A) Example of prompt embedding. $l = 3$, $\alpha = 2/3$. (B) The mechanism of a trained Transformer (2) to implement ICL.

where $\boldsymbol{\mu}_a, \boldsymbol{\mu}_b$ are two different patterns in $\{\boldsymbol{\mu}_j\}_{j=1}^{M_1}$ and are the decisive patterns for task f.

From (7), the task f outputs label +1 (or -1) if the IDR pattern is $\boldsymbol{\mu}_a$ (or $\boldsymbol{\mu}_b$). If the data contains neither of these two patterns, the label is random.

Out-of-Domain Data and Tasks. Assume there are M_1' *out-of-domain-relevant* (*ODR*) patterns $\{\boldsymbol{\mu}_j'\}_{j=1}^{M_1'}$ and M_2' *out-of-domain-irrelevant* (*ODI*) patterns $\{\boldsymbol{\nu}_k'\}_{k=1}^{M_2'}$. Any data $\boldsymbol{x}$ dawn from $\mathcal{D}'$ can be generated by

$$\boldsymbol{x} = \boldsymbol{\mu}_j' + \kappa' \boldsymbol{\nu}_k' \tag{8}$$

where $j \in [M_1']$ and $k \in [M_2']$ are arbitrarily selected, and $\kappa' \sim U(K', K')$ for $K' = \mathcal{O}(1)$. We use ODR(x):= $\boldsymbol{\mu}_j'$ to denote the ODR pattern of $\boldsymbol{x}$.

The set of out-of-domain tasks $\mathcal{T}'$ contains $M_1'(M_1' - 1)$ binary classification problems that are defined in the same fashion as Definition 3.1, with the only difference of using $\{\boldsymbol{\mu}_j'\}_{j=1}^{M_1'}$ rather than $\{\boldsymbol{\mu}_j\}_{j=1}^{M_1}$ to determine labels.

Prompt Construction for Training and Testing. Let l_{tr} and l_{ts} denote the length of training and testing contexts, respectively.

Training prompt embedding: Given an input-label pair $\boldsymbol{x}_{query}$ and $f(\boldsymbol{x}_{query})$ for training, the context inputs $\boldsymbol{x}_i$ in $\boldsymbol{P}$ in (1) are constructed as follows. The IDR pattern is selected from $\{\boldsymbol{\mu}_j\}_{j=1}^{M_1}$ following a categorical distribution parameterized by α, where where $\alpha = \Theta(1) \in (0, 1]$. Specifically, each of $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ (the decisive patterns of task f) is selected with probability $\alpha/2$, and each of these other $M_1 - 2$ patterns elected with probability $(1-\alpha)/(M_1 - 2)$. The context labels are determined by task f.

Testing prompt embedding: The context inputs for the testing query can be selected following a wide range of prompt selection methods (Liu et al., 2022b; Rubin et al., 2022; Wu et al., 2023b). Given an in-domain (or out-of-domain) task f that has decisive patterns $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ (or $\boldsymbol{\mu}_a'$ and $\boldsymbol{\mu}_b'$), we only assume at least $\alpha'/2$ ($\alpha' \in (0, 1]$) fraction of context inputs contain the same IDR (or ODR) pattern as the query.

For the label embedding $\boldsymbol{y}_i$ for both training and testing, $\text{Embd}(+1) = \boldsymbol{q}$, $\text{Embd}(-1) = -\boldsymbol{q}$, where $\boldsymbol{q} \in \mathbb{R}^{d_\mathcal{Y}}$. Hence, $\boldsymbol{y}_i \in \{\boldsymbol{q}, -\boldsymbol{q}\}$ for $i \in [l_{tr}]$ or $i \in [l_{ts}]$.

3.3. In-Domain and Out-of-Domain Generalization With Sample Complexity Analysis

In order for the learned model $F(\Psi)$ to generalize all tasks in $\mathcal{T}$ through ICL, the training tasks in $\mathcal{T}_{tr}$ should uniformly cover all the possibilities of IDR patterns and labels, as stated by the following condition,

Condition 3.2. *For any given $j \in [M_1]$ and either label $+1$ or -1, the number of tasks in $\mathcal{T}_{tr}$ that map $\boldsymbol{\mu}_j$ to that label is $|\mathcal{T}_{tr}|/M_1 (\geq 1)$.*

Note that Condition 3.2 is easy to meet, and $|\mathcal{T}_{tr}|$ does not have to be large. In fact, $|\mathcal{T}_{tr}|$ can be as small as M_1. For example, let the i-th task function ($i \in [M_1 - 1]$) in $\mathcal{T}_{tr}$ map the queries with $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_{i+1}$ as IDR patterns to $+1$ and -1, respectively. The M_1-th task function maps $\boldsymbol{\mu}_{M_1}$ and $\boldsymbol{\mu}_1$ to $+1$ and -1, respectively. We can easily verify $\mathcal{T}_{tr}$ satisfies Condition 3.2 in this case.

Following (Shi et al., 2021; Karp et al., 2021; Li et al., 2023a), we assume the training labels are balanced, i.e., $\big||\{n : z^n = +1\}| - |\{n : z^n = -1\}|\big| = \mathcal{O}(\sqrt{N})$. The next theorem states the training and in-domain generalization.

Theorem 3.3. *(In-Domain Generalization) Suppose Condition 3.2 holds. For any $\epsilon > 0$, when (i) the number of neurons in $\boldsymbol{W}_O$ satisfies $m \geq \Omega(M_1^2 \log M_1)$, (ii) batch size $B > \Omega(\max\{\epsilon^{-2}, M_1\} \cdot \log M_1)$, (iii) the lengths of training and testing contexts are*

$$l_{tr} \geq \max\{\Omega(\log M_1/\alpha), \Omega(1/(\beta^2\alpha))\},\ l_{ts} \geq \alpha'^{-1}, \tag{9}$$

(iv) and the number of iterations satisfies

$$T = \Theta(\eta^{-1} M_1 \alpha^{-\frac{2}{3}} \beta^{-2/3} \sqrt{\log M_1}), \tag{10}$$

with step size $\eta \leq 1$ and $N = BT$ samples, then with a high probability, the returned model satisfies that

$$\underset{\boldsymbol{x}_{query} \sim \mathcal{D}, f \in \mathcal{T}}{\mathbb{E}} [\ell(\Psi; \boldsymbol{P}, z)] \leq \mathcal{O}(\epsilon). \tag{11}$$

Theorem 3.3 characterizes the sufficient condition on the model size, the required number of iterations, and the number of prompt embedding and label pairs, such that the trained model achieves an in-domain generalization error of $\mathcal{O}(\epsilon)$. Theorem 3.3 includes three major insights:

1. *In-domain generalization capability using a diminishing fraction of training tasks*. Because $\mathcal{T}_{tr}$ can satisfy Condition

3.2 even when $|\mathcal{T}_{tr}| = M_1$, then the number of training tasks is only a fraction $(M_1-1)^{-1/2}$ of the total number of in-domain tasks in $\mathcal{T}$.

2. (*Context length*) The required length of training and testing contexts increase in the order of α^{-1} and ${\alpha'}^{-1}$, respectively, which implies that a longer context is needed when the fraction of IDR patterns in the context is small.

3. (*Convergence and sample complexity*) The required number of iterations and the training samples is proportional to $\alpha^{-2/3}$. This indicates that a larger fraction of the IDR pattern in the context leads to more efficient convergence and generalization.

Based on the in-domain result, we can also investigate the properties of out-of-domain generalization.

Theorem 3.4. *(Out-of-Domain Generalization)*

Suppose Condition 3.2 and conditions (i)-(iv) in Theorem 3.3 hold. For any $\boldsymbol{\mu}'_1, \cdots, \boldsymbol{\mu}'_{M_1}, \boldsymbol{\nu}'_1, \boldsymbol{\nu}'_{M_2}$ that are pairwise orthogonal and $\|\boldsymbol{\mu}'_j\| = \|\boldsymbol{\nu}'_k\| = \beta$, if

$$\boldsymbol{\mu}'_j \in \left\{\sum_{i=1}^{M_1} k_{j,i}\boldsymbol{\mu}_i \Big| S_j := \sum_{i=1}^{M_1} k_{j,i} \geq 1, k_{j,i} \in \mathbb{R}\right\}, \quad (12)$$

and $\boldsymbol{\nu}'_k \in span\{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \cdots, \boldsymbol{\nu}_{M'_2}\}$, $j \in [M'_1]$, $k \in [M'_2]$, then with high probability, the learned model can achieve an out-of-domain generalization error of

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{query}\sim\mathcal{D}', f\in\mathcal{T}'}[\ell(\Psi; \boldsymbol{P}, z)] \leq \mathcal{O}(\epsilon). \quad (13)$$

Remark 3.5. Theorem 3.4 indicates that a one-layer Transformer can generalize well in context, even in the presence of distribution shifts between the training and testing data. The conditions for a favorable generalization encompass the following: (1) the ODR patterns are linear combinations of IDR patterns with a summation of coefficients ≥ 1, and each ODI pattern is in the subspace spanned by IDI patterns; (2) the testing prompt is long enough, which is linear in ${\alpha'}^{-1}$, to include context inputs involving ODR patterns.

Remark 3.6. (Comparison with existing ICL analysis) (Huang et al., 2023) analyzes the generalization performance of ICL on unseen tasks under a similar data model that includes decisive and indecisive patterns. However, (Huang et al., 2023) only analyzes in-domain unseen tasks, while our results also apply to one type of out-of-domain tasks through data shift. To the best of our knowledge, only (Zhang et al., 2023a) studies out-of-domain generalization under the setup of linear regression problems with Gaussian inputs. They conclude that, under this setup, the covariate shift, i.e., the difference between the training and testing data distributions $\mathcal{D}$ and $\mathcal{D}'$, does not guarantee generalization. We consider classification problems under a data model different from (Zhang et al., 2023a). We provide the out-of-domain generalization guarantee for one type of distribution between $\mathcal{D}$ and $\mathcal{D}'$.

3.4. ICL With Magnitude-Based Model Pruning

Theorem 3.7. *Let $\boldsymbol{r}_i$ be the i-row of $\boldsymbol{W}_O\boldsymbol{W}_V$, $i \in [m]$. Suppose Condition 3.2 and conditions (i)-(iv) in Theorem 3.3 hold, then there exists $\mathcal{L} \subset [m]$ with $|\mathcal{L}| = \Omega(m)$ s.t.,*

$$\begin{aligned}\|\boldsymbol{r}_i^{(T)}\| &\geq \Omega(1), i \in \mathcal{L},\\ \|\boldsymbol{r}_i^{(T)}\| &\leq (1/\sqrt{M_2}), i \in \mathcal{L}^c,\end{aligned} \quad (14)$$

where $\mathcal{L}^c$ is the complementary set of $\mathcal{L}$. Then, for any $\epsilon > 0$ and any in- or out-of-domain $\boldsymbol{x}_{query} \sim \mathcal{D}$ (or $\mathcal{D}'$) and corresponding $f \in \mathcal{T}$ (or $\mathcal{T}'$), pruning all neurons $i \in \mathcal{L}^c$ leads to a generalization error

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{query}, f}[\ell(\Psi_{\mathcal{L}^c}; \boldsymbol{P}, z)] \leq \mathcal{O}(\epsilon + M_1^{-1/2}), \quad (15)$$

where $\Psi_{\mathcal{L}^c}$ represents the model weights after removing neurons in $\mathcal{L}^c$ in $\boldsymbol{W}_O$. In contrast, pruning $\mathcal{S} \subset \mathcal{L}$ with size $|\mathcal{S}| = Rm$, where $R \in (0, 1)$ and is a constant, and $\alpha' \geq \Omega(M_1^{-0.5})$ results in a generalization error of

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{query}, f}[\ell(\Psi_{\mathcal{S}}; \boldsymbol{P}, z)] \geq \Omega(R + (\alpha' M_1)^{-1}). \quad (16)$$

Remark 3.8. Theorem 3.7 proves that a constant fraction of neurons in $\mathcal{L}$ in the trained MLP layer has large weights, while the remaining ones in $\mathcal{L}^c$ have small weights. Pruning neurons with a smaller magnitude leads to almost the same generalization result as that of the unpruned Ψ. However, pruning neurons with a larger magnitude cause an increasing generalization error as the pruning ratio R increases. Theorem 3.7 indicates that in our setup, magnitude-based pruning on $\boldsymbol{W}_O$ does not hurt the model's ICL capability.

4. The Mechanism of ICL by the Trained Transformer

Here, we provide a detailed discussion about how the generalization performance in Theorems 3.3 and 3.4 are achieved. We first introduce novel properties of the self-attention layer and the MLP layer of the learned Transformer to implement ICL in Sections 4.1 and 4.2. The high-level proof idea of Theorems 3.3 and 3.4 is presented in Appendix D.1.

4.1. Self-Attention Selects Contexts With the Same IDR/ODR Pattern as the Query

We first show the learned self-attention layer promotes context examples that share the same IDR/ODR pattern as the query. Specifically, for any vector $\boldsymbol{p} \in \mathbb{R}^{d_{\mathcal{X}}+d_{\mathcal{Y}}}$ that includes input $\boldsymbol{x}$ and the corresponding output embedding $\boldsymbol{y}$. We use XDR($\boldsymbol{p}$) to represent the relevant pattern, which is the IDR($\boldsymbol{x}$) for in-domain data and ODR($\boldsymbol{x}$) for out-of-domain data. Then

Proposition 4.1. *The trained model after being updated by T (characterized in (10)) iterations satisfies that, for any $(\boldsymbol{p}, \boldsymbol{W}) \in \{(\boldsymbol{p}_{query}, \boldsymbol{W}_Q^{(T)}), (\boldsymbol{p}_i, \boldsymbol{W}_K^{(T)})\}_{i=1}^l$,*

$$\|[XDR(\boldsymbol{p})^\top, \boldsymbol{0}^\top]\boldsymbol{W}\boldsymbol{p}\| \geq \Omega(\sqrt{\log M_1}), \quad (17)$$

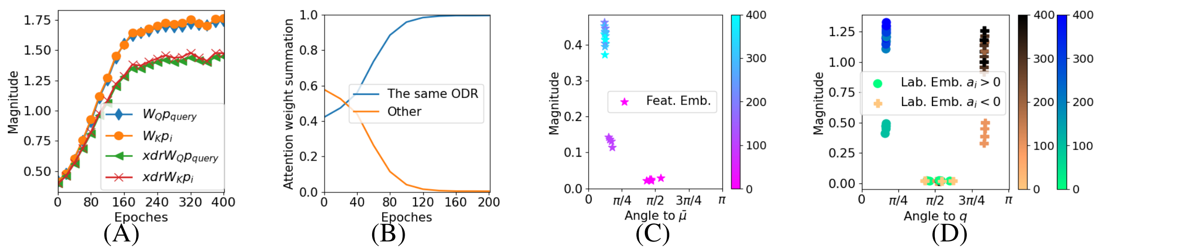


Figure 2. The properties of the trained model. (A) The average norm of $\boldsymbol{W}_Q\boldsymbol{p}_{query}$, $\boldsymbol{W}_K\boldsymbol{p}_i$, $[XDR(\boldsymbol{p}_{query})^\top/\beta, \boldsymbol{0}^\top]\cdot\boldsymbol{W}_Q\boldsymbol{p}_{query}$, and $[XDR(\boldsymbol{p}_i)^\top/\beta, \boldsymbol{0}^\top]\boldsymbol{W}_K\boldsymbol{p}_i$. (B) The attention weight summation on contexts with the same ODR pattern as the query and other contexts. (C) The magnitude of the feature embedding of 5 neurons in $\boldsymbol{W}_O\boldsymbol{W}_V$ and their angles to $\bar{\boldsymbol{\mu}}$ in 400 epochs. (D) The magnitude of the label embedding of 10 neurons in $\boldsymbol{W}_O\boldsymbol{W}_V$ and their angles to $\boldsymbol{q}$ in 400 epochs. We choose 5 neurons for $a_i > 0$ and 5 for $a_i < 0$.

$$\|[\boldsymbol{a}^\top, \boldsymbol{0}^\top]\boldsymbol{W}\boldsymbol{p}\| \leq \mathcal{O}(\sqrt{\log M_1}(1/M_1 + 1/M_2)), \quad (18)$$

$$\|[\boldsymbol{b}^\top, \boldsymbol{0}^\top]\boldsymbol{W}\boldsymbol{p}\| \leq \mathcal{O}(\sqrt{\log M_1}(1/M_1 + 1/M_2)), \quad (19)$$

where $\boldsymbol{a}$ is any IDR (or ODR) pattern that is different from XDR($\boldsymbol{p}$) for in-domain (or out-of-domain) tasks, $\boldsymbol{b}$ is any IDI (or ODI) pattern, and $\boldsymbol{0}$ is an all-zero vector in $\mathbb{R}^{m_a - d_\mathcal{X}}$.

Remark 4.2. Proposition 4.1 indicates that the self-attention layer parameters $\boldsymbol{W}_Q^{(T)}$ and $\boldsymbol{W}_K^{(T)}$ in the returned model projects $\boldsymbol{p}_{query}$ or context embeddings $\boldsymbol{p}_i$ mainly to the directions of the corresponding IDR pattern for in-domain data or ODR pattern for out-of-domain data. This can be deduced by combining (17), (18), and (19), since components of $\boldsymbol{W}\boldsymbol{p}$ in other directions rather than $[XDR(\boldsymbol{p})^\top, \boldsymbol{0}^\top]$ are relatively smaller. Hence, Proposition 4.1 implies that the learned $\boldsymbol{W}_Q^{(T)}$ and $\boldsymbol{W}_K^{(T)}$ remove the effect of IDI/ODI patterns. Meanwhile, (17) states that the $\boldsymbol{W}_Q^{(T)}$ and $\boldsymbol{W}_K^{(T)}$ enlarge the magnitude of the IDR or ODR patterns from $\Theta(1)$ to $\Theta(\sqrt{\log M_1})$, given that the $\boldsymbol{W}_Q^{(0)}$ and $\boldsymbol{W}_K^{(0)}$ are initialized with a scalar $\delta = \Theta(1)$.

Proposition 4.1 enables us to compute the attention map of the trained model. Therefore, we have the following.

Corollary 4.3. *For any testing query embedding $\boldsymbol{p}_{query} = [\boldsymbol{x}_{query}^\top, \boldsymbol{0}^\top]^\top$, let $\mathcal{N}_* \in [l]$ be the set of indices of context inputs that share the same IDR (or ODR) pattern as the in-domain (or out-of-domain) $\boldsymbol{x}_{query}$. Then, for any constant $C > 1$, by definition in (2), it holds that*

$$\sum_{s\in\mathcal{N}_*} attn(\Psi; \boldsymbol{P}, i) \geq 1 - \Theta(1/M_1^C). \quad (20)$$

Remark 4.4. Corollary 4.3 shows that after training, the attention weights become concentrated on contexts in $\mathcal{N}_*$. This means that the learned self-attention layer only selects some crucial contexts that share the same IDR/ODR pattern as the query rather than all samples uniformly or randomly.

4.2. MLP Neurons Distinguish Label Embeddings Rather Than Feature Embeddings.

We next show that the trained MLP layer can distinguish the label embeddings for data from different classes.

Proposition 4.5. *Let $\boldsymbol{r}_i$ introduced in Theorem 3.7 be $(\boldsymbol{r}_{i_{d_\mathcal{X}}}^\top, \boldsymbol{r}_{i_{d_\mathcal{Y}}}^\top, \boldsymbol{r}_i'^\top)$ where $\boldsymbol{r}_{i_{d_\mathcal{X}}} \in \mathbb{R}^{d_\mathcal{X}}$, $\boldsymbol{r}_{i_{d_\mathcal{Y}}} \in \mathbb{R}^{d_\mathcal{Y}}$, and $\boldsymbol{r}_i' \in \mathbb{R}^{m_b - d_\mathcal{X} - d_\mathcal{Y}}$. Then, for any $i \in \mathcal{L}$*

$$\boldsymbol{r}_{i_{d_\mathcal{X}}}^{(T)}\bar{\boldsymbol{\mu}}/(\|\boldsymbol{r}_{i_{d_\mathcal{X}}}^{(T)}\| \cdot \|\bar{\boldsymbol{\mu}}\|) \geq 1 - \Theta(1)/M_2, \quad (21)$$

$$\boldsymbol{r}_{i_{d_\mathcal{Y}}}^{(T)}\boldsymbol{q}_e/(\|\boldsymbol{r}_{i_{d_\mathcal{Y}}}^{(T)}\| \cdot \|\boldsymbol{q}_e\|) \geq 1 - \Theta(1)/M_1, \quad (22)$$

where $\bar{\boldsymbol{\mu}} = \sum_{k=1}^{M_1} \boldsymbol{\mu}_k^\top/M_1$, $\boldsymbol{q}_e = \boldsymbol{q}$ if $a_i > 0$ and $\boldsymbol{q}_e = -\boldsymbol{q}$ if $a_i < 0$, where a_i is the i-th entry of $\boldsymbol{a}$ in (1).

Remark 4.6. Proposition 4.5 demonstrates that neurons with indices in $\mathcal{L}$ have the following two properties. (P1) The first $d_\mathcal{X}$ entries of all the corresponding row vectors in $\boldsymbol{W}_O^{(T)}\boldsymbol{W}_V^{(T)}$ approximate the average of all IDR patterns $\boldsymbol{\mu}_j$, $j \in [M_1]$. (P2) The next $d_\mathcal{Y}$ entries of the ith row of $\boldsymbol{W}_O^{(T)}\boldsymbol{W}_V^{(T)}$ approximates the label embedding $\boldsymbol{q}$ when $a_i > 0$ and approximates $-\boldsymbol{q}$ when $a_i < 0$. (P1) indicates that the output layer focuses on all IDR patterns equally rather than any IDI pattern. (P2) indicates that the MLP layer can distinguish label embeddings for different classes.

5. Numerical Experiments

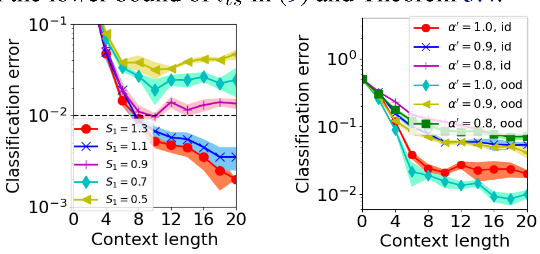
Data Generation We verify our theoretical findings using data generated as described in Section 2. Let $d_\mathcal{X} = d_\mathcal{Y} = 30$, $\beta = 3$, $K' = 5$, $K = 0.5$. The in-context binary classification error is evaluated by $\mathbb{E}_{(\boldsymbol{x},y)}[\Pr(y \cdot F(\Psi; \boldsymbol{P}) < 0)]$ for $\boldsymbol{x}$ following either $\mathcal{D}$ or $\mathcal{D}'$ and $\boldsymbol{P}$ constructed in (1). If not otherwise specified, we set $M_1 = 6$, $M_2 = 24$. For out-of-domain generalization, $M_1' = 3$, $\boldsymbol{\nu}_i' = \boldsymbol{\nu}_i$ for $i \in [M_2']$. $\boldsymbol{\mu}_1' = 0.3 \cdot (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + a\boldsymbol{\mu}_5 + b\boldsymbol{\mu}_6$. $\boldsymbol{\mu}_2' = \sqrt{2}/2 \cdot (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$. $\boldsymbol{\mu}_3' = \sqrt{2}/2 \cdot (\boldsymbol{\mu}_3 + \boldsymbol{\mu}_4)$. For testing, we select contexts with the two decisive patterns with $\alpha'/2$ probability each and others with $(1 - \alpha')/(M_1' - 2)$ probability each to keep the context outputs balanced.

Model and Training Setup: The models we use include both the one-layer Transformer defined in (2) and the 3-layer 2-head real-world model GPT-2 (Radford et al., 2019) following (Bai et al., 2023; Wu et al., 2023a). If not otherwise specified, we set $\alpha = 0.8$, $l_{tr} = 20$ for training. The training tasks are formulated as follows to satisfy Condition

3.2. Define $\boldsymbol{a}_i = \boldsymbol{a}_{i+M_1} = \boldsymbol{\mu}_i$ for $i \in [M_1]$, and then the $((k-1)\cdot M_1 + j)$-th task function maps the queries with $\boldsymbol{a}_j$ and $\boldsymbol{a}_{j+k}$ as IDR patterns to $+1$ and -1, respectively, for $j \in [M_1]$ and $k \in [U]$. For the one-layer Transformer, we use $U = 1$ and $m_a = m_b = 60$. Hence, $|\mathcal{T}_{tr}| = 6$, and there are $|\mathcal{T}\backslash\mathcal{T}_{tr}| = 24$ in-domain unseen tasks. For GPT-2, $U = 4$. Then, $|\mathcal{T}_{tr}| = 24$, $|\mathcal{T}\backslash\mathcal{T}_{tr}| = 6$. Note that we evaluate in-domain generalization error only on unseen tasks $\mathcal{T}\backslash\mathcal{T}_{tr}$, which is generally an upper bound of that defined in (4) after sufficient training.

5.1. Experiments on the Generalization of ICL

We first verify the sufficient condition (12) for out-of-domain generalization. From the selection of $\boldsymbol{\mu}'$'s, we know that $S_1 = a + b$, $S_2 = S_3 = \sqrt{2}$. We vary a and b while satisfying $a^2 + b^2 + 2 \cdot 0.3^2 = 1$. Figure 3 (A) shows that the out-of-domain classification error archives < 0.01 when $S_1 \geq 1$ and deviates from 0 when $S_1 < 1$, which justifies the necessity of condition (12). We then investigate how the context length is affected by α', i.e., the fraction of contexts with the same IDR/ODR pattern as the query. Figure 3 (B) indicates that a longer testing context length is needed when α' is smaller for in- or out-of-domain, which is consistent with the lower bound of l_{ts} in (9) and Theorem 3.4.

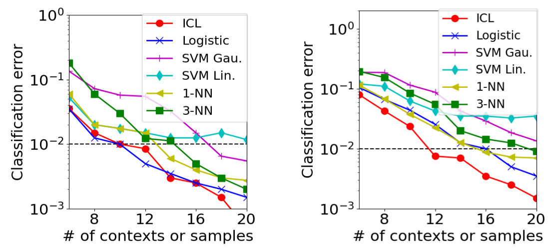


(A) (B)

Figure 3. Out-of-domain ICL classification error on GPT-2 with (a) different S_1 on GPT-2 (b) different α' for in-domain (id) and out-of-domain (ood) generalization.

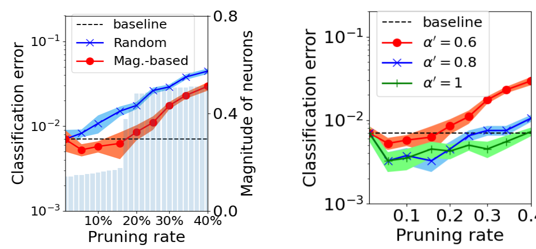
We then compare ICL with other machine learning algorithms for classification, where contexts are used as training samples for these methods. Figure 4 (A) and (B) show that when $\alpha' = 0.8$, the advance of ICL over other algorithms is not significant, while when $\alpha' = 0.6$, ICL is the most sample-efficient for a small generalization error. Thus, ICL can remove irrelevant data and is more robust to random noise in labels than other learning algorithms.

We also investigate the effect of pruning techniques on ICL. Let $\alpha = 0.6$. Figure 5 (A) shows that magnitude-based pruning does not hurt out-of-domain generalization if the pruning rate is lower than around 15%, which is the ratio of $\boldsymbol{W}_O$ neurons with a small magnitude. The generalization error increases as the pruning rate increases when pruning neurons with large weights. This is consistent with Theorem 3.7 and Remark 3.8. Figure 5 (B) justifies the impact of α' in Theorem 3.7 that larger α' can improve the performance of the pruned model.



(A) (B)

Figure 4. Binary classification performance of using ICL, logistic regression (Logistic), SVM with Gaussian kernel (SVM Gau.), SVM with linear kernel (SVM Lin.), 1-nearest neighbor (1-NN), and 3-nearest neighbor (3-NN) with one-layer Transformer when (A) $\alpha' = 0.8$ (B) $\alpha' = 0.6$.



(A) (B)

Figure 5. (A) Out-of-domain classification error (left y-axis for curves) with model pruning of the trained $\boldsymbol{W}_O$ using baseline (no pruning), random pruning, and magnitude-based pruning (Mag.-based), and the magnitude of each neuron of $\boldsymbol{W}_O$ (right y-axis for light blue bars) (B) Out-of-domain classification error when varying α'. These two are implemented on a one-layer Transformer.

5.2. Experiments on the Mechanism of ICL

We examine our findings regarding the mechanism of ICL in Section 4 using a one-layer Transformer formulated in (2). In Figure 2 (A) and (B), we consider out-of-domain data with $a = b = 0.64$. Figure 2 (A) shows that for any query $\boldsymbol{p}_{query}$ (or context example $\boldsymbol{p}_i$ for $i \in l_{ts}$), the norm of $[XDR(\boldsymbol{p})^\top, \boldsymbol{0}^\top]\boldsymbol{W}_Q\boldsymbol{p}_{query}$ (or $[XDR(\boldsymbol{p})^\top, \boldsymbol{0}^\top]\boldsymbol{W}_K\boldsymbol{p}_i$) is close to the norm of $\boldsymbol{W}_Q\boldsymbol{p}_{query}$ (or $\boldsymbol{W}_K\boldsymbol{p}_i$). This implies that the components of $\boldsymbol{W}_Q\boldsymbol{p}_{query}$ (or $\boldsymbol{W}_K\boldsymbol{p}_i$) in directions other than $[XDR(\boldsymbol{p})^\top, \boldsymbol{0}^\top]$ are small, which is consistent with (18) and (19) in Proposition 4.1. Moreover, these norms increase from initialization during training, which justifies (17). Figure 2 (B) depicts the concentration of attention on contexts in $\mathcal{N}_*$ after training. This verifies Corollary 4.3. Figure 2 (C) and (D) jointly verify Proposition 4.5. The color bars represent the epochs of training. We can observe that except for some neurons, $\boldsymbol{r}_{i_{d_\mathcal{X}}}$ grows to be close to the direction of $\bar{\boldsymbol{\mu}}$ with a larger magnitude in Figure 2 (C). Moreover, Figure 2 (D) shows for $a_i > 0$ (or $a_i < 0$), $\boldsymbol{r}_{i_{d_\mathcal{Y}}}$ becomes close to $\boldsymbol{q}$ (or $-\boldsymbol{q}$) with a large magnitude.

6. Conclusion

This paper provides theoretical analyses of the training dynamics of Transformers with nonlinear attention and nonlinear MLP, and the resulting ICL capability for new tasks with possible data shift. This paper also provides a theoretical justification for magnitude-based pruning to reduce inference costs while maintaining the ICL capability. Future directions include designing practical prompt selection algorithms and model pruning methods based on the obtained insights, as well as investigating ICL on generation tasks.

Acknowledgements

This work was supported by IBM through the IBM-Rensselaer Future of Computing Research Collaboration. We thank Dr. Hui Wan (*hui.wan@gmail.com*) at Google Cloud AI, USA, for the valuable discussion. We thank all anonymous reviewers for their constructive comments.

Impact Statement

This paper aims to explore the mechanisms of transformer-based neural networks in the context of in-context learning. The primary focus is on the mathematical analysis of generalization errors. To the best of our knowledge, no potential societal consequences are associated with our work.

References

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Allen-Zhu, Z., Li, Y., and Liang, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6155–6166, 2019a.

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

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901, 2020.

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

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 10836–10846, 2019.

Chen, S. and Li, Y. Provably learning a multi-head attention layer. *arXiv preprint arXiv:2402.04084*, 2024.

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. A generalized neural tangent kernel analysis for two-layer neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Cheng, X., Chen, Y., and Sra, S. Transformers implement functional gradient descent to learn non-linear functions in context. *arXiv preprint arXiv:2312.06528*, 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Chowdhury, M. N. R., Zhang, S., Wang, M., Liu, S., and Chen, P.-Y. Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks. In *International Conference on Machine Learning*, 2023.

Chowdhury, M. N. R., Wang, M., Maghraoui, K. E., Wang, N., Chen, P.-Y., and Carothers, C. A provably effective method for pruning experts in fine-tuned sparse mixture-of-experts. *arXiv preprint arXiv:2405.16646*, 2024.

Cui, Y., Ren, J., He, P., Tang, J., and Xing, Y. Superiority of multi-head attention in in-context linear regression. *arXiv preprint arXiv:2401.17426*, 2024.

Daniely, A. and Malach, E. Learning parities with neural networks. *Advances in Neural Information Processing Systems*, 33:20356–20365, 2020.

Deora, P., Ghaderi, R., Taheri, H., and Thrampoulidis, C. On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*, 2023.

Frantar, E. and Alistarh, D. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pp. 10323–10337. PMLR, 2023.

Fu, H., Chi, Y., and Liang, Y. Guaranteed recovery of one-hidden-layer neural networks via cross entropy. *IEEE Transactions on Signal Processing*, 68:3225–3235, 2020.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Guo, T., Hu, W., Mei, S., Wang, H., Xiong, C., Savarese, S., and Bai, Y. How do transformers learn in-context beyond simple functions? a case study on learning with representations. *arXiv preprint arXiv:2310.10616*, 2023.

Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

Huang, Y., Cheng, Y., and Liang, Y. In-context convergence of transformers. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Jelassi, S., Sander, M., and Li, Y. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

Karp, S., Winston, E., Li, Y., and Singh, A. Local signal adaptivity: Provable feature learning in neural networks beyond kernels. *Advances in Neural Information Processing Systems*, 34:24883–24897, 2021.

Li, H., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. Generalization guarantee of training graph convolutional networks with graph topology sampling. In *International Conference on Machine Learning*, pp. 13014–13051. PMLR, 2022a.

Li, H., Zhang, S., and Wang, M. 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)*, pp. 37–42. IEEE, 2022b.

Li, H., Wang, M., Liu, S., and Chen, P.-Y. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=jClGv3Qjhb.

Li, H., Wang, M., Ma, T., Liu, S., ZHANG, Z., and Chen, P.-Y. What improves the generalization of graph transformer? a theoretical dive into self-attention and positional encoding. In *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*, 2023b. URL https://openreview.net/forum?id=BaxFC3z9R6.

Li, H., Zhang, S., Zhang, Y., Wang, M., Liu, S., and Chen, P.-Y. How does promoting the minority fraction affect generalization? a theoretical study of one-hidden-layer neural network on group imbalance. *IEEE Journal of Selected Topics in Signal Processing*, 2024.

Li, S. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70, 2010.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, 2023c.

Li, Y., Li, Y., and Risteski, A. How do transformers learn topic structure: Towards a mechanistic understanding. *arXiv preprint arXiv:2303.04245*, 2023d.

Liu, A., Swayamdipta, S., Smith, N. A., and Choi, Y. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6826–6847, 2022a.

Liu, J., Shen, D., Zhang, Y., Dolan, W. B., Carin, L., and Chen, W. 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*, pp. 100–114, 2022b.

Liu, Z., Wang, J., Dao, T., Zhou, T., Yuan, B., Song, Z., Shrivastava, A., Zhang, C., Tian, Y., Re, C., et al. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pp. 22137–22176. PMLR, 2023.

Lucy, L. and Bamman, D. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pp. 48–55, 2021.

Luo, J.-H., Wu, J., and Lin, W. Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision*, pp. 5058–5066, 2017.

Luo, Y., Li, H., Shi, L., and Wu, X.-M. Enhancing graph transformers with hierarchical distance structural encoding, 2024.

Ma, X., Fang, G., and Wang, X. LLM-pruner: On the structural pruning of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=J8Ajf9WfXP.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Molchanov, P., Tyree, S., Karras, T., Aila, T., and Kautz, J. Pruning convolutional neural networks for resource efficient inference. In *International Conference on Learning Representations*, 2016.

OpenAI. Gpt-4 technical report. *OpenAI*, 2023.

Oymak, S., Rawat, A. S., Soltanolkotabi, M., and Thrampoulidis, C. On the role of attention in prompt-tuning. *arXiv preprint arXiv:2306.03435*, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Rubin, O., Herzig, J., and Berant, J. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, 2022.

Shi, Z., Wei, J., and Liang, Y. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. In *International Conference on Learning Representations*, 2021.

Sun, J., Li, H., and Wang, M. How do skip connections affect graph convolutional networks with graph sampling? a theoretical analysis on generalization, 2024. URL https://openreview.net/forum?id=J2pMoN2pon.

Sun, M., Liu, Z., Bair, A., and Kolter, J. Z. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.

Tarzanagh, D. A., Li, Y., Thrampoulidis, C., and Oymak, S. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023a.

Tarzanagh, D. A., Li, Y., Zhang, X., and Oymak, S. Max-margin token selection in attention mechanism. *CoRR*, 2023b.

Tian, Y., Wang, Y., Chen, B., and Du, S. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *arXiv preprint arXiv:2305.16380*, 2023a.

Tian, Y., Wang, Y., Zhang, Z., Chen, B., and Du, S. S. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. In *Conference on Parsimony and Learning (Recent Spotlight Track)*, 2023b.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Wen, W., Wu, C., Wang, Y., Chen, Y., and Li, H. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Wu, J., Zou, D., Chen, Z., Braverman, V., Gu, Q., and Bartlett, P. L. How many pretraining tasks are needed for in-context learning of linear regression? *arXiv preprint arXiv:2310.08391*, 2023a.

Wu, Z., Wang, Y., Ye, J., and Kong, L. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. *ACL*, 2023b.

Yang, H. and Wang, Z. On the neural tangent kernel analysis of randomly pruned neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1513–1553. PMLR, 2023.

Yang, H., Liang, Y., Guo, X., Wu, L., and Wang, Z. Theoretical characterization of how neural network pruning affects its generalization. *arXiv preprint arXiv:2301.00335*, 2023.

Yang, K., Ackermann, J., He, Z., Feng, G., Zhang, B., Feng, Y., Ye, Q., He, D., and Wang, L. Do efficient transformers really save computation? *arXiv preprint arXiv:2402.13934*, 2024.

Zhang, R., Frei, S., and Bartlett, P. L. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023a.

Zhang, S., Wang, M., Liu, S., Chen, P.-Y., and Xiong, J. Why lottery ticket wins? a theoretical perspective of sample complexity on sparse neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhang, S., Li, H., Wang, M., Liu, M., Chen, P.-Y., Lu, S., Liu, S., Murugesan, K., and Chaudhury, S. On the convergence and sample complexity analysis of deep q-networks with ϵ-greedy exploration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Zhang, S., Wang, M., Chen, P.-Y., Liu, S., Lu, S., and Liu, M. Joint edge-model sparse learning is provably efficient for graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023c.

Zhang, Y., Li, H., Yao, Y., Chen, A., Zhang, S., Chen, P.-Y., Wang, M., and Liu, S. Visual prompting reimagined: The power of activation prompts, 2024. URL https://openreview.net/forum?id=0b328CMwn1.

Zhong, K., Song, Z., Jain, P., Bartlett, P. L., and Dhillon, I. S. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4140–4149, 2017. URL https://arxiv.org/pdf/1706.03175.pdf.

Zou, D. and Gu, Q. An improved analysis of training over-parameterized deep neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, 2019.

A. Proof Sketch

We partially include the proof backbone in Section 4 when introducing the mechanism of the trained Transformer. We elaborate more about our proof intuition of Theorem 3.3 in the following.

We first briefly introduce our proof intuition of Theorem 3.3. We first respectively build Lemmas D.5, D.6, and D.7 to characterize gradient updates for $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$, $\boldsymbol{W}_V$ and $\boldsymbol{W}_O$. These Lemmas are based on an observation that a constant fraction of neurons in $\boldsymbol{W}_O$ can always be activated (Lemma D.9, D.10) to avoid the non-smoothness of Relu activation. The orthogonality of patterns and Definition 3.1 enable the self-attention layer to learn in-domain-relevant (IDR) patterns rather than in-domain-irrelevant (IDI) patterns and select contexts with the same IDR pattern as the query. Then, to develop Theorem 3.3, we use Lemma D.5 to show that the attention weights converge to be close to 1 when $\eta T \geq \Omega(M_1\sqrt{\log M_1})$. Next, we compute the network output according to the label embedding using Lemma D.6 and D.7. Finally, we derive the required number of iterations to make the generalization error $\mathcal{O}(\epsilon)$ by concentration inequalities.

We then would like to specify how we handle the Relu activation in the gradient of $\boldsymbol{W}_K$, $\boldsymbol{W}_Q$. We also want to clarify that how we can handle the training dynamics related to the softmax is different from (Huang et al., 2023) although we both derive a sparse attention distribution. Inspired by the intuition of feature-learning analyses for two-layer Relu networks (Brutzkus & Globerson, 2021; Shi et al., 2021; Zhang et al., 2023c; Allen-Zhu & Li, 2023), we initialize the model such that at least a constant fraction of the neurons of $\boldsymbol{W}_O$ are activated (Lemma D.9), which are called lucky neurons as in (Zhang et al., 2023c; Li et al., 2023a). We prove that these lucky neurons are always activated (Lemma D.10) and grow with an increasing magnitude and two fixed directions of the label embedding along the training (Lemma D.7). Then, we can show that the gradient growths of $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$ can be lower bounded by contributions from these lucky neurons. Therefore, we are able to characterize the gradient updates of $\boldsymbol{W}_K$ and $\boldsymbol{W}_Q$ given a dynamic $\boldsymbol{W}_O$. This process is different from (Huang et al., 2023) since (Huang et al., 2023) does not include Relu MLP, so there is no need to study lucky neurons. Besides, we use Hinge loss, while their training loss is logistic loss, which leads to more training phases, as a difference in the training dynamics between us.

B. Addition Discussions and Extensions

B.1. The Motivation to Study NONLINEAR Transformers

The reasons we study nonlinear Transformers in this work are as follows. First, nonlinear Transformers for ICL, which are different from linear Transformers, are common in practice but less explored in theory. Nonlinear attention and nonlinear MLP are default components of standard Transformers (Vaswani et al., 2017) and are widely applied in large language models for implementing ICL in practice. Existing works show that nonlinear Transformers exhibit their empirical advantages when learning nonlinear functions (Cheng et al., 2023) or conducting dynamic programming tasks (Yang et al., 2024). However, state-of-the-art theoretical works (Zhang et al., 2023a; Huang et al., 2023; Wu et al., 2023a) ignore the nonlinearities (partially) to simplify the analysis or the presentation. Second, the analysis of nonlinear Transformers is quite different from that of Transformers without nonlinearities. For example, softmax attention has a different derivative from linear attention, which includes nonlinear exponential terms and needs a more complicated computation of the gradient updates. Relu MLP provides several non-differential points, which makes the loss landscape more challenging to analyze.

B.2. The Discussion on Single/Multi-Head Attention

There are several reasons why we only study single-head attention in the main body of the paper. First, all the previous theoretical works studying the optimization and generalization of Transformers on ICL (Zhang et al., 2023a; Huang et al., 2023; Wu et al., 2023a) only consider single-head attention in the network. Some concurrent works consider multi-head attention, but they either do not study ICL (Deora et al., 2023; Chen & Li, 2024) or do not involve convergence/generalization guarantee (Cui et al., 2024). Hence, the question of how the ICL ability on unseen tasks and out-of-domain data is obtained by training is still unexplored. Our theoretical analysis studies the convergence and generalization of ICL using Transformers with softmax attention and Relu MLP, involving generalization on unseen tasks and OOD data. Second, our empirical experiments on GPT-2 in Figure 3 are conducted with two heads to verify our theoretical findings, which means some theoretical conclusions hold in Transformers with multiple heads.

However, our analysis for single-head attention can be extended to multi-head attention to some degree. Consider a multi-head attention layer where the layer output is a concatenation of the output of each head, i.e.,

$$\Big\|_{h=1}^{H} \sum_{i=1}^{l} \boldsymbol{W}_{V_h}\boldsymbol{p}_i \cdot \text{softmax}({\boldsymbol{p}_i}^\top \boldsymbol{W}_{K_h}^\top \boldsymbol{W}_{Q_h}\boldsymbol{p}_{query}). \tag{23}$$

The overall conclusion will remain the same, given the same data formulation and initialization on each head because of the orthogonality of patterns. Specifically, we can still show that each attention head selects contexts with the same in-domain-relevant (IDR) pattern as the query. The MLP layer will still make predictions based on the label embedding, as suggested in Section 4.2. We will leave the analysis on multi-head attention with more general settings as future directions.

B.3. Extension to Multiple Patterns for One Class

We can extend our analysis to the case that several orthogonal IDR patterns correspond to the label $+1$, while some other orthogonal IDR patterns correspond to the label -1. Then, as long as there is always a context input that shares the same IDR/ODR pattern as the query, we can still prove that the self-attention layer selects contexts with the same IDR/ODR pattern as the query. Furthermore, we can show the MLP layer makes predictions based on the label embedding. Therefore, the mechanism remains the same as the current setting in the manuscript, where one pattern corresponds to one pattern. We will leave other cases where the data formulation is different in future works.

The reason why we use our current setting in the main body of the paper is to simplify the presentation while emphasizing our major contributions of analyzing optimization and generalization of nonlinear Transformers both in-domain and out-of-domain. As the first work on this problem, as far as we know, we believe our data formulation keeps the necessary complexity.

B.4. Additional Related Works

We introduce other existing theoretical works on learning and generalization of neural networks in this section. Some works (Zhong et al., 2017; Fu et al., 2020; Li et al., 2022b; Zhang et al., 2023b; Li et al., 2024) 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 (Jacot et al., 2018; Allen-Zhu et al., 2019a;b; Cao & Gu, 2019; Zou & Gu, 2019; Chen et al., 2020; Li et al., 2022a; Sun et al., 2024) considers strongly overparameterized networks to linearize the neural network around the initialization. The generalization performance is independent of the feature distribution. (Daniely & Malach, 2020; Shi et al., 2021; Karp et al., 2021; Brutzkus & Globerson, 2021; Zhang et al., 2023c; Li et al., 2023a; Zhang et al., 2024; Chowdhury et al., 2023; 2024) investigate the generalization of neural networks assuming a data model consisting of discriminative patterns and background patterns. Our analysis belongs to the last line of research.

C. Additional Experiments and the Algorithm

We first present the training algorithm introduced in Section 2.3.

Algorithm 1 Training with Stochastic Gradient Descent (SGD)

1: **Hyperparameters:** The step size η, the number of iterations T, batch size B.

2: **Initialization:** Each entry of $\boldsymbol{W}_O^{(0)}$ and $\boldsymbol{a}^{(0)}$ from $\mathcal{N}(0,\xi^2)$ and Uniform$(\{+1/\sqrt{m}, -1/\sqrt{m}\})$, respectively. $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$ and $\boldsymbol{W}_V$ are initialized such that all diagonal entries of $\boldsymbol{W}_V^{(0)}$, and the first $d_\mathcal{X}$ diagonal entries of $\boldsymbol{W}_Q^{(0)}$ and $\boldsymbol{W}_K^{(0)}$ are set as δ with $\delta \in (0, 0.2]$.

3: **Training by SGD:** For each iteration, we independently sample $\boldsymbol{x}_{query} \sim \mathcal{D}$, $f \in \mathcal{T}_{tr}$ to form a batch of training prompt and labels $\{\boldsymbol{P}^n, z^n\}_{n\in\mathcal{B}_t}$ as introduced in Section 3.2. Each IDR pattern is sampled equally likely in each batch. For each $t = 0, 1, \cdots, T-1$ and $\boldsymbol{W}^{(t)} \in \Psi^{(t)}$

$$\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta \cdot \frac{1}{B}\sum_{n\in\mathcal{B}_t} \nabla_{\boldsymbol{W}^{(t)}} \ell(\Psi^{(t)}; \boldsymbol{P}^n, z^n). \tag{24}$$

4: **Output:** $\boldsymbol{W}_O^{(T)}, \boldsymbol{W}_V^{(T)}, \boldsymbol{W}_K^{(T)}, \boldsymbol{W}_Q^{(T)}$.

Then, we introduce additional experiments to verify our theory.

C.1. The impact of α

We choose $\alpha = 0.6$ and use a one-layer Transformer as in (2). Figure 6 shows that the required length of the training prompt is linear in α^{-1}, while the required number of training iterations is linear in $\alpha^{-2/3}$, which verify the theoretical findings in (9) and (10).

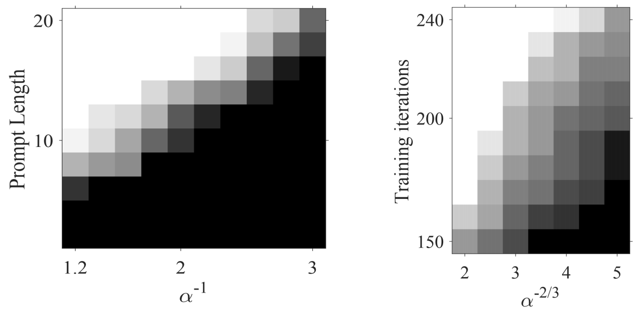


Figure 6. The prompt length against α, and the required number of training iterations against α.

C.2. The required number of training tasks

We choose $\alpha = 0.6$ and use a one-layer Transformer as in (2). For a given $\mathcal{T}$, we first generate a set of tasks that satisfies Condition 3.2 as follows. Define $\boldsymbol{a}_i = \boldsymbol{a}_{i+M_1} = \boldsymbol{\mu}_i$ for $i \in [M_1]$, and then the j-th task function map the queries with $\boldsymbol{a}_j$ and $\boldsymbol{a}_{j+1}$ as IDR patterns to $+1$ and -1, respectively, for $j \in [M_1]$. Then, we get a task set $\mathcal{T}_{tr0}$ with $|\mathcal{T}_{tr0}| = M_1$. Then, we vary the number of training tasks in the way that (1) we sample within $\mathcal{T}_{tr0}$ to get a set $\mathcal{T}_{tr}$ with $|\mathcal{T}_{tr}| \leq M_1$ (2) we sample within $\mathcal{T}\backslash\mathcal{T}_{tr0}$ to get a set $\mathcal{T}_{tr}'$, and $\mathcal{T}_{tr} = \mathcal{T}_{tr}' \cup \mathcal{T}_{tr0}$ such that $|\mathcal{T}_{tr}| \geq M_1$. Figure 7 shows that for any M_1, the generalization error is significant as long as $|\mathcal{T}|_{tr} < M_1$, while the generalization error reaches around 0 as long as $|\mathcal{T}|_{tr} \geq M_1$ and $\mathcal{T}$ covers all the possibilities of IDR patterns and labels. This verifies that Condition 3.2 can be met with a fraction of $(M_1 - 1)^{-1/2}$ of total number of in-domain tasks.

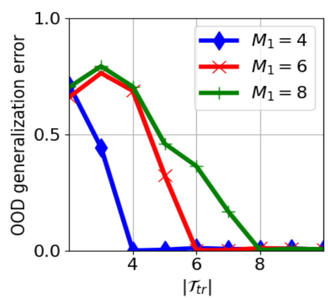


Figure 7. The required number of training tasks for different M_1.

D. Proofs of the Main Theorems

We first provide several useful definitions and key lemmas for the proof of the main theorems. Table 2 shows a summary of notations used in the proof.

D.1. Proof Overview of Main Theorems

This section illustrates how Corollary 4.3 and Proposition 4.5 contribute to the final in- and out-of-domain generalization performance of ICL.

The establishment of generalization

1. (*Self-Attention*) We can deduce from Corollary 4.3 that, for a query with IDR pattern $\boldsymbol{\mu}_j$ ($j \in [M_1]$) and label $+1$, the weighted summation of contexts and the query by the attention score, i.e., $\sum_{i=1}^{l} \boldsymbol{p}_i \text{attn}(\Psi; \boldsymbol{P}, i)$, is close to $[\boldsymbol{\mu}_j^\top, \boldsymbol{q}^\top]^\top$. This is because as long as the training/testing prompt length satisfies (9), large attention weights are assigned on $\boldsymbol{p}_i$ of which

Table 2. Summary of Notations

Notations	Annotation
$\boldsymbol{x}_s^n$, $\boldsymbol{y}_s^n$	$\boldsymbol{x}_s^n$ is the data for classification. $\boldsymbol{y}_s^n$ is the embedding of the label for $\boldsymbol{x}_s^n$.
$\boldsymbol{P}^n$, z^n	$\boldsymbol{P}^n$ is a prompt that consists of l pairs of $\boldsymbol{x}_s^n$ and $\boldsymbol{y}_s^n$, $s \in [l]$. The last column of $\boldsymbol{P}^n$ contains $\boldsymbol{p}_{query}^n$, which is the query of $\boldsymbol{P}^n$. $z^n \in \{+1, -1\}$ is the binary label of $\boldsymbol{p}_{query}^n$, which is also the label of $\boldsymbol{P}^n$ when we formulate the problem as a supervised learning problem.
$F(\Psi; \boldsymbol{P}^n)$, $\ell(\Psi; \boldsymbol{P}^n, z^n)$	$F(\Psi; \boldsymbol{P}^n)$ is the Transformer output for $\boldsymbol{P}^n$ with Ψ as the parameter. $\ell(\Psi; \boldsymbol{P}^n, z^n)$ is the loss function value given $\boldsymbol{P}^n$ and the corresponding label z^n.
$\boldsymbol{p}_s^n$, $\boldsymbol{\mu}_j$, $\boldsymbol{\nu}_k$	$\boldsymbol{p}_s^n$ is the s-th example with the corresponding label in $\boldsymbol{P}^n$. If $s = query$, $\boldsymbol{p}_s^n$ is the query. $\boldsymbol{\mu}_j$ and $\boldsymbol{\nu}_k$ are the IDR and IDI patterns in the feature embedding of $\boldsymbol{p}_s^n$ as the corresponding coefficients, respectively.
$\boldsymbol{q}$	$\boldsymbol{q}$ is the label space embedding.
M_1, M_2, M	M_1 is the number of IDR patterns. M_2 is the number of IDI patterns. $M = M_1 + M_2$.
α, a	α is the probability of selecting examples that contain either of the two decisive IDR patterns in each $\boldsymbol{P}^n$. $a = 1/\lvert a_i \rvert$ where a_i is the entry of each neuron in $\boldsymbol{W}_O$. $a = m$.
κ, κ', K, K', β	κ and κ' are the coefficients of the IDI pattern and the ODI pattern in the input $\boldsymbol{x}$, respectively. κ and κ' follow uniform distribution $U(-K, K)$ and $U(-K', K')$ with $K \leq 1/2$ and $K' \leq \mathcal{O}(1)$, respectively. β is the norm of in-/out-of-domain-(ir)relevant (IDR/ODR/IDI/ODI) patterns.
$\mathcal{W}_n, \mathcal{U}_n$	The sets of lucky neurons. $\mathcal{W}_n$ is the set of neurons of $\boldsymbol{W}_O$ that can activate the terms inside Relu$(\cdot)$ in $F(\Psi; \boldsymbol{P}^n)$ for $z^n = +1$ at initialization. $\mathcal{U}_n$ is the set of neurons of $\boldsymbol{W}_O$ that can activate the Relu part of $F(\Psi; \boldsymbol{P}^n)$ for $z^n = -1$ at initialization.
$\mathcal{W}, \mathcal{U}$	$\mathcal{W} = \cup_{n \in [N]} \mathcal{W}_n$. $\mathcal{U} = \cup_{n \in [N]} \mathcal{U}_n$
$\mathcal{N}_j^n$	The set of examples in $\boldsymbol{P}^n$ that contains $\boldsymbol{\mu}_j$ as the IDR pattern.
γ_t	γ_t is the summation of attention weight on examples that have different IDR patterns from the query.
ζ_t	ζ_t is smallest positive value inside the Relu$(\cdot)$ in $F(\Psi; \boldsymbol{P}^n)$ for all the $\boldsymbol{W}_O$ neuron and all $n \in [N]$.
$\mathcal{B}_b$	$\mathcal{B}_b$ is the SGD batch at the b-th iteration.
l_{tr}	l_{tr} is the prompt length of the training data.
l_{ts}	l_{ts} is the prompt length of the testing data.
$\mathcal{O}()$, $\Omega()$, $\Theta()$	We follow the convention that $f(x) = O(g(x))$ (or $\Omega(g(x))$, $\Theta(g(x))$)) means that $f(x)$ increases at most, at least, or in the order of $g(x)$, respectively.
$\gtrsim$, $\lesssim$	$f(x) \gtrsim g(x)$ (or $f(x) \lesssim g(x)$) means that $f(x) \geq \Omega(g(x))$ (or $f(x) \lesssim \mathcal{O}(g(x))$).

the IDR pattern is $\boldsymbol{\mu}_j$, and the label embedding is $\boldsymbol{q}$ by (20). Similarly, if its label is -1, the weighted summation of contexts and the query outputs $[\boldsymbol{\mu}_j^\top, -\boldsymbol{q}^\top]^\top$.

2. (*MLP*) By Proposition 4.5, we know that a large enough proportion of positive (or negative) neurons $i \in [m]$ have the label embedding of $\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)} \boldsymbol{W}_V^{(T)}$ close to $\pm\boldsymbol{q}$ (22). They can thus map the weighted summation of contexts and the query by attention with $+\boldsymbol{q}$ (or $-\boldsymbol{q}$) to positive (or negative) values. This leads to a correct prediction in-domain (Theorem 3.3).

3. (*Out-of-Domain Generalization*) Since Corollary 4.3 also applies to ODR patterns, then for a query with an ODR pattern $\boldsymbol{\mu}'_j$, $j \in [M'_1]$, the resulting weighted summation of contexts and the query is close to $[{\boldsymbol{\mu}'_j}^\top, \boldsymbol{q}^\top]^\top$ or $[{\boldsymbol{\mu}'_j}^\top, -\boldsymbol{q}^\top]^\top$. Then, by combining (21), (22) and the condition on ODR pattern characterized in (12), we can ensure that the MLP layer produces a desired prediction out of the domain (Theorem 3.4).

D.2. Preliminaries

Lemma D.1. *(Multiplicative Chernoff bounds, Theorem D.4 of (Mohri et al., 2018)) 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* $\hat{p} = \frac{1}{m}\sum_{i=1}^m X_i$*:*

$$\Pr(\hat{p} \geq (1+\gamma)p) \leq e^{-\frac{mp\gamma^2}{3}}, \tag{25}$$

$$\Pr(\hat{p} \leq (1-\gamma)p) \leq e^{-\frac{mp\gamma^2}{2}}. \tag{26}$$

Definition D.2. (Vershynin, 2010) We say X is a sub-Gaussian random variable with sub-Gaussian norm $K > 0$, if $(\mathbb{E}|X|^p)^{\frac{1}{p}} \leq K\sqrt{p}$ for all $p \geq 1$. In addition, the sub-Gaussian norm of X, denoted $\|X\|_{\psi_2}$, is defined as $\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-\frac{1}{2}}(\mathbb{E}|X|^p)^{\frac{1}{p}}$.

Lemma D.3. *((Vershynin, 2010) Proposition 5.1, Hoeffding's inequality) Let* $X_1, X_2, \cdots, X_N$ *be independent centered sub-gaussian random variables, and let* $K = \max_i \|\boldsymbol{X}_i\|_{\psi_2}$*. Then for every* $\boldsymbol{a} = (a_1, \cdots, a_N) \in \mathbb{R}^N$ *and every* $t \geq 0$*, we have*

$$\Pr\Big(\Big|\sum_{i=1}^N a_i X_i\Big| \geq t\Big) \leq e \cdot \exp\Big(-\frac{ct^2}{K^2\|\boldsymbol{a}\|^2}\Big), \tag{27}$$

where $c > 0$ *is an absolute constant.*

Definition D.4. For any data index n and iteration t, we can find i such that $\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n) > 0$ by the initialization with high probability. Define

1. $\zeta_{i,n,t} := \boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)$.
2. $\zeta_{i,t} = \min_n\{\zeta_{i,n,t}\}$.
3. $\zeta_t = \min_i\{\zeta_{i,t}\}$.
4. $\gamma_{t,n} = 1 - \sum_{s\in\mathcal{N}_*^n}\text{softmax}((\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n)^\top(\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n))$.
5. $\gamma_t = \max_{n\in[N]}\{\gamma_{t,n}\}$.

Lemma D.5. *(gradient updates of* $\boldsymbol{W}_Q$ *and* $\boldsymbol{W}_K$*) By the SGD training method described in Section 2.3, we have the following equations. Given the definition of in-/out-of-domain data as in (1) and the in-/out-of-domain data distribution* $\mathcal{D}$ *in (6) and* $\mathcal{D}'$ *in (8), we study the gradient updates in the directions of queries or contexts. Note that we require* $m \gtrsim M_1^2$*,* $B \gtrsim M_1\log M_1$*,* $l = l_{tr} \gtrsim 1$*,* $\beta \in [1, O(1)]$*.*
We first consider the case when the feature embeddings of the query $\boldsymbol{x}_{query}$ *and the example* $\boldsymbol{x}_q, q \in [l]$ *are* $\boldsymbol{\mu}_j$*. The label embedding is* $\boldsymbol{0}$ *for the query and* $\pm\boldsymbol{q}$ *for non-query examples. Then, for any* $l, a \in [M_1]$*,* $k \in [M_2]$*,* $t_0 \geq 1$*, where* $\boldsymbol{\mu}_l$ *forms a task in* $\mathcal{T}_{tr}$ *with* $\boldsymbol{\mu}_j$ *and* $\boldsymbol{\mu}_a$ *does not,*

$$(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\sum_{b=0}^{t_0}\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top, \boldsymbol{0}^\top)^\top \gtrsim \eta\frac{1}{M_1}\sum_{b=0}^{t_0}\zeta_b\delta\gamma_b\beta^4, \tag{28}$$

$$\begin{aligned}&\Big|(\boldsymbol{\mu}_l^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{t_0-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top, \boldsymbol{0}^\top)^\top\Big|\\ \lesssim & e^{-\Theta((\frac{nt_0}{M_1})^2)}\Big|(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\sum_{b=0}^{t_0-1}\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top, \boldsymbol{0}^\top)^\top\Big|,\end{aligned} \tag{29}$$

$$\begin{aligned}&\Big|(\boldsymbol{\mu}_a^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{t_0-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top, \boldsymbol{0}^\top)^\top\Big|\\ \lesssim & \frac{1}{M_1}\Big|(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\sum_{b=0}^{t_0-1}\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top, \boldsymbol{0}^\top)^\top\Big|,\end{aligned} \tag{30}$$

$$\Big|(\boldsymbol{\nu}_k^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q^{(t)}}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top,\mathbf{0}^\top)^\top\Big|\lesssim\frac{1}{M_2}\Big|(\boldsymbol{\mu}_j^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q^{(t)}}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top,\mathbf{0}^\top)^\top\Big|, \tag{31}$$

$$(\boldsymbol{\mu}_j^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K^{(t)}}\Big|_{t=t_0+1}\boldsymbol{p}_q\gtrsim\eta\frac{1}{M_1}\sum_{b=0}^{t_0}\zeta_b\delta\gamma_b\beta^4, \tag{32}$$

$$\begin{aligned}&\Big|(\boldsymbol{\mu}_l^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K^{(t)}}\Big|_{t=t_0}\boldsymbol{p}_q\Big|\\ \lesssim& e^{-\Theta((\frac{\eta t_0}{M_1})^2)}\Big|(\boldsymbol{\mu}_j^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K^{(t)}}\Big|_{t=t_0}\boldsymbol{p}_q\Big|,\end{aligned} \tag{33}$$

$$\Big|(\boldsymbol{\mu}_a^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K^{(t)}}\Big|_{t=t_0}\boldsymbol{p}_q\Big|\lesssim\frac{1}{M_1}\Big|(\boldsymbol{\mu}_j^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K^{(t)}}\Big|_{t=t_0}\boldsymbol{p}_q\Big|, \tag{34}$$

$$\Big|(\boldsymbol{\nu}_k^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K^{(t)}}\Big|_{t=t_0}\boldsymbol{p}_q\Big|\lesssim\frac{1}{M_2}\Big|(\boldsymbol{\mu}_j^\top,\boldsymbol{q}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K^{(t)}}\Big|_{t=t_0}\boldsymbol{p}_q\Big|. \tag{35}$$

In the above, equations (28), (29), (30), and (31) characterize the directions of gradient updates of $\boldsymbol{W}_Q$ when projected with $(\boldsymbol{x}_{query}^\top,\mathbf{0})^\top$. Similarly, equations (32), (33), (34), and (35) characterize the directions of gradient updates of $\boldsymbol{W}_K$ when projected with $\boldsymbol{p}_q, q\in[l]$.

Lemma D.6. *(gradient updates of $\boldsymbol{W}_V$) For $\boldsymbol{p}_j^n$ defined in (1) and $t_0\geq 1$, if $l=l_{tr}\gtrsim\max\{1,\frac{1}{\alpha\beta^2}\}$ and $BT\gtrsim\Theta(M_1^2)$, $B\gtrsim M_1$, we have that for $\boldsymbol{p}_j$ of which the corresponding label embedding is $\boldsymbol{q}$,*

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_V^{(b)}}\boldsymbol{p}_j\\ =&\eta\sum_{b=0}^{t_0}(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}),\end{aligned} \tag{36}$$

where

$$-V_i(b)\gtrsim\beta^2(1-\gamma_t)/a,\quad i\in\mathcal{W}_n, \tag{37}$$

$$-V_i(b)\leq\frac{1}{\beta^2+1}V_j(b),\quad i\in\mathcal{U}_n, j\in\mathcal{W}_n, \tag{38}$$

$$|V_i(b)|\lesssim\sqrt{\frac{\log B}{B}}\cdot\frac{1}{a},\quad i\notin\mathcal{W}_n\cup\mathcal{U}_n. \tag{39}$$

If the corresponding label embedding is $-\boldsymbol{q}$, we have the that (36) holds with

$$-V_i(b)\gtrsim\beta^2(1-\gamma_t)/a,\quad i\in\mathcal{U}_n, \tag{40}$$

$$-V_i(b)\leq\frac{1}{\beta^2+1}V_j(b),\quad i\in\mathcal{W}_n, j\in\mathcal{U}_n, \tag{41}$$

$$|V_i(b)|\lesssim\sqrt{\frac{\log B}{B}}\cdot\frac{1}{a},\quad i\notin\mathcal{W}_n\cup\mathcal{U}_n. \tag{42}$$

We can also derive

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_V^{(t)}}(\boldsymbol{\nu}_k^\top,\mathbf{0}^\top)^\top\\ =:&\eta\sum_{b=0}^{t_0}(\sum_{i\in\mathcal{W}_n}V_i'(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\mathcal{U}_n}V_i'(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i'(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}),\end{aligned} \tag{43}$$

where

$$|V_i'(b)| \leq |V_i(b)| \cdot \frac{1}{M_2}. \tag{44}$$

Lemma D.7. *(gradient updates of $\boldsymbol{W}_O$) We are given $\Theta(1) \geq \beta \geq 1$ and $m \gtrsim M_1^2$, $BT \gtrsim M_1 \log M_1$, $B \gtrsim M_1$, $t = t_0 \geq \Theta(1)$. Denote the set of examples that share the same IDR pattern as $\boldsymbol{p}_{query}^n$ as $\mathcal{B}_b^n$ in the b-th iteration. For $i \in \mathcal{W}$, $b \neq a$, and $\boldsymbol{p}_{query}^n$ corresponding to $\boldsymbol{q}$ and $\boldsymbol{\mu}_a$,*

$$\eta \frac{1}{|\mathcal{B}_b^n|} \sum_{n \in \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)}} (\boldsymbol{\mu}_a^\top, \boldsymbol{q}^\top)^\top = \delta(\beta^2+1)\frac{\alpha\eta}{2a}(1+\frac{\eta^2 m}{a^2})^{t_0}, \tag{45}$$

$$\eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)}} (\boldsymbol{\mu}_a^\top, \boldsymbol{q}^\top)^\top \gtrsim \delta(\beta^2+1)\frac{\alpha\eta t_0}{2a}. \tag{46}$$

For $i \in \mathcal{U}$, $b \neq a$, and $\boldsymbol{p}_{query}^n$ corresponding to $\boldsymbol{q}$ and $\boldsymbol{\mu}_a$,

$$\eta \frac{1}{|\mathcal{B}_b^n|} \sum_{n \in \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)}} (\boldsymbol{\mu}_a^\top, -\boldsymbol{q}^\top)^\top = \delta(\beta^2+1)\frac{\alpha\eta}{2a}(1+\frac{\eta^2 m}{a^2})^{t_0}, \tag{47}$$

$$\begin{aligned}&\eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}} (\boldsymbol{\mu}_a^\top, -\boldsymbol{q}^\top)^\top \\ \gtrsim &\delta(\beta^2+1)\frac{\alpha\eta t_0}{2a}.\end{aligned} \tag{48}$$

For $i \in \mathcal{W} \cup \mathcal{U}$ and $c \in [M_2]$,

$$\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)}\| \gtrsim \sqrt{M_1}\delta(\beta^2+1)^{\frac{1}{2}}\frac{\alpha\eta t_0}{2a}, \tag{49}$$

$$\eta \frac{1}{B} \sum_{b=0}^{t_0} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}} (\boldsymbol{\nu}_c^\top, \pm\boldsymbol{q}^\top)^\top \leq \frac{1}{M_2}\eta \frac{1}{B} \sum_{b=0}^{t_0} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}} (\boldsymbol{\mu}_b^\top, \boldsymbol{q}^\top)^\top. \tag{50}$$

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

$$\eta \frac{1}{B} \sum_{b=0}^{t_0} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}} (\boldsymbol{\mu}_a^\top, \pm\boldsymbol{q}^\top)^\top \leq \eta t \sqrt{\frac{\log B}{B}}\frac{1}{a}. \tag{51}$$

Definition D.8. Define

$$\boldsymbol{V}^n(t) := \sum_{s=1}^{l+1} \boldsymbol{W}_V^{(t)} \boldsymbol{p}_s^n \text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{query}^n), \tag{52}$$

for $\boldsymbol{P}^n$. Let $\boldsymbol{W}_{O_{(i,\cdot)}} = (\boldsymbol{O}_{i,1}, \boldsymbol{O}_{i,2}, \boldsymbol{0}^\top)$ where $\boldsymbol{O}_{i,1} \in \mathbb{R}^{d_\mathcal{X}}, \boldsymbol{O}_{i,2} \in \mathbb{R}^{d_\mathcal{Y}}$. Let $\boldsymbol{V}^n(t) = (\boldsymbol{V}_{n,1}(t)^\top, \boldsymbol{V}_{n,2}(t)^\top, \boldsymbol{0}^\top)^\top$ where $\boldsymbol{V}_{i,1}(t) \in \mathbb{R}^{d_\mathcal{X}}, \boldsymbol{V}_{i,2}(t) \in \mathbb{R}^{d_\mathcal{Y}}$. Define $\mathcal{W}_n, \mathcal{U}_n$ as the sets of lucky neurons such that

$$\mathcal{W}_n = \{i : \boldsymbol{O}_{i,1}^{(0)} \boldsymbol{V}_{n,1}(0) > 0, \boldsymbol{O}_{i,2}^{(0)} \boldsymbol{V}_{n,2}(0) > 0, a_i > 0\}, \tag{53}$$

$$\mathcal{U}_n = \{i : \boldsymbol{O}_{i,1}^{(0)} \boldsymbol{V}_{n,1}(0) > 0, \boldsymbol{O}_{i,2}^{(0)} \boldsymbol{V}_{n,2}(0) > 0, a_i < 0\}. \tag{54}$$

Define

$$\mathcal{N}_j^{n,i} = \{i : i \in [l+1], \boldsymbol{x}_i^n = \boldsymbol{\mu}_j + \kappa_i^n \boldsymbol{\nu}_k + \boldsymbol{n}_i^n, k \in [M_2]\}, \tag{55}$$

$$\mathcal{M}_k^{n,i} = \{i : i \in [l+1], \boldsymbol{x}_i^n = \boldsymbol{\mu}_j + \kappa_i^n \boldsymbol{\nu}_k + \boldsymbol{n}_i^n, j \in [M_1]\}, \tag{56}$$

as the set of example inputs with $\boldsymbol{\mu}_j$ as the IDR patterns or with $\boldsymbol{\nu}_k$ as the IDI patterns, respectively.

$$\mathcal{W} = \bigcup_{n=1}^N \mathcal{W}_n, \quad \mathcal{U} = \bigcup_{n=1}^N \mathcal{U}_n. \tag{57}$$

Lemma D.9. *By the definition of lucky neurons in (53) and (54), and the initialization described in Section 2.3, the number of lucky neurons $|\mathcal{W}_n|$, $|\mathcal{U}_n|$ satisfies*

$$|\mathcal{W}_n|,\ |\mathcal{U}_n| \geq \Omega(m). \tag{58}$$

Hence,

$$|\mathcal{W}|,\ |\mathcal{U}| \geq \Omega(m). \tag{59}$$

Lemma D.10. *Under the condition that $m \gtrsim M_1^2 \log M_1$, we have the following results.*

1. *When $t \geq 0$, for $V^n(t)$ where $\boldsymbol{p}_{query}^n$ corresponds to the label $+1$,*

$$\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)} \boldsymbol{V}^n(t)] = 1, i \in \mathcal{W}_n, \tag{60}$$

for $V^n(t)$ where $\boldsymbol{p}_{query}^n$ corresponds to the label -1,

$$\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)} \boldsymbol{V}^n(t)] = 1, i \in \mathcal{U}_n. \tag{61}$$

2. *When $t \geq \Theta(1)$, for $i \in \mathcal{W}$, we have that for $V^n(t)$ where $\boldsymbol{p}_{query}^n$ corresponds to the label $+1$,*

$$\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)} \boldsymbol{V}^n(t)] = 1. \tag{62}$$

For $i \in \mathcal{U}$, we have that for $V^n(t)$ where $\boldsymbol{p}_{query}^n$ corresponds to the label -1,

$$\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)} \boldsymbol{V}^n(t)] = 1. \tag{63}$$

Lemma D.11. *With in-domain tasks defined in Definition 3.1 and Condition 3.2, the number of training tasks should satisfy $|\mathcal{T}_{tr}| \geq M_1$ to make Condition 3.2 hold.*

D.3. Proof of Theorem 3.3

Proof. We first look at the required length of the prompt. Define m_i as the corresponding IDR pattern in the i-th demonstration. Consider the categorical 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 D.1, we can obtain

$$\Pr\left(\frac{1}{l_{tr}}\sum_{i=1}^{l_{tr}} \mathbb{1}[m_i = \boldsymbol{\mu}_a] \leq (1-c)\frac{\alpha}{2}\right) \leq e^{-l_{tr}c^2\frac{\alpha}{2}} = M_1^{-C}, \tag{64}$$

for some $c \in (0,1)$ and $C > 0$. Hence, with a high probability, combining the condition $l_{tr} \geq (\alpha\beta^2)^{-1}$ in Lemma D.6,

$$l_{tr} \gtrsim \max\left\{\Omega\left(\frac{2\log M_1}{\alpha}\right), \Omega\left(\frac{1}{\alpha\beta^2}\right)\right\}. \tag{65}$$

By the condition in Lemma D.5, we have that

$$B \geq \Omega(M_1 \log M_1). \tag{66}$$

We know that there exists gradient noise caused by imbalanced IDR patterns in each batchTherefore, by Hoeffding's inequality (27), for any $\boldsymbol{W} \in \Psi$,

$$\Pr\left(\Big\|\frac{1}{|\mathcal{B}_b|}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\Psi;\boldsymbol{P}^n,z^n)}{\partial \boldsymbol{W}} - \mathbb{E}\left[\frac{\partial \ell(\Psi;\boldsymbol{P}^n,z^n)}{\partial \boldsymbol{W}}\right]\Big\| \geq \Big|\mathbb{E}\left[\frac{\partial \ell(\Psi;\boldsymbol{P}^n,z^n)}{\partial \boldsymbol{W}}\right]\epsilon\right) \leq e^{-B\epsilon^2} \leq M_1^{-C}, \tag{67}$$

if $B \gtrsim \epsilon^{-2}\log M_1$. Therefore, we require

$$B \gtrsim \max\{\epsilon^{-2}, M_1\}\log M_1. \tag{68}$$

(a) We have that for i such that $a_i > 0$ but $i \notin \mathcal{W}$ by the definition of the Relu function,

$$a_i \text{Relu}(\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)} \sum_{s=1}^{l+1} (\boldsymbol{W}_V^{(T)} \boldsymbol{p}_s^n) \text{softmax}((\boldsymbol{W}_K^{(T)} \boldsymbol{p}_s^n)^\top (\boldsymbol{W}_Q^{(T)} \boldsymbol{p}_{query}^n))) \geq 0. \tag{69}$$

(b) Furthermore, we have that for $i \in \mathcal{W}$, and for $\boldsymbol{p}_s^n$ that shares the same IDR pattern as $\boldsymbol{p}_{query}^n$, with a high probability of $1 - M_1^{-C}$,

$$\begin{aligned} &\eta \sum_{b=0}^{T-1} \boldsymbol{W}_{O_{(j,\cdot)}}^{(T)} \sum_{j \in \mathcal{W}_n} \boldsymbol{W}_{O_{(j,\cdot)}}^{(b)}{}^\top \\ \geq &\eta \sum_{b=0}^{T-1} M_1 \delta (\beta^2 + 1)^{\frac{1}{2}} \frac{\alpha \eta T}{2a} \cdot \delta (\beta^2+1)^{\frac{1}{2}} \frac{\alpha \eta b}{2a} \\ \gtrsim &\delta^2 (\beta^2 + 1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2}, \end{aligned} \tag{70}$$

where the first step comes from (49) in Lemma D.7, and the second step is by $\sum_{b=0}^{T-1} b = \Theta(T^2)$. Then, we can obtain

$$\begin{aligned} &\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)} \boldsymbol{W}_V^{(T)} \boldsymbol{p}_s^n \\ =&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)} (\delta \boldsymbol{p}_s^n + \sum_{b=0}^{T-1} \eta (\sum_{i \in \mathcal{W}_n} V_i(b) \boldsymbol{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \boldsymbol{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top) \\ \gtrsim &\delta^2 (\beta^2+1) \frac{\alpha \eta T}{2a} + \delta^2 (\beta^2 + 1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2}, \end{aligned} \tag{71}$$

where the first step is by (36) in Lemma D.6, and the last step comes from Lemma D.7.

Therefore, by combining Lemma D.9 and Lemma D.10, we have that

$$\begin{aligned} &\sum_{i \in \mathcal{W}} a_i \text{Relu}(\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)} \sum_{s=1}^{l+1} (\boldsymbol{W}_V^{(T)} \boldsymbol{p}_s^n) \text{softmax}((\boldsymbol{W}_K^{(T)} \boldsymbol{p}_s^n)^\top (\boldsymbol{W}_Q^{(T)} \boldsymbol{p}_{query}^n))) \\ \gtrsim &(1 - \gamma_T) \cdot \Big(\delta^2 (\beta^2 + 1) \frac{\alpha \eta T}{2a} + \delta^2 (\beta^2+1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2} \Big), \end{aligned} \tag{72}$$

when γ_T is order-wise smaller than 1.We next give a bound for γ_T, which is give by Definition D.4,

$$\gamma_T \geq 1 - \sum_{s \in \mathcal{N}_*^n} \text{softmax}((\boldsymbol{W}_K^{(T)} \boldsymbol{p}_s^n)^\top (\boldsymbol{W}_Q^{(T)} \boldsymbol{p}_{query}^n)), \tag{73}$$

from Defition D.4 for $\boldsymbol{\mu}_j$ as the IDR pattern in $\boldsymbol{p}_{query}^n$. We can tell from (72) and Definition D.4,

$$\zeta_b \gtrsim \delta^2 (\beta^2 + 1) \frac{\alpha \eta T}{2a} + \delta^2 (\beta^2 + 1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2}. \tag{74}$$

Then, if $\zeta_T \gtrsim 1$ and $T \gtrsim M_1$, by Lemma D.5, with high probability,

$$\begin{aligned} &(\boldsymbol{W}_K^{(T)} \boldsymbol{p}_s^n)^\top (\boldsymbol{W}_Q^{(T)} \boldsymbol{p}_{query}^n) \\ \gtrsim &(\eta \frac{1}{M_1} \sum_{b=0}^{T-1} \zeta_b \delta \gamma_b \beta^2 + \delta)^2 \\ \gtrsim &(\eta \sum_{b=0}^{T-1} \gamma_b \beta^2 \cdot (\delta^2 \frac{\alpha \eta T}{2a} + \delta^2 (\beta^2 + 1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2}) + \delta)^2 \\ := &(\eta \sum_{b=0}^{T-1} \gamma_b \cdot \Delta_T + \delta)^2, \end{aligned} \tag{75}$$

where the first step comes from the fact that the gradient update projections of $\boldsymbol{W}_Q$ and $\boldsymbol{W}_K$ onto queries or examples are close to the corresponding IDR pattern the most by Lemma D.5. In the last inequality of (75), we only consider the term related to T and γ_b. For any $\boldsymbol{p}_l^n$ that shares a different IDR pattern as $\boldsymbol{p}_{query}^n$, we have

$$(\boldsymbol{W}_K^{(T)}\boldsymbol{p}_l^n)^\top(\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}^n)\lesssim\frac{1}{M_1}(\boldsymbol{W}_K^{(T)}\boldsymbol{p}_s^n)^\top(\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}^n), \tag{76}$$

by Lemma D.5. Then, given the definition of softmax,

$$\begin{aligned}&\sum_{s\in\mathcal{N}_j^n}\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ \geq&\frac{\sum_{s\in\mathcal{N}_j^n}e^{\Theta(\delta^2)+(\eta\sum_{b=0}^{T-1}\gamma_b\cdot\Delta_T)^2}}{\sum_{s\in\mathcal{N}_j^n}e^{\Theta(\delta^2)+(\eta\sum_{b=0}^{T-1}\gamma_b\cdot\Delta_T)^2}+\sum_{s\in[l]-\mathcal{N}_j^n}e^{\frac{1}{M_1}(\eta\sum_{b=0}^{T-1}\gamma_b\cdot\Delta_T)^2}}\\ \geq&1-\frac{2-\alpha}{\alpha}e^{-(\eta\sum_{b=0}^{T-1}\gamma_b\cdot\Delta_T)^2},\end{aligned}\tag{77}$$

where the first step is by (75) and (76). Combining with (73), we can derive

$$\begin{aligned}\gamma_T\leq&\frac{2-\alpha}{\alpha}e^{-(\eta\sum_{b=0}^{T-1}\gamma_b\cdot\Delta_T)^2}=\frac{2-\alpha}{\alpha}e^{-(\eta\sum_{b=0}^{T-2}\gamma_b\cdot\Delta_T)^2}\cdot e^{-\eta^2\Delta_T^2(2\gamma_{T-1}\sum_{b=0}^{T-2}\gamma_b+\gamma_{T-1}^2)}\\ =&\gamma_{T-1}\cdot e^{-\eta^2\Delta_T^2(2\gamma_{T-1}\sum_{b=0}^{T-2}\gamma_b+\gamma_{T-1}^2)}.\end{aligned}\tag{78}$$

When T is large, γ_T is approaching zero. Hence, the equality of (78) is close to being achieved, in which case,

$$\gamma_T\approx\gamma_{T-1}\cdot e^{-\eta^2\Delta_T^2(2\gamma_{T-1}\sum_{b=0}^{T-2}\gamma_b+\gamma_{T-1}^2)}.\tag{79}$$

We can observe that when $\sum_{b=0}^{t_0-1}\eta\gamma_b\Delta_T\geq\sqrt{\log M_1}$, γ_{t_0} reaches $\Theta(1/M_1\cdot\frac{2-\alpha}{\alpha})$. Similarly, when $\sum_{b=0}^{t_0'-1}\eta\gamma_b\Delta_T\leq\sqrt{\log C}$ for some $C>1$, $\gamma_{t_0'}$ is still $\Theta(1)$, which indicates $t_0'\lesssim\eta^{-1}M_1\sqrt{\log C}$ if we only care about η and M_1 as variables. Therefore, we require that the final T satisfies $T\gtrsim\eta^{-1}M_1\sqrt{\log M_1}$.

(c) We next look at i where $a_i<0$. If $i\in\mathcal{U}$, 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,

$$\begin{aligned}&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n\\ \lesssim&-\delta^2\frac{\alpha\eta T}{2a}-\delta^2\alpha^2\frac{(\eta T)^3M_1}{a^2},\end{aligned}\tag{80}$$

if $\beta\leq\Theta(1)$. Hence,

$$\text{Relu}(\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n)\text{softmax}((\boldsymbol{W}_K^{(T)}\boldsymbol{p}_s^n)^\top(\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}^n)))=0.\tag{81}$$

(d) If $i\notin\mathcal{W}\cup\mathcal{U}$ and $s\in\mathcal{W}$, we have,

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n\lesssim\frac{1}{\sqrt{M_1}}\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^n,\tag{82}$$

by Lemma D.6 and $B\gtrsim M_1$. The final lower bound of $F(\Psi;\boldsymbol{P}^n)$ is based on the lower bound introduced by $i\in\mathcal{W}$. Then, combining (69), (72), (81), and (82), we can derive

$$\begin{aligned}&F(\Psi;\boldsymbol{P}^n)\\ \gtrsim&(1-\gamma_T)\cdot\Big(\delta^2(\beta^2+1)\frac{\alpha\eta T}{2a}+\delta^2(\beta^2+1)\alpha^2\frac{(\eta T)^3M_1}{a^2}\Big).\end{aligned}\tag{83}$$

Therefore, as long as

$$T=\Theta(\eta^{-1}M_1\delta^{-2/3}\beta^{-2/3}\alpha^{-2/3}\sqrt{\log M_1}),\tag{84}$$

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

$$F(\Psi; \boldsymbol{P}^n) \geq 1. \tag{85}$$

Similarly, we can derive that for $z^n = -1$,

$$F(\Psi; \boldsymbol{P}^n) \leq -1. \tag{86}$$

Then, we study in-domain generalization. By (67), for any given testing prompt embedding $\boldsymbol{P}$ with $z = +1$, we have

$$F(\Psi; \boldsymbol{P}) \geq 1 - \epsilon, \tag{87}$$

and if $z = -1$,

$$F(\Psi; \boldsymbol{P}) \leq -1 + \epsilon. \tag{88}$$

Therefore,

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{query} \sim \mathcal{D}, f \in \mathcal{T}} [\ell(\Psi; \boldsymbol{P}, y)] \leq \epsilon. \tag{89}$$

□

D.4. Proof of Theorem 3.4

Proof. Note that we require that the fraction of contexts with the same ODR pattern as the query is at least α'. Since we need that there exists at least one context that contains the same ODR pattern as the query, we have

$$l_{ts} \geq \frac{1}{\alpha'}. \tag{90}$$

Consider ${\boldsymbol{p}_{query}^n}'$ such that the label is $+1$. Let $\boldsymbol{\mu}_j' = \sum_{j=1}^{M_1} c_j \boldsymbol{\mu}_j$ where $\sum_{j=1}^{M_1} c_j^2 = 1$ and $\boldsymbol{\nu}_k' = \sum_{j=1}^{M_2} g_j \boldsymbol{\nu}_j$ where $\sum_{j=1}^{M_2} g_j^2 = 1$. Following the derivation of (74) and (75), we have that for $s \in \mathcal{N}^n$,

$$\begin{aligned} &(\boldsymbol{W}_K^{(T)} {\boldsymbol{p}_s^n}')^\top \boldsymbol{W}_Q^{(T)} {\boldsymbol{p}_{query}^n}' \\ \gtrsim & \sum_{j=1}^{M_1} c_j^2 \cdot (\eta \sum_{b=0}^{T-1} \gamma_b \beta^2 \cdot (\delta^2 \frac{\alpha \eta T}{2a} + \delta^2 (\beta+1)^2 \alpha^2 \frac{(\eta T)^3 M_1}{a^2}))^2 \\ \gtrsim & \log M_1. \end{aligned} \tag{91}$$

For ODR patterns, by Proposition 4.1, we have for $\boldsymbol{p}_l^n$ that has a different ODR pattern than $\boldsymbol{p}_{query}^n$,

$$(\boldsymbol{W}_K^{(T)} \boldsymbol{p}_l^n)^\top (\boldsymbol{W}_Q^{(T)} \boldsymbol{p}_{query}^n) \lesssim (\frac{1}{M_1} + \frac{1}{M_2}) (\boldsymbol{W}_K^{(T)} \boldsymbol{p}_s^n)^\top (\boldsymbol{W}_Q^{(T)} \boldsymbol{p}_{query}^n). \tag{92}$$

Therefore, by similarly defining $\mathcal{N}_j^n = \{{\boldsymbol{p}_s^n}' : \text{The testing-relevant pattern of } \boldsymbol{P}^n \text{ is } \boldsymbol{\mu}_j'\}$, we can derive

$$\sum_{s \in \mathcal{N}_j^n} \text{softmax}((\boldsymbol{W}_K^{(T)} {\boldsymbol{p}_s^n}')^\top (\boldsymbol{W}_Q^{(T)} {\boldsymbol{p}_{query}^n}')) \geq 1 - \frac{2 - \alpha'}{\alpha'} \Theta(\frac{1}{M_1}) \geq 1 - \frac{2}{\alpha'} \Theta(\frac{1}{M_1}). \tag{93}$$

Note that for ${\boldsymbol{p}_s^n}' = \sum_{i=1}^{M_1} c_i \boldsymbol{\mu}_i + \sum_{j=1}^{M_2} {\kappa_s^n}' g_j \boldsymbol{\nu}_j$, when $M_1 \geq M_2$, we can find a set of $\boldsymbol{\mu}_j + {\kappa_s^n}' \boldsymbol{\nu}_j$ from $j = 1$ to $j = M_2$ with g_j as the coefficients. When $M_1 < M2 = \Theta(M_1)$, we can find a set of $\boldsymbol{\mu}_t + {\kappa_s^n}' \boldsymbol{\nu}_j$ from $j = 1$ to $j = M_2$ with $t \in [M_1]$, g_j as the coefficients likewise. The remaining $\boldsymbol{\mu}_i$ has coefficients of which the summation is smaller than 1.

Therefore, we have that for a certain $i \in \mathcal{W}$ and $\boldsymbol{p}_s^n$ where the corresponding label-space embedding is $\boldsymbol{q}$,

$$\begin{aligned}
&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\boldsymbol{p}_s^{n\prime}\\
=&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}(\sum_{i=1}^{M_1}c_i\boldsymbol{\mu}_i^\top+\kappa_s^{n\prime}\sum_{j=1}^{M_2}g_j\boldsymbol{\nu}_j^\top+\boldsymbol{o}_s^{n\top},\boldsymbol{q}^\top)^\top\\
\gtrsim&\sum_{i=1}^{M_1}c_i(\delta^2\beta^2\frac{\alpha\eta T}{2a}+\delta^2(\beta^2\alpha^2\frac{(\eta T)^3M_1}{a^2})\\
&+(\delta^2\frac{\alpha\eta T}{2a}+\delta^2\alpha^2\frac{(\eta T)^3M_1}{a^2})(1-\epsilon)\\
\geq&(\delta^2(\beta^2+1)\frac{\alpha\eta T}{2a}+\delta^2(\beta^2+1)\alpha^2\frac{(\eta T)^3M_1}{a^2})\cdot(1-\epsilon),
\end{aligned}\tag{94}$$

where the first equality comes from the definition of $\boldsymbol{p}_s^{n\prime}$. The first inequality of (94) is derived from (67). The last inequality is by the condition $\sum_{i=1}^{M_1}c_i\geq 1$. Therefore, we can derive that

$$\begin{aligned}
F(\Psi;\boldsymbol{P}^{n\prime})\gtrsim&(1-\gamma_T)(\delta^2(\beta^2+1)\frac{\alpha\eta T}{2a}+\delta^2(\beta^2+1)\alpha^2\frac{(\eta T)^3M_1}{a^2})\quad\cdot(1-\epsilon)\\
\geq&1-\epsilon,
\end{aligned}\tag{95}$$

where the first step is by following (85), and the remaining steps are from basic mathematical computation. Likewise, for ${\boldsymbol{p}_{query}^n}'$ such that the label is -1, we can obtain

$$F(\Psi;\boldsymbol{P}^{n\prime})<-(1-\epsilon).\tag{96}$$

Therefore, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{query}\sim\mathcal{D}',f\in\mathcal{T}'}[\ell(\Psi;\boldsymbol{P},y)]\leq\epsilon.\tag{97}$$

□

D.5. Proof of Theorem 3.7

Proof. We cover the proof in the proof of Proposition 4.5. Please see Section E.4 for more details. □

E. Proofs of Key Lemmas and propositions

E.1. Proof of Lemma D.11

Proof. We first show that if $|\mathcal{T}_{tr}|<M_1$, Condition 3.2 cannot hold. Then, We show that there exists $\mathcal{T}_{tr}$ with $|\mathcal{T}_{tr}|\geq M_1$ such that Condition 3.2 holds.

(1) If $|\mathcal{T}_{tr}<M_1|$, then $|\mathcal{T}_{tr}|/M_1<1$, which is contradict to $|\mathcal{T}_{tr}|/M_1\geq 1$ in Condition 3.2.

(2) The following example satisfies $|\mathcal{T}_{tr}|\geq M_1$. In this example, the i-th task function ($i\in[M_1]$) in $\mathcal{T}_{tr}$ maps the query with $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_{i+1}$ as IDR patterns to $+1$ and -1, respectively, where we denote $\boldsymbol{\mu}_{M_1+1}:=\boldsymbol{\mu}_1$. Hence, the numbers of tasks that map $\boldsymbol{\mu}_j$ to $+1$ and -1 are both 1 for any $j\in[M_1]$. In this case, $|\mathcal{T}_{tr}|=M_1$.

□

E.2. Proof of Proposition 4.1

Proof. We first show the results for in-domain patterns.
(1) We investigate the results about $\boldsymbol{W}_Q$ and then $\boldsymbol{W}_K$. For $\boldsymbol{p}_{query}$ with $\boldsymbol{\mu}_j$ as the IDR pattern and $\boldsymbol{a}\in$

$\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_{M_1}\}\backslash\{\boldsymbol{\mu}_j\}$, by (30), we have

$$\begin{aligned}&(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}\\ =&(\boldsymbol{a}^\top, \boldsymbol{0}^\top)(\boldsymbol{W}_Q^{(0)}+\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=b}\boldsymbol{p}_{query}\\ \lesssim&\frac{1}{M_1}\cdot(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=b}\boldsymbol{p}_{query},\end{aligned}\tag{98}$$

if $\boldsymbol{a}$ does not form a task in $\mathcal{T}_{tr}$ with $\boldsymbol{\mu}_j$. If $\boldsymbol{a}$ forms a task in $\mathcal{T}_{tr}$ with $\boldsymbol{\mu}_j$, and $\eta T=\Theta(M_1\sqrt{\log M_1})$, by (29)

$$\begin{aligned}&(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}\\ \lesssim&\frac{1}{M_1}\cdot(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=b}\boldsymbol{p}_{query}.\end{aligned}\tag{99}$$

For $\boldsymbol{a}\perp\boldsymbol{\mu}_j$ but $\boldsymbol{a}\notin\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_{M_1}\}$, by (31), we have

$$\begin{aligned}&(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)^\top\\ =&(\boldsymbol{a}^\top, \boldsymbol{0}^\top)(\boldsymbol{W}_Q^{(0)}+\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=b})\boldsymbol{p}_{query}\\ \lesssim&\frac{1}{M_2}\cdot(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=b}\boldsymbol{p}_{query}.\end{aligned}\tag{100}$$

By (28) and the initialization, we have

$$(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}\geq(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=b}\boldsymbol{p}_{query}\gtrsim\sqrt{\log M_1}+\delta\gtrsim\sqrt{\log M_1},\tag{101}$$

where the $\sqrt{\log M_1}$ in the second step comes from that $\eta T\geq\Theta(M_1)\sqrt{\log M_1}$. Hence, by combining (98), (99), and (101), we can derive that

$$\begin{aligned}\|\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}^n\|\lesssim&\sqrt{1+\frac{1}{M_1^2}\cdot M_1+\frac{1}{M_1^2}+\frac{1}{M_2^2}\cdot M_2}(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=b}\boldsymbol{p}_{query}\\ =&\sqrt{1+\frac{1}{M_1}+\frac{1}{M_2}}(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_Q^{(t)}}\Big|_{t=b}\boldsymbol{p}_{query},\end{aligned}\tag{102}$$

where in the first step, the first $1/M_1^2\cdot M_1$ comes from (98) with M_1-1 choices of $\boldsymbol{a}$. The second $1/M_1^2$ comes from (99), i.e., $(1/M_1)^2\cdot\Theta(1)$ since there are only a constant number of such cases. The third $1/M_2^2\cdot M_2$ is from (100) with M_2 choices of $\boldsymbol{a}$. Therefore, by (101) and (102), for $\boldsymbol{a}\in\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_{M_1}\}\backslash\{\boldsymbol{\mu}_j\}$, we have

$$(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}\lesssim\frac{\sqrt{\log M_1}}{M_1}.\tag{103}$$

For $\boldsymbol{a}\perp\boldsymbol{\mu}_j$ but $\boldsymbol{a}\notin\{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_{M_1}\}$,

$$(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}\lesssim\frac{\sqrt{\log M_1}}{M_2}.\tag{104}$$

For $\boldsymbol{W}_K$, we can make derivations following the above steps. For $\boldsymbol{p}_q$ with $\boldsymbol{\mu}_j$ as the IDR pattern and $\boldsymbol{a} \in \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_{M_1}\}\backslash\{\boldsymbol{\mu}_j\}$, by (34), we have

$$\begin{aligned}&(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_K^{(T)}\boldsymbol{p}_q\\ \lesssim&\frac{1}{M_1}\cdot(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_K^{(t)}}\Big|_{t=b}\boldsymbol{p}_q,\end{aligned}\tag{105}$$

if $\boldsymbol{a}$ does not form a task in $\mathcal{T}_{tr}$ with $\boldsymbol{\mu}_j$. If $\boldsymbol{a}$ forms a task in $\mathcal{T}_{tr}$ with $\boldsymbol{\mu}_j$, and $\eta T = \Theta(M_1\sqrt{\log M_1})$, by (33),

$$\begin{aligned}&(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_K^{(T)}\boldsymbol{p}_q\\ \lesssim&\frac{1}{M_1}\cdot(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_K^{(t)}}\Big|_{t=b}\boldsymbol{p}_q.\end{aligned}\tag{106}$$

For $\boldsymbol{a} \perp \boldsymbol{\mu}_j$ but $\boldsymbol{a} \notin \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_{M_1}\}$, by (35), we have

$$\begin{aligned}&(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_K^{(T)}\boldsymbol{p}_q\\ \lesssim&\frac{1}{M_2}\cdot(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_K^{(t)}}\Big|_{t=b}\boldsymbol{p}_q.\end{aligned}\tag{107}$$

By (32) and the initialization, we have

$$(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\boldsymbol{W}_K^{(T)}\boldsymbol{p}_q \geq \sqrt{\log M_1} + \delta \geq \sqrt{\log M_1}.\tag{108}$$

Hence, by combining (105), (106), and (107), we can derive that

$$\|\boldsymbol{W}_K^{(T)}\boldsymbol{p}_i^n\| \lesssim \sqrt{1 + \frac{1}{M_1} + \frac{1}{M_1^2} + \frac{1}{M_2}}(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{b=0}^{T-1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n; \Psi)}{\partial \boldsymbol{W}_K^{(t)}}\Big|_{t=b}(\boldsymbol{\mu}_j^\top, \pm\boldsymbol{q})^\top.\tag{109}$$

Therefore, by (108) and (109), for $\boldsymbol{a} \in \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_{M_1}\}\backslash\{\boldsymbol{\mu}_j\}$, we have

$$(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_K^{(T)}\boldsymbol{p}_q \lesssim \frac{\sqrt{\log M_1}}{M_1}.\tag{110}$$

For $\boldsymbol{a} \perp \boldsymbol{\mu}_j$ but $\boldsymbol{a} \notin \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_{M_1}\}$,

$$(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_K^{(T)}\boldsymbol{p}_q \lesssim \frac{\sqrt{\log M_1}}{M_2}.\tag{111}$$

(2) For out-of-domain patterns, we have the following derivation. Let $\boldsymbol{\mu}_j' = \sum_{i=1}^{M_1} k_{ji}\boldsymbol{\mu}_i$ where $\sum_{i=1}^{M_1} k_{ji} \geq 1$ and $\sum_{i=1}^{M_1} k_{ji}^2 = 1$. Then, for a query $\boldsymbol{p}_{query}'$, of which the corresponding ODR pattern is $\boldsymbol{\mu}_j'$, we have that by (28), (29), (30), and (31),

$$|(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}'| \geq |k_j|(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query}(1 - \frac{\Theta(1)}{M_1} - \frac{\Theta(1)}{M_2}),\tag{112}$$

$$|(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}'| \leq |k_j|(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}(1 + \frac{\Theta(1)}{M_1} + \frac{\Theta(1)}{M_2}),\tag{113}$$

for any $\boldsymbol{p}_{query}$ with $\boldsymbol{\mu}_j$ as the IDR pattern. Meanwhile,

$$|(\boldsymbol{\nu}_k^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}'| \leq \frac{1}{M_2}(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}.\tag{114}$$

Therefore,

$$\|\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}'\| \geq \sqrt{\log M_1}(1 - \frac{\Theta(1)}{M_1} - \frac{\Theta(1)}{M_2}) \gtrsim \sqrt{\log M_1},\tag{115}$$

$$\|\boldsymbol{W}_Q^{(t)}\boldsymbol{p}'_{query}\| \leq \sqrt{\log M_1}(1+\frac{\Theta(1)}{M_1}+\frac{\Theta(1)}{M_2}) \lesssim \sqrt{\log M_1}, \tag{116}$$

$$|({\boldsymbol{\mu}'_j}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}\boldsymbol{p}'_{query}| \gtrsim \sum_{i=1}^{M_1}|k_{ji}|\sqrt{\log M_1} \geq \sqrt{\log M_1}. \tag{117}$$

For $\boldsymbol{\mu}_a \in \{\boldsymbol{\mu}'_1, \cdots, \boldsymbol{\mu}'_{M'_1}\}\backslash\{\boldsymbol{\mu}'_j\}$, let $\boldsymbol{\mu}_a = \sum_{i=1}^{M_1} k_{ai}\boldsymbol{\mu}_i$, we have

$$(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query} \lesssim (\frac{1}{M_1}+\frac{1}{M_2})\sum_{i=1}^{M_1}|k_{ai}k_{ji}| \leq \sqrt{\log M_1}(\frac{1}{M_1}+\frac{1}{M_2}), \tag{118}$$

where the first step is by (112) and (113), and the second step is by Cauchy-Schwarz inequality given that $\sum_{i=1}^{M_1} k_{ji}^2 = \sum_{i=1} k_{ai}^2 = 1$. For $\boldsymbol{a} \perp \boldsymbol{\mu}'_j$ but $\boldsymbol{a} \notin \{\boldsymbol{\mu}'_1, \cdots, \boldsymbol{\mu}'_{M'_1}\}$,

$$(\boldsymbol{a}^\top, \boldsymbol{0}^\top)\boldsymbol{W}_Q^{(T)}\boldsymbol{p}_{query} \lesssim \sqrt{\log M_1}(\frac{1}{M_1}+\frac{1}{M_2}). \tag{119}$$

Likewise, we can derive the conclusion for the testing context with $\boldsymbol{W}_K^{(T)}$.

□

E.3. Proof of Corollary 4.3

Proof. From (75) to (79), we can derive the conclusion for IDR patterns. For ODR patterns, from (93), we can obtain the conclusion. Note that $\frac{2-\alpha}{\alpha} = \Theta(1)$ since $\alpha = \Theta(1)$. □

E.4. Proof of Proposition 4.5

Proof. For any i, we denote $\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)} = (\boldsymbol{O}_{i,1}^{(b)}, \boldsymbol{O}_{i,2}^{(b)}, \boldsymbol{0}^\top)$ where ${\boldsymbol{O}_{i,1}^{(b)}}^\top \in \mathbb{R}^{d_\mathcal{X}}$ and ${\boldsymbol{O}_{i,2}^{(b)}}^\top \in \mathbb{R}^{d_\mathcal{Y}}$. Following the derivation of (71), we can obtain that for $s \in [l]$ or s is the query,

$$\begin{aligned}
&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}(\boldsymbol{p}_s^n, \boldsymbol{0}^\top)^\top\\
=&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}(\delta(\boldsymbol{p}_s^n, \boldsymbol{0}^\top)^\top + \sum_{b=0}^{T-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{O}_{i,1}^{(b)} + \sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{O}_{i,1}^{(b)} + \sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{O}_{i,1}^{(b)})^\top)\\
\gtrsim&\delta^2\beta^2\frac{\alpha\eta T}{2a} + \delta^2\beta^2\alpha^2\frac{(\eta T)^3 M_1}{a^2},
\end{aligned} \tag{120}$$

for any $j \in [M_1]$, and

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}(\kappa_i^n\boldsymbol{\nu}_k^\top, \boldsymbol{0}^\top)^\top \leq \frac{1}{M_2}\cdot\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)^\top, \tag{121}$$

for $k \in [M_2]$ by (44) and (50) in Lemma D.6 and D.7, respectively. Then, we have

$$\begin{aligned}
\frac{(\frac{1}{M_1}\sum_{j=1}^{M_1}\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}}{\|(\frac{1}{M_1}\sum_{j=1}^{M_1}\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\|\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\|} \geq & \frac{1}{\sqrt{1+\frac{1}{M_2^2}\cdot M_2}}\\
\geq & 1-\frac{\Theta(1)}{M_2},
\end{aligned} \tag{122}$$

because $BT \geq \Theta(M_1^2)$. For any $i \in \mathcal{W}$,

$$\begin{aligned}
&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}(\boldsymbol{0}^\top, \boldsymbol{q}^\top)^\top\\
=&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}(\delta(\boldsymbol{0}^\top, \boldsymbol{q}^\top)^\top + \sum_{b=0}^{T-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{O}_{i,2}^{(b)} + \sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{O}_{i,2}^{(b)} + \sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{O}_{i,2}^{(b)})^\top)\\
\gtrsim&\delta^2\frac{\alpha\eta T}{2a} + \delta^2\alpha^2\frac{(\eta T)^3 M_1}{a^2}.
\end{aligned} \tag{123}$$

Note that by gradient updates of $\boldsymbol{W}_O$ and $\boldsymbol{W}_V$, there are no gradient components perpendicular to $\boldsymbol{p}$ except some Gaussian noise. Hence,

$$\begin{aligned}\frac{(\boldsymbol{0}^\top,\boldsymbol{q}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}}{\|(\boldsymbol{0}^\top,\boldsymbol{q}^\top)\|\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\|}\geq&\frac{1}{\sqrt{1+\xi}}\\ \geq&1-\frac{\Theta(1)}{M_1}.\end{aligned}\tag{124}$$

For any $i\in\mathcal{U}$,

$$\begin{aligned}&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}(\boldsymbol{0}^\top,-\boldsymbol{q}^\top)^\top\\ =&\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}(\delta(\boldsymbol{0}^\top,-\boldsymbol{q}^\top)^\top+\sum_{b=0}^{T-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{O}_{i,2}^{(b)}+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{O}_{i,2}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{O}_{i,2}^{(b)})^\top)\\ \gtrsim&\delta^2\frac{\alpha\eta T}{2a}+\delta^2\alpha^2\frac{(\eta T)^3M_1}{a^2}.\end{aligned}\tag{125}$$

Similarly to (124), we have

$$\frac{(\boldsymbol{0}^\top,-\boldsymbol{q}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}}{\|(\boldsymbol{0}^\top,-\boldsymbol{q}^\top)\|\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\|}\geq1-\frac{\Theta(1)}{M_1}.\tag{126}$$

Hence, for $i\in\mathcal{W}\cup\mathcal{U}$,

$$\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\|\gtrsim\beta^{-1}=\Omega(1).\tag{127}$$

By (82), we have that for $i\notin\mathcal{W}\cup\mathcal{U}$,

$$\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(T)}\boldsymbol{W}_V^{(T)}\|\lesssim\sqrt{\frac{1}{M_1}+\frac{1}{M_2^2}\cdot M_2}=\frac{1}{\sqrt{M_2}},\tag{128}$$

where $1/M_1$ is the square of (82). $1/M_2^2$ is the square of the scaling in RHS of (121), and M_2 is the number of IDI patterns.

If we prune all neurons $i\notin\mathcal{W}\cup\mathcal{U}$, we have that

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}_{query},f}[\ell(\Psi;\boldsymbol{P},y)]\leq&1-(1-\frac{2}{\alpha'M_1})\frac{1-\epsilon}{(1-\frac{2}{\alpha M_1})}(1-\frac{1}{\sqrt{M_1}})\\ \leq&1-(1-\frac{2}{\alpha'M_1})(1-\epsilon)(1-\frac{1}{\sqrt{M_1}})\\ \leq&1-(1-\frac{2}{\alpha'M_1}-\epsilon-\frac{1}{\sqrt{M_1}})\\ \leq&\mathcal{O}(\epsilon+\frac{1}{\sqrt{M_1}}+\frac{1}{\alpha'M_1})\\ \leq&\mathcal{O}(\epsilon+\frac{1}{\sqrt{M_1}}),\end{aligned}\tag{129}$$

where the first step combines (83), (82), and $2/(\alpha M_1)$ and $2/(\alpha' M_1)$ comes from (77) and (93). The last step comes from $\alpha'\geq M_1^{-1/2}$. Meanwhile, if we prune R fraction of neurons in $\mathcal{W}\cup\mathcal{U}$, given (67), we have for the trained model Ψ,

$$F(\Psi;\boldsymbol{P}^n)\leq(1+\epsilon)(1-R)\cdot\frac{(1-\frac{2}{\alpha'M_1})}{(1-\frac{2}{\alpha M_1})}.\tag{130}$$

Then,

$$\begin{aligned}\underset{\boldsymbol{x}_{query},f}{\mathbb{E}}[\ell(\Psi;\boldsymbol{P},y)] \geq& 1-(1-\frac{2}{\alpha' M_1})\frac{1+\epsilon}{(1-\frac{2}{\alpha M_1})}(1-R)\\ \geq& 1-(1-\frac{2}{\alpha' M_1})(1+\epsilon)(1+\frac{4}{\alpha M_1})(1-R)\\ =&1-(1-R-\frac{2}{\alpha' M_1}+\frac{2R}{\alpha' M_1})(1+\epsilon+\frac{4}{\alpha M_1}+\frac{4\epsilon}{\alpha M_1})\\ \geq& R+\frac{2}{\alpha' M_1}-\frac{2R}{\alpha' M_1}-(1-R-\frac{2}{\alpha' M_1})(\epsilon+\frac{4+4\epsilon}{\alpha M_1})\\ \geq&\Omega(R+\frac{1}{\alpha' M_1}),\end{aligned} \tag{131}$$

where the second step is by $(1-x)^{-1}\leq 1+2x$ for small $x>0$, and the last step is by $R=\Theta(1)$.

□

E.5. Proof of Lemma D.5

Proof. 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 D.6 and D.7.

(a) From the training loss function, by basic mathematical computation, we can obtain

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_Q}\\ =&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial F(\boldsymbol{p}_{query}^n)}\frac{\partial F(\boldsymbol{p}_{query}^n)}{\partial \boldsymbol{W}_Q}\\ =&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\ &\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\\ &\cdot\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\boldsymbol{W}_K(\boldsymbol{p}_s^n-\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top\Big)\\ =&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\ &\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\\ &\cdot(\boldsymbol{W}_K\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\boldsymbol{W}_K\boldsymbol{p}_r^n){\boldsymbol{p}_{query}}^\top\Big).\end{aligned} \tag{132}$$

If $t=0$, we have that

$$(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n)^\top(\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)={\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n. \tag{133}$$

When $z^n=+1$, let $\boldsymbol{x}_{query}^n$ be a noisy version of $\boldsymbol{\mu}_j+\kappa_{query}^n\boldsymbol{\nu}_k$ where $j\in\{1,2,\cdots,M_1\}$ and $k\in\{1,2,\cdots,M_2\}$. Define m_i as the corresponding IDR pattern in the i-th demonstration. Consider the categorical distribution where the probability of selecting $\boldsymbol{\mu}_q$ is $\alpha/2$. We know there exists a $\boldsymbol{\mu}_j$ such that the probability of selecting $\boldsymbol{\mu}_j$ is also $\alpha/2$. Selecting other $\boldsymbol{\mu}_t$ for $t\neq p,j$ has a probability of $(1-\alpha)/(M_1-2)$. Selecting any IDI pattern $\boldsymbol{\nu}_k$ has a probability of $1/M_2$. By the Chernoff bound of Bernoulli distribution in Lemma D.1, given $l\geq\Theta(\max\{M_1,M_2\}\log M_1)$, we can obtain

$$\Pr\left(\sum_{i=1}^l\mathbb{1}[m_i=\boldsymbol{\mu}_j]\leq l\cdot\frac{\alpha}{2}\right)\leq e^{-C\log M_1}=M_1^{-C}, \tag{134}$$

$$\Pr\left(\sum_{i=1}^l \mathbb{1}[m_i=\boldsymbol{\mu}_s]\leq l\cdot\frac{\alpha}{2}\right)\leq e^{-C\log M_1}=M_1^{-C}, \tag{135}$$

$$\Pr\left(\sum_{i=1}^l \mathbb{1}[m_i=\boldsymbol{\mu}_t]\geq l\cdot\frac{1}{M_1}\right)\leq e^{-C\log M_1\cdot M_1\cdot\frac{1}{M_1}}=M_1^{-C}, \tag{136}$$

$$\Pr\left(\sum_{i=1}^l \mathbb{1}[m_i=\boldsymbol{\nu}_k]\geq l\cdot\frac{1}{M_2}\right)\leq e^{-C\log M_1\cdot M_2\cdot\frac{1}{M_2}}=M_1^{-C}, \tag{137}$$

for some $C>0$. Therefore, since that $\frac{1}{\sqrt{M_2}}\cdot e^{\delta^2}\lesssim\frac{\alpha}{2}=\Theta(1)$,

$$\begin{aligned}\sum_{s\in\mathcal{N}_j^{n,i}\cap\mathcal{M}_k^{n,i}} e^{\delta^2(\beta\cdot\beta+1)}&\leq l\cdot\frac{1}{\sqrt{M_2}}e^{\delta^2}\cdot e^{\delta^2(\beta\cdot\beta)}\\&\lesssim l\cdot\frac{\alpha}{2}\cdot e^{\delta^2(\beta\cdot\beta)}\\&\lesssim\sum_{s\in\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}} e^{\delta^2(\beta\cdot\beta)}.\end{aligned}\tag{138}$$

Similarly,

$$\begin{aligned}\sum_{s\in\mathcal{M}_k^{n,i}-\mathcal{N}_j^{n,i}} e^{\delta^2}&\leq l\cdot\frac{1}{\sqrt{M_2}}e^{\delta^2}\cdot e^{\delta^2(\beta\cdot\beta)}\\&\lesssim l\cdot\frac{\alpha}{2}\cdot e^{\delta^2(\beta\cdot\beta)}\\&\lesssim\sum_{s\in[l]-\mathcal{N}_j^{n,i}-\mathcal{N}_k^{n,i}} e^{\delta^2},\end{aligned}\tag{139}$$

where the last step is by the fact that there exists $\boldsymbol{\mu}_s$ for $p\in\{1,2,\cdots,M_2\}\backslash\{j\}$ such that selecting $\boldsymbol{\mu}_s$ has a probability of $\alpha/2$. Let $i\in\mathcal{W}$, $s\in\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}$, then

$$\begin{aligned}&\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\\geq& e^{\delta^2(\beta\cdot\beta)}\cdot(\sum_{s\in\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}} e^{\delta^2(\beta\cdot\beta)}+\sum_{s\in\mathcal{N}_j^{n,i}\cap\mathcal{M}_k^{n,i}} e^{\delta^2(\beta\cdot\beta+1)}\\&+\sum_{s\in[l]-\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}} e^{\delta^2}+\sum_{s\in\mathcal{M}_k^{n,i}-\mathcal{N}_j^{n,i}} e^{\delta^2})^{-1}\\\gtrsim&\frac{e^{\delta^2(\beta\cdot\beta)}}{\sum_{s\in\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}} e^{\delta^2(\beta\cdot\beta)}+\sum_{s\in[l]-\mathcal{N}_j^{n,i}-\mathcal{N}_k^{n,i}} e^{\delta^2}},\end{aligned}\tag{140}$$

where the second step is by (138) and (139). Similarly, for $s\in\mathcal{N}_j^{n,i}\cap\mathcal{M}_k^{n,i}$,

$$\begin{aligned}&\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\\gtrsim&\frac{e^{\delta^2(\beta\cdot\beta)}}{\sum_{s\in\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}} e^{\delta^2(\beta\cdot\beta)}+\sum_{s\in[l]-\mathcal{N}_j^{n,i}-\mathcal{N}_k^{n,i}} e^{\delta^2}}.\end{aligned}\tag{141}$$

For $s\in\mathcal{M}_k^{n,i}-\mathcal{N}_j^{n,i}$,

$$\begin{aligned}&\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\\lesssim&\frac{e^{\delta^2}}{\sum_{s\in\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}} e^{\delta^2(\beta\cdot\beta)}+\sum_{s\in[l]-\mathcal{N}_j^{n,i}-\mathcal{N}_k^{n,i}} e^{\delta^2}}.\end{aligned}\tag{142}$$

For $s \in [l]-\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}$,

$$\begin{aligned}&\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ \lesssim&\frac{e^{\delta^2}}{\sum_{s\in\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}}e^{\delta^2(\beta\cdot\beta)}+\sum_{s\in[l]-\mathcal{N}_j^{n,i}-\mathcal{N}_k^{n,i}}e^{\delta^2}}.\end{aligned} \tag{143}$$

By (53) and (54) in Definition D.8, we have that for $i \in \mathcal{W}_n$,

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)>0. \tag{144}$$

Then we derive

$$\begin{aligned}&\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n\\ =&\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\\ =&\Big(\sum_{r\in\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}}+\sum_{r\in\mathcal{N}_j^{n,i}\cap\mathcal{M}_k^{n,i}}+\sum_{r\in\mathcal{M}_k^{n,i}-\mathcal{N}_j^{n,i}}+\sum_{r\in[l]-\mathcal{N}_j^{n,i}-\mathcal{M}_k^{n,i}}\Big)\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n).\end{aligned} \tag{145}$$

One can observe that

$$\begin{aligned}&\sum_{s\in\mathcal{N}_j^{n,i}}\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\\ =&\sum_{s\in\mathcal{N}_j^{n,i}}\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-(\sum_{r\in\mathcal{N}_j^{n,i}}+\sum_{r\notin\mathcal{N}_j^{n,i}})\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\\ =&\sum_{r\notin\mathcal{N}_j^{n,i}}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\cdot\sum_{s\in\mathcal{N}_j^{n,i}}\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n\\ &-\sum_{s\in\mathcal{N}_j^{n,i}}\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\cdot\sum_{r\notin\mathcal{N}_j^{n,i}}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n+\boldsymbol{n}\\ =&\sum_{s\in\mathcal{N}_j^{n,i}}\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\cdot\sum_{r\notin\mathcal{N}_j^{n,i}}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}(\boldsymbol{p}_s^n-\boldsymbol{p}_r^n)+\boldsymbol{n}.\end{aligned} \tag{146}$$

Hence, by Definition D.4,

$$1>\sum_{r\in[l]-\mathcal{N}_j^{n,i}}\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\geq\gamma_t>0. \tag{147}$$

Since that the feature space embedding of $({\boldsymbol{p}_r^n}^\top,\boldsymbol{0}^\top)^\top$ are orthogonal to $\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n$ for $r\in[l]-\mathcal{N}_j^{n,i}$, we have that with high probability, for $s\in\mathcal{N}_j^{n,i}$,

$$\begin{aligned}&({\boldsymbol{x}_s^n}^\top,\boldsymbol{0}^\top)\sum_{r\in[l]-\mathcal{N}_j^{n,i}}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\\ \geq&\gamma_t\beta^2\delta,\end{aligned} \tag{148}$$

where γ_t comes from the definition. β is from the definition of the data. Meanwhile, for r such that $\boldsymbol{\mu}_r$ is the IDR pattern with the probability of $\alpha/2$ to be selected,

$$\begin{aligned}&\Big|(\boldsymbol{x}_r^{n\top},\boldsymbol{0}^\top)\sum_r \text{softmax}(\boldsymbol{p}_r^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\Big|\\ \leq&(\boldsymbol{x}_s^{n\top},\boldsymbol{0}^\top)\sum_{r\in[l]-\mathcal{N}_j^{n,i}} \text{softmax}(\boldsymbol{p}_r^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\cdot\frac{1-\alpha}{1-\alpha/2},\end{aligned} \tag{149}$$

where $(1-\alpha)/(1-\alpha/2)$ comes from the fraction of attention weights on $\boldsymbol{\mu}_r$ in $[l]-\mathcal{N}_j^{n,i}$. If $\boldsymbol{\mu}_r$ is the pattern that does not decide the label of the current $\boldsymbol{P}^n$, we have

$$\begin{aligned}&\Big|(\boldsymbol{x}_r^{n\top},\boldsymbol{0}^\top)\sum_r \text{softmax}(\boldsymbol{p}_r^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\Big|\\ \leq&\frac{\gamma_t}{l},\end{aligned} \tag{150}$$

where l in the denominator comes from that with high probability, at most 1 $\boldsymbol{\mu}_r$ appears in one data for a certain r. Therefore, for $i\in\mathcal{W}_n$, we denote that $\zeta_{i,n}'=\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s\in\mathcal{N}_j^{n,i}}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^{n(t)})\text{softmax}(\boldsymbol{p}_s^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)$. Then, if $\zeta_{i,n}'>0$, we have that for $\boldsymbol{\mu}_q$ that has the same IDR pattern as $\boldsymbol{x}_{query}$,

$$\begin{aligned}&(\boldsymbol{\mu}_q^\top,\boldsymbol{0}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s\in\mathcal{N}_j^{n,i}}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^{n(t)})\text{softmax}(\boldsymbol{p}_s^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}(\boldsymbol{p}_r^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\boldsymbol{p}_{query}^{n}{}^\top(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\\ \geq&\zeta_{i,n}'\delta\beta^4\gamma_t,\end{aligned} \tag{151}$$

where γ_t comes from that $\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\boldsymbol{p}_s^n$ is much larger in average, $i\in\mathcal{W}_n$, if than other $\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\boldsymbol{p}_t^n$ if $s\in\mathcal{N}_*^n$ while $s\notin\mathcal{N}_*^n$. For j such that $\boldsymbol{\mu}_j$ is the IDR pattern with the probability of $\alpha/2$ to be selected but different from q,

$$\begin{aligned}&(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s\in\mathcal{N}_j^{n,i}}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^{n})\text{softmax}(\boldsymbol{p}_s^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}(\boldsymbol{p}_r^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\boldsymbol{p}_{query}{}^\top(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\\ \leq&\frac{1-\alpha}{1-\alpha/2}(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s\in\mathcal{N}_j^{n,i}}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^{n(t)})\text{softmax}(\boldsymbol{p}_s^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n).\end{aligned} \tag{152}$$

For $\boldsymbol{\mu}_j$ that does not decide the label of the current $\boldsymbol{P}^n$, we have

$$\begin{aligned}&(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s\in\mathcal{N}_j^{n,i}}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^{n})\text{softmax}(\boldsymbol{p}_s^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}(\boldsymbol{p}_r^{n\top}\boldsymbol{W}_K^{(t)\top}\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\boldsymbol{p}_{query}^{n}{}^\top(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\\ \leq&\frac{\zeta_{i,n}'\delta\gamma_t\beta^4}{l}.\end{aligned} \tag{153}$$

To deal with $s\in[l]-\mathcal{N}_j^{n,i}$, we cover this part when summing up all the neurons. Since that each entry of $\boldsymbol{W}_{O_{(i,\cdot)}}$ follows $\mathcal{N}(0,\xi^2)$, we have

$$\Pr(\|\boldsymbol{W}_{O_{(i,1:d_\mathcal{X})}}\boldsymbol{x}_{query}^n\|\leq\beta\xi)\leq\beta\xi, \tag{154}$$

by the standard property of Gaussian distribution. Meanwhile, by Hoeffding's inequality (27),

$$\Pr(\|\boldsymbol{W}_{O_{(i,1:d_{\mathcal{X}})}}\boldsymbol{x}_{query}^n\| \geq \beta\xi\log M_1) \leq M_1^{-C}, \tag{155}$$

for some $C > 1$. Hence, with a high probability, by Hoeffding's inequality (27),

$$\begin{aligned}\Big|\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,1:d_{\mathcal{X}})}}^{(t)}\boldsymbol{x}_{query}^n\Big| \lesssim & \frac{|\mathcal{W}_n|}{m}\Phi(0)\beta\xi + \beta\xi\cdot\frac{\log M_1}{\sqrt{m}}\\ \lesssim & \beta\xi,\end{aligned} \tag{156}$$

$$\begin{aligned}\Big|\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,1:d_{\mathcal{X}})}}^{(t)}\boldsymbol{x}_{query}^n\Big| \gtrsim & \frac{|\mathcal{W}_n|}{m}\Phi(0)\beta\xi - \beta\xi\cdot\frac{\log M_1}{\sqrt{m}}\\ \gtrsim & \beta\xi.\end{aligned} \tag{157}$$

For p such that the probability of selecting $\boldsymbol{\mu}_p$ is $\alpha/2$, we have

$$\Big|\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,1:d_{\mathcal{X}})}}^{(t)}\boldsymbol{\mu}_p\Big| \lesssim \Big|\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,1:d_{\mathcal{X}})}}^{(t)}\boldsymbol{\mu}_j\Big|\cdot e^{-\delta\beta^2}. \tag{158}$$

We have that for $z^n = 1$, we can then derive that for $s \in \mathcal{N}_j^{n,i}$, by Definition D.4, for $\boldsymbol{\mu}_q$ which is the IDR pattern of $\boldsymbol{x}_{query}$,

$$\begin{aligned}&(\boldsymbol{\mu}_q^\top, \boldsymbol{0}^\top)\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n - \sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top(\boldsymbol{x}_{query}^\top, \boldsymbol{0}^\top)^\top\\ \geq&\zeta_{i,t}\delta\gamma_t(1-e^{-\delta^2\beta^2})\beta^4,\end{aligned} \tag{159}$$

where $\zeta_{i,t}$ is used as a lower bound after taking an average of $i \in \mathcal{W}_n$. Similarly, for j such that $\boldsymbol{\mu}_j$ has a probability of $\alpha/2$ to be selected, but different from q we have

$$\begin{aligned}&(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n - \sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top(\boldsymbol{x}_{query}^\top, \boldsymbol{0}^\top)^\top\\ \lesssim&e^{-\delta^2\beta^2}\cdot(\boldsymbol{\mu}_q^\top, \boldsymbol{0}^\top)\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n - \sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top(\boldsymbol{x}_{query}^\top, \boldsymbol{0}^\top)^\top.\end{aligned} \tag{160}$$

For $j \in [l] - \mathcal{N}_q^{n,i}$ with probability of $(1-\alpha)/(M_1-2)$ to be selected, with high probability, at most 1 example has $\boldsymbol{\mu}_j$ in each data. Then,

$$\begin{aligned}
&(\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\\
\leq&\frac{1}{l}\cdot(\boldsymbol{\mu}_q^\top, \boldsymbol{0}^\top)\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top.
\end{aligned}\tag{161}$$

If $i \in \mathcal{U}_n$, since that $z^n = 1$, the indicator by the Relu activation returns zero. Hence, we do not need to compute this case. If $i \notin \mathcal{W}_n \cup \mathcal{U}_n$, by the uniform distribution of a_i, we have that, for $\boldsymbol{\mu}_q$ which is the IDR pattern of $\boldsymbol{x}_{query}$,

$$\begin{aligned}
&({\boldsymbol{\mu}_q}^\top, \boldsymbol{0}^\top)\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top\Big)(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\\
\leq&\sqrt{\frac{\log m}{m}}\cdot(\boldsymbol{\mu}_q^\top, \boldsymbol{0}^\top)\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top,
\end{aligned}\tag{162}$$

where $\sqrt{\log m/m}$ is because a_i can be either $+1$ and -1 following a uniform distribution in this case. For $\boldsymbol{\mu}_j$ that has a probability of $\alpha/2$ to be selected but different from $\boldsymbol{\mu}_q$, we have

$$\begin{aligned}
&({\boldsymbol{\mu}_j}^\top, \boldsymbol{0}^\top)\sum_{i\notin\mathcal{W}_t\cup\mathcal{U}}a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top\Big)(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\\
\leq&e^{-\delta\beta^2}\sqrt{\frac{\log m}{m}}\cdot(\boldsymbol{\mu}_q^\top, \boldsymbol{0}^\top)\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top.
\end{aligned}\tag{163}$$

For $\boldsymbol{x}_j^n$ with $\boldsymbol{\mu}_j$ that has a probability of $(1-\alpha)/(M_1-2)$ to be selected, we have

$$\begin{aligned}
&(\boldsymbol{\mu}_j{}^\top, \boldsymbol{0}^\top)\sum_{i\notin\mathcal{W}_t\cup\mathcal{U}} a_i \mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\boldsymbol{p}_{query}{}^\top\Big)(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\\
\leq&\frac{1}{l}\cdot\sqrt{\frac{\log m}{m}}\cdot(\boldsymbol{\mu}_q^\top,\boldsymbol{0}^\top)\frac{1}{m}\sum_{i\in\mathcal{W}_n}\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top.
\end{aligned}\tag{164}$$

Therefore, by (159), (162), (163), and (164), we have that for one $\boldsymbol{x}_{query}$,

$$\begin{aligned}
&\Big|\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)(\boldsymbol{x}_{query}{}^\top,\boldsymbol{0}^\top)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top\Big)(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)\Big|\\
\geq&\eta\frac{1}{BM_1}\sum_{n\in\mathcal{B}_b}\frac{1}{m}\sum_{i\in\mathcal{W}_n}\zeta_{i,t}\delta\gamma_t(1-e^{-\delta^2\beta^2}\sqrt{\frac{\log m}{m}}-\frac{1}{l}\sqrt{\frac{\log m}{m}})\beta^4\\
\gtrsim&\eta\frac{1}{M_1}\zeta_t\delta\gamma_t\beta^4,
\end{aligned}\tag{165}$$

as long as

$$m\gtrsim 1,\tag{166}$$

and

$$B\gtrsim M_1\log M_1,\tag{167}$$

to ensure that

$$\Pr\left(\sum_{n=1}^B\mathbb{1}[m_n=\boldsymbol{\mu}_j]\leq B(1-c)\cdot\frac{1}{M_1}\right)\leq e^{-c^2B\cdot\frac{1}{M_1}}=e^{-c\log M_1}=M_1^{-C},\tag{168}$$

for some $c \in (0,1)$ and $C > 1$, where m_i denotes the IDR pattern in the query of the n-th data. Meanwhile, for $j \in [l] - \mathcal{N}_q^{n,i}$ that has a IDR pattern which forms a task in $\mathcal{T}_{tr}$ with the IDR pattern of $\boldsymbol{x}_{query}$, more indicators of $i \in \mathcal{U}_n$ is activated.

$$\begin{aligned}
&-\Big|\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)({\boldsymbol{\mu}_j}^\top,\boldsymbol{0}^\top)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top\Big)(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\Big|\\
\geq&-\frac{1}{2}e^{-\delta^2\beta^2}\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)({\boldsymbol{\mu}_q}^\top,\boldsymbol{0}^\top)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top\Big)(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top.
\end{aligned}\tag{169}$$

For other $j \in [l] - \mathcal{N}_q^{n,i}$,

$$\begin{aligned}
&\Big|\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)({\boldsymbol{\mu}_j}^\top,\boldsymbol{0}^\top)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top\Big)(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\Big|\\
\leq&\frac{1}{M_1}\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)({\boldsymbol{\mu}_q^n}^\top,\boldsymbol{0}^\top)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\
&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\
&\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}^n}^\top\Big)(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top,
\end{aligned}\tag{170}$$

where M_1 comes from the fact that the softmax value between $\boldsymbol{p}_{query}^n$ and $\boldsymbol{p}_r^n$ with $\boldsymbol{\mu}_j$ as the IDR pattern of $\boldsymbol{p}_r^n$ is $\Theta(1-\gamma_t)/M_1$ in average of $B \gtrsim M_1\log M_1$ samples. Then, by combining (165), (169), and (170), we have

$$\begin{aligned}
&(\boldsymbol{\mu}_q^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\\
\gtrsim&\eta\frac{1}{M_1}\zeta_t\delta\gamma_t\beta^4.
\end{aligned}\tag{171}$$

By (169) and (170), we have that for $\boldsymbol{\mu}_j$ which forms a task in $\mathcal{T}_{tr}$ with the $\boldsymbol{\mu}_q$,

$$\begin{aligned}
&-\Big|(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\Big|\\
\geq&-\frac{1}{2}e^{-\delta^2\beta^2}\Big|(\boldsymbol{\mu}_q^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\Big|.
\end{aligned}\tag{172}$$

For $\boldsymbol{\mu}_j$ which does not form a task in $\mathcal{T}_{tr}$ with the $\boldsymbol{\mu}_q$,

$$\begin{aligned}&\Big|(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\Big|\\ \leq&\frac{1}{M_1}\Big|(\boldsymbol{\mu}_q^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\Big|.\end{aligned}\tag{173}$$

Similarly, for $k\in[M_2]$,

$$\begin{aligned}&\Big|(\boldsymbol{\nu}_k^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\Big|\\ \lesssim&\frac{1}{M_2}\Big|(\boldsymbol{\mu}_q^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q}\Big|_{t=0}(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\Big|,\end{aligned}\tag{174}$$

where M_2 comes from that for ν_k that is added to $\boldsymbol{\mu}_j$, the contribution of gradient is $1/M_2$ times of replacing $\boldsymbol{\nu}_k$ with $\boldsymbol{\mu}_j$. Hence, $1/M_2\cdot(1+1/M_1\cdot M_1)=2/M_2=\Theta(1/M_2)$. For the label embedding, we have

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q}\Big|_{t=0}[:,d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}]\\ =&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\ &\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n){\boldsymbol{p}_{query}}^\top\Big)[:,d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}].\\ =&\boldsymbol{0}.\end{aligned}\tag{175}$$

We then have

$$\begin{aligned}&\Big|\boldsymbol{q}^\top\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_l^n)\geq 0]\\ &\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\\ &\cdot(\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\boldsymbol{W}_K^{(t)}\boldsymbol{p}_r^n)\Big)[d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}]\Big|\\ =&0.\end{aligned}\tag{176}$$

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 (165) and (169). Note that the indicator of $\mathcal{W}_n$ will not change along the training. The only difference is the changes in ζ_t and γ_t. Thus, we can obtain that for $\boldsymbol{\mu}_q$ with the same IDR pattern as $\boldsymbol{x}_{query}$,

$$\begin{aligned}&(\boldsymbol{\mu}_q^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_Q}\Big|_{t=t_0+1}(\boldsymbol{x}_{query}^\top,\boldsymbol{0}^\top)^\top\\ \gtrsim&\eta\frac{1}{M_1}\sum_{b=0}^{t_0}\zeta_b\delta\gamma_b\beta^4\\ \gtrsim&\eta\frac{1}{M_1}\sum_{b=0}^{t_0}\zeta_b\delta\gamma_b\beta^4,\end{aligned}\tag{177}$$

as long as (167) holds. We also have

$$\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0+1}[:,d_\mathcal{X}+1:d_\mathcal{X}+d_\mathcal{Y}]=\mathbf{0}. \tag{178}$$

Similarly, for $j\neq q$ and $j\in[M_1]$ where $\boldsymbol{\mu}_l$ does not form a task in $\mathcal{T}_{tr}$ with the $\boldsymbol{\mu}_q$,

$$\begin{aligned}&\Big|(\boldsymbol{\mu}_j^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top,\mathbf{0}^\top)^\top\Big|\\ \lesssim&\frac{1}{M_1}\Big|(\boldsymbol{\mu}_q^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top,\mathbf{0}^\top)^\top\Big|.\end{aligned} \tag{179}$$

For $j\neq q$ and $j\in[M_1]$ where $\boldsymbol{\mu}_j$ forms a task in $\mathcal{T}_{tr}$ with $\boldsymbol{\mu}_q$,

$$\begin{aligned}&-\Big|(\boldsymbol{\mu}_j^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top,\mathbf{0}^\top)^\top\Big|\\ \gtrsim&-e^{-\delta^2\beta^2-(\eta\frac{1}{M_1}\sum_{b=0}^{t_0-1}\zeta_b\delta\gamma_b\beta^2)^2}\Big|(\boldsymbol{\mu}_q^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top,\mathbf{0}^\top)^\top\Big|\\ \gtrsim&-e^{-\Theta(\frac{\eta t_0}{M_1})^2}\Big|(\boldsymbol{\mu}_q^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top,\mathbf{0}^\top)^\top\Big|,\end{aligned} \tag{180}$$

where the first step comes from the fact that a negative gradient update makes the softmax value of $\boldsymbol{\mu}_l$ much smaller. The last step is obtained in the order related to η,t,M_1 as variables. Meanwhile, for $k\in[M_2]$,

$$\begin{aligned}&\Big|(\boldsymbol{\nu}_k^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top,\mathbf{0}^\top)^\top\Big|\\ \lesssim&\frac{1}{M_2}\Big|(\boldsymbol{\mu}_q^\top,\mathbf{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_Q}\Big|_{t=t_0}(\boldsymbol{x}_{query}^\top,\mathbf{0}^\top)^\top\Big|.\end{aligned} \tag{181}$$

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

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n,\Psi)}{\partial \boldsymbol{W}_K}\\ =&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\ &\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query})\boldsymbol{W}_Q^\top\boldsymbol{p}_{query}^n\\ &\cdot(\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\boldsymbol{p}_r^n)^\top\Big).\end{aligned} \tag{182}$$

If we investigate $\boldsymbol{W}_K^{(t)}\boldsymbol{p}_s^n$, we can tell that the output is a weighed summation of multiple $\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n$. Similarly, the output of $\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n$ 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 $\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n$ only contains the contribution from the feature space embeddings at the initialization. One difference is that since that $\boldsymbol{q}$ appears with $1/2$ probability in all $\boldsymbol{p}_s^n$,

$$\begin{aligned}&\Big|\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\&\cdot\Big(\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query})\boldsymbol{W}_Q^\top\boldsymbol{p}_{query}^n\\&\cdot(\boldsymbol{p}_s^n-\sum_{r=1}^{l+1}\text{softmax}({\boldsymbol{p}_r^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\boldsymbol{p}_r^n)^\top\Big)[d_{\mathcal{X}}+1:d_{\mathcal{X}}+d_{\mathcal{Y}}]\boldsymbol{q}\Big|\\\leq&\sqrt{\frac{\log B}{B}}(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K}\Big|_{t=t_0+1}(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)^\top.\end{aligned}\tag{183}$$

Following the steps in Part (a), we can obtain that for $\boldsymbol{\mu}_q$ as the IDR pattern of $\boldsymbol{x}_q$, $e\in[l]$,

$$\begin{aligned}&(\boldsymbol{\mu}_q^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K}\Big|_{t=t_0+1}(\boldsymbol{x}_e^\top,\boldsymbol{0}^\top)^\top\\\gtrsim&\eta\frac{1}{M_1}\sum_{b=0}^{t_0}\zeta_b\delta\gamma_b\beta^2,\end{aligned}\tag{184}$$

and combining (183),

$$\begin{aligned}&(\boldsymbol{\mu}_q^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K}\Big|_{t=t_0+1}\boldsymbol{p}_q\\\gtrsim&\eta\frac{1}{M_1}\sum_{b=0}^{t_0}\zeta_b\delta\gamma_b\beta^2(1-\sqrt{\frac{\log B}{B}})\\\gtrsim&\eta\frac{1}{M_1}\sum_{b=0}^{t_0}\zeta_b\delta\gamma_b\beta^2,\end{aligned}\tag{185}$$

where the last step holds as long as (167). For $\boldsymbol{\mu}_j$ which forms a task in $\mathcal{T}_{tr}$ with the $\boldsymbol{\mu}_q$,

$$\begin{aligned}&-\Big|(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K}\Big|_{t=t_0}\boldsymbol{p}_q\Big|\\\gtrsim&-e^{-\Theta(\frac{\eta t_0}{M_1})^2}\Big|(\boldsymbol{\mu}_q^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K}\Big|_{t=t_0}\boldsymbol{p}_q\Big|.\end{aligned}\tag{186}$$

For $\boldsymbol{\mu}_j$ which does not form a task in $\mathcal{T}_{tr}$ with the $\boldsymbol{\mu}_q$,

$$\begin{aligned}&\Big|(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K}\Big|_{t=t_0}\boldsymbol{p}_q\Big|\\\leq&\frac{1}{M_1}\Big|(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K}\Big|_{t=t_0}\boldsymbol{p}_q\Big|.\end{aligned}\tag{187}$$

Meanwhile, for $k\in[M_2]$, similar to (181),

$$\begin{aligned}&\Big|(\boldsymbol{\nu}_k^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K}\Big|_{t=t_0}\boldsymbol{p}_q\Big|\\\leq&\frac{1}{M_2}\Big|(\boldsymbol{\mu}_j^\top,\boldsymbol{0}^\top)\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_K}\Big|_{t=t_0}\boldsymbol{p}_q\Big|.\end{aligned}\tag{188}$$

$\square$

E.6. Proof of Lemma D.6

Proof.

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_V}\\&=\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial F(\boldsymbol{p}_{query}^n)}\frac{\partial F(\boldsymbol{p}_{query}^n)}{\partial\boldsymbol{W}_V}\\&=\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i=1}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\&\cdot\boldsymbol{W}_{O_{(i,\cdot)}}^\top\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top.\end{aligned}\tag{189}$$

Let $\boldsymbol{x}_i^n$ and $\boldsymbol{x}_j^n$ correspond to IDR patterns $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$, respectively. For $\boldsymbol{p}_{query}^n$ which corresponds to the IDR feature $\boldsymbol{\mu}_a$,

$$\begin{aligned}\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top({\boldsymbol{x}_i^n}^\top,\boldsymbol{q}^\top)^\top&\gtrsim\beta^2(1-\gamma_t)\cdot 1-\frac{1}{\frac{\alpha}{2}l}\\&\gtrsim\beta^2(1-\gamma_t)\cdot 1,\end{aligned}\tag{190}$$

where the first step holds since that by (137), with high probability, no other $\boldsymbol{x}_k^n$ where $k\neq l+1$ shares the same IDI pattern as $\boldsymbol{x}_{query}^n$. The last step holds if

$$l_{tr}\gtrsim\frac{1}{\alpha\beta^2}.\tag{191}$$

Meanwhile, by a different IDR pattern of $\boldsymbol{x}_j^n$,

$$\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top({\boldsymbol{x}_j^n}^\top,\boldsymbol{q}^\top)^\top\lesssim\beta^2\gamma_t.\tag{192}$$

When $t=0$, for all $i\in\mathcal{W}_n$, we have that by Lemma D.10, for $\boldsymbol{p}_{query}^n$ that corresponds to $\boldsymbol{\mu}_a$,

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)>0.\tag{193}$$

Therefore, for any $\boldsymbol{p}_j^n=({\boldsymbol{x}_j^n}^\top,{\boldsymbol{y}_j^n}^\top)^\top$ where $f^{(n)}(\tilde{\boldsymbol{x}}_j^n)=+1$, and

$$\boldsymbol{x}_j^n=\boldsymbol{\mu}_a+\kappa_j^n\boldsymbol{\nu}_b,\tag{194}$$

we have

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_V^{(t)}}\boldsymbol{p}_j^n\\=&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{k=1}^m a_k\mathbb{1}[\boldsymbol{W}_{O_{(k,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\&\cdot\boldsymbol{W}_{O_{(k,\cdot)}}^\top\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top\boldsymbol{p}_j^n.\end{aligned}\tag{195}$$

We then have that by combining (190) and (192),

$$\begin{aligned}&-\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i\in\mathcal{W}_n}a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\&\cdot\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top\boldsymbol{p}_j^n\\\gtrsim&\eta\beta^2(1-\gamma_t).\end{aligned}\tag{196}$$

Since that for $\boldsymbol{p}_s^n$ and $\boldsymbol{p}_j^n$ with different label embeddings, their inner product is smaller than $-1+\beta$ if they share the same IDR pattern, or smaller than -1 if they share different IDR patterns. On average, in a batch, this product is close to -1. Hence

$$\begin{aligned}
&-\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i\in\mathcal{U}_n}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(k,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\
&\cdot\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top\boldsymbol{p}_j^n\\
\leq&\frac{1}{\beta^2+1}\cdot(-\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i\in\mathcal{W}_n}a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\
&\cdot\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top\boldsymbol{p}_j^n).
\end{aligned}\tag{197}$$

Meanwhile, since that

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}^m a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\
&\cdot\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top\boldsymbol{p}_j^n\\
\lesssim&\sqrt{\frac{\log B}{B}}\cdot\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{j\in\mathcal{W}_n}^m a_j\mathbb{1}[\boldsymbol{W}_{O_{(j,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\
&\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top\boldsymbol{p}_j^n,
\end{aligned}\tag{198}$$

where $\sqrt{\log B/B}$ is because that z^n is selected from $\{+1,-1\}$ with equal probability. Hence, we can denote and derive that when $t=t_0+1$,

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_V^{(b)}}\boldsymbol{p}_j^n\\
=&\eta\sum_{b=0}^{t_0}(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}),
\end{aligned}\tag{199}$$

where

$$-V_i(b)\gtrsim\beta^2(1-\gamma_t)1/a,\quad i\in\mathcal{W}_n,\tag{200}$$

$$-V_i(b)\leq\frac{1}{\beta^2+1}V_j(b),\quad i\in\mathcal{U}_n,j\in\mathcal{W}_n,\tag{201}$$

$$|V_i(b)|\lesssim\sqrt{\frac{\log B}{B}}\cdot\frac{1}{a},\quad i\notin\mathcal{W}_n\cup\mathcal{U}_n.\tag{202}$$

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

$$\boldsymbol{x}_j^n=\boldsymbol{\mu}_a+\kappa_j^n\boldsymbol{\nu}_b,\tag{203}$$

we have

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_V^{(t)}}\boldsymbol{p}_j^n\\=&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{k=1}^m a_k\mathbb{1}[\boldsymbol{W}_{O_{(k,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\&\cdot\boldsymbol{W}_{O_{(k,\cdot)}}^\top\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top\boldsymbol{p}_j^n,\end{aligned}\tag{204}$$

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\sum_{b=0}^{t_0}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_V^{(b)}}\boldsymbol{p}_j^n\\=&\eta\sum_{b=0}^{t_0}(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}\cup\mathcal{U}}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}),\end{aligned}\tag{205}$$

where

$$-V_i(b)\gtrsim\beta^2(1-\gamma_t)1/a,\quad i\in\mathcal{U}_n,\tag{206}$$

$$-V_i(b)\leq\frac{1}{\beta^2+1}V_j(b),\quad i\in\mathcal{W}_n,j\in\mathcal{U}_n,\tag{207}$$

$$|V_i(b)|\lesssim\sqrt{\frac{\log B}{B}}\cdot\frac{1}{a},\quad i\notin\mathcal{W}_n\cup\mathcal{U}_n.\tag{208}$$

We can also derive

$$\begin{aligned}&\eta\frac{1}{B}\sum_{b=0}^{t_0}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_V^{(t)}}(\boldsymbol{v}_k^\top,\boldsymbol{0}^\top)^\top\\=&\eta\frac{1}{B}\sum_{b=0}^{t_0}\sum_{n\in\mathcal{B}_b}(-z^n)\sum_{k=1}^m a_k\mathbb{1}[\boldsymbol{W}_{O_{(k,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\&\cdot\boldsymbol{W}_{O_{(k,\cdot)}}^\top\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n){\boldsymbol{p}_s^n}^\top(\boldsymbol{\nu}_k^\top,\boldsymbol{0}^\top)^\top\\:=&\eta\sum_{b=0}^{t_0}(\sum_{i\in\mathcal{W}_n}V_i'(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\mathcal{U}_n}V_i'(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}\cup\mathcal{U}}V_i'(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}),\end{aligned}\tag{209}$$

where

$$|V_i'(b)|\leq|V_i(b)|\cdot\frac{1}{M_2},\tag{210}$$

since that $1/M_2$ fraction of $\boldsymbol{p}_s^n$ has $\boldsymbol{\nu}_K$ as the IDI pattern in average.

□

E.7. Proof of Lemma D.7

Proof.

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}\\
=&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial F(\boldsymbol{p}_{query}^n)}\frac{\partial F(\boldsymbol{p}_{query}^n)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}\\
=&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\
&\cdot\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n).
\end{aligned}\tag{211}$$

We have that

$$\begin{aligned}
&\boldsymbol{W}_V^{(t)}\boldsymbol{p}_s^n\\
=&\delta({\boldsymbol{p}_s^n}^\top,\boldsymbol{0}^\top)^\top+\sum_{b=0}^{t-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top.
\end{aligned}\tag{212}$$

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

$$\boldsymbol{x}_s^n=\boldsymbol{\mu}_a+\kappa_s^n\boldsymbol{\nu}_b.\tag{213}$$

When $t=0$, we can obtain that for $i\in\mathcal{W}_n$ and $\boldsymbol{\mu}_a$ as the IDR pattern of $\boldsymbol{p}_{query}^n$,

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top\\
=&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{1}{a}\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)(\delta{\boldsymbol{p}_s^n}^\top(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top+\sum_{b=0}^{t-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\\
&+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top)\\
\geq&\frac{\alpha\eta}{2a}\delta(\beta^2+1).
\end{aligned}\tag{214}$$

Then, we have the following results by Lemma D.6, and the magnitude of $\boldsymbol{\mu}_a$, $\boldsymbol{\mu}_b$, and $\boldsymbol{q}$,

$$\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,-\boldsymbol{q}^\top)^\top\leq\frac{\beta^2-1}{\beta^2+1}\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top,\tag{215}$$

$$\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_b^\top,\boldsymbol{q}^\top)^\top\leq\frac{1}{\beta^2+1}\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top,\tag{216}$$

$$\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_b^\top,-\boldsymbol{q}^\top)^\top\leq-\frac{1}{\beta^2+1}\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top,\tag{217}$$

$$\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\nu}_c^\top,\boldsymbol{0}^\top)^\top\leq\frac{1}{M_2}\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top.\tag{218}$$

Denote the set of data that share one same IDR pattern as $\boldsymbol{p}_{query}^n$ as $\mathcal{B}_b^n$ in the b-th iteration. Therefore, when $t=1$, we have

$$\begin{aligned}
&\eta\frac{1}{|\mathcal{B}_b^n|}\sum_{n\in\mathcal{B}_b^n}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top\\
=&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{1}{a}\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)(\delta{\boldsymbol{p}_s^n}^\top(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top+\sum_{b=0}^{t-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\\
&+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top\\
\geq&\frac{\alpha}{2}(\frac{\eta}{a}(\delta(\beta^2+1)-\frac{\eta}{a}\cdot\eta\cdot\sqrt{\frac{\log B}{B}}\frac{m}{a}\cdot\xi\log M_1\\
&+\frac{\eta}{a}\cdot\eta\frac{m}{a}(\beta^2(1-\gamma_t))(\frac{\eta}{a}\delta(\beta^2+1)-\xi))\\
\gtrsim&\frac{\alpha}{2}(\frac{\eta}{a}(\delta(\beta^2+1)+\frac{\eta}{a}\cdot\eta\frac{m}{a}(\beta^2(1-\gamma_t))\frac{\eta}{a}\delta(\beta^2+1))\\
\gtrsim&\delta(\beta^2+1)\frac{\alpha\eta}{2a}(1+\frac{\eta^2 m}{a^2}),
\end{aligned}\tag{219}$$

where the first inequality comes from that the update in the previous step makes the output of $\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}$ for $i\in\mathcal{W}_n$ be positive. The second step holds when $B\gtrsim M_1$. We also have

$$\begin{aligned}
&\eta\frac{1}{|\mathcal{B}_b-\mathcal{B}_b^n|}\sum_{n\in\mathcal{B}_b-\mathcal{B}_b^n}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top\\
\gtrsim&\delta(\beta^2+1)\frac{\alpha\eta}{2a}(1+\frac{\eta^2 m}{a^2}).
\end{aligned}\tag{220}$$

For $i\in\mathcal{U}_n$, we also have

$$\begin{aligned}
&\eta\frac{1}{|\mathcal{B}_b^n|}\sum_{n\in\mathcal{B}_b^n}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,-\boldsymbol{q}^\top)^\top\\
\gtrsim&\delta(\beta^2+1)\frac{\alpha\eta}{2a}(1+\frac{\eta^2 m}{a^2}),
\end{aligned}\tag{221}$$

$$\begin{aligned}
&\eta\frac{1}{|\mathcal{B}_b-\mathcal{B}_b^n|}\sum_{n\in\mathcal{B}_b-\mathcal{B}_b^n}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,-\boldsymbol{q}^\top)^\top\\
\gtrsim&\delta(\beta^2+1)\frac{\alpha\eta}{2a}(1+\frac{\eta^2 m}{a^2}),
\end{aligned}\tag{222}$$

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

$$\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}({\boldsymbol{p}_j^n}^\top,\boldsymbol{0})^\top\leq\eta\sqrt{\frac{\log B}{B}}\frac{1}{a}.\tag{223}$$

Suppose that the conclusion holds when $t\leq t_0$. Then when $t=t_0+1$, we have that for $i\in\mathcal{W}_n$, $b\neq a$, and $\boldsymbol{p}_{query}^n$

corresponding to $\boldsymbol{q}$ and $\boldsymbol{\mu}_a$,

$$\begin{aligned}
&\eta\frac{1}{|\mathcal{B}_b^n|}\sum_{n\in\mathcal{B}_b^n}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}(\boldsymbol{\mu}_a^\top, \boldsymbol{q}^\top)^\top\\
=&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{1}{a}\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)(\delta{\boldsymbol{p}_s^n}^\top(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top+\sum_{b=0}^{t-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\\
&+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top(\boldsymbol{\mu}_a^\top,\boldsymbol{q}^\top)^\top\\
\gtrsim&\delta(\beta^2+1)\frac{\alpha\eta}{2a}+\frac{\alpha\eta}{2a}\cdot\eta\sum_{b=0}^{t_0}\delta(\beta^2+1)\frac{\eta m}{a^2}(1+\frac{\eta^2 m}{a^2})^b\\
=&\delta(\beta^2+1)\frac{\alpha\eta}{2a}(1+\frac{\eta^2 m}{a^2}\cdot\frac{(1+\frac{\eta^2 m}{a^2})^{t_0+1}-1}{\frac{\eta^2 m}{a^2}})\\
=&\delta(\beta^2+1)\frac{\alpha\eta}{2a}(1+\frac{\eta^2 m}{a^2})^{t_0+1},
\end{aligned}\tag{224}$$

where the first inequality is by plugging the condition in the induction. The last two steps come from basic mathematical computation. Then,

$$\begin{aligned}
&\eta\frac{1}{|\mathcal{B}_b^n|}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b^n}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}(\boldsymbol{\mu}_a^\top, \boldsymbol{q}^\top)^\top\\
\gtrsim&\delta(\beta^2+1)\sum_{b=0}^{t_0+1}\frac{\alpha\eta}{2a}(1+\frac{\eta^2 m}{a^2})^b\\
\gtrsim&\delta(\beta^2+1)\frac{\alpha\eta}{2a}(t_0+1),
\end{aligned}\tag{225}$$

where lower bound in the last step is also a tight estimation of the second to last step if $\eta^2 Tm/a^2\ll 1$. Then, we have

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b^n}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top, \boldsymbol{q}^\top)^\top\\
\gtrsim&\frac{1}{M_1}\cdot\delta\beta^2(\beta^2+1)\frac{\alpha\eta}{2a}(t_0+1).
\end{aligned}\tag{226}$$

By Lemma D.10, when $t\geq\Theta(1)$, we have

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b-\mathcal{B}_b^n}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_b^\top, \boldsymbol{q}^\top)^\top\\
\gtrsim&\frac{M_1-1}{M_1}\cdot\delta(\beta^2+1)\frac{\alpha\eta}{2a}(t_0+1).
\end{aligned}\tag{227}$$

Hence,

$$\begin{aligned}
&\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_b^\top, \boldsymbol{q}^\top)^\top\\
\gtrsim&\delta(\beta^2+1)\frac{\alpha\eta}{2a}(t_0+1),
\end{aligned}\tag{228}$$

which holds for $i\in\cup_{n\in[N]}\mathcal{W}_n=\mathcal{W}$. Meanwhile,

$$\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\nu}_c^\top, \boldsymbol{0}^\top)^\top\leq\frac{1}{BT}\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}\frac{\partial \ell(\tilde{\boldsymbol{P}}^n, z^n;\Psi)}{\partial \boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_b^\top, \boldsymbol{q}^\top)^\top.\tag{229}$$

For $i\in\mathcal{U}_n$ and $\boldsymbol{p}_{query}^n$ corresponding to $-\boldsymbol{q}$ and $\boldsymbol{\mu}_a$, similarly to (226), (227), and (228), we have

$$\begin{aligned}&\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b^n}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,-\boldsymbol{q}^\top)^\top\\ \gtrsim&\frac{1}{M_1}\cdot\delta(\beta^2+1)\frac{\alpha\eta}{2a}(t_0+1),\end{aligned}\tag{230}$$

and when $t\geq\Theta(1)$,

$$\begin{aligned}&\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b-\mathcal{B}_b^n}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_b^\top,-\boldsymbol{q}^\top)^\top\\ \gtrsim&\frac{M_1-1}{M_1}\delta(\beta^2+1)\frac{\alpha\eta}{2a}(t_0+1),\end{aligned}\tag{231}$$

$$\begin{aligned}&\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_b^\top,-\boldsymbol{q}^\top)^\top\\ \gtrsim&\delta(\beta^2+1)\frac{\alpha\eta}{2a}(t_0+1),\end{aligned}\tag{232}$$

which also holds for $i\in\cup_{n\in[N]}\mathcal{U}_n=\mathcal{U}$. Meanwhile,

$$\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\nu}_c^\top,\pm\boldsymbol{q}^\top)^\top\leq\frac{1}{M_2}\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_b^\top,\boldsymbol{q}^\top)^\top.\tag{233}$$

Then, for $i\in\mathcal{W}_n\cup\mathcal{U}_n$,

$$\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0+1)}\|\gtrsim\sqrt{M_1}\delta(\beta^2+1)^{\frac{1}{2}}\frac{\alpha\eta}{2a}(t_0+1).\tag{234}$$

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

$$\eta\frac{1}{B}\sum_{b=0}^{t_0+1}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_{O_{(i,\cdot)}}}(\boldsymbol{\mu}_a^\top,-\boldsymbol{q}^\top)^\top\leq\eta\sqrt{\frac{\log B(t_0+1)}{B(t_0+1)}}\frac{1}{a}.\tag{235}$$

□

E.8. Proof of Lemma D.9

Proof. We know that the Gaussian initialization of $\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}$ generates a uniform distribution on the $d_\mathcal{X}-1$-sphere for the first $d_\mathcal{X}$ dimensions. Therefore,

$$\Pr(i\in\mathcal{W}_n)=A_{d_\mathcal{X}}^{cap}(\phi)/A_{d_\mathcal{X}},\tag{236}$$

where $A_{d_\mathcal{X}}$ is the surface area of an $d_\mathcal{X}-1$-sphere. $A_{d_\mathcal{X}}^{cap}(\phi)$ is the surface area of a $d_\mathcal{X}-1$-spherical cap with ϕ as the colatitude angle. By Equation 1 in (Li, 2010), we have

$$\Pr(i\in\mathcal{W}_n)=\frac{1}{2}I_{\sin^2\phi}(\frac{d_\mathcal{X}-1}{2},\frac{1}{2})=\frac{\int_0^{\sin^2\phi}t^{\frac{d_\mathcal{X}-3}{2}}(1-t)^{-\frac{1}{2}}dt}{2\int_0^1 t^{\frac{d_\mathcal{X}-3}{2}}(1-t)^{-\frac{1}{2}}dt},\tag{237}$$

where $I_\cdot(\cdot,\cdot)$ is the regularized incomplete beta function. Since that

$$\phi\leq\pi/2-\Theta(1/M_1),\tag{238}$$

to avoid concentration error of $\Theta(\sqrt{1/m})$ if $m \gtrsim M_1^2$, we have that when $d_\mathcal{X} = M_1 + M_2 = M = \Theta(M)$,

$$\begin{aligned}
&\frac{\int_0^{\sin^2\phi} t^{\frac{d_\mathcal{X}-3}{2}}(1-t)^{-\frac{1}{2}}dt}{\int_0^1 t^{\frac{d_\mathcal{X}-3}{2}}(1-t)^{-\frac{1}{2}}dt}\\
\geq&\frac{\int_0^{\cos^2 1/M} t^{\frac{d_\mathcal{X}-3}{2}}(1-t)^{-\frac{1}{2}}dt}{\int_0^1 t^{\frac{d_\mathcal{X}-3}{2}}(1-t)^{-\frac{1}{2}}dt}\\
\geq&1-\frac{\int_{1-1/M^2}^1 t^{\frac{d_\mathcal{X}-3}{2}}(1-t)^{-\frac{1}{2}}dt}{\int_0^1 t^{\frac{d_\mathcal{X}-3}{2}}(1-t)^{-\frac{1}{2}}dt}\\
\geq&1-\frac{\int_{1-1/M^2}^1 (1-t)^{-\frac{1}{2}}dt}{\int_{1-1/M}^1 \Theta(1)\cdot(1-t)^{-\frac{1}{2}}dt}\\
=&1-\frac{\frac{2}{M}}{\Theta(1)\cdot(\frac{2}{\sqrt{M}}-\frac{2}{M})}\\
\geq&\Theta(1),
\end{aligned} \tag{239}$$

where the second inequality comes from that $\cos^2(1/M) = (1+\cos(2/M))/2 \geq 1-1/M^2 \geq 1-1/M$, and the third to last step is by $(1-1/M)^{\frac{d_\mathcal{X}-3}{2}} \geq \Theta(1)$, and the last step is by $M \geq \Theta(1)$. For the second $d_\mathcal{Y}$ dimensions of $\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}$, we can derive a similar result by replacing $d_\mathcal{X}$ with $d_\mathcal{Y}$ in (239). This implies that

$$|\mathcal{W}_n| \geq \Omega(1)\cdot\Omega(1)\cdot m \geq \Omega(m). \tag{240}$$

Likewise, the conclusion holds for $\mathcal{U}_n$. Since that $\mathcal{W} = \cup_{n\in[N}\mathcal{W}_n, \mathcal{U} = \cup_{n\in[N}\mathcal{U}_n$, we have

$$|\mathcal{W}|, |\mathcal{U}| \geq \Omega(m). \tag{241}$$

□

E.9. Proof of Lemma D.10

Proof. We prove this lemma in two steps. In the first step, we prove the conclusion by replacing $\mathcal{W}$ with $\mathcal{W}_n$, and replacing $\mathcal{U}$ with $\mathcal{U}_n$. We will also cover the proof of $\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}{}^\top > 0$ and $\sum_{b=0}^{t-1}\eta\sum_{i\in\mathcal{W}_n} V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\boldsymbol{W}_{O_{i,\cdot}}^{(t)}{}^\top > 0$ in the induction as a support. In the second step, we prove the results for $\mathcal{W}$ and $\mathcal{U}$.
(1) When $t=0$. For any $i \in \mathcal{W}_n$, we have that by definition of $\mathcal{W}_n$

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}\boldsymbol{V}^n(0) > 0, \tag{242}$$

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}\boldsymbol{W}_{O_{i,\cdot}}^{(t)}{}^\top > 0. \tag{243}$$

Hence, the conclusion holds. When $t=1$, we have

$$\begin{aligned}
&\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\boldsymbol{V}^n(t)\\
=&\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}{}^\top + \eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{1}{a}\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)(\delta{\boldsymbol{p}_s^n}^\top\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}{}^\top + \sum_{b=0}^{t-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\\
&+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}{}^\top).
\end{aligned} \tag{244}$$

By (211) and definition of $\boldsymbol{W}_n$, we have

$$\begin{aligned}&\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(+z^n)a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top\boldsymbol{W}_K^\top\boldsymbol{W}_Q\boldsymbol{p}_{query}^n)\geq 0]\\&\cdot\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t-1)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t-1)}}^\top\boldsymbol{W}_Q^{(t-1)}\boldsymbol{p}_{query}^n)\\>&0.\end{aligned}\tag{245}$$

Hence,

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}{\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top=\sum_{b=0}^{t-1}\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}{\boldsymbol{W}_{O_{i,\cdot}}^{(t)}}^\top>0,\tag{246}$$

and

$$\sum_{b=0}^{t-1}\eta\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}{\boldsymbol{W}_{O_{i,\cdot}}^{(t)}}^\top>0.\tag{247}$$

By the gradient update when $t=0$, we know that the largest component in the feature embedding is the IDR pattern for $\boldsymbol{p}_{query}^n$, and the label embedding is close to being in the direction of the label embedding of $\boldsymbol{p}_{query}^n$. Hence,

$$\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{1}{a}\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\delta{\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top>0.\tag{248}$$

Denote θ_n^i as the angle between the feature embeddings of $\boldsymbol{V}^n(0)$ and $\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}$. Since that the feature embedding of $\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}$ is initialized uniformed on the $d_\mathcal{X}-1$-sphere, we have $\mathbb{E}[\theta_n^i]=0$. By Hoeffding's inequality (27), we have

$$\Big\|\frac{1}{|\mathcal{W}_n|}\sum_{i\in\mathcal{W}_n}\theta_n^i-\mathbb{E}[\theta_n^i]\Big\|=\Big\|\frac{1}{|\mathcal{W}_n|}\sum_{i\in\mathcal{W}_n}\theta_n^i\Big\|\leq\sqrt{\frac{\log M_1}{m}},\tag{249}$$

with probability of at least $1-M_1^{-10}$. When $m\gtrsim M^2\log M_1$, we can obtain that

$$\Big\|\frac{1}{|\mathcal{W}_n|}\sum_{i\in\mathcal{W}_n}\theta_n^i-\mathbb{E}[\theta_n^i]\Big\|\leq\Theta(\frac{1}{M_1}).\tag{250}$$

Therefore, for $i\in\mathcal{W}_n$, as long as $m\gtrsim M_1^2\log M_1$, we have

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}\sum_{b=0}^{t-1}\sum_{i\in\mathcal{W}_n}{\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}}^\top>0.\tag{251}$$

Given $B\gtrsim M_1\log M_1$, by Lemma D.6, and combining (251), we have that

$$\begin{aligned}&\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{1}{a}\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\sum_{b=0}^{t-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\\&+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top{\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top\\>&0.\end{aligned}\tag{252}$$

Therefore, the conclusion holds when $t=1$.
Suppose that the conclusion holds when $t\leq t_0$. When $t=t_0+1$, by (244), we can check that

$$\begin{aligned}&\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}{\boldsymbol{W}_{O_{i,\cdot}}^{(t)}}^\top\\ =&\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)}({\boldsymbol{W}_{O_{i,\cdot}}^{(0)}}^\top+\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(+z^n)a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}\sum_{s=1}^{l+1}(\boldsymbol{W}_V\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t-1)}}^\top\boldsymbol{W}_Q^{(t-1)}\boldsymbol{p}_{query}^n)\geq 0]\\ &\cdot\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t-1)}\boldsymbol{p}_s^n)\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t-1)}}^\top\boldsymbol{W}_Q^{(t-1)}\boldsymbol{p}_{query}^n)\\ >&0+0=0,\end{aligned}\tag{253}$$

where the second 0 comes from (248) and the conditions that such conclusion in (253) holds when $t\leq t_0$. Combining (228) and the fact that the weighted summation of $\boldsymbol{p}_s^n$ is close to be in the direction of $\boldsymbol{\mu}_j$ and $\boldsymbol{q}$ in the feature label embeddings, respectively, where $\boldsymbol{\mu}_j$ is the IDR pattern of the $\boldsymbol{p}_{query}$, we have

$$\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{1}{a}\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t)}}^\top\boldsymbol{W}_Q^{(t)}\boldsymbol{p}_{query}^n)\delta{\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top>0,\tag{254}$$

$$\begin{aligned}&\sum_{b=0}^{t-1}\eta\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}{\boldsymbol{W}_{O_{i,\cdot}}^{(t)}}^\top\\ =&(\sum_{b=0}^{t_0-1}\eta\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\eta\sum_{i\in\mathcal{W}_n}V_i(t_0)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)})(\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)}+\quad\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}\Big|t=t_0)^\top\\ >&0+(\sum_{b=0}^{t_0-1}\eta\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\eta\sum_{i\in\mathcal{W}_n}V_i(t_0)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)})\quad\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}\frac{\partial\ell(\tilde{\boldsymbol{P}}^n,z^n;\Psi)}{\partial\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}\Big|t=t_0^\top\\ =&(\sum_{b=0}^{t_0-1}\eta\sum_{i\in\mathcal{W}_n}V_i(b){\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)}}^{(b)}+\eta\sum_{i\in\mathcal{W}_n}V_i(t_0)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)})\eta\frac{1}{B}\sum_{n\in\mathcal{B}_b}(-z^n)a_i\mathbb{1}[\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)}\sum_{s=1}^{l+1}(\boldsymbol{W}_V^{(t_0)}\boldsymbol{p}_s^n)\\ &\cdot\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t_0)}}^\top\boldsymbol{W}_Q^{(t_0)}\boldsymbol{p}_{query}^n)\geq 0]\cdot\sum_{s=1}^{l+1}\text{softmax}({\boldsymbol{p}_s^n}^\top{\boldsymbol{W}_K^{(t_0)}}^\top\boldsymbol{W}_Q^{(t_0)}\boldsymbol{p}_{query}^n)(\delta({\boldsymbol{p}_s^n}^\top,\boldsymbol{0}^\top)^\top\\ &+\sum_{b=0}^{t_0-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top),\\ >&0,\end{aligned}\tag{255}$$

where the first step is by the formula of the gradient descent, and the second step is by $(\sum_{b=0}^{t_0-1}\eta\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\eta\sum_{i\in\mathcal{W}_n}V_i(t_0)\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)})\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)}>0$ from the induction steps. The last step comes from the fact that $\|\sum_{b=0}^{t_0-1}\eta\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\|^2>0$ and $\sum_{b=0}^{t_0-1}\eta\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}\cdot\eta\sum_{i\in\mathcal{W}_n}V_i(t_0){\boldsymbol{W}_{O_{(i,\cdot)}}^{(t_0)}}^\top>0$ by the induction, and that $V_i(t)$ for $i\notin\mathcal{W}_n$ is much smaller than that in $\mathcal{W}_n$ given $B\gtrsim M_1$. Then,

$$\sum_{b=0}^{t-1}\eta(\sum_{i\in\mathcal{W}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\in\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}+\sum_{i\notin\mathcal{W}_n\cup\mathcal{U}_n}V_i(b)\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top{\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top>0,\tag{256}$$

since the norm $\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}$ for $i\notin\mathcal{W}_n$ is no larger than that for $i\in\mathcal{W}_n$. Combining (253), (254), and (256), we have

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}\boldsymbol{V}^n(t)>0.\tag{257}$$

Hence, we finish the induction.
(2) When $t \gtrsim \Theta(\beta)$, for $i \in \mathcal{W}$, by checking (244), we can deduce that

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)} {\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top + \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{a} \sum_{s=1}^{l+1} \text{softmax}({\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_K^{(t)}}^\top \boldsymbol{W}_Q^{(t)} \boldsymbol{p}_{query}^n) \delta {\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top > 0, \tag{258}$$

since that the accumulated label embedding term of $\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}$ contributed positively to ${\boldsymbol{p}_s^n}^\top {\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top$ and is larger than that of the feature embedding contribution by (154) and (155) (the gradient updates is close in the direction of the IDR pattern of $\boldsymbol{p}_{query}^n$ when $m \gtrsim M_1^2$). Since that $\|\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)} (\boldsymbol{\mu}_j^\top, \boldsymbol{0}^\top)^\top\| \leq \beta\xi$ for any $j \in [M_1]$, the effect of $\boldsymbol{W}_{O_{(i,\cdot)}}^{(0)} {\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top$ to the sign is much smaller than the remaining terms in (258). Hence, we show (258).
Then, since that the label embedding of $\boldsymbol{W}_{O_{(i,\cdot)}}$, $\boldsymbol{W}_{O_{(j,\cdot)}}$ are both close to $\boldsymbol{q}$ for $i, j \in \mathcal{W}$, and that the feature embedding of $\boldsymbol{W}_{O_{(i,\cdot)}}, i \in \mathcal{W}_n$ is close to the IDR pattern of $\boldsymbol{p}_{query}^n$, which is not the negative direction of the feature embedding of $\boldsymbol{W}_{O_{(j,\cdot)}}, j \in \mathcal{W}_{n'}$, we have for $j \in \mathcal{W} \backslash \mathcal{W}_n$,

$$\sum_{b=0}^{t-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) {\boldsymbol{W}_{O_{(i,\cdot)}}^{(b)}}^\top {\boldsymbol{W}_{O_{(j,\cdot)}}^{(t)}}^\top > 0. \tag{259}$$

Given that $V_i(t)$ for $i \notin \mathcal{W}_n$ is much smaller than that in $\mathcal{W}_n$ given $B \gtrsim M_1$ and the norm $\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}$ for $i \notin \mathcal{W}_n$ is no larger than that for $i \in \mathcal{W}_n$, we have

$$\sum_{b=0}^{t-1} \eta (\sum_{i \in \mathcal{W}_n} V_i(b) \boldsymbol{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \boldsymbol{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \boldsymbol{W}_{O_{(i,\cdot)}}^{(b)})^\top {\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)}}^\top > 0. \tag{260}$$

Therefore, we can derive that for $i \in \mathcal{W}$,

$$\boldsymbol{W}_{O_{(i,\cdot)}}^{(t)} \boldsymbol{V}^n(t) > 0. \tag{261}$$

□