

**OPTIMIZATION AND GENERALIZATION ANALYSIS OF  
ADVANCED NEURAL NETWORKS AND LEARNING  
ALGORITHMS**

**Hongkang Li**

Submitted in Partial Fullfillment of the Requirements  
for the Degree of

*DOCTOR OF PHILOSOPHY*

Approved by:  
Meng Wang, Chair  
Tianyi Chen  
Ali Tajer  
Pin-Yu Chen



*Department of Electrical, Computer, and Systems Engineering*  
Rensselaer Polytechnic Institute  
Troy, New York

[December 2024]

© Copyright 2024

by

Hongkang Li

All Rights Reserved

## CONTENTS

LIST OF TABLES . . . . .	x
LIST OF FIGURES . . . . .	xi
ACKNOWLEDGMENT . . . . .	xiv
ABSTRACT . . . . .	xvi
1. INTRODUCTION . . . . .	1
1.1 Learning with Advanced Neural Networks . . . . .	2
1.1.1 Generalization and Sample Complexity Analysis of Learning Vision Transformers . . . . .	3
1.1.2 Theoretical Understanding of Graph Transformers . . . . .	3
1.2 The Provable Effectiveness of Learning Algorithms . . . . .	4
1.2.1 Generalization of One-Hidden-Layer Neural Networks with Group Imbalance . . . . .	5
1.2.2 Learning with Graph Convolutional Network Using Graph Topology Sampling . . . . .	6
1.2.3 Training Dynamics and Generalization of In-Context Learning . . . . .	6
1.2.4 Generalization and Mechanism of Chain-of-Thought with Nonlinear Transformers . . . . .	7
2. A THEORETICAL UNDERSTANDING OF SHALLOW VISION TRANSFORMERS: LEARNING, GENERALIZATION, AND SAMPLE COMPLEXITY . . . . .	9
2.1 Introduction . . . . .	9
2.1.1 Background and Related Works . . . . .	11
2.2 Problem Formulation and the Learning Algorithm . . . . .	12
2.3 Theoretical Results . . . . .	13
2.3.1 Main Theoretical Insights . . . . .	13
2.3.2 Data Model . . . . .	14
2.3.3 Formal Theoretical Results . . . . .	15
2.4 Numerical Experiments . . . . .	19
2.4.1 Experiments on Synthetic Datasets . . . . .	19
2.4.2 Experiments on Image Classification Datasets . . . . .	22
2.5 Conclusion . . . . .	24

3. WHAT IMPROVES THE GENERALIZATION OF GRAPH TRANSFORMER? A THEORETICAL DIVE INTO SELF-ATTENTION AND POSITIONAL ENCODING	25
3.1 Introduction . . . . .	25
3.2 Related Works . . . . .	26
3.3 Problem Formulation and the Learning Algorithm . . . . .	27
3.4 Theoretical Results . . . . .	30
3.4.1 Theoretical Insights . . . . .	30
3.4.2 Data Model Assumptions . . . . .	30
3.4.3 Main Theoretical Results for Graph Transformers . . . . .	32
3.4.4 What Does Self-Attention Improve? A Comparison with GCN . . . . .	34
3.4.5 How Does Positional Encoding Guide the Graph Learning Process? .	35
3.4.6 Proof Sketch . . . . .	36
3.5 Numerical Experiments . . . . .	37
3.5.1 Experiments on Synthetic Data . . . . .	37
3.5.2 Experiments on Real-world Dataset . . . . .	40
3.6 Conclusion, Limitation, and Future Works . . . . .	41
4. HOW DOES PROMPTING THE MINORITY FRACTION AFFECT GENERALIZATION? A THEORETICAL STUDY OF ONE-HIDDEN-LAYER NEURAL NETWORK ON GROUP IMBALANCE	42
4.1 Introduction . . . . .	42
4.2 Background and Related Work . . . . .	44
4.3 Problem Formulation and Algorithm . . . . .	46
4.4 Main Theoretical Results . . . . .	47
4.4.1 Theoretical Insights . . . . .	50
4.4.2 Proof Idea and Technical Novelty . . . . .	51
4.5 Numerical Experiments . . . . .	53
4.5.1 Experiments on Synthetic Datasets . . . . .	53
4.5.2 Image Classification on Dataset CelebA . . . . .	57
4.6 Conclusions and Future Directions . . . . .	58
5. GENERALIZATION GUARANTEE OF TRAINING GRAPH CONVOLUTIONAL NETWORKS WITH GRAPH TOPOLOGY SAMPLING	59
5.1 Introduction . . . . .	59
5.1.1 Related Works . . . . .	61
5.1.2 Notations . . . . .	61

5.2	Training GCNs with Topology Sampling: Formulation and Main Components	62
5.3	Main Algorithmic and Theoretical Results . . . . .	64
5.3.1	Informal Key Theoretical Findings . . . . .	64
5.3.2	Graph Topology Sampling Strategy . . . . .	65
5.3.3	The Algorithmic Framework of Training GCNs . . . . .	68
5.3.4	Generalization Guarantee . . . . .	69
5.4	Numerical Results . . . . .	72
5.4.1	Sample Complexity and Neural Network Width with Respect to the Effective Adjacency Matrix . . . . .	73
5.4.2	Graph Sampling Affects the Effective Adjacency Matrix . . . . .	74
5.5	Conclusion . . . . .	75
6.	HOW DO NONLINEAR TRANSFORMERS LEARN AND GENERALIZE IN IN-CONTEXT LEARNING? . . . . .	77
6.1	Introduction . . . . .	77
6.1.1	Major Contributions of This Work . . . . .	78
6.1.2	Related Work . . . . .	79
6.2	Problem Formulation . . . . .	80
6.2.1	Training to Enhance ICL Capability . . . . .	80
6.2.2	Generalization Evaluation . . . . .	82
6.2.3	Training Algorithm . . . . .	82
6.2.4	Model Pruning . . . . .	83
6.3	Theoretical Results . . . . .	83
6.3.1	Main Theoretical Insights . . . . .	83
6.3.2	The Modeling of Training Data and Tasks . . . . .	84
6.3.3	In-Domain and Out-of-Domain Generalization with Sample Complexity Analysis . . . . .	86
6.3.4	ICL with Magnitude-Based Model Pruning . . . . .	88
6.4	The Mechanism of ICL by the Trained Transformer . . . . .	89
6.4.1	Self-Attention Selects Contexts with the Same IDR/ODR Pattern as the Query . . . . .	90
6.4.2	MLP Neurons Distinguish Label Embeddings Rather Than Feature Embeddings. . . . .	91
6.5	Numerical Experiments . . . . .	92
6.5.1	Experiments on the Generalization of ICL . . . . .	92
6.5.2	Experiments on the Mechanism of ICL . . . . .	94
6.6	Conclusion . . . . .	95

7. TRAINING NONLINEAR TRANSFORMERS FOR CHAIN-OF-THOUGHT INFERENCE: A THEORETICAL GENERALIZATION ANALYSIS . . . . .	96
7.1 Introduction . . . . .	96
7.1.1 Major Contributions . . . . .	97
7.1.2 Related Works . . . . .	98
7.2 Problem Formulation . . . . .	98
7.2.1 Training to Acquire the Chain-of-Thought Ability . . . . .	99
7.2.2 Training Algorithm . . . . .	100
7.2.3 Chain-of-Thought Inference . . . . .	100
7.2.4 In-Context Learning Inference . . . . .	101
7.3 Theoretical Results . . . . .	102
7.3.1 Main Theoretical Insights . . . . .	102
7.3.2 The Formulation of Data and Tasks . . . . .	103
7.3.3 The Sample Complexity Analysis of the Training Stage . . . . .	105
7.3.4 CoT Generalization Guarantee . . . . .	107
7.3.5 ICL Generalization and Comparison with CoT . . . . .	108
7.4 The Mechanism of CoT and the Proof Sketch . . . . .	109
7.4.1 Transformers Implement CoT by Attending to the Most Similar Examples Every Step . . . . .	109
7.4.2 An Overview of the Proof . . . . .	110
7.5 Numerical Experiments . . . . .	111
7.6 Conclusion, Limitations, and Future Works . . . . .	114
8. CONCLUSION AND FUTURE DIRECTIONS . . . . .	115
REFERENCES . . . . .	116
APPENDICES . . . . .	136
A. APPENDIX OF CHAPTER 2 . . . . .	136
A.1 Comparison with Two Related Works . . . . .	136
A.1.1 Comparison with (Allen-Zhu & Li, 2023) . . . . .	136
A.1.2 Comparison with (Jelassi et al., 2022) . . . . .	137
A.2 Preliminaries . . . . .	137
A.3 Proof of the Main Theorem and Propositions . . . . .	140
A.4 Proof of Lemma A.3.1 . . . . .	148
A.4.1 Proof of Claim 1 of Lemma A.3.1 . . . . .	148

A.4.2	Proof of Claim 2 of Lemma A.3.1 . . . . .	156
A.4.3	Proof of Claim 3 of Lemma A.3.1 . . . . .	169
A.5	Other Useful Lemmas . . . . .	171
A.6	Extension to More General Cases . . . . .	174
A.6.1	Extension to Multi-Classification . . . . .	174
A.6.2	Extension to A More General Data Model . . . . .	176
A.6.3	Extension to Multi-Head Networks . . . . .	178
A.6.4	Extension to Skip Connections and Normalization . . . . .	178
B.	APPENDIX OF CHAPTER 3 . . . . .	180
B.1	Additional Experiments . . . . .	180
B.1.1	Verifying Assumptions Made on the Graph Data Model . . . . .	180
B.1.2	Experiments on Synthetic Dataset . . . . .	182
B.1.3	Experiments on Real-world Datasets . . . . .	182
B.2	Preliminaries . . . . .	184
B.3	Key Lemmas and Proof of the Main Theorems . . . . .	189
B.4	Useful Lemmas . . . . .	199
B.5	Extension of Our Analysis and Additional Discussion . . . . .	227
B.5.1	Assumption on the Pre-Trained Model . . . . .	227
B.5.2	Extension to Other Positional Encodings . . . . .	227
B.5.3	Extension of the Analysis on GAT . . . . .	229
B.5.4	Extension to Graph Classification Problems . . . . .	229
B.5.5	Extension to Multi-Classification . . . . .	230
B.5.6	Comparision with Other Frameworks of Analysis . . . . .	231
C.	APPENDIX OF CHAPTER 4 . . . . .	233
C.1	Algorithm . . . . .	233
C.1.1	Numerical Evaluation of Tensor Initialization . . . . .	235
C.2	Preliminaries of the Main Proof . . . . .	236
C.3	Proof of Theorem 2 and Corollary 4.4.1 . . . . .	255
C.4	Proof of Lemma C.3.1 . . . . .	265
C.4.1	Useful Lemmas in the Proof of Lemma C.3.1 . . . . .	266
C.4.2	Proof of Lemma C.3.1 . . . . .	267
C.4.3	Proof of Lemma C.4.1 . . . . .	269
C.4.4	Proof of Lemma C.4.2 . . . . .	270
C.4.5	Proof of Lemma C.4.3 . . . . .	271

C.4.6	Proof of Lemma C.4.4 . . . . .	274
C.4.7	Proof of Lemma C.4.5 . . . . .	280
C.5	Proof of Lemma C.3.2 . . . . .	286
C.5.1	A Useful Lemma Used in the Proof . . . . .	286
C.5.2	Proof of Lemma C.3.2 . . . . .	290
C.6	Proof of Lemma C.3.3 . . . . .	292
C.6.1	Useful Lemmas in the Proof . . . . .	293
C.6.2	Proof of Lemma C.3.3 . . . . .	295
C.6.3	Proof of Lemma C.6.1 . . . . .	296
C.6.4	Proof of Lemma C.6.2 . . . . .	299
C.6.5	Proof of Lemma C.6.3 . . . . .	300
D.	APPENDIX OF CHAPTER 5 . . . . .	302
D.1	Preliminaries . . . . .	302
D.1.1	Symmetric Graph Sampling Method . . . . .	304
D.2	Node Classification for Three Layers . . . . .	306
D.2.1	Lemmas . . . . .	307
E.	APPENDIX OF CHAPTER 6 . . . . .	346
E.1	Proof Sketch . . . . .	346
E.2	Addition Discussions and Extensions . . . . .	347
E.2.1	The Motivation to Study Nonlinear Transformers . . . . .	347
E.2.2	The Discussion on Single/Multi-Head Attention . . . . .	347
E.2.3	Extension to Multiple Patterns for One Class . . . . .	348
E.2.4	Additional Related Works . . . . .	348
E.3	Additional Experiments and the Algorithm . . . . .	349
E.3.1	The Impact of ALPHA . . . . .	349
E.3.2	The Required Number of Training Tasks . . . . .	349
E.4	Proofs of the Main Theorems . . . . .	350
E.4.1	Proof Overview of Main Theorems . . . . .	350
E.4.2	Preliminaries . . . . .	352
E.4.3	Proof of Theorem 6.3.3 . . . . .	358
E.4.4	Proof of Theorem 6.3.4 . . . . .	362
E.4.5	Proof of Theorem 6.3.7 . . . . .	364
E.5	Proofs of Key Lemmas and Propositions . . . . .	364
E.5.1	Proof of Lemma E.4.11 . . . . .	364

E.5.2	Proof of Proposition 6.4.1 . . . . .	365
E.5.3	Proof of Corollary 6.4.3 . . . . .	368
E.5.4	Proof of Proposition 6.4.5 . . . . .	369
E.5.5	Proof of Lemma E.4.5 . . . . .	371
E.5.6	Proof of Lemma E.4.6 . . . . .	388
E.5.7	Proof of Lemma E.4.7 . . . . .	392
E.5.8	Proof of Lemma E.4.9 . . . . .	397
E.5.9	Proof of Lemma E.4.10 . . . . .	398
F.	APPENDIX OF CHAPTER 7 . . . . .	403
F.1	Additional Discussions . . . . .	403
F.1.1	The Motivation to Study One-Layer Single-Head Transformers . . . . .	403
F.1.2	The Motivation of the Data and Task Formulation . . . . .	404
F.1.3	The Discussion of Positional Encoding . . . . .	404
F.2	Algorithms . . . . .	405
F.3	Preliminaries . . . . .	405
F.4	Proof of Main Theorems . . . . .	410
F.4.1	Proof of Theorem 3 . . . . .	410
F.4.2	Proof of Theorem 4 . . . . .	413
F.4.3	Proof of Theorem 5 . . . . .	417
F.4.4	Proof of Proposition 3 . . . . .	419
F.5	Proof of Lemmas . . . . .	419
F.5.1	Proof of Lemma F.3.5 . . . . .	419
F.5.2	Proof of Lemma F.3.6 . . . . .	433
F.5.3	Proof of Lemma F.3.7 . . . . .	436

## LIST OF TABLES

2.1	Some important notations. . . . .	15
3.1	Some important notations. . . . .	33
4.1	Impact of GMM parameters on the learning performance in sample regimes. ©2024 IEEE. . . . .	49
5.1	Parameter choices for Algorithm 2. . . . .	70
6.1	Comparison with existing works about training analysis and generalization guarantee of in-context learning. . . . .	79
A.1	Summary of notations. . . . .	138
B.1	The fraction of discriminative nodes in each class of Cora. . . . .	184
B.2	The fraction of discriminative nodes in each class of PubMed. . . . .	184
B.3	The fraction of discriminative nodes in each class of Actor. . . . .	184
B.4	The fraction of discriminative nodes in each class of PascalVOC-SP-1G. . . . .	184
B.5	The fraction of nodes satisfying $\Delta_n(z_m) > 0$ . . . . .	185
B.6	The statistics of datasets. . . . .	185
B.7	Summary of notations. . . . .	186
B.8	The fraction of discriminative nodes in each class of Actor. . . . .	187
C.1	Summary of notations. ©2024 IEEE. . . . .	234
D.1	Summary of notations. . . . .	307
D.2	Full parameter choices for three-layer GCN. . . . .	308
E.1	Summary of notations. . . . .	351
E.2	Summary of notations (Continued). . . . .	352
F.1	Summary of notations. . . . .	407
F.2	Summary of notations (Continued). . . . .	408

## LIST OF FIGURES

2.1	The impact of (a) $\alpha_*$ and (b) $\sigma$ on sample complexity . . . . .	20
2.2	The number of iterations against $\alpha_*^{-1}$ . . . . .	21
2.3	Comparison between ViT and CNN. . . . .	21
2.4	Concentration of attention weights. . . . .	22
2.5	Impact of token sparsification on testing loss. . . . .	22
2.6	(a) Test accuracy when $N$ and $\alpha_*$ change. (b) Relationship of sample complexity against $\alpha^*$ . (c) Test accuracy when token sparsification removes spurious correlations. . . . .	24
3.1	Graph Transformers in (3.1). . . . .	28
3.2	Example of the winning margin. Node $n$ has a non-discriminative feature $\mu_3$ and label +1. Then $\Delta_n(1) = -2$ , and $\Delta_n(2) = 3$ . . . . .	32
3.3	The impact of (a) $\gamma_d$ and (b) $\epsilon_S$ on the sample complexity of GT. . . . .	38
3.4	The test Hinge loss against the number of epochs for different $\epsilon_0$ . . . . .	39
3.5	Concentration of attention weights. . . . .	39
3.6	Sample complexity against $\gamma_d$ . . . . .	39
3.7	The required # of iterations against $\gamma_d$ . . . . .	39
3.8	The values of entries of $\mathbf{b}$ and the test accuracy of PE-based sampling. Left to right: PubMed, Actor, PascalVOC-SP-1G. . . . .	40
3.9	Test accuracy of GT with/without PE and GCN when the number of labeled nodes varies. Left to right: PubMed, Actor, PascalVOC-SP-1G. . . . .	40
4.1	Group imbalance experiment. (a) Binary classification on CelebA dataset using Gaussian augmentation to control the minority group co-variance. (b) Test accuracy against the augmented noise level. ©2024 IEEE. . . . .	43
4.2	The sample complexity when the feature dimension changes. ©2024 IEEE. . .	54
4.3	The sample complexity (a) when one mean changes, (b) when one co-variance changes. ©2024 IEEE. . . . .	54
4.4	(a) The convergence rate with different $\mu_1$ . (b) The convergence rate with different $\Sigma$ . ©2024 IEEE. . . . .	55
4.5	(a) Convergence rate when the number of neurons $K$ changes. (b) The relative error of the learned model when $n$ changes. ©2024 IEEE. . . . .	56

4.6	(a) The cross-entropy test loss when the co-variance of the minority group changes. (b) The cross-entropy test loss when the mean of the minority group changes. ©2024 IEEE.	56
4.7	The test loss (cross entropy loss) of synthetic data with different $\lambda_2$ values. (a) Group 2 has a smaller level of co-variance. (b) Group 2 has a larger level of co-variance. ©2024 IEEE.	57
4.8	The test accuracy on CelebA dataset has opposite trends when the minority group fraction increases. (a) Male group is the minority (b) Female group is the minority. ©2024 IEEE.	58
5.1	The testing error when $ \Omega $ and $\ \mathbf{A}^*\ _\infty$ change. $m = 500$ .	74
5.2	The testing error when $m$ and $\ \mathbf{A}^*\ _\infty$ change. $ \Omega  = 1500$ .	74
5.3	Generalization performance of learned GCNs on datasets generated from different $\hat{\mathbf{A}}$ by (a) our graph sampling strategy and (b) FastGCN. $\mathbf{A}$ is very unbalanced.	75
5.4	Generalization performance of learned GCNs on datasets generated from different $\mathbf{A}^*$ by (a) our graph sampling strategy and (b) FastGCN. $\mathbf{A}$ is balanced.	76
6.1	(a) Example of prompt embedding. $l = 3$ , $\alpha = 2/3$ . (b) The mechanism of a trained Transformer (6.2) to implement ICL. Part I: The attention layer assigns the largest attention score (0.8) on $\boldsymbol{\mu}_1 - 0.3\boldsymbol{\nu}_5$ , which has the same IDR pattern as the query. Then the weighted sum of input tokens is close to $(\boldsymbol{\mu}_1^\top, \mathbf{q}^\top)^\top$ by the trained attention layer. Part II: The neurons in $\mathbf{W}_O \mathbf{W}_V$ with a large magnitude are aligned with $\bar{\boldsymbol{\mu}}$ and $\pm \mathbf{q}$ in the first $d_X$ and the rest $d_Y$ dimensions, respectively. Then the prediction is based on the part of $\pm \mathbf{q}$ that varies for different queries rather than the part of $\bar{\boldsymbol{\mu}}$ that is universal for all IDR patterns.	83
6.2	The properties of the trained model. (a) The average norm of $\mathbf{W}_Q \mathbf{p}_{query}$ , $\mathbf{W}_K \mathbf{p}_i$ , $[XDR(\beta^{-1} \cdot \mathbf{p}_{query})^\top, \mathbf{0}^\top] \mathbf{W}_Q \mathbf{p}_{query}$ , and $[XDR(\mathbf{p}_i)^\top / \beta, \mathbf{0}^\top] \mathbf{W}_K \mathbf{p}_i$ . (b) The attention weight summation on contexts with the same ODR pattern as the query and other contexts. (c) The magnitude of the first $d_X$ dimensions of 5 neurons in $\mathbf{W}_O \mathbf{W}_V$ and their angles to $\bar{\boldsymbol{\mu}}$ in 400 epochs. (d) The magnitude of the rest $d_Y$ dimensions of 10 neurons in $\mathbf{W}_O \mathbf{W}_V$ and their angles to $\mathbf{q}$ in 400 epochs. We choose 5 neurons for $a_i > 0$ and 5 for $a_i < 0$ .	89
6.3	Out-of-domain ICL classification error on GPT-2 with (a) different $S_1$ on GPT-2 (b) different $\alpha'$ for in-domain (id) and out-of-domain (ood) generalization.	93
6.4	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) $\alpha' = 0.6$ .	93

6.5	(a) Out-of-domain classification error (left y-axis for curves) with model pruning of the trained $\mathbf{W}_O$ using baseline (no pruning), random pruning, and magnitude-based pruning (Mag.-based), and the magnitude of each neuron of $\mathbf{W}_O$ (right y-axis for light blue bars) (b) Out-of-domain classification error when varying $\alpha'$ . These two are implemented on a one-layer Transformer. . . . .	94
7.1	An example of a two-step inference. . . . .	106
7.2	Concentration of attention weights for CoT inference. . . . .	109
7.3	CoT testing error with different (a) $\alpha'$ (b) $\tau^f$ (c) $\rho^f$ . . . . .	112
7.4	ICL testing error with different (a) $\alpha'$ (b) $\tau_o^f$ (c) $\rho_o^f$ . . . . .	112
7.5	Comparison between CoT and ICL w./w.o. Condition 7.3.2. . . . .	112
7.6	Training dynamics of Transformers for CoT. . . . .	113
7.7	Training dynamics of Transformers. (a) Layer 1, Head 2 (b) Layer 2 Head 2 (c) Layer 3 Head 2. . . . .	113
B.1	Eigenvalues of the covariance matrix of the feature matrix of all classes of Cora. . . . .	182
B.2	Eigenvalues of the covariance matrix of the feature matrix of all classes of PubMed. . . . .	182
B.3	Eigenvalues of the covariance matrix of the feature matrix of all classes of Actor. . . . .	183
B.4	Eigenvalues of the covariance matrix of the feature matrix of all classes of PascalVOC-SP-1G. . . . .	183
B.5	Normalized $\bar{\Delta}(z)$ for Cora, PubMed, Actor, and PascalVOC-SP-1G. . . . .	185
B.6	The required number of iterations against $\gamma_d^{-2}$ (a) Graph Transformer (b) GCN. . . . .	185
B.7	The values of entries of $\mathbf{b}$ and the test accuracy of PE-based sampling for Ogbn-Arxiv. . . . .	186
B.8	Test accuracy of GT with/without PE and GCN when the number of label nodes varies for Ogbn-Arxiv. . . . .	186
C.1	Comparison between tensor initialization, a random initialization near $\mathbf{W}^*$ , and an arbitrary random initialization. ©2024 IEEE. . . . .	237
D.1	(a) Dependency between $\mathbf{a}_s \mathbf{X}$ and $\mathbf{a}_p \mathbf{X}$ (b) Dependency between $y_i$ and $y_j$ . . . . .	344
E.1	The prompt length against $\alpha$ , and the required number of training iterations against $\alpha$ . . . . .	350
E.2	The required number of training tasks for different $M_1$ . . . . .	350
F.1	CoT mechanism with standard PE of (a) Layer 1 (b) Layer 2 (c) Layer 3. . . . .	405

## ACKNOWLEDGMENT

I would like to take this opportunity to express my gratitude to everyone who has supported me throughout my life and studies during my Ph.D at RPI.

I want to thank my girlfriend Chenyi Kuang for her companionship and support over the past five years. We went from graduating from undergrad in China together to completing our Ph.D. together in USA—you are truly irreplaceable. I'd also like to thank our cat Crusoe for keeping me company and helping me keep up my daily exercise.

I want to express my gratitude to my parents for their care and upbringing, as well as their support for my studies abroad. I am also grateful to all my family members, especially my grandmother. Thank you to all my former classmates for their help and support along the way.

I am especially grateful to my advisor, Prof. Meng Wang, for her guidance and support throughout my Ph.D. Her extensive knowledge and patience helped me grow from an ordinary undergraduate into an independent researcher. The positive lab environment she fostered allowed me to balance life and studies, making my five years in Troy a wonderful experience. I would also like to thank my collaborators, Dr. Pin-Yu Chen, Prof. Sijia Liu, Dr. Songtao Lu, Dr. Xiaodong Cui, Dr. Hui Wan, and Prof. Jinjun Xiong, who provided valuable advice on experiments, theory, and writing, helping me improve the quality of my papers. Dr. Pin-Yu Chen, in particular, served as a member of my thesis committee. I am also grateful to my other committee members Prof. Ali Tajer and Prof. Tianyi Chen, whose feedback greatly contributed to the writing and presentation of my thesis.

I want to thank the lab members for their support. Dr. Shuai Zhang provided substantial guidance on various aspects of my early research. Dr. Ming Yi, Dr. Wenting Li, and Dr. Ren Wang kindly gave me a lot of advice in life. I have also benefited greatly from discussions with other group members, including Nowaz, Yating, Jiawei, Niranjan, Ishtiaq, and Amir. Additionally, I am grateful to several RPI undergraduates, Ruisi Jian, Haolin Xiong, and Peilin Lai, who contributed to my research. I also want to acknowledge the help of collaborators and students from outside RPI. Yihua has been a longstanding collaborator, making significant contributions to the experiments in our joint papers. Discussions with Yingcong, Zaixi, Yu, Ruiqi, and Yao-Chieh were also very helpful in broadening my perspective and generating ideas.

I would like to thank the ECSE staff, especially Kelly, for all the efforts made to facilitate students' research and lives. I am also especially grateful to the ECSE department and RPI for their financial support, which enabled me to attend various conferences.

Finally, I would like to thank everyone else who has helped me along the way, even if their names were not mentioned above.

## ABSTRACT

In recent years, deep learning has undergone rapid advancement. A notable trend in this development is the enhancement of learning efficiency for large foundation models, giving rise to numerous advanced learning algorithms designed for advanced neural networks. However, there remains a limited theoretical understanding of these algorithms and deep models. This thesis addresses this gap by delving into the optimization and generalization in advanced neural networks.

The first part of the thesis focuses on the theoretical investigation of the basic block of advanced neural networks. The first work is to study one-layer single-head Vision Transformers (ViT), which is a self-attention layer followed by a two-layer perceptron. This work provides the sample complexity and the required number of iterations to achieve zero generalization on a binary classification task based on a data model where each data contains several label-relevant and label-irrelevant tokens. The sample complexity bound implies that a larger fraction of label-relevant tokens, a smaller token noise level, and a smaller initial model error can enhance the generalization. The theoretical finding also verifies the general intuition about the success of attention by proving that training using stochastic gradient descent (SGD) generates a sparse attention map focused on label-relevant tokens. Moreover, we also conclude that proper token sparsification can improve performance by removing label-irrelevant or noisy tokens, including spurious correlations. We then explore the generalization of Graph Transformer, a developing architecture originated from Transformers for graph learning. This work is based on a graph data model with discriminative nodes that determine node labels and non-discriminative nodes that are class-irrelevant. The theoretical results quantitatively characterize the sample complexity and number of iterations for convergence dependent on the fraction of discriminative nodes, the dominant patterns, and the fraction of erroneous labels. Meanwhile, we show that self-attention and positional encoding lead to generalization by making the attention map sparse and promoting the core neighborhood. This explains the superior feature representation of Graph Transformers compared with GCN and Graph Transformer without positional encoding.

The second part of this thesis is to study modern algorithms on basic neural models. The first work studies learning with the group imbalance issue on a one-hidden-layer fully-connected neural network with a mixture of Gaussian input. This work quantifies the

impact of individual groups on learning performance. The theoretical results include that when all group-level co-variance is in the medium regime and all mean are close to zero, we can achieve a small sample complexity, a fast training rate, and a high average and group-level testing accuracy. Moreover, it is shown that increasing the fraction of the minority group in the training data does not necessarily improve the generalization performance of the minority group. The second is the graph topology sampling with a three-layer Graph Convolutional Network (GCN), which not only consists of three-layer networks but also includes graph information in the model. This work characterizes sufficient conditions for graph topology sampling, such that GCN training leads to a diminishing generalization error on a semi-supervised node classification task. The sample complexity result explicitly depicts the impact of graph structures and topology sampling on the generalization performance. The third work is to study in-context learning (ICL), an inference method using pairs of testing data and labels as a prompt to make predictions without fine-tuning the model. We theoretically quantify the required number of training prompts and iterations and the length and distribution of the testing prompts for a desired ICL capability on unseen tasks with and without data distribution shifts. The training dynamics analysis also characterizes how different components in the learned Transformers contribute to the ICL performance. Moreover, this work proves that proper magnitude-based pruning has a minimal impact on performance while reducing inference costs. The last work is about Chain-of-Thought (CoT), a prompting method that incorporates multiple intermediate steps into each context example. This work establishes the first theoretical analysis of training nonlinear Transformers to obtain the CoT generalization ability by quantifying the required training samples and iterations. This work next theoretically characterizes the conditions for an accurate inference output by CoT when the provided reasoning examples contain noises and are not always accurate. Meanwhile, ICL, i.e., one-step CoT without intermediate steps, may fail to provide an accurate output when CoT does.

# CHAPTER 1

## INTRODUCTION

In the era of rapid advancements in artificial intelligence, the continuous evolution of neural networks and complicated learning algorithms forms a critical foundation driving progress across various domains. As researchers dive deeper into deep learning, optimization and generalization emerge as essential challenges. These factors are vital not only for understanding the underlying mechanisms of neural network architectures but also for enhancing the performance and scalability of learning methods. While there has been substantial research on basic neural network structures and fundamental algorithms, much remains unexplored in areas involving deeper networks and complicated data models and algorithms. Addressing these problems is crucial for advancing the understanding and application of neural networks in terms of computational efficiency and model interpretability, particularly in real-world scenarios where complexity and the amount of data continue to escalate. This thesis seeks to concentrate on these advanced, relatively uncharted territories within ML research, focusing on analyzing optimization techniques and generalization capabilities in deeper and more complex neural networks. By doing so, this thesis aims to contribute new insights and methods that can better align neural network research with the demands of increasingly diverse and challenging applications.

---

Portions of this chapter have previously appeared as: H. Li, M. Wang, S. Liu, and P.-Y. Chen, “A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity,” in *Proc. Int. Conf. Learn. Represent.*, May 2023, Paper 3368.

Portions of this chapter have previously appeared as: H. Li, M. Wang, T. Ma, S. Liu, Z. Zhang, and P.-Y. Chen, “What improves the generalization of graph transformer? A theoretical dive into self-attention and positional encoding,” in *Proc. Int. Conf. Mach. Learn.*, Jul. 2024, pp. 28784–28829.

Portions of this chapter have previously appeared as: H. Li, S. Zhang, Y. Zhang, M. Wang, S. Liu, and P.-Y. Chen, “How does promoting the minority fraction affect generalization? A theoretical study of one-hidden-layer neural network on group imbalance,” *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 216–231, Mar. 2024. ©2024 IEEE.

Portions of this chapter have previously appeared as: H. Li, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, “Generalization guarantee of training graph convolutional networks with graph topology sampling,” in *Proc. Int. Conf. Mach. Learn.*, Jul. 2022, pp. 13014–13051.

Portions of this chapter have previously appeared as: H. Li, M. Wang, S. Lu, X. Cui, and P.-Y. Chen, “How do nonlinear transformers learn and generalize in in-context learning?” in *Proc. Int. Conf. Mach. Learn.*, Jul. 2024, pp. 28734–28783.

Portions of this chapter have previously appeared as: H. Li, M. Wang, S. Lu, X. Cui, and P.-Y. Chen, “Training nonlinear transformers for chain-of-thought inference: A theoretical generalization analysis,” 2024, *arXiv:2410.02167*.

Portions of this chapter have previously appeared as: H. Li, S. Zhang, M. Wang, “Learning and generalization of one-hidden-layer neural networks, going beyond standard gaussian data,” In *Annu. Conf. Inf. Sci. Syst.*, Mar. 2022, pp. 37–42.

## 1.1 Learning with Advanced Neural Networks

In recent years, advanced neural network models based on Transformers [1] have gradually replaced traditional convolutional neural networks (CNNs) and recurrent neural networks (RNNs), achieving impressive results in fields such as computer vision, natural language processing, and recommendation systems. Transformers have become the mainstream framework for these tasks. In particular, Transformer-based large language models have demonstrated remarkable capabilities in tasks like language generation and complex reasoning. These experimental advances have sparked a need for a deeper theoretical understanding of Transformers, specifically addressing the question:

*Why can Transformers be trained to achieve outstanding learning and generalization abilities on different tasks?*

Focusing on Vision Transformers [2] and Graph Transformers [3], [4], [5], we investigate this problem of learning vision and graph data for classification problems.

**The mechanism of attention and positional encoding.** A Transformer layer is typically composed of multiple self-attention heads and a multi-layer perceptron (MLP). Positional encoding is usually added to the model within the attention layer. Compared to traditional neural networks, aside from shared MLP structures, normalization, and residual blocks, one unique feature of Transformers is the use of self-attention layers and positional encoding. This raises another question: what roles do self-attention and positional encoding play in the learning process of the model? This thesis studies the mechanism of attention and positional encoding by characterizing the training dynamics of these modules.

**The training dynamics and generalization analysis.** The softmax activation in the attention layer introduces new nonlinear components to the model, making the loss landscape of the training problem more complex and challenging to analyze. Some existing works consider analysis within the Neural Tangent Kernel (NTK) setting, assuming very wide networks so that the optimal parameters can be found near the initialization. However, such approaches rely on linearized approximations of nonlinear models, which are unrealistic. Other studies consider a setting where a ground-truth model determines the model output. In this case, gradient training near the ground-truth parameters allows for a characterization of the loss landscape and global convergence. However, this type of work does not extend to the overparameterized setting. To better characterize the training dynamics of models,

we adopt a feature learning analysis framework, which, by assuming low-dimensional and sparse structures within the data, enables easier analysis of key representations related to attention mechanisms through gradient updates. Beyond convergence analysis, we leverage the bounded model outputs to prove a diminishing generalization gap, forming the basis of our generalization analysis.

The following two subsections introduce our work on analyzing Vision Transformers [6] and Graph Transformers [7], respectively.

### 1.1.1 Generalization and Sample Complexity Analysis of Learning Vision Transformers

Vision Transformer (ViT) [2] now serves as the backbone of many vision tasks due to its superior performance. Existing theoretical works only concentrate on the expressive power and Turing-completeness [8], [9], [10], [11], [12], [13], [14], [15], [16] with statistical guarantees [17], [18] of ViTs.

Our work studies the conditions where a one-layer ViT, i.e., a self-attention layer followed by a two-layer perceptron, achieves desirable generalization. We focus on a binary classification problem on structured data, where tokens with discriminative patterns determine the label from a majority vote, while tokens with non-discriminative patterns do not affect the labels. We explicitly characterize the sample complexity to achieve a desirable generalization performance. The highlights of the technical contributions include: (1) proposing a new analytical framework to tackle the non-convex optimization and generalization for shallow ViTs; (2) theoretically depicting the evolution of the attention map during the training and characterizing how “attention” is paid to different tokens during the training; (3) providing a theoretical explanation for the improved generalization using the “token sparsification” technique, an efficient method to improve the computational complexity by removing redundant tokens of data. The results show that if removing class-irrelevant, highly noisy tokens, and/or spurious correlated tokens, then the sample complexity is reduced while achieving the same testing accuracy.

### 1.1.2 Theoretical Understanding of Graph Transformers

Graph Transformer [3], [4], [5] is a new architecture that incorporates Transformer into graph learning. It is designed specifically to handle graph data by constructing positional

embeddings that capture important graph information and using nodes as input tokens for the Transformer model. Some critical theoretical aspects of Graph Transformers remain much less explored due to the complicated architecture. Existing works either study Graph Transformers by comparing their expressive power with conventional graph neural networks [5], [19] or explain the success of positioned encoding by topology and/or spectral theory [4], [20], [21] without any optimization or generalization analysis.

We focus on a basic block of Graph Transformer where there is one self-attention layer with relative positional encoding, followed by a two-layer perceptron. The goal of this work is to study (1) how a Graph Transformer can achieve adequate generalization, and (2) the advantage of self-attention and positional encoding in graph learning. We consider a semi-supervised binary node classification problem on structured graph data, where each node contains either a discriminative or a non-discriminative pattern, and the ground truth label is determined by the dominant discriminative pattern in the core neighborhood. Our contributions are as follows. First, we develop a novel framework to analyze the optimization and generalization of shallow graph transformers. Second, we characterize the advantages of the self-attention. Third, this work demonstrates that positional encoding enhances the generalization by promoting the core neighborhood.

## 1.2 The Provable Effectiveness of Learning Algorithms

With the development of deep learning models, the research focus has shifted beyond only learning performance to address other issues within deep learning systems. For example, people observe that larger models and more data lead to increased computational and memory costs during training and inference. Additionally, an imbalance in training data can cause models to exhibit preferences for certain inputs or outputs, raising concerns about fairness in deep learning systems. Our works analyze algorithms that concentrate on these issues.

**Efficient learning methods.** To improve computational efficiency, various efficient deep learning algorithms for different scenarios have been proposed. For instance, in the training of graph neural networks, numerous graph sampling methods have been developed to address the memory challenges of large-scale graphs. These methods involve removing certain edges or nodes from the graph in each training iteration. Subsection 1.2.2 presents our analysis of graph topology sampling [22]. Furthermore, we have explored a new paradigm in efficient deep learning within large language models: In-Context Learning (ICL). This

approach generates correct outputs by inputting appropriate prompts into the pre-trained model, without the need for extensive parameter fine-tuning. The prompt design can include input-label pairs or input-output examples with intermediate reasoning steps, the latter being known as Chain-of-Thought (CoT). Subsections 1.2.3 and 1.2.4 discuss the theoretical analysis of ICL [23] and CoT [24], which also includes an analysis of model pruning during in-context inference.

**Data imbalance issue in deep learning.** Imbalanced data usually means that a few classes of data make up the majority of the dataset, while the majority of classes have few data samples, which typically refers to many underrepresented minority groups, such as women, ethnic minorities, people with disabilities, and so on. Training on such imbalanced data may result in a deep learning model that is unfair to minority groups. Many approaches, such as data augmentation and oversampling, have been proposed to improve performance in imbalanced learning. However, there is still a lack of theoretical analysis explaining how these methods affect the generalization performance of the model across different groups. Subsection 1.2.1 delves into this issue [25] in more detail.

### 1.2.1 Generalization of One-Hidden-Layer Neural Networks with Group Imbalance

Group imbalance [26], [27], [28], [29], [30] refers to that a well-trained model with high average accuracy may have significant errors on the minority group that infrequently appears in the data. Empirical methods to alleviate group imbalance include data augmentation, reweighting minority groups, etc. However, no existing works justify the validity of these methods from a theoretical perspective.

This work theoretically characterizes the impact of empirical risk minimization on group imbalance. Assuming data from Gaussian mixture distribution, where samples of each group are generated from a Gaussian distribution with an arbitrary mean vector and co-variance matrix, this chapter quantifies the impact of individual groups on the sample complexity, the training convergence rate, and the average and group-level test error. Our key results include (1) Medium-range group-level co-variance enhances the learning performance. When a group-level co-variance deviates from the medium regime, the algorithm converges slower, and both the average and group-level test error increases. (2) Group-level mean shifts from zero hurt the learning performance. When a group-level mean deviates from zero, the sample

complexity increases, the learning performance degrades with higher sample complexity, slower convergence in training, and worse average and group-level generalization performance. (3) Increasing the fraction of the minority group in the training data does not always improve its generalization performance. This is because generalization performance is also affected by the mean and co-variance of individual groups.

### 1.2.2 Learning with Graph Convolutional Network Using Graph Topology Sampling

Graph convolutional neural networks (GCNs) are a common model for graph learning, which aggregate each node’s embedding with its neighboring nodes’ embedding in each layer. GCNs have demonstrated great empirical advantage in many areas [31], [32], [33], [34], which is often achieved at a cost of high computational and memory costs, especially for large graphs. To alleviate the computational cost in training, various graph topology sampling methods have been proposed to only aggregate the node embeddings from a selected subset of neighbors.

This work studies the question: Under what conditions does a GCN learned with graph topology sampling achieve satisfactory generalization? The existing work [35] only analyzes the convergence of graph sampling without a generalization guarantee. We provide the first generalization analysis of training GCNs with graph topology sampling on semi-supervised node classification problems. The contributions can be summarized as (1) We propose a training framework that implements both stochastic gradient descent (SGD) and graph topology sampling, and the learned three-layer GCN model is guaranteed to approach the best generalization performance of a large class of target functions. (2) we explicitly characterize the impact of graph topology sampling on the generalization performance through the proposed effective adjacency matrix. We show that learning with topology sampling performs the same generalization as training GCNs using the effective adjacency matrix. (3) the results indicate that the required number of labeled nodes is a polynomial of the infinity norm of the effective adjacency matrix and the maximum node degree.

### 1.2.3 Training Dynamics and Generalization of In-Context Learning

In-context learning (ICL) refers to adding input-label pairs before the query as context examples for the current task, thereby activating the pre-trained model’s predictive abilities

for that task. This approach eliminates the need to finetune the pre-trained model, which significantly saves computational resources. It leverages the knowledge learned by the pre-trained model, which is analogous to the human ability to deduce reasoning patterns from a limited number of examples. Existing works theoretically study ICL by characterizing (1) the existence of a Transformer with proper parameters to implement gradient descent or its variant, or (2) the convergence and generalization analysis of linear Transformers on linear regression tasks only.

We aim to study how a Transformer can be trained to perform ICL in and out of domain successfully and efficiently. The key differences compared with previous works include that (1) we study nonlinear Transformers with softmax self-attention and Relu MLP layer instead of linear Transformer, and (2) we establish the ICL generalization analysis on distribution-shifted data on classification tasks. Specifically, by focusing on a group of binary classification tasks, we theoretically show that training a nonlinear one-layer single-head Transformer using training prompts from a subset of these tasks can return a model with the ICL generalization capability to tasks that are unseen during the training with possible data distribution shifts. We quantify the required number of training data, iterations, the length of prompts, and the resulting ICL performance. The training dynamics analysis allows us to characterize the mechanism of the ICL capability of Transformers on classification tasks, which is aligned with the empirical observations of the “induction head”. After training, the attention weights are concentrated on contexts that share the same relevant pattern as the query, and the MLP layer makes predictions based on the label embedding of the attended-to examples. Furthermore, we theoretically justify the validity of magnitude-based pruning in ICL inference by showing that removing small trained neurons has little effect on the generalization. This is crucial to improve the inference efficiency.

#### **1.2.4 Generalization and Mechanism of Chain-of-Thought with Nonlinear Transformers**

Chain-of-thought (CoT) can be viewed as an advanced prompting method ICL. It obtains reasoning capabilities for a given task by including examples with multiple intermediate steps before the query. With these additional steps, CoT can demonstrate stronger reasoning abilities than ICL across various tasks, such as arithmetic reasoning, symbolic reasoning, and commonsense reasoning. However, due to the more complex implementation and more

input information involved, there is limited theoretical understanding of CoT. Existing works concentrate on the expressive power of CoT in handling reasoning tasks, while no works theoretically answer why a Transformer can acquire generalization-guaranteed CoT ability on multi-step reasoning tasks by training from data with gradient-based methods.

By formulating reasoning tasks as a transition between patterns, our work provides the first theoretical analysis of the training dynamics of nonlinear Transformers to achieve CoT ability. We prove that the trained model has guaranteed CoT generalization performance on new tasks with distribution shifts from the training tasks, even when noisy and erroneous context examples exist in the testing prompt. We theoretically quantify the required number of training samples and iterations to learn a desirable model and the required number of context examples for successful CoT reasoning. The training dynamics analysis leads to a theoretical understanding of the CoT mechanism, i.e., transformers implement CoT by attending to the most similar examples at each inference step. Moreover, we provide a theoretical explanation for why CoT outperforms ICL in some cases. Specifically, a successful ICL needs an additional condition that the fraction of correct input-label examples in the testing prompt must be dominant, while the success of CoT does not depend on this condition.

# CHAPTER 2

## A THEORETICAL UNDERSTANDING OF SHALLOW VISION TRANSFORMERS: LEARNING, GENERALIZATION, AND SAMPLE COMPLEXITY

### 2.1 Introduction

As the backbone of Transformers [1], the self-attention mechanism [36] computes the feature representation by globally modeling long-range interactions within the input. Transformers have demonstrated tremendous empirical success in numerous areas, including natural language processing [37], [38], recommendation system [39], [40], [41], and reinforcement learning [42], [43], [44]. Starting from the advent of Vision Transformer (ViT) [2], Transformer-based models [45], [46], [47], [48] gradually replace convolutional neural network (CNN) architectures and become prevalent in vision tasks. Various techniques have been developed to train ViT efficiently. Among them, token sparsification [49], [50], [51], [52], [53] removes redundant tokens (image patches) of data to improve the computational complexity while maintaining a comparable learning performance. For example, [51], [52] prune tokens following criteria designed based on the magnitude of the attention map. Despite the remarkable empirical success, one fundamental question about training Transformers is still vastly open, which is

*Under what conditions does a Transformer achieve satisfactory generalization?*

Some recent works analyze Transformers theoretically from the perspective of proved Lipschitz constant of self-attention [54], [55], properties of the neural tangent kernel [56], [57] and expressive power and Turing-completeness [8], [9], [10], [11], [12], [13], [14], [15], [16] with statistical guarantees [17], [18]. [14] showed a model complexity for the function approximation of the self-attention module. [15] provided sufficient and necessary conditions for multi-head self-attention structures to simulate convolution layers. None of these works, however, characterize the generalization performance of the learned model theoretically. Only [12] theoretically proved that a single self-attention head can represent a sparse function of the

---

Portions of this chapter have previously appeared as: H. Li, M. Wang, S. Liu, and P.-Y. Chen, “A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity,” in *Proc. Int. Conf. Learn. Represent.*, May 2023, Paper 3368.

input with a sample complexity for a generalization gap between the training loss and the test loss, but no discussion is provided regarding what algorithm to train the Transformer to achieve a desirable loss.

**Contributions:** To the best of our knowledge, this chapter provides the first learning and generalization analysis of training a basic shallow Vision Transformer using stochastic gradient descent (SGD). This chapter focuses on a binary classification problem on structured data, where tokens with discriminative patterns determine the label from a majority vote, while tokens with non-discriminative patterns do not affect the labels. We train a ViT containing a self-attention layer followed by a two-layer perceptron using SGD from a proper initial model. This chapter explicitly characterizes the required number of training samples to achieve a desirable generalization performance, referred to as the sample complexity. Our sample complexity bound is positively correlated with the inverse of the fraction of label-relevant tokens, the token noise level, and the error from the initial model, indicating a better generalization performance on data with fewer label-irrelevant patterns and less noise from a better initial model. The highlights of our technical contributions include:

**First, this chapter proposes a new analytical framework to tackle the non-convex optimization and generalization for shallow ViTs.** Due to the more involved non-convex interactions of learning parameters and diverse activation functions across layers, the ViT model, i.e., a three-layer neural network with one self-attention layer, considered in this chapter is more complicated to analyze than three-layer CNNs considered in [58], [59], the most complicated neural network model that has been analyzed so far for across-layer nonconvex interactions. We consider a structured data model with relaxed assumptions from existing models and establish a new analytical framework to overcome the new technical challenges to handle ViTs.

**Second, this chapter theoretically depicts the evolution of the attention map during the training and characterizes how “attention” is paid to different tokens during the training.** Specifically, we show that under the structured data model, the learning parameters of the self-attention module grow in the direction that projects the data to the label-relevant patterns, resulting in an increasingly sparse attention map. This insight provides a theoretical justification of the magnitude-based token pruning methods such as [51], [52] for efficient learning.

**Third, we provide a theoretical explanation for the improved generalization**

**using token sparsification.** We quantitatively show that if a token sparsification method can remove class-irrelevant and/or highly noisy tokens, then the sample complexity is reduced while achieving the same testing accuracy. Moreover, token sparsification can also remove spurious correlations to improve the testing accuracy [60], [61]. This insight provides a guideline in designing token sparsification and few-shot learning methods for Transformer [62], [63].

### 2.1.1 Background and Related Works

**Efficient ViT learning.** To alleviate the memory and computation burden in training [2], [45], [64], various acceleration techniques have been developed other than token sparsification. [65] identifies the importance of different dimensions in each layer of ViTs and then executes model pruning. [66], [67], [68] quantize weights and inputs to compress the learning model. [69] studies automated progressive learning that automatically increases the model capacity on-the-fly. Moreover, modifications of attention modules, such as the network architecture based on local attention [47], [48], [70], can simplify the computation of global attention for acceleration.

**Theoretical analysis of learning and generalization of neural networks.** One line of research [71], [72], [73], [74], [75], [76] analyzes the generalization performance when the number of neurons is smaller than the number of training samples. The neural-tangent-kernel (NTK) analysis [77], [58], [78], [79], [80], [81], [82], [83], [22] considers strongly overparameterized networks and eliminates the nonconvex interactions across layers by linearizing the neural network around the initialization. The generalization performance is independent of the feature distribution and cannot explain the advantages of self-attention modules.

**Neural network learning on structured data.** [84] provide the generalization analysis of a fully-connected neural network when the data comes from separated distributions. [85], [86], [87], [88], [89] study fully connected networks and convolutional neural networks assuming that data contains discriminative patterns and background patterns. [90] illustrates the robustness of adversarial training by introducing the feature purification mechanism, in which neural networks with non-linear activation functions can memorize the data-dependent features. [91] extends this framework to the area of self-supervised contrastive learning. All these works consider one-hidden-layer neural networks without self-attention.

**Notations:** Vectors are in bold lowercase, and matrices and tensors are in bold uppercase. Scalars are in normal fonts. Sets are in calligraphy font. For instance,  $\mathbf{Z}$  is a matrix, and  $\mathbf{z}$  is a vector.  $z_i$  denotes the  $i$ -th entry of  $\mathbf{z}$ , and  $Z_{i,j}$  denotes the  $(i,j)$ -th entry of  $\mathbf{Z}$ .  $[K]$  ( $K > 0$ ) denotes the set including integers from 1 to  $K$ . 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.

## 2.2 Problem Formulation and the Learning Algorithm

We study a binary classification problem<sup>1</sup> following the common setup in [2], [45], [46]. Given  $N$  training samples  $\{(\mathbf{X}^n, y^n)\}_{n=1}^N$  generated from an unknown distribution  $\mathcal{D}$  and a fair initial model, the goal is to find an improved model that maps  $\mathbf{X}$  to  $y$  for any  $(\mathbf{X}, y) \sim \mathcal{D}$ . Here each data point contains  $L$  tokens  $\mathbf{x}_1^n, \mathbf{x}_2^n, \dots, \mathbf{x}_L^n$ , i.e.,  $\mathbf{X}^n = [\mathbf{x}_1^n, \dots, \mathbf{x}_L^n] \in \mathbb{R}^{d \times L}$ , where each token is  $d$ -dimensional and unit-norm.  $y^n \in \{+1, -1\}$  is a scalar. A token can be an image patch [2]. We consider a general setup that also applies to token sparsification, where some tokens are set to zero to reduce the computational time. Let  $\mathcal{S}^n \subseteq [L]$  denote the set of indices of remaining tokens in  $\mathbf{X}^n$  after sparsification. Then  $|\mathcal{S}^n| \leq L$ , and  $\mathcal{S}^n = [L]$  without token sparsification.

Learning is performed over a basic shallow Vision Transformer, a neural network with a single-head self-attention layer and a two-layer fully connected network, as shown in (2.1). This is a simplified model of practical Vision Transformers [2] to avoid unnecessary complications in analyzing the most critical component of ViTs, the self-attention.

$$F(\mathbf{X}^n) = \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \mathbf{a}_{(l)}^\top \text{Relu}(\mathbf{W}_O \mathbf{W}_V \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n)), \quad (2.1)$$

where the queue weights  $\mathbf{W}_Q$  in  $\mathbb{R}^{m_b \times d}$ , the key weights  $\mathbf{W}_K$  in  $\mathbb{R}^{m_b \times d}$ , and the value weights  $\mathbf{W}_V$  in  $\mathbb{R}^{m_a \times d}$  in the attention unit are multiplied with  $\mathbf{X}^n$  to obtain the queue vector  $\mathbf{W}_Q \mathbf{X}^n$ , the key vector  $\mathbf{W}_K \mathbf{X}^n$ , and the value vector  $\mathbf{W}_V \mathbf{X}^n$ , respectively [1].  $\mathbf{W}_O$  is in  $\mathbb{R}^{m \times m_a}$  and  $\mathbf{A} = (\mathbf{a}_{(1)}, \mathbf{a}_{(2)}, \dots, \mathbf{a}_L)$  where  $\mathbf{a}_{(l)} \in \mathbb{R}^m$ ,  $l \in [L]$  are the hidden-layer and output-layer weights of the two-layer perceptron, respectively.  $m$  is the number of neurons in the hidden layer.  $\text{Relu} : \mathbb{R}^m \rightarrow \mathbb{R}^m$  where  $\text{Relu}(\mathbf{x}) = \max\{\mathbf{x}, 0\}$ .  $\text{softmax} : \mathbb{R}^L \rightarrow \mathbb{R}^L$  where  $\text{softmax}(\mathbf{x}) = (e^{x_1}, e^{x_2}, \dots, e^{x_L}) / \sum_{i=1}^L e^{x_i}$ . Let  $\psi = (\mathbf{A}, \mathbf{W}_O, \mathbf{W}_V, \mathbf{W}_K, \mathbf{W}_Q)$  denote the set

---

<sup>1</sup>Extension to multi-classification is briefly discussed in Section A.6.1.

of parameters to train. The training problem minimizes the empirical risk  $f_N(\psi)$ ,

$$\min_{\psi} : f_N(\psi) = \frac{1}{N} \sum_{n=1}^N \ell(\mathbf{X}^n, y^n; \psi), \quad (2.2)$$

where  $\ell(\mathbf{X}^n, y^n; \psi)$  is the Hinge loss function, i.e.,

$$\ell(\mathbf{X}^n, y^n; \psi) = \max\{1 - y^n \cdot F(\mathbf{X}^n), 0\}. \quad (2.3)$$

The generalization performance of a learned model  $\psi$  is evaluated by the population risk  $f(\psi)$ , where

$$f(\psi) = f(\mathbf{A}, \mathbf{W}_O, \mathbf{W}_V, \mathbf{W}_K, \mathbf{W}_Q) = \mathbb{E}_{(\mathbf{X}, y) \sim \mathcal{D}}[\max\{1 - y \cdot F(\mathbf{X}), 0\}]. \quad (2.4)$$

The training problem (2.2) is solved via a mini-batch stochastic gradient descent (SGD), as summarized in Algorithm 3. At iteration  $t$ ,  $t = 0, 1, 2, \dots, T - 1$ , the gradient is computed using a mini-batch  $\mathcal{B}_t$  with  $|\mathcal{B}_t| = B$ . The step size is  $\eta$ .

Similar to [2], [45], [46],  $\mathbf{W}_V^{(0)}$ ,  $\mathbf{W}_Q^{(0)}$ , and  $\mathbf{W}_K^{(0)}$  come from an initial model. Every entry of  $\mathbf{W}_O$  is generated from  $\mathcal{N}(0, \xi^2)$ . Every entry of  $\mathbf{a}_l^{(0)}$  is sampled from  $\{+\frac{1}{\sqrt{m}}, -\frac{1}{\sqrt{m}}\}$  with equal probability.  $\mathbf{A}$  does not update during the training<sup>2</sup>.

## 2.3 Theoretical Results

### 2.3.1 Main Theoretical Insights

Before formally introducing our data model and main theory, we first summarize the major insights. We consider a data model where tokens are noisy versions of *label-relevant* patterns that determine the data label and *label-irrelevant* patterns that do not affect the label.  $\alpha_*$  is the fraction of label-relevant tokens.  $\sigma$  represents the initial model error, and  $\tau$  characterizes the token noise level.

**(P1). A Convergence and sample complexity analysis of SGD to achieve zero generalization error.** We prove SGD with a proper initialization converges to a model

---

<sup>2</sup>It is common to fix the output layer weights as the random initialization in the theoretical analysis of neural networks, including NTK [58], [79], model recovery [71], and feature learning [87], [90] type of approaches. The optimization problem here of  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$ ,  $\mathbf{W}_V$ , and  $\mathbf{W}_O$  with non-linear activations is still highly non-convex and challenging.

with zero generalization error. The required number of iterations is proportional to  $1/\alpha_*$  and  $1/(\Theta(1) - \sigma - \tau)$ . Our sample complexity bound is linear in  $\alpha_*^{-2}$  and  $(\Theta(1) - \sigma - \tau)^{-2}$ . Therefore, the learning performance is improved, in the sense of a faster convergence and fewer training samples to achieve a desirable generalization, with a larger fraction of label-relevant patterns, a better initial model, and less token noise.

**(P2). A theoretical characterization of increased sparsity of the self-attention module during training.** We prove that the attention weights, which are softmax values of each token in the self-attention module, become increasingly sparse during the training, with non-zero weights concentrated at label-relevant tokens. This formally justifies the general intuition that the attention layer makes the neural network focus on the most important part of data.

**(P3). A theoretical guideline of designing token sparsification methods to reduce sample complexity.** Our sample complexity bound indicates that the required number of samples to achieve zero generalization can be reduced if a token sparsification method removes some label-irrelevant tokens (reducing  $\alpha_*$ ), or tokens with large noise (reducing  $\sigma$ ), or both. This insight provides a guideline to design proper token sparsification methods.

**(P4). A new theoretical framework to analyze the nonconvex interactions in shallow ViTs.** This chapter develops a new framework to analyze ViTs based on a more general data model than existing works like [88, 87, 91]. Compared with the nonconvex interactions in three-layer feedforward neural networks, analyzing ViTs has technical challenges that the softmax activation is highly non-linear, and the gradient computation on token correlations is complicated. We develop new tools to handle this problem by exploiting structures in the data and proving that SGD iterations increase the magnitude of label-relevant tokens only rather than label-irrelevant tokens. This theoretical framework is of independent interest and can be potentially applied to analyze different variants of Transformers and attention mechanisms.

### 2.3.2 Data Model

There are  $M$  ( $2 < M < m_a, m_b$ ) distinct patterns  $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_M\}$  in  $\mathbb{R}^d$ , where  $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$  are *discriminative patterns* that determine the binary labels, and the remaining  $M - 2$  patterns  $\boldsymbol{\mu}_3, \boldsymbol{\mu}_4, \dots, \boldsymbol{\mu}_M$  are *non-discriminative patterns* that do not affect the labels. Let

$\kappa = \min_{1 \leq i \neq j \leq M} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| > 0$  denote the minimum distance between patterns. Each token  $\mathbf{x}_l^n$  of  $\mathbf{X}^n$  is a noisy version of one of the patterns, i.e.,

$$\min_{j \in [M]} \|\mathbf{x}_l^n - \boldsymbol{\mu}_j\| \leq \tau, \quad (2.5)$$

and the noise level  $\tau < \kappa/4$ . We take  $\kappa - 4\tau$  as  $\Theta(1)$  for the simplicity of presentation.

The label  $y^n$  is determined by the tokens that correspond to discriminative patterns through a majority vote. If the number of tokens that are noisy versions of  $\boldsymbol{\mu}_1$  is larger than the number of tokens that correspond to  $\boldsymbol{\mu}_2$  in  $\mathbf{X}^n$ , then  $y^n = 1$ . In this case that the label  $y^n = 1$ , the tokens that are noisy  $\boldsymbol{\mu}_1$  are referred to as *label-relevant* tokens, and the tokens that are noisy  $\boldsymbol{\mu}_2$  are referred to as *confusion* tokens. Similarly, if there are more tokens that are noisy  $\boldsymbol{\mu}_2$  than those that are noisy  $\boldsymbol{\mu}_1$ , the former are label-relevant tokens, the latter are confusion tokens, and  $y^n = -1$ . All other tokens that are not label-relevant are called label-irrelevant tokens.

Let  $\alpha_*$  and  $\alpha_\#$  as the average fraction of the label-relevant and the confusion tokens over the distribution  $\mathcal{D}$ , respectively. We consider a balanced dataset. Let  $\mathcal{D}_+ = \{(\mathbf{X}^n, y^n) | y^n = +1, n \in [N]\}$  and  $\mathcal{D}_- = \{(\mathbf{X}^n, y^n) | y^n = -1, n \in [N]\}$  denote the sets of positive and negative labels, respectively. Then  $|\mathcal{D}_+| - |\mathcal{D}_-| = O(\sqrt{N})$ .

Our model is motivated by and generalized from those used in the state-of-art analysis of neural networks on structured data [84, 88, 87]. All the existing models require that only one discriminative pattern exists in each sample, i.e., either  $\boldsymbol{\mu}_1$  or  $\boldsymbol{\mu}_2$ , but not both, while our model allows both patterns to appear in the same sample.

### 2.3.3 Formal Theoretical Results

Before presenting our main theory below, we first characterize the behavior of the initial model through Assumption 2.3.1. Some important notations are summarized in Table 2.1.

**Table 2.1:** Some important notations.

$\sigma$	Initialization error for value vectors	$\delta$	Initialization error for query and key vectors
$\kappa$	Minimum of $\ \boldsymbol{\mu}_i - \boldsymbol{\mu}_j\ $ for any $i, j \in [M], i \neq j$ .	$\tau$	Token noise level
$M$	Total number of patterns	$m$	The number of neurons in $\mathbf{W}_O$
$\alpha_*$	Average fraction of label-relevant tokens	$\alpha_\#$	Average fraction of confusion tokens

**Assumption 2.3.1.** Assume  $\max(\|\mathbf{W}_V^{(0)}\|, \|\mathbf{W}_K^{(0)}\|, \|\mathbf{W}_Q^{(0)}\|) \leq 1$  without loss of generality. There exist three (not necessarily different) sets of orthonormal bases  $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M\}$ ,  $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_M\}$ , and  $\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_M\}$ , where  $\mathbf{p}_l \in \mathbb{R}^{m_a}$ ,  $\mathbf{q}_l, \mathbf{r}_l \in \mathbb{R}^{m_b}$ ,  $\forall l \in [M]$ ,  $\mathbf{q}_1 = \mathbf{r}_1$ , and  $\mathbf{q}_2 = \mathbf{r}_2$ <sup>3</sup> such that

$$\|\mathbf{W}_V^{(0)} \boldsymbol{\mu}_j - \mathbf{p}_j\| \leq \sigma, \quad \|\mathbf{W}_K^{(0)} \boldsymbol{\mu}_j - \mathbf{q}_j\| \leq \delta, \quad \text{and } \|\mathbf{W}_Q^{(0)} \boldsymbol{\mu}_j - \mathbf{r}_j\| \leq \delta. \quad (2.6)$$

hold for some  $\sigma = O(1/M)$  and  $\delta < 1/2$ .

Assumption 2.3.1 characterizes the distance of query, key, and value vectors of patterns  $\{\boldsymbol{\mu}_j\}_{j=1}^M$  to orthonormal vectors. The requirement on  $\delta$  is minor because  $\delta$  can be in the same order as  $\|\boldsymbol{\mu}_j\|$ .

*Theorem 1* (Generalization of ViT). Suppose Assumption 2.3.1 holds;  $\tau \leq \min(\sigma, \delta)$ ; a sufficiently large model with

$$m \gtrsim M^2 \log N, \quad (2.7)$$

the average fraction of label-relevant patterns satisfies

$$\alpha_* \geq \frac{\alpha_\#}{e^{-(\delta+\tau)}(1 - (\sigma + \tau))}, \quad (2.8)$$

, and the mini-batch size and the number of sampled tokens of each data  $\mathbf{X}^n$ ,  $n \in [N]$  satisfy

$$B \geq \Omega(1), \quad |\mathcal{S}^n| \geq \Omega(1) \quad (2.9)$$

Then, after  $T$  number of iterations such that

$$T = \Theta(\eta^{-3/5} \alpha_*^{-1}) \quad (2.10)$$

, as long as the number of training samples  $N$  satisfies

$$N \geq \Omega\left(\frac{1}{(\alpha_* - c'(1 - \zeta) - c''(\sigma + \tau))^2}\right) \quad (2.11)$$

for some constant  $c', c'' > 0$ , and  $\zeta \gtrsim 1 - \eta^{10}$ , with a probability of at least 0.99, the returned

---

<sup>3</sup>The condition  $\mathbf{q}_1 = \mathbf{r}_1$  and  $\mathbf{q}_2 = \mathbf{r}_2$  is to eliminate the trivial case that the initial attention value is very small. This condition can be relaxed but we keep this form to simplify the representation.

model achieves zero generalization error as

$$f(\mathbf{A}^{(0)}, \mathbf{W}_O^{(T)}, \mathbf{W}_V^{(T)}, \mathbf{W}_K^{(T)}, \mathbf{W}_Q^{(T)}) = 0 \quad (2.12)$$

Theorem 1 characterizes under what condition of the data the neural network with self-attention in (2.1) trained with Algorithm 3 can achieve zero generalization error. To show that the self-attention layer can improve the generalization performance by reducing the required sample complexity to achieve zero generalization error, we also quantify the sample complexity when there is no self-attention layer in the following proposition.

*Proposition 1* (Generalization without self-attention). Suppose assumptions in Theorem 1 hold. When there is no self-attention layer, i.e.,  $\mathbf{W}_K$  and  $\mathbf{W}_Q$  are not updated during the training, if  $N$  satisfies

$$N \geq \Omega\left(\frac{1}{(\alpha_*(\alpha_* - \sigma - \tau))^2}\right) \quad (2.13)$$

then after  $T$  iterations with  $T$  in (2.10), the returned model achieves zero generalization error as

$$f(\mathbf{A}^{(0)}, \mathbf{W}_O^{(T)}, \mathbf{W}_V^{(T)}, \mathbf{W}_K^{(0)}, \mathbf{W}_Q^{(0)}) = 0 \quad (2.14)$$

*Remark 1.* (Advantage of the self-attention layer) Because  $m \gg m_a, m_b, d$ , the number of trainable parameter remains almost the same with or without updating the attention layer. Combining Theorem 1 and Proposition 1, we can see that with the additional self-attention layer, the sample complexity<sup>4</sup> is reduced by a factor  $1/\alpha_*^2$  with an approximately equal number of network parameters.

*Remark 2.* (Generalization improvement by token sparsification). (2.11) and (2.10) show that the sample complexity  $N$  and the required number of iterations  $T$  scale with  $1/\alpha_*^2$  and  $1/\alpha_*$ , respectively. Then, increasing  $\alpha_*$ , the fraction of label-relevant tokens, can reduce the sample complexity and speed up the convergence. Similarly,  $N$  and  $T$  scale with  $1/(\Theta(1) - \tau)^2$  and  $1/(\Theta(1) - \tau)$ . Then decreasing  $\tau$ , the noise in the tokens, can also improve the generalization. Note that a properly designed token sparsification method can both increase  $\alpha_*$  by removing label-irrelevant tokens and decrease  $\tau$  by removing noisy tokens,

---

<sup>4</sup>The sample complexity bounds in (2.11) and (2.13) are sufficient but not necessary. Thus, rigorously speaking, one can not compare two cases based on sufficient conditions only. In our analysis, however, these two bounds are derived with exactly the same technique with the only difference in handling the self-attention layer. Therefore, we believe it is fair to compare these two bounds to show the advantage of ViT.

thus improving the generalization performance.

*Remark 3.* (Impact of the initial model) The initial model  $\mathbf{W}_V^{(0)}$ ,  $\mathbf{W}_K^{(0)}$ ,  $\mathbf{W}_Q^{(0)}$  affects the learning performance through  $\sigma$  and  $\delta$ , both of which decrease as the initial model is improved. Then from (2.11) and (2.10), the sample complexity reduces and the convergence speeds up for a better initial model.

Proposition 2 shows that the attention weights are increasingly concentrated on label-relevant tokens during the training. Proposition 2 is a critical component in proving Theorem 1 and is of independent interest.

*Proposition 2.* The attention weights for each token become increasingly concentrated on those correlated with tokens of the label-relevant pattern during the training, i.e.,

$$\sum_{i \in \mathcal{S}_*^n} \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n)_i = \sum_{i \in \mathcal{S}_*^n} \frac{\exp(\mathbf{x}_i^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n)}{\sum_{r \in \mathcal{S}^n} \exp(\mathbf{x}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n)} \rightarrow 1 - \eta^C \quad (2.15)$$

at a sublinear rate of  $O(1/t)$  when  $t$  is large for a large  $C > 0$  and all  $l \in \mathcal{S}^n$  and  $n \in [N]$ .

Proposition 2 indicates that only label-relevant tokens are highlighted by the learned attention of ViTs, while other tokens have less weight. This provides a theoretical justification of magnitude-based token sparsification methods.  $\text{softmax}(\cdot)_i$  in (2.15) denotes the  $i$ -th entry of  $\text{softmax}(\cdot)$ .

**Proof idea sketch:** The main proof idea is to show that the SGD updates scale up value, query, and key vectors of discriminative patterns, while keeping the magnitude of the projections of non-discriminative patterns and the initial model error almost unchanged. To be more specific, by Lemma A.5.1, A.5.2, we can identify two groups of neurons in the hidden layer  $\mathbf{W}_O$ , where one group only learns the positive pattern, and the other group only learns the negative pattern. Claim 1 of Lemma A.3.1 states that during the SGD updates, the neuron weights in these two groups evolve in the direction of projected discriminative patterns,  $\mathbf{p}_1$  and  $\mathbf{p}_2$ , respectively. Meanwhile, Claim 2 of Lemma A.3.1 indicates that  $\mathbf{W}_K$  and  $\mathbf{W}_Q$  update in the direction of increasing the magnitude of the query and key vectors of label-relevant tokens from 1 to  $\Theta(\log T)$ , such that the attention weights correlated with label-relevant tokens gradually become dominant. Moreover, by Claim 3 of Lemma A.3.1, the update of  $\mathbf{W}_V$  increases the magnitude of the value vectors of label-relevant tokens, by adding partial neuron weights of  $\mathbf{W}_O$  that are aligned with the value vectors to these vectors.

Due to the above properties during the training, one can simplify the training process to show that the output of neural network (2.1) changes linearly in the iteration number  $t$ . From the above analysis, we can develop the sample complexity and the required number of iterations for the zero generalization guarantee.

**Technical novelty:** Our proof technique is inspired by the feature learning technique in analyzing fully connect networks and convolution neural networks [86], [88]. Our paper makes new technical contributions from the following aspects. First, we provide a new framework of studying the nonconvex interactions of multiple weight matrices in a shallow ViT while other feature learning works [86], [88], [87], [90], [91], [89] only study one trainable weight matrix in the hidden layer of a two-layer network. Second, we analyze the updates of the self-attention module with the softmax function during the training, while other papers either ignore this issue without exploring convergence analysis [12] or oversimplify the analysis by applying the neural-tangent-kernel (NTK) method that considers impractical over-parameterization and updates the weights only around initialization. [56], [57], [58], [79]. Third, we consider a more general data model, where discriminative patterns of multiple classes can exist in the same data sample, but the data models in [88], [87] require one discriminative pattern only in each sample.

## 2.4 Numerical Experiments

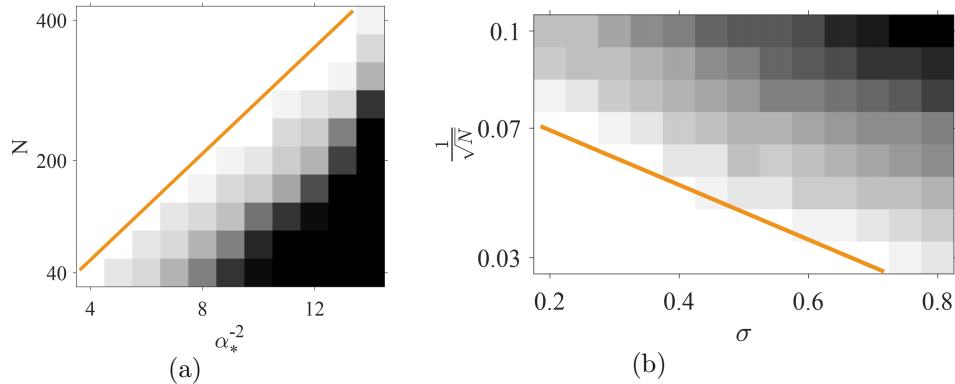
### 2.4.1 Experiments on Synthetic Datasets

We first verify the theoretical bounds in Theorem 1 on synthetic data. We set the dimension of data and attention embeddings to be  $d = m_a = m_b = 10$ . Let  $c_0 = 0.01$ . Let the total number of patterns  $M = 5$ , and  $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_M\}$  be a set of orthonormal bases. To satisfy Assumption 2.3.1, we generate every token that is a noisy version of  $\boldsymbol{\mu}_i$  from a Gaussian distribution  $\mathcal{N}(\boldsymbol{\mu}_i, c_0^2 \cdot \mathbf{I})$  with the mean  $\boldsymbol{\mu}_i$  and covariance  $c_0^2 \mathbf{I}$ , where  $\mathbf{I} \in \mathbb{R}^d$  is the identity matrix.  $\mathbf{W}_Q^{(0)} = \mathbf{W}_Q^{(0)} = \delta^2 \mathbf{I}/c_0^2$ ,  $\mathbf{W}_V^{(0)} = \sigma^2 \mathbf{U}/c_0^2$ , and each entry of  $\mathbf{W}_O^{(0)}$  follows  $\mathcal{N}(0, \xi^2)$ , where  $\mathbf{U}$  is an  $m_a \times m_a$  orthonormal matrix, and  $\xi = 0.01$ . The number of neurons  $m$  of  $\mathbf{W}_O$  is 1000. We set the ratio of different patterns the same among all the data for simplicity.

**Sample complexity and convergence rate:** We first study the impact of the fraction of the label-relevant patterns  $\alpha_*$  on the sample complexity. Let the number of tokens after sparsification be  $|\mathcal{S}^n| = 100$ , the initialization error  $\sigma = 0.1$ , and  $\delta = 0.2$ . The fraction of

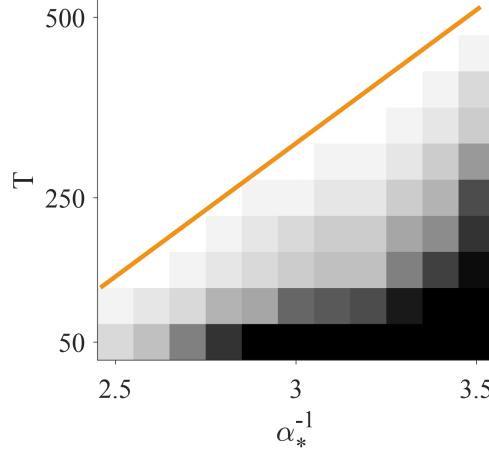
non-discriminative patterns is fixed to be 0.5. We implement 20 independent experiments with the same  $\alpha_*$  and  $N$  and record the Hinge loss values of the testing data. An experiment is successful if the testing loss is smaller than  $10^{-3}$ . Figure 2.1 (a) shows the success rate of these experiments. A black block means that all the trials fail. A white block means that they all succeed. The sample complexity is indeed almost linear in  $\alpha_*^{-2}$ , as predicted in 2.11. We next explore the impact on  $\sigma$ . Set  $\alpha_* = 0.3$  and  $\alpha_\# = 0.2$ . The number of tokens after sparsification is fixed at 50 for all the data. Figure 2.1 (b) shows that  $1/\sqrt{N}$  is linear in  $\Theta(1) - \sigma$ , matching our theoretical prediction in (2.11). The result on the noise level  $\tau$  is similar to Figure 2.1 (b), and we skip it here. In Figure 2.2, we verify the number of iterations  $T$  against  $\alpha_*^{-1}$  in (2.10) where we set  $\sigma = 0.1$  and  $\delta = 0.4$ .

**Advantage of self-attention:** To verify Proposition 1, we compare the performance on ViT in 2.1 and on the same network with  $\mathbf{W}_K$  and  $\mathbf{W}_Q$  fixed during the training, i.e., a three-layer CNN. Compared with ViT, the number of trainable parameters in CNN is reduced by only 1%. Figure 2.3 shows the sample complexity of CNN is almost linear in  $\alpha_*^{-4}$  as predicted in (2.13). Compared with Figure 2.2 (a), the sample complexity significantly increases for small  $\alpha_*$ , indicating a much worse generalization of CNN.



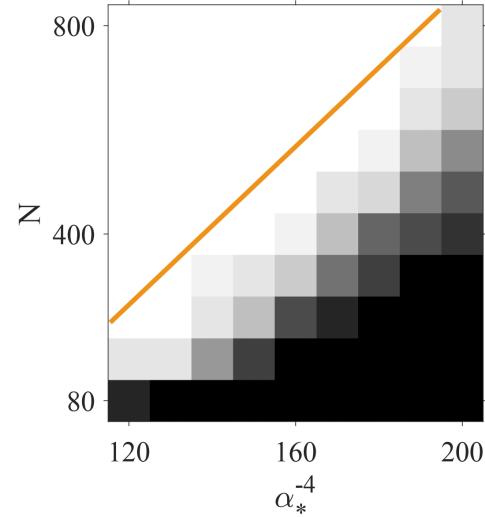
**Figure 2.1: The impact of (a)  $\alpha_*$  and (b)  $\sigma$  on sample complexity.**

**Attention map:** We then evaluate the evolution of the attention map during the training. Let  $|\mathcal{S}^n| = 50$  for all  $n \in [N]$ . The number of training samples is  $N = 200$ .  $\sigma = 0.1$ ,  $\delta = 0.2$ ,  $\alpha_* = 0.5$ ,  $\alpha_\# = 0.05$ . In Figure 2.4, the red line with asterisks shows that the sum of attention weights on label-relevant tokens, i.e., the left side of (2.15) averaged over all  $l$ , indeed increases to be close to 1 when the number of iterations increases. Correspondingly, the sum of attention weights on other tokens decreases to be close to 0, as shown in the blue



**Figure 2.2:** The number of iterations against  $\alpha_*^{-1}$ .

line with squares. This verifies Lemma 2 on a sparse attention map.



**Figure 2.3:** Comparison between ViT and CNN.

**Token sparsification:** We verify the improvement by token sparsification in Figure 2.5. The experiment is duplicated 20 times. The number of training samples  $N = 80$ . Let  $|\mathcal{S}^n| = 50$  for all  $n \in [N]$ . Set  $\sigma = 0.1$ ,  $\delta = 0.5$ ,  $\alpha_* = 0.6$ ,  $\alpha_\# = 0.05$ . If we apply random sampling over all tokens, the performance cannot be improved as shown in the red curve because  $\alpha_*$  and  $\sigma$  do not change. If we remove either label-irrelevant tokens or tokens with significant noise, the testing loss decreases, as indicated in the blue and black curves. This justifies our insight **P3** on token sparsification.

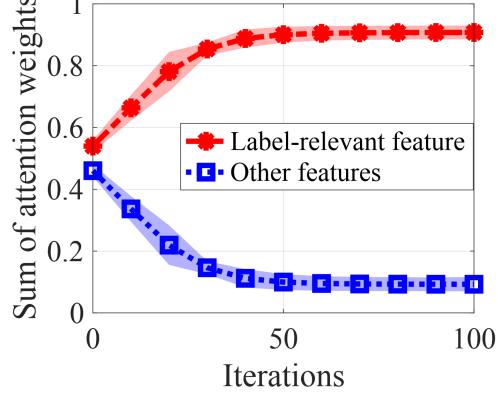


Figure 2.4: Concentration of attention weights.

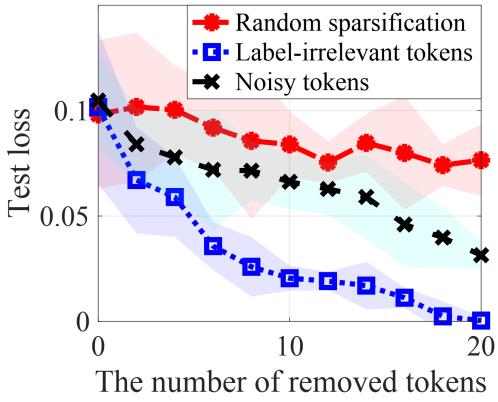


Figure 2.5: Impact of token sparsification on testing loss.

#### 2.4.2 Experiments on Image Classification Datasets

**Dataset:** To characterize the effect of label-relevant and label-irrelevant tokens on generalization, following the setup of image integration in [87], we adopt an image from CIFAR-10 dataset as the label-relevant image pattern and integrate it with a noisy background image from the IMAGENET Plants synset [87], [92], which plays the role of label-irrelevant feature. Specifically, we randomly cut out a region with size  $26 \times 26$  in the IMAGENET image and replace it with a resized CIFAR-10 image.

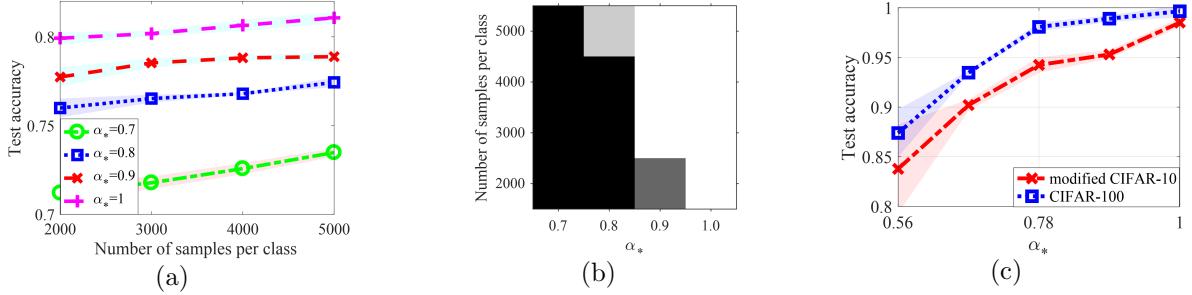
**Architecture:** Experiments are implemented on a deep ViT model. Following [2], the network architecture contains 5 blocks, where we have a 4-head self-attention layer and a one-layer perceptron with skip connections and Layer-Normalization in each block.

We first evaluate the impact on generalization of token sparsification that removes label-irrelevant patterns to increase  $\alpha_*$ . We consider a ten-classification problem where in both the training and testing datasets, the images used for integration are randomly selected

from CIFAR-10 and IMAGENET. The number of samples for training and testing is  $50K$  and  $10K$ , respectively. A pre-trained model from CIFAR-100 is used as the initial model with the output layer randomly initialized. Without token sparsification, the fraction of class-relevant tokens is  $\alpha_* \approx 0.66$ .  $\alpha_* = 1$  implies all background tokens are removed. Figure 2.6 (a) indicates that a larger  $\alpha_*$  by removing more label-irrelevant tokens leads to higher test accuracy. Moreover, the test performance improves with more training samples. These are consistent with our sample complexity analysis in (2.11). Figure 2.6 (b) presents the required sample complexity to learn a model with desirable test accuracy. We run 10 independent experiments for each pair of  $\alpha_*$  and  $N$ , and the experiment is considered a success if the learned model achieves a test accuracy of at least 77.5%.

We then evaluate the impact of token sparsification on removing spurious correlations [28], as well as the impact of the initial model. We consider a binary classification problem that differentiates “bird” and “airplane” images. To introduce spurious correlations in the training data, 90% of bird images in the training data are integrated into the IMAGENET plant background, while only 10% of airplane images have the plant background. The remaining training data are integrated into a clean background by zero padding. Therefore, the label “bird” is spuriously correlated with the class-irrelevant plant background. The testing data contain 50% birds and 50% airplanes, and each class has 50% plant background and 50% clean background. The numbers of training and testing samples are  $10K$  and  $2K$ , respectively. We initialize the ViT using two pre-trained models. The first one is pre-trained with CIFAR-100, which contains images of 100 classes not including birds and airplanes. The other initial model is trained with a modified CIFAR-10 with 500 images per class for a total of eight classes, excluding birds and airplanes. The pre-trained model on CIFAR-100 is a better initial model because it is trained on a more diverse dataset with more samples.

In Figure 2.6 (c), the token sparsification method removes the tokens of the added background, and the corresponding  $\alpha_*$  increases. Note that removing background in the training dataset also reduces the spurious correlations between birds and plants. Figure 2.6 (c) shows that from both initial models, the testing accuracy increases when more background tokens are removed. Moreover, a better initial model leads to a better testing performance. This is consistent with Remarks 2 and 3.



**Figure 2.6:** (a) Test accuracy when  $N$  and  $\alpha_*$  change. (b) Relationship of sample complexity against  $\alpha^*$ . (c) Test accuracy when token sparsification removes spurious correlations.

## 2.5 Conclusion

This paper provides a novel theoretical generalization analysis of shallow ViTs. Focusing on a data model with label-relevant and label-irrelevant tokens, this paper explicitly quantifies the sample complexity as a function of the fraction of label-relevant tokens and the token noise projected by the initial model. It proves that the learned attention map becomes increasingly sparse during the training, where the attention weights are concentrated on those of label-relevant tokens. Our theoretical results also offer a guideline on designing proper token sparsification methods to improve the test performance.

This paper considers a simplified but representative Transformer architecture to theoretically examine the role of self-attention layer as the first step. One future direction is to analyze more practical architectures such as those with skip connection, local attention layers, and Transformers in other areas. We see no ethical or immediate negative societal consequence of our work.

# CHAPTER 3

## WHAT IMPROVES THE GENERALIZATION OF GRAPH TRANSFORMER? A THEORETICAL DIVE INTO SELF-ATTENTION AND POSITIONAL ENCODING

### 3.1 Introduction

Graph Transformers [3], [4], [5] were developed for graph machine learning as a response to the impressive performance of Transformers demonstrated in various domains [1], [37], [38], [2], [40]. It is designed specifically to handle graph data by constructing positional embeddings that capture important graph information and using nodes as input tokens for the Transformer model. Empirical results have shown that Graph Transformers (GT) outperform classical graph neural networks (GNN), such as graph convolutional networks (GCN), in graph-level learning tasks such as molecular property prediction [93], [4], [94], image classification [21], [20], as well as node-level tasks like document analysis [95], [96], [97], [98], [19], semantic segmentation [20], [99], and social network analysis [100], [3], [101].

Despite the notable empirical advancements, some critical theoretical aspects of Graph Transformers remain much less explored. These include fundamental inquiries such as:

- *Under what conditions can a Graph Transformer achieve adequate generalization?*
- *What is the advantage of self-attention and positional encoding in graph learning?*

Some recent works [5], [19] theoretically study GTs by comparing their expressive power with other graph neural networks without self-attention. Meanwhile, other studies [4], [20], [21] explain the design of positional encoding (PE) in terms of graph topology and spectral theory. However, these analyses only establish the existence of a desired GT model, rather than its achievability through practical learning methods. Additionally, none of the existing works have theoretically examined the generalization of GTs, which is essential to explain their superior performance and guide the model and algorithm design.

To the best of our knowledge, this chapter presents the first learning and generalization analysis of a basic shallow GT trained using stochastic gradient descent (SGD). We focus on

---

Portions of this chapter have previously appeared as: H. Li, M. Wang, T. Ma, S. Liu, Z. Zhang, and P.-Y. Chen, “What improves the generalization of graph transformer? A theoretical dive into self-attention and positional encoding,” in *Proc. Int. Conf. Mach. Learn.*, Jul. 2024, pp. 28784–28829.

a semi-supervised binary node classification problem on structured graph data, where each node feature corresponds to either a discriminative or a non-discriminative pattern, and each ground truth node label is determined by the dominant discriminative pattern in the core neighborhood. We explicitly characterize the required number of training samples, i.e., the sample complexity, and the number of SGD iterations to achieve a desired generalization error. Our sample complexity bound indicates that graphs with a larger fraction of discriminative nodes tend to have superior generalization performance. Moreover, our analysis reveals that better generalization performance can be achieved by using graph sampling methods that prioritize class-relevant nodes. Our **technical contributions** are highlighted below:

**First, this chapter establishes a novel framework for the optimization and generalization analysis of shallow GTs.** We consider a shallow GT model with non-convex interactions across layers, including learnable self-attention and PE parameters, and Relu, softmax activation functions, while the state-of-the-art works on GNNs [102], [103], [89] exclude attention layers due to such difficulties. This chapter develops a novel and extendable feature-learning framework for analyzing the optimization and generalization of GTs.

**Secondly, this chapter theoretically characterizes the benefits of the self-attention layer of GTs.** Our analysis shows that self-attention evolves in a way that promotes class-relevant nodes during training. Thus, a GT trained produces a sparse attention map. Compared with GCNs without self-attention, GTs have a lower sample complexity and faster convergence rate for better generalization.

**Third, this chapter theoretically demonstrates that positional embedding improves the generalization by promoting the nodes in the core neighborhood.** Different from the state-of-the-art theoretical studies on Transformers that either ignore PE in analyzing generalization [6], [104], [103] or only characterize the expressive power of PE [20], [21], this chapter analyzes the generalization of a GT with a trainable relative positional embedding and proves that, with no prior knowledge, positional embedding trained with SGD can identify and promote the core neighborhood. This, in turn, leads to fewer training iterations and a smaller sample complexity.

## 3.2 Related Works

**Theoretical study on GTs.** Previous research has applied tools of topology theory, spectral theory, and expressive power to explain the success of GTs. For example, [5], [19]

illustrates that proper weights of the Transformer layer can represent basic operations of popular GNN models and capture more multi-hop information. [20] explains the necessity of PEs in distinguishing links that cannot be learned by 1-Weisfeiler-Leman test. [4], [21] depict that the PE can measure the physical interactions between nodes and reconstruct the raw graph as a bijection.

**Theoretical analyses of GNNs.** The works in [35], [105] characterize the expressive power of GNNs by studying the Weisfeiler-Leman test, inter-nodal distances, and graph biconnectivity. [106, 35, 107] analyze the stability of training GCNs. References [108], [109], [110], [74] characterize the generalization gap via concentration bound for transductive learning or dependent variables. In [22], [89], the authors explore the generalization of GNNs with node sampling.

**Learning neural networks on structured data.** [86], [88], [90], [89], [111] study one-hidden-layer fully-connected networks or convolutional neural networks given data containing discriminative and background patterns. This framework is extended to self-supervised learning and ensemble learning [91], [112], [113]. The learning and generalization of one-layer single-head Transformers are studied in [114], [6], [115], [116], [117], [23], [118] based on the spatial or pattern-space association between tokens.

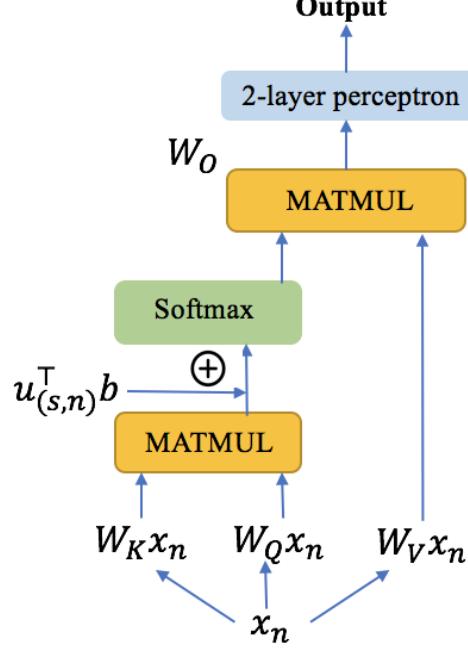
### 3.3 Problem Formulation and the Learning Algorithm

Let  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  denote an un-directed graph, where  $\mathcal{V}$  is the set of nodes with size  $|\mathcal{V}| = N$  and  $\mathcal{E}$  is the set of edges.  $\mathbf{X} \in \mathbb{R}^{d \times N}$  denotes the matrix of the features of  $N$  nodes, where the  $n$ -th column of  $\mathbf{X}$ , denoted by  $\mathbf{x}_n \in \mathbb{R}^d$ , represents the feature of node  $n$ . Assume  $\|\mathbf{x}_n\| = 1$  for all nodes without loss of generality. We study a binary node classification problem<sup>5</sup>. The label of node  $n$  is  $y_n \in \{+1, -1\}$ . Let  $\mathcal{L} \subset \mathcal{V}$  denote the set of labeled nodes. Given  $\mathbf{X}$  and labels in  $\mathcal{L}$ , the objective of semi-supervised learning for node classification is to predict the unknown labels in  $\mathcal{V} - \mathcal{L}$ . The learning process is implemented on a basic one-layer Graph Transformer in (3.1)<sup>6</sup>, which includes a single-head self-attention layer and a

---

<sup>5</sup>Extension to graph classification and multi-classification is briefly discussed in Appendix B.5.4 and B.5.5.

<sup>6</sup>Since the queries and keys are normalized, we remove the  $\sqrt{m_a}$  scaling in the softmax function as in [6], [104], [119], [115].



**Figure 3.1: Graph Transformers in (3.1).**

two-layer perceptron with a relative positional embedding.

$$\begin{aligned}
 F(\mathbf{x}_n) = & \mathbf{a}^\top \text{Relu}(\mathbf{W}_O \sum_{s \in \mathcal{T}^n} \mathbf{W}_V \mathbf{x}_s \\
 & \cdot \text{softmax}_n((\mathbf{W}_K \mathbf{x}_s)^\top \mathbf{W}_Q \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b})), \tag{3.1}
 \end{aligned}$$

where  $\mathbf{x}_n, \mathbf{x}_s \in \mathbb{R}^d$  and  $\mathcal{T}^n$  is the set of nodes for the aggregation computation of node  $n$ .  $\text{softmax}_n(g(s, n)) = \exp(g(s, n)) / \sum_{j \in \mathcal{T}^n} \exp(g(j, n))$  if we denote  $g(s, n) = \mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b}$ .  $\mathbf{W}_K \in \mathbb{R}^{m_a \times d}$ ,  $\mathbf{W}_Q \in \mathbb{R}^{m_a \times d}$ , and  $\mathbf{W}_V \in \mathbb{R}^{m_b \times d}$  are key, query, and value parameters to compute the self-attention representation by multiplying  $\mathbf{X}$ .  $\mathbf{W}_O \in \mathbb{R}^{m \times m_b}$  and  $\mathbf{a} \in \mathbb{R}^m$  are the hidden and output weights in the two-layer feedforward network. We define the one-hot distance vector  $\mathbf{u}_{(s,n)} \in \mathbb{R}^Z$ , where the non-zero index reflects the *truncated distance* between nodes  $s$  and  $n$ . It is an indicator of the shortest-path distance (SPD) between nodes. Then,

$$\mathbf{u}_{(s,n)} = \begin{cases} \mathbf{e}_i, & \text{if SPD of } s, n \text{ is } i - 1 \text{ and } i \leq Z, \\ \mathbf{e}_Z, & \text{if SPD of } s, n \text{ is } i - 1 \text{ and } i > Z, \end{cases} \tag{3.2}$$

where  $\mathbf{e}_i$  is the  $i$ -th standard basis in  $\mathbb{R}^Z$ . This architecture originates from [1] and is widely used in [4], [100], [98], [20] for node classification on graphs. The PE  $\mathbf{u}_{(s,n)}^\top \mathbf{b}$  is motivated by

[5], [20], [21], [120], [121], which is one of the most commonly used PEs in GTs.<sup>7</sup>

Denote  $\psi = (\mathbf{a}, \mathbf{W}_O, \mathbf{W}_V, \mathbf{W}_K, \mathbf{W}_Q, \mathbf{b})$  as the set of parameters to train. The semi-supervised learning problem solves the following empirical risk minimization problem  $f_N(\psi)$ ,

$$\begin{aligned} \min_{\psi} : f_N(\psi) &= \frac{1}{|\mathcal{L}|} \sum_{n \in \mathcal{L}} \ell(\mathbf{x}_n, y_n; \psi), \\ \ell(\mathbf{x}_n, y_n; \psi) &= \max\{1 - y_n \cdot F(\mathbf{x}_n), 0\}, \end{aligned} \quad (3.3)$$

where  $\ell(\mathbf{x}_n, y_n; \psi)$  is the Hinge loss function. Assume  $(\mathbf{x}_n, y_n)$  are identically distributed but *dependent* samples drawn from some unknown distribution  $\mathcal{D}$ . The sample dependence results from the dependence of node labels on neighboring node features. The test/generalization performance of a learned model  $\psi$  is evaluated by the population risk  $f(\psi)$ , where

$$\begin{aligned} f(\psi) &= f(\mathbf{a}, \mathbf{W}_O, \mathbf{W}_V, \mathbf{W}_K, \mathbf{W}_Q, \mathbf{b}) \\ &= \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} [\max\{1 - y \cdot F(\mathbf{x}), 0\}]. \end{aligned} \quad (3.4)$$

**Training Algorithm:** The training problem (3.3) is solved via a stochastic gradient descent (SGD), as summarized in Algorithm 4. At each iteration  $t$ , the gradient is computed using a batch  $\mathcal{B}_t$  with  $|\mathcal{B}_t| = B$  and step size  $\eta$  with all parameters in  $\psi$  except  $\mathbf{a}$ . At iteration  $t$ , we uniformly sample a subset  $\mathcal{S}^{n,t}$  of nodes from the whole graph for aggregation of each node  $n$ .

Following the framework “pre-training & fine-tuning” for node classification using [122], [95], [97], [123], we set  $\mathbf{W}_V^{(0)}$ ,  $\mathbf{W}_Q^{(0)}$ , and  $\mathbf{W}_K^{(0)}$  come from an initial model. Every entry of  $\mathbf{W}_O^{(0)}$  is generated from  $\mathcal{N}(0, \xi^2)$ . Every entry of  $\mathbf{a}^{(0)}$  is sampled from  $\{+1/\sqrt{m}, -1/\sqrt{m}\}$  with equal probability.  $\mathbf{b}^{(0)} = \mathbf{0}$ .  $\mathbf{a}$  is fixed during the training<sup>8</sup>.

---

<sup>7</sup>As the first work on the generalization of GT, we mainly study this PE for simplicity of the presentation. The analytical framework is extendable to GTs with other PEs. We briefly introduce the formulation and analysis of absolute PE, such as Laplacian vectors and node degree, in Appendix B.5.2.

<sup>8</sup>It is common to fix the output layer weights as the random initialization in the theoretical analysis of neural networks, including NTK [58], [79] and feature learning [87], [90], [6] type of approaches. The optimization problem of  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$ ,  $\mathbf{W}_V$ ,  $\mathbf{W}_O$ , and  $\mathbf{b}$  with non-linear activations is still highly non-convex and challenging.

## 3.4 Theoretical Results

### 3.4.1 Theoretical Insights

Before formally introducing our data model in Section 3.4.2 and the formal theoretical results in Section 3.4.3, we first summarize our key insights. We consider a data model where node features are *discriminative* patterns that directly determine the node labels and *non-discriminative* patterns that do not affect the labels.  $\gamma_d$  is the fraction of discriminative nodes, The node labels are determined by a majority vote of discriminative patterns in a so-called *core neighborhood*. A small  $\epsilon_S$  corresponds to a clear-cutting vote in sampled nodes in the core neighborhood.  $\sigma$  and  $\delta$  are the initial model error.  $\epsilon_0$  is the fraction of labels that are inconsistent with structural information.

**(P1). A new theoretical framework of a convergence and generalization analysis using SGD for GT.** This chapter develops a new framework to analyze GTs based on a more general graph data model than existing works like [89]. We show that with a proper initialization, the learning model converges with a desirable generalization error. The sample complexity bound is linear in  $\gamma_d^{-2}$ ,  $(\Theta(1) - \epsilon_S)^{-2}$ . The required number of iterations is proportional to  $(1 - 2\epsilon_0)^{-1/2}$  and  $(\Theta(1) - \delta)^{-1/2}$ . The result indicates that a larger fraction of discriminative nodes and a smaller confusion ratio improve the sample complexity. A smaller fraction of inconsistent labels and smaller embedding noises accelerate the convergence.

**(P2). Self-attention helps GTs perform better than Graph convolutional networks.** We theoretically illustrate that the attention weights, i.e., softmax values of each node in the self-attention module, become increasingly sparse during the training and are concentrated at discriminative nodes. GTs can then learn more distinguishable representations for different classes, outperforming GCNs.

**(P3) Positional embedding promotes the core neighborhood.** We prove that starting from zero initialization, the positional embedding eventually finds the core neighborhood and assigns nodes in the core neighborhood with higher weights, which improves the generalization.

### 3.4.2 Data Model Assumptions

Each node feature  $\mathbf{x}_n$  is one of  $M$  ( $2 \leq M < m_a, m_b$ ) distinct patterns  $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_M\}$  in  $\mathbb{R}^d$ , i.e.,  $\mathbf{x}_n = \boldsymbol{\mu}_j, \forall n \in \mathcal{V}$  and for a certain  $j \in [M_1]$ .  $\boldsymbol{\mu}_1$  and  $\boldsymbol{\mu}_2$  are two *discriminative patterns* that correspond to the label 1 and  $-1$ , respectively. All other patterns  $\boldsymbol{\mu}_3, \boldsymbol{\mu}_4, \dots, \boldsymbol{\mu}_M$

are referred to as *non-discriminative patterns* that do not determine the labels. Let  $\kappa = \min_{1 \leq i \neq j \leq M} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| > 0$  denote the minimum distance between different patterns. Denote the set of nodes that contain  $\boldsymbol{\mu}_l$  as  $\mathcal{D}_l$ ,  $l \in [M]$ , and  $\cup_{l=1}^M \mathcal{D}_l = \mathcal{V}$ . Let  $\gamma_d = |\mathcal{D}_1 \cup \mathcal{D}_2|/|\mathcal{V}| = \Theta(1)$  represent the fraction of nodes that contain discriminative patterns<sup>9</sup>. We assume the dataset is balanced, i.e., the gap between the numbers of positive and negative labels is at most  $O(\sqrt{N})$ .

If node  $n$  has the label  $y^n = 1$ , the nodes in  $\mathcal{D}_1$  are called *class-relevant* nodes for node  $n$ , and nodes in  $\mathcal{D}_2$  called *confusion* nodes for node  $n$ . Conversely, if  $y^n = -1$ ,  $\mathcal{D}_2$  and  $\mathcal{D}_1$  are class-relevant and confusion nodes for node  $n$ , respectively. We use notations  $\mathcal{D}_*^n$  and  $\mathcal{D}_\#^n$  for the class-relevant and confusion nodes for node  $n$  without specifying  $\mathcal{D}_1$  and  $\mathcal{D}_2$ . We define *distance- $z$  neighborhood* of node  $n$ , denoted by  $\mathcal{N}_z^n$ , as the set of nodes that are away from node  $n$  with distance  $z$ . The average winning margin of each node  $n$  and the *core* distance  $z_m$  are defined as follows.

**Definition 3.4.1.** The winning margin for each node  $n$  of distance- $z$  and the average winning margin for all the nodes of distance- $z$  are defined as

$$\Delta_n(z) = |\mathcal{D}_*^n \cap \mathcal{N}_z^n| - |\mathcal{D}_\#^n \cap \mathcal{N}_z^n|, \quad \bar{\Delta}(z) = \frac{1}{N} \sum_{n \in \mathcal{V}} \Delta_n(z), \quad (3.5)$$

for any  $z \in [Z-1]$ . The core distance is defined as

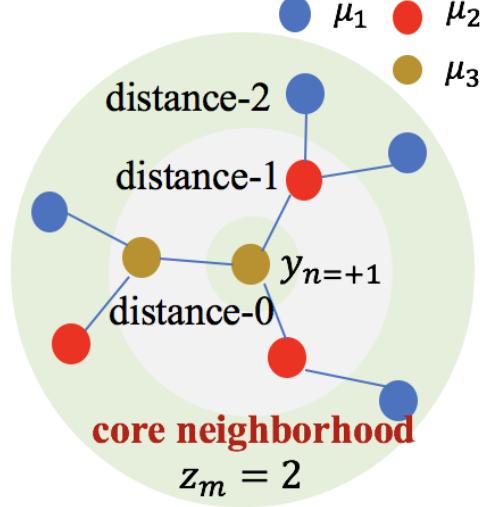
$$z_m = \arg \max_{z \in [Z-1]} \bar{\Delta}(z). \quad (3.6)$$

**Assumption 3.4.2.** There exists  $\mathcal{V}_d \subseteq \mathcal{V}$  with  $|\mathcal{V}_d|/|\mathcal{V}| \geq 1 - \epsilon_0$  ( $\epsilon_0 \in (0, 1)$ ) such that  $\Delta_n(z_m) > 0$  holds for all  $n \in \mathcal{V}_d$ .

Figure 3.2 provides an example of a winning margin. Assumption 3.4.2 indicates that the node label  $y^n$  for every node  $n \in \mathcal{V}_d$  is consistent with a majority voting of  $\boldsymbol{\mu}_1$  and  $\boldsymbol{\mu}_2$  patterns in the core neighborhood  $\mathcal{N}_{z_m}$ , i.e., if  $y^n = 1$  (or  $y^n = -1$ ), then there are more nodes that correspond to  $\boldsymbol{\mu}_1$  (or  $\boldsymbol{\mu}_2$ ) in  $\mathcal{N}_{z_m}^n$ . This assumption is verified in Table B.5 of Appendix B.1.1. We can deduce that  $\epsilon_0 < 0.05$  is a small value in three real-world datasets.

---

<sup>9</sup>The pattern of each node  $n \in \mathcal{V}$  follows a categorical distribution with probability  $(\nu_1, \nu_2, \dots, \nu_M)$ , where  $\nu_1 = \nu_2 = \gamma_d/2$  and  $\nu_3 + \nu_4 + \dots + \nu_M = 1 - \gamma_d$



**Figure 3.2: Example of the winning margin.** Node  $n$  has a non-discriminative feature  $\mu_3$  and label  $+1$ . Then  $\Delta_n(1) = -2$ , and  $\Delta_n(2) = 3$ .

We also assume  $|\mathcal{N}_{z_m}^n|$  is not too small to facilitate the sampling. We set  $|\mathcal{N}_{z_m}^n| \geq N/\text{poly}(Z)$  for all  $n$  to avoid a trivial size of the core neighborhood.

Assumption B.2.1 in Appendix A.2 requires the pre-trained model maps the query, key, and value embeddings to be close to orthogonal vectors with an error of  $\sigma < O(1/M)$  for queries and keys and  $\delta < 0.5$  for values. It is the same as Assumption 1 in [6]. Such assumptions on the orthogonality of embeddings or data are widely employed in state-of-the-art generalization analysis for Transformers [115, 104].<sup>10</sup>

### 3.4.3 Main Theoretical Results for Graph Transformers

We define *confusion ratio*  $\epsilon_S$  as the average fraction of confusion nodes in the distance- $z_m$  neighborhood over all iterations and all labeled nodes. Some notations are summarized in Table 3.1.

**Definition 3.4.3.** The confusion ratio  $\epsilon_S$  is

$$\epsilon_S = \mathbb{E}_{t \geq 0, n \in (\cup_{l=3}^M \mathcal{D}_l) \cap \mathcal{L}} \frac{|\mathcal{S}_\#^{n,t} \cap \mathcal{N}_{z_m}^n|}{|(\mathcal{S}_*^{n,t} \cup \mathcal{S}_\#^{n,t}) \cap \mathcal{N}_{z_m}^n|}, \quad (3.7)$$

where  $\mathcal{S}_*^{n,t}$  and  $\mathcal{S}_\#^{n,t}$  denote the sampled class-relevant and confusion nodes in  $\mathcal{T}^n$  for

<sup>10</sup>We conduct experiments to verify the existence of discriminative nodes and the core neighborhood with four real-world datasets in Appendix B.1.1. We also show Assumption 3.4.2 and B.2.1 are not strong by comparing existing works in Appendix B.5.1.

**Table 3.1: Some important notations.**

$\mathcal{V}$	The set of all the nodes	$\mathcal{D}_*^n, \mathcal{D}_{\#}^n$	Sets of class-relevant nodes and confusion nodes for node $n$
$\mathcal{L}$	The set of labeled nodes	$\mathcal{T}^n$	The set of nodes for aggregation for $n$
$\mathcal{D}_l$	The set of nodes of the pattern $\mu_l$	$\mathcal{S}_*^{n,t}, \mathcal{S}_{\#}^{n,t}$	Sampled class-relevant and confusion nodes out of $\mathcal{T}^n$ at iteration $t$
$\gamma_d$	The fraction of discriminative nodes	$\bar{\Delta}(z)$	Average winning margin of all nodes at the distance- $z$ neighborhood
$\mathcal{N}_z^n$	Distance- $z$ neighborhood of node $n$	$z_m$	The core distance that has the largest winning margin
$\epsilon_{\mathcal{S}}$	confusion ratio, the average fraction of confusion nodes in sampled nodes of distance- $z_m$ neighborhood		

node  $n$  in training iteration  $t$ , respectively.

We then introduce our major theoretical results.

**Theorem 3.4.4.** (*Generalization Guarantee of Graph Transformers*) *As long as for any  $\epsilon \in (0, 1)$ , the model with  $m \geq \Omega(M^2 \log N)$ , and the batch size  $B \geq \Omega(\epsilon^{-2} \log N)$  and the number of sampled nodes  $|\mathcal{S}^{n,t}|$  for each iteration  $t$  larger than  $\Omega(1)$ . Then, after  $T$  iterations such that*

$$T = \Theta(\eta^{-1/2}(1 - 2\epsilon_0)^{-1/2}(1 - \delta)^{-1/2}), \quad (3.8)$$

*as long as the number of known labels satisfies*

$$|\mathcal{L}| \geq \max\{\Omega\left(\frac{(1 + \delta_{z_m}^2) \cdot \log N}{(1 - 2\epsilon_{\mathcal{S}}(1 - \gamma_d) - \sigma)^2}\right), BT\}, \quad (3.9)$$

*where  $\delta_{z_m} = \max_{n \in \mathcal{V}} |\mathcal{N}_{z_m}^n|$  measures the maximum number of nodes in distance- $z_m$  neighborhood, for some  $\epsilon_{\mathcal{S}} \in (0, 1/2)$  and  $\epsilon_0 \in (0, 1/2)$ , then with a probability of at least 0.99, the returned model trained by Algorithm 4 achieves a desirable testing loss as*

$$f(\psi) \leq 2\epsilon_0 + \epsilon. \quad (3.10)$$

**Remark 1.** (Generalization improvement by good graph properties) Theorem 3.4.4 shows that given all required conditions and an  $\epsilon_0$  fraction of inconsistent labels in testing, the trained model can achieve a diminishing testing loss  $2\epsilon_0 + \epsilon$ . The first term in (3.9) dominates when  $\epsilon_0$  is not very close to<sup>11</sup>  $1/2$ , i.e., the fraction of inconsistent labels is small.

<sup>11</sup>The exact condition is when  $\epsilon_0 < 1/2 - \delta_{z_m}^{-4}\epsilon^{-4}/2$ .

Then the sample complexity in (3.9) scales with  $1/\gamma_d^2$ ,  $(1 - \epsilon_S)^{-2}$  and  $(\Theta(1) - \sigma)^{-2}$ . Hence, a larger fraction of nodes of discriminative patterns (a larger  $\gamma_d$ ), a smaller fraction of confusion patterns in the core neighborhood (a smaller  $\epsilon_S$ ), a smaller embedding noise (a smaller  $\sigma$ ) can reduce the sample complexity. The required number of iterations also reduces with a smaller fraction of inconsistent labels  $\epsilon_0$  and the embedding noise  $\sigma$ .

**Remark 2.** (Impact of graph sampling) A graph sampling method that can sample more class-relevant nodes in the distance- $z_m$  neighborhood can improve the learning by reducing  $\epsilon_S$ .

### 3.4.4 What Does Self-Attention Improve? A Comparison with GCN

We show that the attention weights become concentrated on class-relevant nodes in Lemma 3.4.5. It increases the distance between output vectors from different classes, which in turn improves the test accuracy. In contrast, Theorem 3.4.6 shows that without the self-attention layer, GCN requires more iterations and training samples.

**Lemma 3.4.5.** (*Sparse attention map*) *The attention weights for each node become increasingly concentrated on those correlated with class-relevant nodes during the training, i.e.,*

$$\begin{aligned} & \sum_{i \in \mathcal{S}_*^{n,t}} \text{softmax}_n(\mathbf{x}_i^\top \mathbf{W}_K^{(t)} \mathbf{W}_Q^{(t)} \mathbf{x}_n + \mathbf{u}_{(i,n)}^\top \mathbf{b}^{(t)}) \\ & \rightarrow \begin{cases} 1 - \eta^C, & n: \text{discriminative}, \\ 1 - \epsilon_S - \eta^C, & n: \text{non-discriminative}, \end{cases} \end{aligned} \quad (3.11)$$

at a sublinear rate of  $O(1/t)$  as  $t$  increases for a large  $C > 0$  and all  $n \in \mathcal{V}$ .

Lemma 3.4.5 indicates that the outputs of the self-attention layer for all nodes, which are weighted summations of value vectors, evolve in the direction of the class-relevant value features along the training. Then it promotes learning class-relevant features while ignoring other features. Lemma 3.4.5 is a generalization of Proposition 2 in [6], which considers a shallow ViT with one self-attention layer without positional embedding or graph structure. Here, we extend the analysis to node classification on graphs with PE.

Theorem 3.4.6 indicates that without the self-attention layer, the resulting GCN requires more training iterations and samples to achieve the desired generalization, even if

the core distance  $z_m$  is known, and the learning is performed on the core neighborhood only. Specifically,

**Theorem 3.4.6.** (*Generalization of GCN*) *When fixing  $\mathbf{W}_K = \mathbf{W}_Q = 0$  and  $\mathbf{b} = 0$  in (3.1), and all  $\mathcal{S}^{n,t}$  ( $n \in \mathcal{L}$ ) and  $\mathcal{T}^n$  ( $n \in \mathcal{V} - \mathcal{L}$ ) are subsets of  $\mathcal{N}_{z_m}^n$ , the resulting GCN [32, 124] learning on the core neighborhood  $\mathcal{N}_{z_m}^n$  can achieve a desirable generalization of  $2\epsilon_0 + \epsilon$  with the same condition in Theorem 3.4.4, but the number of iterations and the sample complexity should satisfy*

$$T = \Theta(\eta^{-1/2}(1 - 2\epsilon_0)^{-1/2}\gamma_d^{-2}(1 - \delta)^{-1/2}), \quad (3.12)$$

$$|\mathcal{L}| \geq \max\{\Omega((\gamma_d^2 - \sigma)^{-2}(1 + \delta_{z_m}^2)\log N), BT\}, \quad (3.13)$$

When  $m \gg m_a, m_b$ , i.e., the number of parameters is almost the same for GCN and GT, Theorem 3.4.6 shows that GCN requires  $\Theta(\gamma_d^{-2})$  times more training samples and iterations<sup>12</sup> to achieve desirable testing loss than those using GT in (3.9) and (3.8), respectively. This explains the advantage of using self-attention layers as in insight (P2).

### 3.4.5 How Does Positional Encoding Guide the Graph Learning Process?

In this section, we study how PE affects learning performance. Our insight is that the learnable parameter for the PE promotes the core neighborhood for classification and, thus, improves the sample complexity and required number of iterations for generalization. To see this, first, Lemma 3.4.7 shows that the largest entry in  $\mathbf{b}^{(T)}$  indeed corresponds to the core distance  $z_m$ . Therefore, PE “attracts the attention” of GT to the  $z_m$ -distance neighborhood. Then, Theorem 3.4.8 indicates that learning with the positional embedding has the same generalization performance as an artificial learning process when the core neighborhood  $\mathcal{N}_{z_m}^n$  is known, and the learning is performed on  $\mathcal{N}_{z_m}^n$  only.

**Lemma 3.4.7.** *Starting from  $\mathbf{b}^{(0)} = 0$ , if  $T$  satisfies (3.8), the returned model trained by Algorithm 4 satisfies*

$$b_{z_m}^{(T)} - b_z^{(T)} \geq \Omega(\gamma_d(\bar{\Delta}(z_m) - \bar{\Delta}(z))), \quad (3.14)$$

---

<sup>12</sup>All the sample complexity and iteration bounds in this chapter are obtained based on sufficient conditions for desirable generalization. Rigorously speaking, necessary conditions are also required to compare the generalization of different network architectures. However, necessary conditions are rarely considered in the literature due to technical challenges. Here, we still believe it is a fair comparison of sufficient conditions because we employ the same tools to analyze different neural network architectures.

Lemma 3.4.7 shows that  $b_{z_m}$  is the largest one among all  $1 \leq z \leq Z - 1$  because  $\bar{\Delta}(z_m)$  is the largest by (3.6). Because the softmax function employs  $e^{b_z}$  when computing the attention map, nodes at the  $z_m$ -distance neighborhood dominate the attention weights.

**Theorem 3.4.8.** *(The equivalent effect of the positional embedding)<sup>13</sup> when  $\mathbf{b} = 0$  in (3.1), and all  $\mathcal{S}^{n,t}$  ( $n \in \mathcal{L}$ ) and  $\mathcal{T}^n$  ( $n \in \mathcal{V} - \mathcal{L}$ ) are subsets of  $\mathcal{N}_{z_m}^n$ , a desirable generalization can be achieved when the sample complexity and the number of iterations satisfy (3.9) and (3.8) in Theorem 3.4.4.*

The learning process described in Theorem 3.4.8 is artificial because  $z_m$  is generally unknown. Theorem 3.4.8 shows that learning with position embedding has an equivalent generalization performance to learning from the core neighborhood  $\mathcal{N}_{z_m}^n$  only.

### 3.4.6 Proof Sketch

The main proof idea of Theorem 3.4.4 is to unveil a joint learning mechanism of GTs for our graph data model: (i) identifying discriminative features and the core neighborhood using PE and (ii) determining the labels of non-discriminative nodes through a majority vote in the core neighborhood by self-attention. Several lemmas are introduced to support the proof.

Specifically, by supportive Lemmas B.3.5 and B.3.6, we first characterize two groups of neurons that respectively activate the self-attention layer output of  $\mu_1$  and  $\mu_2$  nodes from initialization. Then, Lemma B.3.1 shows that the neurons of  $\mathbf{W}_O$  in these two groups grow along the two directions of the discriminative pattern embeddings. Lemma 3 indicates that the updates of  $\mathbf{W}_V$  consist of neuron weights from these two groups. Meanwhile, Lemma 2 states that  $\mathbf{W}_Q$  and  $\mathbf{W}_K$  evolve to promote the magnitude of query and key embeddings of discriminative nodes. Lemma B.3.3 depicts the training trajectory of the learning parameter of PE that emphasizes the core neighborhood. Different from the proof in [6, 104, 116, 119] that does not consider PE and graph structure, we make the proof of each lemma tractable by studying gradient growth per distance- $z$  neighborhood for each  $z$  rather than directly characterizing the gradient growth over the whole graph. Such a technique enables a dynamic tracking of per-parameter gradient updates. As a novel aspect, we prove Lemma B.3.3 by

---

<sup>13</sup>We discuss the application of Theorem 3.4.8 to analyze the generalization of one-layer GAT in Appendix B.5.3.

showing that its most significant gradient component is proportional to the average winning margin in the core neighborhood.

**Proof of Theorem 3.4.4** We can build the generalization guarantee in Theorem 3.4.4 from the above. First, Lemma 2 and B.3.3 collaborate to illustrate that attention weights correlated with class-relevant nodes become close to 1 when  $\eta t = \Theta(1)$ . Second, we compute the network output by Lemmas B.3.1 and 3. By enforcing the output to be either  $\geq 1$  or  $\leq -1$  to achieve  $\epsilon_0$  Hinge loss, we derive the sample complexity bound and the required number of iterations by concentration inequalities.

**The proof of Theorem 3.4.6 and 3.4.8** follow a similar idea as Theorem 3.4.4. When the self-attention layer weights are fixed at 0 in Theorem 3.4.6, since that  $\gamma_d = \Theta(1)$  and a given core neighborhood still ensure non-trivial attention weights correlated with class-relevant nodes along the training, the updates of  $\mathbf{W}_O$  and  $\mathbf{W}_V$  are order-wise the same as Lemmas B.3.1 and 3. Then, we can apply Lemmas B.3.1 and 3 to derive the required number of samples and iterations for desirable generalization. Likewise, given a known core neighborhood in Theorem 3.4.8, the remaining parameters follow the same order-wise update as Lemmas B.3.1, 2 and 3. Hence, Theorems 3.4.6 and 3.4.8 can be proved.

## 3.5 Numerical Experiments

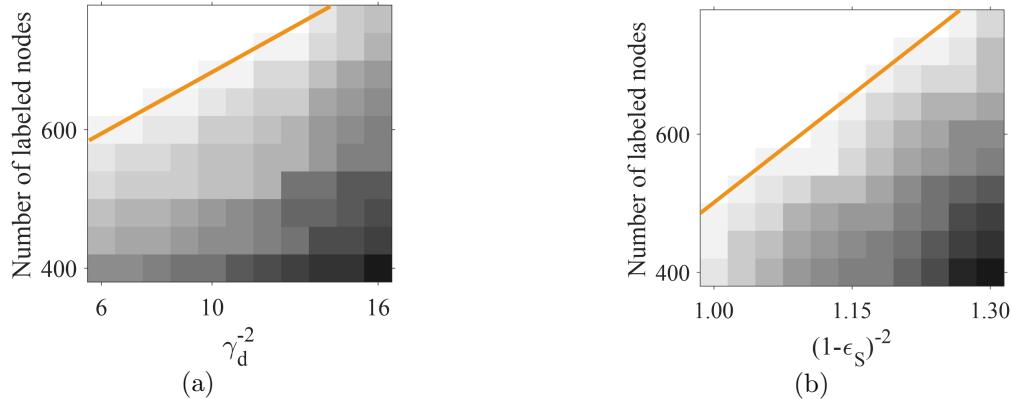
### 3.5.1 Experiments on Synthetic Data

**Graph data generation:** The graph contains 1000 nodes in total.  $M = 10$ ,  $\boldsymbol{\mu}_1$  to  $\boldsymbol{\mu}_M$  are selected as orthonormal vectors in  $\mathbb{R}^d$ , where  $d$  is 20. Node features that correspond to pattern  $\boldsymbol{\mu}_i$  are sampled from Gaussian distributions  $\mathcal{N}(\boldsymbol{\mu}_i, c_0^2 \cdot \mathbf{I})$ , where  $c_0 = 0.01$ , and  $\mathbf{I} \in \mathbb{R}^d$  is the identity matrix.  $\gamma_d/2$  fraction of nodes are selected as noisy versions of class-discriminative  $\boldsymbol{\mu}_1$  and  $\boldsymbol{\mu}_2$ , respectively. The remaining nodes are evenly distributed among other non-discriminative  $M - 2$  patterns.  $\gamma_d = 0.4$  unless otherwise specified. Our graph construction method is motivated by and extends from that in [89]. Every non-discriminative node is labeled with +1 or -1 with equal probability. If labeled +1, that non-discriminative node is randomly connected with  $120 \cdot (1 - \epsilon_S)$  nodes of  $\boldsymbol{\mu}_1$  and  $120 \cdot \epsilon_S$  of  $\boldsymbol{\mu}_2$  for some  $\epsilon_S$  in  $[0, 1/2]$ . If labeled -1, it is randomly connected with  $120 \cdot (1 - \epsilon_S)$  nodes of  $\boldsymbol{\mu}_2$  and  $120 \cdot \epsilon_S$  of  $\boldsymbol{\mu}_1$ . We also add edges among  $\boldsymbol{\mu}_1$  nodes themselves, and edges among  $\boldsymbol{\mu}_2$  nodes themselves to make each node degree at least 120. There is no edge between  $\boldsymbol{\mu}_1$  nodes and  $\boldsymbol{\mu}_2$  nodes. The ground-truth label for  $\boldsymbol{\mu}_1$  or  $\boldsymbol{\mu}_2$  nodes is +1 or -1, respectively.  $\epsilon_0 = 0$  if not otherwise

specified.

**Learner network and algorithm:** The learner network is a one-layer GT defined in equation 3.1. Set dimensions of embeddings to be  $m_a = m_b = 20$ . The number of neurons  $m$  of  $\mathbf{W}_O$  is 400.  $\delta = 0.2$ ,  $\sigma = 0.1$ , and  $\xi = 0.01$ .  $\mathbf{W}_Q^{(0)} = \mathbf{W}_Q^{(0)} = \delta^2 \mathbf{I} / c_0^2$ ,  $\mathbf{W}_V^{(0)} = \sigma^2 \mathbf{U} / c_0^2$ , where each entry of  $\mathbf{W}_O^{(0)}$  follows  $\mathcal{N}(0, \xi^2)$ .  $\mathbf{U}$  is an  $m_a \times m_a$  orthonormal matrix. The step size  $\eta = 0.01$ .  $\mathcal{S}^{n,t}$  contains node  $n$  and 60 uniformly sampled nodes from distance-1 and distance-2 neighborhood for each node  $n$  at iteration  $t$ .

**Sample complexity and convergence rate:** We first study the impact of the fraction  $\gamma_d$  of discriminative nodes on the sample complexity. Let  $\epsilon_S = 0.05$ . We implement 20 independent experiments with the same  $\gamma_d$  and  $|\mathcal{L}|$  while randomly generating graph structure, node features, and sampled labels. An experiment is successful if the Hinge testing loss is smaller than  $10^{-3}$ . A black block means all the trials fail, while a white block means they all succeed. Figure 3.3 (a) shows that the sample complexity is indeed almost linear in  $\gamma_d^{-2}$ , as indicated in 3.9. We next set  $\gamma_d = 0.4$  and vary  $\epsilon_S$ . Figure 3.3 (b) shows that the sample complexity is linear in  $(1 - \epsilon_S)^{-2}$ , which is consistent with our result in (3.9). We then change  $\epsilon_0$  and evaluate the prediction error when the number of training iterations changes, when  $\gamma_d = 0.4$ ,  $\epsilon_S = 0$ , and  $|\mathcal{L}| = 400$ . Figure 3.4 shows that a larger  $\epsilon_0$  requires more iterations to converge, and the convergent testing loss is around  $2\epsilon_0$ , which is consistent with (3.10).



**Figure 3.3: The impact of (a)  $\gamma_d$  and (b)  $\epsilon_S$  on the sample complexity of GT.**

**Attention map and comparison with GCN:** We then verify the sparsity of the attention map during the training. Let  $|\mathcal{L}| = 400$ ,  $\gamma_d = 0.2$ . In Figure 3.5, the blue circled line shows the summation of attention weights on class-relevant nodes averaged over all

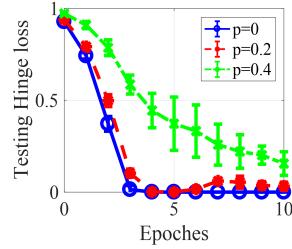


Figure 3.4: The test Hinge loss against the number of epochs for different  $\epsilon_0$ .

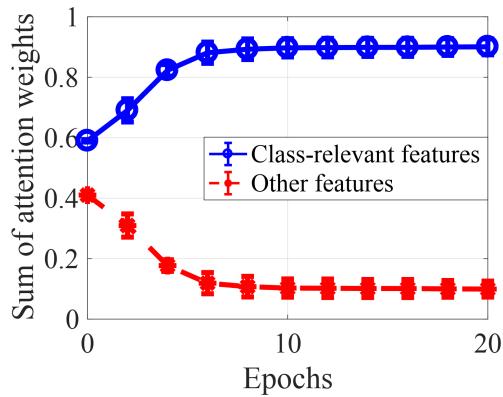


Figure 3.5: Concentration of attention weights.

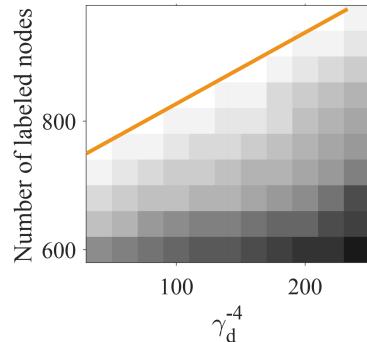


Figure 3.6: Sample complexity against  $\gamma_d$ .

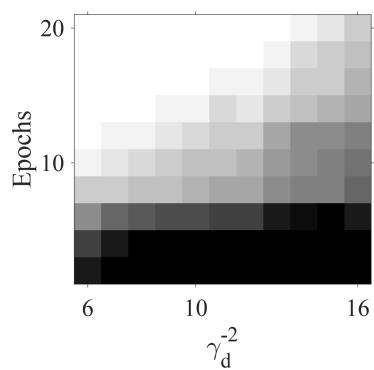
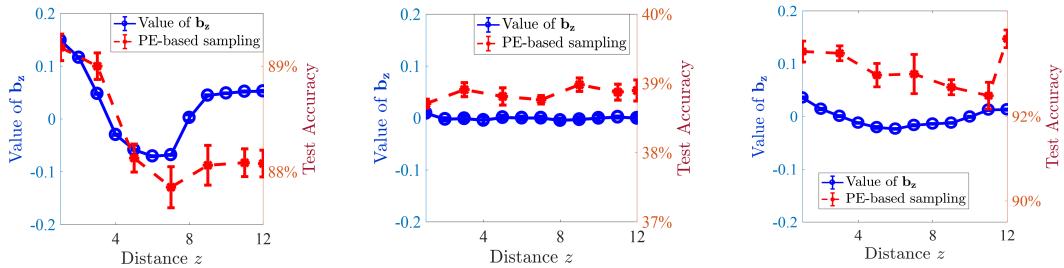


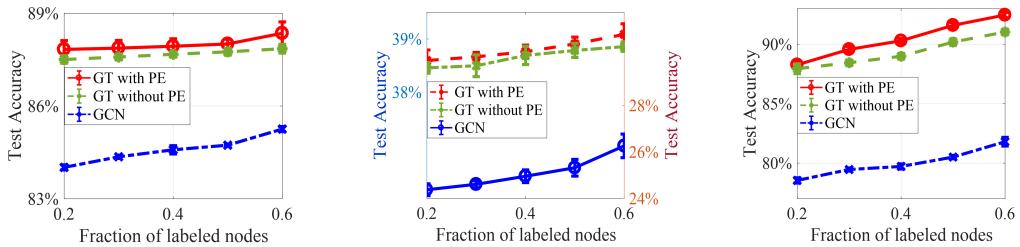
Figure 3.7: The required # of iterations against  $\gamma_d$ .

labeled nodes increases to be close to 1 during training, which justifies (3.11), since when  $\epsilon_S = 0$ , the left side of (3.11) converges to  $1 - \eta^C$  for  $C > 0$  for all nodes. Meanwhile, the summation of attention weights on other nodes decreases to be close to 0, as shown in the red dotted line. We also compare the performance on GT in (3.1) and a one-layer GCN with a similar architecture, and  $\mathbf{W}_K$  and  $\mathbf{W}_Q$  being 0,  $\epsilon_S = 0.2$ . Figures 3.6 and 3.7 show the sample complexity and the required number of iterations of GCN are almost linear in  $\gamma_d^{-4}$  and  $\gamma_d^{-2}$ , consistent with theoretical results in (3.13) and (3.12), respectively. In contrast, the theoretical sample complexity and the number of iterations of GT are respectively linear in  $\gamma_d^{-2}$  (also see Figure 3.3) and independent of  $\gamma_d$ , which are order-wise smaller than GCN.

### 3.5.2 Experiments on Real-world Dataset



**Figure 3.8: The values of entries of  $b$  and the test accuracy of PE-based sampling. Left to right: PubMed, Actor, PascalVOC-SP-1G.**



**Figure 3.9: Test accuracy of GT with/without PE and GCN when the number of labeled nodes varies. Left to right: PubMed, Actor, PascalVOC-SP-1G.**

**Dataset and neural network model:** We evaluate node classification tasks on three benchmarks, a seven-classification citation graph PubMed [32], a five-classification Actor co-occurrence graph [125], and a four-classification computer vision graph PascalVOC-SP-1G [126], which are a homophilous, heterophilous, and a long-range graph, respectively. Please

refer to Appendix B.1 for detailed information on these datasets and results on large-scale dataset Ogbn-Arxiv [127]. The network contains four layers of four-head Transformer blocks. We implement the SPD-based PE as defined in (3.2) with  $Z = 20$  and uniformly sample 20 nodes across the whole graph for feature aggregation of each node during every iteration.

**Success of PE:** The blue circled lines in Figure 3.8 show the average values of each dimension of the last-layer learned PE vector  $\mathbf{b}^{(T)}$  in these three datasets. We additionally train multiple models with the same setup, except that only distance- $z$  nodes are used for training and label prediction, i.e.,  $\mathcal{S}^{n,t}$  (for all labeled nodes  $n$  and iteration  $t$ ) and  $\mathcal{T}^n$  (for all unlabeled nodes  $n$ ) belong to  $\mathcal{N}_z^n$ .  $|\mathcal{S}^{n,t}|$  is still 20. The red dashed curves show the test accuracy of these models. One can see that the test accuracy of these models has a similar trend as that of  $b_z$  values. This justifies the success of PE and the existence of a core neighborhood defined in Definition 3.4.1.

**Comparison of GTs with/without PE and GCN.** We use a four-layer GCN defined in [32]. The model size of GCN is slightly larger than GT by  $\leq 10\%$ . Figure 3.9 shows that GT with PE has a better performance than that without PE and is better than GCN. This verifies Theorem 3.4.6 and discussions in Section 3.4.5.

### 3.6 Conclusion, Limitation, and Future Works

This chapter presents a novel theoretical analysis of Graph Transformers by explicitly characterizing the required sample complexity and the number of training steps to achieve a desirable generalization for node classification tasks. The analysis is based on a new graph data model that includes class-discriminative features that determine classes and class-irrelevant features, as well as a core neighborhood that determines the labels based on a majority vote of class-discriminative features. This chapter shows that the sample complexity and iterations are reduced when the fraction of class-discriminative nodes increases and/or the sampled nodes have a clear-cutting vote in the core neighborhood. This chapter also proves that attention weights are concentrated on those of class-relevant nodes, and the positional embedding promotes the core neighborhood. All the theoretical results are centered on simplified shallow Transformer architectures, while experimental results on real-world datasets and deep neural network architectures support our theoretical findings. Future direction includes theoretically analyzing and designing other models with milder assumptions and devising better graph sampling methods.

# CHAPTER 4

## HOW DOES PROMPTING THE MINORITY FRACTION AFFECT GENERALIZATION? A THEORETICAL STUDY OF ONE-HIDDEN-LAYER NEURAL NETWORK ON GROUP IMBALANCE

### 4.1 Introduction

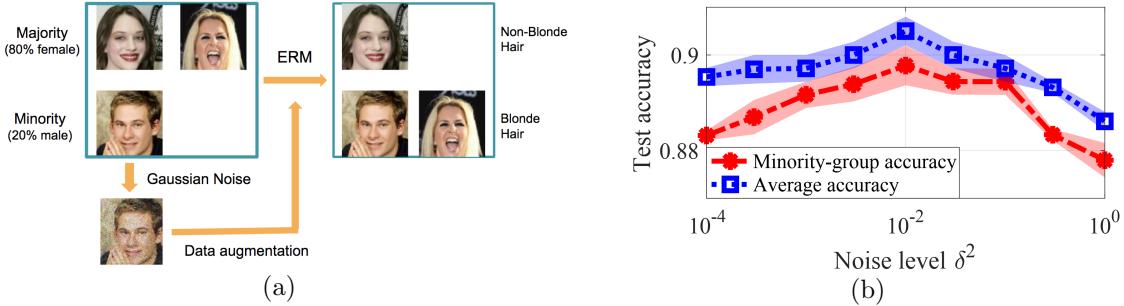
Training neural networks with empirical risk minimization (ERM) is a common practice to reduce the average loss of a machine learning task evaluated on a dataset. However, recent findings [26], [27], [28], [29], [30] have shown empirical evidence about a critical challenge of ERM, known as *group imbalance*, where a well-trained model that has high average accuracy may have significant errors on the minority group that infrequently appears in the data. Moreover, the group attributes that determine the majority and minority groups are usually hidden and unknown during the training. The training set can be augmented by data augmentation methods [128] with varying performance, such as cropping and rotation [129], noise injection [130], and generative adversarial network (GAN)-based methods [131].

As ERM is a prominent method and enjoys great empirical success, it is important to characterize the impact of ERM on group imbalance theoretically. However, the technical difficulty of analyzing the nonconvex ERM problem of neural networks results from the concatenation of nonlinear functions across layers, and the existing generalization analyses of ERM often make overly simplistic assumptions and only focus on the average generalization performance. For example, the neural tangent kernel type of analysis [79],[78],[58],[80], [77],[22] linearizes the neural network around the random initialization to remove the non-convex interactions across layers. The generalization bounds are independent of the feature distribution and cannot be exploited to analyze the impact of individual groups. [84] provides the sample complexity analysis when the data comes from the mixtures of well-separated

---

Portions of this chapter have previously appeared as: H. Li, S. Zhang, Y. Zhang, M. Wang, S. Liu, and P.-Y. Chen, “How does promoting the minority fraction affect generalization? A theoretical study of one-hidden-layer neural network on group imbalance,” *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 216–231, Mar. 2024. ©2024 IEEE.

Portions of this chapter have previously appeared as: H Li, S Zhang, M Wang, “Learning and generalization of one-hidden-layer neural networks, going beyond standard gaussian data,” In *Annu. Conf. Inf. Sci. Syst.*, Mar. 2022. pp. 37-42.



**Figure 4.1: Group imbalance experiment. (a) Binary classification on CelebA dataset using Gaussian augmentation to control the minority group co-variance. (b) Test accuracy against the augmented noise level.**  
©2024 IEEE.

distributions but still cannot characterize the learning performance of individual groups. Another line of works [132],[133],[84],[134],[6] considers one-hidden-layer neural networks because the ERM problem is already highly nonconvex, and the analytical complexity increases tremendously when the number of hidden layers increases. In these works, the input features are usually assumed to be i.i.d. samples drawn from the standard Gaussian distribution, and this data model cannot differentiate the majority and minority groups.

**Contribution:** To the best of our knowledge, *this chapter provides the first theoretical characterization of both the average and group-level generalization of a one-hidden-layer neural network trained by ERM on data generated from a mixture of distributions*. This chapter considers the binary classification problem with the cross entropy loss function, with training data generated by a ground-truth neural network with known architecture and unknown weights. The optimization problem is challenging due to a high non-convexity from the multi-neuron architecture and the non-linear sigmoid activation.

Assuming the features follow a Gaussian Mixture Model (GMM), where samples of each group are generated from a Gaussian distribution with an arbitrary mean vector and co-variance matrix, this chapter quantifies the impact of individual groups on the sample complexity, the training convergence rate, and the average and group-level test error. The training algorithm is the gradient descent following a tensor initialization and converges linearly. Our key results include

- (1) *Medium-range group-level co-variance enhances the learning performance.* When a group-level co-variance deviates from the medium regime, the learning performance degrades

in terms of higher sample complexity, slower convergence in training, and worse average and group-level generalization performance. As shown in Figure 4.1(a), we introduce Gaussian augmentation to control the co-variance level of the minority group in the CelebA dataset [135]. The learned model achieves the highest test accuracy when the co-variance is at the medium level, see Figure 4.1(b). Another implication is that the diverse performance of different data augmentation methods might partially result from the different group-level co-variance introduced by these methods. Furthermore, although our setup does not directly model the batch normalization approach [136] that modifies the mean and variance in each layer to achieve fast and stable convergence, our result provides a theoretical insight that co-variance indeed affects the learning performance.

(2) *Group-level mean shifts from zero hurt the learning performance.* When a group-level mean deviates from zero, the sample complexity increases, the algorithm converges slower, and both the average and group-level test error increases. Thus, the learning performance is improved if each distribution is zero-mean. This chapter provides a similar theoretical insight to practical tricks such as whitening [137], subgroup shift [138], [139], population shift [140], [141] and the pre-processing of making data zero-mean [142], that data mean affects the learning performance.

(3) *Increasing the fraction of the minority group in the training data does not always improve its generalization performance.* The generalization performance is also affected by the mean and co-variance of individual groups. In fact, increasing the fraction of the minority group in the training data can have a completely opposite impact in different datasets.

## 4.2 Background and Related Work

### **Improving the minority-group performance with known group attributes.**

With known group attributes, distributionally robust optimization (DRO) [29] minimizes the worst-group training loss instead of solving ERM. DRO is more computationally expensive than ERM and does not always outperform ERM in the minority-group test error. Spurious correlations [28] can be viewed as one reason of group imbalance, where strong associations between labels and irrelevant features exist in training samples. Different from the approaches that address spurious correlations, such as down-sampling the majority [143], [144], up-weight the minority group [145], and removing spurious features [146], [147], this chapter does not require the special model of spurious correlations and any group attribute information.

**Imbalance learning and long-tailed learning** focus on learning from imbalanced data with a long-tailed distribution, which means that a few classes of the data make up the majority of the dataset, while the majority of classes have little data samples [148], [149], [150], [151], [152], [153], [154], [155], [156]. Some works [149], [156] claimed that naively increasing the number of the minority does not always improve the generalization. Therefore, some recent works develop novel oversampling and data augmentation methods [153], [152], [155] that can promote the minority fraction by generating diverse and context-rich minority data. However, there are very limited theoretical explanations of how these techniques affect the generalization.

**Generalization performance with the standard Gaussian input for one-hidden-layer neural networks.** [157],[158], [159],[160] consider infinite training samples. [71] characterize the sample complexity of fully connected neural networks with smooth activation functions. [161], [75] extend to the non-smooth ReLU activation for fully-connected and convolutional neural networks, respectively. [72] analyzes the cross entropy loss function for binary classification problems. [74] analyzes the generalizability of graph neural networks for both regression and binary classification problems. One-hidden-layer case of neural network pruning and self-training are also studied in [162] and [163], respectively.

**Theoretical characterization of learning performance from other input distributions for one-hidden-layer neural networks.** [164] analyzes the training loss with a single Gaussian with an arbitrary co-variance. [165] quantifies the SGD evolution trained on the Gaussian mixture model. When the hidden layer only contains one neuron, [132] analyzes rotationally invariant distributions. With an infinite number of neurons and an infinite input dimension, [134] analyzes the generalization error based on the mean-field analysis for distributions like Gaussian Mixture with the same mean. [133] considers inputs with low-dimensional structures. No sample complexity is provided in all these works.

**Notations:**  $\mathbf{Z}$  is a matrix with  $Z_{i,j}$  as the  $(i,j)$ -th entry.  $\mathbf{z}$  is a vector with  $z_i$  as the  $i$ -th entry.  $[K]$  denotes the set including integers from 1 to  $K$ .  $\mathbf{I}_d$  and  $\mathbf{e}_i$  represent the identity matrix in  $\mathbb{R}^{d \times d}$  and the  $i$ -th standard basis vector, respectively.  $\delta_i(\mathbf{Z})$  denotes the  $i$ -th largest singular value of  $\mathbf{Z}$ . The matrix norm  $\|\mathbf{Z}\| = \delta_1(\mathbf{Z})$ .  $f(x) = O(g(x))$  (or  $\Omega(g(x))$ ,  $\Theta(g(x))$ ) means that  $f(x)$  increases at most, at least, or in the order of  $g(x)$ , respectively.

### 4.3 Problem Formulation and Algorithm

We consider the classification problem with an unbalanced dataset using fully connected neural networks over  $n$  independent training examples  $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$  from a data distribution. The learning algorithm is to minimize the empirical risk function via gradient descent (GD). In what follows, we will present the data model and neural network model considered in this chapter.

**Data Model.** Let  $\mathbf{x} \in \mathbb{R}^d$  and  $y \in \mathbb{R}$  denote the input feature and label, respectively. We consider an unbalanced dataset that consists of  $L$  ( $L \geq 2$ ) groups of data, where the feature  $\mathbf{x}$  in the group  $l$  ( $l \in [L]$ ) is drawn from a multi-variate Gaussian distribution with mean  $\boldsymbol{\mu}_l \in \mathbb{R}^d$ , and covariance  $\boldsymbol{\Sigma}_l \in \mathbb{R}^{d \times d}$ . Specifically,  $\mathbf{x}$  follows the Gaussian mixture model (GMM) [166], [167], [168], denoted as  $\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ .  $\lambda_l \in (0, 1)$  is the probability of sampling from distribution- $l$  and represents the expected fraction of group- $l$  data.  $\sum_{l=1}^L \lambda_l = 1$ . Group  $l$  is defined as a minority group if  $\lambda_l$  is less than  $1/L$ . We use  $\Psi = \{\lambda_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \forall l\}$  to denote all parameters of the mixture model<sup>14</sup>. We consider binary classification with label  $y$  generated by a ground-truth neural network with unknown weights  $\mathbf{W}^* = [\mathbf{w}_1^*, \dots, \mathbf{w}_K^*] \in \mathbb{R}^{d \times K}$  and sigmoid activation<sup>15</sup>. function  $\phi(x) = \frac{1}{1+\exp(-x)}$ , where<sup>16</sup>

$$\mathbb{P}(y = 1 | \mathbf{x}) = H(\mathbf{W}^*, \mathbf{x}) := \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*\top} \mathbf{x}). \quad (4.1)$$

**Learning model.** Learning is performed over a neural network that has the same architecture as in (4.1), which is a one-hidden-layer fully connected neural network<sup>17</sup> with its weights denoted by  $\mathbf{W} \in \mathbb{R}^{d \times K}$ . Given  $n$  training samples  $\{\mathbf{x}_i, y_i\}_{i=1}^n$  where  $\mathbf{x}_i$  follows the GMM model, and  $y_i$  is from (4.1), we aim to find the model weights via solving the empirical

---

<sup>14</sup>In practice,  $\Psi$  can be estimated by the EM algorithm and the moment-based method [166]. The EM algorithm returns model parameters within Euclidean distance  $O((\frac{d}{n})^{\frac{1}{2}})$  when the number of mixture components  $L$  is known. When  $L$  is unknown, one usually over-specifies an estimate  $\bar{L} > L$ , then the estimation error by the EM algorithm scales as  $O((\frac{d}{n})^{\frac{1}{4}})$ . Please refer to [169],[170],[171] for details.

<sup>15</sup>The results can be generalized to any activation function  $\phi$  with bounded  $\phi$ ,  $\phi'$  and  $\phi''$ , where  $\phi'$  is even. Examples include  $\tanh$  and  $\text{erf}$ .

<sup>16</sup>Our data model is reduced to logistic regression in the special case that  $K = 1$ . We mainly study the more challenging case when  $K > 1$ , because the learning problem becomes highly non-convex when there are multiple neurons in the network.

<sup>17</sup>All the weights in the second layer are assumed to be fixed to facilitate the analysis. This is a standard assumption in theoretical generalization analysis [161], [72], [74].

risk minimization (ERM), where  $f_n(\mathbf{W})$  is the empirical risk,

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times K}} f_n(\mathbf{W}) := \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{W}; \mathbf{x}_i, y_i), \quad (4.2)$$

where  $\ell(\mathbf{W}; \mathbf{x}_i, y_i)$  is the cross-entropy loss function, i.e.,

$$\begin{aligned} \ell(\mathbf{W}; \mathbf{x}_i, y_i) &= -y_i \cdot \log(H(\mathbf{W}, \mathbf{x}_i)) \\ &\quad - (1 - y_i) \cdot \log(1 - H(\mathbf{W}, \mathbf{x}_i)). \end{aligned} \quad (4.3)$$

Note that for any permutation matrix  $\mathbf{P}$ ,  $\mathbf{W}\mathbf{P}$  corresponds permuting neurons of a network with weights  $\mathbf{W}$ . Therefore,  $H(\mathbf{W}, \mathbf{x}) = H(\mathbf{W}\mathbf{P}, \mathbf{x})$ , and  $f_n(\mathbf{W}\mathbf{P}) = f_n(\mathbf{W})$ . The estimation is considered successful if one finds any column permutation of  $\mathbf{W}^*$ .

The average generalization performance of a learned model  $\mathbf{W}$  is evaluated by the average risk

$$\bar{f}(\mathbf{W}) = \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \ell(\mathbf{W}; \mathbf{x}_i, y_i), \quad (4.4)$$

and the generalization performance on group  $l$  is evaluated by the group- $l$  risk

$$\bar{f}_l(\mathbf{W}) = \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \ell(\mathbf{W}; \mathbf{x}_i, y_i). \quad (4.5)$$

**Training Algorithm.** Our algorithm starts from an initialization  $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$  computed based on the tensor initialization method (Subroutine 1 in Appendix) and then updates the iterates  $\mathbf{W}_t$  using gradient descent with the step size<sup>18</sup>  $\eta_0$ . The computational complexity of tensor initialization is  $O(Knd)$ . The per-iteration complexity of the gradient step is  $O(Knd)$ . We defer the details of Algorithm 1 in Appendix.

## 4.4 Main Theoretical Results

We will formally present our main theory below, and the insights are summarized in Section 4.4.1. For the convenience of presentation, some quantities are defined here, and all of them can be viewed as constant. Define  $\sigma_{\max} = \max_{l \in [L]} \{\|\boldsymbol{\Sigma}_l\|^{\frac{1}{2}}\}$ ,  $\sigma_{\min} = \min_{l \in [L]} \{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\}$ .

---

<sup>18</sup>Algorithm 1 employs a constant step size. One can potentially speed up the convergence, i.e., reduce  $v$ , by using a variable step size. We leave the corresponding theoretical analysis for future work.

---

**Algorithm 1** Our ERM learning algorithm
 

---

- 1: **Input:** Training data  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ , the step size  $\eta_0 = O\left(\left(\sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\right)^{-1}\right)$ , the total number of iterations  $T$
- 2: **Initialization:**  $\mathbf{W}_0 \leftarrow$  Tensor initialization method via Subroutine 1
- 3: **Gradient Descent:** for  $t = 0, 1, \dots, T - 1$

$$\begin{aligned} \mathbf{W}_{t+1} &= \mathbf{W}_t - \eta_0 \cdot \frac{1}{n} \sum_{i=1}^n (\nabla l(\mathbf{W}, \mathbf{x}_i, y_i) + \nu_i) \\ &= \mathbf{W}_t - \eta_0 \left( \nabla f_n(\mathbf{W}) + \frac{1}{n} \sum_{i=1}^n \nu_i \right) \end{aligned} \quad (4.6)$$

- 4: **Output:**  $\mathbf{W}_T$
- 

Let  $\tau = \sigma_{\max}/\sigma_{\min}$ . We assume  $\tau = \Theta(1)$ <sup>19</sup>, indicating that  $\sigma_{\max}$  and  $\sigma_{\min}$  are in the same order. Let  $\delta_i(\mathbf{W}^*)$  denote the  $i$ -th largest singular value of  $\mathbf{W}^*$ . Let  $\kappa = \frac{\delta_1(\mathbf{W}^*)}{\delta_K(\mathbf{W}^*)}$ , and define  $\eta = \prod_{i=1}^K (\delta_i(\mathbf{W}^*)/\delta_K(\mathbf{W}^*))$ .

*Theorem 2.* There exist  $\epsilon_0 \in (0, \frac{1}{4})$  and positive value functions  $\mathcal{B}(\Psi)$  (sample complexity parameter),  $q(\Psi)$  (convergence rate parameter), and  $\mathcal{E}_w(\Psi)$ ,  $\mathcal{E}(\Psi)$ ,  $\mathcal{E}_l(\Psi)$  (generalization parameters) such that as long as the sample size  $n$  satisfies

$$n \geq n_{\text{sc}} := \text{poly}(\epsilon_0^{-1}, \kappa, \eta, \tau, K, \delta_1(\mathbf{W}^*)) \mathcal{B}(\Psi) d \log^2 d, \quad (4.7)$$

we have that with probability at least  $1 - d^{-10}$ , the iterates  $\{\mathbf{W}_t\}_{t=1}^T$  returned by Algorithm 1 with step size  $\eta_0 = O\left(\left(\sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\right)^{-1}\right)$  converge linearly with a statistical error to a critical point  $\widehat{\mathbf{W}}_n$  with the rate of convergence  $v$ , i.e.,

$$\|\mathbf{W}_t - \widehat{\mathbf{W}}_n\|_F \leq v(\Psi)^t \|\mathbf{W}_0 - \widehat{\mathbf{W}}_n\|_F + \frac{\eta_0 \xi}{1 - v(\Psi)} \sqrt{dK \log n/n}, \quad (4.8)$$

$$v(\Psi) = 1 - K^{-2} q(\Psi), \quad (4.9)$$

where  $\xi \geq 0$  is the upper bound of the entry-wise additive noise in the gradient computation.

---

<sup>19</sup>This is a mild assumption because many practical datasets can be approximated with a minor loss of information by a low-rank dataset which has  $\tau$  being  $\Theta(1)$ . It is verified in Section A.1 in the Appendix using the CelebA dataset.

**Table 4.1: Impact of GMM parameters on the learning performance in sample regimes. ©2024 IEEE.**

	$\Sigma_l$ changes		$\mu_l$ changes	$\lambda_l$ changes, const. $\ \Sigma_j\ $ 's, equal $\ \mu_j\ $ 's	
	$\ \Sigma_l\  = o(1)$	$\ \Sigma_l\  = \Omega(1)$		if $\ \Sigma_l\  = \sigma_{\min}^2$	if $\ \Sigma_l\  = \sigma_{\max}^2$
$\mathcal{B}(\Psi)$ , sample compl. $n_{sc}$	$O( \Sigma_l ^{-3})$	$O\ \Sigma_l\ ^3)$	$O(\text{poly}(\ \mu_l\ ))$	$O(\frac{1}{(1+\lambda_l)^2})$	$O(1) - \frac{\Theta(1)}{(1+\lambda_l)^2}$
conv. rate $v(\Psi) \propto -q(\Psi)$	$1 - \Theta(\ \Sigma_l\ ^3)$	$1 - \Theta(\frac{1}{1+\ \Sigma_l\ })$	$1 - \Theta(\frac{1}{\ \mu_l\ ^2+1})$	$\Theta(\frac{1}{1+\lambda_l})$	$1 - \Theta(\frac{1}{1+\lambda_l})$
$\mathcal{E}_w(\Psi)$ , $\ \widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\ _F$	$O(1) - \Theta(\ \Sigma_l\ ^3)$	$O(\sqrt{\ \Sigma_l\ })$	$O(1 + \ \mu_l\ )$	$O(\frac{1}{1+\sqrt{\lambda_l}})$	$O(1 + \sqrt{\lambda_l})$
$\mathcal{E}(\Psi)$ , average risk $\bar{f}$	$O(1) - \Theta(\ \Sigma_l\ ^3)$	$O(\ \Sigma_l\ )$	$O(1 + \ \mu_l\ ^2)$	$O(\frac{1}{1+\lambda_l})$	$O(1) - \frac{\Theta(1)}{1+\lambda_l}$
$\mathcal{E}_l(\Psi)$ , group-l risk $\bar{f}_l$	$O(1) - \Theta(\ \Sigma_l\ ^3)$	$O(\ \Sigma_l\ )$	$O(1 + \ \mu_l\ ^2)$	$O(\frac{1}{1+\sqrt{\lambda_l}})$	$O(1 + \sqrt{\lambda_l})$

Moreover, there exists a permutation matrix  $\mathbf{P}^*$  such that

$$\begin{aligned} \|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}^*\|_F &\leq \mathcal{E}_w(\Psi) \cdot \text{poly}(\kappa, \eta, \tau, \delta_1(\mathbf{W}^*)) \\ &\quad \cdot \Theta\left(K^{\frac{5}{2}}(1 + \xi) \cdot \sqrt{d \log n/n}\right). \end{aligned} \quad (4.10)$$

The average population risk  $\bar{f}$  and the group-l risk  $\bar{f}_l$  satisfy

$$\bar{f} \leq \mathcal{E}(\Psi) \cdot \text{poly}(\kappa, \eta, \tau, \delta_1(\mathbf{W}^*)) \cdot \Theta\left(K^{\frac{5}{2}}(1 + \xi) \cdot \sqrt{d \log n/n}\right) \quad (4.11)$$

$$\bar{f}_l \leq \mathcal{E}_l(\Psi) \cdot \text{poly}(\kappa, \eta, \tau, \delta_1(\mathbf{W}^*)) \cdot \Theta\left(K^{\frac{5}{2}}(1 + \xi) \cdot \sqrt{d \log n/n}\right) \quad (4.12)$$

The closed-form expressions of  $\mathcal{B}$ ,  $q$ ,  $\mathcal{E}_w$ ,  $\mathcal{E}$ , and  $\mathcal{E}_l$  are in Section D of the supplementary material and skipped here. The quantitative impact of the GMM model parameters  $\Psi$  on the learning performance varies in different regimes and can be derived from Theorem 2. The following corollary summarizes the impact of  $\Psi$  on the learning performance in some sample regimes.

**Corollary 4.4.1.** *When we vary one parameter of group  $l$  for any  $l \in [L]$  of the GMM*

model  $\Psi$  and fix all the others, the learning performance degrades in the sense that the sample complexity  $n_{sc}$ , the convergence rate  $v$ ,  $\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F$ , average risk  $\bar{f}$  and group- $l$  risk  $\bar{f}_l$  all increase (details summarized in Table 4.1), as long as any of the following conditions happens,

(i)  $\|\Sigma_l\|$  approaches 0; (ii)  $\|\Sigma_l\|$  increases from some constant; (iii)  $\|\mu_l\|$  increases from 0,

(iv)  $\lambda_l$  decreases, provided that  $\|\Sigma_l\| = \sigma_{\min}^2$ , i.e., group  $l$  has the smallest group-level co-variance, where  $\|\Sigma_j\|$  are all constants, and  $\|\mu_i\| = \|\mu_j\|$  for all  $i, j \in [L]$ .

(v)  $\lambda_l$  increases, provided that  $\|\Sigma_l\| = \sigma_{\max}^2$ , i.e., group  $l$  has the largest group-level co-variance, where  $\|\Sigma_j\|$  are all constants, and  $\|\mu_i\| = \|\mu_j\|$  for all  $i, j \in [L]$ .

To the best of our knowledge, Theorem 2 provides the first characterization of the sample complexity, learning rate, and generalization performance under the Gaussian mixture model. It also firstly characterizes the per-group generalization performance in addition to the average generalization.

#### 4.4.1 Theoretical Insights

We summarize the crucial implications of Theorem 2 and Corollary 4.4.1 as follows.

**(P1). Training convergence and generalization guarantee.** The iterates  $\mathbf{W}_t$  converge to a critical point  $\widehat{\mathbf{W}}_n$  linearly, and the distance between  $\widehat{\mathbf{W}}_n$  and  $\mathbf{W}^* \mathbf{P}^*$  is  $O(\sqrt{d \log n / n})$  for a certain permutation matrix  $\mathbf{P}^*$ . When the computed gradients contain noise, there is an additional error term of  $O(\xi \sqrt{d \log n / n})$ , where  $\xi$  is the noise level ( $\xi = 0$  for noiseless case). Moreover, the average risk of all groups and the risk of each individual group are both  $O((1 + \xi) \sqrt{d \log n / n})$ .

**(P2). Sample complexity.** For a given GMM, the sample complexity is  $\Theta(d \log^2 d)$ , where  $d$  is the feature dimension. This result is in the same order as the sample complexity for the standard Gaussian input in [72] and [71]. Our bound is almost order-wise optimal with respect to  $d$  because the degree of freedom is  $dK$ . The additional multiplier of  $\log^2 d$  results from the concentration bound in the proof technique. We focus on the dependence on the feature dimension  $d$  and treat the network width  $K$  as constant. The sample complexity in [72] and [71] is also  $d \cdot \text{poly}(K, \log d)$ .

**(P3). Learning performance is improved at a medium regime of group-level co-variance.** On the one hand, when  $\|\Sigma_l\|$  is  $\Omega(1)$ , the learning performance degrades as

$\|\Sigma_l\|$  increases in the sense that the sample complexity  $n_{sc}$ , the convergence rate  $v$ , the estimation error of  $\mathbf{W}^*$ , the average risk  $\bar{f}$ , and the group- $l$  risk  $\bar{f}_l$  all increase. This is due to the saturation of the loss and gradient when the samples have a large magnitude. On the other hand, when  $\|\Sigma_l\|$  is  $o(1)$ , the learning performance also degrades when  $\|\Sigma_l\|$  approaches zero. The intuition is that in this regime, the input data are concentrated on a few vectors, and the optimization problem does not have a benign landscape.

**(P4). Increasing the fraction of the minority group data does not always improve the generalization**, while the performance also depends on the mean and covariance of individual groups. Take  $\|\Sigma_j\| = \Theta(1)$  for all group  $j$ , and  $\|\mu_j\|$  is the same for all  $j$  as an example (columns 5 and 6 of Table 4.1). When  $\|\Sigma_l\|$  is the smallest among all groups, increasing  $\lambda_l$  improves the learning performance. When  $\|\Sigma_l\|$  is the largest among all groups, increasing  $\lambda_l$  actually degrades the performance. The intuition is that from (P3), the learning performance is enhanced at a medium regime of group-level co-variance. Thus, increasing the fraction of a group with a medium level of co-variance improves the performance, while increasing the fraction of a group with large co-variance degrades the learning performance. Similarly, when augmenting the training data, an argumentation method that introduces medium variance could improve the learning performance, while an argumentation method that introduces a significant level of variance could hurt the learning performance.

**(P5). Group-level mean shifts from zero degrade the learning performance.** The learning performance degrades as  $\|\mu_l\|$  increases. An intuitive explanation of the degradation is that some training samples have a significant large magnitude such that the sigmoid function saturates.

#### 4.4.2 Proof Idea and Technical Novelty

**Proof Idea.** Different from the analysis of logistic regression for generalized linear models, our paper deals with more technical challenges of nonconvex optimization due to the multi-neuron architecture, the GMM model, and a more complicated activation and loss. The establishment of Theorem 2 consists of three key lemmas.

**Lemma 4.4.2.** (*informal version*) *As long as the number of training samples is larger than  $\Omega(dK^5 \log^2 d)$ , the empirical risk function is strongly convex in the neighborhood of  $\mathbf{W}^*$  (or a permutation of  $\mathbf{W}^*$ ). The size of the convex region is characterized by the Gaussian mixture distribution.*

The main proof idea of Lemma 4.4.2 is to show that the nonconvex empirical risk  $f_n(\mathbf{W})$  in a small neighborhood around  $\mathbf{W}^*$  (or any permutation  $\mathbf{W}^*\mathbf{P}$ ) is almost convex with a sufficiently large  $n$ . The difficulty is to find a positive lower bound of the smallest singular value of  $\nabla^2 \bar{f}(\mathbf{W})$ , which should also be a function of the GMM. Then, we can obtain  $\nabla^2 f_n(\mathbf{W})$  from  $\nabla^2 \bar{f}(\mathbf{W})$  by concentration inequalities.

**Lemma 4.4.3.** (*informal version*) *If initialized in the convex region, the gradient descent algorithm converges linearly to a critical point  $\widehat{\mathbf{W}}_n$ , which is close to  $\mathbf{W}^*$  (or any permutation of  $\mathbf{W}^*$ ), and the distance is diminishing as the number of training samples increases.*

Given the locally strong convexity, Lemma 4.4.3 provides the linear convergence to a critical point. The convergence rate is determined by the GMM.

**Lemma 4.4.4.** (*informal version*) *Tensor Initialization Method initializes  $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$  around  $\mathbf{W}^*$  (or a permutation of  $\mathbf{W}^*$ ).*

The idea of tensor initialization is to first find quantities (see  $\mathbf{Q}_j$  in Definition 1) in the supplementary material) which are proven to be functions of tensors of  $\mathbf{w}_i^*$ . Then the method approximates these quantities numerically using training samples and then applies the tensor decomposition method on the estimated quantities to obtain  $\mathbf{W}_0$ , which is an estimation of  $\mathbf{W}^*$ .

Combining the above three lemmas together, one can derive the required sample complexity and the upper bound of  $\bar{f}$  and  $\bar{f}_l$  in (4.7), (4.11), and (4.12), respectively. The idea is first to compute the sample complexity bound such that the tensor initialization method initializes  $\mathbf{W}_O$  in the local convex region by Lemma 4.4.4. Then the final sample complexity is obtained by comparing two sample complexities from Lemma 4.4.2 and 4.4.4.

By further looking into the order of the terms  $\mathcal{B}(\Psi)$ ,  $v(\Psi)$ ,  $\mathcal{E}(\Psi)$ ,  $\mathcal{E}_w(\Psi)$ , and  $\mathcal{E}_l(\Psi)$  in several cases of  $\Psi$ , Theorem 2 leads to Corollary 4.4.1. To be more specific, we only vary parameters  $\Sigma_l$ , or  $\boldsymbol{\mu}_l$ , or  $\lambda_l$  following the cases in Table 4.1, while fixing all other parameters of  $\Psi$ . We apply the Taylor expansion to approximate the terms and derive error bounds with the Lipschitz smoothness of the loss function.

**Technical Novelty.** Our algorithmic and analytical framework is built upon some recent works on the generalization analysis of one-hidden-layer neural networks, see, e.g., [71], [161], [72], [74], [162], which assume that  $\mathbf{x}_i$  follows the standard Gaussian distribution and

cannot be directly extended to GMM. This chapter makes new technical contributions from the following aspects.

**First, we characterize the local convex region near  $\mathbf{W}^*$  for the GMM model.**

To be more specific, we explicitly characterize the positive lower bound of the smallest singular value of  $\nabla^2 \bar{f}(\mathbf{W})$  with respect to  $\Psi$ , while existing results either only hold for standard Gaussian data [71], [72], [162], [163], or can only show  $\nabla^2 \bar{f}(\mathbf{W})$  is positive definite regardless the impact of  $\Psi$  [79].

**Second, new tools, including matrix concentration bounds are developed to explicitly quantify the impact of  $\Psi$  on the sample complexity.**

**Third, we investigate and provide the order of the bound for sample complexity, convergence rate, generalization error, average risk, and group- $l$  risk in terms of  $\Psi$  for the first time** in the line of research of model estimation [71], [72], [74], [162], [163], which is also a novel result for the case of Gaussian inputs.

**Fourth, we design and analyze new tensors for the mixture model to initialize properly,** while the previous tensor methods in [71], [161], [72], [74] utilize the rotation invariant property that only holds for zero mean Gaussian.

## 4.5 Numerical Experiments

### 4.5.1 Experiments on Synthetic Datasets

We first verify the theoretical bounds in Theorem 2 on synthetic data. Each entry of  $\mathbf{W}^* \in \mathbb{R}^{d \times K}$  is generated from  $\mathcal{N}(0, 1)$ . The training data  $\{\mathbf{x}_i, y_i\}_{i=1}^n$  is generated using the GMM model and (4.1). If not otherwise specified,  $L = 2$ ,  $d = 5$ , and  $K = 3^{20}$ . To reduce the computational time, we randomly initialize near  $\mathbf{W}^*$  instead of computing the tensor initialization<sup>21</sup>.

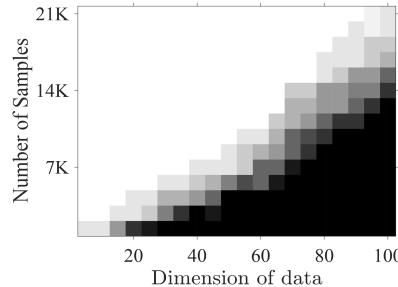
**Sample complexity.** We first study the impact of  $d$  on the sample complexity. Let  $\boldsymbol{\mu}_1 = \mathbf{1}$  in  $\mathbb{R}^d$  and let  $\boldsymbol{\mu}_2 = \mathbf{0}$ . Let  $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$ .  $\lambda_1 = \lambda_2 = 0.5$ . We randomly initialize  $M$  times and let  $\widehat{\mathbf{W}}_n^{(m)}$  denote the output of Algorithm 1 in the  $m$ th trial. Let  $\bar{\mathbf{W}}_n$  denote the

---

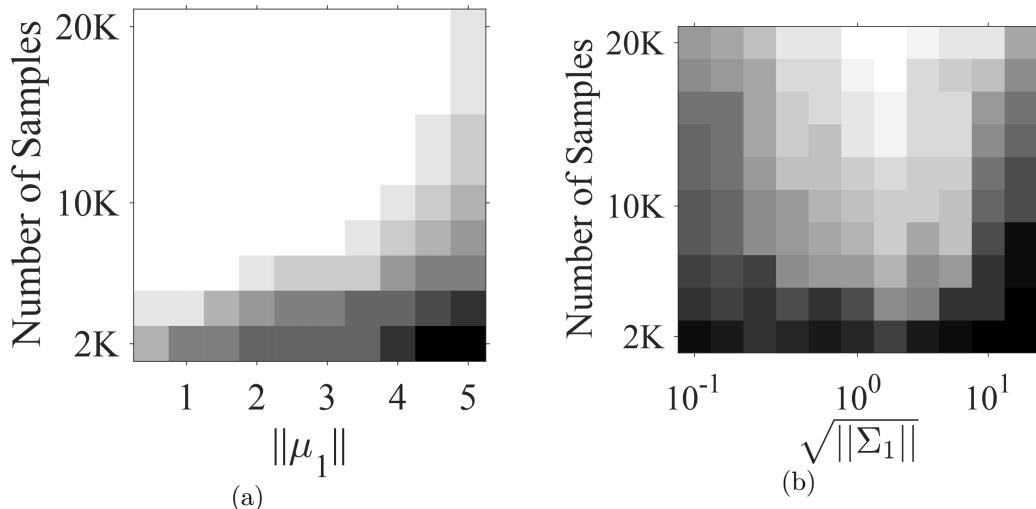
<sup>20</sup>Like [71], [161], [72], we consider a small-sized network in synthetic experiments to reduce the computational time, especially for computing the sample complexity in Figure 4.3. Our results hold for large networks too.

<sup>21</sup>The existing methods based on tensor initialization all use random initialization in synthetic experiments to reduce the computational time. See [72], [161], [74], [163] as examples. We compare tensor initialization and local random initialization numerically in Section B of the supplementary material and show that they have the same performance.

mean values of all  $\widehat{\mathbf{W}}_n^{(m)}$ , and let  $V_W = \sqrt{\sum_{m=1}^M \|\widehat{\mathbf{w}}_n^m - \bar{\mathbf{W}}_n\|^2/M}$  denote the variance. An experiment is successful if  $V_W \leq 10^{-3}$  and fails otherwise.  $M$  is set to 20. For each pair of  $d$  and  $n$ , 20 independent sets of  $\mathbf{W}^*$  and the corresponding training samples are generated. Figure 4.2 shows the success rate of these independent experiments. A black block means that all the experiments fail. A white block means that they all succeed. The sample complexity is indeed almost linear in  $d$ , as predicted by (4.7).



**Figure 4.2:** The sample complexity when the feature dimension changes. ©2024 IEEE.

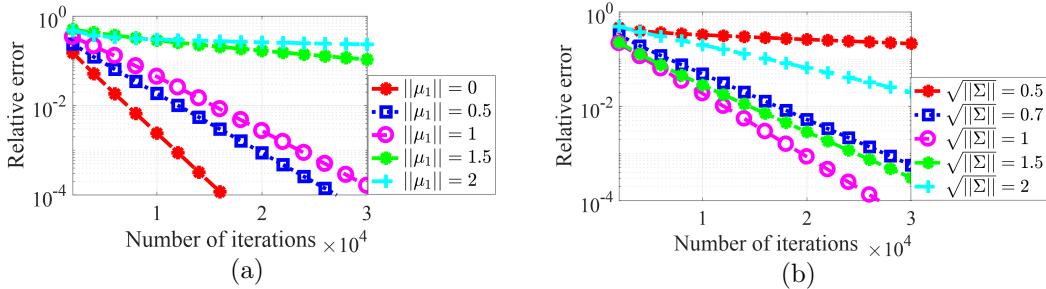


**Figure 4.3:** The sample complexity (a) when one mean changes, (b) when one co-variance changes. ©2024 IEEE.

We next study the impact on the sample complexity of the GMM model. In Figure 4.3 (a),  $\Sigma_1 = \Sigma_2 = \mathbf{I}$ , and let  $\boldsymbol{\mu}_1 = \mu \cdot \mathbf{1}$ ,  $\boldsymbol{\mu}_2 = -\mathbf{1}$ .  $\|\boldsymbol{\mu}_1\|$  varies from 0 to 5. Figure 4.3(a) shows that when the mean increases, the sample complexity increases. In Figure 4.3 (b), we fix  $\boldsymbol{\mu}_1 = \mathbf{1}$ ,  $\boldsymbol{\mu}_2 = -\mathbf{1}$ , and let  $\Sigma_1 = \sigma^2 \mathbf{I}$  and  $\Sigma_2 = \mathbf{I}$ .  $\sigma$  varies from  $10^{-1}$  to  $10^1$ . The sample

complexity increases both when  $\|\Sigma_1\|$  increases and when  $\|\Sigma_1\|$  approaches zero. All results match predictions in Corollary 4.4.1.

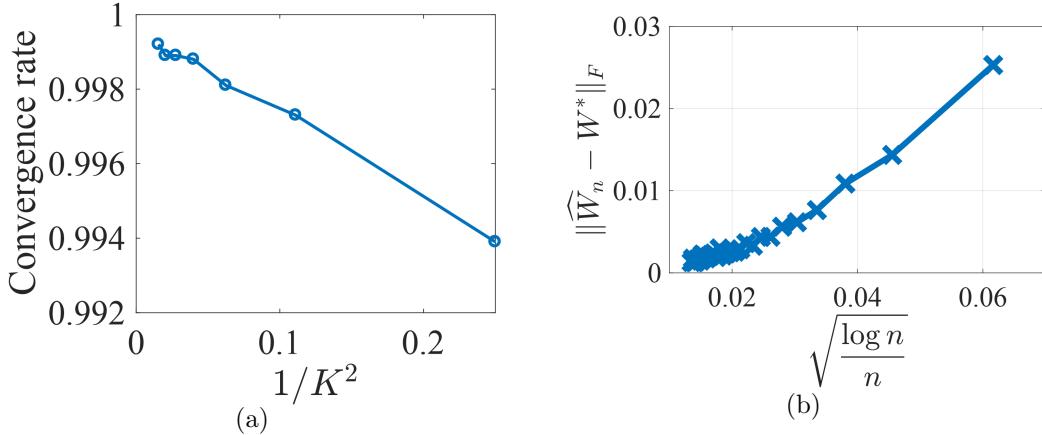
**Convergence analysis.** We next study the convergence rate of Algorithm 1. Figure 4.4(a) shows the impact of  $\|\mu_1\|$ .  $\lambda_1 = \lambda_2 = 0.5$ ,  $\mu_1 = -\mu_2 = C \cdot \mathbf{1}$  for a positive  $C$ , and  $\Sigma_1 = \Sigma_2 = \Lambda^\top D \Lambda$ . Here  $\Lambda$  is generated by computing the left-singular vectors of a  $d \times d$  random matrix from the Gaussian distribution.  $D = \text{diag}(1, 1.1, 1.2, 1.3, 1.4)$ .  $n = 1 \times 10^4$ . Algorithm 1 always converges linearly when  $\|\mu_1\|$  changes. Moreover, as  $\|\mu_1\|$  increases, Algorithm 1 converges slower. Figure 4.4 (b) shows the impact of the variance of the Gaussian mixture model.  $\lambda_1 = \lambda_2 = 0.5$ ,  $\mu_1 = \mathbf{1}$ ,  $\mu_2 = -\mathbf{1}$ ,  $\Sigma_1 = \Sigma_2 = \Sigma = \sigma^2 \cdot \Lambda^\top D \Lambda$ .  $n = 5 \times 10^4$ . We change  $\|\Sigma\|$  by changing  $\sigma$ . Among the values we test, Algorithm 1 converges fastest when  $\|\Sigma\| = 1$ . The convergence rate slows down when  $\|\Sigma\|$  increases or decreases from 1. All results are consistent with the predictions in Corollary 4.4.1. We then study the impact of  $K$  on the convergence rate.  $\lambda_1 = \lambda_2 = 0.5$ ,  $\mu_1 = \mathbf{1}$ ,  $\mu_2 = -\mathbf{1}$ ,  $\Sigma_1 = \Sigma_2 = \mathbf{I}$ . Figure 4.5 (a) shows that, as predicted by (4.9), the convergence rate is linear in  $-1/K^2$ .



**Figure 4.4:** (a) The convergence rate with different  $\mu_1$ . (b) The convergence rate with different  $\Sigma$ . ©2024 IEEE.

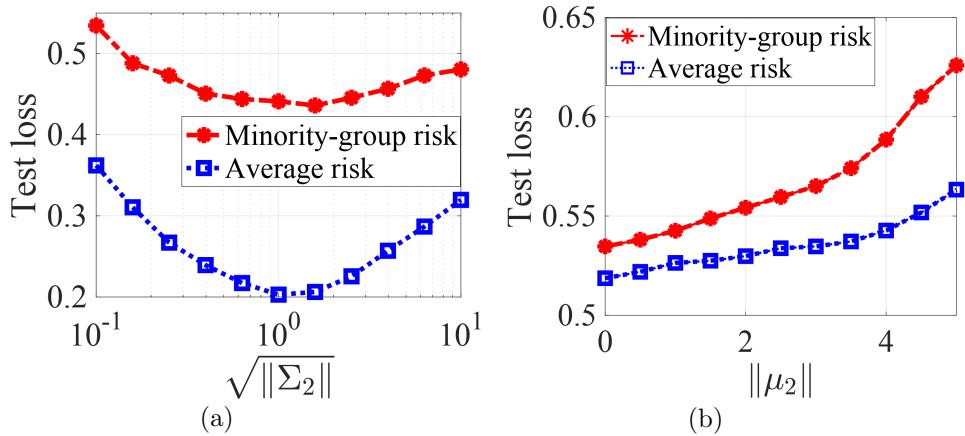
**Average and group-level generalization performance.** The distance between  $\widehat{\mathbf{W}}_n$  returned by Algorithm 1 and  $\mathbf{W}^*$  is measured by  $\|\widehat{\mathbf{W}}_n - \mathbf{W}^*\|_F$ .  $n$  ranges from  $2 \times 10^3$  to  $6 \times 10^4$ .  $\Sigma_1 = \Sigma_2 = 9\mathbf{I}$ ,  $\mu_1 = \mathbf{1}$ ,  $\mu_2 = -\mathbf{1}$ . Each point in Figure 4.5 (b) is averaged over 20 experiments of different  $\mathbf{W}^*$  and training set. The error is indeed linear in  $\sqrt{\log(n)/n}$ , as predicted by (4.8).

We evaluate the impact of one mean/co-variance of the minority group on the generalization.  $n = 2 \times 10^4$ . Let  $\lambda_1 = 0.8$ ,  $\lambda_2 = 0.2$ ,  $\mu_1 = 2 \cdot \mathbf{1}$ ,  $\Sigma_1 = \mathbf{I}$ . First, we let  $\mu_2 = (\mu_2 - 2) \cdot \mathbf{1}$  and  $\Sigma_2 = \mathbf{I}$ . Figure 4.6 (b) shows that both the average risk and the group-2 risk increase as  $\mu_2$  increases, consistent with (P5). Then we set  $\mu_2 = -2 \cdot \mathbf{1}$ ,  $\Sigma_2 = \sigma_2^2 \cdot \mathbf{I}$ . Figure 4.6 (a)



**Figure 4.5:** (a) Convergence rate when the number of neurons  $K$  changes. (b) The relative error of the learned model when  $n$  changes. ©2024 IEEE.

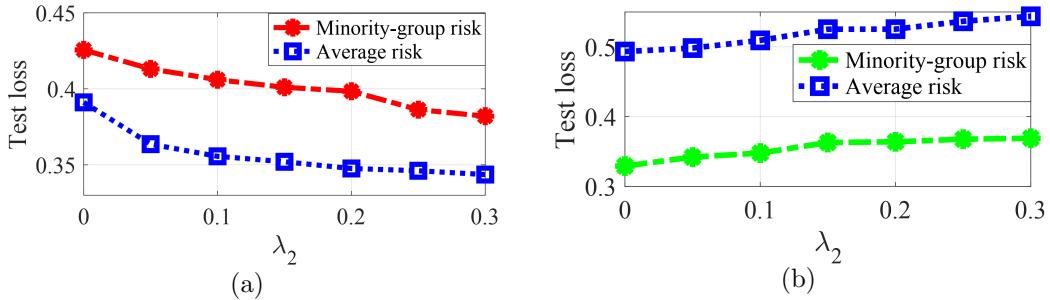
indicates that both the average and the group-2 risk will first decrease and then increase as  $\|\Sigma\|_2$  increases, consistent with (P3).



**Figure 4.6:** (a) The cross-entropy test loss when the co-variance of the minority group changes. (b) The cross-entropy test loss when the mean of the minority group changes. ©2024 IEEE.

Next, we study the impact of increasing the fraction of the minority group.  $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$ . Let group 2 be the minority group. In Figure 4.7 (a),  $\Sigma_1 = 10 \cdot \mathbf{I}$  and  $\Sigma_2 = \mathbf{I}$ , the minority group has a smaller level of co-variance. Then when  $\lambda_2$  increases from 0 to 0.5, both the average and group-2 risk decease. In Figure 4.7 (b),  $\Sigma_1 = \mathbf{I}$  and  $\Sigma_2 = 10 \cdot \mathbf{I}$ , and the minority group has a higher-level of co-variance. Then when  $\lambda_2$  increases from 0 to 0.3, both the average and group-2 risk increase. As predicted by insight (P4), increasing  $\lambda_2$  does not

necessarily improve the generalization of group 2.



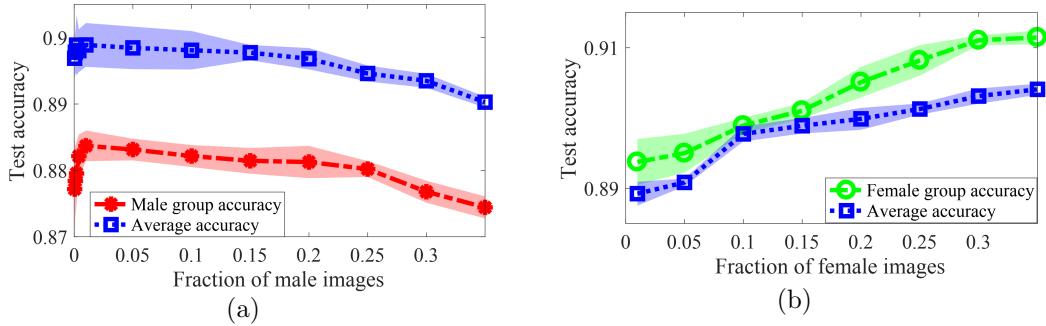
**Figure 4.7:** The test loss (cross entropy loss) of synthetic data with different  $\lambda_2$  values. (a) Group 2 has a smaller level of co-variance. (b) Group 2 has a larger level of co-variance. ©2024 IEEE.

#### 4.5.2 Image Classification on Dataset CelebA

We choose the attribute “blonde hair” as the binary classification label. ResNet 9 [172] is selected to be the learning model here because it was applied in many simple computer vision tasks [173], [174]. To study the impact of co-variance, we pick 4000 female (majority) and 1000 male (minority) images and implement Gaussian data augmentation to create additional 300 images for the male group. Specifically, we select 300 out of 1000 male images and add i.i.d. noise drawn from  $\mathcal{N}(0, \delta^2)$  to every entry. The test set includes 500 male and 500 female images. Figure 4.1 shows that when  $\delta^2$  increases, i.e., when the co-variance of the minority group increases, both the minority-group and average test accuracy increase first and then decrease, coinciding with our insight (P3).

Then we fix the total number of training data to be 5000 and vary the fractions of the two groups. From Figure 4.8(a)<sup>22</sup> and (b), we observe opposite trends if we increase the fraction of the minority group in the training data with the male being the minority and the female being the minority. The norm of covariance of the male and female group in the feature space is 5.1833 and 4.9716, respectively. This is consistent with Insight (P4). Due to space limit, our results on the CIFAR10 dataset are deferred to Section A in the supplementary material.

<sup>22</sup>In Figure 4.8(a), when the minority fraction is less than 0.01, the minority group distribution is almost removed from the Gaussian mixture model. Then the  $O(1)$  constants in the last column of Table 4.1 have some minor changes, and the order-wise analyses do not reflect the minor fluctuations in this regime.



**Figure 4.8:** The test accuracy on CelebA dataset has opposite trends when the minority group fraction increases. (a) Male group is the minority (b) Female group is the minority. ©2024 IEEE.

## 4.6 Conclusions and Future Directions

This paper provides a novel theoretical framework for characterizing neural network generalization with group imbalance. The group imbalance is formulated using the Gaussian mixture model. This paper explicitly quantifies the impact of each group on the sample complexity, convergence rate, and the average and the group-level generalization. The learning performance is enhanced when the group-level covariance is at a medium regime, and the group-level mean is close to zero. Moreover, increasing the fraction of minority group does not guarantee improved group-level generalization.

One future direction is to extend the analysis to multiple-hidden-layer neural networks and multi-class classification. Because of the concatenation of nonlinear activation functions, the analysis of the landscape of the empirical risk and the design of a proper initialization is more challenging and requires the development of new tools. Another future direction is to analyze other robust training methods, such as DRO. We see no ethical or immediate negative societal consequence of our work.

# CHAPTER 5

## GENERALIZATION GUARANTEE OF TRAINING GRAPH CONVOLUTIONAL NETWORKS WITH GRAPH TOPOLOGY SAMPLING

### 5.1 Introduction

Graph convolutional neural networks (GCNs) aggregate the embedding of each node with the embedding of its neighboring nodes in each layer. GCNs can model graph-structured data more accurately and compactly than conventional neural networks and have demonstrated great empirical advantage in text analysis [31], [32], [175], [34], computer vision [176], [177], [178], recommendation systems [179], [180], physical reasoning [181], [182], and biological science [183]. Such empirical success is often achieved at a cost of higher computational and memory costs, especially for large graphs, because the embedding of one node depends recursively on the neighbors. To alleviate the exponential increase of computational cost in training deep GCNs, various graph topology sampling methods have been proposed to only aggregate the embeddings of a selected subset of neighbors in training GCNs. Node-wise neighbor-sampling methods such as GraphSAGE [31], VRGCN [184], and Cluster-GCN [185] sample a subset of neighbors for each node. Layer-wise importance sampling methods such as FastGCN [186] and LADIES [187] sample a fixed number of nodes for each layer based on the estimate of node importance. Another line of works such as [188], [189], [190] employ graph sparsification or pruning to reduce the computational and memory cost. Surprisingly, these sampling methods often have comparable or even better testing performance compared to training with the original graph in many empirical studies [186], [190].

In contrast to the empirical success, the theoretical foundation of training GCNs with graph sampling is much less investigated. Only [35] analyzes the convergence rate of graph sampling, but no generalization analysis is provided. One fundamental question about training GCNs is still vastly open, which is:

*Under what conditions does a GCN learned with graph topology sampling achieve satisfactory generalization?*

---

Portions of this chapter have previously appeared as: H. Li, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, “Generalization guarantee of training graph convolutional networks with graph topology sampling,” in *Proc. Int. Conf. Mach. Learn.*, Jul. 2022, pp. 13014–13051.

**Our contributions:** To the best of our knowledge, this chapter provides the first generalization analysis of training GCNs with graph topology sampling. We focus on semi-supervised node classification problems where, with all node features and partial node labels, the objective is to predict unknown node labels. We summarize our contributions from the following dimensions.

*First*, this chapter proposes a training framework that implements both stochastic gradient descent (SGD) and graph topology sampling, and the learned GCN model with Rectified Linear Unit (ReLU) activation is guaranteed to approach the best generalization performance of a large class of target functions. Moreover, as the number of labeled nodes and the number of neurons increase, the class of target function enlarges, indicating improved generalization.

*Second*, this chapter explicitly characterizes the impact of graph topology sampling on the generalization performance through the proposed *effective adjacency matrix*  $\mathbf{A}^*$  of a directed graph that models the node correlations.  $\mathbf{A}^*$  depends on both the given normalized graph adjacency matrix in GCNs and the graph sampling strategy. We provide the general insights that (1) if a node is sampled with a low frequency, its impact on other nodes is reduced in  $\mathbf{A}^*$  compared with  $\mathbf{A}$ ; (2) graph sampling on a highly-unbalanced  $\mathbf{A}$ , where some nodes have a dominating impact in the graph, results in a more balanced  $\mathbf{A}^*$ . Moreover, these insights apply to other graph sampling methods such as FastGCN [186].

We show that learning with topology sampling has the same generalization performance as training GCNs using  $\mathbf{A}^*$ . Therefore, a satisfactory generalization can still be achieved even when the number of sampled nodes is small, provided that the resulting  $\mathbf{A}^*$  still characterizes the data correlations properly. This is the first theoretical explanation of the empirical success of graph topology sampling.

*Third*, this chapter shows that the required number of labeled nodes, referred to as the sample complexity, is a polynomial of  $\|\mathbf{A}^*\|_\infty$  and the maximum node degree, where  $\|\cdot\|_\infty$  measures the maximum absolute row sum. Moreover, our sample complexity is only logarithmic in the number of neurons  $m$  and consistent with the practical over-parameterization of GCNs, in contrast to the loose bound of  $\text{poly}(m)$  in [74] in the restrictive setting of two-layer (one-hidden-layer) GCNs without graph topology sampling.

### 5.1.1 Related Works

**Generalization analyses of GCNs without graph sampling.** Some recent works analyze GCNs trained on the original graph. [191], [35] characterize the expressive power of GCNs. [192] analyzes the convergence of gradient descent in training linear GCNs. [193], [108], [109], [110] characterize the generalization gap, which is the difference between the training error and testing error, through Rademacher complexity. [106], [35], [107] analyze the generalization gap of training GCNs using SGD via the notation of algorithmic stability.

To analyze the training error and generalization performance simultaneously, [194] uses the neural tangent kernel (NTK) approach, where the neural network width is infinite and the step size is infinitesimal, shows that the training error is zero, and characterizes the generalization bound. [74] proves that gradient descent can learn a model with zero population risk, provided that all data are generated by an unknown target model. The result in [74] is limited to two-layer GCNs and requires a proper initialization in the local convex region of the optimal solution.

**Generalization analyses of feed-forward neural networks.** The NTK approach was first developed to analyze fully connected neural networks (FCNNs), see, e.g., [77]. The works of [71], [72], [76] analyze one-hidden-layer neural networks with Gaussian input data. [195] analyzes multi-layer FCNNs but focuses on training the last layer only, while the changes in the hidden layers are negligible. [58] provides the optimization and generalization of three-layer FCNNs. Our proof framework is built upon [58] but makes two important technical contributions. First, this chapter provides the first generalization analysis of graph topology sampling in training GCNs, while [58] considers FCNNs with neither graph topology nor graph sampling. Second, [58] considers i.i.d. training samples, while this chapter considers semi-supervised GCNs where the training data are correlated through graph convolution.

### 5.1.2 Notations

Vectors are in bold lowercase, matrices and tensors in are bold uppercase. Scalars are in normal fonts. For instance,  $\mathbf{Z}$  is a matrix, and  $\mathbf{z}$  is a vector.  $z_i$  denotes the  $i$ -th entry of  $\mathbf{z}$ , and  $Z_{i,j}$  denotes the  $(i, j)$ -th entry of  $\mathbf{Z}$ .  $[K]$  ( $K > 0$ ) denotes the set including integers from 1 to  $K$ .  $\mathbf{I}_d \in \mathbb{R}^{d \times d}$  and  $\mathbf{e}_i$  represent the identity matrix in  $\mathbb{R}^{d \times d}$  and the  $i$ -th standard basis

vector, respectively. We denote the column  $\ell_p$  norm for  $\mathbf{W} \in \mathbb{R}^{d \times N}$  (for  $p \geq 1$ ) as

$$\|\mathbf{W}\|_{2,p} = \left( \sum_{i \in [m]} \|\mathbf{w}_i\|_2^p \right)^{\frac{1}{p}} \quad (5.1)$$

Hence,  $\|\mathbf{W}\|_{2,2} = \|\mathbf{W}\|_F$  is the Frobenius norm of  $\mathbf{W}$ . We use  $\mathbf{w}_i$  ( $\tilde{\mathbf{w}}_i$ ) to denote the  $i$ -th column (row) vector of  $\mathbf{W}$ . We follow the convention that  $f(x) = O(g(x))$  (or  $\Omega(g(x))$ ,  $\Theta(g(x))$ ) means that  $f(x)$  increases at most (or at least, or in the same, respectively,) order of  $g(x)$ . With high probability (w.h.p.) means with probability  $1 - e^{-c \log^2(m_1, m_2)}$  for a sufficient large constant  $c$  where  $m_1$  and  $m_2$  are the number of neurons in the two hidden layers.

**Function complexity.** For any smooth function  $\phi(z)$  with its power series representation as  $\phi(z) = \sum_{i=0}^{\infty} c_i z^i$ , define two useful parameters as follows,

$$\mathcal{C}_{\epsilon}(\phi, R) = \sum_{i=0}^{\infty} \left( (C^* R)^i + \left( \frac{\sqrt{\log(1/\epsilon)}}{\sqrt{i}} C^* R \right)^i \right) |c_i| \quad (5.2)$$

$$\mathcal{C}_s(\phi, R) = C^* \sum_{i=0}^{\infty} (i+1)^{1.75} R^i |c_i| \quad (5.3)$$

where  $R \geq 0$  and  $C^*$  is a sufficiently large constant. These two quantities are used in the model complexity and sample complexity, which represent the required number of model parameters and training samples to learn  $\phi$  up to  $\epsilon$  error, respectively. Many population functions have bounded complexity. For instance, if  $\phi(z)$  is  $\exp(z)$ ,  $\sin(z)$ ,  $\cos(z)$  or polynomials of  $z$ , then  $\mathcal{C}_{\epsilon}(\phi, O(1)) \leq O(\text{poly}(1/\epsilon))$  and  $\mathcal{C}_s(\phi, O(1)) \leq O(1)$ .

The main notations are summarized in Table D.1 in Appendix.

## 5.2 Training GCNs with Topology Sampling: Formulation and Main Components

**GCN setup.** Let  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$  denote an un-directed graph, where  $\mathcal{V}$  is the set of nodes with size  $|\mathcal{V}| = N$  and  $\mathcal{E}$  is the set of edges. Let  $\tilde{\mathbf{A}} \in \{0, 1\}^{N \times N}$  be the adjacency matrix of  $\mathcal{G}$  with added self-connections. Let  $\mathbf{D}$  be the degree matrix with diagonal elements  $D_{i,i} = \sum_j \tilde{A}_{i,j}$  and zero entries otherwise.  $\mathbf{A}$  denotes the normalized adjacency matrix with  $\mathbf{A} = \mathbf{D}^{-\frac{1}{2}} \tilde{\mathbf{A}} \mathbf{D}^{-\frac{1}{2}}$ . Let  $\mathbf{X} \in \mathbb{R}^{N \times d}$  denote the matrix of the features of  $N$  nodes, where the  $n$ -th row of  $\mathbf{X}$ , denoted by  $\tilde{\mathbf{x}}_n \in \mathbb{R}^{1 \times d}$ , represents the feature of node  $n$ . Assume  $\|\tilde{\mathbf{x}}_n\| = 1$

for all  $n$  without loss of generality.  $y_n \in \mathcal{Y}$  represents the label of node  $n$ , where  $\mathcal{Y}$  is a set of all labels.  $y_n$  depends on not only  $\mathbf{x}_n$  but the neighbors. Let  $\Omega \subset \mathcal{V}$  denote the set of labeled nodes. Given  $\mathbf{X}$  and labels in  $\Omega$ , the objective of semi-supervised node-classification is to predict the unknown labels in  $\mathcal{V}/\Omega$ .

**Learner network** We consider the setting of training a three-layer GCN  $F : \mathbb{R}^N \times \mathbb{R}^{N \times d} \rightarrow \mathbb{R}^{1 \times K}$  with

$$\begin{aligned} F_{\mathbf{A}}(\mathbf{e}_g, \mathbf{X}; \mathbf{W}, \mathbf{V}) &= \mathbf{e}_g^\top \mathbf{A} \sigma(\mathbf{r} + \mathbf{B}_2) \mathbf{C} \quad \text{and} \\ \mathbf{r} &= \mathbf{A} \sigma(\mathbf{A} \mathbf{X} \mathbf{W} + \mathbf{B}_1) \mathbf{V} \end{aligned} \tag{5.4}$$

where  $\sigma(x) = \max(x, 0)$  is the ReLU activation function,  $\mathbf{W} \in \mathbb{R}^{d \times m_1}$  and  $\mathbf{V} \in \mathbb{R}^{m_1 \times m_2}$  represent the weights of  $m_1$  and  $m_2$  hidden nodes in the first and second layer, respectively.  $\mathbf{B}_1 \in \mathbb{R}^{N \times m_1}$  and  $\mathbf{B}_2 \in \mathbb{R}^{m_1 \times m_2}$  represent the bias matrices.  $\mathbf{C} \in \mathbb{R}^{m \times K}$  is the output weight vector.  $\mathbf{e}_g \in \mathbb{R}^N$  belongs to  $\{\mathbf{e}_i\}_{i=1}^N$  and selects the index of the node label. We write  $F$  as  $F_{\mathbf{A}}(\mathbf{e}_g, \mathbf{X}; \mathbf{W}, \mathbf{V})$ , because we only update  $\mathbf{W}$  and  $\mathbf{V}$  in training, and  $\mathbf{A}$  represents the graph topology. Note that in conventional GCNs such as [32],  $\mathbf{C}$  is a learnable parameter, and  $\mathbf{B}_1$  and  $\mathbf{B}_2$  can be zero. Here for the analytical purpose, we consider a slightly different model where  $\mathbf{C}$ ,  $\mathbf{B}_1$  and  $\mathbf{B}_2$  are fixed as randomly selected values.

Consider a loss function  $L : \mathbb{R}^{1 \times k} \times \mathcal{Y} \rightarrow \mathbb{R}$  such that for every  $y \in \mathcal{Y}$ , the function  $L(\cdot, y)$  is nonnegative, convex, 1-Lipschitz continuous and 1-Lipschitz smooth and  $L(0, y) \in [0, 1]$ . This includes both the cross-entropy loss and the  $\ell_2$ -regression loss (for bounded  $\mathcal{Y}$ ). The learning problem solves the following empirical risk minimization problem:

$$\min_{\mathbf{W}, \mathbf{V}} L_\Omega(\mathbf{W}, \mathbf{V}) = \frac{1}{|\Omega|} \sum_{i \in \Omega} L(F_{\mathbf{A}}(\mathbf{e}_i, \mathbf{X}; \mathbf{W}, \mathbf{V}), y^i) \tag{5.5}$$

where  $L_\Omega$  is the empirical risk of the labeled nodes in  $\Omega$ . The trained weights are used to estimate the unknown labels on  $\mathcal{V}/\Omega$ . Note that the results in this chapter are *distribution-free*, and no assumption is made on the distributions of  $\tilde{x}_n$  and  $y_n$ .

**Training with SGD.** In practice, (5.5) is often solved by gradient type of methods, where in iteration  $t$ , the currently estimations are updated by subtracting the product of a positive step size and the gradient of  $L_\Omega$  evaluated at the current estimate. To reduce the computational complexity in estimating the gradient, an SGD method is often employed to compute the gradient of the risk of a randomly selected subset of  $\Omega$  rather than using the

whole set  $\Omega$ .

However, due to the recursive embedding of neighboring features in GCNs, see the concatenations of  $\mathbf{A}$  in (5.4), the computation and memory cost of computing the gradient can be high. Thus, graph topology sampling methods have been proposed to further reduce the computational cost.

**Graph topology sampling.** A node sampling method randomly removes a subset of nodes and the incident edges from  $\mathcal{G}$  in each iteration independently, and the embedding aggregation is based on the reduced graph. Mathematically, in iteration  $s$ , replace  $\mathbf{A}$  in (5.4) with<sup>23</sup>  $\mathbf{A}^s = \mathbf{A}\mathbf{P}^s$ , where  $\mathbf{P}^s$  is a diagonal matrix, and the  $i$ th diagonal entry is 0, if node  $i$  is removed in iteration  $s$ . The non-zero diagonal entries of  $\mathbf{P}^s$  are selected differently based on different sampling methods. Because  $\mathbf{A}^s$  is much more sparse than  $\mathbf{A}$ , the computation and memory cost of embedding neighboring features is significantly reduced.

This chapter will analyze the generalization performance, i.e., the prediction accuracy of unknown labels, of our algorithm framework that implements both SGD and graph topology sampling to solve (5.5). The details of our algorithm are discussed in Section 5.3.2-5.3.3, and the generalization performance is presented in Section 5.3.4.

## 5.3 Main Algorithmic and Theoretical Results

### 5.3.1 Informal Key Theoretical Findings

We first summarize the main insights of our results before presenting them formally.

**1. A provable generalization guarantee of GCNs beyond two layers and with graph topology sampling.** The learned GCN by our Algorithm 2 can approach the best performance of label prediction using a large class of target functions. Moreover, the prediction performance improves when the number of labeled nodes and the number of neurons  $m_1$  and  $m_2$  increase. This is the first generalization performance guarantee of training GCNs with graph topology sampling.

**2. The explicit characterization of the impact of graph sampling through the effective adjacency matrix  $\mathbf{A}^*$ .** We show that training with graph sampling returns a

---

<sup>23</sup>Here we use the same sampled matrix  $\mathbf{A}^s$  in all three layers in (5.4) to simplify the representation. Our analysis applies to the more general setting that each layer uses a different sampled adjacency matrix, i.e., the three  $\mathbf{A}$  matrices in (5.4) are replaced with  $\mathbf{A}^{s(1)} = \mathbf{A}\mathbf{P}^{s(1)}$ ,  $\mathbf{A}^{s(2)} = \mathbf{A}\mathbf{P}^{s(2)}$ ,  $\mathbf{A}^{s(3)} = \mathbf{A}\mathbf{P}^{s(3)}$ , respectively, as in [187], [196], where  $\mathbf{P}^{s(1)}$ ,  $\mathbf{P}^{s(2)}$ , and  $\mathbf{P}^{s(3)}$  are independently sampled following the same sampling strategy.

model that has the same label prediction performance as that of a model trained by replacing  $\mathbf{A}$  with  $\mathbf{A}^*$  in (5.4), where  $\mathbf{A}^*$  depends on both  $\mathbf{A}$  and the graph sampling strategy. As long as  $\mathbf{A}^*$  can characterize the correlation among nodes properly, the learned GCN maintains a desirable prediction performance. This explains the empirical success of graph topology sampling in many datasets.

**3. The explicit sample complexity bound on graph properties.** We provide explicit bounds on the sample complexity and the required number of neurons, both of which grow as the node correlation increase. Moreover, the sample complexity depends on the number of neurons only logarithmically, which is consistent with the practical over-parameterization. To the best of our knowledge, [74] is the only existing work that provides a sample complexity bound based on the graph topology, but in the non-practical and restrictive setting of two-layer GCNs. Moreover, the sample complexity bound by [74] is polynomial in the number of neurons.

**4. Tackling the non-convex interaction of weights between different layers.** The convexity plays a critical role in many exiting analyses of GCNs. For instance, the analyses in [74] require a special initialization in the local convex region of the global minimum, and the results only apply to two-layer GCNs. The NTK approach in [194] considers the limiting case that the interactions across layers are negligible. Here, we directly address the non-convex interaction of weights  $\mathbf{W}$  and  $\mathbf{V}$  in both algorithmic design and theoretical analyses.

### 5.3.2 Graph Topology Sampling Strategy

Here we describe our graph topology sampling strategy using  $\mathbf{A}^s$ , which we randomly generate to replace  $\mathbf{A}$  in the  $s$ th SGD iteration. Although our method is motivated for analysis and different from the existing graph sampling strategies, our insights generalize to other sampling methods like FastGCN [186]. The outline of our algorithmic framework of training GCNs with graph sampling is deferred to Section 5.3.3.

Suppose the node degrees in  $\mathcal{G}$  can be divided into  $L$  groups with  $L \geq 1$ , where the degrees of nodes in group  $l$  are in the order of  $d_l$ , i.e., between  $cd_l$  and  $Cd_l$  for some constants  $c \leq C$ , and  $d_l$  is order-wise smaller than  $d_{l+1}$ , i.e.,  $d_l = o(d_{l+1})$ . Let  $N_l$  denote the number of nodes in group  $l$ .

**Graph sampling strategy**<sup>24</sup>. We consider a group-wise uniform sampling strategy, where  $S_l$  out of  $N_l$  nodes are sampled uniformly from each group  $l$ . For all unsampled nodes, we set the corresponding diagonal entries of a diagonal matrix  $\mathbf{P}^s$  to be zero. If node  $i$  is sampled in this iteration and belongs to group  $l$  for any  $i$  and  $l$ , the  $i$ th diagonal entry of  $\mathbf{P}^s$  is set as  $p_l^* N_l / S_l$  for some non-negative constant  $p_l^*$ . Then  $\mathbf{A}^s = \mathbf{A}\mathbf{P}^s$ .  $N_l / S_l$  can be viewed as the scaling to compensate for the unsampled nodes in group  $l$ .  $p_l^*$  can be viewed as the scaling to reflect the impact of sampling on nodes with different importance that will be discussed in detail soon.

**Effective adjacency matrix  $\mathbf{A}^*$  by graph sampling.** To analyze the impact of graph topology sampling on the learning performance, we define the effective adjacency matrix as follows:

$$\mathbf{A}^* = \mathbf{A}\mathbf{P}^* \quad (5.6)$$

where  $\mathbf{P}^*$  is a diagonal matrix defined as

$$\mathbf{P}_{ii}^* = p_l^* \quad \text{if node } i \text{ belongs to degree group } l \quad (5.7)$$

Therefore, compared with  $\mathbf{A}$ , all the columns with indices corresponding to group  $l$  are scaled by a factor of  $p_l^*$ . We will formally analyze the impact of graph topology sampling on the generalization performance in Section 5.3.4, but an intuitive understanding is that our graph sampling strategy effectively changes the normalized adjacency matrix  $\mathbf{A}$  in the GCN network model (5.4) to  $\mathbf{A}^*$ .

$\mathbf{A}^*$  can be viewed as an adjacency matrix of a weighted directed graph  $\mathcal{G}'$  that reflects the node correlations, where each un-directed edge in  $\mathcal{G}$  corresponds to two directed edges in  $\mathcal{G}'$  with possibly different weights.  $\mathbf{A}_{ji}^*$  measures the impact of the feature of node  $i$  on the label of node  $j$ . If  $p_l^*$  is in the range of  $(0, 1)$ , the corresponding entries of columns with indices in group  $l$  in  $\mathbf{A}^*$  are smaller than those in  $\mathbf{A}$ . That means the impact of a node in group  $l$  on all other nodes is reduced from those in  $\mathbf{A}$ . Conversely, if  $p_l^* > 1$ , then the impact of nodes in group  $l$  in  $\mathbf{A}^*$  is enhanced from that in  $\mathbf{A}$ .

### Parameter selection and insights

---

<sup>24</sup>Here we discuss asymmetric sampling as a general case. The special case of symmetric sampling is introduced in Section D.1.1

(1) The scaling factor  $p_l^*$  should satisfy

$$0 \leq p_l^* \leq \frac{c_1}{L\psi_l}, \quad \forall l \quad (5.8)$$

for a positive constant  $c_1$  that can be sufficiently large.  $\psi_l$  is defined as follows,

$$\psi_l := \frac{\sqrt{d_L d_l} \mathbf{N}_l}{\sum_{i=1}^L d_i \mathbf{N}_i} \quad \forall l \in [L] \quad (5.9)$$

Note that (5.8) is a minor requirement for most graphs. To see this, suppose  $L$  is a constant, and every  $N_l$  is in the order of  $N$ . Then  $\psi_l$  is less than  $O(1)$  for all  $l$ . Thus, all constant values of  $p_l^*$  satisfy (5.8) with  $\psi_l$  from (5.9). A special example is that  $p_l^*$  are all equal, i.e.,  $\mathbf{A}^* = c_2 \mathbf{A}$  for some constant  $c_2$ . Because one can scale  $\mathbf{W}$  and  $\mathbf{V}$  by  $1/c_2$  in (5.4) without changing the results,  $\mathbf{A}^*$  is equivalent to  $\mathbf{A}$  in this case.

The upper bound in (5.9) only becomes active in highly unbalanced graphs where there exists a dominating group  $\hat{l}$  such that  $\sqrt{d_{\hat{l}}} N_{\hat{l}} \gg \sqrt{d_l} N_l$  for all other  $l$ . Then the upper bound of  $p_{\hat{l}}^*$  is much smaller than those for other  $p_l^*$ . Therefore, the columns of  $\mathbf{A}^*$  that correspond to group  $\hat{l}$  are scaled down more significantly than other columns, indicating that the impact of group  $\hat{l}$  is reduced more significantly than other groups in  $\mathbf{A}^*$ . Therefore, the takeaway is that **graph topology sampling reduces the impact of dominating nodes more than other nodes, resulting in a more balanced  $\mathbf{A}^*$  compared with  $\mathbf{A}$** .

(2) The number of sampled nodes shall satisfy

$$\frac{S_l}{N_l} \geq (1 + \frac{c_1 \text{poly}(\epsilon)}{L p_l^* \psi_l})^{-1} \quad \forall l \in [L] \quad (5.10)$$

where  $\epsilon$  is a small positive value. The sampling requirement in (5.10) has two takeaways. **First, the higher-degree groups shall be sampled more frequently than lower-degree groups.** To see this, consider a special case that  $p_l^* = 1$ , and  $N_l = N/L$  for all  $l$ . Then (5.10) indicates that  $S_l$  is larger in a group  $l$  with a larger  $d_l$ . This intuition is the same as FastGCN [186], which also samples high-degree nodes with a higher probability in many cases. Therefore, the insights from our graph sampling method also apply to other sampling methods such as FastGCN. We will show the connection to FastGCN empirically in Section 5.4.2. **Second, reducing the number of samples in group  $l$  corresponds to reducing the impact of group  $l$  in  $\mathbf{A}^*$ .** To see this, note that decreasing  $p_l^*$  reduces the right-hand

---

**Algorithm 2** Training with SGD and graph topology sampling

---

1: **Input:** Normalized adjacency matrix  $\mathbf{A}$ , node features  $\mathbf{X}$ , known node labels in  $\Omega$ , the step size  $\eta$ , the number of inner iterations  $T_w$ , the number of outer iterations  $T$ ,  $\sigma_w$ ,  $\sigma_v$ ,  $\lambda_w$ ,  $\lambda_v$ .  
2: Initialize  $\mathbf{W}^{(0)}$ ,  $\mathbf{V}^{(0)}$ ,  $\mathbf{B}_1$ ,  $\mathbf{B}_2$ ,  $\mathbf{C}$ .  
3:  $\mathbf{W}_0 = 0$ ,  $\mathbf{V}_0 = 0$ .  
4: **for**  $t = 0, 1, \dots, T - 1$  **do**  
5:   Apply noisy SGD with step size  $\eta$  on the stochastic objective  $\hat{L}_\Omega(\lambda_t; \mathbf{W}, \mathbf{V})$  in (5.11) for  $T_w$  steps. To generate the stochastic objective in each step  $s$ , randomly sample a batch of labeled nodes  $\Omega^s$  from  $\Omega$ ; generate  $\mathbf{A}^s$  using graph sampling; randomly generate  $\mathbf{W}^\rho$ ,  $\mathbf{V}^\rho$  and  $\Sigma$ .  
   Let the starting point be  $\mathbf{W} = \mathbf{W}_t$ ,  $\mathbf{V} = \mathbf{V}_t$  and suppose it reaches  $\mathbf{W}_{t+1}$  and  $\mathbf{V}_{t+1}$ .  
6:    $\lambda_{t+1} = \lambda_t \cdot (1 - \eta)$ .  
7: **end for**  
8: **Output:**  

$$\mathbf{W}^{(out)} = \sqrt{\lambda_{T-1}}(\mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma)$$
  

$$\mathbf{V}^{(out)} = \sqrt{\lambda_{T-1}}(\mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T).$$


---

side of (5.10).

### 5.3.3 The Algorithmic Framework of Training GCNs

Because (5.5) is non-convex, solving it directly using SGD can get stuck at a bad local minimum in theory. The main idea in the theoretical analysis to address this non-convexity is to add weight decay and regularization in the objective of (5.5) such that with a proper regularization, any second-order critical point is *almost* a global minimum.

Specifically, for initialization, entries of  $\mathbf{W}^{(0)}$  are i.i.d. from  $\mathcal{N}(0, \frac{1}{m_1})$ , and entries of  $\mathbf{V}^{(0)}$  are i.i.d. from  $\mathcal{N}(0, \frac{1}{m_2})$ .  $\mathbf{B}_1$  (or  $\mathbf{B}_2$ ) is initialized to be an all-one vector multiplying a row vector with i.i.d. samples from  $\mathcal{N}(0, \frac{1}{m_1})$  (or  $\mathcal{N}(0, \frac{1}{m_2})$ ). Entries of  $\mathbf{C}$  are drawn i.i.d. from  $\mathcal{N}(0, 1)$ .

In each outer loop  $t = 0, \dots, T - 1$ , we use noisy SGD<sup>25</sup> with step size  $\eta$  for  $T_w$  iterations to minimize the stochastic objective function  $\hat{L}_\Omega$  in (5.11) with some fixed  $\lambda_{t-1}$ , where  $\lambda_0 = 1$ ,

---

<sup>25</sup>Noisy SGD is vanilla SGD plus Gaussian perturbation. It is a common trick in the theoretical analyses of non-convex optimization [197] and is not needed in practice.

and the weight decays with  $\lambda_{t+1} = (1 - \eta)\lambda_t$ .

$$\begin{aligned}\hat{L}_\Omega(\lambda_t; \mathbf{W}, \mathbf{V}) \\ = & L_\Omega(\sqrt{\lambda_t}(\mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}\Sigma), \sqrt{\lambda_t}(\mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma\mathbf{V})) \\ & + \lambda_w \|\sqrt{\lambda_t}\mathbf{W}\|_{2,4}^4 + \lambda_v \|\sqrt{\lambda_t}\mathbf{V}\|_F^2\end{aligned}\quad (5.11)$$

$\hat{L}_\Omega(\lambda_t; \mathbf{W}, \mathbf{V})$  is stochastic because in each inner iteration  $s$ , (1) we randomly sample a subset  $\Omega^s$  of labeled nodes; (2) we randomly sample  $\mathbf{A}^s$  from the graph topology sampling method in Section 5.3.2; (3)  $\mathbf{W}^\rho$  and  $\mathbf{V}^\rho$  are small perturbation matrices with entries i.i.d. drawn from  $\mathcal{N}(0, \sigma_w^2)$  and  $\mathcal{N}(0, \sigma_v^2)$ , respectively; and (4)  $\Sigma \in \mathbb{R}^{m_1 \times m_1}$  is a random diagonal matrix with diagonal entries uniformly drawn from  $\{1, -1\}$ .  $\mathbf{W}^\rho$  and  $\mathbf{V}^\rho$  are standard Gaussian smoothing in the literature of theoretical analyses of non-convex optimization, see, e.g. [197], and are not needed in practice.  $\Sigma$  is similar to the practical Dropout [198] technique that randomly masks out neurons and is also introduced for the theoretical analysis only.

The last two terms in (5.11) are additional regularization terms for some positive  $\lambda_w$  and  $\lambda_v$ . As shown in [58],  $\|\cdot\|_{2,4}$  is used for the analysis to drive the weights to be evenly distributed among neurons. The practical regularization  $\|\cdot\|_F$  has the same effect in empirical results, while the theoretical justification is open.

Algorithm 2 summarizes the algorithm with the parameter selections in Table 5.1. Let  $\mathbf{W}^{out}$  and  $\mathbf{V}^{out}$  denote the returned weights. We use  $F_{\mathbf{A}^*}(\mathbf{e}_i, \mathbf{X}; \mathbf{W}^{out}, \mathbf{V}^{out})$  to predict the label of node  $i$ . This might sound different from the conventional practice which uses  $\mathbf{A}$  in predicting unknown labels. However, note that  $\mathbf{A}^*$  only differs from  $\mathbf{A}$  by a column-wise scaling as from (5.6). Moreover,  $\mathbf{A}^*$  can be set as  $\mathbf{A}$  in many practical datasets based on our discussion after (5.9). Here we use the general form of  $\mathbf{A}^*$  for the purpose of analysis.

We remark that our framework of algorithm and analysis can be easily applied to the simplified setup of two-layer GCNs. The resulting algorithm is much simplified to a vanilla SGD plus graph topology sampling. All the additional components above are introduced to address the non-convex interaction of  $\mathbf{W}$  and  $\mathbf{V}$  theoretically and may not be needed for practical implementation. We skip the discussion of two-layer GCNs in this chapter.

### 5.3.4 Generalization Guarantee

Our formal generalization analysis shows that our learning method returns a GCN model that approaches the minimum prediction error that can be achieved by the best

**Table 5.1: Parameter choices for Algorithm 2.**

$\lambda_v$	$2\epsilon_0 m_2/m_1^{1-0.01}$	$\sigma_v$	$1/m_2^{1/2+0.01}$
$\lambda_w$	$2\epsilon_0 m_1^{3-0.002}/C_0^4$	$\sigma_w$	$1/m_1^{1-0.01}$
$C$	$C_\epsilon(\phi, \ \mathbf{A}^*\ _\infty) \sqrt{\ \mathbf{A}^*\ _\infty^2 + 1}$	$C'$	$10C\sqrt{p_2}$
$C''$	$C_\epsilon(\Phi, C') \sqrt{\ \mathbf{A}^*\ _\infty^2 + 1}$	$C_0$	$\tilde{O}(p_1^2 p_2 K^2 C C'')$

function in a large concept class of target functions, which have two important properties: (1) the prediction error decreases as size of the function class increases; and (2) the concept class uses  $\mathbf{A}^*$  in (5.6) as the adjacency matrix of the graph topology. Therefore, the result implies that if  $\mathbf{A}^*$  accurately captures the correlations among node features and labels, the learned GCN model can achieve a small prediction error of unknown labels. Moreover, no other functions in a large concept class can perform better than the learned GCN model. To formalize the results, we first define the target functions as follows.

**Concept class and target function  $F^*$ .** Consider a concept class consisting of target functions  $F^* : \mathbb{R}^N \times \mathbb{R}^{N \times d} \rightarrow \mathbb{R}^{1 \times K}$ :

$$\begin{aligned} F_{\mathbf{A}^*}^*(\mathbf{e}_g, \mathbf{X}) &= \mathbf{e}_g^\top \mathbf{A}^* (\Phi(\mathbf{r}_1) \odot \mathbf{r}_2) \mathbf{C}^*, \\ \text{where } \mathbf{r}_1 &= \mathbf{A}^* \phi_1(\mathbf{A}^* \mathbf{X} \mathbf{W}_1^*) \mathbf{V}_1^*, \\ \mathbf{r}_2 &= \mathbf{A}^* \phi_2(\mathbf{A}^* \mathbf{X} \mathbf{W}_2^*) \mathbf{V}_2^*, \end{aligned} \tag{5.12}$$

where  $\phi_1, \phi_2, \Phi: \mathbb{R} \rightarrow \mathbb{R}$  all infinite-order smooth<sup>26</sup>. The parameters  $\mathbf{W}_1^*, \mathbf{W}_2^* \in \mathbb{R}^{d \times p_2}$ ,  $\mathbf{V}_1^*, \mathbf{V}_2^* \in \mathbb{R}^{p_2 \times p_1}$ ,  $\mathbf{C}^* \in \mathbb{R}^{p_1 \times k}$  satisfy that every column of  $\mathbf{W}_1^*, \mathbf{W}_2^*, \mathbf{V}_1^*, \mathbf{V}_2^*$  is unit norm, and the maximum absolute value of  $\mathbf{C}^*$  is at most 1. The effective adjacency matrix  $\mathbf{A}^*$  is defined in (5.6). Define

$$\mathcal{C}_\epsilon(\phi, R) = \max(\mathcal{C}_\epsilon(\phi_1, R), \mathcal{C}_\epsilon(\phi_2, R)), \tag{5.13}$$

$$\mathcal{C}_s(\phi, R) = \max(\mathcal{C}_s(\phi_1, R), \mathcal{C}_s(\phi_2, R)). \tag{5.14}$$

We focus on target functions where the function complexity  $\mathcal{C}_\epsilon(\Phi, R)$ ,  $\mathcal{C}_s(\Phi, R)$ ,  $\mathcal{C}_\epsilon(\phi, R)$ ,  $\mathcal{C}_s(\phi, R)$ , defined in (5.2)-(5.3), (5.13)-(5.14), as well as  $p_1$  and  $p_2$ , are all bounded.

<sup>26</sup>When  $\Phi$  is operated on a matrix  $\mathbf{r}_1$ ,  $\Phi(\mathbf{r}_1)$  means applying  $\Phi$  on each entry of  $\mathbf{r}_1$ . In fact, our results still hold for a more general case that a different function  $\Phi_j$  is applied to every entry of the  $j$ th column of  $\mathbf{r}_1$ ,  $j \in [p_2]$ . We keep the simpler model to have a more compact representation. The similar arguments hold for  $\phi_1, \phi_2$ .

(5.12) is more general than GCNs. If  $\mathbf{r}_2$  is a constant matrix, (5.12) models a GCN, where  $\mathbf{W}_1^*$  and  $\mathbf{V}_1^*$  are weight matrices in the first and second layer, respectively, and  $\phi_1$  and  $\Phi$  are the activation functions in each layer.

**Modeling the prediction error of unknown labels.** We will show that the learned GCN by our method performs almost the same as the best function in the concept class in (5.12) in predicting unknown labels. Because the practical datasets usually contain noise in features and labels, we employ a probabilistic model to model the data. Note that our result is distribution-free, and the following distributions are introduced for the presentation of the results.

Specifically, let  $\mathcal{D}_{\tilde{x}_n}$  denote the distribution from which the feature  $\tilde{x}_n$  of node  $n$  is drawn. For example, when the noise level is low,  $\mathcal{D}_{\tilde{x}_n}$  can be a distribution centered at the observed feature of node  $n$  with a small variance. Similarly, let  $\mathcal{D}_{y_n}$  denote the distribution from which the label  $y_n$  at node  $n$  is drawn. Let  $\mathbf{e}_g$  be uniformly selected from  $\{\mathbf{e}_i\}_{i=1}^N \in \mathbb{R}^N$ . Let  $\mathcal{D}$  denote the concatenation of these distributions of a data point

$$z = (\mathbf{e}_g, \mathbf{X}, y) \in \mathbb{R}^N \times \mathbb{R}^{N \times d} \times \mathcal{Y}. \quad (5.15)$$

Then the given feature matrix  $\mathbf{X}$  and partial labels in  $\Omega$  can be viewed as  $|\Omega|$  identically distributed but *correlated* samples from  $\mathcal{D}$ . The correlation results from the fact that the label of node  $i$  depends on not only the feature of node  $i$  but also neighboring features. This model of correlated samples is different from the conventional assumption of i.i.d. samples in supervised learning and makes our analyses more involved.

Let

$$\text{OPT}_{\mathbf{A}^*} = \min_{\mathbf{W}_1^*, \mathbf{W}_2^*, \mathbf{V}_1^*, \mathbf{V}_2^*, \mathbf{C}^*} \mathbb{E}_{(\mathbf{e}_g, \mathbf{X}, y) \sim \mathcal{D}} L(F_{\mathbf{A}^*}^*(\mathbf{e}_g, \mathbf{X}), y) \quad (5.16)$$

be the smallest population risk achieved by the best target function (over the choices of  $\mathbf{W}_1^*$ ,  $\mathbf{W}_2^*$ ,  $\mathbf{V}_1^*$ ,  $\mathbf{V}_2^*$ ,  $\mathbf{C}^*$ ) in the concept class  $F_{\mathbf{A}^*}^*$  in (5.12).  $\text{OPT}_{\mathbf{A}^*}$  measures the average loss of predicting the unknown labels if the estimates are computed using the best target function in (5.12). Clearly,  $\text{OPT}_{\mathbf{A}^*}$  decreases as the size of the concept increases, i.e., when  $p_1$  and  $p_2$  increase. Moreover, if  $\mathbf{A}^*$  indeed models the node correlations accurately,  $\text{OPT}_{\mathbf{A}^*}$  can be very small, indicating a desired generalization performance. We next show that the population risk of the learned GCN model by our method can be arbitrarily close to  $\text{OPT}_{\mathbf{A}^*}$ .

**Theorem 5.3.1.** *For every  $\epsilon_0 \in (0, \frac{1}{100}]$ , every  $\epsilon \in (0, (K p_1 p_2^2 \mathcal{C}_s(\Phi, p_2 \mathcal{C}_s(\phi, O(1))) \mathcal{C}_s(\phi, O(1)))$*

$\cdot \|\mathbf{A}^*\|_\infty^2)^{-1}\epsilon_0)$ , as long as

$$\begin{aligned} m_1 &= m_2 = m \\ &\geq \text{poly}\left(\mathcal{C}_\epsilon(\Phi, \mathcal{C}_\epsilon(\phi, O(1))), p_2, \|\mathbf{A}^*\|_\infty, \frac{1}{\epsilon}\right) \end{aligned} \quad (5.17)$$

$$\begin{aligned} |\Omega| &\geq \Theta(\epsilon_0^{-2} \|\mathbf{A}^*\|_\infty^8 K^6 (1 + p_1^4 p_2^5 \mathcal{C}_\epsilon(\Phi, \sqrt{p_2} \mathcal{C}_\epsilon(\phi, O(1)))) \\ &\quad \cdot \mathcal{C}_\epsilon(\phi, O(1)) (\|\mathbf{A}^*\|_\infty + 1)^4) (1 + \delta)^4 \log N \log m), \end{aligned} \quad (5.18)$$

(5.8) and (5.10) hold, there is a choice  $\eta = 1/\text{poly}(\|\mathbf{A}^*\|_\infty, K, m)$  and  $T = \text{poly}(\|\mathbf{A}^*\|_\infty, K, m)$  such that with probability at least 0.99,

$$\begin{aligned} &\mathbb{E}_{(\mathbf{e}_g, \mathbf{X}, y) \in \mathcal{D}} L(F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}; \mathbf{W}^{(out)}, \mathbf{V}^{(out)}), y) \\ &\leq (1 + \epsilon_0) \text{OPT}_{\mathbf{A}^*} + \epsilon_0, \end{aligned} \quad (5.19)$$

where  $\mathbf{A}^*$  is the effective adjacency matrix in (5.12).

Theorem 5.3.1 shows that the required sample complexity is polynomial in  $\|\mathbf{A}^*\|$  and  $\delta$ , where  $\delta$  is the maximum node degree without self-connections in  $\mathbf{A}$ . Note that condition (5.8) implies that  $\|\mathbf{A}^*\|_\infty$  is  $O(1)$ . Then as long as  $\delta$  is  $O(N^\alpha)$  for some small  $\alpha$  in  $(0, 1)$ , say  $\alpha = 1/5$ , then one can accurately infer the unknown labels from a small percentage of labeled nodes. Moreover, our sample complexity is sufficient but not necessary. It is possible to achieve a desirable generalization performance if the number of labeled nodes is less than the bound in (5.18).

Graph topology sampling affects the generalization performance through  $\mathbf{A}^*$ . From the discussion in Section 5.3.2, graph sampling reduces the node correlation in  $\mathbf{A}^*$ , especially for dominating nodes. The generalization performance does not degrade when  $\text{OPT}_{\mathbf{A}^*}$  is small, i.e., the resulting  $\mathbf{A}^*$  is sufficient to characterize the node correlation in a given dataset. That explains the empirical success of graph sampling in many datasets.

## 5.4 Numerical Results

To unveil how our theoretical results are aligned with GCN's generalization performance in experiments, we will focus on numerical evaluations on synthetic data where we can control target functions and compare with  $\mathbf{A}^*$  explicitly. We also evaluate both our graph sampling method and FastGCN [186] to validate that insights for our graph sampling method also

apply to FastGCN.

We generate a graph  $\mathcal{G}$  with  $N = 2000$  nodes.  $\mathcal{G}$  has two degree groups. Group 1 has  $N_1$  nodes, and every node degree approximately equals  $d_1$ . Group 2 has  $N_2$  nodes, and every node degree approximately equals  $d_2$ . The edges between nodes are randomly selected.  $\mathbf{A}$  is the normalized adjacency matrix of  $\mathcal{G}$ .

The node labels are generated by the target function

$$y = (\sin(\hat{\mathbf{A}}\mathbf{X}\mathbf{W}^*) \odot \tanh(\hat{\mathbf{A}}\mathbf{X}\mathbf{W}^*))\mathbf{C}^*, \quad (5.20)$$

where  $\hat{\mathbf{A}} \in \mathbb{R}^{N \times N}$ ,  $\mathbf{X} \in \mathbb{R}^{N \times d}$ ,  $\mathbf{W}^* \in \mathbb{R}^{d \times p}$  and  $\mathbf{C}^* \in \mathbb{R}^{p \times K}$ . The feature dimension  $d = 10$ ,  $p = 10$ , and  $K = 2$ .  $\mathbf{X}$ ,  $\mathbf{W}^*$  and  $\mathbf{C}^*$  are all randomly generated with each entry i.i.d. from  $\mathcal{N}(0, 1)$ .

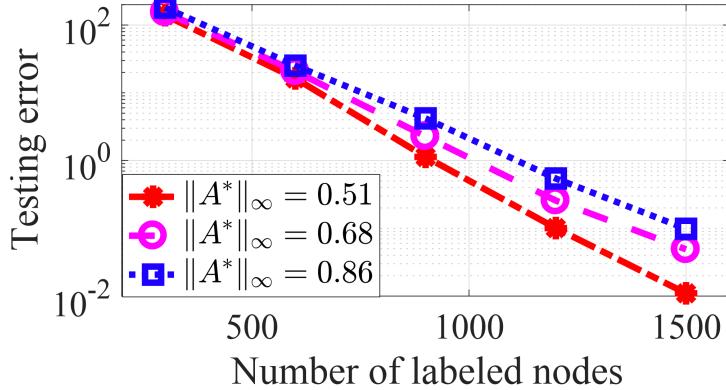
We consider a regression task with the  $\ell_2$ -regression loss function. A three-layer GCN as defined in (5.4) with  $m$  neurons in each hidden layer is trained on a randomly selected set  $\Omega$  of labeled nodes. The rest  $N - |\Omega|$  labels are used for testing. The learning rate  $\eta = 10^{-3}$ . The mini-batch size is 5, and the dropout rate as 0.4. The total number of iterations is  $TT_w = 4|\Omega|$ . Our graph topology sampling method samples  $S_1 = 0.9N_1$  and  $S_2 = 0.9N_2$  nodes for both groups in each iteration.

#### 5.4.1 Sample Complexity and Neural Network Width with Respect to the Effective Adjacency Matrix

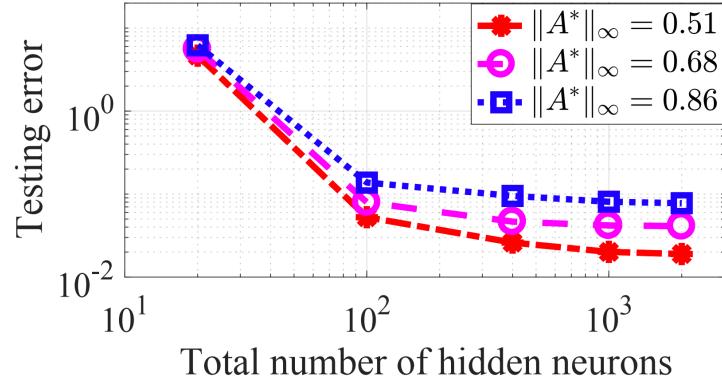
We fix  $N_1 = 100$ ,  $N_2 = 1900$  and vary  $\mathbf{A}$  by changing node degrees  $d_1$  and  $d_2$ . In the graph topology sampling method,  $p_1^* = 0.7$  and  $p_2^* = 0.3$ . For every fixed  $\mathbf{A}$ , the effective adjacency matrix  $\mathbf{A}^*$  is computed based on (5.6) using  $p_1^*$  and  $p_2^*$ . Synthetic labels are generated based on (5.20) using  $\mathbf{A}^*$  as  $\hat{\mathbf{A}}$ .

Figure 5.1 shows the testing error decreases as the number of labeled nodes  $|\Omega|$  increases, when the number of neurons per layer  $m$  is fixed as 500. Moreover, as  $\|\mathbf{A}^*\|_\infty$  increases, the required number of labeled nodes increases to achieve the same level of testing error. This verifies our sample complexity bound in (5.18).

Figure 5.2 shows the testing error decreases as  $m$  increases when  $|\Omega|$  is fixed as 1500. Moreover, as  $\|\mathbf{A}^*\|_\infty$  increases, a larger  $m$  is needed to achieve the same level of testing error. This verifies our bound on the number of neurons in (5.17).



**Figure 5.1:** The testing error when  $|\Omega|$  and  $\|A^*\|_\infty$  change.  $m = 500$ .



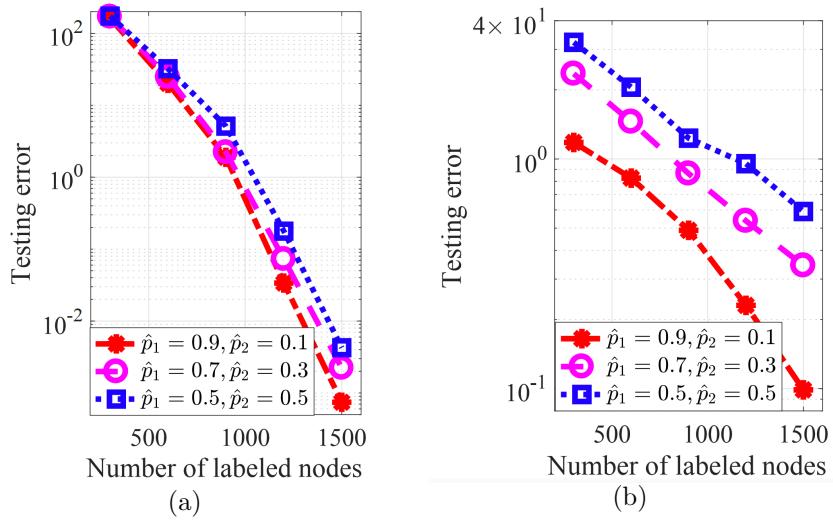
**Figure 5.2:** The testing error when  $m$  and  $\|A^*\|_\infty$  change.  $|\Omega| = 1500$ .

#### 5.4.2 Graph Sampling Affects the Effective Adjacency Matrix

Here we fix  $\mathbf{A}$  and the graph sampling strategy, and evaluate the prediction performance on datasets generated by (5.20) using different  $\hat{\mathbf{A}}$ . We generate  $\hat{\mathbf{A}}$  from  $\hat{\mathbf{A}} = \mathbf{A}\hat{\mathbf{P}}$ , where  $\hat{\mathbf{P}}$  is a diagonal matrix with  $\hat{\mathbf{P}}_{ii} = \hat{p}_1$  for nodes  $i$  in group 1 and  $\hat{\mathbf{P}}_{ii} = \hat{p}_2$  for nodes  $i$  in group 2. We vary  $\hat{p}_1$  and  $\hat{p}_2$  to generate three different datasets from (5.20). We consider both our graph sampling method in Section 5.3.2 and FastGCN [186].

In Figure 5.3,  $N_1 = 100$  and  $N_2 = 1900$ .  $d_1 = 10$  and  $d_2 = 1$ . Figure 5.3(a) shows the testing performance of a learned GCN by Algorithm 1, where  $p_1^* = 0.9$  and  $p_2^* = 0.1$ . the method indeed performs the best on Dataset 1 when  $\hat{\mathbf{A}}$  is generated using  $\hat{p}_1 = 0.9$  and  $\hat{p}_2 = 0.1$ , in which case  $\mathbf{A}^* = \hat{\mathbf{A}}$ . This verifies our theoretical result that graph sampling affects  $\mathbf{A}^*$  in the target functions, i.e., it achieves the best performance if  $\mathbf{A}^*$  is the same as  $\hat{\mathbf{A}}$  in the target function.

Fig. 5.3 (b) shows the performance on the same three datasets where in each iteration



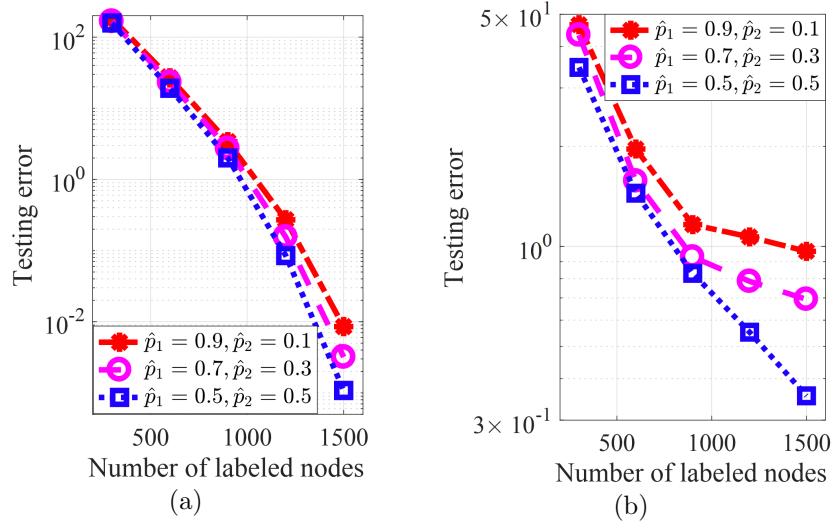
**Figure 5.3: Generalization performance of learned GCNs on datasets generated from different  $\hat{\mathbf{A}}$  by (a) our graph sampling strategy and (b) FastGCN.  $\mathbf{A}$  is very unbalanced.**

of Algorithm 1, the graph sampling strategy is replaced with FastGCN [186]. The method also performs the best in Dataset 1 when  $\mathbf{A}^*$  is generated using  $\hat{p}_1 = 0.9$  and  $\hat{p}_2 = 0.1$ . The reason is that the graph topology is highly unbalanced in the sense that  $\sqrt{d_2}N_2 \gg \sqrt{d_1}N_1$ , which means group 2 has a much higher impact on other nodes in group 1 in  $\mathbf{A}$ . The graph sampling reduces the impact of group 2 nodes more significantly than group 1 nodes, as discussed in Section 5.3.2.

To further illustrate this, in Figure 5.4 we change the graph topology by setting  $N_1 = 1000$  and  $N_2 = 1000$ , and all the other settings remain the same. In this case, the graph is balanced because  $\sqrt{d_2}N_2$  and  $\sqrt{d_1}N_1$  are in the same order. We generate different datasets using the new  $\mathbf{A}$  following the same method and evaluate the performance of both our graph sampling method and FastGCN. Both methods perform the best in Dataset 3 when  $\hat{\mathbf{A}}$  is generated using  $\hat{p}_1 = 0.5$  and  $\hat{p}_2 = 0.5$ . That is because on a balanced graph, graph sampling reduces the impact of both groups equally.

## 5.5 Conclusion

This chapter provides a new theoretical framework for explaining the empirical success of graph sampling in training GCNs. It quantifies the impact of graph sampling explicitly through the effective adjacency matrix and provides generalization and sample complexity



**Figure 5.4: Generalization performance of learned GCNs on datasets generated from different  $A^*$  by (a) our graph sampling strategy and (b) FastGCN.  $A$  is balanced.**

analyses. One future direction is to develop active graph sampling strategies based on the presented insights and analyze its generalization performance. Other potential extension includes the construction of statistical-model-based characterization of  $A^*$  and fitness to real-world data, and the generalization analysis of deep GCNs, graph auto-encoders, and jumping knowledge networks.

# CHAPTER 6

## HOW DO NONLINEAR TRANSFORMERS LEARN AND GENERALIZE IN IN-CONTEXT LEARNING?

### 6.1 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 [38], PaLM [199], LLaMa [200], as well as versatile visual and multi-modal models such as CLIP [201], DALL-E [202], and GPT-4. One intriguing capability exhibited by certain large language models (LLMs) is known as “**in-context learning**” (ICL) [38]. Given a pre-trained model  $F(\Psi)$ , parameterized by weights  $\Psi$ , the conventional approach fine-tunes  $\Psi$  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. [203] is the first paper to mathematically formulate ICL. Briefly speaking, to predict  $f(\mathbf{x}_{\text{query}})$  of a query input  $\mathbf{x}_{\text{query}}$  for a new task represented by the label function  $f$ , ICL augments  $\mathbf{x}_{\text{query}}$  by  $l$  example input-output pairs  $(\mathbf{x}_i, f(\mathbf{x}_i))_{i=1}^l$ . 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 [204, 205], natural language inference [206, 205], text generation [38, 207], etc.

In parallel, model pruning [208], [209] 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 [210] and reconstruction error minimization [211], magnitude-based pruning [209] is the most popular approach due to its simplicity and demonstrated promising empirical results. A few recent works [212], [213], [214], [215] 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:

---

Portions of this chapter have previously appeared as: H. Li, M. Wang, S. Lu, X. Cui, and P.-Y. Chen, “How do nonlinear transformers learn and generalize in in-context learning?” in *Proc. Int. Conf. Mach. Learn.*, Jul. 2024, pp. 28734–28783.

*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 [216], [217]. Specifically, [216] investigate the generalization gap and stability of ICL. [217] explore the training and generalization of ICL with Transformers, especially with distribution shifts during inference. [218] studies the required number of pre-training tasks for a desirable ICL property. [219] 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 [217], [218] or nonlinear activation in the multilayer perceptron (MLP) [219], [217], [218] or cannot characterize how to train a model to achieve the desirable ICL capability with distribution-shifted data [219], [216], [218].

### 6.1.1 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 [219]. 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.

**Table 6.1: Comparison with existing works about training analysis and generalization guarantee of in-context learning.**

Theoretical Works	Nonlinear Attention	Nonlinear MLP	Training Analysis	Distribution-Shifted Data	Tasks
[216]	✓	✓			linear regression
[217]			✓	✓	linear regression
[219]	✓		✓		linear regression
[218]			✓		linear regression
Ours	✓	✓	✓	✓	classification

#### 6.1.2 Related Work

**Expressive power of ICL** Some existing works study the expressive power of Transformers to implement algorithms via ICL. [220], [221] demonstrate that Transformers conduct

gradient descent during the forward pass of Transformers with prompts as inputs. [222], [223] extend the conclusion to preconditioned and functional gradient descent via ICL. [203], [224], [225] 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. [114], [6], [7], [118] study the generalization of one-layer Transformer by assuming spatial association or the majority voting of tokens. [116] delve into how one-layer Transformers learn semantic structure. [115] depict the trajectory of prompt tuning of attention networks. [119], [226] characterize that the gradient updates of the prompt or weights converge to a max-margin SVM solution. [104], [227] 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, [162] study the sample complexity of training a pruned network with a given sparse ground truth weight. [228] investigate the neural tangent kernel of the pruned model. [89], [229] 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.

## 6.2 Problem Formulation

This work studies the optimization and generalization of binary classification problems for in-context learning. Consider a query  $\mathbf{x}_{query}$  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  $\mathbf{x}_{query} \in \mathbb{R}^{d_x}$  through a task  $f$  that is randomly chosen from  $\mathcal{T}$ , i.e.,  $z = f(\mathbf{x}_{query}) \in \{+1, -1\}, f \in \mathcal{T}$ .

### 6.2.1 Training to Enhance ICL Capability

Following the framework of training for ICL in [203], [220], [224], 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  $\mathbf{P}$  for the query  $\mathbf{x}_{query}$ , and its label  $f(\mathbf{x}_{query})$ . We start by formulating  $\mathbf{P}$  and then introduce the learning model in this section.

Following [221], [217], [219], the prompt embedding  $\mathbf{P}$  of query  $\mathbf{x}_{query}$  is formulated as:

$$\begin{aligned} \mathbf{P} &= \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_l & \mathbf{x}_{query} \\ \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_l & \mathbf{0} \end{pmatrix} \\ &:= (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{query}) \in \mathbb{R}^{(d_x+d_y) \times (l+1)}, \end{aligned} \quad (6.1)$$

where the last column of  $\mathbf{P}$ , denoted by  $\mathbf{p}_{query}$ , includes the query  $\mathbf{x}_{query}$  with padding zeros, and the first  $l$  columns are the contexts for  $\mathbf{x}_{query}$ . We respectively call  $\mathbf{x}_i$  and  $\mathbf{y}_i$ ,  $i \in [l]$  *context* inputs and outputs, where  $l$  is also known as the prompt length. Let  $\text{Embd}(\cdot)$  be the embedding function of each context output.  $\mathbf{y}_i \in \mathbb{R}^{d_y}$  in (6.1) is defined as  $\mathbf{y}_i = \text{Embd}(f(\mathbf{x}_i))$ . Hence,  $\mathbf{P}$  is a function of  $f$ . The first  $d_x$  dimensions of  $\mathbf{p}_i$  are referred to as the feature embedding, while the last  $d_y$  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

$$\begin{aligned} F(\Psi; \mathbf{P}) &= \mathbf{a}^\top \text{Relu}(\mathbf{W}_O \sum_{i=1}^l \mathbf{W}_V \mathbf{p}_i \cdot \text{attn}(\Psi; \mathbf{P}, i)), \\ \text{attn}(\Psi; \mathbf{P}, i) &= \text{softmax}((\mathbf{W}_K \mathbf{p}_i)^\top \mathbf{W}_Q \mathbf{p}_{query}), \end{aligned} \quad (6.2)$$

where  $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{m_a \times (d_x+d_y)}$ ,  $\mathbf{W}_V \in \mathbb{R}^{m_b \times (d_x+d_y)}$  are the embedding matrices for queries, keys, and values, respectively, and  $\mathbf{W}_O \in \mathbb{R}^{m \times m_b}$  and  $\mathbf{a} \in \mathbb{R}^m$  are parameters in the MLP layer. Here,  $\text{softmax}((\mathbf{W}_K \mathbf{p}_i)^\top \mathbf{W}_Q \mathbf{p}_{query}) = e^{(\mathbf{W}_K \mathbf{p}_i)^\top \mathbf{W}_Q \mathbf{p}_{query}} / \sum_{j=1}^l e^{(\mathbf{W}_K \mathbf{p}_j)^\top \mathbf{W}_Q \mathbf{p}_{query}}$ .  $\Psi := \{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O, \mathbf{a}\}$  denotes the set of all model weights. Typically,  $\min(m_a, m_b) > d_x + d_y$ .

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

$$\min_{\Psi} R_N(\Psi) := \frac{1}{N} \sum_{n=1}^N \ell(\Psi; \mathbf{P}^n, z^n), \quad (6.3)$$

using  $N$  pairs of prompt embedding and label pairs  $\{\mathbf{P}^n, z^n\}_{n=1}^N$ . For the  $n$ -th pair,  $\mathbf{x}_{query}^n$  and the context input  $\mathbf{x}_i^n$  are all sampled from an unknown distribution  $\mathcal{D}$ , the task  $f^n$  is sampled from  $\mathcal{T}$ , and  $\mathbf{P}^n$  is constructed following (6.1). The loss function is a Hinge

loss, i.e.,  $\ell(\Psi; \mathbf{P}^n, z^n) = \max\{0, 1 - z^n \cdot F(\Psi; \mathbf{P}^n)\}$ , where  $F(\Psi; \mathbf{P}^n)$  is defined in (6.2). 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 (6.3) is a *multi-task learning* problem when  $|\mathcal{T}_{tr}| > 1$ .

### 6.2.2 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<sup>27</sup>

$$\mathbb{E}_{\mathbf{x}_{query} \sim \mathcal{D}, f \in \mathcal{T}} [\ell(\Psi; \mathbf{P}, z)], \quad (6.4)$$

where  $\mathbf{P}$  is defined in (6.1). Note that the in-domain performance includes the testing performance on *unseen* tasks in  $\mathcal{T} \setminus \mathcal{T}_{tr}$  that do not appear in the training samples.

**Out-of-Domain Generalization:** Suppose the testing queries  $\mathbf{x}_{query}$  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

$$\mathbb{E}_{\mathbf{x}_{query} \sim \mathcal{D}', f \in \mathcal{T}'} [\ell(\Psi; \mathbf{P}, z)]. \quad (6.5)$$

### 6.2.3 Training Algorithm

The model is trained using stochastic gradient descent (SGD) with step size  $\eta$  with batch size  $B$ , summarized in Algorithm 2 in Appendix E.3.  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$  and  $\mathbf{W}_V$  are initialized such that all diagonal entries of  $\mathbf{W}_V^{(0)}$ , and the first  $d_X$  diagonal entries of  $\mathbf{W}_Q^{(0)}$  and  $\mathbf{W}_K^{(0)}$  are set as  $\delta$  with  $\delta \in (0, 0.2]$ , and all other entries are 0. Each entry of  $\mathbf{W}_O^{(0)}$  is generated from  $\mathcal{N}(0, \xi^2)$ ,  $\xi = 1/\sqrt{m}$  and each entry of  $\mathbf{a}$  is uniformly sampled from  $\{1/m, -1/m\}$ . Besides,  $\mathbf{a}$  does not update during training.

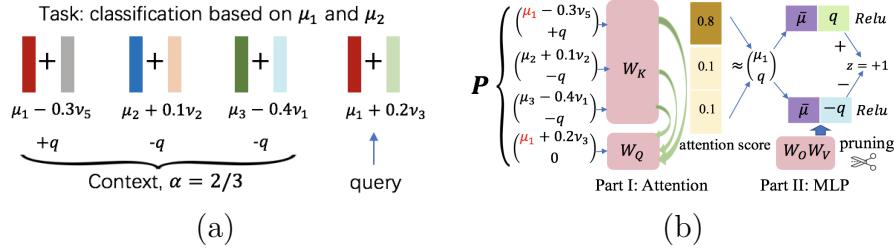
---

<sup>27</sup>In terms of evaluating generalization on unseen tasks, (6.4) is almost equivalent to replacing  $f \in \mathcal{T}$  with  $f \in \mathcal{T} \setminus \mathcal{T}_{tr}$  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 6.3.2). Therefore, an  $\mathcal{O}(\epsilon)$  generalization error on  $f \in \mathcal{T}$  can indeed reflect an  $\mathcal{O}(\epsilon)$  generalization error on  $f \in \mathcal{T} \setminus \mathcal{T}_{tr}$

### 6.2.4 Model Pruning

We also consider the case that the learned model  $\Psi$  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  $\mathbf{W}_O$ , resulting in the reduced matrix size of  $(m - |\mathcal{S}|) \cdot m_b$ .

## 6.3 Theoretical Results



**Figure 6.1:** (a) Example of prompt embedding.  $l = 3$ ,  $\alpha = 2/3$ . (b) The mechanism of a trained Transformer (6.2) to implement ICL. Part I: The attention layer assigns the largest attention score (0.8) on  $\mu_1 - 0.3\nu_5$ , which has the same IDR pattern as the query. Then the weighted sum of input tokens is close to  $(\mu_1^\top, q^\top)^\top$  by the trained attention layer. Part II: The neurons in  $W_O W_V$  with a large magnitude are aligned with  $\bar{\mu}$  and  $\pm q$  in the first  $d_x$  and the rest  $d_y$  dimensions, respectively. Then the prediction is based on the part of  $\pm q$  that varies for different queries rather than the part of  $\bar{\mu}$  that is universal for all IDR patterns.

We first summarize the main insights in Section 6.3.1. Section 6.3.2 formally presents our analysis model. Section 6.3.3 presents the formal theoretical results on the learning performance and the resulting ICL generalization. Section 6.3.4 provides the theoretical result that magnitude-based pruning on the out layer does not hurt ICL performance.

### 6.3.1 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  $\beta^{-1}$  and  $\alpha^{-1}$ , where  $\beta$  represents the norm of relevant patterns, and  $\alpha$  denotes the fraction of context inputs with the same in-domain-relevant pattern as the query. A higher  $\alpha$  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.

### 6.3.2 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  $\{\boldsymbol{\nu}_k\}_{k=1}^{M_2}$  ( $M_1 + M_2 = d_x$ ) in  $\mathbb{R}^{d_x}$ , where these  $M_1 + M_2$  patterns are pairwise orthogonal, and  $\|\boldsymbol{\mu}_j\| = \|\boldsymbol{\nu}_k\| = \beta \geq 1$  ( $\beta$  is a constant) for  $j \in [M_1], k \in [M_2]$ . Each in-domain data  $\mathbf{x}$  drawn from  $\mathcal{D}$  is generated by

$$\mathbf{x} = \boldsymbol{\mu}_j + \kappa \boldsymbol{\nu}_k, \quad (6.6)$$

where  $j \in [M_1]$  and  $k \in [M_2]$  are arbitrarily selected.  $\kappa$  follows a uniform distribution  $U(-K, K)$ ,  $K \leq 1/2$ . Denote  $\text{IDR}(\mathbf{x}) := \boldsymbol{\mu}_j$  as the IDR pattern in data  $\mathbf{x}$ . Our data assumption originates from recent feature learning works on deep learning [113], [6], [115] for

language and vision data. To the best of our knowledge, only [219] theoretically analyzes the performance of ICL with softmax attention, assuming all  $\mathbf{x}$  are orthogonal to each other. Our assumption in (6.6) is more general than that in [219].

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 6.3.1.** (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

$$f(\mathbf{x}) = \begin{cases} +1, & \text{IDR}(\mathbf{x}) = \boldsymbol{\mu}_a, \\ -1, & \text{IDR}(\mathbf{x}) = \boldsymbol{\mu}_b, \\ \text{random from } \{+1, -1\}, & \text{otherwise,} \end{cases} \quad (6.7)$$

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 (6.7), 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  $\mathbf{x}$  drawn from  $\mathcal{D}'$  can be generated by

$$\mathbf{x} = \boldsymbol{\mu}'_j + \kappa' \boldsymbol{\nu}'_k \quad (6.8)$$

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  $\text{ODR}(\mathbf{x}) := \boldsymbol{\mu}'_j$  to denote the ODR pattern of  $\mathbf{x}$ .

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 6.3.1, 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  $\mathbf{x}_{query}$  and  $f(\mathbf{x}_{query})$  for training, the context inputs  $\mathbf{x}_i$  in  $\mathbf{P}$  in (6.1) are constructed as follows. The IDR pattern is selected from  $\{\boldsymbol{\mu}_j\}_{j=1}^{M_1}$  following a categorical distribution parameterized by  $\alpha$ , 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 [204], [230], [205]. 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  $\mathbf{y}_i$  for both training and testing,  $\text{Emb}(+1) = \mathbf{q}$ ,  $\text{Emb}(-1) = -\mathbf{q}$ , where  $\mathbf{q} \in \mathbb{R}^{d_y}$ . Hence,  $\mathbf{y}_i \in \{\mathbf{q}, -\mathbf{q}\}$  for  $i \in [l_{tr}]$  or  $i \in [l_{ts}]$ .

### 6.3.3 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 6.3.2.** 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 6.3.2 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 6.3.2 in this case.

Following [86], [87], [6], we assume the training labels are balanced, i.e.,  $\left| |\{n : z^n = +1\}| - |\{n : z^n = -1\}| \right| = \mathcal{O}(\sqrt{N})$ . The next theorem states the training and in-domain generalization.

**Theorem 6.3.3. (In-Domain Generalization)** Suppose Condition 6.3.2 holds. For any  $\epsilon > 0$ , when (i) the number of neurons in  $\mathbf{W}_O$  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

$$l_{tr} \geq \max\{\Omega(\log M_1/\alpha), \Omega(1/(\beta^2\alpha))\}, \quad l_{ts} \geq \alpha'^{-1}, \quad (6.9)$$

(iv) and the number of iterations satisfies

$$T = \Theta(\eta^{-1} M_1 \alpha^{-\frac{2}{3}} \beta^{-2/3} \sqrt{\log M_1}), \quad (6.10)$$

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

$$\mathbb{E}_{\mathbf{x}_{query} \sim \mathcal{D}, f \in \mathcal{T}} [\ell(\Psi; \mathbf{P}, z)] \leq \mathcal{O}(\epsilon). \quad (6.11)$$

Theorem 6.3.3 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 6.3.3 includes three major insights:

1. *In-domain generalization capability using a diminishing fraction of training tasks.*

Because  $\mathcal{T}_{tr}$  can satisfy Condition 6.3.2 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  $\alpha^{-1}$  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 6.3.4. (Out-of-Domain Generalization)**

Suppose Condition 6.3.2 and conditions (i)-(iv) in Theorem 6.3.3 hold. For any  $\boldsymbol{\mu}'_1, \dots, \boldsymbol{\mu}'_{M_1}, \boldsymbol{\nu}'_1, \boldsymbol{\nu}'_{M_2}$  that are pairwise orthogonal and  $\|\boldsymbol{\mu}'_j\| = \|\boldsymbol{\nu}'_k\| = \beta$ , if

$$\boldsymbol{\mu}'_j \in \left\{ \sum_{i=1}^{M_1} k_{j,i} \boldsymbol{\mu}_i \mid S_j := \sum_{i=1}^{M_1} k_{j,i} \geq 1, k_{j,i} \in \mathbb{R} \right\}, \quad (6.12)$$

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

$$\mathbb{E}_{\mathbf{x}_{query} \sim \mathcal{D}', f \in \mathcal{T}'} [\ell(\Psi; \mathbf{P}, z)] \leq \mathcal{O}(\epsilon). \quad (6.13)$$

*Remark 6.3.5.* Theorem 6.3.4 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  $\geq 1$ , 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 6.3.6.* (Comparison with existing ICL analysis) [219] analyzes the generalization performance of ICL on unseen tasks under a similar data model that includes decisive and indecisive patterns. However, [219] 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 [217] 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 [217]. We provide the out-of-domain generalization guarantee for one type of distribution between  $\mathcal{D}$  and  $\mathcal{D}'$ .

### 6.3.4 ICL with Magnitude-Based Model Pruning

**Theorem 6.3.7.** Let  $\mathbf{r}_i$  be the  $i$ -row of  $\mathbf{W}_O \mathbf{W}_V$ ,  $i \in [m]$ . Suppose Condition 6.3.2 and conditions (i)-(iv) in Theorem 6.3.3 hold, then there exists  $\mathcal{L} \subset [m]$  with  $|\mathcal{L}| = \Omega(m)$  s.t.,

$$\begin{aligned} \|\mathbf{r}_i^{(T)}\| &\geq \Omega(1), \quad i \in \mathcal{L}, \\ \|\mathbf{r}_i^{(T)}\| &\leq (1/\sqrt{M_2}), \quad i \in \mathcal{L}^c, \end{aligned} \tag{6.14}$$

where  $\mathcal{L}^c$  is the complementary set of  $\mathcal{L}$ . Then, for any  $\epsilon > 0$  and any in- or out-of-domain  $\mathbf{x}_{query} \sim \mathcal{D}$  (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

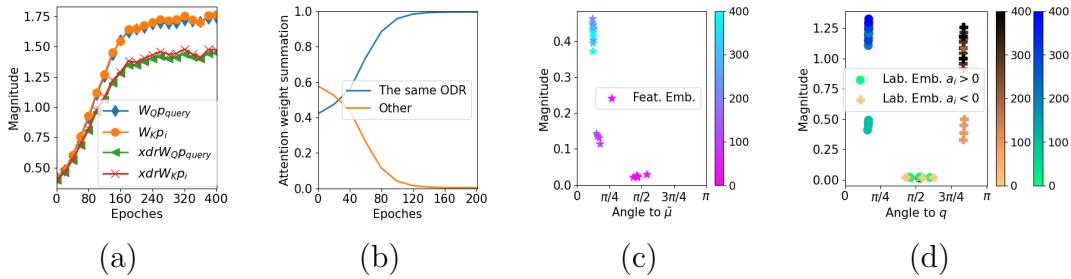
$$\mathbb{E}_{\mathbf{x}_{query}, f} [\ell(\Psi_{\mathcal{L}^c}; \mathbf{P}, z)] \leq \mathcal{O}(\epsilon + M_2^{-1/2}), \tag{6.15}$$

where  $\Psi_{\mathcal{L}^c}$  represents the model weights after removing neurons in  $\mathcal{L}^c$  in  $\mathbf{W}_O$ . 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

$$\mathbb{E}_{x_{query}, f} [\ell(\Psi_S; \mathbf{P}, z)] \geq \Omega(R + (\alpha' M_1)^{-1}). \quad (6.16)$$

*Remark 6.3.8.* Theorem 6.3.7 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  $\Psi$ . However, pruning neurons with a larger magnitude cause an increasing generalization error as the pruning ratio  $R$  increases. Theorem 6.3.7 indicates that in our setup, magnitude-based pruning on  $\mathbf{W}_O$  does not hurt the model's ICL capability.



**Figure 6.2: The properties of the trained model.** (a) The average norm of  $\mathbf{W}_Q p_{query}$ ,  $\mathbf{W}_K p_i$ ,  $[XDR(\beta^{-1} \cdot p_{query})^\top, \mathbf{0}^\top] \mathbf{W}_Q p_{query}$ , and  $[XDR(p_i)^\top / \beta, \mathbf{0}^\top] \mathbf{W}_K p_i$ . (b) The attention weight summation on contexts with the same ODR pattern as the query and other contexts. (c) The magnitude of the first  $d_x$  dimensions of 5 neurons in  $\mathbf{W}_O \mathbf{W}_V$  and their angles to  $\bar{\mu}$  in 400 epochs. (d) The magnitude of the rest  $d_y$  dimensions of 10 neurons in  $\mathbf{W}_O \mathbf{W}_V$  and their angles to  $q$  in 400 epochs. We choose 5 neurons for  $a_i > 0$  and 5 for  $a_i < 0$ .

## 6.4 The Mechanism of ICL by the Trained Transformer

Here, we provide a detailed discussion about how the generalization performance in Theorems 6.3.3 and 6.3.4 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 6.4.1 and 6.4.2. The high-level proof idea of Theorems 6.3.3 and 6.3.4 is presented in Appendix E.4.1.

### 6.4.1 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  $\mathbf{p} \in \mathbb{R}^{d_x+d_y}$  that includes input  $\mathbf{x}$  and the corresponding output embedding  $\mathbf{y}$ . We use  $XDR(\mathbf{p})$  to represent the relevant pattern, which is the IDR( $\mathbf{x}$ ) for in-domain data and ODR( $\mathbf{x}$ ) for out-of-domain data. Then

**Proposition 6.4.1.** *The trained model after being updated by  $T$  (characterized in (6.10)) iterations satisfies that, for any  $(\mathbf{p}, \mathbf{W}) \in \{(\mathbf{p}_{query}, \mathbf{W}_Q^{(T)}), \{(\mathbf{p}_i, \mathbf{W}_K^{(T)})\}_{i=1}^l\}$ ,*

$$\|[XDR(\mathbf{p})^\top, \mathbf{0}^\top] \mathbf{W} \mathbf{p}\| \geq \Omega(\sqrt{\log M_1}), \quad (6.17)$$

$$\|[\mathbf{a}^\top, \mathbf{0}^\top] \mathbf{W} \mathbf{p}\| \leq \mathcal{O}(\sqrt{\log M_1}(1/M_1 + 1/M_2)), \quad (6.18)$$

$$\|[\mathbf{b}^\top, \mathbf{0}^\top] \mathbf{W} \mathbf{p}\| \leq \mathcal{O}(\sqrt{\log M_1}(1/M_1 + 1/M_2)), \quad (6.19)$$

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

*Remark 6.4.2.* Proposition 6.4.1 indicates that the self-attention layer parameters  $\mathbf{W}_Q^{(T)}$  and  $\mathbf{W}_K^{(T)}$  in the returned model projects  $\mathbf{p}_{query}$  or context embeddings  $\mathbf{p}_i$  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 (6.17), (6.18), and (6.19), since components of  $\mathbf{W} \mathbf{p}$  in other directions rather than  $[XDR(\mathbf{p})^\top, \mathbf{0}^\top]$  are relatively smaller. Hence, Proposition 6.4.1 implies that the learned  $\mathbf{W}_Q^{(T)}$  and  $\mathbf{W}_K^{(T)}$  remove the effect of IDI/ODI patterns. Meanwhile, (6.17) states that the  $\mathbf{W}_Q^{(T)}$  and  $\mathbf{W}_K^{(T)}$  enlarge the magnitude of the IDR or ODR patterns from  $\Theta(1)$  to  $\Theta(\sqrt{\log M_1})$ , given that the  $\mathbf{W}_Q^{(0)}$  and  $\mathbf{W}_K^{(0)}$  are initialized with a scalar  $\delta = \Theta(1)$ .

Proposition 6.4.1 enables us to compute the attention map of the trained model. Therefore, we have the following.

**Corollary 6.4.3.** *For any testing query embedding  $\mathbf{p}_{query} = [\mathbf{x}_{query}^\top, \mathbf{0}^\top]^\top$ , 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)  $\mathbf{x}_{query}$ . Then, for any constant  $C > 1$ , by definition in (6.2), it holds that

$$\sum_{s \in \mathcal{N}_*} attn(\Psi; \mathbf{P}, i) \geq 1 - \Theta(1/M_1^C). \quad (6.20)$$

*Remark 6.4.4.* Corollary 6.4.3 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.

#### 6.4.2 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 6.4.5.** Let  $\mathbf{r}_i$  introduced in Theorem 6.3.7 be  $(\mathbf{r}_{i_{d_x}}, \mathbf{r}_{i_{d_y}})$  where  $\mathbf{r}_{i_{d_x}} \in \mathbb{R}^{1 \times d_x}$ ,  $\mathbf{r}_{i_{d_y}} \in \mathbb{R}^{1 \times d_y}$ . Then, for any  $i \in \mathcal{L}$ ,

$$\mathbf{r}_{i_{d_x}}^{(T)} \bar{\boldsymbol{\mu}} / (\|\mathbf{r}_{i_{d_x}}^{(T)}\| \cdot \|\bar{\boldsymbol{\mu}}\|) \geq 1 - \Theta(1)/M_2, \quad (6.21)$$

$$\mathbf{r}_{i_{d_y}}^{(T)} \mathbf{q}_e / (\|\mathbf{r}_{i_{d_y}}^{(T)}\| \cdot \|\mathbf{q}_e\|) \geq 1 - \Theta(1)/M_1, \quad (6.22)$$

where  $\bar{\boldsymbol{\mu}} = \sum_{k=1}^{M_1} \boldsymbol{\mu}_k^\top / M_1$ ,  $\mathbf{q}_e = \mathbf{q}$  if  $a_i > 0$  and  $\mathbf{q}_e = -\mathbf{q}$  if  $a_i < 0$ , where  $a_i$  is the  $i$ -th entry of  $\mathbf{a}$  in (6.1).

*Remark 6.4.6.* Proposition 6.4.5 demonstrates that neurons with indices in  $\mathcal{L}$  have the following two properties. (P1) The first  $d_x$  entries of all the corresponding row vectors in  $\mathbf{W}_O^{(T)} \mathbf{W}_V^{(T)}$  approximate the average of all IDR patterns  $\boldsymbol{\mu}_j$ ,  $j \in [M_1]$ . (P2) The next  $d_y$  entries of the  $i$ th row of  $\mathbf{W}_O^{(T)} \mathbf{W}_V^{(T)}$  approximates the label embedding  $\mathbf{q}$  when  $a_i > 0$  and approximates  $-\mathbf{q}$  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.

## 6.5 Numerical Experiments

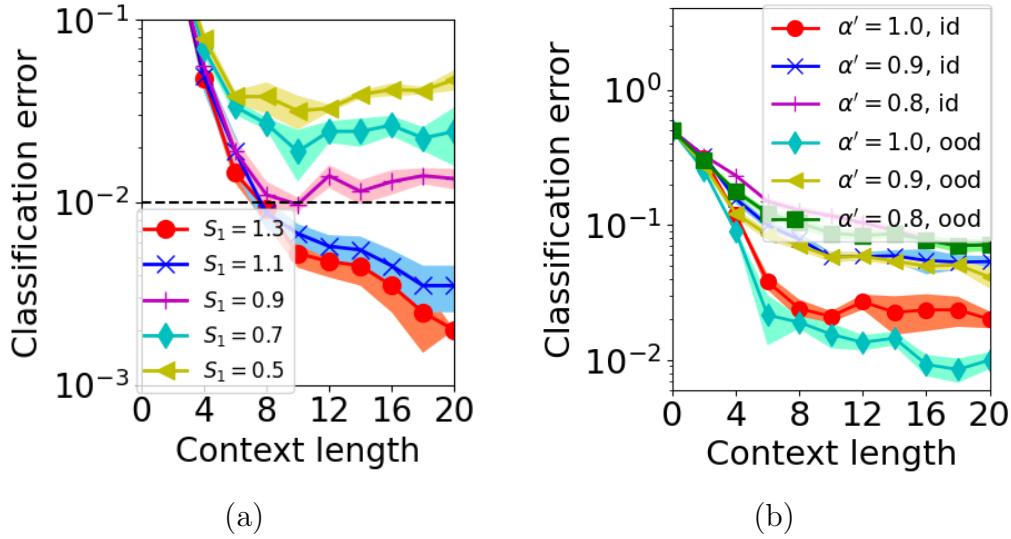
**Data Generation** We verify our theoretical findings using data generated as described in Section 6.2. Let  $d_x = d_y = 30$ ,  $\beta = 3$ ,  $K' = 5$ ,  $K = 0.5$ . The in-context binary classification error is evaluated by  $\mathbb{E}_{(\mathbf{x},y)}[\Pr(y \cdot F(\Psi; \mathbf{P}) < 0)]$  for  $\mathbf{x}$  following either  $\mathcal{D}$  or  $\mathcal{D}'$  and  $\mathbf{P}$  constructed in (6.1). 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  $i \in [M'_1]$ .  $\boldsymbol{\mu}'_1 = 0.3 \cdot (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + a\boldsymbol{\mu}_5 + b\boldsymbol{\mu}_6$ .  $\boldsymbol{\mu}'_2 = \sqrt{2}/2 \cdot (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ .  $\boldsymbol{\mu}'_3 = \sqrt{2}/2 \cdot (\boldsymbol{\mu}_3 + \boldsymbol{\mu}_4)$ . 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 (6.2) and the 3-layer 2-head real-world model GPT-2 following [224], [218]. If not otherwise specified, we set  $\alpha = 0.8$ ,  $l_{tr} = 20$  for training. The training tasks are formulated as follows to satisfy Condition 6.3.2. Define  $\mathbf{a}_i = \mathbf{a}_{i+M_1} = \boldsymbol{\mu}_i$  for  $i \in [M_1]$ , and then the  $((k-1) \cdot M_1 + j)$ -th task function maps the queries with  $\mathbf{a}_j$  and  $\mathbf{a}_{j+k}$  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  $|\mathcal{T} \setminus \mathcal{T}_{tr}| = 24$  in-domain unseen tasks. For GPT-2,  $U = 4$ . Then,  $|\mathcal{T}_{tr}| = 24$ ,  $|\mathcal{T} \setminus \mathcal{T}_{tr}| = 6$ . Note that we evaluate in-domain generalization error only on unseen tasks  $\mathcal{T} \setminus \mathcal{T}_{tr}$ , which is generally an upper bound of that defined in (6.4) after sufficient training.

### 6.5.1 Experiments on the Generalization of ICL

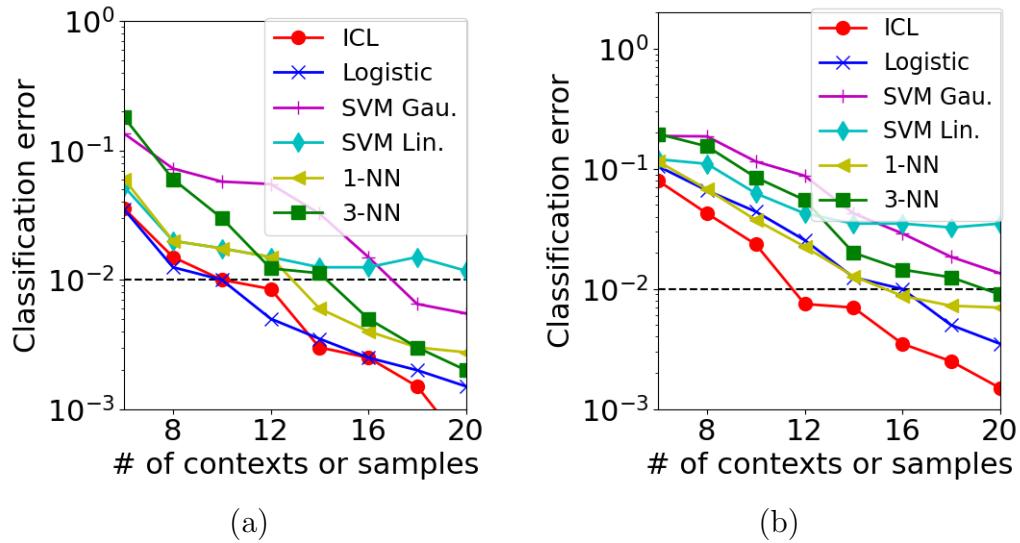
We first verify the sufficient condition (6.12) for out-of-domain generalization. From the selection of  $\boldsymbol{\mu}'$ 's, we know that  $S_1 = a + b$ ,  $S_2 = S_3 = \sqrt{2}$ . We vary  $a$  and  $b$  while satisfying  $a^2 + b^2 + 2 \cdot 0.3^2 = 1$ . Figure 6.3 (A) shows that the out-of-domain classification error achieves  $< 0.01$  when  $S_1 \geq 1$  and deviates from 0 when  $S_1 < 1$ , which justifies the necessity of condition (6.12). We then investigate how the context length is affected by  $\alpha'$ , i.e., the fraction of contexts with the same IDR/ODR pattern as the query. Figure 6.3 (B) indicates that a longer testing context length is needed when  $\alpha'$  is smaller for in- or out-of-domain, which is consistent with the lower bound of  $l_{ts}$  in (6.9) and Theorem 6.3.4.

We then compare ICL with other machine learning algorithms for classification, where contexts are used as training samples for these methods. Figure 6.4 (A) and (B) show that when  $\alpha' = 0.8$ , the advance of ICL over other algorithms is not significant, while when



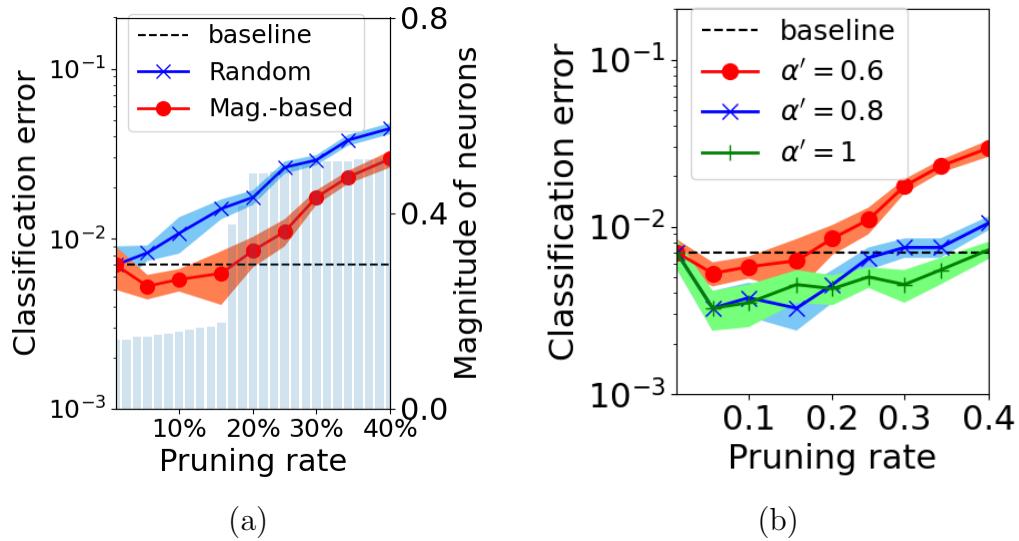
**Figure 6.3:** Out-of-domain ICL classification error on GPT-2 with (a) different  $S_1$  on GPT-2 (b) different  $\alpha'$  for in-domain (id) and out-of-domain (ood) generalization.

$\alpha' = 0.6$ , 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.



**Figure 6.4:** 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)  $\alpha' = 0.6$ .

We also investigate the effect of pruning techniques on ICL. Let  $\alpha = 0.6$ . Figure 6.5 (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  $\mathbf{W}_O$  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 6.3.7 and Remark 6.3.8. Figure 6.5 (B) justifies the impact of  $\alpha'$  in Theorem 6.3.7 that larger  $\alpha'$  can improve the performance of the pruned model.



**Figure 6.5:** (a) Out-of-domain classification error (left y-axis for curves) with model pruning of the trained  $\mathbf{W}_O$  using baseline (no pruning), random pruning, and magnitude-based pruning (Mag.-based), and the magnitude of each neuron of  $\mathbf{W}_O$  (right y-axis for light blue bars) (b) Out-of-domain classification error when varying  $\alpha'$ . These two are implemented on a one-layer Transformer.

### 6.5.2 Experiments on the Mechanism of ICL

We examine our findings regarding the mechanism of ICL in Section 6.4 using a one-layer Transformer formulated in (6.2). In Figure 6.2 (A) and (B), we consider out-of-domain data with  $a = b = 0.64$ . Figure 6.2 (A) shows that for any query  $\mathbf{p}_{query}$  (or context example  $\mathbf{p}_i$  for  $i \in l_{ts}$ ), the norm of  $[XDR(\mathbf{p})^\top, \mathbf{0}^\top] \mathbf{W}_Q \mathbf{p}_{query}$  (or  $[XDR(\mathbf{p})^\top, \mathbf{0}^\top] \mathbf{W}_K \mathbf{p}_i$ ) is close to the norm of  $\mathbf{W}_Q \mathbf{p}_{query}$  (or  $\mathbf{W}_K \mathbf{p}_i$ ). This implies that the components of  $\mathbf{W}_Q \mathbf{p}_{query}$  (or  $\mathbf{W}_K \mathbf{p}_i$ ) in directions other than  $[XDR(\mathbf{p})^\top, \mathbf{0}^\top]$  are small, which is consistent with (6.18) and (6.19) in Proposition 6.4.1. Moreover, these norms increase from initialization during training, which

justifies (6.17). Figure 6.2 (B) depicts the concentration of attention on contexts in  $\mathcal{N}_*$  after training. This verifies Corollary 6.4.3. Figure 6.2 (C) and (D) jointly verify Proposition 6.4.5. The color bars represent the epochs of training. We can observe that except for some neurons,  $\mathbf{r}_{i_{d_x}}$  grows to be close to the direction of  $\bar{\mu}$  with a larger magnitude in Figure 6.2 (C). Moreover, Figure 6.2 (D) shows for  $a_i > 0$  (or  $a_i < 0$ ),  $\mathbf{r}_{i_{d_y}}$  becomes close to  $\mathbf{q}$  (or  $-\mathbf{q}$ ) with a large magnitude.

## 6.6 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.

# CHAPTER 7

## TRAINING NONLINEAR TRANSFORMERS FOR CHAIN-OF-THOUGHT INFERENCE: A THEORETICAL GENERALIZATION ANALYSIS

### 7.1 Introduction

Transformer-based large-scale foundation models, such as GPT-3 [38], GPT-4, LLaMa [200], [231], and Sora [232], have demonstrated remarkable success across various tasks, including natural language processing [38], [231], multimodal learning [201], and image/video generation [232]. What is more surprising is that large language models (LLMs) demonstrate reasoning ability through the so-called “Chain-of-Thought” (CoT) method [233]. The objective is to let a pre-trained LLM generate  $K$  steps of reasoning given input query  $\mathbf{x}_{query}$  without any fine-tuning. To achieve that, the input  $\mathbf{x}_{query}$  is augmented with  $l$  examples  $\{\mathbf{x}_i, \{\mathbf{y}_{i,j}\}_{j=1}^K\}_{i=1}^l$  of a certain  $K$ -step reasoning task, where each  $\mathbf{x}_i$  is the input with  $\mathbf{y}_{i,j}$  as the  $j$ -th reasoning step, and  $\mathbf{y}_{i,K}$  is the final output. A pre-trained model then takes the resulting augmented input, referred to as a *prompt*, and outputs the corresponding reasoning steps  $\{\mathbf{z}_j\}_{j=1}^K$  for  $\mathbf{x}_{query}$ , or simply outputs  $\mathbf{z}_K$ . CoT can be viewed as an extended and more intelligent method than the previous in-context learning (ICL) method, where only input-label pairs  $\{\mathbf{x}_i, \mathbf{y}_{i,K}\}_{i=1}^l$  are augmented in the prompt to predict  $\mathbf{z}_K$  with the pre-trained model.

Inspired by the outstanding empirical performance of CoT in arithmetic reasoning [234], [235], [236], symbolic reasoning [235], [237], and commonsense reasoning [234], [236], there have been some recent works [238], [239], [240], [241], [242] on the theoretical understanding of CoT. These works investigate CoT from the perspective of expressive power, i.e., they construct the Transformer architecture that is proven to have the CoT ability. They also demonstrate empirically that supervised training on pairs of CoT prompts and corresponding outputs can lead to models with CoT ability. However, none of these results theoretically address the question of why a Transformer can obtain generalization-guaranteed CoT ability by training from data with gradient-based methods. Meanwhile, another line of research [217], [219], [218], [23] aims to unveil the reasons behind the ICL ability of Transformers through

---

Portions of this chapter have previously appeared as: H. Li, M. Wang, S. Lu, X. Cui, and P.-Y. Chen, “Training nonlinear transformers for chain-of-thought inference: A theoretical generalization analysis,” 2024, *arXiv:2410.02167*.

characterizing the training dynamics of a Transformer in the supervised setting. These analyses are specifically applicable to ICL. Therefore, a theoretical question still remains less explored, i.e.,

*Why can a Transformer be trained to generalize on multi-step reasoning tasks via CoT?*

### 7.1.1 Major Contributions

Following [238], [239], [240], [241], [242], we train the model in a supervised setting using prompt and label pairs. This chapter provides the first theoretical analysis of the training dynamics of nonlinear Transformers to achieve CoT ability. We prove that the learned model has guaranteed CoT ability for new tasks with distribution shifts from the training tasks, even when there exist noisy and erroneous context examples in the prompt. We theoretically characterize the required number of training samples and iterations needed to train a desirable model and the number of context examples required for successful CoT reasoning with a generalization guarantee. Moreover, we provide a theoretical explanation for why CoT outperforms ICL in some cases. Our main technical contributions are as follows:

**1. A quantitative analysis of how the training can enable the CoT ability:**

We theoretically analyze the training dynamics on a one-layer single-head attention-only Transformer and quantify the required number of context examples in each training sample, the total number of training samples, and the number of training iterations needed to acquire CoT ability. We illustrate that the CoT ability results from the property that the attention values of the learned model are concentrated on testing context examples with the same input patterns as the testing query during each reasoning step.

**2. A quantitative analysis of how context examples affect CoT performance:**

We characterize the required number of context examples in the testing prompt for successful CoT reasoning when noise and error exist in contexts. Our quantitative bounds are consistent with the intuition that more accurate context examples and more similar examples to the query improve CoT accuracy.

**3. A theoretical characterization of why CoT outperforms ICL:** We provide a quantitative analysis of the requirements for successful ICL reasoning with our studied trained model. We show that successful ICL requires an additional condition that the prompt has a dominant number of correct input-label examples, while the success of CoT does not depend on this condition. This can be viewed as one of the possible reasons why CoT outperforms

ICL.

### 7.1.2 Related Works

**Expressive power of CoT** [238] proves the existence of a Transformer that can learn a multi-layer perceptron (MLP). They interpret CoT as first filtering important tokens and then making predictions by ICL. They also establish the required number of context examples for a desired prediction with the constructed Transformer. [239], [240], [243] show that Transformers with CoT are more expressive than Transformers without CoT. [241], [242] show the superiority of standard Transformers in some reasoning tasks compared with recurrent neural networks and linear Transformers.

**Theoretical analysis of ICL** As a simplified one-step version of CoT, ICL has gained much attention from the theoretical community. [203], [220], [224], [225] demonstrate that Transformers are expressive to conduct many machine learning algorithms in context. [220], [221], [222], [223], [244] especially show the existence of Transformers to implement gradient descent and its variants with different input prompts. [217], [219], [218], [23] explore the training dynamics and generalization of ICL on single-attention Transformers. [245], [246] provably show the superiority of multi-head attention over single-head attention to achieve ICL ability.

**Training and Generalization of Transformers** There have been several recent works about the optimization and generalization analysis of Transformers. [114], [116], [115], [6], [7], [247], [118], [248] study the generalization of one-layer Transformers by assuming spatial association, semantic/contextual structure, or the majority voting of tokens in the data. [115], [119], [226], [104], [227], [249], [250], [251], [252] investigate the training dynamics or loss landscape of Transformers for the next token prediction by assuming infinitely long input sequences, causal structure/Markov Chain of data, or a proper prediction head. [253], [254] analyze the optimization and generalization of multi-head attention networks.

## 7.2 Problem Formulation

We study the problem of learning and generalization of  $K$ -steps reasoning tasks. Each task  $f = f_K \circ \dots \circ f_2 \circ f_1$  is a composition of functions  $\{f_i\}_{i=1}^K$  and outputs labels  $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_K$  for the input  $\mathbf{x}_{query}$ . During the  $k$ -th reasoning step,  $k \in [K]$ , the label is  $\mathbf{z}_k = f_k(\mathbf{z}_{k-1})$ , where  $\mathbf{z}_0 := \mathbf{x}_{query}$ .

### 7.2.1 Training to Acquire the Chain-of-Thought Ability

Following theoretical analysis [239], [240], [242] and empirical works like process supervision [255], we first investigate the training on a Transformer model to obtain the CoT ability in evaluating new data and tasks. It is a supervised learning setting on pairs of prompts and labels. Different from the testing prompt that includes examples and only  $\mathbf{x}_{query}$ , the training prompt includes multiple  $K$ -steps reasoning examples and a  $(k - 1)$ -step reasoning of  $\mathbf{x}_{query}$  for any  $k$  in  $[K]$ , and the label for this prompt is  $\mathbf{z}_k$ . Specifically,

**Training Prompt and Label for CoT.** For every prompt and output pair from a task  $f = f_K \circ \dots \circ f_2 \circ f_1$ , we construct a prompt  $\mathbf{P}$  that include the query input  $\mathbf{z}_{k-1}$  by prepending  $l_{tr}$  reasoning examples and the first  $k - 1$  steps of the reasoning query. The prompt  $\mathbf{P}$  of the query input  $\mathbf{z}_{k-1}$  is formulated as:

$$\mathbf{P} = (\mathbf{E}_1, \mathbf{E}_2, \dots, \mathbf{E}_{l_{tr}}, \mathbf{Q}_k) \in \mathbb{R}^{2d_x \times (l_{tr}K+k)},$$

$$\text{where } \mathbf{E}_i = \begin{pmatrix} \mathbf{x}_i & \mathbf{y}_{i,1} & \cdots & \mathbf{y}_{i,K-1} \\ \mathbf{y}_{i,1} & \mathbf{y}_{i,2} & \cdots & \mathbf{y}_{i,K} \end{pmatrix}, \quad \mathbf{Q}_k = \begin{pmatrix} \mathbf{z}_0 & \mathbf{z}_1 & \cdots & \mathbf{z}_{k-2} & \mathbf{z}_{k-1} \\ \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_{k-1} & \mathbf{0} \end{pmatrix}, i \in [l_{tr}], \quad (7.1)$$

where  $\mathbf{E}_i$  is the  $i$ -th context example, and  $\mathbf{Q}_k$  is the first  $k$  steps of the reasoning query for any  $k$  in  $[K]$ . We have  $\mathbf{y}_{i,k} = f_k(\mathbf{y}_{i,k-1})$  and  $\mathbf{z}_k = f_k(\mathbf{z}_{k-1})$  for  $i \in [l_{tr}]$ ,  $k \in [K]$  with a notation  $\mathbf{y}_{i,0} := \mathbf{x}_i$ . Let  $\mathbf{p}_s$  and  $\mathbf{p}_{query}$  be the  $s$ -th column and the last column of  $\mathbf{P}$ , respectively, for  $s \in [l_{tr}K + k - 1]$ .  $\mathbf{x}_i, \mathbf{y}_{i,k}, \mathbf{z}_j \in \mathbb{R}^{d_x}$  for  $i \in [l_{tr}]$  and  $j, k \in [K]$ . We respectively call  $\mathbf{x}_i$  and  $\mathbf{y}_{i,k}$  *context* inputs and outputs of the  $k$ -th step of the  $i$ th context example. For simplicity of presentation, we denote  $\mathbf{z}$  as the label of  $\mathbf{P}$ , which is indeed  $\mathbf{z}_k$  for (7.1). All the notations are summarized in Table A.1 in Appendix.

The **learning model** is a single-head, one-layer attention-only Transformer. We consider positional encoding  $\{\mathbf{c}_k\}_{k=1}^K \in \mathbb{R}^{2d_x}$ . Following theoretical works [114], [248], [250], we add the positional encoding to each  $\mathbf{p}_i$  by  $\tilde{\mathbf{p}}_i = \mathbf{p}_i + \mathbf{c}_{(i \bmod K)}$ ,  $i \in [K(l_{tr} + 1)]$ .  $\tilde{\mathbf{p}}_{query}$  is also defined by adding the corresponding  $\mathbf{c}_k$  to  $\mathbf{p}_{query}$ . Mathematically, given a prompt  $\mathbf{P}$  defined in (7.1) with  $\text{len}(P)$  (which is at most  $K(l_{tr} + 1)$ ) denoting the number of columns, it can be written as

$$F(\Psi; \mathbf{P}) = \sum_{i=1}^{\text{len}(P)-1} \mathbf{W}_V \tilde{\mathbf{p}}_i \cdot \text{softmax}((\mathbf{W}_K \tilde{\mathbf{p}}_i)^\top \mathbf{W}_Q \tilde{\mathbf{p}}_{query}), \quad (7.2)$$

where  $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{m \times (2d_x)}$ ,  $\mathbf{W}_V \in \mathbb{R}^{d_x \times (2d_x)}$  are the embedding matrices for queries,

and values, respectively.  $\Psi := \{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V\}$  is the set of all model weights<sup>28</sup>. Typically,  $m > 2d_{\mathcal{X}}$ .

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

$$\min_{\Psi} R_N(\Psi) := \frac{1}{N} \sum_{n=1}^N \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n), \quad (7.3)$$

using  $N$  prompt and label pairs  $\{\mathbf{P}^n, \mathbf{z}^n\}_{n=1}^N$ . For the  $n$ -th sample,  $\mathbf{x}_{query}^n$  and the context input  $\mathbf{x}_i^n$  are all sampled from an unknown distribution  $\mathcal{D}$ , the training task  $f^n$  is sampled from  $\mathcal{T}$ ,  $k$  is randomly selected from 1 to  $K$ , and  $\mathbf{P}^n$  is constructed following (7.1). The loss function is squared loss, i.e.,  $\ell(\Psi; \mathbf{P}^n, \mathbf{z}^n) = 1/2 \cdot \|\mathbf{z}^n - F(\Psi; \mathbf{P}^n)\|^2$ , where  $F(\Psi; \mathbf{P}^n)$  is defined in (7.2).

### 7.2.2 Training Algorithm

For simplicity of analysis, we let  $\mathbf{W} = \mathbf{W}_K^\top \mathbf{W}_Q$  and  $\mathbf{W}_V = (\mathbf{0}_{d_{\mathcal{X}} \times d_{\mathcal{X}}} \ \mathbf{I}_{d_{\mathcal{X}}}) \in \mathbb{R}^{d_{\mathcal{X}} \times (2d_{\mathcal{X}})}$  as [114], [219], [217], [248]. Let  $\{\mathbf{c}_k\}_{k=1}^K$  be a set of orthonormal vectors. The model is trained using stochastic gradient descent (SGD) with step size  $\eta$  with batch size  $B$ , summarized in Algorithm 3 in Appendix F.2. Each entry of  $\mathbf{W}^{(0)}$  is generated from  $\mathcal{N}(0, \xi^2)$  for a tiny  $\xi > 0$ .  $\mathbf{W}_V$  is fixed during the training. The fraction of prompts with  $\mathbf{z}_{k-1}$  as the query input is  $1/K$  by uniform sampling for any  $k \in [K]$  in each batch.

### 7.2.3 Chain-of-Thought Inference

We then consider another  $K$ -steps reasoning task  $f \in \mathcal{T}'$ , whose target is to predict labels  $\{\mathbf{z}_k\}_{k=1}^K$  given the input query  $\mathbf{x}_{query}$ .  $\mathcal{T}'$  is the set of testing tasks, and  $\mathcal{T}' \neq \mathcal{T}$ .

**Testing Prompt for CoT.** The testing prompt  $\mathbf{P}$  is composed of  $l_{ts}$  ( $\leq l_{tr}$ ) context examples of  $K$  steps plus a query, which is constructed as

$$\mathbf{P} = (\mathbf{E}_1, \mathbf{E}_2, \dots, \mathbf{E}_{l_{ts}}, \mathbf{p}_{query}) \in \mathbb{R}^{(2d_{\mathcal{X}}) \times (l_{ts}K+1)}, \mathbf{p}_{query} = (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top, \quad (7.4)$$

where  $\mathbf{E}_i$  follows the form in (7.1) for  $i \in [l_{ts}]$ .

We follow the CoT-I/O scheme formulated in [238], [239], [240], [241], [256] as the inference method. Specifically, for a  $K$ -step CoT with  $l_{ts}$  examples on a certain  $f \in \mathcal{T}'$ , given

---

<sup>28</sup>We focus on a one-layer single-head Transformer motivated by recent advancements and current state in Transformer and CoT analysis. Please see Appendix F.1.1 for discussion.

the testing prompt  $\mathbf{P}$  defined in (7.4), let  $\mathbf{P}_1 = \mathbf{P}$  and  $\mathbf{P}_0$  be the first  $K \cdot l_{ts}$  columns of  $\mathbf{P}$ . When we use CoT prompting for prediction in the  $k$ -th step, we first generate the output  $\mathbf{v}_k$ ,  $k \in [K]$  via greedy decoding by feeding the  $k$ -th step prompt  $\mathbf{P}_k$  to the trained model  $\Psi$  obtained from (7.3). The greedy decoding scheme means outputting the most probable token from the discrete set  $\mathcal{Y}$  of all possible outputs, as stated in (7.5).

$$\mathbf{v}_k = \arg \min_{\mathbf{u} \in \mathcal{Y}} \frac{1}{2} \|F(\Psi; \mathbf{P}_k) - \mathbf{u}\|^2, \text{ (greedy decoding)} \quad (7.5)$$

Then, we use the output  $\mathbf{v}_k$  to update  $\mathbf{P}_k$  and use  $\mathbf{v}_k$  as the query input to form the input prompt  $\mathbf{P}_{k+1}$  for the next step, which is computed as

$$\begin{aligned} \mathbf{P}_k &= (\mathbf{P}_{k-1} \ \mathbf{q}_k) \in \mathbb{R}^{(2d_{\mathcal{X}}) \times (Kl_{ts} + k)}, \quad \mathbf{P}_{k+1} = (\mathbf{P}_k \ \mathbf{q}_{k+1}) \in \mathbb{R}^{(2d_{\mathcal{X}}) \times (Kl_{ts} + k + 1)}, \\ \text{where } \mathbf{q}_k &= (\mathbf{v}_{k-1}^\top \ \mathbf{v}_k^\top)^\top, \quad \mathbf{q}_{k+1} = (\mathbf{v}_k^\top \ \mathbf{0}^\top)^\top, \end{aligned} \quad (7.6)$$

where  $\mathbf{q}_k$  is the  $k$ -th step reasoning column for the query. The model finally outputs  $\mathbf{v}_1, \dots, \mathbf{v}_K$  as CoT result for query  $\mathbf{x}_{query}$  by (7.5). The CoT process is summarized in Algorithm 4 of Appendix F.2.

When  $K \geq 2$ , following [238], [239], [240], [241], the **CoT generalization error** given the testing query  $\mathbf{x}_{query}$ , the testing data distribution  $\mathcal{D}'$ , and the labels  $\{\mathbf{z}_k\}_{k=1}^K$  on a  $K$ -steps testing task  $f \in \mathcal{T}'$  is defined as

$$\bar{R}_{CoT, \mathbf{x}_{query} \sim \mathcal{D}', f \in \mathcal{T}'}^f(\Psi) = \mathbb{E}_{\mathbf{x}_{query} \sim \mathcal{D}'} \left[ \frac{1}{K} \sum_{k=1}^K \mathbb{1}[\mathbf{z}_k \neq \mathbf{v}_k] \right], \quad (7.7)$$

which measures the average error between the output and the label of each reasoning step. A zero CoT generalization error indicates correct generations in all  $K$  steps.

#### 7.2.4 In-Context Learning Inference

The ICL inference on a  $K$ -steps reasoning task  $f \in \mathcal{T}'$  only predicts the final-step label by perpending examples of input and label pairs before the query. ICL can be viewed as a one-step CoT without intermediate steps. Here, we evaluate the ICL performance of the trained model.

**Testing Prompt for ICL.** Mathematically, ICL is implemented by constructing a

prompt  $\mathbf{P}$  as below,

$$\mathbf{P} = (\mathbf{E}_1, \dots, \mathbf{E}_{l_{ts}}, \mathbf{p}_{query}), \text{ where } \mathbf{p}_{query} = \begin{pmatrix} \mathbf{x}_{query} \\ \mathbf{0} \end{pmatrix}, \mathbf{E}_i = \begin{pmatrix} \mathbf{x}_i & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{y}_{i,K} & \mathbf{0} & \cdots & \mathbf{0} \end{pmatrix} \quad (7.8)$$

$\mathbf{P} \in \mathbb{R}^{(2d_X) \times (l_{ts}K+1)}$ ,  $\mathbf{E}_i \in \mathbb{R}^{(2d_X) \times K}$  for  $i \in [l_{ts}]$ . Note that in the ICL setting,  $\mathbf{E}_i$  only has input  $\mathbf{x}_i$  and the  $K$ -step output  $\mathbf{y}_{i,K}$  but does not include any intermediate labels. We pad zeros in  $\mathbf{E}_i$  so that its dimension is the same as  $\mathbf{E}_i$  in (7.1) for the inference with the same model as for CoT. The ICL output is  $\mathbf{v} = \arg \min_{\mathbf{u} \in \mathcal{Y}} \frac{1}{2} \|F(\Psi; \mathbf{P}) - \mathbf{u}\|^2$ , following (7.5). The **ICL generalization error** is

$$\bar{R}_{ICL, \mathbf{x}_{query} \sim \mathcal{D}', f \in \mathcal{T}'}^f(\Psi) = \mathbb{E}_{\mathbf{x}_{query} \sim \mathcal{D}'} [\mathbb{1}[\mathbf{z}_K \neq \mathbf{v}]], \quad (7.9)$$

which measures the error between the one-step reasoning output and the final step label.

## 7.3 Theoretical Results

We first summarize the main theoretical insights in Section 7.3.1. Then, we introduce the formulation of data and tasks in Section 7.3.2. Sections 7.3.3, 7.3.4, and 7.3.5, respectively characterize the training analysis of the Transformer and generalization using CoT and ICL with the trained model.

### 7.3.1 Main Theoretical Insights

We consider the setup that the model is trained using samples generated from tasks in  $\mathcal{T}$  that operate on  $M$  orthonormal training-relevant (TRR) patterns, while both CoT and ICL are evaluated on tasks in  $\mathcal{T}'$  that operate on  $M'$  orthonormal testing-relevant (TSR) patterns that belong to the span of TRR patterns. We consider the general setup that the context examples in the prompt for CoT and ICL testing are both noisy, i.e., TSR patterns with additive noise, and partially inaccurate, i.e., the reasoning in some examples contains incorrect steps. Our main insights are as follows.

**P1. Training Dynamics of Nonlinear Transformer towards CoT.** We theoretically analyze the training dynamics on a one-layer single-head attention-only Transformer to acquire the CoT generalization ability and characterize the required number of training samples and iterations. Theorem 3 shows that to learn a model with guaranteed CoT ability,

the required number of context examples in each training sample and the total number of training samples/iterations are linear in  $\alpha^{-1}$  and  $\alpha^{-2}$ , respectively, where  $\alpha$  is the fraction of context examples with inputs that share the same TRR patterns as the query. This is consistent with the intuition that the CoT performance is enhanced if more context examples are similar to the query. Moreover, the attention values of the learned model are proved to be concentrated on testing context examples that share similar input TSR patterns as the testing query during each of the reasoning steps (Proposition 3), which is an important property that leads to the success of the CoT generalization.

**P2. Guaranteed CoT Generalization.** To achieve zero CoT error on tasks in  $\mathcal{T}'$  with the learned model, Theorem 4 shows that the required number of context examples, where noise and errors are present, for task  $f$  in the testing prompt is proportional to  $(\alpha' \tau^f \rho^f)^{-2}$ , where  $\alpha'$  is the fraction of context examples with inputs that share the same TSR patterns as the query, the constant  $\tau^f$  in  $(0, 1)$  measures the fraction of accurate context examples, and a larger constant  $\rho^f$  in  $(0, 1)$  reflects a higher reasoning accuracy in each step of the examples. This result formally characterizes the intuition that more accurate context examples and more similar examples to the query improve the CoT accuracy.

**P3. CoT outperforms ICL.** In Theorem 5, We theoretically show that the required number of testing context examples for ICL to be successful has a similar form to that for CoT in Theorem 4, but with an additional requirement (Condition 1) that the fraction of correct input-label examples in the testing prompt must be dominant. Because not all testing cases satisfy this requirement, our result provides one explanation for why CoT sometimes outperforms ICL.

### 7.3.2 The Formulation of Data and Tasks

**Training data and tasks:** Consider  $M$  training-relevant (TRR) patterns  $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_M$ , which form an orthonormal set  $\mathcal{M} = \{\boldsymbol{\mu}_i\}_{i=1}^M$ .  $M = \Theta(d)$ ,  $M \leq d$ .  $(\boldsymbol{\mu}_i^\top, 0_{d_x}^\top)^\top \perp \mathbf{c}_k$  for  $i \in [M'], k \in [K]$ .

Every training prompt  $\mathbf{P}$  in (7.1) contains the query and training examples from the same training task  $f$  in the set of training tasks  $\mathcal{T}$ . Specifically, each training task  $f$  is a composition of  $K$  functions  $f = f_K \circ \dots \circ f_2 \circ f_1$  where each function  $f_k$  belongs to a function set  $\mathcal{F}$ . The  $k$ -th step label of the query is  $\mathbf{z}_k = f_k(\mathbf{z}_{k-1})$  given the  $k$ -th step input  $\mathbf{z}_{k-1}$  with  $\mathbf{z}_k \in \mathcal{M}$ ,  $k \in [K]$ . Moreover, the  $k$ -th step label of the  $i$ -th ( $i \in [l_{tr}]$ ) context example

is  $\mathbf{y}_{i,k} = f_k(\mathbf{y}_{i,k-1})$  given the  $k - 1$ th step input  $\mathbf{y}_{i,k-1}, k \in [K]$  with  $\mathbf{x}_i, \mathbf{y}_{i,k} \in \mathcal{M}$ , where  $\mathbf{y}_{i,0} := \mathbf{x}_i$ <sup>29</sup>. We assume that  $f_k(\mathbf{x}) \neq f_{k'}(\mathbf{x}')$  if and only if either  $\mathbf{x} \neq \mathbf{x}'$  or  $f_k \neq f_{k'}$ .

**Training prompt:** Consider a training prompt  $\mathbf{P}$  on task  $f \in \mathcal{T}$  defined in (7.1) with the query input  $\mathbf{z}_{k-1}, k \in [K]$ . Let  $\alpha \in (0, 1 - c]$  for some constant  $c > 0$ <sup>30</sup> denote the fraction of context examples with input sharing the same TRR pattern as the query input.

**Testing task and query:** Consider  $M'$  testing-relevant (TSR) patterns  $\boldsymbol{\mu}'_1, \boldsymbol{\mu}'_2, \dots, \boldsymbol{\mu}'_{M'}$ , which form an orthonormal set  $\mathcal{M}' = \{\boldsymbol{\mu}'_i\}_{i=1}^{M'}$ .  $M' \leq M$ . We also have  $\boldsymbol{\mu}'_i \perp \mathbf{c}_k$  for  $i \in [M'], k \in [K]$ . Let  $\mathcal{T}'$  denote the set of testing tasks, which all operate on patterns in  $\mathcal{M}'$  rather than  $\mathcal{M}$  in training tasks in  $\mathcal{T}$ . Every testing task  $f = f_K \circ \dots \circ f_2 \circ f_1 \in \mathcal{T}'$  is a composition of  $K$  functions. The reasoning for the testing query is considered to be *noiseless* and *accurate*. That means,

$$\mathbf{z}_k \in \mathcal{M}' \text{ for all } k \in \{0\} \cup [K], \text{ and } \mathbf{z}_k = f_k(\mathbf{z}_{k-1}), \mathbf{z}_0 = \mathbf{x}_{query}. \quad (7.10)$$

**Testing prompt:** We consider the general setup that testing examples are *noisy* and *erroneous*. By noisy examples, we mean all inputs and outputs of each step are noisy versions of TSR patterns, i.e.,

$$\mathbf{x}_i, \mathbf{y}_{i,k} \in \{\mathbf{b} \in \mathbb{R}^d \mid \mathbf{b} = \boldsymbol{\mu}'_j + \boldsymbol{\delta}, j \in [M'], \boldsymbol{\delta} \perp \mathcal{M}', \|\boldsymbol{\delta}\| \leq \sqrt{2}/2\}, \quad (7.11)$$

with noise  $\boldsymbol{\delta} \neq 0$  for  $i \in [Kl_{ts}^f], k \in [K]$ . Denote  $\text{TSR} : \mathbb{R}^d \mapsto \mathbb{Z}^+$  as a function that outputs the index of the TSR pattern of the noisy input. We consider the case that at least an  $\alpha'$  fraction of context examples where the TSR pattern of the input  $\mathbf{y}_{s,1}, s \in [l_{ts}^f]$  is the same as  $\mathbf{x}_{query}$ .

By erroneous examples, we mean that the reasoning steps in test examples may contain errors. To formally model this, we define the **step-wise transition matrices**  $\{\mathbf{A}_k^f\}_{k=1}^K \in \mathbb{R}^{M' \times M'}$  such that  $\mathbf{A}_k^f$  represents the reasoning probabilities of step  $k$  in test examples. Specifically, there exists some constant  $\rho^f$  in  $(0, 1)$  such that for all  $s \in [l_{ts}^f], k \in [K]$ , the  $i, j$ -th entry of  $\mathbf{A}_k^f$  satisfies

---

<sup>29</sup>The formulation of  $f$  is motivated by recent theoretical works on model training or ICL with Transformers. Please see Appendix F.1.2 for details.

<sup>30</sup>This is to prevent the trivial case that the model only learns the positional encoding but not the TRR patterns when  $\alpha$  becomes arbitrarily close to 1.

$$A_{k(i,j)}^f = \Pr(\text{TSR}(\mathbf{y}_{s,k}) = j | \text{TSR}(\mathbf{y}_{s,k-1}) = i), \quad (7.12)$$

and  $A_{k(i,j^*)}^f \geq 1/(1 - \rho^f) \cdot A_{k(i,j)}^f, \forall j \in [M'],$  where  $\boldsymbol{\mu}'_{j^*} = f_k(\boldsymbol{\mu}'_i),$

Note that (7.12) characterizes a general case in inference that for any given  $k$ , in the  $k$ -th reasoning step of the test example, the  $k$ -th step output is a noisy version of the true label with the highest probability, which guarantees that the examples are overall informative in the  $k$ -th step. This requirement is intuitive because otherwise, these examples would overall provide inaccurate information on the  $k$ -th step reasoning. Moreover, (7.12) models the general case that, with some probability, the  $k$ -step reasoning is inaccurate in the examples.  $\rho^f$  is referred to as the **primacy** of the step-wise transition matrices.  $\rho^f$  reflects the difference in the probability of correct reasoning and incorrect reasoning in each step, and a larger  $\rho^f$  indicates a larger probability of accurate reasoning.

Let  $\mathbf{B}^f = \prod_{k=1}^K \mathbf{A}_k^f$  be the  **$K$ -step transition matrix**. Then  $\mathbf{B}_{(i,j)}^f$  is the probability that the  $K$ -th step output is a noisy version of  $\boldsymbol{\mu}'_j$ , when the input is a noisy version of  $\boldsymbol{\mu}'_i$  in the testing example. We similarly define  $\rho_o^f$  in  $(0, 1)$  as the primacy of  $\mathbf{B}^f$ , where

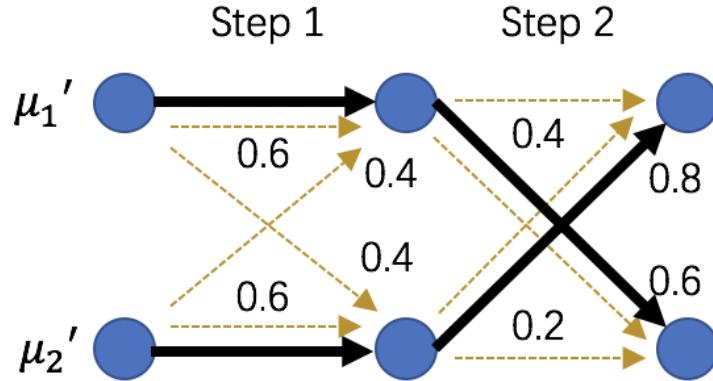
$$B_{(i,j^*)}^f \geq 1/(1 - \rho_o^f) \cdot B_{(i,j)}^f, \quad \forall j \in [M'], \quad j^* = \arg \max_{j \in [M']} B_{(i,j)}^f. \quad (7.13)$$

*Example 1.* Consider a simple two-step inference example with  $K = 2$ ,  $\boldsymbol{\mu}'_1, \boldsymbol{\mu}'_2$  as the TSR pattern, and  $\boldsymbol{\delta} = 0$  in inputs and outputs of every step, as shown in Figure 7.1. The black solid arrows denote the correct inference process, where  $f_1(\boldsymbol{\mu}'_1) = \boldsymbol{\mu}'_1, f_1(\boldsymbol{\mu}'_2) = \boldsymbol{\mu}'_2, f_2(\boldsymbol{\mu}'_1) = \boldsymbol{\mu}'_2,$  and  $f_1(\boldsymbol{\mu}'_2) = \boldsymbol{\mu}'_1.$  Hence,  $\boldsymbol{\mu}'_1 \rightarrow \boldsymbol{\mu}'_1 \rightarrow \boldsymbol{\mu}'_2$  and  $\boldsymbol{\mu}'_2 \rightarrow \boldsymbol{\mu}'_2 \rightarrow \boldsymbol{\mu}'_1$  are two inference sequences under the function  $f.$  The testing examples contain errors and follow the transition matrices  $\mathbf{A}_1^f$  and  $\mathbf{A}_2^f$  (brown dashed arrows). We let  $\mathbf{A}_1^f = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}, \mathbf{A}_2^f = \begin{pmatrix} 0.4 & 0.6 \\ 0.8 & 0.2 \end{pmatrix},$  which results in  $\mathbf{B}^f = \begin{pmatrix} 0.56 & 0.44 \\ 0.64 & 0.36 \end{pmatrix}.$

### 7.3.3 The Sample Complexity Analysis of the Training Stage

We first characterize the convergence and the testing performance of the model during the training stage with sample complexity analysis in Theorem 3.

*Theorem 3.* For any  $\epsilon > 0$ , when (i) the number of context examples in every training sample



**Figure 7.1: An example of a two-step inference.**

is

$$l_{tr} \geq \Omega(\alpha^{-1}), \quad (7.14)$$

(ii) the number of iterations satisfies

$$T \geq \Omega(\eta^{-1}\alpha^{-2}K^3 \log \frac{K}{\epsilon} + \eta^{-1}MK(\alpha^{-1} + \epsilon^{-1})), \quad (7.15)$$

and (iii) the training tasks and samples are selected such that every TRR pattern is equally likely in every inference step and in each training batch<sup>31</sup> with batch size  $B \geq \Omega(\max\{\epsilon^{-2}, M\} \cdot \log M)$ , the step size  $\eta < 1$  and  $N = BT$  samples, then with a high probability, the returned model guarantees

$$\mathbb{E}_{\mathbf{x}_{query} \in \mathcal{M}, f \in \mathcal{T}} [\ell(\Psi; \mathbf{P}, \mathbf{z})] \leq \mathcal{O}(\epsilon). \quad (7.16)$$

Theorem 3 indicates that with long enough training prompts and a sufficient number of iterations and samples for training, a one-layer Transformer can achieve a diminishing loss of  $\mathcal{O}(\epsilon)$  on data following the same distribution as training examples. The results indicate that (i) the required number of context examples is proportional to  $\alpha^{-1}$ ; (ii) the required number of iterations and samples increases as  $M$  and  $\alpha^{-2}$  increases. As a sanity check, these bounds are consistent with the intuition that it will make the training stage more time- and sample-consuming if the number of TRR patterns increases or the fraction of prompt examples that share the same TRR pattern as the query decreases.

---

<sup>31</sup>Our analysis assumes that the whole set of  $\mathcal{M}$  is achievable uniformly in each step and training batch. This condition is to ensure a balanced gradient update among all TRR patterns, as used in [23] for ICL.

### 7.3.4 CoT Generalization Guarantee

In this section, we first define two quantities,  $\tau^f$ , and  $\tau_o^f$  for each testing task  $f \in \mathcal{T}'$  based on the formulation of testing data and tasks in Section 7.3.2. These two quantities are used to characterize the CoT and ICL generalization in Theorems 4 and 5, respectively.

**Definition 7.3.1.** For  $f = f_K \circ \dots \circ f_1 \in \mathcal{T}'$ , we define the **min-max trajectory transition probability** as:

$$\tau^f = \min_{i \in [M']} \prod_{k=1}^K A_{k(\text{TSR}(f_{k-1} \circ \dots \circ f_0(\boldsymbol{\mu}'_i)), \text{TSR}(f_k \circ \dots \circ f_0(\boldsymbol{\mu}'_i)))}^f, \text{ where } f_0(\boldsymbol{\mu}'_i) := \boldsymbol{\mu}'_i, \forall i \in [M'], \quad (7.17)$$

which measures the minimum probability of the most probable  $K$ -step reasoning trajectory over the initial TSR pattern. We also define the **min-max input-label transition probability** as

$$\tau_o^f = \min_{i \in [M']} \max_{j \in [M']} B_{i,j}^f, \quad (7.18)$$

which measures the minimum probability of the most probable output over the initial TSR pattern.

For instance, in Example 1 after (7.13),  $\tau^f = \min\{0.36, 0.48\} = 0.36$ ,  $\tau_o^f = \min\{0.56, 0.64\} = 0.56$ .

*Theorem 4* (CoT generalization). Given a trained model that satisfies conditions (i) to (iii) in Theorem 3, as long as (iv)

$$\boldsymbol{\mu}'_j \in \text{span}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_M), \quad (7.19)$$

for  $j \in [M']$ , and (v) the number of testing examples for every task  $f \in \mathcal{T}'$  is

$$l_{ts}^f \geq \Omega((\alpha' \tau^f \rho^f)^{-2} \log M), \quad (7.20)$$

we have  $\bar{R}_{CoT, \mathbf{x}_{query} \in \mathcal{M}', f \in \mathcal{T}'}^f(\Psi) = 0$ .

*Remark 4.* Theorem 4 characterizes the sufficient conditions for a trained one-layer Transformer to generate all  $K$ -steps reasoning correctly by CoT for a task  $f$  in  $\mathcal{T}'$ . First, the TSR patterns of a new task in  $\mathcal{T}'$  should be linear combinations of TRR patterns in the training tasks in  $\mathcal{T}$ . Second, the number of context examples should be in the order of  $\alpha'^{-2}$ ,  $\rho_s^{f-2}$ , and  $\tau^{f-2}$ . One can equivalently interpret the decrease in the number of required context examples

to achieve zero CoT error as an improvement of the CoT accuracy with fixed context length. Then, when the fraction  $\alpha'$  of contexts where the TSR pattern of the first step input is the same as the query increases, the contexts become more informative for the query. Thus, the CoT accuracy increases. When  $\rho^f$  and  $\tau^f$  increase, the reasoning labels in the context examples are more likely to be accurate based on their definitions in (7.12) and (7.17), then the CoT accuracy is improved.

### 7.3.5 ICL Generalization and Comparison with CoT

Because only input-label pairs are used as context examples without intermediate reasoning steps for ICL, then the input-label pairs in context examples should be accurate on average. Otherwise, the context examples are not informative about the task and will lead to the failure of ICL. We formulate this requirement as Condition 7.3.2.

**Condition 7.3.2.** For the testing task  $f = f_K \circ \dots \circ f_1 \in \mathcal{T}'$ , we have that for any  $i \in [M']$ ,

$$\text{TSR}(f(\boldsymbol{\mu}'_i)) = \arg \max_{j \in [M']} B_{(i,j)}^f. \quad (7.21)$$

Condition 7.3.2 requires that in a context example, if the input TSR is  $\boldsymbol{\mu}'_i$ , then the output TSR needs to be  $f(\boldsymbol{\mu}'_i)$  with the largest probability over all other TSR patterns. It is intuitive that the success of ICL requires this condition. Note that although (7.12) indicates that,  $A_k^f(i, j^*)$  achieves the largest value for all  $j$  when  $\boldsymbol{\mu}'_{j^*} = f_k(\boldsymbol{\mu}'_i)$  for every  $k$  and  $i$ , (7.12) does not always lead to (7.21). Example 1 demonstrates a case where Condition 7.3.2 does not hold. Given the input  $\boldsymbol{\mu}'_1$ , the correct inference output shall be  $\boldsymbol{\mu}'_2$ , but  $\mathbf{B}^f$  indicates that the most probable output is  $\boldsymbol{\mu}'_1$ .

Our result of the ICL generalization is stated as follows.

*Theorem 5* (ICL generalization). Given a trained model that satisfies conditions (i) to (iii) of Theorem 3 and (7.19), for the testing task  $f \in \mathcal{T}'$ ,

- a. if Condition 7.3.2 does not hold, then  $\bar{R}_{ICL, \mathbf{x}_{query} \in \mathcal{M}', f \in \mathcal{T}'}^f(\Psi) \geq \Omega(1)$ ;
- b. if Condition 7.3.2 holds, we have  $\bar{R}_{ICL, \mathbf{x}_{query} \in \mathcal{M}', f \in \mathcal{T}'}^f(\Psi) = 0$ , as long as the number of testing examples is

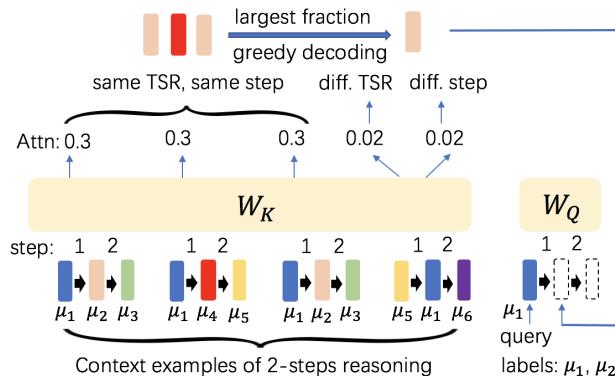
$$l_{ts}^f \geq \Omega((\alpha' \tau_o^f \rho_o^f)^{-2} \log M). \quad (7.22)$$

*Remark 5* (Comparison between CoT and ICL). Theorem 5(a) formally states that, Condition 7.3.2 is necessary for a successful ICL generalization. Because Condition 7.3.2 is not required for CoT generalization, CoT performs better than ICL if Condition 7.3.2 fails<sup>32</sup>. Theorem 5(b) characterizes that when Condition 7.3.2 holds, a desired ICL generalization needs a testing prompt length linear in  $\alpha'^{-2}$ ,  $\rho_o^{f^{-2}}$ , and  $\tau_o^{f^{-2}}$  for the testing task  $f \in \mathcal{T}'$ . This result is the counterpart of the requirement (7.20) for the CoT generalization, indicating that more context examples with the same TSR pattern as the query and more accurate context examples improve ICL generalization.

Ref. [238] also shows the advantage of CoT over ICL to learn MLP functions, but in a different setting from ours, where our studied tasks operate on patterns. More importantly, this chapter characterizes the CoT and ICL performance theoretically when the testing task has a distribution shift from training tasks (TRR patterns to TSR patterns), and the testing examples contain errors, while [238] only empirically evaluates the CoT and ICL performance with noisy examples.

## 7.4 The Mechanism of CoT and the Proof Sketch

### 7.4.1 Transformers Implement CoT by Attending to the Most Similar Examples Every Step



**Figure 7.2: Concentration of attention weights for CoT inference.**

In this section, we characterize the key mechanism of a properly trained one-layer Transformer to implement CoT on a  $K$ -steps reasoning task via training dynamics analysis

<sup>32</sup>Our insight of the comparison between CoT and ICL still holds when we evaluate CoT generalization only by the final step output. This is because a successful CoT generalization in Theorem 4 on all reasoning steps already ensures a satisfactory CoT generalization on the final step.

of the attention layer, as demonstrated in Figure 7.2. This is different from the mechanism study in [238], [239] by constructing a model that can conduct CoT. We have the following proposition for the trained model.

*Proposition 3.* Let  $\mathcal{S}_k^*$  denote the index set of the context columns of the testing prompt  $\mathbf{P}$  in (7.4) that (a) correspond to the  $k$ -th step in a context example and (b) share the same TSR pattern in the  $k$ -th input as the  $k$ -th input  $\mathbf{v}_{k-1}$  of the query,  $k \in [K]$ . Given a trained model that satisfies conditions (i) to (iii) of Theorem 3 and (7.19) and (7.20) after  $T$  iterations, we have

$$\sum_{i \in \mathcal{S}_k^*} \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W}^{(T)} \tilde{\mathbf{q}}_k) \geq 1 - \epsilon, \text{ where } \tilde{\mathbf{p}}_i = \mathbf{p}_i + \mathbf{c}_{(i \mod K)}, \tilde{\mathbf{q}}_k = \mathbf{q}_k + \mathbf{c}_k, \quad (7.23)$$

with  $\mathbf{q}_k$  defined in (7.6). Moreover, for any  $f \in \mathcal{T}'$ , the  $k$ -th step output  $\mathbf{v}_k$  given  $\mathbf{x}_{\text{query}} = \boldsymbol{\mu}'_i$  satisfies,

$$\mathbf{v}_k = f_k \circ \dots \circ f_1(\boldsymbol{\mu}'_i). \quad (7.24)$$

Proposition 3 first illustrates that, when conducting the  $k$ -th step reasoning of the query for any  $k \in [K']$ , the trained model assigns dominant attention weights on the prompt columns that are also the  $k$ -th step reasoning of examples and share the same TSR pattern in the  $k$ -th step input as the query. Then, given a sufficient number of testing context examples by (7.20), it is ensured that the fraction of the correct TSR pattern is the largest in the output of each step by (7.12). Subsequently, the generation by greedy decoding (7.5) is correct in each step, leading to a successful CoT generalization.

#### 7.4.2 An Overview of the Proof

The technical challenges of the proof are concentrated on Theorem 3, where the property of the trained model is derived. The proof of Theorem 3 is built upon three Lemmas, which characterize the **two stages of the training dynamics**. Specifically, Lemmas F.3.5 and F.3.6 show that if a training prompt  $\mathbf{P}$  includes the first  $k$  steps of the reasoning query, then the attention weights on columns of  $\mathbf{P}$  with a different step from the query decrease to be close to zero in the first stage. Lemma F.3.7 computes the gradient updates in the second stage, where the attention weights on columns in  $\mathbf{P}$  that correspond to step  $k$  and have the same TRR pattern as the query gradually become dominant. Theorem 3 unveils this training process by showing the required number of training iterations and sample complexity.

To prove Theorem 4, we first compute the required number of context examples for the new task  $f \in \mathcal{T}'$  so that by concentration inequalities, the number of context examples with accurate TSR is larger than examples with inaccurate TSR patterns in all  $K$  reasoning steps with high probability. Then, due to the linear correlation between TSR and TRR patterns (7.19), we also show that the trained Transformer can attend to context columns with the same TSR pattern as the query. Therefore, the model can make the correct generation in each step. Theorem 5 follows a similar proof idea to Theorem 4, with the difference that the trained model predicts output directly from the input query following  $\mathbf{B}^f$  instead of using  $K$  reasoning steps following  $\mathbf{A}_k^f, k \in [K]$  in CoT. Therefore, Condition 7.3.2 is required for the success of ICL generalization.

## 7.5 Numerical Experiments

**Data Generation and Model setup.** We use synthetic data generated following Sections 7.2 and 7.3.2. Let  $d_{\mathcal{X}} = 30$ ,  $M = 20$ ,  $M' = 10$ ,  $\alpha = 0.4$ . We consider 3-steps tasks for training and testing, i.e.,  $K = 3$ . A reasoning task  $f$  is generated by first sampling a set of numbers of permutations  $\{p_i\}_{i=1}^M$  with  $p_i \in [M]$  and then let  $f_k(\boldsymbol{\mu}_{p_i}) = \boldsymbol{\mu}_{p_{((i+k) \bmod M)}}$  for  $i \in [M], k, j \in [K]$ . The testing noise level is set to be 0.2 for any examples and  $f \in \mathcal{T}'$ . The learning model is a one-layer single-head Transformer defined in (7.2) or a three-layer two-head Transformer. We set  $\tau^f = 0.5$ ,  $\rho^f = 0.8$ ,  $\alpha' = 0.8$  for CoT testing if not otherwise specified.

**Experiments on the generalization of CoT.** We first verify the required number of context examples for a desired CoT generalization on a one-layer Transformer. We investigate the impact of  $\alpha'$ ,  $\tau^f$ , and  $\rho^f$  by varying one and fixing the other two. Figure 7.3 illustrates that more testing examples are needed when  $\alpha'$ ,  $\tau^f$ , or  $\rho^f$  is small, which verifies the trend of the lower bound of  $l_{ts}^f$  in (7.20).

**Experiments on the generalization of ICL and a comparison with CoT.** We then verify the ICL generalization with the trained model. We vary  $\tau_o^f$  and  $\rho_o^f$  by changing  $\tau^f$  and  $\rho^f$ . Figure 7.3 indicates that more testing examples are required when  $\alpha'$ ,  $\tau_o^f$ , or  $\rho_o^f$  is small, which is consistent with our bound in (7.22). We then consider the case where  $\tau_o^f = 0.4$  and  $\rho_o^f = 0.1$  so that the generated testing prompt may not satisfy Condition 7.3.2 depending on the specific choices of  $A_k^f$ 's. Figure 7.5 shows that when Condition 7.3.2 holds, the ICL testing error decreases if the number of contexts increases. However, when Condition 7.3.2

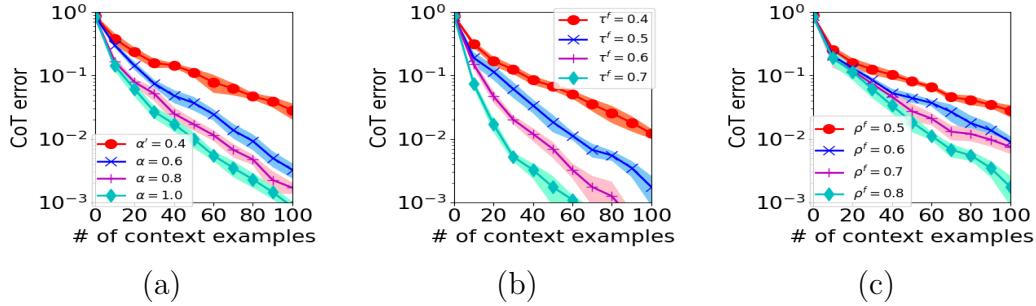


Figure 7.3: CoT testing error with different (a)  $\alpha'$  (b)  $\tau^f$  (c)  $\rho^f$ .

fails, the ICL testing error remains large, irrespective of the number of contexts.

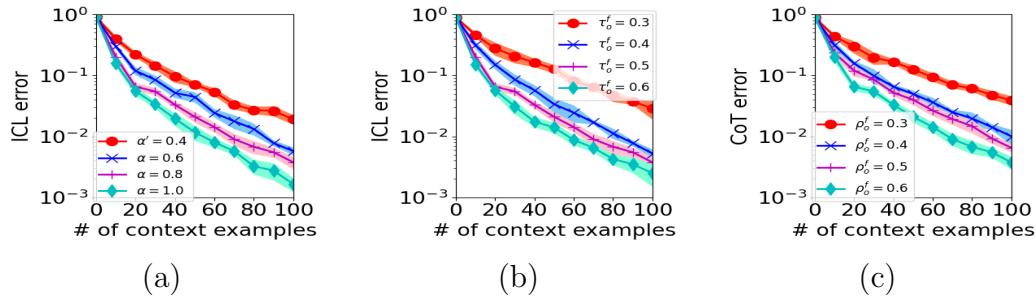
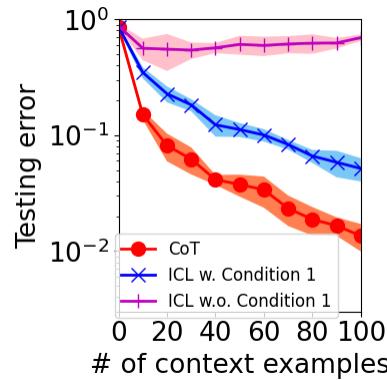


Figure 7.4: ICL testing error with different (a)  $\alpha'$  (b)  $\tau_o^f$  (c)  $\rho_o^f$ .



**Figure 7.5: Comparison between CoT and ICL w./w.o. Condition 7.3.2.**

**Experiments on the training dynamics of CoT.** In Figure 7.6, we compute the total attention weights on four types of testing context columns along the training, which are contexts with the same (or different) TSR pattern and in the same (or different) step as the query. The result shows that the attention weights on contexts that share the same TSR pattern and in the same step as the query increase along the training and converge to around

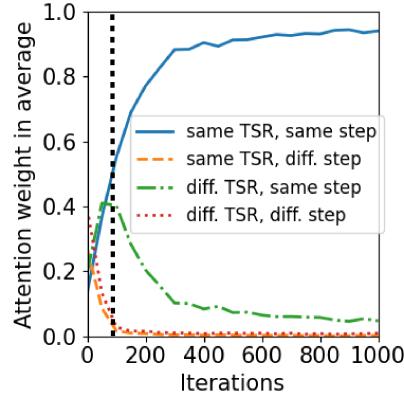


Figure 7.6: Training dynamics of Transformers for CoT.

1. This verifies the mechanism formulated in (7.23). Meanwhile, Figure 7.6 also justifies the two-stage training dynamics proposed in Section 7.4.2, where we add a black vertical dashed line to demonstrate the stage transition boundary. We observe that the attention weights on context columns with a different step, i.e., the red and yellow curves, decrease to zero in the first stage. Then, the attention weights on contexts with the same TSR pattern and the same step as the query, i.e., the blue curve, increase to 1 in the second stage. We also justify the attention mechanism of CoT on a three-layer two-head Transformer with a two-step reasoning task. Figure 7.7 shows that there exists at least one head in each layer of the Transformer that implements CoT as characterized in Proposition 3. This indicates that the CoT mechanism we characterize on one-layer Transformers can be extended to multi-layer multi-head Transformers.

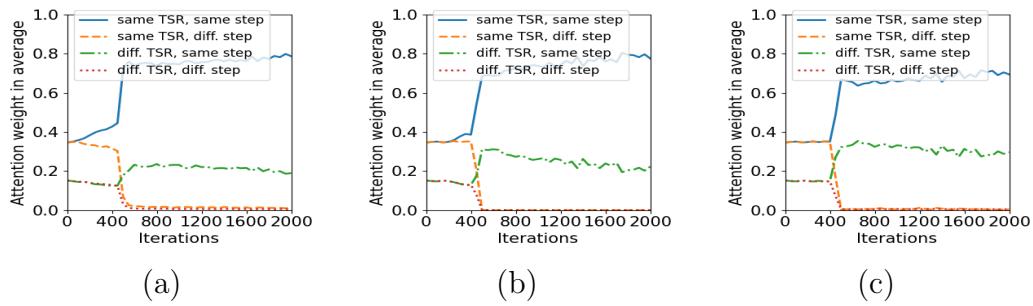


Figure 7.7: Training dynamics of Transformers. (a) Layer 1, Head 2 (b) Layer 2 Head 2 (c) Layer 3 Head 2.

## 7.6 Conclusion, Limitations, and Future Works

This paper theoretically analyzes the training dynamics of Transformers with nonlinear attention, together with the CoT generalization ability of the resulting model on new tasks with noisy and partially inaccurate context examples. We quantitatively characterize and compare the required conditions for the success of CoT and ICL. Although based on a simplified Transformer model and reasoning tasks operating on patterns, this work deepens the theoretical understanding of the CoT mechanism. Future directions include designing efficient prompt-generating methods for CoT and analyzing LLM reasoning on a more complicated data model.

## CHAPTER 8

# CONCLUSION AND FUTURE DIRECTIONS

In this thesis, we conduct a theoretical investigation into the optimization and generalization mechanisms of advanced neural networks, with a particular focus on Transformer-based models, such as Vision Transformers and Graph Transformers. We also analyze how various deep learning algorithms can be effectively applied to different problems. For efficient machine learning algorithms, we examine graph topology sampling in graph convolutional networks, In-Context Learning, and Chain-of-Thought in large language models. Regarding fairness in machine learning, we analyze how data imbalance affects neural network training and generalization.

Future research directions could include analyzing the optimization and generalization mechanisms of Transformer-like models, such as the Mamba model. It has been observed [257] that Mamba can be decomposed into linear attention plus a nonlinear gating mechanism. By analyzing training dynamics, we could gain insights into the data types that models with nonlinear gating can learn well, leading to a comparative understanding of Transformers and Mamba models.

Another direction involves investigating the performance of Transformers on more complex reasoning tasks and settings. For instance, we could explore how Chain-of-Thought performs in complex causal structures, beyond the pattern transition reasoning tasks [24] analyzed by us. Moreover, we could study the effectiveness of training with reasoning examples generated through self-taught methods or similar approaches.

## REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
- [2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Represent.*, 2020, paper 1909.
- [3] V. P. Dwivedi and X. Bresson, “A generalization of transformer networks to graphs,” in *AAAI Workshop Deep Learn. Graphs: Methods Appl.*, 2021, pp. 1–8.
- [4] D. Kreuzer, D. Beaini, W. Hamilton, V. Létourneau, and P. Tossou, “Rethinking graph transformers with spectral attention,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 21 618–21 629.
- [5] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu, “Do transformers really perform badly for graph representation?” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 28 877–28 888.
- [6] H. Li, M. Wang, S. Liu, and P.-Y. Chen, “A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity,” in *Proc. Int. Conf. Learn. Represent.*, 2023, paper 3368.
- [7] H. Li, M. Wang, T. Ma, S. Liu, Z. Zhang, and P.-Y. Chen, “What improves the generalization of graph transformers? a theoretical dive into the self-attention and positional encoding,” in *Proc. Int. Conf. Mach. Learn.*, vol. 235, 2024, pp. 28 784–28 829.
- [8] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and L. Kaiser, “Universal transformers,” in *Proc. Int. Conf. Learn. Represent.*, 2018, paper 910.
- [9] C. Yun, S. Bhojanapalli, A. S. Rawat, S. Reddi, and S. Kumar, “Are transformers universal approximators of sequence-to-sequence functions?” in *Proc. Int. Conf. Learn. Represent.*, 2019, paper 1020.
- [10] S. Bhattacharya, K. Ahuja, and N. Goyal, “On the ability and limitations of transformers to recognize formal languages,” 2020, *arXiv:2009.11264*.
- [11] S. Bhattacharya, A. Patel, and N. Goyal, “On the computational power of transformers and its implications in sequence modeling,” 2020, *arXiv:2006.09286*.
- [12] B. L. Edelman, S. Goel, S. Kakade, and C. Zhang, “Inductive biases and variable creation in self-attention mechanisms,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 5793–5831.

- [13] Y. Dong, J.-B. Cordonnier, and A. Loukas, “Attention is not all you need: Pure attention loses rank doubly exponentially with depth,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2793–2803.
- [14] V. Likhoshesterstov, K. Choromanski, and A. Weller, “On the expressive power of self-attention matrices,” 2021, *arXiv:2106.03764*.
- [15] J.-B. Cordonnier, A. Loukas, and M. Jaggi, “On the relationship between self-attention and convolutional layers,” in *Proc. Int. Conf. Learn. Represent.*, 2019, paper 2040.
- [16] Y. Levine, N. Wies, O. Sharir, H. Bata, and A. Shashua, “Limits to depth efficiencies of self-attention,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22 640–22 651.
- [17] C. Snell, R. Zhong, D. Klein, and J. Steinhardt, “Approximating how single head attention learns,” 2021, *arXiv:2103.07601*.
- [18] C. Wei, Y. Chen, and T. Ma, “Statistically meaningful approximation: a case study on approximating turing machines with transformers,” 2021, *arXiv:2107.13163*.
- [19] J. Chen, K. Gao, G. Li, and K. He, “NAGphormer: A tokenized graph transformer for node classification in large graphs,” in *Proc. Int. Conf. Learn. Represent.*, 2023, paper 3965.
- [20] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini, “Recipe for a general, powerful, scalable graph transformer,” 2022, *arXiv:2205.12454*.
- [21] R. B. Gabrielsson, M. Yurochkin, and J. Solomon, “Rewiring with positional encodings for GNNs,” 2022, *arXiv:2201.12674*.
- [22] H. Li, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, “Generalization guarantee of training graph convolutional networks with graph topology sampling,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 13 014–13 051.
- [23] H. Li, M. Wang, S. Lu, X. Cui, and P.-Y. Chen, “How do nonlinear transformers learn and generalize in in-context learning?” in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 28 734–28 783.
- [24] ——, “Training nonlinear transformers for chain-of-thought inference: A theoretical generalization analysis,” 2024, *arXiv:2410.02167*.
- [25] H. Li, S. Zhang, Y. Zhang, M. Wang, S. Liu, and P.-Y. Chen, “How does promoting the minority fraction affect generalization? a theoretical study of one-hidden-layer neural network on group imbalance,” *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 216–231, Mar. 2024.
- [26] J. Buolamwini and T. Gebru, “Gender shades: intersectional accuracy disparities in commercial gender classification,” in *Proc. Conf. Fairness, Accountability and Transparency*, 2018, pp. 77–91.

- [27] T. McCoy, E. Pavlick, and T. Linzen, “Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference,” in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2019, pp. 3428–3448.
- [28] S. Sagawa, A. Raghunathan, P. W. Koh, and P. Liang, “An investigation of why overparameterization exacerbates spurious correlations,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 8346–8356.
- [29] S. Sagawa\*, P. W. Koh\*, T. B. Hashimoto, and P. Liang, “Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization,” in *Proc. Int. Conf. Learn. Represent.*, 2020, paper 1796.
- [30] H. Yao, Y. Wang, S. Li, L. Zhang, W. Liang, J. Zou, and C. Finn, “Improving out-of-distribution robustness via selective augmentation,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 25407–25437.
- [31] W. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1024–1034.
- [32] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.
- [33] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” in *Proc. Int. Conf. Learn. Represent.*, 2018, paper 164.
- [34] N. Peng, H. Poon, C. Quirk, K. Toutanova, and W.-t. Yih, “Cross-sentence n-ary relation extraction with graph lstms,” *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 101–115, Apr. 2017.
- [35] W. Cong, M. Ramezani, and M. Mahdavi, “On provable benefits of depth in training graph convolutional networks,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9936–9949.
- [36] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, *arXiv:1409.0473*.
- [37] J. D. M.-W. C. Kenton and L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguist.*, 2019, pp. 4171–4186.
- [38] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.
- [39] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, “Deep interest network for click-through rate prediction,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 1059–1068.

- [40] Q. Chen, H. Zhao, W. Li, P. Huang, and W. Ou, “Behavior sequence transformer for e-commerce recommendation in alibaba,” in *Proc. 1st Int. Workshop Deep Learn. Pract. High-Dimensional Sparse Data*, 2019, pp. 1–4.
- [41] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang, “Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer,” in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 1441–1450.
- [42] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, “Decision transformer: Reinforcement learning via sequence modeling,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15 084–15 097.
- [43] M. Janner, Q. Li, and S. Levine, “Offline reinforcement learning as one big sequence modeling problem,” in *Proc. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1273–1286.
- [44] Q. Zheng, A. Zhang, and A. Grover, “Online decision transformer,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 27 042–27 059.
- [45] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10 347–10 357.
- [46] Z.-H. Jiang, Q. Hou, L. Yuan, D. Zhou, Y. Shi, X. Jin, A. Wang, and J. Feng, “All tokens matter: Token labeling for training better vision transformers,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 18 590–18 602.
- [47] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in *Int. Conf. Comput. Vis.*, 2021, pp. 568–578.
- [48] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.
- [49] Z. Pan, B. Zhuang, J. Liu, H. He, and J. Cai, “Scalable vision transformers with hierarchical pooling,” in *Int. Conf. Comput. Vis.*, 2021, pp. 377–386.
- [50] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, “Dynamicvit: Efficient vision transformers with dynamic token sparsification,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13 937–13 949.
- [51] Y. Liang, G. Chongjian, Z. Tong, Y. Song, J. Wang, and P. Xie, “Not all patches are what you need: Expediting vision transformers via token reorganizations,” in *Proc. Int. Conf. Learn. Represent.*, 2022, paper 6168.
- [52] Y. Tang, K. Han, Y. Wang, C. Xu, J. Guo, C. Xu, and D. Tao, “Patch slimming for efficient vision transformers,” in *Int. Conf. Comput. Vis.*, 2022, pp. 12 165–12 174.

- [53] H. Yin, A. Vahdat, J. M. Alvarez, A. Mallya, J. Kautz, and P. Molchanov, “A-vit: Adaptive tokens for efficient vision transformer,” in *Int. Conf. Comput. Vis.*, 2022, pp. 10 809–10 818.
- [54] R. T. d. C. James Vuckovic, Baratin Aristide, “A mathematical theory of attention,” 2020, *arXiv:2007.02876*.
- [55] H. Kim, G. Papamakarios, and A. Mnih, “The lipschitz constant of self-attention,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5562–5571.
- [56] J. Hron, Y. Bahri, J. Sohl-Dickstein, and R. Novak, “Infinite attention: Nngp and ntk for deep attention networks,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4376–4386.
- [57] G. Yang, “Tensor programs ii: Neural tangent kernel for any architecture,” 2020, *arXiv:2006.14548*.
- [58] Z. Allen-Zhu, Y. Li, and Y. Liang, “Learning and generalization in overparameterized neural networks, going beyond two layers,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2019, pp. 6155–6166.
- [59] Z. Allen-Zhu and Y. Li, “What can resnet learn efficiently, going beyond kernels?” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 9017–9028.
- [60] T. Likhomanenko, Q. Xu, G. Synnaeve, R. Collobert, and A. Rogozhnikov, “Cape: Encoding relative positions with continuous augmented positional embeddings,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 16 079–16 092.
- [61] C. Zhu, W. Ping, C. Xiao, M. Shoeybi, T. Goldstein, A. Anandkumar, and B. Catanzaro, “Long-short transformer: Efficient transformers for language and vision,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 17 723–17 736.
- [62] Y. He, W. Liang, D. Zhao, H.-Y. Zhou, W. Ge, Y. Yu, and W. Zhang, “Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning,” in *Int. Conf. Comput. Vis.*, 2022, pp. 9119–9129.
- [63] J. Guibas, M. Mardani, Z. Li, A. Tao, A. Anandkumar, and B. Catanzaro, “Efficient token mixing for transformers via adaptive fourier neural operators,” in *Proc. Int. Conf. Learn. Represent.*, 2022, paper 4047.
- [64] Z. Wang, W. Jiang, Y. M. Zhu, L. Yuan, Y. Song, and W. Liu, “Dynamixer: a vision mlp architecture with dynamic mixing,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 22 691–22 701.
- [65] M. Zhu, Y. Tang, and K. Han, “Vision transformer pruning,” in *KDD 2021 Workshop Model Mining*, 2021, pp. 1–4.
- [66] Z. Liu, Y. Wang, K. Han, W. Zhang, S. Ma, and W. Gao, “Post-training quantization for vision transformer,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 28 092–28 103.

- [67] Y. Lin, T. Zhang, P. Sun, Z. Li, and S. Zhou, “Fq-vit: Post-training quantization for fully quantized vision transformer,” in *Int. Joint Conf. Artif. Intell.*, 2022, pp. 1173–1179.
- [68] Z. Li, T. Yang, P. Wang, and J. Cheng, “Q-vit: Fully differentiable quantization for vision transformer,” 2022, *arXiv:2201.07703*.
- [69] C. Li, B. Zhuang, G. Wang, X. Liang, X. Chang, and Y. Yang, “Automated progressive learning for efficient training of vision transformers,” in *Int. Conf. Comput. Vis.*, 2022, pp. 12 486–12 496.
- [70] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, “Twins: Revisiting the design of spatial attention in vision transformers,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9355–9366.
- [71] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon, “Recovery guarantees for one-hidden-layer neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 4140–4149.
- [72] H. Fu, Y. Chi, and Y. Liang, “Guaranteed recovery of one-hidden-layer neural networks via cross entropy,” *IEEE Trans. Signal Process.*, vol. 68, pp. 3225–3235, Feb. 2020.
- [73] K. Zhong, Z. Song, and I. S. Dhillon, “Learning non-overlapping convolutional neural networks with multiple kernels,” 2017, *arXiv:1711.03440*.
- [74] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, “Fast learning of graph neural networks with guaranteed generalizability: One-hidden-layer case,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11 268–11 277.
- [75] S. Zhang, M. Wang, J. Xiong, S. Liu, and P.-Y. Chen, “Improved linear convergence of training cnns with generalizability guarantees: A one-hidden-layer case,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2622–2635, Jun. 2020.
- [76] H. Li, S. Zhang, and M. Wang, “Learning and generalization of one-hidden-layer neural networks, going beyond standard gaussian data,” in *Annu. Conf. Inf. Sci. Syst.*, 2022, pp. 37–42.
- [77] A. Jacot, F. Gabriel, and C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8571–8580.
- [78] Z. Allen-Zhu, Y. Li, and Z. Song, “A convergence theory for deep learning via over-parameterization,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 242–252.
- [79] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, “Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 322–332.

- [80] Y. Cao and Q. Gu, “Generalization bounds of stochastic gradient descent for wide and deep neural networks,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2019, pp. 10 836–10 846.
- [81] D. Zou and Q. Gu, “An improved analysis of training over-parameterized deep neural networks,” in *Proc. Conf. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 2053–2062.
- [82] S. S. Du, X. Zhai, B. Poczos, and A. Singh, “Gradient descent provably optimizes over-parameterized neural networks,” in *Proc. Int. Conf. Learn. Represent.*, 2019, paper 729.
- [83] Z. Chen, Y. Cao, Q. Gu, and T. Zhang, “A generalized neural tangent kernel analysis for two-layer neural networks,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 13 363–13 373.
- [84] Y. Li and Y. Liang, “Learning overparameterized neural networks via stochastic gradient descent on structured data,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2018, pp. 8157–8166.
- [85] A. Daniely and E. Malach, “Learning parities with neural networks,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 20 356–20 365.
- [86] Z. Shi, J. Wei, and Y. Liang, “A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features,” in *Proc. Int. Conf. Learn. Represent.*, 2021, paper 1676.
- [87] S. Karp, E. Winston, Y. Li, and A. Singh, “Local signal adaptivity: Provable feature learning in neural networks beyond kernels,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24 883–24 897.
- [88] A. Brutzkus and A. Globerson, “An optimization and generalization analysis for max-pooling networks,” in *Proc. Conf. Uncertainty Artif. Intell.*, 2021, pp. 1650–1660.
- [89] S. Zhang, M. Wang, P.-Y. Chen, S. Liu, S. Lu, and M. Liu, “Joint edge-model sparse learning is provably efficient for graph neural networks,” in *Proc. Int. Conf. Learn. Represent.*, 2023, paper 1109.
- [90] Z. Allen-Zhu and Y. Li, “Feature purification: How adversarial training performs robust deep learning,” in *Annu. Symp. Found. Comput. Sci.*, 2022, pp. 977–988.
- [91] Z. Wen and Y. Li, “Toward understanding the feature learning process of self-supervised contrastive learning,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11 112–11 122.
- [92] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

- [93] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, and J. Huang, “Self-supervised graph transformer on large-scale molecular data,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12 559–12 571.
- [94] Z. Wu, P. Jain, M. Wright, A. Mirhoseini, J. E. Gonzalez, and I. Stoica, “Representing long-range context for graph neural networks with global attention,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13 266–13 279.
- [95] H. Zhang and J. Zhang, “Text graph transformer for document classification,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 8322–8327.
- [96] Z. Hu, Y. Dong, K. Wang, and Y. Sun, “Heterogeneous graph transformer,” in *Proc. Int. World Wide Web Conf.*, 2020, pp. 2704–2710.
- [97] Z. Hu, Y. Dong, K. Wang, K.-W. Chang, and Y. Sun, “Gpt-gnn: generative pre-training of graph neural networks,” in *Proc. Int. Conf. Knowl. Discovery Data Mining*, 2020, pp. 1857–1867.
- [98] Z. ZHANG, Q. Liu, Q. Hu, and C.-K. Lee, “Hierarchical graph transformer with adaptive node sampling,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2022, pp. 21 171–21 183.
- [99] M. S. Hussain, M. J. Zaki, and D. Subramanian, “Global self-attention as a replacement for graph convolution,” in *Proc. ACM SIGKDD Int. Conf. Knowl. discovery Data mining*, 2022, pp. 655–665.
- [100] J. Zhao, C. Li, Q. Wen, Y. Wang, Y. Liu, H. Sun, X. Xie, and Y. Ye, “Gophormer: ego-graph transformer for node classification,” 2021, *arXiv:2110.13094*.
- [101] D. Chen, L. O’Bray, and K. Borgwardt, “Structure-aware transformer for graph representation learning,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 3469–3489.
- [102] S. Maskey, R. Levie, Y. Lee, and G. Kutyniok, “Generalization analysis of message passing neural networks on large random graphs,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 4805–4817.
- [103] H. Tang and Y. Liu, “Towards understanding the generalization of graph neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 33 674–33 719.
- [104] Y. Tian, Y. Wang, B. Chen, and S. Du, “Scan and snap: Understanding training dynamics and token composition in 1-layer transformer,” 2023, *arXiv:2305.16380*.
- [105] B. Zhang, S. Luo, L. Wang, and D. He, “Rethinking the expressive power of GNNs via graph biconnectivity,” in *Proc. Int. Conf. Learn. Represent.*, 2023, paper 453.
- [106] S. Verma and Z.-L. Zhang, “Stability and generalization of graph convolutional neural networks,” in *Proc. Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 1539–1548.
- [107] X. Zhou and H. Wang, “The generalization error of graph convolutional networks may enlarge with more layers,” *Neurocomputing*, vol. 424, pp. 97–106, Feb. 2021.

- [108] R. Liao, R. Urtasun, and R. Zemel, “A pac-bayesian approach to generalization bounds for graph neural networks,” in *Proc. Int. Conf. Learn. Represent.*, 2021, paper 2733.
- [109] V. Garg, S. Jegelka, and T. Jaakkola, “Generalization and representational limits of graph neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3419–3430.
- [110] K. Oono and T. Suzuki, “Optimization and generalization analysis of transduction through gradient boosting and application to multi-scale graph neural networks,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18 917–18 930.
- [111] M. N. R. Chowdhury, S. Zhang, M. Wang, S. Liu, and P.-Y. Chen, “Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 6074–6114.
- [112] Z. Wen and Y. Li, “The mechanism of prediction head in non-contrastive self-supervised learning,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2022, pp. 24 794–24 809.
- [113] Z. Allen-Zhu and Y. Li, “Towards understanding ensemble, knowledge distillation and self-distillation in deep learning,” in *Proc. Int. Conf. Learn. Represent.*, 2023, paper 5565.
- [114] S. Jelassi, M. Sander, and Y. Li, “Vision transformers provably learn spatial structure,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 37 822–37 836.
- [115] S. Oymak, A. S. Rawat, M. Soltanolkotabi, and C. Thrampoulidis, “On the role of attention in prompt-tuning,” 2023, *arXiv:2306.03435*.
- [116] Y. Li, Y. Li, and A. Risteski, “How do transformers learn topic structure: Towards a mechanistic understanding,” 2023, *arXiv:2303.04245*.
- [117] H. Li, M. Wang, S. Lu, H. Wan, X. Cui, and P.-Y. Chen, “Transformers as multi-task feature selectors: Generalization analysis of in-context learning,” in *Proc. NeurIPS Workshop Math. Mod. Mach. Learn.*, 2023, paper 70.
- [118] Y. Luo, “Transformers for capturing multi-level graph structure using hierarchical distances,” 2023, *arXiv:2308.11129*.
- [119] D. Ataee Tarzanagh, Y. Li, X. Zhang, and S. Oymak, “Max-margin token selection in attention mechanism,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 48 314–48 362.
- [120] Q. Wu, W. Zhao, Z. Li, D. P. Wipf, and J. Yan, “Nodeformer: A scalable graph structure learning transformer for node classification,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27 387–27 401.
- [121] Z. Zhang, X. Wang, C. Guan, Z. Zhang, H. Li, and W. Zhu, “Autogt: Automated graph transformer architecture search,” in *Proc. Int. Conf. Learn. Represent.*, 2023, paper 1210.

- [122] J. Zhang, H. Zhang, C. Xia, and L. Sun, “Graph-bert: Only attention is needed for learning graph representations,” 2020, *arXiv:2001.05140*.
- [123] Y. Liu, S. Yang, C. Lei, G. Wang, H. Tang, J. Zhang, A. Sun, and C. Miao, “Pre-training graph transformer with multimodal side information for recommendation,” in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 2853–2861.
- [124] H. Nt and T. Maehara, “Revisiting graph neural networks: All we have is low-pass filters,” 2019, *arXiv:1905.09550*.
- [125] E. Chien, J. Peng, P. Li, and O. Milenkovic, “Adaptive universal generalized pagerank graph neural network,” in *Proc. Int. Conf. Learn. Represent.*, 2021, paper 2940.
- [126] V. P. Dwivedi, L. Rampasek, M. Galkin, A. Parviz, G. Wolf, A. T. Luu, and D. Beaini, “Long range graph benchmark,” in *Proc. Conf. Neural Inf. Process. Syst. Datasets and Benchmarks Track*, 2022, paper 136.
- [127] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, “Open graph benchmark: Datasets for machine learning on graphs,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22118–22133.
- [128] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *J. Big Data*, vol. 6, no. 1, pp. 1–48, Jul. 2019.
- [129] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
- [130] F. J. Moreno-Barea, F. Strazzeri, J. M. Jerez, D. Urda, and L. Franco, “Forward noise adjustment scheme for data augmentation,” in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2018, pp. 728–734.
- [131] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [132] S. S. Du, J. D. Lee, and Y. Tian, “When is a convolutional filter easy to learn?” in *Proc. Int. Conf. Learn. Represent.*, 2018, paper 88.
- [133] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, “When do neural networks outperform kernel methods?” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 14820–14830.
- [134] S. Mei, A. Montanari, and P.-M. Nguyen, “A mean field view of the landscape of two-layer neural networks,” *Proc. Natl Acad. Sci.*, vol. 115, no. 33, pp. E7665–E7671, Jun. 2018.
- [135] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Int. Conf. Comput. Vis.*, 2015, pp. 3730–3738.

- [136] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
- [137] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, “Efficient backprop,” in *Neural Netw.: Tricks Trade.*, 1998, pp. 9–50.
- [138] L. M. Koch, C. M. Schürch, A. Gretton, and P. Berens, “Hidden in plain sight: Subgroup shifts escape OOD detection,” in *Proc. Med. Imaging Deep Learn.*, 2022, pp. 726–740.
- [139] J. Ma, J. Deng, and Q. Mei, “Subgroup generalization and fairness of graph neural networks,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2021, pp. 1048–1061.
- [140] A. Biswas and S. Mukherjee, “Ensuring fairness under prior probability shifts,” in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2021, pp. 414–424.
- [141] S. Giguere, B. Metevier, Y. Brun, P. S. Thomas, S. Niekum, and B. C. da Silva, “Fairness guarantees under demographic shift,” in *Proc. Int. Conf. Learn. Represent.*, 2022, paper 1094.
- [142] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [143] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications,” *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.
- [144] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018.
- [145] J. Byrd and Z. Lipton, “What is the effect of importance weighting in deep learning?” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 872–881.
- [146] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel, “Counterfactual fairness in text classification through robustness,” in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2019, pp. 219–226.
- [147] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 325–333.
- [148] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
- [149] N. Sarafianos, X. Xu, and I. A. Kakadiaris, “Deep imbalanced attribute classification using visual attention aggregation,” in *Eur. Conf. Comput. Vis.*, 2018, pp. 680–697.
- [150] S. S. Mullick, S. Datta, and S. Das, “Generative adversarial minority oversampling,” in *Int. Conf. Comput. Vis.*, 2019, pp. 1695–1704.

- [151] J. Kim, J. Jeong, and J. Shin, “M2m: Imbalanced classification via major-to-minor translation,” in *Int. Conf. Comput. Vis.*, 2020, pp. 13 896–13 905.
- [152] P. Chu, X. Bian, S. Liu, and H. Ling, “Feature space augmentation for long-tailed data,” in *Eur. Conf. Comput. Vis.*, 2020, pp. 694–710.
- [153] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, “Metasaug: Meta semantic augmentation for long-tailed visual recognition,” in *Int. Conf. Comput. Vis.*, 2021, pp. 5212–5221.
- [154] C. Fang, H. He, Q. Long, and W. J. Su, “Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training,” *Proc. Nat. Acad. Sci.*, vol. 118, no. 43, Aug. 2021.
- [155] L. Yang, H. Jiang, Q. Song, and J. Guo, “A survey on long-tailed visual recognition,” *Int. J. Comput. Vis.*, vol. 130, no. 7, pp. 1837–1872, May 2022.
- [156] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi, “The majority can help the minority: Context-rich minority oversampling for long-tailed classification,” in *Int. Conf. Comput. Vis.*, 2022, pp. 6887–6896.
- [157] S. S. Du, J. D. Lee, Y. Tian, A. Singh, and B. Poczos, “Gradient descent learns one-hidden-layer cnn: Don’t be afraid of spurious local minima,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1338–1347.
- [158] R. Ge, J. D. Lee, and T. Ma, “Learning one-hidden-layer neural networks with landscape design,” in *Proc. Int. Conf. Learn. Represent.*, 2018, paper 697.
- [159] Y. Li and Y. Yuan, “Convergence analysis of two-layer neural networks with ReLU activation,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 597–607.
- [160] I. Safran and O. Shamir, “Spurious local minima are common in two-layer relu neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4430–4438.
- [161] X. Zhang, Y. Yu, L. Wang, and Q. Gu, “Learning one-hidden-layer relu networks via gradient descent,” in *Proc. Conf. Artif. Intell. Statist.*, 2019, pp. 1524–1534.
- [162] S. Zhang, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, “Why lottery ticket wins? a theoretical perspective of sample complexity on sparse neural networks,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2707–2720.
- [163] ——, “How unlabeled data improve generalization in self-training? a one-hidden-layer theoretical analysis,” in *Proc. Int. Conf. Learn. Represent.*, 2021, paper 1970.
- [164] Y. Yoshida and M. Okada, “Data-dependence of plateau phenomenon in learning with neural network — statistical mechanical analysis,” in *Proc. Conf. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1722–1730.

- [165] F. Mignacco, F. Krzakala, P. Urbani, and L. Zdeborová, “Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9540–9550.
- [166] D. Hsu and S. M. Kakade, “Learning mixtures of spherical gaussians: moment methods and spectral decompositions,” in *Proc. 4th Conf. Innov. Theor. Comput. Sci.*, 2013, pp. 11–20.
- [167] A. Moitra and G. Valiant, “Settling the polynomial learnability of mixtures of gaussians,” in *Proc. IEEE Annu. Symp. Found. Comput. Sci.*, 2010, pp. 93–102.
- [168] O. Regev and A. Vijayaraghavan, “On learning mixtures of well-separated gaussians,” in *Annu. Symp. Found. Comput. Sci.*, 2017, pp. 85–96.
- [169] N. Ho and X. Nguyen, “Convergence rates of parameter estimation for some weakly identifiable finite mixtures,” *Annu. Statist.*, vol. 44, no. 6, pp. 2726–2755, Dec. 2016.
- [170] R. Dwivedi, N. Ho, K. Khamaru, M. I. Jordan, M. J. Wainwright, and B. Yu, “Singularity, misspecification, and the convergence rate of em,” *Annu. Statist.*, vol. 48, no. 6, pp. 3161–3182, Dec. 2020.
- [171] R. Dwivedi, N. Ho, K. Khamaru, M. Wainwright, M. Jordan, and B. Yu, “Sharp analysis of expectation-maximization for weakly identifiable models,” in *Proc. Conf. Artif. Intell. Statist.*, 2020, pp. 1866–1876.
- [172] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Int. Conf. Comput. Vis.*, 2016, pp. 770–778.
- [173] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer, “Shift: A zero flop, zero parameter alternative to spatial convolutions,” in *Int. Conf. Comput. Vis.*, 2018, pp. 9127–9135.
- [174] A. Dutta, E. H. Bergou, A. M. Abdelmoniem, C.-Y. Ho, A. N. Sahu, M. Canini, and P. Kalnis, “On the discrepancy between the theoretical analysis and practical implementations of compressed communication for distributed deep learning,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 3817–3824.
- [175] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in *Proc. Int. Conf. Learn. Represent.*, 2018, paper 164.
- [176] V. G. Satorras and J. B. Estrach, “Few-shot learning with graph neural networks,” in *Proc. Int. Conf. Learn. Represent.*, 2018, paper 910.
- [177] X. Wang, Y. Ye, and A. Gupta, “Zero-shot recognition via semantic embeddings and knowledge graphs,” in *Int. Conf. Comput. Vis.*, 2018, pp. 6857–6866.
- [178] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, “Relation networks for object detection,” in *Int. Conf. Comput. Vis.*, 2018, pp. 3588–3597.

- [179] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, “Graph convolutional neural networks for web-scale recommender systems,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 974–983.
- [180] R. v. d. Berg, T. N. Kipf, and M. Welling, “Graph convolutional matrix completion,” 2017, *arXiv:1706.02263*.
- [181] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende *et al.*, “Interaction networks for learning about objects, relations and physics,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2016, pp. 4502–4510.
- [182] A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell, and P. Battaglia, “Graph networks as learnable physics engines for inference and control,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4470–4479.
- [183] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, “Convolutional networks on graphs for learning molecular fingerprints,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2224–2232.
- [184] J. Chen, J. Zhu, and L. Song, “Stochastic training of graph convolutional networks with variance reduction,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 942–950.
- [185] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, “Cluster-gcn: an efficient algorithm for training deep and large graph convolutional networks,” in *Proc. Int. Conf. Knowl. Discovery Data Mining*, 2019, pp. 257–266.
- [186] J. Chen, T. Ma, and C. Xiao, “Fastgcn: Fast learning with graph convolutional networks via importance sampling,” in *Proc. Int. Conf. Learn. Represent.*, 2018, paper 613.
- [187] D. Zou, Z. Hu, Y. Wang, S. Jiang, Y. Sun, and Q. Gu, “Layer-dependent importance sampling for training deep and large graph convolutional networks,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 11 249–11 259.
- [188] C. Zheng, B. Zong, W. Cheng, D. Song, J. Ni, W. Yu, H. Chen, and W. Wang, “Robust graph representation learning via neural sparsification,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11 458–11 468.
- [189] J. Li, T. Zhang, H. Tian, S. Jin, M. Fardad, and R. Zafarani, “Sgcn: A graph sparsifier based on graph convolutional networks,” in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2020, pp. 275–287.
- [190] T. Chen, Y. Sui, X. Chen, A. Zhang, and Z. Wang, “A unified lottery ticket hypothesis for graph neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 1695–1706.
- [191] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?” in *Proc. Int. Conf. Learn. Represent.*, 2019, paper 835.

- [192] K. Xu, M. Zhang, S. Jegelka, and K. Kawaguchi, “Optimization of graph neural networks: Implicit acceleration by skip connections and more depth,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11 592–11 602.
- [193] S. Lv, “Generalization bounds for graph convolutional neural networks via rademacher complexity,” 2021, *arXiv:2102.10234*.
- [194] S. S. Du, K. Hou, R. R. Salakhutdinov, B. Poczos, R. Wang, and K. Xu, “Graph neural tangent kernel: Fusing graph neural networks with graph kernels,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2019, pp. 5724–5734.
- [195] A. Daniely, “Sgd learns the conjugate kernel class of the network,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 2422–2430.
- [196] M. Ramezani, W. Cong, M. Mahdavi, A. Sivasubramaniam, and M. Kandemir, “Gcn meets gpu: Decoupling “when to sample” from “how to sample”,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18 482–18 492.
- [197] R. Ge, F. Huang, C. Jin, and Y. Yuan, “Escaping from saddle points—online stochastic gradient for tensor decomposition,” in *Conf. Learn. Theory*, 2015, pp. 797–842.
- [198] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
- [199] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, “Palm: Scaling language modeling with pathways,” 2022, *arXiv:2204.02311*.
- [200] H. Touvron, T. Lavigil, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” 2023, *arXiv:2302.13971*.
- [201] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [202] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.
- [203] S. Garg, D. Tsipras, P. S. Liang, and G. Valiant, “What can transformers learn in-context? a case study of simple function classes,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 30 583–30 598.
- [204] J. Liu, D. Shen, Y. Zhang, W. B. Dolan, L. Carin, and W. Chen, “What makes good in-context examples for gpt-3?” in *Proc. Deep Learn. Inside Out Workshop (DeeLIO)*, 2022, pp. 100–114.

- [205] Z. Wu, Y. Wang, J. Ye, and L. Kong, “Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering,” in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2023, pp. 1423–1436.
- [206] A. Liu, S. Swayamdipta, N. A. Smith, and Y. Choi, “Wanli: Worker and ai collaboration for natural language inference dataset creation,” in *Findings Assoc. Comput. Linguist.: EMNLP*, 2022, pp. 6826–6847.
- [207] L. Lucy and D. Bamman, “Gender and representation bias in gpt-3 generated stories,” in *Proc. 3rd Workshop on Narrative Understanding*, 2021, pp. 48–55.
- [208] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” 2015, *arXiv:1510.00149*.
- [209] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, “Learning structured sparsity in deep neural networks,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 2082–2090.
- [210] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, “Pruning convolutional neural networks for resource efficient inference,” in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–17.
- [211] J.-H. Luo, J. Wu, and W. Lin, “Thinet: A filter level pruning method for deep neural network compression,” in *Int. Conf. Comput. Vis.*, 2017, pp. 5058–5066.
- [212] E. Frantar and D. Alistarh, “Sparsegpt: Massive language models can be accurately pruned in one-shot,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 10 323–10 337.
- [213] X. Ma, G. Fang, and X. Wang, “LLM-pruner: On the structural pruning of large language models,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2023, pp. 21 702–21 720.
- [214] M. Sun, Z. Liu, A. Bair, and J. Z. Kolter, “A simple and effective pruning approach for large language models,” 2023, *arXiv:2306.11695*.
- [215] Z. Liu, J. Wang, T. Dao, T. Zhou, B. Yuan, Z. Song, A. Shrivastava, C. Zhang, Y. Tian, C. Re *et al.*, “Deja vu: Contextual sparsity for efficient llms at inference time,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 22 137–22 176.
- [216] Y. Li, M. E. Ildiz, D. Papailiopoulos, and S. Oymak, “Transformers as algorithms: Generalization and stability in in-context learning,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19 565–19 594.
- [217] R. Zhang, S. Frei, and P. L. Bartlett, “Trained transformers learn linear models in-context,” 2023, *arXiv:2306.09927*.
- [218] J. Wu, D. Zou, Z. Chen, V. Braverman, Q. Gu, and P. L. Bartlett, “How many pretraining tasks are needed for in-context learning of linear regression?” 2023, *arXiv:2310.08391*.

- [219] Y. Huang, Y. Cheng, and Y. Liang, “In-context convergence of transformers,” in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 19 660–19 722.
- [220] E. Akyürek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou, “What learning algorithm is in-context learning? investigations with linear models,” in *Proc. Int. Conf. Learn. Represent.*, 2023, paper 5613.
- [221] J. Von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov, “Transformers learn in-context by gradient descent,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 35 151–35 174.
- [222] K. Ahn, X. Cheng, H. Daneshmand, and S. Sra, “Transformers learn to implement preconditioned gradient descent for in-context learning,” 2023, *arXiv:2306.00297*.
- [223] X. Cheng, Y. Chen, and S. Sra, “Transformers implement functional gradient descent to learn non-linear functions in context,” 2023, *arXiv:2312.06528*.
- [224] Y. Bai, F. Chen, H. Wang, C. Xiong, and S. Mei, “Transformers as statisticians: Provable in-context learning with in-context algorithm selection,” 2023, *arXiv:2306.04637*.
- [225] T. Guo, W. Hu, S. Mei, H. Wang, C. Xiong, S. Savarese, and Y. Bai, “How do transformers learn in-context beyond simple functions? a case study on learning with representations,” 2023, *arXiv:2310.10616*.
- [226] D. A. Tarzanagh, Y. Li, C. Thrampoulidis, and S. Oymak, “Transformers as support vector machines,” 2023, *arXiv:2308.16898*.
- [227] Y. Tian, Y. Wang, Z. Zhang, B. Chen, and S. S. Du, “JoMA: Demystifying multilayer transformers via joint dynamics of MLP and attention,” in *Proc. Int. Conf. Learn. Represent.*, 2024, paper 4127.
- [228] H. Yang and Z. Wang, “On the neural tangent kernel analysis of randomly pruned neural networks,” in *Proc. Conf. Artif. Intell. statist.*, 2023, pp. 1513–1553.
- [229] H. Yang, Y. Liang, X. Guo, L. Wu, and Z. Wang, “Theoretical characterization of how neural network pruning affects its generalization,” 2023, *arXiv:2301.00335*.
- [230] O. Rubin, J. Herzig, and J. Berant, “Learning to retrieve prompts for in-context learning,” in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguist.*, 2022, pp. 2655–2671.
- [231] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” 2023, *arXiv:2307.09288*.
- [232] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao *et al.*, “Sora: A review on background, technology, limitations, and opportunities of large vision models,” 2024, *arXiv:2402.17177*.

- [233] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24 824–24 837.
- [234] L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K.-W. Lee, and E.-P. Lim, “Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models,” in *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2023, pp. 2609–2634.
- [235] Z. Zhang, A. Zhang, M. Li, and A. Smola, “Automatic chain of thought prompting in large language models,” in *Proc. Int. Conf. Learn. Represent.*, 2023, paper 3497.
- [236] X. Wang and D. Zhou, “Chain-of-thought reasoning without prompting,” 2024, *arXiv:2402.10200*.
- [237] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. V. Le *et al.*, “Least-to-most prompting enables complex reasoning in large language models,” in *Proc. Int. Conf. Learn. Represent.*, 2023, paper 929.
- [238] Y. Li, K. Sreenivasan, A. Giannou, D. Papailiopoulos, and S. Oymak, “Dissecting chain-of-thought: Compositionality through in-context filtering and learning,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 22 021–22 046.
- [239] G. Feng, B. Zhang, Y. Gu, H. Ye, D. He, and L. Wang, “Towards revealing the mystery behind chain of thought: a theoretical perspective,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 70 757–70 798.
- [240] Z. Li, H. Liu, D. Zhou, and T. Ma, “Chain of thought empowers transformers to solve inherently serial problems,” in *Proc. Int. Conf. Learn. Represent.*, 2024, paper 8645.
- [241] K. Yang, J. Ackermann, Z. He, G. Feng, B. Zhang, Y. Feng, Q. Ye, D. He, and L. Wang, “Do efficient transformers really save computation?” 2024, *arXiv:2402.13934*.
- [242] K. Wen, X. Dang, and K. Lyu, “Rnns are not transformers (yet): The key bottleneck on in-context retrieval,” 2024, *arXiv:2402.18510*.
- [243] W. Merrill and A. Sabharwal, “The expressive power of transformers with chain of thought,” in *Proc. Int. Conf. Learn. Represent.*, 2024, paper 1377.
- [244] N. Ding, T. Levinboim, J. Wu, S. Goodman, and R. Soricut, “CausalLM is not optimal for in-context learning,” in *Proc. Int. Conf. Learn. Represent.*, 2024, paper 327.
- [245] Y. Cui, J. Ren, P. He, J. Tang, and Y. Xing, “Superiority of multi-head attention in in-context linear regression,” 2024, *arXiv:2401.17426*.
- [246] S. Chen, H. Sheen, T. Wang, and Z. Yang, “Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality,” 2024, *arXiv:2402.19442*.
- [247] H. Li, M. Wang, S. Zhang, S. Liu, and P.-Y. Chen, “Learning on transformers is provable low-rank and sparse: A one-layer analysis,” 2024, *arXiv:2406.17167*.

- [248] Y. Huang, Z. Wen, Y. Chi, and Y. Liang, “Transformers provably learn feature-position correlations in masked image modeling,” 2024, *arXiv:2403.02233*.
- [249] Y. Li, Y. Huang, M. E. Ildiz, A. S. Rawat, and S. Oymak, “Mechanics of next token prediction with self-attention,” in *Proc. Conf. Artif. Intell. Statist.*, 2024, pp. 685–693.
- [250] M. E. Ildiz, Y. Huang, Y. Li, A. S. Rawat, and S. Oymak, “From self-attention to markov models: Unveiling the dynamics of generative transformers,” 2024, *arXiv:2402.13512*.
- [251] E. Nichani, A. Damian, and J. D. Lee, “How transformers learn causal structure with gradient descent,” 2024, *arXiv:2402.14735*.
- [252] A. V. Makkula, M. Bondaschi, A. Girish, A. Nagle, M. Jaggi, H. Kim, and M. Gastpar, “Attention with markov: A framework for principled analysis of transformers via markov chains,” 2024, *arXiv:2402.04161*.
- [253] P. Deora, R. Ghaderi, H. Taheri, and C. Thrampoulidis, “On the optimization and generalization of multi-head attention,” 2023, *arXiv:2310.12680*.
- [254] S. Chen and Y. Li, “Provably learning a multi-head attention layer,” 2024, *arXiv:2402.04084*.
- [255] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, “Let’s verify step by step,” in *Proc. Int. Conf. Learn. Represent.*, 2024, paper 2762.
- [256] J. Park, J. Park, Z. Xiong, N. Lee, J. Cho, S. Oymak, K. Lee, and D. Papailiopoulos, “Can mamba learn how to learn? a comparative study on in-context learning tasks,” 2024, *arXiv:2402.04248*.
- [257] D. Han, Z. Wang, Z. Xia, Y. Han, Y. Pu, C. Ge, J. Song, S. Song, B. Zheng, and G. Huang, “Demystify mamba in vision: A linear attention perspective,” 2024, *arXiv:2405.16605*.
- [258] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” 2010, *arXiv:1011.3027*.
- [259] S. Li, “Concise formulas for the area and volume of a hyperspherical cap,” *Asian J. Math. Statist.*, vol. 4, no. 1, pp. 66–70, Jan. 2010.
- [260] S. Zhang, H. Li, M. Wang, M. Liu, P.-Y. Chen, S. Lu, S. Liu, K. Murugesan, and S. Chaudhury, “On the convergence and sample complexity analysis of deep q-networks with  $\epsilon$ -greedy exploration,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2023, pp. 13 064–13 102.
- [261] Y. Huang, Y. Zeng, Q. Wu, and L. Lü, “Higher-order graph convolutional network with flower-petals laplacians on simplicial complexes,” 2023, *arXiv:2309.12971*.

- [262] H. Pei, B. Wei, K. C.-C. Chang, Y. Lei, and B. Yang, “Geom-gcn: Geometric graph convolutional networks,” in *Proc. Int. Conf. Learn. Represent.*, 2020, paper 2589.
- [263] U. Von Luxburg, “A tutorial on spectral clustering,” *Statist. Comput.*, vol. 17, pp. 395–416, Aug. 2007.
- [264] I. Tolstikhin, G. Blanchard, and M. Kloft, “Localized complexities for transductive learning,” in *Conf. Learn. Theory*, 2014, pp. 857–884.
- [265] P. Esser, L. Chennuru Vankadara, and D. Ghoshdastidar, “Learning theory can (sometimes) explain generalisation in graph neural networks,” in *Annu. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 27043–27056.
- [266] M. Janzamin, H. Sedghi, and A. Anandkumar, “Score function features for discriminative learning: Matrix and tensor framework,” 2014, *arXiv:1412.2863*.
- [267] V. Kuleshov, A. Chaganty, and P. Liang, “Tensor factorization via matrix factorization,” in *Proc. Artif. Intell. Statist.*, 2015, pp. 507–516.
- [268] S. Mei, Y. Bai, and A. Montanari, “The landscape of empirical risk for non-convex losses,” *Annu. Statist.*, vol. 46, no. 6A, pp. 2747–2774, Jul. 2016.
- [269] M. N. R. Chowdhury, M. Wang, K. E. Maghraoui, N. Wang, P.-Y. Chen, and C. Carothers, “A provably effective method for pruning experts in fine-tuned sparse mixture-of-experts,” 2024, *arXiv:2405.16646*.
- [270] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2018.
- [271] A. V. Makkula, M. Bondaschi, C. Ekbote, A. Girish, A. Nagle, H. Kim, and M. Gastpar, “Local to global: Learning dynamics and effect of initialization for transformers,” 2024, *arXiv:2406.03072*.

## APPENDIX A

## APPENDIX OF CHAPTER 2

The appendix contains 6 sections. We first provide a brief discussion about comparisons between our works and other two related works in Section A.1. In Section A.2, we introduce some definitions and assumptions in accordance with the main paper for the ease of the proof in the following. Section A.3 first states a core lemma for the proof, based on which we provide the proof of Theorem 1 and Proposition 1 and 2. Section A.4 gives the proof of Lemma A.3.1 with three subsections to prove its three main claims. Section A.5 shows key lemmas and the proof of lemmas for this chapter. We finally discuss the extension of our analysis in Section A.6, including extension to multi-classification cases, general data model cases, multi-head attention cases, and cases with skip connections in Section A.6.1, A.6.2, A.6.3, and A.6.4, respectively.

### A.1 Comparison with Two Related Works

#### A.1.1 Comparison with (Allen-Zhu & Li, 2023)

[113] studies ensemble learning and knowledge distillation. Its main proof idea is that given large amounts of multi-view data, each single model learns one feature, and then ensemble learning integrates all learned features and, thus, improves over single models. Knowledge distillation applies softmax logits to make use of information learned from the ensemble model. It is analyzed in a similar approach to studying single models. The single models considered in [113] is a two-layer Relu network.

In contrast, in this chapter, we consider a two-layer Relu network with an additional self-attention layer. The network architecture and training algorithm for the self-attention layer is completely different from those for the softmax logit in the knowledge distillation function. In our proof, we analyze the impact of the gradient of  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$ , and  $\mathbf{W}_V$  on different patterns (Claims 2 and 3 of Lemma 2), showing that the training process helps to enlarge the magnitude of label-relevant features. We also show that neurons in  $\mathbf{W}_O$  mainly learn from discriminative patterns (Claim 1 of Lemma 2). Such a learning process is affected by the error in the initial model and the noise in tokens. Please see details in “Proof idea

---

Portions of this appendix previously appeared as: H. Li, M. Wang, S. Liu, and P.-Y. Chen, “A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity,” in *Proc. Int. Conf. Learn. Represent.*, May 2023, Paper 3368.

sketch” and “Technical novelty” in Section 3.3 on Page 7 of the paper. This technique we develop plays a critical role in our analysis of self-attention layers. This technique is novel and did not appear in any existing works.

### A.1.2 Comparison with (Jelassi et al., 2022)

[114] is a concurrent work which theoretically studies Vision Transformers. The major difference between [114] and our work is that we consider different data models and network architectures. In [114], the data model requires spatial association between tokens. The attention map is replaced with position encoding, and the training process of the attention map is simplified to train a linear layer. Our setup models the data mainly based on the category of patterns. We keep the classical structure and training process of self-attention, where  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$ , and  $\mathbf{W}_V$  are trained separately. The required number of samples and iterations are derived as functions of the fraction of label-relevant patterns. In addition, the non-linear activation function they consider is polynomial activation, instead of Relu or Gelu as in practice. Based on these conditions, they are able to study a different and more general labelling function.

## A.2 Preliminaries

We first formally restate the neural network with different notations of loss functions, and the Algorithm 3 of the training steps after token sparsification. The notations used in the Appendix is summarized in Table A.1.

For the network<sup>33</sup>

$$F(\mathbf{X}^n) = \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \mathbf{a}_{(l)}^\top \text{Relu}(\mathbf{W}_O \mathbf{W}_V \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n)) \quad (\text{A.1})$$

The loss function of a single data, a mini-batch, the empirical loss, and the population loss is defined in the following.

$$\text{Loss}(\mathbf{X}^n, y^n) = \max\{1 - y^n \cdot F(\mathbf{X}^n), 0\} \quad (\text{A.2})$$

---

<sup>33</sup>Note that in our proof in the Appendix, we often use the notation  $\text{softmax}(\mathbf{x}_i^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n)$ , which is the same meaning as  $\text{softmax}((\mathbf{X}^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)}) \mathbf{x}_l^n)_i$ .

**Table A.1: Summary of notations.**

$F(\mathbf{X}^n)$ , Loss( $\mathbf{X}^n, y^n$ )	The network output for $\mathbf{X}^n$ and the loss function of a single data.
$\overline{\text{Loss}}_b$ , Loss, Loss	The loss function of a mini-batch, the empirical loss, and the population loss, respectively.
$\mathbf{p}_j(t)$ , $\mathbf{q}_j(t)$ , $\mathbf{r}_j(t)$	The features in value, key, and query vectors at the iteration $t$ for pattern $j$ , respectively. We have $\mathbf{p}_j(0) = \mathbf{p}_j$ , $\mathbf{q}_j(0) = \mathbf{q}_j$ , and $\mathbf{r}_j(0) = \mathbf{r}_j$ .
$\mathbf{z}_j^n(t)$ , $\mathbf{n}_j^n(t)$ , $\mathbf{o}_j^n(t)$	The error terms in the value, key, and query vectors of the $j$ -th token and $n$ -th data compared to their features at iteration $t$ .
$\mathcal{W}_{l,n}(0)$ , $\mathcal{U}_{l,n}(0)$	The set of lucky neurons for the token $l$ of data $n$ .
$\phi_n(t)$ , $\nu_n(t)$ , $p_n(t)$ , $\lambda$	The bounds of value of some attention weights at iteration $t$ . $\lambda$ is the threshold between inner products of tokens from the same pattern and different patterns.
$\mathcal{S}_j^n$ , $\mathcal{S}_*^n$ , $\mathcal{S}_\#^n$	$\mathcal{S}_j^n$ is the set of sampled tokens of pattern $j$ for the $n$ -th data. $\mathcal{S}_*^n$ , $\mathcal{S}_\#^n$ are sets of sampled tokens of the label-relevant pattern and the confusion pattern for the $n$ -th data, respectively.
$\alpha_*$ , $\alpha_\#$ , $\alpha_{nd}$	The mean of fraction of label-relevant tokens, confusion tokens, and non-discriminative tokens, respectively.

$$\overline{\text{Loss}}_b = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \text{Loss}(\mathbf{X}^n, y^n) \quad (\text{A.3})$$

$$\overline{\text{Loss}} = \frac{1}{N} \sum_{n=1}^N \text{Loss}(\mathbf{X}^n, y^n) \quad (\text{A.4})$$

$$\text{Loss} = \mathbb{E}_{(\mathbf{X}, y) \sim \mathcal{D}} [\overline{\text{Loss}}] \quad (\text{A.5})$$

The formal algorithm is as follows. We assume that each entry of  $\mathbf{W}_O^{(0)}$  is randomly initialized from  $\mathcal{N}(0, \xi^2)$  where  $\xi = \frac{1}{\sqrt{M}}$ . Define that  $a_{(l)i}^{(0)}$ ,  $i \in [m]$ ,  $l \in [L]$  is uniformly initialized from  $+\{\frac{1}{\sqrt{m}}, -\frac{1}{\sqrt{m}}\}$  and fixed during the training.  $\mathbf{W}_V$ ,  $\mathbf{W}_K$ , and  $\mathbf{W}_Q$  are initialized from a good pretrained model.

Assumption 2.3.1 can be interpreted as that we initialize  $\mathbf{W}_V$ ,  $\mathbf{W}_K$ , and  $\mathbf{W}_Q$  to be the matrices that can map tokens to orthogonal features with added error terms.

**Assumption A.2.1.** Define  $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M) \in \mathbb{R}^{m_a \times M}$ ,  $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_M) \in \mathbb{R}^{m_b \times M}$  and  $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_M) \in \mathbb{R}^{m_b \times M}$  as three feature matrices, where  $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M\}$ ,  $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_M\}$  and  $\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_M\}$  are three sets of orthonormal bases. Define the noise terms  $\mathbf{z}_j^n(t)$ ,  $\mathbf{n}_j^n(t)$  and  $\mathbf{o}_j^n(t)$  with  $\|\mathbf{z}_j^n(0)\| \leq \sigma + \tau$  and  $\|\mathbf{n}_j^n(0)\|, \|\mathbf{o}_j^n(0)\| \leq \delta + \tau$  for  $j \in [L]$ .  $\mathbf{q}_1 = \mathbf{r}_1$ ,  $\mathbf{q}_2 = \mathbf{r}_2$ . Suppose  $\|\mathbf{W}_V^{(0)}\|, \|\mathbf{W}_K^{(0)}\|, \|\mathbf{W}_Q^{(0)}\| \leq 1$ ,  $\sigma, \tau < O(1/M)$  and

---

**Algorithm 3** Training with SGD

---

- 1: **Input:** Training data  $\{(\mathbf{X}^n, y^n)\}_{n=1}^N$ , the step size  $\eta$ , the total number of iterations  $T$ , batch size  $B$ .
- 2: **Initialization:** Every entry of  $\mathbf{W}_O^{(0)}$  from  $\mathcal{N}(0, \xi^2)$ , and every entry of  $\mathbf{a}_{(l)}^{(0)}$  from Uniform( $\{+\frac{1}{\sqrt{m}}, -\frac{1}{\sqrt{m}}\}$ ).  $\mathbf{W}_V^{(0)}, \mathbf{W}_K^{(0)}$  and  $\mathbf{W}_Q^{(0)}$  from a pre-trained model.
- 3: **Stochastic Gradient Descent:** for  $t = 0, 1, \dots, T-1$  and  $\mathbf{W}^{(t)} \in \{\mathbf{W}_O^{(t)}, \mathbf{W}_V^{(t)}, \mathbf{W}_K^{(t)}, \mathbf{W}_Q^{(t)}\}$

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_t} \nabla_{\mathbf{W}^{(t)}} \ell(\mathbf{X}^n, y^n; \mathbf{W}_O^{(t)}, \mathbf{W}_V^{(t)}, \mathbf{W}_K^{(t)}, \mathbf{W}_Q^{(t)}) \quad (\text{A.6})$$

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

$\delta < 1/2$ . Then, for  $\mathbf{x}_l^n \in \mathcal{S}_j^n$

1.  $\mathbf{W}_V^{(0)} \mathbf{x}_l^n = \mathbf{p}_j + \mathbf{z}_j^n(0)$ .
2.  $\mathbf{W}_K^{(0)} \mathbf{x}_l^n = \mathbf{q}_j + \mathbf{n}_j^n(0)$ .
3.  $\mathbf{W}_Q^{(0)} \mathbf{x}_l^n = \mathbf{r}_j + \mathbf{o}_j^n(0)$ .

Assumption A.2.1 is a straightforward combination of Assumption 2.3.1 and (2.5) by applying the triangle inequality to bound the error terms for tokens.

**Definition A.2.2.** 1.  $\phi_n(t) = \frac{1}{|\mathcal{S}_1^n| e^{\|\mathbf{q}_1(t)\|^2 + (\delta + \tau) \|\mathbf{q}_1(t)\|} + |\mathcal{S}^n| - |\mathcal{S}_1^n|}$ .

2.  $\nu_n(t) = \frac{1}{|\mathcal{S}_1^n| e^{\|\mathbf{q}_1(t)\|^2 - (\delta + \tau) \|\mathbf{q}_1(t)\|} + |\mathcal{S}^n| - |\mathcal{S}_1^n|}$ .

3.  $p_n(t) = |\mathcal{S}_1^n| e^{\|\mathbf{q}_1(t)\|^2 - (\delta + \tau) \|\mathbf{q}_1(t)\|} \nu_n(t)$ .

4.  $\mathcal{S}_*^n = \begin{cases} \mathcal{S}_1^n, & \text{if } y^n = 1 \\ \mathcal{S}_2^n, & \text{if } y^n = -1 \end{cases}, \mathcal{S}_\#^n = \begin{cases} \mathcal{S}_2^n, & \text{if } y^n = 1 \\ \mathcal{S}_1^n, & \text{if } y^n = -1 \end{cases}$

5.  $\alpha_* = \mathbb{E}\left[\frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|}\right], \alpha_\# = \mathbb{E}\left[\frac{|\mathcal{S}_\#^n|}{|\mathcal{S}^n|}\right], \alpha_{nd} = \sum_{l=3}^M \mathbb{E}\left[\frac{|\mathcal{S}_l^n|}{|\mathcal{S}^n|}\right]$ .

**Definition A.2.3.** Define

$$\mathbf{V}_l^n(t) = \mathbf{W}_V^{(t)} \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \quad (\text{A.7})$$

for the token  $l$  of data  $n$ . Define  $\mathcal{W}_{l,n}(0)$ ,  $\mathcal{U}_{l,n}(0)$  as the sets of lucky neurons such that

$$\mathcal{W}_{l,n}(0) = \{i : \mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{V}_{l,n}(0) > 0, l \in \mathcal{S}_1^n\} \quad (\text{A.8})$$

$$\mathcal{U}_{l,n}(0) = \{i : \mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{V}_{l,n}(0) > 0, l \in \mathcal{S}_2^n\} \quad (\text{A.9})$$

**Assumption A.2.4.** For one data  $\mathbf{X}^n$ , if the patch  $i$  and  $j$  correspond to the same feature  $k \in [M]$ , i.e.,  $i \in \mathcal{S}_k^n$  and  $j \in \mathcal{S}_k^n$ , we have

$$\mathbf{x}_i^{n\top} \mathbf{x}_j^n \geq 1 \quad (\text{A.10})$$

If the patch  $i$  and  $j$  correspond to the different feature  $k, l \in [M], k \neq l$  i.e.,  $i \in \mathcal{S}_k^n$  and  $j \in \mathcal{S}_l^n, k \neq l$ , we have

$$\mathbf{x}_i^{n\top} \mathbf{x}_j^n \leq \lambda < 1 \quad (\text{A.11})$$

This assumption is equivalent to the data model by (2.5) since  $\tau < O(1/M)$ . For the simplicity of presentation, we scale up all tokens a little bit to make the threshold of linear separability be 1. We also take  $1 - \lambda$  and  $\lambda$  as  $\Theta(1)$  for the simplicity.

**Definition A.2.5.** [258] 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 A.2.6.** ([258] Proposition 5.1, Hoeffding's inequality) Let  $X_1, X_2, \dots, X_N$  be independent centered sub-gaussian random variables, and let  $K = \max_i \|\mathbf{X}_i\|_{\psi_2}$ . Then for every  $\mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N$  and every  $t \geq 0$ , we have

$$\mathbb{P}\left\{\left|\sum_{i=1}^N a_i X_i\right| \geq t\right\} \leq e \cdot \exp\left(-\frac{ct^2}{K^2 \|\mathbf{a}\|^2}\right) \quad (\text{A.12})$$

where  $c > 0$  is an absolute constant.

### A.3 Proof of the Main Theorem and Propositions

We state Lemma A.3.1 first before we introduce the proof of main theorems. Lemma A.3.1 is the key lemma in our paper to show the training process of our ViT model using SGD. It has three major claims. Claim 1 involves the growth of  $\mathbf{W}_O^{(t)}$  in terms of different directions

of  $\mathbf{p}_l$ ,  $i \in [M]$ . Claim 2 describes the training dynamics of  $\mathbf{W}_Q^{(t)}$  and  $\mathbf{W}_K^{(t)}$  separately to show the tendency to a sparse attention map. Claim 3 studies the gradient update process of  $\mathbf{W}_V^{(t)}$ .

**Lemma A.3.1.** *For  $l \in \mathcal{S}_1^n$  for the data with  $y^n = 1$ , define*

$$\mathbf{V}_l^n(t) = \mathbf{W}_V^{(t)} \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \quad (\text{A.13})$$

We later show that

$$\begin{aligned} \mathbf{V}_l^n(t) &= \sum_{s \in \mathcal{S}_1^n} \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{p}_1 + \mathbf{z}(t) + \sum_{j \neq 1} W_j^n(t) \mathbf{p}_j \\ &\quad - \eta \sum_{b=1}^t \left( \sum_{i \in \mathcal{W}_{l,n}(0)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)\top} + \sum_{i \notin \mathcal{W}_{l,n}(0)} V_i(b) \lambda \mathbf{W}_{O_{(i,\cdot)}}^{(b)\top} \right) \end{aligned} \quad (\text{A.14})$$

where

$$W_l^n(t) \leq \nu_n(t) |\mathcal{S}_j^n| \quad (\text{A.15})$$

$$V_i(t) \lesssim \frac{1}{2B} \sum_{n \in \mathcal{B}_{b+}} -\frac{|\mathcal{S}_1^n|}{a|\mathcal{S}^n|} p_n(t), \quad i \in \mathcal{W}_{l,n}(0) \quad (\text{A.16})$$

$$V_i(t) \gtrsim \frac{1}{2B} \sum_{n \in \mathcal{B}_{b-}} \frac{|\mathcal{S}_2^n|}{a|\mathcal{S}^n|} p_n(t), \quad i \in \mathcal{U}_{l,n}(0) \quad (\text{A.17})$$

$$|V_i(t)| \leq \frac{1}{\sqrt{B}aM}, \quad \text{if } i \text{ is an unlucky neuron.} \quad (\text{A.18})$$

We also have the following claims when  $m \gtrsim M^2 \log N$ :

*Claim 1.* For the lucky neuron  $i \in \mathcal{W}_{l,n}(0)$  and  $b \in [T]$ , we have

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p}_1 \gtrsim \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(t) + \xi \quad (\text{A.19})$$

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p} \lesssim \frac{1}{\sqrt{B}} \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p}_1, \quad \text{for } \mathbf{p} \in \mathcal{P} \setminus \{\mathbf{p}_1\}, \quad (\text{A.20})$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\|^2 \geq \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(t) + \xi \right)^2 \quad (\text{A.21})$$

and for the noise  $\mathbf{z}_l(t)$ ,

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{z}_l(t)\| \leq (\sigma + \tau) \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\| \quad (\text{A.22})$$

For  $i \in \mathcal{U}_{l,n}(0)$ , we also have equations as in (A.19) to (A.22), including

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p}_2 \gtrsim \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) + \xi \quad (\text{A.23})$$

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p} \lesssim \frac{1}{\sqrt{B}} \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p}_1, \quad \text{for } \mathbf{p} \in \mathcal{P} \setminus \{\mathbf{p}_2\}, \quad (\text{A.24})$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\|^2 \geq \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) + \xi \right)^2 \quad (\text{A.25})$$

and for the noise  $\mathbf{z}_l(t)$ ,

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{z}_l(t)\| \leq (\sigma + \tau) \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\| \quad (\text{A.26})$$

For unlucky neurons  $i$  and  $j \in \mathcal{W}_{l,n}(0)$ ,  $k \in \mathcal{U}_{l,n}(0)$ ,  $p \in \mathcal{P}$ , we have

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p} \leq \frac{1}{\sqrt{B}} \min\{\mathbf{W}_{O_{(j,\cdot)}}^{(t)} \mathbf{p}_1, \mathbf{W}_{O_{(k,\cdot)}}^{(t)} \mathbf{p}_2\}, \quad (\text{A.27})$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{z}_l(t)\| \leq (\sigma + \tau) \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\| \quad (\text{A.28})$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\|^2 \leq \frac{1}{B} \min\{\|\mathbf{W}_{O_{(j,\cdot)}}^{(t)}\|^2, \|\mathbf{W}_{O_{(k,\cdot)}}^{(t)}\|^2\} \quad (\text{A.29})$$

*Claim 2.* Given conditions in (2.8), there exists  $K(t), Q(t) > 0$ , where  $t$  is large enough before the end of training, such that for  $j \in \mathcal{S}_*^n$ ,

$$\text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n) \gtrsim \frac{e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|}}{|\mathcal{S}_1^n| e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \quad (\text{A.30})$$

$$\begin{aligned} & \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_j^n) - \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \\ & \gtrsim \frac{|\mathcal{S}^n| - |\mathcal{S}_1^n|}{(|\mathcal{S}_1^n| e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|))^2} e^{\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} \cdot K(t), \end{aligned} \quad (\text{A.31})$$

and for  $j \notin \mathcal{S}_*^n$ , we have

$$\text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n) \lesssim \frac{1}{|\mathcal{S}_1^n| e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - \delta\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \quad (\text{A.32})$$

$$\begin{aligned} & \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n) - \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \\ & \lesssim - \frac{|\mathcal{S}_1^n|}{(|\mathcal{S}_1^n| e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - \delta\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|))^2} e^{\|\mathbf{q}_1(t)\|^2 - \delta\|\mathbf{q}_1(t)\|} \cdot K(t) \end{aligned} \quad (\text{A.33})$$

For  $i = 1, 2$ ,

$$\mathbf{q}_i(t) = \sqrt{\prod_{l=0}^{t-1} (1 + K(l))} \mathbf{q}_i \quad (\text{A.34})$$

$$\mathbf{r}_i(t) = \sqrt{\prod_{l=0}^{t-1} (1 + Q(l))} \mathbf{r}_i \quad (\text{A.35})$$

*Claim 3.* For the update of  $\mathbf{W}_V^{(t)}$ , there exists  $\lambda \leq \Theta(1)$  such that

$$\mathbf{W}_V^{(t)} \mathbf{x}_j^n = \mathbf{p}_1 - \eta \sum_{b=1}^t \left( \sum_{i \in \mathcal{W}_{l,n}(0)} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \top + \sum_{i \notin \mathcal{W}_{l,n}(0)} \lambda V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \top \right) + \mathbf{z}_j(t), \quad j \in \mathcal{S}_1^n \quad (\text{A.36})$$

$$\mathbf{W}_V^{(t)} \mathbf{x}_j^n = \mathbf{p}_2 - \eta \sum_{b=1}^t \left( \sum_{i \in \mathcal{U}_{l,n}(0)} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \top + \sum_{i \notin \mathcal{U}_{l,n}(0)} \lambda V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \top \right) + \mathbf{z}_j(t), \quad j \in \mathcal{S}_2^n \quad (\text{A.37})$$

$$\mathbf{W}_V^{(t+1)} \mathbf{x}_j^n = \mathbf{p}_j - \eta \sum_{b=1}^t \sum_{i=1}^m \lambda V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \top + \mathbf{z}_j(t), \quad j \in [|\mathcal{S}^n|]/(\mathcal{S}_1^n \cup \mathcal{S}_2^n) \quad (\text{A.38})$$

$$\|\mathbf{z}_j(t)\| \leq (\sigma + \tau) \quad (\text{A.39})$$

To prove Theorem 1, we either show  $F(\mathbf{X}^n) > 1$  for  $y^n = 1$  or show  $F(\mathbf{X}^n) < -1$  for  $y^n = -1$ . Take  $y^n = 1$  as an example, the basic idea of the proof is to make use of Lemma A.3.1 to find a lower bound as a function of  $\alpha_*$ ,  $\sigma$ ,  $\tau$ , etc.. The remaining step is to derive conditions on the sample complexity and the required number of iterations in terms of  $\alpha_*$ ,  $\sigma$ , and  $\tau$  such that the lower bound is greater than 1. Given a balanced dataset, these conditions also ensure that  $F(\mathbf{X}^n) < -1$  for  $y^n = -1$ . During the proof, we may need to use some of equations as intermediate steps in the proof of Lemma A.3.1. Since that these equations are not concise for presentation, we prefer not to state them formally in Lemma A.3.1, but still refer to them as useful conclusions. The following is the details of the proof.

### Proof of Theorem 1:

For  $y^n = 1$ , define  $\mathcal{K}_+ = \{i \in [m] : a_i > 0\}$  and  $\mathcal{K}_- = \{i \in [m] : a_i < 0\}$ . We have

$$\begin{aligned} F(\mathbf{X}^n) = & \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i \in \mathcal{W}_{l,n}(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)) + \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i \in \mathcal{K}_+ \setminus \mathcal{W}_{l,n}(0)} \frac{1}{a} \\ & \cdot \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)) - \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i \in \mathcal{K}_-} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)) \end{aligned} \quad (\text{A.40})$$

By Lemma A.3.1, we have that when  $m \gtrsim M^2 \log N$ ,

$$\begin{aligned}
& \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i \in \mathcal{W}_{l,n}(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)) \\
&= \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}_1^n} \sum_{i \in \mathcal{W}_{l,n}(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)) + \sum_{l \notin \mathcal{S}_1^n} \sum_{i \in \mathcal{W}_{l,n}(0)} \frac{1}{m} \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)) \\
&\gtrsim |\mathcal{S}_1^n| \frac{1}{a|\mathcal{S}^n|} \cdot \mathbf{W}_{O_{(i,:)}}^{(t)} \left( \sum_{s \in \mathcal{S}_1^n} \mathbf{p}_s \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_s^n) + \mathbf{z}(t) + \sum_{l \neq s} W_l(u) \mathbf{p}_l \right. \\
&\quad \left. - \eta t \left( \sum_{j \in \mathcal{W}_{l,n}(0)} V_j(t) \mathbf{W}_{O_{(j,:)}}^{(t)\top} + \sum_{j \notin \mathcal{W}_{l,n}(0)} V_j(t) \lambda \mathbf{W}_{O_{(j,:)}}^{(t)\top} \right) \right) |\mathcal{W}_{l,n}(0)| + 0 \\
&\gtrsim \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 |\mathcal{S}_1^n|m}{|\mathcal{S}^n|a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(b) - \sigma - \tau \right) \right. \\
&\quad \cdot p_n(t) + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_1^n|}{a|\mathcal{S}^n|} p_n(b) \\
&\quad \left. \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta (t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) \right)^2 \right)
\end{aligned} \tag{A.41}$$

where the second step comes from (A.14) and the last step is by (A.118). By the definition of  $\mathcal{K}_+^l$ , we have

$$\frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i \in \mathcal{K}_+ \setminus \mathcal{W}_{l,n}(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)) \geq 0 \tag{A.42}$$

We can obtain that

$$\frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i \in \mathcal{U}_{l,n}(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)) \leq \frac{|\mathcal{S}_2^n|}{|\mathcal{S}_1^n|} \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i \in \mathcal{W}_{l,n}(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)) \tag{A.43}$$

Combining (A.118) and (A.120), we can obtain

$$\begin{aligned}
& \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i \in \mathcal{K}_- \setminus \mathcal{U}_{l,n}(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)) \\
&\lesssim \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} \cdot \frac{1}{\sqrt{B}} \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i \in \mathcal{W}_{l,n}(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t))
\end{aligned} \tag{A.44}$$

Note that at the  $T$ -th iteration where  $\eta T = \Theta(1)$ ,

$$\begin{aligned}
& K(t) \\
& \gtrsim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(t) \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 |\mathcal{S}_1^n| m}{|\mathcal{S}^n| a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(b) - \sigma - \tau \right) \right. \\
& \quad \cdot + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_1^n|}{a |\mathcal{S}^n|} p_n(b) \\
& \quad \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta (t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(t) \right)^2 \Big) \\
& \quad \cdot \phi_n(t) (|\mathcal{S}^n| - |\mathcal{S}_1^n|) \|\mathbf{q}_1(t)\|^2 \\
& \gtrsim \frac{\eta}{e^{\|\mathbf{q}_1(t)\|^2 - (\delta + \tau) \|\mathbf{q}_1(t)\|}}
\end{aligned} \tag{A.45}$$

Since that

$$\begin{aligned}
\mathbf{q}_1(T) & \gtrsim (1 + \min_{l=0,1,\dots,T-1} \{K(l)\})^T \\
& \gtrsim (1 + \frac{\eta}{e^{\|\mathbf{q}_1(T)\|^2 - (\delta + \tau) \|\mathbf{q}_1(T)\|}})^T
\end{aligned} \tag{A.46}$$

To find the order-wise lower bound of  $\mathbf{q}_1(T)$ , we need to check the equation

$$\mathbf{q}_1(T) \lesssim (1 + \frac{1}{e^{\|\mathbf{q}_1(T)\|^2 - (\delta + \tau) \|\mathbf{q}_1(T)\|}})^T \tag{A.47}$$

One can obtain

$$\Theta(\log T) = \|\mathbf{q}_1(T)\|^2 = o(T) \tag{A.48}$$

Therefore,

$$p_n(T) \gtrsim \frac{T^C}{T^C + \frac{1-\alpha}{\alpha}} \geq 1 - \frac{1}{\frac{\alpha}{1-\alpha} (\eta^{-1})^C} \geq 1 - \Theta(\eta^C) \tag{A.49}$$

$$\phi_n(T) (|\mathcal{S}^n| - |\mathcal{S}_1^n|) \leq \eta^C \tag{A.50}$$

for some large  $C > 0$ . We require that

$$\begin{aligned}
& \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta(t+1)^2 |\mathcal{S}_1^n|m}{|\mathcal{S}^n|a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(b) - \sigma - \tau \right) \right. \\
& \quad \cdot p_n(t) + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_1^n|}{a|\mathcal{S}^n|} p_n(b) \\
& \quad \cdot \left. \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta(t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) \right)^2 \right) \\
& := a_0 \eta^3 T^5 + a_1 \eta T^2 \\
& > 1,
\end{aligned} \tag{A.51}$$

where the first step is by letting  $a = \sqrt{m}$  and  $\xi = 1/\sqrt{m}$ . We replace  $p_n(b)$  with  $p_n(T)$  because when  $b$  achieves the level of  $T$ ,  $b^{o_1} p_n(b)^{o_2}$  is close to  $b^{o_1}$  for  $o_1, o_2 \geq 0$  by (A.49). Thus,

$$\sum_{b=1}^T b^{o_1} p_n(b)^{o_2} \gtrsim T^{o_1+1} p_n(\Theta(1) \cdot T)^{o_2} \gtrsim T^{o_1+1} p_n(T)^{o_2} \tag{A.52}$$

We also require

$$B \geq \Omega(1), \tag{A.53}$$

for some  $\epsilon_0 > 0$ .

We know that

$$\begin{aligned}
& \left| \frac{1}{N} \sum_{n=1}^N \frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|} p_n(T) (p_n(T) - (\sigma + \tau)) - \mathbb{E} \left[ \frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|} \right] \right| \\
& \leq \left| \frac{1}{N} \sum_{n=1}^N \frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|} p_n(T) (p_n(T) - (\sigma + \tau)) - \mathbb{E} \left[ \frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|} p_n(T) (p_n(T) - (\sigma + \tau)) \right] \right| \\
& \quad + \left| \mathbb{E} \left[ \frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|} (p_n(T) (p_n(T) - (\sigma + \tau)) - 1) \right] \right| \\
& \lesssim \sqrt{\frac{\log N}{N}} + c'(1 - \zeta) + c''((\sigma + \tau))
\end{aligned} \tag{A.54}$$

for  $c' > 0$  and  $c'' > 0$ . We can then have

$$\begin{aligned} t \geq T &= \frac{\eta^{-\frac{3}{5}}}{a_1} = \frac{\eta^{-\frac{3}{5}}}{\frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{N} \sum_{n=1}^N (\frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} \|\mathbf{p}_1\|^2 p_n(t) - (\sigma + \tau)) p_n(t)} \\ &= \Theta\left(\frac{\eta^{-\frac{3}{5}}}{\left(\mathbb{E}\left[\frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|}\right] - \sqrt{\frac{\log N}{N}} - c'(1 - \zeta) - c''(\sigma + \tau)\right)}\right) \\ &= \Theta\left(\frac{\eta^{-\frac{3}{5}}}{\mathbb{E}\left[\frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|}\right]}\right) \end{aligned} \quad (\text{A.55})$$

where

$$\alpha \geq \frac{1 - \alpha_{nd}}{1 + e^{-(\delta + \tau)}(1 - (\sigma + \tau))} \quad (\text{A.56})$$

by (A.138), as long as

$$N \geq \Omega\left(\frac{1}{(\alpha - c'(1 - \zeta) - c''((\sigma + \tau)))^2}\right) \quad (\text{A.57})$$

where  $\zeta \geq 1 - \eta^{10}$ . If there is no mechanism like the self-attention to compute the weight using the correlations between tokens, we have

$$c'(1 - \zeta) = O(\alpha_*(1 - \alpha_*)), \quad (\text{A.58})$$

which can scale up the sample complexity in (A.57) by  $\alpha_*^{-2}$ .

Therefore, we can obtain

$$F(\mathbf{X}^n) > 1 \quad (\text{A.59})$$

Similarly, we can derive that for  $y = -1$ ,

$$F(\mathbf{X}) < -1 \quad (\text{A.60})$$

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

$$\text{Loss}(\mathbf{X}^n, y^n) = 0 \quad (\text{A.61})$$

We also have

$$\text{Loss} = \mathbb{E}_{(\mathbf{X}^n, y^n) \sim \mathcal{D}} [\text{Loss}(\mathbf{X}^n, y^n)] = 0 \quad (\text{A.62})$$

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

### Proof of Proposition 1:

The main proof is the same as the proof of Theorem 1. The only difference is that we need to modify (A.54) as follows

$$\begin{aligned}
& \left| \frac{1}{N} \sum_{n=1}^N \frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|} p_n(T)(p_n(T) - (\sigma + \tau)) - \mathbb{E}\left[\frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|}\right] \right| \\
& \leq \left| \frac{1}{N} \sum_{n=1}^N \frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|} p_n(0)(p_n(0) - (\sigma + \tau)) - \mathbb{E}\left[\frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|} p_n(0)(p_n(T) - (\sigma + \tau))\right] \right| \\
& \quad + \left| \mathbb{E}\left[\frac{|\mathcal{S}_*^n|}{|\mathcal{S}^n|} (p_n(0)(p_n(0) - (\sigma + \tau)) - 1)\right] \right| \\
& \lesssim \sqrt{\frac{\log N}{N}} + |1 - \Theta(\alpha_*^2) + \Theta(\alpha_*)(\sigma + \tau)|
\end{aligned} \tag{A.63}$$

where the first step is because  $p_n(T)$  does not update since  $\mathbf{W}_K^{(t)}$  and  $\mathbf{W}_Q^{(t)}$  are fixed at initialization  $\mathbf{W}_K^{(0)}$  and  $\mathbf{W}_Q^{(0)}$ , and the second step is by  $p_n(0) = \Theta(\alpha_*)$ . Since that

$$\sqrt{\frac{\log N}{N}} + |1 - \Theta(\alpha_*^2) + \Theta(\alpha_*)(\sigma + \tau)| \leq \Theta(1) \cdot \alpha_*, \tag{A.64}$$

we have

$$\begin{aligned}
N & \geq \frac{1}{(\Theta(\alpha_*) - 1 + \Theta(\alpha_*^2) - \Theta(\alpha_*)(\sigma + \tau))^2} \\
& = \Omega\left(\frac{1}{(\alpha_*(\alpha_* - \sigma - \tau))^2}\right)
\end{aligned} \tag{A.65}$$

### Proof of Proposition 2:

It can be easily derived from Claim 2 of Lemma A.3.1, (A.48), and (A.49).

## A.4 Proof of Lemma A.3.1

We prove the whole lemma by a long induction, which is the reason why we prefer to wrap three claims into one lemma. To make it easier to follow, however, we break this Section into three parts to introduce the proof of three claims of Lemma A.3.1 separately.

### A.4.1 Proof of Claim 1 of Lemma A.3.1

Although it looks cumbersome, the key idea of Claim 1 is to characterize the growth of  $\mathbf{W}_O^{(t)}$  in terms of  $\mathbf{p}_l$ ,  $l \in [M]$ . We compare  $\mathbf{W}_{O_{(i,:)}}^{(t+1)} \mathbf{p}_l$  and  $\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{p}_l$  to see the direction of growth

by computing the gradient. One can eventually find that lucky neurons grow the most in directions of  $\mathbf{p}_1$  and  $\mathbf{p}_2$ , i.e., the feature of label-relevant patterns, while unlucky neurons do not change much in magnitude.

We start our proof. At the  $t$ -th iteration ( $t > 1$ ), if  $l \in \mathcal{S}_1^n$ , let

$$\begin{aligned} \mathbf{V}_l^n(t) &= \mathbf{W}_V^{(t)} \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \\ &= \sum_{s \in \mathcal{S}_1} \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{p}_1 + \mathbf{z}(t) + \sum_{j \neq 1} W_j^n(t) \mathbf{p}_j \\ &\quad - \eta \sum_{b=0}^{t-1} \left( \sum_{i \in \mathcal{W}_{l,n}(0)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)\top} + \sum_{i \notin \mathcal{W}_{l,n}(0)} V_i(b) \lambda \mathbf{W}_{O_{(i,\cdot)}}^{(b)\top} \right) \end{aligned} \quad (\text{A.66})$$

,  $l \in [M]$ , where the second step comes from Claim 3 of Lemma A.3.1. We can derive

$$0 < W_l^n(t) \leq \frac{|\mathcal{S}_j^n| e^{\delta \|\mathbf{q}_1(t)\|}}{(|\mathcal{S}^n| - |\mathcal{S}_1^n|) e^{\delta \|\mathbf{q}_1(t)\|} + |\mathcal{S}_1^n| e^{\|\mathbf{q}_1(t)\|^2 - \delta \|\mathbf{q}_1(t)\|}} = \nu_n(t) |\mathcal{S}_j^n| \quad (\text{A.67})$$

which is much smaller than  $\Theta(1)$  when  $t$  is large. This is the reason why we ignore the impact of  $W_l^n(t)$  on  $\eta \sum_{b=0}^{t-1} (\sum_{i \in \mathcal{W}_{l,n}(0)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)\top} + \sum_{i \notin \mathcal{W}_{l,n}(0)} V_i(b) \lambda \mathbf{W}_{O_{(i,\cdot)}}^{(b)\top})$  in (A.66). Hence, by (A.3),

$$\frac{\partial \overline{\text{Loss}}_b}{\partial \mathbf{W}_{O_{(i,\cdot)}}^\top} = -\frac{1}{B} \sum_{n \in \mathcal{B}_b} y^n \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{(l)_i} \mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}} \mathbf{V}_l^n(t) \geq 0] \mathbf{V}_l^n(t)^\top \quad (\text{A.68})$$

Define that for  $j \in [M]$ ,

$$I_4 = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y^n \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{(l)_i} \mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{V}_l^n(t) \geq 0] \sum_{k \in \mathcal{W}_{l,n}(0)} V_k(t) \mathbf{W}_{O_{(k,\cdot)}}^{(t)} \mathbf{p}_j \quad (\text{A.69})$$

$$I_5 = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y^n \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{(l)_i} \mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{V}_l^n(t) \geq 0] \sum_{k \notin \mathcal{W}_{l,n}(0)} V_k(t) \mathbf{W}_{O_{(k,\cdot)}}^{(t)} \mathbf{p}_j, \quad (\text{A.70})$$

and we can then obtain

$$\begin{aligned}
& \left\langle \mathbf{W}_{O_{(i,:)}}^{(t+1)\top}, \mathbf{p}_j \right\rangle - \left\langle \mathbf{W}_{O_{(i,:)}}^{(t)\top}, \mathbf{p}_j \right\rangle \\
&= \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y^n \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{(l)i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t) \geq 0] \mathbf{V}_l^n(t)^\top \mathbf{p}_j \\
&= \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y^n \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{(l)i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t) \geq 0] \mathbf{z}_l(t)^\top \mathbf{p}_j \\
&\quad + \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y^n \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{(l)i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t) \geq 0] \sum_{s \in \mathcal{S}_l} \text{softmax}(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l) \\
&\quad \cdot \mathbf{p}_l^\top \mathbf{p}_j + \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y^n \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{(l)i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t) \geq 0] \sum_{k \neq l} W_l(t) \mathbf{p}_k^\top \mathbf{p}_j + I_4 + I_5 \\
&:= I_1 + I_2 + I_3 + I_4 + I_5,
\end{aligned} \tag{A.71}$$

where

$$I_1 = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y^n \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{(l)i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t) \geq 0] \mathbf{z}_l(t)^\top \mathbf{p}_j \tag{A.72}$$

$$I_2 = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y^n \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{(l)i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t) \geq 0] \sum_{s \in \mathcal{S}_l} \text{softmax}(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l) \mathbf{p}_l^\top \mathbf{p}_j \tag{A.73}$$

$$I_3 = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y^n \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{(l)i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t) \geq 0] \sum_{k \neq l} W_l(t) \mathbf{p}_k^\top \mathbf{p}_j \tag{A.74}$$

We then show the statements in different cases.

(1) When  $j = 1$ , since that  $\Pr(y^n = 1) = \Pr(y^n = -1) = 1/2$ , by Hoeffding's inequality in (A.12), we can derive

$$\Pr \left( \left| \frac{1}{B} \sum_{n \in \mathcal{B}_b} y^n \right| \geq \sqrt{\frac{\log B}{B}} \right) \leq B^{-c} \tag{A.75}$$

$$\Pr \left( \left| \mathbf{z}_l(t)^\top \mathbf{p}_1 \right| \geq \sqrt{((\sigma + \tau))^2 \log m} \right) \leq m^{-c} \tag{A.76}$$

Hence, with probability at least  $1 - (mB)^{-c}$ , we have

$$|I_1| \leq \frac{\eta((\sigma + \tau))}{a} \sqrt{\frac{\log m \log B}{B}} \tag{A.77}$$

For  $i \in \mathcal{W}_{l,n}(0)$ , by Lemma A.5.2, we have

$$\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^L \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_s^n) > 0 \quad (\text{A.78})$$

Denote  $p_n(t) = |\mathcal{S}_1^n| \nu_n(t) e^{\|\mathbf{q}_1(t)\|^2 - 2\delta \|\mathbf{q}_1(t)\|}$ . Hence, for  $k \notin \mathcal{W}_{l,n}(0)$ ,

$$I_2 \gtrsim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} \cdot \frac{1}{a} \|\mathbf{p}_1\|^2 \cdot p_n(t) \quad (\text{A.79})$$

$$I_3 = 0 \quad (\text{A.80})$$

$$I_4 \gtrsim \frac{1}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{\eta^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n| a} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(t) \|\mathbf{p}_1\|^2 \mathbf{W}_{O_{(i,:)}} \mathbf{p}_1 \quad (\text{A.81})$$

$$\begin{aligned} |I_5| &\lesssim \frac{1}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{\eta^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n| a} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n| m}{|\mathcal{S}^n| a} p_n(t) \|\mathbf{p}_1\|^2 \mathbf{W}_{O_{(i,:)}} \mathbf{p}_2 \\ &\quad + \frac{\eta^2 t m}{\sqrt{B} a^2} \mathbf{W}_{O_{(k,:)}} \mathbf{p}_1 \end{aligned} \quad (\text{A.82})$$

Hence, combining (A.77), (A.79), (A.80), (A.81), and (A.82), we can obtain

$$\begin{aligned} &\left\langle \mathbf{W}_{O_{(i,:)}}^{(t+1)\top}, \mathbf{p}_1 \right\rangle - \left\langle \mathbf{W}_{O_{(i,:)}}^{(t)\top}, \mathbf{p}_1 \right\rangle \\ &\gtrsim \frac{\eta}{a} \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(t) - ((\sigma + \tau)) + \frac{\eta t |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(t) \right. \\ &\quad \cdot \mathbf{W}_{O_{(i,:)}} \mathbf{p}_1 (1 - (\sigma + \tau)) - \frac{\eta t |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n| m}{|\mathcal{S}^n| a} p_n(t) \\ &\quad \cdot \mathbf{W}_{O_{(i,:)}} \mathbf{p}_2 (1 + (\sigma + \tau)) - \frac{\eta t m \mathbf{W}_{O_{(k,:)}} \mathbf{p}_1}{\sqrt{B} a^2} \|\mathbf{p}_1\|^2 \Big) \\ &\gtrsim \frac{\eta}{a B} \sum_{n \in \mathcal{B}_b} \left( \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(t) - ((\sigma + \tau)) + \frac{\eta t |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(t) \right. \\ &\quad \cdot \mathbf{W}_{O_{(i,:)}} \mathbf{p}_1 \Big) \|\mathbf{p}_1\|^2 \end{aligned} \quad (\text{A.83})$$

where the last step holds when  $B \geq \Omega(1)$ . Since that  $\mathbf{W}_{O_{(i,:)}}^{(0)} \sim \mathcal{N}(0, \frac{\xi^2}{m_a} \mathbf{I})$ , by the

standard property of Gaussian distribution, we have

$$\Pr(\|\mathbf{W}_{O_{(i,\cdot)}}^{(0)}\| \lesssim \xi) \lesssim \xi \quad (\text{A.84})$$

Therefore, with high probability for all  $i \in [m]$ , we have

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(0)}\| \gtrsim \xi \quad (\text{A.85})$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(0)} \mathbf{p}_1\| \gtrsim \xi \quad (\text{A.86})$$

When  $\eta$  is very small, given  $p_n(t)$  as the order of a constant, (A.83) leads to a PDE on the lower bound of  $\mathbf{W}_{O_{(i,\cdot)}} \mathbf{p}_1$  since the last step of (A.83) is always positive. Denote  $y(t)$  as a lower bound of  $\mathbf{W}_{O_{(i,\cdot)}} \mathbf{p}_1$ , we have

$$\begin{aligned} & \frac{\partial y(t)}{\partial t} \\ &= \Theta\left(\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \left(\frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(t) - (\sigma + \tau)\right) + \frac{\eta t |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) y(t)\right) \end{aligned} \quad (\text{A.87})$$

Therefore, we can derive

$$\begin{aligned} y(t) &= e^{\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t)} \left( \int_{-\infty}^t \frac{1}{aB} \sum_{n \in \mathcal{B}_b} \left(\frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(t) - (\sigma + \tau)\right) \right. \\ &\quad \left. \cdot e^{-\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta u^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t)} du + C_0 \right) \end{aligned} \quad (\text{A.88})$$

Note that

$$\begin{aligned} & \int_{-\infty}^t e^{-\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta u^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t)} du \\ & \leq \int_{-\infty}^{\infty} e^{-\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta u^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t)} du \\ &= \sqrt{2\pi} \cdot \left(\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t)\right)^{-1} \\ &= \Theta(\eta^{-1}) \end{aligned} \quad (\text{A.89})$$

$$\begin{aligned}
& \int_{-\infty}^t e^{-\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta u^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t)} du \\
& \geq \int_{-\infty}^0 e^{-\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta u^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t)} du \\
& = \Theta(\eta^{-1})
\end{aligned} \tag{A.90}$$

Hence,

$$y(0) = \frac{\eta^{-1}}{aB} \sum_{n \in \mathcal{B}_b} \left( \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(t) - (\sigma + \tau) \right) + C_0 = \Theta(\eta^{-1}\xi) + C_0 = \xi \tag{A.91}$$

$$C_0 = \xi(1 - \Theta(\eta^{-1})) \tag{A.92}$$

$$\begin{aligned}
\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_1 & \gtrsim y(t) \\
& \gtrsim e^{\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta(t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t)} \xi \\
& \gtrsim \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta(t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) + \xi
\end{aligned} \tag{A.93}$$

(2) When  $\mathbf{p}_j \in \mathcal{P} \setminus p^+$ , we have

$$I_2 = 0 \tag{A.94}$$

$$|I_3| \leq \frac{1}{B} \sum_{n \in \mathcal{B}_b} \nu_n(t) \frac{\eta |\mathcal{S}_l^n|}{a} \sqrt{\frac{\log m \log B}{B}} \|\mathbf{p}\|^2 \tag{A.95}$$

$$|I_4| \leq \frac{\eta^2}{a} \sum_{b=1}^t \sqrt{\frac{\log m \log B}{B}} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(b) \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}_j \|\mathbf{p}\| \tag{A.96}$$

$$|I_5| \lesssim \frac{\eta^2 t m}{\sqrt{B} a^2} \xi \|\mathbf{p}\|^2 + \frac{\eta^2}{a} \sum_{b=1}^t \sqrt{\frac{\log m \log B}{B}} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n|m}{|\mathcal{S}^n|a} p_n(t) \xi \|\mathbf{p}\| \tag{A.97}$$

with probability at least  $1 - (mB)^{-c}$ . (A.96) comes from (A.21). Then, combining (A.77),

(A.94), (A.95), (A.96) and (A.97), we can obtain

$$\begin{aligned}
& \left| \left\langle \mathbf{W}_{O_{(i,\cdot)}}^{(t+1)\top}, \mathbf{p}_j \right\rangle - \left\langle \mathbf{W}_{O_{(i,\cdot)}}^{(t)\top}, \mathbf{p}_j \right\rangle \right| \\
& \lesssim \frac{\eta}{a} \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{|\mathcal{S}_l^n|}{|\mathcal{S}^n|} |\mathcal{S}_l^n| \nu_n(t) + ((\sigma + \tau)) \right. \\
& \quad \left. + \sum_{b=1}^t \frac{|\mathcal{S}_1^n| p_n(b) \eta m}{|\mathcal{S}^n| a} \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}_j \right) \sqrt{\frac{\log m \log B}{B}} \|\mathbf{p}\|^2 + \frac{\eta^2 t m}{\sqrt{B} a^2} \xi \|\mathbf{p}\| \\
& \lesssim \frac{\eta}{a} \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{|\mathcal{S}_l^n|}{|\mathcal{S}^n|} |\mathcal{S}_l^n| \nu_n(t) + ((\sigma + \tau)) \right. \\
& \quad \left. + \sum_{b=1}^t \frac{|\mathcal{S}_1^n| p_n(b) \eta m}{|\mathcal{S}^n| a} \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}_j \right) \sqrt{\frac{\log m \log B}{B}} \|\mathbf{p}\|^2
\end{aligned} \tag{A.98}$$

Comparing (A.83) and (A.98), we have

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_j \lesssim \frac{1}{\sqrt{B}} \mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_1 \tag{A.99}$$

(3) If  $i \in \mathcal{U}_{l,n}(0)$ , following the derivation of (A.93) and (A.99), we can conclude that

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_2 \gtrsim \frac{\xi}{a B} \sum_{n \in \mathcal{B}_b} \frac{\eta(t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(t) + \xi \tag{A.100}$$

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p} \lesssim \frac{1}{\sqrt{B}} \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p}_2, \quad \text{for } \mathbf{p} \in \mathcal{P} \setminus \{\mathbf{p}_2\}, \tag{A.101}$$

(4) If  $i \notin (\mathcal{W}_{l,n}(0) \cup \mathcal{U}_{l,n}(0))$ ,

$$|I_2 + I_3| \leq \frac{\eta}{a} \sqrt{\frac{\log m \log B}{B}} \|\mathbf{p}\|^2 \tag{A.102}$$

Following (A.96) and (A.97), we have

$$|I_4| \leq \sum_{b=1}^t \frac{\eta^2}{a} \sqrt{\frac{\log m \log B}{B}} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(b) \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}_j \|\mathbf{p}\| \tag{A.103}$$

$$|I_5| \lesssim \frac{\eta^2 t m}{\sqrt{B} a^2} \xi \|\mathbf{p}\|^2 + \sum_{b=1}^t \frac{\eta^2}{a} \sqrt{\frac{\log m \log B}{B}} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n| m}{|\mathcal{S}^n| a} p_n(b) \xi \|\mathbf{p}\| \tag{A.104}$$

Hence, combining (A.102), (A.103), and (A.104), we can obtain

$$\begin{aligned}
& \left| \left\langle \mathbf{W}_{O_{(i,\cdot)}}^{(t+1)\top}, \mathbf{p} \right\rangle - \left\langle \mathbf{W}_{O_{(i,\cdot)}}^{(t)\top}, \mathbf{p} \right\rangle \right| \\
& \lesssim \frac{\eta}{a} \cdot (\|\mathbf{p}\| + (\sigma + \tau) + \sum_{b=1}^t \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| p_n(b) \eta m}{|\mathcal{S}^n| a} \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}_j) \sqrt{\frac{\log m \log B}{B}} \|\mathbf{p}\| \\
& \quad + \frac{\eta^2 t m}{\sqrt{B} a^2} \xi \|\mathbf{p}\|^2 \\
& \lesssim \frac{\eta}{a} \cdot (\|\mathbf{p}\| + (\sigma + \tau) + \sum_{b=1}^t \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| p_n(b) \eta m}{|\mathcal{S}^n| a} \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}_j) \sqrt{\frac{\log m \log B}{B}} \|\mathbf{p}\|,
\end{aligned} \tag{A.105}$$

Comparing (A.83) and (A.105), we have

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_j \lesssim \frac{1}{\sqrt{B}} \mathbf{W}_{O_{(j,\cdot)}}^{(t+1)} \mathbf{p}_1 \tag{A.106}$$

where  $j \in \mathcal{W}_{l,n}(0)$ .

(5) We finally study the bound of  $\mathbf{W}_{O_{(i,\cdot)}}^{(t)}$  and the product with the noise term according to the analysis above.

By (A.39), for the lucky neuron  $i$ , since that the update of  $\mathbf{W}_{O_{(i,\cdot)}}^{(t)}$  lies in the subspace spanned by  $\mathcal{P}$  and  $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M$  all have a unit norm, we can derive

$$\begin{aligned}
\|\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)}\|^2 &= \sum_{l=1}^M (\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_l)^2 \geq (\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_1)^2 \\
&\gtrsim \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta(t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(t) \right)^2
\end{aligned} \tag{A.107}$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{z}_l(t)\| \leq \left| (\sigma + \tau) \|\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)}\| \right| \tag{A.108}$$

For the unlucky neuron  $i$ , we can similarly obtain

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)}\|^2 \leq \frac{1}{B} \|\mathbf{W}_{O_{(j,\cdot)}}^{(t+1)}\|^2 \tag{A.109}$$

where  $j$  is a lucky neuron. The proof of Claim 1 ends here.

### A.4.2 Proof of Claim 2 of Lemma A.3.1

The proof of Claim 2 is one of the most challenging parts in our paper, since that we need to deal with the complicated softmax function. The core idea of proof is that we pay more attention on the changes of label-relevant features in the gradient update, which should be the most crucial factor based on our data model. We then show the attention map converges to be sparse as long as the data model satisfies (2.8).

We first study the gradient of  $\mathbf{W}_Q^{(t+1)}$  in part (a) and the gradient of  $\mathbf{W}_K^{(t+1)}$  in part (b).

(a) By (A.2), we have

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y^n)}{\partial \mathbf{W}_Q} \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y^n)}{\partial F(\mathbf{X}^n)} \frac{F(\mathbf{X}^n)}{\partial \mathbf{W}_Q} \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-y^n) \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i=1}^m a_{(l)i} \mathbb{1}[\mathbf{W}_{O(i,\cdot)} \mathbf{W}_V \mathbf{X} \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \geq 0] \\
&\quad \cdot \left( \mathbf{W}_{O(i,\cdot)} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \right. \\
&\quad \cdot \left. \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{x}_r^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \mathbf{W}_K (\mathbf{x}_s^n - \mathbf{x}_r^n) \mathbf{x}_l^{n\top} \right) \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-y^n) \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i=1}^m a_{(l)i} \mathbb{1}[\mathbf{W}_{O(i,\cdot)} \mathbf{W}_V \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \geq 0] \\
&\quad \cdot \left( \mathbf{W}_{O(i,\cdot)} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \right. \\
&\quad \cdot \left. (\mathbf{W}_K \mathbf{x}_s^n - \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{x}_r^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \mathbf{W}_K \mathbf{x}_r^n) \mathbf{x}_l^{n\top} \right)
\end{aligned} \tag{A.110}$$

For  $r, l \in \mathcal{S}_1^n$ , by (A.30) we have

$$\text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t)} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \gtrsim \frac{e^{\|\mathbf{q}_1(t)\|^2 - (\delta + \tau)\|\mathbf{q}_1(t)\|}}{|\mathcal{S}_1^n| e^{\|\mathbf{q}_1(t)\|^2 - (\delta + \tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \tag{A.111}$$

For  $r \notin \mathcal{S}_1^n$  and  $l \in \mathcal{S}_1^n$ , we have

$$\text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n) \lesssim \frac{1}{|\mathcal{S}_1^n| e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta + \tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \tag{A.112}$$

Therefore, for  $s, r, l \in \mathcal{S}_1^n$ , let

$$\mathbf{W}_K^{(t)} \mathbf{x}_s^n - \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{x}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{W}_K^{(t)} \mathbf{x}_r^n := \beta_1^n(t) \mathbf{q}_1(t) + \beta_2^n(t), \quad (\text{A.113})$$

where

$$\begin{aligned} \beta_1^n(t) &\gtrsim \frac{|\mathcal{S}^n| - |\mathcal{S}_1^n|}{|\mathcal{S}_1^n| e^{\|\mathbf{q}_1(t)\|^2 + (\delta+\tau)\|\mathbf{q}_1(t)\|} + |\mathcal{S}^n| - |\mathcal{S}_1^n|} \\ &:= \phi_n(t)(|\mathcal{S}^n| - |\mathcal{S}_1^n|). \end{aligned} \quad (\text{A.114})$$

$$\beta_1^n(t) \lesssim \nu_n(t)(|\mathcal{S}^n| - |\mathcal{S}_1^n|) \lesssim e^{2(\tau+\delta)\|\mathbf{q}_1(t)\|} \phi_n(t)(|\mathcal{S}^n| - |\mathcal{S}_1^n|) \leq \phi_n(t)(|\mathcal{S}^n| - |\mathcal{S}_1^n|) \quad (\text{A.115})$$

where the last inequality holds when the final iteration  $\log T \leq \Theta(1)$ .

$$\begin{aligned} \beta_2^n(t) &\approx \Theta(1) \cdot \mathbf{o}_j^n(t) + Q_e(t) \mathbf{r}_2(t) + \sum_{l=3}^M \gamma'_l \mathbf{r}_l(t) - \sum_{a=1}^M \sum_{r \in \mathcal{S}_l^n} \text{softmax}(\mathbf{x}_r^\top \mathbf{W}_K^{(t)\top} \\ &\quad \cdot \mathbf{W}_Q^{(t)} \mathbf{x}_l) \mathbf{r}_a(t) \\ &= \Theta(1) \cdot \mathbf{o}_j^n(t) + \sum_{l=1}^M \zeta'_l \mathbf{r}_l(t) \end{aligned} \quad (\text{A.116})$$

for some  $Q_e(t) > 0$  and  $\gamma'_l > 0$ . Here

$$|\zeta'_l| \leq \beta_1^n(t) \frac{|\mathcal{S}_l^n|}{|\mathcal{S}^n| - |\mathcal{S}_1^n|} \quad (\text{A.117})$$

for  $l \geq 2$ . Note that  $|\zeta'_l| = 0$  if  $|\mathcal{S}^n| = |\mathcal{S}_1^n|$ ,  $l \geq 2$ .

For  $i \in \mathcal{W}_{l,n}(0)$ , by Lemma A.5.2, Then we study how large the coefficient of  $\mathbf{q}_1(t)$  in (A.110).

If  $s \in \mathcal{S}_1^n$ , by basic computation given (A.19) to (A.22),

$$\begin{aligned} &\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \\ &\gtrsim \frac{p_n(t)}{|\mathcal{S}_1^n|} \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 |\mathcal{S}_1^n| m}{|\mathcal{S}^n| a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(b) - \sigma - \tau \right) \right. \\ &\quad \cdot + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_1^n|}{a |\mathcal{S}^n|} p_n(b) \\ &\quad \cdot \left. (1 - (\sigma + \tau)) \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta (t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{a} p_n(t) \right)^2 \right) \end{aligned} \quad (\text{A.118})$$

If  $s \in \mathcal{S}_2^n$  and  $j \in \mathcal{S}_1^n$ , from (A.23) to (A.26), we have

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \\ & \lesssim \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{W}_V^{(t)} \mathbf{x}_j^n \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \phi_n(t) \cdot \frac{|\mathcal{S}_1^n|}{p_n(t)} \end{aligned} \quad (\text{A.119})$$

If  $i \in \mathcal{W}_{l,n}(0)$ ,  $s \notin (\mathcal{S}_1^n \cup \mathcal{S}_2^n)$  and  $j \in \mathcal{S}_1^n$ ,

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \\ & \lesssim \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{W}_V^{(t)} \mathbf{x}_j^n \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \phi_n(t) \cdot \frac{|\mathcal{S}_1^n|}{\sqrt{B} p_n(t)} \end{aligned} \quad (\text{A.120})$$

by (A.27) to (A.29).

Hence, for  $i \in \mathcal{W}_{l,n}(0)$ ,  $j \in \mathcal{S}_1^g$ , combining (A.114) and (A.118), we have

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{q}_1(t)^\top \\ & \cdot (\mathbf{W}_K^{(t)} \mathbf{x}_s^n - \sum_{r=1}^L \text{softmax}(\mathbf{x}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{W}_K^{(t)} \mathbf{x}_r^n) \mathbf{x}_l^{n\top} \mathbf{x}_j^g \\ & \gtrsim p_n(t) \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 |\mathcal{S}_1^n| m}{|\mathcal{S}^n| a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(b) - \sigma - \tau \right) \right. \\ & \quad \cdot + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_1^n|}{a |\mathcal{S}^n|} p_n(b) \\ & \quad \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta (t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n| m}{|\mathcal{S}^n| a} p_n(t) \right)^2 \Big) \\ & \quad \cdot \phi_n(t) (|\mathcal{S}^n| - |\mathcal{S}_1^n|) \|\mathbf{q}_1(t)\|^2 \end{aligned} \quad (\text{A.121})$$

The following upper bounds use the lower bound of (A.121) since further results will be the

gap of these terms. For  $i \in \mathcal{U}_{l,n}(0)$  and  $l \in \mathcal{S}_1^n$ ,  $j \in \mathcal{S}_1^g$ , and  $k \in \mathcal{W}_{l,n}(0)$ ,

$$\begin{aligned}
& \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{q}_1(t)^\top \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{x}_s^n - \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{x}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{W}_K^{(t)} \mathbf{x}_r^n) \mathbf{x}_l^{n\top} \mathbf{x}_j^g \\
& \lesssim p_n(t) \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 |\mathcal{S}_2^n| m}{|\mathcal{S}^n| a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n|}{|\mathcal{S}^n|} p_n(b) - \sigma - \tau \right) \right. \\
& \quad \cdot + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_2^n|}{a |\mathcal{S}^n|} p_n(b) \\
& \quad \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta (t+1)^2 |\mathcal{S}_2^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n| m}{|\mathcal{S}^n| a} p_n(t) \right)^2 \Big) \\
& \quad \cdot \phi_n(t) |\mathcal{S}_2^n| \beta_1(t) \|\mathbf{q}_1(t)\|^2
\end{aligned} \tag{A.122}$$

For  $i \notin (\mathcal{W}_{l,n}(0) \cup \mathcal{U}_{l,n}(0))$  and  $l \in \mathcal{S}_1^n$ ,  $j \in \mathcal{S}_1^g$ , and  $k \in \mathcal{W}_{l,n}(0)$ ,

$$\begin{aligned}
& \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{q}_1(t)^\top \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{x}_s^n - \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{x}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{x}_r^n) \mathbf{x}_l^{n\top} \mathbf{x}_j^g \\
& \lesssim \mathbf{W}_{O_{(k,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{q}_1(t)^\top \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{x}_s^n - \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{x}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{W}_K^{(t)} \mathbf{x}_r^n) \mathbf{x}_l^{n\top} \mathbf{x}_j^g \cdot \frac{1}{\sqrt{B}}
\end{aligned} \tag{A.123}$$

To study the case when  $l \notin \mathcal{S}_1^n$  for all  $n \in [N]$ , we need to check all other  $l$ 's. Recall that we focus on the coefficient of  $\mathbf{q}_1(t)$  in this part. Based on the computation in (A.119) and (A.120), we know that the contribution of coefficient from non-discriminative patches is no

more than that from discriminative patches, i.e., for  $l \notin (\mathcal{S}_1^n \cup \mathcal{S}_2^n)$ ,  $n \in [N]$  and  $k \in \mathcal{S}_1^n$ ,

$$\begin{aligned} & \left| \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{q}_1(t)^\top \right. \\ & \quad \cdot \left. (\mathbf{W}_K^{(t)} \mathbf{x}_s^n - \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{W}_K^{(t)} \mathbf{x}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{W}_K^{(t)} \mathbf{x}_r^n) \mathbf{x}_l^{n\top} \mathbf{x}_j^g \right| \\ & \lesssim \left| \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_k^n) \mathbf{q}_1(t)^\top \right. \\ & \quad \cdot \left. (\mathbf{W}_K^{(t)} \mathbf{x}_s^n - \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{W}_K^{(t)} \mathbf{x}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{W}_K^{(t)} \mathbf{x}_r^n) \mathbf{x}_k^{n\top} \mathbf{x}_j^g \right| \end{aligned} \quad (\text{A.124})$$

Similar to (A.121), we have that for  $l \in \mathcal{S}_2^n$ ,  $j \in \mathcal{S}_1^g$ , and  $i \in \mathcal{U}_{l,n}(0)$ ,

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V^{(t)} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{q}_1(t)^\top \\ & \cdot (\mathbf{W}_K^{(t)} \mathbf{x}_s^n - \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{W}_K^{(t)} \mathbf{x}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{W}_K^{(t)} \mathbf{x}_r^n) \mathbf{x}_l^{n\top} \mathbf{x}_j^g \\ & \lesssim \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 |\mathcal{S}_2^n| m}{|\mathcal{S}^n| a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n|}{|\mathcal{S}^n|} p_n(b) \right) \right. \\ & \quad \cdot + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_2^n|}{a |\mathcal{S}^n|} p_n(b) \\ & \quad \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta (t+1)^2 |\mathcal{S}_2^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n| m}{|\mathcal{S}^n| a} p_n(t) \right)^2 \\ & \quad \cdot \beta_1(t) \|\mathbf{q}_1(t)\|^2 \lambda \frac{|\mathcal{S}_\#^n|}{|\mathcal{S}^n| - |\mathcal{S}_*^n|} \end{aligned} \quad (\text{A.125})$$

Therefore, by the update rule,

$$\begin{aligned} \mathbf{W}_Q^{(t+1)} \mathbf{x}_j^n &= \mathbf{W}_Q^{(t)} \mathbf{x}_j^n - \eta \left( \frac{\partial L}{\partial \mathbf{W}_Q} \Big| \mathbf{W}_Q^{(t)} \right) \mathbf{x}_j^n \\ &= \mathbf{r}_1(t) + K(t) \mathbf{q}_1(t) + \Theta(1) \cdot \mathbf{n}_j(t) + \sum_{b=0}^{t-1} |K_e(b)| \mathbf{q}_2(b) + \sum_{l=3}^M \gamma'_l \mathbf{q}_l(t) \\ &= (1 + K(t)) \mathbf{q}_1(t) + \Theta(1) \cdot \mathbf{n}_j(t) + \sum_{b=0}^{t-1} |K_e(b)| \mathbf{q}_2(b) + \sum_{l=3}^M \gamma'_l \mathbf{q}_l(t) \end{aligned} \quad (\text{A.126})$$

where the last step is by the condition that

$$\mathbf{q}_1(t) = k_1(t) \cdot \mathbf{r}_1(t), \quad (\text{A.127})$$

and

$$\mathbf{q}_2(t) = k_2(t) \cdot \mathbf{r}_2(t) \quad (\text{A.128})$$

for  $k_1(t) > 0$  and  $k_2(t) > 0$  from induction, i.e.,  $\mathbf{q}_1(t)$  and  $\mathbf{r}_1(t)$ ,  $\mathbf{q}_1(t)$  and  $\mathbf{r}_1(t)$  are in the same direction, respectively. Define  $qc_t(\mathbf{x}) = \mathbf{x}^\top \mathbf{q}_1(t) / \|\mathbf{q}_1(t)\|$  and denote

$$\begin{aligned} \Delta(l, i) = & a_{(l)_i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}} \mathbf{W}_V \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \geq 0] \\ & \cdot \left( \mathbf{W}_{O_{(i,:)}} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \right. \\ & \cdot \left. (\mathbf{W}_K \mathbf{x}_s^n - \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{x}_r^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \mathbf{W}_K \mathbf{x}_r^n) \mathbf{x}_l^{n\top} \right) \end{aligned} \quad (\text{A.129})$$

We also have

$$\begin{aligned} K(t) & \gtrsim \eta \frac{1}{B} \left( \left| \sum_{n \in \mathcal{B}_b} (-y^n) \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}_1^n} \sum_{i \in \mathcal{W}_{l,n}(0)} qc_t(\Delta(l, i)) \right| - \left| \sum_{n \in \mathcal{B}_b} (-y^n) \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}_1^n} \sum_{i \in \mathcal{U}_{l,n}(0)} qc_t(\Delta(l, i)) \right| \right. \\ & \quad \left. - \left| \sum_{n \in \mathcal{B}_b} (-y^n) \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}_1^n} \sum_{i \notin \mathcal{W}_{l,n}(0) \cup \mathcal{U}_{l,n}(0)} qc_t(\Delta(l, i)) \right| - \left| \sum_{n \in \mathcal{B}_b} (-y^n) \frac{1}{|\mathcal{S}^n|} \right. \right. \\ & \quad \left. \left. \cdot \sum_{l \in \mathcal{S}_2^n} \sum_{i=1}^m qc_t(\Delta(l, i)) \right| - \left| \sum_{n \in \mathcal{B}_b} (-y^n) \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n - \mathcal{S}_1^n - \mathcal{S}_2^n} \sum_{i=1}^m qc_t(\Delta(l, i)) \right| \right) \\ & \gtrsim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 |\mathcal{S}_1^n|m}{|\mathcal{S}^n|a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(b) - \sigma - \tau \right) \right. \\ & \quad \left. + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_1^n|}{a|\mathcal{S}^n|} p_n(b) \right. \\ & \quad \left. \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta (t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) \right)^2 \right) \end{aligned} \quad (\text{A.130})$$

$$\cdot \phi_n(t) (|\mathcal{S}^n| - |\mathcal{S}_1^n|) \|\mathbf{q}_1(t)\|^2$$

$$|\gamma'_l| \lesssim \frac{1}{B} \sum_{n \in \mathcal{B}_b} K(t) \cdot \frac{|\mathcal{S}_l^n|}{|\mathcal{S}^n| - |\mathcal{S}_1^n|} \quad (\text{A.131})$$

$$|K_e(t)| \lesssim \frac{1}{B} \sum_{n \in \mathcal{B}_b} K(t) \cdot \frac{|\mathcal{S}_2^n|}{|\mathcal{S}^n| - |\mathcal{S}_1^n|} \quad (\text{A.132})$$

as long as

$$\begin{aligned} & \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 |\mathcal{S}_1^n|m}{|\mathcal{S}^n|a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(b) - \sigma - \tau \right) \right. \\ & \quad \cdot + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_1^n|}{a|\mathcal{S}^n|} p_n(b) \\ & \quad \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta (t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) \right)^2 \Big) \\ & \quad \cdot \phi_n(t) (|\mathcal{S}^n| - |\mathcal{S}_1^n|) \|\mathbf{q}_1(t)\|^2 \\ & \gtrsim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n|m}{|\mathcal{S}^n|a} \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 |\mathcal{S}_2^n|m}{|\mathcal{S}^n|a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n|}{|\mathcal{S}^n|} p_n(b) \right) \right. \\ & \quad \cdot + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_2^n|}{a|\mathcal{S}^n|} p_n(b) \\ & \quad \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta (t+1)^2 |\mathcal{S}_2^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n|m}{|\mathcal{S}^n|a} p_n(t) \right)^2 \Big) \\ & \quad \cdot \beta_1(t) \|\mathbf{q}_1(t)\|^2 \lambda \frac{|\mathcal{S}_\#^n|}{|\mathcal{S}^n| - |\mathcal{S}_*^n|} \end{aligned} \quad (\text{A.133})$$

To find the sufficient condition for (A.133), we mainly need to compare  $\frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(t)$   $\phi_n(t) (|\mathcal{S}^n| - |\mathcal{S}_1^n|)$  and  $\frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n|}{|\mathcal{S}^n|} \beta_1(t) \lambda \frac{|\mathcal{S}_\#^n|}{|\mathcal{S}^n| - |\mathcal{S}_*^n|}$ .

When  $|\mathcal{S}^n| > |\mathcal{S}_1^n|$ , by (A.115),

$$\phi_n(t) (|\mathcal{S}^n| - |\mathcal{S}_1^n|) \gtrsim \beta_1^n(t) \quad (\text{A.134})$$

From Definition A.2.2, we know

$$1 \geq p_n(t) \geq p_n(0) = \Theta \left( \frac{|\mathcal{S}_1^n| e^{-(\delta+\tau)}}{|\mathcal{S}_1^n| e^{-(\delta+\tau)} + |\mathcal{S}^n| - |\mathcal{S}_1^n|} \right) \geq \Theta(e^{-(\delta+\tau)}) \quad (\text{A.135})$$

Meanwhile, by Hoeffding's inequality (A.12),

$$\begin{aligned} & \left| \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} - \sigma - \tau \right) - (1 - \sigma - \tau) \mathbb{E}\left[\frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|}\right] \right| \\ & \leq \left| \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} - \sigma - \tau \right) - \mathbb{E}\left[\frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|}\right] + \sigma + \tau \right| + \left| (\sigma + \tau)(1 - \mathbb{E}\left[\frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|}\right]) \right| \\ & \leq \sqrt{\frac{\log B}{B}} + \sigma + \tau \end{aligned} \quad (\text{A.136})$$

Therefore, a sufficient function for (A.133) is that

$$e^{-(\delta+\tau)}(1 - \sigma - \tau)(\alpha_* - \sqrt{\frac{\log B}{B}} - \sigma - \tau) \geq \sqrt{\frac{\log B}{B}} + \alpha_\# \quad (\text{A.137})$$

Hence,

$$\alpha_* \geq \frac{1 - \alpha_{nd}}{1 + (1 - (\tau + \sigma))e^{-(\delta+\tau)}} \quad (\text{A.138})$$

if

$$B \geq \Omega(1) \quad (\text{A.139})$$

Then we give a brief derivation of  $\mathbf{W}_Q^{(t+1)} \mathbf{x}_j^n$  for  $j \notin \mathcal{S}_1^n$  in the following.

To be specific, for  $j \in \mathcal{S}_n / (\mathcal{S}_1^n \cup \mathcal{S}_2^n)$ ,

$$\begin{aligned} & \left\langle \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \mathbf{Loss}(\mathbf{X}^n, y^n)}{\partial \mathbf{W}_Q^{(t)}} \mathbf{x}_j^n, \mathbf{q}_1(t) \right\rangle \\ & \gtrsim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p'_n(t) \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 |\mathcal{S}_1^n|m}{|\mathcal{S}^n|a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|}{|\mathcal{S}^n|} p_n(b) - \sigma - \tau \right) \right. \\ & \quad \cdot + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_1^n|}{a|\mathcal{S}^n|} p_n(b) \\ & \quad \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta (t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) \right)^2 \Big) \\ & \quad \cdot \phi_n(t) (|\mathcal{S}^n| - |\mathcal{S}_1^n|) \|\mathbf{q}_1(t)\|^2 \end{aligned} \quad (\text{A.140})$$

where

$$p'_n(t) = \frac{|\mathcal{S}_1^n| e^{\mathbf{q}_1(t)^\top \sum_{b=1}^t K(b) \mathbf{q}_1(0) - (\delta+\tau)] \|\mathbf{q}_1(t)\|}}{|\mathcal{S}_1^n| e^{\mathbf{q}_1(t)^\top \sum_{b=1}^t K(b) \mathbf{q}_1(b) - (\delta+\tau)] \|\mathbf{q}_1(t)\|} + |\mathcal{S}^n| - |\mathcal{S}_1^n|} \quad (\text{A.141})$$

When  $K(b)$  is close to  $0^+$ , we have

$$\prod_{b=1}^t \sqrt{1 + K(b)} \|\mathbf{q}(0)\|^2 \gtrsim e^{\sum_{b=1}^t K(b) \|\mathbf{q}_1(0)\|^2} \geq \sum_{b=1}^t K(b) \|\mathbf{q}_1(0)\|^2 \quad (\text{A.142})$$

where the first step is by  $\log(1 + x) \approx x$  when  $x \rightarrow 0^+$ . Therefore, one can derive that

$$\left\langle \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y^n)}{\partial \mathbf{W}_Q^{(t)}} \mathbf{x}_j^n, \mathbf{q}_1(t) \right\rangle \gtrsim \Theta(1) \cdot K(t) \quad (\text{A.143})$$

Meanwhile, the value of  $p'_n(t)$  will increase to 1 during training, making the component of  $\mathbf{q}_1(t)$  the major part in  $\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y^n)}{\partial \mathbf{W}_Q^{(t)}} \mathbf{x}_j^n$ .

Hence, if  $j \in \mathcal{S}_l^n$  for  $l \geq 3$ ,

$$\mathbf{W}_Q^{(t+1)} \mathbf{x}_j = \mathbf{q}_l(t) + \Theta(1) \cdot \mathbf{n}_j(t) + \Theta(1) \cdot \sum_{b=0}^{t-1} K(b) \mathbf{q}_1(b) + \sum_{l=2}^M \gamma'_l \mathbf{q}_l(t) \quad (\text{A.144})$$

Similarly, for  $j \in \mathcal{S}_2^n$ ,

$$\mathbf{W}_Q^{(t+1)} \mathbf{x}_j = (1 + K(t)) \mathbf{q}_2(t) + \Theta(1) \cdot \mathbf{n}_j(t) + \sum_{b=0}^{t-1} |K_e(b)| \mathbf{q}_1(b) + \sum_{l=2}^M \gamma'_l \mathbf{q}_l(t) \quad (\text{A.145})$$

(b) For the gradient of  $\mathbf{W}_K$ , we have

$$\begin{aligned} \frac{\partial \overline{\text{Loss}}_b}{\partial \mathbf{W}_K} &= \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y^n)}{\partial F(\mathbf{X})} \frac{F(\mathbf{X})}{\partial \mathbf{W}_K} \\ &= \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-y^n) \sum_{l \in \mathcal{S}^n} \sum_{i=1}^m a_{(l)_i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}} \mathbf{W}_V \mathbf{X} \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \geq 0] \\ &\quad \cdot \left( \mathbf{W}_{O_{(i,:)}} \sum_{s \in \mathcal{S}^n} \mathbf{W}_V \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \mathbf{W}_Q^\top \mathbf{x}_l^n \right. \\ &\quad \left. \cdot (\mathbf{x}_s^n - \sum_{r \in \mathcal{S}^n} \text{softmax}(\mathbf{x}_r^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \mathbf{x}_r^n)^\top \right) \end{aligned} \quad (\text{A.146})$$

Hence, for  $j \in \mathcal{S}_1^n$ , we can follow the derivation of (A.126) to obtain

$$\mathbf{W}_K^{(t+1)} \mathbf{x}_j = (1 + Q(t)) \mathbf{q}_1(t) + \Theta(1) \cdot \mathbf{o}_j^n(t) \pm \sum_{b=0}^{t-1} |Q_e(b)|(1 - \lambda) \mathbf{r}_2(b) + \sum_{l=3}^M \gamma'_l \mathbf{r}_l(t), \quad (\text{A.147})$$

where

$$Q(t) \geq K(t)(1 - \lambda) > 0 \quad (\text{A.148})$$

for  $\lambda < 1$  introduced in Assumption A.2.4, and

$$|\gamma_l| \lesssim \frac{1}{B} \sum_{n \in \mathcal{B}_b} Q(t) \cdot \frac{|\mathcal{S}_l^n|}{|\mathcal{S}^n| - |\mathcal{S}_*^n|} \quad (\text{A.149})$$

$$|Q_e(t)| \lesssim \frac{1}{B} \sum_{n \in \mathcal{B}_b} Q(t) \cdot \frac{|\mathcal{S}_\#^n|}{|\mathcal{S}^n| - |\mathcal{S}_*^n|} \quad (\text{A.150})$$

Similarly, for  $j \in \mathcal{S}_2^n$ , we have

$$\mathbf{W}_K^{(t+1)} \mathbf{x}_j \approx (1 + Q(t)) \mathbf{q}_2(t) + \Theta(1) \cdot \mathbf{o}_j^n(t) \pm \sum_{b=0}^{t-1} |Q_e(b)|(1 - \lambda) \mathbf{r}_1(b) + \sum_{l=3}^M \gamma'_l \mathbf{r}_l(t), \quad (\text{A.151})$$

For  $j \in \mathcal{S}_z^n$ ,  $z = 3, 4, \dots, M$ , we have

$$\begin{aligned} & \left| \left\langle \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y^n)}{\partial F(\mathbf{X})} \frac{F(\mathbf{X})}{\partial \mathbf{W}_K} \mathbf{x}_j^n, \mathbf{q}_1(t) \right\rangle \right| \\ & \lesssim \left| \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-y^n) \sum_{l \in \mathcal{S}_1^n} \sum_{i=1}^m a_{(l)i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}} \mathbf{W}_V \mathbf{X} \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_i^n) \geq 0] \right. \\ & \quad \cdot \left. \left( \mathbf{W}_{O_{(i,:)}} \left( \sum_{s \in \mathcal{S}_z^n} + \lambda \sum_{s \in \mathcal{S}_1^n} \right) \mathbf{W}_V \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \right) \|\mathbf{q}_1(t)\|^2 \right| \\ & \leq \lambda |Q_f(t)| \|\mathbf{q}_1(t)\|^2 \end{aligned} \quad (\text{A.152})$$

$$\begin{aligned} & \left| \left\langle \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y^n)}{\partial F(\mathbf{X})} \frac{F(\mathbf{X})}{\partial \mathbf{W}_K} \mathbf{x}_j^n, \mathbf{q}_z(t) \right\rangle \right| \\ & \lesssim \left| \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-y^n) \sum_{l \in \mathcal{S}_z^n} \sum_{i=1}^m a_{(l)i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}} \mathbf{W}_V \mathbf{X} \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_i^n) \geq 0] \right. \\ & \quad \cdot \left. \left( \mathbf{W}_{O_{(i,:)}} \left( \sum_{s \in \mathcal{S}_z^n} + \lambda \sum_{s \in \mathcal{S}_1^n} \right) \mathbf{W}_V \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \right) \|\mathbf{q}_z(t)\|^2 \right| \\ & \leq \lambda |Q_f(t)| \|\mathbf{q}_z(t)\|^2 \end{aligned} \quad (\text{A.153})$$

$$\begin{aligned} \mathbf{W}_K^{(t+1)} \mathbf{x}_j &\approx (1 \pm c_{k_1} \lambda |Q_f(t)|) \mathbf{q}_l(t) + \Theta(1) \cdot \mathbf{o}_j^n(t) \\ &\quad \pm c_{k_2} \lambda \cdot \sum_{b=0}^{t-1} |Q_f(b)| \mathbf{r}_1(b) \pm c_{k_3} \lambda \cdot \sum_{b=0}^{t-1} |Q_f(b)| \mathbf{r}_2(b) + \sum_{i=3}^M \gamma'_i \mathbf{r}_i(t), \end{aligned} \quad (\text{A.154})$$

where  $0 < c_{k_1}, c_{k_2}, c_{k_3} < 1$ , and

$$|Q_f(t)| \lesssim Q(t) \quad (\text{A.155})$$

Therefore, for  $l \in \mathcal{S}_1^n$ , if  $j \in \mathcal{S}_1^n$ ,

$$\begin{aligned} &\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n \\ &\gtrsim (1 + K(t))(1 + Q(t)) \|\mathbf{q}_1(t)\|^2 - (\delta + \tau) \|\mathbf{q}_1(t)\| + \sum_{b=0}^{t-1} K_e(b) \sum_{b=0}^{t-1} Q_e(b) \\ &\quad \cdot \|\mathbf{q}_2(b)\| \|\mathbf{r}_2(b)\| + \sum_{l=3}^M \gamma_l \gamma'_l \|\mathbf{q}_l(t)\| \|\mathbf{r}_l(t)\| \\ &\gtrsim (1 + K(t))(1 + Q(t)) \|\mathbf{q}_1(t)\|^2 - (\delta + \tau) \|\mathbf{q}_1(t)\| \\ &\quad - \sqrt{\sum_{l=3}^M \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} Q(t) \frac{|\mathcal{S}_l^n|}{|\mathcal{S}^n| - |\mathcal{S}_*^n|} \right)^2 \|\mathbf{r}_l(t)\|^2} \\ &\quad \cdot \sqrt{\sum_{l=3}^M \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} K(t) \frac{|\mathcal{S}_l^n|}{|\mathcal{S}^n| - |\mathcal{S}_*^n|} \right)^2 \|\mathbf{q}_l(t)\|^2} \\ &\gtrsim (1 + K(t) + Q(t)) \|\mathbf{q}_1(t)\|^2 - (\delta + \tau) \|\mathbf{q}_1(t)\| \end{aligned} \quad (\text{A.156})$$

where the second step is by Cauchy-Schwarz inequality.

If  $j \notin \mathcal{S}_1^n$ ,

$$\begin{aligned} &\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n \\ &\lesssim Q_f(t) \|\mathbf{q}_1(t)\|^2 + (\delta + \tau) \|\mathbf{q}_1(t)\| \end{aligned} \quad (\text{A.157})$$

Hence, for  $j, l \in \mathcal{S}_1^n$ ,

$$\text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n) \gtrsim \frac{e^{(1+K(t)) \|\mathbf{q}_1(t)\|^2 - (\delta + \tau) \|\mathbf{q}_1(t)\|}}{|\mathcal{S}_1^n| e^{(1+K(t)) \|\mathbf{q}_1(t)\|^2 - (\delta + \tau) \|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \quad (\text{A.158})$$

$$\begin{aligned}
& \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n) - \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \\
& \gtrsim \frac{e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|}}{|\mathcal{S}_1^n| e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \\
& \quad - \frac{e^{\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|}}{|\mathcal{S}_1^n| e^{\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \\
& = \frac{|\mathcal{S}^n| - |\mathcal{S}_1^n|}{(|\mathcal{S}_1^n| e^x + (|\mathcal{S}^n| - |\mathcal{S}_1^n|))^2} e^{\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} (e^{K(t)} - 1) \\
& \geq \frac{|\mathcal{S}^n| - |\mathcal{S}_1^n|}{(|\mathcal{S}_1^n| e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|))^2} e^{\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} \cdot K(t)
\end{aligned} \tag{A.159}$$

where the second to last step is by the Mean Value Theorem with

$$x \in [\|\mathbf{q}_1(t)\|^2 - (\delta + \tau)\|\mathbf{q}_1(t)\|, (1 + K(t))\|\mathbf{q}_1(t)\|^2 - (\delta + \tau)\|\mathbf{q}_1(t)\|] \tag{A.160}$$

We then need to study if  $l \notin (\mathcal{S}_1^n \cup \mathcal{S}_2^n)$  and  $j \in \mathcal{S}_1^n$ , i.e.,

$$\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n \gtrsim (1 + Q(t)) \sum_{b=0}^{t-1} |K(b)| \|\mathbf{q}_1(b)\| - (\delta + \tau) \|\mathbf{q}_1(t)\| \tag{A.161}$$

For  $j, l \notin (\mathcal{S}_1^n \cup \mathcal{S}_2^n)$ ,

$$\begin{aligned}
\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n & \lesssim \pm c_{k_2} \lambda \cdot \sum_{b=0}^{t-1} |Q_f(b)| \mathbf{r}_1(b) \pm c_{k_3} \lambda \cdot \sum_{b=0}^{t-1} |Q_f(b)| \mathbf{r}_2(b) \\
& \quad + (1 \pm c_{k_1} \lambda |Q_f(t)|) \|\mathbf{q}_l(t)\|^2
\end{aligned} \tag{A.162}$$

We know that the magnitude of  $\|\mathbf{q}_1(t)\|$  increases along the training and finally reaches no larger than  $\Theta(\sqrt{\log T})$ . At the final step, we have

$$\sum_{b=0}^{t-1} K(b) \|\mathbf{q}_1(b)\| \geq \frac{T}{e^{\|\mathbf{q}_1(T)\|^2 - (\delta + \tau)\|\mathbf{q}_1(T)\|}} \geq \Theta(\sqrt{\log T}) \tag{A.163}$$

Therefore, when  $t$  is large enough during the training but before the final step of convergence, we have if  $j', l \notin (\mathcal{S}_1^n \cup \mathcal{S}_2^n)$  and  $j \in \mathcal{S}_1^n$ , we can obtain

$$(\mathbf{x}_j^n - \mathbf{x}_{j'}^n)^\top \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n \gtrsim \Theta(1) \cdot ((1 + K(t))\|\mathbf{q}_1(t)\|^2 - (\delta + \tau)\|\mathbf{q}_1(t)\|) \tag{A.164}$$

We can derive the same conclusion for  $j \in \mathcal{S}_2^n$  in (A.164). Therefore, by  $|\mathcal{S}_2| \leq e^{-(\delta+\tau)}(1 -$

$(\sigma - \tau))|\mathcal{S}_1^n|$  in (A.138), we can obtain

$$\begin{aligned} & \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n) \\ & \gtrsim \frac{e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|}}{(|\mathcal{S}_1^n| + |\mathcal{S}_2^n|)e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n| - |\mathcal{S}_2^n|))} \\ & \gtrsim \frac{e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|}}{|\mathcal{S}_1^n|e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \end{aligned} \quad (\text{A.165})$$

$$\begin{aligned} & \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n) - \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \\ & \gtrsim \frac{|\mathcal{S}^n| - |\mathcal{S}_1^n|}{(|\mathcal{S}_1^n|e^{\Theta(1)((1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|)} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|))^2} e^{\Theta(1)(\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|)} \\ & \quad \cdot K(t) \\ & \gtrsim \frac{|\mathcal{S}^n| - |\mathcal{S}_1^n|}{(|\mathcal{S}_1^n|e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|))^2} e^{\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} \cdot K(t) \end{aligned} \quad (\text{A.166})$$

Meanwhile, for  $l \in \mathcal{S}_1^n$  and  $j \notin \mathcal{S}_1^n$ ,

$$\text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n) \lesssim \frac{1}{|\mathcal{S}_1^n|e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \quad (\text{A.167})$$

$$\begin{aligned} & \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t+1)} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l^n) - \text{softmax}(\mathbf{x}_j^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \\ & \lesssim \frac{1}{|\mathcal{S}_1^n|e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \\ & \quad - \frac{1}{|\mathcal{S}_1^n|e^{\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|)} \\ & = - \frac{|\mathcal{S}_1^n|}{(|\mathcal{S}_1^n|e^x + (|\mathcal{S}^n| - |\mathcal{S}_1^n|))^2} e^{\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} (e^{K(t)} - 1) \\ & \leq - \frac{|\mathcal{S}_1^n|}{(|\mathcal{S}_1^n|e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} + (|\mathcal{S}^n| - |\mathcal{S}_1^n|))^2} e^{\|\mathbf{q}_1(t)\|^2 - (\delta+\tau)\|\mathbf{q}_1(t)\|} \cdot K(t) \end{aligned} \quad (\text{A.168})$$

where the second to last step is by the Mean Value Theorem with

$$x \in [\|\mathbf{q}_1(t)\|^2 - (\delta + \tau)\|\mathbf{q}_1(t)\|, (1 + K(t))\|\mathbf{q}_1(t)\|^2 - (\delta + \tau)\|\mathbf{q}_1(t)\|] \quad (\text{A.169})$$

The same conclusion holds if  $l \notin (\mathcal{S}_1^n \cup \mathcal{S}_2^n)$  and  $j \notin \mathcal{S}_1^n$ .

Note that

$$\mathbf{q}_1(t+1) = \sqrt{(1 + K(t))} \mathbf{q}_1(t) \quad (\text{A.170})$$

$$\mathbf{q}_2(t+1) = \sqrt{(1 + K(t))} \mathbf{q}_2(t) \quad (\text{A.171})$$

$$\mathbf{r}_1(t+1) = \sqrt{(1+Q(t))} \mathbf{r}_1(t) \quad (\text{A.172})$$

$$\mathbf{r}_2(t+1) = \sqrt{(1+Q(t))} \mathbf{r}_2(t) \quad (\text{A.173})$$

It can also be verified that this claim holds when  $t = 1$ .

#### A.4.3 Proof of Claim 3 of Lemma A.3.1

The computation of the gradient of  $\mathbf{W}_V$  is straightforward. The gradient would be related to  $\mathbf{W}_O$  by their connections. One still need to study the influence of the gradient on different patterns, where we introduce the discussion for the term  $V_i(t)$ 's.

For the gradient of  $\mathbf{W}_V$ , by (A.3) we have

$$\begin{aligned} & \frac{\partial \overline{\text{Loss}}_b}{\partial \mathbf{W}_V} \\ &= \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y^n)}{\partial F(\mathbf{X}^n)} \frac{\partial F(\mathbf{X}^n)}{\partial \mathbf{W}_V} \\ &= -y \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i=1}^m a_{(l)i}^* \mathbb{1}[\mathbf{W}_{O_{(i,:)}} \mathbf{W}_V \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n) \geq 0] \\ & \quad \cdot \mathbf{W}_{O_{(i,:)}}^\top \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l^n)^\top \mathbf{X}^{n\top} \end{aligned} \quad (\text{A.174})$$

Consider a data  $\{\mathbf{X}^n, y^n\}$  where  $y^n = 1$ . Let  $l \in \mathcal{S}_1^n$

$$\sum_{s \in \mathcal{S}_1^n} \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \geq p_n(t) \quad (\text{A.175})$$

Then for  $j \in \mathcal{S}_1^g$ ,  $g \in [N]$ ,

$$\begin{aligned} & \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y^n)}{\partial \mathbf{W}_V^{(t)}} \Big| \mathbf{W}_V^{(t)} \mathbf{x}_j^g \\ &= \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-y^n) \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \sum_{i=1}^m a_{(l)i} \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s \in \mathcal{S}^n} \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \\ & \quad \mathbf{W}_V^{(t)} \mathbf{x}_s^n \geq 0] \mathbf{W}_{O_{(i,:)}}^\top \sum_{s \in \mathcal{S}^n} \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l) \mathbf{x}_s^{n\top} \mathbf{x}_j^g \\ &= \sum_{i \in \mathcal{W}_{l,n}(0)} V_i(t) \mathbf{W}_{O_{(i,:)}}^\top + \sum_{i \notin \mathcal{W}_{l,n}(0)} \lambda V_i(t) \mathbf{W}_{O_{(i,:)}}^\top, \end{aligned} \quad (\text{A.176})$$

When  $t = 0$ , only  $\mathbf{W}_{O_{(i,:)}}^{(t)}$  from  $i \in \mathcal{W}_{l,n}(0)$  can ensure the indicator be 1 for  $l \in \mathcal{S}_1^n$ . If

$i \in \mathcal{W}_{l,n}(0)$ , by the fact that  $\mathcal{S}_\#^n$  contributes more to  $V_i(t)$  compared to  $\mathcal{S}_l^n$  for  $l \geq 3$  and Assumption A.2.4, we have

$$\begin{aligned} V_i(t) &\lesssim \frac{1}{2B} \sum_{n \in \mathcal{B}_{b+}} -\frac{|\mathcal{S}_1^n|}{a|\mathcal{S}^n|} p_n(t) + \frac{|\mathcal{S}_2^n|}{a|\mathcal{S}^n|} |\lambda| \nu_n(t) (|\mathcal{S}^n| - |\mathcal{S}_1^n|) \\ &\lesssim \frac{1}{2B} \sum_{n \in \mathcal{B}_{b+}} -\frac{|\mathcal{S}_1^n|}{a|\mathcal{S}^n|} p_n(t) \end{aligned} \quad (\text{A.177})$$

Similarly, if  $i \in \mathcal{U}_{l,n}(0)$ ,

$$V_i(t) \gtrsim \frac{1}{2B} \sum_{n \in \mathcal{B}_{b-}} \frac{|\mathcal{S}_2^n|}{a|\mathcal{S}^n|} p_n(t) \quad (\text{A.178})$$

if  $i$  is an unlucky neuron, by Hoeffding's inequality in (A.12), we have

$$\begin{aligned} |V_i(t)| &\leq \frac{1}{\sqrt{B}} \cdot \frac{1}{a} \\ &\lesssim \frac{1}{\sqrt{Ba}} \end{aligned} \quad (\text{A.179})$$

which is smaller than  $V_i(t)$  for  $i \in \mathcal{W}_{l,n}(0)$  and  $i \in \mathcal{U}_{l,n}(0)$  given  $B \geq \Omega(1)$ . When  $t$  is large, since unlucky neurons can activate tokens from both  $+1$  and  $-1$  data, we still have that (A.177), (A.178) and (A.179) hold. Then, for  $i \in \mathcal{W}_{l,n}(0)$ , we have

$$\begin{aligned} &- \eta \sum_{b=1}^t \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \sum_{j \in \mathcal{W}_{l,n}(0)} V_j(b) \mathbf{W}_{O_{(j,\cdot)}}^{(b)} {}^\top \\ &\gtrsim \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{|\mathcal{S}_1^n|}{a|\mathcal{S}^n|} p_n(b) \end{aligned} \quad (\text{A.180})$$

$$\cdot \left( \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta(t+1)^2 |\mathcal{S}_1^n|}{|\mathcal{S}^n|} \frac{1}{4B} \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_1^n|m}{|\mathcal{S}^n|a} p_n(t) \right)^2 \right. \quad (\text{A.181})$$

$$\begin{aligned} &|\eta \sum_{b=1}^t \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \sum_{j \in \mathcal{U}_{l,n}(0)} V_j(b) \mathbf{W}_{O_{(j,\cdot)}}^{(b)} {}^\top| \\ &\lesssim \frac{\eta}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_2^n| p_n(b)m}{|\mathcal{S}^n|a} \|\mathbf{W}_{O_{(i,\cdot)}}^{(b)}\|^2 \|\mathbf{p}_1\|^2 \end{aligned} \quad (\text{A.181})$$

$$-\eta t \mathbf{W}_{O_{(i,\cdot)}} \sum_{j \notin (\mathcal{W}_{l,n}(0) \cup \mathcal{U}_{l,n}(0))} V_j(t) \mathbf{W}_{O_{(j,\cdot)}} {}^\top \lesssim \frac{\eta tm \|\mathbf{p}\|^2}{Ba} \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\|^2 \quad (\text{A.182})$$

Hence,

(1) If  $j \in \mathcal{S}_1^n$  for one  $n \in [N]$ ,

$$\begin{aligned} \mathbf{W}_V^{(t+1)} \mathbf{x}_j^n &= \mathbf{W}_V^{(t)} \mathbf{x}_j^n - \eta \left( \frac{\partial L}{\partial \mathbf{W}_V} \Big| \mathbf{W}_V^{(t)} \right) \mathbf{x}_j^n \\ &= \mathbf{p}_1 - \eta \sum_{b=1}^{t+1} \sum_{i \in \mathcal{W}_{l,n}(0)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} {}^\top - \eta \sum_{b=1}^{t+1} \sum_{i \notin \mathcal{W}_{l,n}(0)} \lambda V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} {}^\top + \mathbf{z}_j(t) \end{aligned} \quad (\text{A.183})$$

(2) If  $j \in \mathcal{S}_2^n$ , we have

$$\begin{aligned} \mathbf{W}_V^{(t+1)} \mathbf{x}_j &= \mathbf{W}_V^{(0)} \mathbf{x}_j^n - \eta \left( \frac{\partial L}{\partial \mathbf{W}_V} \Big| \mathbf{W}_V^{(0)} \right) \mathbf{x}_j^n \\ &= \mathbf{p}_2 - \eta \sum_{b=1}^{t+1} \sum_{i \in \mathcal{U}_{l,n}(0)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} {}^\top - \eta \sum_{b=1}^{t+1} \sum_{i \notin \mathcal{U}_{l,n}(0)} \lambda V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} {}^\top + \mathbf{z}_j(t) \end{aligned} \quad (\text{A.184})$$

(3) If  $j \in \mathcal{S}^n / (\mathcal{S}_1^n \cup \mathcal{S}_2^n)$ , we have

$$\begin{aligned} \mathbf{W}_V^{(t+1)} \mathbf{x}_j^n &= \mathbf{W}_V^{(0)} \mathbf{x}_j^n - \eta \left( \frac{\partial L}{\partial \mathbf{W}_V} \Big| \mathbf{W}_V^{(0)} \right) \mathbf{x}_j^n \\ &= \mathbf{p}_j - \eta \sum_{b=1}^{t+1} \sum_{i=1}^m \lambda V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} {}^\top + \mathbf{z}_j(t) \end{aligned} \quad (\text{A.185})$$

Here

$$\|\mathbf{z}_j(t)\| \leq (\sigma + \tau) \quad (\text{A.186})$$

for  $t \geq 1$ . Note that this claim also holds when  $t = 1$ .

## A.5 Other Useful Lemmas

**Lemma A.5.1.** *The number of lucky neurons at the initialization  $|\mathcal{W}_{l,n}(0)|$ ,  $|\mathcal{U}_{l,n}(0)|$  satisfies*

$$|\mathcal{W}_{l,n}(0)|, |\mathcal{U}_{l,n}(0)| \geq \Omega(m) \quad (\text{A.187})$$

**Proof:**

We know that the Gaussian initialization of  $\mathbf{W}_{O_{(i,\cdot)}}^{(0)}$  generates a uniform distribution on the  $m_a - 1$ -sphere. Therefore,

$$\Pr(i \in \mathcal{W}_{l,n}(0)) = A_{m_a}^{\text{cap}}(\phi) / A_{m_a}, \quad (\text{A.188})$$

where  $A_{m_a}$  is the surface area of an  $m_a - 1$ -sphere.  $A_{m_a}^{cap}(\phi)$  is the surface area of a  $m_a - 1$ -spherical cap with  $\phi$  as the colatitude angle. By Equation 1 in [259], we have

$$\Pr(i \in \mathcal{W}_{l,n}(0)) = \frac{1}{2} I_{\sin^2 \phi} \left( \frac{m_a - 1}{2}, \frac{1}{2} \right) = \frac{\int_0^{\sin^2 \phi} t^{\frac{m_a-3}{2}} (1-t)^{-\frac{1}{2}} dt}{\int_0^1 t^{\frac{m_a-3}{2}} (1-t)^{-\frac{1}{2}} dt}, \quad (\text{A.189})$$

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

$$\phi \leq \pi/2 - \sigma - \tau = \pi/2 - \Theta(1/M), \quad (\text{A.190})$$

we have that

$$\begin{aligned} & \frac{\int_0^{\sin^2 \phi} t^{\frac{m_a-3}{2}} (1-t)^{-\frac{1}{2}} dt}{\int_0^1 t^{\frac{m_a-3}{2}} (1-t)^{-\frac{1}{2}} dt} \\ & \geq \frac{\int_0^{\cos^2 1/M} t^{\frac{m_a-3}{2}} (1-t)^{-\frac{1}{2}} dt}{\int_0^1 t^{\frac{m_a-3}{2}} (1-t)^{-\frac{1}{2}} dt} \\ & \geq 1 - \frac{\int_{1-\Theta(1/M^2)}^1 t^{\frac{m_a-3}{2}} (1-t)^{-\frac{1}{2}} dt}{\int_0^1 t^{\frac{m_a-3}{2}} (1-t)^{-\frac{1}{2}} dt} \\ & \geq 1 - \frac{\int_{1-\Theta(1/M^2)}^1 (1-t)^{-\frac{1}{2}} dt}{\int_0^1 t^{\frac{m_a-3}{2}} dt} \\ & = 1 - \frac{\Theta(\frac{2}{M})}{\frac{1}{\frac{m_a-1}{2}}} \\ & \geq \Theta(1) \end{aligned} \quad (\text{A.191})$$

where the last step is by  $m_a = \Theta(M)$ . This implies that

$$|\mathcal{W}_{l,n}(0)| \geq \Omega(m) \quad (\text{A.192})$$

Likewise, the conclusion holds for  $\mathcal{U}_{l,n}(0)$ .

**Lemma A.5.2.** *Under the condition that  $m \gtrsim M^2 \log N$ , we have the following result.*

*For  $i \in \mathcal{W}_{l,n}(0)$  and  $l \in \mathcal{S}_1^n$ , we have*

$$\mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{V}_l^n(t)] = 1; \quad (\text{A.193})$$

For  $i \in \mathcal{U}_{l,n}(0)$  and  $l \in \mathcal{S}_2^n$ , we have

$$\mathbb{E}[\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t)] = 1; \quad (\text{A.194})$$

**Proof:**

We prove this lemma by induction.

When  $t = 0$ . For  $i \in \mathcal{W}_l^n(0)$  and  $l \in \mathcal{S}_1^n$ , we have that

$$\begin{aligned} & \mathbf{W}_{O_{(i,:)}}^{(0)} \left( \sum_{s \in \mathcal{S}_1^n} \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{p}_1 + \mathbf{z}(0) + \sum_{j \neq 1} W_j^n(0) \mathbf{p}_j \right) \\ & \gtrsim \xi(\Theta(1) - \sigma - \tau) > 0 \end{aligned} \quad (\text{A.195})$$

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

$$\begin{aligned} & \mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t) \\ &= \mathbf{W}_{O_{(i,:)}}^{(t)} \left( \sum_{s \in \mathcal{S}_1^n} \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{p}_1 + \mathbf{z}(t) + \sum_{j \neq 1} W_j^n(t) \mathbf{p}_j \right. \\ & \quad \left. - \eta \sum_{b=0}^{t-1} \left( \sum_{i \in \mathcal{W}_{l,n}(0)} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)\top} + \sum_{i \notin \mathcal{W}_{l,n}(0)} V_i(b) \lambda \mathbf{W}_{O_{(i,:)}}^{(b)\top} \right) \right) \end{aligned} \quad (\text{A.196})$$

Denote  $\theta_{l,n}^i$  as the angle between  $\mathbf{V}_l^n(0)$  and  $\mathbf{W}_{O_{(i,:)}}^{(0)}$ . Since that  $\mathbf{W}_{O_{(i,:)}}^{(0)}$  is initialized uniformed on the  $m_a - 1$ -sphere, we have  $\mathbb{E}[\theta_{l,n}^i] = 0$ . By Hoeffding's inequality (A.12), we have

$$\left\| \frac{1}{|\mathcal{W}_{l,n}(0)|} \sum_{i \in \mathcal{W}_{l,n}(0)} \theta_{l,n}^i - \mathbb{E}[\theta_{l,n}^i] \right\| = \left\| \frac{1}{|\mathcal{W}_{l,n}(0)|} \sum_{i \in \mathcal{W}_{l,n}(0)} \theta_{l,n}^i \right\| \leq \sqrt{\frac{\log N}{m}}, \quad (\text{A.197})$$

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

$$\left\| \frac{1}{|\mathcal{W}_{l,n}(0)|} \sum_{i \in \mathcal{W}_{l,n}(0)} \theta_{l,n}^i - \mathbb{E}[\theta_{l,n}^i] \right\| \leq O\left(\frac{1}{M}\right). \quad (\text{A.198})$$

Therefore, for  $i \in \mathcal{W}_{l,n}(0)$ , we have

$$\mathbf{W}_{O_{(i,:)}} \sum_{b=0}^{t-1} \sum_{i \in \mathcal{W}_{l,n}(0)} \mathbf{W}_{O_{(i,:)}}^{(b)} > 0 \quad (\text{A.199})$$

Similarly, we have that  $\sum_{b=0}^{t-1} \sum_{i \notin \mathcal{W}_{l,n}(0)} \mathbf{W}_{O_{(i,:)}}^{(b)}$  is close to  $-\mathbf{V}_l^n(0)$ . Given that  $\lambda < 1$ , we can approximately acquire that

$$-\mathbf{W}_{O_{(i,:)}}^{(0)} \eta \sum_{b=0}^{t-1} \left( \sum_{i \in \mathcal{W}_{l,n}(0)} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)}{}^\top + \sum_{i \notin \mathcal{W}_{l,n}(0)} V_i(b) \lambda \mathbf{W}_{O_{(i,:)}}^{(b)}{}^\top \right) > 0 \quad (\text{A.200})$$

Since that  $i \in \mathcal{W}_{l,n}(0)$ , we have

$$\mathbf{W}_{O_{(i,:)}}^{(0)} \left( \sum_{s \in \mathcal{S}_1^n} \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{p}_1 + \mathbf{z}(t) + \sum_{j \neq 1} W_j^n(t) \mathbf{p}_j \right) > 0 \quad (\text{A.201})$$

Therefore, we have

$$\mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{V}_l^n(t) > 0 \quad (\text{A.202})$$

Meanwhile, the addition from  $\mathbf{W}_{O_{(i,:)}}^{(0)}$  to  $\mathbf{W}_{O_{(i,:)}}^{(1)}$  is approximately a summation of multiple  $\mathbf{V}_j^n(0)$  such that  $\mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{V}_j^n(0) > 0$  and  $j \in \mathcal{S}_1^n$ . Therefore,  $\mathbf{V}_j^n(0)^\top \mathbf{V}_l^n(0) > 0$ . Therefore, we can obtain

$$\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t) > 0 \quad (\text{A.203})$$

**(2) Suppose that the conclusion holds when  $t = s$ . When  $t = s + 1$ , we can follow the derivation of the case where  $t = 1$ . Although the unit vector of  $\mathbf{W}_{O_{(i,:)}}^{(t)}$  no longer follows a uniform distribution, we know that (A.197) holds since the angle is bounded and has a mean which is very close to  $\mathbf{V}_l^n(0)$ . Then, the conclusion still holds.**

One can develop the proof for  $\mathcal{U}_{l,n}(0)$  following the above steps.

## A.6 Extension to More General Cases

### A.6.1 Extension to Multi-Classification

Consider the classification problem with four classes, we use the label  $y \in \{+1, -1\}^2$  to denote the corresponding class. Similarly to the previous setup, there are four orthogonal discriminative patterns. In the output layer,  $a_{l(i)}$  for the data  $(\mathbf{X}^n, y^n)$  is changed into an  $\mathbb{R}^2$  vector  $\mathbf{a}_{l(i)}$  for  $l \in [|\mathcal{S}^n|]$  and  $i \in [m]$ . Hence, we define

$$\mathbf{F}(\mathbf{X}^n) = \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \mathbf{a}_{l(i)} \text{Relu}(\mathbf{W}_O \mathbf{W}_V \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_Q^\top \mathbf{W}_K \mathbf{x}_l^n)) \quad (\text{A.204})$$

$$F_1(\mathbf{X}^n) = \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{l_{1(i)}} \text{Relu}(\mathbf{W}_O \mathbf{W}_V \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_Q^\top \mathbf{W}_K \mathbf{x}_l^n)) \quad (\text{A.205})$$

$$F_2(\mathbf{X}^n) = \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} a_{l_{2(i)}} \text{Relu}(\mathbf{W}_O \mathbf{W}_V \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_Q^\top \mathbf{W}_K \mathbf{x}_l^n)) \quad (\text{A.206})$$

The dataset  $\mathcal{D}$  can be divided into four groups as

$$\mathcal{D}_1 = \{(\mathbf{X}^n, \mathbf{y}^n) | \mathbf{y}^n = (1, 1)\} \quad (\text{A.207})$$

$$\mathcal{D}_2 = \{(\mathbf{X}^n, \mathbf{y}^n) | \mathbf{y}^n = (1, -1)\} \quad (\text{A.208})$$

$$\mathcal{D}_3 = \{(\mathbf{X}^n, \mathbf{y}^n) | \mathbf{y}^n = (-1, 1)\} \quad (\text{A.209})$$

$$\mathcal{D}_4 = \{(\mathbf{X}^n, \mathbf{y}^n) | \mathbf{y}^n = (-1, -1)\} \quad (\text{A.210})$$

The hinge loss function for data  $(\mathbf{X}^n, \mathbf{y}^n)$  will be

$$\text{Loss}(\mathbf{X}^n, \mathbf{y}^n) = \max\{1 - \mathbf{y}^{n\top} \mathbf{F}(\mathbf{X}^n), 0\} \quad (\text{A.211})$$

We can divide the weights  $\mathbf{W}_{O_{(i,:)}}$  ( $i \in [m]$ ) into two groups, respectively.

$$\mathcal{W}_1 = \{i | \mathbf{a}_{l_{(i)}} = \frac{1}{\sqrt{m}} \cdot (1, 1)\} \quad (\text{A.212})$$

$$\mathcal{W}_2 = \{i | \mathbf{a}_{l_{(i)}} = \frac{1}{\sqrt{m}} \cdot (1, -1)\} \quad (\text{A.213})$$

$$\mathcal{W}_3 = \{i | \mathbf{a}_{l_{(i)}} = \frac{1}{\sqrt{m}} \cdot (-1, 1)\} \quad (\text{A.214})$$

$$\mathcal{W}_4 = \{i | \mathbf{a}_{l_{(i)}} = \frac{1}{\sqrt{m}} \cdot (-1, -1)\} \quad (\text{A.215})$$

Therefore, for  $\mathbf{W}_{O_u}$  in the network (A.204), we have

$$\frac{\partial \text{Loss}(\mathbf{X}^n, \mathbf{y}^n)}{\partial \mathbf{W}_{O_{(i,:)}}^\top} = -y_1^n \frac{\partial F_1(\mathbf{X}^n)}{\partial \mathbf{W}_{O_{1(i,:)}}} - y_2^n \frac{\partial F_2(\mathbf{X}^n)}{\partial \mathbf{W}_{O_{2(i,:)}}} \quad (\text{A.216})$$

where the derivation of  $\frac{\partial F_1(\mathbf{X}^n)}{\partial \mathbf{W}_{O_{1(i,:)}}}$  and  $\frac{\partial F_2(\mathbf{X}^n)}{\partial \mathbf{W}_{O_{2(i,:)}}}$  can be found in the analysis of binary classification above. For any  $i \in \mathcal{W}_2$ , suppose that we only care about  $\mathbf{p}_i, i = 1, 2, 3, 4$  in the gradient.

Following the proof of Claim 1 of Lemma A.3.1, if the data  $(\mathbf{X}^n, y^n) \in \mathcal{D}_2$ , we have

$$\begin{aligned} -\frac{\partial \text{Loss}(\mathbf{X}^n, \mathbf{y}^n)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^\top} &= y_1^n \frac{\partial F_1(\mathbf{X}^n)}{\partial \mathbf{W}_{O_{1(i,\cdot)}}} + y_2^n \frac{\partial F_2(\mathbf{X}^n)}{\partial \mathbf{W}_{O_{2(i,\cdot)}}} \\ &\approx \infty 1 \cdot \frac{1}{\sqrt{m}} \mathbf{p}_2 - 1 \cdot \left(-\frac{1}{\sqrt{m}}\right) \mathbf{p}_2 = \frac{2}{\sqrt{m}} \mathbf{p}_2 \end{aligned} \quad (\text{A.217})$$

$$(\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} - \mathbf{W}_{O_{(i,\cdot)}}^{(t)}) \mathbf{p}_2 \propto \|\mathbf{p}_2\|^2 > 0 \quad (\text{A.218})$$

if  $(\mathbf{X}^n, y^n) \in \mathcal{D}_1$ , we have

$$-\frac{\partial \text{Loss}(\mathbf{X}^n, \mathbf{y}^n)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^\top} \approx \infty 1 \cdot \frac{1}{\sqrt{m}} \mathbf{p}_1 + 1 \cdot \left(-\frac{1}{\sqrt{m}}\right) \mathbf{p}_1 = 0 \quad (\text{A.219})$$

$$(\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} - \mathbf{W}_{O_{(i,\cdot)}}^{(t)}) \mathbf{p}_1 \approx 0 \quad (\text{A.220})$$

if  $(\mathbf{X}^n, y^n) \in \mathcal{D}_3$ , we have

$$-\frac{\partial \text{Loss}(\mathbf{X}^n, \mathbf{y}^n)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^\top} \approx \infty -1 \cdot \frac{1}{\sqrt{m}} \mathbf{p}_3 + 1 \cdot \left(-\frac{1}{\sqrt{m}}\right) \mathbf{p}_3 = -\frac{2}{\sqrt{m}} \mathbf{p}_3 \quad (\text{A.221})$$

$$(\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} - \mathbf{W}_{O_{(i,\cdot)}}^{(t)}) \mathbf{p}_3 \leq 0 \quad (\text{A.222})$$

if  $(\mathbf{X}^n, y^n) \in \mathcal{D}_4$ , we have

$$-\frac{\partial \text{Loss}(\mathbf{X}^n, \mathbf{y}^n)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^\top} \approx \infty -1 \cdot \frac{1}{\sqrt{m}} \mathbf{p}_4 - 1 \cdot \left(-\frac{1}{\sqrt{m}}\right) \mathbf{p}_4 = 0 \quad (\text{A.223})$$

$$(\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} - \mathbf{W}_{O_{(i,\cdot)}}^{(t)}) \mathbf{p}_4 \approx 0 \quad (\text{A.224})$$

By the algorithm,  $\mathbf{W}_{O_{(i,\cdot)}}$  will update along the direction of  $\mathbf{p}_2$  for  $i \in \mathcal{W}_2$ . We can analyze  $\mathbf{W}_V$ ,  $\mathbf{W}_K$  and  $\mathbf{W}_Q$  similarly.

### A.6.2 Extension to A More General Data Model

We generalize the patterns from vectors to sets of vectors. Consider that there are  $M$  ( $2 < M < m_a, m_b$ ) distinct sets  $\{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_M\}$  where  $\mathcal{M}_l = \{\boldsymbol{\mu}_{l,1}, \boldsymbol{\mu}_{l,2}, \dots, \boldsymbol{\mu}_{l,l_m}\}$ ,  $l_m \geq 1$ .  $\mathcal{M}_1, \mathcal{M}_2$  denote sets of discriminative patterns for the binary labels, and  $\mathcal{M}_3, \dots, \mathcal{M}_M$

are sets of non-discriminative patterns.

$$\kappa = \min_{1 \leq i \neq j \leq M, 1 \leq a \leq i_m, 1 \leq b \leq j_m} \|\boldsymbol{\mu}_{i,a} - \boldsymbol{\mu}_{j,b}\| > 0 \quad (\text{A.225})$$

is the minimum distance between patterns of different sets. Each token  $\mathbf{x}_l^n$  of  $\mathbf{X}^n$  is a noisy version of one pattern, i.e.,

$$\min_{j \in [M], b \in [j_m]} \|\mathbf{x}_l^n - \boldsymbol{\mu}_{j,b}\| \leq \tau \quad (\text{A.226})$$

Define that for  $l, s$  corresponding to  $b_1, b_2$ , respectively,

$$\min_{j \in [M], b_1, b_2 \in [j_m]} \|\mathbf{x}_l^n - \mathbf{x}_s^n\| \leq \Delta, \quad (\text{A.227})$$

we have  $2\tau + \Delta < \kappa$ .

To simplify our theoretical analysis, one can similarly rescale all tokens a little bit like in Assumption A.2.4 such that tokens corresponding to patterns in the same pattern set has an inner product larger than 1, while tokens corresponding to patterns from different pattern sets has an inner product smaller than  $\lambda < 1$ .

Assumption 2.3.1 can be modified such that

$$\|\mathbf{W}_V^{(0)} \boldsymbol{\mu}_{j,b} - \mathbf{p}_{j,b}\| \leq \sigma \quad (\text{A.228})$$

$$\|\mathbf{W}_K^{(0)} \boldsymbol{\mu}_{j,b} - \mathbf{q}_{j,b}\| \leq \delta \quad (\text{A.229})$$

$$\|\mathbf{W}_Q^{(0)} \boldsymbol{\mu}_{j,b} - \mathbf{r}_{j,b}\| \leq \delta \quad (\text{A.230})$$

where  $\mathbf{p}_{j,b} \perp \mathbf{p}_{i,a}$  for any  $i, j \in [M]$ ,  $b \in [j_m]$ , and  $a \in [i_m]$ . Likewise,  $\mathbf{q}_{j,b} \perp \mathbf{q}_{i,a}$  for any  $i, j \in [M]$ ,  $b \in [j_m]$ , and  $a \in [i_m]$ .  $\mathbf{r}_{j,b} \perp \mathbf{r}_{i,a}$  for any  $i, j \in [M]$ ,  $b \in [j_m]$ , and  $a \in [i_m]$ .

Therefore, we make sure that the initial query, key and value features from different sets of patterns are still close to be orthogonal to each other. Then, we can follow our main proof idea. To be more specific, for label-relevant tokens  $\mathbf{x}_l^n$ , by computing (A.110) and (A.146),  $\mathbf{W}_K^{(t)} \mathbf{x}_l^n, \mathbf{W}_Q^{(t)} \mathbf{x}_l^n$  will grow in the direction of a fixed linear combination of  $\mathbf{q}_{l,1}, \dots, \mathbf{q}_{l,l_m}$ , and  $\mathbf{r}_{l,1}, \dots, \mathbf{r}_{l,l_m}$ . The coefficient of the linear combination is a function of fractions of different pattern vectors  $\boldsymbol{\mu}_{l,b}$  in  $\mathcal{M}_l$ . One can still derive a sparse attention map with weights of non-discriminative patterns decreasing to be close to zero during the training.

### A.6.3 Extension to Multi-Head Networks

Suppose there are  $H$  heads in total. The network is modified to

$$\begin{aligned} F(\mathbf{X}^n) &= \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \mathbf{a}_{(l)}^\top \text{Relu}(\mathbf{W}_O \left\| \sum_{h=1}^H \sum_{s \in \mathcal{S}^n} \mathbf{W}_{V_h} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_{K_h}^\top \mathbf{W}_{Q_h} \mathbf{x}_s^n) \right.) \\ &= \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \mathbf{a}_{(l)}^\top \text{Relu}\left(\sum_{h=1}^H \mathbf{W}_{O_h} \sum_{s \in \mathcal{S}^n} \mathbf{W}_{V_h} \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_{K_h}^\top \mathbf{W}_{Q_h} \mathbf{x}_s^n)\right) \end{aligned} \quad (\text{A.231})$$

where  $\mathbf{W}_{V_h} \in \mathbb{R}^{m_a \times d}$ ,  $\mathbf{W}_O = (\mathbf{W}_{O_1}, \mathbf{W}_{O_2}, \dots, \mathbf{W}_{O_H}) \in \mathbb{R}^{m \times H m_a}$ , and  $\mathbf{W}_{O_h} \in \mathbb{R}^{m \times m_a}$  for  $h \in [H]$ .

One can make similar assumptions for  $\mathbf{W}_{Q_h}^{(0)}$ ,  $\mathbf{W}_{K_h}^{(0)}$ , and  $\mathbf{W}_{V_h}^{(0)}$  as in Assumption 2.3.1. Note that  $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M\}$  needs to be changed into  $\{\mathbf{p}_{h_1}, \mathbf{p}_{h_2}, \dots, \mathbf{p}_{h_M}\}$ , and the set  $\{\mathbf{p}_{h_1}, \mathbf{p}_{h_2}, \dots, \mathbf{p}_{h_M}\}$  can vary for different  $h \in [H]$ . It is also the same for  $\{\mathbf{q}_{h_1}, \mathbf{q}_{h_2}, \dots, \mathbf{q}_{h_M}\}$  and  $\{\mathbf{r}_{h_1}, \mathbf{r}_{h_2}, \dots, \mathbf{r}_{h_M}\}$  for different  $h \in [H]$ .

Based on the modified assumption with  $H$  heads, the backbone of the proof remains the same. Lucky neurons in  $\mathbf{W}_O$  tend to learn  $(\mathbf{p}_{1_1}^\top, \mathbf{p}_{2_1}^\top, \dots, \mathbf{p}_{H_1}^\top)^\top$  and  $(\mathbf{p}_{1_2}^\top, \mathbf{p}_{2_2}^\top, \dots, \mathbf{p}_{H_2}^\top)^\top$ . Hence, the properties of the Relu activation are almost the same as the single-head case because luck neurons are still activated by either of two label-relevant patterns with a high probability. In fact, one can expect a more stable training process by multiple heads due to a more stable Relu gate for lucky neurons.

### A.6.4 Extension to Skip Connections and Normalization

Consider a basic case where a skip connection is added after the self-attention layer. Let  $m_a = d$ . The network is changed into

$$F(\mathbf{X}^n) = \frac{1}{|\mathcal{S}^n|} \sum_{l \in \mathcal{S}^n} \mathbf{a}_{(l)}^\top \text{Relu}(\mathbf{W}_O \left( \sum_{s \in \mathcal{S}^n} \mathbf{W}_V \mathbf{x}_s^n \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_s^n) + \mathbf{x}_l^n \right)) \quad (\text{A.232})$$

The assumption of  $\mathbf{W}_V^{(0)}$  in Assumption 2.3.1 should be changed into

$$\|(\mathbf{W}_V^{(0)} + \mathbf{I})\boldsymbol{\mu}_j - \mathbf{p}_j\| \leq \sigma, \quad (\text{A.233})$$

while the assumption of  $\mathbf{W}_Q^{(0)}$  and  $\mathbf{W}_K^{(0)}$  remain the same.

One can easily verify that the gradients of  $\mathbf{W}_K$ ,  $\mathbf{W}_Q$ , and  $\mathbf{W}_V$  for (A.232) are almost the

same as those for (2.1) except for the Relu gate. The major differences come from the gradient of  $\mathbf{W}_O$ , which also helps to determine the Relu gate. One needs to redefine

$$\begin{aligned}\mathbf{V}_l^n(t) &= \mathbf{W}_V^{(t)} \mathbf{X}^n \text{softmax}(\mathbf{X}^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) + \mathbf{x}_l^n \\ &= \sum_{s \in \mathcal{S}_1} \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{p}_1 + \mathbf{z}(t) + \sum_{j \neq 1} W_j^n(t)(\mathbf{x}_j^n + \mathbf{x}_l^n) \\ &\quad - \eta \left( \sum_{i \in \mathcal{W}_{l,n}(0)} V_i(t) \mathbf{W}_{O_{(i,:)}}^{(t)\top} + \sum_{i \notin \mathcal{W}_{l,n}(0)} V_i(t) \lambda \mathbf{W}_{O_{(i,:)}}^{(t)\top} \right)\end{aligned}\tag{A.234}$$

for  $l \in \mathcal{S}_1^n$ . The inner product between the lucky neuron and the term  $\sum_{j \neq 1} W_j^n(t)(\mathbf{x}_j^n + \mathbf{x}_l^n)$  can still be upper bounded by the inner product between the lucky neuron and the term  $\sum_{s \in \mathcal{S}_1} \text{softmax}(\mathbf{x}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l^n) \mathbf{p}_1$  given good initialization of  $\mathbf{W}_K$  and  $\mathbf{W}_Q$ . Therefore, (A.232) can be analyzed following our proof techniques.

For layer normalization, one usually use that approach to normalize each data. It is consistent with our normalization of  $\mathbf{x}_l^n$ , which plays an important role in our proof. By normalization, the training process becomes more stable because of the unified norm of all tokens.

## APPENDIX B

## APPENDIX OF CHAPTER 3

The appendix contains five sections. We add some extra experiments in Section B.1. In Section B.2, we introduce some definitions and assumptions in accordance with the main paper for ease of proof. Section B.3 first lists some key lemmas and then provides the proof of Theorem 3.4.4, Theorem 3.4.6, Theorem 3.4.8, Lemma 3.4.5, and Lemma 3.4.7. Section B.4 shows the proof of lemmas of this chapter. We finally add the extension of our analysis and other discussions in Section B.5.

We first briefly introduce some additional related works here on theoretical learning and generalization of neural networks without considering structured data. Some works [71], [72], [76], [260], [25] 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 [77], [58], [78], [80], [81], [83], [22] considers strongly overparameterized networks to linearize the neural network around the initialization. The generalization performance is independent of the feature distribution.

### B.1 Additional Experiments

#### B.1.1 Verifying Assumptions Made on the Graph Data Model

For the assumption on the graph data model, we conduct several experiments to verify this assumption on the real-world dataset Cora, PubMed, Actor, and PascalVOC-SP-1G.

**Existence of discriminative nodes.** We first compute the eigenvalue of the covariance matrix of the feature matrix of data of all classes in Figures B.1, B.2, B.3, and B.4. One can observe that the feature matrix is almost low-rank, which indicates that node features from the same class can be represented by a few eigenvectors. Therefore, for each class, we select the top three eigenvectors and compute the 3-dimensional representations of each node feature with these three eigenvectors. Then, we select all nodes with features that are less than  $\pi/4$  angle away from the mean of 3-dimensional representations as discriminative nodes. Non-discriminative nodes are the remaining nodes of each class. Tables B.1, B.2, B.3 show

---

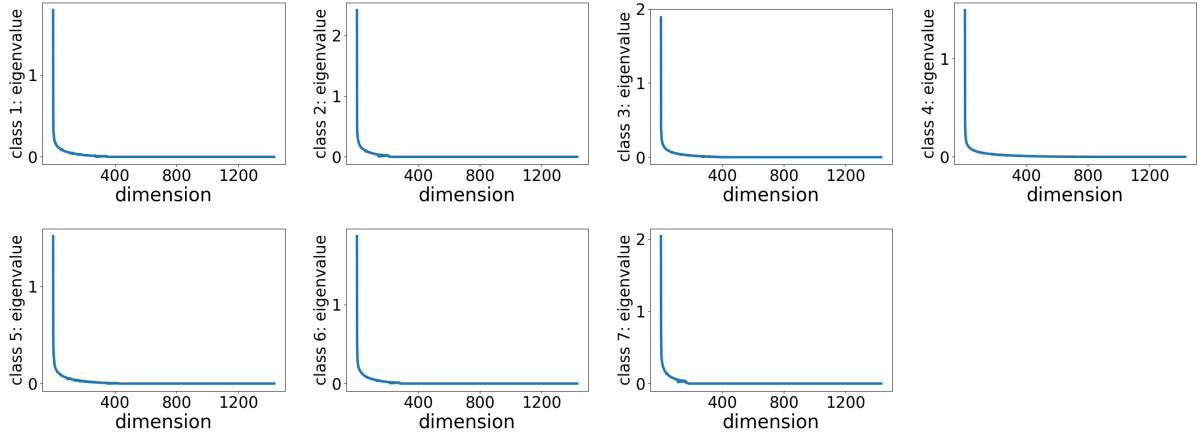
Portions of this appendix previously appeared as: H. Li, M. Wang, T. Ma, S. Liu, Z. Zhang, and P.-Y. Chen, “What improves the generalization of graph transformer? A theoretical dive into self-attention and positional encoding,” in *Proc. Int. Conf. Mach. Learn.*, Jul. 2024, pp. 28784–28829.

the fraction of discriminative nodes in each class. One can see a large fraction of the node features in each class is close to its top three eigenvectors.

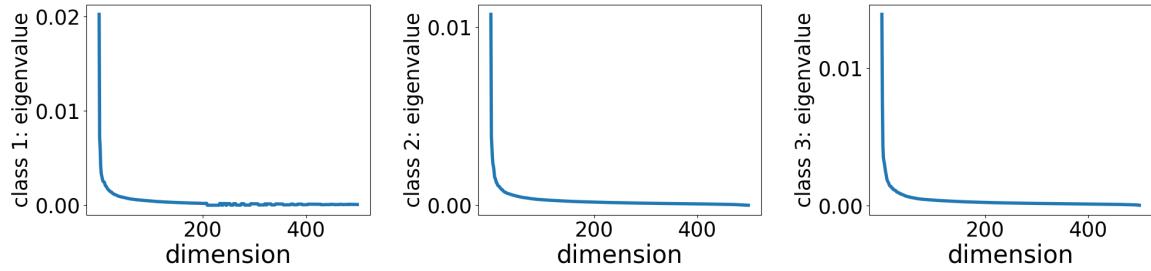
**The core distance (Assumption 3.4.2).** We further probe the core distance of each dataset by computing the fraction of nodes of which the label is aligned with the majority vote of the discriminative nodes in the distance- $z$  neighborhood. To extend the definition from binary classification in our formulation to multi-classification tasks, we use the average number of confusion nodes per class in the distance- $z$  neighborhood as  $|\mathcal{D}_\#^n \cap \mathcal{N}_z^n|$ , the number of confusion nodes in the distance- $z$  neighborhood of node  $n$ . Figure B.5 shows the value of a normalized  $\bar{\Delta}(z)$  for  $z = 1, 2, \dots, 12$ , where  $\bar{\Delta}(z)$  is divided by  $|\mathcal{N}_z^n|$  to control the gap of different numbers of nodes in different neighborhoods. The empirical result indicates that (1) homophilous graphs Cora and PubMed have a decreasing value of the normalized  $\bar{\Delta}(z)$  as  $z$  increases. The gap between the largest and the smallest normalized  $\bar{\Delta}(z)$  is large. This implies the core distance is 1 for Cora and PubMed and is aligned with the PE-based sampling performance of PubMed in Figure 3.8. (2) the heterophilous graph Actor has the largest normalized  $\bar{\Delta}(z)$  at  $z = 1$ , but the difference from other  $z$  is very small. This is consistent with the result in Figure 3.8 where the PE-based sampling has a close performance of less than 0.5% across  $z$ . (3) the long-range graph PascalVOC-SP-1G has the normalized  $\bar{\Delta}(z)$  when  $z = 1$ , but the value when  $z = 12$  is also remarkable. This corresponds to Figure 3.8 where the testing performance of PascalVOC-SP-1G is the highest when  $z = 1$  or  $z = 12$ .

Table B.5 shows that the fractions of nodes satisfying  $\Delta_n(z_m) > 0$  are all greater than 86% for all the four graph datasets. This fraction is especially larger than 95% in Cora, PubMed, and PascalVOC-SP-1G, which indicates a very small  $\epsilon_0 < 0.05$ . A slightly larger  $\epsilon_0 \approx 0.14$  for Actor is consistent with the challenge in training it with the state-of-the-art performance around 42% ([261]), which is consistent with the generalization bound scaling by  $\epsilon_0$  in Theorem 3.4.4.

We then verify the balanced dataset assumption and show a difference of no more than  $O(\sqrt{N})$  could be achieved in practical datasets. Table B.8 shows that for Cora and Actor, this condition holds since the largest gap between the average number of nodes and the number of any class of nodes is smaller than  $O(\sqrt{N}) = 10\sqrt{N}$ .



**Figure B.1:** Eigenvalues of the covariance matrix of the feature matrix of all classes of Cora.



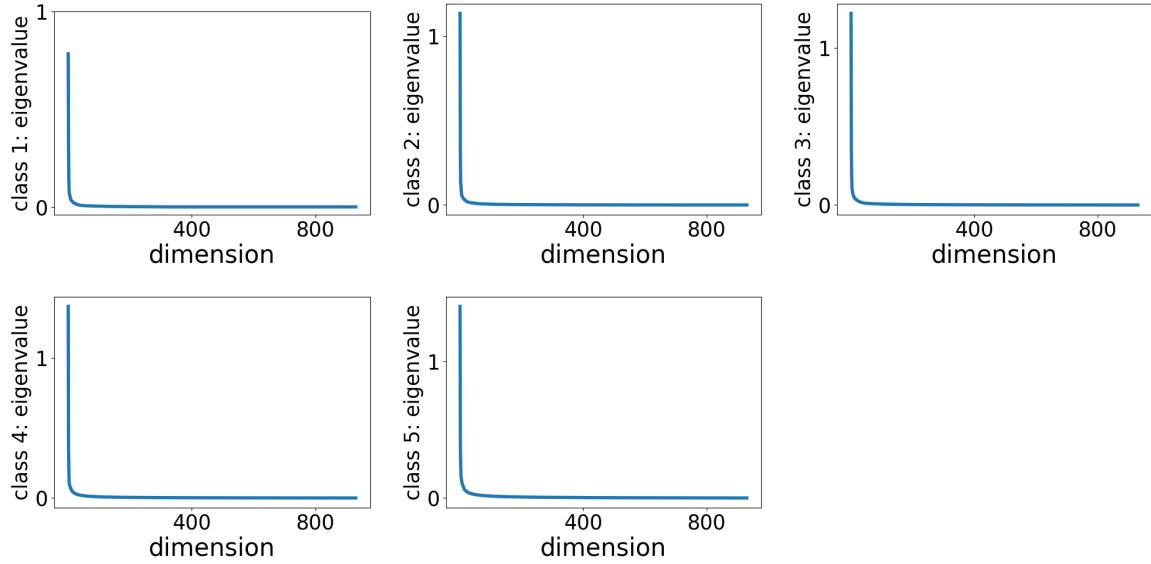
**Figure B.2:** Eigenvalues of the covariance matrix of the feature matrix of all classes of PubMed.

### B.1.2 Experiments on Synthetic Dataset

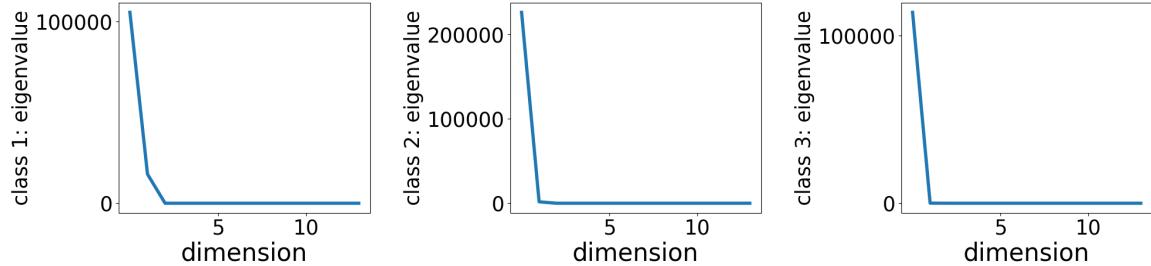
This section compares the required number of iterations for Graph Transformer and GCN by their orders in  $\gamma_d$ . The experiment setup follows Section 3.5.1. We set the number of known labels to be 800. For Graph Transformer,  $\epsilon_S = 0.05$ . For GCN,  $\epsilon_S = 0.2$ . Figure B.6 (a) shows that the required number of iterations is independent of  $\gamma_d$ . In contrast, Figure B.6 (b), which is exactly Figure 3.7 indicates the number of iterations is linear in  $\gamma_d^{-2}$ .

### B.1.3 Experiments on Real-world Datasets

We first add an introduction to the dataset PascalVOC-SP-1G we evaluate. This belongs to the Long Range Graph Benchmark, PascalVOC-SP [126], which is a computer vision dataset for node classification containing 11,355 graphs, 5,443,545 nodes, and 30,777,444 edges in total. Since this dataset is large, we pick the 2nd graph from the whole dataset and name this graph PascalVOC-SP-1G, which contains 479 nodes and 2,718 edges for node



**Figure B.3: Eigenvalues of the covariance matrix of the feature matrix of all classes of Actor.**



**Figure B.4: Eigenvalues of the covariance matrix of the feature matrix of all classes of PascalVOC-SP-1G.**

classification. The dimension of the node feature is 14. The number of classes is 3. Note that the size of the graph is not small compared with WebKB datasets [262], including Cornell, Texas, and Wisconsin, which contain 183, 183, and 251 nodes in each dataset, respectively.

Meanwhile, to verify the scalability of our conclusion, we conduct the experiments on the large-scale graph dataset Ogbn-Arxiv [127], which is a citation network with for node classification. The detailed statistics of these four datasets can be found in Table B.6.

We show the results of the Ogbn-Arxiv in Figure B.7 and B.8, where the dimension of  $\mathbf{b}^{(T)}$  is set to be 5. We still plot  $b_z^{(T)}$  with blue-circled lines for these datasets. Red dashed curves denote the test accuracy of the models learned with nodes all sampled from the distance- $z$  neighborhood for  $z \in \{1, 2, \dots, 5\}$ . The result of Ogbn-Arxiv shows a large  $b_z^{(T)}$

**Table B.1:** The fraction of discriminative nodes in each class of Cora.

class 1	class 2	class 3	class 4	class 5	class 6	class 7
82.05%	88.02%	82.54%	78.12%	78.17%	83.56%	76.11%

**Table B.2:** The fraction of discriminative nodes in each class of PubMed.

class 1	class 2	class 3
82.18%	93.34%	80.48%

when  $z$  is around 1. One can also observe that the testing accuracy using only distance- $z$  nodes has a similar trend as  $b_z^{(T)}$  with the largest accuracy around  $z = 1$ . This is consistent with our conclusions on PubMed from Figure B.7 in Section 3.5.2 since Ogbn-Arxiv and PubMed are both citation networks that are homophilous.

Figure B.8 showcases that for Ogbn-Arxiv, GT with PE has a better performance than that without PE and GCN. The conclusion is consistent with Figure 3.9

## B.2 Preliminaries

We first formally state the Algorithm 4. The notations used in the Appendix is summarized in Table B.7.

The loss function of a single data is defined in the following.

$$\text{Loss}(\mathbf{x}_l, y_l) = \max\{1 - y_l \cdot F(\mathbf{x}_l), 0\}. \quad (\text{B.1})$$

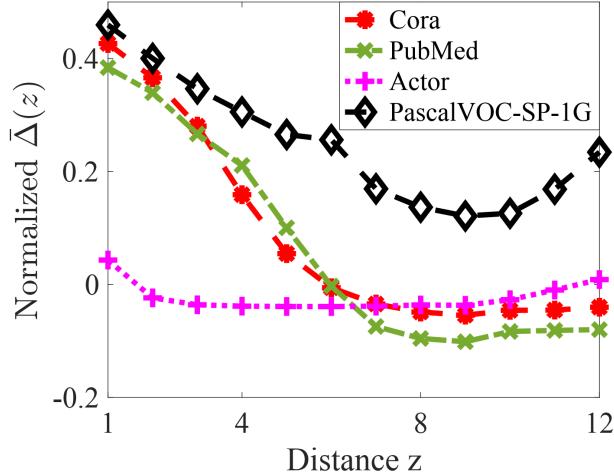
The formal algorithm is as follows. At each iteration  $t$ , the gradient is computed using

**Table B.3:** The fraction of discriminative nodes in each class of Actor.

class 1	class 2	class 3	class 4	class 5
42.09%	53.33%	57.85%	60.93%	64.79%

**Table B.4:** The fraction of discriminative nodes in each class of PascalVOC-SP-1G.

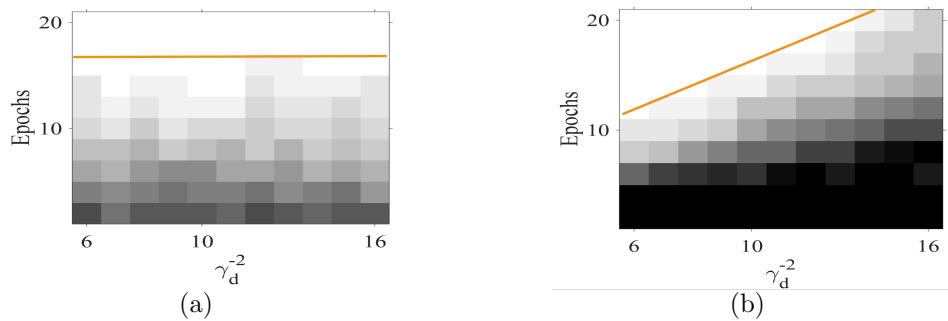
class 1	class 2	class 3
98.62%	100%	100%



**Figure B.5:** Normalized  $\bar{\Delta}(z)$  for Cora, PubMed, Actor, and PascalVOC-SP-1G.

**Table B.5:** The fraction of nodes satisfying  $\Delta_n(z_m) > 0$ .

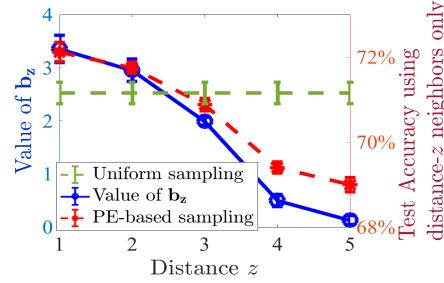
Cora	PubMed	Actor	PascalVOC-SP-1G
95.68%	95.50%	86.31%	98.54%



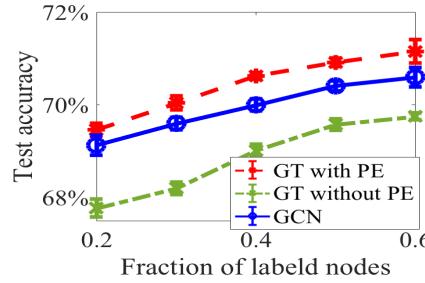
**Figure B.6:** The required number of iterations against  $\gamma_d^{-2}$  (a) Graph Transformer (b) GCN.

**Table B.6:** The statistics of datasets.

Dataset	#Nodes	#Edges	#Classes	#Features	Type
PubMed	19,717	44,324	3	500	Citation network
Actor	7,600	26,659	5	932	Actors in movies
PascalVOC-SP-1G	479	2,718	3	14	Computer vision
Ogbn-Arxiv	169,343	1,166,243	40	128	Citation network



**Figure B.7:** The values of entries of  $b$  and the test accuracy of PE-based sampling for Ogbn-Arxiv.



**Figure B.8:** Test accuracy of GT with/without PE and GCN when the number of label nodes varies for Ogbn-Arxiv.

**Table B.7: Summary of notations.**

$F(\mathbf{x}_l)$ , $\text{Loss}(\mathbf{x}_l, y_l)$	The network output for the node $\mathbf{x}_l$ and the loss function of a single node.
$\mathbf{p}_j(t)$ , $\mathbf{q}_j(t)$ , $\mathbf{r}_j(t)$	The features in value, key, and query vectors at the iteration $t$ for pattern $j$ , respectively. We have $\mathbf{p}_j(0) = \mathbf{p}_j$ , $\mathbf{q}_j(0) = \mathbf{q}_j$ , and $\mathbf{r}_j(0) = \mathbf{r}_j$ .
$\mathbf{z}_j(t)$ , $\mathbf{n}_j(t)$ , $\mathbf{o}_j(t)$	The error terms in the value, key, and query vectors of the $j$ -th node compared to their features at iteration $t$ .
$\mathcal{W}_l(0)$ , $\mathcal{U}_l(0)$	The set of lucky neurons for node $l$ .
$\phi_n(t)$ , $\nu_n(t)$ , $p_n(t)$ , $\lambda$	Approximate value of some attention weights at iteration $t$ . $\lambda$ is the threshold between inner products of tokens from the same pattern and different patterns.
$\mathcal{S}_j^{n,t}$	$\mathcal{S}_j^{n,t}$ is the set of sampled nodes of pattern $j$ at iteration $t$ to compute the aggregation of node $n$ .
$\delta_z$	The maximum number of nodes in distance- $z$ neighborhood for all nodes, which is no larger than $\sqrt{N}$ .

a mini-batch  $\mathcal{B}_t$  with  $|\mathcal{B}_t| = B$  and step size  $\eta$ . We first pre-train  $\mathbf{W}_O$  for  $T_0$  steps and then implement a full training with all parameters in  $\psi$  except  $\mathbf{a}$  for  $T(\geq T_0)$  steps. At iteration  $t$ , we uniformly sample a subset  $\mathcal{S}^{n,t}$  of nodes from the whole graph for aggregation of each

**Table B.8:** The fraction of discriminative nodes in each class of Actor.

	average # of each class	$10\sqrt{N}$	largest gap to the average
Cora	386.86	520.38	431.14
Actor	1520	871.78	667

node  $n$ . We set that  $\mathbf{W}_V^{(0)}$ ,  $\mathbf{W}_Q^{(0)}$ , and  $\mathbf{W}_K^{(0)}$  come from an initial model. Every entry of  $\mathbf{W}_O^{(0)}$  is generated from  $\mathcal{N}(0, \xi^2)$ . Every entry of  $\mathbf{a}^{(0)}$  is sampled from  $\{+1/\sqrt{m}, -1/\sqrt{m}\}$  with equal probability.  $\mathbf{b}^{(0)} = \mathbf{0}$ .  $\mathbf{a}$  does not update during the training.

---

**Algorithm 4** Training with Stochastic Gradient Descent (SGD)

---

- 1: **Input:** Training data  $\{(\mathbf{X}, y_n)\}_{n \in \mathcal{L}}$ , the step size  $\eta$ , the number of iterations  $T$ , batch size  $B$ .
- 2: **Initialization:** Each entry of  $\mathbf{W}_O^{(0)}$  and  $\mathbf{a}^{(0)}$  from  $\mathcal{N}(0, \xi^2)$  and Uniform( $\{+1/\sqrt{m}, -1/\sqrt{m}\}$ ), respectively.  $\mathbf{W}_V^{(0)}$ ,  $\mathbf{W}_K^{(0)}$  and  $\mathbf{W}_Q^{(0)}$  are initialized from a fair model.  $\mathbf{b}^{(0)} = \mathbf{0}$ .
- 3: **Node sampling:** At each iteration  $t$ , sample  $\mathcal{S}^{n,t}$  for each node  $n$  to replace  $\mathcal{T}^n$  in (3.1) when computing the  $\ell(\cdot)$  function in (3.3).
- 4: **Training by SGD:** For  $t = 0, 1, \dots, T-1$  and  $\mathbf{W}^{(t)} \in \{\mathbf{W}_O^{(t)}, \mathbf{W}_V^{(t)}, \mathbf{W}_K^{(t)}, \mathbf{W}_Q^{(t)}, \mathbf{b}^{(t)}\}$

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_t} \nabla_{\mathbf{W}^{(t)}} \ell(\mathbf{x}_n, y_n; \mathbf{a}^{(0)}, \mathbf{W}_O^{(t)}, \mathbf{W}_V^{(t)}, \mathbf{W}_K^{(t)}, \mathbf{W}_Q^{(t)}, \mathbf{b}^{(t)}) \quad (\text{B.2})$$

- 5: **Output:**  $\mathbf{W}_O^{(T)}, \mathbf{W}_V^{(T)}, \mathbf{W}_K^{(T)}, \mathbf{W}_Q^{(T)}, \mathbf{b}^{(T)}$ .
- 

**Assumption B.2.1.** [6] Define  $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M) \in \mathbb{R}^{m_a \times M}$ ,  $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_M) \in \mathbb{R}^{m_b \times M}$  and  $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_M) \in \mathbb{R}^{m_b \times M}$  as three feature matrices, where  $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_M\}$ ,  $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_M\}$  and  $\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_M\}$  are three sets of orthonormal bases. Define the noise terms  $\mathbf{z}_j(t)$ ,  $\mathbf{n}_j(t)$  and  $\mathbf{o}_j(t)$  with  $\|\mathbf{z}_j(0)\| \leq \sigma$  and  $\|\mathbf{n}_j(0)\|, \|\mathbf{o}_j(0)\| \leq \delta$  for  $j \in [L]$ .  $\mathbf{q}_1 = \mathbf{r}_1$ ,  $\mathbf{q}_2 = \mathbf{r}_2$ . Suppose  $\|\mathbf{W}_V^{(0)}\|, \|\mathbf{W}_K^{(0)}\|, \|\mathbf{W}_Q^{(0)}\| \leq 1$ ,  $\sigma < O(1/M)$  and  $\delta < 1/2$ . Then, for  $\mathbf{x}_l \in \mathcal{S}_j^n$

1.  $\mathbf{W}_V^{(0)} \mathbf{x}_l = \mathbf{p}_j + \mathbf{z}_j(0)$ .

2.  $\mathbf{W}_K^{(0)} \mathbf{x}_l = \mathbf{q}_j + \mathbf{n}_j(0)$ .

3.  $\mathbf{W}_Q^{(0)} \mathbf{x}_l = \mathbf{r}_j + \mathbf{o}_j(0)$ .

Assumption B.2.1 is a straightforward combination of Assumption 1 in [6] and the equation  $\min_{j \in [M]} \|\mathbf{x}_n - \boldsymbol{\mu}_j\| = 0, \forall n \in \mathcal{V}$  by applying the triangle inequality to bound the error terms for tokens. We then provide a condition which is equivalent to the equation  $\min_{j \in [M]} \|\mathbf{x}_n - \boldsymbol{\mu}_j\| = 0, \forall n \in \mathcal{V}$ , i.e., if nodes  $i$  and  $j$  correspond to the same pattern  $k \in [M]$ , i.e.,  $i \in \mathcal{D}_k$  and  $j \in \mathcal{D}_k$ , we have  $\mathbf{x}_i^\top \mathbf{x}_j \geq 1$ . If nodes  $i$  and  $j$  correspond to the different feature  $k, l \in [M], k \neq l$  i.e.,  $i \in \mathcal{D}_k$  and  $j \in \mathcal{D}_l, k \neq l$ , we have  $\mathbf{x}_i^\top \mathbf{x}_j \leq \lambda < 1$ . Here, we scale up all nodes a bit to make the threshold of linear separability 1 for the simplicity of presentation.

**Definition B.2.2.** Define

$$\mathbf{V}_n(t) = \mathbf{W}_V^{(t)} \sum_{s \in \mathcal{T}^n} \mathbf{x}_n \text{softmax}_n(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b}^{(t)}). \quad (\text{B.3})$$

for the node  $n$ . Define  $\mathcal{W}_n(0), \mathcal{U}_n(0)$  as the sets of lucky neurons such that

$$\mathcal{W}_n(0) = \{i : \mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{V}_n(0) > 0, l \in \mathcal{S}_1^{n,t}\}, \quad (\text{B.4})$$

$$\mathcal{U}_n(0) = \{i : \mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{V}_n(0) > 0, l \in \mathcal{S}_2^{n,t}\}. \quad (\text{B.5})$$

**Definition B.2.3.** When  $n \in \mathcal{D}_1 \cup \mathcal{D}_2$ , we have

1.  $\phi_n(t) = (\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,t}| e^{\|\mathbf{q}_1(t)\|^2 + \sigma \|\mathbf{q}_1(t)\| + b_z^{(t)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cap \mathcal{S}_*^{n,t}) - \mathcal{S}_1^{n,t}| e^{b_z^{(t)}})^{-1}$ .
2.  $\nu_n(t) = (\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,t}| e^{\|\mathbf{q}_1(t)\|^2 - \sigma \|\mathbf{q}_1(t)\| + b_z^{(t)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cap \mathcal{S}_*^{n,t}) - \mathcal{S}_1^{n,t}| e^{b_z^{(t)}})^{-1}$ .
3.  $p_n(t) = \sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,t}| e^{\|\mathbf{q}_1(t)\|^2 - \sigma \|\mathbf{q}_1(t)\| + b_z^{(t)}} \nu_n(t)$ .

When  $n \notin \mathcal{D}_1 \cup \mathcal{D}_2$ , we have

1.  $\phi_n(t) = (\sum_{z \in \mathcal{Z}} (|\mathcal{N}_z^n \cap \mathcal{S}_*^{n,t}| + |\mathcal{N}_z^n \cap \mathcal{S}_\#^{n,t}|) e^{\|\mathbf{q}_1(t)\|^2 + \sigma \|\mathbf{q}_1(t)\| + b_z^{(t)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cup \mathcal{S}_*^{n,t}) / (\mathcal{S}_1^{n,t} \cup \mathcal{S}_2^{n,t})| e^{b_z^{(t)}})^{-1}$ .
2.  $\nu_n(t) = (\sum_{z \in \mathcal{Z}} (|\mathcal{N}_z^n \cap \mathcal{S}_*^{n,t}| + |\mathcal{N}_z^n \cap \mathcal{S}_\#^{n,t}|) e^{\|\mathbf{q}_1(t)\|^2 - \sigma \|\mathbf{q}_1(t)\| + b_z^{(t)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cup \mathcal{S}_*^{n,t}) / (\mathcal{S}_1^{n,t} \cup \mathcal{S}_2^{n,t})| e^{b_z^{(t)}})^{-1}$ .
3.  $p_n(t) = \sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,t}| e^{\|\mathbf{q}_1(t)\|^2 - \sigma \|\mathbf{q}_1(t)\| + b_z^{(t)}} \nu_n(t)$ .

We then cite useful results of the concentration bounds on sub-gaussian variables.

**Definition B.2.4.** [258] 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 B.2.5.** ([258] Proposition 5.1, hoeffding's inequality) Let  $X_1, X_2, \dots, X_N$  be independent centered sub-gaussian random variables, and let  $K = \max_i \|X_i\|_{\psi_2}$ . Then for every  $\mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N$  and every  $t \geq 0$ , we have

$$\mathbb{P}\left\{\left|\sum_{i=1}^N a_i X_i\right| \geq t\right\} \leq e \cdot \exp\left(-\frac{ct^2}{K^2 \|\mathbf{a}\|^2}\right). \quad (\text{B.6})$$

where  $c > 0$  is an absolute constant.

### B.3 Key Lemmas and Proof of the Main Theorems

We first present our key lemmas, followed by the proof of the main theorems.

For  $l \in \mathcal{S}_1^{n,t}$  for the data with  $y_n = 1$ , define

$$\mathbf{V}_l(t) = \sum_{s \in \mathcal{S}_1^{n,t}} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_n(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b}^{(t)}). \quad (\text{B.7})$$

We later can show that

$$\begin{aligned} \mathbf{V}_l(t) &= \sum_{s \in \mathcal{S}_1^{n,t}} \text{softmax}_n(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b}^{(t)}) \mathbf{p}_1 + \mathbf{z}(t) + \sum_{j \neq 1} W_j(t) \mathbf{p}_j \\ &\quad - \eta \sum_{b=1}^t \left( \sum_{i \in \mathcal{W}_l(b)} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)\top} + \sum_{i \notin \mathcal{W}_l(b)} V_i(b) \lambda \mathbf{W}_{O_{(i,:)}}^{(b)\top} \right). \end{aligned} \quad (\text{B.8})$$

We have the following Lemmas:

**Lemma B.3.1.** For the lucky neuron  $i \in \mathcal{W}_l(0)$  and  $b \in [T]$ , we have that the major component of  $\mathbf{W}_{O_{(i,:)}}^{(t)}$  is in the direction of  $\mathbf{p}_1$ , i.e.,

$$\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{p}_1 \gtrsim \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2 (1 - 2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) + \xi, \quad (\text{B.9})$$

$$\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{p} \lesssim \frac{1}{\sqrt{B}} \mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{p}_1, \quad \text{for } \mathbf{p} \in \{\mathbf{p}_2, \mathbf{p}_3, \dots, \mathbf{p}_M\}, \quad (\text{B.10})$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\|^2 \geq \left(\frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) + \xi\right)^2, \quad (\text{B.11})$$

and for the noise  $\mathbf{z}_l(t)$ ,

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{z}_l(t)\| \leq \sigma \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\|. \quad (\text{B.12})$$

For  $i \in \mathcal{U}_l(0)$ , we also have equations as in (B.9) to (B.12), including

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p}_2 \gtrsim \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) + \xi, \quad (\text{B.13})$$

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p} \lesssim \frac{1}{\sqrt{B}} \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p}_1, \quad \text{for } \mathbf{p} \in \{\mathbf{p}_1, \mathbf{p}_3, \mathbf{p}_4, \dots, \mathbf{p}_M\}, \quad (\text{B.14})$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\|^2 \geq \left(\frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \eta t^2 \frac{\eta t^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) + \xi\right)^2. \quad (\text{B.15})$$

For the noise  $\mathbf{z}_l(t)$ ,

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{z}_l(t)\| \leq \sigma \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\|. \quad (\text{B.16})$$

For unlucky neurons  $i$  and  $j \in \mathcal{W}_l(0)$ ,  $k \in \mathcal{U}_l(0)$ ,  $p \in \mathcal{P}$ , we have

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p} \leq \frac{1}{\sqrt{B}} \min\{\mathbf{W}_{O_{(j,\cdot)}}^{(t)} \mathbf{p}_1, \mathbf{W}_{O_{(k,\cdot)}}^{(t)} \mathbf{p}_2\}, \quad (\text{B.17})$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{z}_l(t)\| \leq \sigma \|\mathbf{W}_{O_{(j,\cdot)}}^{(t)}\|, \quad (\text{B.18})$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\|^2 \leq \frac{1}{B} \min\{\|\mathbf{W}_{O_{(j,\cdot)}}^{(t)}\|^2, \|\mathbf{W}_{O_{(k,\cdot)}}^{(t)}\|^2\}. \quad (\text{B.19})$$

**Lemma B.3.2.** There exists  $K(t), Q(t) > 0$ ,  $t = 0, 1, \dots, T-1$  such that for  $r \in \mathcal{S}_*^{n,t}$ , if  $u_{(r,l)z_0} = 1$ , defining

$$\mathbf{q}_i(t) = \sqrt{\prod_{l=0}^{t-1} (1 + K(l))} \mathbf{q}_i, \quad (\text{B.20})$$

$$\mathbf{r}_i(t) = \sqrt{\prod_{l=0}^{t-1} (1 + Q(l))} \mathbf{r}_i, \quad (\text{B.21})$$

where  $i = 1, 2$ . Then, we have

$$\begin{aligned} & softmax_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t+1)} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t+1)}) \\ & \gtrsim \frac{e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - \delta\|\mathbf{q}_1(t)\| + b_{z_0}^{(t)}}}{\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| e^{(1+K(t))\|\mathbf{q}_1(T)\|^2 - \sigma\|\mathbf{q}_1(T)\| + b_z^{(T)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cap \mathcal{S}^{n,T}) - \mathcal{S}_1^{n,T}| e^{b_z^{(T)}}}. \end{aligned} \quad (\text{B.22})$$

Similarly, for  $r \notin \mathcal{S}_*^{l,t}$ , we have

$$\begin{aligned} & softmax_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t+1)^\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) \\ & \lesssim \frac{e^{b_{z_0}^{(t)}}}{\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| e^{(1+K(t))\|\mathbf{q}_1(T)\|^2 - \sigma\|\mathbf{q}_1(T)\| + b_z^{(T)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cap \mathcal{S}^{n,T}) - \mathcal{S}_1^{n,T}| e^{b_z^{(T)}}}. \end{aligned} \quad (\text{B.23})$$

**Lemma B.3.3.** During the training, we fix  $b_0^{(t)} = b_0^{(0)} = \Omega(1)$ . For  $z \geq 1$ ,

$$\begin{aligned} & b_{z_m}^{(t)} - b_z^{(t)} \\ & \gtrsim \eta \frac{1}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \eta \frac{(1-2\epsilon_0)^3}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m^2}{a^2} \left( \frac{\xi \eta t^2 m}{a^2} \right)^2 \|\mathbf{p}_1\|^2 \cdot \frac{\gamma_d}{2} \\ & \quad \cdot \left( \frac{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_{z_m}^l| - |\mathcal{S}_\#^{l,t} \cap \mathcal{N}_{z_m}^l|}{K|\mathcal{S}^{l,t}|} - \frac{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l| - |\mathcal{S}_\#^{l,t} \cap \mathcal{N}_z^l|}{K|\mathcal{S}^{l,t}|} \right). \end{aligned} \quad (\text{B.24})$$

$$\begin{aligned} b_z^{(t)} & \geq \eta \frac{1}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \eta \frac{(1-2\epsilon_0)^3}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m^2}{a^2} \left( \frac{\xi \eta t^2 m}{a^2} \right)^2 \|\mathbf{p}_1\|^2 \cdot \frac{\gamma_d}{2} \\ & \quad \cdot \frac{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l| - |\mathcal{S}_\#^{l,t} \cap \mathcal{N}_z^l|}{K|\mathcal{S}^{l,t}|}. \end{aligned} \quad (\text{B.25})$$

**Lemma B.3.4.** For the update of  $\mathbf{W}_V^{(t)}$ , there exists  $\lambda \leq \Theta(1)$  such that

$$\mathbf{W}_V^{(t)} \mathbf{x}_j = \mathbf{p}_1 - \eta \sum_{b=1}^t \left( \sum_{i \in \mathcal{W}_n(0)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)^\top} + \sum_{i \notin \mathcal{W}_n(0)} \lambda V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)^\top} \right) + \mathbf{z}_j(t), \quad j \in \mathcal{S}_1^{n,t}, \quad (\text{B.26})$$

$$\mathbf{W}_V^{(t)} \mathbf{x}_j^n = \mathbf{p}_2 - \eta \sum_{b=1}^t \left( \sum_{i \in \mathcal{U}(0)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)^\top} + \sum_{i \notin \mathcal{U}(0)} \lambda V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)^\top} \right) + \mathbf{z}_j(t), \quad j \in \mathcal{S}_2^{n,t}, \quad (\text{B.27})$$

$$\mathbf{W}_V^{(t+1)} \mathbf{x}_j^n = \mathbf{p}_l - \eta \sum_{b=1}^t \sum_{i=1}^m \lambda V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)^\top} + \mathbf{z}_j(t), \quad j \in \mathcal{S}^{n,t} \setminus (\mathcal{S}_1^{n,t} \cup \mathcal{S}_2^{n,t}), \quad (\text{B.28})$$

$$\|\mathbf{z}_j(t)\| \leq \sigma, \quad (\text{B.29})$$

with

$$W_l(t) \leq \nu_n(t)|\mathcal{S}_j^n|, \quad l \in \mathcal{S}_j^n, \quad (\text{B.30})$$

$$V_i(t) \lesssim \frac{1-2\epsilon_0}{2B} \sum_{n \in \mathcal{B}_{b+}} -\frac{1}{a} p_n(t), \quad i \in \mathcal{W}_l(0), \quad (\text{B.31})$$

$$V_i(t) \gtrsim \frac{1-2\epsilon_0}{2B} \sum_{n \in \mathcal{B}_{b-}} \frac{1}{a} p_n(t), \quad i \in \mathcal{U}_l(0), \quad (\text{B.32})$$

$$V_i(t) \geq -\frac{1}{\sqrt{Ba}}, \quad \text{if } i \text{ is an unlucky neuron.} \quad (\text{B.33})$$

**Lemma B.3.5.** [6] If the number of neurons  $m$  is larger enough such that

$$m \geq M^2 \log N, \quad (\text{B.34})$$

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

$$|\mathcal{W}_l(0)|, |\mathcal{U}_l(0)| \geq \Omega(m). \quad (\text{B.35})$$

**Lemma B.3.6.** Under the condition that  $m \gtrsim M^2 \log N$ , we have the following result.

For  $i \in \mathcal{W}_l(0)$  and  $l \in \mathcal{D}_1$ , we have

$$\mathbb{I}[\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_l(t)] = 1; \quad (\text{B.36})$$

For  $i \in \mathcal{U}_l(0)$  and  $l \in \mathcal{D}_2$ , we have

$$\mathbb{I}[\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_l(t)] = 1; \quad (\text{B.37})$$

#### Proof of Theorem 3.4.4:

Denote the set of neurons with positive  $a_i$  as  $\mathcal{K}_+$  and the set of neurons with negative  $a_i$  as  $\mathcal{K}_-$ . For  $y_n = 1$ , recall from (3.11) and Definition B.2.3, we have

$$\begin{aligned} F(\mathbf{x}_n) &= \sum_{i \in \mathcal{W}_n(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_n(t)) + \sum_{i \in \mathcal{K}_+/W_n(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_n(t)) \\ &\quad - \sum_{i \in \mathcal{K}_-} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_n(t)). \end{aligned} \quad (\text{B.38})$$

Therefore,

$$\begin{aligned}
& \sum_{i \in \mathcal{W}_n(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_n(t)) \\
&= \sum_{i \in \mathcal{W}_n(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_n(t)) + \sum_{i \in \mathcal{W}_n(t)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_n(t)) \\
&\gtrsim \frac{1}{a} \cdot \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \left( \sum_{s \in \mathcal{S}_1^{n,t}} \mathbf{p}_s \text{softmax}_n(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_n) + \mathbf{z}(t) + \sum_{l \neq s} W_l(u) \mathbf{p}_l \right. \\
&\quad \left. - \eta t \left( \sum_{j \in \mathcal{W}_n(0)} V_j(t) \mathbf{W}_{O_{(j,\cdot)}}^{(t)\top} + \sum_{j \notin \mathcal{W}_n(0)} V_j(t) \lambda \mathbf{W}_{O_{(j,\cdot)}}^{(t)\top} \right) \right) |\mathcal{W}_n(0)| + 0 \\
&\gtrsim \frac{m}{a} \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta t^2 m}{a^2} \left( \frac{1-2\epsilon_0}{4B} \sum_{n \in \mathcal{B}_b} p_n(b) - \sigma \right) p_n(t) + \eta m \frac{1-2\epsilon_0}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{1}{a} p_n(b) \right. \\
&\quad \left. \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2 (1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) \right)^2 \right), 
\end{aligned} \tag{B.39}$$

where the second step results from the formulation of  $\mathbf{V}_n(t)$  in (B.8) and the last step is by (B.128). Meanwhile, we have

$$\sum_{i \in \mathcal{K}_+ / \mathcal{W}_n(0)} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_n(t)) \geq 0. \tag{B.40}$$

To deal with the upper bound of the third term in (B.38), we have

$$\left| \sum_{i \in \mathcal{K}_-} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_n(t)) \right| \lesssim \sum_{i \in \mathcal{K}_+} \frac{1}{a} \text{Relu}(\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_n(t)). \tag{B.41}$$

Note that at the  $t$ -th iteration,

$$\begin{aligned}
& K(t) \\
&\gtrsim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} \left( \frac{1-2\epsilon_0}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 m}{a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} p_n(b) - \sigma \right) + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)}{a} \right. \\
&\quad \left. \cdot (1-\sigma) \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{(1-2\epsilon_0) \eta (t+1)^2}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) \right)^2 \right) \phi_n(t) (|\mathcal{S}^{l,t}| - |\mathcal{S}_1^{l,t}|) \|\mathbf{q}_1(t)\|^2 \\
&\gtrsim \frac{1}{e^{\|\mathbf{q}_1(t)\|^2 - \delta \|\mathbf{q}_1(t)\|}}.
\end{aligned} \tag{B.42}$$

Since that

$$\begin{aligned}\mathbf{q}_1(T) &\gtrsim (1 + \min_{l=0,1,\dots,T-1} \{K(l)\})^T \\ &\gtrsim (1 + \frac{1}{e^{\|\mathbf{q}_1(T)\|^2 - \delta\|\mathbf{q}_1(T)\|}})^T.\end{aligned}\tag{B.43}$$

To find the order-wise lower bound of  $\mathbf{q}_1(T)$ , we need to check the equation

$$\mathbf{q}_1(T) \lesssim (1 + \frac{1}{e^{\|\mathbf{q}_1(T)\|^2 - \delta\|\mathbf{q}_1(T)\|}})^T.\tag{B.44}$$

One can obtain

$$\Theta(\sqrt{\log T(1-\delta)}) = \mathbf{q}_1(T) \leq \Theta(T).\tag{B.45}$$

We require that

$$\begin{aligned}&\frac{m}{a} \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta T^2 m}{a^2} \left( \frac{1-2\epsilon_0}{4B} \sum_{n \in \mathcal{B}_b} p_n(b) - \sigma \right) p_n(T) + \eta m \frac{1-2\epsilon_0}{2B} \sum_{b=1}^T \sum_{n \in \mathcal{B}_{b+}} \frac{1}{a} p_n(b) \right. \\ &\quad \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta T^2 (1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(T))^2 \right) \\ &:= a_0 \eta^3 T^5 + a_1 \eta T^2, \\ &> 1,\end{aligned}\tag{B.46}$$

where the first step is by letting  $a = \sqrt{m}$  and  $m \gtrsim M^2 \log N$ . We replace  $p_n(b)$  with  $p_n(T)$  because when  $b$  achieves the level of  $T$ ,  $b^{o_1} p_n(b)^{o_2}$  is the same order as  $b^{o_1}$  for  $o_1, o_2 \geq 0$ . Thus,

$$\sum_{b=1}^T b^{o_1} p_n(b)^{o_2} \gtrsim T^{o_1+1} p_n(\Theta(1) \cdot T)^{o_2} \gtrsim T^{o_1+1} p_n(T)^{o_2}.\tag{B.47}$$

We also require

$$B \gtrsim \Theta(1).\tag{B.48}$$

Note that  $p_n(t)$  is dependent on other  $\sum_{z \in \mathcal{Z}} \delta_z$  nodes. Hence, we know that each  $p_n(T)$  is dependent on other  $1 + \sum_{z \in \mathcal{Z}} \delta_z^2$  variables of  $p_j(T)$  for  $j, n \in \mathcal{V}$ . It is easy to find that  $p_n(T)$  is a 1-sub-gaussian random variable because its absolute value is upper bounded by 1. By Lemma 7 in [74], we can obtain

$$\mathbb{E}_{\mathcal{D}}[e^{s(\sum_{n \in \mathcal{L}} p_n(T) - |\mathcal{L}| \mathbb{E}_{\mathcal{D}}[p_n(T)])}] \leq e^{|\mathcal{L}|(1 + \sum_{z \in \mathcal{Z}} \delta_z^2)s^2}.\tag{B.49}$$

When  $\eta T = \Theta(1)$ , we have  $|b_z(T)| = \Theta(1)$  and  $|b_z(T) - b_{z'}(T)| \leq \Theta(1)$ . Therefore, when  $n \in \mathcal{D}_1 \cup \mathcal{D}_2$ , we have

$$\begin{aligned} & p_n(T) \\ &= \frac{\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| e^{\|\mathbf{q}_1(T)\|^2 - \sigma \|\mathbf{q}_1(T)\| + b_z^{(T)}}}{\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| e^{\|\mathbf{q}_1(T)\|^2 - \sigma \|\mathbf{q}_1(T)\| + b_z^{(T)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cap \mathcal{S}^{n,T}) - \mathcal{S}_1^{n,T}| e^{b_z^{(T)}}} \\ &\geq 1 - \eta^C. \end{aligned} \quad (\text{B.50})$$

When  $n \notin \mathcal{D}_1 \cup \mathcal{D}_2$ , we have

$$\begin{aligned} & p_n(T) \\ &= \sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| e^{\|\mathbf{q}_1(T)\|^2 - \sigma \|\mathbf{q}_1(T)\| + b_z^{(T)}} \left( \sum_{z \in \mathcal{Z}} (|\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| + |\mathcal{N}_z^n \cap \mathcal{S}_{\#}^{n,T}|) \right. \\ &\quad \cdot e^{\|\mathbf{q}_1(T)\|^2 - \sigma \|\mathbf{q}_1(T)\| + b_z^{(T)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cup \mathcal{S}^{n,T}) / (\mathcal{S}_1^{n,T} \cup \mathcal{S}_2^{n,T})| e^{b_z^{(T)}})^{-1} \\ &\geq \sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| (T(1 - \delta))^C e^{b_z^{(T)} - b_{z_m}^{(T)}} \left( \sum_{z \in \mathcal{Z}} (|\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| + |\mathcal{N}_z^n \cap \mathcal{S}_{\#}^{n,T}|) \right. \\ &\quad \cdot (T(1 - \delta))^C e^{b_z^{(T)} - b_{z_m}^{(T)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cup \mathcal{S}^{n,T}) / (\mathcal{S}_1^{n,T} \cup \mathcal{S}_2^{n,T})| e^{b_z^{(T)} - b_{z_m}^{(T)}})^{-1} \\ &\geq \frac{|\mathcal{N}_{z_m}^n \cap \mathcal{S}_{\#}^{n,T}|}{(|\mathcal{N}_{z_m}^n \cap \mathcal{S}_*^{n,T}| + |\mathcal{N}_{z_m}^n \cap \mathcal{S}_{\#}^{n,T}|)} - (T(1 - \delta))^{-C} e^{-b_z(T)}. \end{aligned} \quad (\text{B.51})$$

$$\begin{aligned} & \zeta = \mathbb{E}_{\mathcal{D}}[p_n(T)] \\ &\geq (1 - \gamma_d) \cdot \mathbb{E}_{\mathcal{D}} \left[ \frac{\sum_{z \in \mathcal{Z}} (T(1 - \delta))^C |\mathcal{S}_*^{n,T} \cap \mathcal{N}_z^n| e^{b_z(T)}}{\sum_{z \in \mathcal{Z}} (T(1 - \delta))^C |(\mathcal{S}_1^{n,T} \cup \mathcal{S}_2^{n,T}) \cap \mathcal{N}_z^n| e^{b_z(T)} + \Theta(1)} \right] \\ &\quad + \gamma_d \cdot \frac{(T(1 - \delta))^C}{(T(1 - \delta))^C + \Theta(1)} \\ &\geq (1 - \gamma_d)(1 - \epsilon_{\mathcal{S}} - (T(1 - \delta))^{-C} e^{-b_z(T)}) + \gamma_d(1 - \eta^C) \\ &\gtrsim 1 - \epsilon_{\mathcal{S}}(1 - \gamma_d) - \eta^C. \end{aligned} \quad (\text{B.52})$$

Hence, define

$$p_n(T) \geq p'_n(T) := \begin{cases} 1, & \text{if } n \in \mathcal{D}_1 \cup \mathcal{D}_2 \\ \frac{|\mathcal{N}_{z_m}^n \cap \mathcal{S}_{\#}^{n,T}|}{(|\mathcal{N}_{z_m}^n \cap \mathcal{S}_*^{n,T}| + |\mathcal{N}_{z_m}^n \cap \mathcal{S}_{\#}^{n,T}|)}, & \text{if } n \notin \mathcal{D}_1 \cup \mathcal{D}_2. \end{cases} \quad (\text{B.53})$$

Therefore,

$$\begin{aligned} & |\mathbb{E}_{t \geq 0, n \in \mathcal{V}}[p'_n(T)] - 1| \\ & \leq (1 - \gamma_d) \mathbb{E}_{n \notin (\mathcal{D}_1 \cup \mathcal{D}_2)} \left[ \frac{|\mathcal{N}_{z_m}^n \cap \mathcal{S}_\#^{n,T}|}{(|\mathcal{N}_{z_m}^n \cap \mathcal{S}_*^{n,T}| + |\mathcal{N}_{z_m}^n \cap \mathcal{S}_\#^{n,T}|)} \right] = (1 - \gamma_d) \epsilon_{\mathcal{S}}. \end{aligned} \quad (\text{B.54})$$

We can also derive

$$\mathbb{E}_{n \in \mathcal{L}} \left[ (1 - p'_n(T)^2) \right] \leq 2 \mathbb{E}_{\mathcal{D}} \left[ 1 - p'_n(T) \right] \leq 2(1 - \gamma_d) \epsilon_{\mathcal{S}}, \quad (\text{B.55})$$

where the first inequality is by  $1 - (p'_n(T))^2 \leq (1 - p'_n(T))(1 + p'_n(T)) \leq 2(1 - p'_n(T))$ . We have

$$\begin{aligned} & \left| \frac{1}{|\mathcal{L}|} \sum_{n \in \mathcal{L}} (p'_n(T) - \sigma) p'_n(T) - 1 \right| \\ & \leq \left| \frac{1}{|\mathcal{L}|} \sum_{n \in \mathcal{L}} (p'_n(T) - \sigma) p'_n(T) - \mathbb{E}_{n \in \mathcal{L}}[(p'_n(T) - \sigma) p'_n(T)] \right| \\ & \quad + \mathbb{E}_{n \in \mathcal{L}}[|1 - p'_n(T)^2|] + \mathbb{E}_{n \in \mathcal{L}}[\sigma p'_n(T)] \\ & \lesssim \sqrt{\frac{(1 + \delta_{z_m}^2) \cdot \log N}{|\mathcal{L}|}} + 2(1 - \gamma_d) \epsilon_{\mathcal{S}} + \sigma, \end{aligned} \quad (\text{B.56})$$

$$\left| \frac{1}{|\mathcal{L}|} \sum_{n \in \mathcal{L}} p_n(T)^2 - 1 \right| \lesssim \sqrt{\frac{(1 + \delta_{z_m}^2) \cdot \log N}{|\mathcal{L}|}} + 2(1 - \gamma_d) \epsilon_{\mathcal{S}}, \quad (\text{B.57})$$

$$\left| \frac{1}{|\mathcal{L}|} \sum_{n \in \mathcal{L}} p_n(T) - 1 \right| \lesssim \sqrt{\frac{(1 + \delta_{z_m}^2) \cdot \log N}{|\mathcal{L}|}} + (1 - \gamma_d) \epsilon_{\mathcal{S}}. \quad (\text{B.58})$$

We can then have

$$T = \frac{\eta^{-\frac{1}{2}} (1 - \delta)^{-\frac{1}{2}}}{\sqrt{a_1}} = \frac{\eta^{-\frac{1}{2}} (1 - \delta)^{-\frac{1}{2}}}{(1 - 2\epsilon_0)^{\frac{1}{2}}}. \quad (\text{B.59})$$

As long as

$$|\mathcal{L}| \geq \max\{\Omega(\frac{(1 + \delta_{z_m}^2) \cdot \log N}{(1 - 2(1 - \gamma_d)\epsilon_{\mathcal{S}} - \sigma)^2}), BT\}, \quad (\text{B.60})$$

we can obtain

$$F(\mathbf{x}_n) > 1. \quad (\text{B.61})$$

Similarly, we can derive that for  $y_n = -1$ ,

$$F(\mathbf{x}_n) < -1. \quad (\text{B.62})$$

Note that due to the existence of gradient noise by imperfectly balanced training batch, for any  $\mathbf{W} \in \Psi$ ,

$$\Pr \left( \left\| \frac{1}{|\mathcal{B}_b|} \sum_{n \in \mathcal{B}_b} \frac{\partial f_N(\Psi)}{\partial \mathbf{W}} - \mathbb{E} \left[ \frac{\partial f_N(\Psi)}{\partial \mathbf{W}} \right] \right\| \geq \left| \mathbb{E} \left[ \frac{\partial f_N(\Psi)}{\partial \mathbf{W}} \right] \epsilon \right| \right) \leq e^{-B\epsilon^2} \leq N^{-C}, \quad (\text{B.63})$$

if  $B \gtrsim \epsilon^{-2} \log N$  for some  $C > 1$ . Then, the batch size should satisfy  $B \gtrsim \epsilon^{-2} \log N$ . Hence, for all  $n \in \mathcal{V}_d$ ,

$$f_N(\Psi) \leq \epsilon. \quad (\text{B.64})$$

We then have for nodes with actual labels,

$$f(\Psi) \leq 2\epsilon_0 + \epsilon, \quad (\text{B.65})$$

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

#### Proof of Lemma 3.4.5:

This Lemma is proved by (B.50) and (B.51).

#### Proof of Theorem 3.4.6:

The main proof idea is similar to the proof of Theorem 3.4.4. A major difference is that the aggregation matrix does not update, i.e.,  $p_n(t)$  stays at  $t = 0$ . Since that a given core neighborhood and a  $\gamma_d = \Theta(1)$  fraction of discriminative nodes still ensures non-trivial attention weights correlated with class-relevant nodes along the training, the updates of  $\mathbf{W}_O$  and  $\mathbf{W}_V$  are order-wise the same as Lemmas B.3.1 and B.3.4.

Since that

$$p_n(0) = \begin{cases} \frac{\sum_{z \in \mathcal{Z}} |\mathcal{S}_*^{n,t} \cap \mathcal{N}_z^n|}{\sum_{z \in \mathcal{Z}} |\mathcal{S}_*^{n,t} \cap \mathcal{N}_z^n| + \sum_{z \in \mathcal{Z}} (|\mathcal{N}_z^n| - |\mathcal{S}_*^{n,t} \cap \mathcal{N}_z^n|)e^{-1}}, & \text{if } n \in \mathcal{S}_1^{n,t} \cup \mathcal{S}_2^{n,t} \\ \frac{\sum_{z \in \mathcal{Z}} |\mathcal{S}_*^{n,t} \cap \mathcal{N}_z^n|}{\sum_{z \in \mathcal{Z}} (|\mathcal{S}^{n,t}| - |\mathcal{S}^{n,t} \cap \mathcal{N}_z^n|) + \sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_l^{n,t}|_e}, & \text{if } n \notin (\mathcal{S}_1^{n,t} \cup \mathcal{S}_2^{n,t}) \end{cases} \quad (\text{B.66})$$

$$= \Theta(1),$$

there exists  $c_\gamma > 0$ , such that

$$\mathbb{E}[p_n(0)] = \gamma_d \cdot \Theta(\gamma_d) + (1 - \gamma_d)\Theta\left(\frac{\gamma_d}{2}\right) = c_\gamma \gamma_d, \quad (\text{B.67})$$

$$\mathbb{E}[|p_n(0) \pm c_\gamma \gamma_d|^2] \leq \gamma_d \cdot \Theta(\gamma_d^2) + (1 - \gamma_d) \cdot \Theta(|\gamma_d \pm \frac{1}{2}|^2 \gamma_d^2) \leq \Theta(\gamma_d^2). \quad (\text{B.68})$$

Therefore,

$$\begin{aligned} & \left| \frac{1}{|\mathcal{L}|} \sum_{n=1}^N p_n(T)(p_n(T) - \sigma) - c_\gamma^2 \gamma_d^2 \right| \\ & \leq \left| \frac{1}{|\mathcal{L}|} \sum_{n=1}^N p_n(0)(p_n(0) - \sigma) - \mathbb{E}[p_n(0)(p_n(0) - \sigma)] \right| \\ & \quad + \left| \mathbb{E}[|p_n(0)^2 - \sigma p_n(0) - c_\gamma^2 \gamma_d^2|] \right| \\ & \lesssim \sqrt{\frac{\log N}{|\mathcal{L}|}} + \sigma + \sqrt{\mathbb{E}[|p_n(0) + c_\gamma \gamma_d|^2] \cdot \mathbb{E}[|p_n(0) - c_\gamma \gamma_d|^2]} \\ & \lesssim \sqrt{\frac{\log N}{|\mathcal{L}|}} + \sigma + \Theta(\gamma_d^2), \end{aligned} \quad (\text{B.69})$$

where the first step is because  $p_n(T)$  does not update since  $\mathbf{W}_K^{(t)}$  and  $\mathbf{W}_Q^{(t)}$  are fixed at initialization  $\mathbf{W}_K^{(0)}$  and  $\mathbf{W}_Q^{(0)}$ , and the second step is by Cauchy-Schwarz inequality. Since that

$$\sqrt{\frac{\log N}{N}} + \sigma \leq \Theta(\gamma_d^2), \quad (\text{B.70})$$

we have

$$|\mathcal{L}| \geq \Omega\left(\frac{(1 + \delta_{z_m}^2) \log N}{(\gamma_d^2 - \sigma)^2}\right), \quad (\text{B.71})$$

and

$$T = \frac{\eta^{-\frac{1}{2}}}{(1 - 2\epsilon_0)^{\frac{1}{2}}(1 - \delta)^{\frac{1}{2}}\gamma_d^2}. \quad (\text{B.72})$$

### Proof of Lemma 3.4.7:

When  $t = T$ , we have  $\eta T \geq \Theta(1)$ . Since that by Lemma B.3.3 and

$$\left| \frac{1}{B} \sum_{b=1}^T \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{S}_*^{n,T} \cap \mathcal{N}_z^n|}{|\mathcal{S}^{n,T}|} - \frac{1}{B} \sum_{b=1}^T \sum_{n \in \mathcal{B}_b} \frac{|\mathcal{D}_*^n \cap \mathcal{N}_z^n|}{N} \right| \leq \epsilon_*, \quad (\text{B.73})$$

with high probability for some  $1 > \epsilon_* > 0$ , we can derive (3.14).

### Proof of Theorem 3.4.8:

When  $\mathbf{b} = 0$  is fixed during the training, but  $\mathcal{S}^{n,t}$  and  $\mathcal{T}^n$  are subsets of  $\mathcal{N}_{z_m}^n$ , the bound for  $p_n(T)$  is still the same as in (B.50) and (B.51). Given a known core neighborhood in Theorem 3.4.8, the remaining parameters follow the same order-wise update as Lemmas B.3.1, B.3.2 and B.3.4. The remaining proof steps just follow the remaining contents in the proof of Theorem 3.4.4.

## B.4 Useful Lemmas

We prove Lemma B.3.1, B.3.2, B.3.4, and B.3.3 jointly by induction. Lemma B.3.1 first studies the gradient update of lucky neurons in  $\mathcal{W}_l(t)$  in directions of  $\mathbf{p}_1$ ,  $\mathbf{p}_2$ , and other  $\mathbf{p}$ . We divide the updates into several terms and solve each of them. By applying a known result of PDE, we bound the component in the direction of  $\mathbf{p}_1$ , which is the most important one. The updates of other neurons follow the above procedure. Lemma B.3.2 computes the gradient update of  $\mathbf{W}_Q$  and  $\mathbf{W}_K$  in different directions of  $\mathbf{x}_l$ . By controlling the gradient update to be positive in the directions of discriminative nodes, we get a lower bound of  $B$ . Meanwhile, we obtain the update of key and query embeddings. Lemma B.3.4 is derived by considering different components of  $\mathbf{W}_{O_{(i,:)}}$  in the gradient. In proving Lemma B.3.3, we characterize the update of different distance  $z$  in terms of components from different neighborhoods. Combining concentration bounds, we remove the influence on unimportant terms and only retain one part, which represents the update of the average winning margin of the majority vote, i.e., the update of  $\bar{\Delta}(z)$ .

For Lemma B.3.6, We characterize the updates of the lucky neurons to the desired directions to show lucky neurons can activate the self-attention output of discriminative nodes along the training.

### Proof of Lemma B.3.1:

At the  $t$ -th iteration, if  $s \in \mathcal{S}_1^{n,t}$ , we can obtain

$$\begin{aligned} \mathbf{V}_n(t) &= \sum_{s \in \mathcal{S}^{n,t}} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_n(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b}^{(t)}) \\ &= \sum_{s \in \mathcal{S}_1} \text{softmax}_n(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b}^{(t)}) \mathbf{p}_1 + \mathbf{z}(t) + \sum_{j \neq 1} W_j^l(t) \mathbf{p}_j \\ &\quad - \eta \sum_{b=1}^t \left( \sum_{i \in \mathcal{W}_n(0)} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)\top} + \sum_{i \notin \mathcal{W}_l(b)} V_i(b) \lambda \mathbf{W}_{O_{(i,:)}}^{(b)\top} \right), \end{aligned} \quad (\text{B.74})$$

$l \in [M]$ , where the last step comes from Lemma B.3.4. Then we can derive that for  $k \in \mathcal{S}_j^{n,t}$ ,

$$W_k^n(t) \leq \frac{\sum_{z \in \mathcal{Z}} |\mathcal{S}_j^{n,t} \cap \mathcal{N}_z^n| e^{\delta \|\mathbf{q}_1(t)\| + b_z^{(t)}}}{\sum_{z \in \mathcal{Z}} |\mathcal{S}_j^{n,t} \cap \mathcal{N}_z^n| e^{\|\mathbf{q}_1(t)\|^2 - (\sigma + \delta) \|\mathbf{q}_1(t)\| + b_z^{(t)}}} p_n(t), \quad (\text{B.75})$$

which is much smaller than  $\Theta(1)$  when  $t$  is large. This is the reason why we ignore the impact of  $W_l(t)$  on  $\eta \sum_{b=0}^{t-1} (\sum_{i \in \mathcal{W}_l(0)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)\top} + \sum_{i \notin \mathcal{W}_l(0)} V_i(b) \lambda \mathbf{W}_{O_{(i,\cdot)}}^{(b)\top})$ . Hence,

$$\frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{x}_n, y_n)}{\partial \mathbf{W}_{O_{(i)}}^\top} = -\frac{1}{B} \sum_{n \in \mathcal{B}_b} y_n \sum_{l \in \mathcal{S}^{n,t}} a_i \mathbb{1}[\mathbf{W}_{O_{(i)}} \mathbf{V}_l(t) \geq 0] \mathbf{V}_l(t)^\top. \quad (\text{B.76})$$

Denote that for  $j \in [M]$ ,

$$H_4 = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y_n a_i \mathbb{1}[\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_l(t) \geq 0] (-\eta) \sum_{b=1}^t \sum_{k \in \mathcal{W}_l(b)} V_k(b) \mathbf{W}_{O_{(k,\cdot)}}^{(b)} \mathbf{p}_j, \quad (\text{B.77})$$

$$H_4 = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y_n a_i \mathbb{1}[\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_l(t) \geq 0] (-\eta) \sum_{b=1}^t \sum_{k \notin \mathcal{W}_l(b)} V_k(b) \mathbf{W}_{O_{(k,\cdot)}}^{(b)} \mathbf{p}_j, \quad (\text{B.78})$$

and we can then derive

$$\begin{aligned} & \left\langle \mathbf{W}_{O_{(i)}}^{(t+1)\top}, \mathbf{p}_j \right\rangle - \left\langle \mathbf{W}_{O_{(i)}}^{(t)\top}, \mathbf{p}_j \right\rangle \\ &= \frac{1}{B} \sum_{l \in \mathcal{B}_b} \eta y_n a_i \mathbb{1}[\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_l(t) \geq 0] \mathbf{V}_l(t)^\top \mathbf{p}_j \\ &= \frac{1}{B} \sum_{l \in \mathcal{B}_b} \eta y_n a_i \mathbb{1}[\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_l(t) \geq 0] \mathbf{z}_l(t)^\top \mathbf{p}_j \\ &\quad + \frac{1}{B} \sum_{l \in \mathcal{B}_b} \eta y_n a_i \mathbb{1}[\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_l(t) \geq 0] \sum_{s \in \mathcal{S}_l} \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{p}_l^\top \mathbf{p}_j \\ &\quad + \frac{1}{B} \sum_{l \in \mathcal{B}_b} \eta y_n a_i \mathbb{1}[\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_l(t) \geq 0] \sum_{k \neq l} W_l(t) \mathbf{p}_k^\top \mathbf{p}_j + H_4 + H_4 \\ &:= H_1 + H_2 + H_3 + H_4 + H_4, \end{aligned} \quad (\text{B.79})$$

where

$$H_1 = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y_n a_i \mathbb{1}[\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_l(t) \geq 0] \mathbf{z}_l(t)^\top \mathbf{p}_j, \quad (\text{B.80})$$

$$H_2 = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y_n a_i \mathbb{1}[\mathbf{W}_{O_{(i)}}^{(t)} \mathbf{V}_l(t) \geq 0] \sum_{s \in \mathcal{S}_l} \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{p}_l^\top \mathbf{p}_j, \quad (\text{B.81})$$

$$H_3 = \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta y_n a_i \mathbb{1}[\mathbf{W}_{O(i)}^{(t)} \mathbf{V}_l(t) \geq 0] \sum_{j \neq l} W_l(t) \mathbf{p}_j^\top \mathbf{p}_j. \quad (\text{B.82})$$

We then show the statements in different cases.

(1) When  $j = 1$ , since that  $\Pr(y_n = 1) = \Pr(y_n = -1) = 1/2$ , by hoeffding's inequality in (B.6), one can obtain

$$\Pr\left(\left|\frac{1}{B} \sum_{n \in \mathcal{B}_b} y_n\right| \geq \sqrt{\frac{\log B}{B}}\right) \leq B^{-c}, \quad (\text{B.83})$$

$$\Pr\left(\left|\mathbf{z}_l(t)^\top \mathbf{p}_1\right| \geq \sqrt{(\sigma)^2 \log m}\right) \leq m^{-c}. \quad (\text{B.84})$$

Hence, with a high probability, we have

$$|H_1| \leq \frac{\eta(\sigma)}{a} \sqrt{\frac{\log m \log B}{B}}. \quad (\text{B.85})$$

For  $i \in \mathcal{W}_l(0)$ , by the reasoning in (B.127) later, we can obtain

$$\mathbf{W}_{O(i,\cdot)}^{(t)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) > 0. \quad (\text{B.86})$$

Denote  $p_n(t) = |\mathcal{S}_1^{n,t}| \nu_n(t) e^{\|\mathbf{q}_1(t)\|^2 - 2\delta \|\mathbf{q}_1(t)\|}$ . Hence, for  $k \notin \mathcal{W}_l(0)$ ,

$$H_2 \gtrsim \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{a} \|\mathbf{p}_1\|^2 \cdot p_n(t) (1 - 2\epsilon_0), \quad (\text{B.87})$$

$$H_3 = 0, \quad (\text{B.88})$$

$$H_4 \gtrsim \frac{1}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{\eta^2}{a} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) (1 - 2\epsilon_0) \|\mathbf{p}_1\|^2 \left(1 - \epsilon_m - \frac{\sigma M}{\pi}\right) \mathbf{W}_{O(i,\cdot)} \mathbf{p}_1, \quad (\text{B.89})$$

$$\begin{aligned} |H_4| &\lesssim \frac{1}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{\eta^2}{a} \left(1 - \epsilon_m - \frac{\sigma}{\pi}\right) \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{m}{aM} p_n(t) \|\mathbf{p}_1\|^2 \mathbf{W}_{O(i,\cdot)} \mathbf{p}_2 \\ &\quad + \frac{\eta^2 t m}{\sqrt{B} a^2} \mathbf{W}_{O(k,\cdot)} \mathbf{p}_1. \end{aligned} \quad (\text{B.90})$$

Hence, if we combine (B.85), (B.87), (B.88), (B.89), and (B.90), we can derive

$$\begin{aligned}
& \left\langle \mathbf{W}_{O_{(i)}}^{(t+1)\top}, \mathbf{p}_1 \right\rangle - \left\langle \mathbf{W}_{O_{(i)}}^{(t)\top}, \mathbf{p}_1 \right\rangle \\
& \gtrsim \frac{\eta}{a} \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} (p_n(t)(1 - 2\epsilon_0) - \sigma + \eta \sum_{b=1}^t \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t)(1 - \epsilon_m - \frac{\sigma M}{\pi}) \\
& \quad \cdot \mathbf{W}_{O_{(i,\cdot)}} \mathbf{p}_1 (1 - 2\epsilon_0) - \eta \sum_{b=1}^t \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t)(1 - \epsilon_m - \frac{\sigma M}{\pi}) \\
& \quad \cdot \mathbf{W}_{O_{(i,\cdot)}} \mathbf{p}_2 (1 + \sigma) - \frac{\eta t m \mathbf{W}_{O_{(k,\cdot)}} \mathbf{p}_1}{\sqrt{B} a}) \\
& \gtrsim \frac{\eta}{aB} \sum_{n \in \mathcal{B}_b} (p_n(t)(1 - 2\epsilon_0) - \sigma + \frac{\eta t (1 - 2\epsilon_0)}{2B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) \cdot (1 - \epsilon_m - \frac{\sigma M}{\pi})) \\
& \quad \cdot \mathbf{W}_{O_{(i,\cdot)}} \mathbf{p}_1).
\end{aligned} \tag{B.91}$$

Since that  $\mathbf{W}_{O_{(i,\cdot)}}^{(0)} \sim \mathcal{N}(0, \frac{\xi^2 I}{m_a})$ , from the property of Gaussian distribution, we have

$$\Pr(\|\mathbf{W}_{O_{(i,\cdot)}}^{(0)}\| \lesssim \xi) \lesssim \xi. \tag{B.92}$$

Therefore, with high probability for all  $i \in [m]$ , we can derive

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(0)}\| \gtrsim \xi. \tag{B.93}$$

When  $\eta$  is very small, given  $p_n(t)$  as the order of a constant, (B.91) leads to a PDE on the lower bound of  $\mathbf{W}_{O_{(i,\cdot)}} \mathbf{p}_1$  since the last step of (B.91) is always positive. Denote  $y(t)$  as a lower bound of  $\mathbf{W}_{O_{(i,\cdot)}} \mathbf{p}_1$ , we have

$$\begin{aligned}
& \frac{\partial y(t)}{\partial t} \\
& = \Theta\left(\frac{1}{aB} \sum_{n \in \mathcal{B}_b} (p_n(t)(1 - 2\epsilon_0) - \sigma) + \frac{\eta t (1 - 2\epsilon_0)}{2B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) y(t)\right).
\end{aligned} \tag{B.94}$$

Therefore, we can derive

$$\begin{aligned}
y(t) & = e^{\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2 (1 - 2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t)} \left( \int_{-\infty}^t \frac{1}{aB} \sum_{n \in \mathcal{B}_b} (p_n(u)(1 - 2\epsilon_0) - \sigma) \right. \\
& \quad \left. \cdot e^{-\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta u^2 (1 - 2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(u)} du + C_0 \right).
\end{aligned} \tag{B.95}$$

Note that

$$\begin{aligned}
& \int_{-\infty}^t e^{-\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta u^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t)} du \\
& \leq \int_{-\infty}^{\infty} e^{-\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta u^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t)} du \\
& = \sqrt{2\pi} \cdot \left( \frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) \right)^{-1} \\
& = \Theta(\eta^{-1}). \tag{B.96}
\end{aligned}$$

$$\begin{aligned}
& \int_{-\infty}^t e^{-\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta u^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t)} du \\
& \geq \int_{-\infty}^0 e^{-\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta u^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t)} du \\
& = \Theta(\eta^{-1}). \tag{B.97}
\end{aligned}$$

Hence,

$$y(0) = \frac{\eta^{-1}}{aB} \sum_{n \in \mathcal{B}_b} (p_n(t)(1-2\epsilon_0) - \sigma) + C_0 = \Theta(\eta^{-1}\xi) + C_0 = \xi, \tag{B.98}$$

$$C_0 = \xi(1 - \Theta(\eta^{-1})), \tag{B.99}$$

$$\begin{aligned}
& \mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_1 \gtrsim y(t) \\
& \gtrsim e^{\frac{1}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta(t+1)^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t)} \xi \\
& \gtrsim \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta(t+1)^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) + \xi. \tag{B.100}
\end{aligned}$$

(2) When  $\mathbf{p}_j \in \mathcal{P}/p^+$ , we have

$$H_2 = 0, \tag{B.101}$$

$$|H_3| \leq \frac{1}{B} \sum_{n \in \mathcal{B}_b} \nu_n(t) \frac{\eta}{a} \sqrt{\frac{\log m \log B}{B}} \|\mathbf{p}\|^2, \tag{B.102}$$

$$|H_4| \leq \frac{\eta^2}{a} \sum_{b=1}^t \sqrt{\frac{\log m \log B}{B}} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(b) \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}_j. \tag{B.103}$$

For  $k \notin \mathcal{W}_l(0)$ ,

$$|H_5| \lesssim \frac{\eta^2 t m}{\sqrt{B} a^2} \mathbf{W}_{O_{(k,\cdot)}}^\top \mathbf{p}_1 + \frac{\eta^2}{a} \sum_{b=1}^t \sqrt{\frac{\log m \log B}{B}} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}_2, \tag{B.104}$$

with high probability. (B.103) is from (B.11). Then, combining (B.85), (B.101), (B.102), (B.103) and (B.104), we have

$$\begin{aligned} & \left| \left\langle \mathbf{W}_{O_{(i)}}^{(t+1)\top}, \mathbf{p}_j \right\rangle - \left\langle \mathbf{W}_{O_{(i)}}^{(t)\top}, \mathbf{p}_j \right\rangle \right| \\ & \lesssim \frac{\eta}{a} \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} (\nu_n(t) + \sigma \\ & \quad + \sum_{b=1}^t \frac{p_n(b)\eta m}{a} \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}_j) \sqrt{\frac{\log m \log B}{B}}. \end{aligned} \quad (\text{B.105})$$

Comparing (B.91) and (B.105), we have

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_j \lesssim \frac{1}{\sqrt{B}} \mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_1. \quad (\text{B.106})$$

(3) If  $i \in \mathcal{U}_l(0)$ , from the derivation of (B.100) and (B.106), we can obtain

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_2 \gtrsim \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta(t+1)^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) + \xi, \quad (\text{B.107})$$

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_j \lesssim \frac{1}{\sqrt{B}} \mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_2, \quad \text{for } \mathbf{p} \in \mathcal{P}/\mathbf{p}_2. \quad (\text{B.108})$$

(4) If  $i \notin (\mathcal{W}_l(0) \cup \mathcal{U}_{l,n}(0))$ ,

$$|H_2 + H_3| \leq \frac{\eta}{a} \sqrt{\frac{\log m \log B}{B}} \|\mathbf{p}\|^2, \quad (\text{B.109})$$

Following (B.103) and (B.104), we have

$$|H_4| \leq \sum_{b=1}^t \frac{\eta^2}{a} \sqrt{\frac{\log m \log B}{B}} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(b) \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}, \quad (\text{B.110})$$

$$|H_5| \lesssim \frac{\eta^2 t m}{\sqrt{B} a^2} \mathbf{W}_{O_{(k,\cdot)}}^{(t)} \mathbf{p}_1 + \sum_{b=1}^t \frac{\eta^2}{a} \sqrt{\frac{\log m \log B}{B}} \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(b) \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p}_2. \quad (\text{B.111})$$

Thus, combining (B.109), (B.110), and (B.111), we can derive

$$\begin{aligned} & \left| \left\langle \mathbf{W}_{O_{(i,\cdot)}}^{(t+1)\top}, \mathbf{p} \right\rangle - \left\langle \mathbf{W}_{O_{(i,\cdot)}}^{(t)\top}, \mathbf{p} \right\rangle \right| \\ & \lesssim \frac{\eta}{a} \cdot (\|\mathbf{p}\| + \sigma + \sum_{b=1}^t \frac{1}{2B} \sum_{n \in \mathcal{B}_b} \frac{p_n(b)\eta m}{a} \mathbf{W}_{O_{(i,\cdot)}}^\top \mathbf{p}_j) \sqrt{\frac{\log m \log B}{B}}, \end{aligned} \quad (\text{B.112})$$

Comparing (B.91) and (B.112), we can obtain

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_j \lesssim \frac{1}{\sqrt{B}} \mathbf{W}_{O_{(j,\cdot)}}^{(t+1)} \mathbf{p}_1, \quad (\text{B.113})$$

for  $j \in \mathcal{W}_l(0)$ .

(5) In this part, we study the bound of  $\mathbf{W}_{O_{(i,\cdot)}}^{(t)}$  and the product with the noise term according to the analysis above.

By (B.29), for the lucky neuron  $i$ , since that the update of  $\mathbf{W}_{O_{(i,\cdot)}}^{(t)}$  lies in the subspace spanned by  $\mathcal{P}$ , we can obtain

$$\begin{aligned} \|\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)}\|^2 &= \sum_{l=1}^M (\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_l)^2 \geq (\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{p}_1)^2 \\ &\gtrsim \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta(t+1)^2(1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) \right)^2, \end{aligned} \quad (\text{B.114})$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)} \mathbf{z}_l(t)\| \leq \left| \sigma \|\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)}\| \right|. \quad (\text{B.115})$$

For the unlucky neuron  $i$ , we can similarly get

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t+1)}\|^2 \leq \frac{1}{B} \|\mathbf{W}_{O_{(j,\cdot)}}^{(t+1)}\|^2, \quad (\text{B.116})$$

where  $j$  is a lucky neuron. The proof of Lemma B.3.1 finishes here.

### Proof of Lemma B.3.2:

We first study the gradient of  $\mathbf{W}_Q^{(t+1)}$  in part (a) and the gradient of  $\mathbf{W}_K^{(t+1)}$  in part (b).

(a) from (B.1), we can obtain

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{l \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y_n)}{\partial \mathbf{W}_Q} \\
&= \eta \frac{1}{B} \sum_{l \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y_n)}{\partial F(\mathbf{X}^n)} \frac{\partial F(\mathbf{X}^n)}{\partial \mathbf{W}_Q} \\
&= \eta \frac{1}{B} \sum_{l \in \mathcal{B}_b} (-y_n) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l \\
&\quad + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \geq 0] \cdot \left( \mathbf{W}_{O_{(i,:)}} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \right. \\
&\quad \cdot \left. \sum_{r \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) \mathbf{W}_K (\mathbf{x}_s - \mathbf{x}_r) \mathbf{x}_l^\top \right) \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-y_l) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l \\
&\quad + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \geq 0] \left( \mathbf{W}_{O_{(i,:)}} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \right. \\
&\quad \cdot \left. (\mathbf{W}_K \mathbf{x}_s - \sum_{r \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) \mathbf{W}_K \mathbf{x}_r) \mathbf{x}_l^\top \right).
\end{aligned} \tag{B.117}$$

(i) If  $l \in \mathcal{S}_1^{n,t}$  or  $l \in \mathcal{S}_2^{n,t}$ , say  $l \in \mathcal{S}_1^{n,t}$ , we have the following derivation.

At the initial point, we can obtain

$$\mathbf{W}_{O_{(i,:)}}^{(0)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V^{(0)} \mathbf{x}_s \text{softmax}_l((\mathbf{W}_K^{(0)} \mathbf{x}_s)^\top \mathbf{W}_Q^{(0)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(0)}) > 0, \tag{B.118}$$

and

$$\begin{aligned}
& \text{softmax}_l((\mathbf{W}_K^{(0)} \mathbf{x}_s)^\top \mathbf{W}_Q^{(0)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(0)}) \\
& \geq \Omega(1) \cdot \sum_{r \in \mathcal{S}_2^{l,t}} \text{softmax}_l((\mathbf{W}_K^{(0)} \mathbf{x}_r)^\top \mathbf{W}_Q^{(0)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(0)}),
\end{aligned} \tag{B.119}$$

for  $s \in \mathcal{S}_1^{l,t}$ .

For  $r, l \in \mathcal{S}_1^{l,t}$ , if  $u_{(r,l)z_0} = 1$ , by (B.22) we have

$$\begin{aligned}
& \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t)} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) \\
& \gtrsim \frac{e^{\|\mathbf{q}_1(t)\|^2 - \delta \|\mathbf{q}_1(t)\| + b_{z_0}^{(t)}}}{\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| e^{\|\mathbf{q}_1(T)\|^2 - \sigma \|\mathbf{q}_1(T)\| + b_z^{(T)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cap \mathcal{S}^{n,T}) - \mathcal{S}_1^{n,T}| e^{b_z^{(T)}}}.
\end{aligned} \tag{B.120}$$

Likewise, for  $r \notin \mathcal{S}_1^{l,t}$  and  $l \in \mathcal{S}_1^{l,t}$ , we have

$$\begin{aligned} & \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) \\ & \lesssim \frac{e^{b_{z_0}^{(t)}}}{\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| e^{\|\mathbf{q}_1(T)\|^2 - \sigma \|\mathbf{q}_1(T)\| + b_z^{(T)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cap \mathcal{S}^{n,T}) - \mathcal{S}_1^{n,T}| e^{b_z^{(T)}}}. \end{aligned} \quad (\text{B.121})$$

Therefore, for  $s, r, l \in \mathcal{S}_1^{n,t}$ , let

$$\mathbf{W}_K^{(t)} \mathbf{x}_s - \sum_{r \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) \mathbf{W}_K^{(t)} \mathbf{x}_r := \beta_1^l(t) \mathbf{q}_1(t) + \beta_2^l(t), \quad (\text{B.122})$$

where

$$\begin{aligned} \beta_1^l(t) & \gtrsim \frac{\sum_{z \in \mathcal{Z}} (|\mathcal{N}_z^n \cap \mathcal{S}^{l,t}| - |\mathcal{N}_z^l \cap \mathcal{S}_1^{l,t}|) e^{b_z(t)}}{\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^l \cap \mathcal{S}_*^{n,T}| e^{\|\mathbf{q}_1(T)\|^2 - \sigma \|\mathbf{q}_1(T)\| + b_z^{(T)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cap \mathcal{S}^{n,T}) - \mathcal{S}_1^{n,T}| e^{b_z^{(T)}}} \\ & \gtrsim \phi_l(t) (|\mathcal{S}^{l,t}| - |\mathcal{S}_1^{l,t}|), \end{aligned} \quad (\text{B.123})$$

$$\beta_1^l(t) \lesssim e^{2\delta \|\mathbf{q}_1(t)\|} \phi_l(t) (|\mathcal{S}^{l,t}| - |\mathcal{S}_1^{l,t}|) \leq \phi_l(t) (|\mathcal{S}^{l,t}| - |\mathcal{S}_1^{l,t}|). \quad (\text{B.124})$$

Meanwhile,

$$\begin{aligned} \beta_2^l(t) & \approx \Theta(1) \cdot \mathbf{o}_j^l(t) + Q_e(t) \mathbf{r}_2(t) + \sum_{n=3}^M \gamma'_n \mathbf{r}_n(t) - \sum_{a=1}^M \sum_{r \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t)\top} \\ & \quad \mathbf{W}_Q^{(t)} \mathbf{x}_l) \mathbf{r}_a(t) = \Theta(1) \cdot \mathbf{o}_j^l(t) + \sum_{n=1}^M \zeta'_n \mathbf{r}_n(t), \end{aligned} \quad (\text{B.125})$$

for some  $Q_e(t) > 0$  and  $\gamma'_l > 0$ . Here

$$|\zeta'_l| \leq \beta_1^n(t) \frac{|\mathcal{S}_l^{n,t}|}{|\mathcal{S}^{n,t}| - |\mathcal{S}_1^{n,t}|}, \quad (\text{B.126})$$

for  $l \geq 2$ . Note that  $|\zeta'_l| = 0$  if  $|\mathcal{S}^{n,t}| = |\mathcal{S}_1^{n,t}|$ ,  $l \geq 2$ .

For  $i \in \mathcal{W}_l(0)$ , by Lemma B.3.6,

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) > 0. \quad (\text{B.127})$$

Then we study how large the coefficient of  $\mathbf{q}_1(t)$  in (B.117).

If  $s \in \mathcal{S}_1^{l,t}$ , from basic mathematical computation given (B.9) to (B.12),

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \\ & \gtrsim \frac{p_l(t)}{|\mathcal{S}_1^{n,t}|} \left( \frac{1-2\epsilon_0}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi\eta(t+1)^2 m}{a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} p_n(b) - \sigma \right) \right. \\ & \quad \left. + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)}{a} (1-\sigma) \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{(1-2\epsilon_0)\eta(t+1)^2}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_l(t) \right)^2 \right). \end{aligned} \quad (\text{B.128})$$

If  $s \in \mathcal{S}_2^{l,t}$  and  $j \in \mathcal{S}_1^{l,t}$ , from (B.13) to (B.16), we have

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \\ & \lesssim \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{W}_V^{(t)} \mathbf{x}_j \text{softmax}_l(\mathbf{x}_j^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(j,l)}^\top \mathbf{b}^{(t)}) \cdot \phi_n(t) \frac{|\mathcal{S}_1^{n,t}|}{p_l(t)}. \end{aligned} \quad (\text{B.129})$$

If  $i \in \mathcal{W}_l(0)$ ,  $s \notin (\mathcal{S}_1^{l,t} \cup \mathcal{S}_2^{l,t})$ , and  $j \in \mathcal{S}_1^{l,t}$ ,

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \\ & \lesssim \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{W}_V^{(t)} \mathbf{x}_j \text{softmax}_l(\mathbf{x}_j^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(j,l)}^\top \mathbf{b}^{(t)}) \phi_l(t) \cdot \frac{|\mathcal{S}_1^{l,t}|}{p_l(t)}, \end{aligned} \quad (\text{B.130})$$

by (B.17) to (B.19).

Hence, for  $i \in \mathcal{W}_l(0)$ ,  $j \in \mathcal{S}_1^{g,t}$ , combining (B.123) and (B.128), we can obtain

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{q}_1(t)^\top \\ & \quad \cdot (\mathbf{W}_K^{(t)} \mathbf{x}_s - \sum_{r \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) \mathbf{W}_K^{(t)} \mathbf{x}_r) \mathbf{x}_l^\top \mathbf{x}_j \\ & \gtrsim \left( \frac{1-2\epsilon_0}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi\eta(t+1)^2 m}{a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} p_n(b) - \sigma \right) + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)}{a} (1-\sigma) \right. \\ & \quad \left. \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{(1-2\epsilon_0)\eta(t+1)^2}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_l(t) \right)^2 \right) \phi_l(t) (|\mathcal{S}^{l,t}| - |\mathcal{S}_1^{l,t}|) \|\mathbf{q}_1(t)\|^2. \end{aligned} \quad (\text{B.131})$$

For  $i \in \mathcal{U}_l(t)$  and  $l \in \mathcal{S}_1^{l,t}$ ,  $j \in \mathcal{S}_1^{g,t}$ , and  $k \in \mathcal{W}_l(0)$ ,

$$\begin{aligned}
& \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{q}_1(t)^\top \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{x}_s - \sum_{r \in \mathcal{S}^{n,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{W}_K^{(t)} \mathbf{x}_r) \mathbf{x}_l^\top \mathbf{x}_j \\
& \lesssim \left( \frac{1-2\epsilon_0}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi\eta(t+1)^2 m}{a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} p_n(b) - \sigma \right) + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)}{a} (1-\sigma) \right. \\
& \quad \cdot \left. \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{(1-2\epsilon_0)\eta(t+1)^2}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_l(t) \right)^2 \right) \phi_n(t) |\mathcal{S}_2^{n,t}| \cdot \beta_1(t) \|\mathbf{q}_1(t)\|^2. \tag{B.132}
\end{aligned}$$

For  $i \notin (\mathcal{W}_l(t) \cup \mathcal{U}_l(t))$  and  $l \in \mathcal{S}_1^{l,t}$ ,  $j \in \mathcal{S}_1^g$ ,

$$\begin{aligned}
& \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{q}_1(t)^\top \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{x}_s - \sum_{r \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{x}_r) \mathbf{x}_l^\top \mathbf{x}_j \\
& \lesssim \mathbf{W}_{O_{(k,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V^{(t)} \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{q}_1(t)^\top \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{x}_s - \sum_{r \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{x}_r) \mathbf{x}_l^\top \mathbf{x}_j \cdot \frac{1}{\sqrt{B}}. \tag{B.133}
\end{aligned}$$

Therefore, by the update rule,

$$\begin{aligned}
\mathbf{W}_Q^{(t+1)} \mathbf{x}_j &= \mathbf{W}_Q^{(t)} \mathbf{x}_j - \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{\partial \text{Loss}(\mathbf{X}, y_n)}{\partial \mathbf{W}_Q} \Big| \mathbf{W}_Q^{(t)} \right) \mathbf{x}_j \\
&= \mathbf{r}_1(t) + K(t) \mathbf{q}_1(t) + \Theta(1) \cdot \mathbf{n}_j(t) + |K_e|(t) \mathbf{q}_2(t) + \sum_{l=3}^M \gamma'_l \mathbf{q}_l(t) \\
&= (1 + K(t)) \mathbf{q}_1(t) + \Theta(1) \cdot \mathbf{n}_j(t) + |K_e|(t) \mathbf{q}_2(t) + \sum_{l=3}^M \gamma'_l \mathbf{q}_l(t), \tag{B.134}
\end{aligned}$$

where the last step is by

$$\mathbf{q}_1(t) = k_1(t) \cdot \mathbf{r}_1(t), \tag{B.135}$$

and

$$\mathbf{q}_2(t) = k_2(t) \cdot \mathbf{r}_2(t), \tag{B.136}$$

for  $k_1(t) > 0$  and  $k_2(t) > 0$  from induction, i.e.,  $\mathbf{q}_1(t)$  and  $\mathbf{r}_1(t)$ ,  $\mathbf{q}_1(t)$  and  $\mathbf{r}_1(t)$  are from the same direction, respectively. Define  $qc_t(\mathbf{x}) = \mathbf{x}^\top \mathbf{q}_1(t) / \|\mathbf{q}_1(t)\|$  and denote

$$\begin{aligned} \Delta(l, i) = & a_i \mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}) \geq 0] \\ & \cdot \left( \mathbf{W}_{O_{(i,\cdot)}} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b} \right. \\ & \left. \cdot (\mathbf{W}_K \mathbf{x}_s - \sum_{r \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}) \mathbf{W}_K \mathbf{x}_r) \mathbf{x}_l^\top \right). \end{aligned} \quad (\text{B.137})$$

We then have

$$\begin{aligned} K(t) & \gtrsim \eta \frac{1}{B} \left( \left| \sum_{l \in \mathcal{B}_b, l \in \mathcal{S}_1^{l,t}} (-y_l) \sum_{i \in \mathcal{W}_l(0)} qc_t(\Delta(l, i)) \right| - \left| \sum_{l \in \mathcal{B}_b, l \in \mathcal{S}_1^{l,t}} (-y_l) \sum_{i \in \mathcal{U}_{l,n}(0)} qc_t(\Delta(l, i)) \right| \right. \\ & \quad - \left| \sum_{l \in \mathcal{B}_b, l \in \mathcal{S}_1^{l,t}} (-y_l) \sum_{i \notin \mathcal{W}_l(0) \cup \mathcal{U}_{l,n}(0)} qc_t(\Delta(l, i)) \right| - \left| \sum_{l \in \mathcal{B}_b, l \in \mathcal{S}_2^{l,t}} (-y_l) \sum_{i=1}^m qc_t(\Delta(l, i)) \right| \\ & \quad \left. - \left| \sum_{l \in \mathcal{B}_b, l \in \mathcal{S}^{l,t} - \mathcal{S}_1^{l,t} - \mathcal{S}_2^{l,t}} (-y_l) \sum_{i=1}^m qc_t(\Delta(l, i)) \right| \right) \\ & \gtrsim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} \left( \frac{1-2\epsilon_0}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 m}{a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} p_n(b) - \sigma \right) + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)}{a} \right. \\ & \quad \left. \cdot (1-\sigma) \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{(1-2\epsilon_0)\eta(t+1)^2}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_l(t) \right)^2 \right) \phi_l(t) (|\mathcal{S}^{l,t}| - |\mathcal{S}_1^{l,t}|) \|\mathbf{q}_1(t)\|^2 \\ & > 0, \end{aligned} \quad (\text{B.138})$$

$$|\gamma'_l| \lesssim \frac{1}{B} \sum_{n \in \mathcal{B}_b} K(t) \cdot \frac{|\mathcal{S}_l^{n,t}|}{|\mathcal{S}^{n,t}| - |\mathcal{S}_1^{n,t}|}, \quad (\text{B.139})$$

$$|K_e(t)| \lesssim \frac{1}{B} \sum_{n \in \mathcal{B}_b} \lambda \cdot K(t) \cdot \frac{|\mathcal{S}_2^{n,t}|}{|\mathcal{S}^{n,t}| - |\mathcal{S}_1^{n,t}|}, \quad (\text{B.140})$$

as long as

$$\begin{aligned}
& \left( \frac{1-2\epsilon_0}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 m}{a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} p_n(b) - \sigma \right) + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)}{a} (1-\sigma) \right. \\
& \quad \cdot \left. \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{(1-2\epsilon_0)\eta(t+1)^2}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_l(t)^2 \right) \phi_l(t) (|\mathcal{S}^{l,t}| - |\mathcal{S}_1^{l,t}|) \|\mathbf{q}_1(t)\|^2 \right) \\
& \gtrsim \left( \frac{1-2\epsilon_0}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 m}{a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} p_n(b) - \sigma \right) + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)}{a} (1-\sigma) \right. \\
& \quad \cdot \left. \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{(1-2\epsilon_0)\eta(t+1)^2}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_l(t)^2 \right) \phi_l(t) |\mathcal{S}_2^{l,t}| \cdot \beta_1(t) \|\mathbf{q}_1(t)\|^2 \right). \tag{B.141}
\end{aligned}$$

To find the sufficient condition for (B.141), we compare the LHS with two terms of RHS in (B.141). Note that when  $|\mathcal{S}^{n,t}| > |\mathcal{S}_1^{n,t}|$ , by (B.124),

$$\phi_n(t) (|\mathcal{S}^{n,t}| - |\mathcal{S}_1^{n,t}|) \gtrsim \beta_1^n(t). \tag{B.142}$$

Moreover,

$$1 \gtrsim \phi_n(t) |\mathcal{S}_2^{n,t}|. \tag{B.143}$$

For the second term on RHS, we can derive the bound in the same way.

(ii) Then we provide a brief derivation of  $\mathbf{W}_Q^{(t+1)} \mathbf{x}_j$  for  $j \notin (\mathcal{S}_1^{n,t} \cup \mathcal{S}_2^{n,t})$  in the following.

To be specific, for  $j \in \mathcal{S}_n / (\mathcal{S}_1^{n,t} \cup \mathcal{S}_2^{n,t})$ ,

$$\begin{aligned}
& \left\langle \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}, y_n)}{\partial \mathbf{W}_Q^{(t)}} \mathbf{x}_j^n, \mathbf{q}_1(t) \right\rangle \\
& \gtrsim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} \left( \frac{1-2\epsilon_0}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 m}{a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} p'_n(b) - \sigma \right) + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p'_n(b)}{a} (1-\sigma) \right. \\
& \quad \cdot \left. (1-\sigma) \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{(1-2\epsilon_0)\eta(t+1)^2}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p'_n(t)^2 \right) \phi_n(t) (|\mathcal{S}^{l,t}| - |\mathcal{S}_1^{l,t}|) \|\mathbf{q}_1(t)\|^2 \right) \tag{B.144}
\end{aligned}$$

where

$$p'_n(t) = \frac{\sum_{z \in \mathcal{Z}} |\mathcal{S}_1^{n,t} \cap \mathcal{N}_z^n| e^{\mathbf{q}_1(t)^\top \sum_{b=1}^t K(b) \mathbf{q}_1(0) - \delta} \| \mathbf{q}_1(t) \| + b_z^{(t)}}{\sum_{z \in \mathcal{Z}} |(\mathcal{S}_1^{n,t} \cup \mathcal{S}_2^{n,t}) \cap \mathcal{N}_z^n| e^{\mathbf{q}_1(t)^\top \sum_{b=1}^t K(b) \mathbf{q}_1(b) - \delta} \| \mathbf{q}_1(t) \| + b_z^{(t)} + |\mathcal{S}^{n,t}| - |\mathcal{S}_1^{n,t}| - |\mathcal{S}_2^{n,t}|}. \quad (\text{B.145})$$

When  $K(b)$  is close to  $0^+$ , we have

$$\prod_{b=1}^t \sqrt{1 + K(b)} \| \mathbf{q}(0) \|^2 \gtrsim e^{\sum_{b=1}^t K(b) \| \mathbf{q}_1(0) \|^2} \geq \sum_{b=1}^t K(b) \| \mathbf{q}_1(0) \|^2, \quad (\text{B.146})$$

where the first step comes from  $\log(1 + x) \approx x$  when  $x \rightarrow 0^+$ . Therefore, one can derive that

$$\left\langle \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y_n)}{\partial \mathbf{W}_Q^{(t)}} \mathbf{x}_j^n, \mathbf{q}_1(t) \right\rangle \gtrsim \Theta(1) \cdot K(t). \quad (\text{B.147})$$

At the same time, the value of  $p'_n(t)$  will increase to 1 along the training, making the component of  $\mathbf{q}_1(t)$  the major part in  $\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y_n)}{\partial \mathbf{W}_Q^{(t)}} \mathbf{x}_j^n$ . This is also the same for  $\mathbf{q}_2(t)$ .

Hence, if  $j \in \mathcal{S}_l^{n,t}$  for  $l \geq 3$ ,

$$\mathbf{W}_Q^{(t+1)} \mathbf{x}_j = \mathbf{q}_l(t) + \Theta(1) \cdot \mathbf{n}_j(t) + \Theta(1) \cdot K(t) (\mathbf{q}_1(t) + \mathbf{q}_2(t)) + \sum_{l=2}^M \gamma'_l \mathbf{q}_l(t). \quad (\text{B.148})$$

Similarly, for  $j \in \mathcal{S}_2^{n,t}$ ,

$$\mathbf{W}_Q^{(t+1)} \mathbf{x}_j = (1 + K(t) \frac{|\mathcal{S}_2^{n,t}|}{|\mathcal{S}_1^{n,t}|}) \mathbf{q}_2(t) + \Theta(1) \cdot \mathbf{n}_j(t) + \Theta(1) \cdot K(t) \mathbf{q}_1(t) + \sum_{l=2}^M \gamma'_l \mathbf{q}_l(t). \quad (\text{B.149})$$

(b) For the gradient of  $\mathbf{W}_K$ , we have

$$\begin{aligned} & \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{x}_n, y_n)}{\partial F(\mathbf{x}_n)} \frac{\partial F(\mathbf{x}_n)}{\partial \mathbf{W}_K} \\ &= \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-y_n) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O(i,\cdot)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \geq 0] \\ & \quad \cdot \left( \mathbf{W}_{O(i,\cdot)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{W}_Q^\top \mathbf{x}_l \right. \\ & \quad \left. \cdot (\mathbf{x}_s - \sum_{r \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{x}_r)^\top \right). \end{aligned} \quad (\text{B.150})$$

Hence, for  $j \in \mathcal{S}_1^{n,t}$ , we can follow (B.134) to derive

$$\mathbf{W}_K^{(t+1)} \mathbf{x}_j \approx (1 + Q(t)) \mathbf{q}_1(t) + \Theta(1) \cdot \mathbf{o}_j(t) + |Q_e(t)| \mathbf{r}_2(t) + \sum_{l=3}^M \gamma_l' \mathbf{r}_l(t), \quad (\text{B.151})$$

where

$$Q(t) \geq K(t)(1 - \lambda) > 0, \quad (\text{B.152})$$

for  $\lambda < 1$ , and

$$|\gamma_l| \lesssim \frac{1}{B} \sum_{n \in \mathcal{B}_b} Q(t) \cdot \frac{|\mathcal{S}_l^{n,t}|}{|\mathcal{S}^{n,t}| - |\mathcal{S}_*^{n,t}|}, \quad (\text{B.153})$$

$$|Q_e(t)| \lesssim \frac{1}{B} \sum_{n \in \mathcal{B}_b} Q(t) \cdot \frac{|\mathcal{S}_\#^{n,t}|}{|\mathcal{S}^{n,t}| - |\mathcal{S}_*^{n,t}|}. \quad (\text{B.154})$$

Similarly, for  $j \in \mathcal{S}_2^{n,t}$ , we can obtain

$$\mathbf{W}_K^{(t+1)} \mathbf{x}_j \approx (1 + Q(t)) \mathbf{q}_2(t) + \Theta(1) \cdot \mathbf{o}_j(t) + |Q_e(t)| \mathbf{r}_1(t) + \sum_{l=3}^M \gamma_l' \mathbf{r}_l(t), \quad (\text{B.155})$$

For  $j \in \mathcal{S}_l^{n,t}$ ,  $l = 3, 4, \dots, M$ , we can obtain

$$\mathbf{W}_K^{(t+1)} \mathbf{x}_j \approx \mathbf{q}_l(t) + \Theta(1) \cdot \mathbf{o}_j(t) + \Theta(1) \cdot |Q_f(t)| \mathbf{r}_1(t) + \Theta(1) \cdot Q_f(t) \mathbf{r}_2(t) + \sum_{i=3}^M \gamma_i' \mathbf{r}_i(t), \quad (\text{B.156})$$

where

$$|Q_f(t)| \lesssim Q(t). \quad (\text{B.157})$$

Therefore, for  $l \in \mathcal{S}_1^{n,t}$ , if  $j \in \mathcal{S}_1^{n,t}$ ,

$$\begin{aligned}
& \mathbf{x}_j^\top \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l \\
& \gtrsim (1 + K(t))(1 + Q(t)) \|\mathbf{q}_1(t)\|^2 - \delta \|\mathbf{q}_1(t)\| + K_e(t) Q_e(t) \|\mathbf{q}_2(t)\| \|\mathbf{r}_2(t)\| \\
& \quad + \sum_{l=3}^M \gamma_l \gamma_l' \|\mathbf{q}_l(t)\| \|\mathbf{r}_l(t)\| \\
& \gtrsim (1 + K(t))(1 + Q(t)) \|\mathbf{q}_1(t)\|^2 - \delta \|\mathbf{q}_1(t)\| \\
& \quad - \sqrt{\sum_{l=2}^M \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} Q(t) \frac{|\mathcal{S}_l^{n,t}|}{|\mathcal{S}^{n,t}| - |\mathcal{S}_*^{n,t}|} \right)^2 \|\mathbf{r}_l(t)\|^2} \\
& \quad \cdot \sqrt{\sum_{l=2}^M \left( \frac{1}{B} \sum_{n \in \mathcal{B}_b} K(t) \frac{|\mathcal{S}_l^{n,t}|}{|\mathcal{S}^{n,t}| - |\mathcal{S}_*^{n,t}|} \right)^2 \|\mathbf{q}_l(t)\|^2} \\
& \gtrsim (1 + K(t) + Q(t)) \|\mathbf{q}_1(t)\|^2 - \delta \|\mathbf{q}_1(t)\|,
\end{aligned} \tag{B.158}$$

where the second step is from Cauchy-Schwarz inequality.

If  $j \notin \mathcal{S}_1^{n,t}$ ,

$$\begin{aligned}
& \mathbf{x}_j^\top \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l \\
& \lesssim (1 + K(t)) Q_f(t) \|\mathbf{q}_1(t)\|^2 + K_e(t) Q_f(t) \|\mathbf{q}_2(t)\|^2 + \gamma_l \|\mathbf{q}_l(t)\|^2 + \delta \|\mathbf{q}_1(t)\| \\
& \lesssim Q_f(t) \|\mathbf{q}_1(t)\|^2 + \delta \|\mathbf{q}_1(t)\|.
\end{aligned} \tag{B.159}$$

Therefore, for  $r, l \in \mathcal{S}_1^{l,t}$ , if  $u_{(r,l)z_0} = 1$ , we have

$$\begin{aligned}
& \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t+1)}) \\
& \gtrsim \frac{e^{(1+K(t))\|\mathbf{q}_1(t)\|^2 - \delta \|\mathbf{q}_1(t)\| + b_{z_0}^{(t)}}}{\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| e^{(1+K(t))\|\mathbf{q}_1(T)\|^2 - \sigma \|\mathbf{q}_1(T)\| + b_z^{(T)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cap \mathcal{S}^{n,T}) - \mathcal{S}_1^{n,T}| e^{b_z^{(T)}}}.
\end{aligned} \tag{B.160}$$

Similarly, for  $r \notin \mathcal{S}_1^{l,t}$  and  $l \in \mathcal{S}_1^{l,t}$ , we have

$$\begin{aligned}
& \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^{(t+1)\top} \mathbf{W}_Q^{(t+1)} \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) \\
& \lesssim \frac{e^{b_{z_0}^{(t)}}}{\sum_{z \in \mathcal{Z}} |\mathcal{N}_z^n \cap \mathcal{S}_*^{n,T}| e^{(1+K(t))\|\mathbf{q}_1(T)\|^2 - \sigma \|\mathbf{q}_1(T)\| + b_z^{(T)}} + \sum_{z \in \mathcal{Z}} |(\mathcal{N}_z^n \cap \mathcal{S}^{n,T}) - \mathcal{S}_1^{n,T}| e^{b_z^{(T)}}}.
\end{aligned} \tag{B.161}$$

The same conclusion holds if  $l \notin (\mathcal{S}_1^{n,t} \cup \mathcal{S}_2^{n,t})$ .

Hence

$$\mathbf{q}_1(t+1) = \sqrt{(1+K(t))}\mathbf{q}_1(t). \quad (\text{B.162})$$

$$\mathbf{q}_2(t+1) = \sqrt{(1+K(t))}\mathbf{q}_2(t). \quad (\text{B.163})$$

$$\mathbf{r}_1(t+1) = \sqrt{(1+Q(t))}\mathbf{r}_1(t). \quad (\text{B.164})$$

$$\mathbf{r}_2(t+1) = \sqrt{(1+Q(t))}\mathbf{r}_2(t). \quad (\text{B.165})$$

It can also be verified that this Lemma holds when  $t = 1$ .

**Proof of Lemma B.3.3:**

$$\begin{aligned} & \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{x}_n, y_n)}{\partial \mathbf{b}} \\ &= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{x}_n, y_n)}{\partial F(\mathbf{x}_n)} \frac{\partial F(\mathbf{x}_n)}{\partial \mathbf{b}} \\ &= \eta \frac{1}{B} \sum_{l \in \mathcal{B}_b} (-y_l) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O(i,\cdot)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \\ &\geq 0] \cdot \left( \mathbf{W}_{O(i,\cdot)} \cdot \sum_{s \in \mathcal{S}_l} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \right. \\ &\quad \left. \sum_{r \in \mathcal{S}^{n,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \cdot (\mathbf{u}_{(s,l)} - \mathbf{u}_{(r,l)}) \right) \\ &= \eta \frac{1}{B} \sum_{l \in \mathcal{B}_b} (-y_l) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O(i,\cdot)} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \geq 0] \\ &\quad \cdot \left( \mathbf{W}_{O(i,\cdot)} \sum_{s \in \mathcal{S}_l} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) (\mathbf{u}_{(s,l)} - \sum_{r \in \mathcal{S}^{n,t}} \right. \\ &\quad \left. \cdot \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{u}_{(r,l)}) \right). \end{aligned} \quad (\text{B.166})$$

Therefore, we can derive

$$\begin{aligned}
& \mathbf{W}_{O_{(i,\cdot)}} \cdot \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) (u_{(s,l)_z}) \\
& - \sum_{r \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) u_{(r,l)_z}) \\
= & \mathbf{W}_{O_{(i,\cdot)}} \sum_{s \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) (1) \\
& - \sum_{r \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) + \mathbf{W}_{O_{(i,\cdot)}} \sum_{s \in \mathcal{S}^{l,t} - \mathcal{N}_l^z} \mathbf{W}_V \mathbf{x}_s \\
& \cdot \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) (- \sum_{r \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l \\
& + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)})) \\
= & \mathbf{W}_{O_{(i,\cdot)}} \sum_{s \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \sum_{r \in \mathcal{S}^{l,t} - \mathcal{N}_l^z} \\
& \cdot \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) - \mathbf{W}_{O_{(i,\cdot)}} \sum_{s \in \mathcal{S}^{l,t} - \mathcal{N}_l^z} \mathbf{W}_V \mathbf{x}_s \\
& \cdot \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \sum_{r \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l \\
& + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) \\
= & P_1 + P_2 + P_3,
\end{aligned} \tag{B.167}$$

where the second step is by

$$\left( \sum_{s \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^{z-1}} + \sum_{s \in \mathcal{S}^{l,t} - \mathcal{N}_l^{z-1}} \right) \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) = 1. \tag{B.168}$$

Define

$$\begin{aligned}
P_1 = & \sum_{s \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z \cap \mathcal{S}_*^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \\
& \cdot \sum_{r \in (\mathcal{S}^{l,t} - \mathcal{N}_l^z) \cap \mathcal{S}_*^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) - \sum_{s \in (\mathcal{S}^{l,t} - \mathcal{N}_l^z) \cap \mathcal{S}_*^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \\
& \mathbf{W}_V \mathbf{x}_s \cdot \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \\
& \sum_{r \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z \cap \mathcal{S}_*^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}),
\end{aligned} \tag{B.169}$$

$$\begin{aligned}
P_2 = & \sum_{s \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z - \mathcal{S}_*^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}) \sum_{r \in \mathcal{S}^{l,t} - \mathcal{N}_z^l} \\
& \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) - \sum_{s \in (\mathcal{S}^{l,t} - \mathcal{N}_z^l) - \mathcal{S}_*^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \\
& \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}) \sum_{r \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l \\
& + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}),
\end{aligned} \tag{B.170}$$

$$\begin{aligned}
P_3 = & \sum_{s \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z \cap \mathcal{S}_*^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \sum_{r \in (\mathcal{S}^{l,t} - \mathcal{N}_z^l) - \mathcal{S}_*^{l,t}} \\
& \cdot \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) - \sum_{s \in (\mathcal{S}^{l,t} - \mathcal{N}_z^l) \cap \mathcal{S}_*^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \\
& \cdot \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \sum_{r \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z - \mathcal{S}_*^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l \\
& + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}).
\end{aligned} \tag{B.171}$$

Note that  $((\mathcal{S}^{l,t} - \mathcal{N}_l^{z-1}) \cap \mathcal{S}_*^{l,t}) + (\mathcal{S}^{l,t} \cap \mathcal{N}_l^{z-1} \cap \mathcal{S}_*^{l,t}) + (\mathcal{S}^{l,t} - \mathcal{S}_*^{l,t}) = \mathcal{S}^{l,t}$ . For  $s, j \in \mathcal{S}_*^{l,t}$ , by (B.211) and (B.212), we have

$$\|\mathbf{W}_V^{(t)} \mathbf{x}_s - \mathbf{W}_V^{(t)} \mathbf{x}_j^n\| = 0. \tag{B.172}$$

Combining (B.12), we can obtain

$$|P_1| \leq \sigma \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\| \left\| \frac{|\mathcal{S}_1^{n,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} \left( \frac{|\mathcal{S}_1^{n,t}|}{|\mathcal{S}^{l,t}|} - \frac{|\mathcal{S}_1^{n,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} \right) \right\|. \tag{B.173}$$

Let

$$\begin{aligned}
T_1 = & \sum_{s \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z - \mathcal{S}_*^{l,t} - \mathcal{S}_\#^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l) \\
& \cdot \sum_{r \in \mathcal{S}^{l,t} - \mathcal{N}_z^l} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) - \sum_{s \in (\mathcal{S}^{l,t} - \mathcal{N}_z^l) - \mathcal{S}_*^{l,t} - \mathcal{S}_\#^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \\
& \cdot \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l) \sum_{r \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}),
\end{aligned} \tag{B.174}$$

$$\begin{aligned}
T_2 = & \sum_{s \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z \cap \mathcal{S}_{\#}^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \\
& \cdot \sum_{r \in (\mathcal{S}^{l,t} - \mathcal{N}_z^l) \cap \mathcal{S}_{\#}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) - \sum_{s \in (\mathcal{S}^{l,t} - \mathcal{N}_z^l) \cap \mathcal{S}_{\#}^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \\
& \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \sum_{r \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z \cap \mathcal{S}_{\#}^{l,t}} \\
& \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}), \\
T_3 = & \sum_{s \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z \cap \mathcal{S}_{\#}^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \sum_{r \in \mathcal{S}^{l,t} - \mathcal{N}_z^l - \mathcal{S}_{\#}^{l,t}} \\
& \cdot \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) - \sum_{s \in (\mathcal{S}^{l,t} - \mathcal{N}_z^l) \cap \mathcal{S}_{\#}^{l,t}} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \\
& \cdot \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \sum_{r \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z - \mathcal{S}_{\#}^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l \\
& + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}). 
\end{aligned} \tag{B.175}$$

Therefore,

$$P_2 = T_1 + T_2 + T_3, \tag{B.177}$$

$$T_1 \leq \sigma \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\| \cdot \frac{|\mathcal{N}_z^l - \mathcal{S}_*^{l,t} - \mathcal{S}_{\#}^{l,t}|}{|\mathcal{S}^{l,t}|} \frac{|\mathcal{S}^{l,t} - \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|}, \tag{B.178}$$

$$T_2 \leq \sigma \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\| \cdot \frac{|\mathcal{S}_{\#}^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} \left( \frac{|\mathcal{S}_{\#}^{l,t}|}{|\mathcal{S}^{l,t}|} - \frac{|\mathcal{S}_{\#}^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} \right). \tag{B.179}$$

For  $y_n = 1$ ,  $s \in \mathcal{S}_*^{l,t}$  and  $j \in \mathcal{S}_{\#}^{l,t}$ , by (B.211), (B.212), and (B.213), we have

$$\begin{aligned}
& \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}(\mathbf{W}_V^{(t)} \mathbf{x}_s - \mathbf{W}_V^{(t)} \mathbf{x}_j)\| \\
= & \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}(\mathbf{p}_1 - \mathbf{p}_2 + \mathbf{z}(t) - (\eta \sum_{b=1}^t \sum_{i \in \mathcal{W}_m(b)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)}{}^\top - \eta \sum_{b=1}^t \sum_{i \in \mathcal{U}(b)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)}{}^\top)) \\
& - (\eta \sum_{b=1}^t \sum_{i \notin \mathcal{W}_n(b)} \lambda V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)}{}^\top - \eta \sum_{b=1}^t \sum_{i \notin \mathcal{U}(b)} \lambda V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)}{}^\top))\| \\
\gtrsim & \left( \frac{1-2\epsilon_0}{B} \sum_{n \in \mathcal{B}_b} \frac{\xi \eta (t+1)^2 m}{a^2} \left( \frac{1}{4B} \sum_{n \in \mathcal{B}_b} p_n(b) - \sigma \right) \right. \\
& \left. + \eta m \frac{1}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)}{a} (1-\sigma) \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{(1-2\epsilon_0)\eta(t+1)^2}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) \right)^2 \right). 
\end{aligned} \tag{B.180}$$

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}(\mathbf{W}_V^{(t)}\mathbf{x}_s - \mathbf{W}_V^{(t)}\mathbf{x}_j^n)\| \lesssim \frac{\xi\eta t^2 m}{a^2} + \frac{\eta tm}{a} \cdot \left(\frac{\xi\eta t^2 m}{a^2}\right)^2. \quad (\text{B.181})$$

Given  $i \in \mathcal{W}_l(0)$ , with regard to  $P_2$ , we first consider the case when  $t = 0$ . Then with probability at least  $1 - |\mathcal{S}^{n,t}|^{-C} \geq 1 - (MZ)^{-C'}$  for  $C, C' > 0$ , when  $z = z_m$ ,

$$\begin{aligned} & \left| \frac{1}{|\mathcal{S}^{l,t}|} \sum_{j=1}^N \mathbb{1}[j \in \mathcal{D}_i \cap \mathcal{N}_z^l \cap \mathcal{S}^{l,t}] - \mathbb{E}\left[\frac{1}{|\mathcal{S}^{l,t}|} \sum_{j=1}^N \mathbb{1}[j \in \mathcal{D}_i \cap \mathcal{N}_z^l \cap \mathcal{S}^{l,t}]\right] \right| \\ &= \left| \frac{|\mathcal{S}_i^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} - \frac{|\mathcal{D}_i \cap \mathcal{N}_z^l|}{N} \right| \leq \sqrt{\frac{\log |\mathcal{S}^{n,t}|}{|\mathcal{S}^{n,t}|}} \leq \frac{1}{\text{poly}(Z)}, \end{aligned} \quad (\text{B.182})$$

$$\begin{aligned} & \left| \frac{1}{|\mathcal{S}^{l,t}|} \sum_{j=1}^N \mathbb{1}[j \in \mathcal{S}^{l,t} \cap \mathcal{N}_z^l] - \mathbb{E}\left[\frac{1}{|\mathcal{S}^{l,t}|} \sum_{j=1}^N \mathbb{1}[j \in \mathcal{S}^{l,t} \cap \mathcal{N}_z^l]\right] \right| \\ &= \left| \frac{|\mathcal{S}^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} - \frac{|\mathcal{N}_z^l|}{N} \right| \leq \sqrt{\frac{\log |\mathcal{S}^{n,t}|}{|\mathcal{S}^{n,t}|}} \leq \frac{1}{\text{poly}(Z)}. \end{aligned} \quad (\text{B.183})$$

For  $z = z_m$ , if  $y_l = 1$

$$\frac{|(\mathcal{D}_1 \cup \mathcal{D}_2) \cap \mathcal{N}_z^l|}{|(\mathcal{D}_1 \cup \mathcal{D}_2)|} = \frac{|\mathcal{N}_z^l|}{N} \leq \frac{|\mathcal{D}_1 \cap \mathcal{N}_z^l|}{|\mathcal{D}_1|}. \quad (\text{B.184})$$

Therefore, we have

$$\frac{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}_*^{l,t}|} = \frac{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l| + |\mathcal{S}_*^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|} \geq \frac{|\mathcal{N}_z^l|}{|\mathcal{N}_z^l| + |\mathcal{V} - \mathcal{N}_z^l|} - \frac{1}{\text{poly}(Z)}. \quad (\text{B.185})$$

For  $i = 3, 4, \dots, M$ , when  $z = z_m$ , we can derive

$$\frac{|\mathcal{S}_*^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l|} \leq \frac{|\mathcal{V} - \mathcal{N}_z^l|}{|\mathcal{N}_z^l|} + \frac{\Theta(1)}{\text{poly}(Z)} \leq \frac{|\mathcal{S}_i^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}_i^{l,t} \cap \mathcal{N}_z^l|} + \frac{\Theta(1)}{\text{poly}(Z)}, \quad (\text{B.186})$$

$$\frac{|\mathcal{S}_i^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}_i^{l,t} \cap \mathcal{N}_z^l|} \leq \frac{|\mathcal{S}_\#^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}_\#^{l,t} \cap \mathcal{N}_z^l|} + \frac{\Theta(1)}{\text{poly}(Z)}. \quad (\text{B.187})$$

Hence, we have

$$\frac{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} \frac{|\mathcal{S}_i^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}^{l,t}|} - \frac{|\mathcal{S}_i^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} \frac{|\mathcal{S}_*^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}^{l,t}|} \geq -\frac{1}{\text{poly}(z)}, \quad (\text{B.188})$$

$$\frac{|\mathcal{S}_i^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} \frac{|\mathcal{S}_\#^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}^{l,t}|} - \frac{|\mathcal{S}_\#^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} \frac{|\mathcal{S}_i^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}^{l,t}|} \geq -\frac{1}{\text{poly}(z)}. \quad (\text{B.189})$$

Then take the case where  $\mu_1$  is the class-relevant pattern as an example, we have

$$\begin{aligned}
& P_3 + T_3 \\
& \gtrsim \left( (1 - \sigma)^2 \cdot \frac{|\mathcal{S}_1^{l,t} \cap \mathcal{N}_l^z|}{|\mathcal{S}^{l,t}|e} \cdot \frac{|(\mathcal{S}^{l,t} - \mathcal{N}_l^z) \cap \mathcal{S}_2^{l,t}|}{|\mathcal{S}^{l,t}|e} - (1 + \sigma)^2 \cdot \frac{|(\mathcal{S}^{l,t} - \mathcal{N}_z^l) \cap \mathcal{S}_1^n|}{|\mathcal{S}^{l,t}|e} \right. \\
& \quad \left. \cdot \frac{|\mathcal{S}^{l,t} \cap \mathcal{N}_l^z \cap \mathcal{S}_2^{l,t}|}{|\mathcal{S}^{l,t}|e} \right) \cdot \eta \frac{(1 - 2\epsilon_0)^3}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m}{a} \cdot \left( \frac{\xi \eta t^2 m}{a^2} \right)^2 \|\mathbf{p}_1\|^2 + T_4 \quad (\text{B.190}) \\
& \gtrsim (1 - \sigma)^2 \cdot \frac{|\mathcal{S}_1^{l,t}|}{|\mathcal{S}^{l,t}|} \frac{|\mathcal{S}_1^{l,t} \cap \mathcal{N}_z^l| - |\mathcal{S}_2^{l,t} \cap \mathcal{N}_z^l|}{|\mathcal{S}^{l,t}|} \cdot \eta \frac{(1 - 2\epsilon_0)^3}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m}{a} \\
& \quad \left( \frac{\xi \eta t^2 m}{a^2} \right)^2 \|\mathbf{p}_1\|^2,
\end{aligned}$$

given that

$$\begin{aligned}
T_4 := & \sum_{s \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z \cap (\mathcal{S}_{\#}^{l,t} \cup \mathcal{S}_*^{l,t})} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \\
& \cdot \sum_{r \in \mathcal{S}^{l,t} - \mathcal{N}_z^l - \mathcal{S}_{\#}^{l,t} - \mathcal{S}_*^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}) \\
& - \sum_{s \in (\mathcal{S}^{l,t} - \mathcal{N}_z^l) \cap (\mathcal{S}_{\#}^{l,t} \cup \mathcal{S}_*^{l,t})} \mathbf{W}_{O_{(i,\cdot)}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \\
& \cdot \sum_{r \in \mathcal{S}^{l,t} \cap \mathcal{N}_l^z - \mathcal{S}_{\#}^{l,t} - \mathcal{S}_*^{l,t}} \text{softmax}_l(\mathbf{x}_r^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(r,l)}^\top \mathbf{b}^{(t)}), \quad (\text{B.191})
\end{aligned}$$

and

$$|T_4| \leq \sigma \frac{\eta^3 t^5 m^3 \xi^2}{a^5} \|\mathbf{p}\| \cdot \frac{1}{\text{poly}(Z)}. \quad (\text{B.192})$$

One can obtain the opposite conclusion if

$$\begin{aligned}
& \frac{|\mathcal{S}_*^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l|} \geq \frac{|\mathcal{V} - \mathcal{N}_z^l|}{|\mathcal{N}_z^l|} + \frac{\Theta(1)}{\text{poly}(Z)} \geq \frac{|\mathcal{S}_i^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}_i^{l,t} \cap \mathcal{N}_z^l|} + \frac{\Theta(1)}{\text{poly}(Z)} \\
& \geq \frac{|\mathcal{S}_{\#}^{l,t} \cap (\mathcal{V} - \mathcal{N}_z^l)|}{|\mathcal{S}_{\#}^{l,t} \cap \mathcal{N}_z^l|} + \frac{\Theta(1)}{\text{poly}(Z)}. \quad (\text{B.193})
\end{aligned}$$

We can conclude that  $b_z$  will increase during the updates with the condition (B.186) and

decrease with the condition (B.193). When  $t$  is large, given that  $|\mathcal{N}_z^l| = \Theta(|\mathcal{S}^{l,t}|)$ , define

$$K := \max_{z \in \mathcal{Z}} \{b_z^{(t)}\} - \min_{z \in \mathcal{Z}} \{b_z^{(t)}\}. \quad (\text{B.194})$$

Therefore,

$$\begin{aligned} & P_3 + T_3 \\ & \gtrsim \left( (1 - \sigma)^2 \frac{K |\mathcal{S}_1^{l,t} \cap \mathcal{N}_z^l| \cdot |(\mathcal{S}^{l,t} - \mathcal{N}_z^l) \cap \mathcal{S}_2^{l,t}|}{(K |\mathcal{S}^{l,t} \cap \mathcal{N}_z^l| + |\mathcal{S}^{l,t} - \mathcal{N}_z^l|)^2} - (1 + \sigma)^2 \right. \\ & \quad \cdot \left. \frac{|(\mathcal{S}^{l,t} - \mathcal{N}_z^l) \cap \mathcal{S}_1^n| \cdot K |\mathcal{S}^{l,t} \cap \mathcal{N}_z^l \cap \mathcal{S}_2^{l,t}|}{(K |\mathcal{S}^{l,t} \cap \mathcal{N}_z^l| + |\mathcal{S}^{l,t} - \mathcal{N}_z^l|)^2} \right) \\ & \quad \cdot \eta \frac{(1 - 2\epsilon_0)^3}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m}{a} \left( \frac{\xi \eta t^2 m}{a^2} \right)^2 \|\mathbf{p}_1\|^2 \\ & \gtrsim (1 - \sigma)^2 \cdot \frac{K |\mathcal{S}_1^{l,t}| \cdot (|\mathcal{S}_1^{l,t} \cap \mathcal{N}_z^l| - |\mathcal{S}_2^{l,t} \cap \mathcal{N}_z^l|)}{(K |\mathcal{S}^{l,t} \cap \mathcal{N}_z^l| + |\mathcal{S}^{l,t} - \mathcal{N}_z^l|)^2} \cdot \eta \frac{(1 - 2\epsilon_0)^3}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m}{a} \\ & \quad \left( \frac{\xi \eta t^2 m}{a^2} \right)^2 \|\mathbf{p}_1\|^2 \\ & \gtrsim (1 - \sigma)^2 \cdot \frac{|\mathcal{S}_1^{l,t}|}{|\mathcal{S}^{l,t}|} \frac{|\mathcal{S}_1^{l,t} \cap \mathcal{N}_z^l| - |\mathcal{S}_2^{l,t} \cap \mathcal{N}_z^l|}{K |\mathcal{S}^{l,t}|} \cdot \eta \frac{(1 - 2\epsilon_0)^3}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m}{a} \\ & \quad \cdot \left( \frac{\xi \eta t^2 m}{a^2} \right)^2 \|\mathbf{p}_1\|^2. \end{aligned} \quad (\text{B.195})$$

By combining (B.173), (B.178), (B.179), and B.195, we can derive,

$$\begin{aligned} & -\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \mathbf{Loss}(\mathbf{X}^n, y_n)}{\partial b_z} \\ & \gtrsim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \eta \frac{(1 - 2\epsilon_0)^3}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m^2}{a^2} \left( \frac{\xi \eta t^2 m}{a^2} \right)^2 \|\mathbf{p}_1\|^2 \cdot \frac{|\mathcal{S}_*^{l,t}|}{|\mathcal{S}^{l,t}|} \\ & \quad \frac{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l| - |\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l|}{K |\mathcal{S}^{l,t}|}. \end{aligned} \quad (\text{B.196})$$

If  $u_{(s,l)z^*} = 1$ ,

$$\mathbf{u}_{(s,l)}^\top (\mathbf{b}^{(t+1)} - \mathbf{b}^{(t)}) = -\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \mathbf{Loss}(\mathbf{X}^n, y_n)}{\partial b_{z^*}}, \quad (\text{B.197})$$

$$\begin{aligned}
& \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)} \\
& \geq \eta \frac{1}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \eta \frac{(1-2\epsilon_0)^3}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m^2}{a^2} \left( \frac{\xi \eta t^2 m}{a^2} \right)^2 \|\mathbf{p}_1\|^2 \\
& \cdot \frac{\gamma_d}{2} \frac{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l| - |\mathcal{S}_\#^{l,t} \cap \mathcal{N}_z^l|}{K |\mathcal{S}^{l,t}|}.
\end{aligned} \tag{B.198}$$

If we want to compute the difference term  $b_{z_m}^{(t)} - b_z^{(t)}$ , note that we only need to study the differences in  $P_3 + T_3$  given the previous analysis. Since that the term  $\mathbf{W}_{O_{(i,:)}} \mathbf{W}_V \mathbf{x}_s \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)})$  is larger when  $s \in \mathcal{N}_{z_m}^l$ , we can bound the difference  $P_3 + T_3$  using terms in (B.195). To find the lower bound, we apply the result in (B.180) and then directly use the fraction of sampled nodes in different neighborhoods because concentration bounds can control the error. To be more specific, on the one hand, if  $\mathcal{N}_z^l$  is too small for one  $z \in [Z-1]$  and  $l \in \mathcal{V}$ , the left-hand side of (B.188), (B.189), and (B.190) are close to zero, and these three equations still hold. On the other hand, if we want to see whether terms (B.188) and (B.189) with  $z = z_m$  are larger than them with other  $z \neq z_m$ , we have the following derivation. Take (B.188) as an example,

$$\begin{aligned}
& |\mathcal{D}_*^l \cap \mathcal{N}_z^l| |\mathcal{D}_i \cap (\mathcal{V} - \mathcal{N}_z^l)| - |\mathcal{D}_i \cap \mathcal{N}_z^l| |\mathcal{D}_*^l \cap (\mathcal{V} - \mathcal{N}_z^l)| \\
& = |\mathcal{D}_*^l \cap \mathcal{N}_z^l| \cdot |\mathcal{D}_i| - |\mathcal{D}_i \cap \mathcal{N}_z^l| \cdot |\mathcal{D}_*^l|.
\end{aligned} \tag{B.199}$$

$$\begin{aligned}
& |\mathcal{D}_*^l \cap \mathcal{N}_{z_m}^l| |\mathcal{D}_i \cap (\mathcal{V} - \mathcal{N}_{z_m}^l)| - |\mathcal{D}_i \cap \mathcal{N}_{z_m}^l| |\mathcal{D}_*^l \cap (\mathcal{V} - \mathcal{N}_{z_m}^l)| \\
& - (|\mathcal{D}_*^l \cap \mathcal{N}_z^l| |\mathcal{D}_i \cap (\mathcal{V} - \mathcal{N}_z^l)| - |\mathcal{D}_i \cap \mathcal{N}_z^l| |\mathcal{D}_*^l \cap (\mathcal{V} - \mathcal{N}_z^l)|) \\
& = (|\mathcal{D}_*^l \cap \mathcal{N}_{z_m}^l| - |\mathcal{D}_*^l \cap \mathcal{N}_z^l|) \cdot |\mathcal{D}_i| - (|\mathcal{D}_i \cap \mathcal{N}_{z_m}^l| - |\mathcal{D}_i \cap \mathcal{N}_z^l|) \cdot |\mathcal{D}_*^l| \\
& = (|\mathcal{N}_{z_m}^l| - |\mathcal{N}_z^l|) \cdot \frac{\gamma_d}{2} |\mathcal{D}_i| - (|\mathcal{D}_i \cap \mathcal{N}_{z_m}^l| - |\mathcal{D}_i \cap \mathcal{N}_z^l|) \cdot |\mathcal{D}_*^l| \\
& + (|\mathcal{D}_*^l \cap \mathcal{N}_{z_m}^l| - |\mathcal{N}_{z_m}^l| \frac{\gamma_d}{2} - (|\mathcal{D}_*^l \cap \mathcal{N}_z^l| - |\mathcal{N}_z^l| \frac{\gamma_d}{2})) \cdot |\mathcal{D}_i| \\
& = \frac{1}{2} (|\mathcal{D}_*^l \cap \mathcal{N}_{z_m}^l| - |\mathcal{D}_\#^l \cap \mathcal{N}_{z_m}^l| - (|\mathcal{D}_*^l \cap \mathcal{N}_z^l| - |\mathcal{D}_\#^l \cap \mathcal{N}_z^l|)) \cdot |\mathcal{D}_i| \\
& \geq 0,
\end{aligned} \tag{B.200}$$

where the first step is by (B.199), the second step comes from mathematical derivation, the third step is obtained from that  $\mu_i$ ,  $i = 2, 3, \dots, M$  is uniformly distributed in the whole graph, and the last step is by the definition of  $z_m$  in (3.6). We can derive (B.189) in the same

way. Hence,

$$\begin{aligned}
& b_{z_m}^{(t)} - b_z^{(t)} \\
& \gtrsim \eta \frac{1}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \eta \frac{(1-2\epsilon_0)^3}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m^2}{a^2} \left( \frac{\xi \eta t^2 m}{a^2} \right)^2 \|\mathbf{p}_1\|^2 \cdot \frac{\gamma_d}{2} \\
& \quad \cdot \left( \frac{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_{z_m}^l| - |\mathcal{S}_\#^{l,t} \cap \mathcal{N}_{z_m}^l|}{K|\mathcal{S}^{l,t}|} - \frac{|\mathcal{S}_*^{l,t} \cap \mathcal{N}_z^l| - |\mathcal{S}_\#^{l,t} \cap \mathcal{N}_z^l|}{K|\mathcal{S}^{l,t}|} \right).
\end{aligned} \tag{B.201}$$

Note that finally  $\eta T = \Theta(1)$ . Therefore,  $K = \Theta(1)$ .

### Proof of Lemma B.3.4:

For the gradient of  $\mathbf{W}_V$ ,

$$\begin{aligned}
\frac{\partial \overline{\text{Loss}}_b}{\partial \mathbf{W}_V} &= \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y_n)}{\partial F(\mathbf{X}^n)} \frac{\partial F(\mathbf{X}^n)}{\partial \mathbf{W}_V} \\
&= \frac{1}{B} \sum_{n \in \mathcal{B}_b} \sum_{i=1}^m (-y_i) a_i \mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}} \sum_{s \in \mathcal{S}^{l,t}} \mathbf{W}_V \mathbf{x}_l \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l \\
&\quad + \mathbf{u}_{(s,l)}^\top \mathbf{b}) \geq 0] \mathbf{W}_{O_{(i,\cdot)}}^\top \sum_{s \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b})^\top \mathbf{x}_s^\top. 
\end{aligned} \tag{B.202}$$

Consider a node  $n$  where  $y_n = 1$ . Let  $l \in \mathcal{S}_1^{n,t}$

$$\sum_{s \in \mathcal{S}_1^{n,t}} \text{softmax}_n(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b}^{(t)}) \geq p_n(t). \tag{B.203}$$

Then for  $j \in \mathcal{S}_1^{g,t}$ ,  $g \in \mathcal{V}$ ,

$$\begin{aligned}
& \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \text{Loss}(\mathbf{X}^n, y_n)}{\partial \mathbf{W}_V^{(t)}} \Big| \mathbf{W}_V^{(t)} \mathbf{x}_j \\
&= \frac{1}{B} \sum_{l \in \mathcal{B}_b} (-y_l) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{S}^{l,t}} \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \\
&\quad \mathbf{W}_V^{(t)} \mathbf{x}_s \geq 0] \mathbf{W}_{O_{(i,\cdot)}}^{(t)\top} \sum_{s \in \mathcal{S}^{n,t}} \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{x}_s^\top \mathbf{x}_j \\
&= \Theta(1) \cdot \left( \sum_{i \in \mathcal{W}_l(0)} V_i(t) \mathbf{W}_{O_{(i,\cdot)}}^\top + \sum_{i \notin \mathcal{W}_l(0)} \lambda V_i(t) \mathbf{W}_{O_{(i,\cdot)}}^\top \right),
\end{aligned} \tag{B.204}$$

If  $i \in \mathcal{W}_l(0)$ , we have

$$V_i(t) \lesssim \frac{1-2\epsilon_0}{2B} \sum_{n \in \mathcal{B}_{b+}} -\frac{1}{a} p_n(t). \quad (\text{B.205})$$

Similarly, if  $i \in \mathcal{U}_l(t)$ ,

$$V_i(t) \gtrsim \frac{1-2\epsilon_0}{2B} \sum_{n \in \mathcal{B}_{b-}} \frac{1}{a} p_n(t), \quad (\text{B.206})$$

if  $i$  is an unlucky neuron, by hoeffding's inequality in (B.6), we have

$$|V_i(t)| \lesssim \frac{1}{\sqrt{B}} \cdot \frac{1}{a}. \quad (\text{B.207})$$

Therefore, we can derive

$$\begin{aligned} & -\eta \sum_{b=1}^t \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \sum_{j \in \mathcal{W}_l(0)} V_j(b) \mathbf{W}_{O_{(j,\cdot)}}^{(b)} {}^\top \\ & \gtrsim \eta m \frac{1-2\epsilon_0}{2B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_{b+}} \frac{1}{a} p_n(b) \cdot \left( \frac{\xi}{aB} \sum_{n \in \mathcal{B}_b} \frac{\eta t^2 (1-2\epsilon_0)}{4B} \sum_{n \in \mathcal{B}_b} \frac{m}{a} p_n(t) \right)^2, \end{aligned} \quad (\text{B.208})$$

$$\begin{aligned} & |\eta \sum_{b=1}^t \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \sum_{j \in \mathcal{U}_{l,n}(0)} V_j(b) \mathbf{W}_{O_{(j,\cdot)}}^{(b)} {}^\top| \\ & \lesssim \frac{\eta}{B} \sum_{b=1}^t \sum_{n \in \mathcal{B}_b} \frac{p_n(b)m}{a} \|\mathbf{W}_{O_{(i,\cdot)}}^{(b)}\|^2 \|\mathbf{p}_1\|^2, \end{aligned} \quad (\text{B.209})$$

$$-\eta t \mathbf{W}_{O_{(i,\cdot)}} \sum_{j \notin (\mathcal{W}_l(0) \cup \mathcal{U}_{l,n}(0))} V_j(t) \mathbf{W}_{O_{(j,\cdot)}} {}^\top \lesssim \frac{\eta tm \|\mathbf{p}\|^2}{Ba} \|\mathbf{W}_{O_{(i,\cdot)}}^{(t)}\|^2. \quad (\text{B.210})$$

Hence,

(1) If  $j \in \mathcal{S}_1^{n,t}$  for one  $n \in \mathcal{V}$ ,

$$\begin{aligned} \mathbf{W}_V^{(t+1)} \mathbf{x}_j^n &= \mathbf{W}_V^{(t)} \mathbf{x}_j^n - \eta \left( \frac{\partial \text{Loss}(\mathbf{X}^n, y_n)}{\partial \mathbf{W}_V} \Big| \mathbf{W}_V^{(t)} \right) \mathbf{x}_j^n \\ &= \mathbf{p}_1 - \eta \sum_{b=1}^{t+1} \sum_{i \in \mathcal{W}_{(n)b}} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} {}^\top - \eta \sum_{b=1}^{t+1} \sum_{i \notin \mathcal{W}_n(b)} \lambda V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} {}^\top + \mathbf{z}_j(t). \end{aligned} \quad (\text{B.211})$$

(2) If  $j \in \mathcal{S}_2^{n,t}$ , we have

$$\begin{aligned} \mathbf{W}_V^{(t+1)} \mathbf{x}_j &= \mathbf{W}_V^{(0)} \mathbf{x}_j^n - \eta \left( \frac{\partial \text{Loss}(\mathbf{X}^n, y_n)}{\partial \mathbf{W}_V} \Big| \mathbf{W}_V^{(0)} \right) \mathbf{x}_j^n \\ &= \mathbf{p}_2 - \eta \sum_{b=1}^{t+1} \sum_{i \in \mathcal{U}(b)} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)\top} - \eta \sum_{b=1}^{t+1} \sum_{i \notin \mathcal{U}(b)} \lambda V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)\top} + \mathbf{z}_j(t). \end{aligned} \quad (\text{B.212})$$

(3) If  $j \in \mathcal{S}^{n,t}/(\mathcal{S}_1^{n,t} \cup \mathcal{S}_2^{n,t})$ , we have

$$\begin{aligned} \mathbf{W}_V^{(t+1)} \mathbf{x}_j^n &= \mathbf{W}_V^{(0)} \mathbf{x}_j^n - \eta \left( \frac{\partial \text{Loss}(\mathbf{X}^n, y_n)}{\partial \mathbf{W}_V} \Big| \mathbf{W}_V^{(0)} \right) \mathbf{x}_j^n \\ &= \mathbf{p}_k - \eta \sum_{b=1}^{t+1} \sum_{i=1}^m \lambda V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)\top} + \mathbf{z}_j(t). \end{aligned} \quad (\text{B.213})$$

Here

$$\|\mathbf{z}_j(t)\| \leq \sigma. \quad (\text{B.214})$$

for  $t \geq 1$ . Note that this Lemma also holds when  $t = 1$ .

### Proof of Lemma B.3.6:

We prove this lemma by induction.

When  $t = 0$ . For  $i \in \mathcal{W}_l(0)$  and  $l \in \mathcal{D}_1$ , we have that

$$\mathbf{W}_{O_{(i,:)}}^{(0)} \left( \sum_{s \in \mathcal{S}_1^{l,t}} \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + 0) \mathbf{p}_1 + \mathbf{z}(0) + \sum_{j \neq 1} W_j^n(0) \mathbf{p}_j \right) \gtrsim \xi(\Theta(1) - \sigma) > 0. \quad (\text{B.215})$$

$$\gtrsim \xi(\Theta(1) - \sigma) > 0.$$

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

$$\begin{aligned} &\mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}_l^n(t) \\ &= \mathbf{W}_{O_{(i,:)}}^{(t)} \left( \sum_{s \in \mathcal{S}_1^n} \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{p}_1 + \mathbf{z}(t) + \sum_{j \neq 1} W_j^n(t) \mathbf{p}_j \right. \\ &\quad \left. - \eta \sum_{b=0}^{t-1} \left( \sum_{i \in \mathcal{W}_l(0)} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)\top} + \sum_{i \notin \mathcal{W}_l(0)} V_i(b) \lambda \mathbf{W}_{O_{(i,:)}}^{(b)\top} \right) \right). \end{aligned} \quad (\text{B.216})$$

Denote  $\theta_l^i$  as the angle between  $\mathbf{V}_l(0)$  and  $\mathbf{W}_{O_{(i,:)}}^{(0)}$ . Since that  $\mathbf{W}_{O_{(j,:)}}^{(0)}$  is initialized uniformed

on the  $m_a - 1$ -sphere, we have  $\mathbb{E}[\theta_l^i] = 0$ . By hoeffding's inequality (B.6), we have

$$\left\| \frac{1}{|\mathcal{W}_l(0)|} \sum_{i \in \mathcal{W}_l(0)} \theta_l^i - \mathbb{E}[\theta_l^i] \right\| = \left\| \frac{1}{|\mathcal{W}_l(0)|} \sum_{i \in \mathcal{W}_l(0)} \theta_l^i \right\| \leq \sqrt{\frac{\log N}{m}}, \quad (\text{B.217})$$

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

$$\left\| \frac{1}{|\mathcal{W}_l(0)|} \sum_{i \in \mathcal{W}_l(0)} \theta_l^i - \mathbb{E}[\theta_l^i] \right\| \leq O\left(\frac{1}{M}\right). \quad (\text{B.218})$$

Therefore, for  $i \in \mathcal{W}_l(0)$ , we have

$$\mathbf{W}_{O_{(i,\cdot)}} \sum_{b=0}^{t-1} \sum_{i \in \mathcal{W}_l(0)} \mathbf{W}_{O_{(i,\cdot)}}^{(b)} > 0. \quad (\text{B.219})$$

Similarly, we have that  $\sum_{b=0}^{t-1} \sum_{i \notin \mathcal{W}_l(0)} \mathbf{W}_{O_{(i,\cdot)}}^{(b)}$  is close to  $-\mathbf{V}_l^n(0)$ . Given that  $\lambda < 1$ , we can approximately acquire that

$$-\mathbf{W}_{O_{(i,\cdot)}}^{(0)} \eta \sum_{b=0}^{t-1} \left( \sum_{i \in \mathcal{W}_l(0)} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)}{}^\top + \sum_{i \notin \mathcal{W}_l(0)} V_i(b) \lambda \mathbf{W}_{O_{(i,\cdot)}}^{(b)}{}^\top \right) > 0. \quad (\text{B.220})$$

After the first iterations, we know that  $\mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}$  increases the most from  $\mathbf{u}_{(s,l)}^\top \mathbf{b}^{(0)}$  by  $\gamma_d$  fraction of discriminative nodes if  $s \in \mathcal{N}_{z_m}^l$ . Because the softmax is based exponential functions, the most significance increase in  $\mathcal{N}_{z_m}^l$  enlarges  $\sum_{s \in \mathcal{S}_1^{l,t}} \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)})$ . Since that  $i \in \mathcal{W}_n(0)$ , we then have

$$\mathbf{W}_{O_{(i,\cdot)}}^{(0)} \left( \sum_{s \in \mathcal{S}_1^n} \text{softmax}_l(\mathbf{x}_s^\top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{x}_l + \mathbf{u}_{(s,l)}^\top \mathbf{b}^{(t)}) \mathbf{p}_1 + \mathbf{z}(t) + \sum_{j \neq 1} W_j^n(t) \mathbf{p}_j \right) > 0. \quad (\text{B.221})$$

Therefore, we have

$$\mathbf{W}_{O_{(i,\cdot)}}^{(0)} \mathbf{V}_l^n(t) > 0. \quad (\text{B.222})$$

Meanwhile, the addition from  $\mathbf{W}_{O_{(i,\cdot)}}^{(0)}$  to  $\mathbf{W}_{O_{(i,\cdot)}}^{(1)}$  is approximately a summation of multiple  $\mathbf{V}_j(0)$  such that  $\mathbf{W}_{O_{(i,\cdot)}}^{(0)} \mathbf{V}_j(0) > 0$  and  $j \in \mathcal{S}_1^n$ . Therefore,  $\mathbf{V}_j(0)^\top \mathbf{V}_l(0) > 0$ . Therefore, we can obtain

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{V}_l(t) > 0. \quad (\text{B.223})$$

(2) Suppose that the conclusion holds when  $t = s$ . When  $t = s + 1$ , we can follow the derivation of the case where  $t = 1$ . Although the unit vector of  $\mathbf{W}_{O_{(i,:)}}^{(t)}$  no longer follows a uniform distribution, we know that (B.217) holds since the angle is bounded and has a mean which is very close to  $\mathbf{V}_l(0)$ . Then, the conclusion still holds.

One can develop the proof for  $\mathcal{U}_{l,n}(0)$  following the above steps.

## B.5 Extension of Our Analysis and Additional Discussion

### B.5.1 Assumption on the Pre-Trained Model

For the assumption on the pre-trained model, we provide the following discussion.

We would like to clarify that the training problem of the graph transformer (GT) is very challenging to analyze due to its significant non-convexity, and some form of assumptions is needed to facilitate the analysis. In fact, even for the conventional Transformers, the existing state-of-the-art theoretical optimization and generalization analyses all make some assumptions on the data, embedding or initial models, or make further simplifications on the Transformer model. For example, [115] assumes orthogonality on the raw data. [114] simplifies the self-attention layer by only considering the positional encoding (PE). [104], [116] use linear activation in the MLP layer. About the initialization, [6] assumes orthogonality on the initialization of embeddings. [119] requires that the initialization of the query embedding is close to the optimal solution.

The initialization assumption made in our manuscript is the same as [6] but for a GT. We would like to emphasize that our initialization assumption is at least no stronger than the existing initialization assumptions in [6] and [119]. Notably, our work proposes a novel theoretical framework for the training dynamics and generalization of GT for the first time, where the number of trainable parameters is more than the above existing works. Third, although we have assumptions on the initialization for theoretical analysis, our experiments on real-world datasets in Section 3.5.2 are implemented from random initialization. The performance is aligned with our theoretical findings.

### B.5.2 Extension to Other Positional Encodings

Our theoretical analysis is general and can be applied to different positional encodings. Specifically, Theorem 3.4.4 is based on proving these two parts, (i) the success of positional encoding, i.e., the positional encoding can identify the correct structure information (which is

the core neighborhood in our data model), (ii) if structural information is known, analyzing the sample complexity and convergence rate of Graph Transformer. We next discuss the extension of both parts to other positional encoding separately.

For part (i), the success of positional encoding, because different types of positional encoding can learn different types of structure information the best, this analysis needs to be case-by-case for different positional encoding. However, our technique and insight can be potentially useful with some modifications to other positional encodings. For example, Laplacian eigenvectors can essentially divide a graph into several clusters considering its relationship to spectral clustering [263] and would work best for a data model where data labels depend on clusters. Moreover, Random Walk PE can encode structural information such as whether the node is part of an m-long circle [20]. Degree PE [5], one of the standard centrality measures, can capture the local degree information. PE using distance from the centroid of the whole graph [20] can represent global distance information. Our techniques in analyzing the core neighborhood can be useful in analyzing these positional encodings. Similarly to our framework of the core neighborhood, where a large amount of class-relevant nodes is located, one can respectively construct data models for these positional encodings where class-relevant nodes are dominant for nodes within the corresponding structures, such as a cluster, an m-long circle, a certain degree, and a certain distance to the centroid of the whole graph. The remaining step of the generalization analysis is to learn this data model by Graph Transformer using positional encoding, which is elaborated in detail in the next paragraph.

For part (ii), our proof technique can be easily generalized to other positional encodings with some straightforward transformation. Specifically, positional encoding can be divided into absolute and relative positional encodings. What we study in this work belongs to relative positional encodings. Absolute positional encodings can be formulated as a concatenation to the initial node feature, either by their raw definition [4], [20] or by transferring from a bias term [21] ( $\mathbf{W}\mathbf{x} + \mathbf{a} = (\mathbf{W}, \mathbf{W}')(\mathbf{x}^\top, \mathbf{b}'^\top)^\top$ , where the trainable positional encoding  $\mathbf{a}$  is transferred into a fixed augmented feature  $\mathbf{b}'$  and a trainable augmented weight  $\mathbf{W}'$ ). The structural information is then incorporated into the node representation  $(\mathbf{x}, \mathbf{b}')$ . Denote the positional encoding  $\mathbf{b}'$  for a query node  $q$  as  $\mathbf{b}'_q$ . Denote the PE of one core-neighborhood node  $c$  and one other neighboring node  $o$  for this query node as  $\mathbf{b}'_c$ , and  $\mathbf{b}'_o$ , respectively. Suppose all the  $\mathbf{b}'$  are normalized. Then, given that the defined positional encoding  $\mathbf{b}'$  can locate the

core neighborhood, i.e., the distance between  $\mathbf{b}'_c$  and  $\mathbf{b}'_q$  is much smaller than the distance between  $\mathbf{b}'_o$  and  $\mathbf{b}'_q$ , we can deduce that the inner product between  $\mathbf{b}'_c$  and  $\mathbf{b}'_q$  is much larger than the inner product between  $\mathbf{b}'_o$  and  $\mathbf{b}'_q$ . This leads to a dominant attention weight between the query node  $q$  and the core-neighborhood node  $c$  based on the definition of self-attention. Then, one could ignore other neighbors and focus only on core-neighborhood nodes when computing the Graph Transformer output. Then the proof in Theorem 3.4.4 for part (ii) applies directly.

### B.5.3 Extension of the Analysis on GAT

From a high-level understanding, a one-layer GAT can be regarded as a Graph Transformer that only uses distance-1 neighborhood information. Therefore, our Theorem 3.4.8 can be applied to analyze the generalization of a one-layer GAT when its self-attention follows the self-attention mechanism in 3.1 of our manuscript, given the distance-1 neighborhood as the core neighborhood. From a perspective of training dynamics, GATs also share a common mechanism that computes the aggregation based on the similarity between node features as Graph Transformer does, although the attention layer of GAT [175] is different. In this sense, one-layer GAT can generalize as well as Graph Transformer if the graph satisfies that the latent core neighborhood is the distance-1/distance-small neighborhood, such as homophilous graphs. The generalization analysis of using GATs on graphs with a larger distance of core neighborhoods and its comparison with graph transformers needs more effort, and we will leave it as future work.

### B.5.4 Extension to Graph Classification Problems

Since we aim to make a comparison with GCN, which focuses more on node classification tasks, our work also mainly studies node classification. However, our analysis is extendable to graph classification tasks. Consider a supervised-learning binary classification problem on a set of graphs  $\{\mathcal{G}_i\}_{i=1}^N$ . Denote the feature matrix of the graph  $\mathcal{G}_i$  by  $\mathbf{X}_i$ . Following [5], [4], we apply ‘‘Mean’’ or ‘‘Sum’’ as the READOUT function. Hence, we have

$$F(\mathbf{X}_i) = K \sum_{n \in \mathcal{T}^i} \mathbf{a}_n^\top \text{Relu}(\mathbf{W}_O \sum_{s \in \mathcal{T}^i} \mathbf{W}_V \mathbf{x}_s \text{softmax}_n(\mathbf{x}_s^\top \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b})). \quad (\text{B.224})$$

where  $K = 1$  if READOUT is “Sum”, and  $K = 1/|\mathcal{T}_i|$  if READOUT is “Mean”. When we compute the gradients, the only difference is that we sum up or average over all nodes in each graph.

**Data Model** The data model follows from Section 3.4.2. The difference is that the core neighborhood is defined based on the graph label, i.e., we assume the ground truth graph label is determined by the summation/mean of the majority vote of  $\mu_1, \mu_2$  nodes in the core neighborhood for some nodes in each graph. This is motivated by graph classification on social networks, where the connections between the central person and other people in the graph decide the graph label. For example, if  $z_m = 2$  and the distance- $z_m$  neighborhood of nodes in  $\mathcal{R}^i$  determines the label, then for the ground truth graph label  $\tilde{y}_i = 1$ ,  $|\mathcal{D}_1^i \cap (\cup_{j \in \mathcal{R}^i} \mathcal{N}_{z_m}^j)|$  is larger than  $|\mathcal{D}_2^i \cap (\cup_{j \in \mathcal{R}^i} \mathcal{N}_{z_m}^j)|$ , where  $\mathcal{D}_1^i$  and  $\mathcal{D}_2^i$  are the set of  $\mu_1$  nodes and  $\mu_2$  nodes in  $\mathcal{G}_i$ . Such a data model ensures that the graph label comes from the graph structure. Meanwhile, it prevents us from assuming a more trivial model where the number of  $\mu_1$  nodes and  $\mu_2$  nodes in each graph indicates the label and no graph information is used, which is almost the same as that in the ViT work [6]. Hence, when we compute the graph-level output, the distance- $z_m$  neighborhood of nodes in  $\mathcal{R}^i$  still plays a vital role. Then, we can apply the generalization analysis of node classification based on the core neighborhood to the graph classification problem.

### B.5.5 Extension to Multi-Classification

Consider the classification problem with four classes. We use the label  $y \in \{+1, -1\}^2$  to denote the corresponding class. Similarly to the previous setup, there are four orthogonal discriminative patterns. We have  $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2)$ ,  $\mathbf{W}_O = (\mathbf{W}_{O_1}, \mathbf{W}_{O_2})$ ,  $\mathbf{W}_V = (\mathbf{W}_{V_1}, \mathbf{W}_{V_2})$ ,  $\mathbf{W}_K = (\mathbf{W}_{K_1}, \mathbf{W}_{K_2})$ ,  $\mathbf{W}_Q = (\mathbf{W}_{Q_1}, \mathbf{W}_{Q_2})$ , and  $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2)$ . Hence, we define

$$\mathbf{F}(\mathbf{x}_n) = (F_1(\mathbf{x}_n), F_2(\mathbf{x}_n)), \quad (\text{B.225})$$

$$F_1(\mathbf{x}_n) = \mathbf{a}_1^\top \text{Relu}(\mathbf{W}_{O_1} \sum_{s \in \mathcal{T}_1^n} \mathbf{W}_{V_1} \mathbf{x}_s \text{softmax}_n(\mathbf{x}_s^\top \mathbf{W}_{K_1}^\top \mathbf{W}_{Q_1} \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b}_1)), \quad (\text{B.226})$$

$$F_2(\mathbf{x}_n) = \mathbf{a}_2^\top \text{Relu}(\mathbf{W}_{O_2} \sum_{s \in \mathcal{T}_2^n} \mathbf{W}_{V_2} \mathbf{x}_s \text{softmax}_n(\mathbf{x}_s^\top \mathbf{W}_{K_2}^\top \mathbf{W}_{Q_2} \mathbf{x}_n + \mathbf{u}_{(s,n)}^\top \mathbf{b}_2)). \quad (\text{B.227})$$

The dataset  $\mathcal{D}$  can be divided into four groups as

$$\mathcal{A}_1 = \{(\mathbf{X}^n, \mathbf{y}_n) | \mathbf{y}_n = (1, 1)\}, \quad (\text{B.228})$$

$$\mathcal{A}_2 = \{(\mathbf{X}^n, \mathbf{y}_n) | \mathbf{y}_n = (1, -1)\}, \quad (\text{B.229})$$

$$\mathcal{A}_3 = \{(\mathbf{X}^n, \mathbf{y}_n) | \mathbf{y}_n = (-1, 1)\}, \quad (\text{B.230})$$

$$\mathcal{A}_4 = \{(\mathbf{X}^n, \mathbf{y}_n) | \mathbf{y}_n = (-1, -1)\}. \quad (\text{B.231})$$

The hinge loss function for data  $(\mathbf{X}^n, \mathbf{y}_n)$  will be

$$\text{Loss}(\mathbf{x}_n, \mathbf{y}_n) = \max\{1 - \mathbf{y}_n^\top \mathbf{F}(\mathbf{x}_n), 0\}. \quad (\text{B.232})$$

Therefore, when computing the gradient, the problem becomes a binary classification. One can make derivations following the binary case. One notable difference is that we can assume two core neighborhoods for this four-classification problem.

### B.5.6 Comparision with Other Frameworks of Analysis

In this section, we provide a comparison with other frameworks of analysis.

First, we focus on five other frameworks: Rademacher complexity, algorithmic stability, PAC-Bayesian, model recovery, and neural tangent kernel (NTK). Rademacher complexity [264], [109], [265], algorithmic stability [106], and PAC-Bayesian [108] only focus on the generalization gap, which is the difference between the empirical risk and the population risk function, for a given GCN model with arbitrary parameters and the number of layers [108]. When analyzing the training, these works usually consider impractical infinitely wide Graph neural networks, and a performance gap exists between theory and practice. In contrast, our framework involves the convergence analysis of GCN/Graph Transformers using SGD on a class of target functions and the generalization gap with the trained model. The zero generalization we achieve is zero population risk, which means the learned model from the training is guaranteed to have the desired generalization on the testing data. The model recovery framework [74] requires a tensor initialization to locate the initial parameter close to the ground truth weight. For the NTK [194] framework, they need an impractical condition of an infinitely wide network to linearize the model around the random initialization.

Then, we compare existing works on Transformers. As far as we know, the state-of-the-art generalization analysis on other Transformers [6], [119], [115], [104] did not consider any graph-based labelling function and trainable positional encoding, which are crucial and necessary for node classification tasks. However, we cover these in the formulation and provide the training dynamics and generalization analysis accordingly.

## APPENDIX C

### APPENDIX OF CHAPTER 4

#### C.1 Algorithm

We first introduce new notations to be used in this part and summarize key notions in Table C.1.

We write  $f(x) \lesssim (\gtrsim) g(x)$  if  $f(x) \leq (\geq) \Theta(g(x))$ . The gradient and the Hessian of a function  $f(\mathbf{W})$  are denoted by  $\nabla f(\mathbf{W})$  and  $\nabla^2 f(\mathbf{W})$ , respectively.  $\mathbf{A} \succeq 0$  means  $\mathbf{A}$  is a positive semi-definite (PSD) matrix.  $\mathbf{A}^{\frac{1}{2}}$  means that  $\mathbf{A} = (\mathbf{A}^{\frac{1}{2}})^2$ . The outer product of vectors  $\mathbf{z}_i \in \mathbb{R}^{n_i}$ ,  $i \in [l]$ , is defined as  $\mathbf{T} = \mathbf{z}_1 \otimes \cdots \otimes \mathbf{z}_l \in \mathbb{R}^{n_1 \times \cdots \times n_l}$  with  $\mathbf{T}_{j_1 \cdots j_l} = (\mathbf{z}_1)_{j_1} \cdots (\mathbf{z}_l)_{j_l}$ . Given a tensor  $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$  and matrices  $\mathbf{A} \in \mathbb{R}^{n_1 \times d_1}$ ,  $\mathbf{B} \in \mathbb{R}^{n_2 \times d_2}$ ,  $\mathbf{C} \in \mathbb{R}^{n_3 \times d_3}$ , the  $(i_1, i_2, i_3)$ -th entry of the tensor  $\mathbf{T}(\mathbf{A}, \mathbf{B}, \mathbf{C})$  is given by

$$\sum_{i'_1}^{n_1} \sum_{i'_2}^{n_2} \sum_{i'_3}^{n_3} \mathbf{T}_{i'_1, i'_2, i'_3} \mathbf{A}_{i'_1, i_1} \mathbf{B}_{i'_2, i_2} \mathbf{C}_{i'_3, i_3}. \quad (\text{C.1})$$

The method starts from an initialization  $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$  computed based on the tensor initialization method (Subroutine 1) and then updates the iterates  $\mathbf{W}_t$  using gradient descent with the step size  $\eta_0$ . To model the inaccuracy in computing the gradient, an i.i.d. zero-mean noise  $\{\nu_i\}_{i=1}^n \in \mathbb{R}^{d \times K}$  with bounded magnitude  $|(\nu_i)_{jk}| \leq \xi$  ( $j \in [d], k \in [K]$ ) for some  $\xi \geq 0$  are added in (A.6) when computing the gradient of the loss in (4.3).

Our tensor initialization method in Subroutine 1 is extended from [266] and [71]. The idea is to compute quantities ( $\mathbf{Q}_j$  in (C.2)) that are tensors of  $\mathbf{w}_i^*$  and then apply the tensor decomposition method to estimate  $\mathbf{w}_i^*$ . Because  $\mathbf{Q}_j$  can only be estimated from training samples, tensor decomposition does not return  $\mathbf{w}_i^*$  exactly but provides a close approximation, and this approximation is used as the initialization for Algorithm 1. Because the existing method on tensor construction only applies to the standard Gaussian distribution, we exploit the relationship between probability density functions and tensor expressions developed in

---

Portions of this appendix previously appeared as: H. Li, S. Zhang, Y. Zhang, M. Wang, S. Liu, and P.-Y. Chen, “How does promoting the minority fraction affect generalization? A theoretical study of one-hidden-layer neural network on group imbalance,” *IEEE J. Sel. Top. Signal Process.*, vol. 18, no. 2, pp. 216–231, Mar. 2024. ©2024 IEEE.

Portions of this appendix previously appeared as: H. Li, S. Zhang, M. Wang, “Learning and generalization of one-hidden-layer neural networks, going beyond standard gaussian data,” In *Annu. Conf. Inf. Sci. Syst.*, Mar. 2022. pp. 37-42.

**Table C.1: Summary of notations.** ©2024 IEEE.

$\lambda_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, l \in [L]$	The fraction, mean, and covariance of the $l$ -th component in the Gaussian mixture distribution, respectively.
$d, n, K$	The feature dimension, the number of training samples, and the number of neurons, respectively.
$\mathbf{W}^*, \mathbf{W}_t$	$\mathbf{W}^*$ is the ground truth weight. $\mathbf{W}_t$ is the updated weight in the $t$ -th iteration.
$f_n, \bar{f}, \ell$	$f_n$ is the empirical risk function. $\bar{f}$ is the average risk or the population risk function. $\ell$ is the cross-entropy loss function.
$\Psi, \sigma_{\max}, \sigma_{\min}, \tau$	$\Psi$ denotes our Gaussian mixture model $(\lambda_l, \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l, \forall l)$ . $\sigma_{\max} = \max_{l \in [L]} \{\ \boldsymbol{\Sigma}_l\ ^{\frac{1}{2}}\}$ . $\sigma_{\min} = \min_{l \in [L]} \{\ \boldsymbol{\Sigma}_l^{-1}\ ^{-\frac{1}{2}}\}$ . $\tau = \sigma_{\max}/\sigma_{\min}$ .
$\delta_i(\mathbf{W}^*), \eta, \kappa, i \in [K]$	$\delta_i(\mathbf{W}^*)$ is the $i$ -th largest singular value of $\mathbf{W}^*$ . $\eta$ and $\kappa$ are two functions of $\mathbf{W}^*$ .
$\rho(\mathbf{u}, \sigma), \Gamma(\Psi), D_m(\Psi)$	These items are functions of the Gaussian mixture distribution $\Psi$ used to develop our Theorem 2.
$\boldsymbol{\nu}_i, \xi$	$\boldsymbol{\nu}_i$ is the gradient noise. $\xi$ is the upper bound of the noise level.
$\mathbf{Q}_j, j = 1, 2, 3$	$\mathbf{Q}_j$ 's are tensors used in the initialization.
$\mathcal{B}(\Psi)$	A parameter appeared in the sample complexity bound (4.7).
$v(\Psi), q(\Psi)$	$v(\Psi)$ is the convergence rate (4.8). $q(\Psi)$ is a parameter in the definition of $v(\Psi)$ (4.9).
$\mathcal{E}_w(\Psi), \mathcal{E}, \mathcal{E}_l$	Generalization parameters. $\mathcal{E}_w(\Psi)$ appears in the error bound of the model (4.10). $\mathcal{E}(\Psi)$ and $\mathcal{E}_l(\Psi)$ are to characterize the average risk (4.11) and the group-l risk (4.12), respectively.

[266] to design tensors suitable for the Gaussian mixture model. Formally,

**Definition C.1.1.** For  $j = 1, 2, 3$ , we define

$$\mathbf{Q}_j := \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [y \cdot (-1)^j p^{-1}(\mathbf{x}) \nabla^{(j)} p(\mathbf{x})], \quad (\text{C.2})$$

where  $p(\mathbf{x})$ , the probability density function of GMM is defined as

$$p(\mathbf{x}) = \sum_{l=1}^L \lambda_l (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_l|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_l) \boldsymbol{\Sigma}_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l) \right) \quad (\text{C.3})$$

If the Gaussian mixture model is symmetric, the symmetric distribution can be written as

$$\mathbf{x} \sim \begin{cases} \sum_{l=1}^{\frac{L}{2}} \lambda_l (\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) + \mathcal{N}(-\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)) & L \text{ is even} \\ \lambda_1 \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1) + \sum_{l=2}^{\frac{L-1}{2}} \lambda_l (\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) + \mathcal{N}(-\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)) & L \text{ is odd} \end{cases} \quad (\text{C.4})$$

$\mathbf{Q}_j$  is a  $j$ th-order tensor of  $\mathbf{w}_i^*$ , e.g.,  $\mathbf{Q}_3 = \frac{1}{K} \sum_{i=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\phi'''(\mathbf{w}_i^{*\top} \mathbf{x})] \mathbf{w}_i^{*\otimes 3}$ .

These quantifies cannot be directly computed from (C.2) but can be estimated by sample means, denoted by  $\widehat{\mathbf{Q}}_j$  ( $j = 1, 2, 3$ ), from samples  $\{\mathbf{x}_i, y_i\}_{i=1}^n$ . The following assumption guarantees that these tensors are nonzero and can thus be leveraged to estimate  $\mathbf{W}^*$ .

**Assumption C.1.2.** The Gaussian Mixture Model in (C.4) satisfies the following conditions:

1.  $\mathbf{Q}_1$  and  $\mathbf{Q}_3$  are nonzero.
2. If the distribution is not symmetric, then  $\mathbf{Q}_2$  is nonzero.

Assumption C.1.2 is a very mild assumption<sup>34</sup>. Moreover, as indicated in [266], in the rare case that some quantities  $\mathbf{Q}_i$  ( $i = 1, 2, 3$ ) are zero, one can construct higher-order tensors in a similar way as in Definition C.1.1 and then estimate  $\mathbf{W}^*$  from higher-order tensors.

Subroutine 1 describes the tensor initialization method, which estimates the direction and magnitude of  $\mathbf{w}_j^*, j \in [K]$ , separately. The direction vectors are denoted as  $\bar{\mathbf{w}}_j^* = \mathbf{w}_j^*/\|\mathbf{w}_j^*\|$  and the magnitude  $\|\mathbf{w}_j^*\|$  is denoted as  $z_j$ . Lines 2-6 estimate the subspace  $\widehat{\mathbf{U}}$  spanned by  $\{\mathbf{w}_1^*, \dots, \mathbf{w}_K^*\}$  using  $\widehat{\mathbf{Q}}_2$  or, in the case that  $\mathbf{Q}_2 = 0$ , a second-order tensor projected by  $\widehat{\mathbf{Q}}_3$ . Lines 7-8 estimate  $\bar{\mathbf{w}}_j^*$  by employing the KCL algorithm [267]. Lines 9-10 estimate the magnitude  $z_j$ . Finally, the returned estimation of  $\mathbf{W}^*$  is used as an initialization  $\mathbf{W}_0$  for Algorithm 1. The computational complexity of Subroutine 1 is  $O(Knd)$  based on similar calculations as those in [71].

### C.1.1 Numerical Evaluation of Tensor Initialization

Figure C.1 shows the accuracy of the returned model by Algorithm 1. Here  $n = 2 \times 10^5$ ,  $d = 50$ ,  $K = 2$ ,  $\lambda_1 = \lambda_2 = 0.5$ ,  $\boldsymbol{\mu}_1 = -0.3 \cdot \mathbf{1}$  and  $\boldsymbol{\mu}_2 = \mathbf{0}$ . We compare the tensor initialization with a random initialization in a local region  $\{\mathbf{W} \in \mathbb{R}^{d \times K} : \|\mathbf{W} - \mathbf{W}^*\|_F \leq \epsilon\}$ .

---

<sup>34</sup>By mild, we mean given  $L$ , if Assumption 1 is not met for some  $\Psi_0$ , there exists an infinite number of  $\Psi'$  in any neighborhood of  $\Psi_0$  such that Assumption 1 holds for  $\Psi'$ ,

---

**Subroutine 1** Tensor Initialization Method
 

---

- 1: **Input:** Partition  $n$  pairs of data  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$  into three disjoint subsets  $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$
- 2: **if** the Gaussian Mixture distribution is not symmetric **then**
- 3:   Compute  $\hat{\mathbf{Q}}_2$  using  $\mathcal{D}_1$ . Estimate the subspace  $\hat{\mathbf{U}}$  by orthogonalizing the eigenvectors with respect to the  $K$  largest eigenvalues of  $\hat{\mathbf{Q}}_2$
- 4: **else**
- 5:   Pick an arbitrary vector  $\boldsymbol{\alpha} \in \mathbb{R}^d$ , and use  $\mathcal{D}_1$  to compute  $\hat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})$ . Estimate  $\hat{\mathbf{U}}$  by orthogonalizing the eigenvectors with respect to the  $K$  largest eigenvalues of  $\hat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})$ .
- 6: **end if**
- 7: Compute  $\hat{\mathbf{R}}_3 = \hat{\mathbf{Q}}_3(\hat{\mathbf{U}}, \hat{\mathbf{U}}, \hat{\mathbf{U}})$  from data set  $\mathcal{D}_2$
- 8: Employ the KCL algorithm to compute vectors  $\{\hat{\mathbf{v}}_i\}_{i \in [K]}$ , which are the estimates of  $\{\hat{\mathbf{U}}^\top \bar{\mathbf{w}}_i^*\}_{i=1}^K$ . Then the direction vectors  $\{\bar{\mathbf{w}}_i^*\}_{i=1}^K$  can be approximated by  $\{\hat{\mathbf{U}} \hat{\mathbf{v}}_i\}_{i=1}^K$ .
- 9: Compute  $\hat{\mathbf{Q}}_1$  from data set  $\mathcal{D}_3$ .
- 10: Estimate the magnitude  $\hat{\mathbf{z}}$  by solving the optimization problem

$$\hat{\mathbf{z}} = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \frac{1}{2} \left\| \hat{\mathbf{Q}}_1 - \sum_{j=1}^K \alpha_j \bar{\mathbf{w}}_j^* \right\|^2 \quad (\text{C.5})$$

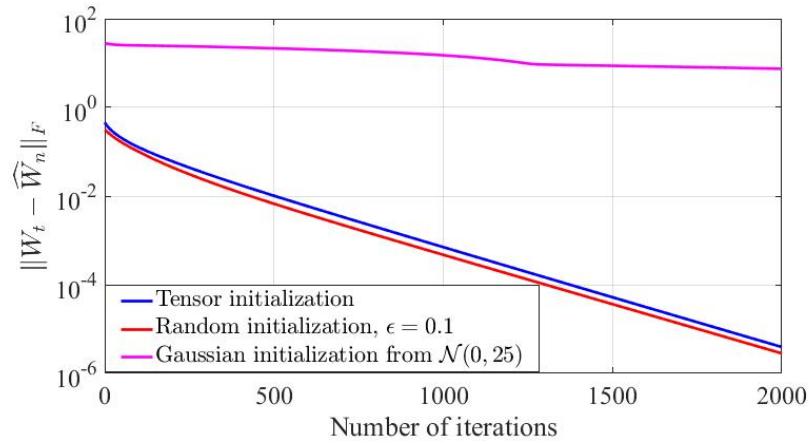
- 11: **Return:** Use  $\hat{z}_j \hat{\mathbf{U}} \hat{\mathbf{v}}_j$  as the  $j$ th column of  $\mathbf{W}_0$ ,  $j \in [K]$ .
- 

Each entry of  $\mathbf{W}^*$  is selected from  $[-0.1, 0.1]$  uniformly. Tensor initialization in Subroutine 1 returns an initial point close to one permutation of  $\mathbf{W}^*$ , with a relative error of 0.65. If the random initialization is also close to  $\mathbf{W}^*$ , e.g.,  $\epsilon = 0.1$ , then the gradient descent algorithm converges to a critical point from both initializations, and the linear convergence rate is the same. We also test a random initialization with each entry drawn from  $\mathcal{N}(0, 25)$ . The initialization is sufficiently far from  $\mathbf{W}^*$ , and the algorithm does not converge. On a MacBook Pro with Intel(R) Core(TM) i5-7360U CPU at 2.30GHz and MATLAB 2017a, it takes 5.52 seconds to compute the tensor initialization. Thus, to reduce the computational time, we consider a random initialization with  $\epsilon = 0.1$  in the experiments instead of computing tensor initialization.

## C.2 Preliminaries of the Main Proof

In this section, we introduce some **definitions** and **properties** that will be used to prove the main results.

First, we define the sub-Gaussian random variable and sub-Gaussian norm.



**Figure C.1:** Comparison between tensor initialization, a random initialization near  $\mathbf{W}^*$ , and an arbitrary random initialization. ©2024 IEEE.

**Definition C.2.1.** 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  $\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-\frac{1}{2}}(\mathbb{E}|X|^p)^{\frac{1}{p}}$ .

Then we define the following three quantities.  $\rho(\boldsymbol{\mu}, \sigma)$  is motivated by the  $\rho$  parameter for the standard Gaussian distribution in [71], and we generalize it to a Gaussian with an arbitrary mean and variance. We define the new quantities  $\Gamma(\Psi)$  and  $D_m(\Psi)$  for the Gaussian mixture model.

**Definition C.2.2.** ( $\rho$ -function). Let  $\mathbf{z} \sim \mathcal{N}(\mathbf{u}, \mathbf{I}_d) \in \mathbb{R}^d$ . Define  $\alpha_q(i, \mathbf{u}, \sigma) = \mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'(\sigma \cdot z_i)z_i^q]$  and  $\beta_q(i, \mathbf{u}, \sigma) = \mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'^2(\sigma \cdot z_i)z_i^q]$ ,  $\forall q \in \{0, 1, 2\}$ , where  $z_i$  and  $u_i$  is the  $i$ -th entry of  $\mathbf{z}$  and  $\mathbf{u}$ , respectively. Define  $\rho(\mathbf{u}, \sigma)$  as

$$\rho(\mathbf{u}, \sigma) = \min_{i, j \in [d], j \neq i} \{(u_j^2 + 1)(\beta_0(i, \mathbf{u}, \sigma) - \alpha_0(i, \mathbf{u}, \sigma)^2), \beta_2(i, \mathbf{u}, \sigma) - \frac{\alpha_2(i, \mathbf{u}, \sigma)^2}{u_i^2 + 1}\} \quad (\text{C.6})$$

**Definition C.2.3.** ( $\Gamma$ -function). With (C.6) and  $\kappa, \eta$  defined in Section 4.3, we define

$$\Gamma(\Psi) = \sum_{l=1}^L \frac{\lambda_l}{\tau^K \kappa^2 \eta} \frac{\|\Sigma_l^{-1}\|^{-1}}{\sigma_{\max}^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right) \quad (\text{C.7})$$

**Definition C.2.4.** (D-function). Given the Gaussian Mixture Model and any positive integer  $m$ , define  $D_m(\Psi)$  as

$$D_m(\Psi) = \sum_{l=1}^L \lambda_l \left( \frac{\|\boldsymbol{\mu}_l\|}{\|\Sigma_l^{-1}\|^{-\frac{1}{2}}} + 1 \right)^m, \quad (\text{C.8})$$

$\rho$ -function is defined to compute the lower bound of the Hessian of the population risk with Gaussian input.  $\Gamma$  function is the weighted sum of  $\rho$ -function under mixture Gaussian distribution. This function is positive and upper bounded by a small value.  $\Gamma$  goes to zero if all  $\|\boldsymbol{\mu}_l\|$  or all  $\sigma_l$  goes to infinity.  $D$ -function is a normalized parameter for the means and variances. It is lower bounded by 1.  $D$ -function is an increasing function of  $\|\boldsymbol{\mu}_l\|$  and a decreasing function of  $\sigma_l$ .

*Property 1.* Given  $\mathbf{W}^* = \mathbf{U}\mathbf{V} \in \mathbb{R}^{d \times k}$ , where  $\mathbf{U} \in \mathbb{R}^{d \times K}$  is the orthogonal basis of  $\mathbf{W}^*$ . For any  $\boldsymbol{\mu} \in \mathbb{R}^d$ , we can find an orthogonal decomposition of  $\boldsymbol{\mu}$  based on the column space of  $\mathbf{W}^*$ , i.e.  $\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathbf{U}} + \boldsymbol{\mu}_{\mathbf{U}^\perp}$ . If we consider the recovery problem of FCN with a dataset of Gaussian Mixture Model, in which  $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_h, \Sigma_h)$  for some  $h \in [L]$ , the problem is equivalent to the problem of FCN with  $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{U}_h}, \Sigma_h)$ . Hence, we can assume without loss of generality that  $\boldsymbol{\mu}_l$  belongs to the column space of  $\mathbf{W}^*$  for all  $l \in [L]$ .

**Proof:**

From (4.1) and (4.3), the recovery problem can be formulated as

$$\min_{\mathbf{W}^*} g(\mathbf{W}^{*\top} \mathbf{x}_i, y_i)$$

For any  $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_h, \Sigma_h)$ ,  $\mathbf{x}_i$  can be written as

$$\mathbf{x}_i = \mathbf{z} + \boldsymbol{\mu}_h$$

where  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \Sigma_h)$ . Therefore,

$$\mathbf{W}^{*\top} \mathbf{x}_i = \mathbf{W}^{*\top} (\mathbf{z} + \boldsymbol{\mu}_h) = \mathbf{W}^{*\top} (\mathbf{z} + \boldsymbol{\mu}_{\mathbf{U}_h} + \boldsymbol{\mu}_{\mathbf{U}_h^\perp}) = \mathbf{W}^{*\top} (\mathbf{z} + \boldsymbol{\mu}_{\mathbf{U}_h})$$

The final step is because  $\mathbf{W}^{*\top} \boldsymbol{\mu}_{\mathbf{U}_h^\perp} = \mathbf{0}$ . So the problem is equivalent to the recovery problem of FCN with  $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{U}_h}, \Sigma_h)$ .

Recall that the gradient noise  $\boldsymbol{\nu}_i \in \mathbb{R}^{d \times K}$  is zero-mean, and each of its entry is upper bounded by  $\xi > 0$ .

*Property 2.* We have that  $\|\boldsymbol{\nu}_i\|_F$  is a sub-Gaussian random variable with its sub-Gaussian norm bounded by  $\xi\sqrt{dK}$ .

**Proof:**

$$(\mathbb{E}\|\boldsymbol{\nu}_i\|_F^p)^{\frac{1}{p}} \leq (\mathbb{E}|\sqrt{dK}\xi|^p)^{\frac{1}{p}} \leq \xi\sqrt{dK} \quad (\text{C.9})$$

We state some general properties of the  $\rho$  function defined in Definition C.2.2 in the following.

*Property 3.*  $\rho(\mathbf{u}, \sigma)$  in Definition C.2.2 satisfies the following properties,

1. **(Positive)**  $\rho(\mathbf{u}, \sigma) > 0$  for any  $\mathbf{u} \in \mathbb{R}^d$  and  $\sigma \neq 0$ .
2. **(Finite limit point for zero mean)**  $\rho(\mathbf{u}, \sigma)$  converges to a positive value function of  $\sigma$  as  $u_i$  goes to 0, i.e.  $\lim_{u_i \rightarrow 0} \rho(\mathbf{u}, \sigma) := \mathcal{C}_m(\sigma)$ .
3. **(Finite limit point for zero variance)** When all  $u_i \neq 0$  ( $i \in [d]$ ),  $\rho(\frac{\mathbf{u}}{\sigma}, \sigma)$  converges to a strictly positive real function of  $\mathbf{u}$  as  $\sigma$  goes to 0, i.e.  $\lim_{\sigma \rightarrow 0} \rho(\frac{\mathbf{u}}{\sigma}, \sigma) := \mathcal{C}_s(\mathbf{u})$ . When  $u_i = 0$  for some  $i \in [d]$ ,  $\lim_{\sigma \rightarrow 0} \rho(\frac{\mathbf{u}}{\sigma}, \sigma) = 0$ .
4. **(Lower bound function of the mean)** When everything else except  $|u_i|$  is fixed,  $\rho(\frac{\mathbf{W}^{*\top}\mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$  is lower bounded by a strictly positive real function,  $\mathcal{L}_m(\frac{(\Lambda\mathbf{W}^*)^\top\Lambda\mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$ , which is monotonically decreasing as  $|u_i|$  increases.
5. **(Lower bound function of the variance)** When everything else except  $\sigma$  is fixed,  $\rho(\frac{\mathbf{W}^{*\top}\mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$  is lower bounded by a strictly positive real function,  $\mathcal{L}_s(\frac{\mathbf{W}^{*\top}\mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$ , which satisfies the following conditions: (a) there exists  $\zeta_{s'} > 0$ , such that  $\sigma^{-1}\mathcal{L}_s(\frac{\mathbf{W}^{*\top}\mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$  is an increasing function of  $\sigma$  when  $\sigma \in (0, \zeta_{s'})$ ; (b) there exists  $\zeta_s > 0$  such that  $\mathcal{L}_s(\frac{\mathbf{W}^{*\top}\mathbf{u}}{\sigma\delta_K(\mathbf{W}^*)}, \sigma\delta_K(\mathbf{W}^*))$  is a decreasing function of  $\sigma$  when  $\sigma \in (\zeta_s, +\infty)$ .

### Proof:

(1) From Cauchy Schwarz's inequality, we have

$$\mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'(\sigma \cdot z_i)] \leq \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'^2(\sigma \cdot z_i)]} \quad (\text{C.10})$$

$$\begin{aligned} \mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'(\sigma \cdot z_i)z_i \cdot z_i] &\leq \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'^2(\sigma \cdot z_i)z_i^2]} \cdot \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[z_i^2]} \\ &= \sqrt{\mathbb{E}_{z_i \sim \mathcal{N}(u_i, 1)}[\phi'^2(\sigma \cdot z_i)z_i^2]} \cdot \sqrt{u_i^2 + 1} \end{aligned} \quad (\text{C.11})$$

The equalities of the (C.10) and (C.11) hold if and only if  $\phi'$  is a constant function. Since that  $\phi$  is the sigmoid function, the equalities of (C.10) and (C.11) cannot hold.

By the definition of  $\rho(\mathbf{u}, \sigma)$  in Definition C.2.2, we have

$$\beta_0(i, \mathbf{u}, \sigma) - \alpha_0^2(i, \mathbf{u}, \sigma) > 0, \quad (\text{C.12})$$

$$\beta_2(i, \mathbf{u}, \sigma) - \frac{\alpha_2^2(i, \mathbf{u}, \sigma)}{u_i^2 + 1} > 0. \quad (\text{C.13})$$

Therefore,

$$\rho(\mathbf{u}, \sigma) > 0 \quad (\text{C.14})$$

(2) We can derive that

$$\begin{aligned} & \lim_{u_i \rightarrow 0} \left( \frac{u_j^2}{\sigma^2} + 1 \right) (\beta_0(i, \mathbf{u}, \sigma) - \alpha_0^2(i, \mathbf{u}, \sigma)) \\ &= \lim_{u_i \rightarrow 0} \left( \frac{u_j^2}{\sigma^2} + 1 \right) \left( \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp(-\frac{\|z_i - u_i\|^2}{2}) dz_i \right. \\ & \quad \left. - \left( \int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp(-\frac{\|z_i - u_i\|^2}{2}) dz_i \right)^2 \right) \\ &= \left( \frac{u_j^2}{\sigma^2} + 1 \right) \left( \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \right. \\ & \quad \left. \exp(-\frac{\|z_i\|^2}{2}) dz_i - \left( \int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp(-\frac{\|z_i\|^2}{2}) dz_i \right)^2 \right), \end{aligned} \quad (\text{C.15})$$

where the first step is by Definition C.2.2, and the second step comes from the limit laws. Similarly, we also have

$$\begin{aligned} & \lim_{u_i \rightarrow 0} \left( \beta_2(i, \mathbf{u}, \sigma) - \frac{1}{u_i^2 + 1} \alpha_2^2(i, \mathbf{u}, \sigma) \right) \\ &= \lim_{u_i \rightarrow 0} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp(-\frac{\|z_i - u_i\|^2}{2}) dz_i \\ & \quad - \left( \frac{1}{u_i^2 + 1} \int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp(-\frac{\|z_i - u_i\|^2}{2}) dz_i \right)^2 \\ &= \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp(-\frac{\|z_i\|^2}{2}) dz_i - \left( \int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \right. \\ & \quad \left. \exp(-\frac{\|z_i\|^2}{2}) dz_i \right)^2 \end{aligned} \quad (\text{C.16})$$

Since that (C.15) and (C.16) are positive due to Jensen's inequality, we can derive that

$\rho(\mathbf{u}, \sigma)$  converges to a positive value function of  $\sigma$  as  $u_i$  goes to 0, i.e.

$$\lim_{u \rightarrow 0} \rho(\mathbf{u}, \sigma) := \mathcal{C}_m(\sigma) \quad (\text{C.17})$$

(3) When all  $u_i \neq 0$  ( $i \in [d]$ ),

$$\begin{aligned}
& \lim_{\sigma \rightarrow 0} \left( \beta_2(i, \frac{\mathbf{u}}{\sigma}, \sigma) - \frac{1}{\frac{u_i^2}{\sigma^2} + 1} \alpha_2^2(i, \frac{\mathbf{u}}{\sigma}, \sigma) \right) \\
&= \lim_{\sigma \rightarrow 0} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i - \frac{u_i}{\sigma}\|^2}{2}\right) dz_i \\
&\quad - \frac{1}{\frac{u_i^2}{\sigma^2} + 1} \left( \int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) z_i^2 (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i - \frac{u_i}{\sigma}\|^2}{2}\right) dz_i \right)^2 \\
&= \lim_{\sigma \rightarrow 0} \int_{-\infty}^{\infty} \phi'^2(u_i \cdot x_i) \frac{u_i^2}{\sigma^2} x_i^2 (2\pi \frac{\sigma^2}{u_i^2})^{-\frac{1}{2}} \exp\left(-\frac{\|x_i - 1\|^2}{2\frac{\sigma^2}{u_i^2}}\right) dx_i \\
&\quad - \frac{1}{\frac{u_i^2}{\sigma^2} + 1} \left( \int_{-\infty}^{\infty} \phi'(u_i \cdot x_i) \frac{u_i^2}{\sigma^2} x_i^2 (2\pi \frac{\sigma^2}{u_i^2})^{-\frac{1}{2}} \exp\left(-\frac{\|x_i - 1\|^2}{2\frac{\sigma^2}{u_i^2}}\right) dx_i \right)^2 \quad z_i = \frac{u_i}{\sigma} x_i \quad (\text{C.18}) \\
&= \lim_{\sigma \rightarrow 0} \phi'^2(u_i) \frac{u_i^2}{\sigma^2} - \frac{1}{\frac{u_i^2}{\sigma^2} + 1} (\phi'(u_i) \frac{u_i^2}{\sigma^2})^2 \\
&= \lim_{\sigma \rightarrow 0} \phi'^2(u_i) \frac{u_i^2}{\sigma^2} \left( 1 - \frac{\frac{u_i^2}{\sigma^2}}{1 + \frac{u_i^2}{\sigma^2}} \right) \\
&= \lim_{\sigma \rightarrow 0} \phi'^2(u_i) \frac{1}{1 + \frac{\sigma^2}{u_i^2}} \\
&= \phi'^2(u_i)
\end{aligned}$$

The first step of (C.18) comes from Definition C.2.2. The second step and the last three steps are derived from some basic mathematical computation and the limit laws. The third step of (C.18) is by the fact that the Gaussian distribution goes to a Dirac delta function when  $\sigma$  goes to 0. Then the integral will take the value when  $x_i = 1$ . Similarly, we can obtain the following

$$\begin{aligned}
& \lim_{\sigma \rightarrow 0} \left( \beta_0(i, \frac{\mathbf{u}}{\sigma}, \sigma) - \alpha_0^2(i, \frac{\mathbf{u}}{\sigma}, \sigma) \right) \\
&= \lim_{\sigma \rightarrow 0} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i - \frac{u_i}{\sigma}\|^2}{2}\right) dz_i \\
&\quad - \left( \int_{-\infty}^{\infty} \phi'(\sigma \cdot z_i) (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\|z_i - \frac{u_i}{\sigma}\|^2}{2}\right) dz_i \right)^2 \\
&= \phi'^2(u_i) - \phi'^2(u_i) = 0
\end{aligned} \tag{C.19}$$

$$\begin{aligned}
& \lim_{\sigma \rightarrow 0} \left( \frac{\partial}{\partial \sigma} \left( \beta_0(i, \frac{\mathbf{u}}{\sigma}, \sigma) - \alpha_0^2(i, \frac{\mathbf{u}}{\sigma}, \sigma) \right) \right) \\
&= \lim_{\sigma \rightarrow 0} \left( \frac{\partial}{\partial \sigma} \left( \int_{-\infty}^{\infty} \phi'^2(x_i) (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\|x_i - u_i\|^2}{2\sigma^2}\right) dx_i \right. \right. \\
&\quad \left. \left. - \left( \int_{-\infty}^{\infty} \phi'(x_i) (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\|x_i - u_i\|^2}{2\sigma^2}\right) dx_i \right)^2 \right) \right) \quad x_i = \sigma \cdot z_i \\
&= \lim_{\sigma \rightarrow 0} \left( \int_{-\infty}^{\infty} \phi'^2(x_i) (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\|x_i - u_i\|^2}{2\sigma^2}\right) (-\sigma^{-1} + \|x_i - u_i\|^2\sigma^{-2}) dx_i \right. \\
&\quad \left. - 2 \left( \int_{-\infty}^{\infty} \phi'(x_i) (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\|x_i - u_i\|^2}{2\sigma^2}\right) dx_i \right) \right. \\
&\quad \left. \cdot \int_{-\infty}^{\infty} \phi'(x_i) (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\|x_i - u_i\|^2}{2\sigma^2}\right) (-\sigma^{-1} + \|x_i - u_i\|^2\sigma^{-2}) dx_i \right) \\
&= \lim_{\sigma \rightarrow 0} \left( \frac{\phi'^2(u_i)}{-\sigma} - 2\phi'(u_i) \frac{\phi'(u_i)}{-\sigma} \right) \\
&= \lim_{\sigma \rightarrow 0} \frac{\phi'^2(u_i)}{\sigma} = +\infty
\end{aligned} \tag{C.20}$$

Therefore, by L'Hopital's rule and (C.19), (C.20), we have

$$\begin{aligned}
& \lim_{\sigma \rightarrow 0} \left( \frac{u_j^2}{\sigma^2} + 1 \right) \left( \beta_0(i, \frac{\mathbf{u}}{\sigma}, \sigma) - \alpha_0(i, \frac{\mathbf{u}}{\sigma}, \sigma) \right) \\
&= \lim_{\sigma \rightarrow 0} \frac{u_i^2}{2\sigma} \frac{\partial}{\partial \sigma} \left( \beta_0(i, \frac{\mathbf{u}}{\sigma}, \sigma) - \alpha_0(i, \frac{\mathbf{u}}{\sigma}, \sigma) \right) \\
&= +\infty
\end{aligned} \tag{C.21}$$

Combining (C.21) and (C.18), we can derive that  $\rho(\frac{\mathbf{u}}{\sigma}, \sigma)$  converges to a positive value function of  $\mathbf{u}$  as  $\sigma$  goes to 0, i.e.

$$\lim_{\sigma \rightarrow 0} \rho\left(\frac{\mathbf{u}}{\sigma}, \sigma\right) := \mathcal{C}_s(\mathbf{u}). \tag{C.22}$$

When  $u_i = 0$  for some  $i \in [d]$ ,  $\lim_{\sigma \rightarrow 0} \left( \frac{u_i^2}{\sigma^2} + 1 \right) \left( \beta_0(j, \frac{\mathbf{u}}{\sigma}, \sigma) - \alpha_0^2(j, \frac{\mathbf{u}}{\sigma}, \sigma) \right) = 0$  by (C.19). Then

from the Definition C.2.2, we have

$$\lim_{\sigma \rightarrow 0} \rho\left(\frac{\mathbf{u}}{\sigma}, \sigma\right) = 0 \quad (\text{C.23})$$

(4) We can define  $\mathcal{L}_m\left(\frac{(\Lambda \mathbf{W}^*)^\top \Lambda \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)$  as

$$\begin{aligned} & \mathcal{L}_m\left(\frac{(\Lambda \mathbf{W}^*)^\top \Lambda \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) \\ &= \min_{v_i \in [0, u_i]} \left\{ \rho\left(\frac{(\Lambda \mathbf{W}^*)^\top \Lambda \mathbf{v}}{\sigma_l \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) : v_j = u_j \text{ for all } j \neq i \right\} \end{aligned} \quad (\text{C.24})$$

Then by this definition, we have

$$0 < \mathcal{L}_m\left(\frac{(\Lambda \mathbf{W}^*)^\top \Lambda \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) \leq \rho\left(\frac{(\Lambda \mathbf{W}^*)^\top \Lambda \mathbf{u}}{\sigma_l \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) \quad (\text{C.25})$$

Meanwhile, for any  $0 \leq u'_i \leq u_i^*$ , since that  $[0, u'_i] \subset [0, u_i^*]$ , we can obtain

$$\mathcal{L}_m\left(\frac{(\Lambda \mathbf{W}^*)^\top \Lambda \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)|_{u_i=u'_i} \geq \mathcal{L}_m\left(\frac{(\Lambda \mathbf{W}^*)^\top \Lambda \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)|_{u_i=u_i^*} \quad (\text{C.26})$$

Hence,  $\mathcal{L}_m\left(\frac{(\Lambda \mathbf{W}^*)^\top \Lambda \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)$  is a strictly positive real function which is monotonically decreasing.

(5) Therefore, we only need to show the condition (a).

When  $(\mathbf{W}^{*\top} \mathbf{u})_i \neq 0$  for all  $i \in [K]$ ,

$$\lim_{\sigma \rightarrow 0} \rho\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) = \mathcal{C}_s(\mathbf{u}) > 0. \quad (\text{C.27})$$

Therefore, there exists  $\zeta_s > 0$ , such that when  $0 < \sigma < \zeta_s$ ,

$$\rho\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) > \frac{\mathcal{C}_s(\mathbf{W}^{*\top} \mathbf{u})}{2}. \quad (\text{C.28})$$

Then we can define

$$\mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) := \frac{\mathcal{C}_s(\mathbf{W}^{*\top} \mathbf{u})}{2\zeta_s} \sigma^2 \quad (\text{C.29})$$

such that  $\sigma^{-1} \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)$  is an increasing function of  $\sigma$  below  $\rho\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)$ .

When  $(\mathbf{W}^*)^\top \mathbf{u}_i = 0$  for some  $i \in [K]$ , then

$$\lim_{\sigma \rightarrow 0} \rho\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) = 0. \quad (\text{C.30})$$

We can define

$$\begin{aligned} & \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) \\ &= \sigma \cdot \min_{v_i \in [u_i, \zeta_{s'}]} \left\{ \rho\left(\frac{\mathbf{W}^{*\top} \mathbf{v}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) : v_j \neq u_j \text{ for all } j \neq i \right\} \end{aligned} \quad (\text{C.31})$$

Then,

$$\begin{aligned} & \sigma^{-1} \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) \\ &= \min_{v_i \in [u_i, \zeta_{s'}]} \left\{ \rho\left(\frac{\mathbf{W}^{*\top} \mathbf{v}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right) : v_j = u_j \text{ for all } j \neq i \right\} \end{aligned} \quad (\text{C.32})$$

For any  $0 \leq u'_i \leq u_i^* < \zeta_{s'}$ , since that  $[u_i^*, \zeta_{s'}] \subset [u'_i, \zeta_{s'}]$ , we can obtain

$$\sigma^{-1} \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)|_{u_i=u'_i} \leq \sigma^{-1} \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)|_{u_i=u_i^*} \quad (\text{C.33})$$

Therefore, we can derive that  $\sigma^{-1} \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)$  is monotonically increasing. Following the steps in (4), we can have that  $\sigma^{-1} \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)$  is a strictly positive real function which is upper bounded by  $\rho\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)$ .

In conclusion, condition (a) is proved.

For condition (b), since that  $\zeta_s > 0$ ,  $\rho\left(\frac{\mathbf{W}^{*\top} \mathbf{u}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)$  is continuous and positive, we can obtain

$$\rho\left(\frac{\mathbf{W}^{*\top} \mathbf{v}}{\sigma \delta_K(\mathbf{W}^*)}, \sigma \delta_K(\mathbf{W}^*)\right)\Big|_{\sigma=\zeta_s} > 0 \quad (\text{C.34})$$

Then condition (b) can be easily proved as in (4).

We then characterize the order of the  $\rho$  function in different cases as follows.

*Property 4.* To specify the order with regard to the distribution parameters,  $\rho(\mathbf{u}, \sigma)$  in Definition C.2.2 satisfies the following properties,

1. (**Small variance**)  $\lim_{\sigma \rightarrow 0^+} \rho(\mathbf{u}, \sigma) = \Theta(\sigma^4)$ .
2. (**Large variance**) For any  $\epsilon > 0$ ,  $\lim_{\sigma \rightarrow \infty} \rho(\mathbf{u}, \sigma) \geq \Theta\left(\frac{1}{\sigma^{3+\epsilon}}\right)$ .

3. **(Large mean)** For any  $\epsilon > 0$ ,  $\lim_{\mu \rightarrow \infty} \rho(\mathbf{u}, \sigma) \geq \Theta(e^{-\frac{\|\mathbf{u}\|^2}{2}}) \frac{1}{\|\mathbf{u}\|^{3+\epsilon}}$ .

**Proof:**

(1)

$$\begin{aligned}
& \beta_0(i, \mathbf{u}, \sigma) - \alpha_0(i, \mathbf{u}, \sigma)^2 \\
&= \mathbb{E}_{z \sim \mathcal{N}(\mu, 1)} [\phi'^2(\sigma \cdot z)] - (\mathbb{E}_{z \sim \mathcal{N}(\mu, 1)} [\phi'(\sigma \cdot z)])^2 \\
&= \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz - \left( \int_{-\infty}^{\infty} \phi'(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \right)^2 \\
&= \int_{-\infty}^{\infty} \left( \frac{1}{4} - \frac{t^2}{16} + \frac{t^4}{96} + \dots \right)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu\sigma)^2}{2\sigma^2}} dt \\
&\quad - \left( \int_{-\infty}^{\infty} \left( \frac{1}{4} - \frac{t^2}{16} + \frac{t^4}{96} + \dots \right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu\sigma)^2}{2\sigma^2}} dt \right)^2 \\
&= \left( \frac{1}{16} - \frac{1}{32}(\mu^2\sigma^2 + \sigma^2) + \frac{7}{768}(3\sigma^4 + 6\mu^2\sigma^4 + \mu^4\sigma^4) + \dots \right) \\
&\quad - \left( \frac{1}{4} - \frac{\mu^2\sigma^2 + \sigma^2}{16} + \frac{3\sigma^4 + 6\mu^2\sigma^4 + \mu^4\sigma^4}{192} + \dots \right)^2 \\
&= \frac{1}{128}\sigma^4 + \frac{\mu^2\sigma^4}{64} + o(\sigma^4), \quad \text{as } \sigma \rightarrow 0^+.
\end{aligned} \tag{C.35}$$

The first step of (C.35) is by Definition C.2.2. The second step and the last steps come from some basic mathematical computation. The third step is from Taylor expansion. Hence,

$$\lim_{\sigma \rightarrow 0^+} (\beta_0(i, \mathbf{u}, \sigma) - \alpha_0(i, \mathbf{u}, \sigma)^2) = \frac{1}{128}\sigma^4 + \frac{\mu^2\sigma^4}{64} + o(\sigma^4) \tag{C.36}$$

Similarly, we can obtain

$$\begin{aligned}
& \beta_2(i, \mathbf{u}, \sigma) - \frac{\alpha_2(i, \mathbf{u}, \sigma)^2}{\mu^2 + 1} \\
&= \mathbb{E}_{z \sim \mathcal{N}(0,1)} [\phi'^2(\sigma \cdot z) z^2] - \frac{(\mathbb{E}_{z \sim \mathcal{N}(0,1)} [\phi'(\sigma \cdot z) z^2])^2}{\mu^2 + 1} \\
&= \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz - \frac{1}{\mu^2 + 1} \left( \int_{-\infty}^{\infty} \phi'(\sigma \cdot z) z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \right)^2 \\
&= \int_{-\infty}^{\infty} \left( \frac{t}{4\sigma} - \frac{t^3}{16\sigma} + \frac{t^5}{96\sigma} \dots \right)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu\sigma)^2}{2\sigma^2}} dt \\
&\quad - \frac{1}{\mu^2 + 1} \left( \int_{-\infty}^{\infty} \left( \frac{t^2}{4\sigma^2} - \frac{t^4}{16\sigma^2} + \frac{t^6}{96\sigma^2} + \dots \right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu\sigma)^2}{2\sigma^2}} dt \right)^2 \\
&= \left( \frac{1+\mu^2}{16} - \frac{3\sigma^2 + 6\mu^2\sigma^2 + \mu^4\sigma^2}{32} + \dots \right) \\
&\quad - \frac{1}{\mu^2 + 1} \left( \frac{1+\mu^2}{4} - \frac{15\sigma^2 + 45\mu^2\sigma^2 + 15\mu^4\sigma^2 + \mu^6\sigma^2}{32} + \dots \right)^2 \\
&= \frac{9}{64}\sigma^2 + \frac{33}{64}\mu^2\sigma^2 + \frac{13}{64}\mu^4\sigma^2 + \frac{1}{64}\mu^6\sigma^2 + o(\sigma^2), \quad \text{as } \sigma \rightarrow 0^+
\end{aligned} \tag{C.37}$$

Hence,

$$\lim_{\sigma \rightarrow 0^+} (\beta_2(i, \mathbf{u}, \sigma) - \frac{\alpha_2(i, \mathbf{u}, \sigma)^2}{\mu^2 + 1}) = \frac{9}{64}\sigma^2 + o(\sigma^2) \tag{C.38}$$

Therefore,

$$\lim_{\sigma \rightarrow 0^+} \rho(\mathbf{u}, \sigma) = \min_{j \in [d], u_j \neq \mu} \{(u_j^2 + 1)\} \frac{1}{128}\sigma^4 \tag{C.39}$$

(2) Note that by some basic mathematical derivation,

$$\begin{aligned}
\int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz &= \int_{-\infty}^{\infty} \frac{1}{(e^{\sigma \cdot z} + e^{-\sigma \cdot z} + 2)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\
&\geq 2 \int_0^{\infty} \frac{1}{16e^{2\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z+\mu)^2}{2}} dz \\
&= \frac{1}{8} e^{2|\mu|\sigma + 2\sigma^2} \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z+2\sigma)^2}{2}} dz \\
&= \frac{1}{8\sqrt{2\pi}} e^{2|\mu|\sigma + 2\sigma^2} \int_{|\mu|+2\sigma}^{\infty} e^{-\frac{t^2}{2}} dt
\end{aligned} \tag{C.40}$$

We then provide the following Claim with its proof to give a lower bound for (C.40).

**Claim:**  $\int_{|\mu|+2\sigma}^{\infty} e^{-\frac{t^2}{2}} dt > e^{-2|\mu|\sigma - 2\sigma^2 - k_1 \log \sigma}$  for  $k_1 > 1$ .

**Proof:** Let

$$f(\sigma) = \int_{|\mu|+2\sigma}^{\infty} e^{-\frac{t^2}{2}} dt - e^{-2|\mu|\sigma - 2\sigma^2 - k_1 \log \sigma}. \tag{C.41}$$

Then,

$$f'(\sigma) = e^{-2\sigma^2}((2|\mu| + 4\sigma + \frac{k_1}{\sigma})\sigma^{-k_1} - 2e^{-\frac{1}{2}\mu^2}). \quad (\text{C.42})$$

It can be easily verified that for a given  $|\mu| \geq 0$ ,  $f'(\sigma) < 0$  when  $\sigma$  is large enough if  $k_1 > 1$ . Combining that  $\lim_{\sigma \rightarrow \infty} f(\sigma) = 0$ , we have  $f(\sigma) > 0$  when  $\sigma$  is large enough by showing the contradiction in the following:

Suppose there is a strictly increasing function  $f(x) > 0$  with  $\lim_{x \rightarrow \infty} f(x) = 0$  when  $x$  is large enough. Then there exists  $x_0 > 0$  such that for any  $\epsilon > 0$ ,  $f(x) < \epsilon$  for  $x > x_0$ . Pick  $\epsilon = f(x_0) > 0$ , then for  $x_1 > x_0$ ,  $f(x_1) > f(x_0) = \epsilon$ . Contradiction!

Similarly, we also have

$$\begin{aligned} \int_{-\infty}^{\infty} \phi'(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz &= \int_{-\infty}^{\infty} \frac{1}{e^{\sigma \cdot z} + e^{-\sigma \cdot z} + 2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &\leq 2 \int_0^{\infty} \frac{1}{e^{\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &= e^{|\mu|\sigma + \frac{1}{2}\sigma^2} \int_0^{\infty} \frac{2}{\sqrt{2\pi}} e^{-\frac{(z+|\mu|+\sigma)^2}{2}} dz \\ &= \frac{2}{\sqrt{2\pi}} e^{|\mu|\sigma + \frac{1}{2}\sigma^2} \int_{|\mu|+\sigma}^{\infty} e^{-\frac{t^2}{2}} dt, \end{aligned} \quad (\text{C.43})$$

and the **Claim**:  $\int_{|\mu|+\sigma}^{\infty} e^{-\frac{t^2}{2}} dt < e^{-|\mu|\sigma - \frac{1}{2}\sigma^2 - k_2 \log \sigma}$  for  $k_2 \leq 1$  to give an upper bound for (C.43).

Therefore, combining (C.40, C.43) and two claims, we have that for any  $\epsilon > 0$ ,

$$\beta_0(i, \mathbf{u}, \sigma) - \alpha_0(i, \mathbf{u}, \sigma)^2 \geq \frac{1}{8\sqrt{2\pi}} \frac{1}{\sigma^{k_1}} - \frac{1}{2\pi} \frac{1}{\sigma^{2k_2}} \gtrsim \frac{1}{\sigma^{1+\epsilon}} \quad (\text{C.44})$$

(The above inequality holds for any  $2k_2 > k_1$  where  $k_1 > 1$  and  $k_2 \leq 1$ .)

Similarly,

$$\begin{aligned} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz &= \int_{-\infty}^{\infty} \frac{z^2}{(e^{\sigma \cdot z} + e^{-\sigma \cdot z} + 2)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &\geq 2 \int_0^{\infty} \frac{z^2}{16e^{2\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z+|\mu|)^2}{2}} dz \\ &= \frac{1}{8\sqrt{2\pi}} e^{|\mu|\sigma + 2\sigma^2} \int_{2|\mu|+2\sigma}^{\infty} (t - 2\sigma)^2 e^{-\frac{t^2}{2}} dt \end{aligned} \quad (\text{C.45})$$

**Claim**:  $\int_{2|\mu|+2\sigma}^{\infty} (t - 2\sigma)^2 e^{-\frac{t^2}{2}} dt \geq e^{-2|\mu|\sigma - 2\sigma^2 - k_1 \log \sigma}$  if  $k_1 > 3$ .

**Proof:** Let

$$f(\sigma) = \int_{|\mu|+2\sigma}^{\infty} (t - 2\sigma)^2 e^{-\frac{t^2}{2}} dt - e^{-2|\mu|\sigma - 2\sigma^2 - k_1 \log \sigma}. \quad (\text{C.46})$$

$$f'(\sigma) = 8\sigma \int_{|\mu|+2\sigma}^{\infty} e^{-\frac{t^2}{2}} dt + e^{-2|\mu|\sigma - 2\sigma^2} (4\sigma^{1-k_1} + k_1\sigma^{-1-k_1} + 2|\mu|\sigma^{-k_1} - 4e^{-\frac{1}{2}\mu^2}). \quad (\text{C.47})$$

We need  $f'(\sigma) < 0$  when  $\sigma$  is large enough. Since that  $f'(\sigma) \rightarrow 0$ ,  $f''(\sigma) \rightarrow 0$  when  $\sigma$  is large, we need  $f''(\sigma) > 0$  and  $f'''(\sigma) < 0$  recursively. Hence,

$$\begin{aligned} f'''(\sigma) = & e^{-2|\mu|\sigma - 2\sigma^2} (64\sigma^{3-k_1} + 96\mu\sigma^{2-k_1} + 16(3k_1 - 3 + \mu^2)\sigma^{1-k_1} \\ & + 8\mu(-\mu^2 - 3 + 6k_1)\sigma^{-k_1} + 4k_1(3k_1 + \mu^2)\sigma^{-1-k_1} + 2k_1(1 + k_1) \\ & \cdot (\mu + 2)\sigma^{-2-k_1} + k_1(1 + k_1)(2 + k_1)\sigma^{-3-k_1} - 16e^{-\frac{1}{2}\mu^2}) < 0 \end{aligned} \quad (\text{C.48})$$

requires  $k_1 > 3$ .

Similarly, we have

$$\begin{aligned} & \int_{-\infty}^{\infty} \phi'(\sigma \cdot z) z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ & \leq 2 \int_0^{\infty} \frac{1}{e^{\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} z^2 e^{-\frac{z^2}{2}} dz = \frac{2}{\sqrt{2\pi}} e^{\frac{1}{2}\sigma^2} \int_{\sigma}^{\infty} (t - \sigma)^2 e^{-\frac{t^2}{2}} dt \end{aligned} \quad (\text{C.49})$$

and the **Claim:**  $\int_{\sigma}^{\infty} (t - \sigma)^2 e^{-\frac{t^2}{2}} dt < e^{-\frac{\sigma^2}{2} - k_2 \log \sigma}$ . Hence,

$$\beta_2(i, \mathbf{u}, \sigma) - \frac{\alpha_2(i, \mathbf{u}, \sigma)^2}{\mu^2 + 1} \geq \frac{1}{8\sqrt{2\pi}} \frac{1}{\sigma^{k_1}} - \frac{2}{\pi(\mu^2 + 1)} \frac{1}{\sigma^{2k_2}} \gtrsim \frac{1}{\sigma^{3.1}} \quad (\text{C.50})$$

(The above inequality holds for any  $2k_2 > k_1$  where  $k_1 > 3$  and  $k_2 < 3$ .)

Therefore, by combining (C.44) and (C.50), for any  $\epsilon > 0$

$$\lim_{\sigma \rightarrow \infty} \rho(\mathbf{u}, \sigma) \geq \Theta\left(\frac{1}{\sigma^{3+\epsilon}}\right). \quad (\text{C.51})$$

(3) Let  $\sigma$  be fixed. For any  $\epsilon > 0$ , following the steps in (2), we can obtain

$$\begin{aligned} \int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz &= \int_{-\infty}^{\infty} \frac{1}{(e^{\sigma \cdot z} + e^{-\sigma \cdot z} + 2)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &\geq 2 \int_0^{\infty} \frac{1}{16e^{2\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z+|\mu|)^2}{2}} dz \\ &= \frac{1}{8\sqrt{2\pi}} e^{2|\mu|\sigma+2\sigma^2} \int_{|\mu|+2\sigma}^{\infty} e^{-\frac{t^2}{2}} dt \\ &\geq \frac{1}{8\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \frac{1}{\mu^{1+\epsilon}} \end{aligned} \quad (\text{C.52})$$

$$\begin{aligned} \int_{-\infty}^{\infty} \phi'(\sigma \cdot z) \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz &= \int_{-\infty}^{\infty} \frac{1}{e^{\sigma \cdot z} + e^{-\sigma \cdot z} + 2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &\leq 2 \int_0^{\infty} \frac{1}{e^{\sigma \cdot z}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \\ &= \frac{2}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \frac{1}{\mu^{1-\epsilon}} \end{aligned} \quad (\text{C.53})$$

Similarly,

$$\int_{-\infty}^{\infty} \phi'^2(\sigma \cdot z) z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \geq \frac{1}{8\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \frac{1}{\mu^{3+\epsilon}} \quad (\text{C.54})$$

$$\int_{-\infty}^{\infty} \phi'(\sigma \cdot z) z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2}} dz \leq \frac{2}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}} \frac{1}{\mu^{3-\epsilon}} \quad (\text{C.55})$$

We can conclude that  $\lim_{\mu \rightarrow \infty} \rho(\mathbf{u}, \sigma) \geq \Theta(e^{-\frac{\|\mathbf{u}\|^2}{2}}) \frac{1}{\|\mathbf{u}\|^{3+\epsilon}}$ .

*Property 5.* If a function  $f(\mathbf{x})$  is an even function, then

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] \quad (\text{C.56})$$

**Proof:**

Denote

$$g(\mathbf{x}) = f(\mathbf{x})(2\pi|\boldsymbol{\Sigma}|^2)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \quad (\text{C.57})$$

By some basic mathematical computation,

$$\begin{aligned}
\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] &= \int_{\mathbf{x} \in \mathbb{R}^d} g(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_d) dx_1 \cdots dx_d \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{\infty}^{-\infty} g(x_1, x_2, \dots, x_d) d(-x_1) dx_2 \cdots dx_d \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(-x_1, x_2, \dots, x_d) dx_1 dx_2 \cdots dx_d \\
&= \int_{\mathbf{x} \in \mathbb{R}^d} g(-\mathbf{x}) d\mathbf{x} \\
&= \int_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) (2\pi|\boldsymbol{\Sigma}|^2)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} + \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} + \boldsymbol{\mu})\right) \\
&= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})]
\end{aligned} \tag{C.58}$$

Therefore, we have

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \boldsymbol{\Sigma})}[f(\mathbf{x})] \tag{C.59}$$

*Property 6.* Under Gaussian Mixture Model  $\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$  where  $\boldsymbol{\Sigma}_l = \text{diag}(\sigma_{l1}^2, \dots, \sigma_{ld}^2)$ , we have the following upper bound.

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}[(\mathbf{u}^\top \mathbf{x})^{2t}] \leq (2t-1)!! \|\mathbf{u}\|^{2t} \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t} \tag{C.60}$$

### Proof:

Note that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}[(\mathbf{u}^\top \mathbf{x})^{2t}] \\
&= \sum_{l=1}^L \lambda_l \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}[(\mathbf{u}^\top \mathbf{x})^{2t}] = \sum_{l=1}^L \lambda_l \mathbb{E}_{y \sim \mathcal{N}(\mathbf{u}^\top \boldsymbol{\mu}_l, \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u})}[y^{2t}],
\end{aligned} \tag{C.61}$$

where the last step is by that  $\mathbf{u}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{u}^\top \boldsymbol{\mu}, \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u})$  for  $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ . By some basic

mathematical computation, we know

$$\begin{aligned}
& \mathbb{E}_{y \sim \mathcal{N}(\mathbf{u}^\top \boldsymbol{\mu}_l, \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u})}[y^{2t}] \\
&= \int_{-\infty}^{\infty} (y - \mathbf{u}^\top \boldsymbol{\mu}_l + \mathbf{u}^\top \boldsymbol{\mu}_l)^{2t} \frac{1}{\sqrt{2\pi \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}}} e^{-\frac{(y - \mathbf{u}^\top \boldsymbol{\mu}_l)^2}{2\mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}}} dy \\
&= \int_{-\infty}^{\infty} \sum_{p=0}^{2t} \binom{2t}{p} (\mathbf{u}^\top \boldsymbol{\mu}_l)^{2t-p} (y - \mathbf{u}^\top \boldsymbol{\mu}_l)^p \frac{1}{\sqrt{2\pi \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}}} e^{-\frac{(y - \mathbf{u}^\top \boldsymbol{\mu}_l)^2}{2\mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}}} dy \\
&= \sum_{p=0}^{2t} \binom{2t}{p} (\mathbf{u}^\top \boldsymbol{\mu}_l)^{2t-p} \cdot \begin{cases} 0 & , p \text{ is odd} \\ (p-1)!! (\mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u})^{\frac{p}{2}} & , p \text{ is even} \end{cases} \quad (C.62) \\
&\leq \sum_{p=0}^{2t} \binom{2t}{p} |\mathbf{u}^\top \boldsymbol{\mu}_l|^{2t-p} (p-1)!! |\mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}|^{\frac{p}{2}} \\
&\leq (2t-1)!! (|\mathbf{u}^\top \boldsymbol{\mu}_l| + |\mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}|^{\frac{1}{2}})^{2t} \\
&\leq (2t-1)!! \|\mathbf{u}\|^{2t} (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}\|^{\frac{1}{2}})^{2t},
\end{aligned}$$

where the second step is by the Binomial theorem. Hence,

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}[(\mathbf{u}^\top \mathbf{x})^{2t}] \leq (2t-1)!! \|\mathbf{u}\|^{2t} \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t} \quad (C.63)$$

*Property 7.* With the Gaussian Mixture Model, we have

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}[\|\mathbf{x}\|^{2t}] \leq d^t (2t-1)!! \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t} \quad (C.64)$$

**Proof:**

$$\begin{aligned}
& \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [||\mathbf{x}||_2^{2t}] \\
&= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\sum_{i=1}^d x_i^2)^t] \\
&= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [d^t (\sum_{i=1}^d \frac{x_i^2}{d})^t] \\
&\leq \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [d^t \sum_{i=1}^d \frac{x_i^{2t}}{d}] \\
&= d^{t-1} \sum_{i=1}^d \sum_{j=1}^L \int_{-\infty}^{\infty} (x_i - \mu_{ji} + \mu_{ji})^{2t} \lambda_j \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp(-\frac{(x_i - \mu_{ji})^2}{2\sigma_{ji}^2}) dx_i \quad (C.65) \\
&= d^{t-1} \sum_{i=1}^d \sum_{j=1}^L \sum_{k=1}^{2t} \binom{2t}{k} \lambda_j |\mu_{ji}|^{2t-k} \cdot \begin{cases} 0 & , \quad k \text{ is odd} \\ (k-1)!! \sigma_{ji}^k, & k \text{ is even} \end{cases} \\
&\leq d^{t-1} \sum_{i=1}^d \sum_{j=1}^L \sum_{k=1}^{2t} \binom{2t}{k} \lambda_j |\mu_{ji}|^{2t-k} \sigma_j^k \cdot (2t-1)!! \\
&= d^{t-1} \sum_{i=1}^d \sum_{j=1}^L \lambda_j (|\mu_{ji}| + \sigma_{ji})^{2t} (2t-1)!! \\
&\leq d^t (2t-1)!! \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t}
\end{aligned}$$

In the 3rd step, we apply Jensen inequality because  $f(x) = x^t$  is convex when  $x \geq 0$  and  $t \geq 1$ . In the 4th step we apply the Binomial theorem and the result of k-order central moment of Gaussian variable.

*Property 8.* Under the Gaussian Mixture Model  $\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$  where  $\boldsymbol{\Sigma}_l = \boldsymbol{\Lambda}_l^\top \mathbf{D}_l \boldsymbol{\Lambda}_l$ , we have the following upper bound.

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\mathbf{u}^\top \mathbf{x})^{2t}] \leq (2t-1)!! \|\mathbf{u}\|^{2t} \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t} \quad (C.66)$$

### Proof:

If  $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ , then  $\mathbf{u}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{u}^\top \boldsymbol{\mu}_l, \mathbf{u}^\top \boldsymbol{\Sigma}_l \mathbf{u}) = \mathcal{N}((\boldsymbol{\Lambda}_l \mathbf{u})^\top \boldsymbol{\Lambda}_l \boldsymbol{\mu}_l, (\boldsymbol{\Lambda}_l \mathbf{u})^\top \mathbf{D}_l (\boldsymbol{\Lambda}_l \mathbf{u}))$ . By

Property 6, we have

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}[(\mathbf{u}^\top \mathbf{x})^{2t}] \leq (2t-1)!! \|\mathbf{u}\|^{2t} (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^{2t} \quad (\text{C.67})$$

Then we can derive the final result.

*Property 9.* The population risk function  $\bar{f}(\mathbf{W})$  is defined as

$$\begin{aligned} \bar{f}(\mathbf{W}) &= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}[f_n(\mathbf{W})] \\ &= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{W}; \mathbf{x}_i, y_i) \right] \\ &= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}[\ell(\mathbf{W}; \mathbf{x}_i, y_i)] \end{aligned} \quad (\text{C.68})$$

For any permutation matrix  $\mathbf{P}$ , where  $\{\pi(j)\}_{j=1}^K$  is the indices permuted by  $\mathbf{P}$ , we have

$$\begin{aligned} H(\mathbf{W}\mathbf{P}, \mathbf{x}) &= \frac{1}{K} \sum_{\pi^*(j)} \phi(\mathbf{w}_{\pi(j)}^\top \mathbf{x}) \\ &= \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^\top \mathbf{x}) \\ &= H(\mathbf{W}, \mathbf{x}) \end{aligned} \quad (\text{C.69})$$

Therefore,

$$\bar{f}(\mathbf{W}) = \bar{f}(\mathbf{W}\mathbf{P}) \quad (\text{C.70})$$

Based on (4.1) and (4.3), we can derive its gradient and Hessian as follows.

$$\frac{\partial \ell(\mathbf{W}; \mathbf{x}, y)}{\partial \mathbf{w}_j} = -\frac{1}{K} \frac{y - H(\mathbf{W})}{H(\mathbf{W})(1 - H(\mathbf{W}))} \phi'(\mathbf{w}_j^\top \mathbf{x}) \mathbf{x} = \zeta(\mathbf{W}) \cdot \mathbf{x} \quad (\text{C.71})$$

$$\frac{\partial^2 \ell(\mathbf{W}; \mathbf{x}, y)}{\partial \mathbf{w}_j \partial \mathbf{w}_l} = \xi_{j,l} \cdot \mathbf{x} \mathbf{x}^\top \quad (\text{C.72})$$

$$\begin{aligned} &\xi_{j,l}(\mathbf{W}) \\ &= \begin{cases} \frac{1}{K^2} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \frac{H(\mathbf{W})^2 + y - 2y \cdot H(\mathbf{W})}{H^2(\mathbf{W})(1 - H(\mathbf{W}))^2}, & j \neq l \\ \frac{1}{K^2} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \frac{H(\mathbf{W})^2 + y - 2y \cdot H(\mathbf{W})}{H^2(\mathbf{W})(1 - H(\mathbf{W}))^2} - \frac{1}{K} \phi''(\mathbf{w}_j^\top \mathbf{x}) \frac{y - H(\mathbf{W})}{H(\mathbf{W})(1 - H(\mathbf{W}))}, & j = l \end{cases} \quad (\text{C.73}) \end{aligned}$$

*Property 10.* With  $D_m(\Psi)$  defined in definition C.2.4, we have

$$(i) \quad D_m(\Psi)D_{2m}(\Psi) \leq D_{3m}(\Psi) \quad (\text{C.74})$$

$$(ii) \quad (D_m(\Psi))^2 \leq D_{2m}(\Psi) \quad (\text{C.75})$$

**Proof:**

To prove (C.74), we can first compare the terms  $\sum_{i=1}^L \lambda_i a_i \sum_{i=1}^L \lambda_i a_i^2$  and  $\sum_{i=1}^L \lambda_i a_i^3$ , where  $a_i \geq 1$ ,  $i \in [L]$  and  $\sum_{i=1}^L \lambda_i = 1$ .

$$\begin{aligned} \sum_{i=1}^L \lambda_i a_i^3 - \sum_{i=1}^L \lambda_i a_i \sum_{i=1}^L \lambda_i a_i^2 &= \sum_{i=1}^L \lambda_i a_i \cdot \left( a_i^2 - \sum_{j=1}^L \lambda_j a_j^2 \right) \\ &= \sum_{i=1}^L \lambda_i a_i \cdot \left( (1 - \lambda_i) a_i^2 - \sum_{1 \leq j \leq L, j \neq i} \lambda_j a_j^2 \right) \\ &= \sum_{i=1}^L \lambda_i a_i \cdot \left( \sum_{1 \leq j \leq L, j \neq i} \lambda_j a_j^2 - \sum_{1 \leq j \leq L, j \neq i} \lambda_j a_j^2 \right) \\ &= \sum_{i=1}^L \lambda_i a_i \cdot \left( \sum_{1 \leq j \leq L, j \neq i} \lambda_j (a_i^2 - a_j^2) \right) \\ &= \sum_{1 \leq i, j \leq L, i \neq j} (\lambda_i \lambda_j a_i (a_i^2 - a_j^2) + \lambda_i \lambda_j a_j (a_j^2 - a_i^2)) \\ &= \sum_{1 \leq i, j \leq L, i \neq j} \lambda_i \lambda_j (a_i - a_j)^2 (a_i + a_j) \geq 0 \end{aligned} \quad (\text{C.76})$$

The second to last step is because we can find the pairwise terms  $\lambda_i a_i \cdot \lambda_j (a_i^2 - a_j^2)$  and  $\lambda_j a_j \cdot \lambda_i (a_j^2 - a_i^2)$  in the summation that can be putted together. From (C.76), we can obtain

$$\sum_{i=1}^L \lambda_i a_i \sum_{i=1}^L \lambda_i a_i^2 \leq \sum_{i=1}^L \lambda_i a_i^3 \quad (\text{C.77})$$

Combining (C.77) and the definition of  $D_m(\Psi)$  in (C.2.4), we can derive (C.74).

Similarly, to prove (C.75), we can first compare the terms  $(\sum_{i=1}^L \lambda_i a_i)^2$  and  $\sum_{i=1}^L \lambda_i a_i^2$ , where

$a_i \geq 1, i \in [L]$  and  $\sum_{i=1}^L \lambda_i = 1$ .

$$\begin{aligned}
\sum_{i=1}^L \lambda_i a_i^2 - (\sum_{i=1}^L \lambda_i a_i)^2 &= \sum_{i=1}^L \lambda_i a_i \cdot (a_i - \sum_{j=1}^L \lambda_j a_j) \\
&= \sum_{i=1}^L \lambda_i a_i \cdot ((1 - \lambda_i)a_i - \sum_{1 \leq j \leq L, j \neq i} \lambda_j a_j) \\
&= \sum_{i=1}^L \lambda_i a_i \cdot \left( \sum_{1 \leq j \leq L, j \neq i} \lambda_j a_i - \sum_{1 \leq j \leq L, j \neq i} \lambda_j a_j \right) \\
&= \sum_{i=1}^L \lambda_i a_i \cdot \left( \sum_{1 \leq j \leq L, j \neq i} \lambda_j (a_i - a_j) \right) \\
&= \sum_{1 \leq i, j \leq L, i \neq j} (\lambda_i \lambda_j a_i (a_i - a_j) + \lambda_i \lambda_j a_j (a_j - a_i)) \\
&= \sum_{1 \leq i, j \leq L, i \neq j} \lambda_i \lambda_j (a_i - a_j)^2 \geq 0
\end{aligned} \tag{C.78}$$

The derivation of (C.78) is close to (C.76). By (C.78) we have

$$(\sum_{i=1}^L \lambda_i a_i)^2 \leq \sum_{i=1}^L \lambda_i a_i^2 \tag{C.79}$$

Combining (C.79) and the definition of  $D_m(\Psi)$  in (C.2.4), we can derive (C.75).

### C.3 Proof of Theorem 2 and Corollary 4.4.1

Theorem 2 is built upon **three lemmas**.

**Lemma C.3.1** shows that with  $O(dK^5 \log^2 d)$  samples, the empirical risk function is strongly convex in the neighborhood of  $\mathbf{W}^*$ .

**Lemma C.3.2** shows that if initialized in the convex region, the gradient descent algorithm converges linearly to a critical point  $\widehat{\mathbf{W}}_n$ , which is close to  $\mathbf{W}^*$ .

**Lemma C.3.3** shows that the Tensor Initialization Method in Subroutine 1 initializes  $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$  in the local convex region. Theorem 1 follows naturally by combining these three lemmas.

This proving approach is built upon those in [72]. One of our major technical contribution is extending Lemmas C.3.1 and C.3.2 to the Gaussian mixture model, while the results

in [72] only apply to Standard Gaussian models. The second major contribution is a new tensor initialization method for Gaussian mixture model such that the initial point is in the convex region (see Lemma C.3.3). Both contributions require the development of new tools, and our analyses are much more involved than those for the standard Gaussian due to the complexity introduced by the Gaussian mixture model.

To present these lemmas, the Euclidean ball  $\mathbb{B}(\mathbf{W}^* \mathbf{P}^*, r)$  is used to denote the neighborhood of  $\mathbf{W}^* \mathbf{P}^*$ , where  $r$  is the radius of the ball.

$$\mathbb{B}(\mathbf{W}^* \mathbf{P}^*, r) = \{\mathbf{W} \in \mathbb{R}^{d \times K} : \|\mathbf{W} - \mathbf{W}^* \mathbf{P}^*\|_F \leq r\} \quad (\text{C.80})$$

The radius of the convex region is

$$r := \Theta\left(\frac{C_3 \epsilon_0 \cdot \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right)}{K^{\frac{7}{2}} \left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^8\right)^{\frac{1}{4}}}\right) \quad (\text{C.81})$$

with some constant  $C_3 > 0$ .

**Lemma C.3.1.** (*Strongly local convexity*) Consider the classification model with FCN (4.1) and the sigmoid activation function. There exists a constant  $C$  such that as long as the sample size

$$\begin{aligned} n \geq & C_1 \epsilon_0^{-2} \cdot \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \right)^2 \\ & \cdot \left( \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right) \right)^{-2} dK^5 \log^2 d \end{aligned} \quad (\text{C.82})$$

for some constant  $C_1 > 0$ ,  $\epsilon_0 \in (0, \frac{1}{4})$ , and any fixed permutation matrix  $\mathbf{P} \in \mathbb{R}^{K \times K}$  we have for all  $\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$ ,

$$\begin{aligned} & \Omega\left(\frac{1-2\epsilon_0}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right) \right) \cdot \mathbf{I}_{dK} \\ & \preceq \nabla^2 f_n(\mathbf{W}) \preceq C_2 \sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\|_\infty + \|\Sigma_l^{\frac{1}{2}}\|)^2 \cdot \mathbf{I}_{dK} \end{aligned} \quad (\text{C.83})$$

with probability at least  $1 - d^{-10}$  for some constant  $C_2 > 0$ .

**Lemma C.3.2.** (*Linear convergence of gradient descent*) Assume the conditions in Lemma C.3.1 hold. Given any fixed permutation matrix  $\mathbf{P} \in \mathbb{R}^{K \times K}$ , if the local convexity of  $\mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$  holds, there exists a critical point in  $\mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$  for some constant  $C_3 > 0$ , and  $\epsilon_0 \in (0, \frac{1}{2})$ , such that

$$\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F \leq O\left(\frac{K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 (1 + \xi)}}{\sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)} \sqrt{\frac{d \log n}{n}}\right) \quad (\text{C.84})$$

If the initial point  $\mathbf{W}_0 \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$ , the gradient descent linearly converges to  $\widehat{\mathbf{W}}_n$ , i.e.,

$$\begin{aligned} & \|\mathbf{W}_t - \widehat{\mathbf{W}}_n\|_F \\ & \leq \left(1 - \Omega\left(\frac{\sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)}{K^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2}\right)\right)^t \|\mathbf{W}_0 - \widehat{\mathbf{W}}_n\|_F \end{aligned} \quad (\text{C.85})$$

with probability at least  $1 - d^{-10}$ .

**Lemma C.3.3.** (*Tensor initialization*) For classification model, with  $D_6(\Psi)$  defined in Definition C.2.4, we have that if the sample size

$$n \geq \kappa^8 K^4 \tau^{12} D_6(\Psi) \cdot d \log^2 d, \quad (\text{C.86})$$

then the output  $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$  satisfies

$$\|\mathbf{W}_0 - \mathbf{W}^* \mathbf{P}^*\| \lesssim \kappa^6 K^3 \cdot \tau^6 \sqrt{D_6(\Psi)} \sqrt{\frac{d \log n}{n}} \|\mathbf{W}^*\| \quad (\text{C.87})$$

with probability at least  $1 - n^{-\Omega(\delta_1^4)}$  for a specific permutation matrix  $\mathbf{P}^* \in \mathbb{R}^{K \times K}$ .

## Proof of Theorem 2

From Lemma C.3.2 and Lemma C.3.3, we know that if  $n$  is sufficiently large such that the initialization  $\mathbf{W}_0$  by the tensor method is in the region  $\mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$ , then the gradient descent method converges to a critical point  $\widehat{\mathbf{W}}_n$  that is sufficiently close to  $\mathbf{W}^*$ . To achieve that,

one sufficient condition is

$$\begin{aligned} \|\mathbf{W}_0 - \mathbf{W}^* \mathbf{P}^*\|_F &\leq \sqrt{K} \|\mathbf{W}_0 - \mathbf{W}^* \mathbf{P}^*\| \leq \kappa^6 K^{\frac{7}{2}} \cdot \tau^6 \sqrt{D_6(\Psi)} \sqrt{\frac{d \log n}{n}} \|\mathbf{W}^* \mathbf{P}\| \\ &\leq \frac{C_3 \epsilon_0 \Gamma(\Psi) \sigma_{\max}^2}{K^{\frac{7}{2}} \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}}} \end{aligned} \quad (\text{C.88})$$

where the first inequality follows from  $\|\mathbf{W}\|_F \leq \sqrt{K} \|\mathbf{W}\|$  for  $\mathbf{W} \in \mathbb{R}^{d \times K}$ , the second inequality comes from Lemma C.3.3, and the third inequality comes from the requirement to be in the region  $\mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$ . That is equivalent to the following condition

$$\begin{aligned} n &\geq C_0 \epsilon_0^{-2} \cdot \tau^{12} \kappa^{12} K^{14} \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{2}} \\ &\quad \cdot (\delta_1(\mathbf{W}^*))^2 D_6(\Psi) \Gamma(\Psi)^{-2} \sigma_{\max}^{-4} \cdot d \log^2 d \end{aligned} \quad (\text{C.89})$$

where  $C_0 = \max\{C_4, C_3^{-2}\}$ . By Definition C.2.4, we can obtain

$$\left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{2}} \leq \sqrt{D_4(\Psi) D_8(\Psi)} \sigma_{\max}^6 \quad (\text{C.90})$$

From Property 10, we have that

$$\begin{aligned} &\sqrt{D_4(\Psi) D_8(\Psi)} D_6(\Psi) \\ &\leq \sqrt{D_{12}(\Psi)} \sqrt{D_{12}(\Psi)} = D_{12}(\Psi) \end{aligned} \quad (\text{C.91})$$

Plugging (C.90), (C.91) into (C.89), we have

$$n \geq C_0 \epsilon_0^{-2} \cdot \kappa^{12} K^{14} (\sigma_{\max} \delta_1(\mathbf{W}^*))^2 \tau^{12} \Gamma(\Psi)^{-2} D_{12}(\Psi) \cdot d \log^2 d \quad (\text{C.92})$$

Considering the requirements on the sample complexity in (C.82), (C.86), and (C.92), (C.92) shows a sufficient number of samples. Taking the union bound of all the failure probabilities in Lemma C.3.1, and C.3.3, (C.92) holds with probability  $1 - d^{-10}$ .

By Property 3.4,  $\rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)$  can be lower bounded by positive and monotonically decreasing functions  $\mathcal{L}_m\left(\frac{(\boldsymbol{\Lambda}_l \mathbf{W}^*)^\top \tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)$  when everything else except  $|\tilde{\boldsymbol{\mu}}_{l(i)}|$  is fixed, or  $\mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)$  when everything else except

$\|\Sigma_l^{\frac{1}{2}}\|$  is fixed. Then, by replacing the lower bound of  $\rho(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  with these two functions in  $\Gamma(\Psi)$ , we can have an upper bound of  $(\sigma_{\max}\delta_1(\mathbf{W}^*))^2\tau^{12}\Gamma(\Psi)^{-2}D_{12}(\Psi)$ , denoted as  $\mathcal{B}(\Psi)$ .

To be more specific, when everything else except  $|\tilde{\boldsymbol{\mu}}_{l(i)}|$  is fixed,  $\mathcal{L}_m(\frac{(\mathbf{A}_l\mathbf{W}^*)^\top\tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\cdot\|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  is plugged in  $\mathcal{B}(\Psi)$ . Then since that  $D_{12}(\Psi)$  and  $\mathcal{L}_m(\frac{(\mathbf{A}_l\mathbf{W}^*)^\top\tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\cdot\|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  are both increasing function of  $|\tilde{\boldsymbol{\mu}}_{l(i)}|$ ,  $\mathcal{B}(\Psi)$  is an increasing function of  $|\tilde{\boldsymbol{\mu}}_{l(i)}|$ .

When everything else except  $\|\Sigma_l^{\frac{1}{2}}\|$  is fixed, if  $\|\Sigma_l^{\frac{1}{2}}\| = \sigma_{\max} > \zeta_s$ , then  $\sigma_{\max}^2\tau^{12}D_{12}(\Psi)$  is an increasing function of  $\|\Sigma_l^{\frac{1}{2}}\|$ . Since that  $\mathcal{L}_s(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  is a decreasing function,  $\mathcal{L}_s(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})^{-2}$  is an increasing function of  $\|\Sigma_l^{\frac{1}{2}}\|$ . Hence,  $\mathcal{B}(\Psi)$  is an increasing function of  $\|\Sigma_l^{\frac{1}{2}}\|$ . Moreover, when all  $\|\Sigma_l^{\frac{1}{2}}\| < \zeta_{s'}$  and go to 0, two decreasing functions of  $\|\Sigma_l^{\frac{1}{2}}\|$ ,  $\sigma_{\max}^2\mathcal{L}_s(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})^{-2}$  and  $D_{12}(\Psi)$  will be the dominant term of  $\mathcal{B}(\Psi)$ . Therefore,  $\mathcal{B}(\Psi)$  increases to infinity as all  $\|\Sigma_l^{\frac{1}{2}}\|$ 's go to 0.

In sum, we can define a universe  $\mathcal{B}(\Psi)$  as:

$$\mathcal{B}(\Psi) = \begin{cases} (\sigma_{\max}\delta_1(\mathbf{W}^*))^2\tau^{12}\left(\sum_{l=1}^L \frac{\lambda_l\|\Sigma_l^{-1}\|^{-1}}{\eta\sigma_{\max}^2}\mathcal{L}_m(\frac{(\mathbf{A}_l\mathbf{W}^*)^\top\tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})\right)^{-2} \\ \cdot D_{12}(\Psi), \text{if } \mathbf{S} \text{ is fixed} \\ (\sigma_{\max}\delta_1(\mathbf{W}^*))^2\tau^{12}\left(\sum_{l=1}^L \frac{\lambda_l\|\Sigma_l^{-1}\|^{-1}}{\eta\sigma_{\max}^2}\mathcal{L}_s(\frac{(\mathbf{A}_l\mathbf{W}^*)^\top\tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})\right)^{-2} \\ \cdot D_{12}(\Psi), \text{if } \mathbf{M} \text{ is fixed} \\ (\sigma_{\max}\delta_1(\mathbf{W}^*))^2\tau^{12}\left(\sum_{l=1}^L \frac{\lambda_l\|\Sigma_l^{-1}\|^{-1}}{\eta\sigma_{\max}^2}\rho(\frac{(\mathbf{A}_l\mathbf{W}^*)^\top\tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})\right)^{-2} \\ \cdot D_{12}(\Psi), \text{otherwise} \end{cases} \quad (\text{C.93})$$

where  $\mathcal{L}_m$ ,  $\mathcal{L}_s$  and  $D_{12}$  are defined in (C.26), (C.31) and Definition C.2.4, respectively.

Hence, we have

$$n \geq \text{poly}(\epsilon_0^{-1}, \kappa, \eta, \tau K) \mathcal{B}(\Psi) \cdot d \log^2 d \quad (\text{C.94})$$

Similarly, by replacing  $\rho(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  with  $\mathcal{L}_m(\frac{(\mathbf{A}_l\mathbf{W}^*)^\top\tilde{\boldsymbol{\mu}}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  when everything else except  $|\tilde{\boldsymbol{\mu}}_{l(i)}|$  is fixed, or  $\mathcal{L}_s(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  (or  $\|\Sigma_l^{-1}\|\mathcal{L}_s(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  for  $\|\Sigma_l^{-1}\|^{-1} \geq 1$ ) when

everything else except  $\|\Sigma_l^{\frac{1}{2}}\|$  is fixed, (C.85) can also be transferred to another feasible upper bound. We denote the modified version of the convergence rate as  $v = 1 - K^{-2}q(\Psi)$ . Since that  $q(\Psi)$  is a ratio between the smallest and the largest singular value of  $\nabla^2 \bar{f}(\mathbf{W}^*)$ , we have  $q(\Psi) \in (0, 1)$ . Hence, we can obtain  $1 - K^{-2}q(\Psi) \in (0, 1)$  by  $K \geq 1$ . When everything else except  $|\tilde{\mu}_{l(i)}|$  is fixed, since that  $\mathcal{L}_m(\frac{(\Lambda_l \mathbf{W}^*)^\top \tilde{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  is monotonically decreasing and  $\sum_{l=1}^L \lambda_l (\|\mu_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2$  is increasing as  $|\tilde{\mu}_{l(i)}|$  increases,  $v$  is an increasing function of  $|\tilde{\mu}_{l(i)}|$  to 1. Similarly, when everything else except  $\|\Sigma_l^{\frac{1}{2}}\|$  is fixed where  $\|\Sigma_l^{\frac{1}{2}}\| \geq \max\{1, \zeta_s\}$ ,  $\frac{1}{\sum_{l=1}^L \lambda_l (\|\mu_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2}$  decreases to 0 as  $\|\Sigma_l\|$  increases. We replace  $\rho(\frac{\mathbf{W}^{*\top} \mu_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  by  $\|\Sigma_l^{-1}\| \mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mu_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}})$  and then

$$\begin{aligned} & \|\Sigma_l^{-1}\|^{-1} \cdot \|\Sigma_l^{-1}\| \mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mu_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}) \\ &= \mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mu_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}) \end{aligned} \quad (\text{C.95})$$

is an decreasing function less than  $\rho(\frac{\mathbf{W}^{*\top} \mu_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}})$ . Therefore,  $v$  is an increasing function of  $\|\Sigma_l^{\frac{1}{2}}\|$  to 1 when  $\|\Sigma_l^{\frac{1}{2}}\| \geq \max\{1, \zeta_s\}$ . When everything else except all  $\|\Sigma_l^{\frac{1}{2}}\| \leq \zeta_{s'}$ 's go to 0, all  $\mathcal{L}_s(\frac{\mathbf{W}^{*\top} \mu_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}})$ 's will decrease and all  $\frac{\|\Sigma_l^{-1}\|^{-1}}{\sum_{l=1}^L \lambda_l (\|\mu_l\|_\infty + \|\Sigma_l^{\frac{1}{2}}\|)^2}$ 's will decrease to 0. Therefore,  $v$  increases to 1.

$q(\Psi)$  can then be defined as

$$q(\Psi) = \begin{cases} \Omega\left(\frac{\sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2} \mathcal{L}_m\left(\frac{(\mathbf{A}_l \mathbf{W}^*)^\top \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right)}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2}\right), \\ \text{if } \mathbf{S} \text{ is fixed} \\ \Omega\left(\frac{\sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2} \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right)}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2}\right), \\ \text{if } \mathbf{M} \text{ is fixed and all } \|\Sigma_l^{\frac{1}{2}}\| \leq \zeta_{s'} \\ \Omega\left(\frac{\lambda_l \frac{1}{\eta\tau^K\kappa^2} \mathcal{L}_s\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right) + \sum_{l \neq i} r(\lambda_l, \boldsymbol{\mu}_l, \Sigma_l, \mathbf{W}^*)}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2}\right), \\ \text{if } \mathbf{M} \text{ is fixed and one } \|\Sigma_i^{\frac{1}{2}}\| \geq \max\{1, \zeta_{s'}\} \\ \Omega\left(\frac{\sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right)}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2}\right), \\ \text{otherwise} \end{cases} \quad (C.96)$$

where  $r(\lambda_l, \boldsymbol{\mu}_l, \Sigma_l, \mathbf{W}^*) = \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right)$ . Note that here the  $\rho(\cdot)$  function is defined in Definition C.2.2.  $\mathcal{L}_m(\cdot)$  and  $\mathcal{L}_s(\cdot)$  are defined in (C.26) and (C.31), respectively.

The bound of  $\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F$  is directly from (C.84). We can derive that

$$\mathcal{E}_w(\Psi) = O\left(\frac{\sqrt{\sum_{j=1}^L \lambda_j (\|\boldsymbol{\mu}_j\| + \|\Sigma_j^{\frac{1}{2}}\|)^2}}{\sum_{j=1}^L \lambda_j \|\Sigma_j^{-1}\|^{-1} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_j}{\delta_K(\mathbf{W}^*) \|\Sigma_j^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_j^{-1}\|^{-\frac{1}{2}}\right)}\right) \quad (C.97)$$

$$\mathcal{E}(\Psi) = O\left(\frac{\sum_{j=1}^L \lambda_j (\|\boldsymbol{\mu}_j\| + \|\Sigma_j^{\frac{1}{2}}\|)^2}{\sum_{j=1}^L \lambda_j \|\Sigma_j^{-1}\|^{-1} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_j}{\delta_K(\mathbf{W}^*) \|\Sigma_j^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_j^{-1}\|^{-\frac{1}{2}}\right)}\right) \quad (C.98)$$

$$\mathcal{E}_l(\Psi) = O\left(\frac{\sqrt{\sum_{j=1}^L \lambda_j (\|\boldsymbol{\mu}_j\| + \|\Sigma_j^{\frac{1}{2}}\|)^2} (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)}{\sum_{j=1}^L \lambda_j \|\Sigma_j^{-1}\|^{-1} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_j}{\delta_K(\mathbf{W}^*) \|\Sigma_j^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_j^{-1}\|^{-\frac{1}{2}}\right)}\right) \quad (C.99)$$

The discussion of the monotonicity of  $\mathcal{E}_w(\Psi)$ ,  $\mathcal{E}(\Psi)$  and  $\mathcal{E}_l(\Psi)$  can follow the analysis of  $q(\Psi)$ . We finish our proof of Theorem 2 here. The parameters  $\mathcal{B}(\Psi)$ ,  $q(\Psi)$ ,  $\mathcal{E}_w(\Psi)$ ,  $\mathcal{E}(\Psi)$ , and  $\mathcal{E}_l(\Psi)$

can be found in C.93, C.96, C.97, C.98, and C.99, respectively.

### Proof of Corollary 4.4.1:

The monotonicity analysis has been included in the proof of Theorem 2. In this part, we specify our proof for the results in Table 4.1. For simplicity, we denote  $\rho_l = \rho\left(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)$ .

When everything else except  $\|\boldsymbol{\Sigma}_l\|^{\frac{1}{2}}$  is fixed, if  $\|\boldsymbol{\Sigma}_l\| = o(1)$ , by some basic mathematical computation, then we have

$$\begin{aligned} n_{sc} &= C_0 \epsilon_0^{-2} \cdot \eta^2 \tau^{12} \kappa^{16} K^{14} \left( \sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{2}} \\ &\quad \cdot (\delta_1(\mathbf{W}^*))^2 D_6(\Psi) \left( \frac{1}{\sum_{l=1}^L \lambda_l \|\boldsymbol{\Sigma}_l^{-1}\|^{-1} \rho_l} \right)^2 \cdot d \log^2 d \\ &\lesssim \text{poly}(\epsilon_0^{-1}, \eta, \tau, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot O\left(\lambda_L \frac{1}{\|\boldsymbol{\Sigma}_L^{\frac{1}{2}}\|^6}\right) \end{aligned} \quad (\text{C.100})$$

$$\begin{aligned} v(\Psi) &= 1 - \frac{\sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \kappa^2} \rho_l}{K^2 (\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2)} \\ &\leq 1 - \frac{\lambda_L}{K^2 \eta \kappa^2 \tau^K} \Theta(\|\boldsymbol{\Sigma}_L\|^3) \end{aligned} \quad (\text{C.101})$$

$$\begin{aligned} &\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}^*\| \\ &\leq O\left(\frac{K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 (1 + \xi)}}{\sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho_l \left(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right)} \sqrt{\frac{d \log n}{n}}\right) \\ &\lesssim \text{poly}(\eta, \kappa, \tau, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O(1 - \|\boldsymbol{\Sigma}_L\|^3) \end{aligned} \quad (\text{C.102})$$

$$\begin{aligned}
\bar{f}_l(\mathbf{W}_t) &= \bar{f}_l(\mathbf{W}_t) - \bar{f}_l(\mathbf{W}^*) \\
&\leq \mathbb{E} \left[ \sum_{k=1}^K \frac{\partial(\bar{f}_l(\mathbf{W}_t))}{\partial \tilde{\mathbf{w}}_k}^\top (\mathbf{w}_{t(k)} - \mathbf{w}_k^*) \right] \\
&\leq \|\mathbf{W}_t - \mathbf{W}^* \mathbf{P}^*\| (\|\boldsymbol{\mu}_l\| + \|\Sigma_l\|^{\frac{1}{2}}) \\
&\lesssim O \left( \frac{\sum_{j=1}^L \sqrt{\lambda_j} (\|\boldsymbol{\mu}_j\| + \|\Sigma_j\|^{\frac{1}{2}})}{\sum_{j=1}^L \lambda_j \|\Sigma_j^{-1}\|^{-1} \rho_j} (\|\boldsymbol{\mu}_j\| + \|\Sigma_j\|^{\frac{1}{2}}) \cdot \sqrt{\frac{d \log n}{n}} \eta \kappa^2 K^2 (1 + \xi) \right) \quad (\text{C.103}) \\
&\lesssim \text{poly}(\eta, \kappa, \tau, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O\left(\frac{1}{1 + \|\Sigma_l\|^3}\right) \\
&\lesssim \text{poly}(\eta, \kappa, \tau, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O(1) - \Theta(\|\Sigma_l\|^3),
\end{aligned}$$

The first inequality of (C.103) is by the Mean Value Theorem. The second inequality of (C.103) is from Property 8, and the third inequality is derived from (C.84, C.85). The last inequality is obtained by the condition that  $\|\Sigma_l\| = o(1)$ . We can similarly have

$$\begin{aligned}
\bar{f}(\mathbf{W}_t) &\leq \mathbb{E} \left[ \sum_{k=1}^K \frac{\partial(\bar{f}(\mathbf{W}_t))}{\partial \tilde{\mathbf{w}}_k}^\top (\mathbf{w}_{t(k)} - \mathbf{w}_k^*) \right] \\
&\lesssim \text{poly}(\eta, \kappa, \tau, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O\left(\frac{1}{1 + \|\Sigma_l\|^3}\right) \quad (\text{C.104}) \\
&\lesssim \text{poly}(\eta, \kappa, \tau, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O(1) - \Theta(\|\Sigma_l\|^3)
\end{aligned}$$

If  $\|\Sigma_l\|^{\frac{1}{2}} = \Omega(1)$ , we have

$$n_{sc} \lesssim \text{poly}(\epsilon_0^{-1}, \eta, \tau, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot O(\|\Sigma_l\|^3) \quad (\text{C.105})$$

$$v(\Psi) \leq 1 - \frac{1}{K^2 \tau^K \eta \kappa^2} \Theta\left(\frac{1}{1 + \|\Sigma_l\|}\right) \quad (\text{C.106})$$

$$\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}^*\|_F \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^{\frac{5}{2}} (1 + \xi) \cdot \sqrt{\|\Sigma_l\|} \quad (\text{C.107})$$

$$\bar{f}_l(\mathbf{W}_t) \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot \|\Sigma_l\| \quad (\text{C.108})$$

$$\bar{f}(\mathbf{W}_t) \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot \|\Sigma_l\| \quad (\text{C.109})$$

When everything is fixed except  $\|\boldsymbol{\mu}_l\|$ , by combining (C.82) and (C.89), we have

$$n_{sc} \lesssim \text{poly}(\epsilon_0^{-1}, \eta, \tau, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot \begin{cases} O(\|\boldsymbol{\mu}_l\|^4), & \text{if } \|\boldsymbol{\mu}_l\| \leq 1 \\ O(\|\boldsymbol{\mu}_l\|^{12}), & \text{if } \|\boldsymbol{\mu}_l\| \geq 1 \end{cases} \quad (\text{C.110})$$

$$v(\Psi) \leq 1 - \frac{1}{K^2 \tau^K \eta \kappa^2} \Theta\left(\frac{1}{1 + \|\boldsymbol{\mu}_l\|^2}\right) \quad (\text{C.111})$$

$$\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}^*\|_F \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^{\frac{5}{2}} (1 + \xi) \cdot (1 + \|\boldsymbol{\mu}_l\|) \quad (\text{C.112})$$

$$\bar{f}_l(\mathbf{W}_t) \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot (1 + \|\boldsymbol{\mu}_l\|^2) \quad (\text{C.113})$$

$$\bar{f}(\mathbf{W}_t) \lesssim \text{poly}(\eta, \tau, \kappa, \delta_K(\mathbf{W}^*)) \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot (1 + \|\boldsymbol{\mu}_l\|^2) \quad (\text{C.114})$$

When everything else is fixed except  $\lambda_1, \lambda_2, \dots, \lambda_L$ , where  $\|\boldsymbol{\Sigma}_j\| = \Omega(1)$ ,  $j \in [L]$  and  $\|\boldsymbol{\mu}_j\| = \|\boldsymbol{\mu}_i\|$ ,  $i, j \in [L]$ , if  $\|\boldsymbol{\Sigma}_l\| \leq \|\boldsymbol{\Sigma}_j\|$ ,  $j \in [L]$ , we have

$$\begin{aligned} n_{sc} &\lesssim \text{poly}(\epsilon_0^{-1}, \eta, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot \frac{(a_1 \lambda_l^2 + a_2 \lambda_l^{\frac{3}{2}} + a_3 \lambda_l + a_4 \lambda_l^{\frac{1}{2}} + a_5)}{(\sum_{j=1}^L \lambda_j \rho_j)^2} \\ &\leq \text{poly}(\epsilon_0^{-1}, \eta, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot \frac{a_5}{(\sum_{j=1}^L \lambda_j \rho_j)^2} \\ &\lesssim \text{poly}(\epsilon_0^{-1}, \eta, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot O((1 + \lambda_l)^{-2}) \end{aligned} \quad (\text{C.115})$$

where  $a_1 = (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|^{\frac{1}{2}})^{12} / \|\boldsymbol{\Sigma}_l\|^3$ ,  $a_2 = (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 (\sum_{j \neq l} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^8)^{\frac{1}{2}} / \|\boldsymbol{\Sigma}_l\|^3$ ,  $a_3 = (\|\boldsymbol{\mu}_l\| / \|\boldsymbol{\Sigma}_l\|^{\frac{1}{2}} + 1)^6$ ,  $(\sum_{j \neq l} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^4 \sum_{j \neq l} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^8)^{\frac{1}{2}} + (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|^{\frac{1}{2}})^6 \sum_{j \neq l} \lambda_j (\|\boldsymbol{\mu}_j\| / \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}} + 1)^6$ ,  $a_4 = \sum_{j \neq l} \lambda_j (\|\boldsymbol{\mu}_j\| / \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}} + 1)^6 (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|^{\frac{1}{2}})^2 (\sum_{j \neq l} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^8)^{\frac{1}{2}}$ ,  $a_5 = (\sum_{j \neq l} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^4 \sum_{j \neq l} \lambda_j (\|\boldsymbol{\mu}_j\| + \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}})^8)^{\frac{1}{2}} \cdot \sum_{j \neq l} \lambda_j (\|\boldsymbol{\mu}_j\| / \|\boldsymbol{\Sigma}_j\|^{\frac{1}{2}} + 1)^6$ . The second step of (C.115) is by  $a_i = O(a_5)$ ,  $i = 1, 2, 3, 4$ .

$$v \leq \frac{1}{K^2 \eta \tau^K \kappa^2} \Theta\left(\frac{1}{1 + \lambda_l}\right) \quad (\text{C.116})$$

$$\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F \leq \text{poly}(\eta, \kappa, \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^{\frac{5}{2}} (1 + \xi) \cdot O\left(\frac{1}{1 + \sqrt{\lambda_l}}\right) \quad (\text{C.117})$$

$$\bar{f}_l(\mathbf{W}_t) \leq \text{poly}(\eta, \kappa, , \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O\left(\frac{1}{1 + \sqrt{\lambda_l}}\right) \quad (\text{C.118})$$

$$\bar{f}(\mathbf{W}_t) \leq \text{poly}(\eta, \kappa, , \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O\left(\frac{1}{1 + \lambda_l}\right) \quad (\text{C.119})$$

If  $\|\Sigma_l\| \geq \|\Sigma_j\|$ ,  $j \in [L]$ , we can similarly derive that

$$\begin{aligned} n_{sc} &\lesssim \text{poly}(\epsilon_0^{-1}, \eta, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot \frac{(a_1 \lambda_l^2 + a_2 \lambda_l^{\frac{3}{2}} + a_3 \lambda_l + a_4 \lambda_l^{\frac{1}{2}} + a_5)}{(\sum_{j=1}^L \lambda_j \rho_j)^2} \\ &\lesssim \text{poly}(\epsilon_0^{-1}, \eta, \kappa, K, \delta_1(\mathbf{W}^*)) \cdot d \log^2 d \cdot (O(1) - \Theta((1 + \lambda_l)^{-2})) \end{aligned} \quad (\text{C.120})$$

$$v \leq 1 - \frac{1}{K^2 \eta \tau^K \kappa^2} \Theta\left(\frac{1}{1 + \lambda_l}\right) \quad (\text{C.121})$$

$$\|\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}\|_F \leq \text{poly}(\eta, \kappa, , \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^{\frac{5}{2}} (1 + \xi) \cdot O(1 + \sqrt{\lambda_l}) \quad (\text{C.122})$$

$$\bar{f}_l(\mathbf{W}_t) \leq \text{poly}(\eta, \kappa, , \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot O(1 + \sqrt{\lambda_l}) \quad (\text{C.123})$$

$$\bar{f}(\mathbf{W}_t) \leq \text{poly}(\eta, \kappa, , \tau, \delta_1(\mathbf{W}^*)) \cdot \sqrt{\frac{d \log n}{n}} K^2 (1 + \xi) \cdot (O(1) - \frac{\Theta(1)}{1 + \lambda_l}) \quad (\text{C.124})$$

## C.4 Proof of Lemma C.3.1

We first state some important lemmas used in proof in Section C.4.1 and describe the proof in Section C.4.2. The proofs of these lemmas are provided in Section C.4.3 to C.4.7 in sequence. The proof idea mainly follows from [72]. Lemma C.4.3 shows the Hessian  $\nabla^2 \bar{f}(\mathbf{W})$  of the population risk function is smooth. Lemma C.4.4 illustrates that  $\nabla^2 \bar{f}(\mathbf{W})$  is strongly convex in the neighborhood around  $\mu^*$ . Lemma C.4.5 shows the Hessian of the empirical risk function  $\nabla^2 f_n(\mathbf{W}^*)$  is close to its population risk  $\nabla^2 \bar{f}(\mathbf{W}^*)$  in the local convex region. Summing up these three lemmas, we can derive the proof of Lemma C.3.1. Lemma C.4.1 is used in the proof of Lemma C.4.4. Lemma C.4.2 is used in the proof of Lemma C.4.5.

The analysis of the Hessian matrix of the population loss in [72] and [71] can not be extended to the Gaussian mixture model. To solve this problem, we develop new tools using some good properties of symmetric distribution and even function. Our approach can also be applied to other activations like tanh or erf. Moreover, if we directly apply the existing

matrix concentration inequalities in these works in bounding the error between the empirical loss and the population loss, the resulting sample complexity bound is loose and cannot reflect the influence of each component of the Gaussian mixture distribution. We develop a new version of Bernstein's inequality (see (C.196)) so that the final bound is  $O(d \log^2 d)$ .

[268] showed that the landscape of the empirical risk is close to that of the population risk when the number of samples is sufficiently large for the special case that  $K = 1$ . Focusing on Gaussian mixture models, our result explicitly shows how the parameters of the input distribution, including the proportion, mean and, variance of each component will affect the error bound between the empirical loss and the population loss in Lemma C.4.5.

#### C.4.1 Useful Lemmas in the Proof of Lemma C.3.1

**Lemma C.4.1.**

$$\mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left[ \left( \sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'(\sigma \cdot x_i) \right)^2 \right] \geq \rho(\boldsymbol{\mu}, \sigma) \|\mathbf{R}\|_F^2 , \quad (\text{C.125})$$

where  $\rho(\boldsymbol{\mu}, \sigma)$  is defined in Definition C.2.2 and  $\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_k) \in \mathbb{R}^{d \times k}$  is an arbitrary matrix.

**Lemma C.4.2.** With the FCN model (4.1) and the Gaussian Mixture Model, for any permutation matrix  $\mathbf{P}$ , for some constant  $C_{12} > 0$ , we have we have

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sup_{\mathbf{W} \neq \mathbf{W}' \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{\|\nabla^2 \ell(\mathbf{W}, \mathbf{x}) - \nabla^2 \ell(\mathbf{W}', \mathbf{x})\|}{\|\mathbf{W} - \mathbf{W}'\|_F} \right] \\ & \leq C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^4} \end{aligned} \quad (\text{C.126})$$

**Lemma C.4.3.** (Hessian smoothness of population loss) In the FCN model (4.1), for some constant  $C_5 > 0$ , for any permutation matrix  $\mathbf{P}$ , we have

$$\begin{aligned} & \|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| \\ & \leq C_5 \cdot K^{\frac{3}{2}} \cdot \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \cdot \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \end{aligned} \quad (\text{C.127})$$

**Lemma C.4.4.** (*Local strong convexity of population loss*) In the FCN model (4.1), for any permutation matrix  $\mathbf{P}$ , if  $\|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \leq r$  for an  $\epsilon_0 \in (0, \frac{1}{4})$ , then for some constant  $C_4 > 0$ ,

$$\begin{aligned} & \frac{4(1-\epsilon_0)}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right) \cdot \mathbf{I}_{dK} \\ & \preceq \nabla^2 \bar{f}(\mathbf{W}) \preceq C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \cdot \mathbf{I}_{dK} \end{aligned} \quad (\text{C.128})$$

**Lemma C.4.5.** In the FCN model (4.1), for any permutation matrix  $\mathbf{P}$ , as long as  $n \geq C' \cdot dK \log dK$  for some constant  $C' > 0$ , we have

$$\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla^2 f_n(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W})\| \leq C_6 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log n}{n}} \quad (\text{C.129})$$

with probability at least  $1 - d^{-10}$  for some constant  $C_6 > 0$ .

### C.4.2 Proof of Lemma C.3.1

From Lemma C.4.4 and C.4.5, with probability at least  $1 - d^{-10}$ ,

$$\begin{aligned} & \nabla^2 f_n(\mathbf{W}) \\ & \succeq \nabla^2 \bar{f}(\mathbf{W}) - \|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 f_n(\mathbf{W})\| \cdot \mathbf{I} \\ & \succeq \Omega\left(\frac{(1-\epsilon_0)}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right)\right) \cdot \mathbf{I} \\ & \quad - O\left(C_6 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log n}{n}}\right) \cdot \mathbf{I} \end{aligned} \quad (\text{C.130})$$

As long as the sample complexity is set to satisfy

$$\begin{aligned} & C_6 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot \sqrt{\frac{dK \log n}{n}} \\ & \leq \frac{\epsilon_0}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right) \cdot \mathbf{I} \end{aligned} \quad (\text{C.131})$$

i.e.,

$$\begin{aligned} n \geq & C_1 \epsilon_0^{-2} \cdot \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^2 \\ & \cdot \left( \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right) \cdot \mathbf{I} \right)^{-2} dK^5 \log^2 d \end{aligned} \quad (\text{C.132})$$

for some constant  $C_1 > 0$ , then we have the lower bound of the Hessian with probability at least  $1 - d^{-10}$ .

$$\nabla^2 f_n(\mathbf{W}) \succeq \Omega\left(\frac{1-2\epsilon_0}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right) \cdot \mathbf{I}\right) \quad (\text{C.133})$$

By (C.128) and (C.129), we can also derive the upper bound as follows,

$$\begin{aligned} \|\nabla^2 f_n(\mathbf{W})\| & \leq \|\nabla^2 \bar{f}(\mathbf{W})\| + \|\nabla^2 f_n(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W})\| \\ & \leq C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 + C_6 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log n}{n}} \\ & \leq C_2 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \end{aligned} \quad (\text{C.134})$$

for some constant  $C_2 > 0$ . Combining (C.133) and (C.134), we have

$$\begin{aligned} & \Omega\left(\frac{1-2\epsilon_0}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}\right) \cdot \mathbf{I}\right) \\ & \preceq \nabla^2 f_n(\mathbf{W}) \preceq C_2 \sum_{l=1}^L \lambda_l (\|\tilde{\boldsymbol{\mu}}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot \mathbf{I} \end{aligned} \quad (\text{C.135})$$

with probability at least  $1 - d^{-10}$ .

### C.4.3 Proof of Lemma C.4.1

Following the proof idea in Lemma D.4 of [71], we have

$$\mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left[ \left( \sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'(\sigma \cdot x_i) \right)^2 \right] = A_0 + B_0 \quad (\text{C.136})$$

$$A_0 = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left( \sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'^2(\sigma \cdot x_i) \cdot \mathbf{x} \mathbf{x}^\top \mathbf{r}_i \right) \quad (\text{C.137})$$

$$B_0 = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left( \sum_{i \neq l} \mathbf{r}_i^\top \phi'(\sigma \cdot x_i) \phi'(\sigma \cdot x_l) \cdot \mathbf{x} \mathbf{x}^\top \mathbf{r}_l \right) \quad (\text{C.138})$$

In  $A_0$ , we know that  $\mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} x_j = 0$ . Therefore, by some basic mathematical computation,

$$\begin{aligned} A_0 &= \sum_{i=1}^k \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left[ \mathbf{r}_i^\top \left( \phi'^2(\sigma \cdot x_i) \left( x_i^2 \mathbf{e}_i \mathbf{e}_i^\top + \sum_{j \neq i} x_i x_j (\mathbf{e}_i \mathbf{e}_j^\top \right. \right. \right. \\ &\quad \left. \left. \left. + \mathbf{e}_j \mathbf{e}_i^\top \right) + \sum_{j \neq i} \sum_{l \neq i} x_j x_l (\mathbf{e}_j \mathbf{e}_l^\top) \right) \right) \mathbf{r}_i \right] \\ &= \sum_{i=1}^k \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left[ \mathbf{r}_i^\top \left( \phi'^2(\sigma \cdot x_i) \left( x_i^2 \mathbf{e}_i \mathbf{e}_i^\top + \sum_{j \neq i} x_j^2 \mathbf{e}_j \mathbf{e}_j^\top \right) \right) \mathbf{r}_i \right] \\ &= \sum_{i=1}^k \left[ \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} [\phi'^2(\sigma \cdot x_i) x_i^2] \mathbf{r}_i^\top \mathbf{e}_i \mathbf{e}_i^\top \mathbf{r}_i \right. \\ &\quad \left. + \sum_{j \neq i} \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} [x_j^2] \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} [\phi'^2(\sigma \cdot x_i)] \mathbf{r}_i^\top \mathbf{e}_j \mathbf{e}_j^\top \mathbf{r}_i \right] \\ &= \sum_{i=1}^k r_{ii}^2 \beta_2(i, \boldsymbol{\mu}, \sigma) + \sum_{i=1}^k \sum_{j \neq i} r_{ij}^2 \beta_0(i, \boldsymbol{\mu}, \sigma) (1 + \mu_j^2) \end{aligned} \quad (\text{C.139})$$

In  $B_0$ ,  $\alpha_1(i, \boldsymbol{\mu}, \sigma) = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} (x_i \phi'(x_i)) = 0$ . By the equation in Page 30 of [71], we have

$$\begin{aligned}
B_0 &= \sum_{i \neq l}^k \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I}_d)} \left[ \mathbf{r}_i^\top \left( \phi'(\sigma \cdot x_i) \phi'(\sigma \cdot x_l) \left( x_i^2 \mathbf{e}_i \mathbf{e}_i^\top + x_l^2 \mathbf{e}_l \mathbf{e}_l^\top + x_i x_l \right. \right. \right. \\
&\quad \left. \left. \left. (\mathbf{e}_i \mathbf{e}_l^\top + \mathbf{e}_l \mathbf{e}_i^\top) + \sum_{j \neq i} x_j x_l \mathbf{e}_j \mathbf{e}_l^\top + \sum_{j \neq l} x_j x_i \mathbf{e}_j \mathbf{e}_i^\top + \sum_{j \neq i, l} \sum_{j' \neq i, l} x_j x_{j'} \mathbf{e}_j \mathbf{e}_{j'}^\top \right) \right) \mathbf{r}_l \right] \quad (\text{C.140}) \\
&= \sum_{i \neq l} r_{ii} r_{li} \alpha_2(i, \boldsymbol{\mu}, \sigma) \alpha_0(l, \boldsymbol{\mu}, \sigma) + \sum_{i \neq l} r_{ij} r_{lj} \alpha_0(i, \boldsymbol{\mu}, \sigma) \alpha_0(l, \boldsymbol{\mu}, \sigma) (1 + \mu_j^2)
\end{aligned}$$

Therefore,

$$\begin{aligned}
A_0 + B_0 &= \sum_{i=1}^k \left( r_{ii} \frac{\alpha_2(i, \boldsymbol{\mu}, \sigma)}{\sqrt{1 + \mu_i^2}} + \sum_{l \neq i} r_{li} \alpha_0(l, \boldsymbol{\mu}, \sigma) \sqrt{1 + \mu_i^2} \right)^2 - \sum_{i=1}^k r_{ii}^2 \frac{\alpha_2^2(i, \boldsymbol{\mu}, \sigma)}{1 + \mu_i^2} \\
&\quad - \sum_{i=1}^k \sum_{l \neq i} r_{li}^2 \alpha_0(l, \boldsymbol{\mu}, \sigma)^2 (1 + \mu_i^2) + \sum_{i=1}^k r_{ii}^2 \beta_2(i, \boldsymbol{\mu}, \sigma) + \sum_{i=1}^k \sum_{j \neq i} r_{ij}^2 \beta_0(i, \boldsymbol{\mu}, \sigma) \\
&\quad (1 + \mu_j^2) \quad (\text{C.141}) \\
&\geq \sum_{i=1}^k r_{ii}^2 \left( \beta_2(i, \boldsymbol{\mu}, \sigma) - \frac{\alpha_2^2(i, \boldsymbol{\mu}, \sigma)}{1 + \mu_i^2} \right) + \sum_{i=1}^k \sum_{j \neq i} r_{ij}^2 \left( \beta_0(i, \boldsymbol{\mu}, \sigma) - \alpha_0^2(i, \boldsymbol{\mu}, \sigma) \right) \\
&\quad (1 + \mu_j^2) \\
&\geq \rho(\boldsymbol{\mu}, \sigma) \|\mathbf{R}\|_F^2
\end{aligned}$$

#### C.4.4 Proof of Lemma C.4.2

Following the equation (92) in Lemma 8 of [72] and by (C.73)

$$\|\nabla^2 \ell(\mathbf{W}) - \nabla^2 \ell(\mathbf{W}')\| \leq \sum_{j=1}^K \sum_{l=1}^K |\xi_{j,l}(\mathbf{W}) - \xi_{j,l}(\mathbf{W}')| \cdot \|\mathbf{x} \mathbf{x}^\top\| \quad (\text{C.142})$$

By Lagrange's inequality, we have

$$|\xi_{j,l}(\mathbf{W}) - \xi_{j,l}(\mathbf{W}')| \leq (\max_k |T_{j,k,l}|) \cdot \|\mathbf{x}\| \cdot \sqrt{K} \|\mathbf{W} - \mathbf{W}'\|_F \quad (\text{C.143})$$

From Lemma C.4.3, we know

$$\max_k |T_{j,k,l}| \leq C_7 \quad (\text{C.144})$$

By Property 7, we have

$$\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [||\mathbf{x}||^{2t}] \leq d^t (2t-1)!! \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^{2t} \quad (\text{C.145})$$

Therefore, for some constant  $C_{12} > 0$

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sup_{\mathbf{W} \neq \mathbf{W}'} \frac{||\nabla^2 \ell(\mathbf{W}) - \nabla^2 \ell(\mathbf{W}')||}{\|\mathbf{W} - \mathbf{W}'\|_F} \right] \leq K^{\frac{5}{2}} \mathbb{E}[||\mathbf{x}||_2^3] \\ & \leq K^{\frac{5}{2}} \sqrt{d \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^2} \sqrt{3d^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^4} \\ & = C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^4} \end{aligned} \quad (\text{C.146})$$

#### C.4.5 Proof of Lemma C.4.3

Let  $\mathbf{a} = (\mathbf{a}_1^\top, \dots, \mathbf{a}_K^\top)^\top \in \mathbb{R}^{dK}$ . Let  $\Delta_{j,l} \in \mathbb{R}^{d \times d}$  be the  $(j, l)$ -th block of  $\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P}) \in \mathbb{R}^{dK \times dK}$ . By definition,

$$||\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})|| = \max_{\|\mathbf{a}\|=1} \sum_{j=1}^K \sum_{l=1}^K \mathbf{a}_j^\top \Delta_{j,l} \mathbf{a}_l \quad (\text{C.147})$$

Denote  $\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_K) \in \mathbb{R}^{K \times K}$ . By the mean value theorem and (C.73),

$$\begin{aligned} \Delta_{j,l} &= \frac{\partial^2 \bar{f}(\mathbf{W})}{\partial \mathbf{w}_j \partial \mathbf{w}_l} - \frac{\partial^2 \bar{f}(\mathbf{W}^* \mathbf{P})}{\partial \mathbf{w}_j^* \partial \mathbf{w}_l^*} = \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \sigma_l^2 \mathbf{I}_d)} [(\xi_{j,l}(\mathbf{W}) - \xi_{j,l}(\mathbf{W}^* \mathbf{P})) \cdot \mathbf{x} \mathbf{x}^\top] \\ &= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sum_{k=1}^K \left\langle \frac{\partial \xi_{j,l}(\mathbf{W}')}{\partial \mathbf{w}'_k}, \mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k \right\rangle \cdot \mathbf{x} \mathbf{x}^\top \right] \\ &= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sum_{k=1}^K \langle T_{j,l,k} \cdot \mathbf{x}, \mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k \rangle \cdot \mathbf{x} \mathbf{x}^\top \right] \end{aligned} \quad (\text{C.148})$$

where  $\mathbf{W}' = \gamma \mathbf{W} + (1-\gamma) \mathbf{W}^* \mathbf{P}$  for some  $\gamma \in (0, 1)$  and  $T_{j,l,k}$  is defined such that  $\frac{\partial \xi_{j,l}(\mathbf{W}')}{\partial \mathbf{w}'_k} = T_{j,l,k} \cdot \mathbf{x} \in \mathbb{R}^d$ . Then we provide an upper bound for  $\xi_{j,l}$ . Since that  $y = 1$  or  $0$ , we first

compute the case in which  $y = 1$ . From (C.73) we can obtain

$$\xi_{j,l}(\mathbf{W}) = \begin{cases} \frac{1}{K^2} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \cdot \frac{1}{H^2(\mathbf{W})}, & j \neq l \\ \frac{1}{K^2} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \cdot \frac{1}{H^2(\mathbf{W})} - \frac{1}{K} \phi''(\mathbf{w}_j^\top \mathbf{x}) \cdot \frac{1}{H(\mathbf{W})}, & j = l \end{cases} \quad (\text{C.149})$$

We can bound  $\xi_{j,l}(\mathbf{W})$  by bounding each component of (C.149). Note that we have

$$\begin{aligned} & \frac{1}{K^2} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \cdot \frac{1}{H^2(\mathbf{W})} \\ & \leq \frac{1}{K^2} \frac{\phi(\mathbf{w}_j^\top \mathbf{x}) \phi(\mathbf{w}_l^\top \mathbf{x})(1 - \phi(\mathbf{w}_j^\top \mathbf{x}))(1 - \phi(\mathbf{w}_l^\top \mathbf{x}))}{\frac{1}{K^2} \phi(\mathbf{w}_j^\top \mathbf{x}) \phi(\mathbf{w}_l^\top \mathbf{x})} \leq 1 \end{aligned} \quad (\text{C.150})$$

$$\frac{1}{K} \phi''(\mathbf{w}_j^\top \mathbf{x}) \cdot \frac{1}{H(\mathbf{W})} \leq \frac{1}{K} \frac{\phi(\mathbf{w}_j^\top \mathbf{x})(1 - \phi(\mathbf{w}_j^\top \mathbf{x}))(1 - 2\phi(\mathbf{w}_j^\top \mathbf{x}))}{\frac{1}{K} \phi(\mathbf{w}_j^\top \mathbf{x})} \leq 1 \quad (\text{C.151})$$

where (C.150) holds for any  $j, l \in [K]$ . The case  $y = 0$  can be computed with the same upper bound by substituting  $(1 - H(\mathbf{W})) = \frac{1}{K} \sum_{j=1}^K (1 - \phi(\mathbf{w}_j^\top \mathbf{x}))$  for  $H(\mathbf{W})$  in (C.149), (C.150) and (C.151). Therefore, there exists a constant  $C_9 > 0$ , such that

$$|\xi_{j,l}(\mathbf{W})| \leq C_9 \quad (\text{C.152})$$

We then need to calculate  $T_{j,l,k}$ . Following the analysis of  $\xi_{j,l}(\mathbf{W})$ , we only consider the case of  $y = 1$  here for simplicity.

$$T_{j,l,k} = \frac{-2}{K^3 H^3(\mathbf{W}')} \phi'(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_l^\top \mathbf{x}) \phi'(\mathbf{w}'_k^\top \mathbf{x}), \quad (\text{C.153})$$

where  $j, l, k$  are not equal to each other, and

$$\begin{aligned} & T_{j,j,k} \\ &= \begin{cases} \frac{-2}{K^3 H^3(\mathbf{W}')} \phi'(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_k^\top \mathbf{x}) + \frac{1}{K^2 H^2(\mathbf{W}')} \phi''(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_k^\top \mathbf{x}), & j \neq k \\ \frac{-2}{K^3 H^3(\mathbf{W}')} (\phi'(\mathbf{w}'_j^\top \mathbf{x}))^3 + \frac{3}{K^2 H^2(\mathbf{W}')} \phi''(\mathbf{w}'_j^\top \mathbf{x}) \phi'(\mathbf{w}'_j^\top \mathbf{x}) - \frac{\phi'''(\mathbf{w}'_j^\top \mathbf{x})}{K H(\mathbf{W}')}, & j = k \end{cases} \quad (\text{C.154}) \end{aligned}$$

$$\begin{aligned}
& \mathbf{a}_j^\top \Delta_{j,l} \mathbf{a}_l \\
&= \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{k=1}^K T_{j,l,k} \langle \mathbf{x}, \mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k \rangle \right) \cdot (\mathbf{a}_j^\top \mathbf{x}) (\mathbf{a}_l^\top \mathbf{x}) \right] \\
&\leq \sqrt{\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sum_{k=1}^K T_{j,k,l}^2 \right] \cdot \mathbb{E} \left[ \sum_{k=1}^K (\langle \mathbf{x}, \mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k \rangle (\mathbf{a}_j^\top \mathbf{x}) (\mathbf{a}_l^\top \mathbf{x}))^2 \right]} \\
&\leq \sqrt{\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sum_{k=1}^K T_{j,k,l}^2 \right]} \sqrt{\sum_{k=1}^K \sqrt{\mathbb{E}((\mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k)^\top \mathbf{x})^4} \sqrt{\mathbb{E}[(\mathbf{a}_j^\top \mathbf{x})^4 (\mathbf{a}_l^\top \mathbf{x})^4]}} \quad (\text{C.155}) \\
&\leq C_8 \sqrt{\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sum_{k=1}^K T_{j,k,l}^2 \right]} \sqrt{\sum_{k=1}^K \|\mathbf{w}_k - \mathbf{W}^* \mathbf{p}_k\|_2^2 \cdot \|\mathbf{a}_j\|_2^2 \cdot \|\mathbf{a}_l\|_2^2} \\
&\quad \cdot \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{1/2}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{1/2}\|)^8 \right)^{1/4}
\end{aligned}$$

for some constant  $C_8 > 0$ . All the three inequalities of (C.155) are derived from Cauchy-Schwarz inequality. Note that we have

$$\begin{aligned}
& \left| \frac{-2}{K^3 H^3(\mathbf{W})} (\phi'(\mathbf{w}_j^\top \mathbf{x}))^2 \phi'(\mathbf{w}_k^\top \mathbf{x}) \right| \\
&\leq \frac{2\phi^2(\mathbf{w}_j^\top \mathbf{x})(1 - \phi(\mathbf{w}_j^\top \mathbf{x}))^2 \phi(\mathbf{w}_k^\top \mathbf{x})(1 - \phi(\mathbf{w}_k^\top \mathbf{x}))}{K^3 \frac{1}{K^3} \phi^2(\mathbf{w}_j^\top \mathbf{x}) \phi(\mathbf{w}_k^\top \mathbf{x})} \quad (\text{C.156}) \\
&= 2(1 - \phi(\mathbf{w}_j^\top \mathbf{x}))^2 (1 - \phi(\mathbf{w}_k^\top \mathbf{x})) \leq 2
\end{aligned}$$

$$\left| \frac{-2}{K^3 H^3(\mathbf{W})} \phi'(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_l^\top \mathbf{x}) \phi'(\mathbf{w}_k^\top \mathbf{x}) \right| \leq 2 \quad (\text{C.157})$$

$$\begin{aligned}
& \left| \frac{3}{K^2 H^2(\mathbf{W})} \phi''(\mathbf{w}_j^\top \mathbf{x}) \phi'(\mathbf{w}_k^\top \mathbf{x}) \right| \\
&\leq \left| \frac{3\phi(\mathbf{w}_j^\top \mathbf{x})(1 - \phi(\mathbf{w}_j^\top \mathbf{x}))(1 - 2\phi(\mathbf{w}_j^\top \mathbf{x})) \phi(\mathbf{w}_k^\top \mathbf{x})(1 - \phi(\mathbf{w}_k^\top \mathbf{x}))}{K^2 \frac{1}{K^2} \phi(\mathbf{w}_j^\top \mathbf{x}) \phi(\mathbf{w}_k^\top \mathbf{x})} \right| \quad (\text{C.158}) \\
&= \left| 3(1 - \phi(\mathbf{w}_j^\top \mathbf{x}))(1 - 2\phi(\mathbf{w}_j^\top \mathbf{x}))(1 - \phi(\mathbf{w}_k^\top \mathbf{x})) \right| \leq 3
\end{aligned}$$

$$\left| \frac{\phi'''(\mathbf{w}_j^\top \mathbf{x})}{K H(\mathbf{W})} \right| \leq \left| \frac{\phi(\mathbf{w}_j^\top \mathbf{x})(1 - \phi(\mathbf{w}_j^\top \mathbf{x}))(1 - 6\phi(\mathbf{w}_j^\top \mathbf{x}) + 6\phi^2(\mathbf{w}_j^\top \mathbf{x}))}{K \frac{1}{K} \phi(\mathbf{w}_j^\top \mathbf{x})} \right| \leq 1 \quad (\text{C.159})$$

Therefore, by combining (C.153), (C.154) and (C.156) to (C.159), we have

$$|T_{j,l,k}| \leq C_7 \quad \Rightarrow \quad T_{j,l,k}^2 \leq C_7^2, \forall j, l, k \in [K], \quad (\text{C.160})$$

for some constants  $C_7 > 0$ . By (C.147), (C.148), (C.155), (C.160) and the Cauchy-Schwarz's Inequality, we have

$$\begin{aligned} & \|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| \\ & \leq C_8 \sqrt{C_7^2 K} \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \\ & \quad \cdot \max_{\|\mathbf{a}\|=1} \sum_{j=1}^K \sum_{l=1}^K \|\mathbf{a}_j\|_2 \|\mathbf{a}_l\|_2 \\ & \leq C_8 \sqrt{C_7^2 K} \cdot \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \cdot \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \quad (\text{C.161}) \\ & \quad \cdot \left( \sum_{j=1}^K \|\mathbf{a}_j\| \right)^2 \\ & \leq C_8 \sqrt{C_7^2 K^3} \cdot \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \cdot \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \end{aligned}$$

Hence, we have

$$\begin{aligned} & \|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| \\ & \leq C_5 K^{\frac{3}{2}} \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \|\mathbf{W} - \mathbf{W}^* \mathbf{P}\|_F \quad (\text{C.162}) \end{aligned}$$

for some constant  $C_5 > 0$ .

#### C.4.6 Proof of Lemma C.4.4

From [72], we know

$$\begin{aligned} \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P}) &\succeq \min_{\|\mathbf{a}\|=1} \frac{4}{K^2} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{j=1}^K \phi'(\mathbf{w}_{\pi^*(j)}^* \mathbf{x}) (\mathbf{a}_{\pi^*(j)}^\top \mathbf{x}) \right)^2 \right] \cdot \mathbf{I}_{dK} \\ &= \min_{\|\mathbf{a}\|=1} \frac{4}{K^2} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{j=1}^K \phi'(\mathbf{w}_j^{*\top} \mathbf{x}) (\mathbf{a}_j^\top \mathbf{x}) \right)^2 \right] \cdot \mathbf{I}_{dK} \end{aligned} \quad (\text{C.163})$$

with  $\mathbf{a} = (\mathbf{a}_1^\top, \dots, \mathbf{a}_K^\top)^\top \in \mathbb{R}^{dK}$ , where  $\mathbf{P}$  is a specific permutation matrix and  $\{\pi^*(j)\}_{j=1}^K$  is the indices permuted by  $\mathbf{P}$ . Similarly,

$$\begin{aligned} &\nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P}) \\ &\preceq \left( \max_{\|\mathbf{a}\|=1} \mathbf{a}^\top \nabla^2 \bar{f}(\mathbf{W}^*) \mathbf{a} \right) \cdot \mathbf{I}_{dK} \preceq C_4 \cdot \max_{\|\mathbf{a}\|=1} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sum_{j=1}^K (\mathbf{a}_{\pi^*(j)}^\top \mathbf{x})^2 \right] \\ &\quad \cdot \mathbf{I}_{dK} \\ &= C_4 \cdot \max_{\|\mathbf{a}\|=1} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sum_{j=1}^K (\mathbf{a}_j^\top \mathbf{x})^2 \right] \cdot \mathbf{I}_{dK} \end{aligned} \quad (\text{C.164})$$

for some constant  $C_4 > 0$ . By applying Property 8, we can derive the upper bound in (C.164) as

$$C_4 \cdot \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sum_{j=1}^K (\mathbf{a}_j^\top \mathbf{x})^2 \right] \cdot \mathbf{I}_{dK} \preceq C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot \mathbf{I}_{dK} \quad (\text{C.165})$$

To find a lower bound for (C.163), we can first transfer the expectation of the Gaussian Mixture Model to the weight sum of the expectations over general Gaussian distributions.

$$\begin{aligned} &\min_{\|\mathbf{a}\|=1} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{j=1}^K \phi'(\mathbf{w}_j^{*\top} \mathbf{x}) (\mathbf{a}_j^\top \mathbf{x}) \right)^2 \right] \\ &= \min_{\|\mathbf{a}\|=1} \sum_{l=1}^L \lambda_l \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{j=1}^K \phi'(\mathbf{w}_j^{*\top} \mathbf{x}) (\mathbf{a}_j^\top \mathbf{x}) \right)^2 \right] \end{aligned} \quad (\text{C.166})$$

Denote  $\mathbf{U} \in \mathbb{R}^{d \times k}$  as the orthogonal basis of  $\mathbf{W}^*$ . For any vector  $\mathbf{a}_i \in \mathbb{R}^d$ , there exists two vectors  $\mathbf{b}_i \in \mathbb{R}^K$  and  $\mathbf{c}_i \in \mathbb{R}^{d-K}$  such that

$$\mathbf{a}_i = \mathbf{U}\mathbf{b}_i + \mathbf{U}_\perp \mathbf{c}_i \quad (\text{C.167})$$

where  $\mathbf{U}_\perp \in \mathbb{R}^{d \times (d-K)}$  denotes the complement of  $\mathbf{U}$ . We also have  $\mathbf{U}_\perp^\top \boldsymbol{\mu}_l = 0$  by Property 1. Plugging (C.167) into RHS of (C.166), and then we have

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{i=1}^K \mathbf{a}_i^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right)^2 \right] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{i=1}^K (\mathbf{U}\mathbf{b}_i + \mathbf{U}_\perp \mathbf{c}_i)^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right)^2 \right] = A + B + C \end{aligned} \quad (\text{C.168})$$

$$A = \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{i=1}^K \mathbf{b}_i^\top \mathbf{U}^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right)^2 \right] \quad (\text{C.169})$$

$$\begin{aligned} C &= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ 2 \left( \sum_{i=1}^K \mathbf{c}_i^\top \mathbf{U}_\perp^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right) \cdot \left( \sum_{i=1}^K \mathbf{b}_i^\top \mathbf{U}^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right) \right] \\ &= \sum_{i=1}^K \sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ 2 \mathbf{c}_i^\top \mathbf{U}_\perp^\top \mathbf{x} \right] \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \mathbf{b}_i^\top \mathbf{U}^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \phi'(\mathbf{w}_j^{*\top} \mathbf{x}) \right] \\ &= \sum_{i=1}^K \sum_{j=1}^K \left[ 2 \mathbf{c}_i^\top \mathbf{U}_\perp^\top \boldsymbol{\mu}_l \right] \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \mathbf{b}_i^\top \mathbf{U}^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \phi'(\mathbf{w}_j^{*\top} \mathbf{x}) \right] = 0 \end{aligned} \quad (\text{C.170})$$

where the last step is by  $\mathbf{U}_\perp^\top \boldsymbol{\mu}_l = 0$  by Property 1.

$$\begin{aligned} B &= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{i=1}^K \mathbf{c}_i^\top \mathbf{U}_\perp^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right)^2 \right] \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [(\mathbf{t}^\top \mathbf{s})^2] \\ &= \sum_{i=1}^K \mathbb{E}[t_i^2 s_i^2] + \sum_{i \neq j} \mathbb{E}[t_i t_j s_i s_j] \\ &= \sum_{i=1}^K \mathbb{E}[t_i^2] \sum_{k=1}^d (\mathbf{U}_\perp)_{ik}^2 \sigma_{lk}^2 + \left( \sum_{i=1}^K \mathbb{E}[t_i^2] (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_i^2 + \sum_{i \neq j} \mathbb{E}[t_i t_j] (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_i \cdot (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_j \right) \\ &= \mathbb{E} \left[ \sum_{i=1}^{d-K} t_i^2 \cdot \sum_{k=1}^d (\mathbf{U}_\perp)_{ik}^2 \sigma_{lk}^2 \right] + \mathbb{E}[(\mathbf{t}^\top \mathbf{U}_\perp^\top \boldsymbol{\mu}_l)^2] = \mathbb{E} \left[ \sum_{i=1}^{d-K} t_i^2 \cdot \sum_{k=1}^d (\mathbf{U}_\perp)_{ik}^2 \sigma_{lk}^2 \right], \end{aligned} \quad (\text{C.171})$$

where the second equality is by defining  $\mathbf{t} = \sum_{i=1}^k \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \mathbf{c}_i \in \mathbb{R}^{d-K}$  and  $\mathbf{s} = \mathbf{U}_\perp^\top \mathbf{x}$ . The last step is by  $\mathbf{U}_\perp^\top \boldsymbol{\mu}_l = 0$ . The 4th step is because that  $s_i$  is independent of  $t_i$ , thus

$$\mathbb{E}[t_i t_j s_i s_j] = \mathbb{E}[t_i t_j] \mathbb{E}[s_i s_j]$$

$$\mathbb{E}[s_i s_j] = \begin{cases} (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_i \cdot (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_j, & \text{if } i \neq j \\ (\mathbf{U}_\perp^\top \boldsymbol{\mu}_l)_i^2 + \sum_{k=1}^d (\mathbf{U}_\perp)_{ik}^2 \sigma_{lk}^2, & \text{if } i = j \end{cases} \quad (\text{C.172})$$

Since  $\left(\sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'(\sigma \cdot x_i)\right)^2$  is an even function for any  $\mathbf{r}_i \in \mathbb{R}^d$ ,  $i \in [k]$ , so from Property 5 we have

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'(\sigma \cdot x_i) \right)^2 \right] = \mathbb{E}_{\mathbf{x} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{i=1}^k \mathbf{r}_i^\top \mathbf{x} \cdot \phi'(\sigma \cdot x_i) \right)^2 \right] \quad (\text{C.173})$$

Combining Lemma C.4.1 and Property 5, we next follow the derivation for the standard Gaussian distribution in Page 36 of [71] and generalize the result to a Gaussian distribution with an arbitrary mean and variance as follows.

$$\begin{aligned}
A &= \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{i=1}^K \mathbf{b}_i^\top \mathbf{U}^\top \mathbf{x} \cdot \phi'(\mathbf{w}_i^{*\top} \mathbf{x}) \right)^2 \right] \\
&\geq \int (2\pi)^{-\frac{K}{2}} |\mathbf{U}^\top \boldsymbol{\Sigma}_l \mathbf{U}|^{-\frac{1}{2}} \left[ \left( \sum_{i=1}^K \mathbf{b}_i^\top \mathbf{z} \cdot \phi'(\mathbf{v}_i^\top \mathbf{z}) \right)^2 \right] \exp \left( -\frac{1}{2} \|\boldsymbol{\Sigma}_l^{-1}\| \|\mathbf{z} - \mathbf{U}^\top \boldsymbol{\mu}_l\|^2 \right) d\mathbf{z} \\
&= \int (2\pi)^{-\frac{K}{2}} |\mathbf{U}^\top \boldsymbol{\Sigma}_l \mathbf{U}|^{-\frac{1}{2}} \left[ \left( \sum_{i=1}^K \mathbf{b}_i^\top \mathbf{V}^{\dagger\top} \mathbf{s} \cdot \phi'(s_i) \right)^2 \right] \\
&\quad \cdot \exp \left( -\frac{1}{2} \|\boldsymbol{\Sigma}_l^{-1}\| \|\mathbf{V}^{\dagger\top} \mathbf{s} - \mathbf{U}^\top \boldsymbol{\mu}_l\|^2 \right) \left| \cdot \det(\mathbf{V}^\dagger) \right| d\mathbf{s} \\
&\geq \int (2\pi)^{-\frac{K}{2}} |\mathbf{U}^\top \boldsymbol{\Sigma}_l \mathbf{U}|^{-\frac{1}{2}} \left[ \left( \sum_{i=1}^k \mathbf{b}_i^\top \mathbf{V}^{\dagger\top} \mathbf{s} \cdot \phi'(s_i) \right)^2 \right] \\
&\quad \cdot \exp \left( -\frac{\|\boldsymbol{\Sigma}_l^{-1}\| \|\mathbf{s} - \mathbf{V}^\top \mathbf{U}^\top \boldsymbol{\mu}_l\|^2}{2\delta_K^2(\mathbf{W}^*)} \right) \left| \cdot \det(\mathbf{V}^\dagger) \right| d\mathbf{s} \\
&\geq \int (2\pi)^{-\frac{K}{2}} |\mathbf{U}^\top \boldsymbol{\Sigma}_l \mathbf{U}|^{-\frac{1}{2}} \left[ \left( \sum_{i=1}^k \mathbf{b}_i^\top \mathbf{V}^{\dagger\top} (\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}) \mathbf{g} \cdot \phi'(\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} g_i) \right)^2 \right] \\
&\quad \cdot \exp \left( -\frac{\|\mathbf{g} - \frac{\sqrt{\|\boldsymbol{\Sigma}_l^{-1}\| \mathbf{W}^{*\top} \boldsymbol{\mu}_l}}{\delta_K(\mathbf{W}^*)}\|^2}{2} \right) \left| \det(\mathbf{V}^\dagger) \right| \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{K}{2}} \delta_K^K(\mathbf{W}^*) d\mathbf{g} \\
&= \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\tau^K \eta} \mathbb{E}_{\mathbf{g}} \left[ \left( \sum_{i=1}^K (\mathbf{b}_i^\top \mathbf{V}^{\dagger\top} \delta_K(\mathbf{W}^*)) \mathbf{g} \cdot \phi'(\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*) \cdot g_i) \right)^2 \right] \\
&\geq \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\tau^K \kappa^2 \eta} \rho \left( \frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*)}, \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*) \right) \|\mathbf{b}\|^2.
\end{aligned} \tag{C.174}$$

The second step is by letting  $\mathbf{z} = \mathbf{U}^\top \mathbf{x} \sim \mathcal{N}(\mathbf{U}^\top \boldsymbol{\mu}_l, \mathbf{U}^\top \boldsymbol{\Sigma} \mathbf{U})$ ,  $\mathbf{y}^\top \mathbf{U}^\top \boldsymbol{\Sigma}_l^{-1} \mathbf{U} \mathbf{y} \leq \|\boldsymbol{\Sigma}_l^{-1}\| \|\mathbf{y}\|^2$  for any  $\mathbf{y} \in \mathbb{R}^K$ . The third step is by letting  $\mathbf{s} = \mathbf{V}^\top \mathbf{z}$ . The last to second step follows from  $\mathbf{g} = \frac{\mathbf{s}}{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*)}$ , where  $\mathbf{g} \sim \mathcal{N}(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*)}, \mathbf{I}_K)$  and the last inequality is by Lemma C.4.1. Similarly, we extend the derivation in Page 37 of [71] for the standard Gaussian distribution to a general Gaussian distribution as follows.

$$\begin{aligned}
B &= \sum_{k=1}^d (\mathbf{U}_\perp)_{ik}^2 \sigma_{lk}^2 \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\|\mathbf{t}\|^2] \\
&\geq \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left( \frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*)}, \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \delta_K(\mathbf{W}^*) \right) \|\mathbf{c}\|^2
\end{aligned} \tag{C.175}$$

Combining (C.168) - (C.171), (C.174) and (C.175), we have

$$\begin{aligned} & \min_{\|\boldsymbol{a}\|=1} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{i=1}^k \boldsymbol{a}_i^\top \boldsymbol{x} \cdot \phi'(\boldsymbol{w}_i^{*\top} \boldsymbol{x}) \right)^2 \right] \\ & \geq \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left( \frac{\boldsymbol{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\boldsymbol{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \right). \end{aligned} \quad (\text{C.176})$$

For the Gaussian Mixture Model  $\boldsymbol{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma})$ , we have

$$\begin{aligned} & \min_{\|\boldsymbol{a}\|=1} \mathbb{E}_{\boldsymbol{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{i=1}^k \boldsymbol{a}_i^\top \boldsymbol{x} \cdot \phi'(\boldsymbol{w}_i^{*\top} \boldsymbol{x}) \right)^2 \right] \\ & \geq \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left( \frac{\boldsymbol{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\boldsymbol{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \right) \end{aligned} \quad (\text{C.177})$$

Therefore,

$$\begin{aligned} & \frac{4}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left( \frac{\boldsymbol{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\boldsymbol{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \right) \cdot \boldsymbol{I}_{dK} \\ & \preceq \nabla^2 \bar{f}(\boldsymbol{W}^* \boldsymbol{P}) \preceq C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot \boldsymbol{I}_{dK} \end{aligned} \quad (\text{C.178})$$

From (C.127) in Lemma C.4.3, since that we have the condition  $\|\boldsymbol{W} - \boldsymbol{W}^* \boldsymbol{P}\|_F \leq r$  and (C.81), we can obtain

$$\begin{aligned} & \|\nabla^2 \bar{f}(\boldsymbol{W}) - \nabla^2 \bar{f}(\boldsymbol{W}^* \boldsymbol{P})\| \\ & \leq C_5 K^{\frac{3}{2}} \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^8 \right)^{\frac{1}{4}} \|\boldsymbol{W} - \boldsymbol{W}^* \boldsymbol{P}\|_F \\ & \leq \frac{4\epsilon_0}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left( \frac{\boldsymbol{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\boldsymbol{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \right), \end{aligned} \quad (\text{C.179})$$

where  $\epsilon_0 \in (0, \frac{1}{4})$ . Then we have

$$\begin{aligned} & \|\nabla^2 \bar{f}(\boldsymbol{W})\| \geq \|\nabla^2 \bar{f}(\boldsymbol{W}^* \boldsymbol{P})\| - \|\nabla^2 \bar{f}(\boldsymbol{W}) - \nabla^2 \bar{f}(\boldsymbol{W}^* \boldsymbol{P})\| \\ & \geq \frac{4(1-\epsilon_0)}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left( \frac{\boldsymbol{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\boldsymbol{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\boldsymbol{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \right) \end{aligned} \quad (\text{C.180})$$

$$\begin{aligned}
\|\nabla^2 \bar{f}(\mathbf{W})\| &\leq \|\nabla^2 \bar{f}(\mathbf{W}^*)\| + \|\nabla^2 \bar{f}(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W}^* \mathbf{P})\| \\
&\leq C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 + \frac{4}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left( \frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-\frac{1}{2}}\|}, \right. \\
&\quad \left. \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-\frac{1}{2}}\| \right) \\
&\lesssim C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2
\end{aligned} \tag{C.181}$$

The last inequality of (C.181) holds since  $C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 = \Omega(\max_l \{\|\boldsymbol{\Sigma}_l\|\})$ ,  $\frac{4}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left( \frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-\frac{1}{2}}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \right) = O\left(\frac{\max_l \{\|\boldsymbol{\Sigma}_l\|\}}{K^2}\right)$  and  $\Omega(\max_l \{\|\boldsymbol{\Sigma}_l\|\}) \geq O\left(\frac{\max_l \{\|\boldsymbol{\Sigma}_l\|\}}{K^2}\right)$ . Combining (C.180) and (C.181), we have

$$\begin{aligned}
&\frac{4(1-\epsilon_0)}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\boldsymbol{\Sigma}_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho \left( \frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-\frac{1}{2}}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\boldsymbol{\Sigma}_l^{-1}\|^{-\frac{1}{2}} \right) \cdot \mathbf{I} \\
&\preceq \nabla^2 \bar{f}(\mathbf{W}) \preceq C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \sigma_l)^2 \cdot \mathbf{I}
\end{aligned} \tag{C.182}$$

#### C.4.7 Proof of Lemma C.4.5

Let  $N_\epsilon$  be the  $\epsilon$ -covering number of the Euclidean ball  $\mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$ . It is known that  $\log N_\epsilon \leq dK \log(\frac{3r}{\epsilon})$  from [258]. Let  $\mathcal{W}_\epsilon = \{\mathbf{W}_1, \dots, \mathbf{W}_{N_\epsilon}\}$  be the  $\epsilon$ -cover set with  $N_\epsilon$  elements. For any  $\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$ , let  $j(\mathbf{W}) = \arg \min_{j \in [N_\epsilon]} \|\mathbf{W} - \mathbf{W}_{j(\mathbf{W})}\|_F \leq \epsilon$  for all  $\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$ .

Then for any  $\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)$ , we have

$$\begin{aligned}
&\|\nabla^2 f_n(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W})\| \\
&\leq \frac{1}{n} \left\| \sum_{i=1}^n [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)] \right\| \\
&\quad + \left\| \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)] \right\| \\
&\quad + \left\| \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)] - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right\|
\end{aligned} \tag{C.183}$$

Hence, we have

$$\mathbb{P}\left(\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla^2 f_n(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W})\| \geq t\right) \leq \mathbb{P}(A_t) + \mathbb{P}(B_t) + \mathbb{P}(C_t) \quad (\text{C.184})$$

where  $A_t$ ,  $B_t$  and  $C_t$  are defined as

$$A_t = \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{1}{n} \left\| \sum_{i=1}^n [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \quad (\text{C.185})$$

$$\begin{aligned} B_t = & \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left\| \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i) \right. \right. \\ & \left. \left. - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \end{aligned} \quad (\text{C.186})$$

$$\begin{aligned} C_t = & \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left\| \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)] \right. \right. \\ & \left. \left. - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \end{aligned} \quad (\text{C.187})$$

Then we bound  $\mathbb{P}(A_t)$ ,  $\mathbb{P}(B_t)$ , and  $\mathbb{P}(C_t)$  separately.

1) **Upper bound on  $\mathbb{P}(B_t)$ .** By Lemma 6 in [72], we obtain

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right\| \\ & \leq 2 \sup_{\mathbf{v} \in \mathbf{V}_{\frac{1}{4}}} \left| \left\langle \mathbf{v}, \left( \frac{1}{n} \sum_{i=1}^n \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right) \mathbf{v} \right\rangle \right| \end{aligned} \quad (\text{C.188})$$

where  $\mathbf{V}_{\frac{1}{4}}$  is a  $\frac{1}{4}$ -cover of the unit-Euclidean-norm ball  $\mathbb{B}(\mathbf{0}, 1)$  with  $\log |\mathbf{V}_{\frac{1}{4}}| \leq dK \log 12$ .

Taking the union bound over  $\mathcal{W}_\epsilon$  and  $\mathbf{V}_{\frac{1}{4}}$ , we have

$$\begin{aligned} \mathbb{P}(B_t) & \leq \mathbb{P}\left(\sup_{\mathbf{W} \in \mathcal{W}_\epsilon, \mathbf{v} \in \mathbf{V}_{\frac{1}{4}}} \left| \frac{1}{n} \sum_{i=1}^n G_i \right| \geq \frac{t}{6} \right) \\ & \leq \exp(dK(\log \frac{3r}{\epsilon} + \log 12)) \sup_{\mathbf{W} \in \mathcal{W}_\epsilon, \mathbf{v} \in \mathbf{V}_{\frac{1}{4}}} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n G_i \right| \geq \frac{t}{6} \right) \end{aligned} \quad (\text{C.189})$$

where  $G_i = \left\langle \mathbf{v}, (\nabla^2 \ell(\mathbf{W}, \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}, \mathbf{x}_i)] \mathbf{v}) \right\rangle$  and  $\mathbb{E}[G_i] = 0$ . Here  $\mathbf{v} = (\mathbf{u}_1^\top, \dots, \mathbf{u}_K^\top)^\top \in \mathbb{R}^{dK}$ .

$$\begin{aligned}
|G_i| &= \left| \sum_{j=1}^K \sum_{l=1}^K \left[ \xi_{j,l} \mathbf{u}_j^\top \mathbf{x} \mathbf{x}^\top \mathbf{u}_l - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\xi_{j,l} \mathbf{u}_j^\top \mathbf{x} \mathbf{x}^\top \mathbf{u}_l) \right] \right| \\
&\leq C_9 \cdot \left[ \sum_{j=1}^K (\mathbf{u}_j^\top \mathbf{x})^2 + \sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right]
\end{aligned} \tag{C.190}$$

for some  $C_9 > 0$ . The first step of (C.190) is by (C.72). The last step is by (C.152) and the Cauchy-Schwarz's Inequality.

$$\begin{aligned}
\mathbb{E}[|G_i|^p] &\leq \sum_{l=1}^p \binom{p}{l} C_9 \cdot \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \left( \sum_{j=1}^K (\mathbf{u}_j^\top \mathbf{x})^2 \right)^l \right] \\
&\quad \cdot \left( \sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^{p-l} \\
&= \sum_{l=1}^p \binom{p}{l} C_9 \cdot \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sum_{l_1+\dots+l_K=l} \frac{l!}{\prod_{j=1}^K l_j!} \prod_{j=1}^K (\mathbf{u}_j^\top \mathbf{x})^{2l_j} \right] \\
&\quad \cdot \left( \sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^{p-l} \\
&= \sum_{l=1}^p \binom{p}{l} C_9 \cdot \left[ \sum_{l_1+\dots+l_K=l} \frac{l!}{\prod_{j=1}^K l_j!} \prod_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^{2l_j} \right] \\
&\quad \cdot \left( \sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^{p-l} \\
&= C_9 \cdot \sum_{l=1}^p \binom{p}{l} \left( \sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^l \\
&\quad \cdot \left( \sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^{p-l} \\
&= C_9 \cdot \left( \sum_{j=1}^K \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\mathbf{u}_j^\top \mathbf{x})^2 \right)^p \\
&\leq C_9 \cdot \left( \sum_{j=1}^K 1!! \|\mathbf{u}_j\|^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^p \\
&\leq C_9 \cdot \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^p
\end{aligned} \tag{C.191}$$

where the first step is by the triangle inequality and the Binomial theorem, and the second step comes from the Multinomial theorem. The second to last inequality in (C.191) results from Property 8. The last inequality is because  $\mathbf{v} \in \mathbf{V}_{\frac{1}{4}}$ ,  $\sum_{j=1}^K \|u_j\|^2 = \|\mathbf{v}\|^2 \leq 1$ .

$$\begin{aligned} \mathbb{E}[\exp(\theta G_i)] &= 1 + \theta \mathbb{E}[G_i] + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}[|G_i|^p]}{p!} \\ &\leq 1 + \sum_{p=2}^{\infty} \frac{|e\theta|^p}{p^p} C_9 \cdot \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^p \\ &\leq 1 + C_9 \cdot |e\theta|^2 \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^2 \end{aligned} \quad (\text{C.192})$$

where the first inequality holds from  $p! \geq (\frac{p}{e})^p$  and (C.191), and the third line holds provided that

$$\max_{p \geq 2} \left\{ \frac{\frac{|e\theta|^{(p+1)}}{(p+1)^{(p+1)}} \cdot \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^{p+1}}{\frac{|e\theta|^p}{p^p} \cdot \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^p} \right\} \leq \frac{1}{2} \quad (\text{C.193})$$

Note that the quantity inside the maximization in (C.193) achieves its maximum when  $p = 2$ , because it is monotonously decreasing. Therefore, (C.193) holds if  $\theta \leq \frac{27}{4e} \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2$ . Then

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n G_i \geq \frac{t}{6}\right) &= \mathbb{P}\left(\exp(\theta \sum_{i=1}^n G_i) \geq \exp\left(\frac{n\theta t}{6}\right)\right) \leq e^{-\frac{n\theta t}{6}} \prod_{i=1}^n \mathbb{E}[\exp(\theta G_i)] \\ &\leq \exp(C_{10}\theta^2 n \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^2 - \frac{n\theta t}{6}) \end{aligned} \quad (\text{C.194})$$

for some constant  $C_{10} > 0$ . The first inequality follows from Markov's Inequality. When  $\theta = \min\left\{\frac{t}{12C_{10}\left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\right)^2}, \frac{27}{4e} \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\right\}$ , we have a modified Bernstein's Inequality for the Gaussian Mixture Model as follows

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n G_i \geq \frac{t}{6}\right) &\leq \exp\left(\max\left\{-\frac{C_{10}nt^2}{144\left(\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\right)^2},\right.\right. \\ &\quad \left.\left.- C_{11}n \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot t\right\}\right) \end{aligned} \quad (\text{C.195})$$

for some constant  $C_{11} > 0$ . We can obtain the same bound for  $\mathbb{P}(-\frac{1}{n} \sum_{i=1}^n G_i \geq \frac{t}{6})$  by replacing  $G_i$  as  $-G_i$ . Therefore, we have

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n G_i\right| \geq \frac{t}{6}\right) &\leq 2 \exp\left(-\max\left\{-\frac{C_{10}nt^2}{144\left(\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2\right)^2},\right.\right. \\ &\quad \left.\left.- C_{11}n \sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot t\right\}\right) \end{aligned} \quad (\text{C.196})$$

Thus, as long as

$$t \geq C_6 \cdot \max\left\{\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log \frac{36r}{\epsilon} + \log \frac{4}{\delta}}{n}}, \frac{dK \log \frac{36r}{\epsilon} + \log \frac{4}{\delta}}{\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 n}\right\} \quad (\text{C.197})$$

for some large constant  $C_6 > 0$ , we have  $\mathbb{P}(B_t) \leq \frac{\delta}{2}$ .

2) **Upper bound on  $\mathbb{P}(A_t)$  and  $\mathbb{P}(C_t)$ .** From Lemma C.4.2, we can obtain

$$\begin{aligned} &\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left| \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x})] \right. \\ &\quad \left. - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x})] \right| \\ &\leq \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left| \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x})] - \right. \\ &\quad \left. \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla^2 \ell(\mathbf{W}; \mathbf{x})] \right| \cdot \left( \|\mathbf{W} - \mathbf{W}_{j(\mathbf{W})}\|_F \right)^{-1} \\ &\quad \cdot \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\mathbf{W} - \mathbf{W}_{j(\mathbf{W})}\|_F \\ &\leq C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \cdot \epsilon} \end{aligned} \quad (\text{C.198})$$

Therefore,  $C_t$  holds if

$$t \geq C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4 \cdot \epsilon} \quad (\text{C.199})$$

We can bound the  $A_t$  as below.

$$\begin{aligned}
& \mathbb{P} \left( \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{1}{n} \left| \left| \sum_{i=1}^n [\nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right| \right| \geq \frac{t}{3} \right) \\
& \leq \frac{3}{t} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{1}{n} \left| \left| \sum_{i=1}^n [\nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i)] \right| \right| \right] \\
& = \frac{3}{t} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left| \left| \nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) \right| \right| \right] \\
& \leq \frac{3}{t} \mathbb{E} \left[ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{\left| \left| \nabla^2 \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i) - \nabla^2 \ell(\mathbf{W}; \mathbf{x}_i) \right| \right|}{\|\mathbf{W} - \mathbf{W}_{j(\mathbf{W})}\|_F} \right] \\
& \cdot \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\mathbf{W} - \mathbf{W}_{j(\mathbf{W})}\|_F \\
& \leq \frac{C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4} \cdot \epsilon}{t},
\end{aligned} \tag{C.200}$$

where the first inequality is by Markov's inequality, and the last inequality comes from Lemma C.4.2. Thus, taking

$$t \geq \frac{C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4} \cdot \epsilon}{\delta} \tag{C.201}$$

ensures that  $\mathbb{P}(A_t) \leq \frac{\delta}{2}$ .

### 3) Final step

Let  $\epsilon = \frac{\delta}{C_{12} \cdot d^{\frac{3}{2}} K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4} \cdot ndK}$  and  $\delta = d^{-10}$ , then from (C.197)

and (C.201) we need

$$\begin{aligned}
t > \max\left\{\frac{1}{ndK}, C_6 \cdot \sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right. \\
&\quad \cdot (dK \log(36rnd^{\frac{25}{2}} K^{\frac{7}{2}} \sqrt{\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4}) \\
&\quad + \log \frac{4}{\delta})^{\frac{1}{2}} n^{-\frac{1}{2}}, \\
&\quad \left. (dK \log(36rnd^{\frac{25}{2}} K^{\frac{7}{2}} \cdot \sqrt{\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^4}) \right. \\
&\quad \left. + \log \frac{4}{\delta})^{\frac{1}{2}} (\sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 n)^{-\frac{1}{2}} \right\} \\
\end{aligned} \tag{C.202}$$

So by setting  $t = \sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log n}{n}}$ , as long as  $n \geq C' \cdot dK \log dK$ , we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla^2 f_n(\mathbf{W}) - \nabla^2 \bar{f}(\mathbf{W})\| \right. \\
&\leq C_6 \cdot \sum_{l=1}^L \lambda_l(\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \sqrt{\frac{dK \log n}{n}} \left. \right) \leq d^{-10}
\end{aligned} \tag{C.203}$$

## C.5 Proof of Lemma C.3.2

We first present a lemma used in proving Lemma C.3.2 in Section C.5.1 and then prove Lemma C.3.2 in Section C.5.2.

### C.5.1 A Useful Lemma Used in the Proof

**Lemma C.5.1.** *If  $r$  is defined in (C.81) for  $\epsilon_0 \in (0, \frac{1}{4})$ , then with probability at least  $1 - d^{-10}$ , we have<sup>35</sup>*

---

<sup>35</sup> $\nabla \tilde{f}_n(\mathbf{W})$  is defined as  $\frac{1}{n} \sum_{i=1}^n (\nabla l(\mathbf{W}, \mathbf{x}_i, y_i) + \nu_i)$  in algorithm 1

$$\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla \tilde{f}_n(\mathbf{W}) - \nabla \tilde{f}(\mathbf{W})\| \leq C_{13} \cdot \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|)^2} \sqrt{\frac{d \log n}{n}} (1 + \xi) \quad (\text{C.204})$$

for some constant  $C_{13} > 0$ , where  $\mathbf{P}$  is a permutation matrix.

**Proof:**

Note that  $\nabla \tilde{f}_n(\mathbf{W}) = \nabla f_n(\mathbf{W}) + \frac{1}{n} \sum_{i=1}^n \nu_i$ ,  $\nabla \tilde{f}(\mathbf{W}) = \nabla \bar{f}(\mathbf{W}) + \mathbb{E}[\nu_i] = \nabla \bar{f}(\mathbf{W})$ . Therefore, we have

$$\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla \tilde{f}_n(\mathbf{W}) - \nabla \tilde{f}(\mathbf{W})\| \leq \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla f_n(\mathbf{W}) - \nabla \bar{f}(\mathbf{W})\| + \left\| \frac{1}{n} \sum_{i=1}^n \nu_i \right\| \quad (\text{C.205})$$

Then, similar to the idea of the proof of Lemma C.4.5, we adopt an  $\epsilon$ -covering net of the ball  $\mathbb{B}(\mathbf{W}^*, r)$  to build a relationship between any arbitrary point in the ball and the points in the covering set. We can then divide the distance between  $\nabla f_n(\mathbf{W})$  and  $\nabla \bar{f}(\mathbf{W})$  into three parts, similar to (C.183). (C.206) to (C.208) can be derived in a similar way as (C.185) to (C.187), with “ $\nabla^2$ ” replaced by “ $\nabla$ ”. Then we need to bound  $\mathbb{P}(A'_t)$ ,  $\mathbb{P}(B'_t)$  and  $\mathbb{P}(C'_t)$  respectively, where  $A'_t$ ,  $B'_t$  and  $C'_t$  are defined below.

$$A'_t = \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{1}{n} \left\| \sum_{i=1}^n [\nabla \ell(\mathbf{W}; \mathbf{x}_i) - \nabla \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \quad (\text{C.206})$$

$$\begin{aligned} B'_t = & \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left\| \frac{1}{n} \sum_{i=1}^n \nabla \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i) \right. \right. \\ & \left. \left. - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \end{aligned} \quad (\text{C.207})$$

$$\begin{aligned} C'_t = & \left\{ \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left\| \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)] \right. \right. \\ & \left. \left. - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}; \mathbf{x}_i)] \right\| \geq \frac{t}{3} \right\} \end{aligned} \quad (\text{C.208})$$

(a) Upper bound of  $\mathbb{P}(B'_t)$ . Applying Lemma 3 in [268], we have

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^n \nabla \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)] \right\| \\ & \leq 2 \sup_{\mathbf{v} \in V_{\frac{1}{2}}} \left| \left\langle \frac{1}{n} \sum_{i=1}^n \nabla \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}_{j(\mathbf{W})}; \mathbf{x}_i)], \mathbf{v} \right\rangle \right| \end{aligned} \quad (\text{C.209})$$

Define  $G'_i = \left\langle \mathbf{v}, (\nabla \ell(\mathbf{W}, \mathbf{x}_i) - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\nabla \ell(\mathbf{W}, \mathbf{x}_i)]) \right\rangle$ . Here  $\mathbf{v} \in \mathbb{R}^d$ . To compute  $\nabla \ell(\mathbf{W}, \mathbf{x}_i)$ , we require the derivation in Property 9. Then we can have an upper bound of  $\zeta(\mathbf{W})$  in (C.71).

$$\zeta(\mathbf{W}) = \begin{cases} \left| -\frac{1}{K} \frac{1}{H(\mathbf{W})} \phi'(\mathbf{w}_j^\top \mathbf{x}) \right| \leq \frac{\phi(\mathbf{w}_j^\top \mathbf{x})(1-\phi(\mathbf{w}_j^\top \mathbf{x}))}{K \cdot \frac{1}{K} \phi(\mathbf{w}_j^\top \mathbf{x})} \leq 1, & y = 1 \\ \left| \frac{1}{K} \frac{1}{1-H(\mathbf{W})} \phi'(\mathbf{w}_j^\top \mathbf{x}) \right| \leq \frac{\phi(\mathbf{w}_j^\top \mathbf{x})(1-\phi(\mathbf{w}_j^\top \mathbf{x}))}{K \cdot \frac{1}{K} (1-\phi(\mathbf{w}_j^\top \mathbf{x}))} \leq 1, & y = 0 \end{cases} \quad (\text{C.210})$$

Then we have an upper bound of  $G'_i$ .

$$\begin{aligned} |G'_i| &= \left| \zeta_{j,l} \mathbf{v}^\top \mathbf{x} - \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\zeta \mathbf{v}^\top \mathbf{x}] \right| \\ &\leq |\mathbf{v}^\top \mathbf{x}| + \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [| \mathbf{v}^\top \mathbf{x} |] \end{aligned} \quad (\text{C.211})$$

Following the idea of (C.191) and (C.192), and by  $\mathbf{v} \in V_{\frac{1}{2}}$ , we have

$$\mathbb{E}[|G'_i|^p] \leq O \left( \left( \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right)^{\frac{p}{2}} \right) \quad (\text{C.212})$$

$$\mathbb{E}[\exp(\theta G'_i)] \leq 1 + O \left( |e\theta^2| \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \right) \quad (\text{C.213})$$

where (C.213) holds if  $\theta \leq \frac{27}{4e} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|)^2}$ . Following the derivation of (C.189) and (C.194) to (C.197), we have

$$\begin{aligned} & \mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^n G'_i \right| \geq \frac{t}{6} \right) \\ & \leq 2 \exp \left( \max \left\{ -\frac{C_{14} n t^2}{144 \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2}, -C_{15} n \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot t} \right\} \right) \end{aligned} \quad (\text{C.214})$$

for some constant  $C_{14} > 0$  and  $C_{15} > 0$ . Moreover, we can obtain  $\mathbb{P}(B'_t) \leq \frac{\delta}{2}$  as long as

$$t \geq C_{13} \cdot \max \left\{ \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{dK \log \frac{18r}{\epsilon} + \log \frac{4}{\delta}}{n}}, \right. \\ \left. \frac{dK \log \frac{18r}{\epsilon} + \log \frac{4}{\delta}}{\sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \cdot n} \right\} \quad (\text{C.215})$$

(b) For the upper bound of  $\mathbb{P}(A'_t)$  and  $\mathbb{P}(C'_t)$ , we can first derive

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sup_{\mathbf{W} \neq \mathbf{W}' \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{\|\nabla \ell(\mathbf{W}, \mathbf{x}) - \nabla \ell(\mathbf{W}', \mathbf{x})\|}{\|\mathbf{W} - \mathbf{W}'\|_F} \right] \\ & \leq \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sup_{\mathbf{W} \neq \mathbf{W}' \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{|\zeta(\mathbf{W}) - \zeta(\mathbf{W}')| \cdot \|\mathbf{x}\|}{\|\mathbf{W} - \mathbf{W}'\|_F} \right] \\ & \leq \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sup_{\mathbf{W} \neq \mathbf{W}' \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \left( \max_{1 \leq j, l \leq K} \{|\xi_{j,l}(\mathbf{W}'')|\} \cdot \|\mathbf{x}\|^2 \sqrt{K} \right. \right. \\ & \quad \left. \left. \|\mathbf{W} - \mathbf{W}'\|_F \right)^{\frac{1}{2}} \|\mathbf{W} - \mathbf{W}'\|_F^{-\frac{1}{2}} \right] \quad (\text{C.216}) \\ & \leq \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \left[ \sup_{\mathbf{W} \neq \mathbf{W}' \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \frac{C_9 \cdot \|\mathbf{x}\|^2 \sqrt{K} \|\mathbf{W} - \mathbf{W}'\|_F}{\|\mathbf{W} - \mathbf{W}'\|_F} \right] \\ & \leq C_9 \cdot 3\sqrt{K}d \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \end{aligned}$$

The first inequality is by (C.71). The second inequality is by the Mean Value Theorem. The third step is by (C.152). The last inequality is by Property 7. Therefore, following the steps in part (2) of Lemma C.4.5, we can conclude that  $C'_t$  holds if

$$t \geq 3C_9 \cdot \sqrt{K}d \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \cdot \epsilon \quad (\text{C.217})$$

Moreover, from (C.201) in Lemma C.4.5 we have that

$$t \geq \frac{18C_9 \cdot \sqrt{K}d \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^2 \cdot \epsilon}{\delta} \quad (\text{C.218})$$

ensures  $\mathbb{P}(A'_t) \leq \frac{\delta}{2}$ . Therefore, let  $\epsilon = \frac{\delta}{18C_9 \cdot \sqrt{K}d \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\|_\infty + \|\boldsymbol{\Sigma}_l\|)^2 \cdot \epsilon \cdot ndK}$ ,  $\delta = d^{-10}$  and  $t = C_{13} \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|)^2} \sqrt{\frac{d \log n}{n}}$ , if  $n \geq C'' \cdot dK \log dK$  for some constant  $C'' > 0$ ,

we have

$$\begin{aligned} & \mathbb{P}\left(\sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla f_n(\mathbf{W}) - \nabla \bar{f}(\mathbf{W})\|\right) \\ & \leq C_{13} \cdot \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l\|)^2} \sqrt{\frac{d \log n}{n}} \leq d^{-10} \end{aligned} \quad (\text{C.219})$$

By Hoeffding's inequality in [258] and Property 2, we have

$$\begin{aligned} & \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \|\nu_i\|_F \geq C_{13} \cdot \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{d K \log n}{n}} \xi\right) \\ & \lesssim \exp(-C_{13}^2 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2 \frac{\xi^2 d K \log n}{d K \xi^2}) \\ & \lesssim d^{-10} \end{aligned} \quad (\text{C.220})$$

Therefore,

$$\begin{aligned} & \sup_{\mathbf{W} \in \mathbb{B}(\mathbf{W}^* \mathbf{P}, r)} \|\nabla \tilde{f}_n(\mathbf{W}) - \nabla \tilde{f}(\mathbf{W})\| \\ & \leq C_{13} \cdot \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{d \log n}{n}} + \frac{1}{n} \sum_{i=1}^n \|\nu_i\| \\ & \leq C_{13} \cdot \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{d \log n}{n}} + \frac{1}{n} \sum_{i=1}^n \|\nu_i\|_F \\ & \leq C_{13} \cdot \sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\boldsymbol{\Sigma}_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{d \log n}{n}} (1 + \xi) \end{aligned} \quad (\text{C.221})$$

### C.5.2 Proof of Lemma C.3.2

Following the proof of Theorem 2 in [72], first, we have Taylor's expansion of  $f_n(\widehat{\mathbf{W}}_n)$

$$\begin{aligned} f_n(\widehat{\mathbf{W}}_n) &= f_n(\mathbf{W}^* \mathbf{P}) + \left\langle \nabla \tilde{f}_n(\mathbf{W}^* \mathbf{P}), \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \right\rangle \\ &\quad + \frac{1}{2} \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \nabla^2 f_n(\mathbf{W}') \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \end{aligned} \quad (\text{C.222})$$

Here  $\mathbf{W}'$  is on the straight line connecting  $\mathbf{W}^* \mathbf{P}$  and  $\widehat{\mathbf{W}}_n$ . By the fact that  $f_n(\widehat{\mathbf{W}}_n) \leq f_n(\mathbf{W}^* \mathbf{P})$ , we have

$$\frac{1}{2} \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \nabla^2 f_n(\mathbf{W}') \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \leq \left| \nabla f_n(\mathbf{W}^* \mathbf{P})^\top \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^* \mathbf{P}) \right| \quad (\text{C.223})$$

From Lemma C.4.4 and Lemma C.5.1, we have

$$\begin{aligned} & \frac{4}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2} \rho\left(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right) \|\widehat{\mathbf{W}}_n - \mathbf{W}^*\mathbf{P}\|_F^2 \\ & \leq \frac{1}{2} \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^*\mathbf{P}) \nabla^2 f_n(\mathbf{W}') \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^*\mathbf{P}) \end{aligned} \quad (\text{C.224})$$

and

$$\begin{aligned} & \left| \nabla \tilde{f}_n(\mathbf{W}^*\mathbf{P})^\top \text{vec}(\widehat{\mathbf{W}}_n - \mathbf{W}^*\mathbf{P}) \right| \\ & \leq \|\nabla \tilde{f}_n(\mathbf{W}^*\mathbf{P})\| \cdot \|\widehat{\mathbf{W}}_n - \mathbf{W}^*\mathbf{P}\|_F \\ & \leq (\|\nabla \tilde{f}_n(\mathbf{W}^*\mathbf{P}) - \nabla \bar{f}(\mathbf{W}^*\mathbf{P})\| + \|\nabla \bar{f}(\mathbf{W}^*\mathbf{P})\|) \cdot \|\widehat{\mathbf{W}}_n - \mathbf{W}^*\mathbf{P}\|_F \\ & \leq O\left(\sqrt{K \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2} \sqrt{\frac{d \log n}{n}} (1 + \xi)\right) \|\widehat{\mathbf{W}}_n - \mathbf{W}^*\mathbf{P}\|_F \end{aligned} \quad (\text{C.225})$$

The second to last step of (C.225) comes from the triangle inequality, and the last step follows from the fact  $\nabla \bar{f}(\mathbf{W}^*\mathbf{P}) = 0$ . Combining (C.223), (C.224) and (C.225), we have

$$\begin{aligned} & \|\widehat{\mathbf{W}}_n - \mathbf{W}^*\mathbf{P}\|_F \\ & \leq O\left(\frac{K^{\frac{5}{2}} \sqrt{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l^{\frac{1}{2}}\|)^2} (1 + \xi)}{\sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta\tau^K\kappa^2} \rho\left(\frac{\mathbf{W}^{*\top}\boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*)\|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right)} \sqrt{\frac{d \log n}{n}}\right) \end{aligned} \quad (\text{C.226})$$

Therefore, we have concluded that there indeed exists a critical point  $\widehat{\mathbf{W}}$  in  $\mathbb{B}(\mathbf{W}^*\mathbf{P}, r)$ . Then we show the linear convergence of Algorithm 1 as below. By the update rule, we have

$$\begin{aligned} \mathbf{W}_{t+1} - \widehat{\mathbf{W}}_n &= \mathbf{W}_t - \eta_0 (\nabla f_n(\mathbf{W}_t) + \frac{1}{n} \sum_{i=1}^n \nu_i) - (\widehat{\mathbf{W}}_n - \eta_0 \nabla f_n(\widehat{\mathbf{W}}_n)) \\ &= \left( \mathbf{I} - \eta_0 \int_0^1 \nabla^2 f_n(\mathbf{W}(\gamma)) d\gamma \right) (\mathbf{W}_t - \widehat{\mathbf{W}}_n) - \frac{\eta_0}{n} \sum_{i=1}^n \nu_i \end{aligned} \quad (\text{C.227})$$

where  $\mathbf{W}(\gamma) = \gamma \widehat{\mathbf{W}}_n + (1 - \gamma) \mathbf{W}_t$  for  $\gamma \in (0, 1)$ . Since  $\mathbf{W}(\gamma) \in \mathbb{B}(\mathbf{W}^*\mathbf{P}, r)$ , by Lemma

C.3.1, we have

$$H_{\min} \cdot \mathbf{I} \preceq \nabla^2 f_n(\mathbf{W}(\gamma)) \leq H_{\max} \cdot \mathbf{I} \quad (\text{C.228})$$

where  $H_{\min} = \Omega\left(\frac{1}{K^2} \sum_{l=1}^L \lambda_l \frac{\|\Sigma_l^{-1}\|^{-1}}{\eta \tau^K \kappa^2} \rho\left(\frac{\mathbf{W}^{*\top} \boldsymbol{\mu}_l}{\delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}}, \delta_K(\mathbf{W}^*) \|\Sigma_l^{-1}\|^{-\frac{1}{2}}\right)\right)$ ,  $H_{\max} = C_4 \cdot \sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l\|)^2$ . Therefore,

$$\begin{aligned} \|\mathbf{W}_{t+1} - \widehat{\mathbf{W}}_n\|_F &= \|\mathbf{I} - \eta_0 \int_0^1 \nabla^2 f_n(\mathbf{W}(\gamma)) \mathbf{d}\gamma\| \cdot \|\mathbf{W}_t - \widehat{\mathbf{W}}_n\|_F + \left\| \frac{\eta_0}{n} \sum_{i=1}^n \nu_i \right\|_F \\ &\leq (1 - \eta_0 H_{\min}) \|\mathbf{W}_t - \widehat{\mathbf{W}}_n\|_F + \left\| \frac{\eta_0}{n} \sum_{i=1}^n \nu_i \right\|_F \end{aligned} \quad (\text{C.229})$$

By setting  $\eta_0 = \frac{1}{H_{\max}} = O\left(\frac{1}{\sum_{l=1}^L \lambda_l (\|\boldsymbol{\mu}_l\| + \|\Sigma_l\|)^2}\right)$ , we obtain

$$\|\widehat{\mathbf{W}}_{t+1} - \widehat{\mathbf{W}}_n\|_F \leq (1 - \frac{H_{\min}}{H_{\max}}) \|\mathbf{W}_t - \widehat{\mathbf{W}}_n\|_F + \frac{\eta_0}{n} \sum_{i=1}^n \|\nu_i\|_F \quad (\text{C.230})$$

Therefore, Algorithm 1 converges linearly to the local minimizer with an extra statistical error.

By Hoeffding's inequality in [258] and Property 2, we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n \|\nu_i\|_F \geq \sqrt{\frac{dK \log n}{n}} \xi\right) \lesssim \exp\left(-\frac{\xi^2 dK \log n}{dK \xi^2}\right) \lesssim d^{-10} \quad (\text{C.231})$$

Therefore, with probability  $1 - d^{-10}$  we can derive

$$\|\widehat{\mathbf{W}}_t - \widehat{\mathbf{W}}_n\|_F \leq (1 - \frac{H_{\min}}{H_{\max}})^t \|\mathbf{W}_0 - \widehat{\mathbf{W}}_n\|_F + \frac{H_{\max} \eta_0}{H_{\min}} \sqrt{\frac{dK \log n}{n}} \xi \quad (\text{C.232})$$

## C.6 Proof of Lemma C.3.3

We need Lemma C.6.1 to Lemma C.6.5, which are stated in Section C.6.1, for the proof of Lemma C.3.3. Section C.6.2 summarizes the proof of Lemma C.3.3. The proofs of Lemma C.6.1 to Lemma C.6.3 are provided in Section C.6.3 to Section C.6.5. Lemma C.6.4 and Lemma C.6.5 are cited from [71]. Although [71] considers the standard Gaussian distribution, the proofs of Lemma C.6.4 and C.6.5 hold for any data distribution. Therefore, these two lemmas can be applied here directly.

The tensor initialization in [71] only holds for the standard Gaussian distribution. We exploit a more general definition of tensors from [266] for the tensor initialization in our algorithm. We also develop new error bounds for the initialization.

### C.6.1 Useful Lemmas in the Proof

**Lemma C.6.1.** *Let  $\mathbf{Q}_2$  and  $\mathbf{Q}_3$  follow Definition C.1.1. Let  $S$  be a set of i.i.d. samples generated from the mixed Gaussian distribution  $\sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ . Let  $\widehat{\mathbf{Q}}_2, \widehat{\mathbf{Q}}_3$  be the empirical version of  $\mathbf{Q}_2, \mathbf{Q}_3$  using data set  $S$ , respectively. Then with a probability at least  $1 - 2n^{-\Omega(\delta_1(\mathbf{W}^*)^4 d)}$ , we have*

$$\|\mathbf{Q}_2 - \widehat{\mathbf{Q}}_2\| \lesssim \sqrt{\frac{d \log n}{n}} \cdot \delta_1(\mathbf{W}^*)^2 \cdot \tau^6 \sqrt{D_2(\Psi) D_4(\Psi)} \quad (\text{C.233})$$

if the mixed Gaussian distribution is not symmetric. We also have

$$\|\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \widehat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\| \lesssim \sqrt{\frac{d \log n}{n}} \cdot \delta_1(\mathbf{W}^*)^2 \cdot \tau^6 \sqrt{D_2(\Psi) D_4(\Psi)} \quad (\text{C.234})$$

for any arbitrary vector  $\boldsymbol{\alpha} \in \mathbb{R}^d$ , if the mixed Gaussian distribution is symmetric.

**Lemma C.6.2.** *Let  $\mathbf{U} \in \mathbb{E}^{d \times K}$  be the orthogonal column span of  $\mathbf{W}^*$ . Let  $\boldsymbol{\alpha}$  be a fixed unit vector and  $\widehat{\mathbf{U}} \in \mathbb{R}^{d \times K}$  denote an orthogonal matrix satisfying  $\|\mathbf{U}\mathbf{U}^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\| \leq \frac{1}{4}$ . Define  $\mathbf{R}_3 = \mathbf{Q}_3(\widehat{\mathbf{U}}, \widehat{\mathbf{U}}, \widehat{\mathbf{U}})$ , where  $\mathbf{Q}_3$  is defined in Definition C.1.1. Let  $\widehat{\mathbf{R}}_3$  be the empirical version of  $\mathbf{R}_3$  using data set  $S$ , where each sample of  $S$  is i.i.d. sampled from the mixed Gaussian distribution  $\sum_{l=1}^L \lambda_l \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ . Then with a probability at least  $1 - n^{-\Omega(\delta^4(\mathbf{W}^*))}$ , we have*

$$\|\widehat{\mathbf{R}}_3 - \mathbf{R}_3\| \lesssim \delta_1(\mathbf{W}^*)^2 \cdot (\tau^6 \sqrt{D_6(\Psi)}) \cdot \sqrt{\frac{\log n}{n}} \quad (\text{C.235})$$

**Lemma C.6.3.** *Let  $\widehat{\mathbf{Q}}_1$  be the empirical version of  $\mathbf{Q}_1$  using dataset  $S$ . Then with a probability at least  $1 - 2n^{-\Omega(d)}$ , we have*

$$\|\widehat{\mathbf{Q}}_1 - \mathbf{Q}_1\| \lesssim (\tau^2 \sqrt{D_2(\Psi)}) \cdot \sqrt{\frac{d \log n}{n}} \quad (\text{C.236})$$

**Lemma C.6.4.** ([71], Lemma E.6) *Let  $\mathbf{Q}_2, \mathbf{Q}_3$  be defined in Definition C.1.1 and  $\widehat{\mathbf{Q}}_2$ ,*

$\widehat{\mathbf{Q}}_3$  be their empirical version, respectively. Let  $\mathbf{U} \in \mathbb{R}^{d \times K}$  be the column span of  $\mathbf{W}^*$ . Assume  $\|\mathbf{Q}_2 - \widehat{\mathbf{Q}}_2\| \leq \frac{\delta_K(\mathbf{Q}_2)}{10}$  for non-symmetric distribution cases and  $\|\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \widehat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\| \leq \frac{\delta_K(\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}))}{10}$  for symmetric distribution cases and any arbitrary vector  $\boldsymbol{\alpha} \in \mathbb{R}^d$ . Then after  $T = O(\log(\frac{1}{\epsilon}))$  iterations, the output of the Tensor Initialization Method 1,  $\widehat{\mathbf{U}}$  will satisfy

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\| \lesssim \frac{\|\widehat{\mathbf{Q}}_2 - \mathbf{Q}_2\|}{\delta_K(\mathbf{Q}_2)} + \epsilon, \quad (\text{C.237})$$

which implies

$$\|(\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top)\mathbf{w}_i^*\| \lesssim \left( \frac{\|\mathbf{Q}_2 - \widehat{\mathbf{Q}}_2\|}{\delta_K(\mathbf{Q}_2)} + \epsilon \right) \|\mathbf{w}_i^*\| \quad (\text{C.238})$$

if the mixed Gaussian distribution is not symmetric. Similarly, we have

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\| \lesssim \frac{\|\widehat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\|}{\delta_K(\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}))} + \epsilon, \quad (\text{C.239})$$

which implies

$$\|(\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top)\mathbf{w}_i^*\| \lesssim \left( \frac{\|\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \widehat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\|}{\delta_K(\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}))} + \epsilon \right) \|\mathbf{w}_i^*\| \quad (\text{C.240})$$

if the mixed Gaussian distribution is symmetric.

**Lemma C.6.5.** ([71], Lemma E.13) Let  $\mathbf{U} \in \mathbb{R}^{d \times K}$  be the orthogonal column span of  $\mathbf{W}^*$ . Let  $\widehat{\mathbf{U}} \in \mathbb{R}^{d \times K}$  be an orthogonal matrix such that  $\|\mathbf{U}\mathbf{U}^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\| \lesssim \gamma_1 \lesssim \frac{1}{\kappa^2 \sqrt{K}}$ . For each  $i \in [K]$ , let  $\widehat{\mathbf{v}}_i$  denote the vector satisfying  $\|\widehat{\mathbf{v}}_i - \widehat{\mathbf{U}}^\top \bar{\mathbf{w}}_i^*\| \leq \gamma_2 \lesssim \frac{1}{\kappa^2 \sqrt{K}}$ . Let  $\mathbf{Q}_1$  be defined in Lemma C.6.3 and  $\widehat{\mathbf{Q}}_1$  be its empirical version. If  $\|\mathbf{Q}_1 - \widehat{\mathbf{Q}}_1\| \leq \gamma_3 \|\mathbf{Q}_1\| \lesssim \frac{1}{4} \|\mathbf{Q}_1\|$ , then we have

$$\left| \|\mathbf{w}_i^*\| - \widehat{\alpha}_i \right| \leq (\kappa^4 K^{\frac{3}{2}} (\gamma_1 + \gamma_2) + \kappa^2 K^{\frac{1}{2}} \gamma_3) \|\mathbf{w}_i^*\| \quad (\text{C.241})$$

### C.6.2 Proof of Lemma C.3.3

By the triangle inequality, we have

$$\begin{aligned}
& \|\mathbf{w}_j^* - \hat{\alpha}_j \hat{\mathbf{U}} \hat{\mathbf{v}}_j\| \\
&= \left\| \mathbf{w}_j^* - \|\mathbf{w}_j^*\| \hat{\mathbf{U}} \hat{\mathbf{v}}_j + \|\mathbf{w}_j^*\| \hat{\mathbf{U}} \hat{\mathbf{v}}_j - \hat{\alpha}_j \hat{\mathbf{U}} \hat{\mathbf{v}}_j \right\| \\
&\leq \left\| \mathbf{w}_j^* - \|\mathbf{w}_j^*\| \hat{\mathbf{U}} \hat{\mathbf{v}}_j \right\| + \left\| \|\mathbf{w}_j^*\| \hat{\mathbf{U}} \hat{\mathbf{v}}_j - \hat{\alpha}_j \hat{\mathbf{U}} \hat{\mathbf{v}}_j \right\| \\
&\leq \|\mathbf{w}_j^*\| \left\| \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{v}}_j \right\| + \left\| \|\mathbf{w}_j^*\| - \hat{\alpha}_j \right\| \|\hat{\mathbf{U}} \hat{\mathbf{v}}_j\| \\
&\leq \|\mathbf{w}_j^*\| \left\| \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* + \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{v}}_j \right\| + \left\| \|\mathbf{w}_j^*\| - \hat{\alpha}_j \right\| \|\hat{\mathbf{U}} \hat{\mathbf{v}}_j\| \\
&\leq \delta_1(\mathbf{W}^*) \left( \left\| \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* \right\| + \left\| \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* - \hat{\mathbf{v}}_j \right\| \right) + \left\| \|\mathbf{w}_j^*\| - \hat{\alpha}_j \right\|
\end{aligned} \tag{C.242}$$

From Lemma C.6.1, Lemma C.6.4,  $\delta_K(\mathbf{Q}_2) \lesssim \delta_K^2(\mathbf{W}^*)$  and  $\delta_K(\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})) \lesssim \delta_K^2(\mathbf{W}^*)$  for any arbitrary vector  $\boldsymbol{\alpha} \in \mathbb{R}^d$ , we have

$$\begin{aligned}
\left\| \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* \right\| &\lesssim \frac{\|\mathbf{Q}_2 - \hat{\mathbf{Q}}_2\|}{\delta_K(\mathbf{Q}_2)} \lesssim \sqrt{\frac{d \log n}{n}} \cdot \frac{\delta_1(\mathbf{W}^*)^2}{\delta_K(\mathbf{W}^*)^2} \cdot \tau^6 \sqrt{D_2(\Psi) D_4(\Psi)} \\
&= \sqrt{\frac{d \log n}{n}} \cdot \kappa^2 \cdot \tau^6 \sqrt{D_2(\Psi) D_4(\Psi)}
\end{aligned} \tag{C.243}$$

if the mixed Gaussian distribution is not symmetric, and

$$\begin{aligned}
& \left\| \bar{\mathbf{w}}_j^* - \hat{\mathbf{U}} \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* \right\| \\
&\lesssim \frac{\|\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \hat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\|}{\delta_K(\mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}))} = \sqrt{\frac{d \log n}{n}} \cdot \kappa^2 \cdot \tau^6 \sqrt{D_2(\Psi) D_4(\Psi)}
\end{aligned} \tag{C.244}$$

if the mixed Gaussian distribution is symmetric. Moreover, we have

$$\left\| \hat{\mathbf{U}}^\top \bar{\mathbf{w}}_j^* - \hat{\mathbf{v}}_j \right\| \leq \frac{K^{\frac{3}{2}}}{\delta_K^2(\mathbf{W}^*)} \|\mathbf{R}_3 - \hat{\mathbf{R}}_3\| \lesssim \kappa^2 \cdot (\tau^6 \sqrt{D_6(\Psi)}) \cdot \sqrt{\frac{K^3 \log n}{n}} \tag{C.245}$$

in which the first step is by Theorem 3 in [267], and the second step is by Lemma C.6.2. By Lemma C.6.5, we have

$$\left\| \|\mathbf{w}_j^*\| - \hat{\alpha}_j \right\| \leq (\kappa^4 K^{\frac{3}{2}} (\gamma_1 + \gamma_2) + \kappa^2 K^{\frac{1}{2}} \gamma_3) \|\mathbf{W}^*\| \tag{C.246}$$

Therefore, taking the union bound of failure probabilities in Lemmas C.6.1, C.6.2, and C.6.3 and by  $D_2(\Psi)D_4(\Psi) \leq D_6(\Psi)$  from Property 10, we have that if the sample size  $n \geq \kappa^8 K^4 \tau^{12} D_6(\Psi) \cdot d \log^2 d$ , then the output  $\mathbf{W}_0 \in \mathbb{R}^{d \times K}$  satisfies

$$\|\mathbf{W}_0 - \mathbf{W}^*\| \lesssim \kappa^6 K^3 \cdot \tau^6 \sqrt{D_6(\Psi)} \sqrt{\frac{d \log n}{n}} \|\mathbf{W}^*\| \quad (\text{C.247})$$

with probability at least  $1 - n^{-\Omega(\delta_1^4(\mathbf{W}^*))}$

### C.6.3 Proof of Lemma C.6.1

From Assumption C.1.2, if the Gaussian Mixture Model is a symmetric probability distribution defined in (C.4), then by Definition C.1.1, we have

$$\begin{aligned} & \|\hat{\mathbf{Q}}_3(\mathbf{I}, \mathbf{I}, \boldsymbol{\alpha}) - \mathbf{Q}_3(\mathbf{I}, \mathbf{I}, \boldsymbol{\alpha})\| \\ &= \left\| \frac{1}{n} \sum_{i=1}^n \left[ y_i \cdot p(\mathbf{x})^{-1} \sum_{l=1}^L \lambda_l (2\pi|\Sigma_l|)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l)\right) \right. \right. \\ & \quad \cdot \left. \left. \left( ((\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1})^{\otimes 3} - ((\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1}) \tilde{\otimes} \Sigma_l^{-1} \right) \right] (\mathbf{I}, \mathbf{I}, \boldsymbol{\alpha}) \right. \\ & \quad \left. - \mathbb{E} \left[ y \cdot p(\mathbf{x})^{-1} \sum_{l=1}^L \lambda_l (2\pi|\Sigma_l|)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l)\right) \right. \right. \\ & \quad \cdot \left. \left. \left( ((\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1})^{\otimes 3} - ((\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1}) \tilde{\otimes} \Sigma_l^{-1} \right) \right] (\mathbf{I}, \mathbf{I}, \boldsymbol{\alpha}) \right\| \end{aligned} \quad (\text{C.248})$$

Following [71],  $\tilde{\otimes}$  is defined such that for any  $\mathbf{v} \in \mathbb{R}^{d_1}$  and  $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2}$ ,

$$\mathbf{v} \tilde{\otimes} \mathbf{Z} = \sum_{i=1}^{d_2} (\mathbf{v} \otimes \mathbf{z}_i \otimes \mathbf{z}_i + \mathbf{z}_i \otimes \mathbf{v} \otimes \mathbf{z}_i + \mathbf{z}_i \otimes \mathbf{z}_i \otimes \mathbf{v}), \quad (\text{C.249})$$

where  $\mathbf{z}_i$  is the  $i$ -th column of  $\mathbf{Z}$ . By Definition C.1.1, we have

$$\begin{aligned}
& \left\| \left[ y \cdot p(\mathbf{x})^{-1} \sum_{l=1}^L \lambda_l (2\pi|\Sigma_l|)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l)\right) \right. \right. \\
& \quad \cdot \left. \left. \left( ((\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1})^{\otimes 3} - ((\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1}) \tilde{\otimes} \Sigma_l^{-1} \right) \right] (\mathbf{I}, \mathbf{I}, \boldsymbol{\alpha}) \right\| \\
& \lesssim \left\| \sum_{l=1}^L \lambda_l (2\pi|\Sigma_l|)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l)\right) \cdot ((\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1})^{\otimes 2} \right. \\
& \quad \left. (\boldsymbol{\alpha}^\top \Sigma_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l)) \left( \sum_{l=1}^L \lambda_l (2\pi|\Sigma_l|)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l)\right) \right)^{-1} \right\| \\
& \lesssim \|\sigma_{\min}^{-6}(\mathbf{x}^\top \boldsymbol{\alpha}) \mathbf{x} \mathbf{x}^\top\|
\end{aligned} \tag{C.250}$$

The first step of (C.250) is because  $(\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l)^{\otimes 2}(\boldsymbol{\alpha}^\top \Sigma_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l))$  is the dominant term of the entire expression, and  $y \leq 1$ . The second step is because the expression can be considered as a normalized weighted summation of  $((\mathbf{x} - \boldsymbol{\mu}_l)\Sigma_l)^{\otimes 2}(\boldsymbol{\alpha}^\top \Sigma_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l))$  and  $(\mathbf{x}^\top \boldsymbol{\alpha}) \mathbf{x} \mathbf{x}^\top$  is its dominant term. Define  $S_m(\mathbf{x}) = (-1)^m \frac{\nabla_{\mathbf{x}}^m p(\mathbf{x})}{p(\mathbf{x})}$ , where  $p(\mathbf{x})$  is the probability density function of the random variable  $\mathbf{x}$ . From Definition C.1.1, we can verify that

$$\mathbf{Q}_j = \mathbb{E}[y \cdot S_m(\mathbf{x})] \quad j \in \{1, 2, 3\} \tag{C.251}$$

Then define  $Gp_i = \langle \mathbf{v}, ([y_i \cdot S_3(\mathbf{x}_i)](\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \mathbb{E}[[y_i \cdot S_3(\mathbf{x}_i)](\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})]\mathbf{v}) \rangle$ , where  $\|\mathbf{v}\| = 1$ , then  $\mathbb{E}[Gp_i] = 0$ . Similar to the proof of (C.190), (C.191), and (C.192) in Lemma C.4.5, we have

$$|Gp_i|^p \lesssim |\sigma_{\min}^{-6}(\mathbf{x}_i^\top \boldsymbol{\alpha})(\mathbf{x}_i^\top \mathbf{v})^2 + \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \Sigma_l)} [\sigma_{\min}^{-6}(\mathbf{x}_i^\top \boldsymbol{\alpha})(\mathbf{x}_i^\top \mathbf{v})^2]|^p \tag{C.252}$$

$$\begin{aligned}
\mathbb{E}[|Gp_i|^p] & \lesssim (\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \Sigma_l)} [\sigma_{\min}^{-6}(\mathbf{x}_i^\top \boldsymbol{\alpha})(\mathbf{x}_i^\top \mathbf{v})^2])^p \\
& \leq \sigma_{\min}^{-6p} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \Sigma_l)} [(\mathbf{x}^\top \boldsymbol{\alpha})^2]^{\frac{p}{2}} \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \Sigma_l)} [(\mathbf{x}^\top \mathbf{v})^4]^{\frac{p}{2}} \\
& \leq \tau^{6p} \sqrt{D_2(\Psi) D_4(\Psi)}^p
\end{aligned} \tag{C.253}$$

$$\begin{aligned}
\mathbb{E}[\exp(\theta Gp_i)] & \lesssim 1 + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}[|Gp_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty} \frac{|e\theta|^p \tau^{6p} (D_2(\Psi) D_4(\Psi))^{\frac{p}{2}}}{p^p} \\
& \lesssim 1 + \theta^2 \tau^{12} D_2(\Psi) D_4(\Psi)
\end{aligned} \tag{C.254}$$

Hence, similar to the derivation of (C.194), we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Gp_i \geq t\right) \leq \exp\left(-n\theta t + C_{16}n\theta^2(\tau^6 \sqrt{D_2(\Psi)D_4(\Psi)})^2\right) \quad (\text{C.255})$$

for some constant  $C_{16} > 0$ . Let  $\theta = \frac{t}{2C_{16}(\tau^6 \sqrt{D_2(\Psi)D_4(\Psi)})^2}$  and  $t = \delta_1^2(\mathbf{W}^*) \cdot (\tau^6 \sqrt{D_2(\Psi)D_4(\Psi)}) \cdot \sqrt{\frac{d \log n}{n}}$ , then we have

$$\|\widehat{\mathbf{Q}}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha}) - \mathbf{Q}_3(\mathbf{I}_d, \mathbf{I}_d, \boldsymbol{\alpha})\| \leq \delta_1(\mathbf{W}^*)^2 \cdot (\tau^6 \sqrt{D_2(\Psi)D_4(\Psi)}) \cdot \sqrt{\frac{d \log n}{n}} \quad (\text{C.256})$$

with probability at least  $1 - 2n^{-\Omega(\delta_1^4(\mathbf{W}^*)d)}$ .

If the Gaussian Mixture Model is not a symmetric distribution which is defined in (C.4), we would have a similar result as follows.

$$\|\widehat{\mathbf{Q}}_2 - \mathbf{Q}_2\| = \left\| \frac{1}{n} \sum_{i=1}^n [y_i \cdot S_2(\mathbf{x})] - \mathbb{E}[y \cdot S_2(\mathbf{x})] \right\| \quad (\text{C.257})$$

$$\|y_i \cdot S_2(\mathbf{x}_i)\| \lesssim \|\sigma_{\min}^{-4} \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*\top} \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\top\| \quad (\text{C.258})$$

Then define  $Gp'_i = \langle \mathbf{v}, ([y_i \cdot S_2(\mathbf{x}_i)] - \mathbb{E}[y_i \cdot S_2(\mathbf{x}_i)] \mathbf{v}) \rangle$ , where  $\|\mathbf{v}\| = 1$ , then  $\mathbb{E}[Gp'_i] = 0$ . Similar to the proof of (C.190), (C.191) and (C.192) in Lemma C.4.5, we have

$$|Gp'_i|^p \lesssim |\sigma_{\min}^{-4}(\mathbf{x}_i^\top \mathbf{v})^2 + \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-4}(\mathbf{x}_i^\top \mathbf{v})^2]|^p \quad (\text{C.259})$$

$$\mathbb{E}[|Gp'_i|^p] \lesssim (\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-4}(\mathbf{x}_i^\top \mathbf{v})^2])^p \leq \tau^{4p} D_2(\Psi)^p \quad (\text{C.260})$$

$$\begin{aligned} \mathbb{E}[\exp(\theta Gp'_i)] &\lesssim 1 + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}[|Gp'_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty} \frac{|e\theta|^p \tau^{4p} D_2(\Psi)^p}{p^p} \\ &\lesssim 1 + \theta^2 \tau^8 D_2(\Psi)^2 \end{aligned} \quad (\text{C.261})$$

Hence, similar to the derivation of (C.194), we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Gp_i \geq t\right) \leq \exp\left(-n\theta t + C_{17}n\theta^2(\tau^4 D_2(\Psi))^2\right) \quad (\text{C.262})$$

for some constant  $C_{17} > 0$ . Let  $\theta = \frac{t}{2C_{17}(\tau^4 D_2(\Psi))^2}$  and  $t = \delta_1^2(\mathbf{W}^*) \cdot (\tau^4 D_2(\Psi)) \cdot \sqrt{\frac{d \log n}{n}}$ , then we have

$$\begin{aligned} \|\hat{\mathbf{Q}}_2 - \mathbf{Q}_2\| &\lesssim \delta_1^2(\mathbf{W}^*) \cdot \tau^4 D_2(\Psi) \cdot \sqrt{\frac{d \log n}{n}} \\ &\lesssim \sqrt{\frac{d \log n}{n}} \cdot \delta_1^2(\mathbf{W}^*) \cdot \tau^6 \sqrt{D_2(\Psi) D_4(\Psi)} \end{aligned} \quad (\text{C.263})$$

with probability at least  $1 - 2n^{-\Omega(\delta_1^4(\mathbf{W}^*)d)}$ .

#### C.6.4 Proof of Lemma C.6.2

We consider each component of  $y = \frac{1}{K} \sum_{i=1}^K \phi(\mathbf{w}_i^{*\top} \mathbf{x})$ .

Define  $\mathbf{T}_i(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^{K \times K \times K}$  such that

$$\mathbf{T}_i(\mathbf{x}) = [\phi(\mathbf{w}_i^{*\top} \mathbf{x}) \cdot S_3(\mathbf{x})](\hat{\mathbf{U}}, \hat{\mathbf{U}}, \hat{\mathbf{U}}) \quad (\text{C.264})$$

We flatten  $\mathbf{T}_i(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^{K \times K \times K}$  along the first dimension to obtain the function  $\mathbf{B}_i(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^{K \times K^2}$ . Similar to the derivation of the last step of Lemma E.8 in [71], we can obtain  $\|\mathbf{T}_i(\mathbf{x})\| \leq \|\mathbf{B}_i(\mathbf{x})\|$ . By (C.248), we have

$$\|\mathbf{B}_i(\mathbf{x})\| \lesssim \sigma_{\min}^{-6} \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*\top} \mathbf{x}_i) (\hat{\mathbf{U}}^\top \mathbf{x})^3 \quad (\text{C.265})$$

Define  $Gr_i = \langle \mathbf{v}, \mathbf{B}_i(\mathbf{x}_i) \rangle - \mathbb{E}[\mathbf{B}_i(\mathbf{x}_i)]\mathbf{v} \rangle$ , where  $\|\mathbf{v}\| = 1$ , so  $\mathbb{E}[Gr_i] = 0$ . Similar to the proof of (C.190), (C.191) and (C.192) in Lemma C.4.5, we have

$$|Gr_i|^p \lesssim |\sigma_{\min}^{-6} (\mathbf{v}^\top \hat{\mathbf{U}}^\top \mathbf{x})^3 + \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-6} (\mathbf{v}^\top \hat{\mathbf{U}}^\top \mathbf{x})^3]|^p \quad (\text{C.266})$$

$$\mathbb{E}[|Gr_i|^p] \lesssim (\mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} [\sigma_{\min}^{-6} (\mathbf{v}^\top \hat{\mathbf{U}}^\top \mathbf{x})^3])^p \lesssim \tau^{6p} \sqrt{D_6(\Psi)}^p \quad (\text{C.267})$$

$$\begin{aligned} \mathbb{E}[\exp(\theta Gr_i)] &\lesssim 1 + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}[|Gr_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty} \frac{|e\theta|^p \tau^{6p} D_6(\Psi)^{\frac{p}{2}}}{p^p} \\ &\leq 1 + \theta^2 (\tau^{12} \sqrt{D_6(\Psi)})^2 \end{aligned} \quad (\text{C.268})$$

Hence, similar to the derivation of (C.194), we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Gr_i \geq t\right) \leq \exp\left(-n\theta t + C_{18}\theta^2 (\tau^6 \sqrt{D_6(\Psi)})^2\right) \quad (\text{C.269})$$

for some constant  $C_{18} > 0$ . Let  $\theta = \frac{t}{C_{18}(\tau^6 \sqrt{D_6(\Psi)})^2}$  and  $t = \delta_1^2(\mathbf{W}^*) \cdot (\tau^6 \sqrt{D_6(\Psi)}) \cdot \sqrt{\frac{\log n}{n}}$ , then we have

$$\|\hat{\mathbf{R}}_3 - \mathbf{R}_3\| \lesssim \delta_1(\mathbf{W}^*)^2 \cdot (\tau^6 \sqrt{D_6(\Psi)}) \cdot \sqrt{\frac{\log n}{n}} \quad (\text{C.270})$$

with probability at least  $1 - 2n^{-\Omega(\delta_1^4(\mathbf{W}^*))}$ .

### C.6.5 Proof of Lemma C.6.3

From Definition C.1.1, we have

$$\|\hat{\mathbf{Q}}_1 - \mathbf{Q}_1\| = \left\| \frac{1}{n} \sum_{i=1}^n [y_i \cdot S_1(\mathbf{x}_i)] - \mathbb{E}[y \cdot S_1(\mathbf{x})] \right\|. \quad (\text{C.271})$$

Based on Definition C.1.1,

$$\begin{aligned} & \left\| [y_i \cdot S_1(\mathbf{x}_i)] \right\| \\ & \lesssim \left\| \frac{\sum_{l=1}^L \lambda_l \lambda_l (2\pi \prod_{k=1}^d \sigma_{lk}^2)^{-\frac{d}{2}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l) \Sigma_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l)) \cdot (\mathbf{x} - \boldsymbol{\mu}_l) \Sigma_l^{-1}}{\sum_{l=1}^L \lambda_l \lambda_l (2\pi \prod_{k=1}^d \sigma_{lk}^2)^{-\frac{d}{2}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_l) \Sigma_l^{-1}(\mathbf{x} - \boldsymbol{\mu}_l))} \right\| \\ & \lesssim \left\| \sigma_{\min}^{-2} \frac{1}{K} \sum_{j=1}^K \phi(\mathbf{w}_j^{*\top} \mathbf{x}_i) \mathbf{x}_i \right\| \end{aligned} \quad (\text{C.272})$$

Define  $Gq_i = \langle \mathbf{v}, ([y_i \cdot S_1(\mathbf{x}_i)] - \mathbb{E}[[y_i \cdot S_1(\mathbf{x}_i)]] \mathbf{v}) \rangle$ , where  $\|\mathbf{v}\| = 1$ , so  $\mathbb{E}[Gq_i] = 0$ . Similar to the proof of (C.190), (C.191), and (C.192) in Lemma C.4.5, we have

$$|Gq_i|^p \lesssim \left| \sigma_{\min}^{-2} (\mathbf{x}_i^\top \mathbf{v}) + \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \Sigma_l)} [\sigma_{\min}^{-2} (\mathbf{x}_i^\top \mathbf{v})] \right|^p \quad (\text{C.273})$$

$$\mathbb{E}[|Gq_i|^p] \lesssim \left( \mathbb{E}_{\mathbf{x} \sim \sum_{l=1}^L \mathcal{N}(\boldsymbol{\mu}_l, \Sigma_l)} [\sigma_{\min}^{-2} (\mathbf{x}_i^\top \mathbf{v})] \right)^p \leq \tau^{2p} \sqrt{D_2(\Psi)}^p \quad (\text{C.274})$$

$$\begin{aligned}\mathbb{E}[\exp(\theta Gq_i)] &\lesssim 1 + \sum_{p=2}^{\infty} \frac{\theta^p \mathbb{E}[|Gq_i|^p]}{p!} \lesssim 1 + \sum_{p=2}^{\infty} \frac{|e\theta|^p \tau^{2p} D_2(\Psi)^{\frac{p}{2}}}{p^p} \\ &\leq 1 + \theta^2 (\tau^2 \sqrt{D_2(\Psi)})^2\end{aligned}\tag{C.275}$$

Hence, similar to the derivation of (C.194), we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Gq_i \geq t\right) \leq \exp\left(-n\theta t + C_{19}\theta^2 (\tau^2 \sqrt{D_2(\Psi)})^2\right)\tag{C.276}$$

for some constant  $C_{19} > 0$ . Let  $\theta = \frac{t}{C_{19}(\tau^2 \sqrt{D_2(\Psi)})^2}$  and  $t = (\tau^2 \sqrt{D_2(\Psi)}) \cdot \sqrt{\frac{d \log n}{n}}$ , then we have

$$||\hat{\mathbf{Q}}_1 - \mathbf{Q}_1|| \lesssim (\tau^2 \sqrt{D_2(\Psi)}) \cdot \sqrt{\frac{d \log n}{n}}\tag{C.277}$$

with probability at least  $1 - 2n^{-\Omega(d)}$ .

## APPENDIX D

### APPENDIX OF CHAPTER 5

#### D.1 Preliminaries

**Lemma D.1.1.**  $\|\tilde{\mathbf{a}}_n \mathbf{X}\| \leq \|\mathbf{A}\|_\infty$ .

Proof:

$$\begin{aligned}
\|\tilde{\mathbf{a}}_n \mathbf{X}\| &= \left\| \sum_{k=1}^N a_{n,k} \tilde{\mathbf{x}}_k \right\| \\
&= \left\| \sum_{k=1}^N \frac{a_{n,k}}{\sum_{k=1}^N a_{n,k}} \tilde{\mathbf{x}}_k \right\| \cdot \sum_{k=1}^N a_{n,k} \\
&\leq \sum_{k=1}^N \frac{a_{n,k}}{\sum_{k=1}^N a_{n,k}} \|\tilde{\mathbf{x}}_k\| \cdot \|\mathbf{A}\|_\infty \\
&= \|\mathbf{A}\|_\infty
\end{aligned} \tag{D.1}$$

where the second to last step is by the convexity of  $\|\cdot\|$ .

**Lemma D.1.2.** *Given a graph  $\mathcal{G}$  with  $L(\geq 1)$  groups of nodes, where the group  $i$  with node degree  $d_i$  is denoted as  $\mathcal{N}_i$ . Suppose that in iteration  $t$ ,  $\mathbf{A}^t$  (or any of  $\mathbf{A}^{t(1)}$ ,  $\mathbf{A}^{t(2)}$ ,  $\mathbf{A}^{t(3)}$  in the general setting) is generated from the sampling strategy in Section 5.3.2, if the number of sampled nodes satisfies  $l_i \geq |\mathcal{N}_i|/(1 + \frac{c_1 \text{poly}(\epsilon)}{L p_i^* \Psi_i})$ , we have*

$$\|\mathbf{A}^t - \mathbf{A}^*\|_\infty \leq \text{poly}(\epsilon) \tag{D.2}$$

**Proof:**

From Section 5.3.2, we can rewrite that

$$\tilde{\mathbf{a}}_n^t = \begin{cases} \frac{|\mathcal{N}_k|}{l_k} p_k^* A_{n,j}, & \text{if the nodes } n, j \text{ are connected and } j \text{ is selected and } j \in \mathcal{N}_k \\ 0, & \text{else} \end{cases} \tag{D.3}$$

---

Portions of this appendix previously appeared as: H. Li, M. Wang, S. Liu, P.-Y. Chen, and J. Xiong, “Generalization guarantee of training graph convolutional networks with graph topology sampling,” in *Proc. Int. Conf. Mach. Learn.*, Jul. 2022, pp. 13014–13051.

$$\tilde{\mathbf{a}}^*_n = \begin{cases} p_k^* A_{n,j}, & \text{if the nodes } n, j \text{ are connected and } j \in \mathcal{N}_k \\ 0, & \text{else} \end{cases} \quad (\text{D.4})$$

Let  $\mathbf{A}^* = (\tilde{\mathbf{a}}_1^*, \tilde{\mathbf{a}}_2^*, \dots, \tilde{\mathbf{a}}_n^*)^\top$ . Since that we need that  $\sum_{j=1}^N A_{n,j}^* \leq O(1)$ , we require

$$p_i^* \sum_{j \in \mathcal{N}_i} A_{n,j} \leq O(1/L), \text{ holds for any } i \in [L], n \in [N] \quad (\text{D.5})$$

We first roughly compute the ratio of edges that one node is connected to the nodes in another group. For the node with degree  $\deg(i)$ , it has  $\deg(i) - 1$  open edges except the self-connection. Hence, the group with degree  $\deg(j)$  has  $(\deg(j) - 1)|\mathcal{N}_j|$  open edges except self-connections in total. Therefore, the ratio of the edges connected to the group  $j$  to all groups is

$$\frac{(\deg(j) - 1)|\mathcal{N}_j|}{\sum_{l=1}^L (\deg(l) - 1)|\mathcal{N}_l|} \approx \frac{d_j |\mathcal{N}_j|}{\sum_{l=1}^L d_l |\mathcal{N}_l|} \quad (\text{D.6})$$

Define

$$\Psi(n, i) = \sqrt{\frac{d_n}{d_i}} \cdot \frac{d_i |\mathcal{N}_i|}{\sum_{l=1}^L d_l |\mathcal{N}_l|} \quad (\text{D.7})$$

Then, as long as

$$p_i^* \sum_{j \in \mathcal{N}_i} A_{n,j} \approx p_i^* \frac{1}{\sqrt{d_i d_n}} \cdot \frac{d_i |\mathcal{N}_i|}{\sum_{l=1}^L d_l |\mathcal{N}_l|} d_n \lesssim p_i^* \Psi(n, i) \leq O(1/L) \quad (\text{D.8})$$

i.e.,

$$p_i^* \leq \frac{c_1}{L \cdot \max_{n \in [L]} \{\Psi(n, i)\}} = \frac{c_1}{L \cdot \Psi(L, i)} = \frac{c_1}{L} \sqrt{\frac{d_L}{d_L}} \frac{\sum_{l=1}^L d_l |\mathcal{N}_l|}{d_L |\mathcal{N}_L|} \quad (\text{D.9})$$

for some constant  $c_1 > 0$ , we can obtain that  $\|\mathbf{A}^*\|_\infty \leq O(1)$ . Since that

$$\sum_{j \in \mathcal{S}_k} A_{n,j} \approx \frac{1}{\sqrt{d_i d_n}} \cdot \frac{d_i |\mathcal{N}_i|}{\sum_{l=1}^L d_l |\mathcal{N}_l|} d_n \frac{l_k}{|\mathcal{N}_k|} \approx \sum_{j \in \mathcal{N}_k} A_{n,j} \frac{l_k}{|\mathcal{N}_k|} \quad (\text{D.10})$$

$$\sum_{j \notin \mathcal{S}_k} A_{n,j} \approx \frac{1}{\sqrt{d_i d_n}} \cdot \frac{d_i |\mathcal{N}_i|}{\sum_{l=1}^L d_l |\mathcal{N}_l|} d_n \left(1 - \frac{l_k}{|\mathcal{N}_k|}\right) \approx \sum_{j \in \mathcal{N}_k} A_{n,j} \left(1 - \frac{l_k}{|\mathcal{N}_k|}\right), \quad (\text{D.11})$$

the difference between  $\tilde{\mathbf{a}}_n^t$  and  $\tilde{\mathbf{a}}_{n,n}^*$  can then be derived as

$$\begin{aligned}
& \|\tilde{\mathbf{a}}_n^t - \tilde{\mathbf{a}}_{n,n}^*\|_1 \\
&= \left| \sum_{k=1}^L \sum_{j \in \mathcal{S}_k} A_{n,j} p_k^* \left( \frac{|\mathcal{N}_k|}{l_k} - 1 \right) + \sum_{k=1}^L \sum_{j \notin \mathcal{S}_k} A_{n,j} p_k^* \right| \\
&\lesssim \sum_{k=1}^L \left( p_k^* \left( \frac{|\mathcal{N}_k|}{l_k} - 1 \right) \frac{l_k}{|\mathcal{N}_k|} \sum_{j \in \mathcal{N}_k} A_{n,j} + \left( 1 - \frac{l_k}{|\mathcal{N}_k|} \right) p_k^* \sum_{j \in \mathcal{N}_k} A_{n,j} \right) \\
&\lesssim \text{poly}(\epsilon) \sum_{k=1}^L \frac{1}{L\Psi(L,k)} \sum_{j \in \mathcal{N}_k} A_{n,j} \\
&:= \text{poly}(\epsilon) \Gamma(\mathbf{A}^*)
\end{aligned} \tag{D.12}$$

where the first inequality is by (D.10, D.11) and the second inequality holds as long as  $l_i \geq |\mathcal{N}_i|/(1 + \frac{c_1 \text{poly}(\epsilon)}{L p_i^* \Psi(L,i)})$ . Combining (D.8), we have

$$\sum_{i=1}^L p_i^* \sum_{j \in \mathcal{N}_i} A_{n,j} \lesssim \sum_{i=1}^L \frac{1}{L\Psi(L,i)} \sum_{j \in \mathcal{N}_i} A_{n,j} = \Gamma(\mathbf{A}^*) \leq O(1) \tag{D.13}$$

Hence, (D.12) can be bounded by  $\text{poly}(\epsilon)$ .

### D.1.1 Symmetric Graph Sampling Method

We provide and discuss a symmetric graph sampling method in this section. The insights behind this version of sampling strategy is the same as in Section 5.3.2.

Similar to the asymmetric construction in Section 5.3.2, we consider a group-wise uniform sampling strategy, where  $S_l$  nodes are sampled uniformly from  $N_l$  nodes. For all unsampled nodes, we set the corresponding diagonal entries of a diagonal matrix  $\mathbf{P}^s$  to be zero. If node  $i$  is sampled in this iteration and belongs to group  $l$  for any  $i$  and  $l$ , the  $i$ th diagonal entry of  $\mathbf{P}^s$  is set as  $\sqrt{p_l^* N_l / S_l}$  for some non-negative constant  $p_l^*$ . Then  $\mathbf{A}^s = \mathbf{P}^s \mathbf{A} \mathbf{P}^s$ .

Based on this symmetric graph sampling method, we define the effective adjacency matrix as

$$\mathbf{A}^* = \mathbf{P}^* \mathbf{A} \mathbf{P}^*, \tag{D.14}$$

where  $\mathbf{P}^*$  is a diagonal matrix defined as

$$\mathbf{P}_{ii}^* = \sqrt{p_l^*} \quad \text{if node } i \text{ belongs to degree group } l \quad (\text{D.15})$$

The scaling factor  $p_l^*$  should satisfy

$$0 \leq p_l^* \leq \frac{c_2}{L^2 \psi_l^2}, \quad \forall l \quad (\text{D.16})$$

for a positive constant  $c_2$  that can be sufficiently large.  $\psi_l$  is defined in (5.9). The number of sampled nodes shall satisfy

$$\frac{S_l}{N_l} \geq (1 + \frac{c_2 \text{poly}(\epsilon)}{L \sqrt{p_l^*} \psi_l})^{-2} \quad \forall l \in [L] \quad (\text{D.17})$$

where  $\epsilon$  is a small positive value.

**Lemma D.1.3.** *Given a graph  $\mathcal{G}$  with  $L (\geq 1)$  groups of nodes, where the group  $i$  with node degree  $d_i$  is denoted as  $\mathcal{N}_i$ . Suppose  $\mathbf{A}^t$  (or any of  $\mathbf{A}^{t(1)}, \mathbf{A}^{t(2)}, \mathbf{A}^{t(3)}$  in the general setting) is generated from the sampling strategy in Section D.1.1, if the number of sampled nodes satisfies  $l_i \geq |\mathcal{N}_i| / (1 + \frac{c_2 \text{poly}(\epsilon)}{L p_i^* \Psi_i})$ , then we have*

$$\|\mathbf{A}^t - \mathbf{A}^*\|_\infty \leq \text{poly}(\epsilon) \quad (\text{D.18})$$

**Proof:**

From Section D.1.1, we can rewrite that

$$\tilde{\mathbf{a}}_n^t = \begin{cases} \sqrt{\frac{|\mathcal{N}_k||\mathcal{N}_u|}{l_k l_u}} p_k^* p_u^* A_{n,j}, & \text{if } n, j \text{ are connected, } j \text{ is selected and } n \in \mathcal{N}_u, j \in \mathcal{N}_k \\ 0, & \text{else} \end{cases} \quad (\text{D.19})$$

$$\tilde{\mathbf{a}}_n^* = \begin{cases} \sqrt{p_k^* p_u^*} A_{n,j}, & \text{if the nodes } n, j \text{ are connected and } n \in \mathcal{N}_u, j \in \mathcal{N}_k \\ 0, & \text{else} \end{cases} \quad (\text{D.20})$$

Then  $\mathbf{A}^* = (\tilde{\mathbf{a}}_1^{*\top}, \tilde{\mathbf{a}}_2^{*\top}, \dots, \tilde{\mathbf{a}}_n^{*\top})^\top$ . Then, for  $n \in \mathcal{N}_u$ , as long as

$$\begin{aligned} \sum_{j \in |\mathcal{N}_i|} \sqrt{p_i^* p_u^*} A_{n,j} &\approx \sqrt{p_u^* p_i^*} \frac{1}{\sqrt{d_i d_n}} \cdot \frac{d_i |\mathcal{N}_i|}{\sum_{l=1}^L d_l |\mathcal{N}_l|} d_n \\ &\lesssim \sqrt{p_u^* p_i^*} \Psi(n, i) \leq \sqrt{p_i^*} \Psi(n, i) \leq O(1/L) \end{aligned} \quad (\text{D.21})$$

i.e.,

$$\sqrt{p_i^*} \leq \frac{c_2}{L \cdot \max_{n \in [L]} \{\Psi(n, i)\}} = \frac{c_1}{L \cdot \Psi(L, i)} = \frac{c_2}{L} \sqrt{\frac{d_i}{d_L} \frac{\sum_{l=1}^L d_l |\mathcal{N}_l|}{d_i |\mathcal{N}_i|}} \quad (\text{D.22})$$

for some constant  $c_2 > 0$ , we can obtain that  $\|\mathbf{A}^*\|_\infty \leq O(1)$ .

The difference between  $\tilde{\mathbf{a}}_n^t$  and  $\tilde{\mathbf{a}}_n^*$  can then be derived as

$$\begin{aligned} &\|\tilde{\mathbf{a}}_n^t - \tilde{\mathbf{a}}_n^*\|_1 \\ &= \left| \sum_{k=1}^L \sum_{j \in \mathcal{S}_k} A_{n,j} \sqrt{p_u^* p_k^*} \left( \sqrt{\frac{|\mathcal{N}_k||\mathcal{N}_u|}{l_k l_u}} - 1 \right) + \sum_{k=1}^L \sum_{j \notin \mathcal{S}_k} A_{n,j} \sqrt{p_u^* p_k^*} \right| \\ &\approx \left| \sum_{k=1}^L \sum_{j \in \mathcal{N}_k} A_{n,j} \sqrt{p_u^* p_k^*} \left( \sqrt{\frac{|\mathcal{N}_k||\mathcal{N}_u|}{l_k l_u}} - 1 \right) \frac{l_k}{|\mathcal{N}_k|} + \sum_{k=1}^L \sum_{j \in \mathcal{N}_k} A_{n,j} \sqrt{p_u^* p_k^*} \left( 1 - \frac{l_k}{|\mathcal{N}_k|} \right) \right| \\ &\lesssim \text{poly}(\epsilon) \end{aligned} \quad (\text{D.23})$$

as long as  $l_i \geq |\mathcal{N}_i| / (1 + \frac{c_2 \text{poly}(\epsilon)}{L \sqrt{p_i^*} \Psi(L, i)})^2$ .

## D.2 Node Classification for Three Layers

In the whole proof, we consider a more general target function compared to (5.12). We write  $F^* : \mathbb{R}^N \times \mathbb{R}^{N \times d} \rightarrow \mathbb{R}^K$ :

$$\begin{aligned} F_{\mathbf{A}^*}^* &= (f_1^*, f_2^*, \dots, f_K^*), \\ f_r^*(\mathbf{e}_g, \mathbf{X}) &= \mathbf{e}_g^\top \sum_{k \in [p_1]} c_{k,r}^* \Phi \left( \mathbf{A}^* \sum_{j \in [p_2]} v_{1,k,j}^* \phi_{1,j}(\mathbf{A}^* \mathbf{X} \mathbf{w}_{1,j}^*) \right) \\ &\quad \odot \left( \mathbf{A}^* \sum_{l \in [p_2]} v_{2,k,l}^* \phi_{2,l}(\mathbf{A}^* \mathbf{X} \mathbf{w}_{2,l}^*) \right), \end{aligned} \quad (\text{D.24})$$

$\forall r \in [K]$ , where each  $\phi_{1,j}$ ,  $\phi_{2,l}$ ,  $\Phi_i : \mathbb{R} \rightarrow \mathbb{R}$  is infinite-order smooth.

Table D.1 shows some important notations used in our theorem and algorithm. Table D.2 gives the full parameter choices for the three-layer GCN.  $\text{ploy}(\log(m_1 m_2))$  in the following

analysis.

**Table D.1: Summary of notations.**

$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$	$\mathcal{G}$ is an un-directed graph consisting of a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$ .
$N$	The total number of nodes in a graph.
$\mathbf{A} = \mathbf{D}^{-\frac{1}{2}} \tilde{\mathbf{A}} \mathbf{D}^{-\frac{1}{2}}$	$\mathbf{A} \in \mathbb{R}^{N \times N}$ is the normalized adjacency matrix computed by the degree matrix $\mathbf{D}$ and the initial adjacency matrix $\tilde{\mathbf{A}}$ .
$\mathbf{A}^*$	The effective adjacency matrix.
$\mathbf{A}^t$	The sampled adjacency matrix using our sampling strategy in Section 5.3.2 at the $t$ -th iteration.
$e_g, \mathbf{X}, y_n$	$e_g$ belongs to $\{\mathbf{e}_i\}_{i=1}^N$ and selects the index of the node label. $\mathbf{X} \in \mathbb{R}^{N \times d}$ is the feature matrix. $y_n$ is the label of the $n$ -th node.
$m_1, m_2$	$m_1, m_2$ are the number of neurons in the first and second hidden layer, respectively.
$\mathbf{W}, \mathbf{V}, \mathbf{B}_1, \mathbf{B}_2$	$\mathbf{X} \in \mathbb{R}^{N \times d}$ is the data matrix. $\mathbf{W}, \mathbf{V}$ are the weight matrices of the first and second hidden layer, respectively. $\mathbf{B}_1, \mathbf{B}_2$ are the corresponding bias matrices.
$\mathbf{W}^{(0)}, \mathbf{V}^{(0)}$	$\mathbf{W}^{(0)}$ and $\mathbf{V}^{(0)}$ are random initializations of $\mathbf{W}$ and $\mathbf{V}$ , respectively.
$\mathbf{W}^\rho, \mathbf{V}^\rho$	$\mathbf{W}^\rho$ and $\mathbf{V}^\rho$ are two random matrices used for Gaussian smoothing.
$\Sigma$	The Dropout technique.
$\Omega, \Omega^t$	$\Omega$ is the set of labeled nodes and $\Omega_t$ is the batch of labeled nodes at the $t$ -th iteration.
$T, T_w, \eta, \lambda_t$	In Algorithm 2, $T$ is the number of outer iterations for the weight decay step, while $T_w$ is the number of inner iterations for the SGD steps. $\eta$ is the step size and $\lambda_t$ is the weight decay coefficient at the $t$ -th iteration.
$L, d_l, S_l, N_l$	$L$ is the number of node groups in a graph. $d_l$ is the order-wise degree in the $l$ -th group. $N_l$ is the number of nodes in group $l$ .
$S_l$	The number of nodes we sample in group $l$ .

### D.2.1 Lemmas

**Function Approximation** To show that the target function can be learnt by the learner network with the Relu function, a good approach is to firstly find a function  $h(\cdot)$  such that the  $\phi$  functions in the target function can be approximated by  $h(\cdot)$  with an indicator function. In this section, Lemma D.2.1 provides the existence of such  $h(\cdot)$  function. Lemma D.2.2 and D.2.3 are two supporting lemmas to prove Lemma D.2.1.

**Table D.2:** Full parameter choices for three-layer GCN.

$\tau'_v$	$m_1^{1/2-0.005}/(\sqrt{\epsilon_0}m_2^{1/2})$
$\tau'_w$	$C_0/(\epsilon_0^{1/4}m_1^{3/4-0.005})$
$\tau_v$	$m_1^{1/2-0.001}/m_2^{1/2} \gg \tau'_v$
$\tau_w$	$1/m_1^{3/4-0.01} \gg \tau'_w$
$\lambda_v$	$2/(\tau'_v)^2$
$\lambda_w$	$2/(\tau'_w)^4$
$\sigma_v$	$1/m_2^{1/2+0.01}$
$\sigma_w$	$\sigma_w = 1/m_1^{1-0.01}$
$C$	$\mathcal{C}_\epsilon(\phi, \ \mathbf{A}\ _\infty) \sqrt{\ \mathbf{A}\ _\infty^2 + 1}$
$C'$	$10C\sqrt{p_2}$
$C''$	$\mathcal{C}_\epsilon(\Phi, C') \sqrt{\ \mathbf{A}\ _\infty^2 + 1}$
$C_0$	$\tilde{O}(p_1^2 p_2 K^2 C C'')$
$\epsilon_c$	1

**Lemma D.2.1.** For every smooth function  $\phi$ , every  $\epsilon \in (0, \frac{1}{\mathcal{C}(\phi, a)\sqrt{a^2+1}})$ , there exists a function  $h : \mathbb{R}^2 \rightarrow [-\mathcal{C}_\epsilon(\phi, a)\sqrt{a^2+1}, \mathcal{C}_\epsilon(\phi, a)\sqrt{a^2+1}]$  that is also  $\mathcal{C}_\epsilon(\phi, a)\sqrt{a^2+1}$ -Lipschitz continuous on its first coordinate with the following two (equivalent) properties:

(a) For every  $x_1 \in [-a, a]$  where  $a > 0$ :

$$\left| \mathbb{E} \left[ \mathbb{1}_{\alpha_1 x_1 + \beta_1 \sqrt{a^2 - x_1^2} + b_0 \geq 0} h(\alpha_1, b_0) \right] - \phi(x_1) \right| \leq \epsilon$$

where  $\alpha_1, \beta_1, b_0 \sim \mathcal{N}(0, 1)$  are independent random variables.

(b) For every  $\mathbf{w}^*, \mathbf{x} \in \mathbb{R}^d$  with  $\|\mathbf{w}^*\|_2 = 1$  and  $\|\mathbf{x}\| \leq a$ :

$$\left| \mathbb{E} \left[ \mathbb{1}_{\mathbf{w}^\top \mathbf{x} + b_0 \geq 0} h(\mathbf{w}^\top \mathbf{w}^*, b_0) \right] - \phi(\mathbf{w}^{*\top} \mathbf{x}) \right| \leq \epsilon$$

where  $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$  is an  $d$ -dimensional Gaussian,  $b_0 \sim \mathcal{N}(0, 1)$ .

Furthermore, we have  $\mathbb{E}_{\alpha_1, b_0 \sim \mathcal{N}(0, 1)} [h(\alpha_1, b_0)^2] \leq (\mathcal{C}_s(\phi, a))^2(a^2 + 1)$ .

(c) For every  $\mathbf{w}^*, \mathbf{x} \in \mathbb{R}^d$  with  $\|\mathbf{w}^*\|_2 = 1$ , let  $\tilde{\mathbf{w}} = (\mathbf{w}, b_0) \in \mathbb{R}^{d+1}$ ,  $\tilde{\mathbf{x}} = (\mathbf{x}, 1) \in \mathbb{R}^{d+1}$  with  $\|\tilde{\mathbf{x}}\| \leq \sqrt{a^2 + 1}$ , then we have

$$\left| \mathbb{E} \left[ \mathbb{1}_{\tilde{\mathbf{w}}^\top \tilde{\mathbf{x}} \geq 0} h(\tilde{\mathbf{w}}[1:d]^\top \mathbf{w}^*, \tilde{\mathbf{w}}[d+1]) \right] - \phi(\mathbf{w}^{*\top} \tilde{\mathbf{x}}[1:d]) \right| \leq \epsilon$$

where  $\tilde{\mathbf{w}} \sim \mathcal{N}(0, \mathbf{I}_{d+1})$  is an  $d$ -dimensional Gaussian.

We also have  $\mathbb{E}_{\tilde{\mathbf{w}} \in \mathcal{N}(0, \mathbf{I}_{d+1})} [h(\tilde{\mathbf{w}}[1:d]^\top \mathbf{w}^*, \tilde{\mathbf{w}}[d+1])^2] \leq (\mathcal{C}_s(\phi, a))^2(a^2 + 1)$ .

**Proof:**

Firstly, since we can assume  $\mathbf{w}^* = (1, 0, \dots, 0)$  without loss of generality by rotating  $\mathbf{x}$  and  $\mathbf{w}$ , it can be derived that  $\mathbf{x}$ ,  $\mathbf{w}$ ,  $\mathbf{w}^*$  are equivalent to that they are two-dimensional. Therefore, proving Lemma D.2.1b suffices in showing Lemma D.2.1a.

Let  $\mathbf{w}_0 = (\alpha, \beta)$ ,  $\mathbf{x} = (x_1, \sqrt{t^2 - x_1^2})$  where  $\alpha$  and  $\beta$  are independent. Following the idea of Lemma 6.3 in [58], we use another randomness as an alternative, i.e., we write  $\mathbf{x}^\perp = (\sqrt{t^2 - x_1^2}, -x_1)$ ,  $\mathbf{w}_0 = \alpha \frac{\mathbf{x}}{t} + \beta \frac{\mathbf{x}^\perp}{t} \sim \mathcal{N}(0, \mathbf{I})$ . Then  $\mathbf{w}_0 \mathbf{X} = t\alpha$ . Let  $\alpha_1 = w_{01} = \alpha \frac{x_1}{t} + \beta \sqrt{1 - \frac{x_1^2}{t^2}}$ , where  $\alpha, \beta \sim \mathcal{N}(0, 1)$ . Hence,  $\alpha_1 \sim \mathcal{N}(0, 1)$ .

We first use Lemma D.2.2 to fit  $\phi(x_1)$ . By Taylor expansion, we have

$$\begin{aligned} \phi(x_1) &= c_0 + \sum_{i=1, \text{ odd } i}^{\infty} c_i x_1^i + \sum_{i=2, \text{ even } i}^{\infty} c_i x_1^i \\ &= c_0 + \sum_{i=1}^{\infty} c'_i \mathbb{E}_{\alpha, \beta \sim \mathcal{N}(0, 1)} [h_i(\alpha_1) \mathbb{1}[q_i(b_0)] \mathbb{1}[\mathbf{w}_0 \mathbf{X} + b_0 \geq 0]] \end{aligned} \quad (\text{D.25})$$

where  $h_i(\cdot)$  is the Hermite polynomial defined in Definition A.5 in [58], and

$$c'_i = \frac{c_i}{p'_i}, \quad |c'_i| \leq \frac{200i^2|c_i| \sqrt{t^2 + 1}}{(i-1)!!} \frac{1}{t^{1-i}} \quad \text{and} \quad q_i(b_0) = \begin{cases} |b_0| \leq t/(2i), & i \text{ is odd} \\ 0 < -b_0 \leq t/(2i), & i \text{ is even} \end{cases} \quad (\text{D.26})$$

Let  $B_i = 100i^{\frac{1}{2}} + 10\sqrt{\log(\frac{1}{\epsilon} \frac{\sqrt{t^2+1}}{t^{1-i}})}$ . Define  $\hat{h}_i(\alpha_1) = h_i(\alpha_1) \cdot \mathbb{1}[|\alpha_1| \leq B_i] + h_i(\text{sign}(\alpha_1)B_i) \cdot \mathbb{1}[|\alpha_1| > B_i]$  as the truncated version of the Hermite polynomial. Then we have

$$\phi(x_1) = c_0 + R(x_1) + \sum_{i=1}^{\infty} c'_i \mathbb{E}_{\alpha, \beta \sim \mathcal{N}(0, 1)} [\hat{h}_i(\alpha_1) \mathbb{1}[q_i(b_0)] \mathbb{1}[\mathbf{w}_0 \mathbf{X} + b_0 \geq 0]],$$

where

$$\begin{aligned} R(x_1) &= \sum_{i=1}^{\infty} c'_i \mathbb{E}_{\alpha, \beta \sim \mathcal{N}(0, 1)} \left[ (h_i(\alpha_1) \cdot \mathbb{1}[|\alpha_1| > B_i] - h_i(\text{sign}(\alpha_1)B_i \cdot \mathbb{1}[|\alpha| > B_i])) \right. \\ &\quad \left. \mathbb{1}[q_i(b_0)] \mathbb{1}[\mathbf{w}_0 \mathbf{X} + b_0 \geq 0] \right] \end{aligned} \quad (\text{D.27})$$

Define

$$h(\alpha_1, b_0) = c_0 + \sum_{i=1}^{\infty} c'_i \cdot \hat{h}_i(\alpha_1) \cdot \mathbb{1}[q_i(b_0)]$$

Then by Lemma D.2.3, we have

$$|\mathbb{E}_{\alpha, \beta, b_0 \sim \mathcal{N}(0,1)}[\mathbb{1}[\mathbf{w}_0 \mathbf{X} + b_0 \geq 0] \cdot h(\alpha_1, b_0) - \phi(x_1)| \leq |R(x_1)| \leq \frac{\epsilon}{4}$$

We also have

$$\begin{aligned} \mathbb{E}_{\alpha_1, b_0 \sim \mathcal{N}(0,1)}[h(\alpha_1, b_0)^2] &\leq (\epsilon^2 + c_0^2) + O(1) \cdot \sum_{i=1}^{\infty} \frac{i! \cdot |c_i|^2 i^3}{((i-1)!!)^2} \cdot \left(\frac{\sqrt{t^2+1}}{t^{1-i}}\right)^2 \\ &\leq (\epsilon^2 + c_0^2) + \sum_{i=1}^{\infty} i^{3.5} \cdot |c_i|^2 \cdot \left(\frac{\sqrt{t^2+1}}{t^{1-i}}\right)^2 \\ &\leq (\epsilon^2 + c_0^2) + \left(\sum_{i=0}^{\infty} (i+1)^{1.75} \cdot |c_i| \cdot t^i \sqrt{t^2+1}\right)^2 \\ &\leq \mathcal{C}_s(\phi, t)^2 (t^2 + 1) \end{aligned} \tag{D.28}$$

**Lemma D.2.2.** Denote  $h_i(x)$  as the degree- $i$  Hermite polynomial as in Definition A.5 in [58]. For every integer  $i \geq 1$ , there exists constant  $p'_i$  with  $|p'_i| \geq \frac{t^{1-i}}{\sqrt{t^2+1}} \frac{(i-1)!!}{100i^2}$  such that

$$\text{for even } i : \quad x_1^i = \frac{1}{p'_i} \mathbb{E}_{\mathbf{w}_0 \sim \mathcal{N}(0, \mathbf{I}), b_0 \sim \mathcal{N}(0,1)}[h_i(\alpha_1) \mathbb{1}[\alpha \geq -\frac{b_0}{t}] \mathbb{1}[0 < -b_0 \leq \frac{t}{2i}]] \tag{D.29}$$

$$\text{for odd } i : \quad x_1^i = \frac{1}{p'_i} \mathbb{E}_{\mathbf{w}_0 \sim \mathcal{N}(0, \mathbf{I}), b_0 \sim \mathcal{N}(0,1)}[h_i(\alpha_1) \mathbb{1}[\alpha \geq -\frac{b_0}{t}] \mathbb{1}[|b_0| \leq \frac{t}{2i}]] \tag{D.30}$$

for  $\|\mathbf{x}\| \leq t$ .

### Proof:

For even  $i$ , by Lemma A.6 in [58], we have

$$\mathbb{E}_{\mathbf{w}_0 \sim \mathcal{N}(0, \mathbf{I}), b_0 \sim \mathcal{N}(0,1)}[h_i(\alpha_1) \mathbb{1}[\alpha \geq -\frac{b_0}{t}] \mathbb{1}[0 < -b_0 \leq \frac{t}{2i}]] = \mathbb{E}_{b_0 \sim \mathcal{N}(0,1)}[p_i \cdot \mathbb{1}[0 < -b_0 \leq \frac{t}{2i}]] \cdot \frac{x_1^i}{t^i}$$

, where

$$p_i = (i-1)!! \frac{\exp(-b_0^2/(2t^2))}{\sqrt{2\pi}} \sum_{r=1, r \text{ odd}}^{i-1} \frac{(-1)^{\frac{i-1-r}{2}}}{r!!} \binom{i/2-1}{(r-1)/2} (-b_0/t)^r$$

Define  $c_r = \frac{(-1)^{\frac{i-1-r}{2}}}{r!!} \binom{i/2-1}{(r-1)/2}$ . Then  $\text{sign}(c_r) = -\text{sign}(c_{r+2})$ . We can derive

$$\left| \frac{c_r(-b_0/t)^r}{c_{r-2}(-b_0/t)^{r-2}} \right| = \left| \left(\frac{b_0}{t}\right)^2 \frac{i+1-r}{r(r-1)} \right| \leq \frac{1}{4i} \leq \frac{1}{4}$$

Therefore,

$$\left| \sum_{r=1, r \text{ odd}}^{i-1} c_r (-b_0/t)^r \right| \geq \frac{3}{4} |b_0/t|$$

$$\begin{aligned} & |\mathbb{E}_{b_0 \sim \mathcal{N}(0,1)}[p_i \cdot \mathbb{1}[0 \leq -b_0/t \leq 1/(2i)]]| \cdot t^{-i} \\ & \geq |\mathbb{E}_{b_0 \sim \mathcal{N}(0,1)}[(i-1)!! \frac{\exp(-b_0^2/2t^2)}{\sqrt{2\pi}} \cdot \frac{3}{4} |b_0/t| \cdot \mathbb{1}[0 \leq -b_0/t \leq 1/(2i)]]| \cdot t^{-i} \\ & = t^{-i} \cdot \int_{-\frac{t}{2i}}^0 (i-1)!! \frac{\exp(-\frac{b_0^2}{2}(1+\frac{1}{t^2}))}{2\pi} \cdot \frac{3}{4} \left(-\frac{b_0}{t}\right) db_0 \\ & = t^{-i} \cdot \frac{t}{t^2+1} \exp\left(-\frac{b_0^2}{2}(1+\frac{1}{t^2})\right) (i-1)!! \frac{3}{8\pi} \Big|_{-\frac{t}{2i}}^0 \\ & = t^{-i} \frac{t}{t^2+1} (i-1)!! \frac{3}{8\pi} \left(1 - \exp\left(-\frac{t^2+1}{8i^2}\right)\right) \\ & \geq t^{1-i} \frac{(i-1)!!}{100i^2} \end{aligned} \tag{D.31}$$

For odd  $i$ , similarly by Lemma A.6 in [58], we can obtain

$$\mathbb{E}_{w_0 \sim \mathcal{N}(0, I), b_0 \sim \mathcal{N}(0, 1)}[h(\alpha_1) \mathbb{1}[\alpha \geq -\frac{b_0}{t}] \mathbb{1}[|b_0| \leq \frac{t}{2i}]] = \mathbb{E}_{b_0 \sim \mathcal{N}(0,1)}[p_i \cdot \mathbb{1}[|b_0| \leq \frac{t}{2i}]] \cdot \frac{x_1^i}{t^i}$$

, where

$$p_i = (i-1)!! \frac{\exp(-b_0^2/(2t^2))}{\sqrt{2\pi}} \sum_{r=1, r \text{ even}}^{i-1} \frac{(-1)^{\frac{i-1-r}{2}}}{r!!} \binom{i/2-1}{(r-1)/2} (-b_0/t)^r$$

Then we also have

$$\left| \frac{c_r(-b_0/t)^r}{c_{r-2}(-b_0/t)^{r-2}} \right| = \left| \left(\frac{b_0}{t}\right)^2 \frac{i+1-r}{r(r-1)} \right| \leq \frac{1}{4i} \leq \frac{1}{4}$$

Therefore,

$$\left| \sum_{r=1, r \text{ odd}}^{i-1} c_r (-b_0/t)^r \right| \geq \frac{3}{4} |c_0| = \frac{3}{4} \frac{(\frac{i}{2}-1)!}{\pi(\frac{1}{2} \cdot \frac{3}{2} \cdots \frac{i-1}{2})} \geq \frac{3}{4} \frac{1}{\pi^{\frac{i-1}{2}}} \geq \frac{3}{2\pi i}$$

$$\begin{aligned}
& |\mathbb{E}_{b_0 \sim \mathcal{N}(0,1)}[p_i \cdot \mathbb{1}[|b_0|/t \leq 1/(2i)]]| \cdot t^{-i} \\
& \geq t^{-i} \cdot |\mathbb{E}_{b_0 \sim \mathcal{N}(0,1)}[(i-1)!! \frac{\exp(-b_0^2/2t^2)}{\sqrt{2\pi}} \cdot \frac{3}{2\pi i} \cdot \mathbb{1}[|b_0|/t \leq 1/(2i)]]| \\
& = t^{-i} \cdot \int_{-\frac{t}{2i}}^{\frac{t}{2i}} (i-1)!! \frac{\exp(-\frac{b_0^2}{2}(1+\frac{1}{t^2}))}{2\pi} \cdot \frac{3}{2\pi i} db_0 \\
& = t^{-i} \cdot (i-1)!! \frac{3}{4\pi^2 i} \cdot \frac{t}{\sqrt{t^2+1}} \cdot \sqrt{2\pi} \cdot \left( 2\Phi(\frac{\sqrt{t^2+1}}{2i}) - 1 \right) \\
& = t^{-i} \cdot (i-1)!! \frac{3}{4\pi^2 i} \cdot \frac{t}{\sqrt{t^2+1}} \cdot \sqrt{2\pi} \cdot \frac{2\Phi(\frac{\sqrt{t^2+1}}{2}) - 1}{i} \\
& \geq \frac{t^{1-i}}{\sqrt{t^2+1}} \frac{(i-1)!!}{100i^2}
\end{aligned} \tag{D.32}$$

**Lemma D.2.3.** For  $B_i = 100i^{1/2} + 10\sqrt{\log(t^{i-1}\sqrt{t^2+1}/\epsilon_i^2)}$  where  $\epsilon_i^2 = t^{i-1}\sqrt{t^2+1}\epsilon^2$ , we have

1.  $\sum_{i=1}^{\infty} |c'_i| \cdot |\mathbb{E}_{x \sim \mathcal{N}(0,1)}[|h_i(x)| \cdot \mathbb{1}[|x| \geq b]]| \leq \frac{\epsilon}{8}\sqrt{t^2+1}$
2.  $\sum_{i=1}^{\infty} |c'_i| \cdot |\mathbb{E}_{x \sim \mathcal{N}(0,1)}[|h_i(b)| \cdot \mathbb{1}[|x| \geq b]]| \leq \frac{\epsilon}{8}\sqrt{t^2+1}$
3.  $\sum_{i=1}^{\infty} |c'_i| \mathbb{E}_{z \in \mathcal{N}(0,1)}[|h_i(z)| \mathbb{1}[|z| \leq B_i]] \leq \mathcal{C}_\epsilon(\phi, t)\sqrt{t^2+1}$
4.  $\sum_{i=1}^{\infty} |c'_i| \mathbb{E}_{z \in \mathcal{N}(0,1)}[|\frac{d}{dz}h_i(z)| \mathbb{1}[|z| \leq B_i]] \leq \mathcal{C}_\epsilon(\phi, t)\sqrt{t^2+1}$

**Proof:**

By the definition of Hermite polynomial in Definition A.5 in [58], we have

$$h_i(x) \leq \sum_{j=1}^{\lfloor i/2 \rfloor} \frac{|x|^{i-2j} i^{2j}}{j!}$$

Combining (D.26), we can obtain

$$|c'_i h_i(x)| \leq O(1) |c_i| \frac{\sqrt{t^2+1}}{t^{1-i}} \frac{i^4}{i!!} \sum_{j=1}^{\lfloor i/2 \rfloor} \frac{|x|^{i-2j} i^{2j}}{j!} \tag{D.33}$$

(1) Let  $b = 100i^{\frac{1}{2}}\theta_i$  and  $\theta_i = 1 + \frac{\sqrt{\log(\frac{1}{\epsilon_i^2} \frac{\sqrt{t^2+1}}{t^{1-i}})}}{10\sqrt{i}}$  for  $\epsilon_i^2 = \frac{\sqrt{t^2+1}}{t^{1-i}}\epsilon^2$  where  $i \geq 1$ , then we have

$$\begin{aligned} (\theta_i \cdot e^{-10^2\theta_i^2})^i &= \left( \left( 1 + \frac{\sqrt{\log(\frac{1}{\epsilon_i^2} \frac{\sqrt{t^2+1}}{t^{1-i}})}}{10\sqrt{i}} \right) \cdot e^{-10^2} e^{-2 \cdot 10^2 \frac{\sqrt{\log(\frac{1}{\epsilon_i^2} \frac{\sqrt{t^2+1}}{t^{2-i}})}}{10\sqrt{i}}} \cdot e^{-10^2 \frac{\log(\frac{1}{\epsilon_i^2} \frac{\sqrt{t^2+1}}{t^{2-i}})}{100i}} \right)^i \\ &= \epsilon_i^2 \frac{t^{1-i}}{\sqrt{t^2+1}} \cdot e^{-10^2 i} \cdot \left( 1 + \frac{\sqrt{\log(\frac{1}{\epsilon_i^2} \frac{\sqrt{t^2+1}}{t^{1-i}})}}{10\sqrt{i}} \right) e^{-2 \cdot 10^2 \frac{\sqrt{\log(\frac{1}{\epsilon_i^2} \frac{\sqrt{t^2+1}}{t^{1-i}})}}{10\sqrt{i}}} \\ &\leq \frac{\epsilon_i^2}{100000^i} \frac{t^{1-i}}{\sqrt{t^2+1}} \end{aligned} \quad (\text{D.34})$$

where the second step comes from that  $(1+s) \cdot e^{-2 \cdot 10^4 \cdot s} \leq 1$  for any  $s > 0$ . Combining the equation C.6, C.7 in [58] and (D.34), we can derive

$$\begin{aligned} &\sum_{i=1}^{\infty} |c'_i| \cdot \mathbb{E}_{x \sim \mathcal{N}(0,1)} [|h_i(z)| \cdot \mathbb{1}[|x| \geq b]] \\ &\leq \sum_{i=1}^{\infty} O(1) |c_i| \frac{\sqrt{t^2+1}}{t^{1-i}} \frac{i^4}{i!!} \cdot i^{\frac{i}{2}} \cdot 1200^i \cdot (\theta_i \cdot e^{-10^2\theta_i^2})^i \\ &\leq \frac{\epsilon}{8} \sqrt{t^2+1} \end{aligned} \quad (\text{D.35})$$

for any  $\epsilon > 0$  and  $t \leq O(1)$ .

(b) Similarly, following (D.34) and (D.35), we have

$$\sum_{i=1}^{\infty} |c'_i| \cdot |\mathbb{E}_{x \sim \mathcal{N}(0,1)} [|h_i(b)| \cdot \mathbb{1}[|x| \geq b]]| \leq \sum_{i=1}^{\infty} O(1) \frac{\sqrt{t^2+1}}{t^{1-i}} |c_i| \frac{i^4}{i!!} \cdot e^{-\frac{b^2}{2}} (3b)^i \leq \frac{\epsilon}{8} \sqrt{t^2+1}$$

(c) Similar to (D.33),

$$\begin{aligned} \sum_{i=1}^{\infty} |c'_i| \mathbb{E}_{z \in \mathcal{N}(0,1)} [|h_i(z)| \mathbb{1}[|z| \leq B_i]] &\leq O(1) \sum_{i=1}^{\infty} |c_i| \frac{i^4}{i!!} \sum_{j=0}^{\lfloor i/2 \rfloor} \frac{B_i^{i-2j} i^{2j}}{j!} t^{i-1} \sqrt{t^2+1} \\ &\leq \sum_{i=1}^{\infty} |c_i| (O(1)\theta_i)^i t^{i-1} \sqrt{t^2+1} \\ &\leq \mathcal{C}_\epsilon(\phi, t) \sqrt{t^2+1}, \end{aligned} \quad (\text{D.36})$$

where the step follows from Claim C.2 (c) in [58].

(d) Since we have

$$\left| \frac{d}{dx} h_i(x) \right| \leq \sum_{j=0}^{\lfloor i/2 \rfloor} |x|^{i-2j} i^{2j} \quad (\text{D.37})$$

by Definition A.5 in [58], we can derive

$$\sum_{i=1}^{\infty} |c'_i| \mathbb{E}_{z \in \mathcal{N}(0,1)} \left[ \left| \frac{d}{dz} h_i(z) \right| \mathbb{1}[|z| \leq B_i] \right] \leq \mathcal{C}_\epsilon(\phi, t) \sqrt{t^2 + 1} \quad (\text{D.38})$$

**Existence of A Good Pseudo Network** We hope to find some good pseudo network that can approximate the target network. In such a pseudo network, the activation  $\mathbb{1}_{x \geq 0}$  is replaced by  $\mathbb{1}_{x^{(0)} \geq 0}$  where  $x^{(0)}$  is the value at the random initialization. We can define a pseudo network without bias as

$$\begin{aligned} & g_r^{(0)}(\mathbf{q}, \mathbf{A}, \mathbf{W}, \mathbf{V}, \mathbf{B}) \\ &= \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{i \in [m_2]} c_{i,r} \mathbb{1}_{r_{n,i} + B_{2(n,i)} \geq 0} \sum_{j=1}^N a_{n,j} \sum_{l \in [m_1]} v_{i,l} \mathbb{1}_{\mathbf{a}_j \mathbf{X} \mathbf{w}_l + B_{1(j,l)} \geq 0} \mathbf{a}_j \mathbf{X} \mathbf{w}_l \end{aligned} \quad (\text{D.39})$$

Lemma D.2.4 shows the target function can be approximated by the pseudo network with some parameters. Lemma D.2.5 to D.2.8 provides how the existence of such a pseudo network is developed step by step.

**Lemma D.2.4.** *For every  $\epsilon \in (0, \frac{1}{K\|\mathbf{q}\|_1 p_1 p_2^2 \mathcal{C}_s(\Phi, p_2 \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty)) \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1}}),$  there exists*

$$M = \text{poly}(\mathcal{C}_\epsilon(\Phi, \sqrt{p_2} \mathcal{C}_\epsilon(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1}), 1/\epsilon)$$

$$C = \mathcal{C}_\epsilon(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1} \quad (\text{D.40})$$

$$C' = 10C\sqrt{p_2} \quad (\text{D.41})$$

$$C'' = \mathcal{C}_\epsilon(\Phi, C') \sqrt{\|\mathbf{A}\|_\infty^2 + 1} \quad (\text{D.42})$$

$$C_0 = \tilde{O}(p_1^2 p_2 K^2 C C'') \quad (\text{D.43})$$

such that with high probability, there exists  $\widehat{\mathbf{W}}, \widehat{\mathbf{V}}$  with  $m_1, m_2 \geq M,$

$$\|\widehat{\mathbf{W}}\|_{2,\infty} \leq \frac{C_0}{m_1}, \quad \|\widehat{\mathbf{V}}\|_{2,\infty} \leq \frac{\sqrt{m_1}}{m_2}$$

such that

$$\mathbb{E}_{(\mathbf{X}, y) \in \mathcal{D}} \left[ \sum_{r=1}^K |f_r^*(\mathbf{q}, \mathbf{A}, \mathbf{X}) - g_r^{(0)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \widehat{\mathbf{W}}, \widehat{\mathbf{V}})| \right] \leq \epsilon$$

$$\mathbb{E}_{(\mathbf{X}, y) \in \mathcal{D}} [|L(G^{(0)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \widehat{\mathbf{W}}, \widehat{\mathbf{V}}))|] \leq OPT + \epsilon$$

**Proof:**

For each  $\phi_{2,j}$ , we can construct  $h_{\phi,j} : \mathbb{R}^2 \rightarrow [-C, C]$  where  $C = \mathcal{C}_\epsilon(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1}$  using Lemma D.2.1 satisfying

$$\mathbb{E}[h_{\phi,j}(\mathbf{w}_{2,j}^* \top \mathbf{w}_i^{(0)}, B_{1(n,i)}^{(0)}) \mathbb{1}_{\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_i^{(0)} + B_{1(n,i)} \geq 0}] = \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}) \pm \epsilon \quad (\text{D.44})$$

for  $i \in [m_1]$ . Consider any arbitrary  $\mathbf{b} \in \mathbb{R}^{m_1}$  with  $v_i \in \{-1, 1\}$ . Define

$$\widehat{\mathbf{W}} = \frac{(C_0 C''/C)^{\frac{1}{2}}}{\epsilon_c^2 m_1} (v_i \sum_{j \in [p_2]} v_{2,j}^* h_{\phi,j}(\mathbf{w}_{2,j}^* \top \mathbf{w}_i^{(0)}, B_{1(i)}^{(0)}) \mathbf{e}_d)_{i \in [m_1]} \quad (\text{D.45})$$

$$\widehat{\mathbf{V}} = (C_0 C''/C)^{-\frac{1}{2}} \sum_{k \in [p_1]} \frac{c_k^*}{m_2} (\mathbf{v} h(\sqrt{m_2} \sum_{j \in [p_2]} v_{1,j}^* \alpha_{i,j}, B_{2(i)}^{(0)}) \sum_{r=1}^K c_{i,r})_{i \in [m_2]} \quad (\text{D.46})$$

Then,

$$\begin{aligned} & g_r^{(0)}(\mathbf{q}, \mathbf{A}, \widehat{\mathbf{W}}, \widehat{\mathbf{V}}, \mathbf{B}) \\ &= \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{i \in [m_1]} c_{i,r} \mathbb{1}_{\mathbf{r}_{n,i} + B_{2(n,i)} \geq 0} \sum_{i' \in [m_2]} \sum_{j=1}^N a_{n,j} \mathbb{1}_{\mathbf{a}_j \mathbf{X} \mathbf{w}_i^{(0)} + B_{1(j,i)} \geq 0} \mathbf{a}_j \mathbf{X} \widehat{\mathbf{W}}_i \widehat{\mathbf{V}}_{i,i'} \\ &= \sum_{k \in [p_1]} \frac{c_k^*}{m_2 \epsilon_c^2} \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{i \in [m_1]} c_{i,r}^2 \mathbb{1}_{\mathbf{r}_{n,i} + B_{2(n,i)} \geq 0} h(\sqrt{m_2} \sum_{j \in [p_2]} v_{1,j}^* \alpha_{i,j}, B_{2(i)}^{(0)}) \sum_{j=1}^N a_{n,j} \\ &\quad \sum_{l \in [p_2]} v_{2,l}^* \phi_{2,l}(\mathbf{a}_j \mathbf{X} \mathbf{w}_{2,l}^*) \\ &= \sum_{k \in [p_1]} \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n c_k^* \Phi \left( \sum_{j \in [p_2]} v_{1,j}^* \sum_{m=1}^N a_{m,n} \phi_{1,j}(\mathbf{a}_m \mathbf{X} \mathbf{w}_{1,j}^*) \right) \sum_{j=1}^N a_{n,j} \sum_{l \in [p_2]} v_{2,l}^* \phi_{2,l}(\mathbf{a}_j \mathbf{X} \mathbf{w}_{2,l}^*) \\ &\quad \pm O(p_1 p_2^2 \mathcal{C}_s(\Phi, p_2 \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty)) \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1} \epsilon) \\ &= \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{k \in [p_1]} c_k^* \Phi \left( \tilde{\mathbf{a}}_n \sum_{j \in [p_2]} v_{1,j}^* \phi_{1,j}(\mathbf{A} \mathbf{X} \mathbf{w}_{1,j}^*) \right) \tilde{\mathbf{a}}_n \sum_{l \in [p_2]} v_{2,l}^* \phi_{2,l}(\mathbf{A} \mathbf{X} \mathbf{w}_{2,l}^*) \\ &\quad \pm O(\|\mathbf{q}\|_1 p_1 p_2^2 \mathcal{C}_s(\Phi, p_2 \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty)) \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1} \epsilon) \end{aligned} \quad (\text{D.47})$$

where the first step comes from definition of  $g^{(0)}$ , the second step is derived from (D.45) and (D.46) and the second to last step is by Lemma D.2.8.

**Lemma D.2.5.** *For every smooth function  $\phi$ , every  $\mathbf{w}^* \in \mathbb{R}^d$  with  $\|\mathbf{w}^*\| = 1$ , for every  $\epsilon \in (0, \frac{1}{C_s(\phi, \|\mathbf{A}\|_\infty)\sqrt{\|\mathbf{A}\|_\infty^2 + 1}})$ , there exists real-valued functions  $\rho(\mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)})$ ,  $J(\tilde{\mathbf{a}}_n \mathbf{X}, \mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)})$ ,  $R(\tilde{\mathbf{a}}_n \mathbf{X}, \mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)})$  and  $\phi_\epsilon(\tilde{\mathbf{a}}_n \mathbf{X})$  such that for every  $\mathbf{X}$*

$$\begin{aligned} r_{n,1}(\mathbf{X}) = & \rho(\mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}) \sum_{j=1}^N a_{j,n} \phi_\epsilon(\mathbf{a}_j \mathbf{X} \mathbf{w}^*) + J(\mathbf{X}, \mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}) \\ & + R(\mathbf{X}, \mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}) \end{aligned} \quad (\text{D.48})$$

Moreover, letting  $C = C_\epsilon(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1}$  be the complexity of  $\phi$ , and if  $v_{1,i} \sim \mathcal{N}(0, \frac{1}{m_2})$  and  $w_{i,j}^{(0)}, \mathbf{B}_{1(n)}^{(0)} \sim \mathcal{N}(0, \frac{1}{m_1})$  are at random initialization, then we have

1. for every fixed  $\tilde{\mathbf{a}}_n \mathbf{X}$ ,  $\rho(\mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)})$  is independent of  $J(\tilde{\mathbf{a}}_n \mathbf{X}, \mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)})$ .
2.  $\rho(\mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}) \sim \mathcal{N}(0, \frac{1}{100C^2 m_2})$ .
3.  $|\phi_\epsilon(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_i^*) - \phi(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_i^*)| \leq \epsilon$
4. with high probability,  $|R(\mathbf{X}, \mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)})| \leq \tilde{O}(\frac{\|\mathbf{A}\|_\infty}{\sqrt{m_1 m_2}})$ ,  $|J(\mathbf{X}, \mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)})| \leq \tilde{O}(\frac{\|\mathbf{A}\|_\infty(1 + \|\mathbf{A}\|_\infty)}{\sqrt{m_2}})$  and  $\mathbb{E}[J(\mathbf{X}, \mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)})] = 0$ .

With high probability, we also have

$$\tilde{\rho}(v_1^{(0)}) \sim \mathcal{N}(0, \frac{\tau}{C^2 m_2})$$

$$\mathcal{W}_2(\rho|_{\mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}}, \tilde{\rho}) \leq \tilde{O}(\frac{1}{C \sqrt{m_2}})$$

**Proof:**

By Lemma D.2.1, we have

$$\mathbb{E}_{\mathbf{w}_i^{(0)} \sim \mathcal{N}(0, \frac{\mathbf{I}}{m_1}), b_{1(n,i)} \sim \mathcal{N}(0, \frac{1}{m_1})} [h(\sqrt{m_1} \mathbf{w}_i^{(0)\top} \mathbf{w}^*, b_{1(n,i)}) \mathbb{1}[\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_i^{(0)} + b_{1(n,i)} \geq 0]] = \frac{\phi_\epsilon(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_i^*)}{C}$$

with

$$|\phi_\epsilon(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}^*) - \phi(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}^*)| \leq \epsilon$$

and  $|h(\sqrt{m_1} \mathbf{w}_i^{(0)\top} \mathbf{w}^*, b_{1(n,i)})| \in [0, 1]$ . Note that here the  $h$  function is rescaled by  $1/C$ .

Then, applying Lemma A.4 of [58], we define

$$I_i = I(h(\sqrt{m_1} \mathbf{w}_i^{(0)\top} \mathbf{w}^*, B_{1(n,i)})) \subset [-2, 2]$$

$$\begin{aligned} S &= \{i \in [m_1] : \sqrt{m_2} v_{1,i}^{(0)} \in I_i\} \\ s_i &= s(h(\sqrt{m_1} \mathbf{w}_i^{(0)\top} \mathbf{w}^*, B_{1(n,i)}), \sqrt{m_2} v_{1,i}^{(0)}) \\ u_i &= \begin{cases} \frac{s_i}{\sqrt{|S|}}, & \text{if } i \in S \\ 0, & \text{if } s \notin S \end{cases} \end{aligned}$$

where  $u_i$ ,  $i \in [m_1]$  is independent of  $\mathbf{W}^{(0)}$ . We can write

$$\mathbf{W}^{(0)} = \alpha \mathbf{e}_d \mathbf{u}^\top + \boldsymbol{\beta},$$

where  $\alpha = \mathbf{u}^\top \mathbf{e}_d^\top \mathbf{W}^{(0)} \sim \mathcal{N}(0, 1/m_1)$  and  $\boldsymbol{\beta} \in \mathbb{R}^{d \times m_1}$  are two independent random variables given  $\mathbf{u}$ . We know  $\alpha$  is independent of  $\mathbf{u}$ . Since each  $i \in S$  with probability  $\tau$ , we know with high probability,

$$|S| = \tilde{\Theta}(\tau m_1) \quad (\text{D.49})$$

Since  $\alpha = \sum_{i \in S} u_i [\mathbf{e}_d^\top \mathbf{W}^{(0)}]_i$  and  $|u_i [\mathbf{e}_d^\top \mathbf{W}^{(0)}]_i| \leq \tilde{O}(1/\sqrt{m_1 |S|})$ , by (D.49) and the Wasserstein distance bound of central limit theorem we know there exists  $g \sim \mathcal{N}(0, \frac{1}{m_1})$  such that

$$\mathcal{W}_2(\alpha |_{\mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}}, g) \leq \tilde{O}\left(\frac{1}{\sqrt{\tau m_1}}\right)$$

Then,

$$\begin{aligned} r_{n,1}(\mathbf{X}) &= \sum_{j=1}^N a_{j,n} \sum_{i=1}^{m_1} v_{i,1}^{(0)} \sigma(\mathbf{a}_j \mathbf{X} \mathbf{w}_1^{(0)} + B_{1(n,i)}^{(0)}) \\ &= \sum_{j=1}^N a_{j,n} \sum_{i \notin S} v_{i,1}^{(0)} \sigma(\mathbf{a}_j \mathbf{X} \mathbf{w}_1^{(0)} + B_{1(n,i)}^{(0)}) + \sum_{j=1}^N a_{j,n} \sum_{i \in S} v_{i,1}^{(0)} \sigma(\mathbf{a}_j \mathbf{X} \mathbf{w}_1^{(0)} \\ &\quad + B_{1(n,i)}^{(0)}) \\ &= J_1 + \sum_{j=1}^N a_{j,n} \sum_{i \in S} v_{i,1}^{(0)} \sigma(\mathbf{a}_j \mathbf{X} \mathbf{w}_1^{(0)} + B_{1(n,i)}^{(0)}) \end{aligned} \quad (\text{D.50})$$

$$\begin{aligned}
& r_{n,1}(\mathbf{X}) - J_1 \\
&= \sum_{j=1}^N a_{j,n} \sum_{i \in S} v_{i,1}^{(0)} \mathbb{1}[\mathbf{a}_j \mathbf{X} \mathbf{w}_1^{(0)} + B_{1(n,i)}^{(0)}] \frac{s_i}{2\sqrt{|S|}} \alpha + \sum_{j=1}^N a_{j,n} \sum_{i \in S} v_{i,1}^{(0)} \mathbb{1}[\mathbf{a}_j \mathbf{X} \mathbf{w}_1^{(0)} \\
&\quad + B_{1(n,i)}^{(0)}] \cdot (\mathbf{a}_j \mathbf{X} \boldsymbol{\beta}_i + B_{1(n,i)}^{(0)}) \\
&= P_1 + P_2
\end{aligned} \tag{D.51}$$

Here, we know that since

$$\mathbb{E}[v_{i,1}^{(0)} \sigma(\mathbf{a}_j \mathbf{X} \mathbf{w}_1^{(0)} + B_{1(n,i)}^{(0)})] = \mathbb{E}[v_{i,1}^{(0)}] \cdot \mathbb{E}[\sigma(\mathbf{a}_j \mathbf{X} \mathbf{w}_1^{(0)} + B_{1(n,i)}^{(0)})] = 0 \tag{D.52}$$

Hence,

$$\mathbb{E}[J_1] = \mathbb{E}\left[\sum_{j=1}^N a_{j,n} \sum_{i \notin S} v_{i,1}^{(0)} \sigma(\mathbf{a}_j \mathbf{X} \mathbf{w}_1^{(0)} + B_{1(n,i)}^{(0)})\right] = 0 \tag{D.53}$$

Then we can derive

$$P_1 = \sum_{j=1}^N a_{j,n} \sum_{i \in S} \mathbb{1}[\mathbf{a}_j \mathbf{X} \mathbf{w}_1^{(0)} + B_{1(n,j)}^{(0)}] \frac{\alpha}{2\sqrt{|S|m_2}} h(\sqrt{m_1} \mathbf{w}_i^{(0)\top} \mathbf{w}^*, B_{1(n,i)}) + R_1 \tag{D.54}$$

where  $|R_1| \leq \tilde{O}(\sqrt{\frac{|S|}{m_1 m_2}})$ . We write  $P_3 = \frac{P_1 - R_1}{\alpha}$ . Then,

$$\begin{aligned}
& |P_3 - \frac{\sqrt{|S|}}{\sqrt{m_2} C} \sum_{j=1}^N a_{j,n} \phi_\epsilon(\mathbf{a}_j \mathbf{X} \mathbf{w}^*)| \leq \tilde{O}(\|\mathbf{A}\|_\infty \frac{1}{\sqrt{m_2}}) \\
& |\frac{C\sqrt{m_2}}{\sqrt{\tau m_1}} P_1 - \sum_{j=1}^N a_{j,n} \phi_\epsilon(\mathbf{a}_j \mathbf{X} \mathbf{w}^*)| \leq \tilde{O}(\|\mathbf{A}\|_\infty \frac{C}{\sqrt{\tau m_1}})
\end{aligned}$$

Define

$$\rho(\mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}) = \frac{\sqrt{\tau m_1}}{C\sqrt{m_2}} \alpha \sim \mathcal{N}(0, \frac{\tau}{C^2 m_2})$$

Then,

$$P_1 = \rho(\mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}) \cdot \sum_{j=1}^N a_{j,n} \phi_\epsilon(\mathbf{a}_j \mathbf{X} \mathbf{w}^*) + R_1 + R_2(\mathbf{X}, \mathbf{v}_1^{(0)}, \mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)})$$

where  $|R_2| \leq \tilde{O}(\frac{1}{\sqrt{m_1 m_2}})$ .

We can also define

$$\tilde{\rho}(v_1^{(0)}) = \frac{\sqrt{\tau m_1}}{C\sqrt{m_2}}g \sim \mathcal{N}(0, \frac{\tau}{C^2 m_2})$$

Therefore,

$$\mathcal{W}_2(\rho|_{\mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}}, \tilde{\rho}) \leq \tilde{O}\left(\frac{1}{C\sqrt{m_2}}\right)$$

Meanwhile,

$$\mathbf{a}_j \mathbf{X} \mathbf{w}_i^{(0)} = \alpha \frac{s_i}{\sqrt{|S|}} \mathbf{a}_j \mathbf{X} \mathbf{e}_d + \mathbf{a}_j \mathbf{X} \boldsymbol{\beta}_i + B_{1(n,i)}^{(0)} = \mathbf{a}_j \mathbf{X} \boldsymbol{\beta}_i + B_{1(n,i)}^{(0)} \pm \tilde{O}\left(\frac{1}{\sqrt{|S|m_1}}\right)$$

we have

$$\begin{aligned} P_2 &= \sum_{j=1}^N a_{j,n} \sum_{i \in S} v_{i,1}^{(0)} \mathbb{1}[\mathbf{a}_j \mathbf{X} \boldsymbol{\beta}_i + b_{1(n,j)}^{(0)}] (\mathbf{a}_j \mathbf{X} \boldsymbol{\beta}_i + b_{1(n,i)}^{(0)}) + R_3 = J_2 + R_3 \\ \mathbb{E}[J_2] &= 0 \end{aligned} \tag{D.55}$$

with  $|R_3| \leq \tilde{O}\left(\frac{\|\mathbf{A}\|_\infty}{\sqrt{m_1 m_2}}\right)$ .

Let  $J = J_1 + J_2$ ,  $R = R_1 + R_2 + R_3$ . Then, w.h.p.,  $\mathbb{E}[J] = 0$ ,  $|J| \leq \tilde{O}\left(\frac{\|\mathbf{A}\|_\infty(1+\|\mathbf{A}\|_\infty)}{\sqrt{m_2}}\right)$ ,  $|R| \leq \tilde{O}\left(\frac{\|\mathbf{A}\|_\infty}{\sqrt{m_1 m_2}}\right)$ .

**Lemma D.2.6.** For every  $\epsilon \in (0, \frac{1}{\mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1}})$ , there exists real-valued functions  $\phi_{1,j,\epsilon}(\cdot)$  such that

$$|\phi_{1,j,\epsilon}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{1,j}^*) - \phi_{1,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{1,j}^*)| \leq \epsilon$$

for  $j \in [p_2]$ . Denote by

$$C = \mathcal{C}_\epsilon(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1}, \quad C' = 10C\sqrt{p_2}, \quad \phi_{1,j,\epsilon}(\mathbf{a}_j \mathbf{X} \mathbf{w}_{1,i}^*) = \frac{1}{C'} \phi_{1,j,\epsilon}(\mathbf{a}_j \mathbf{X} \mathbf{w}_{1,i}^*)$$

For every  $i \in [m_2]$ , there exist independent Gaussians

$$\alpha_{i,j} \sim \mathcal{N}(0, \frac{1}{m_2}), \quad \beta_i(\mathbf{X}) \sim \mathcal{N}(0, \frac{1}{m_2}),$$

satisfying

$$\mathcal{W}_2(r_{n,i}(\mathbf{X}), \sum_{j \in [p_2]} \alpha_{i,j} \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\mathbf{a}_m \mathbf{X} \mathbf{w}_{1,i}^*) + C_i \beta_i(\mathbf{X})) \leq \tilde{O}\left(\frac{p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}} \sqrt{m_2}}\right)$$

**Proof:**

Define  $p_2S$  many chunks of the first layer with each chunk corresponding to a set  $S_{j,l}$ , where  $|S_{j,l}| = m_1/(p_2S)$  for  $j \in [p_2]$  and  $l \in [S]$ , such that

$$\mathcal{S}_{j,l} = \{(j-1)\frac{m_1}{p_2} + (l-1)\frac{m_1}{p_2}S + k | k \in [\frac{m_1}{p_2S}]\} \subset [m_1]$$

By Lemma D.2.5, we have

$$\begin{aligned} r_{n,i}(\mathbf{X}) &= \sum_{j \in [p_2], l \in [S]} \rho(\mathbf{v}_i^{(0)}[j, l], \mathbf{W}^{(0)}[j, l], \mathbf{B}_{1(n)}^{(0)}[j, l]) \sum_{m=1}^N a_{m,n} \phi_\epsilon(\mathbf{a}_m \mathbf{X} \mathbf{w}_{1,j}^*) \\ &\quad + \sum_{j \in [p_2], l \in [S]} J_j(\mathbf{X}, \mathbf{v}_i^{(0)}[j, l], \mathbf{W}^{(0)}[j, l], \mathbf{B}_{1(n)}^{(0)}[j, l]) + R_j(\mathbf{X}, \mathbf{v}_i^{(0)}[j, l], \\ &\quad \mathbf{W}^{(0)}[j, l], \mathbf{B}_{1(n)}^{(0)}[j, l]), \end{aligned} \quad (\text{D.56})$$

where  $\rho(\mathbf{v}_i^{(0)}[j, l], \mathbf{W}^{(0)}[j, l], \mathbf{B}_{1(n)}^{(0)}[j, l]) \sim \mathcal{N}(0, \frac{1}{100C^2m_2p_2S})$ . Then  $\rho_j = \sum_{l \in [S]} \rho_{j,l} \sim \mathcal{N}(0, \frac{1}{C'^2m_2})$  for  $C' = 10C\sqrt{p_2}$ . Define

$$J_j^S(\mathbf{X}) = \sum_{l \in [S]} J_j(\mathbf{X}, \mathbf{v}_i^{(0)}[j, l], \mathbf{W}^{(0)}[j, l], \mathbf{B}_{1(n)}^{(0)}[j, l])$$

$$R_j^S(\mathbf{X}) = \sum_{l \in [S]} R_j(\mathbf{X}, \mathbf{v}_i^{(0)}[j, l], \mathbf{W}^{(0)}[j, l], \mathbf{B}_{1(n)}^{(0)}[j, l])$$

Then there exists Gaussian random variables  $\beta_j(\mathbf{X})$  and  $\beta'(\mathbf{X}) = \sum_{i \in [p_2]} \beta_i(\mathbf{X})$  such that

$$\mathcal{W}_2(J_j^S(\mathbf{X}), \beta_j(\mathbf{X})) \leq \frac{\|\mathbf{A}\|_\infty(1 + \|\mathbf{A}\|_\infty)}{\sqrt{m_2 p S}}$$

$$\mathcal{W}_2(r_{n,i}(\mathbf{X}), \sum_{j \in [p_2]} \rho_j \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\mathbf{a}_m \mathbf{X} \mathbf{w}_{1,j}^*) + \beta'(\mathbf{X})) \leq \tilde{O}\left(\frac{Sp_2}{\sqrt{m_1 m_2}} + \frac{\sqrt{p_2} \|\mathbf{A}\|_\infty(1 + \|\mathbf{A}\|_\infty)}{m_2 S}\right)$$

We know there exists a positive constant  $C_i$  such that  $\beta'/C_i \sim \mathcal{N}(0, \frac{1}{m_2})$ . Let  $\alpha_{i,j} = C' \rho_j$ ,  $\beta'_i = \beta'/C_i$ . Notice that  $\mathbb{E}[\sum_{l \in [S], i \in [p_2]} [J_j^2(\mathbf{X}, \mathbf{v}_i^{(0)}[j, l], \mathbf{W}^{(0)}[j, l], b_1^{(0)}[j, l])] = \tilde{O}(\|\mathbf{A}\|_\infty^2(1 + \|\mathbf{A}\|_\infty)^2/m_2)$ . Hence, we have

$$C_i \leq \tilde{O}(\|\mathbf{A}\|_\infty(1 + \|\mathbf{A}\|_\infty))$$

Let  $S = (m_1/p_2)^{\frac{1}{3}}$ , we can obtain

$$\mathcal{W}_2(r_{n,i}(\mathbf{X}), \sum_{j \in [p_2]} \alpha_{i,j} \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\mathbf{a}_m \mathbf{X} \mathbf{w}_{1,i}^*) + C_i \beta_i(\mathbf{X})) \leq \tilde{O}\left(\frac{p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}} \sqrt{m_2}}\right)$$

**Lemma D.2.7.** *There exists function  $h : \mathbb{R}^2 \rightarrow [-C'', C'']$  for  $C'' = \mathcal{C}_\epsilon(\Phi, C') \sqrt{\|\mathbf{A}\|_\infty^2 + 1}$  such that*

$$\begin{aligned} & \mathbb{E}[\mathbb{1}_{r_{n,i}(\mathbf{X}) + b_{2(n,i)}^{(0)} \geq 0} h(\sqrt{m_2} \sum_{j \in [p_2]} v_{1,j}^* \alpha_{i,j}, b_{2(n,i)}^{(0)}) (\sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*))] \\ &= \Phi(\sum_{j \in [p_2]} v_{1,j}^* \sum_{m=1}^N a_{m,n} \phi_{1,j}) \sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*) \pm \tilde{O}(p_2^2 \mathcal{C}_s(\Phi, p_2 \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty))) \\ & \quad \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1} \end{aligned} \quad (\text{D.57})$$

**Proof:**

Choose  $\mathbf{w} = (\alpha_{i,1}, \dots, \alpha_{i,p_2}, \beta_i)$ ,  $\mathbf{x} = (\sum_{m=1}^N a_{m,n} \phi_{1,1,\epsilon}, \dots, \sum_{m=1}^N a_{m,n} \phi_{1,p_2,\epsilon}, C_i)$  and  $\mathbf{w}^* = (v_{1,1}^*, \dots, v_{1,p_2}^*, 0)$ . Then,  $\|\mathbf{x}\| \leq O(\|\mathbf{A}\|_\infty^2 + \|\mathbf{A}\|_\infty)$ . By Lemma D.2.1, there exists  $h : \mathbb{R}^2 \rightarrow [-C'', C'']$  for  $C'' = \mathcal{C}_s(\Phi, C') \sqrt{\|\mathbf{A}\|_\infty^2 + 1}$  such that

$$\begin{aligned} & \mathbb{E}[\mathbb{1}_{\mathbf{w}^\top \mathbf{X} + b_{2(n,i)}^{(0)} \geq 0} h(\sqrt{m_2} \mathbf{w}^\top \mathbf{w}^*, b_{2(n,i)}^{(0)}) (\sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*))] \\ &= \mathbb{E}_{\alpha_i, \beta_i} [\mathbb{1}_{\sum_{j \in [p_2]} \alpha_{i,j} \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{1,i}^*) + C_i \beta_i' + b_{2(n,i)}^{(0)} \geq 0} h(\sqrt{m_2} \mathbf{w}^\top \mathbf{w}^*, b_{2(n,i)}^{(0)}) \\ & \quad (\sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*))] \\ &= \Phi(C' \sum_{j \in [p_2]} v_{1,j}^* \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}) \sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*) \pm \epsilon C''' \end{aligned} \quad (\text{D.58})$$

where

$$C''' = \sup \left| \sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*) \right| \leq p_2 \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1}$$

By Lemma D.2.6, we know

$$\mathcal{W}_2(r_{n,i}(\mathbf{X}), \sum_{j \in [p_2]} \alpha_{i,j} \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\mathbf{a}_m \mathbf{X} \mathbf{w}_{1,i}^*) + C_i \beta_i(\mathbf{X})) \leq \tilde{O}\left(\frac{p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}} \sqrt{m_2}}\right)$$

Denote  $\mathcal{H} = \{i \in [m_1] : |\sum_{j \in [p_2]} \alpha_{i,j} \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{1,i}^*) + C_i \beta'_i| \geq \tilde{O}(\frac{2p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}} \sqrt{m_2}})\}$ . Then, for every  $i \in [\mathcal{H}]$ , we have that

$$\mathbb{1}_{r_{n,i}(\mathbf{X})+b_{2(n,i)}^{(0)} \geq 0} = \mathbb{1}_{\sum_{j \in [p_2]} \alpha_{i,j} \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{1,i}^*) + C_i \beta'_i + b_{2(n,i)}^{(0)} \geq 0} \quad (\text{D.59})$$

$$\begin{aligned} & \Pr \left( \left| \sum_{j \in [p_2]} \alpha_{i,j} \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{1,i}^*) + C_i \beta'_i \right| \leq \tilde{O}\left(\frac{2p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}} \sqrt{m_2}}\right) \right) \\ & \leq \tilde{O}\left(\frac{2p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}} \sqrt{m_2}}\right) \cdot \sqrt{m_2} = \tilde{O}\left(\frac{2p_2^{\frac{2}{3}}}{m_1^{\frac{1}{6}}}\right), \end{aligned} \quad (\text{D.60})$$

which implies with probability at least  $1 - 2p_2^{2/3}/m_1^{1/6}$ , (D.59) holds. Therefore,

$$\begin{aligned} & \mathbb{E}[\mathbb{1}_{r_{n,i}(\mathbf{X})+b_{2(n,i)}^{(0)} \geq 0} h(\sqrt{m_2} \sum_{j \in [p_2]} v_{1,j}^* \alpha_{i,j}, b_{2(n,i)}^{(0)}) (\sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*))] \\ & = \mathbb{E}[\mathbb{1}_{\sum_{j \in [p_2]} \alpha_{i,j} \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{1,i}^*) + C_i \beta'_i + b_{2(n,i)}^{(0)} \geq 0} h(\sqrt{m_2} \sum_{j \in [p_2]} v_{1,j}^* \alpha_{i,j}, b_{2(n,i)}^{(0)}) \\ & \quad \cdot (\sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*))] \pm \mathbb{E}[\mathbb{1}_{r_{n,i}(\mathbf{X})+b_{2(n,i)}^{(0)} \geq 0} \\ & \neq \mathbb{1}_{\sum_{j \in [p_2]} \alpha_{i,j} \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{1,i}^*) + C_i \beta'_i + b_{2(n,i)}^{(0)} \geq 0} \cdot O(C'''C'')] \\ & = \mathbb{E}[\mathbb{1}_{\sum_{j \in [p_2]} \alpha_{i,j} \sum_{m=1}^N a_{m,n} \phi_{1,j,\epsilon}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{1,i}^*) + C_i \beta'_i + b_{2(n,i)}^{(0)} \geq 0} h(\sqrt{m_2} \sum_{j \in [p_2]} v_{1,j}^* \alpha_{i,j}, b_{2(n,i)}^{(0)}) \\ & \quad \cdot (\sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*))] \pm \tilde{O}\left(\frac{2p_2^{2/3}}{m_1^{1/6}} C'''C''\right) \\ & = \Phi(\sum_{j \in [p_2]} v_{1,j}^* \sum_{m=1}^N a_{m,n} \phi_{1,j}) \sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\mathbf{x}) \pm \tilde{O}(p_2^2 \mathcal{C}_s(\Phi, p_2 \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty)) \\ & \quad \cdot \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1} \cdot \epsilon), \end{aligned} \quad (\text{D.61})$$

where the first step is by Lemma D.2.6, the second step is by (D.59) and (D.60) and the last step comes from (D.58) and  $m_1 \geq M$ .

**Lemma D.2.8.**

$$\begin{aligned}
& \frac{1}{m_2} \mathbb{E} \left[ \sum_{i=1}^{m_2} \frac{c_{i,l}^2}{\epsilon_c^2} \mathbb{1}_{r_{n,i}(\mathbf{X}) + b_{2(n,i)}^{(0)} \geq 0} h(\sqrt{m_2} \sum_{j \in [p_2]} v_{1,j}^* \alpha_{i,j}, b_{2(n,i)}^{(0)}) (\sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*)) \right] \\
& = \Phi \left( \sum_{j \in [p_2]} v_{1,j}^* \sum_{m=1}^N a_{m,n} \mathbf{a}_m \mathbf{X} \boldsymbol{\delta} \phi_{1,j} \right) \sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*) \\
& \pm \tilde{O}(p_2^2 \mathcal{C}_s(\Phi, p_2 \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty)) \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1} \cdot \epsilon)
\end{aligned} \tag{D.62}$$

**Proof:**

Recall  $\tilde{\rho}(\mathbf{v}_1^{(0)}) \sim \mathcal{N}(0, \frac{\tau}{C^2 m_2})$ . Define  $\tilde{\rho}_{j,l} = \tilde{\rho}(\mathbf{v}_1^{(0)}[j, l])$ . Therefore,

$$\mathcal{W}_2(\rho_{j,l} |_{\mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}}, \tilde{\rho}_{j,l}) \leq \tilde{O}\left(\frac{1}{C' \sqrt{m_2} S}\right) \tag{D.63}$$

$$\mathcal{W}_2(\rho_j |_{\mathbf{W}^{(0)}, \mathbf{B}_{1(n)}^{(0)}}, \tilde{\rho}_j) \leq \tilde{O}\left(\frac{1}{C' \sqrt{m_2}}\right) \tag{D.64}$$

where  $\tilde{\rho}_j = \sum_{l \in [S]} \rho_{j,l}$ . We then define  $\tilde{\alpha}_{i,j} = C' \tilde{\rho}_j$

Next modify  $r_{n,i}(\mathbf{X})$ . Define

$$\tilde{r}_{n,i}(\mathbf{X}) = \frac{\sum_{m=1}^N a_{m,n} \sum_{j \in [m_1]} v_{j,i}^{(0)} \sigma(\mathbf{a}_m \mathbf{X} \mathbf{w}_i^{(0)} + b_{1(n,i)}^{(0)})}{\|\mathbf{u}\|} \mathbb{E}[\|\mathbf{u}\|]$$

where  $u = (\sigma(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_j^{(0)} + b_{1(n,j)}^{(0)}))_{j \in [m_1]}$ . By definition, we know

$$\tilde{r}_{n,i} \sim \mathcal{N}(0, \frac{\|\tilde{\mathbf{a}}_n\|_\infty^2 \mathbb{E}[\|\mathbf{u}\|^2]}{m_2})$$

Then we have

$$\mathcal{W}_2(r_{n,i}(\mathbf{X}), \tilde{r}_{n,i}(\mathbf{X})) \leq \tilde{O}\left(\frac{\|\mathbf{A}\|_\infty \sqrt{\|\mathbf{A}\|_\infty^2 + 1}}{\sqrt{m_2}}\right) \tag{D.65}$$

Combining (D.63), (D.64), (D.65) and Lemma D.2.7, we have

$$\begin{aligned}
& \frac{1}{m_2} \mathbb{E} \left[ \sum_{i=1}^{m_2} \frac{c_{i,l}^2}{\epsilon_c^2} \mathbb{1}_{r_{n,i}(\mathbf{X}) + b_{2(n,i)}^{(0)} \geq 0} h(\sqrt{m_2} \sum_{j \in [p_2]} v_{1,j}^* \alpha_{i,j}, b_{2(n,i)}^{(0)}) (\sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*)) \right] \\
& = \Phi \left( \sum_{j \in [p_2]} v_{1,j}^* \sum_{m=1}^N a_{m,n} \phi_{1,j} \right) \sum_{j \in [p_2]} v_{2,j}^* \phi_{2,j}(\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_{2,j}^*) \pm \tilde{O}(p_2^2 \mathcal{C}_s(\Phi, p_2 \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty))) \\
& \quad \cdot \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty) (\sqrt{\|\mathbf{A}\|_\infty^2 + 1} \cdot \epsilon)
\end{aligned} \tag{D.66}$$

**Coupling** This section illustrates the coupling between the real and pseudo networks. We first define diagonal matrices  $\mathbf{D}_{n,w}$ ,  $\mathbf{D}_{n,w} + \mathbf{D}_{n,w}''$ ,  $\mathbf{D}_{n,w} + \mathbf{D}_{n,w}'$  for node  $n$  as the sign of Relu's in the first layer at weights  $\mathbf{W}^{(0)}$ ,  $\mathbf{W}^{(0)} + \mathbf{W}^\rho$  and  $\mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}'$ , respectively. We also define diagonal matrices  $\mathbf{D}_{n,v}$ ,  $\mathbf{D}_{n,v} + \mathbf{D}_{n,v}''$ ,  $\mathbf{D}_{n,v} + \mathbf{D}_{n,v}'$  for node  $n$  as the sign of Relu's in the second layer at weights  $\{\mathbf{W}^{(0)}, \mathbf{V}^{(0)}\}$ ,  $\{\mathbf{W}^{(0)} + \mathbf{W}^\rho, \mathbf{V}^{(0)} + \mathbf{V}^\rho\}$  and  $\{\mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}', \mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}'\}$ , respectively. For every  $l \in [K]$ , we then introduce the pseudo network and its semi-bias, bias-free version as

$$g_l(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}, \mathbf{V}) = \mathbf{q}^\top \mathbf{A} ((\mathbf{A}(\mathbf{A} \mathbf{X} \mathbf{W} + \mathbf{B}_1) \odot (\mathbf{D}_w + \mathbf{D}'_w) \mathbf{V} + \mathbf{B}_2) \odot (\mathbf{D}_v + \mathbf{D}'_v)) \mathbf{c}_l \tag{D.67}$$

$$g_l^{(b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}, \mathbf{V}) = \mathbf{q}^\top \mathbf{A} ((\mathbf{A}(\mathbf{A} \mathbf{X} \mathbf{W} + \mathbf{B}_1) \odot (\mathbf{D}_w + \mathbf{D}'_w) \mathbf{V}) \odot (\mathbf{D}_v + \mathbf{D}'_v)) \mathbf{c}_l \tag{D.68}$$

$$g_l^{(b,b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}, \mathbf{V}) = \mathbf{q}^\top \mathbf{A} ((\mathbf{A}(\mathbf{A} \mathbf{X} \mathbf{W}) \odot (\mathbf{D}_w + \mathbf{D}'_w) \mathbf{V}) \odot (\mathbf{D}_v + \mathbf{D}'_v)) \mathbf{c}_l \tag{D.69}$$

Lemma D.2.9 gives the final result of coupling with added Drop-out noise. Lemma D.2.10 states the sparse sign change in Relu and the function value changes of pseudo network by some update. To be more specific, Lemma D.2.11 shows that the sign pattern can be viewed as fixed for the smoothed objective when a small update is introduced to the current weights. Lemma D.2.12 proves the bias-free pseudo network can also approximate the target function.

**Lemma D.2.9.** *Let  $F_A = (f_1, f_2, \dots, f_K)$ . With high probability, we have for any  $\|\mathbf{W}'\|_{2,4} \leq \tau_w$ ,  $\|\mathbf{V}'\|_F \leq \tau_v$ , such that*

$$\begin{aligned}
& f_l(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}' \Sigma, \mathbf{V}^{(0)} + \Sigma \mathbf{V}') \\
& = \mathbf{q}^\top \mathbf{A} ((\mathbf{A}(\mathbf{A} \mathbf{X} \mathbf{W}^{(0)} + \mathbf{B}_1^{(0)}) \odot \mathbf{D}_{w,x}^{(0)} \mathbf{V}^{(0)} + \mathbf{B}_2^{(0)}) \odot \mathbf{D}_{v,x}^{(0)}) \mathbf{c} + \mathbf{q}^\top \mathbf{A} (\mathbf{A}((\mathbf{A} \mathbf{X} \mathbf{W}')) \\
& \quad \odot \mathbf{D}_{w,x}^{(0)} \mathbf{V}') \odot \mathbf{D}_{v,x}^{(0)}) \mathbf{c}_l \pm \tilde{O}(\tau_v \frac{\sqrt{m_2}}{\sqrt{m_1}} + m_1^{\frac{9}{5}} \tau_w^{\frac{16}{5}} \sqrt{m_2} + \tau_w^{\frac{8}{5}} m_1^{\frac{9}{10}}) \cdot \|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty^2,
\end{aligned} \tag{D.70}$$

where we use  $\mathbf{D}_{\mathbf{w}, \mathbf{x}}^{(0)}$  and  $\mathbf{D}_{\mathbf{v}, \mathbf{x}}^{(0)}$  to denote the sign matrices at random initialization  $\mathbf{W}^{(0)}$ ,  $\mathbf{V}^{(0)}$  and we let  $\mathbf{D}_{\mathbf{w}, \mathbf{x}}^{(0)} + \mathbf{D}'_{\mathbf{w}, \mathbf{x}}$ ,  $\mathbf{D}_{\mathbf{v}, \mathbf{x}}^{(0)} + \mathbf{D}'_{\mathbf{v}, \mathbf{x}}$  be the sign matrices at  $\mathbf{W} + \mathbf{W}'\Sigma$ ,  $\mathbf{V} + \Sigma\mathbf{V}'$ .

**Proof:**

Since  $\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_i^{(0)} + B_{1(n,i)}^{(0)} = \tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{w}}_i^{(0)}$  where  $\tilde{\mathbf{w}}_i^{(0)} = (\mathbf{w}_i^{(0)}, B_{1(n,i)}^{(0)}) \in \mathbb{R}^{d+1}$  and  $\tilde{\mathbf{X}} = (\mathbf{X}, \mathbf{1}) \in \mathbb{R}^{N \times (d+1)}$ , we can ignore the bias term for simplicity. Define

$$\mathbf{Z} = \mathbf{A}(\mathbf{AXW}^{(0)}) \odot \mathbf{D}_{\mathbf{w}, \mathbf{x}}^{(0)}$$

$$\mathbf{Z}_1 = \mathbf{A}(\mathbf{AXW}'\Sigma) \odot \mathbf{D}_{\mathbf{w}, \mathbf{x}}^{(0)}$$

$$\mathbf{Z}_2 = \mathbf{A}(\mathbf{AX}(\mathbf{W}^{(0)} + \mathbf{W}')\Sigma) \odot \mathbf{D}'_{\mathbf{w}, \mathbf{x}}$$

Then by Fact C.9 in [58] we have

$$\begin{aligned} \|\mathbf{Z}_n \Sigma \mathbf{V}'\|_2^2 &\leq \sum_{i=1}^{m_2} (\mathbf{Z}_n \Sigma \mathbf{V}'_i)^2 \leq \sum_{i=1}^{m_2} \tilde{O}(\|\mathbf{Z}_n\|_\infty^2 \cdot \|\mathbf{V}'_i\|_2^2) \\ &\leq \tilde{O}(\|\mathbf{A}\|_\infty^2 m_1^{-1} \tau_v^2) \end{aligned} \quad (\text{D.71})$$

Therefore, we have  $\|\mathbf{Z}_n \Sigma \mathbf{V}'\|_2 \leq \tilde{O}(\|\mathbf{A}\|_\infty m_1^{-\frac{1}{2}} \tau_v)$ .

Let  $s$  be the total number of sign changes in the first layer caused by adding  $\mathbf{W}'$ . Note that the total number of coordinated  $i$  such that  $|\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_i^{(0)}| \leq s'' = \frac{2\tau_w}{s^{\frac{1}{4}}}$  is at most  $s'' m_1^{\frac{3}{2}}$  with high probability. Since  $\|\mathbf{W}'\|_{2,4} \leq \tau_w$ , we must have  $s \leq \tilde{O}(s'' m_1^{\frac{3}{2}}) = \tilde{O}(\frac{\tau_w}{s^{\frac{1}{4}}} m_1^{\frac{3}{2}})$ . Therefore,  $\|\mathbf{Z}_{2,n}\|_0 \leq s = \tilde{O}(\tau_w^{\frac{4}{5}} m_1^{\frac{6}{5}})$ . Then,

$$\begin{aligned} \|\mathbf{Z}_{2,n}\|_2 &= \|(\mathbf{A}(\mathbf{AX}(\mathbf{W}^{(0)} + \mathbf{W}'\Sigma)) \odot \mathbf{D}'_{\mathbf{w}, \mathbf{x}})_n\| \\ &\leq \left( s \cdot \sum_{(\mathbf{D}'_{\mathbf{w}, \mathbf{x}})_n \neq 0} (\mathbf{A} \mathbf{AX} \mathbf{W}^{(0)} + \mathbf{A} \mathbf{AX} \mathbf{W}' \Sigma)_{n,i}^4 \right)^{\frac{1}{4}} \\ &\leq \left( s \cdot \sum_{(\mathbf{D}'_{\mathbf{w}, \mathbf{x}})_n \neq 0} (\mathbf{A} \mathbf{AX} \mathbf{W}' \Sigma)_{n,i}^4 \right)^{\frac{1}{4}} \\ &\leq s^{\frac{1}{4}} \|\mathbf{A}\|_\infty \tau_w \\ &\leq \tilde{O}(\tau_w^{\frac{6}{5}} m_1^{\frac{3}{10}} \|\mathbf{A}\|_\infty) \end{aligned} \quad (\text{D.72})$$

Then we have

$$\|\mathbf{Z}_{2,n} \Sigma \mathbf{V}'\|_2 \leq \tilde{O}(\tau_v \tau_w^{\frac{6}{5}} m_1^{\frac{3}{10}} \|\mathbf{A}\|_\infty)$$

With high probability, we have

$$\begin{aligned}
& \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{i=1}^{m_2} c_{i,l} (\sigma(r_{n,i} + r'_{n,i}) - \sigma(r_{n,i})) \leq \tilde{O}(\|\mathbf{q}\| \sqrt{m_2}) \|r'_{n,i}\| \\
f_l(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}'\Sigma, \mathbf{V}^{(0)} + \Sigma\mathbf{V}') \\
&= \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{i=1}^{m_2} c_{i,l} \sigma((\mathbf{Z} + \mathbf{Z}_1 + \mathbf{Z}_2)_n^\top (\mathbf{V}_i + (\Sigma)_i \mathbf{V}')) \\
&= \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{i=1}^{m_2} c_{i,l} \sigma((\mathbf{Z}_n + \mathbf{Z}_{1,n} + \mathbf{Z}_{2,n})^\top \mathbf{V}_i + \mathbf{Z}_{1,n}^\top (\Sigma \mathbf{V}')_i) \pm \tilde{O}(\|\mathbf{q}\| \|\mathbf{A}\|_\infty \frac{\sqrt{m_2}}{\sqrt{m_1}} \tau_v \\
&\quad + \sqrt{m_2} \|\mathbf{q}\| \|\mathbf{A}\|_\infty \tau_w^{\frac{6}{5}} m_1^{\frac{3}{10}})
\end{aligned} \tag{D.73}$$

We consider the difference between

$$A_1 = \mathbf{q}^\top \mathbf{A} (((\mathbf{Z} + \mathbf{Z}_1 + \mathbf{Z}_2) \mathbf{V}^{(0)} + \mathbf{Z}_1 \Sigma \mathbf{V}') \odot (\mathbf{D}_{v,x}^{(0)} + \mathbf{D}_{v,x}'')) \mathbf{c}_l$$

$$A_2 = \mathbf{q}^\top \mathbf{A} (((\mathbf{Z} + \mathbf{Z}_1 + \mathbf{Z}_2) \mathbf{V}^{(0)} + \mathbf{Z}_1 \Sigma \mathbf{V}') \odot \mathbf{D}_{v,x}^{(0)}) \mathbf{c}_l$$

where  $\mathbf{D}_{v,x}''$  is the diagonal sign change matrix from  $\mathbf{Z}\mathbf{V}^{(0)}$  to  $(\mathbf{Z} + \mathbf{Z}_1 + \mathbf{Z}_2)\mathbf{V}^{(0)} + \mathbf{Z}_1\Sigma\mathbf{V}'$ .

The difference includes three terms.

$$\|\mathbf{Z}_{1,n} \mathbf{V}^{(0)}\|_\infty \leq \tilde{O}(\|\mathbf{A}\|_\infty m_1^{\frac{1}{4}} \tau_w m_2^{-\frac{1}{2}}) \tag{D.74}$$

$$\|\mathbf{Z}_{2,n} \mathbf{V}^{(0)}\|_\infty \leq \tilde{O}(\|\mathbf{Z}_{2,n}\| m_2^{-\frac{1}{2}} \sqrt{s}) \leq \tilde{O}(\|\mathbf{A}\|_\infty m_1^{\frac{9}{10}} \tau_w^{\frac{8}{5}} m_2^{-\frac{1}{2}}) \tag{D.75}$$

$$\|\mathbf{Z}_{1,n} \Sigma \mathbf{V}'\| \leq \tilde{O}(\tau_v \|\mathbf{A}\|_\infty \tau_w m_1^{\frac{1}{4}}) \tag{D.76}$$

where (D.74) is by Fact C.9 in [58]. Then we have

$$|A_1 - A_2| \leq \|\mathbf{q}^\top \mathbf{A}\|_1 \cdot \tilde{O}(m_2^{\frac{3}{2}} \|\mathbf{A}\|_\infty^2 (m_1^{\frac{1}{4}} \tau_w m_2^{-\frac{1}{2}} + m_1^{\frac{9}{10}} \tau_w^{\frac{8}{5}} m_2^{-\frac{1}{2}})^2 + m_2^{\frac{1}{2}} \tau_v^{\frac{4}{3}} \|\mathbf{A}\|_\infty^{\frac{4}{3}} \tau_w^{\frac{4}{3}} m_1^{\frac{1}{3}})$$

From  $A_2$  to our goal

$$A_3 = \mathbf{q}^\top \mathbf{A} (\mathbf{Z}\mathbf{V}^{(0)} \odot \mathbf{D}_{v,x}^{(0)}) \mathbf{c}_l + \mathbf{q}^\top \mathbf{A} (\mathbf{A} (\mathbf{A}\mathbf{X}\mathbf{W}' \odot \mathbf{D}_{w,x}^{(0)} \mathbf{V}') \odot \mathbf{D}_{v,x}^{(0)}) \mathbf{c}_l$$

There are two more terms.

$$|\mathbf{q}^\top \mathbf{A}(\mathbf{Z}_2 \mathbf{V}^{(0)} \odot \mathbf{D}_{\mathbf{v}, \mathbf{x}}^{(0)}) \mathbf{c}_l| \leq \tilde{O}(\|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{Z}_{2,n}\| \sqrt{s}) \leq \tilde{O}(\|\mathbf{q}\|_1 \|\mathbf{A}\|_\infty \tau_w^{\frac{8}{5}} m_1^{\frac{9}{10}})$$

$$|\mathbf{q}^\top \mathbf{A}(\mathbf{Z}_1 \mathbf{V}^{(0)} \odot \mathbf{D}_{\mathbf{v}, \mathbf{x}}^{(0)}) \mathbf{c}_l| \leq \tilde{O}(\|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty \tau_w m_1^{\frac{1}{4}})$$

Therefore, we have

$$|A_2 - A_3| \leq \tilde{O}(\|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty \tau_w^{\frac{8}{5}} m_1^{\frac{9}{10}} + \|\mathbf{q}\|_1 \|\mathbf{A}\|_\infty \tau_w m_1^{\frac{1}{4}} + \|\mathbf{q}\|_1 \tau_v \|\mathbf{A}\|_\infty \tau_w m_1^{\frac{1}{4}})$$

Finally, we have

$$\begin{aligned} & f_l(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}' \Sigma, \mathbf{V}^{(0)} + \Sigma \mathbf{V}') \\ &= \mathbf{q}^\top \mathbf{A}(\mathbf{Z} \mathbf{V}^{(0)} \odot \mathbf{D}_{\mathbf{v}, \mathbf{x}}^{(0)}) \mathbf{c}_l + \mathbf{q}^\top \mathbf{A}(\mathbf{A}(\mathbf{A} \mathbf{X} \mathbf{W}' \odot \mathbf{D}_{\mathbf{w}, \mathbf{x}}^{(0)} \mathbf{V}') \odot \mathbf{D}_{\mathbf{v}, \mathbf{x}}^{(0)}) \mathbf{c}_l \quad (\text{D.77}) \\ &\pm \tilde{O}(\tau_v \frac{\sqrt{m_2}}{\sqrt{m_1}} + m_1^{\frac{9}{5}} \tau_w^{\frac{16}{5}} \sqrt{m_2} + \tau_w^{\frac{8}{5}} m_1^{\frac{9}{10}}) \cdot \|\mathbf{q}\|_1 \|\mathbf{A}\|_\infty \end{aligned}$$

**Lemma D.2.10.** Suppose  $\tau_v \in (0, 1]$ ,  $\tau_w \in [\frac{1}{m_1^{\frac{3}{2}}}, \frac{1}{m_1^{\frac{1}{2}}}]$ ,  $\sigma_w \in [\frac{1}{m_1^{\frac{3}{2}}}, \frac{\tau_w}{m_1^{\frac{1}{4}}}]$ ,  $\sigma_v \in (0, \frac{1}{m_2^{\frac{1}{2}}})$ . The perturbation matrices satisfies  $\|\mathbf{W}'\|_{2,4} \leq \tau_w$ ,  $\|\mathbf{V}'\|_F \leq \tau_v$ ,  $\|\mathbf{W}''\|_{2,4} \leq \tau_w$ ,  $\|\mathbf{V}''\|_F \leq \tau_v$  and random diagonal matrix  $\Sigma$  has each diagonal entry i.i.d. drawn from  $\{\pm 1\}$ . Then with high probability, we have

(1) Sparse sign change

$$\|\mathbf{D}'_{n,\mathbf{w}}\|_0 \leq \tilde{O}(\tau_w^{\frac{4}{5}} m_1^{\frac{6}{5}})$$

$$\|\mathbf{D}'_{n,\mathbf{v}}\|_0 \leq \tilde{O}(m_2^{\frac{3}{2}} \sigma_v (\|\mathbf{A}\|_\infty + \|\mathbf{A}\|_\infty \tau_w m_1^{\frac{1}{4}}) + m_2 \|\mathbf{A}\|_\infty^{\frac{2}{3}} (\|\mathbf{A}\|_\infty \tau_v + \|\mathbf{A}\|_\infty \tau_w m_1^{\frac{1}{4}} (1 + \tau_v))^{\frac{2}{3}})$$

(2) Cross term vanish

$$\begin{aligned} & g_r(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}' + \eta \mathbf{W}'' \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}' + \eta \Sigma \mathbf{V}'') \\ &= g_r(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}', \mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}') + g_r^{(b,b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \eta \mathbf{W}'' \Sigma, \eta \Sigma \mathbf{V}'') \quad (\text{D.78}) \\ &\quad + g'_r \end{aligned}$$

for every  $r \in [K]$ , where  $\mathbb{E}_\Sigma[g'_r] = 0$  and  $|g'_r| \leq \eta \|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty^2 \tau_v$ .

**Proof:**

(1) We first consider the sign changes by  $\mathbf{W}^\rho$ . Since  $\tilde{\mathbf{a}}_n \mathbf{X} \mathbf{w}_i^{(0)} + B_{1(n,i)}^{(0)} = \tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{w}}_i^{(0)}$  where

$\tilde{\mathbf{w}}_i^{(0)} = (\mathbf{w}_i^{(0)}, B_{1(n,i)}^{(0)}) \in \mathbb{R}^{d+1}$  and  $\tilde{\mathbf{X}} = (\mathbf{X}, \mathbf{1}) \in \mathbb{R}^{N \times (d+1)}$ , we can ignore the bias term for simplicity. We have

$$\tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{w}}_i^{(0)} \sim \mathcal{N}(0, \frac{\|\tilde{\mathbf{a}}_n \tilde{\mathbf{X}}\|^2}{m_1})$$

$$\tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{w}}_i^\rho \sim \mathcal{N}(0, \|\tilde{\mathbf{a}}_n \tilde{\mathbf{X}}\|^2 \sigma_w^2)$$

Therefore,

$$\frac{\tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{w}}_i^{(0)}}{\tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{w}}_i^\rho} \sim p(z) = \frac{1}{\pi(\sigma_w \sqrt{m_1} z^2 + \frac{1}{\sigma_w \sqrt{m_1}})}$$

$$\begin{aligned} \Pr[|\tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{w}}_i^{(0)}| \leq |\tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{w}}_i^\rho|] &= \Pr[|z| \leq 1] \\ &= \int_{-1}^1 \frac{1}{\pi(\sigma_w \sqrt{m_1} z^2 + \frac{1}{\sigma_w \sqrt{m_1}})} dz \\ &= \int_{-(\sigma_w^2 m_1)^{\frac{1}{2}}}^{(\sigma_w^2 m_1)^{\frac{1}{2}}} \frac{1}{\pi(t^2 + 1)} dt \\ &= \frac{2}{\pi} \arctan \sigma_w \sqrt{m_1} \\ &\leq \tilde{O}(\sigma_w \sqrt{m_1}) \end{aligned} \tag{D.79}$$

Then, we have

$$\begin{aligned} \|\mathbf{D}_{n,w}''\|_0 &\leq \tilde{O}(\sigma_w m_1^{\frac{3}{2}}) \\ \|\tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{W}}^{(0)} \mathbf{D}_{n,w}''\|_2 &\leq \tilde{O}(\|\tilde{\mathbf{a}}_n \tilde{\mathbf{X}}\| \sigma_w^{\frac{3}{2}} m_1^{\frac{3}{4}}) \end{aligned}$$

We then consider the sign changes by  $\mathbf{W}'$ . Let  $s = \|\mathbf{D}'_{n,w} - \mathbf{D}_{n,w}''\|_0$  be the total number of sign changes in the first layer caused by adding  $\mathbf{W}'$ . Note that the total number of coordinated  $i$  such that  $|\tilde{\mathbf{a}}_n \tilde{\mathbf{X}}(\tilde{\mathbf{W}}^{(0)} + \tilde{\mathbf{W}}^\rho)_i| \leq s'' = \frac{2\tau_w}{s^{\frac{1}{4}}}$  is at most  $s'' m_1^{\frac{3}{2}}$  with high probability. Since  $\|\mathbf{W}'\|_{2,4} \leq \tau_w$ , we must have

$$s \leq \tilde{O}(s'' m_1^{\frac{3}{2}}) = \tilde{O}\left(\frac{\tau_w}{s^{\frac{1}{4}}} m_1^{\frac{3}{2}}\right)$$

$$\|\mathbf{D}'_{n,w} - \mathbf{D}_{n,w}''\|_0 = s \leq \tilde{O}(\tau_w^{\frac{4}{5}} m_1^{\frac{6}{5}})$$

$$\|\tilde{\mathbf{a}}_n \tilde{\mathbf{X}}(\tilde{\mathbf{W}}^{(0)} + \tilde{\mathbf{W}}^\rho)(\mathbf{D}'_{n,w} - \mathbf{D}_{n,w}'')\|_2 \leq \tilde{O}(s^{\frac{1}{4}} \tau_w) \leq \tilde{O}(\tau_w^{\frac{6}{5}} m_1^{\frac{3}{10}})$$

To sum up, we have

$$\|\mathbf{D}'_{n,w}\|_0 \leq \tilde{O}(\sigma_w m_1^{\frac{3}{2}} + \tau_w^{\frac{4}{5}} m_1^{\frac{6}{5}}) \leq \tilde{O}(\tau_w^{\frac{4}{5}} m_1^{\frac{6}{5}})$$

Denote  $\mathbf{z}_{n,0} = \tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{W}}^{(0)} \mathbf{D}_{n,w}$  and  $\mathbf{z}_{n,2} = \tilde{\mathbf{a}}_n \tilde{\mathbf{X}} (\tilde{\mathbf{W}}^{(0)} + \tilde{\mathbf{W}}^\rho + \mathbf{W}') (\mathbf{D}_{n,w} + \mathbf{D}'_{n,w}) - \tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{W}}^{(0)} \mathbf{D}_{n,w}$ . With high probability, we know

$$\begin{aligned} \|\mathbf{z}_{n,2}\| &\leq \|\tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \mathbf{W}'\| + \|\tilde{\mathbf{a}}_n \tilde{\mathbf{X}} \tilde{\mathbf{W}}^\rho\| \\ &\leq \tilde{O}(m_1^{\frac{1}{4}} \tau_w \|\mathbf{A}\|_\infty + \|\mathbf{A}\|_\infty \sigma_w m_1^{\frac{1}{2}}) \\ &\leq \tilde{O}(\|\mathbf{A}\|_\infty \tau_w m_1^{\frac{1}{4}}) \end{aligned} \quad (\text{D.80})$$

Denote  $\mathbf{Z}_0 = (\mathbf{z}_{1,0}^\top, \dots, \mathbf{z}_{N,0}^\top)^\top \in \mathbb{R}^{N \times m_1}$ ,  $\mathbf{Z}_2 = (\mathbf{z}_{1,2}^\top, \dots, \mathbf{z}_{N,2}^\top)^\top \in \mathbb{R}^{N \times m_1}$ . The sign change in the second layer is from  $\tilde{\mathbf{a}}_n \mathbf{Z}_0 \mathbf{V}^{(0)}$  to  $\tilde{\mathbf{a}}_n (\mathbf{Z}_0 + \mathbf{Z}_2) (\mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}')$ . We have

$$\|\tilde{\mathbf{a}}_n (\mathbf{Z}_0 + \mathbf{Z}_2) \mathbf{V}^\rho\|_\infty \leq \tilde{O}(\sigma_v \|\mathbf{A}\|_\infty (\|\mathbf{z}_{1,0}\| + \|\mathbf{z}_{1,2}\|))$$

$$\|\tilde{\mathbf{a}}_n (\mathbf{Z}_0 + \mathbf{Z}_2) \mathbf{V}' + \tilde{\mathbf{a}}_n \mathbf{Z}_2 \mathbf{V}^{(0)}\|_2 \leq \tilde{O}(\|\mathbf{A}\|_\infty ((\|\mathbf{z}_{1,0}\| + \|\mathbf{z}_{1,2}\|) \tau_v + \|\mathbf{z}_{1,2}\|))$$

Combining  $\|\mathbf{z}_{1,0}\| \leq \tilde{O}(\|\mathbf{A}\|_\infty)$ , by Claim C.8 in [58] we have

$$\|\mathbf{D}'_{n,v}\|_0 \leq \tilde{O}(m_2^{\frac{3}{2}} \sigma_v (\|\mathbf{A}\|_\infty^2 + \|\mathbf{A}\|_\infty^2 \tau_w m_1^{\frac{1}{4}}) + m_2 \|\mathbf{A}\|_\infty^{\frac{2}{3}} (\|\mathbf{A}\|_\infty \tau_v + \|\mathbf{A}\|_\infty \tau_w m_1^{\frac{1}{4}} (1 + \tau_v))^{\frac{2}{3}})$$

(2) Diagonal Cross terms.

Denote  $\mathbf{D}_w = (\text{diag}(\mathbf{D}_{1,w})^\top, \dots, \text{diag}(\mathbf{D}_{N,m_1})^\top)^\top \in \mathbb{R}^{N \times m_1}$  and define  $\mathbf{D}'_w$ ,  $\mathbf{D}''_w$ ,  $\mathbf{D}_v$ ,  $\mathbf{D}'_v$ ,  $\mathbf{D}''_v$  accordingly.

Recall

$$g_r(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}, \mathbf{V}) = \mathbf{q}^\top \mathbf{A} ((\mathbf{A}(\mathbf{A} \mathbf{X} \mathbf{W} + \mathbf{B}_1) \odot (\mathbf{D}_w + \mathbf{D}'_w) \mathbf{V} + \mathbf{B}_2) \odot (\mathbf{D}_v + \mathbf{D}'_v)) \mathbf{c}_r$$

$$g_r^{(b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}, \mathbf{V}) = \mathbf{q}^\top \mathbf{A} ((\mathbf{A}(\mathbf{A} \mathbf{X} \mathbf{W} + \mathbf{B}_1) \odot (\mathbf{D}_w + \mathbf{D}'_w) \mathbf{V}) \odot (\mathbf{D}_v + \mathbf{D}'_v)) \mathbf{c}_r$$

$$g_r^{(b,b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}, \mathbf{V}) = \mathbf{q}^\top \mathbf{A} ((\mathbf{A}(\mathbf{A} \mathbf{X} \mathbf{W}) \odot (\mathbf{D}_w + \mathbf{D}'_w) \mathbf{V}) \odot (\mathbf{D}_v + \mathbf{D}'_v)) \mathbf{c}_r$$

Then

$$\begin{aligned}
& g_r(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}' + \eta \mathbf{W}'' \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}' + \eta \Sigma \mathbf{V}'') \\
&= g_r(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}', \mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}') + g_r^{(b,b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \eta \mathbf{W}'' \Sigma, \eta \Sigma \mathbf{V}'') \\
&\quad + g_r^{(b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}', \eta \Sigma \mathbf{V}'') + g_r^{(b,b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \eta \mathbf{W}'' \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho \\
&\quad + \mathbf{V}'),
\end{aligned} \tag{D.81}$$

where the last two terms are the error terms. We know that

$$\|\mathbf{W}^{(0)}\| \leq \max_{\|\mathbf{a}\|=1} \|\mathbf{a}^\top \mathbf{W}^{(0)}\| \leq \max_{\|\mathbf{a}\|=1} \sqrt{\sum_{i=1}^{m_1} (\mathbf{a}^\top \mathbf{w}_i^{(0)})^2} \leq \max_{\|\mathbf{a}\|=1} \sqrt{\sum_{i=1}^{m_1} \left(\frac{1}{\sqrt{m_1}}\right)^2} = 1$$

Therefore,

$$\begin{aligned}
& |g_r^{(b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}', \eta \Sigma \mathbf{V}'')| \\
&= \eta \sum_{n=1}^N \sum_{i=1}^{m_2} \mathbf{q}^\top \mathbf{a}_n c_{i,r} D_{n,\mathbf{v}_i} \sum_{k=1}^N a_{n,k} \sum_{l=1}^{m_1} (\Sigma \mathbf{V})''_{l,i} D_{k,\mathbf{w}_l} (\mathbf{a}_k \mathbf{X} (\mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}'))_l \\
&\quad + \mathbf{B}_{1(k,l)}) \\
&= \eta \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{k=1}^N a_{n,k} ((\mathbf{a}_k \mathbf{X} (\mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}')) + \mathbf{B}_{1k}) \odot \mathbf{D}_{k,\mathbf{w}} \Sigma \mathbf{V}'' \odot \mathbf{D}_{n,\mathbf{v}} \mathbf{c}_r \\
&\leq \eta \|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty (\|\mathbf{A}\|_\infty (m^{-\frac{1}{2}} + \sigma_w + \tau_w) + m^{-\frac{1}{2}}) \tau_v m_2^{\frac{1}{2}} \\
&\leq \tilde{O}(\eta \|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty^2 \tau_v m_2^{\frac{1}{2}} m_1^{-\frac{1}{2}}),
\end{aligned} \tag{D.82}$$

where the last step is by the value selection of  $\sigma_w$ ,  $\tau_w$  and  $\tau_v$ .

$$\begin{aligned}
& |g_r^{(b,b)}(\mathbf{q}, \mathbf{A}, \eta \mathbf{W}'' \Sigma, \mathbf{X}, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}')| \\
&= |\eta \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{k=1}^N a_{n,k} (\mathbf{a}_k \mathbf{X} \mathbf{W}'' \Sigma \odot (\mathbf{D}_{k,w} + \mathbf{D}_{k,w}')) (\mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}'') \odot (\mathbf{D}_{n,v} \\
&\quad + \mathbf{D}_{n,v})' \mathbf{c}_r| \\
&\leq |\eta \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{k=1}^N a_{n,k} (\mathbf{a}_k \mathbf{X} \mathbf{W}'' \Sigma \odot (\mathbf{D}_{k,w} + \mathbf{D}_{k,w})) \mathbf{V}' \odot (\mathbf{D}_{n,v} + \mathbf{D}_{n,v})' \mathbf{c}_r| \\
&\quad + |\eta \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{k=1}^N a_{n,k} (\mathbf{a}_k \mathbf{X} \mathbf{W}'' \Sigma \odot (\mathbf{D}_{k,w} + \mathbf{D}_{k,w})) (\mathbf{V}^{(0)} + \mathbf{V}^\rho) \odot \mathbf{D}_{n,v} \mathbf{c}_r| \\
&\quad + |\eta \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{k=1}^N a_{n,k} (\mathbf{a}_k \mathbf{X} \mathbf{W}'' \Sigma \odot (\mathbf{D}_{k,w} + \mathbf{D}_{k,w})) (\mathbf{V}^{(0)} + \mathbf{V}^\rho) \odot \mathbf{D}_{n,v}' \mathbf{c}_r| \\
&\leq |\eta \| \mathbf{q}^\top \mathbf{A} \|_1 \| \mathbf{A} \|_\infty^2 \tau_w \tau_v m_2^{\frac{1}{2}}| + 2 |\eta \| \mathbf{q}^\top \mathbf{A} \|_1 \| \mathbf{A} \|_\infty^2 \tau_w m_1^{\frac{1}{2}}| \\
&\leq \tilde{O}(|\eta \| \mathbf{q}^\top \mathbf{A} \|_1 \| \mathbf{A} \|_\infty^2 \tau_w m_1^{\frac{1}{2}}|)
\end{aligned} \tag{D.83}$$

**Lemma D.2.11.** Denote

$$\begin{aligned}
P_{\rho,\eta} &= F_{\mathbf{A}}(\mathbf{q}, \mathbf{X}, \mathbf{W} + \mathbf{W}^\rho + \eta \mathbf{W}'' \Sigma, \mathbf{V} + \mathbf{V}^\rho + \eta \Sigma \mathbf{V}'') \\
&= \mathbf{q}^\top \mathbf{A} (\mathbf{A} (\mathbf{A} \mathbf{X} (\mathbf{W} + \mathbf{W}^\rho + \eta \mathbf{W}'' \Sigma) + \mathbf{B}_1) \odot \mathbf{D}_{w,\rho,\eta} (\mathbf{V} + \mathbf{V}^\rho + \eta \Sigma \mathbf{V}'') \\
&\quad \odot \mathbf{D}_{v,\rho,\eta}) \mathbf{c}_r
\end{aligned} \tag{D.84}$$

$$\begin{aligned}
P'_{\rho,\eta} &= G(\mathbf{q}, \mathbf{A}, \mathbf{X}, \mathbf{W} + \mathbf{W}^\rho + \eta \mathbf{W}'' \Sigma, \mathbf{V} + \mathbf{V}^\rho + \eta \Sigma \mathbf{V}'') \\
&= \mathbf{q}^\top \mathbf{A} (\mathbf{A} (\mathbf{A} \mathbf{X} (\mathbf{W} + \mathbf{W}^\rho + \eta \mathbf{W}'' \Sigma) + \mathbf{B}_1) \odot \mathbf{D}_{w,\rho} (\mathbf{V} + \mathbf{V}^\rho + \eta \Sigma \mathbf{V}'') \\
&\quad \odot \mathbf{D}_{v,\rho}) \mathbf{c}_r
\end{aligned} \tag{D.85}$$

There exists  $\eta_0 = \frac{1}{\text{poly}(m_1, m_2)}$  such that for every  $\eta \leq \eta_0$ , for every  $\mathbf{W}''$ ,  $\mathbf{V}''$  that satisfies  $\|\mathbf{W}''\|_{2,\infty} \leq \tau_{w,\infty}$ ,  $\|\mathbf{V}''\|_{2,\infty} \leq \tau_{v,\infty}$ , we have

$$\mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho} \left[ \frac{|P_{\rho,\eta} - P'_{\rho,\eta}|}{\eta^2} \right] = \tilde{O}(\mathbf{q}^\top \mathbf{A} \mathbf{1} \| \mathbf{A} \|^4 \left( \frac{\tau_{w,\infty}^2}{\sigma_w} m_1 + \frac{(\tau_{w,\infty}^2 + \tau_{v,\infty}^2 m_1^{-1})}{\sigma_v} m_2 \right)) + O_p(\eta),$$

where  $O_p$  hides polynomial factor of  $m_1$  and  $m_2$ .

**Proof:**

$$\begin{aligned}
& P_{\rho,\eta} - P'_{\rho,\eta} \\
&= \mathbf{q}^\top \mathbf{A} (\mathbf{A}(\mathbf{AX}(\mathbf{W} + \mathbf{W}^\rho + \eta \mathbf{W}'' \Sigma) + \mathbf{B}_1) \odot (\mathbf{D}_{\mathbf{w},\rho,\eta} - \mathbf{D}_{\mathbf{w},\rho})) (\mathbf{V} + \mathbf{V}^\rho + \eta \Sigma \mathbf{V}'') \\
&\quad \odot \mathbf{D}_{\mathbf{v},\rho}) \mathbf{c}_r + \mathbf{q}^\top \mathbf{A} (\mathbf{A}(\mathbf{AX}(\mathbf{W} + \mathbf{W}^\rho + \eta \mathbf{W}'' \Sigma) + \mathbf{B}_1) \odot \mathbf{D}_{\mathbf{w},\rho,\eta} (\mathbf{V} + \mathbf{V}^\rho \\
&\quad + \eta \Sigma \mathbf{V}'') (\mathbf{D}_{\mathbf{v},\rho,\eta} - \mathbf{D}_{\mathbf{v},\rho})) \mathbf{c}_r
\end{aligned} \tag{D.86}$$

We write

$$\begin{aligned}
\mathbf{Z} &= \mathbf{A}(\mathbf{AX}(\mathbf{W} + \mathbf{W}^\rho + \eta \mathbf{W}'' \Sigma) + \mathbf{B}_1) \odot \mathbf{D}_{\mathbf{w},\rho} \\
\mathbf{Z}' &= \mathbf{A}(\mathbf{AX}(\mathbf{W} + \mathbf{W}^\rho + \eta \mathbf{W}'' \Sigma) + \mathbf{B}_1) \odot \mathbf{D}_{\mathbf{w},\rho,\eta}
\end{aligned}$$

Since for all  $n \in [N]$ ,  $\|\eta(\mathbf{A}\mathbf{AX}\mathbf{W}''\Sigma)_n\|_\infty \leq \eta\|\mathbf{A}\|_\infty \tau_{\mathbf{w},\infty}$ , we have

$$\|\mathbf{Z}'_n\|_\infty \leq \eta\|\mathbf{A}\|_\infty \tau_{\mathbf{w},\infty}$$

$$\Pr_{\mathbf{W}^\rho} [Z'_{n,i} \neq 0] \leq \tilde{O} \left( \frac{\eta\|\mathbf{A}\|_\infty \tau_{\mathbf{w},\infty}}{\sigma_w} \right), \quad i \in [m_1]$$

Then we have

$$\Pr[\|\mathbf{Z}'_n\|_0 \geq 2] \leq O_p(\eta^2)$$

Then we only need to consider the case  $\|\mathbf{Z}'_n\|_0 = 1$ . Let  $Z'_{n,n_i} \neq 0$ . Then the first term in (D.86),  $\mathbf{q}^\top \mathbf{A} (\mathbf{Z}'(\mathbf{V} + \mathbf{V}^\rho + \eta \Sigma \mathbf{V}'') \odot \mathbf{D}_{\mathbf{v},\rho}) \mathbf{c}_r$  should be dealt with separately.

The term  $\mathbf{q}^\top \mathbf{A} (\mathbf{Z}' \eta \Sigma \mathbf{V}'' \odot \mathbf{D}_{\mathbf{v},\rho}) \mathbf{c}_r$  contributes to  $O_p(\eta^3)$  to the whole term.

Then we have

$$\|\mathbf{q}^\top \mathbf{A} (\mathbf{Z}' \eta (\mathbf{V} + \mathbf{V}^\rho) \odot \mathbf{D}_{\mathbf{v},\rho}) \mathbf{c}_r\| \leq \tilde{O}(\eta \|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty \tau_{\mathbf{w},\infty})$$

We also have that

$$\tilde{O} \left( \left( \frac{\eta\|\mathbf{A}\|_\infty \tau_{\mathbf{w},\infty}}{\sigma_w} m_1 \right)^N \right) \leq \tilde{O} \left( \frac{\eta\|\mathbf{A}\|_\infty \tau_{\mathbf{w},\infty}}{\sigma_w} m_1 \right) \leq 1$$

Therefore, the contribution to the first term is  $\tilde{O}(\eta^2 \|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty^2 \frac{\tau_{\mathbf{w},\infty}^2}{\sigma_w} m_1) + O_p(\eta^3)$ .

Denote

$$\begin{aligned}
\boldsymbol{\delta} &= \mathbf{A}(\mathbf{AX}(\mathbf{W} + \mathbf{W}^\rho + \eta \mathbf{W}'' \Sigma) + \mathbf{B}_1) \odot \mathbf{D}_{\mathbf{w},\rho,\eta} (\mathbf{V} + \mathbf{V}^\rho + \eta \Sigma \mathbf{V}'') \\
&\quad - \mathbf{A}(\mathbf{AX}(\mathbf{W} + \mathbf{W}^\rho) + \mathbf{B}_1) \odot \mathbf{D}_{\mathbf{w},\rho} (\mathbf{V} + \mathbf{V}^\rho)
\end{aligned} \tag{D.87}$$

$\delta \in \mathbb{R}^{m_2}$  has the following terms:

1.  $\mathbf{Z}'(\mathbf{V} + \mathbf{V}^\rho + \eta \Sigma \mathbf{V}'')$ . We have its n-th row norm bounded by  $O_p(\eta)$ .
2.  $\mathbf{Z}\eta\Sigma\mathbf{V}''$ . We have its n-th row infinity norm bounded by  $\tilde{O}(\|\mathbf{A}\|_\infty \eta \tau_{v,\infty} m_1^{-\frac{1}{2}})$ .
3.  $\mathbf{A}(\mathbf{AX}\eta\mathbf{W}''\Sigma\odot\mathbf{D}_{w,\rho})(\mathbf{V} + \mathbf{V}^\rho)$ , of which the n-th row infinity is bounded by  $\tilde{O}(\|\mathbf{A}\|_\infty \eta \tau_{w,\infty})$ .
4.  $\mathbf{A}(\mathbf{AX}\eta^2\mathbf{W}''\Sigma \odot \mathbf{D}_{w,\rho,\eta}\Sigma\mathbf{V}'')$ . Bounded by  $O_p(\eta^2)$ .

Therefore,

$$\|\delta_n\|_\infty \leq \tilde{O}(\|\mathbf{A}\|_\infty \eta (\tau_{v,\infty} m_1^{-\frac{1}{2}} + \tau_{w,\infty})) + O_p(\eta^2)$$

Similarly, we can derive that the contribution to the second term is

$$\tilde{O}(\eta^2 \|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty^2 \frac{(\tau_{w,\infty}^2 + \tau_{v,\infty}^2 m_1^{-1})}{\sigma_v} m_2) + O_p(\eta^3).$$

**Lemma D.2.12.** Let  $\mathbf{F}_A^* = (f_1^*, \dots, f_K^*)$ . Perturbation matrices  $\mathbf{W}', \mathbf{V}'$  satisfy

$$\|\mathbf{W}'\|_{2,4} \leq \tau_w, \quad \|\mathbf{V}'\|_F \leq \tau_v$$

There exists  $\widehat{\mathbf{W}}$  and  $\widehat{\mathbf{V}}$  such that

$$\|\widehat{\mathbf{W}}\|_{2,\infty} \leq \frac{C_0}{m_1}, \quad \|\widehat{\mathbf{V}}\|_{2,\infty} \leq \frac{K\sqrt{m_1}}{m_2}$$

$$\mathbb{E}[\sum_{r=1}^K |f_r^*(\mathbf{q}, \mathbf{A}, \mathbf{X}, \widehat{\mathbf{W}}, \widehat{\mathbf{V}}) - g_r^{(b,b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \widehat{\mathbf{W}}, \widehat{\mathbf{V}})|] \leq \epsilon$$

$$\mathbb{E}[G^{(b,b)}(\mathbf{q}, \mathbf{A}, \mathbf{X}, \widehat{\mathbf{W}}, \widehat{\mathbf{V}})] \leq OPT + \epsilon$$

### Proof:

By Lemma D.2.10, we have

$$\|\mathbf{D}'_{n,w}\|_0 \leq \tilde{O}(\tau_w^{\frac{4}{5}} m_1^{\frac{6}{5}}) \ll \tilde{O}(m_1)$$

$$\begin{aligned} \|\mathbf{D}'_{n,v}\|_0 &\leq \tilde{O}(m_2^{\frac{3}{2}} \sigma_v (\|\mathbf{A}\|_\infty^2 + \|\mathbf{A}\|_\infty^2 \tau_w m_1^{\frac{1}{4}}) + m_2 \|\mathbf{A}\|_\infty^{\frac{2}{3}} (\|\mathbf{A}\|_\infty \tau_v \\ &\quad + \|\mathbf{A}\|_\infty \tau_w m_1^{\frac{1}{4}} (1 + \tau_v))^{\frac{2}{3}}) \\ &\leq \tilde{O}(m_2 \|\mathbf{A}\|_\infty^2 (\epsilon/C_0)^{\Theta(1)}) \end{aligned} \tag{D.88}$$

Applying Lemma D.2.9, we know

$$\begin{aligned}
& \mathbf{q}^\top \mathbf{A}((\mathbf{A}(\mathbf{AX}\widehat{\mathbf{W}}) \odot (\mathbf{D}'_{\mathbf{w}})\widehat{\mathbf{V}}) \odot (\mathbf{D}_{\mathbf{v}}))\mathbf{c}_r \\
&= \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{k=1}^N a_{n,k} ((\mathbf{a}_k \mathbf{X}\widehat{\mathbf{W}}) \odot \mathbf{D}_{k,\mathbf{w}}') \widehat{\mathbf{V}} \odot \mathbf{D}_{n,\mathbf{v}} \mathbf{c}_r \\
&\leq \|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty^2 m_1^{\frac{3}{10}} \frac{C_0}{m_1} \frac{K\sqrt{m_1}}{m_2} \cdot m_2 \\
&\leq \epsilon
\end{aligned} \tag{D.89}$$

$$\begin{aligned}
& \mathbf{q}^\top \mathbf{A}((\mathbf{A}(\mathbf{AX}\widehat{\mathbf{W}}) \odot (\mathbf{D}_{\mathbf{w}})\widehat{\mathbf{V}}) \odot (\mathbf{D}'_{\mathbf{v}}))\mathbf{c}_r \\
&= \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{k=1}^N a_{n,k} ((\mathbf{a}_k \mathbf{X}\widehat{\mathbf{W}}) \odot \mathbf{D}_{k,\mathbf{w}}) \widehat{\mathbf{V}} \odot \mathbf{D}'_{n,\mathbf{v}} \mathbf{c}_r \\
&\leq \|\mathbf{q}^\top \mathbf{A}\|_1 \|\mathbf{A}\|_\infty^2 m_1^{\frac{1}{2}} \frac{C_0}{m_1} \frac{K\sqrt{m_1}}{m_2} \cdot m_2 \cdot \|\mathbf{A}\|_\infty^2 \left(\frac{\epsilon}{C_0}\right)^{\Theta(1)} \\
&\leq \epsilon
\end{aligned} \tag{D.90}$$

Then, the conclusion can be derived.

**Optimization** This section states the optimization process and convergence performance of the algorithm. Lemma D.2.13 shows that during the optimization, either there exists an updating direction that decreases the objective, or weight decay decreases the objective. Lemma D.2.14 provides the convergence result of the algorithm.

Define

$$\begin{aligned}
& L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \\
&= \frac{1}{\Omega^t} \sum_{i=1}^{|\Omega^t|} \mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho, \Sigma'} [L(\lambda_t F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}; \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_t \Sigma', \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma' \mathbf{V}_t), y_i)] \\
&\quad + R(\sqrt{\lambda_t} \mathbf{W}_t, \sqrt{\lambda_t} \mathbf{V}_t)
\end{aligned} \tag{D.91}$$

where

$$R(\sqrt{\lambda} \mathbf{W}_t, \sqrt{\lambda} \mathbf{V}_t) = \lambda_v \|\sqrt{\lambda} \mathbf{V}_t\|_F^2 + \lambda_w \|\sqrt{\lambda} \mathbf{W}_t\|_{2,4}^2$$

**Lemma D.2.13.** For every  $\epsilon_0 \in (0, 1)$ ,  $\epsilon \in (0, \frac{\epsilon_0}{K \|\mathbf{A}\|_\infty p_1 p_2^2 \mathcal{C}_s(\Phi, p_2 \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty)) \mathcal{C}_s(\phi, \|\mathbf{A}\|_\infty) \sqrt{\|\mathbf{A}\|_\infty^2 + 1}})$  and  $\gamma \in (0, \frac{1}{4}]$ , consider any  $\mathbf{W}_t, \mathbf{V}_t$  with

$$L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \in [(1 + \gamma) OPT + \Omega(\mathbf{q}^\top \mathbf{A}^* \mathbf{1} \|\mathbf{A}^*\|_\infty^4 \epsilon_0 / \gamma), \tilde{O}(1)]$$

With high probability on random initialization, there exists  $\widehat{\mathbf{W}}$ ,  $\widehat{\mathbf{V}}$  with  $\|\widehat{\mathbf{W}}\|_F \leq 1$ ,  $\|\widehat{\mathbf{V}}\|_F \leq 1$  such that for every  $\eta \in (0, \frac{1}{\text{poly}(m_1, m_2)})$ ,

$$\begin{aligned} & \min\{\mathbb{E}_{\Sigma}[L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t + \sqrt{\eta}\widehat{\mathbf{W}}\Sigma, \mathbf{V}_t + \sqrt{\eta}\Sigma\widehat{\mathbf{V}})], \\ & L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, (1-\eta)\lambda_t, \mathbf{W}_t, \mathbf{V}_t)\} \\ & \leq (1 - \eta\gamma/4)L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \end{aligned} \quad (\text{D.92})$$

### Proof:

Recall the pseudo network and the real network for every  $r \in [K]$  as

$$\begin{aligned} & g_r(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \mathbf{W}', \mathbf{V}') \\ &= \mathbf{q}^\top \mathbf{A}^* (\mathbf{A}^* (\mathbf{A}^* \mathbf{X} (\mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}') + \mathbf{B}_1) \odot \mathbf{D}_{w,\rho,t} (\mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}') \odot \mathbf{D}_{v,\rho,t}) \mathbf{c}_r \end{aligned} \quad (\text{D.93})$$

$$\begin{aligned} & f_r(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \mathbf{W}', \mathbf{V}') \\ &= \mathbf{q}^\top \mathbf{A}^* (\mathbf{A}^* (\mathbf{A}^* \mathbf{X} (\mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}') + \mathbf{B}_1) \odot \mathbf{D}_{w,\rho,\mathbf{W}'} (\mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}') \\ & \quad \odot \mathbf{D}_{v,\rho,\mathbf{V}'}) \mathbf{c}_r \end{aligned} \quad (\text{D.94})$$

where  $\mathbf{D}_{w,\rho,t}$  and  $\mathbf{D}_{v,\rho,t}$  are the diagonal matrices at weights  $\mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_t$  and  $\mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}_t$ .  $\mathbf{D}_{w,\rho,\mathbf{W}'}$  and  $\mathbf{D}_{v,\rho,\mathbf{V}'}$  are the diagonal matrices at weights  $\mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}'$  and  $\mathbf{V}^{(0)} + \mathbf{V}^\rho + \mathbf{V}'$ .

Denote  $G(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \mathbf{W}', \mathbf{V}') = (g_1, \dots, g_K)$ ,  $F_{\mathbf{A}^*}(\mathbf{q}, \mathbf{X}, \mathbf{W}', \mathbf{V}') = (f_1, \dots, f_K)$ .

As long as  $L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \leq \tilde{O}(1)$ , according to C.32 to C.34 in [58], we have

$$\lambda_w \|\sqrt{\lambda_t} \widehat{\mathbf{W}}\|_{2,4}^4 \leq \epsilon_0$$

$$\lambda_v \|\sqrt{\lambda_t} \widehat{\mathbf{V}}\|_F^2 \leq \epsilon_0$$

$$\|\widehat{\mathbf{W}}\|_F \ll 1$$

$$\|\widehat{\mathbf{V}}\|_F \ll 1$$

The we need to study an update direction

$$\widetilde{\mathbf{W}} = \mathbf{W}_t + \sqrt{\eta} \widehat{\mathbf{W}} \Sigma$$

$$\tilde{\mathbf{V}} = \mathbf{V}_t + \sqrt{\eta} \Sigma \hat{\mathbf{V}}$$

**Changes in Regularizer.** Note that here  $\mathbf{W}_t \in \mathbb{R}^{d \times m_1}$ ,  $\mathbf{V}_t \in \mathbb{R}^{m_1 \times m_2}$ ,  $\Sigma \in \mathbb{R}^{m_1 \times m_1}$ . We know that

$$\mathbb{E}_{\Sigma}[\|\mathbf{V}_t + \sqrt{\eta} \Sigma \hat{\mathbf{V}}\|_F^2] = \|\mathbf{V}_t\|_F^2 + \eta \|\hat{\mathbf{V}}\|_F^2$$

$$\mathbb{E}_{\Sigma}[\|\mathbf{W}_t + \sqrt{\eta} \hat{\mathbf{W}} \Sigma\|_{2,4}^4] = \sum_{i \in [m_1]} \mathbb{E}[\|\mathbf{w}_{t,i} + \sqrt{\eta} \hat{\mathbf{W}} \Sigma_i\|_2^4]$$

For each term  $i \in [m_1]$ , we can bound

$$\|\mathbf{w}_{t,i} + \sqrt{\eta} \hat{\mathbf{W}} \Sigma_i\|_2^2 = \|\mathbf{w}_{t,i}\|_2^2 + \eta \|\hat{\mathbf{W}} \Sigma_i\|_2^2 + 2\sqrt{\eta} \mathbf{w}_{t,i}^\top \hat{\mathbf{W}} \Sigma_i$$

$$\begin{aligned} \|\mathbf{w}_{t,i} + \sqrt{\eta} \hat{\mathbf{W}} \Sigma_i\|_2^4 &= \|\mathbf{w}_{t,i}\|_2^4 + \eta^2 \|\hat{\mathbf{W}} \Sigma_i\|_2^4 + 4\eta \|\mathbf{w}_{t,i}^\top \hat{\mathbf{W}} \Sigma_i\|^2 + 2\eta \|\mathbf{w}_{t,i}\|_2^2 \|\hat{\mathbf{W}} \Sigma_i\|_2^2 \\ &\leq \|\mathbf{w}_{t,i}\|_2^4 + 6\eta \|\mathbf{w}_{t,i}\|_2^2 \|\hat{\mathbf{W}} \Sigma_i\|_2^2 + O_p(\eta^2) \end{aligned} \quad (\text{D.95})$$

Therefore, by Cauchy-Schwarz inequality, we have

$$\mathbb{E}_{\Sigma}[\|\mathbf{W}_t + \sqrt{\eta} \hat{\mathbf{W}} \Sigma\|_{2,4}^4] \leq \|\mathbf{W}_t\|_{2,4}^4 + 6\eta \|\mathbf{W}_t\|_{2,4}^2 \|\hat{\mathbf{W}}\|_{2,4}^2 + O_p(\eta^2)$$

Therefore, by  $\lambda_w \|\sqrt{\lambda_t} \mathbf{W}_t\|_{2,4}^4 \leq R(\sqrt{\lambda_t} \mathbf{W}_t, \sqrt{\lambda_t} \mathbf{V}_t)$ , we have

$$\begin{aligned} \mathbb{E}[R(\sqrt{\lambda_t} \tilde{\mathbf{W}}, \sqrt{\lambda_t} \tilde{\mathbf{V}})] &\leq R(\sqrt{\lambda_t} \mathbf{W}_t, \sqrt{\lambda_t} \mathbf{V}_t) + 6\eta \sqrt{\epsilon_0} \sqrt{R(\sqrt{\lambda_t} \mathbf{W}_t, \sqrt{\lambda_t} \mathbf{V}_t)} + \eta \epsilon_0 \\ &\leq R(\sqrt{\lambda_t} \mathbf{W}_t, \sqrt{\lambda_t} \mathbf{V}_t) + \frac{1}{4} \eta R(\sqrt{\lambda_t} \mathbf{W}_t, \sqrt{\lambda_t} \mathbf{V}_t) + 143\eta \epsilon_0 \end{aligned} \quad (\text{D.96})$$

**Changes in Objective.** Recall that here  $\hat{\mathbf{W}}$  and  $\hat{\mathbf{V}}$  satisfy  $\tau_{\mathbf{w}, \infty} \leq \frac{1}{m_1^{1000}}$  and  $\tau_{\mathbf{v}, \infty} \leq \frac{1}{m_2^{2000}}$ .

By Lemma D.2.11, we have for every  $r \in [K]$

$$\begin{aligned} &\mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho} [|f_r(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \mathbf{W} + \mathbf{W}^\rho + \tilde{\mathbf{W}} \Sigma, \mathbf{V} + \mathbf{V}^\rho + \Sigma \tilde{\mathbf{V}}) \\ &\quad - g_r(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \mathbf{W} + \mathbf{W}^\rho + \tilde{\mathbf{W}} \Sigma, \mathbf{V} + \mathbf{V}^\rho + \Sigma \tilde{\mathbf{V}})|] \\ &\leq \tilde{O}(\|\mathbf{q}^\top \mathbf{A}^*\|_1 \|\mathbf{A}^*\|_\infty^2 \epsilon_0 \eta) + O_p(\eta^{1.5}) \end{aligned} \quad (\text{D.97})$$

By Lemma D.2.10, we have

$$\begin{aligned}
G(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{V}}) &= G(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \mathbf{W}_t, \mathbf{V}_t) + \eta G^{(b,b)}(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \widehat{\mathbf{W}}\Sigma, \Sigma\widehat{\mathbf{V}}) + \sqrt{\eta}G' \\
&= F_{\mathbf{A}^*}(\mathbf{q}, \mathbf{X}, \mathbf{W}_t, \mathbf{V}_t) + \eta G^{(b,b)}(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \widehat{\mathbf{W}}\Sigma, \Sigma\widehat{\mathbf{V}}) + \sqrt{\eta}G' \quad (\text{D.98}) \\
&= F_{\mathbf{A}^*}(\mathbf{q}, \mathbf{X}, \mathbf{W}_t, \mathbf{V}_t) + \eta G^{(b,b)}(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \widehat{\mathbf{W}}, \widehat{\mathbf{V}}) + \sqrt{\eta}G'
\end{aligned}$$

where  $\mathbb{E}_{\Sigma}[G'] = 0$  and  $|G'| \leq \epsilon$  with high probability. By C.38 in [58], we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho, \Sigma}[L(\lambda_t F_{\mathbf{A}^*}(\mathbf{q}, \mathbf{X}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{V}}), y)] \\
&\leq \mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho}[L(\lambda_t F_{\mathbf{A}^*}(\mathbf{q}, \mathbf{X}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{V}}) + \eta F_{\mathbf{A}^*}^*(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \widetilde{\mathbf{W}}, \widetilde{\mathbf{V}}), y)] \quad (\text{D.99}) \\
&\quad + O(\|\mathbf{q}^\top \mathbf{A}^*\|_1 \|\mathbf{A}^*\|_\infty^2 \epsilon_0 \eta) + O_p(\eta^{1.5})
\end{aligned}$$

Following C.40 in [58], we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho}[L(\lambda_t F_{\mathbf{A}^*}(\mathbf{q}, \mathbf{X}, \mathbf{W}_t, \mathbf{V}_t) + \eta F_{\mathbf{A}^*}^*(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \mathbf{W}_t, \mathbf{V}_t), y)] \\
&\leq (1 - \eta)(2L(\lambda_t F_{\mathbf{A}^*}(\mathbf{q}, \mathbf{X}, \mathbf{W}_t, \mathbf{V}_t), y) - L((1 - \eta)\lambda_t F_{\mathbf{A}^*}(\mathbf{q}, \mathbf{X}, \mathbf{W}_t, \mathbf{V}_t), y)) \quad (\text{D.100}) \\
&\quad + \eta L(F_{\mathbf{A}^*}^*, y) + O_p(\eta^2)
\end{aligned}$$

**Putting all of them together.** Denote

$$c_1 = \frac{1}{|\Omega^t|} \sum_{i=1}^{|\Omega^t|} \mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho, \Sigma, \Sigma'}[L(\lambda_t F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \widetilde{\mathbf{W}}\Sigma', \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma'\widetilde{\mathbf{V}}), y_i)] \quad (\text{D.101})$$

$$c'_1 = \mathbb{E}_{\Sigma}[L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \widetilde{\mathbf{W}}, +\widetilde{\mathbf{V}})] = c_1 + \mathbb{E}_{\Sigma}[R(\sqrt{\lambda_t} \widetilde{\mathbf{W}}, \sqrt{\lambda} \widetilde{\mathbf{V}})] \quad (\text{D.102})$$

$$\begin{aligned}
c_2 &= \frac{1}{|\Omega^t|} \sum_{i=1}^{|\Omega^t|} \mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho}[L((1 - \eta)\lambda_t F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_t\Sigma', \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma'\mathbf{V}_t), y_i)] \\
&\quad + \Sigma'\mathbf{V}_t), y_i)]
\end{aligned} \quad (\text{D.103})$$

$$c'_2 = L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, (1 - \eta)\lambda_t, \mathbf{W}_t, \mathbf{V}_t) = c_2 + R(\sqrt{(1 - \eta)\lambda_t} \mathbf{W}_t, \sqrt{(1 - \eta)\lambda_t} \mathbf{V}_t) \quad (\text{D.104})$$

$$c_3 = \frac{1}{|\Omega^t|} \sum_{i=1}^{|\Omega^t|} \mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho}[L(\lambda_t F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_t\Sigma', \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma'\mathbf{V}_t), y_i)] \quad (\text{D.105})$$

$$c'_3 = L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) = c_3 + R(\sqrt{\lambda_t} \mathbf{W}_t, \sqrt{\lambda} \mathbf{V}_t) \quad (\text{D.106})$$

Then, following from C.38 to C.42 in [58], we have

$$c'_1 \leq (1 - \eta)(2c'_3 - c'_2) + \frac{\eta\gamma}{4}c'_3 + \eta(OPT + O(\|\mathbf{q}^\top \mathbf{A}^*\|_1 \|\mathbf{A}^*\|_\infty^4 \epsilon_0 / \gamma)) + O_p(\eta^{1.5}), \quad (\text{D.107})$$

which implies

$$\min\{c'_1, c'_2\} \leq (1 - \eta\frac{1}{2} + \frac{\eta\gamma}{8})c'_3 + \eta\frac{1}{2}OPT + O(\|\mathbf{q}^\top \mathbf{A}^*\|_1 \|\mathbf{A}^*\|_\infty^2 \eta \epsilon_0 / \gamma) + O_p(\eta^{1.5})$$

Note that the equation C.35 of [58], i.e, (D.96) in this work, is modified as

$$c'_1 - c_1 \leq (1 + \frac{\eta\gamma}{4})(c'_3 - c_3) + O(\eta \epsilon_0 / \gamma) \quad (\text{D.108})$$

if  $\gamma < \epsilon_0/\Omega(1)$ . As long as  $c'_3 \geq (1 + \gamma)OPT + \Omega(\|\mathbf{q}^\top \mathbf{A}^*\|_1 \|\mathbf{A}^*\|_\infty^2)$ , we have

$$\min\{c'_1, c'_2\} \leq (1 - \eta\frac{\gamma}{4})c'_3$$

**Lemma D.2.14.** *Note that the three sampled aggregation matrices in a three-layer learner network can be different. We denote them as  $\mathbf{A}^{t(1)}$ ,  $\mathbf{A}^{t(2)}$  and  $\mathbf{A}^{t(3)}$ . Let  $\mathbf{W}_t$ ,  $\mathbf{V}_t$  be the updated weights trained using  $\mathbf{A}^*$  and let  $\mathbf{W}'_t$ ,  $\mathbf{V}'_t$  be the updated weights trained using  $\mathbf{A}^{t(i)}$ ,  $i \in [3]$ . With probability at least 99/100, the algorithm converges in  $TT_w = \text{poly}(m_1, m_2)$  iterations to a point with  $\eta \in (0, \frac{1}{\text{poly}(m_1, m_2, \|\mathbf{A}^*\|_\infty, K)})$*

$$L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \leq (1 + \gamma)OPT + \epsilon_0$$

If

$$\begin{aligned} & L'(\mathbf{A}^{t(1)}, \mathbf{A}^{t(2)}, \mathbf{A}^{t(3)}, \lambda_t, \mathbf{W}'_t, \mathbf{V}'_t) \\ &= \frac{1}{|\Omega^t|} \sum_{i=1}^{|\Omega^t|} \mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho, \Sigma} [L(\lambda_t F_{\mathbf{A}^{t(1)}, \mathbf{A}^{t(2)}, \mathbf{A}^{t(3)}}(\mathbf{q}, \mathbf{X}_i, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}'_t \Sigma', \mathbf{V}^{(0)} + \mathbf{V}^\rho \\ & \quad + \Sigma' \mathbf{V}'_t), y_i)] + R(\sqrt{\lambda_t} \mathbf{W}'_t, \sqrt{\lambda_t} \mathbf{V}'_t), \end{aligned} \quad (\text{D.109})$$

where

$$F_{\mathbf{A}^{t(1)}, \mathbf{A}^{t(2)}, \mathbf{A}^{t(3)}}(\mathbf{q}, \mathbf{X}, \mathbf{W}, \mathbf{V}) = \mathbf{q}^\top \mathbf{A}^{t(3)} \sigma(\mathbf{A}^{t(2)} \sigma(\mathbf{A}^{t(1)} \mathbf{X} \mathbf{W} + \mathbf{B}_1) \mathbf{V} + \mathbf{B}_2) \mathbf{C}, \quad (\text{D.110})$$

we also have

$$\begin{aligned}
& L'(\mathbf{A}^{t(1)}, \mathbf{A}^{t(2)}, \mathbf{A}^{t(3)}, \lambda_{T-1}, \mathbf{W}'_T, \mathbf{V}'_T) \\
& \leq L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_T, \mathbf{W}_T, \mathbf{V}_T) + \lambda_{T-1} \cdot O(\text{poly}(\epsilon)) \\
& \leq (1 + \gamma)OPT + \epsilon_0
\end{aligned} \tag{D.111}$$

**Proof:**

By Lemma D.2.13, we know that as long as  $L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \in [(1 + \gamma)OPT + \Omega(\mathbf{q}^\top \mathbf{A}^* \mathbf{1} \|\mathbf{A}^*\|_\infty^4 \epsilon_0 / \gamma), \tilde{O}(1)]$ , then there exists  $\|\widehat{\mathbf{W}}\|_F \leq 1, \|\widehat{\mathbf{V}}\|_F \leq 1$  such that either

$$\begin{aligned}
& \mathbb{E}_{\Sigma, \Sigma'} [L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t \Sigma' + \sqrt{\eta} \widehat{\mathbf{W}} \Sigma \Sigma', \Sigma' \mathbf{V}_t + \sqrt{\eta} \Sigma' \Sigma \widehat{\mathbf{V}})] \\
& \leq (1 - \eta\gamma/4) L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t)
\end{aligned} \tag{D.112}$$

or

$$L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, (1 - \eta)\lambda_t, \mathbf{W}_t, \mathbf{V}_t) \leq (1 - \eta\gamma/4) L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \tag{D.113}$$

Denote  $\mathbf{W} = \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_t \Sigma' + \sqrt{\eta} \widehat{\mathbf{W}} \Sigma \Sigma', \mathbf{V} = \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma' \mathbf{V}_t + \sqrt{\eta} \Sigma' \Sigma \widehat{\mathbf{V}}$ . Note that

$$\frac{\partial L}{\partial \mathbf{w}_j} = \sum_{i=1}^K \frac{\partial L}{\partial f_i} \frac{\partial f_i}{\partial \mathbf{w}_j} \tag{D.114}$$

$$\begin{aligned}
& \frac{\partial}{\partial \mathbf{w}_j} f_r(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_t \Sigma' + \sqrt{\eta} \widehat{\mathbf{W}} \Sigma \Sigma', \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma' \mathbf{V}_t \\
& + \sqrt{\eta} \Sigma' \Sigma \widehat{\mathbf{V}}) \\
& = \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n \sum_{i=1}^{m_2} c_{i,r} \mathbb{1}_{r_{n,i} + B_{2(n,i)} \geq 0} \sum_{k=1}^N a_{n,k} v_{j,i} \mathbb{1}_{\tilde{\mathbf{a}}^* n \mathbf{X} \mathbf{w}_k + B_{1(n,k)} \geq 0} (\tilde{\mathbf{a}}^* n \mathbf{X})^\top,
\end{aligned} \tag{D.115}$$

which implies  $\frac{\partial F}{\partial \mathbf{w}_t}, \frac{\partial^2 F}{\partial \mathbf{w}_t^2}, \frac{\partial^3 F}{\partial \mathbf{w}_t^3}$  are summations of  $\mathbb{1}, \delta, \delta'$  functions and their multiplications. It can be found that no  $\delta(x)\delta'(x)$ ,  $\delta(x)^2$  or  $\delta'^2(x)$  exist in these terms. Therefore, by  $\int_{-\infty}^{\infty} \delta(t)f(t)dt = f(0)$  and  $\int_{-\infty}^{\infty} \delta'(t)f(t)dt = -f'(0)$ , we can obtain that the value of the third-order derivative w.r.t.  $\mathbf{W}^\rho$  of  $\mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho, \Sigma} [L(\lambda_t F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_t \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_t), y)]$  is proportional to  $\text{poly}(\|\mathbf{A}^*\|_\infty, K)$ , some certain value of the probability density function of  $\mathbf{W}^\rho$  and its derivative, i.e.,  $\text{poly}(\sigma_w^{-1})$ . Similarly, the value of the third-order derivative w.r.t.  $\mathbf{W}^\rho$  of  $\mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho, \Sigma} [L(\lambda_t F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_t \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_t), y)]$  is polynomially depend on  $\sigma_v^{-1}$  and  $\|\mathbf{A}^*\|_\infty$ . By the value selection of  $\sigma_w$  and  $\sigma_v$ , we can conclude that  $L'$  is  $B = \text{poly}(m_1, m_2, \|\mathbf{A}^*\|_\infty, K)$  second-order smooth.

By Fact A.8 in [58], it satisfies with  $\eta \in (0, \frac{1}{\text{poly}(m_1, m_2, \|\mathbf{A}^*\|_\infty, K)})$

$$\lambda_{\min}(\nabla^2 L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_{t-1}, \mathbf{W}_t, \mathbf{V}_t)) < -\frac{1}{(m_1 m_2)^8} \quad (\text{D.116})$$

Meanwhile, for  $t \geq 1$ , by the escape saddle point theorem of Lemma A.9 in [58], we know with probability at least  $1 - p$ ,  $\lambda_{\min}(\nabla^2 L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_{t-1}, \mathbf{W}_t, \mathbf{V}_t)) > -\frac{1}{(m_1 m_2)^8}$  holds. Choosing  $p = \frac{1}{100T}$ , then this holds for  $t = 1, 2, \dots, T$  with probability at least 0.999. Therefore, for  $t = 1, 2, \dots, T$ , the first case cannot happen, i.e., as long as  $L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \geq (1 + \gamma)OPT + \Omega(\mathbf{q}^\top \mathbf{A}^* \mathbf{1} \|\mathbf{A}^*\|_\infty^4 \epsilon_0 / \gamma)$ ,

$$L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, (1 - \eta)\lambda_t, \mathbf{W}_t, \mathbf{V}_t) \leq (1 - \eta\gamma/4)L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \quad (\text{D.117})$$

On the other hand, for  $t = 1, 2, \dots, T - 1$ , as long as  $L' \leq \tilde{O}(1)$ , by Lemma A.9 in [58], we have

$$L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_{t+1}, \mathbf{V}_{t+1}) \leq L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) + (m_1 m_2)^{-1} \quad (\text{D.118})$$

By  $L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_1, \mathbf{W}_0, \mathbf{V}_0) \leq \tilde{O}(1)$  with high probability, we have  $L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \leq \tilde{O}(1)$  with high probability for  $t = 1, 2, \dots, T$ . Therefore, after  $T = \tilde{\Theta}(\eta^{-1} \cdot \log \frac{\log m}{\epsilon_0})$  rounds of weight decay, we have  $L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \leq (1 + \gamma)OPT + \Omega(\mathbf{q}^\top \mathbf{A}^* \mathbf{1} \|\mathbf{A}^*\|_\infty^4 \epsilon_0 / \gamma)$ . Rescale down  $\epsilon_0$  and we can obtain our final result.

Consider  $L'(\mathbf{A}^{t(1)}, \mathbf{A}^{t(2)}, \mathbf{A}^{t(3)}, \lambda_t, \mathbf{W}'_t, \mathbf{V}'_t)$ . Let  $\mathbf{w}_i, \mathbf{v}_i$  be the output weights updated with all the aggregation matrices equal to  $\mathbf{A}^*$ , and let  $\mathbf{w}'_i, \mathbf{v}'_i$  be the output weights updated with our sampling strategy in Section 5.3.2. We know that

$$\begin{aligned} \|\mathbf{w}_i - \mathbf{w}'_i\| &\lesssim \sum_{t=0}^{T-1} \left\| \eta \sum_{l=0}^{T_w-1} \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} \right. \\ &\quad \left. \mathbb{1}[\mathbf{a}^*_n \sigma(\mathbf{A}^* \mathbf{X} \mathbf{W}) \mathbf{v}_i \geq 0] \sum_{k=1}^N a_{n,k}^* v_{j,i} \mathbb{1}[\mathbf{a}^*_k \mathbf{X} \mathbf{w}_j \geq 0] (\mathbf{a}^*_k - \mathbf{a}_k^{t(1)}) \mathbf{X} \right\| \quad (\text{D.119}) \\ &\leq \frac{1}{\text{poly}(m_1, m_2)} \cdot \frac{1}{\text{poly}(\epsilon)} \text{poly}(m_1, m_2) \epsilon_c \|\mathbf{A}^*\|_\infty \cdot \text{poly}(\epsilon) = O(\epsilon) \end{aligned}$$

$$\begin{aligned}
\|\mathbf{v}_i - \mathbf{v}'_i\| &\lesssim \sum_{t=0}^{T-1} \left\| \eta \sum_{l=0}^{T_w-1} \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} \mathbb{1}[\mathbf{a}^* \mathbf{a}_n \sigma(\mathbf{A}^* \mathbf{X} \mathbf{W}) \mathbf{v}_i \geq 0] (\mathbf{a}^* \mathbf{a}_n \sigma(\mathbf{A}^* \mathbf{X} \mathbf{W}) \right. \\
&\quad \left. - \mathbf{a}_n^{t(2)} \sigma(\mathbf{A}^{t(1)} \mathbf{X} \mathbf{W}') ) \right\| \\
&\leq \frac{1}{\text{poly}(m_1, m_2)} \cdot \frac{1}{\text{poly}(\epsilon)} \text{poly}(m_1, m_2) \epsilon_c \|\mathbf{A}^*\|_\infty \cdot \text{poly}(\epsilon) = O(\epsilon)
\end{aligned} \tag{D.120}$$

With a slight abuse of notation, for  $r \in [K]$ , we denote

$$f_r(q, A^{t(1)}, A^{t(2)}, A^{t(3)}, X, W'_t, V'_t) = q^\top A^{t(3)} \sigma(A^{t(2)} \sigma(A^{t(1)} X W + B_1) V + B_2) c_r \quad (\text{D.121})$$

The difference between  $f_r(\mathbf{q}, \mathbf{A}^*, \mathbf{X}, \mathbf{W}_t, \mathbf{V}_t)$  and  $f_r(\mathbf{q}, \mathbf{A}^{t(1)}, \mathbf{A}^{t(2)}, \mathbf{A}^{t(3)}, \mathbf{X}, \mathbf{W}'_t, \mathbf{V}'_t)$  is caused by  $\|\mathbf{A}^* - \mathbf{A}^{t(1)}\|_\infty$ ,  $\|\mathbf{A}^* - \mathbf{A}^{t(2)}\|_\infty$ ,  $\|\mathbf{A}^* - \mathbf{A}^{t(3)}\|_\infty$ ,  $\mathbf{w}_i^{(t)} - \mathbf{w}_i^{(t)'}^\top$  and  $\mathbf{v}_i^{(t)} - \mathbf{v}_i^{(t)'}^\top$ . Following the proof in Lemma D.1.2, we can easily obtain that if  $|p_l - p_l^*| \leq p_l^* \cdot O(\text{poly}(\epsilon))$  and  $l_i \geq |\mathcal{N}_i| / (1 + \frac{c_1 \cdot \text{poly}(\epsilon)}{p_l^* L \Phi(L, i)})$ , it can be derived that  $\|\mathbf{A}^* - \mathbf{A}^{(1)}\|_\infty \leq O(\text{poly}(\epsilon))$ ,  $\|\mathbf{A}^* - \mathbf{A}^{(2)}\|_\infty \leq O(\text{poly}(\epsilon))$  and  $\|\mathbf{A}^* - \mathbf{A}^{(3)}\|_\infty \leq O(\text{poly}(\epsilon))$ . Then, by (D.119) and (D.120), we have

$$\begin{aligned}
& |\mathbf{q}^\top \mathbf{A}^* \sigma(\mathbf{A}^* \sigma(\mathbf{A}^* \mathbf{XW}) \mathbf{V}) \mathbf{c}_r - \mathbf{q}^\top \mathbf{A}^* \sigma(\mathbf{A}^{(2)} \sigma(\mathbf{A}^{(1)} \mathbf{XW}') \mathbf{V}') \mathbf{c}_r| \\
& \leq \left| \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} | \sigma(\mathbf{a}_n^* \sigma(\mathbf{A}^* \mathbf{XW}) \mathbf{v}_i) - \sigma(\mathbf{a}_n^{(2)} \sigma(\mathbf{A}^{(1)} \mathbf{XW}') \mathbf{v}'_i) | \right| \\
& \leq \left| \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} | \mathbf{a}_n^* \sigma(\mathbf{A}^* \mathbf{XW}) \mathbf{v}_i - \mathbf{a}_n^{(2)} \sigma(\mathbf{A}^{(1)} \mathbf{XW}') \mathbf{v}'_i | \right| \\
& \leq \left| \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} | (\mathbf{a}_n^* - \mathbf{a}_n^{(2)}) \sigma(\mathbf{A}^* \mathbf{XW}) \mathbf{v}_i | \right| + \left| \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} | \mathbf{a}_n^{(2)} (\sigma(\mathbf{A}^* \mathbf{XW}) \mathbf{v}_i - \sigma(\mathbf{A}^{(1)} \mathbf{XW}') \mathbf{v}'_i) | \right| \\
& \leq \left| \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} | (\mathbf{a}_n^* - \mathbf{a}_n^{(2)}) \sigma(\mathbf{A}^* \mathbf{XW}) \mathbf{v}_i | \right| + \left| \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} | \cdot \mathbf{a}_n^{(2)} ((\sigma(\mathbf{A}^* \mathbf{XW}) - \sigma(\mathbf{A}^{(1)} \mathbf{XW}')) \mathbf{v}_i + \sigma(\mathbf{A}^{(1)} \mathbf{XW}') (\mathbf{v}_i - \mathbf{v}'_i)) | \right| \\
& \leq \left| \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} | (\mathbf{a}_n^* - \mathbf{a}_n^{(2)}) \sigma(\mathbf{A}^* \mathbf{XW}) \mathbf{v}_i | \right| + \left| \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} | \cdot \mathbf{a}_n^{(2)} \sigma(\mathbf{A}^{(1)} \mathbf{XW}') \cdot (\mathbf{v}_i - \mathbf{v}'_i) | \right| + \left| \sum_{n=1}^N \mathbf{q}^\top \mathbf{a}_n^* \sum_{i=1}^{m_2} c_{i,r} | \sum_{k=1}^N a''_{n,k} \sum_{l=1}^{m_1} v_{i,l} | \right| \\
& \quad \cdot | (\mathbf{a}_k^* - \mathbf{a}'_k) \mathbf{Xw}_l + \mathbf{a}'_k \mathbf{X} (\mathbf{w}_l - \mathbf{w}'_l) | \Big| \\
& \leq O(\text{poly}(\epsilon)).
\end{aligned} \tag{D.122}$$

Hence,

$$\begin{aligned}
& |\mathbf{e}_g^\top \mathbf{A}^* \sigma(\mathbf{A}^* \sigma(\mathbf{A}^* \mathbf{X} \mathbf{W}) \mathbf{V}) \mathbf{c}_r - \mathbf{e}_g^\top \mathbf{A}^{(3)} \sigma(\mathbf{A}^{(2)} \sigma(\mathbf{A}^{(1)} \mathbf{X} \mathbf{W}') \mathbf{V}') \mathbf{c}_r| \\
& \leq |\mathbf{e}_g^\top \mathbf{A}^* \sigma(\mathbf{A}^* \sigma(\mathbf{A}^* \mathbf{X} \mathbf{W}) \mathbf{V}) \mathbf{c}_r - \mathbf{e}_g^\top \mathbf{A}^* \sigma(\mathbf{A}^{(2)} \sigma(\mathbf{A}^{(1)} \mathbf{X} \mathbf{W}') \mathbf{V}') \mathbf{c}_r| \\
& \quad + |\mathbf{e}_g^\top (\mathbf{A}^* - \mathbf{A}^{(3)}) \sigma(\mathbf{A}^{(2)} \sigma(\mathbf{A}^{(1)} \mathbf{X} \mathbf{W}) \mathbf{V}) \mathbf{c}_r| \\
& \leq O(\text{poly}(\epsilon)).
\end{aligned} \tag{D.123}$$

which implies

$$\begin{aligned}
& L'(\mathbf{A}^{t(1)}, \mathbf{A}^{t(2)}, \mathbf{A}^{t(3)}, \lambda_{T-1}, \mathbf{W}'_T, \mathbf{V}'_T) \\
& \leq L'(\mathbf{A}^*, \mathbf{A}^*, \mathbf{A}^*, \lambda_T, \mathbf{W}_T, \mathbf{V}_T) + \lambda_{T-1} \cdot O(\text{poly}(\epsilon)) \\
& \leq (1 + \gamma) OPT + \epsilon_0
\end{aligned} \tag{D.124}$$

### Proof of Theorem 5.3.1:

By Lemma D.2.14, we have that the algorithm converges in  $TT_w$  iterations to a point

$$L'(\mathbf{A}^{t(1)}, \mathbf{A}^{t(2)}, \mathbf{A}^{t(3)}, \lambda_t, \mathbf{W}_t, \mathbf{V}_t) \leq (1 + \gamma) OPT + \epsilon_0$$

We know w.h.p., among  $\tilde{O}(1/\epsilon_0^2)$  choices of  $j$ ,

$$\begin{aligned}
& \min_j \{ \mathbb{E}_{\mathbf{W}^\rho, \mathbf{V}^\rho, \Sigma, z \in \Omega} L(\lambda_{T-1} F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^{\rho,j} + \mathbf{W}_T \Sigma, \\
& \mathbf{V}^{(0)} + \mathbf{V}^{\rho,j} + \Sigma \mathbf{V}_T) \} \\
& \leq (1 + \gamma) OPT + \epsilon_0
\end{aligned} \tag{D.125}$$

Then we have

$$\|\mathbf{W}_T\|_{2,4} \leq \epsilon_0^{\frac{1}{4}} \tau'_w \tag{D.126}$$

$$\|\mathbf{V}_T\|_F \leq \epsilon_0^{\frac{1}{2}} \tau'_v \tag{D.127}$$

By Lemma D.2.9, we know that

$$\begin{aligned}
& f_r(\mathbf{e}_g, \mathbf{A}^*, \mathbf{X}_i, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T, \mathbf{B}) \\
& = f_r(\mathbf{e}_g, \mathbf{A}^*, \mathbf{X}_i, \mathbf{W}^{(0)} + \mathbf{W}^\rho, \mathbf{V}^{(0)} + \mathbf{V}^\rho, \mathbf{B}) + g_r^{(b,b)}(\mathbf{e}_g, \mathbf{A}^*, \mathbf{X}_i, \mathbf{W}_T, \mathbf{V}_T, \mathbf{B}) \pm \frac{\epsilon}{K}
\end{aligned} \tag{D.128}$$

Denote  $\mathbf{r}' = \mathbf{A}^* \sigma(\mathbf{A}^* \mathbf{X}(\mathbf{W}^{(0)} + \mathbf{W}^\rho) + \mathbf{B}_1)(\mathbf{V}^{(0)} + \mathbf{V}^\rho)$ . Then,

$$\|\mathbf{r}'\| \leq \|\mathbf{A}^*\| (\|\mathbf{A}^*\|_\infty \cdot \tilde{O}(1)) \cdot \tilde{O}(1) \leq \|\mathbf{A}^*\|_\infty \quad (\text{D.129})$$

Therefore,

$$\begin{aligned} & |f_r(\mathbf{e}_g, \mathbf{A}^*, \mathbf{X}_i, \mathbf{W}^{(0)} + \mathbf{W}^\rho, \mathbf{V}^{(0)} + \mathbf{V}^\rho, \mathbf{B})| \\ &= |\mathbf{e}_g^\top \mathbf{A}^* \sigma(\mathbf{r}' + \mathbf{B}_2) \mathbf{c}_r| \\ &\leq \tilde{O}(\|\mathbf{A}^*\|_\infty (\|\mathbf{A}^*\|_\infty + 1) \epsilon_c) \end{aligned} \quad (\text{D.130})$$

We also have

$$\begin{aligned} & |g_r^{(b,b)}(\mathbf{e}_g, \mathbf{A}^*, \mathbf{X}_i, \mathbf{W}_T, \mathbf{V}_T, \mathbf{B})| \\ &\leq |\mathbf{e}_g^\top \mathbf{A}^* \mathbf{A}^* (\mathbf{A}^* \mathbf{X} \mathbf{W}_T \odot \mathbf{D}_{w,x}^{(0)} \mathbf{V}_T) \odot \mathbf{D}_{v,x}^{(0)} \mathbf{c}_r| \\ &\leq \|\mathbf{A}^*\|_\infty^2 \tau'_v \tau'_w m_1^{\frac{1}{4}} \sqrt{m_2} \epsilon_c \\ &\leq C_0 \|\mathbf{A}^*\|_\infty^2 \end{aligned} \quad (\text{D.131})$$

Hence,

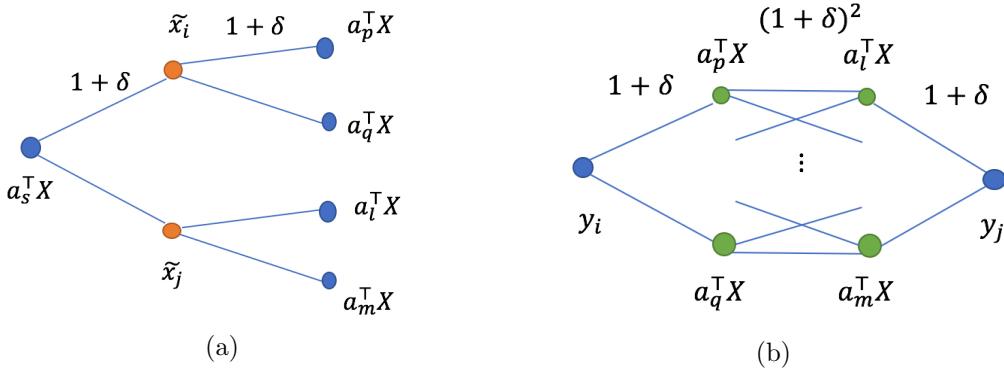
$$f_r(\mathbf{e}_g, \mathbf{A}^*, \mathbf{X}_i, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T, \mathbf{B}) \leq \tilde{O}(\|\mathbf{A}^*\|_\infty^2 (\epsilon_c + C_0)) \quad (\text{D.132})$$

Combining (D.121, D.123), we can obtain

$$\begin{aligned} & f_r(\mathbf{e}_g, \mathbf{A}^{t(1)}, \mathbf{A}^{t(2)}, \mathbf{A}^{t(3)}, \mathbf{X}_i, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T, \mathbf{B}) \\ &\leq \tilde{O}(\|\mathbf{A}^*\|_\infty^2 (\epsilon_c + C_0)) \end{aligned} \quad (\text{D.133})$$

as long as  $\|\mathbf{A}^* - \mathbf{A}^{t(1)}\|_\infty \leq \text{poly}(\epsilon)$ ,  $\|\mathbf{A}^* - \mathbf{A}^{t(2)}\|_\infty \leq \text{poly}(\epsilon)$  and  $\|\mathbf{A}^* - \mathbf{A}^{t(3)}\|_\infty \leq \text{poly}(\epsilon)$ . For any given  $\{\mathbf{X}_i, y_i\}_{i=1}^{|\Omega|}$ , the dependency between  $y_i, y_j$ , where  $i, j \in |\Omega|$  can be considered in two steps. Figure D.1(a) shows  $\mathbf{a}_i \mathbf{X}$  is dependent with at most  $(1 + \delta)^2 \mathbf{a}_j \mathbf{X}'s$ . This is because each  $\mathbf{a}_i \mathbf{X}$  is determined by at most  $(1 + \delta)$  row vector  $\tilde{\mathbf{x}}_l's$ , while each  $\tilde{\mathbf{x}}_l$  is contained by at most  $(1 + \delta)$   $\mathbf{a}_p \mathbf{X}'s$ . Similarly,  $y_i$  is determined by at most  $(1 + \delta) \mathbf{a}_p \mathbf{X}'s$  and by Figure D.1(b) we can find  $y_i$  is dependent with at most  $(1 + \delta)^4 y_j$  (including  $y_i$ ). Since the matrix  $\mathbf{A}^*$  shares the same non-zero entries with  $\mathbf{A}$ , the output with  $\mathbf{A}^*$  indicates the same dependence.

Denote  $u_i = 1/|\Omega^t| \sum_{i=1}^{|\Omega^t|} |L(\lambda_{T-1} F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T), y_i)| - \mathbb{E}_{(\mathbf{e}_g, \mathbf{X}, y) \in \mathcal{D}} [L(\lambda_{T-1} F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T), y_i)]$ . Then,  $\mathbb{E}[u_i] = 0$ .



**Figure D.1:** (a) Dependency between  $a_s^T X$  and  $a_p^T X$  (b) Dependency between  $y_i$  and  $y_j$ .

Since that  $L$  is 1-lipschitz smooth and  $L(\mathbf{0}^K, y) \in [0, 1]$ , we have

$$\begin{aligned} & |L(\lambda_{T-1} F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T), y_i) - L(\mathbf{0}^K, y_i)| \\ & \leq \|(\lambda_{T-1} F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T), y_i) - (\mathbf{0}^K, y_i)\| \quad (\text{D.134}) \\ & \leq \tilde{O}(\sqrt{K} \|\mathbf{A}^*\|_\infty^2 (\epsilon_c + C_0)) \end{aligned}$$

Then,

$$\begin{aligned} |u_i| & \leq 2\sqrt{K} \|\mathbf{A}^*\|_\infty^2 (\epsilon_c + C_0) \\ \mathbb{P}(|u_i| \geq t) & \leq 1 \leq \exp\left(1 - \frac{t^2}{4K \|\mathbf{A}^*\|_\infty^4 (\epsilon_c + C_0)^2}\right) \quad (\text{D.135}) \end{aligned}$$

Then,  $u_i$  is a sub-Gaussian random variable. We have  $\mathbb{E} e^{su_i} \leq e^{\|\mathbf{A}^*\|_\infty^4 (\epsilon_c + C_0)^2 s^2}$ . By Lemma 7 in [74], we have

$$\mathbb{E} e^{s \sum_{i=1}^{|\Omega|} u_i} \leq e^{(1+\delta)^4 K \|\mathbf{A}^*\|_\infty^4 (\epsilon_c + C_0)^2 |\Omega| s^2}$$

Therefore,

$$\mathbb{P}\left(\left|\sum_{i=1}^{|\Omega|} \frac{1}{|\Omega|} u_i\right| \geq k\right) \leq \exp(\|\mathbf{A}^*\|_\infty^4 (\epsilon_c + C_0)^2 K (1 + \delta)^4 |\Omega| s^2 - |\Omega| ks) \quad (\text{D.136})$$

for any  $s > 0$ . Let  $s = \frac{k}{2\|\mathbf{A}^*\|_\infty^4 (\epsilon_c + C_0)^2 K (1 + \delta)^4}$ ,  $k = \|\mathbf{A}^*\|_\infty^4 (\epsilon_c + C_0)^2 K \sqrt{\frac{(1 + \delta)^4 \log N}{|\Omega|}}$ , we can obtain

$$\mathbb{P}\left(\left|\sum_{i=1}^{|\Omega|} \frac{1}{|\Omega|} u_i\right| \geq k\right) \leq \exp(-\|\mathbf{A}^*\|_\infty^4 (\epsilon_c + C_0)^2 K \log N) \leq N^{-K} \quad (\text{D.137})$$

Therefore, with probability at least  $1 - N^{-K}$ , we have

$$\begin{aligned} & \left| \mathbb{E}_{(\mathbf{e}_g, \mathbf{X}, y) \sim \mathcal{D}} [L(\lambda_{T-1} F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T), y_i)] \right. \\ & \left. - \frac{1}{|\Omega^t|} \sum_{i=1}^{|\Omega^t|} L(\lambda_{T-1} F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T), y_i) \right| \quad (\text{D.138}) \\ & \leq \epsilon_0 \end{aligned}$$

as long as  $|\Omega| \geq \tilde{\Theta}(\epsilon_0^{-2} \|\mathbf{A}^*\|_\infty^8 (1 + p_1^4 p_2^5 \mathcal{C}_\epsilon(\phi, \|\mathbf{A}^*\|_\infty) \mathcal{C}_\epsilon(\Phi, \sqrt{p_2} \mathcal{C}_\epsilon(\phi, \|\mathbf{A}^*\|_\infty))) (\|\mathbf{A}^*\|_\infty + 1)^4 K^6 (1 + \delta)^4 \log N)$ , i.e.,

$$\begin{aligned} & \mathbb{E}_{(\mathbf{e}_g, \mathbf{X}, y) \sim \mathcal{D}} [L(\lambda_{T-1} F_{\mathbf{A}^*}(\mathbf{e}_g, \mathbf{X}, \mathbf{W}^{(0)} + \mathbf{W}^\rho + \mathbf{W}_T \Sigma, \mathbf{V}^{(0)} + \mathbf{V}^\rho + \Sigma \mathbf{V}_T), y_i)] \\ & \leq (1 + \gamma) OPT + \epsilon_0 \quad (\text{D.139}) \\ & \leq (1 + \epsilon_0) OPT + \epsilon_0 \end{aligned}$$

## APPENDIX E

### APPENDIX OF CHAPTER 6

#### E.1 Proof Sketch

We partially include the proof backbone in Section 6.4 when introducing the mechanism of the trained Transformer. We elaborate more about our proof intuition of Theorem 6.3.3 in the following.

We first briefly introduce our proof intuition of Theorem 6.3.3. We first respectively build Lemmas E.4.5, E.4.6, and E.4.7 to characterize gradient updates for  $\mathbf{W}_Q$  and  $\mathbf{W}_K$ ,  $\mathbf{W}_V$  and  $\mathbf{W}_O$ . These Lemmas are based on an observation that a constant fraction of neurons in  $\mathbf{W}_O$  can always be activated (Lemma E.4.9, E.4.10) to avoid the non-smoothness of Relu activation. The orthogonality of patterns and Definition 6.3.1 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 6.3.3, we use Lemma E.4.5 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 E.4.6 and E.4.7. 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  $\mathbf{W}_K$ ,  $\mathbf{W}_Q$ . We also want to clarify that how we can handle the training dynamics related to the softmax is different from [219] although we both derive a sparse attention distribution. Inspired by the intuition of feature-learning analyses for two-layer Relu networks [88], [86], [89], [113], we initialize the model such that at least a constant fraction of the neurons of  $\mathbf{W}_O$  are activated (Lemma E.4.9), which are called lucky neurons as in [89], [6]. We prove that these lucky neurons are always activated (Lemma E.4.10) and grow with an increasing magnitude and two fixed directions of the label embedding along the training (Lemma E.4.7). Then, we can show that the gradient growths of  $\mathbf{W}_Q$  and  $\mathbf{W}_K$  can be lower bounded by contributions from these lucky neurons. Therefore, we are able to characterize the gradient updates of  $\mathbf{W}_K$  and  $\mathbf{W}_Q$  given a dynamic  $\mathbf{W}_O$ . This process is different from [219] since [219] does not include Relu MLP, so there is no need to study lucky neurons. Besides, we use

---

Portions of this appendix previously appeared as: H. Li, M. Wang, S. Lu, X. Cui, and P.-Y. Chen, “How do nonlinear transformers learn and generalize in in-context learning?” in *Proc. Int. Conf. Mach. Learn.*, Jul. 2024, pp. 28734–28783.

Hinge loss, while their training loss is logistic loss, which leads to more training phases, as a difference in the training dynamics between us.

## E.2 Addition Discussions and Extensions

### E.2.1 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 [1] 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 [223] or conducting dynamic programming tasks [241]. However, state-of-the-art theoretical works [217], [219], [218] 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.

### E.2.2 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 [217], [219], [218] only consider single-head attention in the network. Some concurrent works consider multi-head attention, but they either do not study ICL [253], [254] or do not involve convergence/generalization guarantee [245]. 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 6.3 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.,

$$\left\| \sum_{h=1}^H \sum_{i=1}^l \mathbf{W}_{V_h} \mathbf{p}_i \cdot \text{softmax}(\mathbf{p}_i^\top \mathbf{W}_{K_h}^\top \mathbf{W}_{Q_h} \mathbf{p}_{query}) \right\|. \quad (\text{E.1})$$

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 6.4.2. We will leave the analysis on multi-head attention with more general settings as future directions.

### E.2.3 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.

### E.2.4 Additional Related Works

We introduce other existing theoretical works on learning and generalization of neural networks in this section. Some works [71], [72], [76], [260], [25] 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 [77], [58], [78], [80], [81],

[83], [22] considers strongly overparameterized networks to linearize the neural network around the initialization. The generalization performance is independent of the feature distribution. [85], [86], [87], [88], [89], [6], [111], [269] 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.

### E.3 Additional Experiments and the Algorithm

We first present the training algorithm introduced in Section 6.2.3.

---

#### Subroutine 2 Training with Stochastic Gradient Descent (SGD)

---

- 1: **Hyperparameters:** The step size  $\eta$ , the number of iterations  $T$ , batch size  $B$ .
- 2: **Initialization:** Each entry of  $\mathbf{W}_O^{(0)}$  and  $\mathbf{a}^{(0)}$  from  $\mathcal{N}(0, \xi^2)$  and Uniform( $\{+1/\sqrt{m}, -1/\sqrt{m}\}$ ), respectively.  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$  and  $\mathbf{W}_V$  are initialized such that all diagonal entries of  $\mathbf{W}_V^{(0)}$ , and the first  $d_X$  diagonal entries of  $\mathbf{W}_Q^{(0)}$  and  $\mathbf{W}_K^{(0)}$  are set as  $\delta$  with  $\delta \in (0, 0.2]$ .
- 3: **Training by SGD:** For each iteration, we independently sample  $\mathbf{x}_{query} \sim \mathcal{D}$ ,  $f \in \mathcal{T}_{tr}$  to form a batch of training prompt and labels  $\{\mathbf{P}^n, z^n\}_{n \in \mathcal{B}_t}$  as introduced in Section 6.3.2. Each IDR pattern is sampled equally likely in each batch. For each  $t = 0, 1, \dots, T-1$  and  $\mathbf{W}^{(t)} \in \Psi^{(t)}$

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_t} \nabla_{\mathbf{W}^{(t)}} \ell(\Psi^{(t)}; \mathbf{P}^n, z^n). \quad (\text{E.2})$$

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

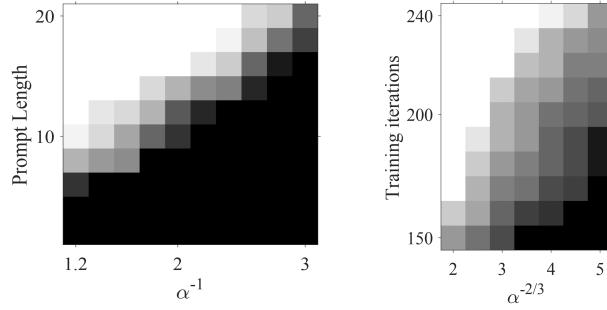
Then, we introduce additional experiments to verify our theory.

#### E.3.1 The Impact of ALPHA

We choose  $\alpha = 0.6$  and use a one-layer Transformer as in (6.2). Figure E.1 shows that the required length of the training prompt is linear in  $\alpha^{-1}$ , while the required number of training iterations is linear in  $\alpha^{-2/3}$ , which verify the theoretical findings in (6.9) and (6.10).

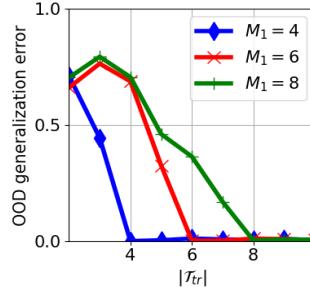
#### E.3.2 The Required Number of Training Tasks

We choose  $\alpha = 0.6$  and use a one-layer Transformer as in (6.2). For a given  $\mathcal{T}$ , we first generate a set of tasks that satisfies Condition 6.3.2 as follows. Define  $\mathbf{a}_i = \mathbf{a}_{i+M_1} = \boldsymbol{\mu}_i$  for  $i \in [M_1]$ , and then the  $j$ -th task function map the queries with  $\mathbf{a}_j$  and  $\mathbf{a}_{j+1}$  as IDR patterns



**Figure E.1:** The prompt length against  $\alpha$ , and the required number of training iterations against  $\alpha$ .

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  $\mathcal{T} \setminus \mathcal{T}_{tr0}$  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 E.2 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 6.3.2 can be met with a fraction of  $(M_1 - 1)^{-1/2}$  of total number of in-domain tasks.



**Figure E.2:** The required number of training tasks for different  $M_1$ .

## E.4 Proofs of the Main Theorems

We first provide several useful definitions and key lemmas for the proof of the main theorems. Tables E.1 and E.2 show a summary of notations used in the proof.

### E.4.1 Proof Overview of Main Theorems

This section illustrates how Corollary 6.4.3 and Proposition 6.4.5 contribute to the final in- and out-of-domain generalization performance of ICL.

**Table E.1: Summary of notations.**

Notations	Annotation
$\mathbf{x}_s^n, \mathbf{y}_s^n$	$\mathbf{x}_s^n$ is the data for classification. $\mathbf{y}_s^n$ is the embedding of the label for $\mathbf{x}_s^n$ .
$\mathbf{P}^n, z^n$	$\mathbf{P}^n$ is a prompt that consists of $l$ pairs of $\mathbf{x}_s^n$ and $\mathbf{y}_s^n$ , $s \in [l]$ . The last column of $\mathbf{P}^n$ contains $\mathbf{p}_{query}^n$ , which is the query of $\mathbf{P}^n$ . $z^n \in \{+1, -1\}$ is the binary label of $\mathbf{p}_{query}^n$ , which is also the label of $\mathbf{P}^n$ when we formulate the problem as a supervised learning problem.
$F(\Psi; \mathbf{P}^n), \ell(\Psi; \mathbf{P}^n, z^n)$	$F(\Psi; \mathbf{P}^n)$ is the Transformer output for $\mathbf{P}^n$ with $\Psi$ as the parameter. $\ell(\Psi; \mathbf{P}^n, z^n)$ is the loss function value given $\mathbf{P}^n$ and the corresponding label $z^n$ .
$\mathbf{p}_s^n, \boldsymbol{\mu}_j, \boldsymbol{\nu}_k$	$\mathbf{p}_s^n$ is the $s$ -th example with the corresponding label in $\mathbf{P}^n$ . If $s = query$ , $\mathbf{p}_s^n$ is the query. $\boldsymbol{\mu}_j$ and $\boldsymbol{\nu}_k$ are the IDR and IDI patterns in the feature embedding of $\mathbf{p}_s^n$ as the corresponding coefficients, respectively.
$\mathbf{q}$	$\mathbf{q}$ 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. $M = M_1 + M_2$ .
$\alpha, a$	$\alpha$ is the probability of selecting examples that contain either of the two decisive IDR patterns in each $\mathbf{P}^n$ . $a = 1/ a_i $ where $a_i$ is the entry of each neuron in $\mathbf{W}_O$ . $a = m$ .
$\kappa, \kappa', K, K', \beta$	$\kappa$ and $\kappa'$ are the coefficients of the IDI pattern and the ODI pattern in the input $\mathbf{x}$ , respectively. $\kappa$ and $\kappa'$ follow uniform distribution $U(-K, K)$ and $U(-K', K')$ with $K \leq 1/2$ and $K' \leq \mathcal{O}(1)$ , respectively. $\beta$ 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 $\mathbf{W}_O$ that can activate the terms inside $\text{Relu}(\cdot)$ in $F(\Psi; \mathbf{P}^n)$ for $z^n = +1$ at initialization. $\mathcal{U}_n$ is the set of neurons of $\mathbf{W}_O$ that can activate the Relu part of $F(\Psi; \mathbf{P}^n)$ for $z^n = -1$ at initialization.
$\mathcal{W}, \mathcal{U}$	$\mathcal{W} = \cup_{n \in [N]} \mathcal{W}_n$ . $\mathcal{U} = \cup_{n \in [N]} \mathcal{U}_n$

### The establishment of generalization

1. (*Self-Attention*) We can deduce from Corollary 6.4.3 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.,  $\sum_{i=1}^l \mathbf{p}_i \text{attn}(\Psi; \mathbf{P}, i)$ , is close to  $[\boldsymbol{\mu}_j^\top, \mathbf{q}^\top]^\top$ . This is because as long as the training/testing prompt length satisfies (6.9), large attention weights are assigned on  $\mathbf{p}_i$  of which the IDR pattern is  $\boldsymbol{\mu}_j$ , and the label embedding is  $\mathbf{q}$  by (6.20). Similarly, if its

**Table E.2: Summary of notations (Continued).**

Notations	Annotation
$\mathcal{N}_j^n$	The set of examples in $\mathbf{P}^n$ that contains $\boldsymbol{\mu}_j$ as the IDR pattern.
$\gamma_t$	$\gamma_t$ is the summation of attention weight on examples that have different IDR patterns from the query.
$\zeta_t$	$\zeta_t$ is smallest positive value inside the $\text{Relu}(\cdot)$ in $F(\Psi; \mathbf{P}^n)$ for all the $\mathbf{W}_O$ 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 $f(x) = \mathcal{O}(g(x))$ (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 $f(x) \lesssim \mathcal{O}(g(x))$ ).

label is  $-1$ , the weighted summation of contexts and the query outputs  $[\boldsymbol{\mu}_j^\top, -\mathbf{q}^\top]^\top$ .

2. (*MLP*) By Proposition 6.4.5, we know that a large enough proportion of positive (or negative) neurons  $i \in [m]$  have the label embedding of  $\mathbf{W}_{O_{(i,:)}}^{(T)}, \mathbf{W}_V^{(T)}$  close to  $\pm \mathbf{q}$  (6.22). They can thus map the weighted summation of contexts and the query by attention with  $+\mathbf{q}$  (or  $-\mathbf{q}$ ) to positive (or negative) values. This leads to a correct prediction in-domain (Theorem 6.3.3).

3. (*Out-of-Domain Generalization*) Since Corollary 6.4.3 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  $[\boldsymbol{\mu}'_j^\top, \mathbf{q}^\top]^\top$  or  $[\boldsymbol{\mu}'_j^\top, -\mathbf{q}^\top]^\top$ . Then, by combining (6.21), (6.22) and the condition on ODR pattern characterized in (6.12), we can ensure that the MLP layer produces a desired prediction out of the domain (Theorem 6.3.4).

#### E.4.2 Preliminaries

**Lemma E.4.1.** (*Multiplicative Chernoff bounds, Theorem D.4 of [270]*) Let  $X_1, \dots, X_m$  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  $\hat{p} = \frac{1}{m} \sum_{i=1}^m X_i$ :

$$\Pr(\hat{p} \geq (1 + \gamma)p) \leq e^{-\frac{mp\gamma^2}{3}}, \quad (\text{E.3})$$

$$\Pr(\hat{p} \leq (1 - \gamma)p) \leq e^{-\frac{mp\gamma^2}{2}}. \quad (\text{E.4})$$

**Definition E.4.2.** [258] 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 E.4.3.** ([258] Proposition 5.1, Hoeffding's inequality) Let  $X_1, X_2, \dots, X_N$  be independent centered sub-gaussian random variables, and let  $K = \max_i \|\mathbf{X}_i\|_{\psi_2}$ . Then for every  $\mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N$  and every  $t \geq 0$ , we have

$$\Pr \left( \left| \sum_{i=1}^N a_i X_i \right| \geq t \right) \leq e \cdot \exp \left( -\frac{ct^2}{K^2 \|\mathbf{a}\|^2} \right), \quad (\text{E.5})$$

where  $c > 0$  is an absolute constant.

**Definition E.4.4.** For any data index  $n$  and iteration  $t$ , we can find  $i$  such that  $\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) > 0$  by the initialization with high probability. Define

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

**Lemma E.4.5.** (gradient updates of  $\mathbf{W}_Q$  and  $\mathbf{W}_K$ ) By the SGD training method described in Section 6.2.3, we have the following equations. Given the definition of in-/out-of-domain data as in (6.1) and the in-/out-of-domain data distribution  $\mathcal{D}$  in (6.6) and  $\mathcal{D}'$  in (6.8), 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  $\mathbf{x}_{query}$  and the example  $\mathbf{x}_q, q \in [l]$  are  $\boldsymbol{\mu}_j$ . The label embedding is  $\mathbf{0}$  for the query and  $\pm \mathbf{q}$  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,

$$(\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \sum_{b=0}^{t_0} \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \gtrsim \eta \frac{1}{M_1} \sum_{b=0}^{t_0} \zeta_b \delta \gamma_b \beta^4, \quad (\text{E.6})$$

$$\begin{aligned} & \left| (\boldsymbol{\mu}_l^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{t_0-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\ & \lesssim e^{-\Theta((\frac{\eta t_0}{M_1})^2)} \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \sum_{b=0}^{t_0-1} \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right|, \end{aligned} \quad (\text{E.7})$$

$$\begin{aligned} & \left| (\boldsymbol{\mu}_a^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{t_0-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\ & \lesssim \frac{1}{M_1} \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \sum_{b=0}^{t_0-1} \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right|, \end{aligned} \quad (\text{E.8})$$

$$\begin{aligned} & \left| (\boldsymbol{\nu}_k^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\ & \lesssim \frac{1}{M_2} \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right|, \end{aligned} \quad (\text{E.9})$$

$$(\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=t_0+1} \mathbf{p}_q \gtrsim \eta \frac{1}{M_1} \sum_{b=0}^{t_0} \zeta_b \delta \gamma_b \beta^4, \quad (\text{E.10})$$

$$\begin{aligned} & \left| (\boldsymbol{\mu}_l^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=t_0} \mathbf{p}_q \right| \\ & \lesssim e^{-\Theta((\frac{\eta t_0}{M_1})^2)} \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=t_0} \mathbf{p}_q \right|, \end{aligned} \quad (\text{E.11})$$

$$\begin{aligned} & \left| (\boldsymbol{\mu}_a^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=t_0} \mathbf{p}_q \right| \\ & \lesssim \frac{1}{M_1} \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=t_0} \mathbf{p}_q \right|, \end{aligned} \quad (\text{E.12})$$

$$\begin{aligned} & \left| (\boldsymbol{\nu}_k^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=t_0} \mathbf{p}_q \right| \\ & \lesssim \frac{1}{M_2} \left| (\boldsymbol{\mu}_j^\top, \mathbf{q}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=t_0} \mathbf{p}_q \right|. \end{aligned} \quad (\text{E.13})$$

In the above, equations (E.6), (E.7), (E.8), and (E.9) characterize the directions of gradient updates of  $\mathbf{W}_Q$  when projected with  $(\mathbf{x}_{query}^\top, \mathbf{0})^\top$ . Similarly, equations (E.10), (E.11), (E.12), and (E.13) characterize the directions of gradient updates of  $\mathbf{W}_K$  when projected with  $\mathbf{p}_q, q \in [l]$ .

**Lemma E.4.6.** (gradient updates of  $\mathbf{W}_V$ ) For  $\mathbf{p}_j^n$  defined in (6.1) 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  $\mathbf{p}_j$  of which the corresponding label embedding is  $\mathbf{q}$ ,

$$\begin{aligned} & \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \sum_{b=0}^{t_0} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_V^{(b)}} \mathbf{p}_j \\ &= \eta \sum_{b=0}^{t_0} \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \right), \end{aligned} \quad (\text{E.14})$$

where

$$-V_i(b) \gtrsim \beta^2(1 - \gamma_t)/a, \quad i \in \mathcal{W}_n, \quad (\text{E.15})$$

$$-V_i(b) \leq \frac{1}{\beta^2 + 1} V_j(b), \quad i \in \mathcal{U}_n, j \in \mathcal{W}_n, \quad (\text{E.16})$$

$$|V_i(b)| \lesssim \sqrt{\frac{\log B}{B}} \cdot \frac{1}{a}, \quad i \notin \mathcal{W}_n \cup \mathcal{U}_n. \quad (\text{E.17})$$

If the corresponding label embedding is  $-\mathbf{q}$ , we have the that (E.14) holds with

$$-V_i(b) \gtrsim \beta^2(1 - \gamma_t)/a, \quad i \in \mathcal{U}_n, \quad (\text{E.18})$$

$$-V_i(b) \leq \frac{1}{\beta^2 + 1} V_j(b), \quad i \in \mathcal{W}_n, j \in \mathcal{U}_n, \quad (\text{E.19})$$

$$|V_i(b)| \lesssim \sqrt{\frac{\log B}{B}} \cdot \frac{1}{a}, \quad i \notin \mathcal{W}_n \cup \mathcal{U}_n. \quad (\text{E.20})$$

We can also derive

$$\begin{aligned} & \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_V^{(t)}} (\boldsymbol{\nu}_k^\top, \mathbf{0}^\top)^\top \\ &= \eta \sum_{b=0}^{t_0} \left( \sum_{i \in \mathcal{W}_n} V'_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V'_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V'_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \right), \end{aligned} \quad (\text{E.21})$$

where

$$|V'_i(b)| \leq |V_i(b)| \cdot \frac{1}{M_2}. \quad (\text{E.22})$$

**Lemma E.4.7.** (gradient updates of  $\mathbf{W}_O$ ) 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  $\mathbf{p}_{query}^n$  as  $\mathcal{B}_b^n$  in the  $b$ -th iteration. For  $i \in \mathcal{W}$ ,  $b \neq a$ , and  $\mathbf{p}_{query}^n$  corresponding to  $\mathbf{q}$  and  $\boldsymbol{\mu}_a$ ,

$$\eta \frac{1}{|\mathcal{B}_b^n|} \sum_{n \in \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^{(t_0)}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top = \delta(\beta^2 + 1) \frac{\alpha \eta}{2a} (1 + \frac{\eta^2 m}{a^2})^{t_0}, \quad (\text{E.23})$$

$$\eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^{(t_0)}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top \gtrsim \delta(\beta^2 + 1) \frac{\alpha \eta t_0}{2a}. \quad (\text{E.24})$$

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

$$\eta \frac{1}{|\mathcal{B}_b^n|} \sum_{n \in \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^{(t_0)}} (\boldsymbol{\mu}_a^\top, -\mathbf{q}^\top)^\top = \delta(\beta^2 + 1) \frac{\alpha \eta}{2a} (1 + \frac{\eta^2 m}{a^2})^{t_0}, \quad (\text{E.25})$$

$$\begin{aligned} & \eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^{(b)}} (\boldsymbol{\mu}_a^\top, -\mathbf{q}^\top)^\top \\ & \gtrsim \delta(\beta^2 + 1) \frac{\alpha \eta t_0}{2a}. \end{aligned} \quad (\text{E.26})$$

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

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(t_0)}\| \gtrsim \sqrt{M_1} \delta(\beta^2 + 1)^{\frac{1}{2}} \frac{\alpha \eta t_0}{2a}, \quad (\text{E.27})$$

$$\eta \frac{1}{B} \sum_{b=0}^{t_0} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^{(b)}} (\boldsymbol{\nu}_c^\top, \pm \mathbf{q}^\top)^\top \leq \frac{1}{M_2} \eta \frac{1}{B} \sum_{b=0}^{t_0} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^{(b)}} (\boldsymbol{\mu}_b^\top, \mathbf{q}^\top)^\top. \quad (\text{E.28})$$

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

$$\eta \frac{1}{B} \sum_{b=0}^{t_0} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^{(b)}} (\boldsymbol{\mu}_a^\top, \pm \mathbf{q}^\top)^\top \leq \eta t \sqrt{\frac{\log B}{B}} \frac{1}{a}. \quad (\text{E.29})$$

**Definition E.4.8.** Define

$$\mathbf{V}^n(t) := \sum_{s=1}^{l+1} \mathbf{W}_V^{(t)} \mathbf{p}_s^n \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n), \quad (\text{E.30})$$

for  $\mathbf{P}^n$ . Let  $\mathbf{W}_{O_{(i,\cdot)}} = (\mathbf{O}_{i,1}, \mathbf{O}_{i,2}, \mathbf{0}^\top)$  where  $\mathbf{O}_{i,1} \in \mathbb{R}^{d_x}, \mathbf{O}_{i,2} \in \mathbb{R}^{d_y}$ . Let  $\mathbf{V}^n(t) =$

$(\mathbf{V}_{n,1}(t)^\top, \mathbf{V}_{n,2}(t)^\top, \mathbf{0}^\top)^\top$  where  $\mathbf{V}_{i,1}(t) \in \mathbb{R}^{d_x}, \mathbf{V}_{i,2}(t) \in \mathbb{R}^{d_y}$ . Define  $\mathcal{W}_n, \mathcal{U}_n$  as the sets of lucky neurons such that

$$\mathcal{W}_n = \{i : \mathbf{O}_{i,1}^{(0)} \mathbf{V}_{n,1}(0) > 0, \mathbf{O}_{i,2}^{(0)} \mathbf{V}_{n,2}(0) > 0, a_i > 0\}, \quad (\text{E.31})$$

$$\mathcal{U}_n = \{i : \mathbf{O}_{i,1}^{(0)} \mathbf{V}_{n,1}(0) > 0, \mathbf{O}_{i,2}^{(0)} \mathbf{V}_{n,2}(0) > 0, a_i < 0\}. \quad (\text{E.32})$$

Define

$$\mathcal{N}_j^{n,i} = \{i : i \in [l+1], \mathbf{x}_i^n = \boldsymbol{\mu}_j + \kappa_i^n \boldsymbol{\nu}_k + \mathbf{n}_i^n, k \in [M_2]\}, \quad (\text{E.33})$$

$$\mathcal{M}_k^{n,i} = \{i : i \in [l+1], \mathbf{x}_i^n = \boldsymbol{\mu}_j + \kappa_i^n \boldsymbol{\nu}_k + \mathbf{n}_i^n, j \in [M_1]\}, \quad (\text{E.34})$$

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

$$\mathcal{W} = \bigcup_{n=1}^N \mathcal{W}_n, \quad \mathcal{U} = \bigcup_{n=1}^N \mathcal{U}_n. \quad (\text{E.35})$$

**Lemma E.4.9.** *By the definition of lucky neurons in (E.31) and (E.32), and the initialization described in Section 6.2.3, the number of lucky neurons  $|\mathcal{W}_n|, |\mathcal{U}_n|$  satisfies*

$$|\mathcal{W}_n|, |\mathcal{U}_n| \geq \Omega(m). \quad (\text{E.36})$$

Hence,

$$|\mathcal{W}|, |\mathcal{U}| \geq \Omega(m). \quad (\text{E.37})$$

**Lemma E.4.10.** *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  $\mathbf{p}_{query}^n$  corresponds to the label +1,

$$\mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{V}^n(t)] = 1, i \in \mathcal{W}_n, \quad (\text{E.38})$$

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

$$\mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{V}^n(t)] = 1, i \in \mathcal{U}_n. \quad (\text{E.39})$$

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

$$\mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{V}^n(t)] = 1. \quad (\text{E.40})$$

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

$$\mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{V}^n(t)] = 1. \quad (\text{E.41})$$

**Lemma E.4.11.** *With in-domain tasks defined in Definition 6.3.1 and Condition 6.3.2, the number of training tasks should satisfy  $|\mathcal{T}_{tr}| \geq M_1$  to make Condition 6.3.2 hold.*

### E.4.3 Proof of Theorem 6.3.3

*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 E.4.1, we can obtain

$$\Pr\left(\frac{1}{l_{tr}} \sum_{i=1}^{l_{tr}} \mathbb{1}[m_i = \boldsymbol{\mu}_a] \leq (1 - c)\frac{\alpha}{2}\right) \leq e^{-l_{tr}c^2\frac{\alpha}{2}} = M_1^{-C}, \quad (\text{E.42})$$

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 E.4.6,

$$l_{tr} \gtrsim \max\left\{\Omega\left(\frac{2\log M_1}{\alpha}\right), \Omega\left(\frac{1}{\alpha\beta^2}\right)\right\}. \quad (\text{E.43})$$

By the condition in Lemma E.4.5, we have that

$$B \geq \Omega(M_1 \log M_1). \quad (\text{E.44})$$

We know that there exists gradient noise caused by imbalanced IDR patterns in each batch. Therefore, by Hoeffding's inequality (E.5), for any  $\mathbf{W} \in \Psi$ ,

$$\begin{aligned} \Pr\left(\left\|\frac{1}{|\mathcal{B}_b|} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, z^n)}{\partial \mathbf{W}} - \mathbb{E}\left[\frac{\partial \ell(\Psi; \mathbf{P}^n, z^n)}{\partial \mathbf{W}}\right]\right\| \geq \left|\mathbb{E}\left[\frac{\partial \ell(\Psi; \mathbf{P}^n, z^n)}{\partial \mathbf{W}}\right]\right| \epsilon\right) \\ \leq e^{-B\epsilon^2} \leq M_1^{-C}, \end{aligned} \quad (\text{E.45})$$

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

$$B \gtrsim \max\{\epsilon^{-2}, M_1\} \log M_1. \quad (\text{E.46})$$

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

$$a_i \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(T)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(T)} \mathbf{p}_s^n) \text{softmax}((\mathbf{W}_K^{(T)} \mathbf{p}_s^n)^\top (\mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n))) \geq 0. \quad (\text{E.47})$$

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

$$\begin{aligned} & \eta \sum_{b=0}^{T-1} \mathbf{W}_{O_{(j,:)}}^{(T)} \sum_{j \in \mathcal{W}_n} \mathbf{W}_{O_{(j,:)}}^{(b)} {}^\top \\ & \geq \eta \sum_{b=0}^{T-1} M_1 \delta (\beta^2 + 1)^{\frac{1}{2}} \frac{\alpha \eta T}{2a} \cdot \delta (\beta^2 + 1)^{\frac{1}{2}} \frac{\alpha \eta b}{2a} \\ & \gtrsim \delta^2 (\beta^2 + 1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2}, \end{aligned} \quad (\text{E.48})$$

where the first step comes from (E.27) in Lemma E.4.7, and the second step is by  $\sum_{b=0}^{T-1} b = \Theta(T^2)$ . Then, we can obtain

$$\begin{aligned} & \mathbf{W}_{O_{(i,:)}}^{(T)} \mathbf{W}_V^{(T)} \mathbf{p}_s^n \\ &= \mathbf{W}_{O_{(i,:)}}^{(T)} (\delta \mathbf{p}_s^n + \sum_{b=0}^{T-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \right)^\top) \quad (\text{E.49}) \\ &\gtrsim \delta^2 (\beta^2 + 1) \frac{\alpha \eta T}{2a} + \delta^2 (\beta^2 + 1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2}, \end{aligned}$$

where the first step is by (E.14) in Lemma E.4.6, and the last step comes from Lemma E.4.7.

Therefore, by combining Lemma E.4.9 and Lemma E.4.10, we have that

$$\begin{aligned} & \sum_{i \in \mathcal{W}} a_i \text{Relu}(\mathbf{W}_{O_{(i,:)}}^{(T)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(T)} \mathbf{p}_s^n) \text{softmax}((\mathbf{W}_K^{(T)} \mathbf{p}_s^n)^\top (\mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n))) \\ & \gtrsim (1 - \gamma_T) \cdot \left( \delta^2 (\beta^2 + 1) \frac{\alpha \eta T}{2a} + \delta^2 (\beta^2 + 1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2} \right), \end{aligned} \quad (\text{E.50})$$

when  $\gamma_T$  is order-wise smaller than 1. We next give a bound for  $\gamma_T$ , which is given by Definition E.4.4,

$$\gamma_T \geq 1 - \sum_{s \in \mathcal{N}_*^n} \text{softmax}((\mathbf{W}_K^{(T)} \mathbf{p}_s^n)^\top (\mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n)), \quad (\text{E.51})$$

from Definition E.4.4 for  $\mu_j$  as the IDR pattern in  $\mathbf{p}_{query}^n$ . We can tell from (E.50) and Definition

E.4.4,

$$\zeta_b \gtrsim \delta^2(\beta^2 + 1) \frac{\alpha\eta T}{2a} + \delta^2(\beta^2 + 1)\alpha^2 \frac{(\eta T)^3 M_1}{a^2}. \quad (\text{E.52})$$

Then, if  $\zeta_T \gtrsim 1$  and  $T \gtrsim M_1$ , by Lemma E.4.5, with high probability,

$$\begin{aligned} & (\mathbf{W}_K^{(T)} \mathbf{p}_s^n)^\top (\mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n) \\ & \gtrsim \left( \eta \frac{1}{M_1} \sum_{b=0}^{T-1} \zeta_b \delta \gamma_b \beta^2 + \delta \right)^2 \\ & \gtrsim \left( \eta \sum_{b=0}^{T-1} \gamma_b \beta^2 \cdot \left( \delta^2 \frac{\alpha\eta T}{2a} + \delta^2(\beta^2 + 1)\alpha^2 \frac{(\eta T)^3 M_1}{a^2} \right) + \delta \right)^2 \\ & := \left( \eta \sum_{b=0}^{T-1} \gamma_b \cdot \Delta_T + \delta \right)^2, \end{aligned} \quad (\text{E.53})$$

where the first step comes from the fact that the gradient update projections of  $\mathbf{W}_Q$  and  $\mathbf{W}_K$  onto queries or examples are close to the corresponding IDR pattern the most by Lemma E.4.5. In the last inequality of (E.53), we only consider the term related to  $T$  and  $\gamma_b$ . For any  $\mathbf{p}_l^n$  that shares a different IDR pattern as  $\mathbf{p}_{query}^n$ , we have

$$(\mathbf{W}_K^{(T)} \mathbf{p}_l^n)^\top (\mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n) \lesssim \frac{1}{M_1} (\mathbf{W}_K^{(T)} \mathbf{p}_s^n)^\top (\mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n), \quad (\text{E.54})$$

by Lemma E.4.5. Then, given the definition of softmax,

$$\begin{aligned} & \sum_{s \in \mathcal{N}_j^n} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\ & \geq \frac{\sum_{s \in \mathcal{N}_j^n} e^{\Theta(\delta^2) + (\eta \sum_{b=0}^{T-1} \gamma_b \cdot \Delta_T)^2}}{\sum_{s \in \mathcal{N}_j^n} e^{\Theta(\delta^2) + (\eta \sum_{b=0}^{T-1} \gamma_b \cdot \Delta_T)^2} + \sum_{s \in [l] - \mathcal{N}_j^n} e^{\frac{1}{M_1}(\eta \sum_{b=0}^{T-1} \gamma_b \cdot \Delta_T)^2}} \\ & \geq 1 - \frac{2 - \alpha}{\alpha} e^{-(\eta \sum_{b=0}^{T-1} \gamma_b \cdot \Delta_T)^2}, \end{aligned} \quad (\text{E.55})$$

where the first step is by (E.53) and (E.54). Combining with (E.51), we can derive

$$\begin{aligned} \gamma_T & \leq \frac{2 - \alpha}{\alpha} e^{-(\eta \sum_{b=0}^{T-1} \gamma_b \cdot \Delta_T)^2} = \frac{2 - \alpha}{\alpha} e^{-(\eta \sum_{b=0}^{T-2} \gamma_b \cdot \Delta_T)^2} \cdot e^{-\eta^2 \Delta_T^2 (2\gamma_{T-1} \sum_{b=0}^{T-2} \gamma_b + \gamma_{T-1}^2)} \\ & = \gamma_{T-1} \cdot e^{-\eta^2 \Delta_T^2 (2\gamma_{T-1} \sum_{b=0}^{T-2} \gamma_b + \gamma_{T-1}^2)}. \end{aligned} \quad (\text{E.56})$$

When  $T$  is large,  $\gamma_T$  is approaching zero. Hence, the equality of (E.56) is close to being

achieved, in which case,

$$\gamma_T \approx \gamma_{T-1} \cdot e^{-\eta^2 \Delta_T^2 (2\gamma_{T-1} \sum_{b=0}^{T-2} \gamma_b + \gamma_{T-1}^2)}. \quad (\text{E.57})$$

We can observe that when  $\sum_{b=0}^{t_0-1} \eta \gamma_b \Delta_T \geq \sqrt{\log M_1}$ ,  $\gamma_{t_0}$  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  $\eta$  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  $\mathbf{p}_s^n$  is  $\mathbf{q}$ , the summation of corresponding softmax value is  $1 - \gamma_T$ . Furthermore,

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(T)} \mathbf{W}_V^{(T)} \mathbf{p}_s^n \\ & \lesssim -\delta^2 \frac{\alpha \eta T}{2a} - \delta^2 \alpha^2 \frac{(\eta T)^3 M_1}{a^2}, \end{aligned} \quad (\text{E.58})$$

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

$$\text{Relu}(\mathbf{W}_{O_{(i,\cdot)}}^{(T)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(T)} \mathbf{p}_s^n) \text{softmax}((\mathbf{W}_K^{(T)} \mathbf{p}_s^n)^\top (\mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n))) = 0. \quad (\text{E.59})$$

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

$$\mathbf{W}_{O_{(i,\cdot)}}^{(T)} \mathbf{W}_V^{(T)} \mathbf{p}_s^n \lesssim \frac{1}{\sqrt{M_1}} \mathbf{W}_{O_{(i,\cdot)}}^{(T)} \mathbf{W}_V^{(T)} \mathbf{p}_s^n, \quad (\text{E.60})$$

by Lemma E.4.6 and  $B \gtrsim M_1$ . The final lower bound of  $F(\Psi; \mathbf{P}^n)$  is based on the lower bound introduced by  $i \in \mathcal{W}$ .

Then, combining (E.47), (E.50), (E.59), and (E.60), we can derive

$$\begin{aligned} & F(\Psi; \mathbf{P}^n) \\ & \gtrsim (1 - \gamma_T) \cdot \left( \delta^2 (\beta^2 + 1) \frac{\alpha \eta T}{2a} + \delta^2 (\beta^2 + 1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2} \right). \end{aligned} \quad (\text{E.61})$$

Therefore, as long as

$$T = \Theta(\eta^{-1} M_1 \delta^{-2/3} \beta^{-2/3} \alpha^{-2/3} \sqrt{\log M_1}), \quad (\text{E.62})$$

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

$$F(\Psi; \mathbf{P}^n) \geq 1. \quad (\text{E.63})$$

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

$$F(\Psi; \mathbf{P}^n) \leq -1. \quad (\text{E.64})$$

**Then, we study in-domain generalization.** By (E.45), for any given testing prompt embedding  $\mathbf{P}$  with  $z = +1$ , we have

$$F(\Psi; \mathbf{P}) \geq 1 - \epsilon, \quad (\text{E.65})$$

and if  $z = -1$ ,

$$F(\Psi; \mathbf{P}) \leq -1 + \epsilon. \quad (\text{E.66})$$

Therefore,

$$\mathbb{E}_{\mathbf{x}_{query} \sim \mathcal{D}, f \in \mathcal{T}} [\ell(\Psi; \mathbf{P}, y)] \leq \epsilon. \quad (\text{E.67})$$

□

#### E.4.4 Proof of Theorem 6.3.4

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

$$l_{ts} \geq \frac{1}{\alpha'}. \quad (\text{E.68})$$

Consider  $\mathbf{p}_{query}^n'$  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 (E.52) and (E.53), we have

that for  $s \in \mathcal{N}^n$ ,

$$\begin{aligned} & (\mathbf{W}_K^{(T)} \mathbf{p}_s^{n'})^\top \mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n \\ & \gtrsim \sum_{j=1}^{M_1} c_j^2 \cdot \left( \eta \sum_{b=0}^{T-1} \gamma_b \beta^2 \cdot \left( \delta^2 \frac{\alpha \eta T}{2a} + \delta^2 (\beta+1)^2 \alpha^2 \frac{(\eta T)^3 M_1}{a^2} \right) \right)^2 \\ & \gtrsim \log M_1. \end{aligned} \quad (\text{E.69})$$

For ODR patterns, by Proposition 6.4.1, we have for  $\mathbf{p}_l^n$  that has a different ODR pattern than  $\mathbf{p}_{query}^n$ ,

$$(\mathbf{W}_K^{(T)} \mathbf{p}_l^n)^\top (\mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n) \lesssim \left( \frac{1}{M_1} + \frac{1}{M_2} \right) (\mathbf{W}_K^{(T)} \mathbf{p}_s^n)^\top (\mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n). \quad (\text{E.70})$$

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

$$\sum_{s \in \mathcal{N}_j^n} \text{softmax}((\mathbf{W}_K^{(T)} \mathbf{p}_s^{n'})^\top (\mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n)) \geq 1 - \frac{2 - \alpha'}{\alpha'} \Theta\left(\frac{1}{M_1}\right) \geq 1 - \frac{2}{\alpha'} \Theta\left(\frac{1}{M_1}\right). \quad (\text{E.71})$$

Note that for  $\mathbf{p}_s^{n'} = \sum_{i=1}^{M_1} c_i \boldsymbol{\mu}_i + \sum_{j=1}^{M_2} \kappa_s^{n'} g_j \boldsymbol{\nu}_j$ , 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  $M_1 < M_2 = \Theta(M_1)$ , 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  $\mathbf{p}_s^n$  where the corresponding label-space embedding is  $\mathbf{q}$ ,

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(T)} \mathbf{W}_V^{(T)} \mathbf{p}_s^{n'} \\ & = \mathbf{W}_{O_{(i,\cdot)}}^{(T)} \mathbf{W}_V^{(T)} \left( \sum_{i=1}^{M_1} c_i \boldsymbol{\mu}_i^\top + \kappa_s^{n'} \sum_{j=1}^{M_2} g_j \boldsymbol{\nu}_j^\top + \mathbf{o}_s^{n\top}, \mathbf{q}^\top \right)^\top \\ & \gtrsim \sum_{i=1}^{M_1} c_i \left( \delta^2 \beta^2 \frac{\alpha \eta T}{2a} + \delta^2 (\beta^2 + 1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2} \right) \\ & \quad + \left( \delta^2 \frac{\alpha \eta T}{2a} + \delta^2 \alpha^2 \frac{(\eta T)^3 M_1}{a^2} \right) (1 - \epsilon) \\ & \geq \left( \delta^2 (\beta^2 + 1) \frac{\alpha \eta T}{2a} + \delta^2 (\beta^2 + 1) \alpha^2 \frac{(\eta T)^3 M_1}{a^2} \right) \cdot (1 - \epsilon), \end{aligned} \quad (\text{E.72})$$

where the first equality comes from the definition of  $\mathbf{p}_s^{n'}$ . The first inequality of (E.72) is derived from (E.45). The last inequality is by the condition  $\sum_{i=1}^{M_1} c_i \geq 1$ . Therefore, we can derive that

$$\begin{aligned} F(\Psi; \mathbf{P}^{n'}) &\gtrsim (1 - \gamma_T)(\delta^2(\beta^2 + 1)\frac{\alpha\eta T}{2a} + \delta^2(\beta^2 + 1)\alpha^2\frac{(\eta T)^3 M_1}{a^2}) \cdot (1 - \epsilon) \\ &\geq 1 - \epsilon, \end{aligned} \quad (\text{E.73})$$

where the first step is by following (E.63), and the remaining steps are from basic mathematical computation. Likewise, for  $\mathbf{p}_{query}^n$ ' such that the label is  $-1$ , we can obtain

$$F(\Psi; \mathbf{P}^{n'}) < -(1 - \epsilon). \quad (\text{E.74})$$

Therefore, we have

$$\mathbb{E}_{x_{query} \sim \mathcal{D}', f \in \mathcal{T}'} [\ell(\Psi; \mathbf{P}, y)] \leq \epsilon. \quad (\text{E.75})$$

□

#### E.4.5 Proof of Theorem 6.3.7

*Proof.* We cover the proof in the proof of Proposition 6.4.5. Please see Section E.5.4 for more details. □

### E.5 Proofs of Key Lemmas and Propositions

#### E.5.1 Proof of Lemma E.4.11

*Proof.* We first show that if  $|\mathcal{T}_{tr}| < M_1$ , Condition 6.3.2 cannot hold. Then, We show that there exists  $\mathcal{T}_{tr}$  with  $|\mathcal{T}_{tr}| \geq M_1$  such that Condition 6.3.2 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 6.3.2.

(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.5.2 Proof of Proposition 6.4.1

*Proof.* We first show the results for in-domain patterns.

(1) We investigate the results about  $\mathbf{W}_Q$  and then  $\mathbf{W}_K$ . For  $\mathbf{p}_{query}$  with  $\boldsymbol{\mu}_j$  as the IDR pattern and  $\mathbf{a} \in \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{M_1}\} \setminus \{\boldsymbol{\mu}_j\}$ , by (E.8), we have

$$\begin{aligned} & (\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} \mathbf{p}_{query} \\ &= (\mathbf{a}^\top, \mathbf{0}^\top) (\mathbf{W}_Q^{(0)} + \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=b}) \mathbf{p}_{query} \\ &\lesssim \frac{1}{M_1} \cdot (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=b} \mathbf{p}_{query}, \end{aligned} \quad (\text{E.76})$$

if  $\mathbf{a}$  does not form a task in  $\mathcal{T}_{tr}$  with  $\boldsymbol{\mu}_j$ . If  $\mathbf{a}$  forms a task in  $\mathcal{T}_{tr}$  with  $\boldsymbol{\mu}_j$ , and  $\eta T = \Theta(M_1 \sqrt{\log M_1})$ , by (E.7)

$$\begin{aligned} & (\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} \mathbf{p}_{query} \\ &\lesssim \frac{1}{M_1} \cdot (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=b} \mathbf{p}_{query}. \end{aligned} \quad (\text{E.77})$$

For  $\mathbf{a} \perp \boldsymbol{\mu}_j$  but  $\mathbf{a} \notin \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{M_1}\}$ , by (E.9), we have

$$\begin{aligned} & (\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top)^\top \\ &= (\mathbf{a}^\top, \mathbf{0}^\top) (\mathbf{W}_Q^{(0)} + \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=b}) \mathbf{p}_{query} \\ &\lesssim \frac{1}{M_2} \cdot (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=b} \mathbf{p}_{query}. \end{aligned} \quad (\text{E.78})$$

By (E.6) and the initialization, we have

$$\begin{aligned} & (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} \mathbf{p}_{query} \\ &\geq (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=b} \mathbf{p}_{query} \gtrsim \sqrt{\log M_1} + \delta \gtrsim \sqrt{\log M_1}, \end{aligned} \quad (\text{E.79})$$

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 (E.76), (E.77), and (E.79), we can derive that

$$\begin{aligned} & \| \mathbf{W}_Q^{(T)} \mathbf{p}_{query}^n \| \\ & \lesssim \sqrt{1 + \frac{1}{M_1^2} \cdot M_1 + \frac{1}{M_1^2} + \frac{1}{M_2^2} \cdot M_2 (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=b} \mathbf{p}_{query}} \quad (\text{E.80}) \\ & = \sqrt{1 + \frac{1}{M_1} + \frac{1}{M_2}} (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q^{(t)}} \Big|_{t=b} \mathbf{p}_{query}, \end{aligned}$$

where in the first step, the first  $1/M_1^2 \cdot M_1$  comes from (E.76) with  $M_1 - 1$  choices of  $\mathbf{a}$ . The second  $1/M_1^2$  comes from (E.77), 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 (E.78) with  $M_2$  choices of  $\mathbf{a}$ . Therefore, by (E.79) and (E.80), for  $\mathbf{a} \in \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{M_1}\} \setminus \{\boldsymbol{\mu}_j\}$ , we have

$$(\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} \mathbf{p}_{query} \lesssim \frac{\sqrt{\log M_1}}{M_1}. \quad (\text{E.81})$$

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

$$(\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} \mathbf{p}_{query} \lesssim \frac{\sqrt{\log M_1}}{M_2}. \quad (\text{E.82})$$

For  $\mathbf{W}_K$ , we can make derivations following the above steps. For  $\mathbf{p}_q$  with  $\boldsymbol{\mu}_j$  as the IDR pattern and  $\mathbf{a} \in \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{M_1}\} \setminus \{\boldsymbol{\mu}_j\}$ , by (E.12), we have

$$\begin{aligned} & (\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_K^{(T)} \mathbf{p}_q \\ & \lesssim \frac{1}{M_1} \cdot (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=b} \mathbf{p}_q, \end{aligned} \quad (\text{E.83})$$

if  $\mathbf{a}$  does not form a task in  $\mathcal{T}_{tr}$  with  $\boldsymbol{\mu}_j$ . If  $\mathbf{a}$  forms a task in  $\mathcal{T}_{tr}$  with  $\boldsymbol{\mu}_j$ , and  $\eta T = \Theta(M_1 \sqrt{\log M_1})$ , by (E.11),

$$\begin{aligned} & (\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_K^{(T)} \mathbf{p}_q \\ & \lesssim \frac{1}{M_1} \cdot (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=b} \mathbf{p}_q. \end{aligned} \quad (\text{E.84})$$

For  $\mathbf{a} \perp \boldsymbol{\mu}_j$  but  $\mathbf{a} \notin \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{M_1}\}$ , by (E.13), we have

$$\begin{aligned} & (\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_K^{(T)} \mathbf{p}_q \\ & \lesssim \frac{1}{M_2} \cdot (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=b} \mathbf{p}_q. \end{aligned} \quad (\text{E.85})$$

By (E.10) and the initialization, we have

$$(\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \mathbf{W}_K^{(T)} \mathbf{p}_q \geq \sqrt{\log M_1} + \delta \geq \sqrt{\log M_1}. \quad (\text{E.86})$$

Hence, by combining (E.83), (E.84), and (E.85), we can derive that

$$\begin{aligned} & \|\mathbf{W}_K^{(T)} \mathbf{p}_i^n\| \\ & \lesssim \sqrt{1 + \frac{1}{M_1} + \frac{1}{M_1^2} + \frac{1}{M_2}} (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{b=0}^{T-1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K^{(t)}} \Big|_{t=b} (\boldsymbol{\mu}_j^\top, \pm \mathbf{q})^\top. \end{aligned} \quad (\text{E.87})$$

Therefore, by (E.86) and (E.87), for  $\mathbf{a} \in \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_{M_1}\} \setminus \{\boldsymbol{\mu}_j\}$ , we have

$$(\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_K^{(T)} \mathbf{p}_q \lesssim \frac{\sqrt{\log M_1}}{M_1}. \quad (\text{E.88})$$

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

$$(\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_K^{(T)} \mathbf{p}_q \lesssim \frac{\sqrt{\log M_1}}{M_2}. \quad (\text{E.89})$$

(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  $\mathbf{p}'_{query}$ , of which the corresponding ODR pattern is  $\boldsymbol{\mu}'_j$ , we have that by (E.6), (E.7), (E.8), and (E.9),

$$|(\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} \mathbf{p}'_{query}| \geq |k_j| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} \mathbf{p}_{query} \left(1 - \frac{\Theta(1)}{M_1} - \frac{\Theta(1)}{M_2}\right), \quad (\text{E.90})$$

$$|(\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(t)} \mathbf{p}'_{query}| \leq |k_j| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(t)} \mathbf{p}_{query} \left(1 + \frac{\Theta(1)}{M_1} + \frac{\Theta(1)}{M_2}\right), \quad (\text{E.91})$$

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

$$|(\boldsymbol{\nu}_k^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(t)} \mathbf{p}'_{query}| \leq \frac{1}{M_2} (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(t)} \mathbf{p}_{query}. \quad (\text{E.92})$$

Therefore,

$$\|\mathbf{W}_Q^{(t)} \mathbf{p}'_{query}\| \geq \sqrt{\log M_1} \left(1 - \frac{\Theta(1)}{M_1} - \frac{\Theta(1)}{M_2}\right) \gtrsim \sqrt{\log M_1}, \quad (\text{E.93})$$

$$\|\mathbf{W}_Q^{(t)} \mathbf{p}'_{query}\| \leq \sqrt{\log M_1} \left(1 + \frac{\Theta(1)}{M_1} + \frac{\Theta(1)}{M_2}\right) \lesssim \sqrt{\log M_1}, \quad (\text{E.94})$$

$$|(\boldsymbol{\mu}_j'^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} \mathbf{p}'_{query}| \gtrsim \sum_{i=1}^{M_1} |k_{ji}| \sqrt{\log M_1} \geq \sqrt{\log M_1}. \quad (\text{E.95})$$

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

$$(\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} \mathbf{p}_{query} \lesssim \left(\frac{1}{M_1} + \frac{1}{M_2}\right) \sum_{i=1}^{M_1} |k_{ai} k_{ji}| \leq \sqrt{\log M_1} \left(\frac{1}{M_1} + \frac{1}{M_2}\right), \quad (\text{E.96})$$

where the first step is by (E.90) and (E.91), and the second step is by Cauchy-Schwarz inequality given that  $\sum_{i=1}^{M_1} k_{ji}^2 = \sum_{i=1}^{M_1} k_{ai}^2 = 1$ . For  $\mathbf{a} \perp \boldsymbol{\mu}'_j$  but  $\mathbf{a} \notin \{\boldsymbol{\mu}'_1, \dots, \boldsymbol{\mu}'_{M'_1}\}$ ,

$$(\mathbf{a}^\top, \mathbf{0}^\top) \mathbf{W}_Q^{(T)} \mathbf{p}_{query} \lesssim \sqrt{\log M_1} \left(\frac{1}{M_1} + \frac{1}{M_2}\right). \quad (\text{E.97})$$

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

□

### E.5.3 Proof of Corollary 6.4.3

*Proof.* From (E.53) to (E.57), we can derive the conclusion for IDR patterns. For ODR patterns, from (E.71), we can obtain the conclusion. Note that  $\frac{2-\alpha}{\alpha} = \Theta(1)$  since  $\alpha = \Theta(1)$ . □

#### E.5.4 Proof of Proposition 6.4.5

*Proof.* For any  $i$ , we denote  $\mathbf{W}_{O_{(i,:)}}^{(b)} = (\mathbf{O}_{i,1}^{(b)}, \mathbf{O}_{i,2}^{(b)}, \mathbf{0}^\top)$  where  $\mathbf{O}_{i,1}^{(b)\top} \in \mathbb{R}^{d_x}$  and  $\mathbf{O}_{i,2}^{(b)\top} \in \mathbb{R}^{d_y}$ .

Following the derivation of (E.49), we can obtain that for  $s \in [l]$  or  $s$  is the query,

$$\begin{aligned} & \mathbf{W}_{O_{(i,:)}}^{(T)} \mathbf{W}_V^{(T)} (\mathbf{p}_s^n, \mathbf{0}^\top)^\top \\ &= \mathbf{W}_{O_{(i,:)}}^{(T)} (\delta(\mathbf{p}_s^n, \mathbf{0}^\top)^\top + \sum_{b=0}^{T-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{O}_{i,1}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{O}_{i,1}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{O}_{i,1}^{(b)\top} \right)) \quad (\text{E.98}) \\ &\gtrsim \delta^2 \beta^2 \frac{\alpha \eta T}{2a} + \delta^2 \beta^2 \alpha^2 \frac{(\eta T)^3 M_1}{a^2}, \end{aligned}$$

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

$$\mathbf{W}_{O_{(i,:)}}^{(T)} \mathbf{W}_V^{(T)} (\kappa_i^n \boldsymbol{\nu}_k^\top, \mathbf{0}^\top)^\top \leq \frac{1}{M_2} \cdot \mathbf{W}_{O_{(i,:)}}^{(T)} \mathbf{W}_V^{(T)} (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top)^\top, \quad (\text{E.99})$$

for  $k \in [M_2]$  by (E.22) and (E.28) in Lemma E.4.6 and E.4.7, respectively. Then, we have

$$\begin{aligned} & \frac{(\frac{1}{M_1} \sum_{j=1}^{M_1} \boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \mathbf{W}_{O_{(i,:)}}^{(T)} \mathbf{W}_V^{(T)}}{\|(\frac{1}{M_1} \sum_{j=1}^{M_1} \boldsymbol{\mu}_j^\top, \mathbf{0}^\top)\| \| \mathbf{W}_{O_{(i,:)}}^{(T)} \mathbf{W}_V^{(T)} \|} \geq \frac{1}{\sqrt{1 + \frac{1}{M_2^2} \cdot M_2}} \\ & \geq 1 - \frac{\Theta(1)}{M_2}, \end{aligned} \quad (\text{E.100})$$

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

$$\begin{aligned} & \mathbf{W}_{O_{(i,:)}}^{(T)} \mathbf{W}_V^{(T)} (\mathbf{0}^\top, \mathbf{q}^\top)^\top \\ &= \mathbf{W}_{O_{(i,:)}}^{(T)} (\delta(\mathbf{0}^\top, \mathbf{q}^\top)^\top + \sum_{b=0}^{T-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{O}_{i,2}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{O}_{i,2}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{O}_{i,2}^{(b)\top} \right)) \quad (\text{E.101}) \\ &\gtrsim \delta^2 \frac{\alpha \eta T}{2a} + \delta^2 \alpha^2 \frac{(\eta T)^3 M_1}{a^2}. \end{aligned}$$

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

$$\begin{aligned} & \frac{(\mathbf{0}^\top, \mathbf{q}^\top) \mathbf{W}_{O_{(i,:)}}^{(T)} \mathbf{W}_V^{(T)}}{\|(\mathbf{0}^\top, \mathbf{q}^\top)\| \| \mathbf{W}_{O_{(i,:)}}^{(T)} \mathbf{W}_V^{(T)} \|} \geq \frac{1}{\sqrt{1 + \xi}} \\ & \geq 1 - \frac{\Theta(1)}{M_1}. \end{aligned} \quad (\text{E.102})$$

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

$$\begin{aligned}
& \mathbf{W}_{O_{(i,\cdot)}}^{(T)} \mathbf{W}_V^{(T)} (\mathbf{0}^\top, -\mathbf{q}^\top)^\top \\
&= \mathbf{W}_{O_{(i,\cdot)}}^{(T)} (\delta(\mathbf{0}^\top, -\mathbf{q}^\top)^\top + \sum_{b=0}^{T-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{O}_{i,2}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{O}_{i,2}^{(b)} \right. \\
&\quad \left. + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{O}_{i,2}^{(b)\top} \right)) \\
&\gtrsim \delta^2 \frac{\alpha \eta T}{2a} + \delta^2 \alpha^2 \frac{(\eta T)^3 M_1}{a^2}.
\end{aligned} \tag{E.103}$$

Similarly to (E.102), we have

$$\frac{(\mathbf{0}^\top, -\mathbf{q}^\top) \mathbf{W}_{O_{(i,\cdot)}}^{(T)} \mathbf{W}_V^{(T)}}{\|(\mathbf{0}^\top, -\mathbf{q}^\top)\| \|\mathbf{W}_{O_{(i,\cdot)}}^{(T)} \mathbf{W}_V^{(T)}\|} \geq 1 - \frac{\Theta(1)}{M_1}. \tag{E.104}$$

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

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(T)} \mathbf{W}_V^{(T)}\| \gtrsim \beta^{-1} = \Omega(1). \tag{E.105}$$

By (E.60), we have that for  $i \notin \mathcal{W} \cup \mathcal{U}$ ,

$$\|\mathbf{W}_{O_{(i,\cdot)}}^{(T)} \mathbf{W}_V^{(T)}\| \lesssim \sqrt{\frac{1}{M_1} + \frac{1}{M_2^2} \cdot M_2} = \frac{1}{\sqrt{M_2}}, \tag{E.106}$$

where  $1/M_1$  is the square of (E.60).  $1/M_2^2$  is the square of the scaling in RHS of (E.99), and  $M_2$  is the number of IDI patterns.

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

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}_{query,f}} [\ell(\Psi; \mathbf{P}, y)] &\leq 1 - \left(1 - \frac{2}{\alpha' M_1}\right) \frac{1 - \epsilon}{\left(1 - \frac{2}{\alpha M_1}\right)} \left(1 - \frac{1}{\sqrt{M_1}}\right) \\
&\leq 1 - \left(1 - \frac{2}{\alpha' M_1}\right) (1 - \epsilon) \left(1 - \frac{1}{\sqrt{M_1}}\right) \\
&\leq 1 - \left(1 - \frac{2}{\alpha' M_1} - \epsilon - \frac{1}{\sqrt{M_1}}\right) \\
&\leq \mathcal{O}(\epsilon + \frac{1}{\sqrt{M_1}} + \frac{1}{\alpha' M_1}) \\
&\leq \mathcal{O}(\epsilon + \frac{1}{\sqrt{M_1}}),
\end{aligned} \tag{E.107}$$

where the first step combines (E.61), (E.60), and  $2/(\alpha M_1)$  and  $2/(\alpha' M_1)$  comes from (E.55) and (E.71). 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 (E.45), we have for the trained model  $\Psi$ ,

$$F(\Psi; \mathbf{P}^n) \leq (1 + \epsilon)(1 - R) \cdot \frac{(1 - \frac{2}{\alpha' M_1})}{(1 - \frac{2}{\alpha M_1})}. \quad (\text{E.108})$$

Then,

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_{query}, f} [\ell(\Psi; \mathbf{P}, y)] &\geq 1 - (1 - \frac{2}{\alpha' M_1}) \frac{1 + \epsilon}{(1 - \frac{2}{\alpha M_1})} (1 - R) \\ &\geq 1 - (1 - \frac{2}{\alpha' M_1})(1 + \epsilon)(1 + \frac{4}{\alpha M_1})(1 - R) \\ &= 1 - (1 - R - \frac{2}{\alpha' M_1} + \frac{2R}{\alpha' M_1})(1 + \epsilon + \frac{4}{\alpha M_1} + \frac{4\epsilon}{\alpha M_1}) \\ &\geq R + \frac{2}{\alpha' M_1} - \frac{2R}{\alpha' M_1} - (1 - R - \frac{2}{\alpha' M_1})(\epsilon + \frac{4 + 4\epsilon}{\alpha M_1}) \\ &\geq \Omega(R + \frac{1}{\alpha' M_1}), \end{aligned} \quad (\text{E.109})$$

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.5.5 Proof of Lemma E.4.5

*Proof.* We first study the gradient of  $\mathbf{W}_Q^{(t+1)}$  in part (a) and the gradient of  $\mathbf{W}_K^{(t+1)}$  in part (b). The proof is derived with a framework of induction combined with Lemma E.4.6 and E.4.7.

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

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial F(\mathbf{p}_{query}^n)} \frac{\partial F(\mathbf{p}_{query}^n)}{\partial \mathbf{W}_Q} \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O(i,\cdot)} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \\
&\quad \cdot \left( \mathbf{W}_{O(i,\cdot)} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \right. \\
&\quad \cdot \left. \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{W}_K (\mathbf{p}_s^n - \mathbf{p}_r^n) \mathbf{p}_{query}^n{}^\top \right) \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O(i,\cdot)} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \\
&\quad \cdot \left( \mathbf{W}_{O(i,\cdot)} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \right. \\
&\quad \cdot \left. (\mathbf{W}_K \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{W}_K \mathbf{p}_r^n) \mathbf{p}_{query}^n{}^\top \right). 
\end{aligned} \tag{E.110}$$

If  $t = 0$ , we have that

$$(\mathbf{W}_K^{(t)} \mathbf{p}_s^n)^\top (\mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) = \mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n. \tag{E.111}$$

When  $z^n = +1$ , let  $\mathbf{x}_{query}^n$  be a noisy version of  $\boldsymbol{\mu}_j + \kappa_{query}^n \boldsymbol{\nu}_k$  where  $j \in \{1, 2, \dots, M_1\}$  and  $k \in \{1, 2, \dots, M_2\}$ . 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 E.4.1, given  $l \geq \Theta(\max\{M_1, M_2\} \log M_1)$ , we can obtain

$$\Pr \left( \sum_{i=1}^l \mathbb{1}[m_i = \boldsymbol{\mu}_j] \leq l \cdot \frac{\alpha}{2} \right) \leq e^{-C \log M_1} = M_1^{-C}, \tag{E.112}$$

$$\Pr\left(\sum_{i=1}^l \mathbb{1}[m_i = \boldsymbol{\mu}_s] \leq l \cdot \frac{\alpha}{2}\right) \leq e^{-C \log M_1} = M_1^{-C}, \quad (\text{E.113})$$

$$\Pr\left(\sum_{i=1}^l \mathbb{1}[m_i = \boldsymbol{\mu}_t] \geq l \cdot \frac{1}{M_1}\right) \leq e^{-C \log M_1 \cdot M_1 \cdot \frac{1}{M_1}} = M_1^{-C}, \quad (\text{E.114})$$

$$\Pr\left(\sum_{i=1}^l \mathbb{1}[m_i = \boldsymbol{\nu}_k] \geq l \cdot \frac{1}{M_2}\right) \leq e^{-C \log M_1 \cdot M_2 \cdot \frac{1}{M_2}} = M_1^{-C}, \quad (\text{E.115})$$

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

$$\begin{aligned} \sum_{s \in \mathcal{N}_j^{n,i} \cap \mathcal{M}_k^{n,i}} e^{\delta^2(\beta \cdot \beta + 1)} &\leq l \cdot \frac{1}{\sqrt{M_2}} e^{\delta^2} \cdot e^{\delta^2(\beta \cdot \beta)} \\ &\lesssim l \cdot \frac{\alpha}{2} \cdot e^{\delta^2(\beta \cdot \beta)} \\ &\lesssim \sum_{s \in \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} e^{\delta^2(\beta \cdot \beta)}. \end{aligned} \quad (\text{E.116})$$

Similarly,

$$\begin{aligned} \sum_{s \in \mathcal{M}_k^{n,i} - \mathcal{N}_j^{n,i}} e^{\delta^2} &\leq l \cdot \frac{1}{\sqrt{M_2}} e^{\delta^2} \cdot e^{\delta^2(\beta \cdot \beta)} \\ &\lesssim l \cdot \frac{\alpha}{2} \cdot e^{\delta^2(\beta \cdot \beta)} \\ &\lesssim \sum_{s \in [l] - \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} e^{\delta^2}, \end{aligned} \quad (\text{E.117})$$

where the last step is by the fact that there exists  $\boldsymbol{\mu}_s$  for  $p \in \{1, 2, \dots, M_2\} \setminus \{j\}$  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

$$\begin{aligned} &\text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\ &\geq e^{\delta^2(\beta \cdot \beta)} \cdot \left( \sum_{s \in \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} e^{\delta^2(\beta \cdot \beta)} + \sum_{s \in \mathcal{N}_j^{n,i} \cap \mathcal{M}_k^{n,i}} e^{\delta^2(\beta \cdot \beta + 1)} \right. \\ &\quad \left. + \sum_{s \in [l] - \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} e^{\delta^2} + \sum_{s \in \mathcal{M}_k^{n,i} - \mathcal{N}_j^{n,i}} e^{\delta^2} \right)^{-1} \\ &\gtrsim \frac{e^{\delta^2(\beta \cdot \beta)}}{\sum_{s \in \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} e^{\delta^2(\beta \cdot \beta)} + \sum_{s \in [l] - \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} e^{\delta^2}}, \end{aligned} \quad (\text{E.118})$$

where the second step is by (E.116) and (E.117). Similarly, for  $s \in \mathcal{N}_j^{n,i} \cap \mathcal{M}_k^{n,i}$ ,

$$\begin{aligned} & \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\ & \gtrsim \frac{e^{\delta^2(\beta\cdot\beta)}}{\sum_{s \in \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} e^{\delta^2(\beta\cdot\beta)} + \sum_{s \in [l] - \mathcal{N}_j^{n,i} - \mathcal{N}_k^{n,i}} e^{\delta^2}}. \end{aligned} \quad (\text{E.119})$$

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

$$\begin{aligned} & \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\ & \lesssim \frac{e^{\delta^2}}{\sum_{s \in \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} e^{\delta^2(\beta\cdot\beta)} + \sum_{s \in [l] - \mathcal{N}_j^{n,i} - \mathcal{N}_k^{n,i}} e^{\delta^2}}. \end{aligned} \quad (\text{E.120})$$

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

$$\begin{aligned} & \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\ & \lesssim \frac{e^{\delta^2}}{\sum_{s \in \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} e^{\delta^2(\beta\cdot\beta)} + \sum_{s \in [l] - \mathcal{N}_j^{n,i} - \mathcal{N}_k^{n,i}} e^{\delta^2}}. \end{aligned} \quad (\text{E.121})$$

By (E.31) and (E.32) in Definition E.4.8, we have that for  $i \in \mathcal{W}_n$ ,

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) > 0. \quad (\text{E.122})$$

Then we derive

$$\begin{aligned} & \mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n \\ & = \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \\ & = \left( \sum_{r \in \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} + \sum_{r \in \mathcal{N}_j^{n,i} \cap \mathcal{M}_k^{n,i}} + \sum_{r \in \mathcal{M}_k^{n,i} - \mathcal{N}_j^{n,i}} + \sum_{r \in [l] - \mathcal{N}_j^{n,i} - \mathcal{M}_k^{n,i}} \right) \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \\ & \quad \cdot \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \mathbf{W}_K^{(t)} \mathbf{p}_r^n). \end{aligned} \quad (\text{E.123})$$

One can observe that

$$\begin{aligned}
& \sum_{s \in \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \\
& \quad \cdot \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \\
&= \sum_{s \in \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - (\sum_{r \in \mathcal{N}_j^{n,i}} + \sum_{r \notin \mathcal{N}_j^{n,i}}) \text{softmax}(\mathbf{p}_r^{n\top} \\
& \quad \cdot \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \\
&= \sum_{r \notin \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot \sum_{s \in \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \quad (\text{E.124}) \\
& \quad \cdot \mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{s \in \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot \sum_{r \notin \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \\
& \quad \cdot \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot \mathbf{W}_K^{(t)} \mathbf{p}_r^n + \mathbf{n} \\
&= \sum_{s \in \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot \sum_{r \notin \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\
& \quad \cdot \mathbf{W}_K^{(t)} (\mathbf{p}_s^n - \mathbf{p}_r^n) + \mathbf{n}.
\end{aligned}$$

Hence, by Definition E.4.4,

$$1 > \sum_{r \in [l] - \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \geq \gamma_t > 0. \quad (\text{E.125})$$

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

$$\begin{aligned}
& (\mathbf{x}_s^{n\top}, \mathbf{0}^\top) \sum_{r \in [l] - \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \\
& \geq \gamma_t \beta^2 \delta,
\end{aligned} \quad (\text{E.126})$$

where  $\gamma_t$  comes from the definition.  $\beta$  is from the definition of the data. Meanwhile, for  $r$  such that  $\mu_r$  is the IDR pattern with the probability of  $\alpha/2$  to be selected,

$$\begin{aligned} & \left| (\mathbf{x}_r^{n\top}, \mathbf{0}^\top) \sum_r \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \right| \\ & \leq (\mathbf{x}_s^{n\top}, \mathbf{0}^\top) \sum_{r \in [l] - \mathcal{N}_j^{n,i}} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \\ & \quad \cdot \frac{1 - \alpha}{1 - \alpha/2}, \end{aligned} \quad (\text{E.127})$$

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

$$\begin{aligned} & \left| (\mathbf{x}_r^{n\top}, \mathbf{0}^\top) \sum_r \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \right| \\ & \leq \frac{\gamma_t}{l}, \end{aligned} \quad (\text{E.128})$$

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

$$\begin{aligned} & (\mu_q^\top, \mathbf{0}^\top) \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{N}_j^{n,i}} (\mathbf{W}_V^{(t)} \mathbf{p}_s^{n(t)}) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\ & \quad \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^n \top (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \\ & \geq \zeta'_{i,n} \delta \beta^4 \gamma_t, \end{aligned} \quad (\text{E.129})$$

where  $\gamma_t$  comes from that  $\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p}_s^n$  is much larger in average,  $i \in \mathcal{W}_n$ , if than other  $\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{p}_t^n$  if  $s \in \mathcal{N}_*^n$  while  $s \notin \mathcal{N}_*^n$ . For  $j$  such that  $\mu_j$  is the IDR pattern with the probability of  $\alpha/2$

to be selected but different from  $q$ ,

$$\begin{aligned}
& (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{N}_j^{n,i}} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^n \top (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \quad (\text{E.130}) \\
& \leq \frac{1-\alpha}{1-\alpha/2} (\mathbf{x}_{query}^\top, \mathbf{0}^\top) \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{N}_j^{n,i}} (\mathbf{W}_V^{(t)} \mathbf{p}_s^{n(t)}) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n).
\end{aligned}$$

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

$$\begin{aligned}
& (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \sum_{s \in \mathcal{N}_j^{n,i}} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^n \top (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \quad (\text{E.131}) \\
& \leq \frac{\zeta'_{i,n} \delta \gamma_t \beta^4}{l}.
\end{aligned}$$

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  $\mathbf{W}_{O_{(i,\cdot)}}$  follows  $\mathcal{N}(0, \xi^2)$ , we have

$$\Pr(\|\mathbf{W}_{O_{(i,1:d_X)}} \mathbf{x}_{query}^n\| \leq \beta\xi) \leq \beta\xi, \quad (\text{E.132})$$

by the standard property of Gaussian distribution. Meanwhile, by Hoeffding's inequality (E.5),

$$\Pr(\|\mathbf{W}_{O_{(i,1:d_X)}} \mathbf{x}_{query}^n\| \geq \beta\xi \log M_1) \leq M_1^{-C}, \quad (\text{E.133})$$

for some  $C > 1$ . Hence, with a high probability, by Hoeffding's inequality (E.5),

$$\begin{aligned}
\left| \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,1:d_X)}}^{(t)} \mathbf{x}_{query}^n \right| & \lesssim \frac{|\mathcal{W}_n|}{m} \Phi(0) \beta \xi + \beta \xi \cdot \frac{\log M_1}{\sqrt{m}} \\
& \lesssim \beta \xi,
\end{aligned} \quad (\text{E.134})$$

$$\begin{aligned}
\left| \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,1:d_X)}}^{(t)} \mathbf{x}_{query}^n \right| & \gtrsim \frac{|\mathcal{W}_n|}{m} \Phi(0) \beta \xi - \beta \xi \cdot \frac{\log M_1}{\sqrt{m}} \\
& \gtrsim \beta \xi.
\end{aligned} \quad (\text{E.135})$$

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

$$\left| \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,1:d_X)}}^{(t)} \boldsymbol{\mu}_p \right| \lesssim \left| \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,1:d_X)}}^{(t)} \boldsymbol{\mu}_j \right| \cdot e^{-\delta\beta^2}. \quad (\text{E.136})$$

We have that for  $z^n = 1$ , we can then derive that for  $s \in \mathcal{N}_j^{n,i}$ , by Definition E.4.4, for  $\boldsymbol{\mu}_q$  which is the IDR pattern of  $\mathbf{x}_{query}$ ,

$$\begin{aligned} & (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\ & \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^n{}^\top (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \quad (\text{E.137}) \\ & \geq \zeta_{i,t} \delta \gamma_t (1 - e^{-\delta^2 \beta^2}) \beta^4, \end{aligned}$$

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

$$\begin{aligned} & (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\ & \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^n{}^\top (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \quad (\text{E.138}) \\ & \lesssim e^{-\delta^2 \beta^2} \cdot (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\ & \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^n{}^\top (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top. \end{aligned}$$

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  $\mu_j$  in each data. Then,

$$\begin{aligned}
& (\mu_j^\top, \mathbf{0}^\top) \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^n \top (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \\
& \leq \frac{1}{l} \cdot (\mu_q^\top, \mathbf{0}^\top) \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^n \top (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top. \tag{E.139}
\end{aligned}$$

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  $\mu_q$  which is the IDR pattern of  $\mathbf{x}_{query}$ ,

$$\begin{aligned}
& (\mu_q^\top, \mathbf{0}^\top) \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_l^n) \geq 0] \\
& \cdot \left( \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \right. \\
& \left. \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^n \top \right) (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \tag{E.140} \\
& \leq \sqrt{\frac{\log m}{m}} \cdot (\mu_q^\top, \mathbf{0}^\top) \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^n \top (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top,
\end{aligned}$$

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

$$\begin{aligned}
& (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \sum_{i \notin \mathcal{W}_t \cup \mathcal{U}} a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_l^n) \geq 0] \\
& \cdot \left( \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \right. \\
& \cdot \left. (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^{n\top} \right) (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \\
& \leq e^{-\delta\beta^2} \sqrt{\frac{\log m}{m}} \cdot (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \\
& \cdot \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^{n\top} \\
& \cdot (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top.
\end{aligned} \tag{E.141}$$

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

$$\begin{aligned}
& (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \sum_{i \notin \mathcal{W}_t \cup \mathcal{U}} a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_l^n) \geq 0] \\
& \cdot \left( \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \right. \\
& \cdot \left. (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^{n\top} \right) (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \\
& \leq \frac{1}{l} \cdot \sqrt{\frac{\log m}{m}} \cdot (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \frac{1}{m} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \\
& \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^{n\top} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top.
\end{aligned} \tag{E.142}$$

Therefore, by (E.137), (E.140), (E.141), and (E.142), we have that for one  $\mathbf{x}_{query}$ ,

$$\begin{aligned}
& \left| \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) (\mathbf{x}_{query}^\top, \mathbf{0}^\top) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \cdot \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \right. \\
& \left. \mathbf{W}_Q^{(t)} \mathbf{p}_l^n) \geq 0] \cdot \left( \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \right. \right. \\
& \left. \left. \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^{n\top} \right) (\mathbf{x}_{query}^\top, \mathbf{0}^\top) \right| \quad (\text{E.143}) \\
& \geq \eta \frac{1}{BM_1} \sum_{n \in \mathcal{B}_b} \frac{1}{m} \sum_{i \in \mathcal{W}_n} \zeta_{i,t} \delta \gamma_t (1 - e^{-\delta^2 \beta^2} \sqrt{\frac{\log m}{m}} - \frac{1}{l} \sqrt{\frac{\log m}{m}}) \beta^4 \\
& \gtrsim \eta \frac{1}{M_1} \zeta_t \delta \gamma_t \beta^4,
\end{aligned}$$

as long as

$$m \gtrsim 1, \quad (\text{E.144})$$

and

$$B \gtrsim M_1 \log M_1, \quad (\text{E.145})$$

to ensure that

$$\Pr \left( \sum_{n=1}^B \mathbb{1}[m_n = \mu_j] \leq B(1-c) \cdot \frac{1}{M_1} \right) \leq e^{-c^2 B \cdot \frac{1}{M_1}} = e^{-c \log M_1} = M_1^{-C}, \quad (\text{E.146})$$

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  $\mathbf{x}_{query}$ , more indicators of  $i \in \mathcal{U}_n$  is activated.

$$\begin{aligned}
& - \left| \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \cdot \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)})^\top \right. \\
& \cdot \mathbf{W}_Q^{(t)} \mathbf{p}_l^n) \geq 0] \left( \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)})^\top \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n \right. \\
& \cdot \left. \left. (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^n \top \mathbf{W}_K^{(t)})^\top \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n \mathbf{p}_{query}^n \top \right) (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\
& \geq - \frac{1}{2} e^{-\delta^2 \beta^2} \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \cdot \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)})^\top \right. \\
& \cdot \left. \mathbf{W}_K^{(t)} \mathbf{p}_l^n \geq 0] \cdot \left( \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)})^\top \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n \right. \right. \\
& \cdot \left. \left. (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^n \top \mathbf{W}_K^{(t)})^\top \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n \mathbf{p}_{query}^n \top \right) (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top. \right. \tag{E.147}
\end{aligned}$$

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

$$\begin{aligned}
& \left| \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \cdot \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)})^\top \right. \\
& \cdot \mathbf{W}_Q^{(t)} \mathbf{p}_l^n) \geq 0] \left( \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)})^\top \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n \right. \\
& \cdot \left. \left. (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^n \top \mathbf{W}_K^{(t)})^\top \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n \mathbf{p}_{query}^n \top \right) (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\
& \leq \frac{1}{M_1} \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) (\boldsymbol{\mu}_q^{n\top}, \mathbf{0}^\top) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \cdot \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)})^\top \right. \\
& \cdot \mathbf{W}_Q^{(t)} \mathbf{p}_l^n) \geq 0] \left( \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)})^\top \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n \right. \\
& \cdot \left. \left. (\mathbf{W}_K^{(t)} \mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^n \top \mathbf{W}_K^{(t)})^\top \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n \mathbf{p}_{query}^n \top \right) (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top, \right. \tag{E.148}
\end{aligned}$$

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

(E.143), (E.147), and (E.148), we have

$$\begin{aligned} & (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \\ & \gtrsim \eta \frac{1}{M_1} \zeta_t \delta \gamma_t \beta^4. \end{aligned} \quad (\text{E.149})$$

By (E.147) and (E.148), we have that for  $\boldsymbol{\mu}_j$  which forms a task in  $\mathcal{T}_{tr}$  with the  $\boldsymbol{\mu}_q$ ,

$$\begin{aligned} & - \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\ & \geq - \frac{1}{2} e^{-\delta^2 \beta^2} \left| (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right|. \end{aligned} \quad (\text{E.150})$$

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

$$\begin{aligned} & \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\ & \leq \frac{1}{M_1} \left| (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right|. \end{aligned} \quad (\text{E.151})$$

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

$$\begin{aligned} & \left| (\boldsymbol{\nu}_k^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\ & \lesssim \frac{1}{M_2} \left| (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right|, \end{aligned} \quad (\text{E.152})$$

where  $M_2$  comes from that for  $\boldsymbol{\nu}_k$  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

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=0} [:, d_{\mathcal{X}} + 1 : d_{\mathcal{X}} + d_{\mathcal{Y}}] \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^{(t) \top} \mathbf{W}_Q^{(t)} \mathbf{p}_l^n) \geq 0] \\
&\quad \cdot \left( \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^{(t) \top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \cdot (\mathbf{W}_K^{(t)} \mathbf{p}_s^n \right. \\
&\quad \left. - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n \top} \mathbf{W}_K^{(t) \top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n) \mathbf{p}_{query}^{n \top} \right) [:, d_{\mathcal{X}} + 1 : d_{\mathcal{X}} + d_{\mathcal{Y}}]. \\
&= 0.
\end{aligned} \tag{E.153}$$

We then have

$$\begin{aligned}
& \left| \mathbf{q}^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^{(t) \top} \mathbf{W}_Q^{(t)} \mathbf{p}_l^n) \geq 0] \right. \\
&\quad \cdot \left( \mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^{(t) \top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \right. \\
&\quad \left. \left. - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n \top} \mathbf{W}_K^{(t) \top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \mathbf{W}_K^{(t)} \mathbf{p}_r^n \right) \right) [d_{\mathcal{X}} + 1 : d_{\mathcal{X}} + d_{\mathcal{Y}}] \right| \\
&= 0.
\end{aligned} \tag{E.154}$$

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 (E.143) and (E.147). Note that the indicator of  $\mathcal{W}_n$  will not change along the training. The only difference is the changes in  $\zeta_t$  and  $\gamma_t$ . Thus, we can obtain that for  $\boldsymbol{\mu}_q$  with the same IDR pattern as  $\mathbf{x}_{query}$ ,

$$\begin{aligned}
& (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=t_0+1} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \\
& \gtrsim \eta \frac{1}{M_1} \sum_{b=0}^{t_0} \zeta_b \delta \gamma_b \beta^4 \\
& \gtrsim \eta \frac{1}{M_1} \sum_{b=0}^{t_0} \zeta_b \delta \gamma_b \beta^4,
\end{aligned} \tag{E.155}$$

as long as (E.145) holds. We also have

$$\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=t_0+1}[:, d_{\mathcal{X}} + 1 : d_{\mathcal{X}} + d_{\mathcal{Y}}] = \mathbf{0}. \quad (\text{E.156})$$

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  $\boldsymbol{\mu}_q$ ,

$$\begin{aligned} & \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\ & \lesssim \frac{1}{M_1} \left| (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right|. \end{aligned} \quad (\text{E.157})$$

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

$$\begin{aligned} & - \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\ & \gtrsim - e^{-\delta^2 \beta^2 - (\eta \frac{1}{M_1} \sum_{b=0}^{t_0-1} \zeta_b \delta \gamma_b \beta^2)^2} \left| (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \quad (\text{E.158}) \\ & \gtrsim - e^{-\Theta(\frac{\eta t_0}{M_1})^2} \left| (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right|, \end{aligned}$$

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  $\eta, t, M_1$  as variables. Meanwhile, for  $k \in [M_2]$ ,

$$\begin{aligned} & \left| (\boldsymbol{\nu}_k^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right| \\ & \lesssim \frac{1}{M_2} \left| (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_Q} \Big|_{t=t_0} (\mathbf{x}_{query}^\top, \mathbf{0}^\top)^\top \right|. \end{aligned} \quad (\text{E.159})$$

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

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n, \Psi)}{\partial \mathbf{W}_K} \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \cdot \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \\
&\quad \cdot \left( \mathbf{W}_{O_{(i,:)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{W}_Q^\top \mathbf{p}_{query}^n \right. \\
&\quad \left. \cdot (\mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_r^n)^\top \right). 
\end{aligned} \tag{E.160}$$

If we investigate  $\mathbf{W}_K^{(t)} \mathbf{p}_s^n$ , we can tell that the output is a weighed summation of multiple  $\mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n$ . Similarly, the output of  $\mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n$  is a weighed summation of multiple  $\mathbf{W}_K^{(t)} \mathbf{p}_s^n$ . Given the initialization  $\mathbf{W}_Q^{(0)}$  and  $\mathbf{W}_K^{(0)}$ , the update of  $\mathbf{W}_K^{(t)} \mathbf{p}_s^n$  and  $\mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n$  only contains the contribution from the feature space embeddings at the initialization. One difference is that since that  $\mathbf{q}$  appears with 1/2 probability in all  $\mathbf{p}_s^n$ ,

$$\begin{aligned}
& \left| \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \cdot \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \right. \\
&\quad \cdot \left( \mathbf{W}_{O_{(i,:)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{W}_Q^\top \mathbf{p}_{query}^n \right. \\
&\quad \left. \left. \cdot (\mathbf{p}_s^n - \sum_{r=1}^{l+1} \text{softmax}(\mathbf{p}_r^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_r^n)^\top \right) [d_{\mathcal{X}} + 1 : d_{\mathcal{X}} + dy] \mathbf{q} \right| \\
&\leq \sqrt{\frac{\log B}{B}} (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K} \Big|_{t=t_0+1} (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top)^\top.
\end{aligned} \tag{E.161}$$

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

$$\begin{aligned}
& (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K} \Big|_{t=t_0+1} (\mathbf{x}_e^\top, \mathbf{0}^\top)^\top \\
&\gtrsim \eta \frac{1}{M_1} \sum_{b=0}^{t_0} \zeta_b \delta \gamma_b \beta^2,
\end{aligned} \tag{E.162}$$

and combining (E.161),

$$\begin{aligned}
& (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K} \Big|_{t=t_0+1} \mathbf{p}_q \\
& \gtrsim \eta \frac{1}{M_1} \sum_{b=0}^{t_0} \zeta_b \delta \gamma_b \beta^2 \left(1 - \sqrt{\frac{\log B}{B}}\right) \\
& \gtrsim \eta \frac{1}{M_1} \sum_{b=0}^{t_0} \zeta_b \delta \gamma_b \beta^2,
\end{aligned} \tag{E.163}$$

where the last step holds as long as (E.145). For  $\boldsymbol{\mu}_j$  which forms a task in  $\mathcal{T}_{tr}$  with the  $\boldsymbol{\mu}_q$ ,

$$\begin{aligned}
& - \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K} \Big|_{t=t_0} \mathbf{p}_q \right| \\
& \gtrsim - e^{-\Theta(\frac{\eta t_0}{M_1})^2} \left| (\boldsymbol{\mu}_q^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K} \Big|_{t=t_0} \mathbf{p}_q \right|.
\end{aligned} \tag{E.164}$$

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

$$\begin{aligned}
& \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K} \Big|_{t=t_0} \mathbf{p}_q \right| \\
& \leq \frac{1}{M_1} \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K} \Big|_{t=t_0} \mathbf{p}_q \right|.
\end{aligned} \tag{E.165}$$

Meanwhile, for  $k \in [M_2]$ , similar to (E.159),

$$\begin{aligned}
& \left| (\boldsymbol{\nu}_k^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K} \Big|_{t=t_0} \mathbf{p}_q \right| \\
& \leq \frac{1}{M_2} \left| (\boldsymbol{\mu}_j^\top, \mathbf{0}^\top) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_K} \Big|_{t=t_0} \mathbf{p}_q \right|.
\end{aligned} \tag{E.166}$$

□

### E.5.6 Proof of Lemma E.4.6

*Proof.*

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_V} \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial F(\mathbf{p}_{query}^n)} \frac{\partial F(\mathbf{p}_{query}^n)}{\partial \mathbf{W}_V} \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i=1}^m a_i \mathbb{1}[\mathbf{W}_{O_{(i,:)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \\
&\quad \cdot \mathbf{W}_{O_{(i,:)}}^\top \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n \top}.
\end{aligned} \tag{E.167}$$

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

$$\begin{aligned}
& \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n \top} (\mathbf{x}_i^{n \top}, \mathbf{q}^\top)^\top \gtrsim \beta^2 (1 - \gamma_t) \cdot 1 - \frac{1}{\frac{\alpha}{2} l} \\
& \gtrsim \beta^2 (1 - \gamma_t) \cdot 1,
\end{aligned} \tag{E.168}$$

where the first step holds since that by (E.115), with high probability, no other  $\mathbf{x}_k^n$  where  $k \neq l + 1$  shares the same IDI pattern as  $\mathbf{x}_{query}^n$ . The last step holds if

$$l_{tr} \gtrsim \frac{1}{\alpha \beta^2}. \tag{E.169}$$

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

$$\sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n \top} (\mathbf{x}_j^{n \top}, \mathbf{q}^\top)^\top \lesssim \beta^2 \gamma_t. \tag{E.170}$$

When  $t = 0$ , for all  $i \in \mathcal{W}_n$ , we have that by Lemma E.4.10, for  $\mathbf{p}_{query}^n$  that corresponds to  $\boldsymbol{\mu}_a$ ,

$$\mathbf{W}_{O_{(i,:)}}^{(t)} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n \top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) > 0. \tag{E.171}$$

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

$$\mathbf{x}_j^n = \boldsymbol{\mu}_a + \kappa_j^n \boldsymbol{\nu}_b, \quad (\text{E.172})$$

we have

$$\begin{aligned} & \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_V^{(t)}} \mathbf{p}_j^n \\ &= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{k=1}^m a_k \mathbb{1}[\mathbf{W}_{O_{(k,\cdot)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \\ & \quad \cdot \mathbf{W}_{O_{(k,\cdot)}}^\top \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n\top} \mathbf{p}_j^n. \end{aligned} \quad (\text{E.173})$$

We then have that by combining (E.168) and (E.170),

$$\begin{aligned} & -\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i \in \mathcal{W}_n} a_i \mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \\ & \quad \cdot \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n\top} \mathbf{p}_j^n \\ & \gtrsim \eta \beta^2 (1 - \gamma_t). \end{aligned} \quad (\text{E.174})$$

Since that for  $\mathbf{p}_s^n$  and  $\mathbf{p}_j^n$  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

$$\begin{aligned} & -\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i \in \mathcal{U}_n} a_i \mathbb{1}[\mathbf{W}_{O_{(k,\cdot)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \\ & \quad \cdot \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n\top} \mathbf{p}_j^n \\ & \leq \frac{1}{\beta^2 + 1} \cdot (-\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i \in \mathcal{W}_n} a_i \mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \\ & \quad \cdot \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n\top} \mathbf{p}_j^n). \end{aligned} \quad (\text{E.175})$$

Meanwhile, since that

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n}^m a_i \mathbb{1}[\mathbf{W}_{O(i,\cdot)} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n)] \geq 0 \\
& \cdot \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n\top} \mathbf{p}_j^n \\
& \lesssim \sqrt{\frac{\log B}{B}} \cdot \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{j \in \mathcal{W}_n}^m a_j \mathbb{1}[\mathbf{W}_{O(j,\cdot)} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \\
& \cdot \mathbf{W}_Q \mathbf{p}_{query}^n)] \geq 0] \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n\top} \mathbf{p}_j^n,
\end{aligned} \tag{E.176}$$

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  $t = t_0 + 1$ ,

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \sum_{b=0}^{t_0} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_V^{(b)}} \mathbf{p}_j^n \\
& = \eta \sum_{b=0}^{t_0} \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O(i,\cdot)}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O(i,\cdot)}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O(i,\cdot)}^{(b)} \right),
\end{aligned} \tag{E.177}$$

where

$$-V_i(b) \gtrsim \beta^2 (1 - \gamma_t) 1/a, \quad i \in \mathcal{W}_n, \tag{E.178}$$

$$-V_i(b) \leq \frac{1}{\beta^2 + 1} V_j(b), \quad i \in \mathcal{U}_n, j \in \mathcal{W}_n, \tag{E.179}$$

$$|V_i(b)| \lesssim \sqrt{\frac{\log B}{B}} \cdot \frac{1}{a}, \quad i \notin \mathcal{W}_n \cup \mathcal{U}_n. \tag{E.180}$$

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

$$\mathbf{x}_j^n = \boldsymbol{\mu}_a + \kappa_j^n \boldsymbol{\nu}_b, \tag{E.181}$$

we have

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_V^{(t)}} \mathbf{p}_j^n \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{k=1}^m a_k \mathbb{1}[\mathbf{W}_{O_{(k,\cdot)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \\
&\quad \cdot \mathbf{W}_{O_{(k,\cdot)}}^\top \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n\top} \mathbf{p}_j^n,
\end{aligned} \tag{E.182}$$

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \sum_{b=0}^{t_0} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_V^{(b)}} \mathbf{p}_j^n \\
&= \eta \sum_{b=0}^{t_0} \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \notin \mathcal{W} \cup \mathcal{U}} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \right),
\end{aligned} \tag{E.183}$$

where

$$-V_i(b) \gtrsim \beta^2 (1 - \gamma_t) 1/a, \quad i \in \mathcal{U}_n, \tag{E.184}$$

$$-V_i(b) \leq \frac{1}{\beta^2 + 1} V_j(b), \quad i \in \mathcal{W}_n, j \in \mathcal{U}_n, \tag{E.185}$$

$$|V_i(b)| \lesssim \sqrt{\frac{\log B}{B}} \cdot \frac{1}{a}, \quad i \notin \mathcal{W}_n \cup \mathcal{U}_n. \tag{E.186}$$

We can also derive

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{b=0}^{t_0} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_V^{(t)}} (\mathbf{v}_k^\top, \mathbf{0}^\top)^\top \\
&= \eta \frac{1}{B} \sum_{b=0}^{t_0} \sum_{n \in \mathcal{B}_b} (-z^n) \sum_{k=1}^m a_k \mathbb{1}[\mathbf{W}_{O_{(k,\cdot)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \\
&\quad \cdot \mathbf{W}_{O_{(k,\cdot)}}^\top \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \mathbf{p}_s^{n\top} (\boldsymbol{\nu}_k^\top, \mathbf{0}^\top)^\top \\
&:= \eta \sum_{b=0}^{t_0} \left( \sum_{i \in \mathcal{W}_n} V'_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V'_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \notin \mathcal{W} \cup \mathcal{U}} V'_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \right),
\end{aligned} \tag{E.187}$$

where

$$|V'_i(b)| \leq |V_i(b)| \cdot \frac{1}{M_2}, \tag{E.188}$$

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

□

### E.5.7 Proof of Lemma E.4.7

*Proof.*

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

We have that

$$\begin{aligned}
& \mathbf{W}_V^{(t)} \mathbf{p}_s^n \\
&= \delta(\mathbf{p}_s^{n\top}, \mathbf{0}^\top)^\top + \sum_{b=0}^{t-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O(i,\cdot)}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O(i,\cdot)}^{(b)} \right. \\
&\quad \left. + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O(i,\cdot)}^{(b)} \right)^\top.
\end{aligned} \tag{E.190}$$

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

$$\mathbf{x}_s^n = \boldsymbol{\mu}_a + \kappa_s^n \boldsymbol{\nu}_b. \tag{E.191}$$

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

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{a} \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) (\delta \mathbf{p}_s^{n\top} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top + \sum_{b=0}^{t-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \right. \\
&\quad \left. \cdot \mathbf{W}_{O(i,\cdot)}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O(i,\cdot)}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O(i,\cdot)}^{(b)} \right)^\top (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top) \\
&\geq \frac{\alpha \eta}{2a} \delta(\beta^2 + 1).
\end{aligned} \tag{E.192}$$

Then, we have the following results by Lemma E.4.6, and the magnitude of  $\boldsymbol{\mu}_a$ ,  $\boldsymbol{\mu}_b$ , and  $\mathbf{q}$ ,

$$\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}} (\boldsymbol{\mu}_a^\top, -\mathbf{q}^\top)^\top \leq \frac{\beta^2 - 1}{\beta^2 + 1} \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top, \quad (\text{E.193})$$

$$\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}} (\boldsymbol{\mu}_b^\top, \mathbf{q}^\top)^\top \leq \frac{1}{\beta^2 + 1} \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top, \quad (\text{E.194})$$

$$\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}} (\boldsymbol{\mu}_b^\top, -\mathbf{q}^\top)^\top \leq -\frac{1}{\beta^2 + 1} \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top, \quad (\text{E.195})$$

$$\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}} (\boldsymbol{\nu}_c^\top, \mathbf{0}^\top)^\top \leq \frac{1}{M_2} \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top. \quad (\text{E.196})$$

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

$$\begin{aligned} & \eta \frac{1}{|\mathcal{B}_b^n|} \sum_{n \in \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top \\ &= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{a} \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) (\delta \mathbf{p}_s^{n\top} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top + \sum_{b=0}^{t-1} \eta (\sum_{i \in \mathcal{W}_n} V_i(b) \\ &\quad \cdot \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)})^\top (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top) \\ &\geq \frac{\alpha}{2} \left( \frac{\eta}{a} (\delta(\beta^2 + 1) - \frac{\eta}{a} \cdot \eta \cdot \sqrt{\frac{\log B}{B}} \frac{m}{a} \cdot \xi \log M_1 \right. \\ &\quad \left. + \frac{\eta}{a} \cdot \eta \frac{m}{a} (\beta^2(1 - \gamma_t)) \left( \frac{\eta}{a} \delta(\beta^2 + 1) - \xi \right) \right) \\ &\gtrsim \frac{\alpha}{2} \left( \frac{\eta}{a} (\delta(\beta^2 + 1) + \frac{\eta}{a} \cdot \eta \frac{m}{a} (\beta^2(1 - \gamma_t)) \frac{\eta}{a} \delta(\beta^2 + 1)) \right) \\ &\gtrsim \delta(\beta^2 + 1) \frac{\alpha \eta}{2a} \left( 1 + \frac{\eta^2 m}{a^2} \right), \end{aligned} \quad (\text{E.197})$$

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

$$\begin{aligned} & \eta \frac{1}{|\mathcal{B}_b - \mathcal{B}_b^n|} \sum_{n \in \mathcal{B}_b - \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top \\ &\gtrsim \delta(\beta^2 + 1) \frac{\alpha \eta}{2a} \left( 1 + \frac{\eta^2 m}{a^2} \right). \end{aligned} \quad (\text{E.198})$$

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

$$\begin{aligned} & \eta \frac{1}{|\mathcal{B}_b^n|} \sum_{n \in \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}} (\boldsymbol{\mu}_a^\top, -\mathbf{q}^\top)^\top \\ & \gtrsim \delta(\beta^2 + 1) \frac{\alpha\eta}{2a} \left(1 + \frac{\eta^2 m}{a^2}\right), \end{aligned} \quad (\text{E.199})$$

$$\begin{aligned} & \eta \frac{1}{|\mathcal{B}_b - \mathcal{B}_b^n|} \sum_{n \in \mathcal{B}_b - \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}} (\boldsymbol{\mu}_a^\top, -\mathbf{q}^\top)^\top \\ & \gtrsim \delta(\beta^2 + 1) \frac{\alpha\eta}{2a} \left(1 + \frac{\eta^2 m}{a^2}\right), \end{aligned} \quad (\text{E.200})$$

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

$$\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}^{(t)}} (\mathbf{p}_j^{n\top}, \mathbf{0})^\top \leq \eta \sqrt{\frac{\log B}{B}} \frac{1}{a}. \quad (\text{E.201})$$

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  $\mathbf{p}_{query}^n$  corresponding to  $\mathbf{q}$  and  $\boldsymbol{\mu}_a$ ,

$$\begin{aligned} & \eta \frac{1}{|\mathcal{B}_b^n|} \sum_{n \in \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}^{(t)}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top \\ & = \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{a} \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) (\delta \mathbf{p}_s^{n\top} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top + \sum_{b=0}^{t-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \right. \\ & \quad \cdot \mathbf{W}_{O(i,\cdot)}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O(i,\cdot)}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O(i,\cdot)}^{(b)})^\top (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top) \\ & \gtrsim \delta(\beta^2 + 1) \frac{\alpha\eta}{2a} + \frac{\alpha\eta}{2a} \cdot \eta \sum_{b=0}^{t_0} \delta(\beta^2 + 1) \frac{\eta m}{a^2} \left(1 + \frac{\eta^2 m}{a^2}\right)^b \\ & = \delta(\beta^2 + 1) \frac{\alpha\eta}{2a} \left(1 + \frac{\eta^2 m}{a^2}\right) \cdot \frac{(1 + \frac{\eta^2 m}{a^2})^{t_0+1} - 1}{\frac{\eta^2 m}{a^2}} \\ & = \delta(\beta^2 + 1) \frac{\alpha\eta}{2a} \left(1 + \frac{\eta^2 m}{a^2}\right)^{t_0+1}, \end{aligned} \quad (\text{E.202})$$

where the first inequality is by plugging the condition in the induction. The last two steps

come from basic mathematical computation. Then,

$$\begin{aligned}
& \eta \frac{1}{|\mathcal{B}_b^n|} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}^{(t)}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top \\
& \gtrsim \delta(\beta^2 + 1) \sum_{b=0}^{t_0+1} \frac{\alpha \eta}{2a} \left(1 + \frac{\eta^2 m}{a^2}\right)^b \\
& \gtrsim \delta(\beta^2 + 1) \frac{\alpha \eta}{2a} (t_0 + 1),
\end{aligned} \tag{E.203}$$

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

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}^{(t)}} (\boldsymbol{\mu}_a^\top, \mathbf{q}^\top)^\top \\
& \gtrsim \frac{1}{M_1} \cdot \delta \beta^2 (\beta^2 + 1) \frac{\alpha \eta}{2a} (t_0 + 1).
\end{aligned} \tag{E.204}$$

By Lemma E.4.10, when  $t \geq \Theta(1)$ , we have

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b - \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}^{(t)}} (\boldsymbol{\mu}_b^\top, \mathbf{q}^\top)^\top \\
& \gtrsim \frac{M_1 - 1}{M_1} \cdot \delta(\beta^2 + 1) \frac{\alpha \eta}{2a} (t_0 + 1).
\end{aligned} \tag{E.205}$$

Hence,

$$\begin{aligned}
& \eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}^{(t)}} (\boldsymbol{\mu}_b^\top, \mathbf{q}^\top)^\top \\
& \gtrsim \delta(\beta^2 + 1) \frac{\alpha \eta}{2a} (t_0 + 1),
\end{aligned} \tag{E.206}$$

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

$$\eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}^{(t)}} (\boldsymbol{\nu}_c^\top, \mathbf{0}^\top)^\top \leq \frac{1}{BT} \eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}^{(t)}} (\boldsymbol{\mu}_b^\top, \mathbf{q}^\top)^\top. \tag{E.207}$$

For  $i \in \mathcal{U}_n$  and  $\mathbf{p}_{query}^n$  corresponding to  $-\mathbf{q}$  and  $\boldsymbol{\mu}_a$ , similarly to (E.204), (E.205), and (E.206),

we have

$$\begin{aligned} & \eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}} (\boldsymbol{\mu}_a^\top, -\mathbf{q}^\top)^\top \\ & \gtrsim \frac{1}{M_1} \cdot \delta(\beta^2 + 1) \frac{\alpha\eta}{2a} (t_0 + 1), \end{aligned} \quad (\text{E.208})$$

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

$$\begin{aligned} & \eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b - \mathcal{B}_b^n} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}} (\boldsymbol{\mu}_b^\top, -\mathbf{q}^\top)^\top \\ & \gtrsim \frac{M_1 - 1}{M_1} \delta(\beta^2 + 1) \frac{\alpha\eta}{2a} (t_0 + 1), \end{aligned} \quad (\text{E.209})$$

$$\begin{aligned} & \eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}} (\boldsymbol{\mu}_b^\top, -\mathbf{q}^\top)^\top \\ & \gtrsim \delta(\beta^2 + 1) \frac{\alpha\eta}{2a} (t_0 + 1), \end{aligned} \quad (\text{E.210})$$

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

$$\begin{aligned} & \eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}} (\boldsymbol{\nu}_c^\top, \pm \mathbf{q}^\top)^\top \\ & \leq \frac{1}{M_2} \eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}} (\boldsymbol{\mu}_b^\top, \mathbf{q}^\top)^\top. \end{aligned} \quad (\text{E.211})$$

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

$$\|\mathbf{W}_{O(i,\cdot)}^{(t_0+1)}\| \gtrsim \sqrt{M_1} \delta(\beta^2 + 1)^{\frac{1}{2}} \frac{\alpha\eta}{2a} (t_0 + 1). \quad (\text{E.212})$$

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

$$\eta \frac{1}{B} \sum_{b=0}^{t_0+1} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O(i,\cdot)}} (\boldsymbol{\mu}_a^\top, -\mathbf{q}^\top)^\top \leq \eta \sqrt{\frac{\log B(t_0 + 1)}{B(t_0 + 1)}} \frac{1}{a}. \quad (\text{E.213})$$

□

### E.5.8 Proof of Lemma E.4.9

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

$$\Pr(i \in \mathcal{W}_n) = A_{d_{\mathcal{X}}}^{\text{cap}}(\phi)/A_{d_{\mathcal{X}}}, \quad (\text{E.214})$$

where  $A_{d_{\mathcal{X}}}$  is the surface area of an  $d_{\mathcal{X}} - 1$ -sphere.  $A_{d_{\mathcal{X}}}^{\text{cap}}(\phi)$  is the surface area of a  $d_{\mathcal{X}} - 1$ -spherical cap with  $\phi$  as the colatitude angle. By Equation 1 in [259], we have

$$\Pr(i \in \mathcal{W}_n) = \frac{1}{2} I_{\sin^2 \phi} \left( \frac{d_{\mathcal{X}} - 1}{2}, \frac{1}{2} \right) = \frac{\int_0^{\sin^2 \phi} t^{\frac{d_{\mathcal{X}}-3}{2}} (1-t)^{-\frac{1}{2}} dt}{2 \int_0^1 t^{\frac{d_{\mathcal{X}}-3}{2}} (1-t)^{-\frac{1}{2}} dt}, \quad (\text{E.215})$$

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

$$\phi \leq \pi/2 - \Theta(1/M_1), \quad (\text{E.216})$$

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

$$\begin{aligned} & \frac{\int_0^{\sin^2 \phi} t^{\frac{d_{\mathcal{X}}-3}{2}} (1-t)^{-\frac{1}{2}} dt}{\int_0^1 t^{\frac{d_{\mathcal{X}}-3}{2}} (1-t)^{-\frac{1}{2}} dt} \\ & \geq \frac{\int_0^{\cos^2 1/M} t^{\frac{d_{\mathcal{X}}-3}{2}} (1-t)^{-\frac{1}{2}} dt}{\int_0^1 t^{\frac{d_{\mathcal{X}}-3}{2}} (1-t)^{-\frac{1}{2}} dt} \\ & \geq 1 - \frac{\int_{1-1/M^2}^1 t^{\frac{d_{\mathcal{X}}-3}{2}} (1-t)^{-\frac{1}{2}} dt}{\int_0^1 t^{\frac{d_{\mathcal{X}}-3}{2}} (1-t)^{-\frac{1}{2}} dt} \\ & \geq 1 - \frac{\int_{1-1/M^2}^1 (1-t)^{-\frac{1}{2}} dt}{\int_{1-1/M}^1 \Theta(1) \cdot (1-t)^{-\frac{1}{2}} dt} \\ & = 1 - \frac{\frac{2}{M}}{\Theta(1) \cdot (\frac{2}{\sqrt{M}} - \frac{2}{M})} \\ & \geq \Theta(1), \end{aligned} \quad (\text{E.217})$$

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  $\mathbf{W}_{O_{(\cdot,\cdot)}}^{(0)}$ , we can derive a similar result by

replacing  $d_{\mathcal{X}}$  with  $d_{\mathcal{Y}}$  in (E.217). This implies that

$$|\mathcal{W}_n| \geq \Omega(1) \cdot \Omega(1) \cdot m \geq \Omega(m). \quad (\text{E.218})$$

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

$$|\mathcal{W}|, |\mathcal{U}| \geq \Omega(m). \quad (\text{E.219})$$

□

### E.5.9 Proof of Lemma E.4.10

*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  $\mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{W}_{O_{(i,:)}}^{(t)} \top > 0$  and  $\sum_{b=0}^{t-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \mathbf{W}_{O_{(i,:)}}^{(t)} \top > 0$  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  $\mathcal{W}_n$

$$\mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{V}^n(0) > 0, \quad (\text{E.220})$$

$$\mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{W}_{O_{(i,:)}}^{(t)} \top > 0. \quad (\text{E.221})$$

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

$$\begin{aligned} & \mathbf{W}_{O_{(i,:)}}^{(t)} \mathbf{V}^n(t) \\ &= \mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{W}_{O_{(i,:)}}^{(t)} \top + \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{a} \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) (\delta \mathbf{p}_s^n \top \mathbf{W}_{O_{(i,:)}}^{(t)} \top \\ & \quad + \sum_{b=0}^{t-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \right) \top \mathbf{W}_{O_{(i,:)}}^{(t)} \top). \end{aligned} \quad (\text{E.222})$$

By (E.189) and definition of  $\mathbf{W}_n$ , we have

$$\begin{aligned} & \mathbf{W}_{O_{(i,\cdot)}}^{(0)} \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (+z^n) a_i \mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^\top \mathbf{W}_Q \mathbf{p}_{query}^n) \geq 0] \\ & \cdot \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t-1)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t-1)\top} \mathbf{W}_Q^{(t-1)} \mathbf{p}_{query}^n) \\ & > 0. \end{aligned} \quad (\text{E.223})$$

Hence,

$$\mathbf{W}_{O_{(i,\cdot)}}^{(0)} \mathbf{W}_{O_{(i,\cdot)}}^{(t)}{}^\top = \sum_{b=0}^{t-1} \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \mathbf{W}_{O_{(i,\cdot)}}^{(t)}{}^\top > 0, \quad (\text{E.224})$$

and

$$\sum_{b=0}^{t-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \mathbf{W}_{O_{(i,\cdot)}}^{(t)}{}^\top > 0. \quad (\text{E.225})$$

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

$$\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{a} \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \delta \mathbf{p}_s^{n\top} \mathbf{W}_{O_{(i,\cdot)}}^{(t)}{}^\top > 0. \quad (\text{E.226})$$

Denote  $\theta_n^i$  as the angle between the feature embeddings of  $\mathbf{V}^n(0)$  and  $\mathbf{W}_{O_{(i,\cdot)}}^{(0)}$ . Since that the feature embedding of  $\mathbf{W}_{O_{(i,\cdot)}}^{(0)}$  is initialized uniformed on the  $d_{\mathcal{X}} - 1$ -sphere, we have  $\mathbb{E}[\theta_n^i] = 0$ . By Hoeffding's inequality (E.5), we have

$$\left\| \frac{1}{|\mathcal{W}_n|} \sum_{i \in \mathcal{W}_n} \theta_n^i - \mathbb{E}[\theta_n^i] \right\| = \left\| \frac{1}{|\mathcal{W}_n|} \sum_{i \in \mathcal{W}_n} \theta_n^i \right\| \leq \sqrt{\frac{\log M_1}{m}}, \quad (\text{E.227})$$

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

$$\left\| \frac{1}{|\mathcal{W}_n|} \sum_{i \in \mathcal{W}_n} \theta_n^i - \mathbb{E}[\theta_n^i] \right\| \leq \Theta\left(\frac{1}{M_1}\right). \quad (\text{E.228})$$

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

$$\mathbf{W}_{O_{(i,:)}}^{(0)} \sum_{b=0}^{t-1} \sum_{i \in \mathcal{W}_n} \mathbf{W}_{O_{(i,:)}}^{(b)}^\top > 0. \quad (\text{E.229})$$

Given  $B \gtrsim M_1 \log M_1$ , by Lemma E.4.6, and combining (E.229), we have that

$$\begin{aligned} & \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{a} \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \sum_{b=0}^{t-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \right. \\ & \left. + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,:)}}^{(b)} \right)^\top \mathbf{W}_{O_{(i,:)}}^{(t)} \\ & > 0. \end{aligned} \quad (\text{E.230})$$

Therefore, the conclusion holds when  $t = 1$ .

Suppose that the conclusion holds when  $t \leq t_0$ . When  $t = t_0 + 1$ , by (E.222), we can check that

$$\begin{aligned} & \mathbf{W}_{O_{(i,:)}}^{(0)} \mathbf{W}_{O_{(i,:)}}^{(t)\top} \\ &= \mathbf{W}_{O_{(i,:)}}^{(0)} (\mathbf{W}_{O_{(i,:)}}^{(0)\top} + \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (+z^n) a_i \mathbb{1} [\mathbf{W}_{O_{(i,:)}} \sum_{s=1}^{l+1} (\mathbf{W}_V \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t-1)\top} \\ & \quad \mathbf{W}_Q^{(t-1)} \mathbf{p}_{query}^n) \geq 0] \cdot \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t-1)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t-1)\top} \mathbf{W}_Q^{(t-1)} \mathbf{p}_{query}^n) \\ & > 0 + 0 = 0, \end{aligned} \quad (\text{E.231})$$

where the second 0 comes from (E.226) and the conditions that such conclusion in (E.231) holds when  $t \leq t_0$ . Combining (E.206) and the fact that the weighted summation of  $\mathbf{p}_s^n$  is close to be in the direction of  $\boldsymbol{\mu}_j$  and  $\mathbf{q}$  in the feature label embeddings, respectively, where  $\boldsymbol{\mu}_j$  is the IDR pattern of the  $\mathbf{p}_{query}$ , we have

$$\eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{a} \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^n \top \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \delta \mathbf{p}_s^n \top \mathbf{W}_{O_{(i,:)}}^{(t)\top} > 0, \quad (\text{E.232})$$

$$\begin{aligned}
& \sum_{b=0}^{t-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \mathbf{W}_{O_{i,\cdot}}^{(t)} \top \\
& = \left( \sum_{b=0}^{t_0-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \eta \sum_{i \in \mathcal{W}_n} V_i(t_0) \mathbf{W}_{O_{(i,\cdot)}}^{(t_0)} \right) (\mathbf{W}_{O_{(i,\cdot)}}^{(t_0)} \\
& \quad + \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^{(t)}} \Big| t = t_0) \top \\
& > 0 + \left( \sum_{b=0}^{t_0-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \eta \sum_{i \in \mathcal{W}_n} V_i(t_0) \mathbf{W}_{O_{(i,\cdot)}}^{(t_0)} \right) \\
& \quad \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\tilde{\mathbf{P}}^n, z^n; \Psi)}{\partial \mathbf{W}_{O_{(i,\cdot)}}^{(t)}} \Big| t = t_0 \tag{E.233} \\
& = \left( \sum_{b=0}^{t_0-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(t_0)} \stackrel{(b)}{=} + \eta \sum_{i \in \mathcal{W}_n} V_i(t_0) \mathbf{W}_{O_{(i,\cdot)}}^{(t_0)} \right) \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (-z^n) a_i \mathbb{1}[\mathbf{W}_{O_{(i,\cdot)}}^{(t_0)} \\
& \quad \sum_{s=1}^{l+1} (\mathbf{W}_V^{(t_0)} \mathbf{p}_s^n) \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t_0)\top} \mathbf{W}_Q^{(t_0)} \mathbf{p}_{query}^n) \geq 0] \cdot \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t_0)\top} \\
& \quad \mathbf{W}_Q^{(t_0)} \mathbf{p}_{query}^n) (\delta(\mathbf{p}_s^{n\top}, \mathbf{0}^\top)^\top + \sum_{b=0}^{t_0-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \right. \\
& \quad \left. + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \right)^\top), \\
& > 0,
\end{aligned}$$

where the first step is by the formula of the gradient descent, and the second step is by  $(\sum_{b=0}^{t_0-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \eta \sum_{i \in \mathcal{W}_n} V_i(t_0) \mathbf{W}_{O_{(i,\cdot)}}^{(t_0)}) \mathbf{W}_{O_{(i,\cdot)}}^{(t_0)} > 0$  from the induction steps. The last step comes from the fact that  $\|\sum_{b=0}^{t_0-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)}\|^2 > 0$  and  $\sum_{b=0}^{t_0-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \cdot \eta \sum_{i \in \mathcal{W}_n} V_i(t_0) \mathbf{W}_{O_{(i,\cdot)}}^{(t_0)\top} > 0$  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,

$$\sum_{b=0}^{t-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \right)^\top \mathbf{W}_{O_{(i,\cdot)}}^{(t)} \top > 0, \tag{E.234}$$

since the norm  $\mathbf{W}_{O_{(i,\cdot)}}^{(t)}$  for  $i \notin \mathcal{W}_n$  is no larger than that for  $i \in \mathcal{W}_n$ . Combining (E.231), (E.232), and (E.234), we have

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{V}^n(t) > 0. \tag{E.235}$$

Hence, we finish the induction.

(2) When  $t \gtrsim \Theta(\beta)$ , for  $i \in \mathcal{W}$ , by checking (E.222), we can deduce that

$$\mathbf{W}_{O_{(i,\cdot)}}^{(0)} \mathbf{W}_{O_{(i,\cdot)}}^{(t)}{}^\top + \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{a} \sum_{s=1}^{l+1} \text{softmax}(\mathbf{p}_s^{n\top} \mathbf{W}_K^{(t)\top} \mathbf{W}_Q^{(t)} \mathbf{p}_{query}^n) \delta \mathbf{p}_s^{n\top} \mathbf{W}_{O_{(i,\cdot)}}^{(t)}{}^\top > 0, \quad (\text{E.236})$$

since that the accumulated label embedding term of  $\mathbf{W}_{O_{(i,\cdot)}}^{(t)}$  contributed positively to  $\mathbf{p}_s^{n\top} \mathbf{W}_{O_{(i,\cdot)}}^{(t)\top}$  and is larger than that of the feature embedding contribution by (E.132) and (E.133) (the gradient updates is close in the direction of the IDR pattern of  $\mathbf{p}_{query}^n$  when  $m \gtrsim M_1^2$ ). Since that  $\|\mathbf{W}_{O_{(i,\cdot)}}^{(0)}(\boldsymbol{\mu}_j^\top, \mathbf{0}^\top)^\top\| \leq \beta\xi$  for any  $j \in [M_1]$ , the effect of  $\mathbf{W}_{O_{(i,\cdot)}}^{(0)} \mathbf{W}_{O_{(i,\cdot)}}^{(t)}{}^\top$  to the sign is much smaller than the remaining terms in (E.236). Hence, we show (E.236).

Then, since that the label embedding of  $\mathbf{W}_{O_{(i,\cdot)}}$ ,  $\mathbf{W}_{O_{(j,\cdot)}}$  are both close to  $\mathbf{q}$  for  $i, j \in \mathcal{W}$ , and that the feature embedding of  $\mathbf{W}_{O_{(i,\cdot)}}, i \in \mathcal{W}_n$  is close to the IDR pattern of  $\mathbf{p}_{query}^n$ , which is not the negative direction of the feature embedding of  $\mathbf{W}_{O_{(j,\cdot)}}, j \in \mathcal{W}_{n'}$ , we have for  $j \in \mathcal{W} \setminus \mathcal{W}_n$ ,

$$\sum_{b=0}^{t-1} \eta \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)}{}^\top \mathbf{W}_{O_{(j,\cdot)}}^{(t)}{}^\top > 0. \quad (\text{E.237})$$

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  $\mathbf{W}_{O_{(i,\cdot)}}^{(t)}$  for  $i \notin \mathcal{W}_n$  is no larger than that for  $i \in \mathcal{W}_n$ , we have

$$\sum_{b=0}^{t-1} \eta \left( \sum_{i \in \mathcal{W}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \in \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} + \sum_{i \notin \mathcal{W}_n \cup \mathcal{U}_n} V_i(b) \mathbf{W}_{O_{(i,\cdot)}}^{(b)} \right)^\top \mathbf{W}_{O_{(j,\cdot)}}^{(t)}{}^\top > 0. \quad (\text{E.238})$$

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

$$\mathbf{W}_{O_{(i,\cdot)}}^{(t)} \mathbf{V}^n(t) > 0. \quad (\text{E.239})$$

□

## APPENDIX F

### APPENDIX OF CHAPTER 7

#### F.1 Additional Discussions

##### F.1.1 The Motivation to Study One-Layer Single-Head Transformers

The reasons we study one-layer single-head attention-only nonlinear Transformers in this work are as follows.

First, it is much more challenging to theoretically analyze the training dynamics and generalization of multi-layer/head Transformers. This is because the loss landscape for multi-layer/head Transformers is highly nonlinear and non-convex due to the interactions between multiple nonlinear functions. The simplified data helps to characterize the gradient updates in different directions for different patterns and steps. Non-orthogonal data make the updates less separable for different inputs, which is more challenging to analyze.

Second, the state-of-the-art theoretical works [6], [23], [219], [271], [250] on optimization and generalization also focus mainly on one-layer Transformers. No existing works study the optimization and generalization of CoT even for one-layer Transformers. Therefore, we plan to focus on the one-layer analysis to obtain more theoretical insights. We leave the theoretical analysis of the multi-layer case as future works.

Third, although we admit the gap between theory and practice, our theory still makes contributions under our settings. Our work is the first one to investigate the optimization and generalization of CoT and characterize the conditions when CoT is better than ICL. We establish the required number of context examples for a successful CoT in terms of how informative and erroneous the prompt is.

We also implement experiments on the attention mechanism for three-layer two-head Transformers on two-step reasoning tasks. Please see Figure 7.7 for details. The findings of all three layers are generally consistent with Proposition 3 for the single-layer single-head case, which indicates that the CoT mechanism we characterize on one-layer Transformers can be extended to multi-layer multi-head Transformers.

---

Portions of this appendix previously appeared as: H. Li, M. Wang, S. Lu, X. Cui, and P.-Y. Chen, “Training nonlinear transformers for chain-of-thought inference: A theoretical generalization analysis,” 2024, *arXiv:2410.02167*.

### F.1.2 The Motivation of the Data and Task Formulation

There are several reasons for using such data formulation.

First, our data formulation of orthogonal patterns, on which the function is based, is widely used in the state-of-the-art theoretical study of model training or ICL on language and sequential data [104], [219], [23], [246]. For example, [219], [23] study ICL on regression or classification tasks, which also use orthogonal patterns as data. Sections 2.1 and 2.2 in [246] consider learning n-gram data in ICL by formulating transitions between orthogonal patterns. Section 3 of [104] also assume orthogonal patterns in Transformer model training, and the generation comes from the orthogonal pattern set. The data formulation we use is consistent with the existing theoretical works.

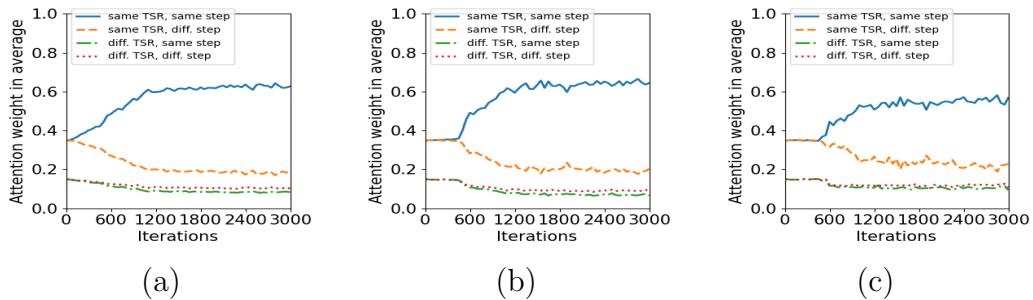
Second, based on this formulation, one can characterize the gradient updates in different directions for different patterns and steps. This enables us to distinguish the impact of different patterns and steps in the convergence analysis of CoT using Transformers. Non-orthogonal data make the model updates less separable for different inputs, which is more challenging to analyze. Moreover, we would like to mention that during the inference, the tokens in testing prompts contain noises as defined in Equation 10. This makes the tokens of different TSR patterns not orthogonal to each other and relaxes our orthogonality condition to some degree.

### F.1.3 The Discussion of Positional Encoding

The positional encoding (PE) we use is simplified for theoretical analysis. The formulation of PE we use is motivated by [248], [251], where each token is added with a PE represented by orthogonal vectors. These works formulate the distribution of the PE to be related to the structure of the data, such as patch-wise association [248], and sparse token selection [251]. Likewise, we follow their intuition to make the PE vary in different steps of our reasoning tasks so that the Transformer can distinguish different steps when making inferences for the query.

Our analysis can be extended to study more general PEs with additional technical work in the future. One possible direction is studying the family of periodic and separable PE. For example, the absolute PE proposed by [1] considers PE as a sinusoid, which is periodic. Such analysis can be made by relaxing the “orthogonality” of PE vectors to a certain “separability” between PE vectors.

We also conduct experiments on a three-layer single-head Transformer with the standard PE proposed in Section 3.5 of [1] for our problem. Figure shows that the blue curve increases to be the largest along the training, which means the attention weights on example steps that share the same TSR pattern and the same step as the query. This indicates that the CoT mechanism of using standard PE is the same as the one proposed in Proposition 3 in our paper. One might note that the scores of the blue curve are not as high as Figure 7.6 in our paper. We guess the reason why the distinction in attention values is more significant in our PE may be the additional orthogonality of our PE and the property that its period is the same as the reasoning length. Nevertheless, the strong similarity between the results on standard PE and our used PE shows the practical significance of our analysis.



**Figure F.1:** CoT mechanism with standard PE of (a) Layer 1 (b) Layer 2 (c) Layer 3.

## F.2 Algorithms

We first present the training algorithm introduced in Section 7.2.2.

We then summarize the algorithm of the CoT inference introduced in Section 7.2.3 as follows.

## F.3 Preliminaries

We first summarize the notations we use in this chapter in Tables F.1 and F.2.

**Lemma F.3.1** (Multiplicative Chernoff bounds, Theorem D.4 of [270]). *Let  $X_1, \dots, X_m$  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  $\hat{p} = \frac{1}{m} \sum_{i=1}^m X_i$ :*

$$\Pr(\hat{p} \geq (1 + \gamma)p) \leq e^{-\frac{mp\gamma^2}{3}}, \quad (\text{F.3})$$

---

**Subroutine 3** Training with Stochastic Gradient Descent (SGD)

---

- 1: **Hyperparameters:** The step size  $\eta$ , the number of iterations  $T$ , batch size  $B$ .
- 2: **Initialization:** Let  $\mathbf{W} = \mathbf{W}_K^\top \mathbf{W}_Q$  and  $\mathbf{W}_V = (\mathbf{0}_{d_X \times d_X} \ \mathbf{I}_{d_X} \ \mathbf{0}_{d_X \times d_\varepsilon})$ . Each entry of  $\mathbf{W}^{(0)}$  is generated from  $\mathcal{N}(0, \xi^2)$  for a small constant  $\xi > 0$ .  $\mathbf{W}_V$  and  $\mathbf{a}$  are fixed during the training.
- 3: **Training by SGD:** For each iteration, we independently sample  $\mathbf{x}_{query} \sim \mathcal{D}$ ,  $f \in \mathcal{T}_{tr}$  to form a batch of training prompt and labels  $\{\mathbf{P}^n, z^n\}_{n \in \mathcal{B}_t}$  as introduced in Section 7.3.2. Each TRR pattern is sampled equally likely in each batch. For each  $t = 0, 1, \dots, T - 1$

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

- 4: **Output:**  $\mathbf{W}^{(T)}$ .
- 

**Subroutine 4** Inference with Chain-of-Thought (CoT)

---

- 1: **Input:**  $\mathbf{z}_0 = \mathbf{v}_0 = \mathbf{x}_{query}$ ,  $\mathbf{P}_0$ , and  $\mathbf{P}_1$ .
- 2: **for**  $k = 1, \dots, K - 1$ , **do**
- 2:

Compute  $\mathbf{v}_k$  by greedy decoding in (7.5). Then update  $\mathbf{P}_k$  and  $\mathbf{P}_{k+1}$  by (7.6).  
(F.2)

- 3: **end for**
  - 4: **Output:**  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{K-1}$ , and  $\mathbf{v}_K$  by (7.5).
- 

$$\Pr(\hat{p} \leq (1 - \gamma)p) \leq e^{-\frac{mp\gamma^2}{2}}. \quad (\text{F.4})$$

**Definition F.3.2** ([258]). 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 F.3.3** ([258] Proposition 5.1, hoeffding's inequality). *Let  $X_1, X_2, \dots, X_N$  be independent centered sub-gaussian random variables, and let  $K = \max_i \|X_i\|_{\psi_2}$ . Then for every  $\mathbf{a} = (a_1, \dots, a_N) \in \mathbb{R}^N$  and every  $t \geq 0$ , we have*

$$\Pr\left(\left|\sum_{i=1}^N a_i X_i\right| \geq t\right) \leq e \cdot \exp\left(-\frac{ct^2}{K^2 \|\mathbf{a}\|^2}\right), \quad (\text{F.5})$$

where  $c > 0$  is an absolute constant.

**Definition F.3.4.** Define that for  $\tilde{\mathbf{p}}_i$  that shares the same TRR/TSR pattern and in the

**Table F.1: Summary of notations.**

Notations	Annotation
$\mathbf{x}_i, \mathbf{y}_{i,k}, \mathbf{x}_{query}, \mathbf{z}_k$	$\mathbf{x}_i$ is the input to the first step of a reasoning example. $\mathbf{y}_{i,k}$ is the $k$ -th step output label of $\mathbf{x}_i$ . $\mathbf{x}_{query}$ is the query input. $\mathbf{z}_k$ the $k$ -th step output label of $\mathbf{x}_{query}$ . $k \in [K]$ .
$\mathbf{P}, \mathbf{p}_{query}, \mathbf{E}_i, \mathbf{Q}_k, \mathbf{v}_k$	$\mathbf{P}$ is a training or testing prompt that consists of multiple training or testing examples and a query. The last column of $\mathbf{P}$ is denoted by $\mathbf{p}_{query}^n$ , which is the query of $\mathbf{P}$ . $\mathbf{E}_i$ is the $i$ -th context example of $\mathbf{P}$ . $\mathbf{Q}_k$ is the first $k$ steps of the reasoning query. $k \in [K]$ . $\mathbf{v}_k$ is the $k$ -th step generation by CoT. $k \in [K]$ .
$\mathbf{c}_i, \tilde{\mathbf{p}}_i, \tilde{\mathbf{p}}_{query}$	$\mathbf{c}_i$ is the positional encoding for the $i$ -th column of the input sequence. $\tilde{\mathbf{p}}_i = \mathbf{p}_i + \mathbf{c}_i$ , where $\mathbf{p}_i$ is the $i$ -th column of $\mathbf{P}$ . $\tilde{\mathbf{p}}_{query}$ is the $\mathbf{p}_i$ of the query column.
$F(\Psi; \mathbf{P}), \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)$	$F(\Psi; \mathbf{P}^n)$ is the Transformer output for $\mathbf{P}$ with $\Psi$ as the parameter. $\ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)$ is the loss function value given $\mathbf{P}^n$ and the corresponding label $\mathbf{z}^n$ .
$\boldsymbol{\mu}_i \in \mathcal{M}, \boldsymbol{\mu}'_i \in \mathcal{M}', \text{TSR}(\cdot)$	$\boldsymbol{\mu}_i$ is the $i$ -th training-relevant (TRR) pattern for $i \in [M]$ . $\boldsymbol{\mu}'_i$ is the $i$ -th testing-relevant (TSR) pattern for $i \in [M']$ . $\mathcal{M}$ and $\mathcal{M}'$ are the set of TRR and TSR patterns, respectively. $\text{TSR}(\cdot)$ is a function that outputs the index of the TSR pattern of the noisy input.
$f_k, f$	$f$ is the task function with $f = f_K \circ \dots \circ f_2 \circ f_1$ for a $K$ -steps reasoning. $f_k$ is the $k$ -th step task function.
$\mathcal{T}, \mathcal{T}', \mathcal{D}, \mathcal{D}'$	$\mathcal{T}$ is the distribution of training tasks, while $\mathcal{T}'$ is the distribution of testing tasks. $\mathcal{D}$ is the training data distribution. $\mathcal{D}'$ is the testing data distribution.
$\alpha, \alpha'$	$\alpha$ (or $\alpha'$ ) is the fraction of context examples with input sharing the same TRR (or TSR) pattern as the query.
$\mathbf{A}_k^f, \mathbf{B}_k^f$	$\mathbf{A}_k^f$ is the step-wise transition matrix at the $k$ -th step for the task $f$ , $k \in [K]$ . $\mathbf{B}_k^f$ is the $K$ -steps transition matrix of the task $f$ .

same step as the query,

$$p_n(t) = \sum_i \text{softmax}(\tilde{\mathbf{p}}_i^n \top \mathbf{W}^{(t)} \tilde{\mathbf{p}}_{query}^n). \quad (\text{F.6})$$

**Lemma F.3.5.** *Given the SGD training scheme described in Section 7.2.2,  $B \geq \Omega(M \log M)$ , and  $l_{tr} \geq \Omega(\alpha^{-1})$ , we have the following results. When  $\mathcal{O}(\eta^{-1}\alpha^{-2}K^3 \log \frac{K}{\epsilon}) \geq t \geq 1$ , for any*

**Table F.2: Summary of notations (Continued).**

Notations	Annotation
$\tau^f, \tau_o^f, \rho^f, \rho_o^f$	$\tau^f$ is the min-max trajectory transition probability for task $f$ . $\tau_o^f$ is the min-max input-label transition probability for task $f$ . $\rho^f$ and $\rho_o^f$ are primacy of the step-wise transition matrices and the $K$ -steps transition matrix, respectively.
$\mathcal{S}_k^*$	The index set of context columns of the prompt that correspond to the $k$ -th step of the example and share the same TSR pattern in the $(k-1)$ -th output as the $(k-1)$ -th output $\mathbf{v}_{k-1}$ of the query.
$p_n(t)$	$p_n(t)$ is the summation of attention weights on context columns that share the same TRR/TSR pattern and in the same step as the query.
$\mathcal{B}_b$	$\mathcal{B}_b$ is the SGD batch at the $b$ -th iteration.
$l_{tr}$	$l_{tr}$ is the universal number of training context examples.
$l_{ts}^f$	$l_{ts}^f$ is the number of testing context examples of the task $f$ .
$\mathcal{O}(), \Omega(), \Theta()$	We follow the convention that $f(x) = \mathcal{O}(g(x))$ (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 $f(x) \lesssim \mathcal{O}(g(x))$ ).

$\mathbf{p}$  as a column of context examples in (7.1), we have

$$\begin{aligned}
& \tilde{\mathbf{p}}^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(t)}} \tilde{\mathbf{p}} \\
& \leq \frac{\eta}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} (1 - p_n(t))^2 (-4p_n(t)(1 + \frac{\alpha^2}{K^2}) + \frac{\alpha^2}{K^2}(1 + \frac{2(K-1)}{K})) \right. \\
& \quad \left. - \frac{\alpha^2}{K^3} (1 - p_n(t))^2 \right). \tag{F.7}
\end{aligned}$$

For any  $\tilde{\mathbf{p}}'$  that shares the same TRR pattern and a different positional encoding as  $\tilde{\mathbf{p}}$ , we

have

$$\begin{aligned}
& \frac{\eta}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} (-4 - (3K - 2)(1 - p_n(t))(1 + \frac{\alpha^2}{K^2})) p_n(t) (1 - p_n(t)) \right. \\
& \quad \left. + \frac{\alpha^2}{K^3} (1 - p_n(t))^2 \right) \\
& \leq \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(t)}} \tilde{\mathbf{p}} \\
& \leq \frac{\eta}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} (-4 - (3K - 2)(1 - p_n(t))(1 + \frac{\alpha^2}{K^2})) p_n(t) (1 - p_n(t)) \right. \\
& \quad \left. + \frac{1}{K} p_n(t) (1 - p_n(t))^2 \cdot (1 + \frac{\alpha^2}{K^2}) \right).
\end{aligned} \tag{F.8}$$

For any  $\tilde{\mathbf{p}}'$  that shares a different TRR pattern but the same positional encoding as  $\tilde{\mathbf{p}}$ , we have

$$\begin{aligned}
& \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} \left( -\frac{\alpha^2}{K^2} + (K - 1 + \frac{(2K - 1)\alpha^2}{K^2}) p_n(t) \right) (1 - p_n(t))^2 \right. \\
& \quad \left. - (1 - p_n(t))^2 \frac{\alpha^2}{K^3} + \frac{1}{K} \cdot (1 - p_n(t))^2 \left( -p_n(t) + (1 - p_n(t)) \frac{\alpha^2}{K^2} \right) \right) \\
& \leq \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(t)}} \tilde{\mathbf{p}} \\
& \leq \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} \left( -\frac{\alpha^2}{K^2} + (K - 1 + \frac{(2K - 1)\alpha^2}{K^2}) p_n(t) \right) (1 - p_n(t))^2 \right. \\
& \quad \left. - (1 - p_n(t))^2 \frac{\alpha^2}{K^3} \right).
\end{aligned} \tag{F.9}$$

For any  $\tilde{\mathbf{p}}'$  that shares a different TRR pattern and a different positional encoding from  $\tilde{\mathbf{p}}$ , we have

$$\begin{aligned}
& \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} p_n(t) (1 - p_n(t))^2 \left( 1 + \frac{(2 - K)\alpha^2}{K^2} \right) + (1 - p_n(t))^2 \cdot \frac{\alpha^2}{K^3} \right) \\
& \leq \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(t)}} \tilde{\mathbf{p}} \\
& \leq \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} p_n(t) (1 - p_n(t))^2 \left( 2 - K + \frac{(2 - K)\alpha^2}{K^2} \right) \right. \\
& \quad \left. + (1 - p_n(t))^2 p_n(t) \left( 1 + \frac{\alpha^2}{K^2} \right) \cdot \frac{1}{K} \right).
\end{aligned} \tag{F.10}$$

**Lemma F.3.6.** Given the SGD training scheme described in Section 7.2.2,  $B \geq \Omega(M \log M)$ ,

and  $l_{tr} \geq \Omega(\alpha^{-1})$ , and

$$t \gtrsim T_1 := \eta^{-1} \alpha^{-2} K^3 \log \frac{K}{\epsilon}, \quad (\text{F.11})$$

we have that if  $\mathbf{p}_{query}$  is in the  $k$ -th step,

$$\sum_{i \in \mathcal{S}_{[K] \setminus k}} \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W}^{(t)} \tilde{\mathbf{p}}_{query}) \leq \epsilon \quad (\text{F.12})$$

where  $\mathcal{S}_{[K] \setminus k}$  means the index set of context columns that are not in the  $k$ -th step.

**Lemma F.3.7.** *Given the SGD training scheme described in Section 7.2.2,  $B \geq \Omega(M \log M)$ , and  $l_{tr} \geq \Omega(\alpha^{-1})$ , we have the following results. When  $t \geq T_1 = \eta^{-1} \alpha^{-2} K^3 \log \frac{K}{\epsilon}$ , for any  $\mathbf{p}$  as a column of context examples in (7.1), we have*

$$\tilde{\mathbf{p}}^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(t)}} \tilde{\mathbf{p}} \leq -\frac{\eta}{2MB} \sum_{n \in \mathcal{B}_b} 4p_n(t)(1 - p_n(t))^2. \quad (\text{F.13})$$

For any  $\tilde{\mathbf{p}}'$  that shares the same TRR pattern and a different positional encoding as  $\tilde{\mathbf{p}}$ , we have

$$\left| \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(t)}} \tilde{\mathbf{p}}' \right| \leq \eta \epsilon. \quad (\text{F.14})$$

For any  $\tilde{\mathbf{p}}'$  that shares a different TRR pattern but the same positional encoding as  $\tilde{\mathbf{p}}$ , we have

$$\left| \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(t)}} \tilde{\mathbf{p}}' \right| \leq \frac{\eta}{2BM} \sum_{n \in \mathcal{B}_b} p_n(b)(1 - p_n(b))^2. \quad (\text{F.15})$$

For any  $\tilde{\mathbf{p}}'$  that shares a different TRR pattern and a different positional encoding from  $\tilde{\mathbf{p}}$ , we have

$$\left| \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(t)}} \tilde{\mathbf{p}}' \right| \leq \eta \epsilon. \quad (\text{F.16})$$

## F.4 Proof of Main Theorems

### F.4.1 Proof of Theorem 3

*Proof.* By the condition in Lemma F.3.5, we have that

$$B \geq \Omega(M \log M). \quad (\text{F.17})$$

We know that there exists gradient noise caused by imbalanced TRR patterns in each batch. Then, by hoeffding's inequality (F.5),

$$\begin{aligned} & \Pr \left( \left\| \frac{1}{|\mathcal{B}_b|} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, z^n)}{\partial \mathbf{W}} - \mathbb{E} \left[ \frac{\partial \ell(\Psi; \mathbf{P}^n, z^n)}{\partial \mathbf{W}} \right] \right\| \geq \left| \mathbb{E} \left[ \frac{\partial \ell(\Psi; \mathbf{P}^n, z^n)}{\partial \mathbf{W}} \right] \epsilon \right| \right) \\ & \leq e^{-B\epsilon^2} \leq M^{-C}, \end{aligned} \quad (\text{F.18})$$

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

$$B \gtrsim \max\{\epsilon^{-2}, M\} \log M. \quad (\text{F.19})$$

By Lemma F.3.7 and Definition F.3.4, for  $\tilde{\mathbf{p}}_i^n$  that share the same TRR pattern and the same positional encoding of  $\tilde{\mathbf{p}}_{query}^n$ ,

$$\frac{p_n(t+1)}{|\mathcal{S}_1^n|} = \text{softmax}(\tilde{\mathbf{p}}_i^n \top \mathbf{W}^{(t+1)} \tilde{\mathbf{p}}_{query}^n) \geq \frac{1}{l} \cdot \frac{1}{\frac{\alpha}{K} + (1 - \frac{1}{K}) \cdot \epsilon + (\frac{1}{K} - \frac{\alpha}{K})e^{-u}}, \quad (\text{F.20})$$

where by (F.136),

$$u \gtrsim \frac{\eta}{KM} \sum_{b=0}^t (1 - p_n(b))^2 p_n(b). \quad (\text{F.21})$$

For  $\tilde{\mathbf{p}}_i^n$  that only share the same positional encoding of  $\tilde{\mathbf{p}}_{query}^n$ ,

$$\text{softmax}(\tilde{\mathbf{p}}_i^n \top \mathbf{W}^{(t+1)} \tilde{\mathbf{p}}_{query}^n) \geq \frac{1}{l} \cdot \frac{1}{\frac{\alpha}{K}e^u + (1 - \frac{1}{K}) \cdot \epsilon + (\frac{1}{K} - \frac{\alpha}{K})}. \quad (\text{F.22})$$

Therefore, to make the attention weights between  $\tilde{\mathbf{p}}_{query}^n$  and  $\tilde{\mathbf{p}}_i^n$  that share the same TRR pattern and the same positional encoding dominant, we need a large enough  $u$ . When  $1 - p_n(b) \geq \Omega(1)$ , we have

$$t \leq T_2 := \eta^{-1} KM \alpha^{-1}. \quad (\text{F.23})$$

When  $1 - p_n(b) \leq O(1)$ ,

$$p_n(t+1) = \frac{e^u}{e^u + \frac{1-\alpha}{\alpha}} \gtrsim 1 - \frac{1-\alpha}{\alpha} e^{-u}, \quad (\text{F.24})$$

and

$$1 - p_n(t+1) \geq \frac{1-\alpha}{\alpha e^u + (1-\alpha)} \gtrsim \frac{1-\alpha}{\alpha} e^{-u}. \quad (\text{F.25})$$

Then, we prove that when  $t$  is large enough,  $u(t) \geq \frac{1}{2} \log \frac{\eta(1-\alpha)^2 t}{\alpha^2 M}$ . We show it by induction. Suppose that the conclusion holds when  $t = t_0$ , then

$$\begin{aligned} u(t) &\geq \frac{\eta}{KM} \sum_{b=0}^{t_0} (1 - p_n(b))^2 p_n(b) + \frac{\eta}{KM} (1 - p_n(t))^2 p_n(t) \\ &\geq \frac{1}{2} \log \frac{(1-\alpha)^2 t}{2\alpha^2 KM} + \frac{\eta}{KM} (1 - p_n(t))^2 p_n(t) \\ &\geq \frac{1}{2} \log \frac{\eta(1-\alpha)^2 (t+1)}{\alpha^2 KM}, \end{aligned} \tag{F.26}$$

where the last step is by

$$\frac{1}{2} \log(1 + \frac{1}{t}) \leq \frac{1}{2t} \leq \frac{\eta}{KM} \cdot (\frac{1-\alpha}{\alpha})^2 e^{-\log \frac{\eta(1-\alpha)^2 t}{\alpha^2 KM}}. \tag{F.27}$$

To make  $(1 - p_n(t))^2 < \epsilon$ , we need

$$(\frac{1-\alpha}{\alpha})^2 e^{-2u} \leq \epsilon. \tag{F.28}$$

Then, we get

$$u \geq \frac{1}{2} \log \frac{1}{\epsilon} + \log \frac{1-\alpha}{\alpha}. \tag{F.29}$$

Therefore, by

$$\frac{1}{2} \log \frac{\eta t}{KM} + \log \frac{1-\alpha}{\alpha} \geq \frac{1}{2} \log \frac{1}{\epsilon} + \log \frac{1-\alpha}{\alpha}, \tag{F.30}$$

we finally obtain

$$t \geq T_3 := \eta^{-1} \epsilon^{-1} KM. \tag{F.31}$$

For  $\tilde{\mathbf{p}}_i^n$  that shares the same TSR pattern as the query, we have that when  $t = T_1$ ,

$$\tilde{\mathbf{p}}_i^n^\top \mathbf{W}^{(t)} \tilde{\mathbf{p}}_{query}^n \geq \log \frac{K}{\epsilon}. \tag{F.32}$$

When  $t = T_1 + T_2 + T_3$ ,

$$\tilde{\mathbf{p}}_i^n^\top \mathbf{W}^{(t)} \tilde{\mathbf{p}}_{query}^n \geq \Theta(1) \cdot \log \frac{K}{\epsilon} = \Theta(\log \frac{K}{\epsilon}). \tag{F.33}$$

Then,

$$\begin{aligned} T &:= T_1 + T_2 + T_3 \\ &= \Theta(\eta^{-1}\alpha^{-2}K^3 \log \frac{K}{\epsilon} + \eta^{-1}MK(\alpha^{-1} + \epsilon^{-1})). \end{aligned} \tag{F.34}$$

Therefore,

$$\mathbb{E}_{\mathbf{x}_{query} \sim \mathcal{D}, f \in \mathcal{T}} [\ell(\Psi; \mathbf{P}, \mathbf{z})] \leq \mathcal{O}(\epsilon). \tag{F.35}$$

□

#### F.4.2 Proof of Theorem 4

*Proof.* We know that  $\alpha'$  is the fraction of examples that share the same TSR pattern as the query. We need that in each step, the number of examples that share the same TSR pattern as the current step of the query is at least 1. Note that the probability of examples where each reasoning step produces the most probable output is

$$\prod_{k=1}^K A_{k(\text{TSR}(f_{k-1} \circ \dots \circ f_0(\boldsymbol{\mu}'_i)), \text{TSR}(f_k \circ \dots \circ f_0(\boldsymbol{\mu}'_i)))}^f, \text{ where } f_0(\boldsymbol{\mu}'_i) := \boldsymbol{\mu}'_i, \forall i \in [M'], \tag{F.36}$$

where the input to the first step has the TSR pattern  $\boldsymbol{\mu}'_i$ . Define  $m_{k(i)}$  as the TSR pattern in the  $k$ -th step output of the  $i$ -th context example by the transition matrix defined in 7.12. Consider that the TSR pattern of the  $k$ -th step label of the testing query is  $\boldsymbol{\mu}'_{q_k}$ , which is also the most probable  $k$ -th step output of the  $k$ -th step of a certain  $\mathbf{x}_i$  with  $\text{TSR}(\mathbf{x}_i) = \text{TSR}(\mathbf{x}_{query}) = q_0$ . Let the TSR pattern of another reasoning process, where for a certain first-step input  $\mathbf{x}_i$  with  $\text{TSR}(\mathbf{x}) = \text{TSR}(\mathbf{x}_{query}) = q_0$ , the  $k$ -th step output is the most probable for  $k \in [K'] \setminus \{h\}$ , while the  $h$ -th step output is the second probable. Denote the TSR pattern of the  $k$ -th step output of  $\mathbf{x}_i$  following this process as  $\boldsymbol{\mu}'_{u_k}$  with  $u_0 = q_0$ . By the Chernoff bound of Bernoulli distribution in Lemma F.3.1, we can obtain

$$\begin{aligned} &\Pr \left( \frac{1}{l_{ts}} \sum_{i=1}^{l_{ts}} \mathbb{1}[m_{k(i)} = \boldsymbol{\mu}'_{q_k}, \forall k \in [K']] \leq (1 - \rho_s^f/2)\alpha' \prod_{k=1}^{K'} A_{k(q_{k-1}, q_k)}^f \right) \\ &\leq e^{-l_{ts}(\rho_s^f)^2 \alpha' \prod_{k=1}^{K'} A_{k(q_{k-1}, q_k)}^f} = M^{-C}, \end{aligned} \tag{F.37}$$

and by Lemma F.3.3,

$$\begin{aligned}
& \Pr \left( \frac{1}{l_{ts}} \sum_{i=1}^{l_{ts}} \mathbb{1}[m_{k(i)} = \boldsymbol{\mu}'_{u_k}, \forall k \in [K']] \geq (1 - \rho_s^f/2) \alpha' \prod_{k=1}^{K'} A_{k(q_{k-1}, q_k)}^f \right) \\
& \leq \Pr \left( \frac{1}{l_{ts}} \sum_{i=1}^{l_{ts}} \mathbb{1}[m_{k(i)} = \boldsymbol{\mu}'_{u_k}, \forall k \in [K']] \geq \alpha' \prod_{k=1}^{K'} A_{k(u_{k-1}, u_k)}^f + t_0 \right) \\
& \leq e^{-l_{ts}t_0^2} = M^{-C},
\end{aligned} \tag{F.38}$$

for some  $c \in (0, 1)$  and  $C > 0$ , where the first step is by the definition of  $\rho_s^f$  in (7.12), and

$$t_0 \lesssim \rho_s^f \alpha' \prod_{k=1}^{K'} A_{k(q_{k-1}, q_k)}^f. \tag{F.39}$$

Hence, with a high probability,

$$\begin{aligned}
l_{ts} & \gtrsim \max \left\{ (\rho_s^f \alpha' \prod_{k=1}^{K'} A_{k(q_{k-1}, q_k)}^f)^{-1} \log M, (\rho_s^f \alpha' \prod_{k=1}^{K'} A_{k(q_{k-1}, q_k)}^f)^{-2} \log M \right\} \\
& \gtrsim (\rho_s^f \alpha' \prod_{k=1}^{K'} A_{k(q_{k-1}, q_k)}^f)^{-2} \log M,
\end{aligned} \tag{F.40}$$

such that the number of examples with the same TSR pattern as the query in each of the total  $K$  steps is at least 1. To make the above condition hold for any TSR pattern of the intermediate step of the query, we need

$$\begin{aligned}
l_{ts} & \gtrsim \max_{q_k \in [M']} (\rho_s^f \alpha' \prod_{k=1}^{K'} A_{k(q_{k-1}, q_k)}^f)^{-2} \log M \\
& = \max_{i \in [M']} (\rho_s^f \alpha' \prod_{k=1}^{K'} A_{k(\text{TSR}(f_{k-1} \circ \dots \circ f_0(\boldsymbol{\mu}'_i)), \text{TSR}(f_k \circ \dots \circ f_0(\boldsymbol{\mu}'_i)))}^f)^{-2} \log M \\
& = (\rho_s^f \alpha' \tau_s^f)^{-2} \log M.
\end{aligned} \tag{F.41}$$

Then, we show the CoT testing error is zero by induction. In the first step, consider  $\mathbf{x}_i = \boldsymbol{\mu}_j + \boldsymbol{\delta}_i$  such that

$$\tilde{\mathbf{p}}_i = \begin{pmatrix} \boldsymbol{\mu}'_j \\ \mathbf{y}_i \end{pmatrix} + \begin{pmatrix} \boldsymbol{\delta}_i \\ \mathbf{0} \end{pmatrix} + \mathbf{c}_i. \tag{F.42}$$

Since that

$$(\boldsymbol{\delta}_i^\top, \mathbf{0}^\top) \mathbf{W}^{(0)} \tilde{\mathbf{p}}_i \lesssim \xi, \quad (\text{F.43})$$

by that each entry of  $\mathbf{W}^{(0)}$  follows  $\mathcal{N}(0, \xi^2)$ , and

$$(\boldsymbol{\delta}_i^\top, \mathbf{0}^\top) \frac{\eta}{B} \sum_{n \in \mathcal{B}_b} \sum_{b=0}^{T-1} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(b)}} \tilde{\mathbf{p}}_{query} = 0, \quad (\text{F.44})$$

we have that for  $\tilde{\mathbf{p}}_i$  that shares the same TSR pattern as the query,

$$\begin{aligned} & \tilde{\mathbf{p}}_i^\top \mathbf{W}^{(T)} \tilde{\mathbf{p}}_{query} \\ &= \tilde{\mathbf{p}}_i^\top (\mathbf{W}^{(0)} + \frac{\eta}{B} \sum_{n \in \mathcal{B}_b} \sum_{b=0}^{T-1} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(b)}}) \tilde{\mathbf{p}}_{query} \\ &= ((\boldsymbol{\mu}'_j^\top, \mathbf{y}_i^\top) + \mathbf{c}_i^\top) (\mathbf{W}^{(0)} + \frac{\eta}{B} \sum_{n \in \mathcal{B}_b} \sum_{b=0}^{T-1} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(b)}}) \tilde{\mathbf{p}}_{query}. \end{aligned} \quad (\text{F.45})$$

Let  $\boldsymbol{\mu}'_j = \sum_{i=1}^{M'} \lambda_{j,i} \boldsymbol{\mu}_i$ . Then, we have

$$\begin{aligned} & \tilde{\mathbf{p}}_i^\top \mathbf{W}^{(T)} \tilde{\mathbf{p}}_{query} \\ &= ((\sum_{i=1}^{M'} \lambda_{j,i} \boldsymbol{\mu}_i^\top, \mathbf{y}_i^\top) + \mathbf{c}_i^\top) (\mathbf{W}^{(0)} + \frac{\eta}{B} \sum_{n \in \mathcal{B}_b} \sum_{b=0}^{T-1} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(b)}}) ((\sum_{i=1}^{M'} \lambda_{j,i} \boldsymbol{\mu}_i^\top, \mathbf{0}^\top) + \mathbf{c}_1)^\top \\ &= \sum_{i=1}^{M'} \lambda_{j,i}^2 ((\boldsymbol{\mu}_i^\top, \mathbf{y}_i^\top) + \mathbf{c}_i^\top) (\mathbf{W}^{(0)} + \frac{\eta}{B} \sum_{n \in \mathcal{B}_b} \sum_{b=0}^{T-1} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(b)}}) ((\boldsymbol{\mu}_i^\top, \mathbf{0}^\top) + \mathbf{c}_1)^\top \\ &\quad + \sum_{i \neq i'} \lambda_{j,i} \lambda_{j,i'} (\boldsymbol{\mu}_i^\top, \mathbf{y}_i^\top, \mathbf{c}_i^\top) (\mathbf{W}^{(0)} + \frac{\eta}{B} \sum_{n \in \mathcal{B}_b} \sum_{b=0}^{T-1} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(b)}}) ((\boldsymbol{\mu}_{i'}^\top, \mathbf{0}^\top) + \mathbf{c}_1)^\top \\ &\geq \Theta(\log \frac{K}{\epsilon}) - \epsilon \\ &= \Theta(\log \frac{K}{\epsilon}), \end{aligned} \quad (\text{F.46})$$

where the second to last step is by Theorem 3. Since the gradient updates for different TRR patterns are very close to each other, we have that if  $\sum_{i=1}^{M'} \lambda_{j,i} \lambda_{k,i} = 0$ ,

$$\begin{aligned} & \sum_{i=1}^{M'} \lambda_{j,i} \lambda_{j,i'} ((\boldsymbol{\mu}_i^\top, \mathbf{y}_i^\top) + \mathbf{c}_i^\top) (\mathbf{W}^{(0)} + \frac{\eta}{B} \sum_{n \in \mathcal{B}_b} \sum_{b=0}^{T-1} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(b)}}) ((\boldsymbol{\mu}_i^\top, \mathbf{0}^\top) + \mathbf{c}_1)^\top \\ & \lesssim \epsilon \log \frac{K}{\epsilon}. \end{aligned} \quad (\text{F.47})$$

Hence, for  $\tilde{\mathbf{p}}_i$  that shares a different TSR pattern with  $\tilde{\mathbf{p}}_i$ ,

$$\tilde{\mathbf{p}}_i^\top \mathbf{W}^{(T)} \tilde{\mathbf{p}}_{query} \lesssim \epsilon \log \frac{K}{\epsilon}. \quad (\text{F.48})$$

Therefore, we can derive that

$$\sum_{i \in \mathcal{S}_1^*} \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W}^{(T)} \tilde{\mathbf{p}}_{query}) \geq 1 - \epsilon, \quad (\text{F.49})$$

where  $\mathcal{S}_1^*$  is the set of the first step of examples that share the same TSR pattern as the query. Then, the first step leads to a correct prediction with zero testing error, since that  $\max_{j \in [M']} A_{k(q_0, j)}$  is the largest to make the correct prediction for  $\mathbf{x}_{query}$  if  $\mathbf{x}_{query} = \boldsymbol{\mu}'_{q_0}$ , i.e.,

$$\mathbf{v}_1 = f_1(\boldsymbol{\mu}'_{q_0}). \quad (\text{F.50})$$

Suppose that the  $k$ -th step generates a zero testing error. Then, for the  $k+1$ -th step, we know that there exists  $\mathbf{p}_j$  that shares the same TSR pattern as  $\mathbf{v}_k$ . Then, we can also derive that

$$\tilde{\mathbf{p}}_j^\top \mathbf{W}^{(T)} ((\mathbf{v}_k^\top, \mathbf{0}^\top)^\top + \mathbf{c}_k^\top)^\top = \Theta(\log \frac{K}{\epsilon}), \quad (\text{F.51})$$

and

$$\sum_{j \in \mathcal{S}_k^*} \text{softmax}(\tilde{\mathbf{p}}_j^\top \mathbf{W}^{(T)} ((\mathbf{v}_{k-1}^\top, \mathbf{v}_k^\top)^\top + \mathbf{c}_k^\top)^\top) \geq 1 - \epsilon. \quad (\text{F.52})$$

Hence, the  $k+1$ -th also makes the correct prediction, i.e.,

$$\mathbf{v}_{k+1} = f_{k+1} \circ \cdots \circ f_1(\boldsymbol{\mu}'_{q_0}), \quad (\text{F.53})$$

where  $\boldsymbol{\mu}'_{q_{k+1}}$  is the TSR pattern of the  $k + 1$ -th step input. Therefore, we show that CoT makes the correct prediction in each step as well as in the final prediction, such that

$$\bar{R}_{CoT, \mathbf{x} \in \mathcal{M}', f \in \mathcal{T}'}^f(\Psi) = 0. \quad (\text{F.54})$$

□

#### F.4.3 Proof of Theorem 5

*Proof.* We know that the positional encodings are the same for the ICL inference in all examples. Hence, similar to (F.49), we can derive that

$$\sum_{i \in \mathcal{S}_K^*} \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W}^{(T)} \tilde{\mathbf{p}}_{query}) \geq 1 - \epsilon, \quad (\text{F.55})$$

where  $\mathcal{S}_K^*$  is the set of the last step output of examples that share the same TSR pattern as the last step output of the query. For  $\mathbf{x}_{query} = \boldsymbol{\mu}'_q$ ,  $q \in [K']$ , we know that the distribution of the corresponding label  $\mathbf{y}$  of  $\mathbf{x}$  with  $\text{TSR}(\mathbf{x}) = q$  follows the  $q$ -th row of the  $K$ -steps transition matrix  $B^f$ . Let  $F(\Psi; \mathbf{P}) = \sum_{i=1}^{M'} \lambda_i^P \boldsymbol{\mu}'_i$ . Hence, based on the output scheme of ICL as stated in Section 7.2.3, we have that

$$\mathbf{v} = \arg \min_{\mathbf{y} \in \mathcal{M}'} \frac{1}{2} \|F(\Psi; \mathbf{P}) - \mathbf{y}\|^2 = \boldsymbol{\mu}_{\arg \max_{i \in [M']} \lambda_i^P}. \quad (\text{F.56})$$

Note that the probability of examples with the most probable final output with  $\boldsymbol{\mu}'_q$  as the TSR pattern of the input is

$$B_{(q, \text{TSR}(f(\boldsymbol{\mu}'_q)))}. \quad (\text{F.57})$$

To ensure that the number of examples with the same TSR pattern as the query that generates the most probable output is at least 1, we compute the following,

$$\begin{aligned} & \Pr \left( \frac{1}{l_{ts}} \sum_{i=1}^{l_{ts}} \mathbb{1}[m_i = \boldsymbol{\mu}'_{q_1}] \leq (1 - \rho_o^f/2) \alpha' B_{(q, \text{TSR}(f(\boldsymbol{\mu}'_q)))} \right) \\ & \leq e^{-l_{ts} \rho_o^f \alpha' B_{(q, \text{TSR}(f(\boldsymbol{\mu}'_q)))}} = M^{-C}, \end{aligned} \quad (\text{F.58})$$

for some  $c \in (0, 1)$  and  $C > 0$  by the Chernoff bound of Bernoulli distribution in Lemma F.3.1. Here,  $m_i$  is defined as the TSR pattern in the final output of the  $i$ -th context example

by the  $K$ -steps transition matrix defined in 7.13. The TSR pattern of the most probable output of the testing query is  $\boldsymbol{\mu}'_{q_1}$ . Similarly, let the TSR pattern of the second most probable output of the testing query be  $\boldsymbol{\mu}'_{q_2}$ . We also have

$$\begin{aligned} & \Pr \left( \frac{1}{l_{ts}} \sum_{i=1}^{l_{ts}} \mathbb{1}[m_i = \boldsymbol{\mu}'_{q_2}] \geq (1 - \rho_o^f/2) \alpha' B_{(q,q_1)}^f \right) \\ & \leq \Pr \left( \frac{1}{l_{ts}} \sum_{i=1}^{l_{ts}} \mathbb{1}[m_i = \boldsymbol{\mu}'_{q_2}] \geq \alpha' B_{(q,q_2)} + c \cdot \rho_o^f \alpha' B_{(q,q_1)}^f \right) \\ & \leq e^{-l_{ts} \rho_o^f c^2 \alpha' B_{(q,q_1)}} = M^{-C}, \end{aligned} \quad (\text{F.59})$$

by Lemma F.3.3 and (7.13) for some constant  $c > 0$ . Therefore, to make the number of examples with the same TSR pattern in the output as the label of the query be at least 1 for any TSR pattern of the query and the output be the most probable one, we need

$$\begin{aligned} l_{ts}^f & \gtrsim \max \left\{ \left( \rho_o^f \alpha' \min_{i \in [M']} B_{(i, \text{TSR}(f(\boldsymbol{\mu}'_i)))} \right)^{-1} \log M, \left( \rho_o^f \alpha' \min_{i \in [M']} B_{(i, \text{TSR}(f(\boldsymbol{\mu}'_i)))} \right)^{-2} \log M \right\} \\ & = (\rho_o^f \alpha' \tau_o^f)^{-2} \log M. \end{aligned} \quad (\text{F.60})$$

In addition, if Condition 7.3.2 holds such that the most probable output is the actual label, we can derive

$$\bar{R}_{ICL, \mathbf{x} \in \mathcal{M}', f \in \mathcal{T}'}^f(\Psi) = 0. \quad (\text{F.61})$$

When (F.60) holds but Condition 7.3.2 does not, we know that ICL still always produces the most probable output by the  $K$ -steps transition matrix, but such an output is not the label since Condition 7.3.2 fails. Hence,

$$\bar{R}_{ICL, \mathbf{x} \in \mathcal{M}', f \in \mathcal{T}'}^f(\Psi) \geq \Omega(1). \quad (\text{F.62})$$

When both Condition 7.3.2 and (F.60) do not hold, ICL can produce multiple possible outputs with a non-trivial probability, which is decided by the distribution of the prompt instead of the  $K$ -steps transition matrix. This can be seen from that (F.58) and (F.59) both do not hold since (F.60) fails. Then, ICL can produce both the most probable and the second most probable output with a constant probability. Let the TSR pattern of the  $r$ -th most probable output of the testing query be  $\boldsymbol{\mu}'_r$ . Recall that  $F(\Psi; \mathbf{P}) = \sum_{i=1}^{M'} \lambda_i^P \boldsymbol{\mu}'_i$ , we then have that for

some small  $\epsilon > 0$ ,

$$\lambda_{r(q)}^{\boldsymbol{P}} = \frac{|\{i \in [l_{ts}^f] : \mathbf{y}_i = \boldsymbol{\mu}'_r \text{ in } \boldsymbol{P}\}|}{l_{ts}^f} \pm \epsilon. \quad (\text{F.63})$$

Then, by (F.56), the output of the query is  $\boldsymbol{\mu}_{\arg \max_{r \in [M']} \lambda_r}$ . Since that (F.60) does not hold, there exists at least a constant probability of the prompt  $\boldsymbol{P}'$  with the same query as  $\boldsymbol{P}$  such that

$$\lambda_r^{\boldsymbol{P}'} = \frac{|\{i \in [l_{ts}^f] : \mathbf{y}_i = \boldsymbol{\mu}'_r \text{ in } \boldsymbol{P}'\}|}{l_{ts}^f} \pm \epsilon \neq \lambda_r^{\boldsymbol{P}}, \quad (\text{F.64})$$

for some  $r \in [M']$ . Therefore, with a constant probability, the output for the same testing query and the same testing task  $f$  varies. This leads to

$$\bar{R}_{ICL, \mathbf{x} \in \mathcal{M}', f \in \mathcal{T}'}^f(\Psi) \geq \Omega(1). \quad (\text{F.65})$$

□

#### F.4.4 Proof of Proposition 3

*Proof.* This proposition is derived from the proof of Theorem 4. (7.23) comes from (F.52), while (7.24) comes from (F.53), both by induction. □

### F.5 Proof of Lemmas

#### F.5.1 Proof of Lemma F.3.5

*Proof.*

$$\begin{aligned} & \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \boldsymbol{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \\ &= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \boldsymbol{P}^n, \mathbf{z}^n)}{\partial F(\Psi; \boldsymbol{P})} \frac{\partial F(\Psi; \boldsymbol{P})}{\partial \mathbf{W}} \\ &= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (\mathbf{F}(\Psi; \boldsymbol{P}) - \mathbf{z}^n)^\top \sum_{i=1}^l \mathbf{W}_V \tilde{\mathbf{p}}_i \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \\ & \quad \cdot (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_i) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top. \end{aligned} \quad (\text{F.66})$$

When  $t = 0$ , we know that each entry of  $\mathbf{W}^{(0)}$  is generated from the Gaussian distribution  $\mathcal{N}(0, \xi^2)$ . Then,

$$|\tilde{\mathbf{p}}_i^\top \mathbf{W}^{(0)} \tilde{\mathbf{p}}_{query}| = |\sum_{k,j} p_{i,k} p_{query,j} W_{k,j}^{(0)}| \lesssim \xi. \quad (\text{F.67})$$

Hence,

$$\text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W}^{(0)} \tilde{\mathbf{p}}_{query}) \geq \frac{e^{-\Theta(\xi)}}{l \cdot e^{\Theta(\xi)}} = \frac{1}{l} \cdot e^{-\Theta(\xi)}, \quad (\text{F.68})$$

$$\text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W}^{(0)} \tilde{\mathbf{p}}_{query}) \leq \frac{e^{-\Theta(\xi)}}{l \cdot e^{\Theta(\xi)}} = \frac{1}{l} \cdot e^{-\Theta(\xi)}. \quad (\text{F.69})$$

We can obtain

$$F(\Psi; \mathbf{P}) = \sum_{i=1}^l \frac{e^{-\Theta(\xi)}}{l} \mathbf{W}_V \mathbf{p}_i. \quad (\text{F.70})$$

Since that  $\text{PE}(\cdot)$ , and  $\text{TRR}(\cdot)$  denote the positional encoding, and the TSR pattern of the input, respectively, we have that for  $\mathbf{p}$ ,

$$\tilde{\mathbf{p}}^\top \tilde{\mathbf{p}}_{query} = \mathbb{1}[\text{TRR}(\tilde{\mathbf{p}}) = \text{TRR}(\tilde{\mathbf{p}}_{query})] + \mathbb{1}[\text{PE}(\tilde{\mathbf{p}}) = \text{PE}(\tilde{\mathbf{p}}_i)]. \quad (\text{F.71})$$

Given  $\text{lab}(\cdot)$  is the label embedding of the context as the input, we have that for  $\mathbf{p}$ ,

$$\tilde{\mathbf{p}}^\top \tilde{\mathbf{p}}_i = \mathbb{1}[\text{TRR}(\tilde{\mathbf{p}}) = \text{TRR}(\tilde{\mathbf{p}}_i)] + \mathbb{1}[\text{lab}(\tilde{\mathbf{p}}) = \text{lab}(\tilde{\mathbf{p}}_i)] + \mathbb{1}[\text{PE}(\tilde{\mathbf{p}}) = \text{PE}(\tilde{\mathbf{p}}_i)], \quad (\text{F.72})$$

$$(\mathbf{W}_V \tilde{\mathbf{p}})^\top \mathbf{W}_V \tilde{\mathbf{p}}_i = \mathbb{1}[\text{lab}(\tilde{\mathbf{p}}) = \text{lab}(\tilde{\mathbf{p}}_i)]. \quad (\text{F.73})$$

When  $t \geq 1$ , we first consider the case where  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the positional encoding as  $\tilde{\mathbf{p}}_{query}$ . If  $\tilde{\mathbf{p}}$  and  $\tilde{\mathbf{p}}_{query}$  share the same TRR pattern, label pattern, and the positional encoding,

$$\begin{aligned} \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} &\geq 2 \cdot (3 - 3p_n(t) - (1 - p_n(t))) \\ &= 4(1 - p_n(t)), \end{aligned} \quad (\text{F.74})$$

and

$$\tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \leq 2 \cdot (3 - 3p_n(t)) = 6(1 - p_n(t)). \quad (\text{F.75})$$

When  $\tilde{\mathbf{p}}$  and  $\tilde{\mathbf{p}}_{query}$  only share the same positional encoding or the same TRR pattern,

$$2 - 6p_n(t) \geq \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -4p_n(t). \quad (\text{F.76})$$

When  $\tilde{\mathbf{p}}$  and  $\tilde{\mathbf{p}}_{query}$  share both different positional encodings and TRR patterns,

$$-6p_n(t) \geq \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -2 - 4p_n(t). \quad (\text{F.77})$$

Then, we consider the case where  $\tilde{\mathbf{p}}$  only shares the same TRR pattern or the same positional encoding as  $\tilde{\mathbf{p}}_i$ . If  $\tilde{\mathbf{p}}$  and  $\tilde{\mathbf{p}}_{query}$  share the same TRR pattern, label pattern, and the positional encoding,

$$\begin{aligned} 3 - p_n(t) &\geq \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ &\geq 1 \cdot (3 - p_n(t) - (1 - p_n(t))) \\ &= 2. \end{aligned} \quad (\text{F.78})$$

When  $\tilde{\mathbf{p}}$  and  $\tilde{\mathbf{p}}_{query}$  only share the same positional encoding or the same TRR pattern,

$$1 - p_n(t) \geq \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq 0. \quad (\text{F.79})$$

When  $\tilde{\mathbf{p}}$  and  $\tilde{\mathbf{p}}_{query}$  only share both different positional encodings and TRR patterns,

$$-p_n(t) \geq \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -1. \quad (\text{F.80})$$

Note that  $-(1 - p_n(t))p_n(t) + (1 - p_n(t))^2\alpha^2/K^2 < 0$  for  $p_n(t) \in [\alpha/K, \alpha]$ . Then, when  $l \geq \Omega(\alpha^{-1})$  and  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the positional encoding as  $\tilde{\mathbf{p}}_i$ ,

$$\begin{aligned} &(\sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\ &\cdot \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ &\leq -4p_n(t)(1 - p_n(t))^2 - 4p_n(t)(1 - p_n(t))^2 \cdot \frac{\alpha^2}{K^2} \\ &+ \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) (-4p_n(t)) + \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) (1 - p_n(t)) (-2 - 4p_n(t))(K - 1) \\ &= -4p_n(t)(1 - p_n(t))^2 \left( 1 + \frac{\alpha^2}{K^2} \right) + \frac{2}{lK} (1 - \alpha) (- (K - 1) - (K + 1)p_n(t) \\ &+ 2p_n(t)^2(K - 1)). \end{aligned} \quad (\text{F.81})$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the different positional encoding as  $\tilde{\mathbf{p}}_{query}$ . Note that

$$\frac{2}{Kl} \cdot (1 - \alpha) \cdot K(1 - p_n(t)) \lesssim |(-(1 - p_n(t))p_n(t) + (1 - p_n(t))^2 \frac{\alpha^2}{K^2})(1 - p_n(t))|, \quad (\text{F.82})$$

if  $l \geq \Omega(\alpha^{-1})$ . Then,

$$\begin{aligned} & \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\ & \quad \cdot \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ & \leq -0 \cdot p_n(t)(1 - p_n(t)) + (1 - p_n(t))^2 \frac{\alpha^2}{K^2} \cdot (+2) + \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) (-(K-1)) \\ & = 2(1 - p_n(t))^2 \frac{\alpha^2}{K^2} - \frac{K-1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right). \end{aligned} \quad (\text{F.83})$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same positional encoding and the different TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned} & \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\ & \quad \cdot \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_i^\top \tilde{\mathbf{p}} \\ & \leq 0 - (1 - p_n(t))^2 \frac{\alpha^2}{K^2} + \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) (-(K-1)) \\ & = -(1 - p_n(t))^2 \frac{\alpha^2}{K^2} - \frac{K-1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right). \end{aligned} \quad (\text{F.84})$$

Therefore, as long as

$$l \geq \Omega(\alpha^{-1}), \quad (\text{F.85})$$

we have

$$\begin{aligned}
& \tilde{\mathbf{p}}^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \mathbf{p} \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (F(\Psi; \mathbf{P}) - \mathbf{z}^n)^\top \sum_{i=1}^l \mathbf{W}_V \tilde{\mathbf{p}}_i \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \\
&\quad \cdot \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_i^\top \tilde{\mathbf{p}} \\
&\leq \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} (1 - p_n(t))^2 (-4p_n(t)(1 + \frac{\alpha^2}{K^2}) + \frac{2(K-1)\alpha^2}{K^2}) \right. \\
&\quad \cdot \left. + (\frac{1}{K} - \frac{1}{M}) ((1 - p_n(t))^2 \frac{\alpha^2}{K^2}) \right) \\
&= \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} (1 - p_n(t))^2 (-4p_n(t)(1 + \frac{\alpha^2}{K^2}) + \frac{\alpha^2}{K}(1 + \frac{2(K-1)}{K})) \right. \\
&\quad \left. - \frac{\alpha^2}{K^3} (1 - p_n(t))^2 \right). \tag{F.86}
\end{aligned}$$

We then consider the case where  $\tilde{\mathbf{p}}'$  shares a different positional encoding and the same TRR pattern as  $\tilde{\mathbf{p}}$ . Let  $\tilde{\mathbf{p}}$  share the same TRR pattern and the positional encoding as  $\tilde{\mathbf{p}}_{query}$ . If  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  share the same TRR pattern, label pattern, and the positional encoding,

$$\begin{aligned}
2(3 - p_n(t)) &\geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
&\geq 2 \cdot (3 - p_n(t) - (1 - p_n(t))) \\
&= 4. \tag{F.87}
\end{aligned}$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_{query}$  only share the same positional encoding or the same TRR pattern,

$$2(1 - p_n(t)) \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq 0. \tag{F.88}$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share both different positional encodings and TRR patterns,

$$-2p_n(t) \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -2. \tag{F.89}$$

Then, we consider the case where  $\tilde{\mathbf{p}}$  only shares the same TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . If  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  share the same TRR pattern, label pattern, and the positional encoding,

$$\begin{aligned} 3 - p_n(t) &\geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ &\geq 1 \cdot (3 - 3p_n(t) - (1 - p_n(t))) = 2(1 - p_n(t)). \end{aligned} \quad (\text{F.90})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share the same positional encoding or the same TRR pattern,

$$1 - p_n(t) \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -2p_n(t). \quad (\text{F.91})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share both different positional encodings and TRR patterns,

$$-p_n(t) \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -1 - 2p_n(t). \quad (\text{F.92})$$

Next, we consider the case where  $\tilde{\mathbf{p}}$  only shares the same positional encoding as  $\tilde{\mathbf{p}}_{query}$ . If  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  share the same TRR pattern, label pattern, and the positional encoding,

$$3 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq 1 \cdot (3 - (1 - p_n(t))) = 2 + p_n(t). \quad (\text{F.93})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share the same positional encoding or the same TRR pattern,

$$1 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq p_n(t). \quad (\text{F.94})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share both different positional encodings and TRR patterns,

$$0 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -1 + p_n(t). \quad (\text{F.95})$$

Then, when  $l \geq \Omega(\alpha^{-1})$  and  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the positional encoding as

$$\tilde{\mathbf{p}}_{query},$$

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
& \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \leq -4p_n(t)(1-p_n(t)) + \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) (-2K).
\end{aligned} \tag{F.96}$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the different positional encoding as  $\tilde{\mathbf{p}}_{query}$ . Then, by (F.82),

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
& \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \leq -2p_n(t)(1-p_n(t))^2 - 2p_n(t)(1-p_n(t))^2 \cdot \frac{\alpha^2}{K^2} + \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) ((-1 - 2p_n(t))K) \\
& = -2p_n(t)(1-p_n(t))^2 \left( 1 + \frac{\alpha^2}{K^2} \right) + \frac{1}{l}(1-\alpha)(-1 - 2p_n(t)).
\end{aligned} \tag{F.97}$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same positional encoding and the different TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
& \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \leq p_n(t)(1-p_n(t))^2 + p_n(t)(1-p_n(t))^2 \frac{\alpha^2}{K^2} + \frac{1}{l}(1-\alpha)(-1 - 2p_n(t)) \\
& = p_n(t)(1-p_n(t))^2 \left( 1 + \frac{\alpha^2}{K^2} \right) - \frac{1}{l}(1-\alpha)(1 + 2p_n(t)).
\end{aligned} \tag{F.98}$$

Therefore, as long as

$$l \geq \Omega(\alpha^{-1}), \tag{F.99}$$

we have

$$\begin{aligned}
& \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \tilde{\mathbf{p}} \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (\mathbf{F}(\Psi; \mathbf{P}) - \mathbf{z}^n)^\top \sum_{i=1}^l \mathbf{W}_V \tilde{\mathbf{p}}_i \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i \\
&\quad - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
&\leq \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} (-4 - 2(K-1)(1-p_n(t))(1+\frac{\alpha^2}{K^2})) p_n(t)(1-p_n(t)) \right. \\
&\quad \left. + (\frac{1}{K} - \frac{1}{M}) p_n(t)(1-p_n(t))^2 (1+\frac{\alpha^2}{K^2}) \right) \\
&= \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} (-4 - (3K-2)(1-p_n(t))(1+\frac{\alpha^2}{K^2})) p_n(t)(1-p_n(t)) \right. \\
&\quad \left. + \frac{1}{K} p_n(t)(1-p_n(t))^2 (1+\frac{\alpha^2}{K^2}) \right), \tag{F.100}
\end{aligned}$$

and

$$\begin{aligned}
& \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \tilde{\mathbf{p}} \\
&\geq \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} (-4 - (3K-2)(1-p_n(t))(1+\frac{\alpha^2}{K^2})) p_n(t)(1-p_n(t)) \right. \\
&\quad \left. + \frac{1}{K} p_n(t)(1-p_n(t))^2 (1+\frac{\alpha^2}{K^2}) + \frac{1}{K} \cdot (1-p_n(t))^2 (-p_n(t) + (1-p_n(t)) \frac{\alpha^2}{K^2}) \right) \tag{F.101} \\
&= \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} (-4 - (3K-2)(1-p_n(t))(1+\frac{\alpha^2}{K^2})) p_n(t)(1-p_n(t)) \right. \\
&\quad \left. + \frac{\alpha^2}{K^3} (1-p_n(t))^2 \right).
\end{aligned}$$

We next consider the case where  $\tilde{\mathbf{p}}'$  shares a different TRR pattern and the same positional encoding as  $\tilde{\mathbf{p}}$ . Let  $\tilde{\mathbf{p}}$  share the same TRR pattern and the positional encoding as  $\tilde{\mathbf{p}}_{query}$ . If

$\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  share the same TRR pattern, label pattern, and positional encoding,

$$\begin{aligned} 2(3 - p_n(t)) &\geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ &\geq 2 \cdot (3 - p_n(t) - (1 - p_n(t))) \\ &= 4. \end{aligned} \quad (\text{F.102})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share the same positional encoding or the same TRR pattern,

$$2(1 - p_n(t)) \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq 0. \quad (\text{F.103})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share both different positional encodings and TRR patterns,

$$-2p_n(t) \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -2. \quad (\text{F.104})$$

Then, we consider the case where  $\tilde{\mathbf{p}}$  only shares the same TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . If  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  share the same TRR pattern, label pattern, and the positional encoding,

$$\begin{aligned} 3 &\geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ &\geq 1 \cdot (3 - (1 - p_n(t))) = 2 + p_n(t). \end{aligned} \quad (\text{F.105})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share the same positional encoding or the same TRR pattern,

$$1 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq p_n(t). \quad (\text{F.106})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share both different positional encodings and TRR patterns,

$$0 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -1 + p_n(t). \quad (\text{F.107})$$

Next, we consider the case where  $\tilde{\mathbf{p}}$  only shares the same positional encoding as  $\tilde{\mathbf{p}}_{query}$ . If  $\tilde{\mathbf{p}}'$

and  $\tilde{\mathbf{p}}_i$  share the same TRR pattern, label pattern, and the positional encoding,

$$\begin{aligned} 3 - p_n(t) &\geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ &\geq 1 \cdot (3 - p_n(t)) = 2. \end{aligned} \quad (\text{F.108})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share the same positional encoding or the same TRR pattern,

$$1 - p_n(t) \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq 0. \quad (\text{F.109})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share both different positional encodings and TRR patterns,

$$-p_n(t) \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -1. \quad (\text{F.110})$$

Then, when  $l \geq \Omega(\alpha^{-1})$ , and when  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the positional encoding as  $\tilde{\mathbf{p}}_{query}$ , by (F.82),

$$\begin{aligned} &(\sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\ &\cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ &\leq 0 - 2(1 - p_n(t))^2 \frac{\alpha^2}{K^2} + \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) (-2(K-1)). \end{aligned} \quad (\text{F.111})$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the different positional

encoding as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
& \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \leq -p_n(t)(1-p_n(t))(-1+p_n(t)) + p_n(t)(1-p_n(t))^2 \cdot \frac{\alpha^2}{K^2} + \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) \\
& K(-1+p_n(t)) \\
& = p_n(t)(1-p_n(t))^2 \left( \frac{\alpha^2}{K^2} + 1 \right) + \frac{1}{l}(1-\alpha)(-1+p_n(t)).
\end{aligned} \tag{F.112}$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same positional encoding and the different TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
& \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \leq -(1-p_n(t))^2 \frac{\alpha^2}{K^2} - 0 + \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) (-K+1) \\
& = -(1-p_n(t))^2 \frac{\alpha^2}{K^2} - \frac{K-1}{Kl} (1-\alpha).
\end{aligned} \tag{F.113}$$

Therefore, as long as

$$l \geq \Omega(\alpha^{-1}), \tag{F.114}$$

we have

$$\begin{aligned}
& \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \mathbf{p} \\
&= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (F(\Psi; \mathbf{P}) - \mathbf{z}^n)^\top \sum_{i=1}^l \mathbf{W}_V \tilde{\mathbf{p}}_i \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \\
&\quad \cdot \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
&\leq \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} \left( -\frac{\alpha^2}{K^2} + (K-1)(1 + \frac{\alpha^2}{K^2}) p_n(t) \right) (1 - p_n(t))^2 - \left( \frac{1}{K} \right. \right. \\
&\quad \left. \left. - \frac{1}{M} \right) (1 - p_n(t))^2 \frac{\alpha^2}{K^2} \right) \\
&= \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} \left( -\frac{\alpha^2}{K^2} + (K-1 + \frac{(2K-1)\alpha^2}{K^2}) p_n(t) \right) (1 - p_n(t))^2 \right. \\
&\quad \left. - (1 - p_n(t))^2 \frac{\alpha^2}{K^3} \right). \tag{F.115}
\end{aligned}$$

and

$$\begin{aligned}
& \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \mathbf{p} \\
&\geq \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} \left( -\frac{\alpha^2}{K^2} + (K-1 + \frac{(2K-1)\alpha^2}{K^2}) p_n(t) \right) (1 - p_n(t))^2 \right. \\
&\quad \left. - (1 - p_n(t))^2 \frac{\alpha^2}{K^3} + \frac{1}{K} \cdot (1 - p_n(t))^2 (-p_n(t) + (1 - p_n(t)) \frac{\alpha^2}{K^2}) \right). \tag{F.116}
\end{aligned}$$

We next consider the case where  $\tilde{\mathbf{p}}'$  shares a different TRR pattern and a different positional encoding as  $\tilde{\mathbf{p}}$ . Let  $\tilde{\mathbf{p}}$  share the same TRR pattern and the positional encoding as  $\tilde{\mathbf{p}}_{query}$ . If  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  share the same TRR pattern, label pattern, and the positional encoding,

$$6 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq 2 \cdot (3 - (1 - p_n(t))) = 4 + 2p_n(t). \tag{F.117}$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share the same positional encoding or the same TRR pattern,

$$2 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq 2p_n(t). \tag{F.118}$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share both different positional encodings and TRR patterns,

$$0 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -2 + 2p_n(t). \quad (\text{F.119})$$

Then, we consider the case where  $\tilde{\mathbf{p}}$  only shares the same TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . If  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  share the same TRR pattern, label pattern, and the positional encoding,

$$3 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq 1 \cdot (3 - p_n(t) - (1 - p_n(t))) = 2. \quad (\text{F.120})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share the same positional encoding or the same TRR pattern,

$$1 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq 0. \quad (\text{F.121})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share both different positional encodings and TRR patterns,

$$0 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -1. \quad (\text{F.122})$$

Next, we consider the case where  $\tilde{\mathbf{p}}$  only shares the same positional encoding as  $\tilde{\mathbf{p}}_{query}$ . If  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  share the same TRR pattern, label pattern, and the positional encoding,

$$3 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq 1 \cdot (3 - (1 - p_n(t))) = 2 + p_n(t). \quad (\text{F.123})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share the same positional encoding or the same TRR pattern,

$$1 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq p_n(t). \quad (\text{F.124})$$

When  $\tilde{\mathbf{p}}'$  and  $\tilde{\mathbf{p}}_i$  only share both different positional encodings and TRR patterns,

$$0 \geq \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \geq -1 + p_n(t). \quad (\text{F.125})$$

Then, when  $l \geq \Omega(\alpha^{-1})$ , and when  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the positional

encoding as  $\tilde{\mathbf{p}}_{query}$ ,

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
& \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \leq -p_n(t)(1-p_n(t))(-2+2p_n(t)) + (1-p_n(t))^2 \frac{\alpha^2}{K^2} \cdot 2p_n(t) \\
& + \frac{1}{l}(1-\alpha)(-2+2p_n(t)).
\end{aligned} \tag{F.126}$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the different positional encoding as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
& \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \leq 0 + p_n(t)(1-p_n(t))^2 \cdot \frac{\alpha^2}{K^2} \cdot (-1) + \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) (-K) \\
& = -p_n(t)(1-p_n(t))^2 \frac{\alpha^2}{K^2} + \frac{1}{l}(1-\alpha)(-1).
\end{aligned} \tag{F.127}$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same positional encoding and the different TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
& \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \leq -(1-p_n(t))p_n(t)(-1+p_n(t)) + p_n(t)(1-p_n(t))^2 \frac{\alpha^2}{K^2} \\
& + \frac{1}{l} \left( \frac{1}{K} - \frac{\alpha}{K} \right) (-1+p_n(t))K \\
& = (1-p_n(t))^2 p_n(t) \left( 1 + \frac{\alpha^2}{K^2} \right) + \frac{1}{l}(1-\alpha)(-1+p_n(t)).
\end{aligned} \tag{F.128}$$

Therefore, as long as

$$l \geq \Omega(\alpha^{-1}), \quad (\text{F.129})$$

we have

$$\begin{aligned} & \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \mathbf{p} \\ &= \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (F(\Psi; \mathbf{P}) - \mathbf{z}^n)^\top \sum_{i=1}^l \mathbf{W}_V \tilde{\mathbf{p}}_i \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \\ & \quad \cdot \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ &\leq \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} (-p_n(t)(1-p_n(t))(-2+2p_n(t)) + (3-K)(1-p_n(t))^2 \frac{\alpha^2}{K^2} \cdot p_n(t)) \right. \\ & \quad \left. + \left( \frac{1}{K} - \frac{1}{M} \right) (1-p_n(t))^2 p_n(t) \left( 1 + \frac{\alpha^2}{K^2} \right) \right) \\ &= \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} p_n(t)(1-p_n(t))^2 \left( 2 - K + \frac{(2-K)\alpha^2}{K^2} \right) \right. \\ & \quad \left. + (1-p_n(t))^2 p_n(t) \left( 1 + \frac{\alpha^2}{K^2} \right) \cdot \frac{1}{K} \right), \end{aligned} \quad (\text{F.130})$$

and

$$\begin{aligned} & \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \mathbf{p} \\ &\geq \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} p_n(t)(1-p_n(t))^2 \left( 1 + \frac{(2-K)\alpha^2}{K^2} \right) + (1-p_n(t))^2 p_n(t) \right. \\ & \quad \left. \cdot \left( 1 + \frac{\alpha^2}{K^2} \right) \cdot \frac{1}{K} + \frac{1}{K} \cdot (1-p_n(t))^2 (-p_n(t) + (1-p_n(t)) \frac{\alpha^2}{K^2}) \right) \\ &= \eta \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{KM} p_n(t)(1-p_n(t))^2 \left( 1 + \frac{(2-K)\alpha^2}{K^2} \right) + (1-p_n(t))^2 \cdot \frac{\alpha^2}{K^3} \right). \end{aligned} \quad (\text{F.131})$$

□

### F.5.2 Proof of Lemma F.3.6

*Proof.* We can derive that when  $1 - p_n(t) \geq \Omega(1)$ ,  $\tilde{\mathbf{p}}'^\top \mathbf{W}^{(t)} \tilde{\mathbf{p}}$  increases if  $\tilde{\mathbf{p}}$  and  $\tilde{\mathbf{p}}'$  share the same positional encoding. Otherwise,  $\tilde{\mathbf{p}}'^\top \mathbf{W}^{(t)} \tilde{\mathbf{p}}$  decreases. We know that  $p_n(t) \geq \frac{\alpha}{2}$ .

Combining the results in Lemma F.3.5, we can derive that when  $t \geq 1$ ,

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}^{(t)}}. \quad (\text{F.132})$$

Then, for  $\tilde{\mathbf{p}}_i^n$  that share the same TRR pattern and the same positional encoding of  $\tilde{\mathbf{p}}_{query}^n$ ,

$$\begin{aligned} \frac{p_n(t+1)}{|\mathcal{S}_1^n|} &= \text{softmax}(\mathbf{p}_i^{n\top} \mathbf{W}^{(t+1)} \tilde{\mathbf{p}}_{query}^n) \\ &\geq \frac{1}{l} \cdot \frac{1}{\frac{\alpha}{K} + \frac{(K-1)\alpha}{K} \cdot e^{-s_1} + (\frac{1}{K} - \frac{\alpha}{K})((K-1)e^{-s_2} + e^{-s_3})}, \end{aligned} \quad (\text{F.133})$$

where

$$s_1 \geq \eta \sum_{b=0}^t ((1 - p_n(b))^2 \frac{\alpha^2}{K^3} + \frac{\alpha^2}{K^3} (1 - p_n(b))^2) = \eta \sum_{b=0}^t (1 - p_n(b))^2 \frac{2\alpha^2}{K^3}, \quad (\text{F.134})$$

$$s_2 \geq \sum_{b=0}^t (1 - p_n(b))^2 \cdot \frac{2\eta\alpha^2}{K^3}, \quad (\text{F.135})$$

$$\begin{aligned} s_3 &\geq -\frac{\eta}{KM} \sum_{b=0}^t (1 - p_n(b))^2 (-4p_n(b)(1 + \frac{\alpha^2}{K^2}) + \frac{\alpha^2}{K}(1 + \frac{2(K-1)}{K}) + \frac{\alpha^2}{K^2} \\ &\quad - (K-1 + \frac{2K-1}{K^2}\alpha^2)p_n(b))) \\ &\geq \frac{\eta}{KM} \sum_{b=0}^t (1 - p_n(b))^2 (p_n(b)(3 + \frac{\alpha^2}{K^2})(4 + \frac{2K-1}{K^2})), \end{aligned} \quad (\text{F.136})$$

where the last step is by  $Kp_n(b) \geq 4\alpha^2/K^2$  when  $p_n(b) \geq \alpha/K$ . For  $\tilde{\mathbf{p}}_i^n$  that share the same TRR pattern and a different positional encoding of  $\tilde{\mathbf{p}}_{query}^n$ ,

$$\text{softmax}(\tilde{\mathbf{p}}_i^{n\top} \mathbf{W}^{(t+1)} \tilde{\mathbf{p}}_{query}^n) = \frac{1}{l} \cdot \frac{1}{\frac{\alpha}{K}e^{s_1} + \frac{(K-1)\alpha}{K} + (\frac{1}{K} - \frac{\alpha}{K})((K-1)e^{-s_4} + e^{s_5})}, \quad (\text{F.137})$$

where

$$\begin{aligned}
s_4 &\geq - \sum_{b=0}^t \frac{\eta}{M} ((-4 - (3K - 2)(1 - p_n(b))(1 + \frac{\alpha^2}{K^2}))p_n(b)(1 - p_n(b)) \\
&\quad - (2 - K)(1 + \frac{\alpha^2}{K^2})p_n(b)(1 - p_n(b))^2) \\
&= \sum_{b=0}^t \frac{\eta}{M} (4 + 2K(1 - p_n(b))(1 + \frac{\alpha^2}{K^2}))p_n(b)(1 - p_n(b)),
\end{aligned} \tag{F.138}$$

$$s_5 \geq \sum_{b=0}^t (1 - p_n(b))^2 \cdot \frac{2\eta\alpha^2}{K^3}. \tag{F.139}$$

When  $M \geq \Omega(K^4\alpha^{-1})$  and  $t \geq \Omega(\eta^{-1}K^3 \log K\alpha^{-2})$ ,

$$(K - 1)e^{-s_4} + e^{s_5} > K. \tag{F.140}$$

If  $M \geq \Omega(K^4\alpha^{-1})$  and  $t \leq O(\eta^{-1}K^3 \log K\alpha^{-2})$ , we cannot ensure

$$(K - 1)e^{-s_4} + e^{s_5} > K. \tag{F.141}$$

For  $\tilde{\mathbf{p}}_i^n$  that share a different TRR pattern and the same positional encoding of  $\tilde{\mathbf{p}}_{query}^n$ ,

$$\text{softmax}(\tilde{\mathbf{p}}_i^n{}^\top \mathbf{W}^{(t+1)} \tilde{\mathbf{p}}_{query}^n) = \frac{1}{l} \cdot \frac{1}{\frac{\alpha}{K}e^{s_3} + \frac{\alpha}{K} \cdot e^{-s_4} + (\frac{1}{K} - \frac{\alpha}{K})(1 + (K - 1)e^{-s_6})}, \tag{F.142}$$

where

$$s_6 \geq \eta \sum_{b=0}^t \frac{2\alpha^2}{K^3} (1 - p_n(b))^2. \tag{F.143}$$

For  $\tilde{\mathbf{p}}_i^n$  that share a different TRR pattern and a different positional encoding of  $\tilde{\mathbf{p}}_{query}^n$ ,

$$\text{softmax}(\tilde{\mathbf{p}}_i^n{}^\top \mathbf{W}^{(t+1)} \tilde{\mathbf{p}}_{query}^n) = \frac{1}{l} \cdot \frac{1}{\frac{\alpha}{K}e^{s_2} + (\frac{1}{K} - \frac{\alpha}{K})(K - 1 + e^{s_6}) + \frac{\alpha}{K}e^{s_4}}. \tag{F.144}$$

Note that when  $t \lesssim \eta^{-1}\alpha^{-2}K^3$ , for  $\mathbf{p}_{query}^n$  in the  $k$ -th step, we have

$$\sum_{i \in \mathcal{S}_{[K]} \setminus \{k\}} \text{softmax}(\tilde{\mathbf{p}}_i^n{}^\top \mathbf{W}^{(t+1)} \tilde{\mathbf{p}}_{query}^n) \geq \Omega(1), \tag{F.145}$$

for  $\tilde{\mathbf{p}}_i^n$  that share a different positional encoding from  $\tilde{\mathbf{p}}_{query}^n$ . To make the total softmax values on contexts that share a different positional encoding and a different TRR pattern from the query smaller than  $\epsilon$ , we need

$$s_1, s_2, s_6 \gtrsim \log \frac{K}{\epsilon}. \quad (\text{F.146})$$

When  $t$  further increases to be larger than  $\Omega(\eta^{-1}\alpha^{-2}K^3 \log \frac{K}{\epsilon})$ , we also have that the total softmax values on contexts that share a different positional encoding and the same TRR pattern from the query smaller than  $\epsilon$ . Therefore,

$$t \gtrsim T_1 := \eta^{-1}\alpha^{-2}K^3 \log \frac{K}{\epsilon}. \quad (\text{F.147})$$

□

### F.5.3 Proof of Lemma F.3.7

*Proof.* We consider the case when  $t \geq T_1$  given Lemma F.3.6. When  $l \geq \Omega(\alpha^{-1})$ , and when  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the positional encoding as  $\tilde{\mathbf{p}}_{query}$ ,

$$\begin{aligned} & \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\ & \cdot \tilde{\mathbf{p}}^\top \left( \tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r \right) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ & \leq -4p_n(t)(1-p_n(t))^2 + \epsilon \\ & \lesssim -4p_n(t)(1-p_n(t))^2. \end{aligned} \quad (\text{F.148})$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the different positional encoding as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
& \cdot \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \lesssim -0 \cdot p_n(t)(1 - p_n(t)) + \epsilon \\
& \lesssim \epsilon.
\end{aligned} \tag{F.149}$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same positional encoding and the different TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
& \cdot \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \lesssim \epsilon.
\end{aligned} \tag{F.150}$$

Therefore,

$$\begin{aligned}
& \tilde{\mathbf{p}}^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \mathbf{p} \\
& = \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (\mathbf{F}(\Psi; \mathbf{P}) - \mathbf{z}^n)^\top \sum_{i=1}^l \mathbf{W}_V \tilde{\mathbf{p}}_i \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \\
& \quad \cdot \tilde{\mathbf{p}}^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
& \lesssim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \left( \frac{1}{2M} (-4p_n(t)(1 - p_n(t))^2) + \left( \frac{1}{2} - \frac{1}{M} \right) \cdot \epsilon \right) \\
& = -\eta \cdot \frac{1}{2M} \cdot \frac{1}{B} \sum_{n \in \mathcal{B}_b} 4p_n(t)(1 - p_n(t))^2.
\end{aligned} \tag{F.151}$$

We then discuss if  $\tilde{\mathbf{p}}$  and  $\tilde{\mathbf{p}}'$  only share the same TRR pattern. When  $l \geq \Omega(\alpha^{-1})$ , and when

$\tilde{\mathbf{p}}$  shares the same TRR pattern and the positional encoding as  $\tilde{\mathbf{p}}_{query}$ , we can obtain

$$\begin{aligned} & \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\ & \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ & \gtrsim -2(1-p_n(t))^2 p_n(t). \end{aligned} \quad (\text{F.152})$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the different positional encoding as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned} & \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\ & \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ & \gtrsim -(1-p_n(t))(1-p_n(t))p_n(t). \end{aligned} \quad (\text{F.153})$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same positional encoding and the different TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned} & \left| \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \right. \\ & \left. \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \right| \\ & \lesssim \epsilon. \end{aligned} \quad (\text{F.154})$$

Therefore,

$$\begin{aligned}
& \left| \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \mathbf{p} \right| \\
&= \left| \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (\mathbf{F}(\Psi; \mathbf{P}) - \mathbf{z}^n)^\top \sum_{i=1}^l \mathbf{W}_V \tilde{\mathbf{p}}_i \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}^\top \right. \\
&\quad \cdot \left. (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \right| \\
&\leq \eta \epsilon.
\end{aligned} \tag{F.155}$$

We next discuss when  $\tilde{\mathbf{p}}$  only shares the same positional encoding as  $\tilde{\mathbf{p}}'$ . When  $l \geq \Omega(\alpha^{-1})$ , and when  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the positional encoding as  $\tilde{\mathbf{p}}_{query}$ ,

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
&\cdot \tilde{\mathbf{p}}'^\top \left( \tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r \right) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
&\lesssim \epsilon.
\end{aligned} \tag{F.156}$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the different positional encoding as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned}
& \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\
&\cdot \tilde{\mathbf{p}}'^\top \left( \tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r \right) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\
&\lesssim -p_n(t)(1-p_n(t))(-1+p_n(t)) + \frac{1}{M} \\
&\lesssim p_n(t)(1-p_n(t))^2.
\end{aligned} \tag{F.157}$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same positional encoding and the different

TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned} & \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\ & \cdot \tilde{\mathbf{p}}'^\top \left( \tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r \right) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ & \lesssim \epsilon. \end{aligned} \tag{F.158}$$

Therefore,

$$\begin{aligned} & \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \mathbf{p} \\ & = \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (F(\Psi; \mathbf{P}) - \mathbf{z}^n)^\top \sum_{i=1}^l \mathbf{W}_V \tilde{\mathbf{p}}_i \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}^\top \\ & \quad \cdot \left( \tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r \right) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ & \lesssim \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{1}{2M} \cdot p_n(b)(1 - p_n(b))^2. \end{aligned} \tag{F.159}$$

We then consider if  $\tilde{\mathbf{p}}$  shares a different TRR pattern and a different positional encoding as  $\tilde{\mathbf{p}}'$ . When  $l \geq \Omega(\alpha^{-1})$ , and when  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the positional encoding as  $\tilde{\mathbf{p}}_{query}$ ,

$$\begin{aligned} & \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\ & \cdot \tilde{\mathbf{p}}'^\top \left( \tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r \right) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ & \gtrsim \epsilon. \end{aligned} \tag{F.160}$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same TRR pattern and the different positional

encoding as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned} & \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \\ & \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \\ & \gtrsim -(1 - p_n(t)) p_n(t). \end{aligned} \quad (\text{F.161})$$

We next consider the case where  $\tilde{\mathbf{p}}$  shares the same positional encoding and the different TRR pattern as  $\tilde{\mathbf{p}}_{query}$ . Then,

$$\begin{aligned} & \left| \left( \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i - \mathbf{z}^n \right)^\top \sum_{i=1}^l \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{W}_V \tilde{\mathbf{p}}_i \right. \\ & \left. \cdot \tilde{\mathbf{p}}'^\top (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \right| \\ & \lesssim \epsilon. \end{aligned} \quad (\text{F.162})$$

Therefore,

$$\begin{aligned} & \left| \tilde{\mathbf{p}}'^\top \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} \frac{\partial \ell(\Psi; \mathbf{P}^n, \mathbf{z}^n)}{\partial \mathbf{W}} \mathbf{p} \right| \\ & = \left| \eta \frac{1}{B} \sum_{n \in \mathcal{B}_b} (F(\Psi; \mathbf{P}) - \mathbf{z}^n)^\top \sum_{i=1}^l \mathbf{W}_V \tilde{\mathbf{p}}_i \text{softmax}(\tilde{\mathbf{p}}_i^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \tilde{\mathbf{p}}^\top \right. \\ & \left. (\tilde{\mathbf{p}}_i - \sum_{r=1}^l \text{softmax}(\tilde{\mathbf{p}}_r^\top \mathbf{W} \tilde{\mathbf{p}}_{query}) \mathbf{p}_r) \tilde{\mathbf{p}}_{query}^\top \tilde{\mathbf{p}} \right| \\ & \lesssim \eta \epsilon. \end{aligned} \quad (\text{F.163})$$

□