Simple Structures in Neural Networks:

On Expressiveness, Optimization and Data Distribution

by

Lei Chen

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

New York University

May, 2025

Professor Joan Bruna

To all who illuminated this path

"The way ahead is long and has no ending;

yet high and low I'll search with my will unbending"

— Qu Yuan, 340 BC - 278 BC

Acknowledgements

First I would like to thank my advisor, Joan Bruna, for his insightful guidance and unconditional support throughout my PhD study. He always encouraged me to work on truly interesting and important problems. He was also a great cheerleader when our manuscripts were rejected due to randomness in reviewing. Beyond academic research, he is the best role model for being generous to students, peer researchers and family members. It was one of the best decisions of my life to continue my PhD study in his group after my master study. In the application season for the 2020 fall, I felt a possibility of switching my research directions from graph machine learning to general deep learning in the future, and then Joan would be an ideal advisor because of his expertise and ambitiousness in almost all areas in the intersection of mathematics and deep learning. This turns out to be true because I switched my directions twice.

I want to thank my committee members, Alberto Bietti, Andrew Gordon Wilson, Kyunghyun Cho and Rob Nowak. Alberto is a wonderful collaborator sharing his deep understanding of theory and experiments in Large Language Models. Andrew and Kyunghyun have been bringing significant help through discussing on my research, providing internship opportunities and teaching excellent courses at NYU. Rob runs a great lab in Madison, where I spent an unforgettable summer, made many friends and shared research interests.

I want to thank Zhengdao Chen especially for his guidance on my research. He is knowledgeable and modest, implicitly teaching me to keep challenging and improving ideas. I also want to thank Soledad Villar for her help on my research and PhD applications. Zhengdao and Soledad

are wonderful teammates when facing very close deadlines. I also got great support from my lab and office mates, including but not limited to, Aaron Zweig, Carles Domingo-Enrich, Cedric Gerbelot, Cinjon Resnick, David Brandfonbrener, Denny Wu, Evan Dogariu, Karl Otness, Loucas Pillaud-Vivien, Min Jae Song, Noah Amsel, Richard Pang and Samy Jelassi. I would like to thank Alex Damian, Jingfeng Wu and Jeremy Cohen for discussion on optimization.

I want to thank the following external collaborators, Micheal Bronstein, Shunwang Gong and Justin Gilmer. Michael and Shunwang host my visit to London and provided a different view of geometric deep learning. Justin taught me how to empirically analyze optimization problems in Large Language Models.

I want to thank my friends to get me in a good mood every day, including but not limited to, Meng Wang, Guangyu Xue, Xiao Xu, Chengpeng Xue, Fangjun Zhang, Xuan Tang, Jialiang Cao, Yihan He, Junwen Yang and Yiyang Wen. Meng inspired me to switch my PhD major from Engineering to Computer Science, as he did. He is also an amateur poet and singer, bringing joy during difficult times. Yihan, Junwen and Yiyang helped make my decision to work in quantative finance after PhD graduation. In Seattle, I met many great friends to enjoy food and discuss research with, including Bingbing Wen, Yujie Li, Chen Liang, Runlong Zhou, Zhihan Xiong, Qiwen Cui, Haozhe Jiang and Weihang Xu. In Madison, I started precious friendships with Jifang Zhang and Joe Shenouda.

I quit my first PhD program of Civil Engineering at Tsinghua University in 2018. I would like to thank my advisors, Jianguo Nie and Congzhen Xiao, for their understanding of my decision and encouraging me to chaise my interests. This experience has inspired me to make better decisions in my life. When some peers were struggling in taking everything into consideration, I was extremely fortunate to be able to follow my heart.

I want to thank my parents, Shaoming Chen and Cuiqin Xie, for their unconditional support. It was not an easy decision to quit from a decent PhD program and then start my master and PhD study in a foreign country. All they want is to see me happy every day. They are my rock, as they

gave me a name with three rocks in Chinese. Someday, I hope I could be their rock as well.

Finally, I want to thank my love, Yifang Chen. We enrolled in and will graduate from PhD programs at the same time. I cannot imagine how I could enjoy this journey so much without her. The initial condition was quite distant, as she was in Seattle and I was in New York. But we will get a global convergence towards the same place, motivated by the momentum of love, and live happily ever after.

Abstract

In this era of Large Language Models (LLMs) and other giant neural networks, we aim to analyze simplified settings from scratch, as foundational steps towards understanding the functionality of the giant models. We present our understanding from three aspects.

On expressive power, we investigate the function class of simplified graph networks, *i.e.*, Graph-Augmented Multi-layer Perceptrons (GA-MLPs), against the classic Graph Neural Networks (GNNs) using measurements of graph isomorphism testing and counting attributed walks.

On optimization, we theoretically study instabilities from large learning rates in training neural networks, *i.e.*, Edge of Stability. We investigate the conditions of how the loss landscape contains such unstable training trajectories, especially oscillating in a low-dimensional subspace. Then we leverage such property in simple, yet representative, learning problems in a teacher-student style, including two-layer single-neuron homogeneous networks and matrix factorization.

On data distribution of reasoning tasks, we propose a decomposition of next-token prediction into two parts: in-context reasoning and distributional association. We study this decomposition empirically and theoretically in a controlled synthetic setting, and find that feed-forward layers tend to learn simple distributional associations such as bigrams, while attention layers focus on in-context reasoning. Finally, based on such understanding, we provide empirical evidence on how modifying the feed-forward layers can improve the performance of LLMs on certain tasks.

Contents

Dedication iii

Acknowledgments iv

Abstract vii

List of Figures xiii

List of Tables xx

1 Introduction 1

1.1 Expressive Power of Simplified Graph Neural Networks . . . 1

1.2 Simplicity Induced by Optimization . . . 3

1.3 Roles of Feed-Forward Layers and Attention in Transformers . . . 4

2 Expressive Power of Graph Neural Networks 6

2.1 Introduction and Our Contributions . . . 6

2.2 Related Works . . . 8

2.3 Preliminaries . . . 10

2.3.1 Notations . . . 10

2.3.2 Graph Neural Networks (GNNs) . . . 10

2.3.3 Graph-Augmented Multi-Layer Peceptrons (GA-MLPs) . . . 11

2.4 Expressive Power as Graph Isomorphism Tests . . . 12
2.5 Expressive Power as Functions on Rooted Graphs . . . 14
2.6 Experiments . . . 18
2.6.1 Number of equivalence classes of rooted graphs . . . 19
2.6.2 Counting attributed walks . . . 19
2.6.3 Community detection on Stochastic Block Models (SBM) . . . 20
2.7 Conclusions . . . 22

3 Optimization Instabilities in Low-dimensional Space 23
3.1 Introduction and Our Contributions . . . 23
3.1.1 Connections between theoretical results . . . 25
3.1.2 Implications from low-dimension to high-dimension . . . 27
3.2 Related Works . . . 28
3.3 Preliminaries . . . 31
3.4 Stable oscillation on 1-D functions: fixed point of two-step update . . . 32
3.4.1 Existence of fixed points . . . 33
3.4.2 Convergence to fixed points . . . 36
3.5 On a two-layer single-neuron homogeneous network . . . 38
3.6 Matrix Factorization and beyond . . . 41
3.6.1 Observations from Matrix Factorization . . . 41
3.6.2 Implications for more complicated settings . . . 45
3.7 Experiments on MLPs and MNIST . . . 49
3.7.1 2-layer high-dim homogeneous ReLU NNs with planted teacher neurons . 49
3.7.2 3, 4, 5-layer non-homogeneous MLPs on MNIST . . . 51
3.8 Conclusions . . . 52

4 Memorization of Training Distribution in Transformer Modules **58**

4.1 Introduction and Our Contributions . . . 58

4.2 Related Works . . . 60

4.3 Preliminaries . . . 61

4.3.1 Reasoning from Context . . . 61

4.3.2 Truncating Weights with LASER [Sharma et al. 2023] . . . 63

4.4 Two-layer Transformer on Noisy In-context Recall . . . 64

4.4.1 Feed-forward layers store the generic noise . . . 69

4.4.2 Attention attends to in-context targets and avoids noise . . . 70

4.4.3 No feed-forward Layers: value matrix stores generic noise association . . . 72

4.4.4 How Does the Two-layer Model Solve Noisy In-context Recall? . . . 75

4.4.5 Multiple Triggers . . . 82

4.5 Experiments on Pre-trained LLMs . . . 86

4.5.1 An Investigation on GPT-2 Small and Pythia Models . . . 86

4.5.2 The effect of truncating feed-forward layers on GSM8K . . . 89

4.6 Discussion and Limitations . . . 89

A Appendix: Supplementary Materials for Chapter 2 **92**

A.1 GA-MLP with general equivariant graph operators for node feature augmentation 92

A.2 Examples of existing GA-MLP models . . . 95

A.3 Equivalence classes induced by GNNs and GA-MLPs among real graphs . . . 95

A.4 Additional notations . . . 96

A.5 Proof of Proposition 2.2 . . . 97

A.6 Proof of Proposition 2.3 . . . 99

A.7 Proof of Proposition 2.4 . . . 101

A.8 Proof of Proposition 2.6 . . . 102

A.9 Proof of Proposition 2.7 . . . 107
A.10 Proof of Proposition 2.1 . . . 108
A.11 Experiment Details . . . 113

B Appendix: Supplementary Materials for Chapter 3 116

B.1 Additional Results . . . 116
B.2 Additional Experiments . . . 120
B.3 Proof of Theorem 3.1 . . . 123
B.4 Proof of Lemma 3.2 . . . 125
B.5 Proof of Prop 1 . . . 128
B.6 Proof of Theorem 3.3 . . . 131
B.7 Proof of Theorem 3.4 . . . 137
B.8 Proof of Lemma B.2 . . . 142
B.9 Proof of Theorem 3.5 . . . 145
B.10 Proof of Matrix Factorization . . . 174
B.11 Useful lemmas . . . 193
B.12 Illustration of period-2 and period-4 orbits . . . 193

C Appendix: Supplementary Materials for Chapter 4 194

C.1 More Experiments on Pythia . . . 194
C.2 Proof of Theorem 4.1 . . . 197
C.3 Proof for First and Second moments in Lemma C.2 . . . 217
C.4 Proof of Theorem 4.2: Training Dynamics of the Attention Layer . . . 243
C.5 Experiments Setup: Linear Associative Memory . . . 252
C.6 Useful Lemmas . . . 255
C.7 Input Examples for LLMs . . . 258
C.8 Synthetic IOI Task . . . 260

Bibliography **266**

List of Figures

2.1 A pair of graphs that can be distinguished by 2 iterations of the WL test but not by GA-MLPs with $\Omega \subseteq \{A^k : k \in \mathbb{N}\}$, as proved in Appendix A.10. 13

2.2 An illustration of rooted graphs and rooted aggregation trees. *Left*: a pair of graphs, G and G'. *Center*: the rooted graphs of 1 in G and G', $G^{[1]}$ and $G'^{[1]}$. *Right*: the rooted aggregation tree that both $G^{[1]}$ and $G'^{[1]}$ correspond to. 15

2.3 A pair of rooted graphs, $G^{[1]}$ (left) and $G'^{[1]}$ (right), in which blue nodes have node feature 0 and green nodes have node feature 1. They belong to the same equivalence class induced by any GA-MLP with operators that only depend on the graph structure, but different equivalence classes induced by GNNs. In particular, $G^{[1]}$ and $G'^{[1]} \in \mathcal{T}_{2,2,(1,1,3)}$ (defined in Appendix A.8), and $|\mathcal{W}_2(G^{[1]}; (1,1))| = 1$ whereas $|\mathcal{W}_2(G'^{[1]}; (1,1))| = 0$. 17

2.4 Community detection on binary SBM with 5 choices of in- and out-group connectivities, each yielding to a different SNR. Higher overlap means better performance. 21

3.1 Connections between our presented theoretical results. The arrows stand for "implies". LG stands for Local Geometry. MF stands for Matrix Factorization. 26

3.2 **Matrix Factorization**: grid search of $\eta\sigma_1^2$ v.s. $\eta\sigma_2^2$ on whether GD diverges or not. (a) Generic initialization: it verifies the condition $\eta\left(\sigma_1^2 + \sigma_2^2\right) < 2$. (b-c) Quasi-symmetric initialization: it verifies the predicted bound $\eta \cdot \left(\sigma_1^2/\alpha^2 + \sigma_2^2\alpha^2\right) < 2$ in Eq.(3.9) as a sufficient condition. 45

3.3 **Matrix Factorization**: $\sigma_{\max}(\mathbf{Y}), \sigma_{\max}(\mathbf{Z})$ for different η's. For each η, the last 10 iterations are sampled for report, due to periodic and chaotic phenomenon. **Observations**: (1) when $\eta\sigma_1^2 \in (1, 1.38)$, all cases have $\sigma_{\max}(\mathbf{Y}) = \sigma_{\max}(\mathbf{Z})$; (2) when $\eta\sigma_1^2 \in (1, 1.23)$, it converges to a period-2 orbit; (3) when $\eta\sigma_1^2 \in (1.23, 1.28)$, it converges to a period-4 orbit; (4) when $\eta\sigma_1^2 > 1.28$, it is rather chaotic; (5) when $\eta\sigma_1^2 < 1$, there is no oscillation. 46

3.4 Trajectories of minimizing $L(x, y) = 1/2(xy - 1)^2$ with $\eta = 1.08, 0.95$. For $\eta = 1.08$, the manifold $\mathcal{M}$ proposed by Damian et al. [2022b] does not exist. For $\eta = 0.95$, the manifold $\mathcal{M}$ exists, but the projection onto it does not change for the first few steps. 47

3.5 Result of 2-layer 16-neuron teacher-student experiment. 54

3.6 Result of **3-layer** ReLU MLPs on MNIST. Both (c) and (d) are for learning rate as 0.5. 55

3.7 Result of **4-layer** ReLU MLPs on MNIST. 56

3.8 Result of **5-layer** ReLU MLPs on MNIST. 57

4.1 **Distributional association *v.s.* in-context reasoning.** In this work, we decompose tasks of next-token prediction into the distributional and the in-context ones, finding that MLPs learn distributional associations before attention develops in-context reasoning capabilities. Furthermore, truncating MLPs promotes in-context reasoning by weakening distributional associations. See Figure 4.13 for an example of this on the Pythia model [Biderman et al. 2023]. 60

4.2 **Noisy in-context recall**. *Purpose of design*: understand mechanisms of attention and feed-forward layers for tasks with **in-context reasoning** (predict $\bar{y}$) and **distributional association** (predict τ). *Task*: predict tokens $\bar{y}$ *v.s.* τ from a sentence $[\ldots, q, \bar{y}, \ldots, q, \tau, \ldots, q]$ where q is trigger, $\bar{y}$ is sampled target token for a sentence, and τ is a fixed generic token across sentences. *Our findings*: in a two-layer transformer, the second-layer attention (Attn-2) only attends towards target tuples $[q, \bar{y}]$ while the feed-forward layer (FF-2) learns to predict τ. 66

4.3 **Left three**: Average probability of predicting correct and noise tokens, and test loss on clean data ($\alpha = 0$), with different fractions ρ of preserved rank in U_{in} of the second-layer MLP F_2. The full model learns to predict noise with probability around $\alpha = 0.5$, as expected from training data. When F_2 is dropped ($\rho = 0$), the model predicts the correct token $\bar{y}$ with probability ≈ 0.98. **Rightmost**: the FF-2 margin of τ *v.s.* all the other tokens with input as q, *i.e.*, $[\mathbf{W}_U F_2(\mathbf{W}_E(q))]_\tau - \max_{k \leqslant N} [\mathbf{W}_U F_2(\mathbf{W}_E(q))]_k$. It reveals that FF-2 learns trigger-noise association in early steps. 67

4.4 The second-layer attention scores of models trained with noise (left), fine-tuned with noise (right, initialized as a model pre-trained without noise), given the same input. It turns out both models learn to attend to the informative structure "[trigger]+$\bar{y}$" instead of "[trigger]+noise". This implies that the attention in these models is only responsible to predict $\bar{y}$, although the training input and output have noise with probability $\alpha = \Theta(1)$. The fine-tuning setting is in Appendix 4.4.4.1. 68

4.5 Pure-label loss for rank-1,2,3,4 models with $n = 3, \alpha = 0.03$ and $d = 12$ (left) or 8 (right). *Only full models are trained*, and we report low-rank results by conducting SVD in each step without manipulating the training. In both figures, the experiments are run for 20 times to examine the randomness. For each rank, we plot curves of the median, 25% and 75% out of 20 runs. It turns out: i) rank-2 models are very likely to have significantly lower pure-label loss thant full models (rank$\geqslant$ 3), and ii) the larger dimension d has more stable results. 74

4.6 **Left**: first-layer attention attending to the previous token from the current token. **Middle**: logits to predict noise from $\langle F_2(\mathbf{W}_E(i)), \mathbf{W}_U(j)\rangle$ with input $i \in [N+1]$ and output $j \in [N+1]$, where the output channel 2 is set as the noise channel. It turns out, for all input i, the logits on output 2 are large, which matches our construction that, at least for trigger q as input, the output 2 has large logits. **Right**: logits to predict singal from $\langle \mathbf{W}_V^2 \mathbf{W}_E(i), \mathbf{W}_U(j)\rangle$ for input $i \in [N+1]$ and output $j \in [N+1]$. It matches our construction that $i = j$ has large logits. Meanwhile, $i = j = 2$ does not have large logits since 2 is the noise channel. . . . 77

4.7 Fractions of predicting the noise token and the other non-noise tokens with $\alpha = 0.5$. (Left) pretraining steps on noisy data; (right) finetuning steps on noisy data, after pretraining on clean data with $\alpha = 0$. In both cases, the models learn to predict noise with probability nearly 0.5. In the first few (~ 5) steps, the models quickly learn to predict noise with probability close to 1. The fine-tuning setting is in Appendix 4.4.4.1. 78

4.8 Logits of $\langle \mathbf{W}_V^1 \mathbf{W}_E(i)^\top, \mathbf{W}_{QK}^2 \mathbf{W}_E(j)\rangle$ for input i and output j when there is one trigger (left, $q = 1$) and five triggers (right, $q \in Q = \{1, 39, 43, 53, 58\}$). In both cases, the logits only have large values when $i = j = q$, verifies the matching mechanism in Appendix 4.4.4.2. 83

4.9 Test performance of fully dropping F_1, F_2 when both F_1, F_2 are two-layer MLPs. It turns out, while dropping F_2 makes the model predict correctly w.p. near 1, dropping F_1 has the model predict noise with high probability. 84

4.10 Test performance of fully dropping F_1, F_2 when both F_1 is MLPs and F_2 Linear. Both dropping methods turn out to help predict more correctly than the full model. Meanwhile, dropping the MLP F_1 is better with lower test loss. 84

4.11 Test performance of fully dropping F_1, F_2 when both F_1 is Linear and F_2 MLPs. Only dropping F_2 helps predict more correctly. Dropping F_1 makes the model predicting noise more. 85

4.12 Test performance of low-rank truncating of $\mathbf{W}_O^1$ when there is no F_1, F_2. Here ρ is the fraction of preserved rank of $\mathbf{W}_O^1$, where actually we re-parametrize the first-layer value matrix in attention as $\mathbf{W}_O^1\mathbf{W}_V^1 \in \mathbb{R}^{d\times d}$. It turns out the best $\rho = 0.05$ improves the model's prediction a little. Meanwhile, a smaller ρ destroys the model's performance. 86

4.13 **Left**: average probability of tokens [IO], [S] and "the" in 100-sentence IOI task in the prediction by Pythia-1B along training. **Right**: average probability of tokens "Spain" and "the" in a factual task predicted by Pythia-1B along training, with input as "Madrid is located in". In both tasks, the full model learns to predict "the" with high probability starting from ∼10 steps, and then learns to solve the tasks. LASER boosts the probability of correct answers against "the" in both tasks: the average probability ratio of correct answers against "the" improves from 2.3× to 12.3× (in IOI) and from 0.16× to 11.3× (in factual) at 14K steps. 88

4.14 The training loss of approximating the global bigram π_b with various allocations of parameters in MLP and Attentions. For each configuration of total parameters and ratios, we use the corresponding best learning rate after search to train 100 steps. 91

A.1 A pair of graphs with identical node features, G (left) and G' (right), which can be distinguished by 2 iterations of the WL test but not by the GA-MLP with $\Omega \subseteq \{A^k\}_{k\in\mathbb{N}}$. 108

B.1 Running GD around the local minima of $f(x) = \frac{1}{4}(x^2-1)^2$ (left two) and $f(x) = 2\sin(x)$ (right two) with learning rate $\eta = 1.01 > \frac{2}{f''(\bar{x})} = 1$. Stars denote the start points. It turns out both functions allow stable oscillation around the local minima. 121

B.2 Running GD in the teacher-student setting with learning rate $\eta = 2.2 = 1.1d$, trained on 1000 points uniformly sampled from sphere $\mathcal{S}^1$ of $\|x\| = 1$. The teacher neuron is $\tilde{w} = [1, 0]$ and the student neuron is initialized as $w^{(0)} = [0, 0.1]$ with $v^{(0)} = 0.1$. 121

B.3 Symmetric and Quasi-symmetric Matrix factorization: running GD around flat ($\alpha = 1$) and sharp ($\alpha = 0.8$) minima. In both cases, their leading singular values converge to the same period-2 orbit (about 6.1 and 5.3). (Left: Training loss. Middle: Largest singular value of symmetric case. Right: Largest singular values of quasi-symmetric case.) 122

B.4 Running GD on $f(x, y) = \frac{1}{2}(xy-1)^2$ with learning rate $\eta = 1.05$ (top) and $\eta = 1.25$ (bottom). When $\eta = 1.05$, it converges to a period-2 orbit. When $\eta = 1.25$, it converges to a period-4 orbit. In both cases, $|x - y|$ decays sharply. 123

B.5 The convergent orbits of GD on $f(x) = \frac{1}{4}(x^2-1)^2$ with learning rate=1.05, 1.235 and 1.237. The first two smaller learning rates drive to period-2 orbits while the last one goes to an period-4 orbit. The significant bound between period-2 and period-4 is predictable by Taylor expansion around the period-2 orbit, as $\eta = \sqrt{5} - 1 \approx 1.236$. 193

C.1 Average ranking of tokens “the” in the prediction by Pythia-160M/410M/1B along training. The inputs are 30 preposition words (left) and 40 sentences ending with prepositions. It turns out “the” becomes one of top predictions around 10 steps. . 195

C.2 The prediction distributions of Pythia-410M and 1B on the IOI task. The setting is the same as in Fgure 4.13 (left). The evaluated models are the final checkpoints after training. LASER turns out to decrease the probability of ”the” while keeping that of the correct [IO] high. 196

C.3 The prediction distributions of Pythia-1B, 1.4B and 2.8B on more examples of factual recall. Compared with the setting in Figure 4.13 (right), here we use 20 examples in Table C.3. LASER turns out to significantly decrease the probability of ”the” against the correct tokens. 263

C.4 Predicted probability for $c \in$ {“Mary”, “them”, “the”, “John”}. LASER is conducted on input matrices of MLP layers on the layer $l = 9, 10, 11, 12$ of GPT-2 Small. The input is “When Mary and John went to a store, John gave a drink to”. The horizontal is the fraction of perserved rank, $\rho \in [0, 1]$, where $\rho = 1$ stands for the full model. It turns out LASER clearly decreases probability of “the” and “them” when $\rho \in [0.1, 0.8]$ for layer $l = 9, 10, 11$, compared with the full model. 264

C.5 **Synthetic IOI trained with SGD**: test loss and accuracy for transformers with different layers. Dropping the last-layer MLP consistently improves the test accuracies across all models. 264

C.6 **Synthetic IOI trained with Adam**: test loss and accuracy for transformers with layers $L = 3, 4, 5$. Truncating the last-layer MLP’s input weights with $\rho = 0.01$ improves the test performances for $L = 3, 4$, while the model fails to converge for $L = 5$. 265

List of Tables

2.1 The number of equivalence classes of rooted graphs induced by GNN and GA-MLP on node classification datasets with node features removed. 19

2.2 MSE loss divided by label variance for counting attributed walks on the Cora graph and RRG. The models denoted as "+" contain twice as many powers of the operator. 19

4.1 Probabilities of the top-5 next-tokens in Pythia-1B before and after LASER. The input prompt is "Madrid is located in". Probabilities of two generic words, *i.e.*, "the" and "a", drop sharply after LASER, while probabilities of meaningful words increase, especially the target "Spain". 63

4.2 Few-shot accuracy (%) of pretrained and finetuned language models on GSM8K. Truncating MLPs (LASER) improves reasoning performances in few-shot CoT settings while it has worse performance in the standard 8-shot setting. The LASER hyper-parameters are in Appendix C.1.2. 89

A.1 The number of equivalence classes of graphs induced by GNN and GA-MLP on real datasets with node features removed. The last row gives the ground-truth number of isomorphism classes of graphs computed from the implementation of [Ivanov et al. 2019]. 96

A.2 Results for community detection on binary SBM by GA-MLP-$\tilde{A}_{(1)}$ 115

C.1 $\mu(j,k), \sigma^2(j,k), R(j,k)$ for different choices of (j,k) in Lemma C.2. 200
C.2 All lemmas about the seven cases classified by $\bar{y}$ and k. 218
C.3 Inputs and Outputs of Factual Knowledge 261

1 Introduction

Large language models (LLMs) have achieved remarkable success in the past years. However, while these models have shown impressive capabilities on a variety of tasks, they remain largely black boxes. A better understanding of the role of Transformer layers and how they are affected by the training process could enable new monitoring and editing techniques, better training data, and ultimately more reliable LLMs. This requires a rigorous study under the constraint of limited computational resources. The first possibility is to train large models in a more efficient way, such as fine-tuning with low-rank adapters [Hu et al. 2022]. Nevertheless, involving nonlinear dynamics with a large number of pre-trained parameters on complicated data, it is still challenging to understand the training process of large models. The second possibility is to study simplified settings from scratch, as foundational steps towards understanding the functionality of the giant models. In this thesis, we present our understanding on this from three aspects.

1.1 Expressive Power of Simplified Graph Neural Networks

Graphs are ubiquitous in many real-world applications, such as social networks, molecular structures, multi-body systems and transportation systems. As a modern method of learning representations of graphs, Graph Neural Networks (GNNs) have been widely applied to various tasks including node classification, link prediction and graph classification [Kipf and Welling 2016; Hamilton et al. 2017; Xu et al. 2019; Zhang and Chen 2018; You et al. 2018]. A large por-

tion of GNNs belong to the framework of Message Passing [Gilmer et al. 2017], where the node representations are updated by aggregating information from their neighbors iteratively. This neighborhood-aggregation mechanism allows GNNs to learn representations of nodes by combining information from their local neighborhoods, and by increasing the depth of such GNNs, we increase the size of the receptive field, which hopefully captures more global information of the graph.

Based on such a message-passing framework, there are two lines of research that have gained significant attention in the GNN community. The first one is to study the expressive power of GNNs, which refers to their ability to distinguish non-isomorphic graphs or to approximate certain functions on graphs [Xu et al. 2019; Morris et al. 2019; Maron et al. 2018, 2019a; Chen et al. 2020b; Zhang et al. 2023]. The second one is to improve the training and scalability of GNNs, especially for deep GNNs, by addressing issues such as over-smoothing, oversquashing and efficiency [Kipf and Welling 2016; Li et al. 2018b; Oono and Suzuki 2020; Alon and Yahav 2020; Rossi et al. 2020]. However, there is still a gap between these two lines of research. Typically, when studying expressive power, the community references message-passing GNNs as a baseline model, and compares it with other more complex models, such as GNNs with higher-order tensors [Maron et al. 2019a], aggregated permutation-sensitive functions over permutations [Murphy et al. 2019; Chen et al. 2020b]. But these models are often much less scalable than the original message-passing GNNs, and thus not practical for real-world applications. On the contrary, when improving the training and scalability of GNNs, most works focus on simplifying the architecture of GNNs, such as using fewer layers or reducing the number of parameters, without considering the expressive power of these simplified models. Therefore, we aim to bridge this gap by, for the first time, understanding the expressive power of a simplified GNN architecture, which has proven to be both scalable and effective in practice. The details are in Chapter 2.

1.2 Simplicity Induced by Optimization

While the expressive power of general neural networks is one of the main factors of their success, *i.e.*, universal approximation theorem [Cybenko 1989], it always remains meaningful to study where optimization methods lead to among such a large space of functions. If the effect of optimization is not sufficiently considered, more expressive models may have less guarantee of generalization, following theory of classic machine learning [Mohri et al. 2018]. However, overparametrization, *i.e.*, the number of parameters is much larger than the number of training samples, has been shown to be beneficial for generalization in deep learning [Zhang et al. 2016; Neyshabur et al. 2017; Arora et al. 2018]. This direction of studying how optimization affects the learnt functions is termed as *implicit regularization*. Generally this direction has unveiled a series of intriguing results that the learnt functions by certain optimization methods are simple although the parametric model family is quite large. Such kinds of results typically involve two conditions - certain intialization and limited learning rates. Different initialization scales induces different training regimes. The large initialization is often referred to as *lazy learning* [Jacot et al. 2018; Chizat et al. 2019], where the optimization dynamics are close to linear, and the optimization trajectory is close to the initialization. The small initialization is usually about *feature learning*, where informative features are inherent in data, then the optimization trajectory is dominated by gradient from such features, *e.g.*, single-index models [Arous et al. 2021; Damian et al. 2024] imitate neural networks to learn a single neuron buried in high-dimensional space. Regarding optimization methods, majority of results are for gradient descent (GD) with small learning rates, exactly or approximately following the trajectory of gradient flow, *i.e.*, GD with infinitesimal learning rates, then the continuous dynamics are easier to solve following ODEs instead the discrete one.

However, all of these regimes are still limited in some subspace. The lazy learning stays around initialization, the feature learning moves closely to the subspace from sparse features,

and small learning rates are easier to be stuck around a local minima. Meanwhile, small learning rates in the beginning might be not small any more along training [Cohen et al. 2020], since the curvature of the loss landscape may increase during training, then the trajectory becomes sensitive to the choice of learning rates. Therefore, this motivates us to study what happens when the learning rate is large, especially in the regime that theory tools from classical (convex) optimization cannot guarantee convergence. We study the dynamics with large learning rates in several settings, including 1-D functions, two-layer single-neuron homogeneous networks and matrix factorization. While large learning rates help escape from initialization and sharp local minima, the trajectories are still attracted by certain simple structures, such as period-2 orbits and symmetric solutions. The details are in Chapter 3.

1.3 Roles of Feed-Forward Layers and Attention in Transformers

Next-token prediction is a fundamental task in large language models (LLMs), where the model learns to predict the next word in a sequence given the previous words. Such an autoregressive training objective turns out to bring impressive capabilities in various tasks, including language understanding, generation, and even reasoning. However, the underlying mechanisms of how these models achieve such performance remain largely unclear. In particular, the roles of different components in the Transformer architecture, such as feed-forward layers and attention mechanisms, and their connections with training data distribution are still under active investigation.

We start from a simple yet representative definition of data distribution, as a mixture of in-context reasoning and fixed associations. Such a mixture is inspired by an observation in English. Prepositions like “to”, “for”, “above” are often followed by nouns, and nouns likely begin with “the” or “a”, which means there is a strong co-occurrence between prepositions and “the”. We refer

to this co-occurrence as *distributional association*, against in-context reasoning where models need to infer the next word based on the context. Then, we train a two-layer Transformer model on this mixture of data distribution, and find that the feed-forward layers play a crucial role in learning the associations, while the attention layers are essential for in-context reasoning. This observation is consistent with the training dynamics of the model, where the feed-forward layers learn to memorize the associations quickly, while the attention layers gradually capture the in-context reasoning patterns. Moreover, based on this understanding, modifying feed-forward layers in pre-trained LLMs can significantly lower the model's predicted probability for such associations. The details are in Chapter 4.

2 | Expressive Power of Graph Neural Networks

2.1 Introduction and Our Contributions

While multi-layer Graph Neural Networks (GNNs) have gained popularity for their applications in various fields, recently authors have started to investigate what their true advantages over baselines are, and whether they can be simplified. On one hand, GNNs based on neighborhood-aggregation allows the combination of information present at different nodes, and by increasing the depth of such GNNs, we increase the size of the receptive field. On the other hand, it has been pointed out that deep GNNs can suffer from issues including over-smoothing, exploding or vanishing gradients in training as well as bottleneck effects [Kipf and Welling 2016; Li et al. 2018b; Luan et al. 2019; Oono and Suzuki 2020; Rossi et al. 2020; Alon and Yahav 2020].

Recently, a series of models have attempted at relieving these issues of deep GNNs while retaining their benefit of combining information across nodes, using the approach of firstly augmenting the node features by propagating the original node features through powers of graph operators such as the (normalized) adjacency matrix, and secondly applying a node-wise function to the augmented node features, usually realized by a Multi-Layer Perceptron (MLP) [Wu et al. 2019; NT and Maehara 2019; Chen et al. 2019a; Rossi et al. 2020]. Because of the usage of graph operators for augmenting the node features, we will refer to such models as *Graph-Augmented*

MLPs (GA-MLPs). These models have achieved competitive performances on various tasks, and moreover enjoy better scalability since the augmented node features can be computed during preprocessing [Rossi et al. 2020]. Thus, it becomes natural to ask what advantages GNNs have over GA-MLPs.

In this section, we ask whether GA-MLPs sacrifice expressive power compared to GNNs while gaining these advantages. A popular measure of the expressive power of GNNs is their ability to distinguish non-isomorphic graphs [Hamilton et al. 2017; Xu et al. 2019; Morris et al. 2019]. In this section, besides studying the expressive power of GA-MLPs from the viewpoint of graph isomorphism tests, we propose a new perspective that better suits the setting of node-prediction tasks: we analyze the expressive power of models including GNNs and GA-MLPs as node-level functions, or equivalently, as functions on rooted graphs. Under this perspective, we prove an exponential-in-depth gap between the expressive powers of GNNs and GA-MLPs. We illustrate this gap by finding a broad family of user-friendly functions that can be provably approximated by GNNs but not GA-MLPs, based on counting attributed walks on the graph. Moreover, via the task of community detection, we show a lack of flexibility of GA-MLPs, compared to GNNs, to learn the best operators to use.

In summary, our main contributions are:

- Finding graph pairs that several GA-MLPs cannot distinguish while GNNs can, but also proving there exist simple GA-MLPs that distinguish almost all non-isomorphic graphs.
- From the perspective of approximating node-level functions, proving an exponential gap between the expressive power of GNNs and GA-MLPs in terms of the equivalence classes on rooted graphs that they induce.
- Showing that the functions that count a particular type of attributed walk among nodes can be approximated by GNNs but not GA-MLPs both in theory and numerically.
- Through community detection tasks, demonstrating that GNNs have higher flexibility in

learning than GA-MLPs due to the fixed choice of the operator family in the latter.

2.2 Related Works

Depth in GNNs. [Kipf and Welling 2016] observe that the performance of Graph Convolutional Networks (GCNs) degrade as the depth grows too large, and the best performance is achieved with 2 or 3 layers. Along the spectral perspective on GNNs [Bruna et al. 2013; Defferrard et al. 2016; Bronstein et al. 2017; NT and Maehara 2019], [Li et al. 2018b] and [Wu et al. 2019] explain the failure of deep GCNs by the over-smoothing of the node features. [Oono and Suzuki 2020] show an exponential loss of expressive power as the depth in GCNs increases in the sense that the hidden node states tend to converge to Laplacian sub-eigenspaces as the depth increases to infinity. [Alon and Yahav 2020] show an over-squashing effect of deep GNNs, in the sense that the width of the hidden states needs to grow exponentially in the depth in order to retain all information about long-range interactions. In comparison, our work focuses on more general GNNs based on neighborhood-aggregation that are not limited in the hidden state widths, and demonstrates the their *advantage* in expressive power compared to GA-MLP models at finite depth, in terms of distinguishing rooted graphs for node-prediction tasks. On the other hand, there exist examples of useful deep GNNs. [Chen et al. 2019b] apply 30-layer GNNs for community detection problems, which uses a family of multi-scale operators as well as normalization steps [Ioffe and Szegedy 2015; Ulyanov et al. 2016]. Recently, [Li et al. 2019, 2020a] and [Chen et al. 2020a] build deeper GCN architectures with the help of various residual connections [He et al. 2016] and normalization steps to achieve impressive results in standard datasets, which further highlights the need to study the role of depth in GNNs. [Gong et al. 2020] propose geometrically principled connections, which improve upon vanilla residual connections on graph- and mesh-based tasks.

Existing GA-MLP-type models. Motivated by better understanding GNNs as well as enhancing computational efficiency, several models of the GA-MLP type have been proposed and they achieve competitive performances on various datasets. [Wu et al. 2019] propose the Simple Graph Convolution (SGC), which removes the intermediary weights and nonlinearities in GCNs. [Chen et al. 2019a] propose the Graph Feature Network (GFN), which further adds intermediary powers of the normalized adjacency matrix to the operator family and is applied to graph-prediction tasks. [NT and Maehara 2019] propose the Graph Filter Neural Networks (gfNN), which enhances the SGC in the final MLP step. [Rossi et al. 2020] propose Scalable Inception Graph Neural Networks (SIGNs), which augments the operator family with Personalized-PageRank-based [Klicpera et al. 2018, 2019] and triangle-based [Monti et al. 2018; Chen et al. 2019b] adjacency matrices.

Expressive Power of GNNs. [Xu et al. 2019] and [Morris et al. 2019] show that GNNs based on neighborhood-aggregation are no more powerful than the Weisfeiler-Lehman (WL) test for graph isomorphism [Weisfeiler and Leman 1968], in the sense that these GNNs cannot distinguish between any pair of non-isomorphic graphs that the WL test cannot distinguish. They also propose models that match the expressive power of the WL test. Since then, many attempts have been made to build GNN models whose expressive power are not limited by WL [Morris et al. 2019; Maron et al. 2019a; Chen et al. 2019c; Morris and Mutzel 2019; You et al. 2019; Bouritsas et al. 2020; Li et al. 2020b; Flam-Shepherd et al. 2020; Sato et al. 2019, 2020]. Other perspectives for understanding the expressive power of GNNs include function approximation [Maron et al. 2019b; Chen et al. 2019c; Keriven and Peyré 2019], substructure counting [Chen et al. 2020b], Turing universality [Loukas 2020] and the determination of graph properties [Garg et al. 2020]. [Sato 2020] provides a survey on these topics. In this paper, besides studying the expressive power of GA-MLPs along the line of graph isomorphism tests, we propose a new perspective of approximating functions on rooted graphs, which is motivated by node-prediction tasks, and show a gap between GA-MLPs and GNNs that grows exponentially in the size of the receptive field in terms

of the equivalence classes that they induce on rooted graphs.

2.3 Preliminaries

2.3.1 Notations

Let $G = (V, E)$ denote a graph, with V being the vertex set and E being the edge set. Let n denote the number of nodes in G, $A \in \mathbb{R}^{n\times n}$ denote the *adjacency matrix*, $D \in \mathbb{R}^{n\times n}$ denote the diagonal *degree matrix* with $D_{ii} = d_i$ being the degree of node i. We call $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ the *(symmetrically) normalized adjacency matrix*, and $D^{-\alpha}AD^{-\beta}$ a *generalized normalized adjacency matrix* for any $\alpha, \beta \in \mathbb{R}$. Let $X \in \mathbb{R}^{n\times d}$ denote the matrix of node features, where X_i denotes the d-dimensional feature that node i possesses. For a node $i \in V$, let $\mathcal{N}(i)$ denote the set of neighbors of i. We assume that the edges do not possess features. In a node prediction task, the labels are given by $Y \in \mathbb{R}^n$.

For a positive integer K, we let $[K] = \{1, ..., K\}$. We use $\{...\}_m$ to denote a multiset, which allows repeated elements. We say a function $f(K)$ is *doubly-exponential* in K if $\log\log f(K)$ is polynomial in K, and *poly-exponential* in K if $\log f(K)$ is polynomial in K, as K tends to infinity.

2.3.2 Graph Neural Networks (GNNs)

Following the notations in [Xu et al. 2019], we consider K-layer GNNs defined generically as follows. For $k \in [K]$, we compute the hidden node states $H \in \mathbb{R}^{n\times d^{(k)}}$ iteratively as

$$M_i^{(k)} = \text{AGGREGATE}^{(k)}(\{H_j^{(k-1)} : j \in \mathcal{N}(i)\}), \quad H_i^{(k)} = \text{COMBINE}^{(k)}(H_i^{(k-1)}, M_i^{(k)}), \tag{2.1}$$

where we set $H^{(0)} = X$ to be the node features. If a graph-level output is desired, we finally let

$$Z_G = \text{READOUT}(\{H_i^{(K)} : i \in V\}), \tag{2.2}$$

Different choices of the trainable COMBINE, AGGREGATE and READOUT functions result in different GNN models, though usually AGGREGATE and READOUT are chosen to be permutation-invariant. As graph-level functions, it is shown in [Xu et al. 2019] and [Morris et al. 2019] that the maximal expressive power of models of this type coincides with running K iterations of the WL test for graph isomorphism, in the sense that any two non-isomorphic graphs that cannot be distinguished by the latter cannot be distinguished by the K-layer GNNs, either. For this reason, we will not distinguish between GNN and WL in discussions on expressive powers.

2.3.3 Graph-Augmented Multi-Layer Peceptrons (GA-MLPs)

GA-MLPs are models that consist of two steps - first augmenting the node features with some operators based on the graph topology, and then applying a node-wise learnable function. Below we focus on using linear graph operators to augment the node features, while an extension of the definition as well as some of the theoretical results in Section 2.5 to GA-MLPs using general graph operators is given in Appendix A.1. Let $\Omega = \{\omega_1(A), ..., \omega_K(A)\} \subseteq \mathbb{R}^{n\times n}$ be a set of (usually multi-hop) linear operators that are functions of the adjacency matrix, A. Common choices of the operators are powers of the (normalized) adjacency matrix, and several particular choices of Ω that give rise to existing GA-MLP models are listed in Appendix A.2. In its general form, a GA-MLP first computes a series of augmented features via

$$\tilde{X}_k = \omega_k(A) \cdot \varphi(X) \in \mathbb{R}^{n\times \tilde{d}}, \tag{2.3}$$

with $\varphi : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ being a learnable function acting as a feature transformation applied to each node separately. It can be realized by an MLP, e.g. $\varphi(X) = \sigma(XW_1)W_2$, where σ is a nonlinear activation function and W_1, W_2 are trainable weight matrices of suitable dimensions. Next, the

model concatenates $\tilde{X}_1, ..., \tilde{X}_K$ into $\tilde{X} = [\tilde{X}_1, ..., \tilde{X}_K] \in \mathbb{R}^{n \times (K\tilde{d})}$, and computes

$$Z = \rho(\tilde{X}) \in \mathbb{R}^{n \times d'}, \tag{2.4}$$

where $\rho : \mathbb{R}^{K\tilde{d}} \to \mathbb{R}^{d'}$ is also a learnable node-wise function, again usually realized by an MLP. If a graph-level output is desired, we can also add a READOUT function as in (2.2).

A simplified version of the model sets φ to be the identity function, in which case (2.3) and (2.4) can be written together as

$$Z = \rho([\omega_1(A) \cdot X, ..., \omega_K(A) \cdot X]) \tag{2.5}$$

Such a simplification improves computational efficiency since the matrix products $\omega_k(A) \cdot X$ can be pre-computed before training [Rossi et al. 2020]. Since we are mostly interested in an upper bounds on the expressive power of GA-MLPs, we will work with the more general update rule (2.3) in this paper, but the lower-bound result in Proposition 2.2 remains valid even when we restrict to the subset of models where φ is taken to be the identity function.

2.4 Expressive Power as Graph Isomorphism Tests

We first study the expressive power of GA-MLPs via their ability to distinguish non-isomorphic graphs. It is not hard to see that when $\Omega = \{I, \tilde{A}, ..., \tilde{A}^K\}$, where $\tilde{A} = D^{-\alpha}AD^{-\beta}$ for any $\alpha, \beta \in \mathbb{R}$ generalizes the normalized adjacency matrix, this is upper-bounded by the power of $K + 1$ iterations of WL. We next ask whether it can fall strictly below. Indeed, for two common choices of Ω, we can find concrete examples: 1) If Ω consists of integer powers of any normalized adjacency matrix of the form $D^{-\alpha}AD^{-(1-\alpha)}$ for some $\alpha \in [0, 1]$, then it is apparent that the GA-MLP cannot distinguish any pair of *regular graphs* with the same size but different node degrees; 2) If Ω consists of integer powers of the adjacency matrix, A, then the model cannot distinguish between

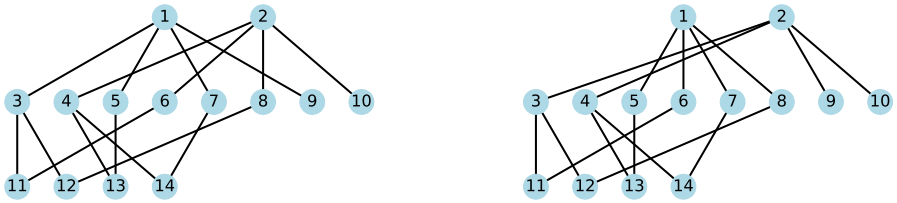


Figure 2.1: A pair of graphs that can be distinguished by 2 iterations of the WL test but not by GA-MLPs with $\Omega \subseteq \{A^k : k \in \mathbb{N}\}$, as proved in Appendix A.10.

the pair of graphs shown in Figure 1, which can be distinguished by 2 iterations of the WL test. The proof of the latter result is given in Appendix A.10. Together, we summarize the results as:

Proposition 2.1. *If $\Omega \subseteq \{\tilde{A}^k : k \in \mathbb{N}\}$, with either $\tilde{A} = A$ or $\tilde{A} = D^{-\alpha}AD^{-(1-\alpha)}$ for some $\alpha \in [0, 1]$, there exists a pair of graphs which can be distinguished by GNNs but not this GA-MLP.*

Nonetheless, if we focus on not particular counterexamples but rather the average performance in distinguishing random graphs, it is not hard for GA-MLPs to reach the same level as WL, which is known to distinguish almost all pairs of random graphs under a uniform distribution [Babai et al. 1980]. Specifically, building on the results in [Babai et al. 1980], we prove in Appendix A.5 that:

Proposition 2.2. *For all $n \in \mathbb{N}_+$, $\exists \alpha_n > 0$ such that any GA-MLP that has $\{D, AD^{-\alpha_n}\} \subseteq \Omega$ can distinguish almost all pairs of non-isomorphic graphs of at most n nodes, in the sense that the fraction of graphs on which such a GA-MLP fails to test isomorphism is $1 - o(1)$ as $n \to \infty$.*

The hypothesis that distinguishing non-isomorphic graphs is not difficult on average for either GNNs or GA-MLPs is further supported by the numerical results provided in Appendix A.3, in which we count the number of equivalence classes that either of them induce on graphs that occur in real-world datasets. This further raises the question of whether graph isomorphism tests along suffice as a criterion for comparing the expressive power of models on graphs, which leads us to the explorations in the next section.

Lastly, we remark that with suitable choices of operators in Ω, it is possible for GA-MLPs to go beyond the power of WL. For example, if Ω contains the *power graph adjacency matrix* introduced

in [Chen et al. 2019b], $\min(A^2, 1)$, then the GA-MLP can distinguish between a hexagon and a pair of triangles, which WL cannot distinguish.

2.5 Expressive Power as Functions on Rooted Graphs

To study the expressive power beyond graph isomorphism tests, we consider the setting of node-wise prediction tasks, for which the final readout step (2.2) is dropped in both GNNs and GA-MLPs. Whether the learning setup is transductive or inductive, we can consider the models as functions on *rooted graphs*, or *egonets* [Preciado and Jadbabaie 2010], which are graphs with one node designated as the root $\{i_1, ..., i_n\}$ is a set of nodes in the graphs $\{G_1, ..., G_n\}$ (not necessarily distinct) and with node-level labels $\{Y_{i_1}, ..., Y_{i_n}\}$ known during training, respectively, then the goal is to fit a function to the input-output pairs $(G_n^{[i_n]}, Y_{i_n})$, where we use $G^{[i]}$ to denote the rooted graph with G being the graph and the node i in G being the root. Thus, we can evaluate the expressive power of GNNs and GA-MLPs by their ability to approximate functions on the space of rooted graphs, which we call $\mathcal{E}$.

To do so, we introduce a notion of induced equivalence relations on $\mathcal{E}$, analogous to the equivalence relations on $\mathcal{G}$ introduced in Appendix A.3. Given a family of functions $\mathcal{F}$ on $\mathcal{E}$, we can define an equivalence relation $\simeq_{\mathcal{E};\mathcal{F}}$ among all rooted graphs such that $\forall G^{[i]}, G'^{[i']} \in \mathcal{E}$, $G^{[i]} \simeq_{\mathcal{E};\mathcal{F}} G'^{[i']}$ if and only if $\forall f \in \mathcal{F}, f(G^{[i]}) = f(G'^{[i']})$. By examining the number and sizes of the induced equivalence classes of rooted graphs, we can evaluate the relative expressive power of different families of functions on $\mathcal{E}$ in a quantitative way.

In the rest of this section, we assume that the node features belong to a finite alphabet $\mathcal{X} \subseteq \mathbb{N}$ and all nodes have degree at most $m \in \mathbb{N}_+$. Firstly, GNNs are known to distinguish neighborhoods up to the *rooted aggregation tree*, which can be obtained by unrolling the neighborhood aggregation steps in the GNNs as well as the WL test [Xu et al. 2019; Morris et al. 2019; Garg et al. 2020]. The *depth-K rooted aggregation tree* of a rooted graph $G^{[i]}$ is a depth-K rooted tree with a

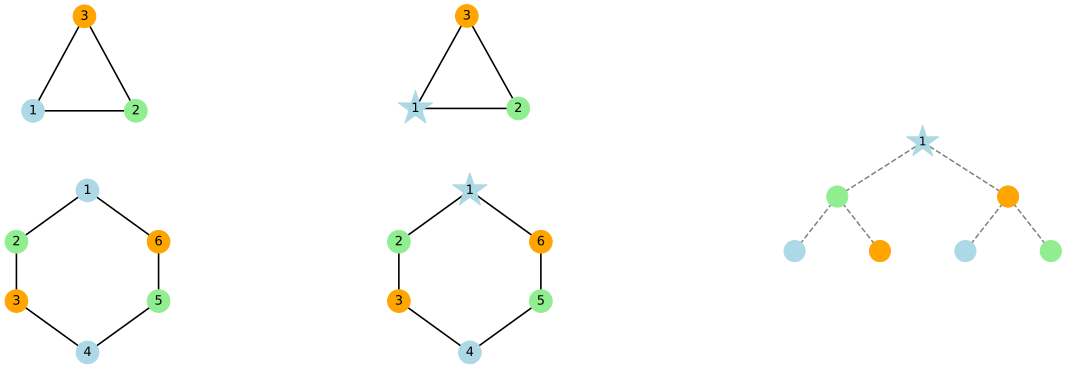


Figure 2.2: An illustration of rooted graphs and rooted aggregation trees. *Left*: a pair of graphs, G and G'. *Center*: the rooted graphs of 1 in G and G', $G^{[1]}$ and $G'^{[1]}$. *Right*: the rooted aggregation tree that both $G^{[1]}$ and $G'^{[1]}$ correspond to.

(possibly many-to-one) mapping from every node in the tree to some node in $G^{[i]}$, where (i) the root of the tree is mapped to node i, and (ii) the children of every node j in the tree are mapped to the neighbors of the node in $G^{[i]}$ to which j is mapped. An illustration of rooted graphs and rooted aggregation trees is given in Figure 2.2. Hence, each equivalence class in $\mathcal{E}$ induced by the family of all depth-K GNNs consists of all rooted graphs that share the same rooted aggregation tree of depth-K. Thus, to estimate the number of equivalence classes on $\mathcal{E}$ induced by GNNs, we need to estimate the number of possible rooted aggregation trees, which is given by Lemma A.6 in Appendix A.6. Thus, we derive the following lower bound on the number of equivalence classes in $\mathcal{E}$ that depth-K GNNs induce:

Proposition 2.3. *Assume that $|\mathcal{X}| \geq 2$ and $m \geq 3$. The total number of equivalence classes of rooted graphs induced by GNNs of depth K grows at least doubly-exponentially in K.*

In comparison, we next demonstrate that the equivalence classes induced by GA-MLPs are more coarsened. To see this, let's first consider the example where we take $\Omega = \{I, \tilde{A}, \tilde{A}^2, ..., \tilde{A}^K\}$, in which $\tilde{A} = D^{-\alpha}AD^{-\beta}$ with any $\alpha, \beta \in \mathbb{R}$ is a generalization of the normalized adjacency matrix.

From formula (2.3), by expanding the matrix product, we have

$$(\tilde{A}^k\varphi(X))_i = \sum_{(i_1,\ldots,i_k)\in\mathcal{W}_k(G^{[i]})} d_i^{-\alpha} d_{i_1}^{-(\alpha+\beta)} \ldots d_{i_{k-1}}^{-(\alpha+\beta)} d_{i_k}^{-\beta} \varphi(X_{i_k}) \,, \tag{2.6}$$

where we define $\mathcal{W}_k(G^{[i]}) = \{(i_1, \ldots, i_k) \subseteq V : A_{i,i_1}, A_{i_1,i_2}, \ldots, A_{i_{k-1},i_k} > 0\}$ to be set of all *walks* of length k in the rooted graph $G^{[i]}$ starting from node i (an illustration is given in Figure 2.3). Thus, the kth augmented feature of node i, $(\tilde{A}^k\varphi(X))_i$, is completely determined by the number of each "type" of walks in $G^{[i]}$ of length k, where the type of a walk, $(i_1, \ldots, i_k)$, is determined jointly by the degree multiset, $\{d_{i_1}, \ldots, d_{i_{k-1}}\}$ as well as the degree and the node feature of the end node, d_{i_k} and X_{i_k}. Hence, to prove an upper bound on the total number of equivalence classes on $\mathcal{E}$ induced by such a GA-MLP, it is sufficient to upper-bound the total number of possibilities of assigning the counts of all types of walks in a rooted graph. This allows us to derive the following result, which we prove in Appendix A.7.

Proposition 2.4. *Fix $\Omega = \{I, \tilde{A}, \tilde{A}^2, \ldots, \tilde{A}^K\}$, where $\tilde{A} = D^{-\alpha}AD^{-\beta}$ for some $\alpha, \beta \in \mathbb{R}$. Then the total number of equivalence classes in $\mathcal{E}$ induced by such GA-MLPs is poly-exponential in K.*

Compared with Proposition 2.3, this shows that the number of equivalence classes on $\mathcal{E}$ induced by such GA-MLPs is exponentially smaller than that by GNNs. In addition, as the other side of the same coin, these results also indicate the complexity of these hypothesis classes. Building on the results in [Chen et al. 2019c, 2020b] on the equivalence between distinguishing non-isomorphic graphs and approximating arbitrary permutation-invariant functions on graphs by neural networks, and by the definition of *VC dimension* [Vapnik and Chervonenkis 1971; Mohri et al. 2018], we conclude that

Corollary 2.5. *The VC dimension of all GNNs of K layers as functions on rooted graphs grows at least doubly-exponentially in K; Fixing $\alpha, \beta \in \mathbb{R}$, the VC dimension of all GA-MLPs with $\Omega = \{I, \tilde{A}, \tilde{A}^2, \ldots, \tilde{A}^K\}$ as functions on rooted graphs is at most poly-exponential in K.*

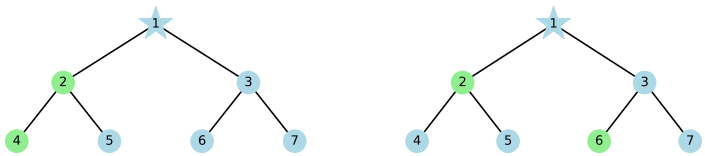


Figure 2.3: A pair of rooted graphs, $G^{[1]}$ (left) and $G'^{[1]}$ (right), in which blue nodes have node feature 0 and green nodes have node feature 1. They belong to the same equivalence class induced by any GA-MLP with operators that only depend on the graph structure, but different equivalence classes induced by GNNs. In particular, $G^{[1]}$ and $G'^{[1]} \in \mathcal{T}_{2,2,(1,1,3)}$ (defined in Appendix A.8), and $|\mathcal{W}_2(G^{[1]}; (1,1))| = 1$ whereas $|\mathcal{W}_2(G'^{[1]}; (1,1))| = 0$.

Meanwhile, for more general operators, we can show that the equivalence classes induced by GA-MLPs are *coarser* than those induced by GNNs at least under some measurements. For instance, the pair of rooted graphs in Figure 2.3 belong to the same equivalence class induced by any GA-MLP (as we prove in Appendix A.8) but different equivalence classes induced by GNNs. Rigorously, we characterize such a gap in expressive power by finding certain equivalence classes in $\mathcal{E}$ induced by GA-MLPs that intersect with many equivalence classes induced by GNNs. In particular, we have the following general result, which we prove in Appendix A.8:

Proposition 2.6. *If Ω is any family of equivariant linear operators on the graph that only depend on the graph topology of at most K hops, then there exist exponentially-in-K many equivalence classes in $\mathcal{E}$ induced by the GA-MLPs with Ω, each of which intersects with doubly-exponentially-in-K many equivalence classes in $\mathcal{E}$ induced by depth-K GNNs, assuming that $|\mathcal{X}| \geq 2$ and $m \geq 3$. Conversely, in constrast, if $\Omega = \{I, \tilde{A}, \tilde{A}^2, ..., \tilde{A}^K\}$, in which $\tilde{A} = D^{-\alpha}AD^{-\beta}$ with any $\alpha, \beta \in \mathbb{R}$, then each equivalence class in $\mathcal{E}$ induced by depth-$(K+1)$ GNNs is contained in one equivalence class induced by the GA-MLPs with Ω.*

In essence, this result establishes that GA-MLP circuits can express fewer (exponentially fewer) functions than GNNs with equivalent receptive field. Taking a step further, we can find explicit functions on rooted graphs that can be approximated by GNNs but not GA-MLPs. In the framework that we have developed so far, this occurs when the image of each equivalence class in $\mathcal{E}$ induced by GNNs under this function contains a single value, whereas the image of some equivalence class in $\mathcal{E}$ induced by GA-MLPs contains multiple values. Inspired by the proofs of

the results above, a natural candidate is the family of functions that count the number of walks of a particular type in the rooted graph. We can establish the following result, which we prove in Appendix A.9:

Proposition 2.7. *For any sequence of node features $\{x_k\}_{k\in\mathbb{N}_+} \subseteq \mathcal{X}$, consider the sequence of functions $f_k(G^{[i]}) := |\mathcal{W}_k(G^{[i]}; (x_1, ..., x_k))|$ on $\mathcal{E}$. For all $k \in \mathbb{N}_+$, the image under f_k of every equivalence class in $\mathcal{E}$ induced by depth-k GNNs contains a single value, while for any GA-MLP using equivariant linear operators that only depend on the graph topology, there exist exponentially-in-k many equivalence classes in $\mathcal{E}$ induced by this GA-MLP whose image under f_k contains exponentially-in-k many values.*

In other words, there exist graph instances where the attributed-walk-counting-function f_k takes different values, yet no GA-MLP model can predict them apart – and there are exponentially many of these instances as the number of hops increases. This suggests the possibility of lower-bounding the average approximation error for certain functions by GA-MLPs under various random graph families, which we leave for future work.

2.6 Experiments

The baseline GA-MLP models we consider has operator family $\Omega = \{I, A, ..., A^K\}$ for a certain K, and we call it GA-MLP-A. In Section 2.6.2 and 2.6.3, we also consider GA-MLPs with $\Omega = \{I, \tilde{A}_{(1)}, ..., \tilde{A}^K_{(1)}\}$ ($\tilde{A}_{(\epsilon)}$ is defined in Appendix A.2), denoted as GA-MLP-$\tilde{A}_{(1)}$. For the experiments in Section 2.6.3, due to the large K as well as the analogy with spectral methods [Chen et al. 2019b], we use instance normalization [Ulyanov et al. 2016]. Further details are described in Appendix A.11.

	Cora		Citeseer		Pubmed	
# Nodes	2708		3327		19717	
K	GNN	GA-MLP	GNN	GA-MLP	GNN	GA-MLP
1	37	37	31	31	82	82
2	1589	756	984	506	8059	3762
3	2301	2158	1855	1550	12814	12014
4	2363	2359	2074	2019	12990	12979
5	2365	2365	2122	2115	12998	12998

Table 2.1: The number of equivalence classes of rooted graphs induced by GNN and GA-MLP on node classification datasets with node features removed.

	Cora		RRG	
Model	Train	Test	Train	Test
GIN	3.98E-6	9.72E-7	3.39E-5	2.61E-4
GA-MLP-A	1.23E-1	1.56E-1	1.75E-2	2.13E-2
GA-MLP-A+	1.87E-2	6.44E-2	1.69E-2	2.13E-2
GA-MLP-$\tilde{A}_{(1)}$	4.22E-1	5.79E-1	1.02E-1	1.58E-1
GA-MLP-$\tilde{A}_{(1)}$+	4.00E-1	5.79E-1	1.12E-1	1.52E-1

Table 2.2: MSE loss divided by label variance for counting attributed walks on the Cora graph and RRG. The models denoted as "+" contain twice as many powers of the operator.

2.6.1 Number of equivalence classes of rooted graphs

Motivated by Propositions 2.3 and 2.4, we numerically count the number of equivalence classes induced by GNNs and GA-MLPs among the rooted graphs found in actual graphs with node features removed. For depth-K GNNs, we implement a WL-like process with hash functions to map the depth-K egonet associated with each node to a string before comparing across nodes. For GA-MLP-A, we compare the augmented features of each egonet computed via (2.3). From the results in Table 2.1, we see that indeed, the number of equivalence classes induced by GA-MLP-A is smaller than that by GNNs, with the highest relative difference occurring at $K = 2$. The contrast is much more visible than their difference in the number of *graph* equivalence classes given in Appendix A.3.

2.6.2 Counting attributed walks

Motivated by Proposition 2.7, we test the ability of GNNs and GA-MLPs to count the number of walks of a particular type in synthetic data. We take graphs from the Cora dataset (with node features removed) as well as generate a random regular graph (RRG) with 1000 nodes and the node degree being 6. We assign node feature *blue* to all nodes with even index and node feature *red* to all nodes with odd index, due to which the node feature is given by 2-dimensional one-hot encoding. On the Cora graph, a node i's label is given by the number of walks of the type

blue→*blue*→*blue* starting from i. On the RRG, the label is given by the number of walks of the type *blue*→*blue*→*blue*→*blue* starting from i. The number of nodes for training and testing is split as 1000/1708 for the Cora graph and 300/700 for the random regular graph. We test four GA-MLP models, two with as many powers of the operator as the walk length and the other two with twice as many operators, and compare their performances against that of the Graph Isomorphism Network (GIN), a GNN model that achieves the expressive power of the WL test [Xu et al. 2019]. From Table 2.2, we see that GIN significantly outperforms GA-MLPs in both training and testing on both graphs, consistent with the theoretical result in Proposition 2.7 that GNNs can count attributed walks while GA-MLPs cannot. Thus, this points out an intuitive task that lies in the gap of expressive power between GNNs and GA-MLPs.

2.6.3 Community detection on Stochastic Block Models (SBM)

We use the task of community detection to illustrate another limitation of GA-MLP models: a lack of flexibility to *learn* the family of operators. SBM is a random graph model in which nodes are partitioned into underlying communities and each edge is drawn independently with a probability that only depends on whether the pair of nodes belong to the same community or not. The task of community detection is then to recover the community assignments from the connectivity pattern. We focus on binary (that is, having two underlying communities) SBM in the sparse regime, where it is known that the hardness of detecting communities is characterized by a signal-to-noise ratio (SNR) that is a function of the in-group and out-group connectivity [Abbe 2017]. We select 5 pairs of in-group and out-group connectivity, resulting in 5 different hardness levels of the task.

Among all different approaches to community detection, spectral methods are particularly worth mentioning here, which usually aim at finding a certain eigenvector of a certain operator that is correlated with the community assignment, such as the second largest eigenvector of the adjacency matrix or the second smallest eigenvector of the Laplacian matrix or the Bethe

Hessian matrix [Krzakala et al. 2013]. In particular, the Bethe Hessian matrix is known to be asymptotically optimal in the hard regime, provided that a data-dependent parameter is known. Note that spectral methods bear close resemblance to GA-MLPs and GNNs. In particular, [Chen et al. 2019b] propose a spectral GNN (sGNN) model for community detection that can be viewed as a learnable generalization of power iterations on a collection of operators. Further details on Bethe Hessian and sGNN are provided in Appendix A.11.

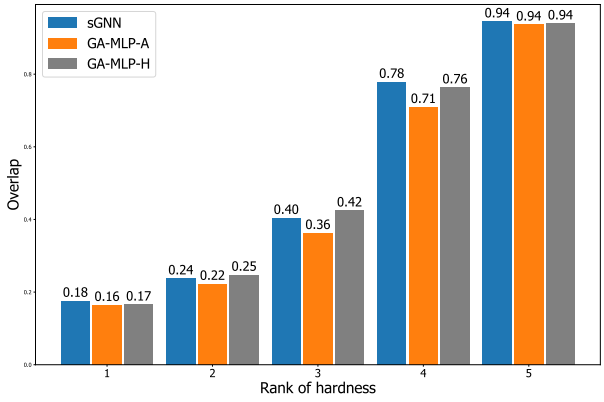


Figure 2.4: Community detection on binary SBM with 5 choices of in- and out-group connectivities, each yielding to a different SNR. Higher overlap means better performance.

We first compare two variants of GA-MLP models: GA-MLP-A with $K = 120$, and GA-MLP-H with Ω generated from the Bethe Hessian matrix with the oracle data-dependent parameter also up to $K = 120$. From Figure 2.4, we see that the latter consistently outperforms the former, indicating the importance of the choice of the operators for GA-MLPs. As reported in Appendix A.11, replacing A by $\tilde{A}_{(1)}$ yields no improvement in performance. Meanwhile, we also test a variant of sGNN that is only based on powers of the A and has the same receptive field as GA-MLP-A (further details given in Appendix A.11). We see that its performance is comparable to that of GA-MLP-H. Thus, this demonstrates a scenario in which GA-MLP with common choices of Ω do not work well, but there exists some choice of Ω that is a priori unknown, with which GA-MLP can achieve good performance. In contrast, a GNN model does not need to rely on the knowledge of such an oracle set of operators, demonstrating its superior capability of learning.

2.7 Conclusions

We have studied the separation in terms of representation power between GNNs and a popular alternative that we coined GA-MLPs. This latter family is appealing due to its computational scalability and its conceptual simplicity, whereby the role of topology is reduced to creating 'augmented' node features then fed into a generic MLP. Our results show that while GA-MLPs can distinguish almost all non-isomorphic graphs, in terms of approximating node-level functions, there exists a gap growing exponentially-in-depth between GA-MLPs and GNNs in terms of the number of equivalence classes of nodes (or rooted graphs) they induce. Furthermore, we find a concrete class of functions that lie in this gap given by the counting of attributed walks. Moreover, through community detection, we demonstrate the lack of GA-MLP's ability to go beyond the fixed family of operators as compared to GNNs. In other words, GNNs possess an inherent ability to discover topological features through learnt diffusion operators, while GA-MLPs are limited to a fixed family of diffusions.

While we do not attempt to provide a decisive answer of whether GNNs or GA-MLPs should be preferred in practice, our theoretical framework and concrete examples help to understand their differences in expressive power and indicate the types of tasks in which a gap is more likely to be seen – those exploiting stronger structures among nodes like counting attributed walks, or those involving the learning of graph operators. That said, our results are purely on the representation side, and disregard optimization considerations; integrating the possible optimization counterparts is an important direction of future improvement. Finally, another open question is to better understand the links between GA-MLPs and spectral methods, and how this can help learning diffusion operators.

3 Optimization Instabilities in Low-dimensional Space

3.1 Introduction and Our Contributions

Given a differentiable objective function $f(\theta)$, where $\theta \in \mathbb{R}^d$ is a high-dimensional parameter vector, the most basic and widely used optimization method is gradient descent (GD), defined as

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta f(\theta^{(t)}), \tag{3.1}$$

where η is the learning rate. For all its widespread application across many different ML setups, a basic question remains: what are the convergence guarantees (even to a local minimiser) under typical objective functions, and how they depend on the (only) hyperaparameter η? In the modern context of large-scale ML applications, an additional key question is not only to understand whether or not GD converges to minimisers, but to *which* ones, since overparametrisation defines a whole manifold of global minimisers, all potentially enjoying drastically different generalisation performance.

The sensible regime to start the analysis is $\eta \to 0$, where GD inherits the local convergence properties of the Gradient Flow ODE via standard arguments from numerical integration. However, in the early phase of training, a large learning rate has been observed to result in better

generalization [LeCun et al. 2012; Bjorck et al. 2018; Jiang et al. 2019; Jastrzebski et al. 2021], where the extent of "large" is measured by comparing the learning rate η and the curvature of the loss landscape, measured with $\lambda(\theta) := \lambda_{\max}\left[\nabla^2_\theta f(\theta)\right]$, the largest eigenvalue of the Hessian with respect to learnable parameters. Although one requires $\sup_\theta \lambda(\theta) < 2/\eta$ to guarantee the convergence of GD [Bottou et al. 2018] to (local) minimisers [1], the work of [Cohen et al. 2020] noticed a remarkable phenomena in the context of neural network training: even in problems where $\lambda(\theta)$ is unbounded (as in NNs), for a fixed η, the curvature $\lambda(\theta^{(t)})$ increases along the training trajectory (3.1), bringing $\lambda(\theta^{(t)}) \geqslant 2/\eta$ [Cohen et al. 2020]. After that, a surprising phenomena is that $\lambda(\theta^{(t)})$ *stably* hovers above $2/\eta$ and the neural network still eventually achieves a decreasing training loss — the so-called "Edge of Stability". We would like to understand and analyse the conditions of such convergence with a large learning rate under a variety models that capture such observed empirical behavior.

Recently, some works have built connections between EoS and implicit bias [Arora et al. 2022; Lyu et al. 2022; Damian et al. 2021, 2022b] in the context of large, overparametrised models such as neural networks. In this setting, GD is expected to converge to a manifold of minimisers, and the question is to what extent a large learning rate 'favors' solutions with small curvature. In essence, these works show that under certain structural assumptions, GD is asymptotically tracking a continuous sharpness-reduction flow, in the limit of small learning rates. Compared with these, we study non-asymptotic properties of GD beyond EoS, by focusing on certain learning problems (*e.g.*, single-neuron ReLU networks and matrix factorization). In particular, we characterize a range of learning rates η *above* the EoS such that GD dynamics hover around minimisers. Moreover, in the matrix factorization setup, where minimisers form a manifold with varying local curvature, our results give a non-asymptotic analogue of the 'Sharpness-Minimisation' arguments from [Arora et al. 2022; Lyu et al. 2022; Damian et al. 2022b].

The straightforward starting point for the local convergence analysis is via Taylor approxi-

[1] One can replace the uniform curvature bound by $\sup_{\theta: f(\theta) \leq f(\theta^{(0)})} \lambda(\theta)$.

mations of the loss function. However, in a quadratic Taylor expansion, gradient descent diverges once $\lambda(\theta) > 2/\eta$ [Cohen et al. 2020], indicating that a higher order Taylor approximation is required. By considering a 1-D function with local minima θ^* of curvature $\lambda^* = \lambda(\theta^*)$, we show the existence of fixed points of two-step updates around the minima with η slightly above the threshold $2/\lambda^*$, provided its high order derivative satisfies mild conditions as in Theorem 3.1, with generalization into matrix factorization in Theorem B.3 and experiments of MLPs in Appendix 3.7.2. A typical example of such functions is $f(x) = \frac{1}{4}(x^2 - \mu)^2$ with $\mu > 0$. Furthermore, we prove that it converges to an orbit of period 2 from a more global initialization rather than the analysis of high-order local approximation.

As it turns out, the analysis of such stable one-dimensional oscillations is sufficiently intrinsic to become useful in higher-dimensional problems. First, we leverage the analysis to a two-layer single-neuron ReLU network, where the task is to learn a teacher neuron with data on a uniform high-dimensional sphere. We show a convergence result under population loss with GD beyond EoS, where the direction of the teacher neuron can be learnt and the norms of two-layer weights stably oscillate, with empirical evidence of 16-neuron networks in Appendix 3.7.1. We then focus on matrix factorization, a canonical non-convex problem whose geometry is characterized by a manifold of minimisers having different local curvature. We provide novel observations of its convergence to period-2 orbit with comprehensive theoretical intuition of the dynamics. Finally, we extend previous works by proposing two models with observations in matrix factorization compatible for future analysis.

3.1.1 Connections between theoretical results

In this section, we discuss the connections between our presented theoretical results, as illustrated in Figure 3.1.

Theorem 3.1 and Lemma 3.2 present (local) intrinsic geometric properties for a 1-D function to allow stable oscillations. Such properties provide us the 1-D function $f(x) = (\mu - x^2)^2$ and,

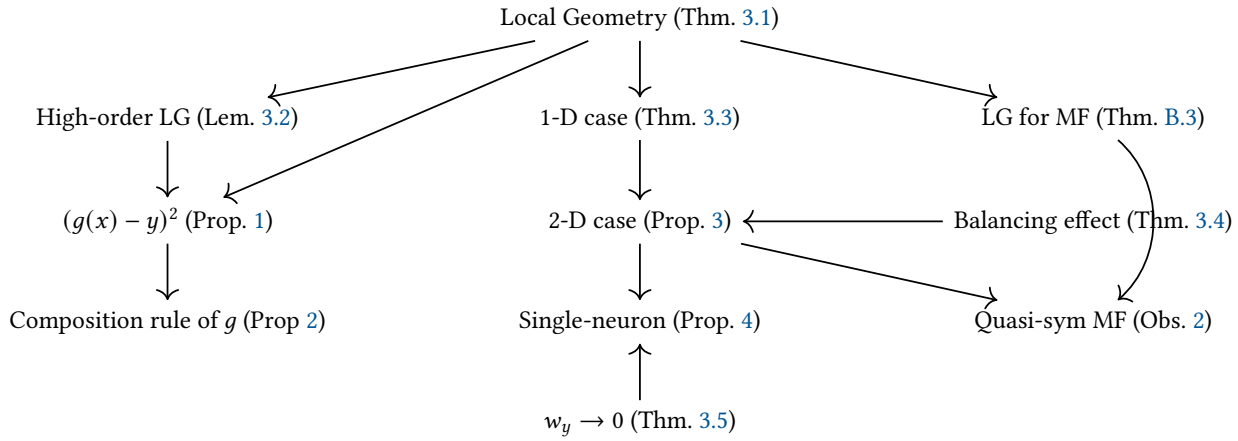


Figure 3.1: Connections between our presented theoretical results. The arrows stand for "implies". LG stands for Local Geometry. MF stands for Matrix Factorization.

furthermore, we generalize the local property to a global convergence result in Theorem 2. Then we are to generalize the 1-D analysis to cases of i) multi-parameter, ii) nonlinear and iii) high-dimension.

(a) **Multi-parameter**. Compared with 1-D $f(x) = (\mu - x^2)^2$, the 2-D function $f(x, y) = (\mu - xy)^2$ can be viewed as the simplest setting of two-layer models. We prove that the 2-D case converges to the region of $x = y$ in Theorem 3.4 in Section 3.4.2, which means it shares the same convergence as the 1-D model. Also, $x = y$ means its sharpness is the flattest.

(b) **Nonlinear**. We extend the 2-D model to a two-layer single-neuron ReLU model in Section 3.5. Although the student neuron can be initialized far from the direction of the teacher neuron, we prove the student neuron converges to the correct direction (as $w_y \to 0$) in Theorem 3.5. Then the problem degenerates to the above 2-D analysis, which means it shares the same convergence with the 2-D, where (v, w_x) corresponds to (x, y) in 2-D.

(c) **High-dimension**. We extend the 2-D model to quasi-symmetric matrix factorization in Section 3.6. Although the parameters are initialized near a sharp minima, GD still walks towards the flattest minima, as shown in Observation 2. At convergence, the top singular

values of $\mathbf{Y}, \mathbf{Z}$ are the same, following the 2-D analysis. So the singular values are in the same period-2 orbit as the 1-D case.

Meanwhile, from Theorem 3.1 and Lemma 3.2, we prove a condition for base models g in regression tasks to allow stable oscillation in Prop 1. Furthermore, we provide a composition rule of two base models to find a more complicated model that allows stable oscillation in Prop 2.

3.1.2 Implications from low-dimension to high-dimension

We would like to emphasize that, although our current simple settings are a little far from practical NNs, it still helps understand the ability of GD at large LRs to discover flat minima in three steps as follows. We include more experiments in Appendix 3.7 to present the following hopes for complicated networks:

(a) By Theorem 3.1, especially its second condition, we wish to discover an intrinsic geometric property around local minima of more complicated models. The key is to investigate the 1-D function at the cross-section of the leading eigenvector and the loss landscape.

- ◆ Theoretical: we prove the 1-D condition holds at any minima for non-trivial matrix factorization, shown as Theorem B.3 in Appendix B.1.2.
- ✻ Empirical: we show the 1-D condition holds around minima of 3,4,5-layer ReLU MLPs on MNIST, shown in Figure 3.6(d), 3.7(d), 3.8(d) in Appendix 3.7.2.

(b) With the above intrinsic geometric property, the next question is whether the training trajectory utilizes this property.

- ◆ Theoretical: in the case of quasi-symmetric matrix factorization, we observe and provide theoretical intuition that the training trajectory follows the leading eigenvector of the Hessian (i.e. the leading component of $\mathbf{X}_0$) in Observation 2, where the only top components of weights are changing in $\omega(\epsilon)$.

- ✻ Empirical: for MLPs on MNIST, we show the almost perfect alignment of the gradient and the top Hessian eigenvector in Figure 3.6(c), 3.7(c), 3.8(c).

(c) The final implication is the implicit bias of EoS after such oscillation. It turns out GD is driven to flatter minima from sharper minima. In the 1-D case, obviously there is nothing about implicit bias since the only thing GD is doing is to approximate the target value. However, an implicit bias from the oscillation appears starting from the 2-D case.

- ◆ Theoretical 1: in the 2-D case in Theorem 3.4, we prove the two learnable parameters x, y will converge to the same values after oscillations of their product xy. Actually in the minimum manifold, smaller $|x - y|$ means a flatter minimizer.
- ◆ Theoretical 2: in the single-neuron ReLU network in Theorem 3.5 and Prop 4, we show the model degenerates to the 2-D case since $w_y \to 0$. The 2-D argument tells that this nonlinear model also walks towards the balanced situation, verified with experiments in Figure B.2.
- ◆ Theoretical 3: in the quasi-symmetric MF in Obs 2, although the initialization is around a sharp minima, GD is still driven towards the flattest minima where $\sigma_{\max}(\mathbf{Y}) = \sigma_{\max}(\mathbf{Z})$.
- ✻ Empirical 1: for 2-layer 16-neuron ReLU network in a student-teacher setting, it turns out learning rate decay after beyond-EoS oscillations drives the model very close to the flattest minima, as shown in Figure 3.5 and in Appendix 3.7.1.
- ✻ Empirical 2: for 3,4,5-layer MLPs on MNIST, larger learning rate drives to a flatter minima, as shown in Figure 3.6(b).

3.2 Related Works

Edge of stability. Cohen et al. [2020] observes a two-stage process in gradient descent, where

the first is loss curvature grows until the sharpness touches the bound $2/\eta$, and the second is the curvature hovers around the bound and training loss still decreases in a macro view regardless of local instability. Gilmer et al. [2021] reports similar observations in stochastic gradient descent and conducts comprehensive experiments of loss sharpness on learning rates, architecture choices and initialization. Lewkowycz et al. [2020] argues that gradient descent would "catapult" into a flatter region if loss landscape around initialization is too sharp.

Some concurrent works [Ahn et al. 2022; Ma et al. 2022; Arora et al. 2022; Damian et al. 2022b] are also theoretically investigating the edge of stability. Ahn et al. [2022] suggests that unstable convergence happens when the loss landscape of neural networks forms a local forward-invariant set near the minima due to some ingredients, such as tanh as the nonlinear activation. Ma et al. [2022] empirically observes a multi-scale structure of loss landscape and, with it as an assumption, shows that gradient descent with different learning rates may stay in different levels. Arora et al. [2022] shows the training provably enters the edge of stability with modified gradient descent or modified loss, and then its associated flow goes to flat regions. Under mild conditions, Damian et al. [2022b] proves that GD beyond EoS follows an optimization trajectory subjected to a sharpness constraint so that a flatter region is found. Note that our learning rate is strictly larger than that of Damian et al. [2022b] so that their proposed manifold does not exists in our settings, as discussed in Section 3.6.2.

Implicit regularization. Due to its theoretical closeness to gradient descent with a small learning rate, gradient flow is a common setting to study the training behavior of neural networks. Barrett and Dherin [2020] suggests that gradient descent is closer to gradient flow with an additional term regularizing the norm of gradients. Through analysing the numerical error of Euler's method, Elkabetz and Cohen [2021] provides theoretical guarantees of a small gap depending on the convexity along the training trajectory. Neither of them fits in the case of our interest, because it is hard to track the parametric gap when $\eta > 1/\lambda$. For instance, in a quadratic

function, the trajectory jumps between the two sides once $\eta > 1/\lambda$. Damian et al. [2021] shows that SGD with label noise is implicitly subjected to a regularizer penalizing sharp minimizers but the learning rate is constraint strictly below the edge of stability threshold.

Balancing effect. Du et al. [2018] proves that gradient flow automatically preserves the norm differences between different layers of a deep homogeneous network. [Ye and Du 2021] shows that gradient descent on matrix factorization with a constant small learning rate still enjoys the auto-balancing property. Also in matrix factorization, Wang et al. [2021] proves that gradient descent with a relatively large learning rate leads to a solution with a more balanced (perhaps not perfectly balanced) solution while the initialization can be in-balanced. In a similar spirit, we extend their finding to a larger learning rate, with which the perfect balance may be achieved in our setting. We estimate our learning rate is strictly larger than theirs [Wang et al. 2021], where they show GD with large learning rates converges to a flat region in the interpolation manifold while the flat region w.r.t. our larger learing rate does not exists so GD is forced to wander around the flattest minima. Note that the implication of balancing effect is to get close to a flatter solution in the global minimum manifold, which may help improve generalization in some common arguments in the community.

Learning a single neuron. Yehudai and Ohad [2020] studies necessary conditions on both the distribution and activation functions to guarantee a one-layer single student neuron aligning with the teacher neuron under gradient descent, SGD and gradient flow. Vardi et al. [2021] extends the investigation into a neuron with a bias term. Vardi and Shamir [2021] empirically studies the training dynamics of a two-layer single neuron, focusing on its implicit bias. In this work, we present a convergence analysis of a two-layer single-neuron ReLU network trained with population loss in a large learning rate beyond the edge of stability.

3.3 Preliminaries

We consider a differentiable objective function $f(\theta)$ with $\theta \in \mathbb{R}^d$, and the GD algorithm from (3.1).

Definition 1. *A differentiable function f is L-gradient Lipschitz if*

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leqslant L \|\theta_1 - \theta_2\|, \quad \forall\, \theta_1, \theta_2. \tag{3.2}$$

The above definition is equivalent to saying that the spectral norm of the Hessian is bounded by L, or the *local curvature* at each point is bounded by L. Then η needs to be bounded by $1/L$ in GD so that it is guaranteed to visit an approximate first-order stationary point [Nesterov 1998]. The perturbed GD requires $\eta = 1/L$ to visit an approximate second-order stationary point [Jin et al. 2021], and stochastic variants share similar assumptions [Ghadimi and Lan 2013; Jin et al. 2021].

However, in practice, such an assumption may be violated, or even impossible to satisfy when $\|\nabla^2 f\|$ is not uniformly bounded. Cohen et al. [2020] observes that, with learning rate η fixed, the largest eigenvalue λ_1 of the loss Hessian of a neural network is below $2/\eta$ at initialization, but it grows above the threshold along training. Such a phenomena is more obvious when the network is deeper or narrower. This reveals the non-smooth nature of the loss landscape of neural networks.

Furthermore, another observation from Cohen et al. [2020] is that once $\lambda_1 \geqslant 2/\eta$, the training loss stops the monotone decreasing. This is not surprising because GD would diverge in a quadratic function with such a large curvature. However, despite of local instability, the training loss still decreases in a longer range of steps, during which the local curvature stays around $2/\eta$. A further phenomena is that, when GD is at the edge of stability, if the learning rate suddenly changes to a smaller value $\eta_s < \eta$, then the local curvature quickly grows to $2/\eta_s$ — indicating

the ability to 'manipulate' the local curvature by adjusting the learning rate.

Besides the analysis of GD, the local curvature itself has also received a lot of attention. Due to the nature of over-parameterization in modern neural networks, the global minimizers of the objective f form a manifold of solutions. There have been active directions to understand the *implicit bias* of GD methods, namely where do they converge to in the manifold, and why some points in the manifold are more preferable than others. For the former question, it is believed that (stochastic) GD prefers flatter minima [Barrett and Dherin 2020; Smith et al. 2021; Damian et al. 2021; Ma and Ying 2021]. For the latter, flatter minima brings better generalization [Hochreiter and Schmidhuber 1997; Li et al. 2018a; Keskar et al. 2016; Ma and Ying 2021; Ding et al. 2022]. It would be meaningful if flatter minima could be obtained via GD with a large learning rate.

More specifically, it has been shown that the eigenvalues of the hessian of a deep homogeneous network could be manipulated to infinity via rescaling the weights of each layer [Elkabetz and Cohen 2021]. Fortunately, gradient flow preserves the difference of norms across layers along the training [Du et al. 2018]. As a result, a balanced initialization induces balanced convergence, while GD would break this balancing effect due to finite learning rate. However, recently it has been observed that GD with large learning rates enjoys a balancing effect [Wang et al. 2021], where it converges to a (not perfect) balanced result despite of imbalanced initialization.

Motivated by the connections of optimization, loss landscape and generalization, we would like to understand the training behavior of gradient descent with a large learning rate, from low-dimensional to representative models.

3.4 Stable oscillation on 1-D functions: fixed point of two-step update

In this section, we provide conditions of existence of fixed points of two-step GD on generic 1-D functions, which are on the third or higher derivatives at the local minima (Theorem 3.1

and Lemma 3.2). More specifically, in the regression setting, these local conditions allow many differentiable non-linear activation functions to the base model (Prop 1), and a composition rule is established to build complicated base models with simple base models (Prop 2).

Within the framework of Theorem 3.1, we identify a specific 1-D function to investigate more: we show the convergence to the fixed points (Theorem 3.3), along with its 2-D extension in Prop 3, serving as the foundation of nonlinear (Section 3.5) and high-dimensional (Section 3.6) cases. Empirical verification of all theorems are provided in Appendix B.2.

3.4.1 Existence of fixed points

Definition 2. *(Period-2 stable oscillation and fixed point of two-step update F_η^2.) Consider GD on a function f in domain Ω. Denote the update rule of GD as $F_\eta(x)$ for $x \in \Omega$ with learning rate η. A **period-2 stable oscillation** is $\exists\ x \in \Omega$ such that $F_\eta(F_\eta(x)) = x$ and x is not a minima of f. Equivalently speaking, $\exists\ x \in \Omega$ is a **fixed point** of the two-step update $F_\eta^2(\cdot) \triangleq F_\eta(F_\eta(\cdot))$.*

Remark 1. It is obvious that fixed points of F_η^2 exist in pairs by the nature of period-2 oscillation.

We initiate our analysis of existence of fixed point of F_η^2 in 1-D. Starting from a condition on general 1-D functions, we look into several specific 1-D functions to verify our arguments. Then, focusing on a function in the form of $f(x) = (x^2 - \mu)^2$, we present the convergence analysis as a foundation for the following discussions. Furthermore, to shed light on the multi-layer setting, we propose a balancing effect on a 2-D function to make a connection to the 1-D analysis.

General 1-D functions. Consider a 1-D function $f(x)$ with a learnable parameter $x \in \mathbb{R}$. The parameter updates following GD with the learning rate η as

$$x^{(t+1)} = F_\eta(x^{(t)}) := x^{(t)} - \eta f'(x^{(t)}). \tag{3.3}$$

Assuming f is differentiable and all derivatives are bounded, the function value in the next step can be approximated by

$$f(x^{(t+1)}) = f(x^{(t)}) - \eta[f'(x^{(t)})]^2\left(1 - \frac{\eta}{2}f''(x^{(t)})\right) + o((x^{(t+1)} - x^{(t)})^2).$$

If $\eta < 2/f''(x^{(t)})$, this approximation reveals that the function monotonically decreases for each step of GD, ignoring higher terms. Such an assumption would guarantee the convergence to a global minimum in a convex function. However, our interest is what happens if $\eta > 2/f''(x)$. For instance, if f is a quadratic function, the second-order derivative f'' is constant. As a result, once $\eta > 2/f''$, GD diverges except when being initialized at the optimum. However, when trained with a large learning rate $\eta > 2/f''(\bar{x})$, there is still some hope for a function to stay around a local minima $\bar{x}$, as stated in the following theorem.

Theorem 3.1. *Consider any 1-D differentiable function $f(x)$ around a local minima $\bar{x}$, satisfying (i) $f^{(3)}(\bar{x}) \neq 0$, and (ii) $3[f^{(3)}]^2 - f''f^{(4)} > 0$ at $\bar{x}$. Then, there exists ϵ with sufficiently small $|\epsilon|$ and $\epsilon \cdot f^{(3)} > 0$ such that: for any point x_0 between $\bar{x}$ and $\bar{x} - \epsilon$, there exists a learning rate η such that $F_\eta^2(x_0) = x_0$, and*

$$\frac{2}{f''(\bar{x})} < \eta < \frac{2}{f''(\bar{x}) - \epsilon \cdot f^{(3)}(\bar{x})}.$$

Remark 1. Here obviously we have $\eta > 2/f''(\bar{x})$ beyond EoS. If we take $f''(x_0) \approx f''(\bar{x}) - \epsilon' f^{(3)}(\bar{x})$ with $\epsilon' \approx \epsilon$, it holds $\eta < \frac{2}{f''(x_0)}$. Symmetrically, it holds $\frac{2}{f''(F_\eta(x_0))} < \frac{2}{f''(\bar{x})}$. Hence, η upper bounded by the EoS at one point in the period-2 orbit.

Remark 2. We prove the key condition, $3[f^{(3)}]^2 - f''f^{(4)} > 0$, in the case of **matrix factorization** around any minima as Theorem B.3 in Appendix B.10.1. Meanwhile, we verify this condition in **multi-layer networks on MNIST**, as shown in Figure 3.6, 3.7, 3.8 in Appendix 3.7.2.

The details of proof are presented in the Appendix B.3. As stated in the Theorem 3.1, we

provide a sufficient condition for existence of fixed point of F_η^2 around a local minima. But still we cannot tell whether or not some functions have it with $f^{(3)}(\bar{x}) = 0$. For instance, a quadratic function does not satisfy this condition since $f^{(3)} = f^{(4)} \equiv 0$ and it diverges when GD is beyond the edge of stability. But for $f(x) = \sin(x)$ around $\bar{x} = -\frac{\pi}{2}$ where $f^{(3)}(\bar{x}) = 0$, it turns out the fixed point exists. Therefore, we extend the argument in Theorem 3.1 to a higher order case in Lemma 3.2. As a result, we verify that the sine function does allow stable oscillation as in Corollary 1, because its lowest order of nonzero derivative (except f'') at the local minima is $f^{(4)}(\bar{x}) < 0$.

Lemma 3.2. *Consider any 1-D differentiable function $f(x)$ around a local minima $\bar{x}$, satisfying that the lowest order non-zero derivative (except the f'') at $\bar{x}$ is $f^{(k)}(\bar{x})$ with $k \geqslant 4$. Then, there exists ϵ with sufficiently small $|\epsilon|$ such that: for any point x_0 between $\bar{x}$ and $\bar{x} - \epsilon$, and*

1. *if k is odd and $\epsilon \cdot f^{(k)}(\bar{x}) > 0, f^{(k+1)}(\bar{x}) < 0$, then there exists $\eta \in (\frac{2}{f''}, \frac{2}{f'' - f^{(k)}\epsilon^{k-2}})$,*
2. *if k is even and $f^{(k)}(\bar{x}) < 0$, then there exists $\eta \in (\frac{2}{f''}, \frac{2}{f'' + f^{(k)}\epsilon^{k-2}})$,*

such that $F_\eta^2(x_0) = x_0$.

The details of proof are presented in the Appendix B.4.

L_2 LOSS ON GENERAL 1-D FUNCTIONS. However, we have to admit that the local conditions above are 1) too abstract to directly write down a meaningful function in this family, or 2) too complicated to compute the higher-order derivatives of a given non-trivial function.

Fortunately, both Theorem 3.1 and Lemma 3.2 provide a guarantee that squared-loss on any base function g provably allows stable oscillation once g satisfies some mild conditions, as stated in Prop 1. Moreover, we provide a straightforward method to build a more complicated model from two simple base models, as stated in Prop 2.

Proposition 1. *Consider a 1-D function $g(x)$, and define the loss function f as $f(x) = (g(x) - y)^2$. Assuming (i) g' is not zero when $g(\bar{x}) = y$, (ii) $g'(\bar{x})g^{(3)}(\bar{x}) < 6[g''(\bar{x})]^2$, then it satisfies the condition in Theorem 3.1 or Lemma 3.2 have a fixed point of F_η^2 around $\bar{x}$.*

This setup covers a broad family of generic non-linear least squares problems, including the base model g being **sine, tanh, high-order monomial, exponential, logarithm, sigmoid, softplus, gaussian**, etc. Many of these nonlinear (activation) functions are widely used in empirical or theoretical deep learning, together with the composition rule (Prop 2), shedding light for future analysis of practical models with these as building blocks.

Proposition 2 (Composition Rule)**.** *Consider two 1-D functions p, q. Assume both $p(x), q(y)$ at $x = \bar{x}, y = p(\bar{x})$ satisfies the conditions of g in Prop 1. Then $q(p(x))$ also satisfies the conditions to have a fixed point of F_η^2 around $x = \bar{x}$.*

In Appendix B.4 and B.5, we provide the proof details of these settings of $g(x)$ as Corollaries 1-8, along with all lemmas and proposition.

After the above discussions on local conditions, a natural question rises up as

Q1: with existence of a fixed point of F_η^2, can iterative runnings of F_η^2 converge to it?

With such a question, we are going to present a careful analysis on $g(x) = x^2$.

3.4.2 Convergence to fixed points

A special 1-D function. Consider $f(x) = \frac{1}{4}(x^2 - \mu)^2$ with $\mu > 0, f^{(3)}(\sqrt{\mu}) = 6\sqrt{\mu}, f''(\sqrt{\mu}) = 2\mu$. Note that this function is more special to us because it can be viewed as a *symmetric scalar factorization* problem subjected to the squared loss. Later we will leverage it to gain insights for asymmetric initialization, two-layer single-neuron networks and matrix factorization. Before that, we would like to show where it converges to when $\eta > \frac{2}{f''(\sqrt{\mu})}$ as follows.

Theorem 3.3. *For $f(x) = \frac{1}{4}(x^2 - \mu)^2$, consider GD with $\eta = K \cdot \frac{1}{\mu}$ where $1 < K < \sqrt{4.5} - 1 \approx 1.121$, and initialized on any point $0 < x_0 < \sqrt{\mu}$. Then it converges to an orbit of period 2, except for a measure-zero initialization where it converges to $\sqrt{\mu}$. More precisely, the period-2 orbit are the solutions $x = \delta_1 \in (0, \sqrt{\mu}), x = \delta_2 \in (\sqrt{\mu}, 2\sqrt{\mu})$ of solving δ in*

$$\eta = \frac{1}{\delta^2 \left(\sqrt{\frac{\mu}{\delta^2} - \frac{3}{4}} + \frac{1}{2} \right)}. \tag{3.4}$$

The details of proof are presented in the Appendix B.6. As shown above, Theorem 3.1 and Theorem 3.3 stand in two different levels: Theorem 3.1 restricts the discussion in a local view because of Taylor approximation, while Theorem 3.3 starts from local convergence and then generalizes it into a global view. However, Theorem 3.1 builds a foundation for Theorem 3.3 because the latter would degenerate to the former when K is extremely close to 1.

A SPECIAL 2-D FUNCTION. Similarly, consider a 2-D function $f(x, y) = \frac{1}{2}(xy - \mu)^2$ under different initialization for x and y, which we would call "in-balanced" initialization. Note that all the global minima in 2-D case form a manifold $\{(x, y) | xy = \mu\}$ while the 1-D case only has two points of global minima. So we need to distinguish all points in the manifold by their sharpness. When $xy = \mu$, the leading eigenvalue of the loss Hessian is $\lambda_1 = (x-y)^2 + 2\mu$. Hence, in the global minima manifold, the local curvature of each point is larger if its two parameters are more imbalanced. Among all these points, the smallest curvature appears to be $\lambda_1 = 2\mu$ when $x = y = \sqrt{\mu}$. In other words, if the learning rate $\eta > 2/2\mu$, all points in the manifold would be too sharp for GD to converge. We would like to investigate the behavior of GD in this case. It turns out the two parameters are driven to a perfect balance although they initialized differently, as follows.

Theorem 3.4. *For $f(x, y) = \frac{1}{2}(xy - \mu)^2$, consider GD with learning rate $\eta = K \cdot \frac{1}{\mu}$. Assume both x and y are always positive during the whole process $\{x_i, y_i\}_{i \geqslant 0}$. In this process, denote a series of all*

points with $xy > \mu$ *as* $\mathcal{P} = \{(x_i, y_i) | x_i y_i > \mu\}_{i \geqslant 0}$. *Then* $|x - y|$ *decays to 0 in* $\mathcal{P}$, *for any* $1 < K < 1.5$.

Theorem 3.4 shows an effect that the two parameters are squeezed to a single variable, which re-directs to our 1-D analysis in Theorem 3.3. Therefore, actually both cases converge to the same orbit when $1 < K < 1.121$, as stated in Prop 3.

Proposition 3. *Follow the setting in Theorem 3.4. Further assume* $1 < K < \sqrt{4.5} - 1 \approx 1.121$. *Then GD converges to an orbit of period 2. The orbit is formally written as* $\{(x = y = \delta_i) | i = 1, 2\}$, *with* $\delta_1 \in (0, \sqrt{\mu}), \delta_2 \in (\sqrt{\mu}, 2\sqrt{\mu})$ *as the solutions of solving* δ *in*

$$\eta = \frac{1}{\delta^2 \left(\sqrt{\frac{\mu}{\delta^2} - \frac{3}{4}} + \frac{1}{2} \right)}.$$

A natural follow-up question is what implications Theorem 3.3 and Prop 3 bring, because 1-D and 2-D is far from the practice of neural networks that contain multi-layer structures, nonlinearity and high dimensions. We precisely incorporate two layers and nonlinearity in Section 3.5, and high dimensions in Section 3.6.

3.5 On a two-layer single-neuron homogeneous network

We denote a two-layer single-neuron network as $f(x; \theta) = v \cdot \sigma(w^\top x)$ where $v \in \mathbb{R}, w \in \mathbb{R}^d$, the set of trained parameters $\theta = (v, w^\top) \in \mathbb{R}^{d+1}$, and the nonlinearity σ is ReLU. We will keep such an order in θ to view it as a vector. The input $x \in \mathbb{R}^d$ is drawn uniformly from a unit sphere $\mathcal{S}^{d-1}$. The parameters are trained by GD subjected to L_2 population loss, as

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t), \quad L(\theta_t) = \mathbb{E}_{x \in \mathcal{S}^{d-1}} \big(f(x; \theta_t) - y \big)^2.$$

We generate labels from a single teacher neuron function, as $y|x = \sigma(\tilde{w}^\top x)$. Hence $\tilde{w}$ is our target neuron to learn. We denote the angle between w and $\tilde{w}$ as $\alpha \geqslant 0$. Note that α is set as non-negative because the loss function is symmetric w.r.t. the angle. Moreover, the rotational symmetry of the population data distribution results in a loss landscape that only depends on w through the angle α and the norm $\|w\|$. Indeed, from the definition, we have

$$\nabla_\theta L = \frac{1}{d}\begin{bmatrix} v\,\|w\|_2^2 - \frac{\|w\|}{\pi}\big(\sin\alpha + (\pi-\alpha)\cos\alpha\big)\,\|\tilde{w}\| \\ v^2 w - \frac{v}{\pi}(\pi - \alpha + \frac{1}{2}\sin 2\alpha)\cdot\tilde{w} - \frac{v}{\pi}(-\frac{1}{2}\cos 2\alpha + \frac{1}{2})\,\|\tilde{w}\|\,\tilde{w}_\perp \end{bmatrix},$$

where we denote $\tilde{w}_\perp$ as the normalized of $w - \mathrm{proj}_{\tilde{w}}\, w$. Consider the Hessian

$$H \triangleq \begin{bmatrix} \partial_v^2 L & \partial_w \partial_v L \\ \partial_v \partial_w L & \partial_w^2 L \end{bmatrix} \overset{\text{if } vw = \tilde{w}}{=\!=} \frac{1}{d}\begin{bmatrix} \|w\|^2 & vw^\top \\ vw & v^2\mathbb{I} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}. \tag{3.5}$$

Hence, in the global minima manifold where $vw = \tilde{w}$, the eigenvalues of the Hessian are $\lambda_1 = \frac{\|w\|^2+v^2}{d}, \lambda_{2\ldots d} = \frac{v^2}{d}, \lambda_{d+1} = 0$. Therefore, the largest eigenvalue λ_1 measures the imbalance (*i.e.*, $|\,\|w\| - v|$) between the two layers again as $\lambda_1 = \frac{(\|w\|-v)^2+2\|\tilde{w}\|}{d}$ similar to the 2-D case in Section 3.4.2. So we would like to investigate where GD converges if $\eta > \frac{2}{2\|\tilde{w}\|/d} = d/\|\tilde{w}\|$ that is too large even for the flattest minima. Note that a key difference between the current case and the previous 2-D analysis is that the current one includes a neuron as a vector and a nonlinear ReLU unit.

From the second row of $\nabla_\theta L$, which is $\nabla_w L$, it is clear that updates of w always stay in the plane spanned by $\tilde{w}$ and $w^{(0)}$. Hence, this problem can be simplified to three variables (v, w_x, w_y) with the target neuron $\tilde{w} = [1, 0]$. The three variables stand for

$$v^{(t)} := v^{(t)}, \quad w_x^{(t)} := \mathrm{proj}_{\tilde{w}}\, w^{(t)},$$
$$w_y^{(t)} := \mathrm{proj}_{\tilde{w}_\perp}\, w^{(t)} = \sqrt{\left\|w^{(t)}\right\|^2 - (w_x^{(t)})^2}.$$

We keep w_y as nonnegative because the loss L is invariant to its sign and our previous notation $\alpha \geqslant 0$ requires a non-negative w_y. Then we show that w_y decays to 0 as follows.

Theorem 3.5. *In the above setting, consider a teacher neuron $\tilde{w} = [1, 0]$ and set the learning rate $\eta = Kd$ with $K \in (1, 1.1]$. Initialize the student as $\left\|w^{(0)}\right\| = v^{(0)} \triangleq \epsilon \in (0, 0.10]$ and $\langle w^{(0)}, \tilde{w} \rangle \geqslant 0$. Then, for $t \geqslant T_1 + 4$, $w_y^{(t)}$ decays as*

$$w_y^{(t)} < 0.1 \cdot (1 - 0.030K)^{t - T_1 - 4}, \quad T_1 \leqslant \left\lceil \log_{2.56} \frac{1.35}{\pi \beta^2} \right\rceil,$$
$$\beta = \left(1 + \frac{1.1}{\pi}\right) \epsilon.$$

The details of proof are presented in the Appendix B.9.

With the guarantee of w_y decaying in the above theorem, the dynamics of the single-neuron ReLU network follow the convergence of the 2-D case in Section 3.4.2, with a convergence result as follows.

Proposition 4. *The single-neuron model in Theorem 3.5 converges to a period-2 orbit where $w_y = 0$ and $(v, w_x) \in \gamma_K$ with $\gamma_K = \{(\delta_1, \delta_1), (\delta_2, \delta_2)\}$. Here $\delta_1 \in (0, 1), \delta_2 \in (1, 2)$ are the solutions δ in*

$$K = \frac{1}{\delta^2 \left(\sqrt{\frac{1}{\delta^2} - \frac{3}{4}} + \frac{1}{2}\right)}. \tag{3.6}$$

Remark 2. Actually this convergence is close to the flattest minima because: if the learning rate decays to infinitesimal after sufficient oscillations, then the trajectory walks towards the flattest minima ($v = w_x = 1, w_y = 0$). Note that we provide an experiment on **16-neuron networks** in Appendix 3.7.1, where GD converges to the period-2 orbit near the flattest minima while being initialized near unbalanced (sharp) minima.

To summarize, the single-neuron model goes through three phases of training dynamics, with an intialization of the angle $\measuredangle(w, \tilde{w})$ as $\frac{\pi}{2}$ at most. First, the angle decreases monotonically but, due to the growth of norms, the absolute deviation w_y still increases. Meanwhile, the imbalance $v - w_x$ stays in a bounded level. Second, w_y starts to decrease and the parameters fall into a basin within four steps. Third, in the basin, w_y decreases exponentially and, after w_y at a reasonable low level, the model approximately follows the dynamic of the 2-D case and the imbalance $v - w_x$ decreases as well, following Theorem 3.4. The model converges to a period-2 orbit as in the 1-D case in Theorem 3.3.

3.6 Matrix Factorization and beyond

In the last two sections, we have presented theoretical results that GD beyond EoS converges to the fixed points of F_η^2 from initialization that is far away. In this section, we address these follow-up questions, by raising observations in Matrix Factorization and discuss whether existing models can explain our observations or not:

Q2: does such a period-2 orbit exist in more complicated settings?

Q3: what does the appropriate model need to cover such oscillation in high-dim problems?

Q4: what will happen if the learning rate grows more?

3.6.1 Observations from Matrix Factorization

Consider a matrix factorization problem, parameterized by learnable weights $\mathbf{Y} \in \mathbb{R}^{d\times d}$, $\mathbf{Z} \in \mathbb{R}^{d\times d}$, and the target matrix is $\mathbf{C} \in \mathbb{R}^{d\times d}$, which is symmetric and positive definite. The loss L is defined as

$$L(\mathbf{Y}, \mathbf{Z}) = \frac{1}{2} \left\| \mathbf{Y}\mathbf{Z}^\top - \mathbf{C} \right\|_F^2 . \tag{3.7}$$

Obviously $\{(\mathbf{Y}, \mathbf{Z}) : \mathbf{Y}\mathbf{Z}^\top = \mathbf{C}\}$ forms a minimum manifold. Although we prove that the necessary 1-D condition holds around minimum as Theorem B.3 (in Appendix B.1.2), which is analogous to Theorem 3.1, it is more attracting to investigate GD in high dimensions. We propose our first observation that Matrix Factorization converges to a period-2 orbit, *i.e.*, fixed points of F_η^2, as follows.

Observation 1 (Matrix Factorization with period-2 orbit)**.** *Consider GD with learning rate η satisfying $\eta\sigma_1^2 \in (1, 1.121)$ and $\eta\left(\sigma_1^2 + \sigma_2^2\right) < 2$ where σ_1^2, σ_2^2 are the first and second largest eigenvalues of $\mathbf{C}$. Then, there exists non-measure-zero initialization, from which GD converges to a period-2 orbit in the form of ($i \in \{1, 2\}$)*

$$\mathbf{Y} = \rho_i u v^\top + \sum_{j=2}^{d} \sigma_{y,j} u_{y,j} v_{y,j}^\top,$$
$$\mathbf{Z} = \rho_i u v^\top + \sum_{j=2}^{d} \sigma_{z,j} u_{z,j} v_{z,j}^\top,$$
$$\mathbf{Y}\mathbf{Z}^\top - \mathbf{C} = (\rho_i^2 - \sigma_1^2) u u^\top,$$

where u is the leading eigenvector of $\mathbf{C}$, v is arbitrary unit vector in $\mathbb{R}^d$, $\{\rho_i\}_{i=1,2}$ are the two positive roots of

$$\eta\sigma_1^2 = \frac{1}{\rho^2\left(\sqrt{\frac{1}{\rho^2} - \frac{3}{4}} + \frac{1}{2}\right)}, \tag{3.8}$$

and the decompositions of $\mathbf{Y}, \mathbf{Z}$ are SVD.

Remark 3. At any minimizer $(\mathbf{X}, \mathbf{Y})$ satisfying $\mathbf{X}\mathbf{Y}^\top = \mathbf{C}$, the largest eigenvalue of loss Hessian w.r.t. parameters is $\sigma_{\max}(\mathbf{X})^2 + \sigma_{\max}(\mathbf{Y})^2$. Consequently, the flattest minima has sharpness as $2\sigma_1^2$, because $\sigma_1^2 = \lambda_{\max}(\mathbf{C}) \leqslant \sigma_{\max}(\mathbf{X})\sigma_{\max}(\mathbf{Y}) \leqslant 0.5\left(\sigma_{\max}(\mathbf{X})^2 + \sigma_{\max}(\mathbf{Y})^2\right)$.

To our knowledge, this observation is beyond all previous results. Damian et al. [2022b] tracks

the trajectory's projection onto the manifold with sharpness $< 2/\eta$. Wang et al. [2021] proposes that GD in a sharper region (sharpness$> 2/\eta$) converges to flatter region (sharpness$< 2/\eta$) for matrix factorization problem. But such a manifold (or flatter region) containing any minimizer does not exist in our setting because $\eta\sigma_1^2 > 1$ makes the flattest minima sharper than $2/\eta$, which means the probability of converging to a stationary point is zero [Ahn et al. 2022].

However, it is difficult to prove Observation 1 rigorously. Meanwhile, general initialization cannot illustrate well the phenomena that GD walks to flatter minima from a sharper one. Therefore, we provide an observation of a limited version of matrix factorization, called *quasi-symmetric*, along with sufficient intuition on its dynamics and careful discussion on what is remaining to prove it.

Definition 3 (Quasi-symmetric Matrix Factorization)**.** *Given a symmetric and positive definite target matrix* $\mathbf{C} \triangleq \mathbf{X}_0\mathbf{X}_0^\top$*, where* $\mathbf{X}_0 = \mathbb{R}^{d\times d}$*. Quasi-symmetric MF is solving the factorization problem with initialization near an unbalanced minima, where the minima is* $(\alpha\mathbf{X}_0, 1/\alpha\mathbf{X}_0)$ *with* $\alpha \neq 1$*.*

Observation 2 (Quasi-symmetric Matrix Factorization with period-2 orbit)**.** *Consider the above quasi-symmetric matrix factorization with learning rate* $\eta \in (1/\sigma_1^2, 1.121/\sigma_1^2)$*. Consider a minima* $(\mathbf{Y}_0 = \alpha\mathbf{X}_0, \mathbf{Z}_0 = 1/\alpha\mathbf{X}_0), \alpha > 0$*. The initialization is around the minimum, as* $\mathbf{Y}_1 = \mathbf{Y}_0 + \Delta\mathbf{Y}_1, \mathbf{Z}_1 = \mathbf{Z}_0 + \Delta\mathbf{Z}_1$*. When*

$$\eta \cdot \max\left\{(\frac{\sigma_1^2}{\alpha^2} + \sigma_2^2\alpha^2, \frac{\sigma_2^2}{\alpha^2} + \sigma_1^2\alpha^2)\right\} \leqslant 2 \tag{3.9}$$

GD would converge to a period-2 orbit γ_η *approximately with error in* $O(\epsilon)$*, formally written as,* $(i = 1, 2)$

$$(\mathbf{Y}_t, \mathbf{Z}_t) \to \gamma_\eta + (\Delta\mathbf{Y}, \Delta\mathbf{Z}), \qquad \|\Delta\mathbf{Y}\|, \|\Delta\mathbf{Z}\| = O(\epsilon),$$
$$\gamma_\eta = \{\left(\mathbf{Y}_0 + (\rho_i - \alpha)\,\sigma_1 u_1 v_1^\top, \mathbf{Z}_0 + (\rho_i - 1/\alpha)\,\sigma_1 u_1 v_1^\top\right)\},$$

where $\rho_1 \in (1,2), \rho_2 \in (0,1)$ *are the same as in Eq.(3.8)*

Remark 4. The intuition on the dynamics in Observation 2 is provided in Appendix B.10.2, along with a discussion on what is missing for rigorous proof for future development. Without loss of generality, assume $\mathbf{X}_0 = \mathrm{diag}([\sigma_1, \sigma_2, \ldots, \sigma_d]) \in \mathbb{R}^{d\times d}$, where $(\mathbf{X}_0)_{i,i} = \sigma_i$ and 0 in all other entries. Intuitively, the dynamics of the system is following

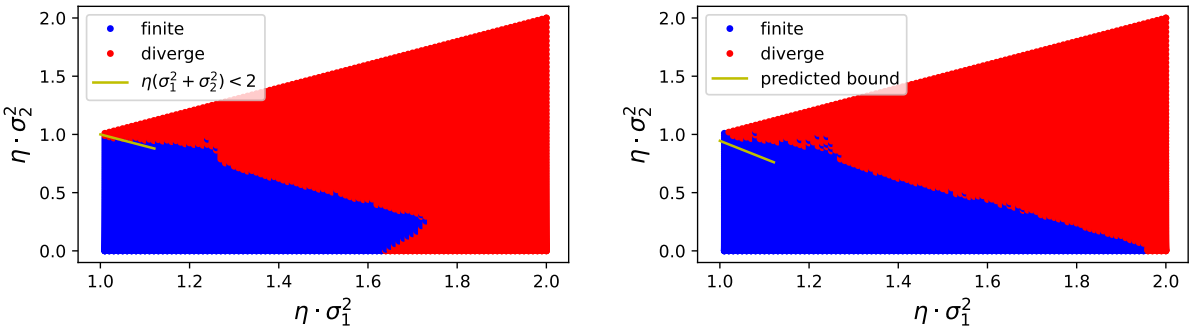
$$\mathrm{Y} = \left[\begin{array}{c|c} \alpha\sigma_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathrm{diag}([\alpha\sigma_i]_{i=2}^d) \end{array}\right] + O(\epsilon) \to \left[\begin{array}{c|c} \rho_i & \mathbf{0} \\ \hline \mathbf{0} & \mathrm{diag}([\alpha\sigma_i]_{i=2}^d) \end{array}\right] + O(\epsilon),$$

$$\mathbf{Z} = \left[\begin{array}{c|c} \alpha\sigma_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathrm{diag}([\sigma_i/\alpha]_{i=2}^d) \end{array}\right] + O(\epsilon) \to \left[\begin{array}{c|c} \rho_i & \mathbf{0} \\ \hline \mathbf{0} & \mathrm{diag}([\sigma_i/\alpha]_{i=2}^d) \end{array}\right] + O(\epsilon),$$

$$\mathrm{YZ}^\top = \left[\begin{array}{c|c} \sigma_1^2 & \mathbf{0} \\ \hline \mathbf{0} & \mathrm{diag}([\sigma_i^2]_{i=2}^d) \end{array}\right] + O(\epsilon) \to \left[\begin{array}{c|c} \rho_i^2 & \mathbf{0} \\ \hline \mathbf{0} & \mathrm{diag}([\sigma_i^2]_{i=2}^d) \end{array}\right].$$

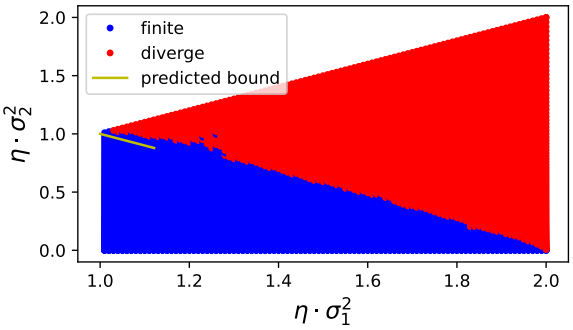
Note that the top singular values of $\mathbf{Y}, \mathbf{Z}$ are always the same in the orbit although it is unbalanced at initialization. A benefit of this is that, if η decays below $1/\sigma_1^2$ after reaching the orbit, it would converge to $\mathbf{Y}, \mathbf{Z}$ with same top singular value σ_1, satisfying $\mathbf{YZ}^\top = \mathbf{C}$.

How tight are Observation 1 and 2? There are two aspects we would like to address: $\eta\sigma_1^2$ and $\eta\sigma_2^2$. The former $\eta\sigma_1^2$ is a natural constraint because it is necessary to carefully set its upper bound in 1-D analysis to contain the oscillation in some finite level set. However, the second $\eta\sigma_2^2$ is novel (and tight) to our knowledge, which is respectively $\eta(\sigma_1^2 + \sigma_2^2) < 2$ in Observation 1 and $\eta \cdot \left(\sigma_1^2/\alpha^2 + \sigma_2^2\alpha^2\right) < 2$ in Observation 2. The tightness of this bound is verified in Figure 3.2, where it approximates the linearity of the empirical boundary between infinite and finite well when $\eta\sigma_1^2 > 1$ slightly. Furthermore, although we do not prefer asserting too much beyond our theorems, the linear trend between $\eta\sigma_1^2$ and $\eta\sigma_2^2$ keeps well when $\eta\sigma_1^2$ goes beyond 1.121 for a long range. Intuitively, We gain the insight of this bound from the analysis of Observation 2 in Appendix B.10.2. More precisely, it appears in Eq.(B.184) to guarantee a transition matrix to be semi-convergent, whose largest absolute eigenvalue is no larger than 1.



(a) Generic init

(b) Quasi-sym init ($\alpha = 0.9$)



(c) Symmetric init ($\alpha = 1$ in Quasi case)

Figure 3.2: Matrix Factorization: grid search of $\eta\sigma_1^2$ v.s. $\eta\sigma_2^2$ on whether GD diverges or not. (a) Generic initialization: it verifies the condition $\eta\left(\sigma_1^2+\sigma_2^2\right)<2$. (b-c) Quasi-symmetric initialization: it verifies the predicted bound $\eta\cdot\left(\sigma_1^2/\alpha^2+\sigma_2^2\alpha^2\right)<2$ in Eq.(3.9) as a sufficient condition.

Is there any other phenomena beyond period-2 orbit when η grows larger? The answer is yes. We conduct experiments of matrix factorization with generic initialization with different η's, as shown in Figure 3.3. It turns out when $\eta\sigma_1^2 \in (1, 1.23)$, it converges to period-2 orbit. When $\eta\sigma_1^2 \in (1.23, 1.28)$, it converges to a period-4 orbit, although the period-2 orbit still exists once $\eta\sigma_1^2 < 1.5$ as shown in Eq.(B.17) (because the existence cannot guarantee convergence, and even local convergence does not hold). When $\eta\sigma_1^2 > 1.28$, it is rather chaotic. However, during most of these, the balancing effect holds, *i.e.*, $\sigma_{\max}(\mathbf{Y}) = \sigma_{\max}(\mathbf{Z})$.

3.6.2 Implications for more complicated settings

Existing models from Ma et al. [2022] and Damian et al. [2022b]. Ma et al. [2022] proposes

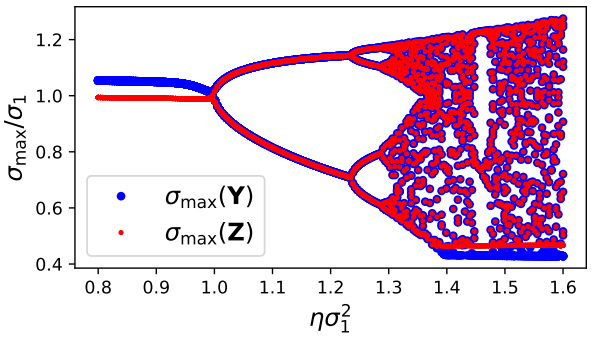


Figure 3.3: Matrix Factorization: $\sigma_{\max}(\mathrm{Y})$, $\sigma_{\max}(\mathbf{Z})$ for different η's. For each η, the last 10 iterations are sampled for report, due to periodic and chaotic phenomenon. **Observations**: (1) when $\eta\sigma_1^2 \in (1, 1.38)$, all cases have $\sigma_{\max}(\mathrm{Y}) = \sigma_{\max}(\mathbf{Z})$; (2) when $\eta\sigma_1^2 \in (1, 1.23)$, it converges to a period-2 orbit; (3) when $\eta\sigma_1^2 \in (1.23, 1.28)$, it converges to a period-4 orbit; (4) when $\eta\sigma_1^2 > 1.28$, it is rather chaotic; (5) when $\eta\sigma_1^2 < 1$, there is no oscillation.

a decomposition of high-dimensional functions into separable functions in eigendirections, in the form of

$$f(\theta) = f_1(p_1^\top \theta) + f_2(p_2^\top \theta) + \cdots + f_d(p_d^\top \theta), \tag{3.10}$$

where $\{p_i \in \mathbb{R}^d\}$ is an orthogonal basis of $\mathbb{R}^d$, $\theta \in \mathbb{R}^d$ is the parameter and each f_i is a function that allows stable oscillation. Within such a framework, all $p_i^\top x$ can stably oscillation since the dynamics is separable in each eigendirection. However, this framework cannot explain the dynamics of matrix factorization, because our experiments in Figure 3.2 have shown that GD will blow up once $\eta\sigma_2 > 1$, which means the eigen-directions associated with σ_1^2 and σ_2^2 cannot be disentangled in this case.

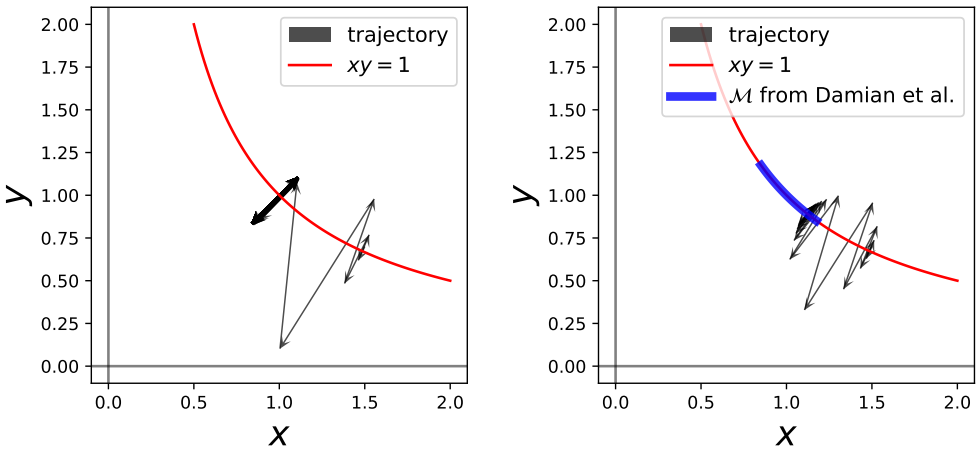
Damian et al. [2022b] proposes to track the trajectory's projection onto manifold $\mathcal{M} = \{\theta : \lambda(\theta) < 2/\eta, \nabla L(\theta) \cdot u(\theta) = 0\}$, where $\lambda(\theta)$ and $u(\theta)$ are the leading eigenvalue and eigenvector of Hessian of loss L. However, such a manifold does not exist in the 2-D case we have studied in Section 3.4.1 because our setting is strictly beyond EoS. Furthermore, in high-order cases, such a manifold containing any minimizer does not exist (Proposition 7).

Proposition 5. *For* $L(x, y) = 1/2(xy - 1)^2$ *with* $\eta > 1$ *on* $\{x > 0, y > 0\}$, *such a manifold* $\mathcal{M}$ *does not exist.*

Proposition 6. *For* $L(x, y) = 1/2(xy - 1)^2$ *with* $\eta < 1$ *on* $\{x > 0, y > 0\}$, $\mathcal{M} = \{(x, y) : xy = 1, x + y < \sqrt{2 + 2/\eta}\}$.

Proposition 7. *For* $L(\{x_i\}_{i=1}^n) = \frac{1}{n}(\prod_{i=1}^n x_i - 1)^2$ *with* $\eta > 1$ *on* $\{x_i > 0,\ \forall i\}$, *such a manifold* $\mathcal{M}$ *containing any minimizer does not exist.*

Moreover, although $\mathcal{M}$ exists when $\eta < 1$ (Proposition 6), the size of $\mathcal{M}$ is limited, which means the trajectory's projection onto it stays unchanged in the early steps, although the trajectory is moving efficiently from sharper region to flatter region, as shown in Figure 3.4(b).



(a) $\eta = 1.08$ (b) $\eta = 0.95$

Figure 3.4: Trajectories of minimizing $L(x, y) = 1/2(xy - 1)^2$ with $\eta = 1.08, 0.95$. For $\eta = 1.08$, the manifold $\mathcal{M}$ proposed by Damian et al. [2022b] does not exist. For $\eta = 0.95$, the manifold $\mathcal{M}$ exists, but the projection onto it does not change for the first few steps.

Two candidate models. From the above discussion, we would like to raise two candidate models to contain the observations from matrix factorization, based on the models proposed in Damian et al. [2022b] and Ma et al. [2022].

Following Damian et al. [2022b], we would like to propose

Definition 4 (Projection onto manifold). $\mathcal{M}_c = \{\theta : \lambda_2(\theta) < 2/\eta, \nabla L(\theta) \cdot u(\theta) = 0\}$, *where* $\lambda_2(\theta)$ *and* $u(\theta)$ *are the second largest eigenvalue and the leading eigenvector of Hessian of loss* L.

The motivation of $\mathcal{M}_c$ is to contain points that have the leading eigenvalue greater than 1. For example, in the case of $1/2(xy-1)^2$, it is $\mathcal{M}_c = \{(x,y) : xy = 1\}$ allowing to track the trajectory walking from sharper region to flatter region. Instead of constraining $\lambda < 2/\eta$, we set $\lambda_2 < 2/\eta$ to make it compatible with our observations in matrix factorization.

The gap between Ma et al. [2022] and observations from matrix factorization is that they assume the orthogonal decomposition of the loss function. However, even in the simplest setting of matrix factorization, this assumption does not hold. Taking a symmetric matrix factorization as an example, we have

$$\begin{aligned} L(\mathbf{X}) &= \frac{1}{4}\left\| \mathbf{X}\mathbf{X}^\top - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\|_F^2 \\ &= \frac{1}{4}\left(\left(\left\|\mathbf{X}_{0,:}\right\|^2 - 1 \right)^2 + \left(\left\|\mathbf{X}_{1,:}\right\|^2 - 1 \right)^2 + 2\left(\langle \mathbf{X}_{0,:}, \mathbf{X}_{1,:} \rangle \right)^2 \right), \end{aligned} \tag{3.11}$$

where the first two terms in the last line are $f_i(p_i^\top x)$ in Eq.(3.10) since the included are orthogonal to each other. However, the last term $\langle \mathbf{X}_{0,:}, \mathbf{X}_{1,:} \rangle$ breaks the separability in the decomposition. Meanwhile, $\left(\langle \mathbf{X}_{0,:}, \mathbf{X}_{1,:} \rangle \right)^2$ is implicitly $\left(\langle \mathbf{X}_{0,:}, \mathbf{X}_{1,:} \rangle - 0 \right)^2$, because $\mathbf{X}_{0,:}, \mathbf{X}_{1,:} \in \mathbb{R}^2$ are expressive enough to form an orthogonal pair satisfying the constraints of norms.

In a similar spirit, we propose an extensive model of Eq.(3.10) [Ma et al. 2022] as follows

$$f(\theta) = \sum_{i=1}^{d} g_i(p_i^\top \theta; a_i) + \sum_{(i,j)\in\mathcal{E}}^{d} h_{ij}(p_i^\top \theta, p_j^\top \theta; b_{ij}), \tag{3.12}$$

where g_i, h_{ij} are functions allowing stable oscillation parameterized by a_i, b_{ij}, $\{p_i\}_{i=1}^d$ are orthogonal basis of $\mathbb{R}^d$ and $\mathcal{E} \subset [d] \times [d]$ is a selected subset of tuples. A simple but effective example to imitate matrix factorization is $g_i(x;a) \triangleq (\|x\|^2 - a)^2$ and $h_{ij}(x,y|b) \triangleq (\langle x, y \rangle - b)^2$ and

$\mathcal{E} = [d] \times [d]$. Intuitively, such a model with larger $|\mathcal{E}|$ allows fewer eigenvalues of Hessian to go beyond $2/\eta$. Conversely, if $\mathcal{E} = \emptyset$, it allows all eigenvalues beyond $2/\eta$, which degenerates to Eq.(3.10) [Ma et al. 2022].

3.7 Experiments on MLPs and MNIST

In this section, we perform two experiments in relatively higher dimension settings. We are to show two observations that coincide with our discussions in the low dimension:

Observation 3. *GD beyond EoS drives to flatter minima.*

Observation 4. *GD beyond EoS is in a similar style with the low dimension.*

3.7.1 2-layer high-dim homogeneous ReLU NNs with planted teacher neurons

We conduct a synthetic experiment in the high-dimension teacher-student framework.

The teacher network is in the form of

$$y|x := f_{\text{teacher}}(x; \tilde{\theta}) = \sum_{i=1}^{16} \mathsf{ReLU}(\mathbf{e}_i^\top x), \tag{3.13}$$

where $x \in \mathbb{R}^{16}$ and $\mathbf{e_i}$ is the i-th vector in the standard basis of $\mathbb{R}^{16}$. The student and the loss are in forms of

$$f(x; \theta) = \sum_{i=1}^{16} v_i \cdot \mathsf{ReLU}(w_i^\top x), \tag{3.14}$$

$$L(\theta; \tilde{\theta}) = \frac{1}{m} \sum_{i}^{16} \left(f(x; \theta) - y|x_i\right)^2. \tag{3.15}$$

Apparently, the global minimum manifold contains the following set $\mathcal{M}$ as (w.l.o.g., ignoring any permutation)

$$\mathcal{M} = \{(v_i, w_i)_{i=1}^{16} \mid \forall i \in [16], w_i = k_i \cdot \mathbf{e}_i, v_i = \frac{1}{k_i}, k_i > 0\}. \tag{3.16}$$

However, different choices of $\{k_i\}_{i=1}^{16}$ induce different extents of sharpness around each minima. Our aim is to show that **GD with a large learning rate beyond the edge of stability drives to the flattest minima from sharper minima**.

Initialization. We initialize all student neurons directionally aligned with the teachers as $w_i \parallel \mathbf{e}_i$ but choose various k_i, as $k_i = 1 + 0.0625(i-1)$. Obviously, such a choice of $\{k_i\}_{i=1}^{16}$ is not at the flattest minima, due to the isotropy of teacher neurons. Also we add small noise to w_i to make the training start closely (but not exactly) from a sharp minima, as

$$w_i = k_i \cdot (\mathbf{e}_i + 0.01\epsilon), \quad \epsilon \sim \mathcal{N}(0, I). \tag{3.17}$$

Data. We uniformly sample 10000 data points from the unit sphere $\mathcal{S}^{15}$.

Training. We run gradient descent with two learning rates $\eta_1 = 0.5, \eta_2 = 2.6$. Later we will show with experiments that the EoS threshold of learning rate is around 2.5, so η_2 is beyond the edge of stability. GD with these two learning rates starts from the same initialization for 100 epochs. Then we extend another 20 epochs with learning rate decay to 0.5 from 2.6 for the learning-rate case.

Results. All results are provided in Figure 3.5. Both Figure 3.5 (a, b) present the gap between these two trajectories, where GD with a small learning rate stays around the sharp minima, while that with a larger one drives to flatter minima. Then GD stably oscillates around the flatter minima.

Meanwhile, from Figure 3.5 (b), when we decrease the learning rate from 2.6 to 0.5 after 100 epochs, GD converges to a nearby minima which is significantly flatter, compared with that of lr=0.5.

Figure 3.5 (c) provides a more detailed view of $\frac{\|w_i\|}{v_i}$ for all 16 neurons. All neurons with lr=0.5 stay at the original ratio k_i^2. But those with lr=2.6 all converge to the same ratio around $k^2 = \frac{\|w\|}{v} = 1.21$, as shown in Figure 3.5 (d). We compute the relationship between the sharpness of global minima in $\mathcal{M}$ and different choices of k, as shown in Figure 3.5 (e, f). Actually, $k^2 = 1.21$ is the best choice of $\{k_i\}_{i=1}^{16}$ such that the minima is the flattest.

Therefore, we have shown that, **in such a setting of high-dimension teacher-student network, GD beyond the edge of stability drives to the flattest minima**.

3.7.2 3, 4, 5-layer non-homogeneous MLPs on MNIST

We conduct an experiment on real data to show that **our finding in the low-dimension setting in Theorem 3.1 is possible to generalize to high-dimensional setting**. More precisely, our goals are to show, when GD is beyond EoS,

1. the oscillation direction (gradient) aligns with the top eigenvector of Hessian.
2. the 1D function at the cross-section of oscillation direction and high-dim loss landscape satisfies the conditions in Theorem 3.1.

Network, dataset and training. We run 3, 4, 5-layer ReLU MLPs on MNIST [LeCun et al. 1998]. The networks have 16 neurons in each layer. To make it easier to compute high-order derivatives, we simplify the dataset by 1) only using 2000 images from class 0 and 1, and 2) only using significant input channels where the standard deviation over the dataset is at least 110, which makes the network input dimension as 79. We train the networks using MSE loss subjected to GD with large learning rates $\eta = 0.5, 0.4, 0.35$ and a small rate $\eta = 0.1$ (for 3-layer).

Note that the larger ones are beyond EoS.

Definition 5 (line search minima)**.** *Consider a function f. Consider learning rate η and a point $x \in domain(f)$. We call $\tilde{x}$ as the line search minima of x if*

$$\tilde{x} = x - c^* \cdot \eta \nabla f(x), \tag{3.18}$$

$$c^* = argmin_{c \in [0,1]} f\left(x - c \cdot \eta \nabla f(x)\right). \tag{3.19}$$

The line search minima $\tilde{x}$ can interpreted as the lowest point on the 1D function induced by the gradient at x. If GD is beyond EoS, $\tilde{x}$ stays in the valley below the oscillation of x.

Results. All results are presented in Figure 3.6, 3.7 and 3.8.

Take the 3-layer as an example. From Figure 3.6 (a, b), GD is beyond EoS during epochs 10-14 and 21-60. For these epochs, cosine similarity between the top Hessian eigenvector v_1 and the gradient is pretty close to 1, as shown in Figure 3.6 (c), which verifies our goal 1.

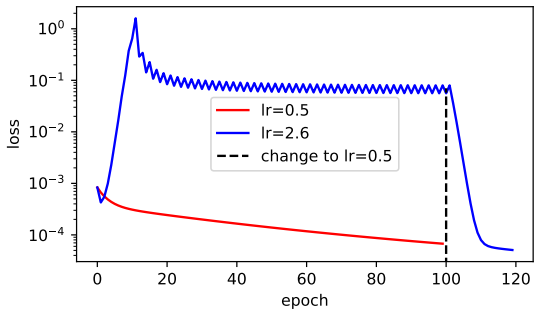
In Figure 3.6 (d), we compute $3[f^{(3)}]^2 - f^{(2)}f^{(4)}$ at line search minima along training, which is required to be positive in Theorem 3.1 to allow stable oscillation. Then it turns out most points have $3[f^{(3)}]^2 - f^{(2)}f^{(4)} > 0$ except a few points, all of which are not in the EoS regime, and these few exceptional points might be due to approximation error to compute the fourth-order derivative since their negativity is quite small. This verifies our goal 2.

Both the above arguments are the same in the cases of 4 and 5 layers as shown in Figure 3.7 and 3.8.

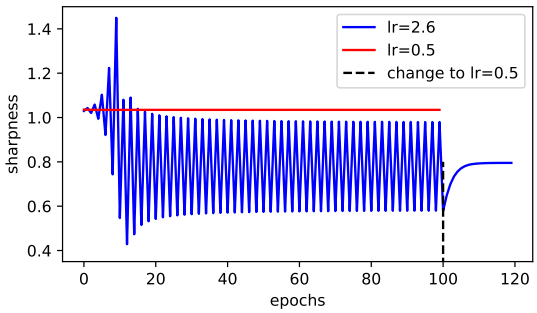
3.8 Conclusions

In this work, we investigate gradient descent with a large step size that crosses the threshold of local stability, via investigating convergence of two-step updates instead of convergence of one-step updates. In the low dimensional setting, we provide conditions on high-order derivatives

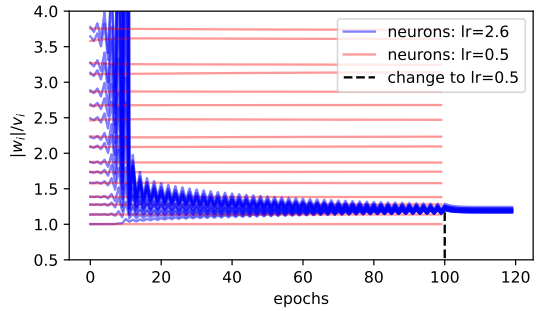
that guarantees the existence of fixed points of two-step updates. For a two-layer single-neuron ReLU network, we prove its convergence to align with the teacher neuron under population loss. For matrix factorization, we prove that the necessary 1-D condition holds around any minima. We provide novel observations of its convergence to period-2 orbit with comprehensive theoretical intuition of the dynamics. Finally, we extend previous works by proposing two models with observations in matrix factorization compatible for future analysis.



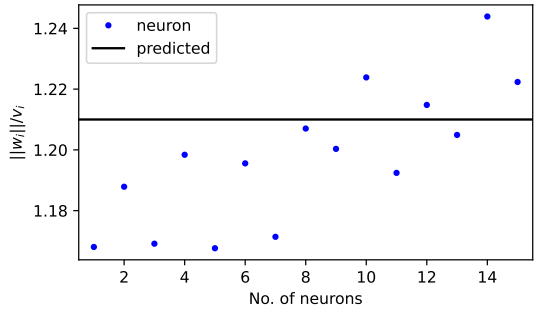
(a) Training loss



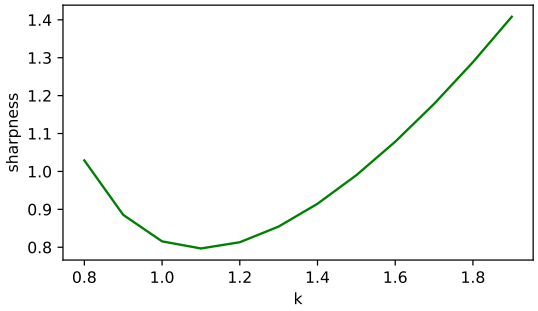
(b) Sharpness



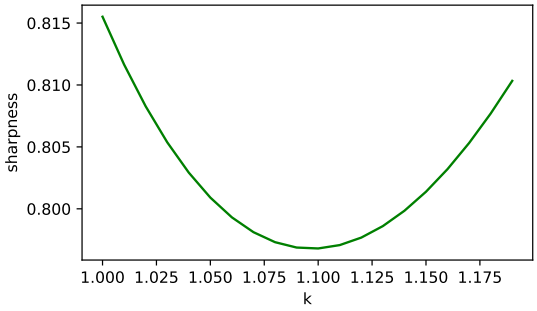
(c) ratio of $\|w_i\|$ and v_i during training



(d) final ratio of $\|w_i\|$ and v_i when lr=2.6

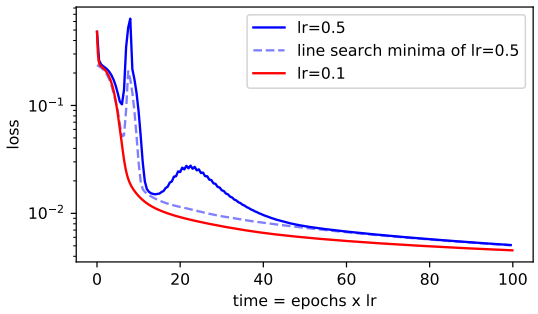


(e) sharpness for different ratio $\sqrt{\frac{\|w_i\|}{v_i}}$

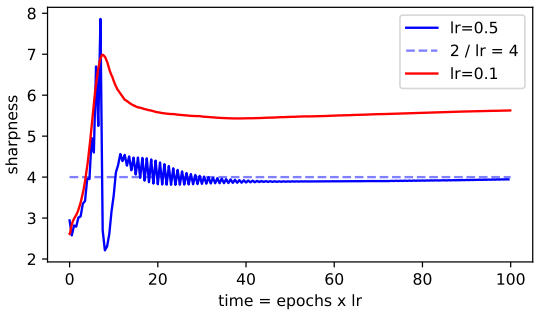


(f) sharpness for different ratio $\sqrt{\frac{\|w_i\|}{v_i}}$ (zoom-in)

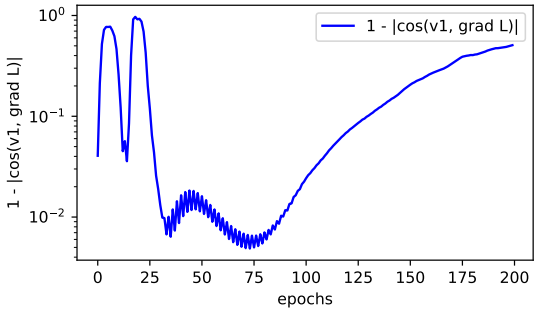
Figure 3.5: Result of 2-layer 16-neuron teacher-student experiment.



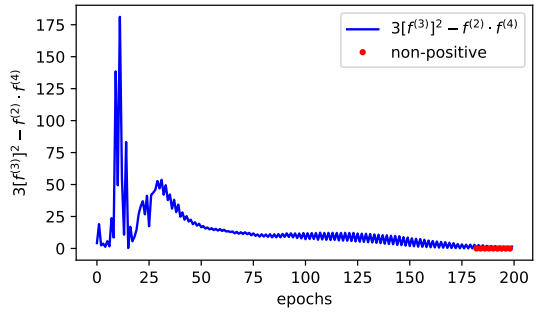
(a) Training loss



(b) Sharpness

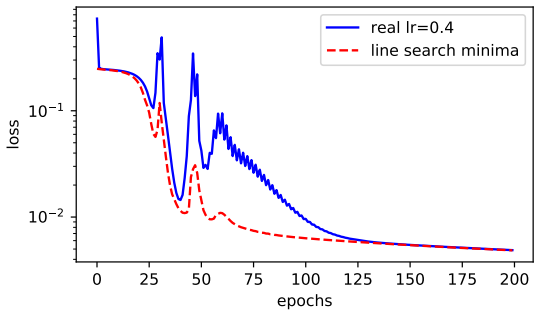


(c) similarity of gradient and top eig-vector v_1

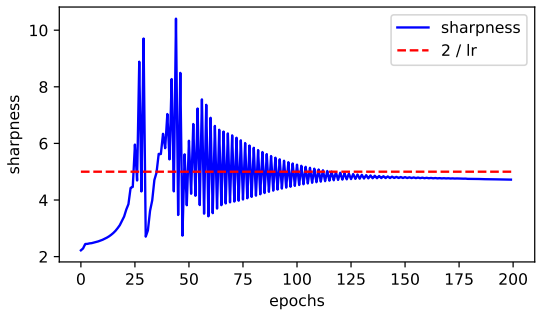


(d) $3[f^{(3)}]^2 - f^{(2)}f^{(4)}$ at line search minima

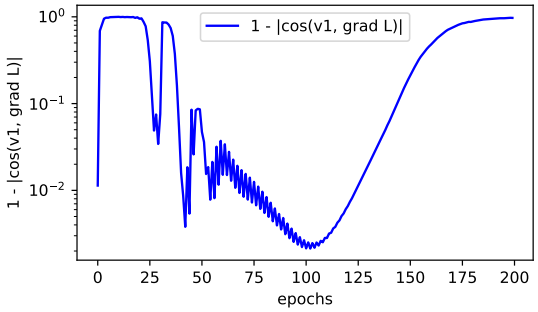
Figure 3.6: Result of **3-layer** ReLU MLPs on MNIST. Both (c) and (d) are for learning rate as 0.5.



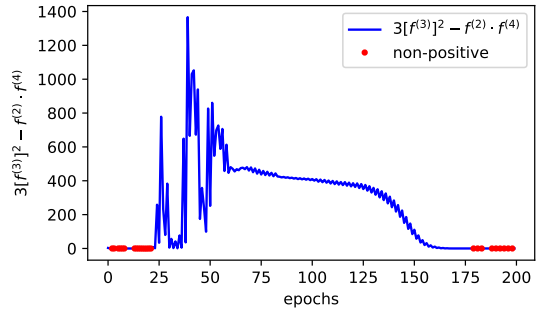
(a) Training loss



(b) Sharpness

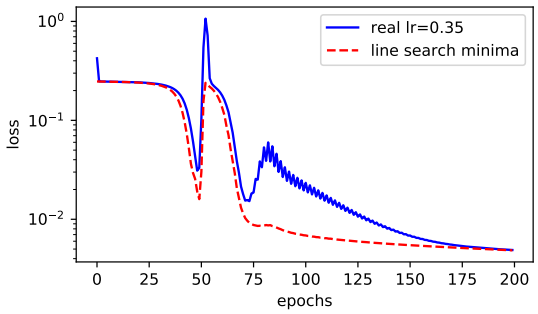


(c) similarity of gradient and top eig-vector v_1

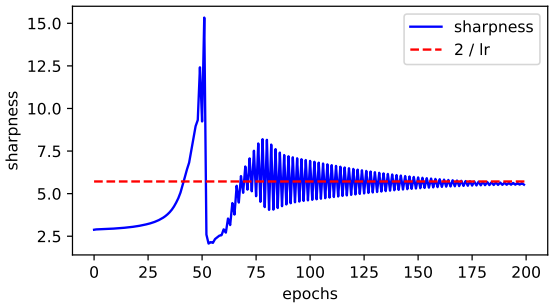


(d) $3[f^{(3)}]^2 - f^{(2)}f^{(4)}$ at line search minima

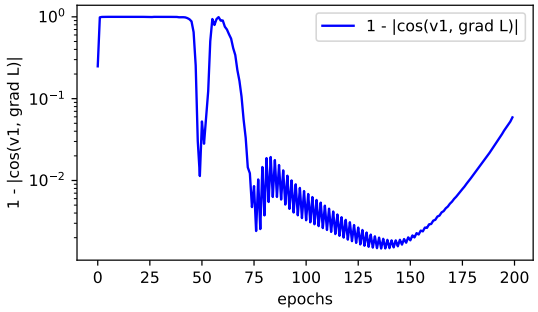
Figure 3.7: Result of **4-layer** ReLU MLPs on MNIST.



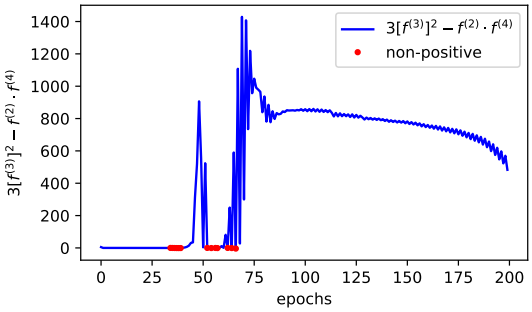
(a) Training loss



(b) Sharpness



(c) similarity of gradient and top eig-vector v_1



(d) $3[f^{(3)}]^2 - f^{(2)}f^{(4)}$ at line search minima

Figure 3.8: Result of **5-layer** ReLU MLPs on MNIST.

4 Memorization of Training Distribution in Transformer Modules

4.1 Introduction and Our Contributions

Large language models (LLMs) have shown impressive capabilities on a variety of tasks, from generating coherent and grammatically correct text, to language understanding and basic mathematical reasoning [Brown et al. 2020; Touvron et al. 2023]. At the heart of this success is the Transformer architecture [Vaswani et al. 2017], which relies on a sequence of self-attention and feed-forward layers to efficiently combine information from the input context and patterns learned from training data. Despite recent progress on interpreting the mechanisms learned by different layers [Meng et al. 2022; Wang et al. 2022], these models remain largely black boxes. A better understanding of the role of Transformer layers and how they are affected by the training process could enable new monitoring and editing techniques, better training data, and ultimately more reliable LLMs.

The task of next-token prediction in language modeling inherently involves different subtasks that may be at odds with each other, as shown in Figure 4.1. For instance, given the context "John gave a book to", the word "the" is a natural and grammatically correct next word to predict,

and relying on global bigram statistics might be enough to predict it given the last word "to". Nonetheless, if another character is present in the context, say Mary, then the name "Mary" may be a better prediction, and this would require a more involved form of reasoning over the context to retrieve this name. In the context of Transformer language models, previous work on interpretability has found that circuits of attention heads seem responsible for such in-context predictions [Wang et al. 2022], while feed-forward layers may be storing more general statistics such as the bigram "to the" or factual knowledge [Geva et al. 2021; Meng et al. 2022; Bietti et al. 2023]. To further strengthen this observation, the recent work [Sharma et al. 2023] found that selectively replacing certain layer weights to their low-rank approximation, particularly late feed-forward layers, may improve performance on various reasoning benchmarks, and observed that the truncated components were often responsible for predicting "generic" tokens such as the word "the".

In this chapter, we provide a finer understanding of these phenomena by studying how such mechanisms arise during training, in particular how simple *distributional associations*, such as the bigram "to the", tend to be localized in feed-forward layers, while attention focuses on in-context reasoning. We first provide a fine-grained study of training dynamics on a synthetic task with two-layer transformers exhibiting similar properties, where the task is in-context recall [Bietti et al. 2023] with additional noise on in-context tokens consisting of a fixed "generic" token:

- In a two-layer model with feed-forward layers (FF), we show that the generic noise token is mainly learned in FF and the attention attends towards correct in-context targets. Removing the feed-forward layers then leads to clean in-context predictions. We provide some theoretical justification through early training steps.
- In a model without FF, we show that the generic noise can be identified in a rank-one subspace of the value matrix in attention block. When the noise level is small, low-rank truncation can filter it out and predict clean outputs.

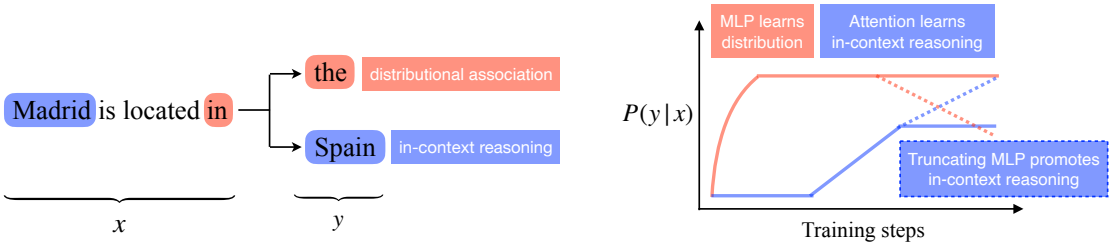


Figure 4.1: Distributional association *v.s.* in-context reasoning. In this work, we decompose tasks of next-token prediction into the distributional and the in-context ones, finding that MLPs learn distributional associations before attention develops in-context reasoning capabilities. Furthermore, truncating MLPs promotes in-context reasoning by weakening distributional associations. See Figure 4.13 for an example of this on the Pythia model [Biderman et al. 2023].

We then investigate such a separation between distributional association and in-context reasoning on pre-trained language models, namely the Pythia family, which has checkpoints available at different training steps [Biderman et al. 2023]. Overall, we provide a useful description of how distributional associations and in-context reasoning mechanisms are learned during training, and tend to be disentangled in different parts of the model, such that selectively removing certain components may lead to better predictions in reasoning tasks.

4.2 Related Works

[Sharma et al. 2023] recently empirically observed that a low-rank approximation of some weights in some pre-trained LLMs can improve reasoning capabilities. Several interpretability works have looked at the role of attention versus feed-forward layers for different tasks. The prominence of feed-forward/MLP layers for storing "global" or "persistent" associations or facts has been observed in [Sukhbaatar et al. 2019; Geva et al. 2021; Meng et al. 2022; Geva et al. 2023]. In contrast, several works have investigated the role of attention heads for "reasoning" or computation over the context, *e.g.*, for simple copying mechanisms with so-called induction heads [Elhage et al. 2021; Olsson et al. 2022; Bietti et al. 2023], or for more complex tasks [Merrill et al. 2022; Wang et al. 2022; Zhang et al. 2022; Liu et al. 2023; Sanford et al. 2024b].

Training dynamics of transformers and attention have been studied in various works [Snell et al. 2021; Jelassi et al. 2022; Li et al. 2023; Oymak et al. 2023; Tian et al. 2023; Bietti et al. 2023; Reddy 2024; Tian et al. 2024; Zhang et al. 2024; Nichani et al. 2024; Edelman et al. 2024]. In particular, the two-layer model and copy task we consider are similar to Bietti et al. [2023], yet their data model does not involve noise on in-context predictions, and they do not study learning of global associations. Chan et al. [2022]; Reddy [2024] study in-context vs. in-weights learning empirically, on different tasks than ours. Cabannes et al. [2024] study training dynamics of linear associative memories, but focuses on deterministic data while our setup has generic noise. Training dynamics were also studied empirically for interpretability [Olsson et al. 2022; Nanda et al. 2023; Quirke et al. 2023; Chen et al. 2024]. Edelman et al. [2022]; Bai et al. [2023]; Abernethy et al. [2024] studied sample complexity of self-attention and in-context learning, but did not consider training dynamics.

4.3 Preliminaries

In this section, we provide some background and motivation on reasoning tasks, and describe the weight truncation technique which we use for ablating weights.

4.3.1 Reasoning from Context

Recent LLMs have shown promising results in more complex "reasoning" tasks which may involve multiple steps of logical or computational processing from context or prompt [Srivastava et al. 2022; Wei et al. 2022; Bubeck et al. 2023; Dziri et al. 2024], as opposed to simple pattern matching or memorization of training data, for instance using learned n-gram predictions.

While it is difficult to clearly separate reasoning from memorization, in this work we will make the simplifying distinction that **in-context reasoning** involves dependencies between *multiple tokens* potentially far away in the context, while we consider **distributional associations** as

simpler predictions that only depend on the *last token*, e.g., through a bigram model. Thus, due to the residual structure of Transformers, reasoning will typically require using attention operations in Transformers over context, while feed-forward layers should suffice for learning distributional associations. We note that our assumption of distributional associations depending only on the last token is mainly for convenience of our analysis, and could be extended to depend on the last token's *residual stream* [Elhage et al. 2021], which may contain additional information from the context. For instance, this could include previous tokens thanks to position-based attention heads [Voita et al. 2019; Elhage et al. 2021; Akyürek et al. 2024], which allows capturing n-grams instead of just bigrams.

Under this definition, we list a few simple examples of reasoning that we will consider below:

- *In-context recall*: when the last token is `a`, we'd like to copy the token that follows previous occurrences of `a` in the context. This `[.. a b .. a]` $\rightarrow$ `b` pattern typically requires a two-layer *induction head* mechanism [Elhage et al. 2021; Bietti et al. 2023; Sanford et al. 2024a];
- *Indirect object identification (IOI)*: we consider contexts of the form "When Mary and John went to the store, John gave the ice cream to" where the prediction should be "Mary" (IO, the indirect object), instead of "John" (S, the subject). Wang et al. [2022] found a circuit of several attention heads that perform this task by copying the name which only occurs once in the context;
- *Factual recall*: sentences of the form "Paul Citroen is a native speaker of" with target "Dutch" as in [Sharma et al. 2023]. While this may be seen as retrieving a distributional association, we will treat it here as reasoning since it involves combining the subject and relation from the context, while a bigram model that only depends on the last token "of" might instead predict the generic word "the".

4.3.2 Truncating Weights with LASER [Sharma et al. 2023]

In order to assess the importance of different weight components for certain predictions, we use the weight truncation technique introduced by Sharma et al. [2023]. They observed that reducing the rank of MLP matrices in certain layers of LLMs effectively brings better performance on several reasoning benchmarks. Their proposed method, Layer-Selective Rank Reduction (LASER), replaces any matrix in the full model by its low-rank approximation with fraction ρ, *i.e.*, a matrix $\mathbf{W} \in \mathbb{R}^{d_{\text{in}}, d_{\text{out}}}$ would be replaced by its rank-$\lfloor \rho \cdot \min\{d_{\text{in}}, d_{\text{out}}\} \rfloor$ approximation via Singular Value Decomposition (SVD). After searching for the best parameters of different models on different datasets, [Sharma et al. 2023] found that applying their method to weight matrices of MLPs on relatively deep layers can enhance in-context reasoning performance on various benchmarks, consistent with our findings. The optimal ρ is smaller than 0.2 for many datasets.

Another observation from [Sharma et al. 2023] is that, when LASER improves the model's prediction on some samples, the full model often predicts "generic" words while the improved model is able to predict the ground-truth answer. For instance, given an input "Madrid is located in", the full model predicts "the" while the truncated model predicts the target "Spain" in Table 4.1. Here, the generic word is consistent with our definition of distributional associations in Section 4.3.1, as it may naturally follow from a bigram distribution conditioned on "in", while the factual answer is more akin to reasoning from context. Thus, we would like to better understand how such a modification of feed-forward layers improves the model from predicting generic words to inferring the answer from context, and how such a gap appears during training.

Table 4.1: Probabilities of the top-5 next-tokens in Pythia-1B before and after LASER. The input prompt is "Madrid is located in". Probabilities of two generic words, *i.e.*, "the" and "a", drop sharply after LASER, while probabilities of meaningful words increase, especially the target "Spain".

	"the"	"Spain"	"a"	"southern"	"northern"
Full	**0.499**	0.079	0.069	0.023	0.021
LASER	0.027	**0.300**	0.002	0.044	0.046

4.4 Two-layer Transformer on Noisy In-context Recall

In this section, we consider simple one- or two-layer transformers on an in-context recall task with added generic token noise, which allows us to study the trade-offs between MLPs and attention layers for storing in-context versus distributional associations, in a controlled setting. We empirically show how transformers solve this task by storing the generic noise token in feed-forward layers, while attention implements the in-context mechanism. We then provide theory showing that feed-forward layers are more likely to store the distributional association (generic token) while attention learns to attend to in-context targets. Finally, we show that when the model has no feed-forward layers, the value matrix in attention stores both in-context and distributional information, in different subspaces.

Data and task. The data model we consider is similar to Bietti et al. [2023], with additional noise. Consider a vocabulary $\mathcal{V} = \{1, 2, \ldots, N, N+1\}$. The token $\tau \triangleq N+1$ is the generic noise token. We fix a *trigger* token $q \in [N]$, which governs in-context recall, and a context length T. Each sequence of tokens $z_{1:T} = [z_1, z_2, \ldots, z_T]$ is generated as follows:

i. Sample a correct *output* token $\bar{y}$ uniformly in $[N]$.

ii. Sample $z_{1:T-1}$ according to the following Markov process (π_u, π_b are distributions on $[N]$ defined later): $z_1 \sim \pi_u(\cdot)$, and

$$z_{t+1}|z_t \sim \begin{cases} \pi_b(\cdot|z_t), & \text{if } z_t \neq q, \\ p_{\alpha,\bar{y}}(\cdot), & \text{otherwise,} \end{cases} \qquad p_{\alpha,\bar{y}}(x) = \begin{cases} 1-\alpha, & \text{if } x = \bar{y}, \\ \alpha, & \text{if } x = \tau, \\ 0, & \text{otherwise.} \end{cases}$$

iii. Set $z_T = q$, and sample the final output $y = z_{T+1} \sim p_{\alpha,\bar{y}}(\cdot)$.

Note that the true $\bar{y}$ varies across sequences, so that the model needs to infer it from context,

e.g., using an induction head as in [Bietti et al. 2023]. Predicting $\bar{y}$ may thus be seen as a basic "reasoning" task, yet when training with $\alpha > 0$, the noisy output also requires the model to learn a distributional trigger-noise association, similar to the "of/in the" bigram discussed in Section 4.3. We also consider using multiple trigger tokens in Appendix 4.4.5 and Figure 4.8.

Two-layer transformer. We consider a simplified two-layer transformer formulated below. The input is a sequence of tokens $z_{1:T} = [z_1, \ldots, z_T] \in [N+1]^T$, and the output is ξ. The embedding matrix $\mathbf{W}_E \in \mathbb{R}^{(N+1)\times d}$ and un-embedding matrix $\mathbf{W}_U \in \mathbb{R}^{(N+1)\times d}$ are fixed at random initialization. The two attention layers have learnable weights $\mathbf{W}_{QK}^1, \mathbf{W}_V^1, \mathbf{W}_{QK}^2, \mathbf{W}_V^2 \in \mathbb{R}^{d\times d}$ with $\sigma(\cdot)$ the softmax on a vector. The two feed-forward layers F_1, F_2 are also learnable, and typically we set them as two-layer MLPs with ReLU activation. We will discuss different architectural choices of F_1, F_2 in Appendix 4.4.5.1. We use the cross-entropy loss to predict $y = z_{T+1}$ from the logits $\xi_T \in \mathbb{R}^{N+1}$.

$$
\begin{aligned}
x_t &\triangleq \mathbf{W}_E(z_t) + p_t, \\
h_t^1 &\triangleq \sum_{s\leqslant t} \left[\sigma(x_t^\top \mathbf{W}_{QK}^1 x_{1:t})\right]_s \cdot \mathbf{W}_V^1 x_s, \\
x_t^1 &\triangleq x_t + h_t^1 + F_1(x_t + h_t^1), \\
h_t^2 &\triangleq \sum_{s\leqslant t} \left[\sigma(x_t^{1\top} \mathbf{W}_{QK}^2 x_{1:t}^1)\right]_s \cdot \mathbf{W}_V^2 x_s^1, \\
x_t^2 &\triangleq x_t^1 + h_t^2 + F_2(x_t^1 + h_t^2), \\
\xi_t &\triangleq \mathbf{W}_U x_t^2.
\end{aligned}
\tag{4.1}
$$

Experimental observations. Following [Bietti et al. 2023], we take π_u and π_b to be the unigram and brigram character-level distributions estimated from the tiny Shakespeare dataset with $N = 65$. The model setup includes $d = 256$ and two-layer MLPs with ReLU for both F_1, F_2. The training setup includes batch size as 512 and the context length $T = 256$. When evaluating trained models, we consider LASER on the input weight U_{in} of F_2. We consider a noise level $\alpha = 0.5$ for training data (though any other constant value would lead to similar observations).

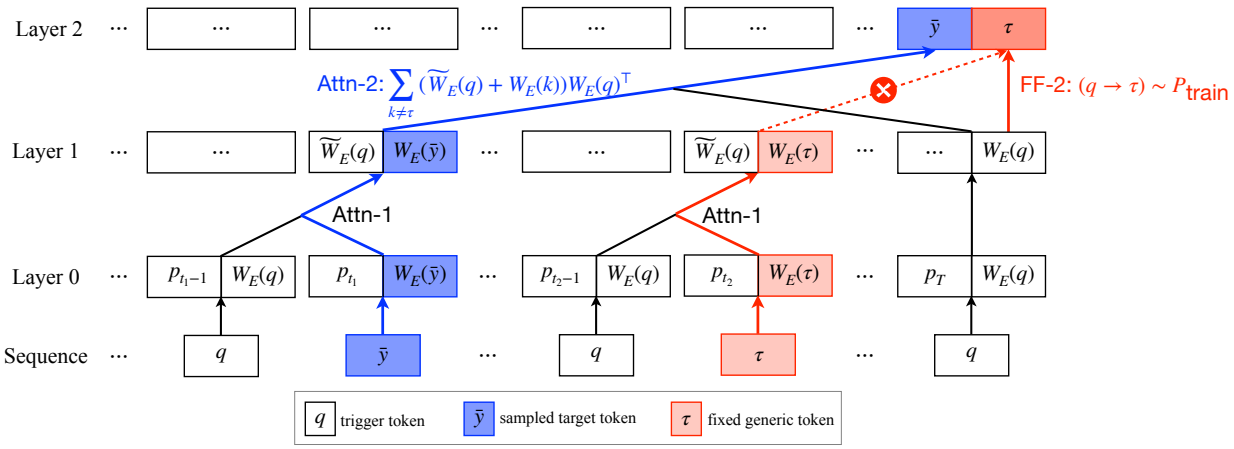


Figure 4.2: Noisy in-context recall. *Purpose of design*: understand mechanisms of attention and feed-forward layers for tasks with **in-context reasoning** (predict $\bar{y}$) and **distributional association** (predict τ). *Task*: predict tokens $\bar{y}$ *v.s.* τ from a sentence $[\ldots, q, \bar{y}, \ldots, q, \tau, \ldots, q]$ where q is trigger, $\bar{y}$ is sampled target token for a sentence, and τ is a fixed generic token across sentences. *Our findings*: in a two-layer transformer, the second-layer attention (Attn-2) only attends towards target tuples $[q, \bar{y}]$ while the feed-forward layer (FF-2) learns to predict τ.

During test time, we set $\alpha = 0$ to compute the test loss, aiming to measure how likely the (full or after-truncation) model predicts the ground-truth $\bar{y}$.

Experimental results are reported in Figure 4.3 and 4.7. The full model predicts noise with probability close to α, which is expected since it is trained to predict the noise token w.p. α. However, when dropping the second-layer MLP F_2, the truncated model predicts the ground-truth $\bar{y}$ with an almost perfect probability ≈ 0.98. This suggests that F_2 is responsible for storing the distributional association "[trigger] + [noise]". Another observation is that the full model first learns to predict the noise with high probability in very early steps, after which it starts learning to predict the correct $\bar{y}$, which resembles the dynamics observed for learning the "to/in the" bigram in Pythia models in Figure 4.13. This suggests that learning the (distributional) trigger-noise association is easier than predicting $\bar{y}$, and we will study this theoretically in Section 4.4.1.

After the distributional noise association is learned, we observe a slower learning of an induction head mechanism, with similar dynamics to Bietti et al. [2023]. Compared to Bietti et al. [2023], we notice that the induction head (i.e., the second layer attention head) filters out the

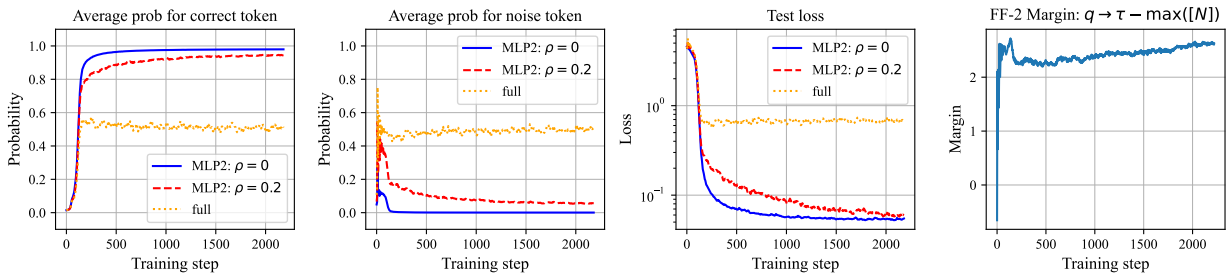


Figure 4.3: Left three: Average probability of predicting correct and noise tokens, and test loss on clean data ($\alpha = 0$), with different fractions ρ of preserved rank in U_{in} of the second-layer MLP F_2. The full model learns to predict noise with probability around $\alpha = 0.5$, as expected from training data. When F_2 is dropped ($\rho = 0$), the model predicts the correct token $\bar{y}$ with probability ≈ 0.98. **Rightmost**: the FF-2 margin of τ *v.s.* all the other tokens with input as q, *i.e.*, $[\mathbf{W}_U F_2(\mathbf{W}_E(q))]_\tau - \max_{k \leqslant N} [\mathbf{W}_U F_2(\mathbf{W}_E(q))]_k$. It reveals that FF-2 learns trigger-noise association in early steps.

noise tokens and only attends to non-noisy output tokens following the trigger, corresponding to the correct $\bar{y}$, as shown in Figure 4.4. We present theoretical understanding for this mechanism in Section 4.4.2. Figure 4.2 and Appendix 4.4.4.2 summarize the roles of all components of the two-layer transformer in this task.

Simplified architecture and data for theoretical analysis. Understanding the full dynamics of the model used in our experiments is out of the scope of the present paper, due to the many moving parts and the complexity of non-linear MLPs. Instead, we focus on a simpler model involving one linear feed-forward layer and one attention layer, and look at the gradient dynamics near initialization. We consider the following simplified 1-layer model. The input $x_t \in \mathbb{R}^d$ at position t is defined as $x_t \triangleq \mathbf{W}_E(z_t) + \widetilde{\mathbf{W}}_E(z_{t-1})$, where $z_t \in [N+1]$ is the token at position t, $\mathbf{W}_E(z_t)$ is its embedding and $\widetilde{\mathbf{W}}_E(z_{t-1})$ is a different embedding of the previous token to a different direction, as in the *previous token head* construction of Bietti et al. [2023], where the value matrix remaps the previous token to a different subspace. We assume all embeddings to be orthogonal (Assumption C.2.1), which requires large enough d, and holds in the infinite-width limit with random embeddings. This model allows us to simplify our analysis by considering a single attention layer with no positional embeddings, while capturing the difficulty of long-range interactions. We note that such a simplification is standard in the in-context learning literature [e.g.,

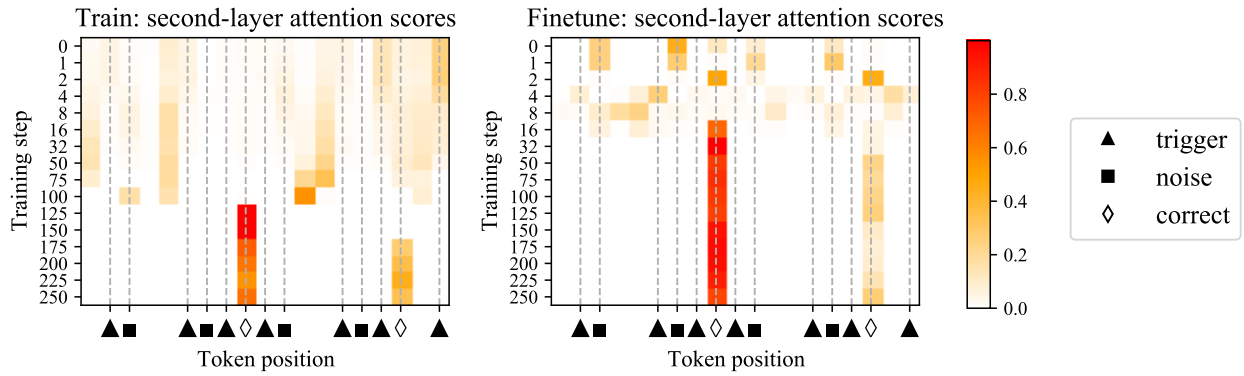


Figure 4.4: The second-layer attention scores of models trained with noise (left), fine-tuned with noise (right, initialized as a model pre-trained without noise), given the same input. It turns out both models learn to attend to the informative structure "[trigger]+$\bar{y}$" instead of "[trigger]+noise". This implies that the attention in these models is only responsible to predict $\bar{y}$, although the training input and output have noise with probability $\alpha = \Theta(1)$. The fine-tuning setting is in Appendix 4.4.4.1.

Akyürek et al. 2023; Mahankali et al. 2024; Zhang et al. 2024], For data generation, π_u and π_b are uniform distributions on $[N]$. Given a sequence of inputs, $x_{1:T} \in \mathbb{R}^{T\times d}$, the output of model is $\xi \triangleq \xi_{\text{attn}} + \xi_{\text{ff}}$ as

$$\begin{aligned} x_t &\triangleq \mathbf{W}_E(z_t) + \widetilde{\mathbf{W}}_E(z_{t-1}) \in \mathbb{R}^d, \\ \phi(x_T, x_{1:T}) &\triangleq \sum_{t\leqslant T} \left[\sigma\left(x_T^\top \mathbf{W}_{QK} x_{1:T}\right)\right]_t \cdot \mathbf{W}_V x_t \in \mathbb{R}^d, \\ \xi_{\text{attn}}(x_{1:T}) &\triangleq \mathbf{W}_U \phi(x_T, x_{1:T}) \in \mathbb{R}^{N+1}, \\ \xi_{\text{ff}}(x_{1:T}) &\triangleq \mathbf{W}_U F(x_T) = \mathbf{W}_U \mathbf{W}_F x_T \in \mathbb{R}^{N+1}, \end{aligned} \tag{4.2}$$

where $\mathbf{W}_U \in \mathbb{R}^{(N+1)\times d}$ is the unembedding matrix, $\phi(s,t)$ is the attention module with query s and context t, and $F(\cdot)$ is a linear feed-forward layer. This architecture is similar to a one-layer transformer, but already highlights the difference between feed-forward and attention layers in a way that we expect to still hold for more layers. In the above parametrization, the learnable matrices are $\mathbf{W}_{QK}, \mathbf{W}_F, \mathbf{W}_V \in \mathbb{R}^{d\times d}$. At initialization, we set $\mathbf{W}_{QK}, \mathbf{W}_F, \mathbf{W}_V = 0$, noting that random initialization in high dimension would lead to similar behaviors thanks to near-orthogonality.

4.4.1 Feed-forward layers store the generic noise

As we saw in Figure 4.3 and 4.7, the model very quickly learns to predict the noise token after a few steps. Then the gap between $\rho = 0$ and 1 in Figure 4.3 suggests that the feed-forward layer F_2 is responsible for storing the distributional association about noise, which is verified in Figure 4.6 (middle). We now provide theoretical justification for this behavior. In particular, we will show that, at initialization, the gradients over the feed-forward parameters are much more informative than the attention gradient, which is dominated by noise unless the sample size is very large. This shows that the feed-forward layer is much more likely to capture the distributional association.

We now look at the first gradient step from initialization, which has commonly been used to understand feature learning and sample complexity in neural networks [Damian et al. 2022a; Ba et al. 2022; Dandi et al. 2023; Oymak et al. 2023; Bietti et al. 2023]. Note that $\mathbf{W}_{QK}$ has no gradient at initialization, so that the gradient of $\mathbf{W}_V$ is most relevant initially [see also Snell et al. 2021; Li et al. 2023; Oymak et al. 2023; Bietti et al. 2023].

Theorem 4.1 (Logits after one gradient step). *Assume $N, T \gg 1, \alpha = \Theta(1)$. For the model in Eq(4.2), consider one gradient step update from zero-initialization on m i.i.d. samples of $z_{1:T}$ with separate learning rates η_f for $\mathbf{W}_F$ and η_v for $\mathbf{W}_V$ (note that the gradient on $\mathbf{W}_{QK}$ is zero). With probability $1 - \delta$, the resulting logits for the feed-forward and attention blocks satisfy, for any test sequence $z_{1:T}$,*

$$\left|\Delta(\xi_{ff}(x_{1:T})) - \eta_f \cdot \alpha\right| \leqslant \eta_f \cdot O\left(\sqrt{\frac{\ln \frac{2(N+1)}{\delta}}{m}}\right),$$

$$\left|\Delta(\xi_{attn}(x_{1:T})) - \frac{\eta_v}{N} \cdot \hat{\alpha}\right| \leqslant \eta_v \cdot O\left(\sqrt{\frac{(\frac{1}{TN} + \frac{1}{N^2}) \ln \frac{2(N+1)}{\delta}}{m}} + \frac{\ln \frac{2(N+1)}{\delta}}{m}\right),$$

where $\Delta(\xi) = \xi_{N+1} - \max_{j \in [N]} \xi_j$ is the margin of predicting the generic noise token and $\hat{\alpha} = (\alpha^2 \hat{q} + \alpha(1 - \hat{q}))$, where $\hat{q} = \frac{1}{T} \sum_{t \leqslant T} \mathbb{1}\{z_t = N + 1\}$ is the fraction of noise tokens in $z_{1:T}$.

The margin $\Delta(\xi)$ reflects how much signal there is in the logits for predicting the noise token, and the theorem provides concentration bounds on the contributions of the updates on $\mathbf{W}_F$ and $\mathbf{W}_V$ to the margin. Note that $\hat{q} \ll 1$ w.h.p. for large N, T, so $\hat{\alpha} \approx \alpha$. We make the following observations:

i. When $m = \tilde{\Omega}(1)$, there is enough signal in $\mathbf{W}_F$ to predict the noise, say with $\eta_f = 1$, and a choice of $\eta_v = O(1)$ will lead to a small but controlled contribution to the prediction from $\mathbf{W}_V$.

ii. When $m = \tilde{\Omega}(N)$, $\mathbf{W}_V$ can also reliably predict the noise by setting $\eta_v = \Theta(N)$ (i.e., with small deviation on the r.h.s.), at the cost of many more samples.

Our result shows that in the initial phase of training, feed-forward layers are more likely to pick up the noise token, leading to a structure of the form $\mathbf{W}_F \approx \mathbf{W}_U(N+1)\mathbf{W}_E(q)^\top$, while attention will be slower due to additional noise and possibly smaller step-sizes. We may then expect the attention layers to focus instead on in-context reasoning, as we observe empirically and discuss next.

4.4.2 Attention attends to in-context targets and avoids noise

When the feed-forward weight learns to predict the noise as shown in Theorem 4.1, Figure 4.4 reveals that the second-layer attention in the two-layer model attends only towards the correct tokens. In contrast, a model pre-trained without noise has second-layer attention attend towards all tokens just after the triggers [Bietti et al. 2023], as observed in the attention pattern at the first step in Figure 4.4(right). Then, after being fine-tuned on noise data, the attention becomes only focused on the correct tokens. Understanding this mechanism requires the analysis of the dynamics of $\mathbf{W}_{QK}$.

Following the simplified model and data distribution in Eq(4.2), we take a step towards understanding how attention "avoids" the noise tokens. Concretely, this mechanism appears because, after the initial training phase when FF learns noise association much faster than the attention,

$\mathbf{W}_V$ has a structure of $\sum_{k\leqslant N+1} \mathbf{W}_U(k)(\mathbf{W}_E(k)+\widetilde{\mathbf{W}}_E(k))^\top$, similar to the non-noise setting in [Bietti et al. 2023]. After such a $\mathbf{W}_V$ is learned, the trigger-label association provides a stronger gradient signal on $\mathbf{W}_{QK}$ than the trigger-noise association. We show this in the following theorem.

Theorem 4.2 (Attention attends to in-context targets). *Assume $N, T \gg 1$ and Assumption C.4.1 hold. Consider the simplified model in Eq(4.2) with infinite samples as $m \to \infty$. After $\mathbf{W}_F$ learns the noise association as in Theorem 4.1, in one step the attention weight $\mathbf{W}_{QK}$ learns to attend to positions $t \in [T]$ where the correct label follows a trigger word, i.e., $z_{t-1} = q, z_t = \bar{y}$.*

More concretely, $\mathbf{W}_{QK}$ has the following structure

$$\xi_{q\to j} - \xi_{k\to l} = \Omega(N^{-3}) > 0, \quad if\, k \neq q,\ \forall\ j, l, \tag{4.3}$$

$$\xi_{q\to j} - \xi_{q\to N+1} = \Omega(N^{-4}) > 0, \quad \forall\ j \leqslant N, \tag{4.4}$$

where $\xi_{i\to j} \triangleq \mathbf{W}_E(q)^\top \mathbf{W}_{QK}(\widetilde{\mathbf{W}}_E(i) + \mathbf{W}_E(j))$ denotes the attention logit for different combinations of $z_{t-1} = i, z_t = j$, with $i, j \leqslant N + 1$.

Note that a set of logits induces a probability distribution via differences between them as $\exp(\xi_i)/\sum_j \exp(\xi_j) = 1/\sum_j \exp(\xi_j - \xi_i)$. Therefore, the above theorem reveals that the attention has two patterns: Eq. (4.3) shows that $\mathbf{W}_{QK}$ prefers attending to locations just after a trigger q, *i.e.*, such that $z_{t-1} = q$, similar to [Bietti et al. 2023], and Eq. (4.4) shows that among all positions that follow a trigger q, $\mathbf{W}_{QK}$ places less attention on the noise token, *i.e.*, $z_t = N + 1$, compared to correct tokens $z_t = \bar{y} \leqslant N$. Such a key difference for attention between noisy and non-noise tasks verifies our experimental observations in Figure 4.4.

4.4.3 No feed-forward Layers: value matrix stores generic noise association

In the above discussion, we've seen separate roles of attention and feed-forward layers play to conduct noisy in-context learning. A natural question is, when there is *no feed-forward layer*, how the attention layer stores both in-context and distributional information. Figure 4.12 indicates that the value matrix stores the noise association in subspace with smaller singular values. In this section, we propose a setting of *linear associative memory with noise* to understand this mechanism.

Unlike Theorem 4.1 and 4.2 showing the separate roles of attention and FF, the attention in a non-FF model has to handle both noise and in-context information once the model is sufficiently trained to reach a global minimum. Due to symmetry from uniformly random sampling $\bar{y}$ from N tokens, we consider passing the output $x \in \mathbb{R}^d$ of the attention to the value matrix $\mathbf{W}_V$ and output matrix $\mathbf{W}_U$ to predict next-token probability $y \in \mathbb{R}^{N+1}$ given $z_{1:T} \in [N+1]^T$ with noise probability of α as follows

$$\begin{aligned} x|\bar{y}, z_{1:T} \triangleq \mathbf{W}_E(\bar{y}) + \overline{\mathbf{W}}(z_{1:T}) \in \mathbb{R}^d, \quad \xi \triangleq \mathbf{W}_U \mathbf{W}_V x \in \mathbb{R}^{N+1}, \\ p_\alpha(y|\bar{y}) = (1-\alpha) \cdot \mathbb{1}\{y = \bar{y}\} + \alpha \cdot \mathbb{1}\{y = N+1\}, \end{aligned} \tag{4.5}$$

where $\overline{\mathbf{W}}(z_{1:T})$ is an aggregate embedding independent of $\bar{y}$. When $T \to \infty$, $\overline{\mathbf{W}}(z_{1:T})$ converges to a fixed embedding $\overline{\mathbf{W}}$ independent of $\bar{y}$, so that we may consider a simplified model $x|\bar{y} \triangleq \mathbf{W}_E(\bar{y}), \xi \triangleq \mathbf{W}x \in \mathbb{R}^{N+1}$ with $\mathbf{W} \in \mathbb{R}^{(N+1)\times d}$, since $\overline{\mathbf{W}}$ only contributes a fixed offset in all logits that can be easily canceled in the softmax predictions. Therefore, we investigate the following *linear associative memory with noise*.

Model and data. Consider a learnable weight matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$ with $d > N$. Consider embeddings for N input tokens as $\{e_i\}_{i=1}^N \subset \mathbb{R}^d$ and embeddings for $(N+1)$ output tokens as

$\{u_i\}_{i=1}^{N+1} \subset \mathbb{R}^d$. Given any pair of input and output tokens, the associative memory model takes the form $f(i, j; \mathbf{W}) \triangleq \langle u_j, \mathbf{W} e_i \rangle,\ \forall\ i, j \in [N] \times [N+1]$, as logits to approximate $p_\alpha(\cdot|i)$ in (4.5). When $k \leqslant d$, we denote the rank-k approximation of f as $f^{(k)}$ by replacing $\mathbf{W}$ with $\mathbf{W}^{(k)}$, where $\mathbf{W}^{(k)}$ is its rank-k approximation.

Experiments. During training, the dataset $\mathcal{D}_\alpha$ is generated with non-zero noise probability $\alpha > 0$. At test time, the dataset $\mathcal{D}_0$ is without noise as $\alpha = 0$, so the computed loss is called **pure-label** loss. The *full* model is trained with Gradient Descent (GD) subjected to cross-entropy loss. The results are reported in Figure 4.5, with more discussions in Appendix C.5.

Low-rank subspace stores noise. In Figure 4.5, the rank-1 subspace corresponding to the smallest non-zero singular value is responsible to store the noise. We prove this mechanism as follows. Note that, here $N = 2$ is for simplicity, which is easy to extend to any $N > 2$.

Theorem 4.3. *Assume Assumptions C.5.1 and C.5.2 hold, considering $N = 2$ and $\alpha \in (0.2, 0.4)$, we train the full model $f(\cdot, \cdot; \mathbf{W})$ with gradient flow. Denote $P(i, j; \mathbf{W})$ as the model's predicted probability for output j conditioned on input i. Then, for $t \to \infty$ and $i \in \{1, 2\}$, we have*

$$P(i, j; \mathbf{W}) = (1 - \alpha) \cdot \mathbb{1}\{j = i\} + \alpha \cdot \mathbb{1}\{j = N + 1\},$$
$$P(i, j; \mathbf{W}^{(1)}) = (1 - \Theta(t^{-1/2})) \cdot \mathbb{1}\{j = i\} + \Theta(t^{-1/2}) \cdot \mathbb{1}\{j = N + 1\}.$$

The above theorem implies, the full model always predicts noise w.p. α, while the rank-1 model eventually predicts correctly without noise, although training is only on the full model with noise. Actually when $N > 2$, the noise is stored in rank-1 subspace and the correct correspondence is stored in rank-$(N-1)$ space. Therefore, this explains how the value matrix stores both in-context and noise information when the model is without FF.

Randomness in experiments. Assume both $\{e_i\}_{i=1}^n$ and $\{u_i\}_{i=1}^c$ are i.i.d. uniformly drawn from sphere $\mathbb{S}^{d-1}$. Also assume the model is initialized as $\mathbf{W}_{i,j} \sim \mathcal{N}(0, \frac{1}{d})$. Due to randomness

from embeddings and model initialization, let's first conduct 20 runs of experiments to obtain significant factors before moving the theoretical argument.

Note that *only full models are trained*, and we track loss for low-rank models by conducting SVD in each step without manipulating training. In Figure 4.5, we illustrate the pure-label loss *v.s.* training steps for models of different ranks, where $n = 3$, $\alpha = 0.03$ and $d = 8$ or 12. It turns out, while the full model (rank$\geqslant 3$) has a constant pure-label loss (~ 0.03, dependent on α), the rank-2 model is very likely to have a significant loss than the full model. Meanwhile, the larger d has more stable results than small d.

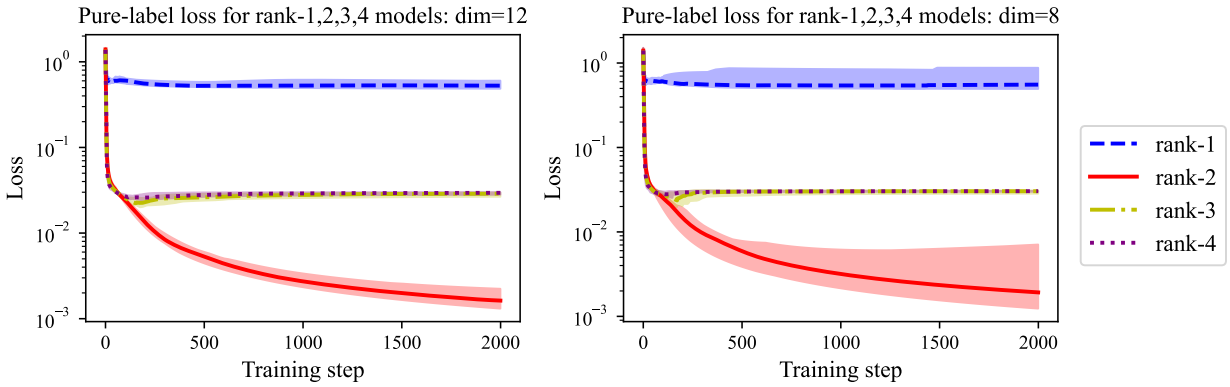


Figure 4.5: Pure-label loss for rank-1,2,3,4 models with $n = 3, \alpha = 0.03$ and $d = 12$ (left) or 8 (right). *Only full models are trained*, and we report low-rank results by conducting SVD in each step without manipulating the training. In both figures, the experiments are run for 20 times to examine the randomness. For each rank, we plot curves of the median, 25% and 75% out of 20 runs. It turns out: i) rank-2 models are very likely to have significantly lower pure-label loss thant full models (rank$\geqslant 3$), and ii) the larger dimension d has more stable results.

Therefore, we can qualify the following important factors for this model:

i. d *v.s.* n, c: when $d \gg n, c$, random drawn embeddings tend to be orthogonal to each other, with inner product in $O(1/\sqrt{d})$. If $n, c = \Omega(d)$, embeddings will be in strong correlations, making the problem extremely difficult to understand. [Cabannes et al. 2024] also discussed about such particle interaction in associative memory.

ii. Low-rank subspace storing the noise. In Figure 4.5, the rank-1 subspace between the full

and rank-2 models is responsible to store the noise, removing which will induce a model ideally predicting the ground-truth without noise. This is understandable if the embeddings are orthogonal, as shown in Theorem 4.3.

iii. α *v.s.* n. When n is large, orthogonal embeddings still induces a low-rank subspace storing the noise, but α decides whether the low-rank subspace corresponds to the smallest singular values of $\mathbf{W}$. If not, it requires more careful manipulation of the spectrum instead of low-rank approximation of $\mathbf{W}$.

4.4.4 How Does the Two-layer Model Solve Noisy In-context Recall?

4.4.4.1 Training settings

In most parts of this work, we consistently train the model with a fixed level of $\alpha > 0$. However, we also present numerical results of **fine-tuning** in Figure 4.7 and 4.4 to show the mechanism of avoiding generic noise token in the second-layer attention. The details of such a fine-tuning setting is as follows.

Fine-tuning: there are two phases of training as

- phase 1 (pre-training): starting from a model with random initialized weights, we train the model on data generated with $\alpha = 0$. This is exactly the same as [Bietti et al. 2023]. At the end of this phase, the second-layer attention is expected to attend *all tokens* after the trigger token, *i.e.*, $t \leqslant T$ such that $z_{t-1} = q$ no matter what z_t is.

- phase 2 (fine-tuning): starting from a model after phase 1, we train all weights in the model on data generated with $\alpha > 0$. At the end of this phase, the second-layer attention learns to avoid the generic noise token, *i.e.*, $t \leqslant T$ such that $z_t = N_1, z_{t-1} = q$, as shown in Figure 4.4.

4.4.4.2 Summarizing: roles of key components in the two-layer transformer

Recall the architecture of two-layer transformers in Section 4.4 as

$$
\begin{aligned}
x_t &\triangleq \mathbf{W}_E(z_t) + p_t, \\
h_t^1 &\triangleq \sum_{s \leqslant t} \left[\sigma(x_t^\top \mathbf{W}_{QK}^1 x_{1:t})\right]_s \cdot \mathbf{W}_V^1 x_s, \\
x_t^1 &\triangleq x_t + h_t^1 + F_1(x_t + h_t^1), \\
h_t^2 &\triangleq \sum_{s \leqslant t} \left[\sigma(x_t^{1\top} \mathbf{W}_{QK}^2 x_{1:t}^1)\right]_s \cdot \mathbf{W}_V^2 x_s^1, \\
x_t^2 &\triangleq x_t^1 + h_t^2 + F_2(x_t^1 + h_t^2), \\
\xi_t &\triangleq \mathbf{W}_U x_t^2.
\end{aligned}
$$

When the task is without noise, *i.e.*, $\alpha = 0$, [Bietti et al. 2023] point out the first-layer attention attends to the previous token through $\mathbf{W}_{QK}^1 = \sum_{t=2}^T p_{t-1} p_t^\top$. Therefore, when $z_t = \bar{y}$ with $z_{t-1} = q$, the output of the first layer is $x_t^1 \approx \mathbf{W}_E(\bar{y}) + \mathbf{W}_V^1 \mathbf{W}_E(q)$. Then they show that the second-layer attention matches such x_t^1 with $z_T = q$ by $\mathbf{W}_{QK}^2 = (\mathbf{W}_V \mathbf{W}_E(q)) \mathbf{W}_E(q)^\top$, through which the information of $\bar{y}$ in x_t^1 is copied to last token as $h_T^2 \approx \mathbf{W}_V^2 \mathbf{W}_E(\bar{y})$. Finally $\mathbf{W}_V^2 = \sum_{z \in [N]} \mathbf{W}_U(z) \mathbf{W}_E(z)^\top$ helps output the correct label of $\bar{y}$.

In our work with noise $\alpha > 0$, the key difference is that there is a fixed probability α for a noise token $N+1$ to appear after each trigger q. This requires $\mathbf{W}_{QK}^2$ to not only match the trigger but also avoid the noise token after trigger. Let's first summarize the whole pipeline of this model for our task.

Roles of key components. The first layer will be basically the same as [Bietti et al. 2023], where $\mathbf{W}_{QK}^1 = \sum_{t=2}^T p_{t-1} p_t^\top$ attends to the previous token. Consider two positions t_1, t_2 with $z_{t_1-1} = z_{t_2-1} = q, z_{t_1} = \bar{y}, z_{t_2} = N+1$, then outputs of the first layer at these two positions are $x_{t_1}^1 \approx \mathbf{W}_E(\bar{y}) + \mathbf{W}_V^1 \mathbf{W}_E(q)$, $x_{t_2}^1 \approx \mathbf{W}_E(N+1) + \mathbf{W}_V^1 \mathbf{W}_E(q)$. Then the second-layer attention $\mathbf{W}_{QK} = (\mathbf{W}_V \mathbf{W}_E(q) - c \cdot \mathbf{W}_E(N+1)) \mathbf{W}_E(q)^\top$ with some positive c makes the attention attend

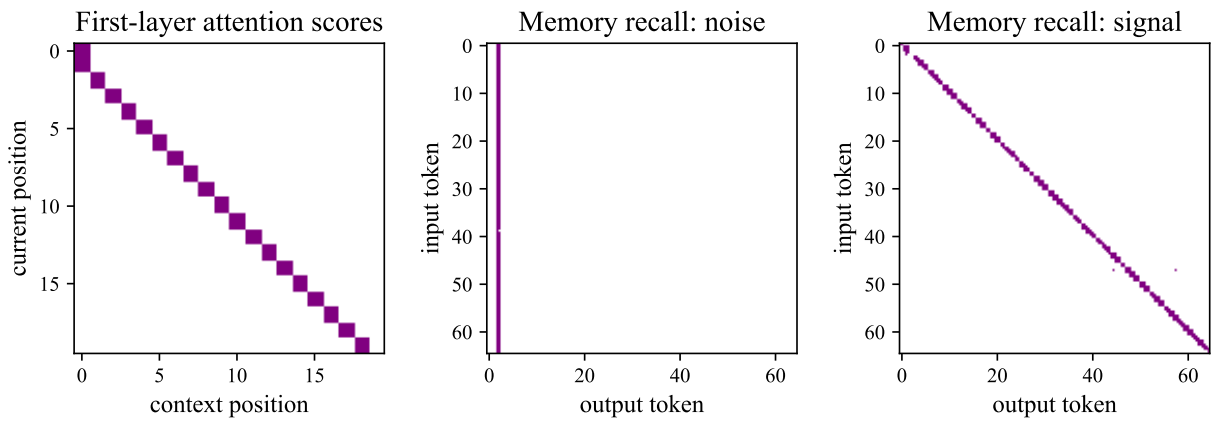


Figure 4.6: Left: first-layer attention attending to the previous token from the current token. **Middle**: logits to predict noise from $\langle F_2(\mathbf{W}_E(i)), \mathbf{W}_U(j)\rangle$ with input $i \in [N+1]$ and output $j \in [N+1]$, where the output channel 2 is set as the noise channel. It turns out, for all input i, the logits on output 2 are large, which matches our construction that, at least for trigger q as input, the output 2 has large logits. **Right**: logits to predict singal from $\langle \mathbf{W}_V^2 \mathbf{W}_E(i), \mathbf{W}_U(j)\rangle$ for input $i \in [N+1]$ and output $j \in [N+1]$. It matches our construction that $i = j$ has large logits. Meanwhile, $i = j = 2$ does not have large logits since 2 is the noise channel.

to t_1 and avoid t_2 simultaneously, matching with the last token $z_T = q$. Therefore, the output of the second-layer attention at T is basically $h_T^2 \approx \mathbf{W}_V^2 \mathbf{W}_E(\bar{y})$. Similar to the noiseless case, $\mathbf{W}_V^2 = \sum_{z\in[N]} \mathbf{W}_U(z)\mathbf{W}_E(z)^\top$ helps output the correct label of $\bar{y}$. Meanwhile, note that x_T^1 actually contains $\mathbf{W}_E(q)$ through x_T, so F_2 is able to predict the noise $N+1$ when seeing a fixed $\mathbf{W}_E(q)$. As a result, combining the two streams from h_T^2 and $F_2(x_T^1)$, the full model is able to predict any $\bar{y}$ w.p. $1-\alpha$ and predict the noise $N+1$ w.p. α.

Evidence. Figure 4.4 illustrates that the second-layer attention learns to attend to $z_{t_1} = \bar{y}$ and avoid $z_{t_2} = N+1$, with Appendix 4.4.4.3 presenting a primitive exploration on how the avoidance is learnt in a simplified setting. Figure 4.6 (left) shows the attention pattern from $\mathbf{W}_{QK}^1$ of attending to the previous token. Figure 4.6 (middle) shows the memory recall of $\mathbf{W}_U(N+1)^\top F_2(\mathbf{W}_E(q))$ to predict the noise. Figure 4.6 (right) illustrates the memory recall of $\mathbf{W}_U(i)^\top \mathbf{W}_V^2 \mathbf{W}_E(i)$ to predict the correct token.

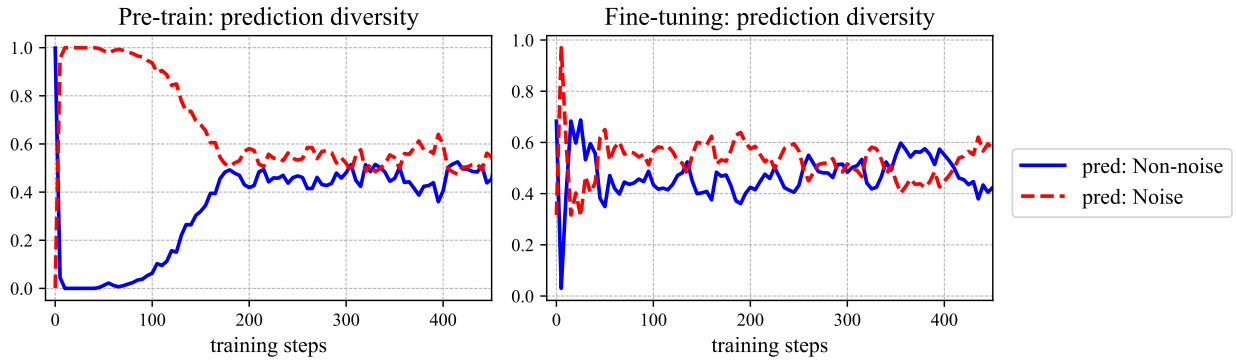


Figure 4.7: Fractions of predicting the noise token and the other non-noise tokens with $\alpha = 0.5$. (Left) pretraining steps on noisy data; (right) finetuning steps on noisy data, after pretraining on clean data with $\alpha = 0$. In both cases, the models learn to predict noise with probability nearly 0.5. In the first few (~ 5) steps, the models quickly learn to predict noise with probability close to 1. The fine-tuning setting is in Appendix 4.4.4.1.

4.4.4.3 How does attention attend less towards the noise token?

We use the same simplified model as in Section 4.4.1 to understand how the second-layer attention learns to avoid the noise. When using the same learning rate $\eta = \eta_v = \eta_f$, Theorem 4.1 implies that the feed-forward $\mathbf{W}_F$ makes the most contribution for predicting the noise after the first-step update. Denote the logits for the noise of the model at time t as $\check{\xi}_t$. The arguments in this section make the following assumptions, which hold at least after the first-step update:

i. $\mathbf{W}_F$ dominates the logits $\check{\xi}_t$ of predicting the noise token, compared with $\mathbf{W}_V$.

ii. Logits for predicting any $k \leqslant N$ is close to 0, which means the predicted probability p_t is approximately $p_t \approx \frac{\exp(\check{\xi}_t)}{N+\exp(\check{\xi}_t)}$.

iii. The predicted probability $p_t < \alpha$.

iv. The attention matrix $\mathbf{W}_{QK}$ is approximately 0, inducing a uniform attention.

v. The dataset has $T, N \gg 1$ and $m \to \infty$, so the gradient is from population loss.

The first assumption holds after the first step from Theorem 4.1 with $\eta_f = \eta_v$.

Then, since $|\mathbf{W}_U(k)^\top(\nabla_{\mathbf{W}_F}L)\mathbf{W}_E(q)| = O(\frac{1}{N})\cdot|\mathbf{W}_U(N+1)^\top(\nabla_{\mathbf{W}_F}L)\mathbf{W}_E(q)|$ for any $k \leqslant N$ in Lemma C.1, the second assumption holds. Meanwhile, the projection of $\nabla_{\mathbf{W}_V}L$ onto any direction in Lemma C.2 is also smaller than $\mathbf{W}_U(N+1)^\top(\nabla_{\mathbf{W}_F}L)\mathbf{W}_E(q)$ by a factor of $O(1/N)$.

Let's check the condition of the third assumption. In the proof of Lemma C.1, the gradient of $\mathbf{W}_F$ has the form of

$$\mathbf{W}_U(N+1)^\top(-\nabla_{\mathbf{W}_F}L)\mathbf{W}_E(q) = \alpha - p_t.$$

This update induces ξ_t to increase by $\eta(\alpha - p_t)$. This implies

$$\xi_t \approx \xi_{t-1} + \eta\left(\alpha - \frac{\exp(\xi_t)}{N+\exp(\xi_t)}\right),\quad \forall\, t \geqslant 1.$$

This sequence $\{\xi_t\}_{t\geqslant 1}$ has stationary point $\xi^* = \log N + \log(\frac{\alpha}{1-\alpha})$. Denoting $\hat{\xi}_t \triangleq \xi_t - \xi^*$ with $\hat{\xi}_1 = -\xi^* < 0$, the iteration becomes

$$\hat{\xi}_{t+1} \approx \hat{\xi}_t + \eta\left(\alpha - \frac{\exp(\hat{\xi}_t)}{\frac{1-\alpha}{\alpha}+\exp(\hat{\xi}_t)}\right).$$

If we would like to have $\hat{\xi}_t$ not hit the positive region by controlling η, it suffices to bound η with any $\hat{\xi} < 0$,

$$\eta \leqslant \frac{\hat{\xi}}{\frac{\exp(\hat{\xi})}{\frac{1-\alpha}{\alpha}+\exp(\hat{\xi})} - \alpha},$$

where RHS is continuous and decreasing on $\xi < 0$ when $\alpha < 0.5$. Hence, we have $\eta \leqslant \frac{1}{\alpha(1-\alpha)}$ evaluated at $\hat{\xi} = 0$ by L'Hospital rule. This bound of η is very strong, since $\eta = O(\log N)$ can still have $\hat{\xi} < 0$ after one step.

The fourth assumption is basically from what we will show at the end of this section, as the second observation.

Then consider the dynamics of $\mathbf{W}_V$, which is much slower than $\mathbf{W}_F$. From the proof of Lemma C.2, the gradient of $\mathbf{W}_V$ satisfies

$$\begin{aligned} \nabla_{\mathbf{W}_V} L &= \mathbb{E}_x \left[\sum_{k=1}^{N+1} (p_{\mathbf{W}}(k|x) - \mathbb{1}\{y=k\}) \mathbf{W}_U(k) \left(\frac{1}{T} \sum_{t=1}^{t} x_t \right)^\top \right], \\ \mathbf{W}_U(N+1)^\top (-\nabla_{\mathbf{W}_V} L) \mathbf{W}_E(k) &\approx \frac{1}{N} \sum_{t \geqslant 1} (\alpha - p_t)(\mathbb{1}\{k \leqslant N\} + \alpha \cdot \mathbb{1}\{k = N+1\}) \\ &\triangleq c \cdot \mathbb{1}\{k \leqslant N\} + c \cdot \alpha \cdot \mathbb{1}\{k = N+1\} = \Theta(\frac{1}{N}), \end{aligned} \tag{4.6}$$

where the projection on $W_E(N+1)$ is always positive and smaller than that on other directions when $p_t < \alpha$. Projections onto other directions $\mathbf{W}_U(j)\mathbf{W}_E(k)^\top, \forall\, j \leqslant N$, are smaller as $\Theta(\frac{1}{N^2})$.

Finally, let's consider the dynamics of $\mathbf{W}_{QK}$. At initialization, $\mathbf{W}_{QK} = 0$ and $\nabla_{\mathbf{W}_{QK}} L = 0$ due to zero initialization of $\mathbf{W}_V$. After one-step, $\mathbf{W}_V$ has such a structure in Eq.(4.6). Then, with $\bar{x}_{1:T} \triangleq \frac{1}{T} \sum_{1 \leqslant t \leqslant T} x_t$ from uniform attention, the gradient of $\mathbf{W}_{QK}$ satisfies

$$\begin{aligned} -\nabla_{\mathbf{W}_{QK}} L &= \mathbb{E}_x \left[\sum_{k=1}^{N} (\mathbb{1}\{y=k\} - p_{\mathbf{W}}(k|x)) \frac{1}{T} \sum_{t=1}^{T} (\mathbf{W}_U(k)^\top \mathbf{W}_V x_t) \cdot (x_t - \bar{x}_{1:T}) \mathbf{W}_E(q)^\top \right] \\ &\approx \sum_{k=1}^{N} \left(\frac{1-\alpha}{N} - \frac{1-p_t}{N} \right) \underbrace{\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} \mathbf{W}_U(k)^\top \mathbf{W}_V x_t \cdot (x_t - \bar{x}_{1:T}) \mathbf{W}_E(q)^\top \right]}_{\triangleq A} \\ &\quad + (\alpha - p_t) \underbrace{\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} (\mathbf{W}_U(N+1)^\top \mathbf{W}_V x_t) \cdot (x_t - \bar{x}_{1:T}) \mathbf{W}_E(q)^\top \right]}_{\triangleq B}. \end{aligned} \tag{4.7}$$

Then, we have

$$\begin{aligned}
\mathbf{W}_E(N+1)^\top B\mathbf{W}_E(q) &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathbf{W}_U(N+1)^\top\mathbf{W}_V x_t)\cdot\mathbf{W}_E(N+1)^\top(x_t-\bar{x}_{1:T})\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(c+c(\alpha-1)\cdot\mathbb{1}\{z_t=N+1\})\cdot\mathbf{W}_E(N+1)^\top(x_t-\bar{x}_{1:T})\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(c(\alpha-1)\cdot\mathbb{1}\{z_t=N+1\})\cdot\mathbf{W}_E(N+1)^\top(x_t-\bar{x}_{1:T})\right] \\
&= \frac{\alpha}{N}\cdot c(\alpha-1)(1-\frac{\alpha}{N}) = \Theta(\frac{1}{N^2}) < 0.
\end{aligned}$$

where (a) is from Eq.(4.6), (b) is due to $\bar{x}_{1:T} = \frac{1}{T}\sum_t x_t$ and note that $c = \Theta(\frac{1}{N})$.

Similarly, we also have

$$\begin{aligned}
\mathbf{W}_E(N+1)^\top A\mathbf{W}_E(q) &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathbf{W}_U(k)^\top\mathbf{W}_V x_t)\mathbf{W}_E(N+1)^\top\cdot(x_t-\bar{x}_{1:T})\right] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\Theta(\frac{1}{N^2})\cdot\mathbb{1}\{z_t=N+1\}\mathbf{W}_E(N+1)^\top\cdot(x_t-\bar{x}_{1:T})\right] = \Theta(\frac{1}{N^3}).
\end{aligned}$$

For any $k \leqslant N$, we have

$$\begin{aligned}
\mathbf{W}_E(k)^\top B\mathbf{W}_E(q) &= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathbf{W}_U(N+1)^\top\mathbf{W}_V x_t)\cdot\mathbf{W}_E(k)^\top(x_t-\bar{x}_{1:T})\right] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(c(\alpha-1)\cdot\mathbb{1}\{z_t=k\})\cdot\mathbf{W}_E(N+1)^\top(x_t-\bar{x}_{1:T})\right] \\
&= \frac{\alpha}{N}\cdot c(\alpha-1)(-\frac{1}{N}) = \Theta(\frac{1}{N^3}) > 0,
\end{aligned}$$

and

$$\mathbf{W}_E(k)^\top A \mathbf{W}_E(q) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathbf{W}_U(k)^\top \mathbf{W}_V x_t)\mathbf{W}_E(k)^\top \cdot (x_t - \bar{x}_{1:T})\right]$$
$$= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\Theta(\frac{1}{N^2}) \cdot \mathbb{1}\{z_t = N+1\}\mathbf{W}_E(k)^\top \cdot (x_t - \bar{x}_{1:T})\right] = \Theta(\frac{1}{N^4}).$$

Combining the above four esimation of projections of A and B with Eq.(4.7), we have

$$\mathbf{W}_E(N+1)^\top(-\nabla_{\mathbf{W}_{QK}}L)\mathbf{W}_E(q) = \Theta(\frac{1}{N^2}) < 0,$$
$$\forall\, k \leqslant N,\ \ \mathbf{W}_E(k)^\top(-\nabla_{\mathbf{W}_{QK}}L)\mathbf{W}_E(q) = \Theta(\frac{1}{N^3}) > 0.$$

Then we have three observations

i. $\mathbf{W}_{QK}$ in this phase avoids the noise token $N+1$ and uniformly attends to all tokens $k \leqslant N$.

ii. The update of $\mathbf{W}_{QK}$ is in $\Theta(\frac{1}{N^2})$, while the update of $\mathbf{W}_F$ is $\Theta(1)$ in Lemma C.1 and that of $\mathbf{W}_V$ is $\Theta(\frac{1}{N})$ in Lemma C.2. These three levels of updating speed also coincide with the assumptions that $\mathbf{W}_F$ dominates first and then $\mathbf{W}_V$ has a micro structure that induces the evolving of $\mathbf{W}_{QK}$.

iii. The current proof for $\mathbf{W}_{QK}$ strongly depends on the fact that the noise token appears less than other token by a factor α in expectation. The proof will have the opposite result if the noise token is made to appear more by manipulating the data distribution. Therefore, we leave a new proof that is robust to such an assumption in data distribution as future work.

4.4.5 Multiple Triggers

In Section 4.4, we assume there is only one fixed trigger $q \in [N]$ for simplicity. Actually the case of multiple triggers has the same mechanism. As discussed by [Bietti et al. 2023] and

Appendix 4.4.4.2, for the case of only one trigger, the second-layer attention has large logits in $\langle \mathbf{W}_V^1 \mathbf{W}_E(i)^\top, \mathbf{W}_{QK}^2 \mathbf{W}_E(j) \rangle$ only for $i = j = q$. For the case of multiple triggers, basically $\langle \mathbf{W}_V^1 \mathbf{W}_E(i)^\top, \mathbf{W}_{QK}^2 \mathbf{W}_E(j) \rangle$ only have large values when $q \in Q$. This is verified in Figure 4.8.

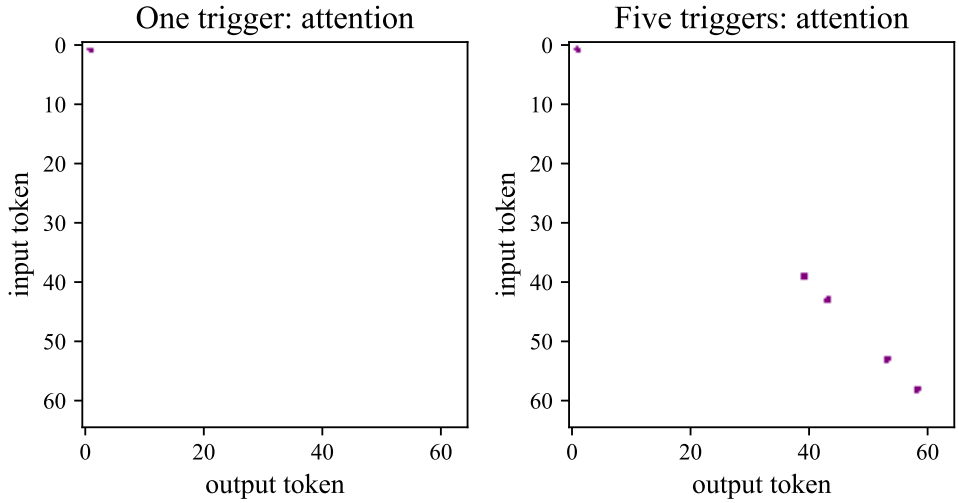


Figure 4.8: Logits of $\langle \mathbf{W}_V^1 \mathbf{W}_E(i)^\top, \mathbf{W}_{QK}^2 \mathbf{W}_E(j) \rangle$ for input i and output j when there is one trigger (left, $q = 1$) and five triggers (right, $q \in Q = \{1, 39, 43, 53, 58\}$). In both cases, the logits only have large values when $i = j = q$, verifies the matching mechanism in Appendix 4.4.4.2.

4.4.5.1 Architectural Choices

In Section 4.4 and Appendix 4.4.4.2, we were focused on experiments with both F_1, F_2 being two-layer ReLU MLPs. Meanwhile, we have also tried other choices of F_1, F_2 and then search for the best truncation method for each architecture. In this section, we would like to summarize our experimental results for better understanding of all modules in the two-layer transformer.

Generally, the feed-forward layer can be two-layer ReLU MLPs, one-layer Linear or "None", where None stands for there is no feed-forward layer so that the value matrices in attention layers are the only weight matrices that transform features.

Both F_1, F_2 are two-layer MLPs. This is our main setting. The best truncation method is to *fully* drop F_2. We also try to fully drop F_1, as reported in Figure 4.9. It turns out fully dropping F_1 makes the model predict the noise with high probability.

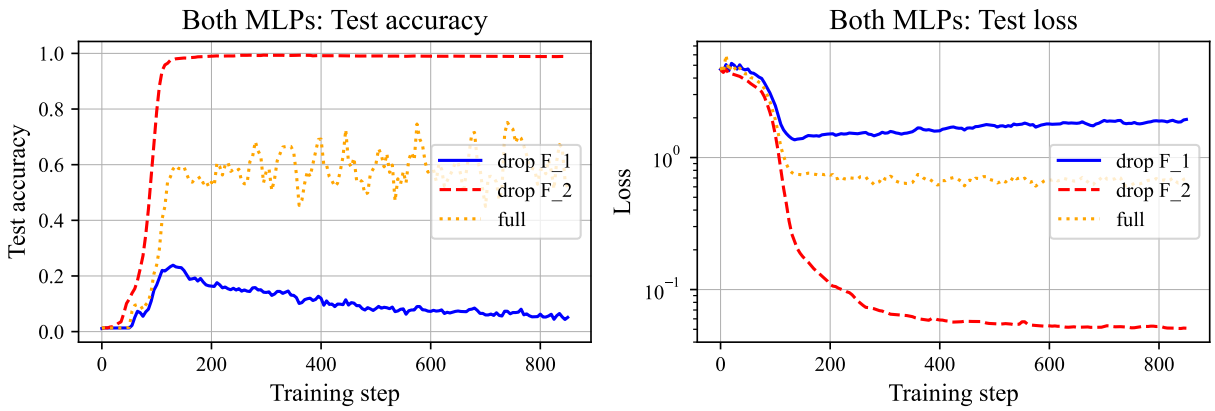


Figure 4.9: Test performance of fully dropping F_1, F_2 when both F_1, F_2 are two-layer MLPs. It turns out, while dropping F_2 makes the model predict correctly w.p. near 1, dropping F_1 has the model predict noise with high probability.

F_1 is MLPs and F_2 is Linear. Figure 4.10 reports the results. Dropping F_1 and F_2 both improve the correct prediction, and dropping F_1 is better with lower test loss. Note that, when test accuracies are near 100%, lower test loss is a better measurement of the prediction quality, because accuracies are taken by argmax over the output logits while test loss are about the exactly predicted probability.

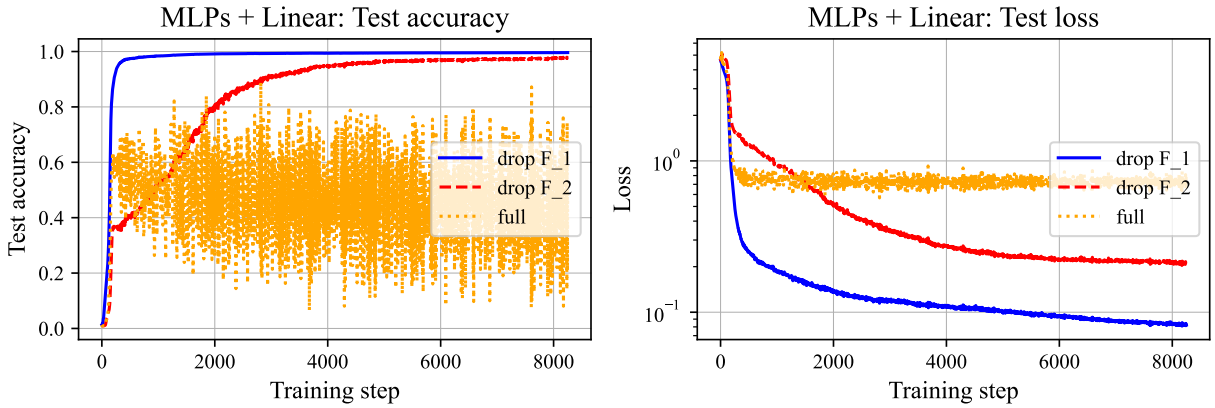


Figure 4.10: Test performance of fully dropping F_1, F_2 when both F_1 is MLPs and F_2 Linear. Both dropping methods turn out to help predict more correctly than the full model. Meanwhile, dropping the MLP F_1 is better with lower test loss.

F_1 is Linear and F_2 is MLPs. Figure 4.11 reports the results. Dropping F_2 improves the

correct prediction while dropping F_1 makes the model predict noise more.

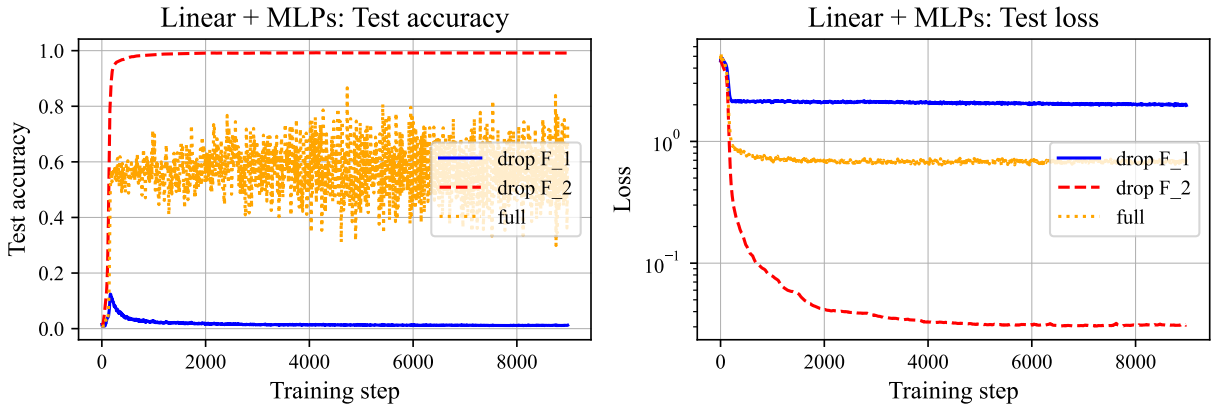


Figure 4.11: Test performance of fully dropping F_1, F_2 when both F_1 is Linear and F_2 MLPs. Only dropping F_2 helps predict more correctly. Dropping F_1 makes the model predicting noise more.

Both F_1 and F_2 are None. Figure 4.12 reports the results. While there is no feed-forward layer any more, low-rank truncating a part $\mathbf{W}_O^1$ of the first-layer matrix improves the model's prediction a little. This implies that, when there is not feed-forward layers, the noise association is possible stored in the first-layer value matrix of attention. Note that the improvement of such low-rank truncation is clearly smaller than *fully* dropping one of feed-forward layers in the previous cases. Meanwhile, a smaller $\rho = 0.01$ destroys the model's performance. This implies fully dropping is not the optimal choice for low-rank truncation of the value matrix, and there is low-rank subspace in it that is useful for predicting the correct tokens. Our discussion of the role of $\mathbf{W}_V^1$ in Appendix 4.4.4.2 is a possible answer to this phenomena.

4.4.5.2 Training Details about Experiments

All of the training is with SGD optimization with learning rate in $\{0.001, 0.03\}$. The batch size is 512. The dimension is 256. The context length is 256. All results in the experiments are stable for any learning rate between 0.001 and 0.03. Each run of experiments is on a single Nvidia Tesla V100 GPU. It takes 3 hours to finish each run for 2K steps, which probably can be optimized a

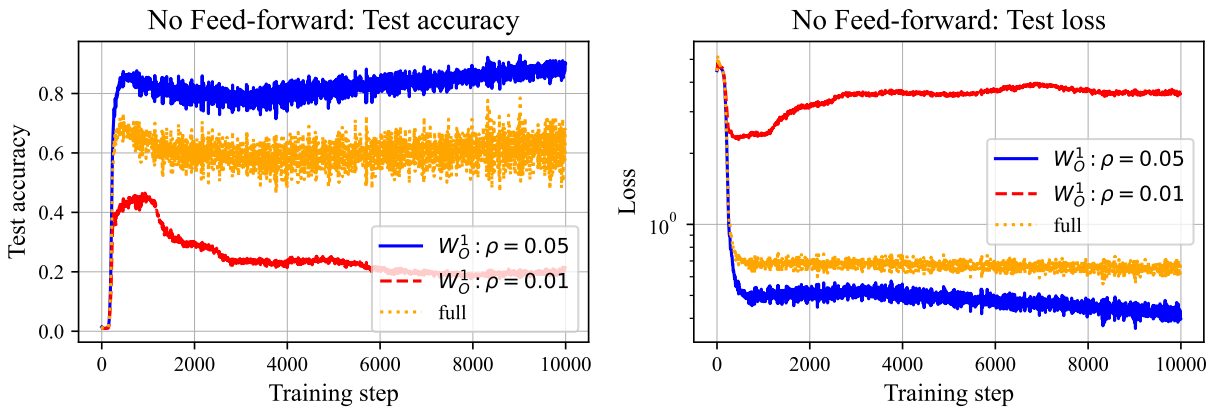


Figure 4.12: Test performance of low-rank truncating of $\mathbf{W}_O^1$ when there is no F_1, F_2. Here ρ is the fraction of preserved rank of $\mathbf{W}_O^1$, where actually we re-parametrize the first-layer value matrix in attention as $\mathbf{W}_O^1\mathbf{W}_V^1 \in \mathbb{R}^{d\times d}$. It turns out the best $\rho = 0.05$ improves the model's prediction a little. Meanwhile, a smaller ρ destroys the model's performance.

lot since we are tracking a lot of measurement along training, not limited to hundreds of possible truncations at each test time.

4.5 Experiments on Pre-trained LLMs

In this section, we empirically investigate how LLMs process distributional vs in-context associations, and how this evolves during training. Meanwhile, we provide numerical results of how much low-rank truncation improves complex reasoning on a real-world reasoning benchmark, GSM8K. Appendix C.8 provides another synethetic IOI dataset that requires counting tokens.

4.5.1 An Investigation on GPT-2 Small and Pythia Models

We consider GPT-2 small and Pythia models on the indirect object identification (IOI) and factual recall tasks described in Section 4.3.1.

Quick demonstration: IOI on GPT2 Small. Different from [Wang et al. 2022], we would like to consider whether a model proposes an output beyond the input x. A quick demonstration

is to consider the IOI task with input x ="When Mary and John went to a store, John gave a drink to"[1]. The top 4 predicted tokens for GPT-2 Small [Radford et al. 2019] on x are ["Mary", "them", "the", "John"]. Although GPT-2 Small successfully predicts Mary (the IO target) instead of John (S), the other two top candidate tokens, *i.e.*, "them" and "the", do not even appear in the context. This prominence of such "generic" words is similar to the factual recall example from Section 4.3.2, and plausibly follows from a distributional associative mechanism conditioned on the preposition "to".

Comprehensive experiment: IOI on Pythia-1B. Now we would like to verify this observation on more models and, more comprehensively, track the behavior of these models along training. We choose to conduct the IOI experiments on Pythia [Biderman et al. 2023], a family of models ranging in sizes from 14M to 12B trained on web data, with hundreds of training checkpoints for each size. We generate an IOI dataset of 100 sentences with random names for [IO] and [S] in each sample. Figure 4.13 reports the test results of Pythia-1B along training. Here LASER is conducted on MLP weights, with parameters given in Appendix C.1.2. LASER boosts the probability ratio of [IO] over "the" from 2.3× to 12.3× at 14K steps.

Factual recall on Pythia-1B. As in Table 4.1, we verify factual recall with input as "Madrid is located in". The full model of Pythia-1B generates "Madrid is located in the north of Spain", while the model after LASER generates "Madrid is located in Spain". We track the probability of predicting "Spain" and "the" along training in Figure 4.13. LASER turns out to boost the probability ratio of "Spain" over "the" from 0.16× to 11.3× at 14K steps. We note that better prompting could avoid the need for LASER in this case (e.g., "Madrid is located in the country of" predicts "Spain"), but increases the context length and thus the inference cost, though this is outside the scope of this paper.

Training dynamics on Pythia. The behavior of the Pythia models on the IOI and factual

[1]Note that here we use "a" store instead of "the" store in the original example of [Wang et al. 2022]. The reason is to rule out the word "the" from the input context.

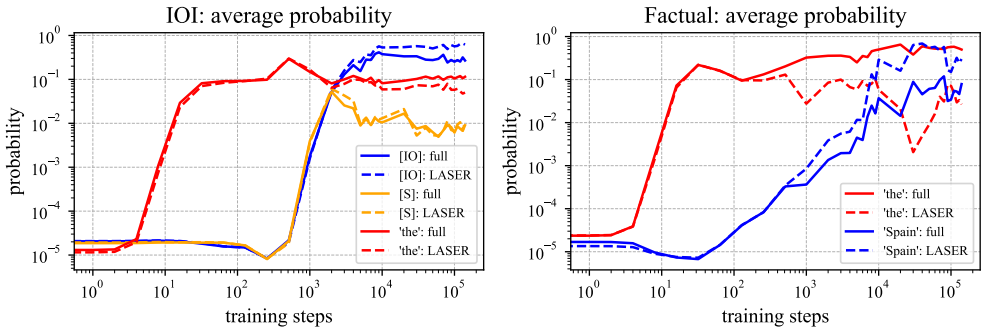


Figure 4.13: Left: average probability of tokens [IO], [S] and "the" in 100-sentence IOI task in the prediction by Pythia-1B along training. **Right**: average probability of tokens "Spain" and "the" in a factual task predicted by Pythia-1B along training, with input as "Madrid is located in". In both tasks, the full model learns to predict "the" with high probability starting from ~10 steps, and then learns to solve the tasks. LASER boosts the probability of correct answers against "the" in both tasks: the average probability ratio of correct answers against "the" improves from 2.3× to 12.3× (in IOI) and from 0.16× to 11.3× (in factual) at 14K steps.

recall tasks during their pre-training process displays several phases, as shown in Figure 4.13. For IOI, we observe:

i. Initialization: all tokens have similar logits since the weights are random initialized.

ii. Between 10 and 1000 steps: the models consistently output "the". They cannot solve IOI task at all, as long as they have almost the same prediction for [IO] and [S]. After 500 steps, [IO] starts the growth towards one of the top predictions.

iii. After 2000 steps: Pythia starts to be able to solve IOI task by preferring [IO] than [S] and "the". Meanwhile, the benefit of LASER appears as enhancing the leading position of [IO].

Therefore, the training process reveals the capacity of predicting "the" is learnt much earlier than predicting [IO]. The reason might be that predicting "the" requires a simpler grammar structure, while predicting [IO] requires a complicated architecture of attention heads of different roles across layer [Wang et al. 2022]. Then we note that the IOI task always has "to" before the masked [IO], which means "to" may be an indicator for the model to predict "the" with non-negligible probability. Similarly, for factual recall we see early learning of the "generic" answer, while the factual answer is learned later. Conceptually, if LLMs are able to write natural text or have been trained sufficiently with natural texts, it is not surprising for the model to predict "the" with high

probability after seeing "to". This is verified in Appendix C.1.1.

4.5.2 The effect of truncating feed-forward layers on GSM8K

As our previous examples of in-context reasoning tasks are too simple for real-world reasoning, we verify whether truncating MLPs improves reasoning on the GSM8K benchmark [Cobbe et al. 2021]. As shown in Table 4.2, LASER improves the few-shot Chain-of-Thought [Wei et al. 2022] reasoning performance on GSM8K when only using 1 or 2 shots, although the performance is worse in the standard 8-shot setting. This suggests that truncating MLPs may help promote in-context reasoning even in more complex settings, perhaps by removing spurious distributional associations.

Table 4.2: Few-shot accuracy (%) of pretrained and finetuned language models on GSM8K. Truncating MLPs (LASER) improves reasoning performances in few-shot CoT settings while it has worse performance in the standard 8-shot setting. The LASER hyper-parameters are in Appendix C.1.2.

	1-shot	2-shot	4-shot	8-shot (standard)
Phi-3 [Abdin et al. 2024]	56.0	72.2	78.2	82.7
Phi-3 + LASER	**66.1**	**74.4**	77.0	82.3
Llama-3.1-8B [AI@Meta 2024]	44.7	50.0	57.6	56.0
Llama-3.1-8B + LASER	**46.1**	**50.7**	55.9	53.8
Llama-3.1-8B-Instruct [AI@Meta 2024]	72.6	74.7	78.5	79.7
Llama-3.1-8B-Instruct + LASER	**73.6**	**75.6**	77.7	77.0

4.6 Discussion and Limitations

In this chapter, we studied the questions of how transformer language models learn to process distributional associations differently than in-context inputs, and how truncating specific weights or layers, particularly feed-forward layers, can help in-context reasoning. While our work provides some initial theoretical understanding of how this may arise on simple controlled settings,

it would be interesting to study how these ideas may extend to more complex tasks where in-context reasoning and distributional knowledge interact in more intricate ways.

Our contribution focuses on understanding the different roles of attention and FF weights in disentangling distributional vs in-context associations, both empirically and theoretically. The application of low-rank truncation is simply a way to verify our claims, and is consistent with the findings in the LASER paper that truncating some FF layers may improve performance on some reasoning tasks.

Nevertheless, our perspective based on distributional associations versus in-context reasoning may be helpful in thinking about how to allocate parameters to feed-forward versus attention layers: for instance, in Figure 4.14 on our synthetic task, we found that for a fixed total parameter budget, models with fewer MLP parameters achieve higher loss on distributional predictions (e.g., non-contextual bigrams) compared to models with more MLP parameters (and fewer attention parameters). These notions may also provide a different way to reason about circuit discovery in mechanistic interpretability from the perspective of training dynamics and properties of the training data. Finally, this disentanglement may inform more effective ways to fine-tune models, e.g., by selectively choosing which layers to fine-tune.

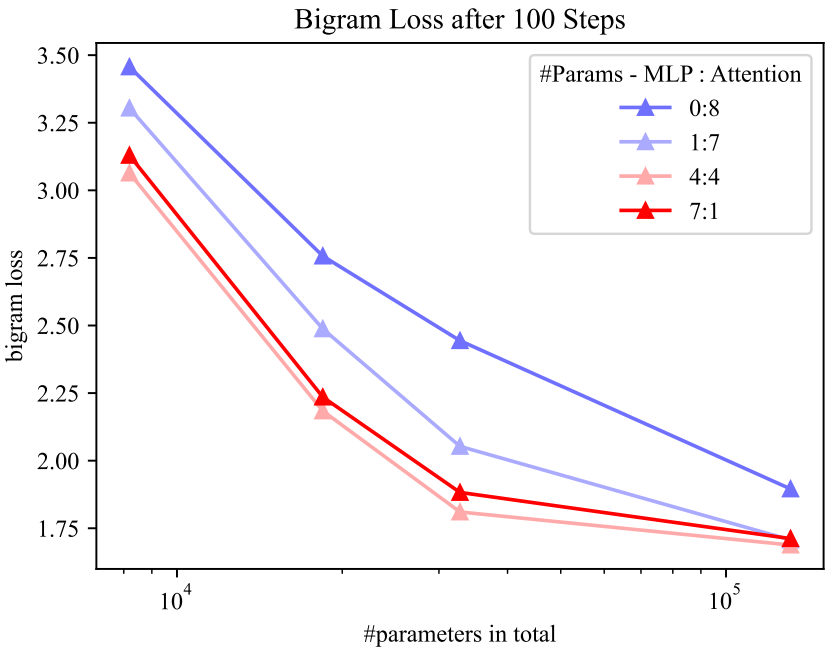


Figure 4.14: The training loss of approximating the global bigram π_b with various allocations of parameters in MLP and Attentions. For each configuration of total parameters and ratios, we use the corresponding best learning rate after search to train 100 steps.

A | Appendix: Supplementary Materials for Chapter 2

A.1 GA-MLP with general equivariant graph operators for node feature augmentation

For a graph $G = (V, E)$ with n nodes, assume without loss of generality that $V = [n]$. Let $\mathbb{S}_n$ denote the set of permutations of n, and $\forall \pi \in \mathbb{S}_n$, it maps a node $i \in [n]$ to $\pi(i) \in [n]$. For $\pi \in \mathbb{S}_n$ and a matrix $M \in \mathbb{R}^{n \times n}$, we use $\pi \star M \in \mathbb{R}^{n \times n}$ to denote the π-permuted version of M, that is, $(\pi \star M)_{i,j} = M_{\pi(i),\pi(j)}$. For $\pi \in \mathbb{S}_n$ and a matrix $Z \in \mathbb{R}^{n \times d}$, we use $\pi \star Z \in \mathbb{R}^{n \times d}$ to denote the π-permuted version of M, that is, $(\pi \star Z)_{i,p} = Z_{\pi(i),p}$.

Below, we define a more general form of GA-MLP models that extend the use of equivariant linear operators for node feature propagation to that of general equivariant graph operators. We first define a map $\omega : \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d'}$, whose first input argument is always the adjacency matrix of a graph, A, and second input argument is a node feature matrix. We say the map satisfies *equivariance* to node permutations if $\forall \pi \in \mathbb{S}_n$, $\forall Z \in \mathbb{R}^{n \times d}$, there is $\omega(\pi \star A, \pi \star Z) = \pi \star \omega(A, Z)$. With a slight abuse of notations, we also use $\omega[A](Z)$ to denote $\omega(A, Z)$, thereby considering $\omega[A] : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d'}$ as an operator on node features. If ω satisfies equivariance to node permutations as defined above, we then call $\omega[A]$ an equivariant graph operator. We can

then define a general (nonlinear) GA-MLP model as

$$\tilde{X} = \omega[A](X)$$
$$Z = \rho(\tilde{X}) \tag{A.1}$$

where ω is an equivariant graph operator, and ρ is a node-wise function.

It is easy to see that

Proposition A.1. *If $\omega[A](X) = m(A) \cdot X$, where $m(\cdot) = \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is an entry-wise function or matrix product or compositions thereof, then $\omega[A]$ is an equivariant graph operator.*

A.1.1 Extending the proof of Proposition 2.6 and 2.7 to general GA-MLPs

An extension of the first half of Proposition 2.6 is

Proposition A.2. *If $\omega[A]$ is an equivariant graph operator, then there exist exponentially-in-K many equivalence classes in $\mathcal{E}$ induced by the general GA-MLPs with $\omega[A]$, each of which intersects with doubly-exponentially-in-K many equivalence classes in $\mathcal{E}$ induced by depth-K GNNs, assuming that $|\mathcal{X}| \geq 2$ and $m \geq 3$.*

Proof: Similar to the proof of Proposition 2.6 given in Appendix A.8, we consider the set of full m-ary rooted trees of depth K, $\mathcal{T}_{m,K,\mathcal{X}}$, that is all rooted trees of depth K in which the nodes have features belonging to the discrete set $\mathcal{X} \subseteq \mathbb{N}$ and all non-leaf nodes have m children. $\mathcal{T}_{m,K,\mathcal{X}}$ is a subset of $\mathcal{E}$, the space of all rooted graphs. Suppose f is a function represented by a general GA-MLP defined in (A.1) with an equivariant graph operator $\omega[A]$. Let V_k denote the set of nodes at depth k of T. Notice the following symmetry among nodes in each V_k: if π is the permutation of a pair of nodes in some V_k for $1 \leq k \leq K$, then $\pi \star A = A$. By the equivariance property of ω,

this implies that

$$\begin{aligned}\omega[A](\pi \star Z) =& \omega[\pi \star A](\pi \star Z) \\ =& \pi \star \omega[A](Z)\end{aligned} \tag{A.2}$$

Let X denote the node feature matrix associated with T, and $\pi \star T$ denote the rooted tree in $\mathcal{T}_{m,K,\mathcal{X}}$ with the same topology (i.e., also a full m-ary rooted tree) but node feature matrix $\pi \star X$. Then, since the root node is not permuted under π, we know that

$$\begin{aligned}f(T) =& \rho\left(\omega[A](X)_{1,:}\right) \\ =& \rho\left((\pi \star \omega[A](X))_{1,:}\right) \\ =& \rho\left(\omega[A](\pi \star X)_{1,:}\right) \\ =& f(\pi \star T)\end{aligned} \tag{A.3}$$

This implies that for two trees T and $T' \in \mathcal{T}_{m,K,\mathcal{X}}$, if $\forall 0 \leq k \leq K, \forall x \in \mathcal{X}$, they satisfy $|\overline{\mathcal{W}}_k(T; x)| = |\overline{\mathcal{W}}_k(T'; x)|$, then $f(T) = f(T')$ for all such f's, and hence T and T' belong to the same equivalence class in $\mathcal{E}$ induced by GA-MLPs. Therefore, by the rest of the argument given in Proposition 2.6, Proposition A.2 can be proven analogously for GA-MLPs with general equivariant graph operators. □

Similarly, Proposition 2.7 can also be extended to

Proposition A.3. *For any sequence of node features $\{x_k\}_{k\in\mathbb{N}_+} \subseteq \mathcal{X}$, consider the sequence of functions $f_k(G^{[i]}) := |\mathcal{W}_k(G^{[i]}; (x_1, ..., x_k))|$ on $\mathcal{E}$. For all $k \in \mathbb{N}_+$, the image under f_k of every equivalence class in $\mathcal{E}$ induced by depth-k GNNs contains a single value, while for any GA-MLP using equivariant graph operators, there exist exponentially-in-k many equivalence classes in $\mathcal{E}$ induced by this GA-MLP whose image under f_k contains exponentially-in-k many values.*

The proof replies on the same extension as described above in the proof of Proposition A.2.

A.2 Examples of existing GA-MLP models

For $\epsilon \in \mathbb{R}$, let $\bar{A}_{(\epsilon)} = A + \epsilon I$, $\bar{D}_{(\epsilon)}$ be the diagonal matrix with $\bar{D}_{(\epsilon),ii} = \sum_j A_{ij} + \epsilon$, and $\tilde{A}_{(\epsilon)} = \bar{D}_{(\epsilon)}^{-1/2} \bar{A}_{(\epsilon)} \bar{D}_{(\epsilon)}^{-1/2}$.

- Simple Graph Convolution [Wu et al. 2019]:

  $\Omega(A) = \{(\tilde{A}_{(1)})^K\}$ for some $K > 0$. In addition, φ is the identity function and $\rho(H) = softmax(HW)$ for some trainable weight matrix W.

- Graph Feature Network [Chen et al. 2019a]:

  $\Omega(A) = \{I, D, \tilde{A}_{(\epsilon)}, ..., (\tilde{A}_{(\epsilon)})^K\}$ for some $K > 0$ and $\epsilon > 0$. In addition, φ is the identity function and ρ is an MLP.

- Scalable Inception Graph Networks [Rossi et al. 2020]:

  $\Omega(A) = \{I\} \cup \Omega_1(A) \cup \Omega_2(A) \cup \Omega_3(A)$, where $\Omega_1(A)$ is a family of simple / normalized adjacency matrices, $\Omega_2(A)$ is a family of Personalized-PageRank-based adjacency matrices, and $\Omega_3(A)$ is a family of triangle-based adjacency matrices. In addition, writing $\tilde{X} = [\tilde{X}_1, ..., \tilde{X}_K]$, there is $Z = \rho(\tilde{X}) = \sigma_1(\sigma_2([\tilde{X}_1 W_1, ..., \tilde{X}_K W_K]) W_{\text{out}})$, with σ_1 and σ_2 being nonlinear activation functions and $W_1, ..., W_K$ and W_{out} being trainable weight matrices of suitable dimensions.

A.3 Equivalence classes induced by GNNs and GA-MLPs among real graphs

Given a space of graphs, $\mathcal{G}$, and a family $\mathcal{F}$ of functions mapping $\mathcal{G}$ to $\mathbb{R}$, $\mathcal{F}$ induces an equivalence relation that we denote by $\simeq_{\mathcal{G};\mathcal{F}}$ among graphs in $\mathcal{G}$ such that for $G_1, G_2 \in \mathcal{G}$, $G_1 \simeq_{\mathcal{G};\mathcal{F}} G_2$ if and only if $\forall f \in \mathcal{F}$, $f(G_1) = f(G_2)$. For example, if $\mathcal{F}$ is powerful enough to distinguish all pairs of non-isomorphic graphs, then each equivalence class under $\simeq_{\mathcal{G},\mathcal{F}}$ contains exactly one graph.

	IMDBBINARY		IMDBMULTI		REDDITBINARY		REDDITMULTI5K		COLLAB	
# Graphs	1000		1500		2000		5000		5000	
K	GNN	GA-MLP	GNN	GA-MLP	GNN	GA-MLP	GNN	GA-MLP	GNN	GA-MLP
1	51	51	49	49	781	781	1365	1365	294	294
2	537	537	387	387	1998	1998	4999	4999	4080	4080
3	537	537	387	387	1998	1998	4999	4999	4080	4080
ground truth	537		387		1998		4999		4080	

Table A.1: The number of equivalence classes of graphs induced by GNN and GA-MLP on real datasets with node features removed. The last row gives the ground-truth number of isomorphism classes of graphs computed from the implementation of [Ivanov et al. 2019].

Thus, by examining the number or sizes of the equivalence classes induced by different families of functions on $\mathcal{G}$, we can evaluate their relative expressive power in a quantitative way.

Hence, we supplement the theoretical result of Proposition 2.2 with the following numerical results on five real-world datasets for graph-predictions. For graphs in each of the two real datasets, we remove their node features and count the total number of equivalence classes among them induced by depth-K GNNs (equivalent to K-iterations of the WL test, as discussed in Section 2.3.2) as well as GA-MLPs with $\Omega = \{I, A, ..., A^K\}$ for different K's. We see from the results in Table A.1 that as soon as $K \geq 2$, the number of equivalence classes induced by GNNs and the GA-MLPs are both close to the total number of graphs up to isomorphism, implying that they are indeed both able to distinguish almost all pairs of non-isomorphic graphs among the ones occurring in these datasets.

A.4 Additional notations

For any $k \in \mathbb{N}_+$ and any rooted graph $G^{[i]} = (V, E, i) \in \mathcal{E}$, define

$$\mathcal{W}_k(G^{[i]}) = \{(i_1, ..., i_k) \subseteq V : A_{i,i_1}, A_{i_1,i_2}, ..., A_{i_{k-1},i_k} > 0\} \tag{A.4}$$

$$\overline{\mathcal{W}}_k(G^{[i]}) = \{(i_1, ..., i_k) \in \mathcal{W}_k(G^{[i]}) : i \neq i_2, i_1 \neq i_3, ..., i_{k-3} \neq i_{k-1}, i_{k-2} \neq i_k\} \tag{A.5}$$

as the sets of *walks* and *non-backtracking walks* of length k in $G^{[i]}$ starting from the root node, respectively. Note that when $G^{[i]}$ is a rooted tree, a non-backtracking walk of length k is a path from the root node to a node at depth k. In addition, for $0 \leq d_1, ..., d_k \leq m$ and $x_1, ..., x_k \in \mathcal{X}$, define the following subsets of $\mathcal{W}_k(G^{[i]})$:

$$\mathcal{W}_k\left(G^{[i]}; (d_1, ..., d_k), x_k\right) = \{(i_1, ..., i_k) \in \mathcal{W}_k(G^{[i]}) : \{d_{i_1}, ..., d_{i_k}\}_m = \{d_1, ..., d_k\}_m, X_{i_k} = x_k\} \quad \text{(A.6)}$$

$$\mathcal{W}_k\left(G^{[i]}; (x_1, ..., x_k)\right) = \{(i_1, ..., i_k) \in \mathcal{W}_k(G^{[i]}) : (X_{i_1}, ..., X_{i_k}) = (x_1, ..., x_k)\} \quad \text{(A.7)}$$

$$\mathcal{W}_k(G^{[i]}; x_k) = \{(i_1, ..., i_k) \in \mathcal{W}_k(G^{[i]}) : X_{i_k} = x_k\} \quad \text{(A.8)}$$

We also define $\overline{\mathcal{W}}_k\left(G^{[i]}; (d_1, ..., d_k), x_k\right)$, $\overline{\mathcal{W}}_k\left(G^{[i]}; (x_1, ..., x_k)\right)$ and $\overline{\mathcal{W}}_k(G^{[i]}; x_k)$ similarly.

A.5 Proof of Proposition 2.2

With node features being identical in the random graphs, we take $X \in \mathbb{R}^{n\times 1}$ to be the all-1 vector. Thus,

$$(DX)_i = d_i\,, \quad \text{(A.9)}$$

and

$$(AD^{-\alpha}X)_i = \sum_{j\in\mathcal{N}(i)} d_j^{-\alpha}\,. \quad \text{(A.10)}$$

Since (2.4) and (2.2) together can approximate arbitrary permutation-invariant functions on multisets [Zaheer et al. 2017], if two graphs $G = (V, E)$ and $G' = (V', E')$ cannot be distinguished by the GA-MLP with an operator family Ω that includes $\{D, AD^{-\alpha}\}$ under any choice of its parameters, it means that the two multisets $\{(d_i, \sum_{j\in\mathcal{N}(i)} d_j^{-\alpha}) : i \in V\}_m = \{(d_{i'}, \sum_{j'\in\mathcal{N}(i')} d_{j'}^{-\alpha}) : i' \in V'\}_m$,

and therefore both of the following hold:

$$\{d_i : i \in V\}_m = \{d_{i'} : i' \in V'\}_m \tag{A.11}$$

$$\{\sum_{j \in \mathcal{N}(i)} d_j^{-\alpha} : i \in V\}_m = \{\sum_{j' \in \mathcal{N}(i')} d_{j'}^{-\alpha} : i' \in V'\}_m \tag{A.12}$$

To see what this means, we need the two following lemmas.

Lemma A.4. *Let $\mathcal{S}_n$ be the set of all multisets consisting of at most n elements, all of which are integers between 0 and n. Consider the function $h_\alpha(S) := \sum_{u \in S} u^{-\alpha}$ defined for multisets S. If $\alpha > \frac{\log n}{\log n - \log(n-1)}$, h_α is an injective function on $\mathcal{S}_n$.*

Proof of Lemma A.4: For h_α to be injective on $\mathcal{S}_n$, it suffices to require that $\forall l \leq n-1$, there is $l^{-\alpha} > n(l+1)^{-\alpha}$, for which it is sufficient to require that $(n-1)^{-\alpha} > n^{-\alpha+1}$, or $\alpha > \frac{\log n}{\log n - \log(n-1)}$. □

Lemma A.5 ([Babai et al. 1980], Theorem 1)**.** *Consider the space of graphs with n vertices, $\mathcal{G}_n$. There is a subset $\mathcal{K}_n \subseteq \mathcal{G}_n$ that contains almost all such graphs (i.e. the fraction converges to 1 as $n \to \infty$) such that the following algorithm yields a unique identifier for every graph $G = (V, E) \in \mathcal{K}_n$:*

Algorithm 1: Set $r = [3 \log n / \log 2]$, and let $\bar{d}(G)$ be the degree of the node in V with the rth largest degree; For each node i in G, define the multiset $\gamma_i = \{d_j : j \in \mathcal{N}(i), d_j > \bar{d}(G)\}_m$; Finally define a multiset associated with G, $F(G) = \{\gamma_i : i \in V\}_m$, which is the output of the algorithm.

In other words, $\forall G, G' \in \mathcal{K}_n$, G and G' are isomorphic if and only if $F(G) = F(G')$ as multisets. In particular, we can choose $\mathcal{K}_n$ such that the top r node degrees of every graph in $\mathcal{K}_n$ are distinct.

Based on these lemmas, we will show that when $\alpha > \frac{\log n}{\log n - \log(n-1)}$ and for $G, G' \in \mathcal{K}_n$, (A.11) and (A.12) together imply that G is isomorphic to G'. To see this, suppose that (A.11) and (A.12) hold. Because of (A.11), we know that G and G' share the same degree sequence, and hence $\bar{d}(G) = \bar{d}(G')$. Because of (A.12), we know that there is a bijective map σ from V to V' such that

$\forall i \in V$,

$$\sum_{j \in \mathcal{N}(i)} d_j^{-\alpha} = \sum_{j' \in \mathcal{N}(i')} d_{j'}^{-\alpha}, \tag{A.13}$$

which, by Lemma A.4, implies that $\{d_j : j \in \mathcal{N}(i)\}_m = \{d_{j'} : j' \in \mathcal{N}(i')\}_m$. We then have $\gamma_i = \{d_j : j \in \mathcal{N}(i)\}_m = \{d_j : j \in \mathcal{N}(i)\}_m \cap (\bar{d}(G), \infty) = \{d_{j'} : j' \in \mathcal{N}(i')\}_m \cap (\bar{d}(G'), \infty) = \gamma_{i'}$, and therefore $F(G) = F(G')$, which implies that G and G' are isomorphic by Lemma A.5. This shows a contradiction. Therefore, if $G, G' \in \mathcal{K}_n$ are not isomorphic, then it cannot be the case that both (A.11) and (A.12) hold, and hence there exists a choice of parameters for the GA-MLP with $\{D, AD^{-\alpha}\} \subseteq \Omega$ that makes it return different outputs when applied to G and G'. This proves Proposition 2.2. □

A.6 Proof of Proposition 2.3

As argued in the main text, to estimate the number of equivalence classes on $\mathcal{E}$ induced by GNNs, we need to estimate the number of possible rooted aggregation trees. In particular, to lower-bound the number of equivalence classes on $\mathcal{E}$ induced by GNNs, we only need to focus on a subset of all possible rooted aggregation trees, namely those in which every node has exactly m children. Letting $\mathcal{T}^{\mathrm{A}}_{m,K,\mathcal{X}}$ denote the set of all rooted aggregation trees of depth K in which each non-leaf node has degree exactly m and the node features belong to $\mathcal{X}$, we will first prove the following lemma:

Lemma A.6. *If* $|\mathcal{X}| \geq 2$, *then* $|\mathcal{T}^{\mathrm{A}}_{m,K,\mathcal{X}}| \geq (m-1)^{(2^K-1)}$.

Note that a rooted *aggregation* tree needs to satisfy the constraint that each of its node must have its parent's feature equal to one of its children's feature, and so this lower bound is not as straightforward to prove as lower-bounding the total number of rooted subtrees. As argued above, this will allow us to derive Proposition 2.3.

Proof of Lemma A.6: Define $\mathcal{B} := \{0, 1\}$. Since $|\mathcal{X}| \geq 2$, we assume without loss of generality that

$\mathcal{B} \subseteq \mathcal{X}$. To prove a lower-bound on the cardinality of $\mathcal{T}^{\mathrm{A}}_{m,K,\mathcal{X}}$, it suffices to restrict our attention to its subset, $\mathcal{T}^{\mathrm{A}}_{m,K} := \mathcal{T}^{\mathrm{A}}_{m,K,\mathcal{B}}$, where all nodes have feature either 0 or 1. Furthermore, it is sufficient to restrict our attention to the subset of $\mathcal{T}^{\mathrm{A}}_{m,K}$ which contain all 2^K possible types of paths of length K from the root to the leaves. Formally, with $\overline{\mathcal{W}}_k$ defined as in Appendix A.4, we let

$$\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K} = \{T \in \mathcal{T}^{\mathrm{A}}_{m,K} : \forall x_1, ..., x_K \in \mathcal{B}, \overline{\mathcal{W}}_K(T; (x_1, ..., x_K)) \geq 1\}\,, \tag{A.14}$$

and it is sufficient to prove a lower bound on the cardinality of $\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K}$. Define $\mathrm{P}_k = \{(x_1, ..., x_k) : x_1, ..., x_k \in \mathcal{B}\}$ to be the set of all binary k-tuples. By the definition of (A.14), we know that $\forall \tau \in \mathrm{P}_K, |\mathcal{W}_K(T; \tau)| \geq 1$. This means that $\forall \tau \in \mathrm{P}_K$, there exists at least one leaf node in T such that the path from the root node to this node consists of a sequence of nodes with features exactly as given by τ. We call any such node a *node under* τ.

We show such a lower bound on the cardinality of $\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K}$ inductively. For the base case, we know that $\tilde{\mathcal{T}}^{\mathrm{A}}_{m,1}$ consists of all binary-featured depth-1 rooted trees with at least 1 leaf node of feature 0 and 1 leaf node of feature 1, and hence $\tilde{\mathcal{T}}^{\mathrm{A}}_{m,1} = 2(m-1)$. Next, we consider the inductive step. For every $K \geq 1$ and every $T \in \tilde{\mathcal{T}}^{\mathrm{A}}_{m,K}$, we can generate rooted aggregation trees belonging to $T \in \tilde{\mathcal{T}}^{\mathrm{A}}_{m,K+1}$ by assigning children of feature 0 or 1 to the leaf nodes of T. First note that, from two non-isomorphic rooted aggregation trees T and $T' \in \tilde{\mathcal{T}}^{\mathrm{A}}_{m,K}$, we obtain non-isomorphic rooted aggregation trees in $\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K+1}$ in this way. Moreover, as we will show next, for every $T \in \tilde{\mathcal{T}}^{\mathrm{A}}_{m,K}$, we can lower-bound the number of distinct rooted aggregation trees belonging to $\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K+1}$ obtained from T in this way.

There are many choices to assign the children. To get a lower-bound on the cardinality of $\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K+1}$, we only need to consider a subset of these choices of assignments, namely, those that assign the same number of children with feature 0 to every node under the same $\tau \in \mathrm{P}_K$. Thus, we let $\bar{q}_{K+1,\tau}$ denote the number of children of feature 0 assigned to every node in τ. Due to the constraint that each node in the rooted aggregation tree must have its parent's feature equal to

one of its children's feature, not all choices of $\{\bar{q}_{K+1,\tau}\}_{\tau\in \mathrm{P}_K}$ lead to legitimate rooted aggregation trees. Nonetheless, when restricting to the choices where $\forall \tau \in \mathrm{P}_K, 1 \leq \bar{q}_{K+1,\tau} \leq m-1$, we see that every leaf node of T gets assigned at least one child of feature 0 and another child of feature 1, thereby satisfying the constraint above whether its parent has feature 0 or 1. Moreover, for such choices, the rooted aggregation tree of depth $K+1$ obtained in this way contains all 2^{K+1} possible paths of length $K+1$, and therefore belongs to $\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K+1}$. Hence, it remains to show a lower bound on how many distinct trees in $\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K+1}$ can be obtained in this way from each T. Since for $\tau, \tau' \in \mathrm{P}_K$ such that $\tau \neq \tau'$, a node under τ is distinguishable from a node under τ', we see that every legitimate choice of the tuple of 2^K integers, $(\bar{q}_{K+1,\tau})_{\tau\in \mathrm{P}_K}$, leads to a distinct rooted aggregation tree of depth $K+1$, and there are $(m-1)^{2^K}$ of these choices. Hence, we have derived that $|\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K+1}| \geq (m-1)^{2^K}|\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K}|$, and therefore $|\tilde{\mathcal{T}}^{\mathrm{A}}_{m,K}| \geq (m-1)^{\sum_{k=1}^{K} 2^k} = (m-1)^{2^K-1}$.

□

A.7 Proof of Proposition 2.4

According to the formula (2.3), by expanding the matrix product, we have

$$
\begin{aligned}
(\tilde{A}^k \varphi(X))_i =& \sum_{(i_1,\ldots,i_k)\in \mathcal{W}_k(G^{[i]})} d_i^{-\alpha} d_{i_1}^{-(\alpha+\beta)} \ldots d_{i_{k-1}}^{-(\alpha+\beta)} d_{i_k}^{-\beta} \varphi(X_{i_k}) \\
=& d_i^{-\alpha} \sum_{\substack{\{\bar{d}_1,\ldots,\bar{d}_{t-1}\}_m,\\ \bar{d}_k, x}} \sum_{\substack{(i,i_1,\ldots,i_k)\in \\ \mathcal{W}_k(G^{[i]};\{\bar{d}_1,\ldots,\bar{d}_{k-1}\}_m,\bar{d}_k,x)}} (\bar{d}_1 \ldots \bar{d}_{k-1})^{-(\alpha+\beta)} \bar{d}_k \varphi(x) \\
=& d_i^{-\alpha} \sum_{\substack{\{\bar{d}_1,\ldots,\bar{d}_{k-1}\}_m,\\ \bar{d}_k, x}} \left((\bar{d}_1 \ldots \bar{d}_{k-1})^{-(\alpha+\beta)} \bar{d}_k \varphi(x) \right) \left| \mathcal{W}_k(G^{[i]}; \{\bar{d}_1, \ldots, \bar{d}_{t-1}\}_m, \bar{d}_k, x) \right| , \qquad \text{(A.15)}
\end{aligned}
$$

with $\mathcal{W}_k(G^{[i]}; \{\bar{d}_1, \ldots, \bar{d}_{t-1}\}_m, \bar{d}_k, x)$ defined in Appendix A.4. Hence, for two different nodes i in G and i' in G' (G and G' can be the same graph), the node-wise outputs of the GA-MLP at i

and i' will be identical if the rooted graphs G^i and $G'^{[i']}$ satisfy $\mathcal{W}_k(G^{[i]}; \{\bar{d}_1, ..., \bar{d}_{k-1}\}_m, \bar{d}_k, x) = \mathcal{W}_k(G'^{[i']}; \{\bar{d}_1, ..., \bar{d}_{k-1}\}_m, \bar{d}_k, x)$ for every combination of choices on the multiset $\{\bar{d}_1, ..., \bar{d}_{k-1}\}_m$, the integer $\bar{d}_k$ and the node feature x, under the constraints of $\bar{d}_1, ..., \bar{d}_k \leq m$ and $x \in \mathcal{X}$. Note that there are at most $\binom{k+m-2}{m-1} \leq (k+m-2)^{m-1}$ possible choices of the multiset $\{\bar{d}_1, ..., \bar{d}_{k-1}\}_m$, m choices of $\bar{d}_k$ and $|\mathcal{X}|$ choices of x, thereby allowing at most $|\mathcal{X}|m(k+m-2)^{m-1}$ possible choices. Because of the constraint

$$\sum_{\substack{\{\bar{d}_1,...,\bar{d}_{k-1}\}_m, \\ \bar{d}_k, x}} \left|\mathcal{W}_k(G_K^{[i]}; \{\bar{d}_1, ..., \bar{d}_{t-1}\}_m, \bar{d}_k, x)\right| = |\mathcal{W}_k(G^{[i]})| \leq m^k \,, \tag{A.16}$$

We see that the total number of equivalence classes on $\mathcal{E}$ induced by such a GA-MLP is upper-bounded by $\binom{m^k+|\mathcal{X}|m(k+m-2)^{m-1}-1}{|\mathcal{X}|m(k+m-2)^{m-1}}$, which is on the order of $O(m^{k^m})$ with k growing and m bounded. Finally, since the total number of equivalence classes induced by multiple operators can be upper-bounded by the product of the number of equivalence classes induced by each operator separately, we derive the proposition as desired.

A.8 Proof of Proposition 2.6

Consider the set of full m-ary rooted trees of depth K, $\mathcal{T}_{m,K,\mathcal{X}}$, that is all rooted trees of depth K in which the nodes have features belonging to the discrete set $\mathcal{X} \subseteq \mathbb{N}$ and all non-leaf nodes have m children. $\mathcal{T}_{m,K,\mathcal{X}}$ is a subset of $\mathcal{E}$, the space of all rooted graphs. If f is a function represented by a GA-MLP using operators of at most K-hop, then for $T \in \mathcal{T}_{m,K,\mathcal{X}}$, we can write

$$f(T) = \rho(\sum_{j \in V} a_j X_j) \,, \tag{A.17}$$

where we denote the node set of T by V and the vectors a_j's depend only on the topological relationship between j and the root node. Let V_k denote the set of nodes at depth k of T. By the

assumption that the operators depend only on the graph topology, and thanks to the topological symmetry of such full m-ary trees among all nodes on the same depth, we have that $\forall 1 \leq k \leq K$ and $\forall j, j' \in V_k$, there is $a_j = a'_j =: a_{[k]}$. Thus, we can write

$$\begin{aligned} f(T) =& \rho\Big(\sum_{0\leq k\leq K}\sum_{j\in V_k} a_{[k]}\phi(X_j)\Big) \\ =& \rho\Big(\sum_{0\leq k\leq K}\sum_{x\in\mathcal{X}} \bar{a}_{[k],x}|\overline{\mathcal{W}}_k(T;x)|\Big) \end{aligned} \tag{A.18}$$

for some other set of coefficients $\bar{a}_{V_k,x}$'s, and where $\overline{\mathcal{W}}_k(T;x)$ is defined in Appendix A.4. In other words, for two trees T and $T' \in \mathcal{T}_{m,K,\mathcal{X}}$, if $\forall 0 \leq k \leq K, \forall x \in \mathcal{X}$, they satisfy $|\overline{\mathcal{W}}_k(T;x)| = |\overline{\mathcal{W}}_k(T';x)|$, then $f(T) = f(T')$ for all such f's, and hence T and T' belong to the same equivalence class in $\mathcal{E}$ induced by GA-MLPs. Thus, for a certain subset of these equivalence classes, we can lower-bound the number of equivalence classes in $\mathcal{E}$ induced by GNNs that they intersect by lower-bounding the number of distinct trees in $\mathcal{T}_{m,K,\mathcal{X}}$ that they contain, because GNNs are able to distinguish non-isomorphic rooted subtrees. In particular, as a lower-bound is sufficient, we restrict attention to the subset of those trees with node features either 0 or 1, that is, trees belonging to $\mathcal{T}_{m,K} := \mathcal{T}_{m,K,\mathcal{B}}$, with $\mathcal{B} := \{0, 1\}$.

In a rooted tree T, $\overline{\mathcal{W}}_k(T;x)$ gives the total number of nodes with feature x at depth k. For integers $q_0, q_1, ..., q_K$ such that $0 \leq q_k \leq m^k, \forall k \leq K$, define

$$\mathcal{T}_{m,K,(q_0,q_1,...,q_K)} = \{T \in \mathcal{T}_{m,K} : \forall k \leq K, |\overline{\mathcal{W}}_k(T;0)| = q_k\}\,, \tag{A.19}$$

that is, the subset of trees whose *per-level-node-counts*, $\{|\overline{\mathcal{W}}_k(T;0)|\}_{k\leq K}$ (and then, therefore, $\{|\overline{\mathcal{W}}_k(T;x)|\}_{k\leq K,x\in\mathcal{B}}$) are given by the tuple $(q_0, q_1, ..., q_k)$. From the argument above, all trees in the same $\mathcal{T}_{m,K,(q_0,q_1,...,q_K)}$ belong to the same equivalence class in $\mathcal{E}$ induced by GA-MLPs. On the other hand, every pair of non-isomorphic trees belong to different equivalence class in $\mathcal{E}$ induced by GNNs. Thus, to show Proposition 2.6, it is sufficient to find sufficiently many choices of

$(q_0, q_1, ..., q_K)$ such that $\mathcal{T}_{m,K,(q_0,q_1,...,q_K)}$ contains sufficiently many non-isomorphic trees. Specifically, we will show the following:

Lemma A.7. *For all integers $q_0, q_1, ..., q_K$ such that $\forall 2 \leq k \leq K$,*

$$2^k - 2^{k-2} \leq q_k \leq \frac{1}{2}m^k \,, \tag{A.20}$$

there is

$$|\mathcal{T}_{m,K,(q_0,q_1,...,q_K)}| \geq 2^{2^{K-1}-1} \tag{A.21}$$

Proof of Lemma A.7: To prove such a lower bound on the cardinality of $\mathcal{T}_{m,K,(q_0,q_1,...,q_K)}$, it is sufficient to prove a lower bound on the cardinality of its subset,

$$\tilde{\mathcal{T}}_{m,K,(q_0,q_1,...,q_K)} = \{T \in \mathcal{T}_{m,K,(q_0,q_1,...,q_K)} : \forall x_1, ..., x_K \in \mathcal{B}, \overline{\mathcal{W}}_K(T; (x_1, ..., x_K)) \geq 1\} \,. \tag{A.22}$$

A similar construction is involved in the proof of Lemma A.6 in Appendix A.6. Then, we will prove this lemma by induction on K. For the base cases, it is obvious that $|\tilde{\mathcal{T}}_{m,0,(0)}| = |\tilde{\mathcal{T}}_{m,0,(1)}| = 1$, and $|\tilde{\mathcal{T}}_{m,1,(0,0)}| = |\tilde{\mathcal{T}}_{m,1,(0,1)}| = |\tilde{\mathcal{T}}_{m,1,(0,2)}| = |\tilde{\mathcal{T}}_{m,1,(1,0)}| = |\tilde{\mathcal{T}}_{m,1,(1,1)}| = |\tilde{\mathcal{T}}_{m,1,(1,2)}| = 1$. We next prove the inductive hypothesis that, for $K \geq 2$ and when $q_0, q_1, ..., q_K$ satisfying (A.20), there is

$$|\tilde{\mathcal{T}}_{m,K,(q_0,q_1,...,q_K)}| \geq 2^{2^{K-2}} \cdot |\tilde{\mathcal{T}}_{m,K-1,(q_0,q_1,...,q_{K-1})}| \,. \tag{A.23}$$

To see this, we will next show that $\forall T \in \tilde{\mathcal{T}}_{m,K-1,(q_0,q_1,...,q_{K-1})}$, we can generate enough number of depth-K trees in $\tilde{\mathcal{T}}_{m,K,(q_0,q_1,...,q_K)}$ by appending children to the leaf nodes of T. Since any two depth-K trees generated from two non-isomorphic depth-$(K-1)$ trees in this way are non-isomorphic, this will allow us to lower-bound the total number of trees in $\tilde{\mathcal{T}}_{m,K,(q_0,q_1,...,q_K)}$.

Consider the set of binary k-tuples, $\mathrm{P}_k = \{(x_1, ..., x_k) : x_1, ..., x_k \in \mathcal{B}\}$, of cardinality 2^k. As $T \in \tilde{\mathcal{T}}_{m,K-1,(q_0,q_1,...,q_{K-1})}$, we know that $\forall \tau \in \mathrm{P}_{K-1}, |\overline{\mathcal{W}}_{K-1}(T; \tau)| \geq 1$. This means that $\forall \tau \in$

P_k, there exists at least one leaf node in T such that the path from the root node to this node consists of a sequence of nodes with features given by τ. We call any such node a *node under* τ. The total number of the children of all nodes under τ is thus $m \cdot |\overline{\mathcal{W}}_{K-1}(T;\tau)| \geq m$. Thus, the total number of children with feature 0 of all nodes under τ is bounded between 0 and $m \cdot |\overline{\mathcal{W}}_{K-1}(T;\tau)|$. Conversely, for any 2^{K-1}-tuple of non-negative integers, $(\bar{q}_{K,\tau})_{\tau\in P_{K-1}}$, which satisfy $\forall\tau \in P_{K-1}, 1 \leq \bar{q}_{K,\tau} \leq m \cdot |\overline{\mathcal{W}}_{K-1}(T;\tau)| - 1$ can be "realized" by at least some depth-K tree T' obtained by appending children to the leaf nodes of T, in the sense that $\forall\tau = (x_1, ..., x_{K-1}) \in P_{K-1}$, there is $\overline{\mathcal{W}}_K(T';(x_1, ..., x_{K-1}, 0)) = \bar{q}_{K,\tau}$ and $\overline{\mathcal{W}}_K(T';(x_1, ..., x_{K-1}, 1)) = m \cdot |\overline{\mathcal{W}}_{K-1}(T;\tau)| - \bar{q}_{K,\tau}$, and hence $T' \in \mathcal{T}_{m,K,(q_0,...,q_{K-1},\bar{q}_K)}$, with $\bar{q}_K = \sum_{\tau\in P_{K-1}} \bar{q}_{K,\tau}$. Because of the requirement that $1 \leq \bar{q}_{K,\tau} \leq m \cdot |\overline{\mathcal{W}}_{K-1}(T;\tau)| - 1$, we further have that $\forall\tau' \in P_K, \overline{\mathcal{W}}_K(T',\tau') \geq 1$, which implies that $T' \in \tilde{\mathcal{T}}_{m,K,(q_0,...,q_{K-1},\bar{q}_K)}$. Therefore, for some fixed q_K, in order to lower-bound the cardinality of $\tilde{\mathcal{T}}_{m,K,(q_0,...,q_K)}$ by that of $\tilde{\mathcal{T}}_{m,K-1,(q_0,...,q_{K-1})}$, it is sufficient to show a lower bound (which is uniform for all $T \in \tilde{\mathcal{T}}_{m,K-1,(q_0,q_1,...,q_{K-1})}$) on the number of 2^{K-1}-tuples, $(q_{K,\tau})_{\tau\in P_{K-1}}$, which satisfy

$$\begin{aligned} q_K &= \sum_{\tau\in P_{K-1}} q_{K,\tau} \\ \forall\tau \in P_{K-1}, 1 \leq& q_{K,\tau} \leq m \cdot |\overline{\mathcal{W}}_{K-1}(T;\tau)| - 1 \end{aligned} \tag{A.24}$$

A simple bound can be obtained in the following way. For every such T, we sort the 2^{K-1}-tuples in P_{K-1} in ascending order of $|\overline{\mathcal{W}}_{K-1}(T;\cdot)|$, and define $P'_{K-1,T}$ to be the subset of the first 2^{K-2} of these elements according to this order. Thus, for example, $\forall\tau \in P'_{K-1,T}, \forall\tau' \in P_{K-1} \setminus P'_{K-1,T}$, there is $|\overline{\mathcal{W}}_{K-1}(T;\tau)| \leq |\overline{\mathcal{W}}_{K-1}(T;\tau')|$. As a consequence, we have $\sum_{\tau\in P'_{K-1,T}} |\overline{\mathcal{W}}_{K-1}(T;\tau)| \leq \sum_{\tau\in P_{K-1}\setminus P'_{K-1,T}} |\overline{\mathcal{W}}_{K-1}(T;\tau)|$, and so $\sum_{\tau\in P'_{K-1,T}} |\overline{\mathcal{W}}_{K-1}(T;\tau)| \leq \frac{1}{2}\sum_{\tau\in P_{K-1}} |\overline{\mathcal{W}}_{K-1}(T;\tau)| = \frac{1}{2}m^{K-1} \leq \sum_{\tau\in P_{K-1}\setminus P'_{K-1,T}} |\overline{\mathcal{W}}_{K-1}(T;\tau)|$.

Lemma A.8. *Let $K \geq 2$ and q_K satisfy (A.20). Then for all choices of the 2^{K-2}-tuple of integers, $(\bar{q}_{K,\tau})_{\tau\in P'_{K-1,T}}$, such that $\forall\tau \in P'_{K-1,T}, \bar{q}_{K,\tau} = 1$ or 2, we can complete it into at least one 2^{K-1}-tuple of integers, $(\bar{q}_{K,\tau})_{\tau\in P_{K-1}}$, which satisfy (A.24).*

Proof of Lemma A.8: For any such 2^{K-2}-tuple, $(\bar{q}_{K,\tau})_{\tau\in \mathrm{P}'_{K-1,T}}$, in order to satisfy the constraints of (A.24), it is sufficient to find another 2^{K-2} integers, $(\bar{q}_{K,\tau})_{\tau\in \mathrm{P}_{K-1}\setminus \mathrm{P}'_{K-1,T}}$, which satisfy

$$\sum_{\tau\in \mathrm{P}_{K-1}\setminus \mathrm{P}'_{K-1,T}} \bar{q}_{K,\tau} = q_K - \sum_{\tau\in \mathrm{P}'_{K-1,T}} \bar{q}_{K,\tau} \tag{A.25}$$

$$\forall \tau \in \mathrm{P}_{K-1}\setminus \mathrm{P}'_{K-1,T}, 1 \le \bar{q}_{K,\tau} \le m\cdot |\overline{\mathcal{W}}_{K-1}(T;\tau)| - 1 \tag{A.26}$$

On one hand, since $\bar{q}_{K,\tau} = 1$ or 2, $\forall \tau \in \mathrm{P}'_{K-1,T}$, there is $q_K - 2^{K-1} \le q_K - \sum_{\tau\in \mathrm{P}'_{K-1,T}} \bar{q}_{K,\tau} \le q_K - 2^{K-2}$. On the other hand, with the only other constraint being (A.26), it is possible to find $(\bar{q}_{K,\tau})_{\tau\in \mathrm{P}_{K-1}\setminus \mathrm{P}'_{K-1,T}}$ such that $\sum_{\tau\in \mathrm{P}_{K-1}\setminus \mathrm{P}'_{K-1,T}}$ equals any integer between 2^{K-2} and

$$m\cdot \sum_{\tau\in \mathrm{P}_{K-1}\setminus \mathrm{P}'_{K-1,T}} |\overline{\mathcal{W}}_{K-1}(T;\tau)| - 2^{K-2},$$

and hence any integer between 2^{K-2} and $m\cdot \frac{1}{2}m^{K-1} - 2^{K-2} = \frac{1}{2}m^K - 2^{K-2}$. Hence, as long as $2^K - 2^{K-2} \le q_K \le \frac{1}{2}m^K$, which is the assumption of (A.20), Lemma A.8 holds true. □

Lemma A.8 implies that $\forall T \in \tilde{\mathcal{T}}_{m,K-1,(q_0,q_1,\ldots,q_{K-1})}$, there are at least $2^{2^{K-2}}$ distinct choices of 2^{K-1}-tuples $(q_{K,\tau})_{\tau\in \mathrm{P}_{K-1}}$ that satisfy the constraint of (A.24), and hence at least $2^{2^{K-2}}$ non-isomorphic trees in $\tilde{\mathcal{T}}_{m,K,(q_0,q_1,\ldots,q_{K-1},q_K)}$ obtained by appending children to the leaf nodes of T. This proves the inductive hypothesis. Hence, we have

$$|\tilde{\mathcal{T}}_{m,K,(q_0,q_1,\ldots,q_K)}| \ge \prod_{k=2}^{K} 2^{2^{k-2}} = 2^{2^{K-1}-1}, \tag{A.27}$$

which implies Lemma A.7. □

Since $m \ge 2$ by assumption, $\frac{1}{2}m^K - (2^K - 2^{K-2})$ grows exponentially in K. This proves Proposition 2.6.

A.9 Proof of Proposition 2.7

Since the number of walks of a particular type that has length at most k is completely determined by the rooted aggregation tree structure of depth k, it is straightforward to see that all egonets in the same equivalence class induced by k iterations of WL (and therefore GNNs of depth k), which yield the same rooted aggregation tree, will get mapped to the same value by f_k.

For the second part of the claim pertaining to GA-MLPs, we assume for simplicity that $\mathcal{X} = \mathcal{B} = \{0, 1\}$, as the extension to the general case is straightforward but demanding heavier notations. Following the strategy in the proof of Proposition 2.6, it is sufficient to find exponentially-in-k many choices of the tuple $(q_0, q_1, ..., q_k)$, with $0 \leq q_k \leq m^k$, such that image of $\mathcal{T}_{m,k,(q_0,q_1,...,q_k)}$ (as defined in (A.19)) under f_k contains exponentially-in-k many values.

To make it simpler to refer to different nodes in the tree, we index each node in a rooted tree by a tuple of natural numbers: for example, the index-tuple $[1, 3, 2]$ refers to the node at depth 3 that is the second children of the third children of the first children of the root. Since there is no intrinsic ordering to different children of the same node, there exist multiple ways of consistently indexing the nodes in a rooted tree. However, to specify a tree, it suffices to specify the node features of all nodes under *one* such way of indexing.

Given $x_1, ..., x_k \in \mathcal{B}$, we consider a set of depth-k full m-ary trees that satisfy the following: $\forall k' \leq k-1$ and $l_1, ..., l_{k'} \in [m]$, $x_{[l_1,l_2,...,l_{k'}]} = x_{k'}$ if $l_1 = 1$ and $\neg x_{k'}$ if $l_1 > 1$. Note that these trees satisfy, for $k' \leq k-1$, $q_{k'} = m^{k'-1}$ if $x_{k'} = 0$ and $q_{k'} = (m-1)m^{k'-1}$ if $x_{k'} = 1$. Thus, $\forall l_2, ..., l_k \in [m]$, the node $[1, l_2, ..., l_k]$ is under the path $\tau = (x_1, ..., x_k)$ if and only if $x_{[1,l_2,...,l_k]} = x_k$, whereas for $l_1 > 1$, the node $[l_1, l_2, ..., l_k]$ is not under the path τ regardless of the feature of $[1, l_2, ..., l_k]$. Therefore, $f_k(G^{[i]}) = |\mathcal{W}_k(G^{[i]}; (x_1, ..., x_k))|$ equals the number of node of feature x_k among the set of m^{k-1} nodes, $\{[1, l_2, ..., l_k]\}_{l_2,...,l_k \in [m]}$. Hence, if for $k' \leq k-1$, we set $q_{k'} = m^{k'-1}$ if $x_{k'} = 0$ and $q_{k'} = (m-1)m^{k'-1}$ if $x_{k'} = 1$, then choosing any q_k between m^{k-1} and $(m-1)m^{k-1}$, we have that for every integer between 0 and m^{k-1}, there exists a tree T in $\mathcal{T}_{m,k,(q_0,...,q_k)}$ such that $f_k(T)$

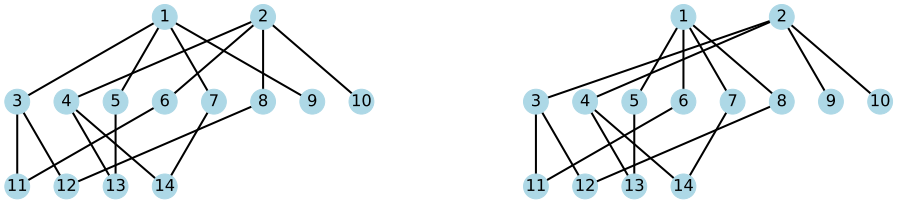


Figure A.1: A pair of graphs with identical node features, G (left) and G' (right), which can be distinguished by 2 iterations of the WL test but not by the GA-MLP with $\Omega \subseteq \{A^k\}_{k\in\mathbb{N}}$.

equals this integer. Since there are $(m-2)m^{k-1}$ choices of q_k (and therefore the tuple $(q_0, ..., q_k)$) and $m^{k-1}+1$ values in the image of $\mathcal{T}_{m,k,(q_0,...,q_k)}$ under f_k, this proves the proposition.

A.10 Proof of Proposition 2.1

We will first prove that the pair of graphs cannot be distinguished by any GA-MLP with $\Omega \subseteq \{A^k\}_{k\in\mathbb{N}}$. Let X and A, X' and A' be the node feature vector and adjacency matrix of the two graphs, G and G', respectively. As these two graphs both contain 14 nodes that have identical features, we have $X, X' \in \mathbb{R}^{14\times 1}$ both being the all-1 vector. Moreover, $\forall i \in [14]$,

$$(A^k X)_i = w_k(i)\,, \quad ((A')^k(X'))_i = w'_k(i) \tag{A.28}$$

where we use $w_k(i)$ and $w'_k(i)$ to denote the numbers of walks (allowing backtracking) of length k starting from node i in graphs G and G', respectively. Thus, to show that any GA-MLP with $\Omega \subseteq \{A^k\}_{k\in\mathbb{N}}$ necessarily returns the same output on G and G', it is sufficient to show that $\forall k \in \mathbb{N}, A^k X = (A')^k(X')$, and therefore sufficient to show that $\forall k \in \mathbb{N}$ and $\forall i \in [14]$, there is $w_k(i) = w'_k(i)$. In fact, we will prove the following lemma:

Lemma A.9. $\forall k \in \mathbb{N}$,

$$w_k(i) = w'_k(i),\ \forall i \in [14] \tag{A.29}$$

$$w_k(1) = w_k(2) \tag{A.30}$$

$$w_k(3) + w_k(9) = w_k(6) + w_k(8) \tag{A.31}$$

$$w_k(5) + w_k(7) = w_k(4) + w_k(10) \tag{A.32}$$

Proof of Lemma A.9: We prove this lemma by induction. For the base case, we have that $w_0(i) = w'_0(i), \forall i \in [14]$. Next, we assume that (A.29) - (A.32) hold for some $k \in \mathbb{N}$ and prove it for $k+1$. A first property to note is that $\forall k \in \mathbb{N}$, $w_{k+1}(i) = \sum_{j \in \mathcal{N}(i)} w_k(j)$ and $w'_{k+1}(i) = \sum_{j \in \mathcal{N}'(i)} w'_k(j)$, where we use $\mathcal{N}(i)$ and $\mathcal{N}'(i)$ to denote the neighborhood of i in G and G', respectively.

To show (A.29) for $k+1$, we look at each node separately:

- $i = 1$

$$\begin{aligned} w_{k+1}(1) &= w_k(3) + w_k(5) + w_k(7) + w_k(9) \\ &= w_k(5) + w_k(6) + w_k(7) + w_k(8) \\ &= w'_k(5) + w'_k(6) + w'_k(7) + w'_k(8) \\ &= w'_{k+1}(1) \end{aligned} \tag{A.33}$$

- $i = 2$

$$\begin{aligned} w_{k+1}(2) &= w_k(4) + w_k(6) + w_k(8) + w_k(10) \\ &= w_k(3) + w_k(4) + w_k(9) + w_k(10) \\ &= w'_k(3) + w'_k(4) + w'_k(9) + w'_k(10) \\ &= w'_{k+1}(2) \end{aligned} \tag{A.34}$$

- $i = 3$

$$\begin{aligned} w_{k+1}(3) =& w_k(1) + w_k(11) + w_k(12) \\ =& w_k(2) + w_k(11) + w_k(12) \\ =& w'_k(2) + w'_k(11) + w'_k(12) \\ =& w'_{k+1}(3) \end{aligned} \tag{A.35}$$

- $i = 4$

$$\begin{aligned} w_{k+1}(4) =& w_k(2) + w_k(13) + w_k(14) \\ =& w'_k(2) + w'_k(13) + w'_k(14) \\ =& w'_{k+1}(4) \end{aligned} \tag{A.36}$$

- $i = 5$

$$\begin{aligned} w_{k+1}(5) =& w_k(1) + w_k(13) \\ =& w'_k(1) + w'_k(13) \\ =& w'_{k+1}(5) \end{aligned} \tag{A.37}$$

- $i = 6$

$$\begin{aligned} w_{k+1}(6) =& w_k(2) + w_k(11) \\ =& w_k(1) + w_k(11) \\ =& w'_k(1) + w'_k(11) \\ =& w'_{k+1}(6) \end{aligned} \tag{A.38}$$

- $i = 7$

$$\begin{aligned} w_{k+1}(7) =& w_k(1) + w_k(13) \\ =& w'_k(1) + w'_k(13) \\ =& w'_{k+1}(7) \end{aligned} \tag{A.39}$$

- $i = 8$

$$\begin{aligned} w_{k+1}(8) =& w_k(2) + w_k(12) \\ =& w_k(1) + w_k(12) \\ =& w'_k(1) + w'_k(12) \\ =& w'_{k+1}(8) \end{aligned} \tag{A.40}$$

- $i = 9$

$$\begin{aligned} w_{k+1}(9) =& w_k(1) \\ =& w_k(2) \\ =& w'_k(2) \\ =& w'_{k+1}(9) \end{aligned} \tag{A.41}$$

- $i = 10$

$$\begin{aligned} w_{k+1}(10) =& w_k(2) \\ =& w'_k(2) \\ =& w'_{k+1}(10) \end{aligned} \tag{A.42}$$

- $i \in \{11, ..., 14\}$

For each of these i's, $\mathcal{N}(i) = \mathcal{N}'(i)$. Therefore,

$$\begin{aligned} w_{k+1}(i) &= \sum_{j \in \mathcal{N}(i)} w_k(j) \\ &= \sum_{j \in \mathcal{N}'(i)} w'_k(j) \\ &= w'_{k+1}(i) \end{aligned} \tag{A.43}$$

Next, for (A.30) - (A.32) at $k+1$,

$$\begin{aligned} w_{k+1}(1) &= w_k(3) + w_k(5) + w_k(7) + w_k(9) \\ &= w_k(4) + w_k(6) + w_k(8) + w_k(10) \\ &= w_{k+1}(2) \end{aligned} \tag{A.44}$$

$$\begin{aligned} w_{k+1}(3) + w_{k+1}(9) &= 2w_k(1) + w_k(11) + w_k(12) \\ &= 2w_k(2) + w_k(11) + w_k(12) \\ &= w_{k+1}(6) + w_{k+1}(8) \end{aligned} \tag{A.45}$$

$$\begin{aligned} w_{k+1}(5) + w_{k+1}(7) &= 2w_k(1) + w_k(13) + w_k(14) \\ &= 2w_k(2) + w_k(13) + w_k(14) \\ &= w_{k+1}(4) + w_{k+1}(10) \end{aligned} \tag{A.46}$$

This proves the inductive hypothethis for $k+1$. □

We next argue that these two graphs can be distinguished by WL in 2 iterations. This is because 2 iterations of WL distinguish neighborhoods up to the depth-2 rooted aggregation trees (as will be defined in Section 2.5), and it is not hard to see that the multiset of depth-2 rooted aggregation trees are different for the two graphs. Note that a depth-2 rooted subtree can be represented by the multiset of the degrees of the depth-1 children. Then for example, the depth-2

rooted aggregation trees of 1 and 2 in G are both $\{3, 2, 2, 1\}_m$, while their rooted aggregation trees in G' are $\{2, 2, 2, 2\}_m$ and $\{3, 3, 1, 1\}_m$, respectively.

A.11 Experiment Details

A.11.1 Specific Architectures

In Section 2.6, we show experiments on several tasks to confirm our theoretical results with several related architectures. Here are some explanations for them:

- **GIN**: Graph Isomorphism Networks proposed by [Xu et al. 2019]. In our experiment of counting attributed walks, we take the depth of GIN as same as the depth of target walks. The number of hidden dimensions is searched in $\{8, 16, 32, 64, 256\}$. The model is trained with the Adam optimizer [Kingma and Ba 2014]. The learning rates are selected from $\{0.1, 0.02, 0.01, 0.005, 0.001\}$. We also train a variant with Jumping Knowledge [Xu et al. 2018].

- **sGNN**: Spectral GNN proposed by [Chen et al. 2019b], which can be viewed as a learnable generalization of power iterations on a collection of operators. While the best performing variant utilizes the non-backtracking operator on the line graph, for a fairer comparison with GA-MLPs, we choose a variant with the base collection of operators being $\{I, A, A^2\}$ on each layer and depth 60, which then has the same receptive field as the chosen GA-MLP models. The model is trained with the Adam optimizer with learning rate selected from $\{0.001, 0.002, 0.004\}$.

- **GA-MLP**: a multilayer perceptron following graph augmented features. For counting attributed walks, we choose the operators from $\{I, \bar{A}^k_{\{\epsilon\}}\}$. The number of hidden dimensions is searched in $\{8, 32, 64, 256\}$. We take the highest order of operators as the twice depth of

target walks at most. For comminity detection, we choose the operators from $\{I, \bar{A}^k_{\{\epsilon\}}, \tilde{H}^k\}$ where $\tilde{H}$ is induced from the Bethe Hessian matrix H. The highest order of operators is searched in $\{30, 60, 120\}$. The number of hidden dimensions is searched in $\{10, 20\}$. On both tasks, the model is trained with the Adam optimizer [Kingma and Ba 2014] with learning rate selected from $\{0.1, 0.02, 0.01, 0.005, 0.001, 0.0001\}$. Additionally, we use Batch Normalization [Ioffe and Szegedy 2015] in community detection after propagating through each operator, following the normalization strategy from [Chen et al. 2019b]. We choose φ to be the identity function.

A.11.2 Bethe Hessian

The Bethe Hessian matrix is defined as

$$H(r) := (r^2 - 1)I - rA + D.$$

with r being a flexible parameter. In SBM, an optimal choice is $r_c = \sqrt{c}$, where c is the average degree. Spectral clustering [Saade et al. 2014] can be performed by computing the eigenvectors associated with the negative eigenvalues of $H(r_c)$ to get clustering information in assortative binary stochastic block model, which is the scenario we consider. In order to utilize power iterations for eigenvector extraction, we induce a new matrix $\tilde{H}$ as

$$\tilde{H} := \kappa I - H(r_c),$$

so that the smallest eigenvalues of H become the largest eigenvalues of $\tilde{H}$. We choose $\kappa = 8$ in our experiments. For GA-MLP-H, we then let $\Omega = \{I, \tilde{H}, ..., \tilde{H}^K\}$.

A.11.3 Results for GA-MLP-$\tilde{A}_{(1)}$ in community detection

Table A.2: Results for community detection on binary SBM by GA-MLP-$\tilde{A}_{(1)}$

Rank of hardness	1	2	3	4	5
Overlap	0.128	0.164	0.262	0.707	0.563

B | Appendix: Supplementary Materials for Chapter 3

B.1 Additional Results

B.1.1 On a 2-D function

Similar to $f(x) = \frac{1}{4}(x^2 - \mu)^2$, consider a 2-D function $f(x, y) = \frac{1}{2}(xy - \mu)^2$. Apparently, if x and y initialize as the same, then $(x^{(t)}, y^{(t)})$ would always align with the 1-D case from the same initialization. Therefore, it is significant to analyze this problem under different initialization for x and y, which we would call "in-balanced" initialization. Meanwhile, another giant difference is that all the global minima in 2-D case form a manifold $\{(x, y)|xy = \mu\}$ while the 1-D case only has two points of global minima. It would be great if we could understand which points in the global minima manifold, or in the whole parameter space, are preferable by GD.

Note that reweighting the two parameters would manipulate the curvature to infinity as in [Elkabetz and Cohen 2021], so the inbalance strongly affects the local curvature. Viewing $f(x)$ as a symmetric scalar factorization problem, we treat $f(x, y)$ as asymmetric scalar factorization. The update rule of GD is

$$x^{(t+1)} := x^{(t)} - \eta(x^{(t)}y^{(t)} - \mu)y^{(t)}, \quad y^{(t+1)} := y^{(t)} - \eta(x^{(t)}y^{(t)} - \mu)x^{(t)}. \tag{B.1}$$

Consider the Hessian as

$$H \triangleq \begin{bmatrix} \partial_x^2 f & \partial_y \partial_x f \\ \partial_x \partial_y f & \partial_y^2 f \end{bmatrix} = \begin{bmatrix} y^2 & 2xy - \mu \\ 2xy - \mu & x^2 \end{bmatrix}. \tag{B.2}$$

When $xy = \mu$, the eigenvalues of H are $\lambda_1 = x^2 + y^2, \lambda_2 = 0$. Note that $\lambda_1 = (x-y)^2 + 2\mu$. Hence, in the global minima manifold, the local curvature of each point is larger if its two parameters are more inbalanced. Among all these points, the smallest curvature appears to be $\lambda_1 = 2\mu$ when $x = y = \sqrt{\mu}$. In other words, if the learning rate $\eta > 2/2\mu$, all points in the manifold would be too sharp for GD to converge. We would like to investigate the behavior of GD in this case. It turns out the two parameters are driven to a perfect balance although they initialized differently, as follows.

Theorem B.1 (Restatement of Theorem 3.4)**.** *For $f(x,y) = \frac{1}{2}(xy-\mu)^2$, consider GD with learning rate $\eta = K \cdot \frac{1}{\mu}$. Assume both x and y are always positive during the whole process $\{x_i, y_i\}_{i \geqslant 0}$. In this process, denote a series of all points with $xy > \mu$ as $\mathcal{P} = \{(x_i, y_i) | x_i y_i > \mu\}$. Then $|x-y|$ decays to 0 in $\mathcal{P}$, for any $1 < K < 1.5$.*

Proof sketch The details of proof are presented in the Appendix B.7. Start from a point $(x^{(t)}, y^{(t)})$ where $x^{(t)} y^{(t)} > \mu$. Because $y^{(t+1)} - x^{(t+1)} = (y^{(t)} - x^{(t)})(1 + \eta(x^{(t)} y^{(t)} - \mu))$, it suffices to show

$$\left| \frac{y^{(t+2)} - x^{(t+2)}}{y^{(t)} - x^{(t)}} \right| = |(1 + \eta(x^{(t)} y^{(t)} - \mu))(1 + \eta(x^{(t+1)} y^{(t+1)} - \mu))| < 1. \tag{B.3}$$

Since $1 + \eta(x^{(t)} y^{(t)} - \mu) > 1$, the analysis of $1 + \eta(x^{(t+1)} y^{(t+1)} - \mu)$ is divided into three cases considering the coupling of $(x^{(t)}, y^{(t)})$, $(x^{(t+1)}, y^{(t+1)})$. □

Remark 5. Actually, for a larger $K \geqslant 1.5$, it is possible for GD to converge to an inbalanced orbit. For instance, Figure 15 in [Wang et al. 2021] shows inbalanced orbits for $f(x) = \frac{1}{2}(xy-1)^2$ with

$K = 1.9$.

Combining with the fact that the probability of GD converging to a stationary point that has sharpness beyond the edge of stability is zero [Ahn et al. 2022], Theorem 3.4 reveals x and y would converge to a perfect balance. Note that this balancing effect is different from that of gradient flow [Du et al. 2018], where the latter states that gradient flow preserves the difference of norms of different layers along training. As a result, in gradient flow, inbalanced initialization induces inbalanced convergence, while in our case inbalanced-initialized weights converge to a perfect balance. Furthermore, Theorem 3.4 shows an effect that the two parameters are squeezed to a single variable, which re-directs to our 1-D analysis in Theorem 3.3. Therefore, actually both cases converge to the same orbit when $1 < K < 1.121$, as stated in Prop 3. Numerical results are presented in Figure B.4.

Proposition 8 (Restatement of Prop 3)**.** *Following the setting in Theorem 3.4. Further assume* $1 < K < \sqrt{4.5} - 1 \approx 1.121$. *Then GD converges to an orbit of period 2. The orbit is formally written as* $\{(x = y = \delta_i) | i = 1, 2\}$, *with* $\delta_1 \in (0, \sqrt{\mu}), \delta_2 \in (\sqrt{\mu}, 2\sqrt{\mu})$ *as the solutions of solving* δ *in*

$$\eta = \frac{1}{\delta^2 \left(\sqrt{\frac{\mu}{\delta^2} - \frac{3}{4}} + \frac{1}{2} \right)}.$$

Remark 6. Actually this convergence is close to the flattest minima because: if the learning rate decays to infinitesimal after sufficient oscillations, then the trajectory walks towards the flattest minima.

However, one thing to notice is that the inbalance at initialization needs to be bounded in Theorem 3.4 because both x and y are assumed to stay positive along the training. More precisely, we have

$$x^{(t+1)} y^{(t+1)} = x^{(t)} y^{(t)} (1 - \eta (x^{(t)} y^{(t)} - \mu))^2 - \eta (x^{(t)} y^{(t)} - \mu)(x^{(t)} - y^{(t)})^2, \tag{B.4}$$

and then $x^{(t+1)}y^{(t+1)} < 0$ when $|x^{(t)} - y^{(t)}|$ is large with $x^{(t)}y^{(t)} > \mu$ fixed. Therefore, we provide a condition to guarantee both x, y positive as follows, with details presented in the Appendix B.8.

Lemma B.2. *In the setting of Theorem 3.4, denote the initialization as* $m = \frac{|y_0 - x_0|}{\sqrt{\mu}}$ *and* $x_0 y_0 > \mu$. *Then, during the whole process, both* x *and* y *will always stay positive, denoting* $p = \frac{4}{\left(m+\sqrt{m^2+4}\right)^2}$ *and* $q = (1+p)^2$, *if*

$$\max\left\{\eta(x_0y_0 - \mu), \frac{4}{27}(1+K)^3 + \left(\frac{2}{3}K^2 - \frac{1}{3}K + \frac{qK^2}{2(K+1)}m^2\right)qm^2 - K\right\} < p.$$

B.1.2 On Matrix Factorization

In this section, we present two additional results of matrix factorization.

B.1.2.1 Asymmetric Case: 1D function at the minima

Before looking into the theorem, we would like to clarify the definition of the loss Hessian. Inherently, we squeeze $\mathbf{X}, \mathbf{Y}$ into a vector $\theta = \mathrm{vec}(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{mp+pq}$, which vectorizes the concatnation. As a result, we are able to represent the loss Hessian w.r.t. θ as a matrix in $\mathbb{R}^{(mp+pq)\times(mp+pq)}$. Meanwhile, the support of the loss landscape is in $\mathbb{R}^{mp+pq}$. Similarly, we use $(\Delta\mathbf{X}, \Delta\mathbf{Y})$ in the same shape of $(\mathbf{X}, \mathbf{Y})$ to denote . In the following theorem, we are to show the leading eigenvector $\Delta \triangleq \mathrm{vec}(\Delta\mathbf{X}, \Delta\mathbf{Y}) \in \mathbb{R}^{mp+pq}$ of the loss Hessian. Since the cross section of the loss landscape and Δ forms a 1D function f_Δ, we would also show the stable-oscillation condition on 1D function holds at the minima of f_Δ.

Theorem B.3. *For a matrix factorization problem, assume* $\mathbf{XY} = \mathbf{C}$. *Consider SVD of both matrices as* $\mathbf{X} = \sum_{i=1}^{\min\{m,p\}} \sigma_{x,i}u_{x,i}v_{x,i}^\top$ *and* $\mathbf{Y} = \sum_{i=1}^{\min\{p,q\}} \sigma_{y,i}u_{y,i}v_{y,i}^\top$, *where both groups of* $\sigma_{\cdot,i}$*'s are in descending order and both top singular values* $\sigma_{x,1}$ *and* $\sigma_{y,1}$ *are unique. Also assume* $v_{x,1}^\top u_{y,1} \neq 0$. *Then the leading*

eigenvector of the loss Hessian is $\Delta = vec(C_1 u_{x,1} u_{y,1}^\top, C_2 v_{x,1} v_{y,1}^\top)$ *with* $C_1 = \frac{\sigma_{y,1}}{\sqrt{\sigma_{x,1}^2+\sigma_{y,1}^2}}, C_2 = \frac{\sigma_{x,1}}{\sqrt{\sigma_{x,1}^2+\sigma_{y,1}^2}}$. *Denote* f_Δ *as the 1D function at the cross section of the loss landscape and the line following the direction of* Δ *passing* $vec(\Delta\mathbf{X}, \Delta\mathbf{Y})$. *Then, at the minima of* f_Δ, *it satisfies*

$$3[f_\Delta^{(3)}]^2 - f_\Delta^{(2)} f_\Delta^{(4)} > 0. \tag{B.5}$$

The proof is provided in Appendix B.10.1. This theorem aims to generalize our 1-D analysis into higher dimension, and it turns out the 1-D condition is sastisfied around any minima for two-layer matrix factorization. In Theorem 3.1 and Lemma 3.2, if such 1-D condition holds, there must exist a period-2 orbit around the minima for GD beyond EoS. However, this is not straightforward to generalize to high dimensions, because 1) directions of leading eigenvectors and (nearby) gradient are not necessarily aligned, and 2) it is more natural and practical to consider initialization *in any direction* around the minima instead of strictly along leading eigenvectors. Therefore, below we present a convergence analysis with initialization near the minima, but in any direction instead.

B.2 Additional Experiments

In Appendix B.2.1, we provide numerical experiments to verify our theorems. Then, we provide additional experiments on MLP and MNIST.

B.2.1 Proven Settings

1-D functions. As discussed in the Section 3.4.1, we have $f(x) = \frac{1}{4}(x^2-1)^2$ satisfying the condition in Theorem 3.1 and $g(x) = 2\sin(x)$ satisfying Lemma 3.2, so we estimate that both f and g allow stable oscillation around the local minima. It turns out GD stably oscillates around

the local minima on both functions, when $\eta > \frac{2}{f''(\bar{x})}$ slightly, as shown in Figure B.1.

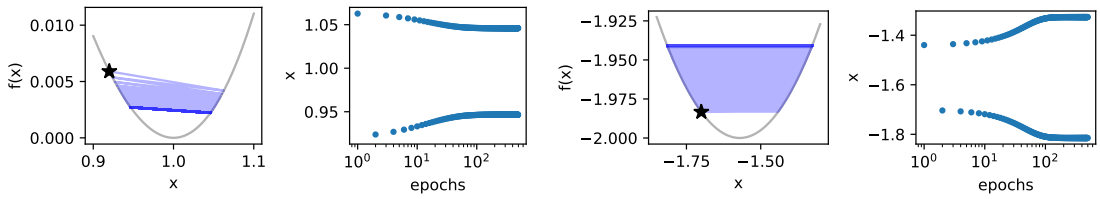


Figure B.1: Running GD around the local minima of $f(x) = \frac{1}{4}(x^2 - 1)^2$ (left two) and $f(x) = 2\sin(x)$ (right two) with learning rate $\eta = 1.01 > \frac{2}{f''(\bar{x})} = 1$. Stars denote the start points. It turns out both functions allow stable oscillation around the local minima.

Two-layer single-neuron model. As discussed in the Section 3.5, with a learning rate $\eta \in (d, 1.1d]$, a single-neuron network $f(x) = v \cdot \sigma(w^\top x)$ is able to align with the direction of the teacher neuron under population loss. We train such a model in empirical loss on 1000 data points uniformly sampled from a sphere $\mathcal{S}^1$, as shown in Figure B.2. The student neuron is initialized orthogonal to the teacher neuron. In the end of training, w_y decays to a small value before the inbalance $|v - w_x|$ decays sharply, which verifies our argument in Section 3.5. With a small w_y, this nonlinear problem degenerates to a 2-D problem on v, w_x. Then, the balanced property makes it align with the 1-D problem where v and w_x converge to a period-2 orbit. Note that the small residuals of $|v - w_x|$ and w_y are due to the difference between population loss and empirical loss.

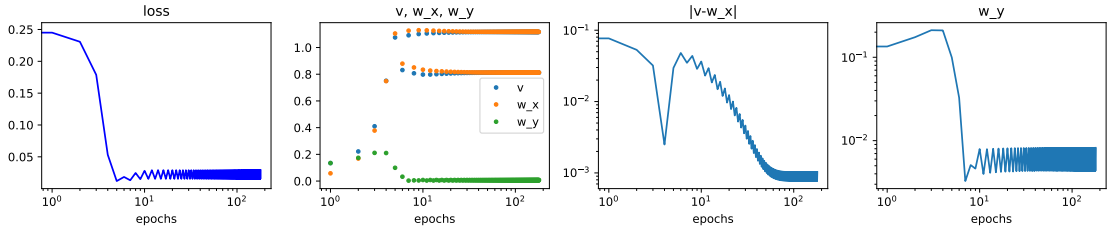


Figure B.2: Running GD in the teacher-student setting with learning rate $\eta = 2.2 = 1.1d$, trained on 1000 points uniformly sampled from sphere $\mathcal{S}^1$ of $\|x\| = 1$. The teacher neuron is $\tilde{w} = [1, 0]$ and the student neuron is initialized as $w^{(0)} = [0, 0.1]$ with $v^{(0)} = 0.1$.

Quasi-symmetric matrix factorization. As discussed in the Section 3.6, with mild assumptions, the quasi-symmetric case stably wanders around the flattest minima. We train GD on a matrix factorization problem with $\mathbf{X}_0\mathbf{X}_0^\top = \mathbf{C} \in \mathbb{R}^{8\times 8}$. The learning rate is 1.02× EoS threshold. Following the setting in Section 3.6, for symmetric case, the training starts near $(\mathbf{X}_0, \mathbf{X}_0)$ and, for quasi-symmetric case, it starts near $(\alpha\mathbf{X}_0, {}^1/_\alpha\mathbf{X}_0)$ with $\alpha = 0.8$, as shown in Figure B.3. Although starting with a re-scaling, the quasi-symmetric case achieves the same top singular values in $\mathbf{Y}$ and $\mathbf{Z}$, which verifies the balancing effect of 2-D functions in Theorem 3.4. Then, the top singular values of both cases converge to the same period-2 orbit, which verifies Observation 2.

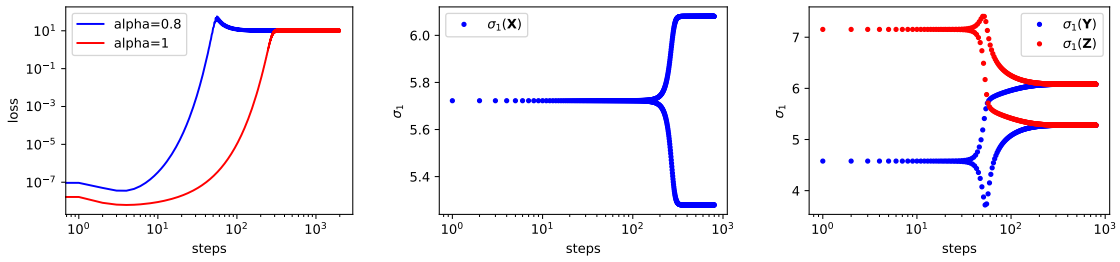
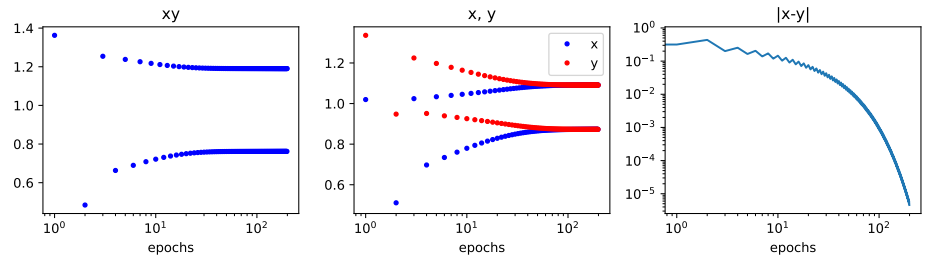


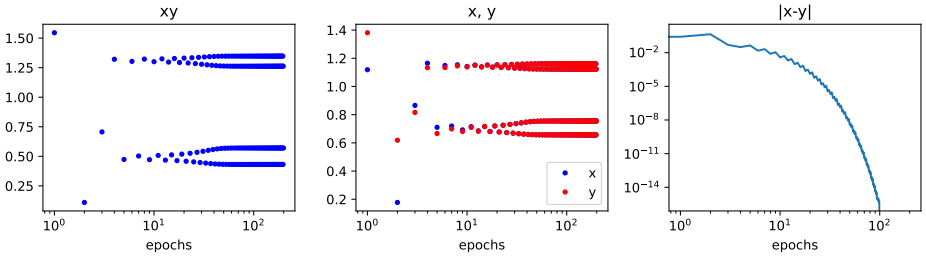
Figure B.3: Symmetric and Quasi-symmetric Matrix factorization: running GD around flat ($\alpha = 1$) and sharp ($\alpha = 0.8$) minima. In both cases, their leading singular values converge to the same period-2 orbit (about 6.1 and 5.3). (Left: Training loss. Middle: Largest singular value of symmetric case. Right: Largest singular values of quasi-symmetric case.)

B.2.2 2-D function

As discussed in the Appendix B.1.1, on the function $f(x, y) = \frac{1}{2}(xy - 1)^2$, we estimate that $|x - y|$ decays to 0 when $\eta \in (1, 1.5)$, as shown in Figure B.4. Since it achieves a perfect balance, the two parameters follows convergence of the corresponding 1-D function $f(x) = \frac{1}{4}(x^2 - 1)^2$. As shown in Figure B.4, xy with $\eta = 1.05$ converges to a period-2 orbit, as stated in the 1-D discussion of Theorem 3.3 while xy with $\eta = 1.25$ converges to a period-4 orbit, which is out of our range in the theorem. But still it falls into the range for balance in Theorem 3.4.



(a) $\eta = 1.05$



(b) $\eta = 1.25$

Figure B.4: Running GD on $f(x, y) = \frac{1}{2}(xy - 1)^2$ with learning rate $\eta = 1.05$ (top) and $\eta = 1.25$ (bottom). When $\eta = 1.05$, it converges to a period-2 orbit. When $\eta = 1.25$, it converges to a period-4 orbit. In both cases, $|x - y|$ decays sharply.

B.3 Proof of Theorem 3.1

Theorem B.4 (Restatement of Theorem 3.1)**.** *Consider any 1-D differentiable function $f(x)$ around a local minima $\bar{x}$, satisfying (i) $f^{(3)}(\bar{x}) \neq 0$, and (ii) $3[f^{(3)}]^2 - f''f^{(4)} > 0$ at $\bar{x}$. Then, there exists ϵ with sufficiently small $|\epsilon|$ and $\epsilon \cdot f^{(3)} > 0$ such that: for any point x_0 between $\bar{x}$ and $\bar{x} - \epsilon$, there exists a learning rate η such that the update rule F_η of GD satisfies $F_\eta(F_\eta(x_0)) = x_0$, and*

$$\frac{2}{f''(\bar{x})} < \eta < \frac{2}{f''(\bar{x}) - \epsilon \cdot f^{(3)}(\bar{x})}.$$

Proof. For simplicity, we assume $f^{(3)}(\bar{x}) > 0$. Imagine a starting point $x_0 = \bar{x} - \epsilon, \epsilon > 0$. We omit $f'(\bar{x}), f''(\bar{x}), f^{(3)}(\bar{x}), f^{(4)}(\bar{x})$ as $f', f'', f^{(3)}, f^{(4)}$. After running two steps of gradient descent, we

have

$$
\begin{aligned}
x_0 &= \bar{x} - \epsilon, \\
f'(x_0) &= f' - f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3 + O(\epsilon^4) \\
&= -f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3 + O(\epsilon^4), \\
x_1 &= x_0 - \eta f'(x_0) = \bar{x} - \epsilon - \eta\big(-f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3\big) + O(\epsilon^4), \\
f'(x_1) &= f'' \cdot (x_1 - \bar{x}) + \frac{1}{2}f^{(3)} \cdot (x_1 - \bar{x})^2 + \frac{1}{6}f^{(4)} \cdot (x_1 - \bar{x})^3 + O(\epsilon^4), \\
x_2 &= x_1 - \eta f'(x_1), \\
\frac{x_2 - x_0}{\eta} &= -\left(-f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3\right) - f'' \cdot \left(-\epsilon - \eta\big(-f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3\big)\right) \\
&\quad - \frac{1}{2}f^{(3)}\left(-\epsilon - \eta\big(-f''\epsilon + \frac{1}{2}f^{(3)}\epsilon^2 - \frac{1}{6}f^{(4)}\epsilon^3\big)\right)^2 - \frac{1}{6}f^{(4)} \cdot (-\epsilon - \eta(-f''\epsilon))^3 + O(\epsilon^4) \\
&= (2f'' - \eta f''f'')\,\epsilon + \left(-\frac{1}{2}f^{(3)} + \frac{1}{2}\eta f''f^{(3)} - \frac{1}{2}f^{(3)}(-1 + \eta f'')^2\right)\epsilon^2 \\
&\quad + \left(\frac{1}{6}f^{(4)} - \frac{1}{6}\eta f''f^{(4)} + \frac{1}{2}(-1 + \eta f'')\eta f^{(3)}f^{(3)} - \frac{1}{6}(-1 + \eta f'')^3 f^{(4)}\right)\epsilon^3 + O(\epsilon^4).
\end{aligned}
$$

When $\eta = \frac{2}{f''}$, it holds

$$
\frac{x_2 - x_0}{\eta} = \left(\frac{1}{2}\eta f^{(3)}f^{(3)} - \frac{1}{3}f^{(4)}\right)\epsilon^3 + O(\epsilon^4), \tag{B.6}
$$

which would be positive if $\frac{1}{2}\eta f^{(3)}f^{(3)} - \frac{1}{3}f^{(4)} = \frac{1}{3f''}(3[f^{(3)}]^2 - f''f^{(4)}) > 0$ and $|\epsilon|$ is sufficiently small.

When $\eta = \frac{2}{f'' - \epsilon \cdot f^{(3)}}$ then $\eta f'' = 2 + 2\frac{f^{(3)}}{f''}\epsilon + O(\epsilon^2)$, it holds

$$
\frac{x_2 - x_0}{\eta} = -2f^{(3)}\epsilon^2 + \left(-\frac{1}{2}f^{(3)} + f^{(3)} - \frac{1}{2}f^{(3)}\right)\epsilon^2 + O(\epsilon^3) = -2f^{(3)}\epsilon^2 + O(\epsilon^3), \tag{B.7}
$$

which is negative when $|\epsilon|$ is sufficiently small.

Therefore, there exists a learning rate $\eta \in (\frac{2}{f''}, \frac{2}{f''-\epsilon \cdot f^{(3)}})$ such that $x_2 = x_0$ due to the continuity of $(x_2 - x_0)$ with respect to η.

The above proof can be generalized to the case of $x_0 = \bar{x} - \epsilon'$ with $\epsilon' \in (0, \epsilon]$ and the learning rate is still bounded as $\eta \in (\frac{2}{f''}, \frac{2}{f''-\epsilon \cdot f^{(3)}})$. □

B.4 Proof of Lemma 3.2

Lemma B.5 (Restatement of Lemma 3.2)**.** *Consider any 1-D differentiable function $f(x)$ around a local minima $\bar{x}$, satisfying that the lowest order non-zero derivative (except the f'') at $\bar{x}$ is $f^{(k)}(\bar{x})$ with $k \geqslant 4$. Then, there exists ϵ with sufficiently small $|\epsilon|$ such that: for any point x_0 between $\bar{x}$ and $\bar{x} - \epsilon$, and*

1. *if k is odd and $\epsilon \cdot f^{(k)}(\bar{x}) > 0, f^{(k+1)}(\bar{x}) < 0$, then there exists $\eta \in (\frac{2}{f''}, \frac{2}{f''-f^{(k)}\epsilon^{k-2}})$,*
2. *if k is even and $f^{(k)}(\bar{x}) < 0$, then there exists $\eta \in (\frac{2}{f''}, \frac{2}{f''+f^{(k)}\epsilon^{k-2}})$,*

such that: the update rule F_η of GD satisfies $F_\eta(F_\eta(x_0)) = x_0$.

Proof. (1) If k is odd, assuming $f^{(k)} > 0$ for simplicity, we have

$$
\begin{aligned}
x_0 &= \bar{x} - \epsilon, \\
f'(x_0) &= -f''\epsilon + \frac{1}{(k-1)!} f^{(k)} \epsilon^{k-1} - \frac{1}{k!} f^{(k+1)} \epsilon^k + O(\epsilon^{k+1}), \\
x_1 &= x_0 - \eta f'(x_0) = \bar{x} - \epsilon + \eta f''\epsilon - \frac{1}{(k-1)!} \eta f^{(k)} \epsilon^{k-1} + \frac{1}{k!} \eta f^{(k+1)} \epsilon^k + O(\epsilon^{k+1}), \\
f'(x_1) &= f'' \cdot (x_1 - \bar{x}) + \frac{1}{(k-1)!} f^{(k)} \cdot (x_1 - \bar{x})^{k-1} + \frac{1}{k!} f^{(k+1)} \cdot (x_1 - \bar{x})^k + O(\epsilon^{k+1}), \\
\frac{x_2 - x_0}{\eta} &= \frac{x_1 - \eta f'(x_1) - x_0}{\eta} = -f'(x_0) - f'(x_1) \\
&= \left(2f'' - \eta f'' f''\right) \epsilon \\
&\quad + \left(-\frac{1}{(k-1)!} f^{(k)} + \frac{1}{(k-1)!} \eta f'' f^{(k)} - \frac{1}{(k-1)!} f^{(k)} \cdot (-1 + \eta f'')^{k-1} \right) \epsilon^{k-1} \\
&\quad + \left(\frac{1}{k!} f^{k+1} - \frac{1}{k!} \eta f'' f^{(k+1)} - \frac{1}{k!} f^{(k+1)} \cdot (-1 + \eta f'')^k \right) \epsilon^k + O(\epsilon^{k+1})
\end{aligned}
$$

When $\eta = \frac{2}{f''}$, it holds

$$
\frac{x_2 - x_0}{\eta} = -\frac{2}{k!} f^{(k+1)} \epsilon^k + O(\epsilon^{k+1}). \tag{B.8}
$$

When $\eta = \frac{2}{f'' - f^{(k)} \epsilon^{k-2}}$ then $\eta f'' = 2 + 2\frac{f^{(k)}}{f''} \epsilon^{k-2} + O(\epsilon^{2k-4})$, then it holds

$$
\frac{x_2 - x_0}{\eta} = -2 f^{(k)} \epsilon^{k-1} + O(\epsilon^k). \tag{B.9}
$$

Since k is odd and $\epsilon \cdot f^{(k)}(\bar{x}) > 0, f^{(k+1)}(\bar{x}) < 0$, the above two estimations of $x_2 - x_0/\eta$ have one positive and one negative exactly. Therefore, due to the continuity of $x_2 - x_0$ wrt η, there exists a learning rate $\eta \in (\frac{2}{f''}, \frac{2}{f'' - f^{(k)} \epsilon^{k-2}})$ such that $x_2 = x_0$.

The above proof can be generalized to any x_0 between $\bar{x}$ and $\bar{x} - \epsilon$ with the same bound for η.

(2) If k is even, we have

$$
\begin{aligned}
x_0 &= \bar{x} - \epsilon, \\
f'(x_0) &= -f''\epsilon - \frac{1}{(k-1)!} f^{(k)} \epsilon^{k-1} + O(\epsilon^k), \\
x_1 &= x_0 - \eta f'(x_0) = \bar{x} - \epsilon + \eta f''\epsilon + \frac{1}{(k-1)!} \eta f^{(k)} \epsilon^{k-1} + O(\epsilon^k), \\
f'(x_1) &= f'' \cdot (x_1 - \bar{x}) + \frac{1}{(k-1)!} f^{(k)} \cdot (x_1 - \bar{x})^{k-1} + O(\epsilon^k), \\
\frac{x_2 - x_0}{\eta} &= \frac{x_1 - \eta f'(x_1) - x_0}{\eta} = -f'(x_0) - f'(x_1) \\
&= (2f'' - \eta f'' f'')\, \epsilon \\
&\quad + \left(\frac{1}{(k-1)!} f^{(k)} - \frac{1}{(k-1)!} \eta f'' f^{(k)} - \frac{1}{(k-1)!} (-1 + \eta f'')^{k-1} \right) \epsilon^{k-1} + O(\epsilon^k).
\end{aligned}
$$

When $\eta = \frac{2}{f''}$, it holds

$$
\frac{x_2 - x_0}{\eta} = -\frac{2}{(k-1)!} f^{(k)} \epsilon^{k-1} + O(\epsilon^k).
$$

When $\eta = \frac{2}{f'' + c \cdot f^{(k)} \epsilon^{k-2}}$ with $c > 0$ as some constant implying $\eta f'' = 2(1 - c \frac{f^{(k)}}{f''} \epsilon^{k-2}) + O(\epsilon^{2k-4})$, then it holds

$$
\frac{x_2 - x_0}{\eta} = 2 \left(c - \frac{1}{(k-1)!} \right) f^{(k)} \epsilon^{k-1} + O(\epsilon^k),
$$

where we then set $c = 1$.

Hence, the above two estimations of $x_2 - x_0 / \eta$ have one positive and one negative with sufficiently small $|\epsilon|$. Therefore, due to the continuity of $x_2 - x_0$, there exists a learning rate $\eta \in (\frac{2}{f''}, \frac{2}{f'' + f^{(k)} \epsilon^{k-2}})$ such that $x_2 = x_0$.

The above proof can be generalized to any x_0 between $\bar{x}$ and $\bar{x} - \epsilon$ with the same bound for η. □

Corollary 1. *$f(x) = \sin(x)$ allows stable oscillation around its local minima $\bar{x}$.*

Proof. Its lowest order nonzero derivative (expect f'') is $f^{(4)}\bar{x} = \sin(\bar{x}) = -1 < 0$ and the order 4 is even. Then Lemma 3.2 gives the result. □

B.5 Proof of Prop 1

Proposition 9 (Restatement of Prop 1)**.** *Consider a 1-D function $g(x)$, and define the loss function f as $f(x) = (g(x) - y)^2$. Assuming (i) g' is not zero when $g(\bar{x}) = y$, (ii) $g'(\bar{x})g^{(3)}(\bar{x}) < 6[g''(\bar{x})]^2$, then it satisfies the condition in Theorem 3.1 or Lemma 3.2 to allow period-2 stable oscillation around $\bar{x}$.*

Proof. From the definition, we have

$$f''(x) = 2[g(x) - y]g''(x) + 2[g'(x)]^2, \tag{B.10}$$

$$f^{(3)}(x) = 2[g(x) - y]g^{(3)}(x) + 6g''(x)g'(x), \tag{B.11}$$

$$f^{(4)}(x) = 2[g(x) - y]g^{(4)}(x) + 6g''(x)g''(x) + 8g'(x)g^{(3)}(x). \tag{B.12}$$

Then at the global minima where $g(x) = y$, we have $f''(x) = 2[g'(x)]^2$ and $f^{(3)}(x) = 6g''(x)g'(x)$. If we assume y is not a trivial value for $g(x)$, which means $g'(x) \neq 0$ at the minima, and g is not linear around the minima (implies $g'' \neq 0$), then f satisfies $f^{(3)}(\bar{x}) \neq 0$ in Theorem 3.1. Meanwhile, we need $3f^{(3)}f^{(3)} - f''f^{(4)} > 0$ as in Theorem 3.1, hence it requires

$$\frac{1}{2g'(x)g'(x)}36g''(x)g''(x)g'(x)g'(x) - \frac{1}{3}\left(6g''(x)g''(x) + 8g'(x)g^{(3)}(x)\right) > 0 \tag{B.13}$$

$$6g''(x)g''(x) > g'(x)g^{(3)}(x). \tag{B.14}$$

The remaining case is, if $g'(x) \neq 0$ and $g'' = 0$ at the minima, it satisfies the condition for Lemma 3.2 with $k = 4$, because $f^{(3)} = 0$ and $f^{(4)} < 0$ due to (B.12, B.14) □

Corollary 2. *$f(x) = (x^2 - 1)^2$ allows stable oscillation around the local minima $\bar{x} = 1$.*

Proof. With $g(x) = x^2$, it has $g'(1) = 2 \neq 0, g''(1) = 2 \neq 0$. All higher order derivatives of g are zero. Then Prop 1 gives the result. □

Corollary 3. *$f(x) = (\sin(x) - y)^2$ allows stable oscillation around the local minima $\bar{x} = \arcsin(y)$ with $y \in (-1, 1)$.*

Proof. With $g(x) = \sin(x)$, it has $g'(\bar{x}) = \cos(\bar{x}) \neq 0, g^{(3)}(\bar{x}) = -\cos(\bar{x})$. We have $g^{(3)}(\bar{x})$ is bounded as $g'g^{(3)} - 6[g'']^2 = -\cos^2(\bar{x}) - 6\sin^2(\bar{x}) < 0$. Then Prop 1 gives the result. □

Corollary 4. *$f(x) = (\tanh(x) - y)^2$ allows stable oscillation around the local minima $\bar{x} = \tanh^{-1}(y)$ with $y \in (-1, 1)$.*

Proof. With $g(x) = \tanh(x)$, it has $g'(\bar{x}) = \operatorname{sech}^2(\bar{x}) \neq 0$, and $g^{(3)}(\bar{x}) = -2\operatorname{sech}^4(\bar{x}) + 4\operatorname{sech}^2(\bar{x})\tanh^2(\bar{x})$ is bounded as

$$g'g^{(3)} - 6[g'']^2 = -2\operatorname{sech}^6 + 4\operatorname{sech}^4\tanh^2 - 24\operatorname{sech}^4\tanh^2 = -2\operatorname{sech}^6 - 20\operatorname{sech}^4\tanh^2 < 0.$$

Then Prop 1 gives the result. □

Corollary 5. *$f(x) = (x^\alpha - y)^2$ (with $k \in \mathbb{Z}, k \geqslant 2$) allows stable oscillation around the local minima $\bar{x} = y^{1/\alpha}$ except $y = 0$.*

Proof. With $g(x) = x^\alpha$, it has $g'(\bar{x}) = \alpha x^{\alpha-1}, g''(\bar{x}) = \alpha(\alpha-1)x^{\alpha-2}, g^{(3)}(\bar{x}) = \alpha(\alpha-1)(\alpha-2)x^{\alpha-3}$. Then we have $g'g^{(3)} - 6[g'']^2 = \alpha^2(\alpha-1)(-5\alpha+4)x^{2\alpha-4} < 0$. Then Prop 1 gives the result. □

Corollary 6. *$f(x) = (\exp(x) - y)^2$ allows stable oscillation around the local minima $\bar{x} = \log y$ for $y > 0$.*

Proof. With $g(x) = \exp x$, it has $g'(\bar{x}) = g''(\bar{x}) = g^{(3)}(\bar{x}) = \exp(\bar{x})$. Then we have $g'g^{(3)} - 6[g'']^2 < 0$. Then Prop 1 gives the result. □

Corollary 7. *$f(x) = (\log(x) - y)^2$ allows stable oscillation around the local minima $\bar{x} = \exp y$.*

Proof. With $g(x) = logx$, it has $g'(\bar{x}) = \frac{1}{\bar{x}}, g''(\bar{x}) = -\frac{1}{\bar{x}^2}, g^{(3)}(\bar{x}) = -\frac{2}{\bar{x}^3}$. Then we have $g'g^{(3)} - 6[g'']^2 < 0$. Then Prop 1 gives the result. □

Corollary 8. *$f(x) = (\frac{1}{1+\exp(-x)} - y)^2$ allows stable oscillation around the local minima $\bar{x} = sigmoid^{-1}(y)$ for $y \in (0, 1)$.*

Proof. With $g(x) = \frac{1}{1+\exp(-x)}$, it has $g'(\bar{x}) = \frac{\exp(-x)}{(\exp(-x)+1)^2}, g''(\bar{x}) = -\frac{\exp(x)(\exp(x)-1)}{(\exp(x)+1)^3}, g^{(3)}(\bar{x}) = \frac{\exp(x)(-4\exp(x)+\exp(2x)+1)}{(\exp(x)+1)^4}$. Then we have $g'g^{(3)} - 6[g'']^2 \propto -4\exp(x)+\exp(2x)+1-6(\exp(x)-1)^2 < 0$. Then Prop 1 gives the result. □

Proposition 10 (Restatement of Prop 2)**.** *Consider two functions f, g. Assume both $f(x), g(y)$ at $x = \bar{x}, y = f(\bar{x})$ satisfies the conditions in Prop 1 to allow stable oscillations. Then $g(f(x))$ allows stable oscillation around $x = \bar{x}$.*

Proof. Denote $F(x) \triangleq g(f(x))$. Then we have

$$
\begin{aligned}
F'(x) &= g'(f(x))f'(x), \\
F''(x) &= g''(f(x))[f'(x)]^2 + g'(f(x))f''(x), \\
F^{(3)}(x) &= g^{(3)}(f(x))[f'(x)]^3 + 3g''(f(x))f'(x)f''(x) + g'(f(x))f^{(3)}(x).
\end{aligned}
$$

Thus, omitting all variables $\bar{x}$ and $f(\bar{x})$ in the derivatives, it holds

$$
\begin{aligned}
F'(\bar{x})F^{(3)}(\bar{x}) - 6[F''(\bar{x})]^2 &= g'f'\left(g^{(3)}(f')^3 + 3g''f'f'' + g'f^{(3)}\right) - 6\left(g''(f')^2 + g'f''\right)^2 \\
&\leqslant -9g'g''(f')^2f'',
\end{aligned}
$$

where the inequality is due to all conditions in Prop 1. So the only problem is whether we can achieve $g'g''f'' > 0$. The good news is that, even if it holds $g'g''f'' < 0$, we can still find functions

to re-represent $g(f(x))$ as $\hat{g}(\hat{f}(x))$ such that $\hat{g}'\hat{g}''\hat{f}'' < 0$ and all other conditions in Prop 1 are satisfied by $\hat{g}, \hat{f}$.

For $g'g''f'' < 0$, construct $\hat{g}(y) \triangleq g(-y), \hat{f}(x) \triangleq -f(x)$. In this sense, it holds $\hat{g}(\hat{f}(\bar{x})) = g(f(\bar{x}))$. It is easy to verify that both $\hat{g}, \hat{f}$ at $y = -f(\bar{x}), x = \bar{x}$ satisfy the conditions in Prop 1, because

$$\hat{g}'(y) = -g'(-y) = -g'(f(\bar{x})), \quad \hat{g}''(y) = g''(-y) = g''(f(\bar{x})), \quad \hat{g}^{(3)}(y) = -g^{(3)}(-y) = -g^{(3)}(f(\bar{x})),$$
$$\hat{f}'(\bar{x}) = -f'(\bar{x}), \quad \hat{f}''(\bar{x}) = -f''(\bar{x}), \quad \hat{f}^{(3)}(y) = -f^{(3)}(\bar{x}).$$

Then, it has $\hat{g}'(y)\hat{g}''(y)\hat{f}''(x) = -g'g''f'' > 0$ at $y = -f(\bar{x}), x = \bar{x}$. Therefore, we have $F'(\bar{x})F^{(3)}(\bar{x}) - 6[F''(\bar{x})]^2 < 0$ and Prop 1 gives the result. □

B.6 Proof of Theorem 3.3

Theorem B.6 (Restatement of Theorem 3.3)**.** *For $f(x) = \frac{1}{4}(x^2 - \mu)^2$, consider GD with $\eta = K \cdot \frac{1}{\mu}$ where $1 < K < \sqrt{4.5} - 1 \approx 1.121$, and initialized on any point $0 < x_0 < \sqrt{\mu}$. Then it converges to an orbit of period 2, except for a measure-zero initialization where it converges to $\sqrt{\mu}$. More precisely, the period-2 orbit are the solutions $x = \delta_1 \in (0, \sqrt{\mu}), x = \delta_2 \in (\sqrt{\mu}, 2\sqrt{\mu})$ of solving δ in*

$$\eta = \frac{1}{\delta^2\left(\sqrt{\frac{\mu}{\delta^2} - \frac{3}{4}} + \frac{1}{2}\right)}. \tag{B.15}$$

Proof. Assume the 2-period orbit is $(\bar{x}_0, \bar{x}_1)$, which means

$$\bar{x}_1 = \bar{x}_0 - \eta \cdot f'(\bar{x}_0) = \bar{x}_0 + \eta \cdot (\mu - \bar{x}_0^2)\bar{x}_0,$$
$$\bar{x}_0 = \bar{x}_1 - \eta \cdot f'(\bar{x}_1) = \bar{x}_1 + \eta \cdot (\mu - \bar{x}_1^2)\bar{x}_1.$$

First, we show the existence and uniqueness of such an orbit when $K \in (1, 1.5]$ via solving

a high-order equation, some roots of which can be eliminated. Then, we conduct an analysis of global convergence by defining a special interval I. GD starting from any point following our assumption will enter I in some steps, and any point in I will back to this interval after two steps of iteration. Finally, any point in I will converge to the orbit $(\bar{x}_0, \bar{x}_1)$.

Before diving into the proof, we briefly show it always holds $x > 0$ under our assumption. If $x_{t-1} > 0$ and $x_t \leqslant 0$, the GD rule reveals $\eta(\mu - x_{t-1}^2) \leqslant -1$ which implies $x_{t-1}^2 \geqslant \mu + \frac{1}{\eta}$. However, the maximum of $x+\eta(\mu-x^2)x$ on $x \in (0, \sqrt{\mu + \frac{1}{\eta}})$ is achieved when $x^2 = \frac{1}{3}(\mu+\frac{1}{\eta})$ so the maximum value is $\sqrt{\frac{1}{3}(\mu + \frac{1}{\eta})}(\frac{2}{3} + \frac{2}{3}\eta\mu) \leqslant 1.4\sqrt{\frac{1}{3}(\mu + \frac{1}{\eta})} < \sqrt{\mu + \frac{1}{\eta}}$. As a result, it always holds $x > 0$.

Part I. Existence and uniqueness of $(\bar{x}_0, \bar{x}_1)$.

In this part, we simply denote both $\bar{x}_0, \bar{x}_1$ as x_0. This means x_0 in all formulas in this part can be interpreted as $\bar{x}_0$ and $\bar{x}_1$. Then the GD update rule tells, for the orbit in two steps,

$$
\begin{aligned}
x_0 \mapsto x_1 &:= x_0 + \eta(\mu - x_0^2)x_0, \\
x_1 \mapsto x_0 &= x_1 + \eta(\mu - x_1^2)x_1,
\end{aligned}
$$

which means

$$
\begin{aligned}
0 &= \eta(\mu - x_0^2)x_0 + \eta\left(\mu - \left(x_0 + \eta(\mu - x_0^2)x_0\right)^2\right)\left(x_0 + \eta(\mu - x_0^2)x_0\right), \\
0 &= \mu - x_0^2 + \left(\mu - \left(x_0 + \eta(\mu - x_0^2)x_0\right)^2\right)\left(1 + \eta(\mu - x_0^2)\right).
\end{aligned}
$$

Denote $z := 1 + \eta(\mu - x_0^2)$, it is equivalent to

$$
\begin{aligned}
0 &= \mu - x_0^2 + (\mu - z^2 x_0^2)z = (z+1)(-x_0^2 z^2 + x_0^2 z + \mu - x_0^2) \\
&= (z+1)\left(-x_0^2(z - \frac{1}{2})^2 + \mu - \frac{3}{4}x_0^2\right).
\end{aligned}
$$

If $z + 1 = 0$, it means $x_1 = -x_0$ which is however out of the range of our discussion on the $x > 0$

domain. So we require $-x_0^2(z-\frac{1}{2})^2+\mu-\frac{3}{4}x_0^2=0$. To ensure the existence of solutions z, it is natural to require

$$\mu-\frac{3}{4}x_0^2\geqslant 0.$$

Then, the solutions are

$$z=\frac{1}{2}\pm\sqrt{\frac{\mu}{x_0^2}-\frac{3}{4}}.$$

However, $z=\frac{1}{2}-\sqrt{\frac{\mu}{x_0^2}-\frac{3}{4}}$ can be ruled out. If it holds, $\eta(\mu-x_0^2)=z-1<-\frac{1}{2}$ which means $x_0^2>\mu+\frac{1}{2\eta}$. Since we restrict $\eta\mu\in(1,1.121]$, it tells $x_0^2>\mu(1+\frac{1}{1.242})$ contradicting with $\mu\geqslant\frac{3}{4}x_0^2$.

Hence, $z=\frac{1}{2}+\sqrt{\frac{\mu}{x_0^2}-\frac{3}{4}}$ is the only reasonable solution, which is saying

$$\eta(\mu-x_0^2)=-\frac{1}{2}+\sqrt{\frac{\mu}{x_0^2}-\frac{3}{4}}.$$

Given a certain η, the above expression is a third-order equation of x_0^2 to solve. Apparently $x_0^2=\mu$ is one trivial solution, since for any learning rate, the gradient descent stays at the global minimum. Then the two other solutions are exactly the orbit $(\bar{x}_0,\bar{x}_1)$, if the equation does have three different roots. This also guarantees the uniqueness of such an orbit.

Assuming $x_0^2\neq\mu$, the above expression can be reformulated as

$$\eta=\frac{1}{x_0^2\left(\sqrt{\frac{\mu}{x_0^2}-\frac{3}{4}}+\frac{1}{2}\right)}. \tag{B.16}$$

One necessary condition for existence is $\mu\geqslant\frac{3}{4}x_0^2$. Note that here x_0 can be both $\bar{x}_0,\bar{x}_1$, one of which is larger than $\sqrt{\mu}$. For simplicity, we assume $\bar{x}_0<\sqrt{\mu}<\bar{x}_1$. Since η from Eq(B.16) is

increasing with x_0^2 when $\mu < x_0^2$, let $x_0^2 = \frac{4}{3}\mu$ and achieve the upper bound as

$$\eta\mu \leqslant \frac{3}{2}, \tag{B.17}$$

which is satisfied by our assumption $1 < \eta\mu < \sqrt{4.5} - 1 \approx 1.121$.

Therefore, we have shown the existence and uniqueness of a period-2 orbit.

Part II. Global convergence to $(\bar{x}_0, \bar{x}_1)$.

The proof structure is as follows:

1. There exists a special interval $I := [x_s, \sqrt{\mu})$ such that any point in I will back to this interval surely after two steps of gradient descent. And $\bar{x}_0 \in I$.

2. Initialized from any point in I, the gradient descent process will converge to $\bar{x}_0$ (every two steps of GD).

3. Initialized from any point between 0 and $\sqrt{\mu}$, the gradient descent process will fall into I in some steps.

(II.1) Consider a function $F_\eta(x) = x + \eta(\mu - x^2)x$ performing one step of gradient descent. Since $F'_\eta(x) = 1 + \eta\mu - 3\eta x^2$, we have $F'_\eta(x) > 0$ for $0 < x^2 < \frac{1}{3}\left(\mu + \frac{1}{\eta}\right)$ and $F'_\eta(x) < 0$ otherwise. It is obvious that the threshold has $x_s^2 := \frac{1}{3}\left(\mu + \frac{1}{\eta}\right) < \mu$. In the other words, for any point on the right of x_s, GD returns a point in a decreasing manner.

To prove anything further, we would like to restrict $\bar{x}_0 \geqslant x_s$, which is

$$\bar{x}_0^2 \geqslant \frac{1}{3}\left(\mu + \frac{1}{\eta}\right) = \frac{1}{3}\left(\mu + \bar{x}_0^2\left(\sqrt{\frac{\mu}{\bar{x}_0^2} - \frac{3}{4}} + \frac{1}{2}\right)\right).$$

Solving this inequality tells

$$\bar{x}_0^2 \geqslant \frac{3 + \sqrt{2}}{7}\mu. \tag{B.18}$$

Consequently, by applying Eq(B.16), we have

$$\eta\mu \leqslant \sqrt{4.5} - 1 \approx 1.121. \tag{B.19}$$

With the above discussion of x_s, we are able to define the special internal $I \coloneqq [x_s, \sqrt{\mu})$. From the definition of F_η, consider a function representing two steps of gradient descent $F_\eta^2(x) \coloneqq F_\eta(F_\eta(x))$. From previous discussion, we know $F_\eta^2(\bar{x}_0) = \bar{x}_0$. What about $F_\eta^2(x_s)$?

It turns out $F_\eta^2(x_s) > x_s$: we have $F_\eta(x_s) = x_s(1 + \eta\mu - \eta x_s^2) = x_s \cdot \frac{2}{3}(1 + \eta\mu)$ and, furthermore, $F_\eta^2(x_s) = F_\eta(x_s \cdot \frac{2}{3}(1 + \eta\mu)) = x_s \cdot \frac{2}{3}(1 + \eta\mu) \cdot \left(1 + \eta\mu - \frac{4}{27}(1 + \eta\mu)^3\right)$. Then we get $F_\eta^2(x_s) > x_s$ because

$$\frac{2}{3}(1 + \eta\mu) \cdot \left(1 + \eta\mu - \frac{4}{27}(1 + \eta\mu)^3\right) > 1 \quad \text{if} \quad \eta\mu \in (1, \sqrt{4.5} - 1). \tag{B.20}$$

Combining the following facts, i) $F_\eta^2(x) - x$ is continous wrt x, ii) $F_\eta^2(x_s) - x_s > 0$, and iii) $F_\eta^2(\bar{x}_0) - \bar{x}_0 = 0$ is the only zero point on $x \in [x_s, \bar{x}_0]$, we can conclude that

$$F_\eta^2(x) > x, \quad \forall x \in [x_s, \bar{x}_0). \tag{B.21}$$

Meanwhile, we can prove $F_\eta^2(x) < x$ for any $x \in (\bar{x}_0, \sqrt{\mu})$. Since $F_\eta^2(\mu) - \mu = 0$ and $F_\eta^2(\bar{x}_0) - \bar{x}_0 = 0$ are the only two zero cases, we only need to show $\exists\, \hat{x} \in (\bar{x}_0, \sqrt{\mu})$, such that $F_\eta^2(\hat{x}) < \hat{x}$. We compute the derivative of $F_\eta^2(x) - x$ at $x^2 = \mu$, which is $\frac{d}{dx} F_\eta^2(x) - x|_{x^2=\mu} = -1 + F'(F(x))F'(x)|_{x^2=\mu} = -1 + [F'(\sqrt{\mu})]^2 = -1 + (1 - 2\eta\mu)^2 > 0$. Then combining it with $F_\eta^2(\bar{x}_0) = \bar{x}_0$, there exists a point $\hat{x} \in (\bar{x}_0, \sqrt{\mu})$ that is very close to $\sqrt{\mu}$ such that $F_\eta^2(\hat{x}) < \hat{x}$. Hence, we can conclude that

$$F_\eta^2(x) < x, \quad \forall x \in (\bar{x}_0, \sqrt{\mu}). \tag{B.22}$$

Since $F_\eta(\cdot)$ is decreasing on $[x_s, \infty)$ and $F_\eta(x) > x_s$ for $x \in [x_s, \sqrt{\mu}]$, it is fair to say $F_\eta^2(x)$ is

increasing on $x \in [x_s, \sqrt{\mu}]$. Hence, we have $F_\eta^2(x) \leqslant F_\eta^2(\bar{x}_0) = \bar{x}_0,\ \forall x \in [x_s, \bar{x}_0]$. And $F_\eta^2(x) \geqslant F_\eta^2(\bar{x}_0) = \bar{x}_0,\ \forall x(\bar{x}_0, \sqrt{\mu})$

Combining the above results, we have

$$F_\eta^2(x) \in (x, \bar{x}_0], \quad \forall x \in [x_s, \bar{x}_0), \tag{B.23}$$

$$F_\eta^2(x) \in [\bar{x}_0, x), \quad \forall x \in (\bar{x}_0, \sqrt{\mu}). \tag{B.24}$$

(II.2) A consequence of Exp(B.23, B.24) is that any point in I will converge to $\bar{x}_0$ with even steps of gradient descent. For simplicity, we provide the proof for $x \in [x_s, \bar{x}_0)$.

Denote $a_0 \in [x_s, \bar{x}_0)$ and $a_n \coloneqq F_\eta^2(a_{n-1}), n \geqslant 1$. The series $\{a_i\}_{i \geqslant 0}$ satisfies

$$\bar{x}_0 \geqslant a_{n+1} > a_n > a_0. \tag{B.25}$$

Since the series is bounded and strictly increasing, it is converging. Assume it is converging to a. If $a < \bar{x}_0$, then

$$\bar{x}_0 \geqslant F_\eta^2(a) > a > F_\eta^2(a_n).$$

Since $F_\eta^2(\cdot)$ is continuous, so $\exists\ \delta > 0$, such that, when $|x - a| < \delta$, we have

$$|F_\eta^2(x) - F_\eta^2(a)| < F_\eta^2(a) - a. \tag{B.26}$$

Since a is the limit, so $\exists\ N > 0$, such that, when $n > N$, $0 < a - F_\eta^2(a_n) < \delta$. So, combining with Exp(B.26), we have

$$|F_\eta^2(F_\eta^2(a_n)) - F_\eta^2(a)| < F_\eta^2(a) - a.$$

But LHS $= F_\eta^2(a) - a_{n+2} > F_\eta^2(a) - a$, so we reach a contradiction.

Hence, we have $\{a_i\}$ converges to $\bar{x}_0$.

(II.3) Obviously, any initialization in $(0, \sqrt{\mu})$ will have gradient descent run into (i) the interval

I, or (ii) the interval on the right of $\sqrt{\mu}$, *i.e.*, $(\sqrt{\mu}, \infty)$. The first case is exactly our target.

Now consider the second case. From the definition of x_s in part III.1, we know $F_\eta(x_s) = \max_{x\in[0,\sqrt{\mu}]} F_\eta(x)$. So it is fair to say this case is $x_n \in (\sqrt{\mu}, F_\eta(x_s)]$. Then the next step will go into the interval I, because

$$F_\eta(x_n) \geqslant F_\eta(F_\eta(x_s)) = F_\eta^2(x_s) > x_s,$$

where the first inequality is from the decreasing property of $F_\eta(\cdot)$ and the second inequality is due to $F_\eta^2(x) > x$ on $x \in [x_s, \bar{x}_0)$. □

B.7 Proof of Theorem 3.4

Theorem B.7 (Restatement of Theorem 3.4)**.** *For $f(x,y) = \frac{1}{2}(xy-\mu)^2$, consider GD with learning rate $\eta = K \cdot \frac{1}{\mu}$. Assume both x and y are always positive during the whole process $\{x_i, y_i\}_{i\geqslant 0}$. In this process, denote a series of all points with $xy > \mu$ as $\mathcal{P} = \{(x_i, y_i) | x_i y_i > \mu\}$. Then $|x-y|$ decays to 0 in $\mathcal{P}$, for any $1 < K < 1.5$.*

Proof. Consider the current step is at (x_t, y_t) with $x_t y_t > \mu$. After two steps of gradient descent, we have

$$x_{t+1} = x_t + \eta(\mu - x_t y_t) y_t \tag{B.27}$$

$$y_{t+1} = y_t + \eta(\mu - x_t y_t) x_t \tag{B.28}$$

$$x_{t+2} = x_{t+1} + \eta(\mu - x_{t+1} y_{t+1}) y_{t+1} \tag{B.29}$$

$$y_{t+2} = y_{t+1} + \eta(\mu - x_{t+1} y_{t+1}) x_{t+1}, \tag{B.30}$$

with which we have the difference evolve as

$$y_{t+1} - x_{t+1} = (y_t - x_t)\left(1 - \eta\left(\mu - x_t y_t\right)\right) \tag{B.31}$$

$$y_{t+2} - x_{t+2} = (y_{t+1} - x_{t+1})\left(1 - \eta\left(\mu - x_{t+1} y_{t+1}\right)\right). \tag{B.32}$$

Meanwhile, we have

$$\begin{aligned} x_{t+1} y_{t+1} &= x_t y_t + \eta\left(\mu - x_t y_t\right)\left(x_t^2 + y_t^2\right) + \eta^2\left(\mu - x_t y_t\right)^2 x_t y_t \\ &= x_t y_t\left(1 + \eta\left(\mu - x_t y_t\right)\right)^2 + \eta\left(\mu - x_t y_t\right)\left(x_t - y_t\right)^2 \end{aligned} \tag{B.33}$$

Note that the second term in Eq(B.33) vanishes when x and y are balanced. When they are not balanced, if $x_t y_t > \mu$, it holds $x_{t+1} y_{t+1} < x_t y_t\left(1 + \eta\left(\mu - x_t y_t\right)\right)^2$. Incorporating this inequality into Eq(B.31, B.32) and assuming $y_t - x_t > 0$, it holds

$$y_{t+2} - x_{t+2} < (y_t - x_t)\left(1 - \eta\left(\mu - x_t y_t\right)\right)\left(1 - \eta\left(\mu - x_t y_t\left(1 + \eta\left(\mu - x_t y_t\right)\right)^2\right)\right). \tag{B.34}$$

To show that $|x - y|$ is decaying as in the theorem, we are to show

1. $y_{t+2} - x_{t+2} < y_t - x_t$
2. $y_{t+2} - x_{t+2} > -(y_t - x_t)$

Note that, although $x_t y_t > \mu$, it is not sure to have $x_{t+2} y_{t+2} > \mu$. However, for any $0 < x_i y_i < \mu$ and $K < 2$, we have

$$\frac{|x_{i+1} - y_{i+1}|}{|x_i - y_i|} = |1 - \eta\left(\mu - x_i y_i\right)| < 1, \tag{B.35}$$

which is saying $|x - y|$ decays until it reaches $xy > \mu$. So it is enough to prove the above two inequalities, whether or not $x_{t+2} y_{t+2} > \mu$.

Part I. To show $y_{t+2} - x_{t+2} < y_t - x_t$

Since we wish to have $y_{t+2} - x_{t+2} < y_t - x_t$, it is sufficient to require

$$(1 - \eta(\mu - x_t y_t))\left(1 - \eta\left(\mu - x_t y_t\left(1 + \eta(\mu - x_t y_t)\right)^2\right)\right) < 1. \tag{B.36}$$

Since we assume $x_{t+1}, y_{t+1} > 0$, Eq (B.27, B.28) tells $\eta(\mu - x_t y_t) > -\min\{\frac{x_t}{y_t}, \frac{y_t}{x_t}\}$, which is equivalent to $1 - \eta(\mu - x_t y_t) < 1 + \min\{\frac{x_t}{y_t}, \frac{y_t}{x_t}\}$.

(I.1) If $\eta(\mu - x_{t+1}y_{t+1}) \geqslant \frac{1}{2}$

Then we have $1 - \eta(\mu - x_{t+1}y_{t+1}) \leqslant \frac{1}{2}$. As a result,

$$\frac{y_{t+2} - x_{t+2}}{y_t - x_t} = (1 - \eta(\mu - x_t y_t))(1 - \eta(\mu - x_{t+1}y_{t+1})) < \left(1 + \min\{\frac{x_t}{y_t}, \frac{y_t}{x_t}\}\right) \times \frac{1}{2} \tag{B.37}$$

$$= \frac{1}{2} + \frac{1}{2}\min\{\frac{x_t}{y_t}, \frac{y_t}{x_t}\} \tag{B.38}$$

(I.2) If $\eta(\mu - x_{t+1}y_{t+1}) < \frac{1}{2}$ **and** $x_{t+1}y_{t+1} \leqslant x_s^2 = \frac{1}{3}\left(\mu + \frac{1}{\eta}\right)$

The second condition reveals

$$\frac{y_{t+2} - x_{t+2}}{y_{t+1} - x_{t+1}} = 1 - \eta(\mu - x_{t+1}y_{t+1}) \leqslant 1 - \eta\left(\mu - \frac{1}{3}\left(\mu + \frac{1}{\eta}\right)\right)$$

$$= \frac{4}{3} - \frac{2}{3}K. \tag{B.39}$$

The first condition is equivalent to $x_{t+1}y_{t+1} > \mu - \frac{1}{2\eta}$. Since the second term in Eq(B.33) is negative, we have

$$x_t y_t (1 + \eta(\mu - x_t y_t))^2 > \mu - \frac{1}{2\eta}, \tag{B.40}$$

with which we would like to find an upper bound of $x_t y_t$.

Denoting $b = x_t y_t$, consider a function $q(b) = b(1 + \eta(\mu - b))^2$. Obviously $q(\mu) = \mu$. Its

derivative is $q'(b) = (1+\eta\mu-\eta b)(1+\eta\mu-3\eta b) < 0$ on the domain of our interest. If we can show an (negative) upper bound for the derivative as $q'(b) < -1$ on a proper domain, then it is fair to say that, from Exp(B.40), $x_t y_t < \mu + \frac{1}{2\eta}$. Then we have

$$\frac{y_{t+1}-x_{t+1}}{y_t - x_t} = 1-\eta(\mu - x_t y_t) < 1-\eta\left(\mu - \left(\mu - \frac{1}{2\eta}\right)\right) = \frac{3}{2}. \tag{B.41}$$

Then, combining Exp(B.41, B.39), it tells

$$\frac{y_{t+2}-x_{t+2}}{y_t - x_t} < 2 - K. \tag{B.42}$$

The remaining is to show $q'(b) < -1$ on a proper domain. We have $q'(b) = (1+\eta\mu-2\eta b)^2 - (\eta b)^2$, which is equal to $1-2\eta\mu < -1$ when $b = \mu$. Meanwhile, the derivative of $q'(b)$ is $q''(b) = -2\eta(\eta b + (1+\eta\mu-2\eta b)) = -2\eta(1+\eta\mu-\eta b)$, which is negative when $b < \mu + \frac{1}{\eta}$. As a result, it always holds $q'(b) < -1$ when $b < \mu + \frac{1}{\eta}$.

(I.3) If $x_{t+1}y_{t+1} \geqslant x_s^2$

Denoting again $b = x_t y_t$, the above inequality in is saying, with $b > \mu$,

$$p(b) = (1-\eta(\mu-b))\left(1-\eta\left(\mu - b\left(1+\eta(\mu-b)\right)^2\right)\right) < 1. \tag{B.43}$$

After expanding $p(\cdot)$, we have

$$p(b) - 1 = \eta(\mu-b)\left(-2+\eta(\mu-b)+2\eta b - \eta^2 b(\mu-b) - \eta^3 b(\mu-b)^2\right).$$

Apparently $p(\mu) = 1$. So it is necessary to investigate whether $p'(b) < 0$ on $b > \mu$, as

$$p'(b) = 2 - 2\eta b + (\mu-b)\left(\eta^2(1+\eta(\mu-b))(-\mu+3b) + \eta^3 b(\mu-b)\right).$$

Since $\eta b > 1$ and $b > \mu$, it is enough to require

$$(1 + \eta\,(\mu - b))\,(-\mu + 3b) + \eta b\,(\mu - b) > 0$$

$$(1 + \eta(\mu - b))(-\mu + b) + \eta b(\mu - b) + 2b(1 + \eta(\mu - b)) > 0.$$

It suffices to show

$$\eta(\mu - b) + 2(1 + \eta(\mu - b)) = 2 + 3\eta(\mu - b) > 0. \tag{B.44}$$

Since $x_{t+1}y_{t+1} \geqslant x_s^2 = \frac{1}{3}\left(\mu + \frac{1}{\eta}\right)$, it holds

$$b\,(1 + \eta(\mu - b))^2 \geqslant \frac{1}{3}\left(\mu + \frac{1}{\eta}\right)$$

$$2 + 3\eta(\mu - b) \geqslant \sqrt{\frac{3\left(\mu + \frac{1}{\eta}\right)}{b}} - 1 > 0,$$

where the last inequality holds because: if $b \geqslant 3\left(\mu + \frac{1}{\eta}\right)$, then $1 + \eta(\mu - b) \leqslant -2\eta\mu - 2 < 0$, which contradicts with the assumption that both x_{t+1}, y_{t+1} are positive. As a result, the above argument gives

$$\frac{y_{t+2} - x_{t+2}}{y_t - x_t} < p(b) < 1 - 2(K - 1)(b - \mu). \tag{B.45}$$

Part II. To show $y_{t+2} - x_{t+2} > -(y_t - x_t)$

Since $x_t y_t > \mu$, we have $1 - \eta(\mu - x_t y_t) > 1$. Combining with $1 - \eta(\mu - x_t y_t) < 2$, it holds

$$\frac{y_{t+1} - x_{t+1}}{y_t - x_t} = 1 - \eta(\mu - x_t y_t) \in (1, 2).$$

So the remaining is to have $\frac{y_{t+2}-x_{t+2}}{y_{t+1}-x_{t+1}} > -0.5$. Actually it is $1-\eta(\mu - x_{t+1}y_{t+1}) \geqslant 1-\eta\mu = 1-K$. Therefore, we have

$$\frac{y_{t+2}-x_{t+2}}{y_t - x_t} > -1 + (3-2K), \tag{B.46}$$

as required.

Part III. To show $y_t - x_t$ converges to 0

From Exp (B.38, B.42, B.45, B.46), we have for points in $\mathcal{P}$, $|y-x|$ is a monotone strictly decreasing sequence lower bounded by 0. Hence it is convergent. Actually it converges to 0. If not, assuming it converges to $\epsilon > 0$, the next point will have the difference as $\tilde{\epsilon} < \epsilon$ as well as all following points. Hence, the contradiction gives the convergence to 0. □

B.8 Proof of Lemma B.2

Lemma B.8 (Restatement of Lemma B.2)**.** *In the setting of Theorem 3.4, denote the initialization as* $m = \frac{|y_0-x_0|}{\sqrt{\mu}}$ *and* $x_0y_0 > \mu$*. Then, during the whole process, both* x *and* y *will always stay positive, denoting* $p = \frac{4}{\left(m+\sqrt{m^2+4}\right)^2}$ *and* $q = (1+p)^2$*, if*

$$\max\left\{\eta(x_0y_0-\mu), \frac{4}{27}(1+K)^3 + \left(\frac{2}{3}K^2 - \frac{1}{3}K + \frac{qK^2}{2(K+1)}m^2\right)qm^2 - K\right\} < p.$$

Proof. Considering $x_ty_t > \mu$, one step of gradient descent returns

$$x_{t+1} = x_t + \eta(\mu - x_ty_t)y_t$$

$$y_{t+1} = y_t + \eta(\mu - x_ty_t)x_t.$$

To have both $x_{t+1} > 0, y_{t+1} > 0$, it suffices to have

$$\eta(x_t y_t - \mu) < \min\left\{\frac{y_t}{x_t}, \frac{x_t}{y_t}\right\}. \tag{B.47}$$

This inequality will be the main target we need to resolve in this proof.

First, we are to show

$$\min\left\{\frac{y_0}{x_0}, \frac{x_0}{y_0}\right\} > \frac{4}{\left(m + \sqrt{m^2+4}\right)^2}.$$

With the difference fixed as $m = (y_0 - x_0)/\sqrt{\mu}$, assuming $y_0 > x_0$, we have $m/y_0 = (1 - x_0/y_0)/\sqrt{\mu}$. if $x_0 y_0$ increases, both x_0 and y_0 increase then m/y_0 decreases, which means x_0/y_0 increases. As a result, we have

$$\min\left\{\frac{y_0}{x_0}, \frac{x_0}{y_0}\right\} > \min\left\{\frac{y_0}{x_0}, \frac{x_0}{y_0}\right\}\Bigg|_{x_0 y_0 = \mu} = \frac{4}{\left(m + \sqrt{m^2+4}\right)^2}.$$

Therefore, at initialization, to have positive x_1 and y_1, it is enough to require

$$\eta(x_0 y_0 - \mu) < \frac{4}{\left(m + \sqrt{m^2+4}\right)^2} \triangleq r.$$

From Theorem 3.4, it is guaranteed that $|x_t - y_t| < |x_0 - y_0|$ with $t \geqslant 2$ until it reaches $x_t y_t > \mu$, with which r is still a good lower bound for $\min\{y_t/x_t, x_t/y_t\}$. So what remains to show is it satisfies $\eta(x_t y_t - \mu) < r$ for the next first time $x_t y_t > \mu$. If this holds, we can always iteratively show, for any $x_t y_t > \mu$ along gradient descent,

$$\eta(x_t y_t - \mu) < r < \min\left\{\frac{y_t}{x_t}, \frac{x_t}{y_t}\right\}.$$

Note that r itself is independent of $x_t y_t$ and all the history, so it is ideal to compute a uniform

upper bound of $\eta(x_t y_t - \mu)$ with any pair of (x_{t-1}, y_{t-1}) satisfying $x_{t-1}y_{t-1} < \mu$. Actually it is possible, since we have $|x_{t-1} - y_{t-1}|$ bounded as in Theorem 3.4.

Assume $x_i y_i > \mu$ and it satisfies the condition of $\eta(x_i y_i - \mu) < r$ and $|x_i - y_i| < |x_0 - y_0|$. As in (B.31), we have

$$\frac{x_{i+1} - y_{i+1}}{x_i - y_i} = 1 - \eta\,(\mu - x_i y_i) \in (1, 1 + r). \tag{B.48}$$

Hence, it suffices to get the maximum value of $g(z)$, with $z \in (0, \mu)$, as

$$g(z) = z\,(1 + \eta(\mu - z))^2 + \eta(\mu - z)(1 + r)^2 (x_0 - y_0)^2, \tag{B.49}$$

which is from (B.33). Denote $\bar{z} = \operatorname{argmax} g(z)$. Obviously $\bar{z} < \frac{1}{3}(\mu + \frac{1}{\eta}) \triangleq z_b$, because the first term of $g(z)$ achieves maximum at $z = \frac{1}{3}(\mu + \frac{1}{\eta})$ and the second term is in a decreasing manner with z. Then let's take the derivative of $g(z)$ as

$$\begin{aligned} g'(z) &= (1 + \eta(\mu - z))\,(1 + \eta\mu - 3\eta z) - \eta(1 + r)^2 (x_0 - y_0)^2 \\ &= (1 + \eta(\mu - z)) \left(1 + \eta\mu - 3\eta z - \frac{\eta(1 + r)^2 (x_0 - y_0)^2}{1 + \eta(\mu - z)}\right), \end{aligned}$$

where the first term is always positive, so we have

$$1 + \eta\mu - 3\eta\bar{z} - \frac{\eta(1 + r)^2 (x_0 - y_0)^2}{1 + \eta(\mu - \bar{z})} = 0, \tag{B.50}$$

which means

$$\bar{z} = \frac{1}{3\eta}\left(1+\eta\mu - \frac{\eta(1+r)^2(x_0-y_0)^2}{1+\eta(\mu-\bar{z})}\right) \tag{B.51}$$

$$> \frac{1}{3\eta}\left(1+\eta\mu - \frac{\eta(1+r)^2(x_0-y_0)^2}{1+\eta(\mu-\frac{1}{3}(\mu+\frac{1}{\eta}))}\right) \tag{B.52}$$

$$= \frac{1}{3}\left(\mu+\frac{1}{\eta} - \frac{3(1+r)^2}{2(\eta+1)}(x_0-y_0)^2\right) \tag{B.53}$$

$$\triangleq z_s, \tag{B.54}$$

where the inequality is from $\bar{z} < \frac{1}{3}(\mu+\frac{1}{\eta})$. As a result, it is safe to say

$$g(z) \leqslant z\left(1+\eta(\mu-z)\right)^2\Big|_{z=z_b} + \eta(\mu-z)(1+r)^2(x_0-y_0)^2\Big|_{z=z_s} \tag{B.55}$$

$$= \frac{4}{27}(1+\eta\mu)^3\cdot\frac{1}{\eta} + \eta(1+r)^2\left(\frac{2}{3}\mu - \frac{1}{3\eta} + \frac{2}{\eta\mu+1}(x_0-y_0)^2\right)(x_0-y_0)^2, \tag{B.56}$$

with which we are able to compute max $\eta(g(z)-\mu)$, which is exactly the final result. □

B.9 Proof of Theorem 3.5

Theorem B.9 (Restatement of Theorem 3.5)**.** *In the above setting, consider a teacher neuron* $\tilde{w} = [1,0]$ *and set the learning rate* $\eta = Kd$ *with* $K \in (1, 1.1]$*. Initialize the student as* $\left\|w^{(0)}\right\| = v^{(0)} \triangleq \epsilon \in (0, 0.10]$ *and* $\langle w^{(0)}, \tilde{w}\rangle \geqslant 0$*. Then, for* $t \geqslant T_1 + 4$*,* $w_y^{(t)}$ *decays as*

$$w_y^{(t)} < 0.1\cdot(1-0.030K)^{t-T_1-4}, \quad T_1 \leqslant \left\lceil \log_{2.56}\frac{1.35}{\pi\beta^2}\right\rceil, \quad \beta = \left(1+\frac{1.1}{\pi}\right)\epsilon.$$

Proof sketch The proof is divided into two stages, depending on whether w_y grows or not. The key is that the change of w_y follows (omitting all superscripts t)

$$\frac{\Delta w_y}{w_y} \propto -vw_x + \frac{1}{\pi}\frac{\frac{w_y}{w_x}}{1+(\frac{w_y}{w_x})^2}, \quad w_y^{(t+1)} = \left|w_y + \Delta w_y\right|. \tag{B.57}$$

where the second term in $\Delta w_y/w_y$ is bounded in $[0, \frac{1}{2\pi}]$. In stage 1 where vw_x is relatively small, we show the growth ratio of w_y is smaller than those of w_x and vw_x, resulting in an upper bound of number of iterations for vw_x to reach $\frac{1}{2\pi}$, so $\max(w_y)$ is bounded too. Although the initialization is balanced as $v^{(0)} = \left\|w^{(0)}\right\|$ for simplicity of proof, $v - w_x$ is also bounded at the end of stage 1. From the beginning of stage 2, thanks to the relatively narrow range of K, we are able to compute the bounds of three variables (including $v - w_x$, vw_x and w_y) and they turn out to fall into a basin in the parameter space after four iterations. In this basin, w_y decays exponentially with a linear rate of 0.97 at most. □

Proof. We restate the update rules as

$$\begin{aligned}
\Delta v^{(t)} &:= v^{(t+1)} - v^{(t)} = Kw_x^{(t)}\left[(-v^{(t)}w_x^{(t)}+1) - v^{(t)}w_y^{(t)}\frac{w_y^{(t)}}{w_x^{(t)}} - \frac{1}{\pi}\left(\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right) - \frac{w_y^{(t)}}{w_x^{(t)}}\right)\right], \\
&= Kw_x^{(t)}\left[(-v^{(t)}w_x^{(t)}+1) - \frac{1}{\pi}\left(\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right) - \frac{w_x^{(t)}w_y^{(t)}}{\left\|w^{(t)}\right\|^2}\right)\right] \\
&\quad + K\frac{(w_y^{(t)})^2}{v^{(t)}}\left(-(v^{(t)})^2 + \frac{v^{(t)}w_y^{(t)}}{\pi\left\|w^{(t)}\right\|^2}\right)
\end{aligned} \tag{B.58}$$

$$\Delta w_x^{(t)} := w_x^{(t+1)} - w_x^{(t)} = Kv^{(t)}\left[(-v^{(t)}w_x^{(t)}+1) - \frac{1}{\pi}\left(\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right) - \frac{w_x^{(t)}w_y^{(t)}}{\left\|w^{(t)}\right\|^2}\right)\right], \tag{B.59}$$

$$\Delta w_y^{(t)} = w_y^{(t)} \cdot K\left(-(v^{(t)})^2 + \frac{v^{(t)}w_y^{(t)}}{\pi\left\|w^{(t)}\right\|^2}\right), \tag{B.60}$$

$$w_y^{(t+1)} = \left|w_y^{(t)} + \Delta w_y^{(t)}\right|. \tag{B.61}$$

For simplicity, we will omit all superscripts of time t unless clarification is necessary. From (B.61),

if the target is to show w_y decaying with a linear rate, it suffices to bound the factor term in (B.60) (by a considerable margin) as

$$-2 < K\left(-v^2 + \frac{vw_y}{\pi \|w\|^2}\right) < 0. \tag{B.62}$$

The technical part is the second inequality of (B.62). If $v, w_x > 0$, it is equivalent to

$$vw_x > \frac{w_x w_y}{\pi \|w\|^2} = \frac{w_x w_y}{\pi(w_x^2 + w_y^2)},$$

where the RHS is smaller than or equal to $\frac{1}{2\pi}$. Hence, $\frac{1}{2\pi}$ is a special threshold with which we will frequently compare vw_x. Another important variable to control is $v - w_x$ that reveals how the two layers are balanced. If it is too large, for the iteration $v^{(t+1)}w_x^{(t+1)}$ may explode as shown in the 2-D case.

The main idea of our proof is that

- Stage 1 with $vw_x \leqslant \frac{w_x w_y}{\pi\|w\|^2}$: in this stage, w_y grows but it grows in a smaller rate than that of v and w_x. Therefore, since we have an upper bound for vw_x to stay in this stage, we are able to compute the upper bound of #iterations to finish this stage, which is T_1 in the theorem. At the end of this stage, both of $v - w_x$ and w_y are bounded under our assumption of initialization.

- Stage 2 with $vw_x > \frac{w_x w_y}{\pi\|w\|^2}$: in this stage, w_y decreases. Since our range of a large learning rate is relatively narrow ($1 < K \leqslant 1.1$), we are able to compute bounds of $vw_x, v - w_x$ and w_y. After eight iterations, it falls into (and stays in) a bounded basin of these three terms, in which w_y decays at least in a linear rate.

Stage 1.

We are to show that, in the last iteration of this stage, there are three facts: 1) $vw_x \leqslant \frac{1}{2\pi}$, 2) $v - w_x \in [-0.017, 0.17]$, and 3) $w_y \leqslant 0.44$.

At initialization, we assume $v^{(0)} = \left\|w^{(0)}\right\|$. Denote $\alpha_0 = \arctan(w_y^{(0)}/w_x^{(0)}) \in [0, \pi/2]$. So for next iteration we have

$$w_y^{(1)} = v^{(0)} \left(1 + K\left(-(v^{(0)})^2 + \frac{1}{\pi}\sin\alpha_0\right)\right), \tag{B.63}$$

$$w_x^{(1)} = v^{(0)} \left[\cos\alpha_0 + K\left(1 - (v^{(0)})^2\cos\alpha_0 + \frac{\cos\alpha_0 \sin\alpha_0 - \alpha_0}{\pi}\right)\right]. \tag{B.64}$$

Apparently $w_y^{(1)}$ increases with α_0 increasing. And

$$\begin{aligned}
\partial_{\alpha_0} w_x^{(1)} &= v^{(0)} \left[-\sin\alpha_0 + K\left((v^{(0)})^2 \sin\alpha_0 + \frac{-\sin^2\alpha_0 + \cos^2\alpha_0 - 1}{\pi}\right)\right] \\
&= v^{(0)} \left[-\sin\alpha_0 + K\left(\left((v^{(0)})^2 - \frac{\sin\alpha_0}{\pi}\right)\sin\alpha_0 + \frac{-\sin^2\alpha_0}{\pi}\right)\right].
\end{aligned}$$

Since in stage 1 it holds $\Delta w_y > 0$ which means $-(v^{(0)})^2 + \frac{1}{\pi}\sin\alpha_0 > 0$ in (B.63). So it follows $\partial_{\alpha_0} w_x^{(1)} \leqslant 0$. Combining the above arguments, we have

$$\begin{aligned}
w_x^{(1)} &\geqslant w_x^{(1)}|_{\alpha_0=\frac{\pi}{2}} = \frac{K}{2} v^{(0)}, \\
w_y^{(1)} &\leqslant w_y^{(1)}|_{\alpha_0=\frac{\pi}{2}} = \left(1 + \frac{K}{\pi} - K(v^{(0)})^2\right) v^{(0)} \leqslant \left(1 + \frac{K}{\pi}\right) v^{(0)}, \\
\frac{w_y^{(1)}}{w_x^{(1)}} &\leqslant \frac{2 + \frac{2K}{\pi}}{K} \leqslant 2.7.
\end{aligned}$$

Regarding $\frac{v}{w_y}$, it has $v^{(0)} \geqslant w_y^{(0)}$ at initialization due to $v^{(0)} = \left\|w^{(0)}\right\|$. From (B.58, B.59, B.60), we have $v\Delta v = w_x \Delta w_x + w_y \Delta w_y$. So it holds $v\Delta v \geqslant y\Delta y$. Meanwhile, $\frac{\Delta w_y}{v} = K(-v w_y + \frac{w_y^2}{\pi\|w\|^2}) \in [0, \frac{K}{\pi}]$. From Lemma B.11, given $v^{(t)} \geqslant w_y^{(t)}$ and $\frac{\Delta w_y}{v} \in [0, 1]$ for any t in this stage, it always holds $v^{(t+1)} \geqslant w_y^{(t+1)}$.

Therefore, it is fair to say

$$\frac{v^{(1)} w_x^{(1)}}{(w_y^{(1)})^2} \geqslant \frac{1}{2.7}.$$

Additionally, to bound the term $vw_y/\|w\|^2$ in Δw_y, we would like to show it always has $vw_y \leqslant \|w\|^2$. At initialization, it naturally holds. Then, for the every next iteration, given it holds in the last iteration, we have

$$
\begin{aligned}
&(v+\Delta v)(w_y+\Delta w_y) - [(w_x+\Delta w_x)^2 + (w_y+\Delta w_y)^2] \\
&= (v + \frac{w_x\Delta w_x + w_y\Delta w_y}{v})(w_y+\Delta w_y) - [(w_x+\Delta w_x)^2 + (w_y+\Delta w_y)^2] \\
&= vw_y + v\Delta w_y + w_x\Delta w_x(\frac{w_y}{v} + \frac{\Delta w_y}{v}) + (w_y\Delta w_y + (\Delta w_y)^2)\frac{w_y}{v} - [(w_x+\Delta w_x)^2 + (w_y+\Delta w_y)^2] \\
&\leqslant vw_y + v\Delta w_y + w_y\Delta w_y\frac{w_y}{v} - (w_x^2 + w_y^2 + 2w_y\Delta w_y + (\Delta w_x)^2) \\
&\leqslant v\Delta w_y + w_y\Delta w_y\frac{w_y}{v} - 2w_y\Delta w_y - (\Delta w_x)^2 \\
&= v\Delta w_y(1-\frac{w_y}{v})^2 - (\Delta w_x)^2 \\
&\leqslant v\Delta w_y - (\Delta w_x)^2
\end{aligned}
$$

where the first equality uses $v\Delta v = w_x\Delta w_x + w_y\Delta w_y$, the first inequality uses the proven $v \geqslant w_y$ and $v \geqslant \Delta w_y$, the second inequality uses the assumption $vw_y \leqslant \|w\|^2$. Now we are to show $v\Delta w_y - (\Delta w_x)^2 \leqslant 0$. We have

$$
\begin{aligned}
v\Delta w_y - (\Delta w_x)^2 &\leqslant Kv^2\frac{w_y^2}{\pi\|w\|^2} - K^2v^2\left(1 - \frac{1}{2\pi} - \gamma^{(t)}\right)^2, \\
\gamma^{(t)} &= \frac{1}{\pi}\left(\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right) - \frac{w_x^{(t)}w_y^{(t)}}{\left\|w^{(t)}\right\|^2}\right).
\end{aligned}
$$

Since we have proven $w_y^{(1)}/w_x^{(1)} \leqslant 2.7$, it is easy to check that

$$
\frac{1}{\pi\left(1 + (\frac{w_x^{(1)}}{w_y^{(1)}})^2\right)} \leqslant (1 - \frac{1}{2\pi} - \gamma^{(1)})^2.
$$

As a result, $v\Delta w_y - (\Delta w_x)^2 \leqslant 0$ at time 1. Furthermore, by checking each term, $v\Delta w_y - (\Delta w_x)^2$

decreases with w_y/w_x decreasing. We will soon show that w_y/w_x itself decreases, by showing the growth ratio of w_x is larger than that of w_y.

Our target lower bound of the growth ratio of w_x is that

$$\frac{\Delta w_x}{w_x} \geqslant 1 - \frac{1}{\pi} - \gamma, \tag{B.65}$$

which is larger than the growth ratio of w_y bounded by $\frac{1}{\pi}$ due to $v\Delta w_y < \|w\|^2$. So it suffices to show $Kv/w_x \geqslant 1$. Assuming $Kv/w_x \geqslant 1$ for the current step, we need to show $Kv^{(t+1)}/w_x^{(t+1)} \geqslant 1$ also holds for the next step. Let's denote

$$A^{(t)} = K\left[(-v^{(t)}w_x^{(t)} + 1) - \frac{1}{\pi}\left(\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right) - \frac{w_x^{(t)}w_y^{(t)}}{\left\|w^{(t)}\right\|^2}\right)\right]. \tag{B.66}$$

Then

$$\begin{aligned}(v + \Delta v) - \frac{1}{K}(w_x + \Delta w_x) &\geqslant v + Aw_x - \frac{w_x}{K} - \frac{Av}{K} \\ &= (v - \frac{w_x}{K})(1 - KA) + v(K - \frac{1}{K})A. \end{aligned} \tag{B.67}$$

If $KA \leqslant 1$, since $K > 1$ and $A > 0$, we have (B.67) as positive, which is what we need. If $KA > 1$, then

$$\begin{aligned}(B.67) &\geqslant (v - \frac{w_x}{K})(1 - K^2) + v(K - \frac{1}{K})A \\ &= ((-K + A)v + w_x)\,(K - \frac{1}{K}),\end{aligned}$$

where the first inequality is due to $A \leqslant K$ and the assumption of $Kv^{(t)}/w_x^{(t)} \geqslant 1$. Then it suffices to show $(-K+A)v + w_x \geqslant (-K + \frac{1}{K})v + w_x \geqslant 0$. Note that $-K + 1/K \in (-0.2, 0]$ when $K \in (1, 1.1]$. It is easy to verify that $v^{(1)} \leqslant 5w_x^{(1)}$. Then, for the next step, we need to show $v^{(t+1)} \leqslant 5w_x^{(t+1)}$

given $v^{(t)} \leqslant 5w^{(t+1)}$. To prove this, we are to bound $v - w_x$, as

$$\begin{aligned} v^{(t+1)} - w_x^{(t+1)} &= (1-A)(v-w) + K\frac{w_y^2}{v}(-v^2 + \frac{vw_y}{\pi \left\|w\right\|^2}) \\ &\leqslant 0.4(v-w) + Kw_y\frac{w_y^2}{\pi \left\|w\right\|^2} \leqslant 0.4(v-w) + \frac{Kw_y}{\pi}, \end{aligned} \tag{B.68}$$

where the first inequality is due to, when $w_y/w_x \leqslant 2.7$,

$$\begin{aligned} A &= K\left[-v^{(t)}w_x^{(t)} + \frac{1}{\pi}\frac{w_x^{(t)}w_y^{(t)}}{\left\|w^{(t)}\right\|^2}\right] + K\left[1 - \frac{1}{\pi}\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right)\right] \\ &\geqslant K\left[1 - \frac{1}{\pi}\arctan\left(\frac{w_y^{(t)}}{w_x^{(t)}}\right)\right] \geqslant 0.6. \end{aligned}$$

We will later show that $v^{(t+1)} - w^{(t+1)} \geqslant -0.1(v^{(t)} - w^{(t)})$. Combining this with (B.68), it is safe to say

$$v^{(t+1)} - w^{(t+1)} \leqslant 0.4(v-w) + \frac{Kw_y}{\pi} \leqslant 0.4 \times 4w + \frac{K \times 5w}{\pi} \leqslant 4w,$$

where the second inequality is due to $v \leqslant 5w$ and $v \geqslant w_y$. Since $w^{(t+1)} \geqslant w^{(t)}$ (due to $A > 0$) in this stage, we have $v^{(t+1)} \leqslant 5w_x^{(t+1)}$.

Combining the above discussion, we have prove (B.65). Obviously, when $w_y/w_x \leqslant 2.7$, RHS of (B.65) is at least 0.55, larger than $1.1/\pi$, which is the upper bound of the $\Delta w_y/w_y$. As a result, w_y/w_x keeps decreasing.

The next step is to show the growing ratio of vw_x is much larger than that of w_y. From (B.59, B.60), it holds

$$\begin{aligned} v^{(t+1)}w_x^{(t+1)} &= (v+\Delta v)(w_x + \Delta w_x) \geqslant vw_x + KA(v^2 + w_x^2) + K^2A^2vw_x \\ &\geqslant vw_x(1+A)^2, \end{aligned}$$

where the first inequality is due to $\Delta w_y \geqslant 0$. It follows $v^{(t+1)}w_x^{(t+1)}/v^{(t)}w_x^{(t)} \geqslant 1.6^2 = 2.56$.

So far, we have shown the following facts: under the defined initialization at time 0, starting from time 1, we have

1. $vw_x \leqslant 1/2\pi$.
2. $\Delta w_x/w_x + 1 \geqslant 1.55$.
3. $\Delta w_y/w_y + 1 \leqslant 1 + K/\pi$.
4. $w_y/w_x \leqslant 2.7$ and keeps decreasing.
5. $v^{(t+1)}w_x^{(t+1)}/v^{(t)}w_x^{(t)} \geqslant 2.56$.
6. $v \geqslant w_y$.
7. $v\Delta w_y < (\Delta w_x)^2$.

Now we are to use the above facts to bound vw_x, w_y and $v - w_x$ to the end of stage 1.

For vw_x, in previous discussion, we have shown that $vw_x \leqslant \frac{1}{2\pi}$. Actually, there is another special value

$$\frac{w_x w_y}{\pi(w_x^2 + w_y^2)} = 0.104 \text{ when } w_y/w_x = 2.7. \tag{B.69}$$

This value is slightly larger than $1/4\pi$. Hence, we would like to split the analysis into three parts: in the **first step of stage 2**,

1. $vw_x \geqslant \frac{1}{2\pi}$.
2. $\frac{1}{4\pi} \leqslant vw_x < \frac{1}{2\pi}$.
3. $vw_x < \frac{1}{4\pi}$.

Note that, although we are discussing the stage 1 in this section, investigating the lower bound of the first step in stage 2 helps calculate the number of iterations in stage 1. Furthermore, it helps bound several variables in stage 1.

Case (I). If $vw_x \geqslant \frac{1}{2\pi}$ in first step of stage 2:

Since we have prove $\frac{v^{(1)}w_x^{(1)}}{(w_y^{(1)})^2} \geqslant 1/2.7$ and $v^{(t+1)}w_x^{(t+1)}/v^{(t)}w_x^{(t)} \geqslant 2.56$, the number of iterations for vw_x to reach $1/2\pi$ is at most

$$T_1 \leqslant \left\lceil \log_{2.56} \frac{\frac{1}{2\pi}}{(w_y^{(1)})^2/2.7} \right\rceil. \tag{B.70}$$

Meanwhile, starting from time 1, the growth ratio of w_y is

$$(w_y + \Delta w_y)/w_y \leqslant 1 + K(-v^2 + 1/\pi) \leqslant 1 + 1.1/\pi - (v^{(1)})^2 \leqslant 1 + 1.1/\pi - (w_y^{(1)})^2, \tag{B.71}$$

where the first inequality is due to $vw_y \leqslant \|w\|^2$, the second is due to $K > 1$ and the third is from $v \geqslant w_y$. Therefore, combining with (B.70), we can bound w_y in the end of stage 1 as

$$w_y \leqslant \left(1 + 1.1/\pi - (w_y^{(1)})^2\right)^{\left\lceil \log_{2.56} \frac{\frac{1}{2\pi}}{(w_y^{(1)})^2/2.7} \right\rceil}. \tag{B.72}$$

Since it initializes as $\left\|w^{(0)}\right\| \leqslant 0.1$, we have $w_y^{(1)} \leqslant 0.1(1 + 1.1/\pi) = 0.135$. Then, it can be verified that, when $w_y^{(1)} \in (0, 0.135]$, it holds

$$w_y \leqslant 0.44. \tag{B.73}$$

The next is to bound $v - w_x$. Combining the update rules of v and w_x in (B.58, B.59), we have

$$\begin{aligned}\Delta(v - w_x) &:= (v^{(t+1)} - w_x^{(t+1)}) - (v^{(t)} - w_x^{(t)}) \\ &= K(v - w_x)\left(vw_x - 1 + \frac{\arctan(w_y/w_x) - \frac{w_x w_y}{\|w\|^2}}{\pi}\right) + K\frac{w_y^2}{v}(-v^2 + \frac{vw_y}{\pi\|w\|^2}). \end{aligned} \tag{B.74}$$

Note that

$$-1 \leqslant vw_x - 1 + \frac{\arctan(w_y/w_x) - \frac{w_x w_y}{\|w\|^2}}{\pi} \leqslant -1 + \frac{\arctan(w_y/w_x)}{\pi}, \tag{B.75}$$

where the left is due to $vw_x > 0$ and , the right is from $\Delta w_y \geqslant 0$. When $w_y/w_x \leqslant 2.7$, the RHS follows $-1 + \frac{\arctan(w_y/w_x)}{\pi} \leqslant -0.6$. So combining both sides tells

$$1 + K\left(vw_x - 1 + \frac{\arctan(w_y/w_x) - \frac{w_x w_y}{\|w\|^2}}{\pi}\right) \in [-K + 1, 0.4] \subset [-0.1, 0.4]. \tag{B.76}$$

Since $\Delta w_y \geqslant 0$, we have $0 \leqslant K\frac{w_y^2}{v}(-v^2 + \frac{vw_y}{\pi\|w\|^2}) \leqslant \frac{K}{\pi}w_y\frac{w_y^2}{\|w\|^2}$. Note that at initialization $w_x^{(0)} \leqslant v^{(0)}$. Then it is easy to verify that

$$-0.01 \leqslant -0.1(v^{(0)} - w^{(0)}) \leqslant v^{(1)} - w^{(1)} \leqslant (1 + \frac{K}{\pi} - \frac{K}{2})v^{(0)} \leqslant 0.082. \tag{B.77}$$

Because the coefficient on the positive side in (B.76) is larger than $0.4 > 0.1$, it is appropriate to

upper bound the $v - w_x$ as

$$\begin{aligned} v - w_x &\leqslant \max\left\{0.082, 0.082 \cdot 0.4^T + \sum_{t=1}^{T} 0.4^{t-1} \frac{K}{\pi} w_y^{(t)} \frac{(w_y^{(t)})^2}{\left\|w^{(t)}\right\|^2}\right\} \\ &\leqslant \max\left\{0.082, 0.082 \cdot 0.4^T + \sum_{t=1}^{T} 0.4^{t-1} \frac{K}{\pi} w_y^{(t)} \frac{1}{1 + \frac{1}{2.7}\left(\frac{1.55}{1+K/\pi}\right)^{2(t-1)}}\right\} \\ &\leqslant \max\left\{0.082, 0.082 \cdot 0.4^T + \sum_{t=1}^{T} 0.4^{t-1} \frac{1.1 \cdot 4.4}{\pi} \frac{1}{1 + \frac{1}{2.7}\left(\frac{1.55}{1+1.1/\pi}\right)^{2(t-1)}}\right\}, \end{aligned}$$

where the second inequality is from the different growth ratios of w_x and w_y. Note that here we take all $T \geqslant 1$ and pick the largest value of RHS to bound w_y. It turns out

$$v - w_x \leqslant 0.17. \tag{B.78}$$

Furthermore, to lower bound $v - w_x$, since obviously $|v - w_x| \leqslant 0.17$, it follows

$$v - w_x \geqslant -0.1 \cdot |v - w_x|_{\max} \geqslant -0.017. \tag{B.79}$$

Case (II). If $\frac{1}{4\pi} \leqslant v w_x < \frac{1}{2\pi}$ in first step of stage 2:

Similar to the discussion in Case (I), we are able to compute the number of iterations for $v w_x$ to reach $1/4\pi$. It is at most

$$T_1 \leqslant \lceil \log_{2.56} \frac{\frac{1}{4\pi}}{(w_y^{(1)})^2/2.7} \rceil. \tag{B.80}$$

Accordingly, w_y is bounded as

$$w_y \leqslant \left(1 + 1.1/\pi - (w_y^{(1)})^2\right)^{\lceil \log_{2.56} \frac{\frac{1}{4\pi}}{(w_y^{(1)})^2/2.7} \rceil} \leqslant 0.37. \tag{B.81}$$

For simplicity, we just keep the bounds for $v - w_x$ as in Case (I), as

$$v - w_x \in [-0.017, 0.17]. \tag{B.82}$$

Case (III). If $vw_x < \frac{1}{4\pi}$ in first step of stage 2:

From the condition, we know $vw_x < \frac{1}{4\pi}$ as well in the last step of stage 1. Since $\Delta w_y > 0$ in stage 1, it tells

$$\frac{1}{\pi}\frac{w_x w_y}{\|w\|^2} < vw_x \leqslant \frac{1}{4\pi}, \tag{B.83}$$

which means

$$\max\{\frac{w_x}{w_y}, \frac{w_y}{w_x}\} \geqslant 2 + \sqrt{3}. \tag{B.84}$$

Since $2 + \sqrt{3} > 2.7$, if $w_y/w_x \geqslant 2 + \sqrt{3}$, then for time 1, $(v^{(1)}, w_x^{(1)}, w_y^{(1)})$ is already in the stage 2. However, it is not possible because $\left\|w^{(0)}\right\| = v^{(0)} \leqslant 0.1$, which means $v^{(1)}w_x^{(1)}$ can not reach $\frac{1}{\pi}\frac{2.7}{1+2.7^2}$.

Therefore, the only possible is $\frac{w_x}{w_y} \geqslant 2 + \sqrt{3}$. In this case, we are able to bound w_y as

$$w_y \leqslant (2 - \sqrt{3})w_x \leqslant (2 - \sqrt{3})\left(\sqrt{\frac{1}{4\pi} + 0.0085^2} + 0.0085\right) \leqslant 0.078, \tag{B.85}$$

where the second inequality is due to $vw_x \leqslant \frac{1}{4\pi}$ and $v - w_x \geqslant -0.017$. Note that here we still use the bound of $v - w_x$ from Case (I), although it is loose somehow but it is enough for our analysis.

We leave the analysis of the bound of number of iterations to the end of this section.

Stage 2.

In the case (I) of stage 1, where the first step in stage 2 is with $vw_x \geqslant \frac{1}{2\pi}$, it has $v - w_x \in$

$[-0.017, 0.17]$ and $w_y \leqslant 0.44$. In the case (II), where the first step of stage 2 is with $vw_x \in [\frac{1}{4\pi}, \frac{1}{2\pi}]$, it has $v - w_x \in [-0.017, 0.17]$ and $w_y \leqslant 0.37$. In the case (III), where the first step of stage 2 is with $vw_x \in [\frac{1}{4\pi}, \frac{1}{2\pi}]$, it has $v - w_x \in [-0.017, 0.17]$ and $w_y \leqslant 0.078$.

To upper bound vw_x in the first step of stage 2, there are two candidates. One is from the case (I),

$$\begin{aligned}
v^{(t+1)}w_x^{(t+1)} &= vw_x\left(1 + K(1 - vw_x - \frac{\arctan(\frac{w_y}{w_x}) - \frac{w_y/w_x}{1+(w_y/w_x)^2}}{\pi})\right)^2 + K\frac{w_x w_y^2}{v}\left(-v^2 + \frac{vw_y}{\pi \|w\|^2}\right) \\
&\quad + K(v - w_x)^2\left(1 + K(1 - vw_x - \frac{\arctan(\frac{w_y}{w_x}) - \frac{w_y/w_x}{1+(w_y/w_x)^2}}{\pi})\right) \\
&\leqslant vw_x\left(1 + K(1 - vw_x)\right)^2 + K\frac{w_x w_y^2}{w_x}\left(-vw_x + \frac{w_x w_y}{\pi \|w\|^2}\right) \\
&\quad + K(v - w_x)^2\left(1 + K(1 - vw_x)\right) \\
&\leqslant \frac{1}{2\pi}\left(1 + 1.1(1 - \frac{1}{2\pi})\right)^2 + 1.1 \cdot 0.44^2\left(-\frac{1}{4\pi} + \frac{1}{2\pi}\right) + 1.1 \cdot 0.17^2\left(1 + 1.1(1 - \frac{1}{2\pi})\right) \\
&\leqslant 0.668, \qquad \text{(B.86)}
\end{aligned}$$

where we use $vw_x \geqslant 1/4\pi$, $x/(1 + x^2) \leqslant 0.5$ for any x.

One is from the case (II),

$$\begin{aligned}
v^{(t+1)}w_x^{(t+1)} &\leqslant vw_x\left(1 + K(1 - vw_x)\right)^2 + K\frac{w_x w_y^2}{w_x}\left(-vw_x + \frac{w_x w_y}{\pi \|w\|^2}\right) \\
&\quad + K(v - w_x)^2\left(1 + K(1 - vw_x)\right) \\
&\leqslant \frac{1}{4\pi}\left(1 + 1.1(1 - \frac{1}{4\pi})\right)^2 + 1.1 \cdot 0.37^2\left(\frac{1}{2\pi}\right) + 1.1 \cdot 0.17^2\left(1 + 1.1(1 - \frac{1}{4\pi})\right) \\
&\leqslant 0.48, \qquad \text{(B.87)}
\end{aligned}$$

where we use $vw_x \leqslant 1/4\pi$, $x/(1 + x^2) \leqslant 0.5$ for any x.

Therefore, we can see that, in the first step of stage 2,

$$vw_x \leqslant 0.668. \tag{B.88}$$

Next we are going to show how the iteration goes in the stage 2. In Case (I), there are three facts:

1. $w_y \leqslant 0.44$.
2. $v - w_x \in [-0.017, 0.17]$.
3. $vw_x \in [\frac{1}{2\pi}, 0.668]$.

Similarly, in Case (II), there are three facts as well:

1. $w_y \leqslant 0.37$.
2. $v - w_x \in [-0.017, 0.17]$.
3. $vw_x \in [\frac{1}{4\pi}, \frac{1}{2\pi}]$.

The main idea is to find a basin that any iteration with the above properties (*i.e.*, in the interval) will converge to and then stay in. The method is to iteratively compute the ranges of the variables for several steps, thanks to the narrow range of K. Before explicitly computing the ranges, let's write down the computing method, depending on whether or not $vw_x \geqslant 1$.

Consider any iteration with $vw_x \in [m_1, m_2], v - w_x \in [d_1, d_2], w_y \leqslant e$, we compute the bounds of $v^{(t+1)}w_x^{(t+1)}, v^{(t+1)} - w_x^{(t+1)}, w_y^{(t+1)}$ in the following process (naturally assuming $d_1 < 0 < d_2$)

1. If $m_1 \geqslant 1$:
   (a) Compute $w_x \geqslant \sqrt{m_1 + (d_2/2)^2} - d_2/2 \triangleq f$.
   (b) Compute $\frac{w_y}{w_x} \leqslant e/f \triangleq g$.

(c) Compute $\frac{\arctan(w_y/w_x)-\frac{w_x w_y}{\|w\|^2}}{\pi} \leqslant \frac{\arctan(g)-g/(1+g^2)}{\pi} \triangleq h$.

(d) Compute $v^{(t+1)}w_x^{(t+1)} \geqslant m_2(1+1.1(1-m_2-h))^2 + 1.1(1-m_2-h)\max\{|d_1|,|d_2|\}^2 - 1.1e^2 m_2$. This is from

$$\begin{aligned} v^{(t+1)}w_x^{(t+1)} &\geqslant vw_x\left(1+K(1-vw_x-h)\right)^2 + K\frac{w_x w_y^2}{v}\left(-v^2+\frac{vw_y}{\pi\|w\|^2}\right) \\ &\quad + K(v-w_x)^2\left(1+K(1-vw_x-h)\right) \\ &\geqslant vw_x\left(1+K(1-vw_x-h)\right)^2 - Kw_y^2\cdot vw_x \\ &\quad + K(v-w_x)^2\left(1+K(1-vw_x-h)\right). \end{aligned}$$

(e) Compute $v^{(t+1)}w_x^{(t+1)} \leqslant m_1(1+1.0(1-m_1))^2$. This is due to $x(1+K(1-x))^2$ decreases with x increasing when $x \geqslant 1$.

(f) Compute $v^{(t+1)}-w_x^{(t+1)} \in [d_1(1+1.1(m_2-1+h)-1.1e^2\cdot(\sqrt{m_2+(d_2/2)^2}+d_2/2)), d_2(1+1.1(m_2-1+h))]$. This is due to

$$\Delta v - \Delta w_x = K(v-w_x)\left(vw_x - 1 + \frac{1}{\pi}(\arctan(\alpha) - \frac{w_x w_y}{\|w\|^2})\right) + K\frac{w_y^2}{v}\left(-v^2+\frac{vw_y}{\pi\|w\|^2}\right),$$

where $vw_x \geqslant 1$, the last term is between $-Kvw_y^2$ and 0.

(g) Compute $w_y^{(t+1)} \leqslant e\cdot\max\{|j_1|,|j_2|\}$, where

$$j_1 = 1 + 1.1\frac{\sqrt{m_1+(d_2/2)^2}+d_2/2}{\sqrt{m_1+(d_2/2)^2}-d_2/2}\cdot(-m_2), \tag{B.89}$$

$$j_2 = 1 + 1.0\frac{\sqrt{m_1+(d_1/2)^2}-d_1/2}{\sqrt{m_1+(d_1/2)^2}+d_1/2}\cdot(-m_1+\frac{1}{2\pi}). \tag{B.90}$$

This is due to

$$\frac{\Delta w_y}{w_y} = K\frac{v}{w_x}(-vw_x + \frac{1}{\pi}\frac{w_x w_y}{\|w\|^2}),$$

then we would like to have the smallest value as $j_1 - 1$ and the largest value as $j_2 - 1$. Since w_y is always non-negative, taking the maximum absolute value gives the upper bound.

2. If $m_2 < 1$:

   (a) Compute $w_x \geqslant \sqrt{m_1 + (d_2/2)^2} - d_2/2 \triangleq f$.

   (b) Compute $\frac{w_y}{w_x} \leqslant e/f \triangleq g$.

   (c) Compute $\frac{\arctan(w_y/w_x) - \frac{w_x w_y}{\|w\|^2}}{\pi} \leqslant \frac{\arctan(g) - g/(1+g^2)}{\pi} \triangleq h$.

   (d) Compute $v^{(t+1)} w_x^{(t+1)} \geqslant \min_{x \in [m_1, m_2]} x(1+1.0(1-x-h))^2 - 1.1e^2 x$. Compared with the case of $m_1 \geqslant 1$, we drop the term $1.1(1 - m_2 - h) \max\{|d_1|, |d_2|\}^2$ because it is possible to have $v - w_x = 0$ in some iterations.

   (e) Compute $v^{(t+1)} w_x^{(t+1)} \leqslant \max_{x \in [m_1, m_2]} x(1 + 1.1(1 - x))^2 + 1.1(1 - x) \max\{|d_1|, |d_2|\}^2$. Compared with the case of $m_1 \geqslant 1$, we add a term depending on the $|v - w_x|_{\max}$ because it enlarges $v w_x$ in the in-balanced case.

   (f) Compute $v^{(t+1)} - w_x^{(t+1)} \in [d_1(1+1.1(m_2-1+h) - 1.1e^2 \cdot (\sqrt{m_2 + (d_2/2)^2} + d_2/2)), d_2(1 + 1.1(m_2 - 1 + h))]$. In fact, a rigorous left bound should include more terms to select a minimum from. Here it is simple because it keeps $1 + K(m_1 - 1) \geqslant 0$ in the following computing, so we do not need to worry about the flipping sign of d_1 and d_2.

   (g) Compute $w_y^{(t+1)} \leqslant e \cdot \max\{|j_1|, |j_2|\}$, where j_1, j_2 are the same with those in the case of $m_1 \geqslant 1$.

Therefore, with the above process, we are able to brutally compute the ranges of $v^{(t+1)} w_x^{(t+1)}$, $v^{(t+1)} - w_x^{(t+1)}$, $w_y^{(t+1)}$ from the current ranges. Note that this process plays a role of building a mapping from one interval to another interval, which covers all points from the source interval. However, it is loose to some extent because gradient descent is a mapping from a point to another

point. The advantage of such a loose method is feasibility of obtaining bounds while losing tightness. To achieve tightness, later we will also include some wisdom in a point-to-point style.

Also note that, a nice way to combine tightness and efficiency in this method is to split and to merge intervals when necessary.

For Case (I):

Now we are to compute the ranges starting from the interval where $I = \{w_y \leqslant 0.44, v - w_x \in [-0.017, 0.17], vw_x \in [\frac{1}{2\pi}, 0.668]\}$. First, we split it into three intervals:

1. $I_1 = \{w_y \leqslant 0.44, v - w_x \in [-0.017, 0.17], vw_x \in [0.213, 0.4]\}$.
2. $I_2 = \{w_y \leqslant 0.44, v - w_x \in [-0.017, 0.17], vw_x \in [0.4, 0.668]\}$.
3. $I_{30} = \{w_y \leqslant 0.44, v - w_x \in [-0.017, 0.17], vw_x \in [\frac{1}{2\pi}, 0.213]\}$.

Then, following the above method with splitting and merging intervals, we have

1. Starting from I_1,
   (a) Step 1: I_1 mapps to $I_3 = \{w_y \leqslant 0.416, v - w_x \in [-0.162, 0.068], vw_x \in [0.55, 1.12131]\}$.
   (b) Step 2: Splitting I_3, we have
      i. $I_4 = \{w_y \leqslant 0.416, v - w_x \in [-0.162, 0.068], vw_x \in [0.55, 0.8]\}$.
      ii. $I_5 = \{w_y \leqslant 0.416, v - w_x \in [-0.162, 0.068], vw_x \in [0.8, 0.9]\}$.
      iii. $I_6 = \{w_y \leqslant 0.416, v - w_x \in [-0.162, 0.068], vw_x \in [0.9, 1.0]\}$.
      iv. $I_7 = \{w_y \leqslant 0.416, v - w_x \in [-0.162, 0.068], vw_x \in [1.0, 1.12131]\}$.

      Then, we have
      i. I_4 mapps to
         $I_8 = \{w_y \leqslant 0.214, v - w_x \in [-0.309, 0.0545], vw_x \in [0.942, 1.25786]\}$.
      ii. I_5 mapps to
         $I_9 = \{w_y \leqslant 0.0966, v - w_x \in [-0.335, 0.0613], vw_x \in [0.880, 1.19649]\}$.

iii. I_6 mapps to

$$I_{10} = \{w_y \leqslant 0.0756, v - w_x \in [-0.362, 0.068], vw_x \in [0.777894, 1.11178]\}.$$

iv. I_7 mapps to

$$I_{11} = \{w_y \leqslant 0.134, v - w_x \in [-0.394, 0.0782], vw_x \in [0.595, 1]\}.$$

(c) Step 3: Splitting and merging I_8, I_9, I_{10}, I_{11}, we have

i. $I_{12} = \{w_y \leqslant 0.134, v - w_x \in [-0.394, 0.078], vw_x \in [0.595, 0.777]\}$.

ii. $I_{13} = \{w_y \leqslant 0.214, v - w_x \in [-0.394, 0.078], vw_x \in [0.777, 1]\}$.

iii. $I_{14} = \{w_y \leqslant 0.214, v - w_x \in [-0.362, 0.068], vw_x \in [1, 1.11178]\}$.

iv. $I_{15} = \{w_y \leqslant 0.214, v - w_x \in [-0.309, 0.061], vw_x \in [1.11178, 1.25786]\}$.

Then, we have

i. I_{12} mapps to

$$I_{16} = \{w_y \leqslant 0.0372, v - w_x \in [-0.317, 0.061], vw_x \in [1.14493, 1.31246]\}.$$

ii. I_{13} mapps to

$$I_{17} = \{w_y \leqslant 0.0432, v - w_x \in [-0.448, 0.078], vw_x \in [0.943633, 1.24393]\}.$$

iii. I_{14} mapps to

$$I_{18} = \{w_y \leqslant 0.0662, v - w_x \in [-0.462, 0.077], vw_x \in [0.77846, 1]\}.$$

iv. I_{15} mapps to

$$I_{20} = \{w_y \leqslant 0.0998, v - w_x \in [-0.456, 0.0785], vw_x \in [0.550, 0.878]\}.$$

2. Starting from I_2,

(a) Step 1: I_2 mapps to $I_{21} = \{w_y \leqslant 0.332, v - w_x \in [-0.205, 0.114], vw_x \in [0.864, 1.25894]\}$

(b) Step 2: Splitting I_{21}, we have

i. $I_{22} = \{w_y \leqslant 0.332, v - w_x \in [-0.205, 0.114], vw_x \in [0.864, 1]\}$.

ii. $I_{23} = \{w_y \leqslant 0.332, v - w_x \in [-0.205, 0.114], vw_x \in [1, 1.125894]\}$.

Then, we have

i. I_{22} mapps to

$$I_{24} = \{w_y \leqslant 0.081, v - w_x \in [-0.336, 0.114], vw_x \in [0.858, 1.14813]\}.$$

ii. I_{23} mapps to

$$I_{25} = \{w_y \leqslant 0.184, v - w_x \in [-0.409, 0.148], vw_x \in [0.463, 1]\}.$$

(c) Step 3: Splitting and merging I_{24}, I_{25}, we have

i. $I_{26} = \{w_y \leqslant 0.184, v - w_x \in [-0.409, 0.148], vw_x \in [0.463, 1]\}$.

ii. $I_{27} = \{w_y \leqslant 0.081, v - w_x \in [-0.336, 0.114], vw_x \in [1, 1.14813]\}$.

Then, we have

i. I_{26} mapps to

$$I_{28} = \{w_y \leqslant 0.083, v - w_x \in [-0.452, 0.148], vw_x \in [0.952783, 1.31778]\}.$$

ii. I_{27} mapps to

$$I_{29} = \{w_y \leqslant 0.034, v - w_x \in [-0.399, 0.133], vw_x \in [0.777, 1]\}.$$

3. Starting from I_{30},

(a) Step 1: I_{30} mapps to $I_{31} = \{w_y \leqslant 0.44, v - w_x \in [-0.124, 0.037], vw_x \in [0.422, 0.767]\}$

(b) Step 2: Splitting I_{31}, we have

i. $I_{32} = \{w_y \leqslant 0.44, v - w_x \in [-0.124, 0.037], vw_x \in [0.422, 0.5]\}$.

ii. $I_{33} = \{w_y \leqslant 0.44, v - w_x \in [-0.124, 0.037], vw_x \in [0.5, 0.6]\}$.

iii. $I_{34} = \{w_y \leqslant 0.44, v - w_x \in [-0.124, 0.037], vw_x \in [0.6, 0.767]\}$.

Then, we have

i. I_{32} mapps to

$$I_{35} = \{w_y \leqslant 0.301, v - w_x \in [-0.218, 0.0185], vw_x \in [0.901, 1.20971]\}.$$

ii. I_{33} mapps to

$$I_{36} = \{w_y \leqslant 0.262, v - w_x \in [-0.245, 0.023], vw_x \in [0.96322, 1.25093]\}.$$

iii. I_{34} mapps to

$$I_{37} = \{w_y \leqslant 0.213, v - w_x \in [-0.288, 0.029], vw_x \in [0.947, 1.25345]\}.$$

(c) Step 3: Splitting and merging I_{35}, I_{36}, I_{37}, we have

i. $I_{38} = \{w_y \leqslant 0.301, v - w_x \in [-0.288, 0.029], vw_x \in [0.901, 1]\}$.

ii. $I_{39} = \{w_y \leqslant 0.301, v - w_x \in [-0.288, 0.029], vw_x \in [1, 1.1]\}$.

iii. $I_{40} = \{w_y \leqslant 0.301, v - w_x \in [-0.288, 0.029], vw_x \in [1.1, 1.25093]\}$.

iv. $I_{41} = \{w_y \leqslant 0.262, v - w_x \in [-0.245, 0.029], vw_x \in [1.25093, 1.25345]\}$.

Then, we have

i. I_{38} mapps to

$$I_{42} = \{w_y \leqslant 0.0404, v - w_x \in [-0.392, 0.029], vw_x \in [0.888, 1.11696]\}.$$

ii. I_{39} mapps to

$$I_{43} = \{w_y \leqslant 0.0740, v - w_x \in [-0.428, 0.033], vw_x \in [0.741, 1]\}.$$

iii. I_{40} mapps to

$$I_{44} = \{w_y \leqslant 0.125, v - w_x \in [-0.482, 0.038], vw_x \in [0.497, 0.891]\}.$$

iv. I_{41} mapps to

$$I_{45} = \{w_y \leqslant 0.109, v - w_x \in [-0.400, 0.038], vw_x \in [0.534, 0.702]\}.$$

(d) Step 4: Splitting and merging $I_{42}, I_{43}, I_{44}, I_{45}$, we have

i. $I_{46} = \{w_y \leqslant 0.125, v - w_x \in [-0.482, 0.038], vw_x \in [0.497, 0.891]\}$.

ii. $I_{47} = \{w_y \leqslant 0.074, v - w_x \in [-0.428, 0.033], vw_x \in [0.891, 1]\}$.

iii. $I_{48} = \{w_y \leqslant 0.041, v - w_x \in [-0.40, 0.029], vw_x \in [1, 1.11696]\}$.

Then, we have

i. I_{46} mapps to

$$I_{49} = \{w_y \leqslant 0.0424, v - w_x \in [-0.442, 0.034], vw_x \in [1.07853, 1.34708]\}.$$

ii. I_{47} mapps to

$I_{50} = \{w_y \leqslant 0.0110, v - w_x \in [-0.435, 0.033], vw_x \in [0.993, 1.13943]\}$.

iii. I_{48} mapps to

$I_{51} = \{w_y \leqslant 0.0109, v - w_x \in [-0.454, 0.033], vw_x \in [0.497, 0.891]\}$.

For Case (II):

Now we are to compute the ranges starting from the interval where $I = \{w_y \leqslant 0.37, v - w_x \in [-0.017, 0.17], vw_x \in [\frac{1}{4\pi}, \frac{1}{2\pi}]\}$. First, we denote it as

1. $I_{52} = \{w_y \leqslant 0.37, v - w_x \in [-0.017, 0.17], vw_x \in [\frac{1}{4\pi}, \frac{1}{2\pi}]$.

Then, following the above method with splitting and merging intervals, we have

1. Starting from I_{52},

   (a) Step 1: I_{52} mapps to $I_{53} = \{w_y \leqslant 0.37, v - w_x \in [-0.079, 0.0271], vw_x \in [0.222, 0.616]\}$.

   (b) Step 2: I_{53} mapps to $I_{54} = \{w_y \leqslant 0.343, v - w_x \in [-0.171, 0.017], vw_x \in [0.621, 1.24894]\}$.

   (c) Step 3: Splitting I_{54}, we have

      i. $I_{55} = \{w_y \leqslant 0.343, v - w_x \in [-0.171, 0.017], vw_x \in [0.621, 1\}$.

      ii. $I_{56} = \{w_y \leqslant 0.343, v - w_x \in [-0.171, 0.017], vw_x \in [1, 1.24894]\}$.

      Then, we have

      i. I_{55} mapps to

      $I_{57} = \{w_y \leqslant 0.150, v - w_x \in [-0.305, 0.017], vw_x \in [0.840, 1.25908]\}$.

      ii. I_{56} mapps to

      $I_{58} = \{w_y \leqslant 0.137, v - w_x \in [-0.367, 0.022], vw_x \in [0.472, 1]\}$.

   (d) Step 4: Splitting and merging I_{57}, I_{58}, we have

      i.

ii. $I_{59} = \{w_y \leqslant 0.150, v - w_x \in [-0.367, 0.022], vw_x \in [0.472, 1\}$.

iii.

iv. $I_{60} = \{w_y \leqslant 0.150, v - w_x \in [-0.305, 0.017], vw_x \in [1, 1.25908\}$.

Then, we have

i. I_{59} mapps to

$I_{61} = \{w_y \leqslant 0.0705, v - w_x \in [-0.393, 0.022], vw_x \in [0.971, 1.304]\}$.

ii. I_{60} mapps to

$I_{62} = \{w_y \leqslant 0.0613, v - w_x \in [-0.421, 0.0219], vw_x \in [0.583, 1]\}$.

For both Cases (I, II):

From $I_{16-20}, I_{28}, I_{29}, I_{49-51}, I_{61}, I_{62}$, we can see that it has fallen into an interval $I_f = \{w_y < 0.1, v - w_x \in [-0.462, 0.148], vw_x \in [0.497, 1.34078]\}$. Something special here is that w_y has been much smaller than w_x. More broadly, let's define an interval I_s generated by $I_g = \{w_y = 0, v - w_x \in [-0.464, 0.148], vw_x \in [1, 1.5]\}$. Here "generated" means

$$I_s = \bigcup_{T \geqslant t} \{(v^{(T)}, w_x^{(T)}, w_y^{(T)}) | (v_t^{(t)}, w_x^{(t)}, w_y^{(t)}) \in I_g\}. \tag{B.91}$$

Then each element $(v, w_x, w_y) \in I_s$ has the following properties:

1. $w_y = 0$.

2. $vw_x \in [0.181, 1.5]$.

3. If $vw_x \leqslant 1$, then $v - w_x \in [-0.735, 0.23]$. If $vw_x > 1$, then $v - w_x \in [-0.474, 0.148]$.

The first property is obvious. The third can be proven as follows: for each element $(v, w_x, w_y) \in I_g$, it has $v^{(t+1)} - w_x^{(t+1)} = (v - w_x)\,(1 + K(vw_x - 1))$, where the ratio $1 + K(vw_x - 1) \in [1, 1 + 1.1(1.5 - 1)]$ when $vw_x \in [1, 1.5]$. Furthermore, in the proven 2-D case, we have shown that "if $vw_x > 1$ with some mild conditions, then $\frac{v^{(t+2)} - w_x^{(t+2)}}{v - w_x} \in (-1, 1)$". Actually it can be tighter as $\frac{v^{(t+2)} - w_x^{(t+2)}}{v - w_x} \in$

$(-0.2, 1)$ because here $K \leqslant 1.1$ while the original bound is for $K \leqslant 1.5$. The condition of bounded $|v - w_x|$ can also be verified, the purpose of which is to keep v, w_x always positive. Then the bound $[-0.2, 1]$ will tell $v - w_x \in [-0.474, 0.148]$ on $vw_x \geqslant 1$, because

$$\frac{0.148}{0.474} > 0.2, \qquad \frac{0.474}{0.148} > 0.2.$$

For the second property, the left bound can be verified as

$$\min_{x\in[1,1.5]} x(1 + 1.1(1 - x))^2 + 1.1(1 - x) \cdot 0.474^2 = \left(x(1 + 1.1(1 - x))^2 + 1.1(1 - x) \cdot 0.474^2\right)\Big|_{x=1.5}$$
$$\geqslant 0.181.$$

The right bound can be verified as

$$\max_{x\in[0,1]} x(1 + 1.1(1 - x))^2 + 1.1(1 - x) * 0.735^2 < 1.5.$$

After proving these three properties, we would like to bound how far I_f is away from I_s. More precisely, the distance is measured by w_y. We are going to show w_y decays exponentially.

Remind the update rules in (B.58, B.59). Denote $\gamma = \frac{1}{\pi}(\arctan(\alpha) - \frac{w_x w_y}{\|w\|^2})$ again and $\delta = K\frac{w_y^2}{v}(-v^2 + \frac{vw_y}{\pi\|w\|^2})$, then it is

$$\Delta v = Kw_x(-vw_x + 1) - Kw_x\gamma + \delta, \tag{B.92}$$

$$\Delta w_x = Kv(-vw_x + 1) - Kv\gamma, \tag{B.93}$$

$$\delta \in [-Kvw_y^2, 0]. \tag{B.94}$$

Note that both γ and δ are very small, so we are to show their effects separately, which is enough to be a good approximation.

Consider an iteration where $v^{(t)}w_x^{(t)} > 1$ and the corresponding $\gamma^{(t)}$. Let's denote $v^{(t+1)}, w_x^{(t+1)}$

as the next parameters with **no corruption** from $\gamma^{(t)}$. Similarly, we denote $\hat{v}^{(t+1)}, \hat{w_x}^{(t+1)}$ are corrupted with $\gamma^{(t)}$. From the 2-D analysis, we know

$$\frac{v^{(t+2)} - w_x^{(t+2)}}{v^{(t)} - w_x^{(t)}} = (1 + K(v^{(t)} w_x^{(t)} - 1))(1 + K(v^{(t+1)} w_x^{(t+1)} - 1)) < 1. \tag{B.95}$$

We would like to show, with a small $\gamma^{(t)}$ and ignoring δ,

$$\frac{\hat{v}^{(t+2)} - \hat{w_x}^{(t+2)}}{v^{(t)} - w_x^{(t)}} = (1 + K(v^{(t)} w_x^{(t)} - 1 + \gamma^{(t)}))(1 + K(\hat{v}^{(t+1)} \hat{w_x}^{(t+1)} - 1 + \gamma^{(t+1)})) \lessapprox 1, \tag{B.96}$$

where $\gamma^{(t+1)}$ is in time $(t+1)$ accordingly. The difference of LHS of the above two expressions turns out to be

$$\begin{aligned}
(B.96) - (B.95) &= K\gamma^{(t)}(1 + K(v^{(t+1)} w_x^{(t+1)} - 1)) \\
&\quad + (1 + K(v^{(t)} w_x^{(t)} - 1))K(\hat{v}^{(t+1)} \hat{w_x}^{(t+1)} - v^{(t+1)} w_x^{(t+1)} + \gamma^{(t+1)}) + O(\gamma^2) \\
&= K\gamma^{(t)}(1 + K(v^{(t+1)} w_x^{(t+1)} - 1)) \\
&\quad + K(1 + K(v^{(t)} w_x^{(t)} - 1))(-K(v^{(t)})^2 \gamma^{(t)} - K(w_x^{(t)})^2 \gamma^{(t)} + \gamma^{(t+1)}) + O(\gamma^2) \\
&\leqslant K\gamma^{(t)}\left(1 + (1 + K(v^{(t)} w_x^{(t)} - 1))\big(- K(v^{(t)})^2 - K(w_x^{(t)})^2 + \frac{\gamma^{(t+1)}}{\gamma^{(t)}}\big)\right) + O(\gamma^2) \\
&\leqslant K\gamma^{(t)}\left(1 + (1 + K(v^{(t)} w_x^{(t)} - 1))\big(- 2Kv^{(t)} w_x^{(t)} + \frac{\gamma^{(t+1)}}{\gamma^{(t)}}\big)\right) + O(\gamma^2).
\end{aligned} \tag{B.97}$$

Since $\frac{\Delta w_x}{w_x} = K\frac{v}{w_x}(-v w_x + 1 - \gamma)$, we have

$$\frac{w_x^{(t+1)}}{w_x^{(t)}} = 1 + K\frac{v^{(t)}}{w_x^{(t)}}(-v^{(t)} w_x^{(t)} + 1 - \gamma^{(t)}) < 1. \tag{B.98}$$

Also we have

$$\frac{\gamma^{(t+1)}}{\gamma^{(t)}} = \frac{\arctan(\frac{w_y^{(t+1)}}{w_x^{(t+1)}}) - \frac{w_x^{(t+1)} w_y^{(t+1)}}{\left\|w^{(t+1)}\right\|^2}}{\arctan(\frac{w_y^{(t)}}{w_x^{(t)}}) - \frac{w_x^{(t)} w_y^{(t)}}{\left\|w^{(t)}\right\|^2}}. \tag{B.99}$$

Since $w_y^{(t+1)} \leqslant w_y^{(t)}$ and

$$\frac{\arctan(mx) - \frac{mx}{1+m^2x^2}}{\arctan(x) - \frac{x}{1+x^2}} \leqslant m^3, \text{ for any } m > 0, x > 0, \tag{B.100}$$

we have

$$\frac{\gamma^{(t+1)}}{\gamma^{(t)}} \leqslant \frac{1}{\left(1 + K\frac{v^{(t)}}{w_x^{(t)}}(-v^{(t)} w_x^{(t)} + 1 - \gamma^{(t)})\right)^3}. \tag{B.101}$$

For general $vw_x \in (1, 1.5]$, (B.101) holds as

$$\frac{\gamma^{(t+1)}}{\gamma^{(t)}} \lessapprox \frac{1}{(1 + 1.1\frac{\sqrt{1+0.074^2}+0.074}{\sqrt{1+0.074^2}-0.074}(-1.5+1))^3} \leqslant 22. \tag{B.102}$$

Since $1 + K(v^{(t)} w_x^{(t)} - 1) \leqslant 1 + 1.1 * 0.5 = 1.55$, it is fair to say

$$(B.96) - (B.95) \lessapprox K\gamma^{(t)}(1 + 1.55 * (-2 + 22)) + O(\gamma^2) = 35.2\gamma^{(t)} + O(\gamma^2). \tag{B.103}$$

Actually $\gamma^{(t)}$ is bounded by

$$\frac{w_y^{(t)}}{w_x^{(t)}} \leqslant \frac{0.099}{\sqrt{1 + 0.074^2} - 0.074} = 0.1066, \tag{B.104}$$

$$\gamma^{(t)} \leqslant \frac{\arctan(x) - \frac{x}{1+x^2}}{\pi} \leqslant 2.6 \times 10^{-4}. \tag{B.105}$$

As a result,

$$(B.96) - (B.95) \lessapprox 0.0084. \tag{B.106}$$

Note that this small value is very easy to cover in (B.95), requiring

$$1 - \frac{v^{(t+2)} - w_x^{(t+2)}}{v^{(t)} - w_x^{(t)}} \geqslant 0.0084, \tag{B.107}$$

except when vw_x is pretty close to 1. When $vw_x \longrightarrow 1$, from the analysis of 2-D case, (derived from the case of $x_{t+1}y_{t+1} \geqslant x_s^2$)

$$1 - \frac{v^{(t+2)} - w_x^{(t+2)}}{v^{(t)} - w_x^{(t)}} \geqslant (2K-2)(v^{(t)}w_x^{(t)} - 1). \tag{B.108}$$

For (B.96) − (B.95), denote a function $p(x)$ as

$$p(x) = 1 + (1+Kx)\left(-2K(x+1) + \frac{1}{\left(1 + K\frac{v}{w_x}(-x)\right)^3}\right), \tag{B.109}$$

where $x = v^{(t)}w_x^{(t)} - 1$ in (B.97, B.101). It is obvious that $p(0) = 1 + (-2K+1) < 0$. When x is small, it turns out

$$p(x) = -2K + 2 + K\left(-2K - 1 + 3\frac{v^{(t)}}{w_x^{(t)}}\right)x + O(x^2) \tag{B.110}$$

As a result, (B.96) − (B.95) < 0 when $vw_x - 1 = o(K-1)$. What if $vw_x - 1 = \Omega(K-1)$? Actually,

we can get a better bound by a more care analysis, as

$$\frac{(B.96)-(B.95)}{K\gamma^{(t)}} \leqslant 1+(1+K(v^{(t)}w_x^{(t)}-1))\big(-K(v^{(t)})^2-K(w_x^{(t)})^2+\frac{\gamma^{(t+1)}}{\gamma^{(t)}}\big) + K\left[v^{(t)}w_x^{(t)}(1+K(1-v^{(t)}w_x^{(t)}))^2-1\right], \tag{B.111}$$

where the last term is due to $v^{(t+1)}w_x^{(t+1)} \leqslant v^{(t)}w_x^{(t)}(1+K(1-v^{(t)}w_x^{(t)}))^2$. Hence, with this bound, by expanding the last term, (B.110) becomes

$$p(x) = -2K+2+K\left(-2K-1+3\frac{v^{(t)}}{w_x^{(t)}}\right)x+K(1-2K)x+O(x^2) \tag{B.112}$$

$$= -2K+2+K\left(-4K+3\frac{v^{(t)}}{w_x^{(t)}}\right)x+O(x^2), \tag{B.113}$$

which is definitely negative because

$$\frac{v^{(t)}}{w_x^{(t)}} \leqslant \frac{\sqrt{1+0.074^2}+0.074}{\sqrt{1+0.074^2}-0.074} < 1.16 < \frac{4}{3}. \tag{B.114}$$

Meanwhile, we are to prove the δ in (B.92) will not make $\tilde{I}_s$ make $v-w_x < -0.474$ starting from $v-w_x \geqslant -0.462$. First, in the region of $\{vw_x \in [1,1.5], v-w_x \leqslant 0.148\}$, we have $Kvw_y^2 \leqslant 1.1\cdot(\sqrt{1.5+0.074^2}+0.074)*0.1^2 \leqslant 0.0144$. Also note that in this region with $v-w_x \geqslant -0.462$, we have

$$\frac{w_y^{(t+1)}}{w_y} \leqslant 1-\frac{\sqrt{1+0.231^2}-0.231}{\sqrt{1+0.231^2}+0.231} = 0.37. \tag{B.115}$$

Hence $Kv(w_y^{(t)})^2+Kv(w_y^{(t+1)})^2 \leqslant 0.0144*(1+0.37^2) = 0.0164$. Since $|v^{(t+2)}-w^{(t+2)}| < |v^{(t)}-w^{(t)}|$ if there is no δ, we shall see that there is no need to discuss the case of $v-w_x \geqslant -0.462+0.0164 = -0.4456$ because it still holds $v^{(t+1)}-w_x^{(t+1)} > -0.462$. When $v^{(t)}-w_x^{(t)} \in [-0.462,-0.4456]$, we

shall see that in (B.36), after adding the term of δ in v,

$$\frac{v^{(t+2)} - w_x^{(t+2)}}{v^{(t)} - w_x^{(t)}} \leqslant 1 - (1 + K(vw_x - 1)) \cdot Kw_x\delta, \tag{B.116}$$

which means the absolute value of $v - w_x$ decays at least by a margin depending on δ. After multiplying the current difference $v^{(t)} - w_x^{(t)}$ on both side, it gives

$$(v^{(t+2)} - w_x^{(t+2)}) - (v^{(t)} - w_x^{(t)}) \geqslant v^{(t)}w_x^{(t)}w_x^{(t)}\delta. \tag{B.117}$$

Note that here $v^{(t+2)} - w_x^{(t+2)}$ does not include $\delta^{(t)}$ and $\delta^{(t+1)}$. As stated above, we have $\frac{\delta^{(t+1)}}{\delta^{(t)}} \leqslant 0.37^2 \leqslant 0.16$ due to the decay of w_y. So it is safe to say $\delta^{(t)} + \delta^{(t+1)} \geqslant 1.16\delta^{(t)}$. Combining with the above inequality, it gives

$$(v^{(t+2)} - w_x^{(t+2)}) - (v^{(t)} - w_x^{(t)}) + \delta^{(t)} + \delta^{(t+1)} \geqslant (v^{(t)}w_x^{(t)}w_x^{(t)} + 1.16)\delta^{(t)}, \tag{B.118}$$

where

$$v^{(t)}w_x^{(t)}w_x^{(t)} + 1.16 \leqslant vw_x \cdot (\sqrt{vw_x + (\frac{0.4456}{2})^2} - \frac{0.4456}{2}) + 1.16 \leqslant 0.6. \tag{B.119}$$

Furthermore, from our previous discussion, $w_y^{(t+2)} < w_y^{(t+2)}$ gives that the sum of (B.118) is bounded by

$$\frac{0.6}{1 - 0.16}\delta^{(t)} \geqslant \frac{0.6}{1 - 0.16} \cdot (-0.0144) \geqslant -0.0103. \tag{B.120}$$

Since $-0.474 - (-0.462) < -0.0103$, we shall see that the term of δ cannot drive $v - w_x < -0.472$. Note that (B.118) shall include a factor (< 1) in front of $\delta^{(t)}$, but we have ignored it to show a more aggressive bound.

Therefore, we are able to say an Interval $\hat{I}_s$ generated by I_f also has the following properties:

for each element $(v, w_x, w_y) \in \hat{I}_s$,

1. $vw_x \in [0.181, 1.5]$.

2. If $vw_x \leqslant 1$, then $v - w_x \in [-0.735, 0.23]$. If $vw_x > 1$, then $v - w_x \in [-0.472, 0.148]$.

Then the decreasing ratio of $\Delta w_y / w_y$ is bounded by

$$\frac{\Delta w_y}{w_y} = K\frac{v}{w_x}\left(-vw_x + \frac{1}{\pi}\frac{w_x w_y}{\|w\|^2}\right) \tag{B.121}$$

$$\in \left[-1.1(\sqrt{1.5 + 0.074^2} + 0.074)^2, -0.030K\right] \tag{B.122}$$

$$= [-1.87, -0.030K]. \tag{B.123}$$

Hence, w_y decays with a linear ratio of 0.97 (or $1 - 0.030K$) at most for Cases (I, II) in stage 2.

For Case (III), in the first step of stage 2, it already has $w_y \leqslant 0.078$ and $v - w_x \in [-0.017, 0.17]$. So surely it will also converge to I_s.

Here we present the time analysis for Case (III) of both stages. The number of iterations in the first stage is apparently similar to that of case (I, II), as

$$T_1 \leqslant \log_{2.56}\left\lceil\frac{2.7\psi}{\beta^2}\right\rceil, \tag{B.124}$$

where $\psi < \frac{1}{4\pi}$ is the value of vw_x in the first step of stage 2. In stage 2, since our target is to find how many steps are necessary to get $vw_x \geqslant 0.181$, so it is

$$v^{(t+1)}w_x^{(t+1)} \geqslant v^{(t)}w_x^{(t)}\left(1 - 0.181 + 1 - \frac{\arctan(2-\sqrt{3}) - \frac{2-\sqrt{3}}{1+(2-\sqrt{3})^2}}{\pi} - 1.1w_y^2\right) \tag{B.125}$$

$$\geqslant 3.28v^{(t)}w_x^{(t)}. \tag{B.126}$$

where obviously it still holds $\frac{w_y}{w_x} \leqslant 2 - \sqrt{3}$ and $w_y^2 < 0.1^2$ in stage 2. Since $3.28 > 2.56$, we have

the total number of steps to have $vw_x > 0.181$ bounded as

$$\begin{aligned}\left\lceil \log_{2.56} \frac{2.7\psi}{\beta^2} \right\rceil + \left\lceil \log_{3.28} \frac{0.181}{\psi} \right\rceil &\leqslant \left\lceil \log_{2.56} \frac{0.675}{\pi\beta^2} \right\rceil + \left\lceil \log_{3.28} \frac{0.181}{\frac{1}{4\pi}} \right\rceil + 2 \\ &\leqslant \left\lceil \log_{2.56} \frac{0.675}{\pi\beta^2} \right\rceil + 3 \\ &< \left\lceil \log_{2.56} \frac{1.35}{\pi\beta^2} \right\rceil + 4,\end{aligned}$$

which is not beyond the bound for Cases (I, II). □

B.10 Proof of Matrix Factorization

Consider a two-layer matrix factorization problem. It's parameterized by learnable weights $\mathbf{X} \in \mathbb{R}^{m\times p}$, $\mathbf{Y} \in \mathbb{R}^{p\times q}$, and the target matrix is $\mathbf{C} \in \mathbb{R}^{m\times q}$. The loss L is defined as

$$L(\mathbf{X}, \mathbf{Y}) = \frac{1}{2} \|\mathbf{XY} - \mathbf{C}\|_F^2 . \tag{B.127}$$

Obviously $\{\mathbf{X}, \mathbf{Y} : \mathbf{XY} = \mathbf{C}\}$ forms a minimum manifold. Focusing on this manifold, our targets are: 1) to prove our condition for stable oscillation on 1D functions holds at the minimum of L for any setting of dimensions, and 2) to provide an observation of walking towards flattest minima with theoretical intuition.

B.10.1 Asymmetric Case: 1D function at the minima

Before looking into the theorem, we would like to clarify the definition of the loss Hessian. Inherently, we squeeze $\mathbf{X}, \mathbf{Y}$ into a vector $\theta = \mathrm{vec}(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{mp+pq}$, which vectorizes the concatnation. As a result, we are able to represent the loss Hessian w.r.t. θ as a matrix in $\mathbb{R}^{(mp+pq)\times(mp+pq)}$. Meanwhile, the support of the loss landscape is in $\mathbb{R}^{mp+pq}$. In the following theorem, we are to show the leading eigenvector $\Delta \triangleq \mathrm{vec}(\Delta\mathbf{X}, \Delta\mathbf{Y}) \in \mathbb{R}^{mp+pq}$ of the loss Hessian. Since the cross section

of the loss landscape and Δ forms a 1D function f_Δ, we would also show the stable-oscillation condition on 1D function holds at the minima of f_Δ.

Theorem B.10. *For a matrix factorization problem, assume* $\mathbf{XY} = \mathbf{C}$. *Consider SVD of both matrices as* $\mathbf{X} = \sum_{i=1}^{\min\{m,p\}} \sigma_{x,i} u_{x,i} v_{x,i}^\top$ *and* $\mathbf{Y} = \sum_{i=1}^{\min\{p,q\}} \sigma_{y,i} u_{y,i} v_{y,i}^\top$, *where both groups of* $\sigma_{\cdot,i}$*'s are in descending order and both top singular values* $\sigma_{x,1}$ *and* $\sigma_{y,1}$ *are unique. Also assume* $v_{x,1}^\top u_{y,1} \neq 0$. *Then the leading eigenvector of the loss Hessian is* $\Delta = vec(C_1 u_{x,1} u_{y,1}^\top, C_2 v_{x,1} v_{y,1}^\top)$ *with* $C_1 = \frac{\sigma_{y,1}}{\sqrt{\sigma_{x,1}^2+\sigma_{y,1}^2}}, C_2 = \frac{\sigma_{x,1}}{\sqrt{\sigma_{x,1}^2+\sigma_{y,1}^2}}$. *Denote* f_Δ *as the 1D function at the cross section of the loss landscape and the line following the direction of* Δ *passing* $vec(\Delta\mathbf{X}, \Delta\mathbf{Y})$. *Then, at the minima of* f_Δ, *it satisfies*

$$3[f_\Delta^{(3)}]^2 - f_\Delta^{(2)} f_\Delta^{(4)} > 0. \tag{B.128}$$

Proof. To obtain the direction of the leading Hessian eigenvector at parameters $(\mathbf{X}, \mathbf{Y})$, consider a small deviation of the parameters as $(\mathbf{X}+\Delta\mathbf{X}, \mathbf{Y}+\Delta\mathbf{Y})$. With $\mathbf{XY} = \mathbf{C}$, evaluate the loss function as

$$L(\mathbf{X}+\Delta\mathbf{X}, \mathbf{Y}+\Delta\mathbf{Y}) = \frac{1}{2} \|\Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y} + \Delta\mathbf{X}\Delta\mathbf{Y}\|_F^2 . \tag{B.129}$$

Expand these terms and split them by orders of $\Delta\mathbf{X}, \Delta\mathbf{Y}$ as follows:

$$\Theta(\|\Delta\mathbf{X}\|^2 + \|\Delta\mathbf{Y}\|^2): \quad \frac{1}{2} \|\Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y}\|_F^2 , \tag{B.130}$$

$$\Theta(\|\Delta\mathbf{X}\|^3 + \|\Delta\mathbf{Y}\|^3): \quad \langle \Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y}, \Delta\mathbf{X}\Delta\mathbf{Y} \rangle, \tag{B.131}$$

$$\Theta(\|\Delta\mathbf{X}\|^4 + \|\Delta\mathbf{Y}\|^4): \quad \frac{1}{2} \|\Delta\mathbf{X}\Delta\mathbf{Y}\|_F^2 . \tag{B.132}$$

From the second-order terms, the leading eigenvector of $\nabla^2 L$ is the solution of

$$\mathrm{vec}(\Delta\mathbf{X}, \Delta\mathbf{Y}) = \underset{\|\Delta\mathbf{X}\|_F^2 + \|\Delta\mathbf{Y}\|_F^2 = 1}{\arg\max} \|\Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y}\|_F^2 . \tag{B.133}$$

Since both the top singular values of $\mathbf{X}, \mathbf{Y}$ are unique, the solution shall have both $\Delta\mathbf{X}, \Delta\mathbf{Y}$ of rank 1. Actually the solution is (here for simplicity we eliminate the sign of both)

$$\Delta\mathbf{X} = \frac{\sigma_{y,1}}{\sqrt{\sigma_{x,1}^2 + \sigma_{y,1}^2}} u_{x,1} u_{y,1}^\top, \quad \Delta\mathbf{Y} = \frac{\sigma_{x,1}}{\sqrt{\sigma_{x,1}^2 + \sigma_{y,1}^2}} v_{x,1} v_{y,1}^\top. \tag{B.134}$$

Equipped with the top eigenvector of Hessian, $\mathrm{vec}(\Delta\mathbf{X}, \Delta\mathbf{Y})$, we consider the 1-D function f_Δ generated by the cross-section of the loss landscape and the eigenvector, passing the minima $(\mathbf{X}, \mathbf{Y})$. Define the function as

$$f_\Delta(\mu) = L(\mathbf{X} + \mu\Delta\mathbf{X}, \mathbf{Y} + \mu\Delta\mathbf{Y}), \quad \mu \in \mathbb{R}. \tag{B.135}$$

Then, around $\mu = 0$, we have

$$f_\Delta(\mu) = \frac{1}{2} \|\Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y}\|_F^2 \cdot \mu^2 + \langle \Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y}, \Delta\mathbf{X}\Delta\mathbf{Y} \rangle \cdot \mu^3 + \frac{1}{2} \|\Delta\mathbf{X}\Delta\mathbf{Y}\|_F^2 \cdot \mu^4. \tag{B.136}$$

Therefore, the several order derivatives of $f_\Delta(\mu)$ at $\mu = 0$ can be obtained from Taylor expansion as

$$f_\Delta^{(2)}(0) = \|\Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y}\|_F^2, \tag{B.137}$$

$$f_\Delta^{(3)}(0) = 6\langle \Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y}, \Delta\mathbf{X}\Delta\mathbf{Y} \rangle, \tag{B.138}$$

$$f_\Delta^{(4)}(0) = 12 \|\Delta\mathbf{X}\Delta\mathbf{Y}\|_F^2. \tag{B.139}$$

Then we compute the condition of stable oscillation of 1-D function as

$$\left[3[f_\Delta^{(3)}]^2 - f_\Delta^{(2)} f_\Delta^{(4)}\right](0) = 108\langle \Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y}, \Delta\mathbf{X}\Delta\mathbf{Y} \rangle^2 - 12 \|\Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y}\|_F^2 \|\Delta\mathbf{X}\Delta\mathbf{Y}\|_F^2 \tag{B.140}$$

$$= 96 \|\Delta\mathbf{X}\mathbf{Y} + \mathbf{X}\Delta\mathbf{Y}\|_F^2 \|\Delta\mathbf{X}\Delta\mathbf{Y}\|_F^2 > 0, \tag{B.141}$$

because all of $\Delta\mathbf{X}\mathbf{Y}, \mathbf{X}\Delta\mathbf{Y}, \Delta\mathbf{X}\Delta\mathbf{Y}$ are parallel to $u_{x,1}v_{y,1}^\top$ and $v_{x,1}^\top u_{y,1} \neq 0$.

□

B.10.2 Quasi-symmetric case: walk towards flattest minima

Observation 5 (Restatement of Observation 2)**.** *Consider the quasi-symmetric matrix factorization with learning rate* $\eta = \frac{1}{\sigma_1^2} + \beta$. *Assume* $0 < \beta\sigma_1^2 < \sqrt{4.5} - 1 \approx 1.121$. *Consider a minimum* $(\mathbf{Y}_0 = \alpha\mathbf{X}_0, \mathbf{Z}_0 = {}^1/_\alpha\mathbf{X}_0), \alpha > 0$. *The initialization is around the minimum, as* $\mathbf{Y}_1 = \mathbf{Y}_0 + \Delta\mathbf{Y}_1, \mathbf{Z}_1 = \mathbf{Z}_0 + \Delta\mathbf{Z}_1$, *with the deviations satisfying* $u_1^\top \Delta\mathbf{Y}_1 v_1 \neq 0, u_1^\top \Delta\mathbf{Z}_1 v_1 \neq 0$ *and* $\|\Delta\mathbf{Y}_1\|, \|\Delta\mathbf{Z}_1\| \leqslant \epsilon$. *The second largest singular value of* $\mathbf{X}_0$ *needs to satisfy*

$$\eta \cdot \max\left\{\left(\frac{\sigma_1^2}{\alpha^2} + \sigma_2^2\alpha^2, \frac{\sigma_2^2}{\alpha^2} + \sigma_1^2\alpha^2\right)\right\} \leqslant 2. \tag{B.142}$$

Then GD would converge to a period-2 orbit γ_η *approximately with error in* $O(\epsilon)$, *formally written as*

$$(\mathbf{Y}_t, \mathbf{Z}_t) \to \gamma_\eta + (\Delta\mathbf{Y}, \Delta\mathbf{Z}), \qquad \|\Delta\mathbf{Y}\|, \|\Delta\mathbf{Z}\| = O(\epsilon), \tag{B.143}$$

$$\gamma_\eta = \left\{\left(\mathbf{Y}_0 + (\rho_i - \alpha)\,\sigma_1 u_1 v_1^\top, \mathbf{Z}_0 + (\rho_i - {}^1/_\alpha)\,\sigma_1 u_1 v_1^\top\right)\right\}, \qquad (i = 1, 2) \tag{B.144}$$

where $\rho_1 \in (1, 2), \rho_2 \in (0, 1)$ *are the two solutions of solving* ρ *in*

$$1 + \beta\sigma_1^2 = \frac{1}{\rho^2\left(\sqrt{\frac{1}{\rho^2} - \frac{3}{4}} + \frac{1}{2}\right)}. \tag{B.145}$$

Remark 7. What is missing for a rigorous proof?

1. Control of error terms in non-asymptotic analysis.
2. Resolving assumptions of spectrum $\mathbf{Q}_{\alpha,\eta,p}(y_t, z_t)$ in early stages.

Proof. Without loss of generality, we assume $\mathbf{X}_0 = \mathrm{diag}([\sigma_1, \sigma_2, \ldots, \sigma_d]) \in \mathbb{R}^{d\times d}$, where $(\mathbf{X}_0)_{i,i} = \sigma_i$ or 0 in all other entries. This can be easily achieved by rotating singular vectors of $\mathbf{X}_0$. Accordingly, we have $\mathrm{Y}_0 = \mathrm{diag}([\sigma_1\alpha, \sigma_2\alpha, \ldots, \sigma_d\alpha]) \in \mathbb{R}^{d\times d}$ and $\mathbf{Z}_0 = \mathrm{diag}([\sigma_1/\alpha, \sigma_2/\alpha, \ldots, \sigma_d/\alpha]) \in \mathbb{R}^{d\times d}$.

Starting from time $t = 1$, we denote the learnable parameter matrices as $\mathrm{Y}_t, \mathbf{Z}_t$, and their deviation as $\Delta\mathrm{Y}_t \triangleq \mathrm{Y}_t - \mathrm{Y}_0, \Delta\mathbf{Z}_t \triangleq \mathbf{Z}_t - \mathbf{Z}_0$. By assumptions, we have $\|\Delta\mathrm{Y}_1\| < \epsilon, \|\Delta\mathbf{Z}_1\| < \epsilon$. Furthermore, we split $\Delta\mathrm{Y}_t, \Delta\mathbf{Z}_t$ as follows,

$$\Delta\mathrm{Y}_t = \left[\begin{array}{c|c} ①_t & ③_t \\ \hline ②_t & ④_t \end{array}\right], \Delta\mathbf{Z}_t = \left[\begin{array}{c|c} ⑤_t & ⑦_t \\ \hline ⑥_t & ⑧_t \end{array}\right], \tag{B.146}$$

$$①_t, ⑤_t \in \mathbb{R}, \quad ②_t, ⑥_t \in \mathbb{R}^{(d-1)\times 1}, \quad ③_t, ⑦_t \in \mathbb{R}^{1\times(d-1)}, \quad ④_t, ⑧_t \in \mathbb{R}^{(d-1)\times(d-1)}. \tag{B.147}$$

Since the update rules of $\mathrm{Y}_t, \mathbf{Z}_t$ are

$$\mathrm{Y}_{t+1} = \mathrm{Y}_t - \eta\left(\Delta\mathrm{Y}_t\mathbf{Z}_0^\top + \mathrm{Y}_0\Delta\mathbf{Z}_t^\top + \Delta\mathrm{Y}_t\Delta\mathbf{Z}_t^\top\right)(\mathbf{Z}_0 + \Delta\mathbf{Z}_t) \tag{B.148}$$

$$\mathbf{Z}_{t+1} = \mathbf{Z}_t - \eta\left(\Delta\mathbf{Z}_t\mathrm{Y}_0^\top + \mathbf{Z}_0\Delta\mathrm{Y}_t^\top + \Delta\mathbf{Z}_t\Delta\mathrm{Y}_t^\top\right)(\mathrm{Y}_0 + \Delta\mathrm{Y}_t) \tag{B.149}$$

The update rules of ① − ⑧ are

$$①_{t+1} = ①_t - \eta \mathbb{I}_1^\top \left(\Delta \mathbf{Y}_t \mathbf{Z}_0^\top + \mathbf{Y}_0 \Delta \mathbf{Z}_t^\top + \Delta \mathbf{Y}_t \Delta \mathbf{Z}_t^\top\right) (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \tag{B.150}$$

$$②_{t+1} = ②_t - \eta \mathbb{I}_{\geqslant 2}^\top \left(\Delta \mathbf{Y}_t \mathbf{Z}_0^\top + \mathbf{Y}_0 \Delta \mathbf{Z}_t^\top + \Delta \mathbf{Y}_t \Delta \mathbf{Z}_t^\top\right) (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \tag{B.151}$$

$$③_{t+1} = ③_t - \eta \mathbb{I}_1^\top \left(\Delta \mathbf{Y}_t \mathbf{Z}_0^\top + \mathbf{Y}_0 \Delta \mathbf{Z}_t^\top + \Delta \mathbf{Y}_t \Delta \mathbf{Z}_t^\top\right) (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \tag{B.152}$$

$$④_{t+1} = ④_t - \eta \mathbb{I}_{\geqslant 2}^\top \left(\Delta \mathbf{Y}_t \mathbf{Z}_0^\top + \mathbf{Y}_0 \Delta \mathbf{Z}_t^\top + \Delta \mathbf{Y}_t \Delta \mathbf{Z}_t^\top\right) (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \tag{B.153}$$

$$⑤_{t+1} = ⑤_t - \eta \mathbb{I}_1^\top \left(\Delta \mathbf{Z}_t \mathbf{Y}_0^\top + \mathbf{Z}_0 \Delta \mathbf{Y}_t^\top + \Delta \mathbf{Z}_t \Delta \mathbf{Y}_t^\top\right) (\mathbf{Y}_0 + \Delta \mathbf{Y}_t) \mathbb{I}_1 \tag{B.154}$$

$$⑥_{t+1} = ⑥_t - \eta \mathbb{I}_{\geqslant 2}^\top \left(\Delta \mathbf{Z}_t \mathbf{Y}_0^\top + \mathbf{Z}_0 \Delta \mathbf{Y}_t^\top + \Delta \mathbf{Z}_t \Delta \mathbf{Y}_t^\top\right) (\mathbf{Y}_0 + \Delta \mathbf{Y}_t) \mathbb{I}_1 \tag{B.155}$$

$$⑦_{t+1} = ⑦_t - \eta \mathbb{I}_1^\top \left(\Delta \mathbf{Z}_t \mathbf{Y}_0^\top + \mathbf{Z}_0 \Delta \mathbf{Y}_t^\top + \Delta \mathbf{Z}_t \Delta \mathbf{Y}_t^\top\right) (\mathbf{Y}_0 + \Delta \mathbf{Y}_t) \mathbb{I}_{\geqslant 2} \tag{B.156}$$

$$⑧_{t+1} = ⑧_t - \eta \mathbb{I}_{\geqslant 2}^\top \left(\Delta \mathbf{Z}_t \mathbf{Y}_0^\top + \mathbf{Z}_0 \Delta \mathbf{Y}_t^\top + \Delta \mathbf{Z}_t \Delta \mathbf{Y}_t^\top\right) (\mathbf{Y}_0 + \Delta \mathbf{Y}_t) \mathbb{I}_{\geqslant 2}, \tag{B.157}$$

where $\mathbb{I}_1 = (\mathbb{I}_d)_{:,1} \in \mathbb{R}^{d\times 1}, \mathbb{I}_{\geqslant 2} = (\mathbb{I}_d)_{:,2:d} \in \mathbb{R}^{d\times(d-1)}$ are the dimension-reduction matrix, defined from blocks of the $d \times d$ identity matrix $\mathbb{I}$. In other words, $\mathbb{I}_1$ (respectively $\mathbb{I}_{\geqslant 2}$) is to pick the first row/column (respectively all remaining rows/columns) from a matrix, which is extracting $①_t - ⑧_t$ from $\Delta \mathbf{Y}_t, \Delta \mathbf{Z}_t$.

Denote $\mathbf{M}_t \triangleq \left(\Delta \mathbf{Y}_t \mathbf{Z}_0^\top + \mathbf{Y}_0 \Delta \mathbf{Z}_t^\top + \Delta \mathbf{Y}_t \Delta \mathbf{Z}_t^\top\right) = \mathbf{Y}_t \mathbf{Z}_t^\top - \mathbf{X}_0 \mathbf{X}_0^\top$.

At initialization, we assume all of $①_1, ②_1, ③_1, ④_1, ⑤_1, ⑥_1, ⑦_1, ⑧_1$ are in $\Theta(\epsilon)$, which means all $\|\mathbb{I}_1 \mathbf{M}_1 \mathbb{I}_1\|, \|\mathbb{I}_{\geqslant 2} \mathbf{M}_1 \mathbb{I}_1\|, \|\mathbb{I}_1 \mathbf{M}_1 \mathbb{I}_{\geqslant 2}\|, \|\mathbb{I}_{\geqslant 2} \mathbf{M}_1 \mathbb{I}_{\geqslant 2}\|$ are in $\Theta(\epsilon)$ as well. Our goal is to show that, as $t \to \infty$,

1. $①_\infty, ⑤_\infty$ are in a period-2 orbit,
2. $②_\infty, ③_\infty, ④_\infty, ⑥_\infty, ⑦_\infty, ⑧_\infty$ are in $\Theta(\epsilon)$,
3. $\|\mathbb{I}_{\geqslant 2} \mathbf{M}_\infty \mathbb{I}_1\|, \|\mathbb{I}_1 \mathbf{M}_\infty \mathbb{I}_{\geqslant 2}\|, \|\mathbb{I}_{\geqslant 2} \mathbf{M}_\infty \mathbb{I}_{\geqslant 2}\|, \left\|\mathbb{I}_1^\top \mathbf{Z}_\infty \mathbf{Z}_\infty^\top \mathbb{I}_{\geqslant 2}\right\|, \left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_\infty \mathbf{Y}_\infty^\top \mathbb{I}_{\geqslant 2}\right\|$ decay to zero.

Then, following the above definitions, we have another representation of

$$\left(\Delta \mathbf{Y}_t \mathbf{Z}_0^\top + \mathbf{Y}_0 \Delta \mathbf{Z}_t^\top + \Delta \mathbf{Y}_t \Delta \mathbf{Z}_t^\top\right),$$

or equivalently its transpose $\left(\Delta \mathbf{Z}_t \mathbf{Y}_0^\top + \mathbf{Z}_0 \Delta \mathbf{Y}_t^\top + \Delta \mathbf{Z}_t \Delta \mathbf{Y}_t^\top\right)$, as

$$\mathbb{I}_1^\top \left(\Delta \mathbf{Y}_t \mathbf{Z}_0^\top + \mathbf{Y}_0 \Delta \mathbf{Z}_t^\top + \Delta \mathbf{Y}_t \Delta \mathbf{Z}_t^\top\right) \mathbb{I}_1 = ①_t \mathbb{I}_1^\top \mathbf{Z}_0^\top \mathbb{I}_1 + \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 ⑤_t + ①_t ⑤_t + ③_t ⑦_t^\top \tag{B.158}$$

$$\mathbb{I}_{\geqslant 2}^\top \left(\Delta \mathbf{Y}_t \mathbf{Z}_0^\top + \mathbf{Y}_0 \Delta \mathbf{Z}_t^\top + \Delta \mathbf{Y}_t \Delta \mathbf{Z}_t^\top\right) \mathbb{I}_1 = ②_t \mathbb{I}_1^\top \mathbf{Z}_0^\top \mathbb{I}_1 + \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} ⑦_t^\top + ②_t ⑤_t + ④_t ⑦_t^\top \tag{B.159}$$

$$\mathbb{I}_1^\top \left(\Delta \mathbf{Y}_t \mathbf{Z}_0^\top + \mathbf{Y}_0 \Delta \mathbf{Z}_t^\top + \Delta \mathbf{Y}_t \Delta \mathbf{Z}_t^\top\right) \mathbb{I}_{\geqslant 2} = ③_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} + \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 ⑥_t^\top + ①_t ⑥_t^\top + ③_t ⑧_t^\top \tag{B.160}$$

$$\mathbb{I}_{\geqslant 2}^\top \left(\Delta \mathbf{Y}_t \mathbf{Z}_0^\top + \mathbf{Y}_0 \Delta \mathbf{Z}_t^\top + \Delta \mathbf{Y}_t \Delta \mathbf{Z}_t^\top\right) \mathbb{I}_{\geqslant 2} = ④_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} + \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} ⑧_t^\top + ②_t ⑥_t^\top + ④_t ⑧_t^\top . \tag{B.161}$$

After substituting with $①_{t+1} - ⑧_{t+1}$, we have

$$
\begin{aligned}
\mathbb{I}_1^\top \mathbf{M}_{t+1} \mathbb{I}_1 = {} & \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0^\top \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right)^\top \mathbf{M}_t \mathbb{I}_1 \\
& - \eta ①_t \mathbb{I}_1^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right)^\top \mathbf{M}_t \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 ⑤_t - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} ⑦_t^\top \\
& - \eta ③_t \mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right)^\top \mathbf{M}_t \mathbb{I}_1 + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right)^\top \mathbf{M}_t \mathbb{I}_1 \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \mathbb{I}_1^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right)^\top \mathbf{M}_t \mathbb{I}_1 \\
= {} & \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0^\top \mathbb{I}_1 \\
& - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t^\top \mathbb{I}_1 \\
& - \eta ①_t \mathbb{I}_1^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 ⑤_t \\
& - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} ⑦_t^\top - \eta ③_t \mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_1 \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_1 \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \mathbb{I}_1^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_1 \\
= {} & \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0^\top \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑤_t \mathbb{I}_1^\top \mathbf{Z}_0^\top \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} ⑥_t \mathbb{I}_1^\top \mathbf{Z}_0^\top \mathbb{I}_1 \\
& - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 ①_t \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 ②_t^\top \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 \\
& - \eta ①_t \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 - \eta ①_t ①_t \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 - \eta ①_t ②_t^\top \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 \\
& - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1 ⑤_t - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑤_t ⑤_t - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} ⑥_t ⑤_t \\
& - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} ⑦_t^\top - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑦_t ⑦_t^\top - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑧_t ⑦_t^\top \\
& - \eta ③_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 - \eta ③_t ④_t^\top \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 - \eta ③_t ③_t^\top \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} + \mathbb{I}_1 ⑦_t + \mathbb{I}_{\geqslant 2} ⑧_t) (\mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top + ③_t^\top \mathbb{I}_1^\top + ④_t^\top \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_1 \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1 + \mathbb{I}_1 ⑤_t + \mathbb{I}_{\geqslant 2} ⑥_t) (\mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top + ①_t \mathbb{I}_1^\top + ②_t^\top \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_1
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_{t+1}\mathbb{I}_1 = {} & \mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0^{\top}\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_t\right)^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& - \eta\textcircled{2}_t\mathbb{I}_1^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_t\right)^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_1\textcircled{5}_t - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_{\geqslant 2}\textcircled{7}_t^{\top} \\
& - \eta\textcircled{4}_t\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_t\right)^{\top}\mathbf{M}_t\mathbb{I}_1 + \eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_t\right)^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& + \eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_1\mathbb{I}_1^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_t\right)^{\top}\mathbf{M}_t\mathbb{I}_1 \\
= {} & \mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbb{I}_1\mathbb{I}_1^{\top}+\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top})(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0^{\top}\mathbb{I}_1 \\
& - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_t\right)^{\top}(\mathbb{I}_1\mathbb{I}_1^{\top}+\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top})\mathbf{M}_t\mathbb{I}_1 \\
& - \eta\textcircled{2}_t\mathbb{I}_1^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_t\right)^{\top}(\mathbb{I}_1\mathbb{I}_1^{\top}+\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top})\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbb{I}_1\mathbb{I}_1^{\top}+\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top})(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_1\textcircled{5}_t \\
& - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbb{I}_1\mathbb{I}_1^{\top}+\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top})(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_{\geqslant 2}\textcircled{7}_t^{\top} - \eta\textcircled{4}_t\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_t\right)^{\top}(\mathbb{I}_1\mathbb{I}_1^{\top}+\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top})\mathbf{M}_t\mathbb{I}_1 \\
& + \eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbb{I}_1\mathbb{I}_1^{\top}+\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top})(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_t\right)^{\top}(\mathbb{I}_1\mathbb{I}_1^{\top}+\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top})\mathbf{M}_t\mathbb{I}_1 \\
& + \eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbb{I}_1\mathbb{I}_1^{\top}+\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top})(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_1\mathbb{I}_1^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_t\right)^{\top}(\mathbb{I}_1\mathbb{I}_1^{\top}+\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top})\mathbf{M}_t\mathbb{I}_1 \\
= {} & \mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0^{\top}\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1\textcircled{5}_t\mathbb{I}_1^{\top}\mathbf{Z}_0^{\top}\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_{\geqslant 2}\textcircled{6}_t\mathbb{I}_1^{\top}\mathbf{Z}_0^{\top}\mathbb{I}_1 \\
& - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\textcircled{3}_t^{\top}\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\textcircled{4}_t^{\top}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& - \eta\textcircled{2}_t\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\textcircled{2}_t\textcircled{1}_t\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\textcircled{2}_t\textcircled{2}_t^{\top}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1\textcircled{5}_t - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1\textcircled{5}_t\textcircled{5}_t - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_{\geqslant 2}\textcircled{6}_t\textcircled{5}_t \\
& - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Z}_0\mathbb{I}_{\geqslant 2}\textcircled{7}_t^{\top} - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1\textcircled{7}_t\textcircled{7}_t^{\top} - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_{\geqslant 2}\textcircled{8}_t\textcircled{7}_t^{\top} \\
& - \eta\textcircled{4}_t\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1 - \eta\textcircled{4}_t\textcircled{3}_t^{\top}\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\textcircled{4}_t\textcircled{4}_t^{\top}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& + \eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Z}_0\mathbb{I}_{\geqslant 2}+\mathbb{I}_1\textcircled{7}_t+\mathbb{I}_{\geqslant 2}\textcircled{8}_t)(\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}+\textcircled{3}_t^{\top}\mathbb{I}_1^{\top}+\textcircled{4}_t^{\top}\mathbb{I}_{\geqslant 2}^{\top})\mathbf{M}_t\mathbb{I}_1 \\
& + \eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1^{\top}+\mathbb{I}_1\textcircled{5}_t+\mathbb{I}_{\geqslant 2}\textcircled{6}_t)(\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1\mathbb{I}_1^{\top}+\textcircled{1}_t\mathbb{I}_1^{\top}+\textcircled{2}_t^{\top}\mathbb{I}_{\geqslant 2}^{\top})\mathbf{M}_t\mathbb{I}_1
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{I}_1^\top \mathbf{M}_{t+1} \mathbb{I}_{\geqslant 2} = {} & \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& - \eta ①_t \mathbb{I}_1^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 ⑥_t^\top - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} ⑧_t^\top \\
& - \eta ③_t \mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \mathbb{I}_1^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
= {} & \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& - \eta ①_t \mathbb{I}_1^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 ⑥_t^\top \\
& - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} ⑧_t^\top - \eta ③_t \mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \mathbb{I}_1^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
= {} & \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑦_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} ⑧_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 ①_t \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 ②_t^\top \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& - \eta ①_t \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta ①_t ①_t \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta ①_t ②_t^\top \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1 ⑥_t^\top - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑤_t ⑥_t^\top - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} ⑥_t ⑥_t^\top \\
& - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} ⑧_t^\top - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑦_t ⑧_t^\top - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} ⑧_t ⑧_t^\top \\
& - \eta ③_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta ③_t ③_t^\top \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta ③_t ④_t^\top \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} + \mathbb{I}_1 ⑦_t + \mathbb{I}_{\geqslant 2} ⑧_t)(\mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top + ③_t^\top \mathbb{I}_1^\top + ④_t^\top \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1^\top + \mathbb{I}_1 ⑤_t + \mathbb{I}_{\geqslant 2} ⑥_t)(\mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top + ①_t \mathbb{I}_1^\top + ②_t^\top \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_{t+1} \mathbb{I}_{\geqslant 2} = {} & \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& - \eta \textcircled{2}_t \mathbb{I}_1^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \textcircled{6}_t^\top - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \textcircled{8}_t^\top \\
& - \eta \textcircled{4}_t \mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} + \eta^2 \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + \eta^2 \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \mathbb{I}_1^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
= {} & \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& - \eta \textcircled{2}_t \mathbb{I}_1^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \textcircled{6}_t^\top \\
& - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \textcircled{8}_t^\top \\
& - \eta \textcircled{4}_t \mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + \eta^2 \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + \eta^2 \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 \mathbb{I}_1^\top (\mathbf{Y}_0 + \Delta \mathbf{Y}_t)^\top (\mathbb{I}_1 \mathbb{I}_1^\top + \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
= {} & \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \textcircled{8}_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 \textcircled{7}_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \textcircled{4}_t^\top \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \textcircled{3}_t^\top \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& - \eta \textcircled{2}_t \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \textcircled{2}_t \textcircled{1}_t \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \textcircled{2}_t \textcircled{3}_t^\top \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1 \textcircled{6}_t^\top - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 \textcircled{5}_t \textcircled{6}_t^\top - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \textcircled{6}_t \textcircled{6}_t^\top \\
& - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} \textcircled{8}_t^\top - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 \textcircled{7} \textcircled{8}_t^\top - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \textcircled{8} \textcircled{8}_t^\top \\
& - \eta \textcircled{4}_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \textcircled{4}_t \textcircled{3}_t^\top \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \textcircled{4}_t \textcircled{4}^\top \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + \eta^2 \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} + \mathbb{I}_1 \textcircled{7}_t + \mathbb{I}_{\geqslant 2} \textcircled{8}_t)(\mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top + \textcircled{3}_t^\top \mathbb{I}_1^\top + \textcircled{4}_t^\top \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + \eta^2 \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t (\mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1^\top + \mathbb{I}_1 \textcircled{5}_t + \mathbb{I}_{\geqslant 2} \textcircled{6}_t)(\mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top + \textcircled{1}_t \mathbb{I}_1^\top + \textcircled{2}_t^\top \mathbb{I}_{\geqslant 2}^\top) \mathbf{M}_t \mathbb{I}_{\geqslant 2}
\end{aligned}
$$

In the following equations, red terms are expected to be $O(1)$ while blue terms are expected

to be $O(\epsilon)$.

$$
\begin{aligned}
①_{t+1} &= ①_t - \eta \mathbb{I}_1^\top \mathbf{M}_t (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 && \text{(B.162)} \\
&= ①_t - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 && \text{(B.163)} \\
&= ①_t - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑤_t - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} ⑥_t, && \text{(B.164)} \\
②_{t+1} &= ②_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_1 && \text{(B.165)} \\
&= ②_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1 - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 ⑤_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} ⑥_t, && \text{(B.166)} \\
③_{t+1} &= ③_t - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} && \text{(B.167)} \\
&= ③_t - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑦_t - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} ⑧_t, && \text{(B.168)} \\
④_{t+1} &= ④_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top (\mathbf{Z}_0 + \Delta \mathbf{Z}_t) \mathbb{I}_{\geqslant 2} && \text{(B.169)} \\
&= ④_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1 ⑦_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} ⑧_t, && \text{(B.170)} \\
⑤_{t+1} &= ⑤_t - \eta \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_1 \mathbb{I}_1^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right) \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right) \mathbb{I}_1 && \text{(B.171)} \\
&= ⑤_t - \eta \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 - \eta \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_1 ①_t - \eta \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2} ②_t, && \text{(B.172)} \\
⑥_{t+1} &= ⑥_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t^\top \mathbb{I}_1 \mathbb{I}_1^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right) \mathbb{I}_1 - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right) \mathbb{I}_1 && \text{(B.173)} \\
&= ⑥_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t^\top \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t^\top \mathbb{I}_1 ①_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2} ②_t, && \text{(B.174)} \\
⑦_{t+1} &= ⑦_t - \eta \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_1 \mathbb{I}_1^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right) \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right) \mathbb{I}_{\geqslant 2} && \text{(B.175)} \\
&= ⑦_t - \eta \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_1 ③_t - \eta \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2} ④_t, && \text{(B.176)} \\
⑧_{t+1} &= ⑧_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t^\top \mathbb{I}_1 \mathbb{I}_1^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right) \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta \mathbf{Y}_t\right) \mathbb{I}_{\geqslant 2} && \text{(B.177)} \\
&= ⑧_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t^\top \mathbb{I}_1 ③_t - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2} ④_t, && \text{(B.178)}
\end{aligned}
$$

By expanding the definition of $\mathbb{I}_1\mathbf{M}_t\mathbb{I}_1$, the update rules of ①$_t$ and ⑤$_t$ are

$$①_{t+1} = ①_t - \eta(①_t\frac{\sigma_1}{\alpha} + \sigma_1\alpha⑤_t + ①_t⑤_t + ③_t⑦_t^\top)(\frac{\sigma_1}{\alpha} + ⑤_t) - \eta\mathbb{I}_1^\top\mathbf{M}_t\mathbb{I}_{\geqslant 2}⑥_t \tag{B.179}$$

$$= ①_t - \eta(①_t\frac{\sigma_1}{\alpha} + \sigma_1\alpha⑤_t + ①_t⑤_t)(\frac{\sigma_1}{\alpha} + ⑤_t) - \eta\mathbb{I}_1^\top\mathbf{M}_t\mathbb{I}_{\geqslant 2}⑥_t - \eta③_t⑦_t^\top(\frac{\sigma_1}{\alpha} + ⑤_t), \tag{B.180}$$

$$⑤_{t+1} = ⑤_t - \eta(①_t\frac{\sigma_1}{\alpha} + \sigma_1\alpha⑤_t + ①_t⑤_t + ③_t⑦_t^\top)(\sigma_1\alpha + ①_t) - \eta\mathbb{I}_1^\top\mathbf{M}_t^\top\mathbb{I}_{\geqslant 2}②_t \tag{B.181}$$

$$= ⑤_t - \eta(①_t\frac{\sigma_1}{\alpha} + \sigma_1\alpha⑤_t + ①_t⑤_t)(\sigma_1\alpha + ①_t) - \eta\mathbb{I}_1^\top\mathbf{M}_t^\top\mathbb{I}_{\geqslant 2}②_t - \eta③_t⑦_t^\top(\sigma_1\alpha + ①_t) \tag{B.182}$$

At initialization $t = 1$, all of $\mathbb{I}_1^\top\mathbf{M}_t\mathbb{I}_{\geqslant 2}, ②_t, ③_t, ⑥_t, ⑦_t$ are in $O(\epsilon)$

Since we have assumed

$$\begin{aligned}\mathbb{I}_{\geqslant 2}^\top\mathbf{M}_{t+1}\mathbb{I}_{\geqslant 2} \approx\ & \mathbb{I}_{\geqslant 2}^\top\mathbf{M}_t\mathbb{I}_{\geqslant 2} - \eta\mathbb{I}_{\geqslant 2}^\top\mathbf{M}_t\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^\top\mathbf{Z}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^\top\mathbf{Z}_0^\top\mathbb{I}_{\geqslant 2} \\ & - \eta\mathbb{I}_{\geqslant 2}^\top\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^\top\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^\top\mathbf{M}_t\mathbb{I}_{\geqslant 2} \\ & + O(\epsilon\cdot\epsilon_t)\end{aligned}$$

$$
\begin{aligned}
\mathbb{I}_1^\top \mathbf{M}_{t+1} \mathbb{I}_{\geqslant 2} \approx\ & \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑦_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 ①_t \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& - \eta ①_t \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta ①_t ①_t \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 \mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1 ⑥_t^\top - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑤_t ⑥_t^\top \\
& - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 ⑦_t ⑧_t^\top \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 (\mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1^\top + ⑤_t)(\mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 + ①_t) \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + O(\epsilon \cdot \epsilon_t) \\
=\ & \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \left(\mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 + ①_t \right)^2 - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} \\
& - \eta \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 \underbrace{\left(\mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1 ⑥_t^\top + ⑤_t ⑥_t^\top + ⑦_t ⑧_t^\top + ⑦_t \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0^\top \mathbb{I}_{\geqslant 2} \right)}_{= \mathbb{I}_1^\top (\mathbf{Z}_0 + \Delta \mathbf{Z}_t)(\mathbf{Z}_0 + \Delta \mathbf{Z}_t)^\top \mathbb{I}_{\geqslant 2}} \\
& + \eta^2 \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 (\mathbb{I}_1^\top \mathbf{Z}_0 \mathbb{I}_1^\top + ⑤_t)(\mathbb{I}_1^\top \mathbf{Y}_0 \mathbb{I}_1 + ①_t) \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
& + O(\epsilon \cdot \epsilon_t)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_{t+1}\mathbb{I}_1 \approx\ & \mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0^{\top}\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1⑤_t\mathbb{I}_1^{\top}\mathbf{Z}_0^{\top}\mathbb{I}_1 \\
& - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}③_t^{\top}\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& - \eta②_t\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta②_t①_t\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1⑤_t - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1⑤_t⑤_t \\
& - \eta④_t③_t^{\top}\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& + \eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1(\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1^{\top} + ⑤_t)(\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1 + ①_t)\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 \\
=\ & \mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1\left(\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1 + ⑤_t\right)^2 - \eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& - \eta\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1\underbrace{\left(\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}③_t^{\top} + ②_t\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1 + ②_t①_t + ④_t③_t^{\top}\right)}_{=\mathbb{I}_{\geqslant 2}^{\top}(\mathbf{Y}_0+\Delta\mathbf{Y}_t)(\mathbf{Y}_0+\Delta\mathbf{Y}_t)^{\top}\mathbb{I}_1} \\
& + \eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1(\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1^{\top} + ⑤_t)(\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1 + ①_t)\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{I}_1^{\top}\mathbf{M}_{t+1}\mathbb{I}_1 \approx\ & \mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0^{\top}\mathbb{I}_1 - \eta\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1⑤_t\mathbb{I}_1^{\top}\mathbf{Z}_0^{\top}\mathbb{I}_1 \\
& - \eta\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1①_t\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& - \eta①_t\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 - \eta①_t①_t\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1 \\
& - \eta\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1⑤_t - \eta\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1⑤_t⑤_t \\
& + \eta^2\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1(\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1 + ⑤_t)(\mathbb{I}_1^{\top}\mathbf{Y}_0\mathbb{I}_1 + ①_t)\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_{t+1}\right)\left(\mathbf{Y}_0+\Delta\mathbf{Y}_{t+1}\right)^{\top}\mathbb{I}_1\\
&=\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_{t}\right)\left(\mathbf{Y}_0+\Delta\mathbf{Y}_{t}\right)^{\top}\mathbb{I}_1\\
&\quad-\eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_{\geqslant 2}③_t^{\top}-\eta(\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}+④_t)\mathbb{I}_{\geqslant 2}^{\top}(\mathbf{Z}_0+\Delta\mathbf{Z}_t)^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&\quad-\eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_1(①_t+\sigma_1\alpha)-\eta②\mathbb{I}_1^{\top}(\mathbf{Z}_0+\Delta\mathbf{Z}_t)^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&\quad+\eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}(\mathbf{Z}_0+\Delta\mathbf{Z}_t)^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&\quad+\eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_1\mathbb{I}_1^{\top}(\mathbf{Z}_0+\Delta\mathbf{Z}_t)^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&=\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_{t}\right)\left(\mathbf{Y}_0+\Delta\mathbf{Y}_{t}\right)^{\top}\mathbb{I}_1\\
&\quad-\eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Z}_0\mathbb{I}_{\geqslant 2}③_t^{\top}-\eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1^{\top}⑦_t③_t^{\top}-\eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_{\geqslant 2}^{\top}⑧_t③_t^{\top}\\
&\quad-\eta(\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}+④_t)\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Z}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1-\eta(\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}+④_t)⑦_t^{\top}\mathbb{I}_1^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&\quad-\eta(\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}+④_t)⑧_t^{\top}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&\quad-\eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1(①_t+\sigma_1\alpha)-\eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1⑤_t(①_t+\sigma_1\alpha)\\
&\quad-\eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_{\geqslant 2}⑥_t(①_t+\sigma_1\alpha)\\
&\quad-\eta②\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1\mathbb{I}_1^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1-\eta②⑤_t\mathbb{I}_1^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1-\eta②⑥_t^{\top}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&\quad+\eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}(\mathbf{Z}_0+\Delta\mathbf{Z}_t)^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&\quad+\eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t(\mathbf{Z}_0+\Delta\mathbf{Z}_t)\mathbb{I}_1\mathbb{I}_1^{\top}(\mathbf{Z}_0+\Delta\mathbf{Z}_t)^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&\approx\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_{t}\right)\left(\mathbf{Y}_0+\Delta\mathbf{Y}_{t}\right)^{\top}\mathbb{I}_1-\eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Z}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&\quad-\eta(①_t+\sigma_1\alpha)(⑤_t+\frac{\sigma_1}{\alpha})\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1\\
&\quad-\eta\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1\underbrace{\left(②\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1+②⑤_t\mathbb{I}_1^{\top}+(\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}+④_t)⑦_t^{\top}\right)}_{=\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1}\\
&\quad+\eta^2\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1(\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1+⑤_t)(\mathbb{I}_1^{\top}\mathbf{Z}_0\mathbb{I}_1+⑤_t)\mathbb{I}_1^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&=\mathbb{I}_{\geqslant 2}^{\top}\left(\mathbf{Y}_0+\Delta\mathbf{Y}_{t}\right)\left(\mathbf{Y}_0+\Delta\mathbf{Y}_{t}\right)^{\top}\mathbb{I}_1-\eta\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Y}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{Z}_0\mathbb{I}_{\geqslant 2}\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t^{\top}\mathbb{I}_1\\
&\quad-\eta\left((①_t+\sigma_1\alpha)(⑤_t+\frac{\sigma_1}{\alpha})+\mathbb{I}_1^{\top}\mathbf{M}_t\mathbb{I}_1(1-\eta(⑤_t+\frac{\sigma_1}{\alpha})^2)\right)\mathbb{I}_{\geqslant 2}^{\top}\mathbf{M}_t\mathbb{I}_1
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{I}_1^\top \left(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t+1}\right)\left(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t+1}\right)^\top \mathbb{I}_{\geqslant 2} \\
&\approx \mathbb{I}_1^\top \left(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t}\right)\left(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t}\right)^\top \mathbb{I}_{\geqslant 2} \\
&\quad - \eta \left((①_t + \sigma_1\alpha)(⑤_t + \frac{\sigma_1}{\alpha}) + \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1 (1 - \eta(①_t + \sigma_1\alpha)^2) \right) \mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2} \\
&\quad - \eta \mathbb{I}_{\geqslant 2}^\top \mathbf{Y}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_{\geqslant 2}^\top \mathbf{Z}_0 \mathbb{I}_{\geqslant 2} \mathbb{I}_1^\top \mathbf{M}_t^\top \mathbb{I}_{\geqslant 2}
\end{aligned}
$$

Therefore, we have built a 4×4 matrix to characterize the dynamics of $\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1$, $\mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2}$, $\mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta\mathbf{Y}_{t+1}\right)\left(\mathbf{Y}_0 + \Delta\mathbf{Y}_{t+1}\right)^\top \mathbb{I}_1$, $\mathbb{I}_1^\top \left(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t+1}\right)\left(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t+1}\right)^\top \mathbb{I}_{\geqslant 2}$ as, for $\forall p \in \{2, 3, \ldots, d\}$

$$
\begin{bmatrix}
\left[\mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta\mathbf{Y}_{t+1}\right)\left(\mathbf{Y}_0 + \Delta\mathbf{Y}_{t+1}\right)^\top \mathbb{I}_1\right]_p \\
\left[\mathbb{I}_1^\top \left(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t+1}\right)\left(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t+1}\right)^\top \mathbb{I}_{\geqslant 2}\right]_p \\
\left[\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_{t+1} \mathbb{I}_1\right]_p \\
\left[\mathbb{I}_1^\top \mathbf{M}_{t+1} \mathbb{I}_{\geqslant 2}\right]_p
\end{bmatrix}
\leftarrow \mathbf{Q}_{\alpha,\eta,p}(y_t, z_t)
\begin{bmatrix}
\left[\mathbb{I}_{\geqslant 2}^\top \left(\mathbf{Y}_0 + \Delta\mathbf{Y}_{t}\right)\left(\mathbf{Y}_0 + \Delta\mathbf{Y}_{t}\right)^\top \mathbb{I}_1\right]_p \\
\left[\mathbb{I}_1^\top \left(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t}\right)\left(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t}\right)^\top \mathbb{I}_{\geqslant 2}\right]_p \\
\left[\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_{t} \mathbb{I}_1\right]_p \\
\left[\mathbb{I}_1^\top \mathbf{M}_{t} \mathbb{I}_{\geqslant 2}\right]_p
\end{bmatrix},
$$

$$
\mathbf{Q}_{\alpha,\eta,p}(y_t, z_t) \triangleq
\begin{bmatrix}
1 & 0 & u_{1,t} & -\eta\sigma_p^2 \\
0 & 1 & -\eta\sigma_p^2 & u_{2,t} \\
-\eta(y_t z_t - \sigma_1^2) & 0 & w_{1,t} & 0 \\
0 & -\eta(y_t z_t - \sigma_1^2) & 0 & w_{2,t}
\end{bmatrix},
$$

$$
y_t \triangleq ①_t + \sigma_1\alpha, \quad z_t \triangleq ⑤_t + {}^{\sigma_1}/_{\alpha_p},
$$

$$
u_{1,t} \triangleq -\eta\left(y_t z_t + (y_t z_t - \sigma_1^2)(1 - \eta z_t^2)\right),
$$

$$
u_{2,t} \triangleq -\eta\left(y_t z_t + (y_t z_t - \sigma_1^2)(1 - \eta y_t^2)\right),
$$

$$
w_{1,t} \triangleq 1 - \eta z_t^2 - \eta\sigma_p^2\alpha^2 + \eta^2 y_t z_t (y_t z_t - \sigma_1^2),
$$

$$
w_{2,t} \triangleq 1 - \eta y_t^2 - \eta^{\sigma_p^2/\alpha^2} + \eta^2 y_t z_t (y_t z_t - \sigma_1^2),
$$

where $[\cdot]_p$ means the p-th value in a vector.

Recall we have y_t, z_t following the training dynamics of minimizing $\frac{1}{2}(\sigma_1^2 - yz)^2$ with learning rate $\eta > \frac{1}{\sigma_1^2}$, where leads to $y = z = \gamma_i$, with γ_i $(i = 1, 2)$ are the two roots of solving the 1-D function (3.4) as δ. We denote their corresponding $\mathbf{Q}$ as $\mathbf{Q}_{\alpha,\eta,p}(\gamma_1, \gamma_1)$ and $\mathbf{Q}_{\alpha,\eta,p}(\gamma_2, \gamma_2)$. We assume that their product $\mathbf{Q}_{\alpha,\eta,p}(\gamma_2, \gamma_2)\mathbf{Q}_{\alpha,\eta,p}(\gamma_1, \gamma_1)$ is diagonalizable with all eigenvalues falling into $(-1, 1)$, which means its infinite power $\lim_{k\to\infty}[\mathbf{Q}_{\alpha,\eta,p}(\gamma_2, \gamma_2)\mathbf{Q}_{\alpha,\eta,p}(\gamma_1, \gamma_1)]^k = 0$. Meanwhile, due to the 2-D analysis of dynamics of GD on $\frac{1}{2}(\sigma_1^2 - yz)^2$, we know $(y_t, z_t) \to \{(\gamma_1, \gamma_1), (\gamma_2, \gamma_2)\}$ exponentially after finite steps. This is equivalent to say, there exists finite t_0, for any $t > t_0$, there exists $i \in \{1, 2\}$, constant C_0 and $\mathbf{R}_t \in \mathbb{R}^{4\times 4}$, such that

$$\mathbf{Q}_{\alpha,\eta,p}(y_{t+1}, z_{t+1})\mathbf{Q}_{\alpha,\eta,p}(y_t, z_t) = \mathbf{Q}_{\alpha,\eta,p}(\gamma_{3-i}, \gamma_{3-i})\mathbf{Q}_{\alpha,\eta,p}(\gamma_i, \gamma_i) + \mathbf{R}_t, \qquad \|\mathbf{R}_t\| \leqslant C_0 r^t, \quad 0 < r < 1.$$

The decay rate r can be estimated via local analysis around the convergence orbit. As a result, it is safe to say $\lim_{t\to\infty} \mathbf{Q}_{\alpha,\eta,p}(y_{2t+1}, z_{2t+1})\mathbf{Q}_{\alpha,\eta,p}(y_{2t}, z_{2t}) = 0$, which means all of $\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1$, $\mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2}$, $\mathbb{I}_{\geqslant 2}^\top (\mathbf{Y}_0 + \Delta\mathbf{Y}_{t+1})(\mathbf{Y}_0 + \Delta\mathbf{Y}_{t+1})^\top \mathbb{I}_1$, $\mathbb{I}_1^\top (\mathbf{Z}_0 + \Delta\mathbf{Z}_{t+1})(\mathbf{Z}_0 + \Delta\mathbf{Z}_{t+1})^\top \mathbb{I}_{\geqslant 2}$ exponentially go to zero.

There is one concern here: what happens before t_0? More concretely, t_0 is dependent of $1/\epsilon$ because it requires more steps (intuitively proportional to $\log 1/\epsilon$) to increase to a certain value from a small ϵ. Assuming $t_0 \sim \log 1/\epsilon$ holds, the product $\{\mathbf{Q}_{\alpha,\eta,p}(y_{2t+1}, z_{2t+1})\mathbf{Q}_{\alpha,\eta,p}(y_{2t}, z_{2t})\}_{t\geqslant 1}$ gives a (loose) upper bound with the norm of products grows exponentially with time $\log 1/\epsilon$, which introduces $1/\epsilon$ to the upper bound of $\left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1\right\|$ and $\left\|\mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2}\right\|$, breaking the assumption of the norms staying in $O(\epsilon)$. Fortunately, there are two aspects to resolve this. Firstly, with initialization ϵ small enough, for a relative long time, $\mathbf{Q}_{\alpha,\eta,p}(y_{2t+1}, z_{2t+1})\mathbf{Q}_{\alpha,\eta,p}(y_{2t}, z_{2t})$ is approximately

having eigenvalues bounded by 1. More precisely, $\mathbf{Q}$ and the product are

$$\mathbf{Q}_{\alpha,\eta,p}(\cdot,\cdot) \approx \begin{bmatrix} 1 & 0 & -\eta\sigma_1^2 & -\eta\sigma_p^2 \\ 0 & 1 & -\eta\sigma_p^2 & -\eta\sigma_1^2 \\ 0 & 0 & 1-\eta^{\sigma_1^2/\alpha^2}-\eta\sigma_p^2\alpha^2 & 0 \\ 0 & 0 & 0 & 1-\eta\sigma_1^2\alpha^2-\eta^{\sigma_p^2/\alpha^2} \end{bmatrix}, \tag{B.183}$$

$$\Lambda(\mathbf{Q}_{\alpha,\eta,p}(\cdot,\cdot)\mathbf{Q}_{\alpha,\eta,p}(\cdot,\cdot)) = \{1, 1, (1-\eta^{\sigma_1^2/\alpha^2}-\eta\sigma_p^2\alpha^2)^2, (1-\eta\sigma_1^2\alpha^2-\eta^{\sigma_p^2/\alpha^2})^2, \} \tag{B.184}$$

where the eigenvalues in Λ are upper bounded by 1, if assuming $\eta(^{\sigma_1^2}/_{\alpha^2} + \sigma_p^2\alpha^2) < 2$ and $1 - \eta\sigma_1^2\alpha^2 - \eta^{\sigma_p^2/\alpha^2} < 2$. As a result, in these steps, $\left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1\right\|$ and $\left\|\mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2}\right\|$ stay in $O(\epsilon)$ due to $\mathbf{Q}_{\alpha,\eta,p}(\cdot,\cdot)\mathbf{Q}_{\alpha,\eta,p}(\cdot,\cdot)$ is a semi-convergent matrix. Secondly, the eigenvectors of $\mathbf{Q}_{\alpha,\eta,p}\mathbf{Q}_{\alpha,\eta,p}$ corresponding to eigenvalue 1 are $[1,0,0,0]^\top$ and $[0,1,0,0]^\top$, which means both $\left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1\right\|$ and $\left\|\mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2}\right\|$ are decaying exponentially. Therefore, it's fair to say smaller ϵ strengthens the assumption of $\left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1\right\|$ and $\left\|\mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2}\right\|$ staying in $O(\epsilon)$ instead of breaking it.

Also note that $\left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_{t+1} \mathbb{I}_{\geqslant 2}\right\| \lessapprox \left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2}\right\| \cdot \max\{|1-\eta\sigma_2\left(\alpha^2 + {}^1/_{\alpha^2}\right)|, |1-\eta\sigma_{d-1}\left(\alpha^2 + {}^1/_{\alpha^2}\right)|\}$, so $\left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_{t+1} \mathbb{I}_{\geqslant 2}\right\|$ decays exponentially.

Since all of $\left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_t \mathbb{I}_1\right\|$, $\left\|\mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2}\right\|$ and $\left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_{t+1} \mathbb{I}_{\geqslant 2}\right\|$ decay exponentially after some steps, all of them are have the sum upper-bounded, which means $\|②_t\|, \|③_t\|, \|④_t\|, \|⑥_t\|, \|⑦_t\|, \|⑧_t\|$ stay in $O(\epsilon)$.

To summarize, it holds

1. $\|②_t\|, \|③_t\|, \|④_t\|, \|⑥_t\|, \|⑦_t\|, \|⑧_t\|$ stay in $O(\epsilon)$.
2. $\left\|\mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_{\geqslant 2}\right\|$ and $\left\|\mathbb{I}_{\geqslant 2}^\top \mathbf{M}_{t+1} \mathbb{I}_{\geqslant 2}\right\|$ decays to zero.
3. $\left\|\mathbb{I}_1^\top \mathbf{M}_t \mathbb{I}_1\right\|$ stays in a period-2 orbit.

□

B.11 Useful lemmas

Lemma B.11. *Assume $a \cdot \Delta a \geqslant b \cdot \Delta b$ and $a \geqslant b$. All of $a, b, \Delta a, \Delta b$ are positive. If $\Delta b \leqslant a$, then $a + \Delta a \geqslant b + \Delta b$.*

Proof. $(a + \Delta a) - (b + \Delta b) \geqslant a + b\frac{\Delta b}{a} - b - \Delta b = (\frac{\Delta b}{a} - 1)(b - a) \geqslant 0.$ □

B.12 Illustration of period-2 and period-4 orbits

In the setting of $f(x) = \frac{1}{4}(x^2-1)^2$, local convergence is guaranteed if $\eta < \sqrt{5}-1 \approx 1.236$ by taylor expansion of F_η^2 around the orbit. Conversely, if the learning rate is larger than it, although the period-2 orbit still exists, GD starting from a point infinitesimally close to the orbit still escapes from it. This is when GD converges to a higher-order orbit.

Figure B.5 precisely shows the effectiveness of such a bound where GD converges to the period-2 orbit when $\eta = 1.235 < \sqrt{5} - 1$ and a period-4 orbit when $\eta = 1.237 > \sqrt{5} - 1$.

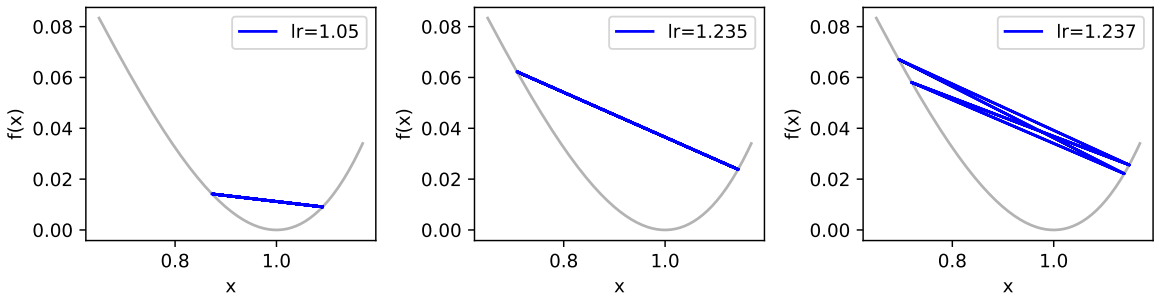


Figure B.5: The convergent orbits of GD on $f(x) = \frac{1}{4}(x^2 - 1)^2$ with learning rate=1.05, 1.235 and 1.237. The first two smaller learning rates drive to period-2 orbits while the last one goes to an period-4 orbit. The significant bound between period-2 and period-4 is predictable by Taylor expansion around the period-2 orbit, as $\eta = \sqrt{5} - 1 \approx 1.236$.

C | Appendix: Supplementary Materials for Chapter 4

C.1 More Experiments on Pythia

C.1.1 Learning Association with Prepositions

We would like to verify our guess about the structure of "to + the" in Pythia in Section 4.5.1. To make the argument generalizable than IOI dataset, we consider a structure of "[preposition] + the", where [preposition] has a pool of 30 prepositions in English, including "to". The input is a raw "[preposition]" or a random sentence ending with "[preposition]", with some examples in Appendix C.7.1. For both kinds of inputs, Pythia-160M/410M/1B turns out to learn the structure of "[preposition] + the" around 10 steps, as shown in Figure C.1.

C.1.2 LASER Parameters for Evaluated LLMs

Following the definition of LASER in Section 4.3.2, we search for the optimal layer, ρ and target weights in Pythia models and GPT-2 Small for each dataset.

IOI on Pythia-410M. The model has 24 layers. The truncation is on the input matrix of MLPs on the 22-th layer with $\rho = 0.02$.

IOI on Pythia-1B. The model has 16 layers. The truncation is on the input matrix of MLPs

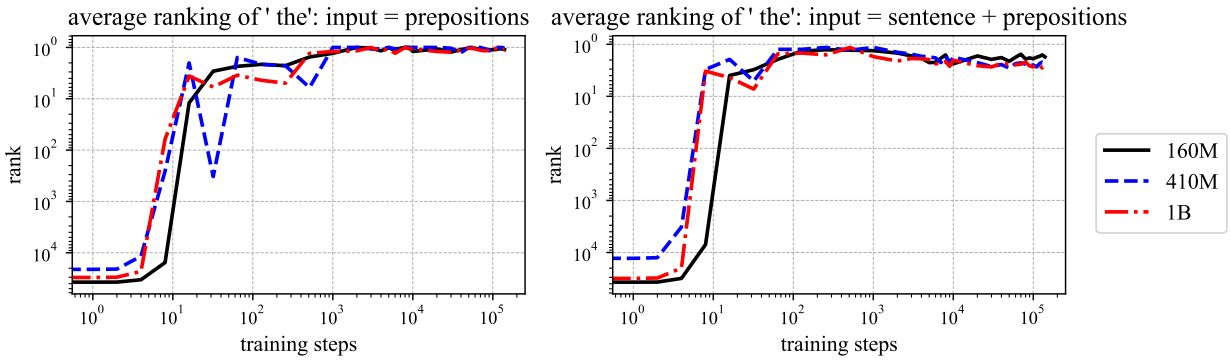


Figure C.1: Average ranking of tokens "the" in the prediction by Pythia-160M/410M/1B along training. The inputs are 30 preposition words (left) and 40 sentences ending with prepositions. It turns out "the" becomes one of top predictions around 10 steps.

on the 11-th layer with $\rho = 0.008$.

Factual recall on Pythia-1B. The truncation is on the input matrix of MLPs on the 16-th layer with $\rho = 0.0125$.

Factual recall on Pythia-1.4B. The model has 24 layers. The truncation is on the input matrix of MLPs on the 24-th layer with $\rho = 0.025$.

Factual recall on Pythia-2.8B. The model has 32 layers. The truncation is on the input matrix of MLPs on the 32-th layer with $\rho = 0.04$.

IOI on GPT2 Small. Related parameters have been contained in Section 4.5.1.

Phi-3 on GSM8K. The model has 32 layers. The truncation is on the output matrix of MLPs on the 28-th layer with $\rho = 0.02$.

Llama3.1-8B(-instruct) on GSM8K. The models have 32 layers. The truncation is on the output of MLPs on the 27-th layer with $\rho = 0.02$.

C.1.3 Other Pythia models on IOI and More Examples of Factual Recall

IOI. In the same setting of Figure 4.13 (left), we plot the prediction distributions of Pythia-410M and 1B on the 100 IOI inputs in Figure C.2. The model checkpoints are the final ones after training. LASER turns out to decrease the probability of "the" while keeping that of the correct [IO] high.

More examples of Factual Recall. In additional to the factual query "Madrid is located in" in Figure 4.13 (right), we consider more such examples in Table C.3. We plot the prediction distributions of Pythia-1B, 1.4B and 2.8B on these inputs in Figure C.3, where LASER significantly lowers the probability of predicting "the" vesus the correct outputs.

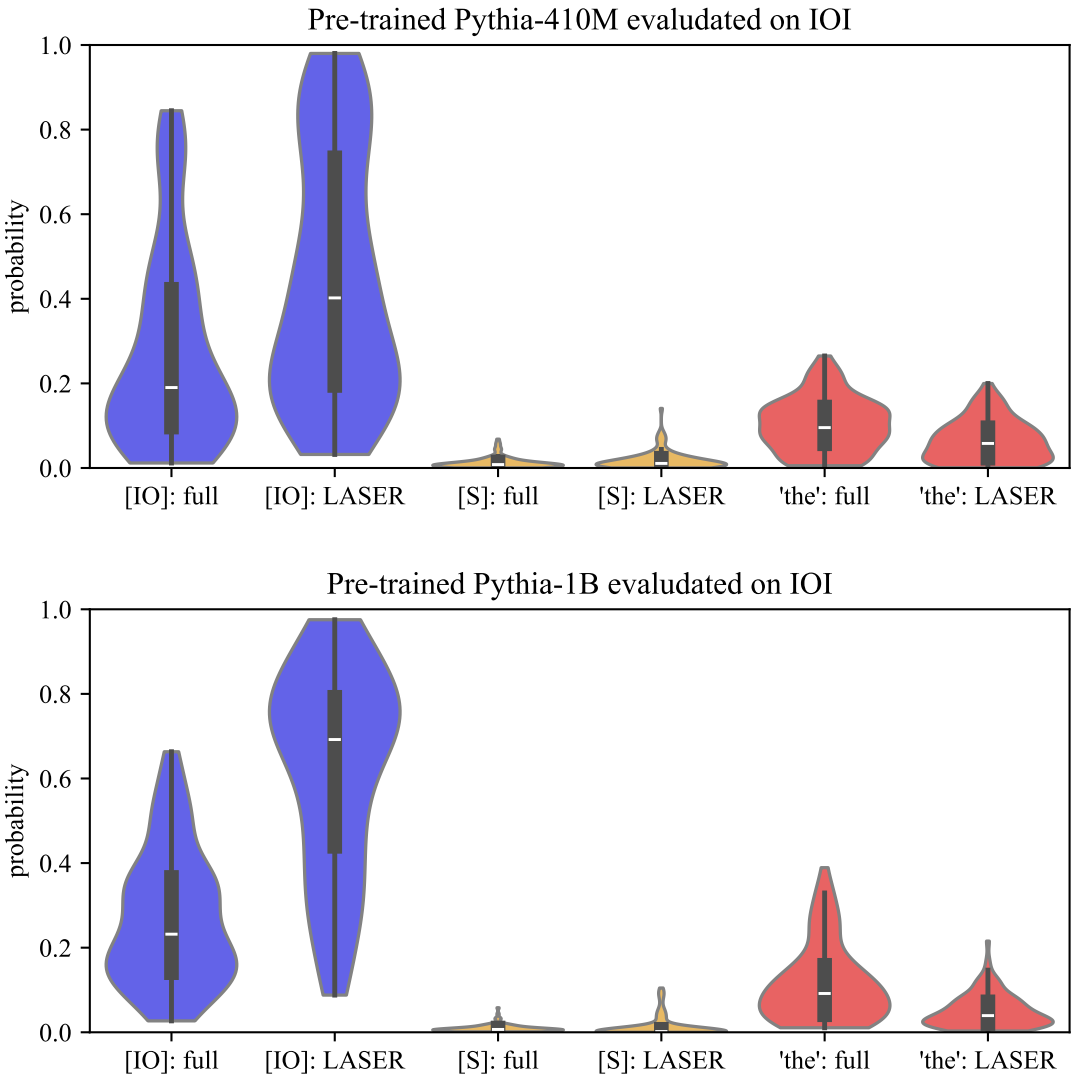


Figure C.2: The prediction distributions of Pythia-410M and 1B on the IOI task. The setting is the same as in Fgure 4.13 (left). The evaluated models are the final checkpoints after training. LASER turns out to decrease the probability of "the" while keeping that of the correct [IO] high.

C.2 Proof of Theorem 4.1

In this section, we will present the expectations and variances of $\nabla_{\mathbf{W}_V}\hat{L}$ and $\nabla_{\mathbf{W}_F}\hat{L}$ with $\mathbf{W}_V = \mathbf{W}_F = 0$ at initialization. The targets are to show:

1. a gap between $\lim_{m\to\infty}\nabla_{\mathbf{W}_V}\hat{L}$ and $\lim_{m\to\infty}\nabla_{\mathbf{W}_F}\hat{L}$ so that a step of GD with large learning rates is enough to learn the noise in $\mathbf{W}_F$, and

2. sample complexity of $\nabla_{\mathbf{W}_V}\hat{L}$ and $\nabla_{\mathbf{W}_F}\hat{L}$ based on expectations and variances.

Assumption C.2.1 (Orthonormal embeddings)**.** *The embeddings $u_k \in \mathbb{R}^d$ are assumed to be orthonormal, i.e., $u_i^\top u_j = \mathbb{1}\{i = j\}$. Meanwhile, if a matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$ is random initialized, it holds $u_i^\top \mathbf{W} u_j = 0$.*

C.2.1 Gradient for the Feed-forward Matrix $\mathbf{W}_F$

Lemma C.1. *Consider zero initialization, $\mathbf{W}_V = \mathbf{W}_F = \mathbf{W}_{QK} = 0$ and $N \gg 1$. Then with probability $1-\delta$, for any $j, k \in [N+1]$, it holds*

$$\begin{aligned}&\left|\mathbf{W}_U(k)^\top(\nabla_{\mathbf{W}_F}\hat{L})\mathbf{W}_E(q) - \mu(k)\right| \\ &\leqslant \sqrt{\frac{4\sigma^2(k)\left(\ln(N+1)+\ln(\frac{2}{\delta})\right)}{m}} + \frac{4R(k)\left(\ln(N+1)+\ln(\frac{2}{\delta})\right)}{m},\end{aligned} \tag{C.1}$$

where $\mu(k), \sigma^2(k), R(k)$ are expectation, variance and range for different choices of $k \in [N]$ as follows:

$$\begin{array}{lll} \mu(N+1) = -\alpha, & \sigma^2(N+1) = \alpha(1-\alpha), & R(N+1) = \max\{\alpha, 1-\alpha\}, \\ \forall\, k \leqslant N: \quad \mu(k) = \frac{1}{N+1} - \frac{1-\alpha}{N}, & \sigma^2(k) = \frac{1-\alpha}{N}, & R(k) = 1. \end{array}$$

Proof. Due to zero initialization, *i.e.*, $\mathbf{W}_V = \mathbf{W}_F = 0$, the current predicted probability is uniform

as $\hat{p}_{\mathbf{W}}(k|x_i) \equiv \frac{1}{N+1}$ for all $i \in [m]$ and $k \in [N+1]$. Therefore, from Lemma C.12, we have

$$\nabla_{\mathbf{W}_F} \hat{L} = \frac{1}{m} \sum_{i=1}^{m} \left[\sum_{k=1}^{N+1} \left(\frac{1}{N+1} - \mathbb{1}\{y_i = k\} \right) \mathbf{W}_U(k) x_{i,T}^\top \right],$$

where $x_{i,T} \in \mathbb{R}^d = \mathbf{W}_E(z_{i,T}) + p_T$ is the input embedding with input token $z_{i,T}$ at position T in sequence i, together with positional encoding p_T for position T. Since $z_{i,T}$ is set to be the trigger q in the data generation process and p_T is assumed to orthogonal to any other vector in $\mathbf{W}_E$ in Assumption C.2.1, we have the following projections for $\nabla_{\mathbf{W}_F} \hat{L}$: $\forall\, k \in [N+1]$,

$$\mathbf{W}_U(k)^\top (\nabla_{\mathbf{W}_F} \hat{L}) \mathbf{W}_E(q) = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{N+1} - \mathbb{1}\{y_i = k\} \right).$$

From the data generation process, it is obvious to get

$$\mathbb{E}_{(x,y)} \left[\frac{1}{N+1} - \mathbb{1}\{y = k\} \right] = \frac{1}{N+1} - \alpha \cdot \mathbb{1}\{k = N+1\} - \frac{1-\alpha}{N} \cdot \mathbb{1}\{k \leqslant N\}. \tag{C.2}$$

Since $\alpha = \Theta(1)$ is much larger than $\frac{1}{N+1}$ when $N \gg 1$, due to law of large numbers, we have the population gradient $\nabla_{\mathbf{W}_F} L$ satisfying

$$\begin{aligned} &\mathbf{W}_U(N+1)^\top (-\nabla_{\mathbf{W}_F} L) \mathbf{W}_E(q) \approx \alpha = \Theta(1), \\ \forall\, k \leqslant N: \quad &\mathbf{W}_U(k)^\top (-\nabla_{\mathbf{W}_F} L) \mathbf{W}_E(q) < 0, \text{ with absolute value in } O(1/N). \end{aligned}$$

The variance of the gradient projection onto $\mathbf{W}_U(N+1)\mathbf{W}_E(q)^\top$ of a single data point follows that of Bernoulli distribution with parameter α, which means

$$\mathrm{Var}\left[\frac{1}{N+1} - \mathbb{1}\{y = N+1\} \right] = \alpha(1-\alpha). \tag{C.3}$$

Similarly, for any $k \leqslant N$, the variance of the gradient projection onto $\mathbf{W}_U(N+1)\mathbf{W}_E(q)^\top$ of a

single data point follows that of Bernoulli distribution with parameter $\frac{1-\alpha}{N}$, which means

$$\mathrm{Var}\left[\frac{1}{N+1} - \mathbb{1}\{y=k\}\right] = \frac{1-\alpha}{N}\left(1 - \frac{1-\alpha}{N}\right) = \Theta(1/N). \tag{C.4}$$

The ranges of the gradient projections' deviation from the expectation are

$$\begin{aligned} &\left|\frac{1}{N+1} - \mathbb{1}\{y=N+1\} - \left(\frac{1}{N+1} - \alpha\right)\right| \leqslant \max\{\alpha, 1-\alpha\}, \\ \forall\, k \leqslant N: \quad &\left|\frac{1}{N+1} - \mathbb{1}\{y=k\} - \left(\frac{1}{N+1} - \frac{1-\alpha}{N}\right)\right| \lessapprox 1. \end{aligned} \tag{C.5}$$

For each choice of $k \in [N+1]$ *individually*, after having the expectation $\mu(k)$, variance $\sigma^2(k)$ and range $R(k)$, by applying Bernstein's inequality, then: for each $k \in [N+1]$, with probability $1-\delta$, it holds

$$\left|\mathbf{W}_U(k)^\top (\nabla_{\mathbf{W}_F} \hat{L}) \mathbf{W}_E(q) - \mu(k)\right| \leqslant \sqrt{\frac{4\sigma^2(k)\ln(\frac{2}{\delta})}{m}} + \frac{4R(k)\ln(\frac{2}{\delta})}{m}.$$

Then by the union bound in probability, we need $(N+1)$ events above to hold at the same time, so we can substitute δ with $\frac{\delta}{N+1}$ to have: with probability $1-\delta$, for any $k \in [N+1]$, it holds

$$\left|\mathbf{W}_U(k)^\top (\nabla_{\mathbf{W}_F} \hat{L}) \mathbf{W}_E(q) - \mu(k)\right| \leqslant \sqrt{\frac{4\sigma^2(k)\left(\ln(N+1) + \ln(\frac{2}{\delta})\right)}{m}} + \frac{4R(k)\left(\ln(N+1) + \ln(\frac{2}{\delta})\right)}{m}. \tag{C.6}$$

□

C.2.2 Gradient for the Value Matrix $\mathbf{W}_V$

Lemma C.2. *Consider zero initialization, $\mathbf{W}_V = \mathbf{W}_F = \mathbf{W}_{QK} = 0$. Then with probability $1-\delta$, for any $j, k \in [N+1]$, it holds*

$$\begin{aligned}&\left|\mathbf{W}_U(j)^\top (\nabla_{\mathbf{W}_V} \hat{L}) \mathbf{W}_E(k) - \mu(j,k)\right| \\ &\leqslant \sqrt{\frac{4\sigma^2(j,k)\left(2\ln(N+1)+\ln(\frac{2}{\delta})\right)}{m}} + \frac{4R(j,k)\left(2\ln(N+1)+\ln(\frac{2}{\delta})\right)}{m},\end{aligned} \tag{C.7}$$

where $\mu(j,k), \sigma^2(j,k), R(j,k)$ are expectation, variance and range for different choices of (j,k) at listed in Table C.1.

Table C.1: $\mu(j,k), \sigma^2(j,k), R(j,k)$ for different choices of (j,k) in Lemma C.2.

j	k	μ	σ^2	R
$N+1$	$N+1$	$-\frac{\alpha^2}{N}$	$\frac{\alpha^2}{TN} + \frac{\alpha^3-\alpha^4}{N^2}$	$\frac{1}{2}$
$N+1$	q	$-\frac{\alpha}{N}$	$\frac{\alpha}{TN} + \frac{\alpha-\alpha^2}{N^2}$	1
$N+1$	$[N]\setminus\{q\}$	$-\frac{\alpha}{N}$	$\frac{\alpha}{TN} + \frac{\alpha-\alpha^2}{N^2}$	1
q	$N+1$	$\frac{2\alpha-1}{N^2}$	$\frac{1}{TN^2} + \frac{\alpha^2-\alpha+1}{N^3}$	$\frac{1}{2}$
q	q	$\frac{2\alpha-1}{\alpha N^2}$	$\frac{\alpha^3-\alpha^2-\alpha+2}{\alpha^3 TN^2} + \frac{\alpha^2-\alpha+1}{\alpha^2 N^3}$	1
q	$[N]\setminus\{q\}$	$\frac{\alpha}{N^2}$	$(2-\alpha)\cdot\left(\frac{1}{TN^2}+\frac{1}{N^3}\right)$	1
$[N]\setminus\{q\}$	$N+1$	$\frac{\alpha^2}{N^2}$	$(2-\alpha)\left(\frac{\alpha}{TN^2}+\frac{\alpha^2}{N^3}\right)$	$\frac{1}{3}$
$[N]\setminus\{q\}$	q	$\frac{\alpha}{N^2}$	$(2-\alpha)\left(\frac{1}{TN^2}+\frac{1}{N^3}\right)$	$\frac{1}{2}$
$[N]\setminus\{q\}$	j	$\frac{-\alpha^2+3\alpha-1}{N^2}$	$\frac{1+(1-\alpha)(2-\alpha)}{TN^2} + \frac{1+(1-\alpha)(2-\alpha)^2}{N^3}$	1
$[N]\setminus\{q\}$	$[N]\setminus\{q,j\}$	$\frac{\alpha}{N^2}$	$(2-\alpha)\left(\frac{1}{TN^2}+\frac{1}{N^3}\right)$	1

Proof. Due to zero initialization, *i.e.*, $\mathbf{W}_V = \mathbf{W}_F = 0$, the current predicted probability is uniform as $\hat{p}_{\mathbf{W}}(k|x_i) \equiv \frac{1}{N+1}$ for all $i \in [m]$ and $k \in [N+1]$. Meanwhile, the attention score is uniform as

$\frac{1}{T}$ for all context positions due to $\mathbf{W}_K = 0$. Therefore, from Lemma C.12, we have

$$\nabla_{\mathbf{W}_F} \hat{L} = \frac{1}{m} \sum_{i=1}^{m} \left[\sum_{k=1}^{N+1} \left(\frac{1}{N+1} - \mathbb{1}\{y_i = k\} \right) \mathbf{W}_U(k) \left(\frac{1}{T} \sum_{t=1}^{T} x_{i,t} \right)^\top \right],$$

where $x_{i,t} \in \mathbb{R}^d = \mathbf{W}_E(z_{i,t}) + p_t$ is the input embedding with input token $z_{i,t}$ at position t in sequence i, together with positional encoding p_t for position t. With the assumption of orthonormality in Assumption C.2.1, we have the projection of $\nabla_{\mathbf{W}_F} \hat{L}$: $\forall\ j, k \in [N+1]$,

$$\mathbf{W}_U(j)^\top (\nabla_{\mathbf{W}_V} \hat{L}) \mathbf{W}_E(k) = \frac{1}{m} \sum_{i=1}^{m} \left[\left(\frac{1}{N+1} - \mathbb{1}\{y_i = j\} \right) \left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{z_{i,t} = k\} \right) \right].$$

Since each sample is drawn i.i.d., it suffices to discuss the expectation and variance of

$$\Gamma_i(j,k) \triangleq \left(\frac{1}{N+1} - \mathbb{1}\{z_{i,T+1} = j\} \right) \left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{z_{i,t} = k\} \right),$$
$$\hat{\Gamma}(j,k) \triangleq \frac{1}{m} \sum_{i=1}^{m} \Gamma_i(j,k),$$

where we use the fact $y_i = z_{i,T+1}$.

Recall that, for each sample in the data generation process, the trigger q is fixed while the correct next token $\bar{y} \sim \text{Uniform}([N])$. Hence, conditioning on $z_{i,T} = q$, it has probability α for $z_{i,T+1} = N+1$ and probability $1-\alpha$ for $z_{i,T+1} = \bar{y}$. This leads to the necessity of discussing whether or not $\bar{y} = k$. Meanwhile, a corner case of $\bar{y} = q$ is also necessary to consider, as this implies an event that increases the counting $\frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{z_{i,t} = q\}$ than the case of $\bar{y} \neq q$.

Therefore, generally there are 10 cases due to different choices of (j,k) as follows:

1. $j = N+1, k = N+1$,

2. $j = N+1, k = q$,

3. $j = N+1, k \in [N] \setminus \{q\}$,

4. $j = q, k = N + 1$,

5. $j = q, k = q$,

6. $j = q, k \in [N] \setminus \{q\}$,

7. $j \in [N] \setminus \{q\}, k = N + 1$,

8. $j \in [N] \setminus \{q\}, k = q$,

9. $j \in [N] \setminus \{q\}, k = j$,

10. $j \in [N] \setminus \{q\}, k \in [N] \setminus \{q, j\}$.

For each $\Gamma_i(j,k)$ *individually*, if we have its expectation $\mu(j,k)$, variance $\sigma^2(j,k)$ and range $R(j,k)$, by applying Bernstein's inequality, then: for each $j, k \in [N+1]$, with probability $1-\delta$, it holds

$$\left|\hat{\Gamma}(j,k) - \mu(j,k)\right| \leqslant \sqrt{\frac{4\sigma^2(j,k)\ln(\frac{2}{\delta})}{m}} + \frac{4R(j,k)\ln(\frac{2}{\delta})}{m}.$$

Then by the union bound in probability, we need $(N+1)^2$ events above to hold at the same time, so we can substitute δ with $\frac{\delta}{(N+1)^2}$ to have: with probability $1-\delta$, for any $j, k \in [N+1]$, it holds

$$\left|\hat{\Gamma}(j,k) - \mu(j,k)\right| \leqslant \sqrt{\frac{4\sigma^2(j,k)\left(2\ln(N+1) + \ln(\frac{2}{\delta})\right)}{m}} + \frac{4R(j,k)\left(2\ln(N+1) + \ln(\frac{2}{\delta})\right)}{m}. \quad \text{(C.8)}$$

As a final step of the proof, now we elaborate the expectation, variance and range of $\Gamma_i(j,k)$ for these 10 cases.

Case 1: $j = N+1, k = N+1$.

There is probability $\frac{1}{N}$ for $\bar{y} = q$ and probability $\frac{N-1}{N}$ for $\bar{y} \neq q$. Hence, we have

$$\mathbb{E}[\Gamma_i(j,k)] = \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y} = q] + \frac{N-1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y} \neq q],$$
$$\mathbb{E}[\Gamma_i(j,k)^2] = \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} = q] + \frac{N-1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} \neq q].$$

From Lemma C.5 and the independence between $\mathbb{1}\{z_{i,T+1} = N+1\}$ and $\sum_{t\leqslant T}\mathbb{1}\{z_{i,t} = k\}$, we have

$$\mathbb{E}[\Gamma_i(j,k)|\bar{y} = q] \approx -\alpha \cdot \frac{1}{N},$$
$$\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} = q] \approx \alpha \cdot \left(\frac{1}{TN} + \frac{1}{N^2}\right),$$

where the second is from

$$\mathbb{E}\left[\left(\frac{1}{N+1} - \mathbb{1}\{z_{i,T+1} = N+1\}\right)^2\right] = (1-\alpha)\cdot\left(\frac{1}{N+1}\right)^2 + \alpha\cdot\left(\frac{1}{N+1} - 1\right)^2 \approx \alpha.$$

Similarly, from Lemma C.8, we have

$$\mathbb{E}[\Gamma_i(j,k)|\bar{y} \neq q] \approx -\alpha \cdot \frac{\alpha}{N},$$
$$\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} \neq q] \approx \alpha \cdot \left(\frac{\alpha}{TN} + \frac{\alpha^2}{N^2}\right).$$

Therefore, it holds

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)] &= \frac{1}{N}\frac{-\alpha}{N} + \frac{N-1}{N}\frac{-\alpha^2}{N} \approx -\frac{\alpha^2}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q] + \frac{N-1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q] \\
&\approx \frac{1}{N}\alpha\cdot\left(\frac{1}{TN}+\frac{1}{N^2}\right) + \frac{N-1}{N}\alpha\cdot\left(\frac{\alpha}{TN}+\frac{\alpha^2}{N^2}\right) \approx \frac{\alpha^2}{TN}+\frac{\alpha^3}{N^2}, \\
\mathrm{Var}[\Gamma_i(j,k)] &= \mathbb{E}[\Gamma_i(j,k)^2] - \mathbb{E}[\Gamma_i(j,k)]^2 \approx \frac{\alpha^2}{TN}+\frac{\alpha^3-\alpha^4}{N^2}.
\end{aligned}$$

The range of $\Gamma_i(j,k)$ is

$$|\Gamma_i(j,k) - \mathbb{E}[\Gamma_i(j,k)]| \leqslant \frac{1}{2},$$

and the extreme case is when half of the sequence is $N+1$ with the rest all being q.

Case 2: $j = N+1, k = q$.

Similar to Case 1, we have $\mathbb{1}\{z_{i,T+1} = N+1\}$ is independent of $\sum_{t\leqslant T}\mathbb{1}\{z_{i,t}=k\}$.

From Lemma C.4, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q] &\approx -\alpha\cdot\frac{1}{\alpha N}, \\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q] &\approx \alpha\cdot\left(\frac{1}{\alpha TN}\left(-1+\frac{2}{\alpha^2}\right)+\frac{1}{\alpha^2N^2}\right).
\end{aligned}$$

From Lemma C.7, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q] &\approx -\alpha\cdot\frac{1}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q] &\approx \alpha\cdot\left(\frac{1}{TN}+\frac{1}{N^2}\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q] + \frac{N-1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q] \approx -\frac{\alpha}{N},\\
\mathbb{E}[\Gamma_i(j,k)^2] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q] + \frac{N-1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q] \approx \frac{\alpha}{TN} + \frac{\alpha}{N^2},\\
\mathrm{Var}[\Gamma_i(j,k)] &= \mathbb{E}[\Gamma_i(j,k)^2] - \mathbb{E}[\Gamma_i(j,k)]^2 \approx \frac{\alpha}{TN} + \frac{\alpha-\alpha^2}{N^2}.
\end{aligned}$$

The range of $\Gamma_i(j,k)$ is

$$|\Gamma_i(j,k) - \mathbb{E}[\Gamma_i(j,k)]| \lessapprox 1,$$

and the extreme case is when $\bar{y} = q$ and the sequence is all q's.

Case 3: $j = N+1, k \in [N] \setminus \{q\}$.

Similar to Case 1, we have $\mathbb{1}\{z_{i,T+1} = N+1\}$ is independent of $\sum_{t\leqslant T}\mathbb{1}\{z_{i,t}=k\}$.

From Lemma C.6, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q] &\approx -\alpha\cdot\frac{1}{N},\\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q] &\approx \alpha\cdot\left(\frac{1}{TN}+\frac{1}{N^2}\right).
\end{aligned}$$

From Lemma C.10, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q] &\approx -\alpha\cdot\frac{1}{N},\\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q] &\approx \alpha\cdot\left(\frac{1}{TN}+\frac{1}{N^2}\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)] &\approx -\alpha \cdot \frac{1}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2] &\approx \alpha \cdot \left(\frac{1}{TN} + \frac{1}{N^2}\right), \\
\mathrm{Var}[\Gamma_i(j,k)] &= \mathbb{E}[\Gamma_i(j,k)^2] - \mathbb{E}[\Gamma_i(j,k)]^2 \approx \frac{\alpha}{TN} + \frac{\alpha - \alpha^2}{N^2}.
\end{aligned}$$

The range of $\Gamma_i(j,k)$ is

$$|\Gamma_i(j,k) - \mathbb{E}[\Gamma_i(j,k)]| \lessapprox 1,$$

and the extreme case is when all of the sequence except the last one is k.

Case 4: $j = q, k = N+1$.

If $\bar{y} \neq q$, we always have $z_{i,T+1} \neq q$ because $z_{i,T+1} \in \{\bar{y}, N+1\}$. If conditioning on $\bar{y} = q$, it has probability $1-\alpha$ for $z_{i,T+1} = q$, independent of $\sum_{t \leqslant T} \mathbb{1}\{z_{i,t} = N+1\}$.

From Lemma C.8, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k) | \bar{y} \neq q] &\approx \frac{1}{N+1} \cdot \frac{\alpha}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2 | \bar{y} \neq q] &\approx \frac{1}{N+1} \cdot \left(\frac{\alpha}{TN} + \frac{\alpha^2}{N^2}\right).
\end{aligned}$$

From Lemma C.5, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k) | \bar{y} = q] &\approx -(1-\alpha) \cdot \frac{1}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2 | \bar{y} = q] &\approx (1-\alpha) \cdot \left(\frac{1}{TN} + \frac{1}{N^2}\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q] + \frac{N-1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q] \approx \frac{2\alpha-1}{N^2},\\
\mathbb{E}[\Gamma_i(j,k)^2] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q] + \frac{N-1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q] \approx \frac{1}{TN^2} + \frac{\alpha^2-\alpha+1}{N^3},\\
\mathrm{Var}[\Gamma_i(j,k)] &= \mathbb{E}[\Gamma_i(j,k)^2] - \mathbb{E}[\Gamma_i(j,k)]^2 \approx \frac{1}{TN^2} + \frac{\alpha^2-\alpha+1}{N^3}.
\end{aligned}$$

The range of $\Gamma_i(j,k)$ is

$$|\Gamma_i(j,k) - \mathbb{E}[\Gamma_i(j,k)]| \lessapprox \frac{1}{2},$$

and the extreme case is when $\bar{y}=q$ and half of the sequence is $N+1$ with the rest all being q.

Case 5: $j=q, k=q$.

Similar to Case 4, if $\bar{y}\neq q$, we always have $z_{i,T+1}\neq q$. If conditioning on $\bar{y}=q$, it has probability $1-\alpha$ for $z_{i,T+1}=q$, independent of $\sum_{t\leqslant T}\mathbb{1}\{z_{i,t}=q\}$.

From Lemma C.7, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q] &\approx \frac{1}{N+1}\cdot\frac{1}{N},\\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q] &\approx \frac{1}{N+1}\cdot\left(\frac{1}{TN}+\frac{1}{N^2}\right).
\end{aligned}$$

From Lemma C.4, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q] &\approx -(1-\alpha)\cdot\frac{1}{\alpha N},\\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q] &\approx (1-\alpha)\cdot\left(\frac{1}{\alpha TN}\left(-1+\frac{2}{\alpha^2}\right)+\frac{1}{\alpha^2N^2}\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q] + \frac{N-1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q] \approx \frac{2\alpha-1}{\alpha N^2},\\
\mathbb{E}[\Gamma_i(j,k)^2] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q] + \frac{N-1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q]\\
&\approx \frac{\alpha^3-\alpha^2-\alpha+2}{\alpha^3 TN^2} + \frac{\alpha^2-\alpha+1}{\alpha^2 N^3},\\
\mathrm{Var}[\Gamma_i(j,k)] &= \mathbb{E}[\Gamma_i(j,k)^2] - \mathbb{E}[\Gamma_i(j,k)]^2 \approx \frac{\alpha^3-\alpha^2-\alpha+2}{\alpha^3 TN^2} + \frac{\alpha^2-\alpha+1}{\alpha^2 N^3}.
\end{aligned}$$

The range of $\Gamma_i(j,k)$ is

$$|\Gamma_i(j,k) - \mathbb{E}[\Gamma_i(j,k)]| \lessapprox 1,$$

and the extreme case is when $\bar{y}=q$ and all of the sequence are q.

Case 6: $j=q, k\in[N]\setminus\{q\}$.

Similar to Case 4, if $\bar{y}\neq q$, we always have $z_{i,T+1}\neq q$. If conditioning on $\bar{y}=q$, it has probability $1-\alpha$ for $z_{i,T+1}=q$, independent of $\sum_{t\leqslant T}\mathbb{1}\{z_{i,t}=k\}$.

Moreover, we need to consider whether $\bar{y}=k$ or not.

From Lemma C.9, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q, k=\bar{y}] &\approx \frac{1}{N+1}\cdot\frac{2-\alpha}{N},\\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q, k=\bar{y}] &\approx \frac{1}{N+1}\cdot\left(\frac{2-\alpha}{TN}+\frac{(2-\alpha)^2}{N^2}\right).
\end{aligned}$$

From Lemma C.10, we have

$$\mathbb{E}[\Gamma_i(j,k)|\bar{y} \neq q, k \in [N] \setminus \{q, \bar{y}\}] \approx \frac{1}{N+1} \cdot \frac{1}{N},$$
$$\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} \neq q, k \in [N] \setminus \{q, \bar{y}\}] \approx \frac{1}{N+1} \cdot \left(\frac{1}{TN} + \frac{1}{N^2}\right).$$

From Lemma C.6, we have

$$\mathbb{E}[\Gamma_i(j,k)|\bar{y} = q] \approx -(1-\alpha) \cdot \frac{1}{N},$$
$$\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} = q] \approx (1-\alpha) \cdot \left(\frac{1}{TN} + \frac{1}{N^2}\right).$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y} = q] + \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y} \neq q, k = \bar{y}] \\
&\quad + \frac{N-2}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y} \neq q, k \in [N] \setminus \{q, \bar{y}\}] \\
&\approx \frac{\alpha}{N^2}, \\
\mathbb{E}[\Gamma_i(j,k)^2] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} = q] + \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} \neq q, k = \bar{y}] \\
&\quad + \frac{N-2}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} \neq q, k \in [N] \setminus \{q, \bar{y}\}] \\
&\approx (2-\alpha) \cdot \left(\frac{1}{TN^2} + \frac{1}{N^3}\right), \\
\mathrm{Var}[\Gamma_i(j,k)] &= \mathbb{E}[\Gamma_i(j,k)^2] - \mathbb{E}[\Gamma_i(j,k)]^2 \approx (2-\alpha) \cdot \left(\frac{1}{TN^2} + \frac{1}{N^3}\right).
\end{aligned}$$

The range of $\Gamma_i(j,k)$ is

$$|\Gamma_i(j,k) - \mathbb{E}[\Gamma_i(j,k)]| \lessapprox 1,$$

and the extreme case is when all of the sequence except the last one are k.

Case 7: $j \in [N] \setminus \{q\}, k = N+1$.

If $\bar{y} \neq j$, we always have $z_{i,T+1} \neq j$ because $z_{i,T+1} \in \{\bar{y}, N+1\}$. If conditioning on $\bar{y} = j$, it has probability $1-\alpha$ for $z_{i,T+1} = j$, independent of $\sum_{t \leqslant T} \mathbb{1}\{z_{i,t} = N+1\}$.

Moreover, in the case of $\bar{y} \neq j$, we need to discuss whether or not $\bar{y} = q$.

From Lemma C.5, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q] &\approx \frac{1}{N+1} \cdot \frac{1}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q] &\approx \frac{1}{N+1} \cdot \left(\frac{1}{TN} + \frac{1}{N^2}\right).
\end{aligned}$$

From Lemma C.8, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q, \bar{y}\neq j] &\approx \frac{1}{N+1} \cdot \frac{\alpha}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q, \bar{y}\neq j] &\approx \frac{1}{N+1} \cdot \left(\frac{\alpha}{TN} + \frac{\alpha^2}{N^2}\right).
\end{aligned}$$

From Lemma C.8, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}=j] &\approx -(1-\alpha) \cdot \frac{\alpha}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=j] &\approx (1-\alpha) \cdot \left(\frac{\alpha}{TN} + \frac{\alpha^2}{N^2}\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q] + \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=j] + \frac{N-2}{N}\mathbb{E}[\Gamma_i(j,k)|y\neq q,\bar{y}\neq j] \\
&\approx \frac{\alpha^2}{N^2}, \\
\mathbb{E}[\Gamma_i(j,k)^2] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q] + \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=j] + \frac{N-2}{N}\mathbb{E}[\Gamma_i(j,k)^2|y\neq q,\bar{y}\neq j] \\
&\approx (2-\alpha)\left(\frac{\alpha}{TN^2}+\frac{\alpha^2}{N^3}\right), \\
\mathrm{Var}[\Gamma_i(j,k)] &= \mathbb{E}[\Gamma_i(j,k)^2] - \mathbb{E}[\Gamma_i(j,k)]^2 \approx (2-\alpha)\left(\frac{\alpha}{TN^2}+\frac{\alpha^2}{N^3}\right).
\end{aligned}$$

The range of $\Gamma_i(j,k)$ is

$$|\Gamma_i(j,k) - \mathbb{E}[\Gamma_i(j,k)]| \lessapprox \frac{1}{3},$$

and the extreme case is when $\bar{y}=j$ and one-third of the sequence are k, where the sequence has a repeated pattern like $[q, j, N+1, q, j, N+1, \dots]$.

Case 8: $j \in [N]\setminus\{q\}, k = q$.

Similar to Case 7, if $\bar{y} \neq j$, we always have $z_{i,T+1} \neq j$. If conditioning on $\bar{y}=j$, it has probability $1-\alpha$ for $z_{i,T+1}=j$, independent of $\sum_{t\leqslant T}\mathbb{1}\{z_{i,t}=N+1\}$.

Moreover, in the case of $\bar{y}\neq j$, we need to discuss whether or not $\bar{y}=q$.

From Lemma C.4, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q] &\approx \frac{1}{N+1}\cdot\frac{1}{\alpha N}, \\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q] &\approx \frac{1}{N+1}\cdot\left(\frac{T}{\alpha N}\left(-1+\frac{2}{\alpha^2}\right)+\frac{T^2}{\alpha^2N^2}\right).
\end{aligned}$$

From Lemma C.7, we have

$$\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q,\bar{y}\neq j]\approx\frac{1}{N+1}\cdot\frac{1}{N},$$
$$\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q,\bar{y}\neq j]\approx\frac{1}{N+1}\cdot\left(\frac{1}{TN}+\frac{1}{N^2}\right).$$

From Lemma C.7, we have

$$\mathbb{E}[\Gamma_i(j,k)|\bar{y}=j]\approx-(1-\alpha)\cdot\frac{1}{N},$$
$$\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=j]\approx(1-\alpha)\cdot\left(\frac{1}{TN}+\frac{1}{N^2}\right).$$

Therefore, we have

$$\begin{aligned}\mathbb{E}[\Gamma_i(j,k)]&=\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q]+\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=j]+\frac{N-2}{N}\mathbb{E}[\Gamma_i(j,k)|y\neq q,\bar{y}\neq j]\\&\approx\frac{\alpha}{N^2},\\\mathbb{E}[\Gamma_i(j,k)^2]&=\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q]+\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=j]+\frac{N-2}{N}\mathbb{E}[\Gamma_i(j,k)^2|y\neq q,\bar{y}\neq j]\\&\approx(2-\alpha)\left(\frac{1}{TN^2}+\frac{1}{N^3}\right),\\\mathrm{Var}[\Gamma_i(j,k)]&=\mathbb{E}[\Gamma_i(j,k)^2]-\mathbb{E}[\Gamma_i(j,k)]^2\approx(2-\alpha)\left(\frac{1}{TN^2}+\frac{1}{N^3}\right).\end{aligned}$$

The range of $\Gamma_i(j,k)$ is

$$|\Gamma_i(j,k)-\mathbb{E}[\Gamma_i(j,k)]|\lessapprox\frac{1}{2},$$

and the extreme case is when $\bar{y}=j$ and half of the sequence are q.

Case 9: $j\in[N]\setminus\{q\},k=j$.

Similar to Case 7, if $\bar{y}\neq j$, we always have $z_{i,T+1}\neq j$. If conditioning on $\bar{y}=j$, it has

probability $1-\alpha$ for $z_{i,T+1}=j$, independent of $\sum_{t\leqslant T}\mathbb{1}\{z_{i,t}=N+1\}$.

Moreover, in the case of $\bar{y}\neq j$, we need to discuss whether or not $\bar{y}=q$.

From Lemma C.6, we have

$$\begin{aligned}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q]&\approx\frac{1}{N+1}\cdot\frac{1}{N},\\ \mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q]&\approx\frac{1}{N+1}\cdot\left(\frac{1}{TN}+\frac{1}{N^2}\right).\end{aligned}$$

From Lemma C.10, we have

$$\begin{aligned}\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q,\bar{y}\neq j]&\approx\frac{1}{N+1}\cdot\frac{1}{N},\\ \mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q,\bar{y}\neq j]&\approx\frac{1}{N+1}\cdot\left(\frac{1}{TN}+\frac{1}{N^2}\right).\end{aligned}$$

From Lemma C.9, we have

$$\begin{aligned}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=j]&\approx-(1-\alpha)\cdot\frac{2-\alpha}{N},\\ \mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=j]&\approx(1-\alpha)\cdot\left(\frac{2-\alpha}{TN}+\frac{(2-\alpha)^2}{N^2}\right).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}[\Gamma_i(j,k)]&=\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q]+\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=j]+\frac{N-2}{N}\mathbb{E}[\Gamma_i(j,k)|y\neq q,\bar{y}\neq j]\\ &\approx\frac{-\alpha^2+3\alpha-1}{N^2},\\ \mathbb{E}[\Gamma_i(j,k)^2]&=\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q]+\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=j]+\frac{N-2}{N}\mathbb{E}[\Gamma_i(j,k)^2|y\neq q,\bar{y}\neq j]\\ &\approx\frac{1+(1-\alpha)(2-\alpha)}{TN^2}+\frac{1+(1-\alpha)(2-\alpha)^2}{N^3},\\ \mathrm{Var}[\Gamma_i(j,k)]&=\mathbb{E}[\Gamma_i(j,k)^2]-\mathbb{E}[\Gamma_i(j,k)]^2\approx\frac{1+(1-\alpha)(2-\alpha)}{TN^2}+\frac{1+(1-\alpha)(2-\alpha)^2}{N^3}.\end{aligned}$$

The range of $\Gamma_i(j,k)$ is

$$|\Gamma_i(j,k) - \mathbb{E}[\Gamma_i(j,k)]| \lessapprox 1,$$

and the extreme case is when $\bar{y} = j$ and all of the sequence are $j = k$.

Case 10: $j \in [N] \setminus \{q\}, k \in [N] \setminus \{q, j\}$.

Similar to Case 7, if $\bar{y} \neq j$, we always have $z_{i,T+1} \neq j$. If conditioning on $\bar{y} = j$, it has probability $1 - \alpha$ for $z_{i,T+1} = j$, independent of $\sum_{t \leqslant T} \mathbb{1}\{z_{i,t} = N+1\}$.

Moreover, in the case of $\bar{y} \neq j$, we need to discuss whether or not $\bar{y} = q$.

From Lemma C.6, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y} = q] &\approx \frac{1}{N+1} \cdot \frac{1}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} = q] &\approx \frac{1}{N+1} \cdot \left(\frac{1}{TN} + \frac{1}{N^2}\right).
\end{aligned}$$

From Lemma C.10, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y} = j] &\approx -(1-\alpha) \cdot \frac{1}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} = j] &\approx (1-\alpha) \cdot \left(\frac{1}{TN} + \frac{1}{N^2}\right).
\end{aligned}$$

From Lemma C.9, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)|\bar{y} = k] &\approx \frac{1}{N+1} \cdot \frac{2-\alpha}{N}, \\
\mathbb{E}[\Gamma_i(j,k)^2|\bar{y} = k] &\approx \frac{1}{N+1} \cdot \left(\frac{2-\alpha}{TN} + \frac{(2-\alpha)^2}{N^2}\right).
\end{aligned}$$

From Lemma C.10, we have

$$\mathbb{E}[\Gamma_i(j,k)|\bar{y}\neq q,\bar{y}\neq j,\bar{y}\neq k]\approx\frac{1}{N+1}\cdot\frac{1}{N},$$
$$\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}\neq q,\bar{y}\neq j,\bar{y}\neq k]\approx\frac{1}{N+1}\cdot\left(\frac{1}{TN}+\frac{1}{N^2}\right).$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}[\Gamma_i(j,k)] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=q]+\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=j]+\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)|\bar{y}=k]\\
&\quad+\frac{N-3}{N}\mathbb{E}[\Gamma_i(j,k)|y\neq q,\bar{y}\neq j]\\
&\approx\frac{\alpha}{N^2},\\
\mathbb{E}[\Gamma_i(j,k)^2] &= \frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=q]+\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=j]+\frac{1}{N}\mathbb{E}[\Gamma_i(j,k)^2|\bar{y}=k]\\
&\quad+\frac{N-3}{N}\mathbb{E}[\Gamma_i(j,k)^2|y\neq q,\bar{y}\neq j]\\
&\approx(2-\alpha)\left(\frac{1}{TN}+\frac{1}{N^2}\right),\\
\mathrm{Var}[\Gamma_i(j,k)] &= \mathbb{E}[\Gamma_i(j,k)^2]-\mathbb{E}[\Gamma_i(j,k)]^2\approx(2-\alpha)\left(\frac{1}{TN^2}+\frac{1}{N^3}\right).
\end{aligned}$$

The range of $\Gamma_i(j,k)$ is

$$|\Gamma_i(j,k)-\mathbb{E}[\Gamma_i(j,k)]|\lessapprox 1,$$

and the extreme case is when $\bar{y}=j$ and all of the sequence except the last are k.

□

C.2.3 Completing the Proof of Theorem 4.1

Theorem C.3 (Restatement of Theorem 4.1)**.** *Assume $N, T \gg 1, \alpha = \Theta(1)$. Consider a one gradient step update from zero-initialization on m i.i.d. samples of $z_{1:T}$ with separate learning rates η_f for $\mathbf{W}_F$ and η_v for $\mathbf{W}_V$ (note that the gradient on $\mathbf{W}_{QK}$ is zero). For a test sequence $z_{1:T}$, the resulting logits for the feed-forward and attention blocks satisfy, with probability $1-\delta$*

$$\left|\Delta(\xi_{ff}(x_{1:T})) - \eta_f \cdot \alpha\right| \leqslant \eta_f \cdot O\left(\sqrt{\frac{\ln\frac{2(N+1)}{\delta}}{m}}\right),$$

$$\left|\Delta(\xi_{attn}(x_{1:T})) - \frac{\eta_v}{N}\cdot(\alpha^2\hat{q} + \alpha(1-\hat{q}))\right| \leqslant \eta_v \cdot O\left(\sqrt{\frac{(\frac{1}{TN}+\frac{1}{N^2})\ln\frac{2(N+1)}{\delta}}{m}} + \frac{\ln\frac{2(N+1)}{\delta}}{m}\right),$$

where $\Delta(\xi) = \xi_{N+1} - \max_{j\in[N]}\xi_j$ is the margin of predicting the noise token and $\hat{q} = \frac{1}{T}\sum_{t\leqslant T}\mathbb{1}\{z_t = N+1\}$.

Proof. For $\mathbf{W}_F$, since the input is always $z_T = q$, the logits will be $[\xi_{\text{ff}}]_k = \mathbf{W}_U(k)^\top \mathbf{W}_F \mathbf{W}_E(q)$, $\forall\, k \in [N+1]$. As $\mathbf{W}_F$ is initialized from 0 and updated by GD with learning rate η_f, after one-step update, we have

$$\xi_{\text{ff}} = \mathbf{W}_U(k)^\top\left(-\eta_f \nabla_{\mathbf{W}_F}\hat{L}\Big|_{\mathbf{W}_F=0}\right)\mathbf{W}_E(q) \in \mathbb{R}^{N+1}.$$

By Lemma C.1, with probability $1-\frac{1}{2}\delta$, we have

$$\left|[\xi_{\text{ff}}]_{N+1} - \eta_f\cdot\alpha\right| \leqslant \eta_f\cdot O\left(\sqrt{\frac{\ln\frac{2(N+1)}{\delta}}{m}}\right),$$

$$\forall\, k \leqslant N,\quad \left|[\xi_{\text{ff}}]_k - \eta_f\cdot\left(\frac{1-\alpha}{N} - \frac{1}{N+1}\right)\right| \leqslant \eta_f\cdot O\left(\sqrt{\frac{\ln\frac{2(N+1)}{\delta}}{Nm}} + \frac{\ln\frac{2(N+1)}{\delta}}{m}\right),$$

and then triangle inequality finishes the proof for ξ_{ff}.

For $\mathbf{W}_V$, since the gradient on $\mathbf{W}_{QK}$ at initialization is zero, $\mathbf{W}_{QK}$ being zero after the first step induces a uniform attention over the input sequence. Consider the input sequence $\{z_i\}_{i=1}^T$, then the logits will be $[\xi_{\text{attn}}]_j = \mathbf{W}_U(j)^\top \mathbf{W}_V \frac{1}{T} \sum_{t=1}^T \mathbf{W}_E(z_t),\ \forall\ j \in [N+1]$.

Then considering the concentration bound of $\mathbf{W}_V$ after one-step update in Lemma C.2, denoting $\Gamma(j,k) = \mathbf{W}_U(j)^\top \mathbf{W}_V \mathbf{W}_E(k)$, we have

$$[\xi_{\text{attn}}]_j = \frac{1}{T} \sum_{t \leqslant T} \Gamma(j, z_t) = \frac{1}{T} \sum_{k \leqslant N+1} n_k \cdot \Gamma(j,k),$$

with concentration bound for each $\Gamma(\cdot,\cdot)$ in Lemma C.2. From Table C.1, note that for all $j = N+1, k \leqslant N$, the expectation and variances are the same, while $k = N+1$ has slightly different expectation and variance (but still in the same order of the others). Hence, denoting $\hat{q} = \frac{1}{T} \sum_{t \leqslant T} \mathbb{1}\{z_t = N+1\}$ dependent of the test sample $z_{1:T}$, we have

$$\left| [\xi_{\text{attn}}(x_{1:T})]_{N+1} - \frac{\eta_v}{N} \cdot (\alpha^2 \hat{q} + \alpha(1-\hat{q})) \right| \leqslant \eta_v \cdot O\left(\sqrt{\frac{(\frac{1}{TN} + \frac{1}{N^2}) \ln \frac{2(N+1)}{\delta}}{m}} + \frac{\ln \frac{2(N+1)}{\delta}}{m} \right).$$

Meanwhile, as the terms in Table C.1 for $j \neq N+1$ always have much smaller mean and variance by a factor $1/N$, using the Bernstein's inequalites for these terms in Lemma C.2 finishes the proof for $\mathbf{W}_V$.

□

C.3 Proof for First and Second moments in Lemma C.2

In this section, we will show the proof of the first and second moments of $\left[\sum_{1 \leqslant t \leqslant T} \mathbb{1}\{z_t = k\} | \cdot \right]$ for all cases. Note that we do not consider $z_T = q$, but including it will not change the results, as $T \gg 1$ and z_T is explicitly fixed as q during data generation in Section 4.4. Generally, there are

three factors to classify the cases as follows:

1. The i.i.d. uniformly sampled correct token $\bar{y} \in [N]$:
   (a) $\bar{y} = q$,
   (b) $\bar{y} \neq q$.
2. The target token $k \in [N+1]$:
   (a) $k = q$,
   (b) $k = N+1$.
   (c) $k \leqslant N, k \neq q, k \neq \bar{y}$,
   (d) (if $\bar{y} \neq q$) $k \leqslant N, k \neq q, k = \bar{y}$,
3. A condition about the token z_0 before the sequence $\{z_t\}_{t \geqslant 1}$:
   (a) $z_0 = q$,
   (b) $z_0 \in [N+1] \setminus \{q\}$.

Note that when z_0 will be implicitly or explicitly considered. When there is no condition on the first token, which means $z_1 \sim \text{Uniform}([N])$, this belongs to Case (3b), *i.e.*, $z_0 \in [N+1] \setminus \{q\}$, following the data generation process.

Table C.2 summarizes all lemmas about the seven cases classified by the first two factors. The third factor about z_0 is explicitly presented in the proof of each corresponding lemma.

Table C.2: All lemmas about the seven cases classified by $\bar{y}$ and k.

	(2a)	(2b)	(2c)	(2d)
(1a)	C.4	C.5	C.6	N/A
(1b)	C.7	C.8	C.10	C.9

C.3.1 When $\bar{y} = q$

Lemma C.4 ($\bar{y} = q, k = q$)**.** *Following the data generation process, assuming $N, T \gg 1$ and $\alpha = \Theta(1)$, if $\bar{y} = q$ and $k = q$, it holds*

$$\begin{aligned}\mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\middle|\bar{y} = q, k = q\right] &\approx \frac{T}{\alpha N},\\ \mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\right)^2\middle|\bar{y} = q, k = q\right] &\approx \frac{T}{\alpha N}\left(-1 + \frac{2}{\alpha^2}\right) + \frac{T^2}{\alpha^2 N^2}.\end{aligned} \tag{C.9}$$

Proof. For simplicity, we omit the condition of $\bar{y} = q, k = q$ in this proof. Denote

$$\begin{aligned}Y(T) &\triangleq \mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\middle|z_0 = q\right],\\ \hat{Y}(T) &\triangleq \mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\middle|z_0 \in [N+1], z_0 \neq q\right].\end{aligned}$$

Then the data generation process implies, $\forall\, T \geqslant 1$,

$$\begin{aligned}Y(T) &= p(z_1 = q|z_0 = q)\cdot(1 + Y(T-1)) + p(z_1 = N+1|z_0 = q)\cdot\hat{Y}(T-1),\\ \hat{Y}(T) &= p(z_1 = q|z_0 \neq q)\cdot(1 + Y(T-1)) + p(z_1 \in [N]\setminus\{q\}|z_0 \neq q)\cdot\hat{Y}(T-1).\end{aligned}$$

The iteration becomes

$$\begin{aligned}Y(T) &= (1-\alpha)\cdot Y(T-1) + \alpha\cdot\hat{Y}(T-1) + 1 - \alpha,\\ \hat{Y}(T) &= \frac{1}{N}\cdot Y(T-1) + \frac{N-1}{N}\cdot\hat{Y}(T-1) + \frac{1}{N}.\end{aligned}$$

This gives

$$
\begin{aligned}
Y(T) - \hat{Y}(T) &= (1 - \alpha - \frac{1}{N})(Y(T-1) - \hat{Y}(T-1)) + 1 - \alpha - \frac{1}{N}, \\
\frac{1}{N} Y(T) + \alpha \hat{Y}(T) &= \frac{1}{N} Y(T-1) + \alpha \hat{Y}(T-1) + \frac{1}{N}.
\end{aligned}
$$

Consider the initialization $Y(0) = \hat{Y}(0) = 0$. This implies

$$
\begin{aligned}
Y(T) - \hat{Y}(T) &= \frac{1 - \alpha - \frac{1}{N}}{\alpha + \frac{1}{N}} \left(1 - \left(1 - \alpha - \frac{1}{N} \right)^T \right), \\
\frac{1}{N} Y(T) + \alpha \hat{Y}(T) &= \frac{1}{N} T.
\end{aligned}
$$

Then we obtain

$$
\begin{aligned}
Y(T) &\approx \frac{1}{\alpha N + 1} (T - \alpha N) + \frac{\alpha}{(\alpha + \frac{1}{N})^2} = \frac{1}{\alpha N + 1} \left(T - \alpha N + \frac{N^2}{\alpha N + 1} \right) \\
&\approx \frac{T}{\alpha N} - 1 + \frac{1}{\alpha^2}, \\
\hat{Y}(T) &\approx \frac{1}{\alpha N + 1} T - \frac{N}{(\alpha N + 1)^2} + \frac{1}{\alpha N + 1} \\
&\approx \frac{T}{\alpha N}.
\end{aligned}
$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$
\mathbb{E} \left[\sum_{t \leqslant T} \mathbb{1}\{z_t = k\} \middle| \bar{y} = q, k = q \right] = \hat{Y}(T) \approx \frac{T}{\alpha N}.
$$

To obtain the expectation of the quadratic term, we similarly denote the following terms with

different z_0:

$$Z(T) \triangleq \mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\right)^2 \middle| z_0 = q\right],$$
$$\hat{Z}(T) \triangleq \mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\right)^2 \middle| z_0 \in [N+1], z_0 \neq q\right].$$

Then the data generation process implies, $\forall\, T \geqslant 1$,

$$Z(T) = p(z_1 = q|z_0 = q)\cdot(1 + 2Y(T-1) + Z(T-1)) + p(z_1 = N+1|z_0 = q)\cdot Z(T-1),$$
$$\hat{Z}(T) = p(z_1 = q|z_0 \neq q)\cdot(1 + 2Y(T-1) + Z(T-1)) + p(z_1 \neq q|z_0 \neq q)\cdot \hat{Z}(T-1),$$

where $2Y(T-1)$ is due to $\mathbb{E}[(1+\sum_{2\leqslant t\leqslant T}\cdot)^2] = 1 + 2\mathbb{E}[\sum_{2\leqslant t\leqslant T}\cdot] + \mathbb{E}[(\sum_{2\leqslant t\leqslant T}\cdot)^2]$.

Then the iteration becomes

$$\begin{aligned} Z(T) &= (1-\alpha)\cdot(1 + 2Y(T-1) + Z(T-1)) + \alpha\cdot\hat{Z}(T-1) \\ &= (1-\alpha)Z(T-1) + \alpha\hat{Z}(T-1) + (1-\alpha)(1+2Y(T-1)), \\ \hat{Z}(T) &= \frac{1}{N}\cdot(1 + 2Y(T-1) + Z(T-1)) + \frac{N-1}{N}\cdot\hat{Z}(T-1) \\ &= \frac{1}{N}Z(T-1) + \frac{N-1}{N}\hat{Z}(T-1) + \frac{1}{N}(1+2Y(T-1)). \end{aligned}$$

This gives

$$\begin{aligned} Z(T) - \hat{Z}(T) &= (1-\alpha-\frac{1}{N})(Z(T-1) - \hat{Z}(T-1)) + (1-\alpha-\frac{1}{N})(1+2Y(T-1)), \\ \frac{1}{N}Z(T) + \alpha\hat{Z}(T) &= \frac{1}{N}Z(T-1) + \alpha\hat{Z}(T-1) + \frac{1}{N}(1+2Y(T-1)). \end{aligned}$$

Considering the initialization $Z(0) = \hat{Z}(0) = 0$, we have

$$\begin{aligned}
Z(T) - \hat{Z}(T) &= \sum_{t \leqslant T-1} (1 - \alpha - \frac{1}{N})^{T-t}(1 + 2Y(t)) \\
&\approx \sum_{t \leqslant T-1} (1 - \alpha - \frac{1}{N})^{T-t} \left(1 + \frac{2t}{\alpha N} - 2 + \frac{2}{\alpha^2}\right) \\
&\approx \left(-1 + \frac{2}{\alpha^2}\right) \frac{1-\alpha}{\alpha} + \frac{2(1-\alpha)}{\alpha^2} \cdot \frac{T}{N}. \\
\frac{1}{N} Z(T) + \alpha \hat{Z}(T) &= \frac{T}{N} + \frac{2}{N} \sum_{1 \leqslant t \leqslant T-1} Y(t) \\
&\approx \frac{T}{N} + \frac{2}{N} \sum_{1 \leqslant t \leqslant T-1} \left(\frac{t}{\alpha N} - 1 + \frac{1}{\alpha^2}\right) \\
&\approx \frac{T}{N} \left(-1 + \frac{2}{\alpha^2}\right) + \frac{T^2}{\alpha N^2}.
\end{aligned}$$

Then we obtain

$$\begin{aligned}
Z(T) &\approx \frac{T}{N} \left(-\frac{3}{\alpha} + \frac{2}{\alpha^2} + \frac{2}{\alpha^3}\right) + \frac{T^2}{\alpha^2 N^2} + \frac{1-\alpha}{\alpha}(\frac{2}{\alpha^2} - 1), \\
\hat{Z}(T) &\approx \frac{T}{\alpha N} \left(-1 + \frac{2}{\alpha^2}\right) + \frac{T^2}{\alpha^2 N^2}.
\end{aligned}$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$\mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| \bar{y} = q, k = q\right] = \hat{Z}(T) \approx \frac{T}{\alpha N} \left(-1 + \frac{2}{\alpha^2}\right) + \frac{T^2}{\alpha^2 N^2}.$$

□

Lemma C.5 ($\bar{y} = q, k = N + 1$). *Following the data generation process, assuming $N, T \gg 1$ and*

$\alpha = \Theta(1)$, *if* $\bar{y} = q$ *and* $k = N + 1$, *it holds*

$$\begin{aligned} \mathbb{E}\left[\sum_{t\leqslant T} \mathbb{1}\{z_t = k\} \middle| \bar{y} = q, k = N+1\right] &\approx \frac{T}{N}, \\ \mathbb{E}\left[\left(\sum_{t\leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| \bar{y} = q, k = N+1\right] &\approx \frac{T}{N} + \frac{T^2}{N^2}. \end{aligned} \tag{C.10}$$

Proof. For simplicity, we omit the condition of $\bar{y} = q, k = N + 1$ in this proof. Denote

$$\begin{aligned} Y(T) &\triangleq \mathbb{E}\left[\sum_{t\leqslant T} \mathbb{1}\{z_t = k\} \middle| z_0 = q\right], \\ \hat{Y}(T) &\triangleq \mathbb{E}\left[\sum_{t\leqslant T} \mathbb{1}\{z_t = k\} \middle| z_0 \in [N+1], z_0 \neq q\right]. \end{aligned}$$

Then the data generation process implies, $\forall\, T \geqslant 1$,

$$\begin{aligned} Y(T) &= p(z_1 = q | z_0 = q) \cdot Y(T-1) + p(z_1 = N+1 | z_0 = q) \cdot (1 + \hat{Y}(T-1)), \\ \hat{Y}(T) &= p(z_1 = q | z_0 \neq q) \cdot Y(T-1) + p(z_1 \in [N] \setminus \{q\} | z_0 \neq q) \cdot \hat{Y}(T-1). \end{aligned}$$

The iteration becomes

$$\begin{aligned} Y(T) &= (1-\alpha) \cdot Y(T-1) + \alpha \cdot \hat{Y}(T-1) + \alpha, \\ \hat{Y}(T) &= \frac{1}{N} \cdot Y(T-1) + \frac{N-1}{N} \cdot \hat{Y}(T-1). \end{aligned}$$

This gives

$$\begin{aligned} Y(T) - \hat{Y}(T) &= (1 - \alpha - \frac{1}{N})(Y(T-1) - \hat{Y}(T-1)) + \alpha, \\ \frac{1}{N} Y(T) + \alpha \hat{Y}(T) &= \frac{1}{N} Y(T-1) + \alpha \hat{Y}(T-1) + \frac{\alpha}{N}. \end{aligned}$$

Consider the initialization $Y(0) = \hat{Y}(0) = 0$. This implies

$$Y(T) - \hat{Y}(T) = \frac{\alpha}{\alpha + \frac{1}{N}} \left(1 - \left(1 - \alpha - \frac{1}{N}\right)^T\right),$$
$$\frac{1}{N} Y(T) + \alpha \hat{Y}(T) = \frac{\alpha}{N} T.$$

Then we obtain

$$Y(T) \approx \frac{T}{N} + 1,$$
$$\hat{Y}(T) \approx \frac{T}{N}.$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$\mathbb{E}\left[\sum_{t \leqslant T} \mathbb{1}\{z_t = k\} \middle| \bar{y} = q, k = N + 1\right] = \hat{Y}(T) \approx \frac{T}{N}.$$

To obtain the expectation of the quadratic term, we similarly denote the following terms with different z_0:

$$Z(T) \triangleq \mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| z_0 = q\right],$$
$$\hat{Z}(T) \triangleq \mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| z_0 \in [N+1], z_0 \neq q\right].$$

Then the data generation process implies, $\forall\, T \geqslant 1$,

$$Z(T) = p(z_1 = q | z_0 = q) \cdot Z(T-1) + p(z_1 = N+1 | z_0 = q) \cdot (1 + 2\hat{Y}(T-1) + \hat{Z}(T-1)),$$
$$\hat{Z}(T) = p(z_1 = q | z_0 \neq q) \cdot Z(T-1) + p(z_1 \neq q | z_0 \neq q) \cdot \hat{Z}(T-1),$$

where $2\hat{Y}(T-1)$ is due to $\mathbb{E}[(1 + \sum_{2 \leqslant t \leqslant T} \cdot)^2] = 1 + 2\mathbb{E}[\sum_{2 \leqslant t \leqslant T} \cdot] + \mathbb{E}[(\sum_{2 \leqslant t \leqslant T} \cdot)^2]$.

Then the iteration becomes

$$
\begin{aligned}
Z(T) &= (1-\alpha)\cdot Z(T-1) + \alpha\cdot(1+2\hat{Y}(T-1)+\hat{Z}(T-1)) \\
&= (1-\alpha)Z(T-1) + \alpha\hat{Z}(T-1) + \alpha(1+2\hat{Y}(T-1)), \\
\hat{Z}(T) &= \frac{1}{N}\cdot Z(T-1) + \frac{N-1}{N}\cdot\hat{Z}(T-1).
\end{aligned}
$$

This gives

$$
\begin{aligned}
Z(T)-\hat{Z}(T) &= (1-\alpha-\frac{1}{N})(Z(T-1)-\hat{Z}(T-1)) + \alpha(1+2\hat{Y}(T-1)), \\
\frac{1}{N}Z(T)+\alpha\hat{Z}(T) &= \frac{1}{N}Z(T-1)+\alpha\hat{Z}(T-1)+\frac{\alpha}{N}(1+2\hat{Y}(T-1)).
\end{aligned}
$$

Considering the initialization $Z(0)=\hat{Z}(0)=0$, we have

$$
\begin{aligned}
Z(T)-\hat{Z}(T) &= \sum_{t\leqslant T-1}\alpha(1-\alpha-\frac{1}{N})^{T-1-t}(1+2\hat{Y}(t)) \\
&\approx \sum_{t\leqslant T-1}\alpha(1-\alpha-\frac{1}{N})^{T-1-t}\left(1+\frac{2t}{N}\right) \\
&\approx \frac{2T}{N}+1, \\
\frac{1}{N}Z(T)+\alpha\hat{Z}(T) &= \frac{\alpha T}{N}+\frac{2\alpha}{N}\sum_{1\leqslant t\leqslant T-1}\hat{Y}(t) \\
&\approx \frac{\alpha T}{N}+\frac{2\alpha}{N}\sum_{1\leqslant t\leqslant T-1}\frac{t}{N} \\
&\approx \frac{\alpha T}{N}+\frac{\alpha T^2}{N^2}.
\end{aligned}
$$

Then we obtain

$$
\begin{aligned}
Z(T) &\approx 3\alpha\frac{T}{N}+\alpha\frac{T^2}{N^2}+\alpha, \\
\hat{Z}(T) &\approx \frac{T}{N}+\frac{T^2}{N^2}.
\end{aligned}
$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$\mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\right)^2 \middle| \bar{y} = q, k = N+1\right] = \hat{Z}(T) \approx \frac{T}{N} + \frac{T^2}{N^2}.$$

□

Lemma C.6 ($\bar{y} = q, k \leqslant N, k \neq q$)**.** *Following the data generation process, assuming $N, T \gg 1$ and $\alpha = \Theta(1)$, if $\bar{y} = q$ and $k \in [N] \setminus \{q\}$, it holds*

$$\begin{aligned}\mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t = k\} \middle| \bar{y} = q, k \in [N]\setminus\{q\}\right] &\approx \frac{T}{N},\\ \mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\right)^2 \middle| \bar{y} = q, k \in [N]\setminus\{q\}\right] &\approx \frac{T}{N} + \frac{T^2}{N^2}.\end{aligned} \tag{C.11}$$

Proof. For simplicity, we omit the condition of $\bar{y} = q, k \in [N] \setminus \{q\}$ in this proof. Denote

$$\begin{aligned}Y(T) &\triangleq \mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t = k\} \middle| z_0 = q\right],\\ \hat{Y}(T) &\triangleq \mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t = k\} \middle| z_0 \in [N+1], z_0 \neq q\right].\end{aligned}$$

Then the data generation process implies, $\forall\, T \geqslant 1$,

$$\begin{aligned}Y(T) &= p(z_1 = q | z_0 = q)\cdot Y(T-1) + p(z_1 = N+1 | z_0 = q)\cdot \hat{Y}(T-1),\\ \hat{Y}(T) &= p(z_1 = q | z_0 \neq q)\cdot Y(T-1)\\ &\quad + p(z_1 \in [N]\setminus\{q\} | z_0 \neq q)\cdot (p(z_1 = k | z_1 \sim \text{Uniform}([N]\setminus\{q\}) + \hat{Y}(T-1)).\end{aligned}$$

The iteration becomes

$$
\begin{aligned}
Y(T) &= (1-\alpha)\cdot Y(T-1) + \alpha \cdot \hat{Y}(T-1), \\
\hat{Y}(T) &= \frac{1}{N}\cdot Y(T-1) + \frac{N-1}{N}\cdot (\hat{Y}(T-1) + \frac{1}{N-1}).
\end{aligned}
$$

This gives

$$
\begin{aligned}
Y(T) - \hat{Y}(T) &= (1-\alpha-\frac{1}{N})(Y(T-1)-\hat{Y}(T-1)) - \frac{1}{N}, \\
\frac{1}{N}Y(T) + \alpha\hat{Y}(T) &= \frac{1}{N}Y(T-1) + \alpha\hat{Y}(T-1) + \frac{\alpha}{N}.
\end{aligned}
$$

Consider the initialization $Y(0) = \hat{Y}(0) = 0$. This implies

$$
\begin{aligned}
Y(T) - \hat{Y}(T) &= \frac{-\frac{1}{N}}{\alpha + \frac{1}{N}}\left(1 - \left(1-\alpha-\frac{1}{N}\right)^T\right), \\
\frac{1}{N}Y(T) + \alpha\hat{Y}(T) &= \frac{\alpha}{N}T.
\end{aligned}
$$

Then we obtain

$$
\begin{aligned}
Y(T) &\approx \frac{T}{N}, \\
\hat{Y}(T) &\approx \frac{T}{N}.
\end{aligned}
$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$
\mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\middle|\bar{y} = q, k = N+1\right] = \hat{Y}(T) \approx \frac{T}{N}.
$$

To obtain the expectation of the quadratic term, we similarly denote the following terms with

different z_0:

$$Z(T) \triangleq \mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t=k\}\right)^2\middle| z_0=q\right],$$
$$\hat{Z}(T) \triangleq \mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t=k\}\right)^2\middle| z_0\in[N+1], z_0\neq q\right].$$

Then the data generation process implies, $\forall\, T\geqslant 1$,

$$\begin{aligned}
Z(T) &= p(z_1=q|z_0=q)\cdot Z(T-1) + p(z_1=N+1|z_0=q)\cdot \hat{Z}(T-1),\\
\hat{Z}(T) &= p(z_1=q|z_0\neq q)\cdot Z(T-1) + p(z_1\neq q|z_0\neq q)\cdot \hat{Z}(T-1)\\
&\quad + p(z_1=k|z_0\neq q)\cdot(1+2\hat{Y}(T-1)),
\end{aligned}$$

where $2\hat{Y}(T-1)$ is due to $\mathbb{E}[(1+\sum_{2\leqslant t\leqslant T}\cdot)^2] = 1+2\mathbb{E}[\sum_{2\leqslant t\leqslant T}\cdot] + \mathbb{E}[(\sum_{2\leqslant t\leqslant T}\cdot)^2]$.

Then the iteration becomes

$$\begin{aligned}
Z(T) &= (1-\alpha)\cdot Z(T-1) + \alpha\cdot\hat{Z}(T-1),\\
\hat{Z}(T) &= \frac{1}{N}\cdot Z(T-1) + \frac{N-1}{N}\cdot\hat{Z}(T-1) + \frac{1}{N}(1+2\hat{Y}(T-1)).
\end{aligned}$$

This gives

$$\begin{aligned}
Z(T)-\hat{Z}(T) &= (1-\alpha-\frac{1}{N})(Z(T-1)-\hat{Z}(T-1)) - \frac{1}{N}(1+2\hat{Y}(T-1)),\\
\frac{1}{N}Z(T)+\alpha\hat{Z}(T) &= \frac{1}{N}Z(T-1)+\alpha\hat{Z}(T-1)+\frac{\alpha}{N}(1+2\hat{Y}(T-1)).
\end{aligned}$$

Considering the initialization $Z(0) = \hat{Z}(0) = 0$, we have

$$\begin{aligned}
Z(T) - \hat{Z}(T) &= -\frac{1}{N} \sum_{t \leqslant T-1} (1 - \alpha - \frac{1}{N})^{T-1-t} (1 + 2\hat{Y}(t)) \\
&\approx -\frac{1}{N} \sum_{t \leqslant T-1} (1 - \alpha - \frac{1}{N})^{T-1-t} \left(1 + \frac{2t}{N}\right) \\
&\approx -\frac{1}{\alpha N} \left(\frac{2T}{N} + 1\right), \\
\frac{1}{N} Z(T) + \alpha \hat{Z}(T) &= \frac{\alpha T}{N} + \frac{2\alpha}{N} \sum_{1 \leqslant t \leqslant T-1} \hat{Y}(t) \\
&\approx \frac{\alpha T}{N} + \frac{2\alpha}{N} \sum_{1 \leqslant t \leqslant T-1} \frac{t}{N} \\
&\approx \frac{\alpha T}{N} + \frac{\alpha T^2}{N^2}.
\end{aligned}$$

Then we obtain

$$\begin{aligned}
Z(T) &\approx \frac{T}{N} + \frac{T^2}{N^2}, \\
\hat{Z}(T) &\approx \frac{T}{N} + \frac{T^2}{N^2}.
\end{aligned}$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$\mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| \bar{y} = q, k \in [N] \setminus \{q\}\right] = \hat{Z}(T) \approx \frac{T}{N} + \frac{T^2}{N^2}.$$

□

C.3.2 When $\bar{y} \neq q$

Lemma C.7 ($\bar{y} \neq q, k = q$)**.** *Following the data generation process, assuming $N, T \gg 1$ and $\alpha = \Theta(1)$, if $\bar{y} \neq q$ and $k = q$, it holds*

$$\begin{aligned} \mathbb{E}\left[\sum_{t \leqslant T} \mathbb{1}\{z_t = k\} \middle| \bar{y} \neq q, k = q\right] &\approx \frac{T}{N}, \\ \mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| \bar{y} \neq q, k = q\right] &\approx \frac{T}{N} + \frac{T^2}{N^2}. \end{aligned} \tag{C.12}$$

Proof. For simplicity, we omit the condition of $\bar{y} \neq q, k = q$ in this proof. Denote

$$\begin{aligned} Y(T) &\triangleq \mathbb{E}\left[\sum_{t \leqslant T} \mathbb{1}\{z_t = k\} \middle| z_0 = q\right], \\ \hat{Y}(T) &\triangleq \mathbb{E}\left[\sum_{t \leqslant T} \mathbb{1}\{z_t = k\} \middle| z_0 \in [N+1], z_0 \neq q\right]. \end{aligned}$$

Then the data generation process implies, $\forall\, T \geqslant 1$,

$$\begin{aligned} Y(T) &= \hat{Y}(T-1), \\ \hat{Y}(T) &= p(z_1 = q | z_0 \neq q) \cdot (1 + Y(T-1)) + p(z_1 \in [N] \setminus \{q\} | z_0 \neq q) \cdot \hat{Y}(T-1). \end{aligned}$$

The iteration becomes

$$\begin{aligned} Y(T) &= \hat{Y}(T-1), \\ \hat{Y}(T) &= \frac{1}{N} \cdot Y(T-1) + \frac{N-1}{N} \cdot \hat{Y}(T-1) + \frac{1}{N}. \end{aligned}$$

This gives

$$
\begin{aligned}
Y(T) - \hat{Y}(T) &= -\frac{1}{N}(Y(T-1) - \hat{Y}(T-1)) - \frac{1}{N}, \\
\frac{1}{N}Y(T) + \hat{Y}(T) &= \frac{1}{N}Y(T-1) + \hat{Y}(T-1) + \frac{1}{N}.
\end{aligned}
$$

Consider the initialization $Y(0) = \hat{Y}(0) = 0$. This implies

$$
\begin{aligned}
Y(T) - \hat{Y}(T) &= \frac{-\frac{1}{N}}{1+\frac{1}{N}}\left(1 - \left(-\frac{1}{N}\right)^T\right), \\
\frac{1}{N}Y(T) + \hat{Y}(T) &= \frac{1}{N}T.
\end{aligned}
$$

Then we obtain

$$
\begin{aligned}
Y(T) &\approx \frac{T}{N}, \\
\hat{Y}(T) &\approx \frac{T}{N}.
\end{aligned}
$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$
\mathbb{E}\left[\sum_{t \leqslant T} \mathbb{1}\{z_t = k\} \middle| \bar{y} \neq q, k = q\right] = \hat{Y}(T) \approx \frac{T}{N}.
$$

To obtain the expectation of the quadratic term, we similarly denote the following terms with different z_0:

$$
\begin{aligned}
Z(T) &\triangleq \mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| z_0 = q\right], \\
\hat{Z}(T) &\triangleq \mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| z_0 \in [N+1], z_0 \neq q\right].
\end{aligned}
$$

Then the data generation process implies, $\forall\, T \geqslant 1$,

$$Z(T) = \hat{Z}(T-1),$$
$$\hat{Z}(T) = p(z_1 = q | z_0 \neq q) \cdot (1 + 2Y(T-1) + Z(T-1)) + p(z_1 \in [N] \setminus \{q\} | z_0 \neq q) \cdot \hat{Z}(T-1),$$

where $2Y(T-1)$ is due to $\mathbb{E}[(1 + \sum_{2 \leqslant t \leqslant T} \cdot)^2] = 1 + 2\mathbb{E}[\sum_{2 \leqslant t \leqslant T} \cdot] + \mathbb{E}[(\sum_{2 \leqslant t \leqslant T} \cdot)^2]$.

Then the iteration becomes

$$\begin{aligned} Z(T) &= \hat{Z}(T-1), \\ \hat{Z}(T) &= \frac{1}{N} Z(T-1) + \frac{N-1}{N} \hat{Z}(T-1) + \frac{1}{N}(1 + 2Y(T-1)). \end{aligned}$$

This gives

$$\begin{aligned} Z(T) - \hat{Z}(T) &= -\frac{1}{N}(Z(T-1) - \hat{Z}(T-1)) - \frac{1}{N}(1 + 2Y(T-1)), \\ \frac{1}{N} Z(T) + \hat{Z}(T) &= \frac{1}{N} Z(T-1) + \hat{Z}(T-1) + \frac{1}{N}(1 + 2Y(T-1)). \end{aligned}$$

Considering the initialization $Z(0) = \hat{Z}(0) = 0$, we have

$$\begin{aligned} Z(T) - \hat{Z}(T) &= -\frac{1}{N} \sum_{t \leqslant T-1} (-\frac{1}{N})^{T-1-t} (1 + 2Y(t)) \\ &\approx -\frac{1}{N} \sum_{t \leqslant T-1} (-\frac{1}{N})^{T-1-t} \left(1 + \frac{2t}{N}\right) \\ &\approx -\frac{1}{N} - \frac{2T}{N^2}, \\ \frac{1}{N} Z(T) + \hat{Z}(T) &= \frac{T}{N} + \frac{2}{N} \sum_{1 \leqslant t \leqslant T-1} Y(t) \\ &\approx \frac{T}{N} + \frac{2}{N} \sum_{1 \leqslant t \leqslant T-1} \frac{t}{N} \\ &\approx \frac{T}{N} + \frac{T^2}{N^2}. \end{aligned}$$

Then we obtain

$$Z(T) \approx \frac{T}{N} + \frac{T^2}{N^2},$$
$$\hat{Z}(T) \approx \frac{T}{N} + \frac{T^2}{N^2}.$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$\mathbb{E}\left[\left(\sum_{t\leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| \bar{y} = q, k \in [N] \setminus \{q\}\right] = \hat{Z}(T) \approx \frac{T}{N} + \frac{T^2}{N^2}.$$

□

Lemma C.8 ($\bar{y} \neq q, k = N+1$)**.** *Following the data generation process, assuming $N, T \gg 1$ and $\alpha = \Theta(1)$, if $\bar{y} \neq q$ and $k = N+1$, it holds*

$$\begin{aligned} \mathbb{E}\left[\sum_{t\leqslant T} \mathbb{1}\{z_t = k\} \middle| \bar{y} \neq q, k = N+1\right] &\approx \frac{\alpha T}{N}, \\ \mathbb{E}\left[\left(\sum_{t\leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| \bar{y} \neq q, k = N+1\right] &\approx \frac{\alpha T}{N} + \frac{\alpha^2 T^2}{N^2}. \end{aligned} \tag{C.13}$$

Proof. For simplicity, we omit the condition of $\bar{y} \neq q, k = N+1$ in this proof. Denote

$$Y(T) \triangleq \mathbb{E}\left[\sum_{t\leqslant T} \mathbb{1}\{z_t = k\} \middle| z_0 = q\right],$$
$$\hat{Y}(T) \triangleq \mathbb{E}\left[\sum_{t\leqslant T} \mathbb{1}\{z_t = k\} \middle| z_0 \in [N+1], z_0 \neq q\right].$$

Then the data generation process implies, $\forall\ T \geqslant 1$,

$$
\begin{aligned}
Y(T) &= \hat{Y}(T-1) + p(z_1 = N+1 | z_0 = q), \\
\hat{Y}(T) &= p(z_1 = q | z_0 \neq q) \cdot Y(T-1) + p(z_1 \in [N] \setminus \{q\} | z_0 \neq q) \cdot \hat{Y}(T-1).
\end{aligned}
$$

The iteration becomes

$$
\begin{aligned}
Y(T) &= \hat{Y}(T-1) + \alpha, \\
\hat{Y}(T) &= \frac{1}{N} \cdot Y(T-1) + \frac{N-1}{N} \cdot \hat{Y}(T-1).
\end{aligned}
$$

This gives

$$
\begin{aligned}
Y(T) - \hat{Y}(T) &= -\frac{1}{N}(Y(T-1) - \hat{Y}(T-1)) + \alpha, \\
\frac{1}{N} Y(T) + \hat{Y}(T) &= \frac{1}{N} Y(T-1) + \hat{Y}(T-1) + \frac{\alpha}{N}.
\end{aligned}
$$

Consider the initialization $Y(0) = \hat{Y}(0) = 0$. This implies

$$
\begin{aligned}
Y(T) - \hat{Y}(T) &= \frac{\alpha}{1 + \frac{1}{N}} \left(1 - \left(-\frac{1}{N} \right)^T \right), \\
\frac{1}{N} Y(T) + \hat{Y}(T) &= \frac{\alpha}{N} T.
\end{aligned}
$$

Then we obtain

$$
\begin{aligned}
Y(T) &\approx \frac{\alpha T}{N} + \alpha, \\
\hat{Y}(T) &\approx \frac{\alpha T}{N}.
\end{aligned}
$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$\mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\middle|\bar{y}\neq q, k = q\right] = \hat{Y}(T) \approx \frac{\alpha T}{N}.$$

To obtain the expectation of the quadratic term, we similarly denote the following terms with different z_0:

$$\begin{aligned}
Z(T) &\triangleq \mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\right)^2\middle| z_0 = q\right],\\
\hat{Z}(T) &\triangleq \mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t = k\}\right)^2\middle| z_0 \in [N+1], z_0 \neq q\right].
\end{aligned}$$

Then the data generation process implies, $\forall\, T \geqslant 1$,

$$\begin{aligned}
Z(T) &= \hat{Z}(T-1) + p(z_1 = N+1|z_0 = q)\cdot(1 + 2\hat{Y}(T-1)),\\
\hat{Z}(T) &= p(z_1 = q|z_0 \neq q)\cdot Z(T-1) + p(z_1 \in [N]\setminus\{q\}|z_0\neq q)\cdot\hat{Z}(T-1),
\end{aligned}$$

where $2\hat{Y}(T-1)$ is due to $\mathbb{E}[(1+\sum_{2\leqslant t\leqslant T}\cdot)^2] = 1 + 2\mathbb{E}[\sum_{2\leqslant t\leqslant T}\cdot] + \mathbb{E}[(\sum_{2\leqslant t\leqslant T}\cdot)^2]$.

Then the iteration becomes

$$\begin{aligned}
Z(T) &= \hat{Z}(T-1) + \alpha(1 + 2\hat{Y}(T-1)),\\
\hat{Z}(T) &= \frac{1}{N}Z(T-1) + \frac{N-1}{N}\hat{Z}(T-1).
\end{aligned}$$

This gives

$$\begin{aligned}
Z(T) - \hat{Z}(T) &= -\frac{1}{N}(Z(T-1) - \hat{Z}(T-1)) + \alpha(1 + 2\hat{Y}(T-1)),\\
\frac{1}{N}Z(T) + \hat{Z}(T) &= \frac{1}{N}Z(T-1) + \hat{Z}(T-1) + \frac{\alpha}{N}(1 + 2\hat{Y}(T-1)).
\end{aligned}$$

Considering the initialization $Z(0) = \hat{Z}(0) = 0$, we have

$$\begin{aligned}
Z(T) - \hat{Z}(T) &= \alpha \sum_{t \leqslant T-1} (-\frac{1}{N})^{T-1-t}(1 + 2\hat{Y}(t)) \\
&\approx \alpha \sum_{t \leqslant T-1} (-\frac{1}{N})^{T-1-t} \left(1 + \frac{2\alpha t}{N}\right) \\
&\approx \frac{2\alpha^2 T}{N} + \alpha, \\
\frac{1}{N} Z(T) + \hat{Z}(T) &= \frac{\alpha T}{N} + \frac{2\alpha}{N} \sum_{1 \leqslant t \leqslant T-1} \hat{Y}(t) \\
&\approx \frac{\alpha T}{N} + \frac{2\alpha}{N} \sum_{1 \leqslant t \leqslant T-1} \frac{\alpha t}{N} \\
&\approx \frac{\alpha T}{N} + \frac{\alpha^2 T^2}{N^2}.
\end{aligned}$$

Then we obtain

$$\begin{aligned}
Z(T) &\approx \frac{T}{N}(2\alpha^2 + \alpha) + \frac{\alpha^2 T^2}{N^2} + \alpha, \\
\hat{Z}(T) &\approx \frac{\alpha T}{N} + \frac{\alpha^2 T^2}{N^2}.
\end{aligned}$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$\mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| \bar{y} = q, k \in [N] \setminus \{q\}\right] = \hat{Z}(T) \approx \frac{\alpha T}{N} + \frac{\alpha^2 T^2}{N^2}.$$

□

Lemma C.9 ($\bar{y} \neq q, k = \bar{y}$). *Following the data generation process, assuming $N, T \gg 1$ and $\alpha =$*

$\Theta(1)$, *if* $\bar{y} \neq q$ *and* $k = \bar{y}$, *it holds*

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\leqslant T} \mathbb{1}\{z_t = k\} \middle| \bar{y} \neq q, k = \bar{y}\right] &\approx (2-\alpha)\frac{T}{N}, \\
\mathbb{E}\left[\left(\sum_{t\leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| \bar{y} \neq q, k = \bar{y}\right] &\approx \frac{(2-\alpha)T}{N} + \frac{(2-\alpha)^2 T^2}{N^2}.
\end{aligned} \tag{C.14}$$

Proof. For simplicity, we omit the condition of $\bar{y} \neq q, k = \bar{y}$ in this proof. Denote

$$\begin{aligned}
Y(T) &\triangleq \mathbb{E}\left[\sum_{t\leqslant T} \mathbb{1}\{z_t = k\} \middle| z_0 = q\right], \\
\hat{Y}(T) &\triangleq \mathbb{E}\left[\sum_{t\leqslant T} \mathbb{1}\{z_t = k\} \middle| z_0 \in [N+1], z_0 \neq q\right].
\end{aligned}$$

Then the data generation process implies, $\forall\ T \geqslant 1$,

$$\begin{aligned}
&Y(T) = \hat{Y}(T-1) + p(z_1 = \bar{y} | z_0 = q), \\
&\hat{Y}(T) = p(z_1 = q | z_0 \neq q) \cdot Y(T-1) + p(z_1 \in [N] \setminus \{q\} | z_0 \neq q) \cdot \hat{Y}(T-1) + p(z_1 = \bar{y} | z_0 \neq q).
\end{aligned}$$

The iteration becomes

$$\begin{aligned}
Y(T) &= \hat{Y}(T-1) + (1-\alpha), \\
\hat{Y}(T) &= \frac{1}{N} \cdot Y(T-1) + \frac{N-1}{N} \cdot \hat{Y}(T-1) + \frac{1}{N}.
\end{aligned}$$

This gives

$$\begin{aligned}
Y(T) - \hat{Y}(T) &= -\frac{1}{N}(Y(T-1) - \hat{Y}(T-1)) + (1 - \alpha - \frac{1}{N}), \\
\frac{1}{N}Y(T) + \hat{Y}(T) &= \frac{1}{N}Y(T-1) + \hat{Y}(T-1) + \frac{2-\alpha}{N}.
\end{aligned}$$

Consider the initialization $Y(0) = \hat{Y}(0) = 0$. This implies

$$Y(T) - \hat{Y}(T) = \frac{1 - \alpha - \frac{1}{N}}{1 + \frac{1}{N}} \left(1 - \left(-\frac{1}{N}\right)^T\right),$$
$$\frac{1}{N} Y(T) + \hat{Y}(T) = \frac{2 - \alpha}{N} T.$$

Then we obtain

$$Y(T) \approx (1 - \alpha) + (2 - \alpha)\frac{T}{N},$$
$$\hat{Y}(T) \approx (2 - \alpha)\frac{T}{N}.$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$\mathbb{E}\left[\sum_{t \leqslant T} \mathbb{1}\{z_t = k\} \middle| \bar{y} \neq q, k = q\right] = \hat{Y}(T) \approx (2 - \alpha)\frac{T}{N}.$$

To obtain the expectation of the quadratic term, we similarly denote the following terms with different z_0:

$$Z(T) \triangleq \mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| z_0 = q\right],$$
$$\hat{Z}(T) \triangleq \mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| z_0 \in [N+1], z_0 \neq q\right].$$

Then the data generation process implies, $\forall\, T \geqslant 1$,

$$Z(T) = \hat{Z}(T-1) + p(z_1 = \bar{y} | z_0 = q) \cdot (1 + 2\hat{Y}(T-1)),$$
$$\hat{Z}(T) = p(z_1 = q | z_0 \neq q) \cdot Z(T-1) + p(z_1 \in [N] \setminus \{q\} | z_0 \neq q) \cdot \hat{Z}(T-1)$$
$$+ p(z_1 = \bar{y} | z_0 \neq q) \cdot (1 + 2\hat{Y}(T-1)),$$

where $2\hat{Y}(T-1)$ is due to $\mathbb{E}[(1+\sum_{2\leqslant t\leqslant T}\cdot)^2] = 1 + 2\mathbb{E}[\sum_{2\leqslant t\leqslant T}\cdot] + \mathbb{E}[(\sum_{2\leqslant t\leqslant T}\cdot)^2]$.

Then the iteration becomes

$$
\begin{aligned}
Z(T) &= \hat{Z}(T-1) + (1-\alpha)(1+2\hat{Y}(T-1)),\\
\hat{Z}(T) &= \frac{1}{N}Z(T-1) + \frac{N-1}{N}\hat{Z}(T-1) + \frac{1}{N}(1+2\hat{Y}(T-1)).
\end{aligned}
$$

This gives

$$
\begin{aligned}
Z(T) - \hat{Z}(T) &= -\frac{1}{N}(Z(T-1) - \hat{Z}(T-1)) + (1-\alpha-\frac{1}{N})(1+2\hat{Y}(T-1)),\\
\frac{1}{N}Z(T) + \hat{Z}(T) &= \frac{1}{N}Z(T-1) + \hat{Z}(T-1) + \frac{2-\alpha}{N}(1+2\hat{Y}(T-1)).
\end{aligned}
$$

Considering the initialization $Z(0) = \hat{Z}(0) = 0$, we have

$$
\begin{aligned}
Z(T) - \hat{Z}(T) &= (1-\alpha-\frac{1}{N})\sum_{t\leqslant T-1}(-\frac{1}{N})^{T-1-t}(1+2\hat{Y}(t))\\
&\approx (1-\alpha-\frac{1}{N})\sum_{t\leqslant T-1}(-\frac{1}{N})^{T-1-t}\left(1+\frac{2(2-\alpha)t}{N}\right)\\
&\approx (1-\alpha)\left(1+\frac{2(2-\alpha)T}{N}\right),\\
\frac{1}{N}Z(T) + \hat{Z}(T) &= \frac{(2-\alpha)T}{N} + \frac{2(2-\alpha)}{N}\sum_{1\leqslant t\leqslant T-1}\hat{Y}(t)\\
&\approx \frac{(2-\alpha)T}{N} + \frac{2(2-\alpha)}{N}\sum_{1\leqslant t\leqslant T-1}\frac{(2-\alpha)t}{N}\\
&\approx \frac{(2-\alpha)T}{N} + \frac{(2-\alpha)^2T^2}{N^2}.
\end{aligned}
$$

Then we obtain

$$
\begin{aligned}
Z(T) &\approx \frac{T}{N}(2-\alpha)(3-2\alpha) + \frac{(2-\alpha)^2T^2}{N^2} + (1-\alpha),\\
\hat{Z}(T) &\approx \frac{(2-\alpha)T}{N} + \frac{(2-\alpha)^2T^2}{N^2}.
\end{aligned}
$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$\mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t=k\}\right)^2\middle|\bar{y}=q,k\in[N]\setminus\{q\}\right]=\hat{Z}(T)\approx\frac{(2-\alpha)T}{N}+\frac{(2-\alpha)^2T^2}{N^2}.$$

□

Lemma C.10 ($\bar{y}\neq q,k\leqslant N,k\neq q,k\neq\bar{y}$)**.** *Following the data generation process, assuming $N,T\gg 1$ and $\alpha=\Theta(1)$, if $\bar{y}\neq q$ and $k\in[N]\setminus\{\bar{y},q\}$, it holds*

$$\begin{aligned}\mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t=k\}\middle|\bar{y}\neq q,k\in[N]\setminus\{\bar{y},q\}\right]&\approx\frac{T}{N},\\ \mathbb{E}\left[\left(\sum_{t\leqslant T}\mathbb{1}\{z_t=k\}\right)^2\middle|\bar{y}\neq q,k\in[N]\setminus\{\bar{y},q\}\right]&\approx\frac{T}{N}+\frac{T^2}{N^2}.\end{aligned}\tag{C.15}$$

Proof. For simplicity, we omit the condition of $\bar{y}\neq q,k\in[N]\setminus\{\bar{y},q\}$ in this proof. Denote

$$\begin{aligned}Y(T)&\triangleq\mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t=k\}\middle|z_0=q\right],\\ \hat{Y}(T)&\triangleq\mathbb{E}\left[\sum_{t\leqslant T}\mathbb{1}\{z_t=k\}\middle|z_0\in[N+1],z_0\neq q\right].\end{aligned}$$

Then the data generation process implies, $\forall\ T\geqslant 1$,

$$\begin{aligned}&Y(T)=\hat{Y}(T-1),\\ &\hat{Y}(T)=p(z_1=q|z_0\neq q)\cdot Y(T-1)+p(z_1\in[N]\setminus\{q\}|z_0\neq q)\cdot\hat{Y}(T-1)+p(z_1=k|z_0\neq q).\end{aligned}$$

The iteration becomes

$$
\begin{aligned}
Y(T) &= \hat{Y}(T-1) + (1-\alpha), \\
\hat{Y}(T) &= \frac{1}{N} \cdot Y(T-1) + \frac{N-1}{N} \cdot \hat{Y}(T-1) + \frac{1}{N}.
\end{aligned}
$$

Note that these two equations are exactly the same as those in Lemma C.7 with same initialization as $Y(0) = \hat{Y}(0) = 0$. Therefore, we have

$$
\begin{aligned}
Y(T) &\approx \frac{T}{N}, \\
\hat{Y}(T) &\approx \frac{T}{N}.
\end{aligned}
$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$
\mathbb{E}\left[\sum_{t \leqslant T} \mathbb{1}\{z_t = k\} \middle| \bar{y} \neq q, k = q\right] = \hat{Y}(T) \approx \frac{T}{N}.
$$

To obtain the expectation of the quadratic term, we similarly denote the following terms with different z_0:

$$
\begin{aligned}
Z(T) &\triangleq \mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| z_0 = q\right], \\
\hat{Z}(T) &\triangleq \mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\}\right)^2 \middle| z_0 \in [N+1], z_0 \neq q\right].
\end{aligned}
$$

Then the data generation process implies, $\forall\, T \geqslant 1$,

$$
\begin{aligned}
Z(T) &= \hat{Z}(T-1), \\
\hat{Z}(T) &= p(z_1 = q | z_0 \neq q) \cdot Z(T-1) + p(z_1 \in [N] \setminus \{q\} | z_0 \neq q) \cdot \hat{Z}(T-1) \\
&\quad + p(z_1 = \bar{k} | z_0 \neq q) \cdot (1 + 2\hat{Y}(T-1)),
\end{aligned}
$$

where $2\hat{Y}(T-1)$ is due to $\mathbb{E}[(1 + \sum_{2 \leqslant t \leqslant T} \cdot)^2] = 1 + 2\mathbb{E}[\sum_{2 \leqslant t \leqslant T} \cdot] + \mathbb{E}[(\sum_{2 \leqslant t \leqslant T} \cdot)^2]$.

Then the iteration becomes

$$
\begin{aligned}
Z(T) &= \hat{Z}(T-1), \\
\hat{Z}(T) &= \frac{1}{N} Z(T-1) + \frac{N-1}{N} \hat{Z}(T-1) + \frac{1}{N}(1 + 2\hat{Y}(T-1)).
\end{aligned}
$$

Again note that, since $Y(T) \approx \hat{Y}(T)$, these two equations are the same as those in Lemma C.7. Therefore, we have

$$
\begin{aligned}
Z(T) &\approx \frac{T}{N} + \frac{T^2}{N^2}, \\
\hat{Z}(T) &\approx \frac{T}{N} + \frac{T^2}{N^2}.
\end{aligned}
$$

Since the data generation process implicitly assumes $z_0 \neq q$, we have the desired expectation as

$$
\mathbb{E}\left[\left(\sum_{t \leqslant T} \mathbb{1}\{z_t = k\} \right)^2 \middle| \bar{y} = q, k \in [N] \setminus \{q\} \right] = \hat{Z}(T) \approx \frac{T}{N} + \frac{T^2}{N^2}.
$$

□

C.4 Proof of Theorem 4.2: Training Dynamics of the Attention Layer

We consider the following simplified 1-layer model for the noisy in-context recall task.

$$\begin{aligned} x_t &\triangleq \mathbf{W}_E(z_t) + \widetilde{\mathbf{W}}_E(z_{t-1}) \in \mathbb{R}^d, \\ \phi(x_T, x_{1:T}) &\triangleq \sum_{t \leqslant T} \left[\sigma\left(x_T^\top \mathbf{W}_{QK} x_{1:T}\right)\right]_t \cdot \mathbf{W}_V x_t \in \mathbb{R}^d, \\ \xi_{\text{attn}}(x_{1:T}) &\triangleq \mathbf{W}_U \phi(x_T, x_{1:T}) \in \mathbb{R}^{N+1}, \\ \xi_{\text{ff}}(x_{1:T}) &\triangleq \mathbf{W}_U F(x_T) = \mathbf{W}_U \mathbf{W}_F x_T \in \mathbb{R}^{N+1}, \end{aligned} \tag{C.16}$$

With zero initialization of $\mathbf{W}_{QK}, \mathbf{W}_V, \mathbf{W}_F$, we analyze the training dynamics of these three matrices in three phases:

1. $\mathbf{W}_F$ learns the noise association in $O(\frac{1}{\eta})$ time,
2. $\mathbf{W}_V$ learns to be identity for all tokens $k \in [N+1]$,
3. $\mathbf{W}_{QK}$ attends to any position t such that $z_{t-1} = q$ and $z_t = \bar{y}$.

Assumption C.4.1. *In this section, we make the following assumptions*

1. *(orthonormal embedding)* $\mathbf{W}_E(i)^\top \mathbf{W}_E(j) = \widetilde{\mathbf{W}}_E(i)^\top \widetilde{\mathbf{W}}_E(j) = \mathbb{1}\{i = j\}$ *and* $\mathbf{W}_E(i)^\top \widetilde{\mathbf{W}}_E(j) = 0$ *for any* $i, j \in [N+1]$.
2. *(Feed-forward learns noise association) After phase 1, the prediction for noise always satisfies* $\hat{p}(N+1|z_{1:T}) = \alpha$ *for any* $z_{1:T} \in [N+1]^{\otimes T}$. *If* $\hat{p}$ *deviates from* α, $\mathbf{W}_F$ *will learn the noise association in a more quick speed than the other weights, so that it is fair to assume* $\hat{p} = \alpha$ *for computing gradients of these weights.*
3. *(Infinite samples)* $m \to \infty$ *so the training loss* L *is population loss.*

4. $\alpha \leqslant 1.5 - \sqrt{5}/2 \approx 0.38$. *This is to ensure the sign* $\mathbf{W}_U(j)^\top(-\nabla_{\mathbf{W}_V}L)\mathbf{W}_E(k) > 0$ *for any* $j = k \leqslant N$ *in (C.18).*

Phase 1: In this phase, the impact of $\widetilde{\mathbf{W}}_E(z_{T-1})$ on $\mathbf{W}_F$ and $\mathbf{W}_V$ is negligible compared with that of $\mathbf{W}_E(z_T)$ because Z_{T-1} is close to uniform in $[N+1]$ while $z_T = q$ is fixed.

Lemma C.1 gives

$$\mathbf{W}_U(k)^\top(-\nabla_{\mathbf{W}_F}L)\mathbf{W}_E(q) = \begin{cases} \Theta(1), & \text{if } k = N+1, \\ \Theta(\frac{1}{N}), & \text{if } k \leqslant N. \end{cases}$$

Lemma C.2 gives

$$\mathbf{W}_U(j)^\top(-\nabla_{\mathbf{W}_V}L)\mathbf{W}_E(k) = \begin{cases} \Theta(\frac{1}{N}), & \text{if } j = N+1, \forall\, k, \\ \Theta(\frac{1}{N^2}), & \text{if } j \leqslant N, \forall\, k. \end{cases} \tag{C.17}$$

Note that the entries of the above projection have the following signs, with details as $-\mu$ in Table C.1,

$$\mathbf{W}_U(j)^\top(-\nabla_{\mathbf{W}_V}L)\mathbf{W}_E(k) \begin{cases} > 0, & \text{if } (j = N+1) \text{ or } (j = k) \text{ or } (j = q, k = N+1), \\ < 0, & \text{otherwise.} \end{cases} \tag{C.18}$$

The arguments in Appendix 4.4.4.3 show

$$\mathbf{W}_E(j)^\top(-\nabla_{\mathbf{W}_{QK}}L)\mathbf{W}_E(q) = \begin{cases} -\Theta(\frac{1}{N^2}), & \text{if } j = N+1, \\ \Theta(\frac{1}{N^3}), & \text{if } j \leqslant N. \end{cases} \tag{C.19}$$

Therefore, during this phase, $\mathbf{W}_F$ learns the noise association with effective graident norm of $\Theta(1)$ as $\mathbf{W}_U(N+1)^\top(-\nabla_{\mathbf{W}_F}L)\mathbf{W}_E(q) = \Theta(1)$. Meanwhile, $\mathbf{W}_F$ moves in the other directions

uniformly in $\Theta(\frac{1}{N})$ as $\mathbf{W}_U(k)^\top(-\nabla_{\mathbf{W}_F}L)\mathbf{W}_E(q) = \Theta(\frac{1}{N})$ for any $k \leqslant N$, which in fact ensures $\hat{p}(k|z_{1:T}) = \frac{1-\hat{p}(N+1|z_{1:T})}{N}$ for any $k \leqslant N$ and $z_{1:T} \in [N+1]^{\otimes T}$.

After $O(\eta^{-1})$ steps in this phase, we have $\hat{p}(N+1|z_{1:T}) = \alpha$ and $\hat{p}(k|z_{1:T}) = \frac{1-\alpha}{N}$ for any $k \leqslant N$ and $z_{1:T}$.

Phase 2: Assume $\hat{p}(N+1|\cdot) = \alpha$ starting from the beginning of this phase as discussed above. Due to symmetry for the rest k channels, we have $\hat{p}(k|\cdot) = \frac{1-\alpha}{N}$. Note that the attention scores in $\phi(\cdot,\cdot)$ are still close to uniform, *i.e.*, $\left[\sigma\left(x_T^\top \mathbf{W}_{QK} x_{1:T}\right)\right]_t \approx \frac{1}{T}$, since the update of $\mathbf{W}_{QK}$ is in $O(N^{-2})$ whose impact on attention scores is also in $O(N^{-2})$ through $\exp(x) \approx 1 + x$ for $x \approx 0$. Then we track the movement of $\mathbf{W}_V$ under these conditions.

Since $m \to \infty$, taking $\bar{x} \triangleq \frac{1}{T}\sum_{i=1}^T x_i$, $\mu_k \triangleq \mathbb{E}[\bar{x}|y=k]$ and $\hat{\mu}_k \triangleq \mathbb{E}[\frac{\hat{p}(k|x)}{p(y=k)}\bar{x}] = \mathbb{E}[\bar{x}]$ since $\hat{p}(k|x) = \alpha\mathbb{1}\{k = N+1\} + \frac{1-\alpha}{N}\mathbb{1}\{k \leqslant N\} = p(y|k)$, Lemma C.12 gives

$$\begin{aligned}
\nabla_{\mathbf{W}_V} L &= \sum_{k=1}^{N+1} p(y=k)\mathbf{W}_U(k)(\mathbb{E}[\bar{x}] - \mathbb{E}[\bar{x}|y=k])^\top \\
&= \sum_{k=1}^{N} p(y=k)\mathbf{W}_U(k)(\mathbb{E}[\bar{x}] - \mathbb{E}[\bar{x}|y=k])^\top \\
&= \sum_{k=1}^{N} \frac{1-\alpha}{N}\mathbf{W}_U(k)(\mathbb{E}[\bar{x}] - \mathbb{E}[\bar{x}|y=k])^\top \\
&= -\frac{1-\alpha}{N^2}\sum_{k=1}^{N} \mathbf{W}_U(k)(\mathbf{W}_E(k) - \overline{\mathbf{W}}_E + \widetilde{\mathbf{W}}_E(k) - \overline{\widetilde{\mathbf{W}}}_E)^\top,
\end{aligned}$$

where the second equality is due to $\mathbb{E}[\bar{x}] = \mathbb{E}[\bar{x}|y = N+1]$ due to $y = N+1$ is uniform for any correct token $\bar{y} \leqslant N$, and the last equality is from

$$\mathbb{E}[\bar{x}] - \mathbb{E}[\bar{x}|y=k] \approx -\frac{1}{N}(\mathbf{W}_E(k) - \overline{\mathbf{W}}_E) - \frac{1}{N}(\widetilde{\mathbf{W}}_E(k) - \overline{\widetilde{\mathbf{W}}}_E)$$

with $\overline{\mathbf{W}}_E = N^{-1}\sum_{i=1}^N \mathbf{W}_E(i)$, $\overline{\widetilde{\mathbf{W}}}_E = N^{-1}\sum_{i=1}^N \widetilde{\mathbf{W}}_E(i)$ because $\mathbb{E}[\bar{x}] = \mathbb{E}_y[\mathbb{E}_x[\bar{x}|y]]$, and the expected number of the tuple $(q, \hat{y})$ in a context length T is $\Theta(\frac{T}{N})$ by comparing Lemma C.9 and C.10.

Therefore, the gradient for $\mathbf{W}_V$ has the following structure

$$\begin{aligned}\mathbf{W}_U(j)^\top(-\nabla_{\mathbf{W}_V}L)\mathbf{W}_E(k) &\approx \frac{1}{N^2}\mathbb{1}\{j=k\}+O\left(\frac{1}{N^3}\right),\forall\, j,k\leqslant N,\\ \mathbf{W}_U(j)^\top(-\nabla_{\mathbf{W}_V}L)\widetilde{\mathbf{W}}_E(k) &\approx \frac{1}{N^2}\mathbb{1}\{j=k\}+O\left(\frac{1}{N^3}\right),\forall\, j,k\leqslant N.\end{aligned} \tag{C.20}$$

Denote steps of phase 1 and phase 2 as t_1 and t_2. Combined with the structure of $\mathbf{W}_V$ in phase 1 as in Eq.(C.17,C.18), ignoring projections that are $O(N^{-3})$ or negative, $\mathbf{W}_V$ has the following structure after phase 2

$$\begin{aligned}\mathbf{W}_U(j)^\top\mathbf{W}_V\mathbf{W}_E(k) &= \begin{cases}\Theta(\eta t_1 N^{-1}), & \text{if } j=N+1,\forall k,\\ \Theta(\eta t_1 N^{-2}+\eta t_2 N^{-2}), & \text{if } j=k\leqslant N,\\ \Theta(\eta t_1 N^{-2}), & \text{if } j=q, k=N+1,\end{cases}\\ \mathbf{W}_U(j)^\top\mathbf{W}_V\widetilde{\mathbf{W}}_E(k) &= \Theta(\eta t_2 N^{-2}), \text{if } j=k\leqslant N.\end{aligned} \tag{C.21}$$

Phase 3: now assume $\mathbf{W}_V$ has the structure in Eq(C.21). The model still predicts $\hat{p}_{\mathbf{W}}(k|z)=\alpha\mathbb{1}\{k=N+1\}+\frac{1-\alpha}{N}\mathbb{1}\{k\leqslant N\}$ because the above projections of $\mathbf{W}_V$ onto $\mathbf{W}_U(j:j\leqslant N)$ is $o(\frac{1}{N})$. Meanwhile, the attention scores are uniform as $\frac{1}{T}$ as $\mathbf{W}_{QK}\approx 0$. Therefore, the gradient of $\mathbf{W}_{QK}$ is

$$\begin{aligned}\nabla_{\mathbf{W}_{QK}}L &= \frac{1}{T}\sum_{k=1}^{N+1}\sum_{t\leqslant T}p(y=k)(\mathbb{E}[(\mathbf{W}_U(k)^\top\mathbf{W}_V x_t)\cdot x_T(x_t-\bar{x})^\top]\\ &\qquad -\mathbb{E}[(\mathbf{W}_U(k)^\top\mathbf{W}_V x_t)\cdot x_T(x_t-\bar{x})^\top|y=k])\\ &= \frac{1-\alpha}{TN}\sum_{k=1}^{N}\sum_{t\leqslant T}(\mathbb{E}[(\mathbf{W}_U(k)^\top\mathbf{W}_V x_t)\cdot x_T(x_t-\bar{x})^\top]\\ &\qquad -\mathbb{E}[(\mathbf{W}_U(k)^\top\mathbf{W}_V x_t)\cdot x_T(x_t-\bar{x})^\top|y=k]),\end{aligned}$$

where $\bar{x}=T^{-1}\sum_{t\leqslant T}x_t$ and the last equality holds due to the condition of $y=N+1$ uniform for any correct token $\hat{y}\leqslant N$. Then, considering the above structure of $\mathbf{W}_V$, we notice that

$\mathbf{W}_U(j)^\top \mathbf{W}_V x_t \approx \beta_1 \mathbb{1}\{z_t = j\} + \beta_2 \mathbb{1}\{z_{t-1} = j\}$ with $\beta_1 = \eta t_1 N^{-2} + \eta t_2 N^{-2}$ and $\beta_2 = \eta t_2 N^{-2}$ for any $j, k \leqslant N$. Here note that we ignore the projection of $j = q, k = N+1$ in Eq(C.21) because $\hat{y} = q$ is with probability $1/N = o(1)$ so that it will not influence much the following derivation.

Plug-in $\mathbf{W}_U(j)^\top \mathbf{W}_V x_t$ and we get

$$\mathbf{W}_E(q)^\top (-\nabla_{\mathbf{W}_{QK}} L)(\mathbf{W}_E(b_1) + \widetilde{\mathbf{W}}_E(b_2)) = \frac{1-\alpha}{TN} \sum_{k \leqslant N} \sum_{t \leqslant T} \mathbb{E}[A^{(t)}_{k,b_1,b_2} | y = k] - \mathbb{E}[A^{(t)}_{k,b_1,b_2}] \tag{C.22}$$

where

$$\begin{aligned} A^{(t)}_{k,b_1,b_2} = {} & (\beta_1 \mathbb{1}\{z_t = k\} + \beta_2 \mathbb{1}\{z_{t-1} = k\}) \\ & \cdot \left(\mathbb{1}\{z_t = b_1\} - \frac{\sum_{s \leqslant T} \mathbb{1}\{z_s = b_1\}}{T} + \mathbb{1}\{z_{t-1} = b_2\} - \frac{\sum_{s \leqslant T} \mathbb{1}\{z_{s-1} = b_2\}}{T} \right). \end{aligned}$$

Now we are to control $\Delta_{k,b_1,b_2} \triangleq \sum_{t \leqslant T} \mathbb{E}[A^{(t)}_{k,b_1,b_2} | y = k] - \mathbb{E}[A^{(t)}_{k,b_1,b_2}]$ for different choices of b_1, b_2. Note that b_1 and b_2 co-exist by sum in $A^{(t)}_{k,b_1,b_2}$, so the additivity of expectation allows us to discuss choices of b_1, b_2 separately and then combine the results. Denote

$$\begin{aligned} B^{(t)}_{k,b_1} &= (\beta_1 \mathbb{1}\{z_t = k\} + \beta_2 \mathbb{1}\{z_{t-1} = k\}) \left(\mathbb{1}\{z_t = b_1\} - \frac{\sum_{s \leqslant T} \mathbb{1}\{z_s = b_1\}}{T} \right), \\ C^{(t)}_{k,b_2} &= (\beta_1 \mathbb{1}\{z_t = k\} + \beta_2 \mathbb{1}\{z_{t-1} = k\}) \left(\mathbb{1}\{z_{t-1} = b_2\} - \frac{\sum_{s \leqslant T} \mathbb{1}\{z_{s-1} = b_2\}}{T} \right). \end{aligned} \tag{C.23}$$

Controlling $\sum_{t \leqslant T} \mathbb{E}[B^{(t)}_{k,b_1} | y = k] - \mathbb{E}[B^{(t)}_{k,b_1}]$:

- If $b_1 = k$, from Lemma C.9 and C.10, we have

$$\mathbb{E}\left[\sum_{t \leqslant T} \beta_1 \mathbb{1}\{z_t = k\} \mathbb{1}\{z_t = k\} \middle| y = k \right] - \mathbb{E}\left[\sum_{t \leqslant T} \beta_1 \mathbb{1}\{z_t = k\} \mathbb{1}\{z_t = k\} \right] = \beta_1 (1-\alpha) \frac{T}{N}.$$

$$\mathbb{E}\left[-\sum_{t\leqslant T}\beta_1\mathbb{1}\{z_t=k\}\frac{\sum_{s\leqslant T}\mathbb{1}\{z_s=k\}}{T}\middle|y=k\right]-\mathbb{E}\left[-\sum_{t\leqslant T}\beta_1\mathbb{1}\{z_t=k\}\frac{\sum_{s\leqslant T}\mathbb{1}\{z_s=k\}}{T}\right]$$
$$=-\mathbb{E}\left[\beta_1T^{-1}(\sum_{s\leqslant T}\mathbb{1}\{z_s=k\})^2|y=k\right]+\mathbb{E}\left[\beta_1T^{-1}(\sum_{s\leqslant T}\mathbb{1}\{z_s=k\})^2\right]$$
$$=\beta_1T^{-1}\left(\frac{T}{N}+\frac{T^2}{N^2}-\frac{(2-\alpha)T}{N}-\frac{(2-\alpha)^2T^2}{N^2}\right)=o\left(\beta_1\frac{T}{N}\right).$$

The terms involving $\mathbb{1}\{z_{t-1}=k\}$ are negligible as $O(T/N^2)$. Therefore, we have

$$\sum_{t\leqslant T}\mathbb{E}[B_{k,k}^{(t)}|y=k]-\mathbb{E}[B_{k,k}^{(t)}]=\beta_1(1-\alpha)\frac{T}{N}. \tag{C.24}$$

- If $b_1\neq k$, all terms are $O(T/N^2)$ because
  - If $b_1\leqslant N$, it holds $p(z_t=b_1|z_{t-1}=k)=1/N$ with the expected number of k in context of length L being $\Theta(T/N)$ from lemmas in Appendix C.3.
  - If $b_1=N+1$, it holds $p(z_t=N+1|z_{t-1}=k)=O(1/N)\cdot\mathbb{1}\{k=q\}$ and the expected number of q in context of length T is $\Theta(T/N)$ from Lemma C.4 and C.7.
  - $\mathbb{E}[\sum_t\mathbb{1}\{z_{t-1}=k\}\#b_1/T|\cdot]=\mathbb{E}[\#k\cdot\#b_1/T]=O(T/N^2)$ no matter it is with condition $y=k$ or not.

Therefore, for any $b_1\neq k$, we have

$$\sum_{t\leqslant T}\mathbb{E}[B_{k,b_1}^{(t)}|y=k]-\mathbb{E}[B_{k,b_1}^{(t)}]=o(T/N). \tag{C.25}$$

Controlling $\sum_{t\leqslant T}\mathbb{E}[C_{k,b_2}^{(t)}|y=k]-\mathbb{E}[C_{k,b_2}^{(t)}]$:

- If $b_2 = q$, Lemma C.7 gives

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t\leqslant T}\beta_1\mathbb{1}\{z_t=k\}\left(\mathbb{1}\{z_{t-1}=q\}-\frac{\#q}{T}\right)\middle|y=k\right]\\
&\quad-\mathbb{E}\left[\sum_{t\leqslant T}\beta_1\mathbb{1}\{z_t=k\}\left(\mathbb{1}\{z_{t-1}=q\}-\frac{\#q}{T}\right)\right]\\
=\;&(1-p(\hat{y}=k))\cdot\mathbb{E}\left[\sum_{t\leqslant T}\beta_1\mathbb{1}\{z_t=k\}\left(\mathbb{1}\{z_{t-1}=q\}-\frac{\#q}{T}\right)\middle|y=k\right]+o\left(\beta_1\frac{T}{N}\right)\\
\approx\;&\beta_1(1-\alpha)\frac{T}{N},
\end{aligned}$$

  where the last equality is from $p(z_t = k|\bar{y} = k, z_{t-1} = q) = 1 - \alpha$.

  All the other terms are negligible with the same reason as above.

  Therefore, we have

$$\sum_{t\leqslant T}\mathbb{E}[C_{k,q}^{(t)}|y=k]-\mathbb{E}[C_{k,q}^{(t)}]=\beta_1(1-\alpha)\frac{T}{N}. \tag{C.26}$$

- If $b_2 = k$, similar to the above discussion about $B_{k,k}$, we have

$$\sum_{t\leqslant T}\mathbb{E}[C_{k,k}^{(t)}|y=k]-\mathbb{E}[C_{k,k}^{(t)}]=\beta_2(1-\alpha)\frac{T}{N}. \tag{C.27}$$

  Note that the key difference is that here we use β_2 instead of β_1, and $\beta_2 < \beta_1$.

- If $b_2 \neq q$ and $b_2 \neq k$, similar to the discussion for Eq(C.25), we have

$$\sum_{t\leqslant T}\mathbb{E}[C_{k,b_2}^{(t)}|y=k]-\mathbb{E}[C_{k,b_2}^{(t)}]=o(T/N). \tag{C.28}$$

Therefore, combining the above results in Eq(C.24, C.25, C.26, C.27, C.28), taking sums of the

corresponding B and C from Eq(C.23) gives

$$\Delta_{k,b_1,b_2} = \begin{cases} \beta_1(1-\alpha)TN^{-1} + \beta_1(1-\alpha)TN^{-1}, & \text{if } b_1 = k, b_2 = q, \\ \beta_1(1-\alpha)TN^{-1} + \beta_2(1-\alpha)TN^{-1}, & \text{if } b_1 = k, b_2 = k, \\ \beta_1(1-\alpha)TN^{-1}, & \text{if } b_1 = k, \text{other } b_2, \\ \beta_1(1-\alpha)TN^{-1}, & \text{if } b_1 \neq k, b_2 = q, \\ \beta_2(1-\alpha)TN^{-1}, & \text{if } b_1 \neq k, b_2 = k, \\ O(TN^{-1}), & \text{otherwise.} \end{cases}$$

To take the summation over all $k \leqslant N$ in Eq(C.22), we discuss the following cases of b_1 and b_2 for $\mathbf{W}_E(q)^\top(-\nabla_{\mathbf{W}_{QK}}L)(\mathbf{W}_E(b_1) + \widetilde{\mathbf{W}}_E(b_2))$.

- If $b_1 \leqslant N, b_1 \neq b_2, b_2 = q$:
  - when $k = b_1$, we take Δ_{k,b_1,b_2} under the condition of $b_1 = k, b_2 = q$.
  - when $k \neq b_1$, we take Δ_{k,b_1,b_2} under the condition of $b_1 \neq k, b_2 = q$. Note that there are $(N-1)$ such k.

  Therefore, it holds

$$\mathbf{W}_E(q)^\top(-\nabla_{\mathbf{W}_{QK}}L)(\mathbf{W}_E(b_1) + \widetilde{\mathbf{W}}_E(q)) = \frac{1-\alpha}{TN}\beta_1(1-\alpha)T(1+N^{-1}). \tag{C.29}$$

- If $b_1 = b_2 = q$:
  - when $k = b_1$, we take Δ_{k,b_1,b_2} under the condition of $b_1 = k, b_2 = k$ to achieve a lower bound of the gap later.
  - when $k \neq b_1$, we take Δ_{k,b_1,b_2} under the condition of $b_1 \neq k, b_2 = q$. Note that there are $(N-1)$ such k.

Therefore, it holds

$$\mathbf{W}_E(q)^\top(-\nabla_{\mathbf{W}_{QK}}L)(\mathbf{W}_E(b_1)+\widetilde{\mathbf{W}}_E(q)) \geqslant \frac{1-\alpha}{TN}\left(\beta_1(1-\alpha)T+\beta_2(1-\alpha)TN^{-1}\right). \quad \text{(C.30)}$$

- If $b_1 = N+1, b_2 = q$: for any $k \leqslant N$, it holds $k \neq b_1$, so we take Δ_{k,b_1,b_2} under the condition of $b_1 \neq k, b_2 = q$. Therefore, it holds

$$\mathbf{W}_E(q)^\top(-\nabla_{\mathbf{W}_{QK}}L)(\mathbf{W}_E(N+1)+\widetilde{\mathbf{W}}_E(q)) = \frac{1-\alpha}{TN}\beta_1(1-\alpha)T. \quad \text{(C.31)}$$

- If $b_2 \neq q, \forall\, b_1$: To get an upper bound of the projection length, we take Δ_{k,b_1,b_2} under the condition of $b_=k, b_2 = k$ or $b_1 \neq k, b_2 = k$. Therefore, it holds

$$\mathbf{W}_E(q)^\top(-\nabla_{\mathbf{W}_{QK}}L)(\mathbf{W}_E(b_1)+\widetilde{\mathbf{W}}_E(b_2)) \leqslant \frac{1-\alpha}{TN}(\beta_1+2\beta_2)(1-\alpha)TN^{-1}. \quad \text{(C.32)}$$

Comparing the above four cases, for any $\bar{y} \leqslant N$, the attention weight $\mathbf{W}_{QK}$ to attend more to $x_t = \mathbf{W}_E(\bar{y})+\widetilde{\mathbf{W}}_E(q)$ than to $x_t = \mathbf{W}_E(N+1)+\widetilde{\mathbf{W}}_E(q)$, with

$$\mathbf{W}_E(q)^\top(-\nabla_{\mathbf{W}_{QK}}L)(\mathbf{W}_E(\bar{y})+\widetilde{\mathbf{W}}_E(q)) - \mathbf{W}_E(q)^\top(-\nabla_{\mathbf{W}_{QK}}L)(\mathbf{W}_E(N+1)+\widetilde{\mathbf{W}}_E(q)) \geqslant \frac{(1-\alpha)^2}{N^2}\beta_2.$$

Meanwhile, any other setting of b_1, b_2 has smaller projection in $(-\nabla_{\mathbf{W}_{QK}}L)$.

In summary, $\mathbf{W}_{QK}$ has the following patterns

1. it learns to attend to indices t such that $z_{t-1} = q$ is the trigger word,
2. when there are multiple t_i's such that $z_{t_i-1=q}$, it learns to attend to those with $z_t = \bar{y}$ more than $z_t = N+1$.

C.5 Experiments Setup: Linear Associative Memory

In Section 4.4, we showed that *fully* truncating a feed-forward layer can be helpful for reasoning. We now present a setting where noisy associations are stored in a rank-one subspace of a layer, so that *intermediate* levels of truncation are more useful to remove noise.

Model and data. We consider a simple associative memory setting where the goal is learn an fixed permutation from input tokens to output tokens (w.l.o.g. taken to be the identity), with a linear model similar to Cabannes et al. [2024]. Consider a learnable weight matrix $\mathbf{W} \in \mathbb{R}^{d\times d}$. Consider embeddings for n input tokens as $\{e_i\}_{i=1}^{n} \subset \mathbb{R}^d$ and embeddings for c output tokens as $\{u_i\}_{i=1}^{c} \subset \mathbb{R}^d$. In contrast to Cabannes et al. [2024], we consider an additional "common noise" output token $c = n + 1$, which is chosen for any input with probability $\alpha \in (0, 1)$. For any input $x \in [n]$, the target distribution $p_\alpha(\cdot|x)$ is defined by

$$p_\alpha(y|x) = (1 - \alpha) \cdot \mathbb{1}\{y = x\} + \alpha \cdot \mathbb{1}\{y = c\}. \tag{C.33}$$

In other words, the last channel (c) for output is the **common noise** with probability α for any input. The training dataset $\mathcal{D}_\alpha$ consists of uniformly distributed inputs $x \in [n]$, and outputs conditionally sampled as $y|x \sim p_\alpha(\cdot|x)$.

Given any pair of input and output tokens, the associative memory model takes the form

$$f(i, j; \mathbf{W}) \triangleq \langle u_j, \mathbf{W}e_i\rangle, \quad \forall\, i, j \in [n] \times [c], \tag{C.34}$$

When $k \leqslant d$, we denote the rank-k approximation of f as $f^{(k)}$ by replacing $\mathbf{W}$ with $\mathbf{W}^{(k)}$, where $\mathbf{W}^{(k)}$ is the rank-k approximation of $\mathbf{W}$.

Training. During training, the dataset $\mathcal{D}_\alpha$ is generated with non-zero noise probability $\alpha > 0$. At test time, the dataset $\mathcal{D}_0$ is without noise as $\alpha = 0$, so the computed loss is called **pure-label**

loss. The model is trained with Gradient Descent (GD) subjected to cross-entropy loss.

C.5.1 Proof of Theorem 4.3

Now we present a theoretical analysis of this problem with some assumptions.

Assumption C.5.1 (Orthonormality)**.** *Embeddings of input and output tokens are orthonormal, i.e.,* $e_i^\top e_j = \mathbb{1}\{i = j\}, \forall\, i, j$ *and* $u_i^\top u_j = \mathbb{1}\{i = j\}, \forall\, i, j$.

Assumption C.5.2 (Initialization)**.** *The learnable matrix* $\mathbf{W}$ *is initialized from* $\mathbf{0}$ *when* $t = 0$.

Theorem C.11 (Restatement of Theorem 4.3)**.** *Assume Assumptions C.5.1 and C.5.2 hold, considering* $n = 2, c = 3$ *and* $\alpha \in (0.2, 0.4)$*, we train the full model* $f(\cdot, \cdot; \mathbf{W})$ *with gradient flow. Denote* $P(i, j; \mathbf{W})$ *as the model's predicted probability for output* j *conditioned on input* i*. Then, for* $t \to \infty$ *and* $i \in \{1, 2\}$*, we have*

$$P(i, j; \mathbf{W}) = (1 - \alpha) \cdot \mathbb{1}\{j = i\} + \alpha \cdot \mathbb{1}\{j = c\},$$
$$P(i, j; \mathbf{W}^{(1)}) = (1 - \Theta(t^{-1/2})) \cdot \mathbb{1}\{j = i\} + \Theta(t^{-1/2}) \cdot \mathbb{1}\{j = c\}.$$

Remark 8. Note that here the assumption $\alpha \in (0.2, 0.4)$ is a technical choice. In experiments, any value $\alpha \in (0, 0.4)$ still has the same result.

Proof. W.l.o.g., we assume the embeddings are standard basis in $\mathbb{R}^d$. For any $\mathbf{W}$, the gradient $\nabla_{\mathbf{W}} L$ can be decomposed as

$$\nabla_{\mathbf{W}} L = \gamma_1 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} + \gamma_2 \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}. \tag{C.35}$$

Since $\mathbf{W}$ initializes from zero, this implies $\mathbf{W}$ can always be decomposed with the same basis

$$\mathbf{W} = \beta_1 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} + \beta_2 \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}. \tag{C.36}$$

Then gradient flow gives the following ODE

$$\begin{aligned} \dot{\beta}_1 = -\gamma_1 &= \frac{\exp(-\beta_1 + \beta_2) - \exp(\beta_1 + \beta_2)}{\exp(-\beta_1 + \beta_2) + \exp(\beta_1 + \beta_2) + \exp(-2\beta_2)} + 1 - \alpha \\ &= \frac{\exp(-2\beta_1) - 1}{\exp(-2\beta_1) + \exp(-\beta_1 - 3\beta_2) + 1} + 1 - \alpha, \\ \dot{\beta}_2 = -\gamma_2 &= \frac{3\exp(-2\beta_2)}{\exp(-\beta_1 + \beta_2) + \exp(\beta_1 + \beta_2) + \exp(-2\beta_2)} - 3\alpha \\ &= \frac{3\exp(-\beta_1 - 3\beta_2)}{\exp(-2\beta_1) + \exp(-\beta_1 - 3\beta_2) + 1} - 3\alpha. \end{aligned} \tag{C.37}$$

Denoting $a = -2\beta_1, b = -\beta_1 - 3\beta_2$, the ODE becomes

$$\begin{aligned} \dot{a} &= \frac{2 - 2\exp(a)}{\exp(a) + \exp(b) + 1} - 2 + 2\alpha, \\ \dot{b} &= \frac{2 - 8\exp(b)}{\exp(a) + \exp(b) + 1} - 2 + 10\alpha. \end{aligned} \tag{C.38}$$

Lemma C.14 gives the solution as, when $t \to \infty$,

$$a \to -\log(t) - \log(1-\alpha)(4-2\alpha), \quad b \to \log\frac{\alpha}{1-\alpha}.$$

For the full model, taking the scores $\mathbf{W}_{1,:}$ of the first input token as an example, we have $\mathbf{W}_{11} = \beta_1 + \beta_2, \mathbf{W}_{12} = -\beta_1 + \beta_2, \mathbf{W}_{13} = -2\beta_2$, so the margins are

$$\mathbf{W}_{11} - \mathbf{W}_{12} = 2\beta_1 = -a, \mathbf{W}_{11} - \mathbf{W}_{13} = \beta_1 + 3\beta_2 = -b.$$

For the rank-1 model (assuming $\beta_1 > \beta_2$), the margins are

$$\mathbf{W}_{11}^{(1)} - \mathbf{W}_{12}^{(1)} = 2\beta_1, \mathbf{W}_{11}^{(1)} - \mathbf{W}_{13}^{(1)} = \beta_1.$$

The proof finishes by computing softmax on the margins. □

C.6 Useful Lemmas

Lemma C.12. *Let p be a data distribution on $(x, y) \in \mathbb{R}^d \times [N]$. Consider training data as m i.i.d. samples $\mathcal{D} \triangleq \{(x_i, y_i)\}_{i=1}^m \subset \mathbb{R}^d \times [N+1]$ from p. Consider the following classification problem, with fixed output embeddings $\mathbf{W}_U$:*

$$\hat{L}(\mathbf{W}) = \frac{1}{m} \sum_{i=1}^{m} [l(y_i, \mathbf{W}_U \mathbf{W} x_i)].$$

The gradients take the following form: denoting $\hat{p}_{\mathbf{W}}(k|x_i)$ as the current predicted probability of class k in $[N+1]$ classes for input x_i,

$$\nabla_{\mathbf{W}} \hat{L}(\mathbf{W}) = \frac{1}{m} \sum_{i=1}^{m} \left[\sum_{k=1}^{N+1} (\hat{p}_{\mathbf{W}}(k|x_i) - \mathbb{1}\{y_i = k\}) \mathbf{W}_U(k) x_i^\top \right].$$

When $m \to \infty$, the above equation becomes

$$\nabla_{\mathbf{W}} L(\mathbf{W}) = \sum_{k=1}^{N+1} p(y = k) \mathbf{W}_U(k) (\hat{\mu}_k - \mu_k)^\top,$$

where $\mu_k \triangleq \mathbb{E}[x|y = k]$ and $\hat{\mu}_k \triangleq \mathbb{E}_x[\frac{\hat{p}_{\mathbf{W}}(k|x)}{p(y=k)} x]$.

Remark 9. This lemma is from Lemma 2 in [Bietti et al. 2023].

Proof. Recall the form of the cross-entropy loss for classification with K classes:

$$l(y, \epsilon) = -\sum_{k=1}^{K} \mathbb{1}\{y = k\} \log \frac{e^{\xi_k}}{\sum_j e^{\xi_j}}.$$

Its derivatives take the form

$$\frac{\partial l}{\partial \xi_k}(y, \xi) = s(\xi)_k - \mathbb{1}\{y = k\},$$

where $s(\xi)_k = \frac{e^{\xi_k}}{\sum_j e^{\xi_j}}$.

The gradient of L is then given by

$$\begin{aligned}
\nabla_{\mathbf{W}} \hat{L}(\mathbf{W}) &= \frac{1}{m} \sum_{i=1}^{m} \left[\sum_{k=1}^{N+1} \frac{\partial l}{\partial \xi_k}(y_i, \mathbf{W}_U \mathbf{W} x_i) \nabla_{\mathbf{W}} (\mathbf{W}_U(k)^\top \mathbf{W} x_i) \right] \\
&= \frac{1}{m} \sum_{i=1}^{m} \left[\sum_{k=1}^{N+1} (\hat{p}_{\mathbf{W}}(k|x_i) - \mathbb{1}\{y_i = k\}) \mathbf{W}_U(k) x_i^\top \right].
\end{aligned}$$

When $m \to \infty$, the above equation becomes

$$\begin{aligned}
\nabla_{\mathbf{W}} L(\mathbf{W}) &= \sum_{k=1}^{N+1} \mathbf{W}_U(k) \mathbb{E}[\hat{p}_{\mathbf{W}}(k|x) x^\top] - \sum_{k=1}^{N+1} \mathbb{E}[\mathbb{1}\{y = k\} \mathbf{W}_U(k) \mathbb{E}[x|y]^\top] \\
&= \sum_{k=1}^{N+1} \mathbf{W}_U(k) \mathbb{E}[\hat{p}_{\mathbf{W}}(k|x) x^\top] - \sum_{j,k} p(y = k) \mathbb{1}\{j = k\} \mathbf{W}_U(k) \mathbb{E}[x|y = j]^\top \\
&= \sum_{k=1}^{N+1} p(y = k) \mathbf{W}_U(k) (\hat{\mu}_k - \mu_k)^\top.
\end{aligned}$$

□

Lemma C.13. *Consider a sequence $\{S_t\}_{t \geqslant 1}$ with $S_t = a^t \cdot t$ where $a \neq 1$. Then $\sum_{1 \leqslant t \leqslant T} S_t = \frac{a(1-a^T)}{(a-1)^2} + \frac{a^{T+1} \cdot T}{a-1}$.*

Proof. Denote $X_t \triangleq \sum_{1 \leqslant t \leqslant T} S_t$. Then we have $a \cdot X_t = \sum_{2 \leqslant t \leqslant T+1} a^t \cdot (t-1)$. Hence, it holds

$(a-1)X_t = -\sum_{2\leqslant t\leqslant T} a^t - a + a^{T+1}\cdot T = -\frac{a(1-a^T)}{1-a} + a^{T+1}\cdot T$. Therefore, we have

$$X_t = \frac{a(1-a^T)}{(a-1)^2} + \frac{a^{T+1}\cdot T}{a-1}.$$

□

Lemma C.14. *Consider the following ODE with with $a(0) = b(0) = 0$ and $\alpha \in (0.2, 0.4)$,*

$$\dot{a} = \frac{2 - 2\exp(a)}{\exp(a) + \exp(b) + 1} - 2 + 2\alpha,$$
$$\dot{b} = \frac{2 - 8\exp(b)}{\exp(a) + \exp(b) + 1} - 2 + 10\alpha.$$

Then, when $t \to \infty$, we have

$$a \to -\log(t) - \log(1-\alpha)(4-2\alpha), \quad b \to \log\frac{\alpha}{1-\alpha}.$$

Proof. The ODE can be re-written as

$$\dot{a} = 2\cdot\frac{(\alpha-2)\exp(a) + (\alpha-1)\exp(b) + \alpha}{\exp(a)+\exp(b)+1} \triangleq \frac{2D}{\exp(a)+\exp(b)+1},$$
$$\dot{b} = 10\cdot\frac{(\alpha-\frac{1}{5})\exp(a) + (\alpha-1)\exp(b) + \alpha}{\exp(a)+\exp(b)+1} \triangleq \frac{10E}{\exp(a)+\exp(b)+1}.$$

At $t = 0$, it holds $\dot{a}(0) < 0, \dot{b}(0) < 0$ since $D = 3\alpha - 3 < 0, E = 3\alpha - \frac{6}{5} < 0$. Hence, a and b start to decrease from $t = 0$. The ending of the decreasing happens when one of D and E gets positive. Let's show D and E will never be positive when $\alpha \in (0.2, 0.4)$ by contradiction.

Assume time T_1 is when one of E and E equals to 0 for the first time. This means $E = 0$, because, for any time t, it always holds $D < E$ since $\exp(a) > 0$ for any $a \in \mathbb{R}$. Then at T_1, we have $\dot{a} < 0, \dot{b} = 0$, which means $\exp(a)$ will decrease for any small time window $\Delta t > 0$ and

$\exp(b)$ stays unchanged. Together with $\alpha > 0.2$, this means it has $E < 0$ again at time $T_1 + \Delta t$. Therefore, it is possible for E to be 0, but E will never be positive. Meanwhile, this also guarantees D will always be negative because $D < E$.

Then, we make an observation that when D is always negative and E is always non-positive, the decreasing nature of a will have $D \approx E$ when $t \to \infty$ by $\exp(a) \approx 0$. This implies $b = \log \frac{\alpha}{1-\alpha}$. Then, by taking $\exp(a) = \beta \cdot t^{-\gamma}$, the ODE gives

$$-\gamma \frac{1}{t} = \frac{(2\alpha - 4)\beta \cdot t^{-\gamma}}{\beta \cdot t^{-\gamma} + \frac{1}{1-\alpha}},$$

which gives $\gamma = 1, \beta = \frac{1}{(1-\alpha)(4-2\alpha)}$.

Therefore, when $t \to \infty$, we have

$$a \to \log\left(\frac{1}{(1-\alpha)(4-2\alpha)} t^{-1}\right), \quad b \to \log \frac{\alpha}{1-\alpha}.$$

□

C.7 Input Examples for LLMs

C.7.1 Examples for Prepositions

For experiments in Appendix C.1.1, we use two synthetic datasets: inputs are 30 prepositions, and inputs are 40 incomplete sentences ending with a preposition.

The 30 prepositions are:

"about", "above", "across", "after", "against", "along", "around", "at", "before", "behind", "below", "beneath", "beside", "between", "by", "during", "for", "from", "in", "inside", "into", "near", "of", "on", "over", "through", "to", "under", "with", "without".

Generated by Claude 3 [Anthropic 2024], the 40 incomplete sentences are:

["Inspired painter gazed at pristine canvas, envisioning next creation about", "Children's delighted squeals filled yard as they frolicked, stumbling across", "Singer inhaled deeply, calming nerves before gracing stage before", "Ominous storm clouds amassed, promising downpour that would soon roll in", "Awestruck trekker admired breathtaking summit vista, looking over", "Rich aroma of freshly roasted beans permeated cozy cafe, enticing during", "With deft sleight of hand, illusionist made coin vanish, leaving spectators in awe without", "Majestic oak stood tall, branches reaching skyward above", "Gentle waves caressed shoreline, soothing rhythm lulling along", "Meticulous investigator scoured crime scene, searching for any evidence left behind", "Radiant sunbeams filtered through sheer curtains, warming hardwood floor beneath", "Concert pianist's nimble fingers glided across ivory keys, room resonating with melody around", "Crickets' evening chorus filled silent field from nearby meadow during", "Jubilant laughter resounded down corridor as jovial group headed towards celebration without", "Struggling poet tapped pen restlessly, seeking words to capture elusive emotion beneath", "Soothing patter of raindrops danced on windowpane, inviting serene relaxation with", "Mouthwatering scent of fresh bread beckoned passersby into cozy bakery without", "Mighty waves thundered against jagged cliffs, echoing roar along rugged shoreline around", "Seasoned trekker carefully navigated winding trail, cautiously avoiding exposed roots and rocks beneath", "Graceful ballerina flowed across stage, movements blending seamlessly with melody during", "Crackling campfire cast dancing shadows across gathered faces around", "Vibrant brush strokes danced across canvas, bold hues bursting into life before", "Photographer framed breathtaking sunset, capturing fleeting beauty over glistening ocean without", "Stern librarian hushed raucous group, reminding them to stay quiet inside", "Ink flowed from author's pen, words brimming with raw passion as page filled during", "Earthy aroma of freshly steeped tea perfumed air, inviting moment of serenity along", "Masterful guitarist's fingers danced nimbly across strings, room alive with haunting melody around", "Meticulous chef artfully garnished plate, adding delicate finishing touches over", "Indomitable marathoner pushed through punishing final stretch, fortitude driving every

stride before", "Engrossed scientist examined specimen's intricate structures through microscope beneath", "Nervous thespian steadied breathing, striding into dazzling spotlight, delivering flawless performance with", "Skilled artist's pencil glided gracefully, deftly capturing subject's essence without", "Weary hiker paused to catch breath, marveling at sweeping panorama from lofty peak above", "Deep in thought, writer drummed fingers, seeking perfect phrasing to convey profound emotion without", "Lost in reverie, violinist swayed gently, fingers dancing across delicate strings during", "Painter's brushstrokes burst into radiant life, canvas ablaze with vivid sunset hues over", "Adept photographer framed picturesque scene, preserving landscape's beauty without", "World-renowned chef meticulously garnished plate, each component strategically placed around", "Dedicated researcher scrutinized specimen under microscope, documenting minute details beneath", "Seasoned actor inhaled deeply, embodying character as bright lights engulfed stage with",].

C.7.2 More Examples of Factual Recall

We consider more examples of factual recall with pairs of input and output shown in Table C.3.

C.8 Synthetic IOI Task

Data and task. Here we consider a synthetic data model similar to the IOI task [Wang et al. 2022], with additional noise. Consider a vocabulary $\mathcal{V} = \{1, 2, \ldots, N, N+1\}$. The token $\tau \triangleq N+1$ is the generic noise token. We fix a *trigger* token $q \in [N]$, which governs in-context recall, and a context length T. Each sequence of tokens $z_{1:T} = [z_1, z_2, \ldots, z_T]$ is generated as follows:

i. Sample a correct *output* token $\bar{y}$ and a different *distractor* token y^D uniformly in $[N]$.

ii. Sample three indices $i_1, i_2, i_3 \in [T-2]$ such that their distances are no smaller than 2. (This is for non-overlapping.)

Table C.3: Inputs and Outputs of Factual Knowledge

Input	Target output
The Great Wall is located in	China
Mount Kilimanjaro is located in	Tanzania
The Nobel Prize is awarded in	Sweden
The Statue of Liberty stands in	New York Harbor
Vatican City is enclosed within	Rome
The Acropolis is situated in	Athens
The Sydney Opera House is located on	Bennelong Point
The Galápagos Islands belong to	Ecuador
The Aurora Borealis can be seen in	Norway
The Amazon River flows through	Brazil
The Andes Mountains extend through	Chile
Machu Picchu is found in	Peru
The Kremlin is located in	Moscow
Uluru is a landmark found in	Australia
Petra is an archaeological city in	Jordan
Angkor Wat is located in	Cambodia
The city of Toronto is in	Canada
The city of Barcelona is in	Spain
The city of Mumbai is in	India
The Eiffel Tower is located in	Paris

iii. Set $z_{i_1} = z_{i_2} = z_{i_3} = q$. Among the three indices $i_1 + 1, i_2 + 1, i_3 + 1$, random select one of them with $z_{i_k+1} = \bar{y}$ with the other two as $z_{i_k+1} = y^D$.

iv. Set $z_T = q$ and sample $z_{T+1} \sim p_{\alpha,\bar{y}}(\cdot)$ with

$$p_{\alpha,\bar{y}}(x) = \begin{cases} 1-\alpha, & \text{if } x = \bar{y}, \\ \alpha, & \text{if } x = \tau, \\ 0, & \text{otherwise.} \end{cases}$$

v. Random fill with tokens from $\mathcal{V} \setminus \{q\}$ into the remaining positions in $[T+1] \setminus \{i_1, i_1+1, i_2, i_2+1, i_3, i_3+1, T, T+1\}$.

The key difference between the above data and noisy in-context recall in Section 4.4 is that, in additional to detecting the tokens $\bar{y}$ and y^D after the trigger q, this task also requires counting to decide which of $\bar{y}$ and y^D appear more. This mechanism is exactly the definition of the correct IO token in [Wang et al. 2022].

Most of the other settings are the same as that in Section 4.4, including the training procedure, the architecture of a transformer layer, dimensionality and the vocabulary size.

Results. Figure C.5 shows the test performance for models with layers $L = 3, 4, 5, 6, 7$, where the models are trained with SGD. **Dropping the last-layer MLP** consistently improves the test performance across all models. Figure C.6 shows the test performance for $L = 3, 4, 5$ trained with Adam [Kingma and Ba 2014]. **Truncating the last MLP's input weights** with $\rho = 0.01$ significantly improves the performance for $L = 3, 4$. We also note that the model fails to converge for $L = 5$, possibly because we do not use any normalization technique in the architecture, so the Adam training is less stable for deep transformers.

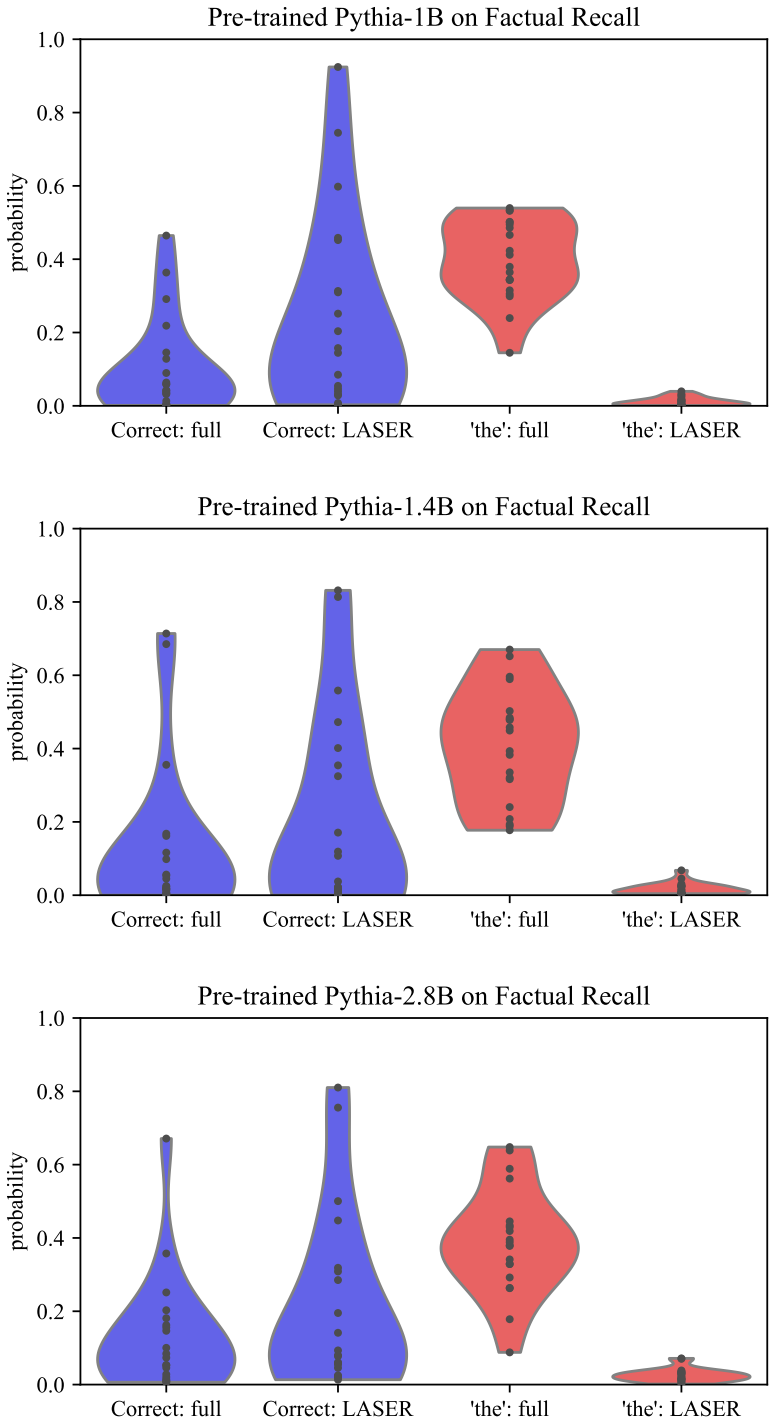


Figure C.3: The prediction distributions of Pythia-1B, 1.4B and 2.8B on more examples of factual recall. Compared with the setting in Figure 4.13 (right), here we use 20 examples in Table C.3. LASER turns out to significantly decrease the probability of "the" against the correct tokens.

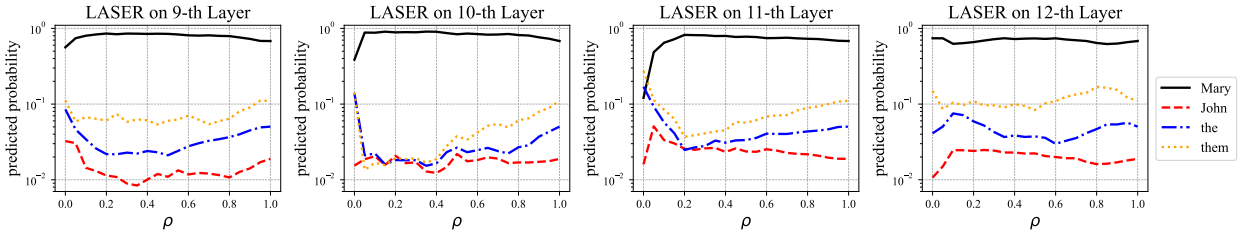


Figure C.4: Predicted probability for $c \in$ {"Mary", "them", "the", "John"}. LASER is conducted on input matrices of MLP layers on the layer $l = 9, 10, 11, 12$ of GPT-2 Small. The input is "When Mary and John went to a store, John gave a drink to". The horizontal is the fraction of perserved rank, $\rho \in [0, 1]$, where $\rho = 1$ stands for the full model. It turns out LASER clearly decreases probability of "the" and "them" when $\rho \in [0.1, 0.8]$ for layer $l = 9, 10, 11$, compared with the full model.

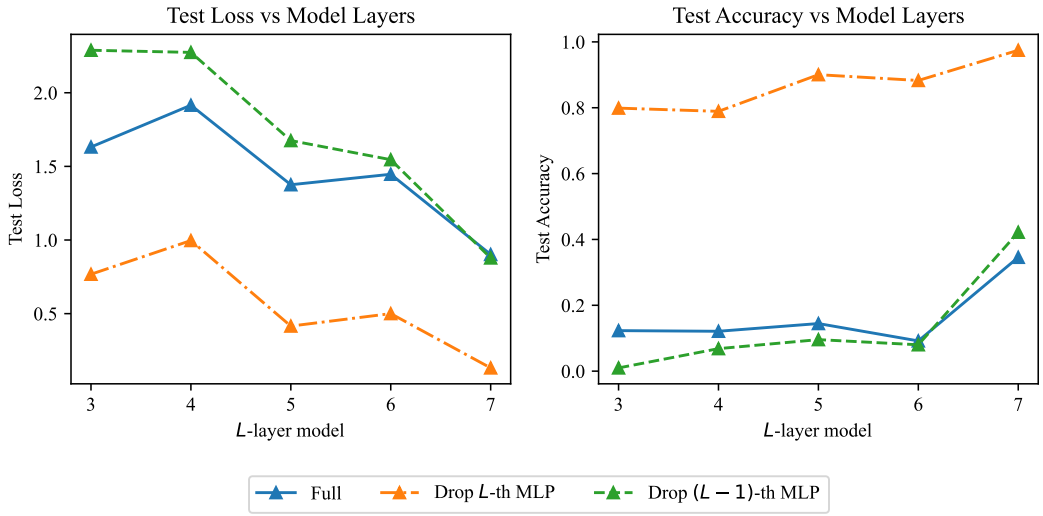


Figure C.5: Synthetic IOI trained with SGD: test loss and accuracy for transformers with different layers. Dropping the last-layer MLP consistently improves the test accuracies across all models.

Figure C.6: Synthetic IOI trained with Adam: test loss and accuracy for transformers with layers $L = 3, 4, 5$. Truncating the last-layer MLP's input weights with $\rho = 0.01$ improves the test performances for $L = 3, 4$, while the model fails to converge for $L = 5$.

Bibliography

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Abernethy, J., Agarwal, A., Marinov, T. V., and Warmuth, M. K. (2024). A mechanism for sample-efficient in-context learning for sparse retrieval tasks. In *International Conference on Algorithmic Learning Theory*.

Ahn, K., Zhang, J., and Sra, S. (2022). Understanding the unstable convergence of gradient descent. *arXiv preprint arXiv:2204.01050*.

AI@Meta (2024). Llama 3 model card.

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

Akyürek, E., Wang, B., Kim, Y., and Andreas, J. (2024). In-context language learning: Arhitectures and algorithms. *arXiv preprint arXiv:2401.12973*.

Alon, U. and Yahav, E. (2020). On the bottleneck of graph neural networks and its practical implications. *arXiv preprint arXiv:2006.05205*.

Anthropic, A. (2024). The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. (2018). Stronger generalization bounds for deep nets via a compression approach. In *International conference on machine learning*, pages 254–263. PMLR.

Arora, S., Li, Z., and Panigrahi, A. (2022). Understanding gradient descent on edge of stability in deep learning. *arXiv preprint arXiv:2205.09745*.

Arous, G. B., Gheissari, R., and Jagannath, A. (2021). Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. (2022). High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*.

Babai, L., Erdos, P., and Selkow, S. M. (1980). Random graph isomorphism. *SIaM Journal on computing*, 9(3):628–635.

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. (2023). Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*.

Barrett, D. and Dherin, B. (2020). Implicit gradient regularization. In *International Conference on Learning Representations*.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. (2023). Pythia: A suite for analyzing large language

models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H., and Bottou, L. (2023). Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*.

Bjorck, N., Gomes, C. P., Selman, B., and Weinberger, K. Q. (2018). Understanding batch normalization. *Advances in neural information processing systems*, 31.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

Bouritsas, G., Frasca, F., Zafeiriou, S., and Bronstein, M. M. (2020). Improving graph neural network expressivity via subgraph isomorphism counting. *arXiv preprint arXiv:2006.09252*.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.

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

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cabannes, V., Simsek, B., and Bietti, A. (2024). Learning associative memories with gradient descent. *arXiv preprint arXiv:2402.18724*.

Chan, S., Santoro, A., Lampinen, A., Wang, J., Singh, A., Richemond, P., McClelland, J., and Hill, F. (2022). Data distributional properties drive emergent in-context learning in transformers.

Chen, A., Schwartz-Ziv, R., Cho, K., Leavitt, M. L., and Saphra, N. (2024). Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. In *International Conference on Learning Representations*.

Chen, M., Wei, Z., Huang, Z., Ding, B., and Li, Y. (2020a). Simple and deep graph convolutional networks. *arXiv preprint arXiv:2007.02133*.

Chen, T., Bian, S., and Sun, Y. (2019a). Are powerful graph neural nets necessary? a dissection on graph classification. *arXiv preprint arXiv:1905.04579*.

Chen, Z., Chen, L., Villar, S., and Bruna, J. (2020b). Can graph neural networks count substructures? *arXiv preprint arXiv:2002.04025*.

Chen, Z., Li, L., and Bruna, J. (2019b). Supervised community detection with line graph neural networks. *Internation Conference on Learning Representations*.

Chen, Z., Villar, S., Chen, L., and Bruna, J. (2019c). On the equivalence between graph isomorphism testing and function approximation with gnns. In *Advances in Neural Information Processing Systems*, pages 15868–15876.

Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems*, 32.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. (2020). Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

Damian, A., Lee, J., and Soltanolkotabi, M. (2022a). Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*.

Damian, A., Ma, T., and Lee, J. D. (2021). Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34.

Damian, A., Nichani, E., and Lee, J. D. (2022b). Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*.

Damian, A., Pillaud-Vivien, L., Lee, J. D., and Bruna, J. (2024). Computational-statistical gaps in gaussian single-index models. *arXiv preprint arXiv:2403.05529*.

Dandi, Y., Krzakala, F., Loureiro, B., Pesce, L., and Stephan, L. (2023). Learning two-layer neural networks, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*.

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852.

Ding, L., Drusvyatskiy, D., and Fazel, M. (2022). Flat minima generalize for low-rank matrix recovery. *arXiv preprint arXiv:2203.03756*.

Du, S. S., Hu, W., and Lee, J. D. (2018). Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems*, 31.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., Welleck, S., West, P., Bhagavatula, C., Le Bras, R., et al. (2024). Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*.

Edelman, B. L., Edelman, E., Goel, S., Malach, E., and Tsilivis, N. (2024). The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. (2022). Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Elkabetz, O. and Cohen, N. (2021). Continuous vs. discrete optimization of deep neural networks. *Advances in Neural Information Processing Systems*, 34.

Flam-Shepherd, D., Wu, T., Friederich, P., and Aspuru-Guzik, A. (2020). Neural message passing on high order paths. *arXiv preprint arXiv:2002.10413*.

Garg, V. K., Jegelka, S., and Jaakkola, T. (2020). Generalization and representational limits of graph neural networks.

Geva, M., Bastings, J., Filippova, K., and Globerson, A. (2023). Dissecting recall of factual associations in auto-regressive language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Geva, M., Schuster, R., Berant, J., and Levy, O. (2021). Transformer feed-forward layers are key-value memories. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Gilmer, J., Ghorbani, B., Garg, A., Kudugunta, S., Neyshabur, B., Cardoze, D., Dahl, G., Nado, Z., and Firat, O. (2021). A loss curvature perspective on training instability in deep learning. *arXiv preprint arXiv:2110.04369*.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272. JMLR. org.

Gong, S., Bahri, M., Bronstein, M. M., and Zafeiriou, S. (2020). Geometrically principled connections in graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11415–11424.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hochreiter, S. and Schmidhuber, J. (1997). Flat minima. *Neural computation*, 9(1):1–42.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Ivanov, S., Sviridov, S., and Burnaev, E. (2019). Understanding isomorphism bias in graph data sets. *arXiv preprint arXiv:1910.12091*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Jastrzebski, S., Arpit, D., Astrand, O., Kerg, G. B., Wang, H., Xiong, C., Socher, R., Cho, K., and Geras, K. J. (2021). Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pages 4772–4784. PMLR.

Jelassi, S., Sander, M., and Li, Y. (2022). Vision transformers provably learn spatial structure. In *Advances in Neural Information Processing Systems*.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. (2019). Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2021). On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29.

Keriven, N. and Peyré, G. (2019). Universal invariant and equivariant graph neural networks. In *Advances in Neural Information Processing Systems*, pages 7092–7101.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Klicpera, J., Bojchevski, A., and Günnemann, S. (2018). Predict then propagate: Graph neural networks meet personalized pagerank. In *International Conference on Learning Representations*.

Klicpera, J., Weißenberger, S., and Günnemann, S. (2019). Diffusion improves graph learning. In *Advances in Neural Information Processing Systems*, pages 13354–13366.

Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L., and Zhang, P. (2013). Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer.

Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., and Gur-Ari, G. (2020). The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*.

Li, G., Muller, M., Thabet, A., and Ghanem, B. (2019). Deepgcns: Can gcns go as deep as cnns? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9267–9276.

Li, G., Xiong, C., Thabet, A., and Ghanem, B. (2020a). Deepergcn: All you need to train deeper gcns. *arXiv preprint arXiv:2006.07739*.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018a). Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31.

Li, P., Wang, Y., Wang, H., and Leskovec, J. (2020b). Distance encoding–design provably more powerful gnns for structural representation learning. *arXiv preprint arXiv:2009.00142*.

Li, Q., Han, Z., and Wu, X.-M. (2018b). Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 3538–3545. Association for the Advancement of Artificial Intelligence.

Li, Y., Li, Y., and Risteski, A. (2023). How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*.

Liu, B., Ash, J. T., Goel, S., Krishnamurthy, A., and Zhang, C. (2023). Transformers learn shortcuts to automata. In *International Conference on Learning Representations*.

Loukas, A. (2020). What graph neural networks cannot learn: depth vs width. In *International Conference on Learning Representations*.

Luan, S., Zhao, M., Chang, X.-W., and Precup, D. (2019). Break the ceiling: Stronger multi-scale deep graph convolutional networks. In *Advances in neural information processing systems*, pages 10945–10955.

Lyu, K., Li, Z., and Arora, S. (2022). Understanding the generalization benefit of normalization layers: Sharpness reduction. *arXiv preprint arXiv:2206.07085*.

Ma, C., Wu, L., and Ying, L. (2022). The multiscale structure of neural network loss functions: The effect on optimization and origin. *arXiv preprint arXiv:2204.11326*.

Ma, C. and Ying, L. (2021). The sobolev regularization effect of stochastic gradient descent. *arXiv preprint arXiv:2105.13462*.

Mahankali, A., Hashimoto, T. B., and Ma, T. (2024). One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *International Conference on Learning Representations (ICLR)*.

Maron, H., Ben-Hamu, H., Serviansky, H., and Lipman, Y. (2019a). Provably powerful graph networks. In *Advances in Neural Information Processing Systems*, pages 2153–2164.

Maron, H., Ben-Hamu, H., Shamir, N., and Lipman, Y. (2018). Invariant and equivariant graph networks.

Maron, H., Fetaya, E., Segol, N., and Lipman, Y. (2019b). On the universality of invariant networks. volume 97 of *Proceedings of Machine Learning Research*, pages 4363–4371, Long Beach, California, USA. PMLR.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*.

Merrill, W., Sabharwal, A., and Smith, N. A. (2022). Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856.

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

Monti, F., Otness, K., and Bronstein, M. M. (2018). Motifnet: a motif-based graph convolutional network for directed graphs. In *2018 IEEE Data Science Workshop (DSW)*, pages 225–228. IEEE.

Morris, C. and Mutzel, P. (2019). Towards a practical k-dimensional weisfeiler-leman algorithm. *arXiv preprint arXiv:1904.01543*.

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. (2019). Weisfeiler and leman go neural: Higher-order graph neural networks. *Association for the Advancement of Artificial Intelligence*.

Murphy, R. L., Srinivasan, B., Rao, V., and Ribeiro, B. (2019). Relational pooling for graph representations. *arXiv preprint arXiv:1903.02541*.

Nanda, N., Chan, L., Liberum, T., Smith, J., and Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*.

Nesterov, Y. (1998). Introductory lectures on convex programming.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. (2017). Exploring generalization in deep learning. *Advances in neural information processing systems*, 30.

Nichani, E., Damian, A., and Lee, J. D. (2024). How transformers learn causal structure with gradient descent. In *International Conference on Learning Representations*.

NT, H. and Maehara, T. (2019). Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. (2022). In-context learning and induction heads. *Transformer Circuits Thread*.

Oono, K. and Suzuki, T. (2020). Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*.

Oymak, S., Rawat, A. S., Soltanolkotabi, M., and Thrampoulidis, C. (2023). On the role of attention in prompt-tuning. In *International Conference on Machine Learning*.

Preciado, V. M. and Jadbabaie, A. (2010). From local measurements to network spectral properties: Beyond degree distributions. In *49th IEEE Conference on Decision and Control (CDC)*, pages 2686–2691. IEEE.

Quirke, L., Heindrich, L., Gurnee, W., and Nanda, N. (2023). Training dynamics of contextual n-grams in language models. *arXiv preprint arXiv:2311.00863*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *Technical report, OpenAI*.

Reddy, G. (2024). The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International Conference on Learning Representations*.

Rossi, E., Frasca, F., Chamberlain, B., Eynard, D., Bronstein, M., and Monti, F. (2020). Sign: Scalable inception graph neural networks. *arXiv preprint arXiv:2004.11198*.

Saade, A., Krzakala, F., and Zdeborová, L. (2014). Spectral clustering of graphs with the bethe hessian. In *Advances in Neural Information Processing Systems*, pages 406–414.

Sanford, C., Hsu, D., and Telgarsky, M. (2024a). One-layer transformers fail to solve the induction heads task. *arXiv preprint arXiv:2408.14332*.

Sanford, C., Hsu, D., and Telgarsky, M. (2024b). Transformers, parallel computation, and logarithmic depth. *arXiv preprint arXiv:2402.09268*.

Sato, R. (2020). A survey on the expressive power of graph neural networks. *arXiv preprint arXiv:2003.04078*.

Sato, R., Yamada, M., and Kashima, H. (2019). Approximation ratios of graph neural networks for combinatorial problems. In *Advances in Neural Information Processing Systems*, pages 4081–4090.

Sato, R., Yamada, M., and Kashima, H. (2020). Random features strengthen graph neural networks. *arXiv preprint arXiv:2002.03155*.

Sharma, P., Ash, J. T., and Misra, D. (2023). The truth is in there: Improving reasoning in language models with layer-selective rank reduction. *arXiv preprint arXiv:2312.13558*.

Smith, S. L., Dherin, B., Barrett, D. G., and De, S. (2021). On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*.

Snell, C., Zhong, R., Klein, D., and Steinhardt, J. (2021). Approximating how single head attention learns. *arXiv preprint arXiv:2103.07601*.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Sukhbaatar, S., Grave, E., Lample, G., Jegou, H., and Joulin, A. (2019). Augmenting self-attention with persistent memory. *arXiv preprint arXiv:1907.01470*.

Tian, Y., Wang, Y., Chen, B., and Du, S. S. (2023). Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In *Advances in Neural Information Processing Systems*.

Tian, Y., Wang, Y., Zhang, Z., Chen, B., and Du, S. (2024). Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.

Vardi, G. and Shamir, O. (2021). Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, pages 4224–4258. PMLR.

Vardi, G., Yehudai, G., and Shamir, O. (2021). Learning a single neuron with bias using gradient descent. *Advances in Neural Information Processing Systems*, 34.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. (2022). Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Wang, Y., Chen, M., Zhao, T., and Tao, M. (2021). Large learning rate tames homogeneity: Convergence and balancing effect. *arXiv preprint arXiv:2110.03677*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*.

Weisfeiler, B. and Leman, A. (1968). The reduction of a graph to canonical form and the algebra which appears therein. *Nauchno-Technicheskaya Informatsia*, 2(9):12-16.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. (2019). Simplifying graph convolutional networks. volume 97 of *Proceedings of Machine Learning Research*, pages 6861–6871, Long Beach, California, USA. PMLR.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). How powerful are graph neural networks? In *International Conference on Learning Representations*.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., and Jegelka, S. (2018). Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, pages 5453–5462.

Ye, T. and Du, S. S. (2021). Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34.

Yehudai, G. and Ohad, S. (2020). Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pages 3756–3786. PMLR.

You, J., Ying, R., and Leskovec, J. (2019). Position-aware graph neural networks. In *International Conference on Machine Learning*, pages 7134–7143.

You, J., Ying, R., Ren, X., Hamilton, W. L., and Leskovec, J. (2018). Graphrnn: A deep generative model for graphs. *CoRR*, abs/1802.08773.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. (2017). Deep sets. In *Advances in neural information processing systems*, pages 3391–3401.

Zhang, B., Luo, S., Wang, L., and He, D. (2023). Rethinking the expressive power of gnns via graph biconnectivity. *arXiv preprint arXiv:2301.09505*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Zhang, M. and Chen, Y. (2018). Link prediction based on graph neural networks. In *Advances in Neural Information Processing Systems*, pages 5165–5175.

Zhang, R., Frei, S., and Bartlett, P. L. (2024). Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55.

Zhang, Y., Backurs, A., Bubeck, S., Eldan, R., Gunasekar, S., and Wagner, T. (2022). Unveiling transformers with lego: a synthetic reasoning task. *arXiv preprint arXiv:2206.04301*.