
Towards Understanding Transformers in Learning Random Walks

Wei Shi

School of Computing and Data Science
The University of Hong Kong
shiwei1997@connect.hku.hk

Yuan Cao

School of Computing and Data Science
The University of Hong Kong
yuancao@hku.hk

Abstract

Transformers have proven highly effective across various applications, especially in handling sequential data such as natural languages and time series. However, transformer models often lack clear interpretability, and the success of transformers has not been well understood in theory. In this paper, we study the capability and interpretability of transformers in learning a family of classic statistical models, namely random walks on circles. We theoretically demonstrate that, after training with gradient descent, a one-layer transformer model can achieve optimal accuracy in predicting random walks. Importantly, our analysis reveals that the trained model is interpretable: the trained softmax attention serves as a token selector, focusing on the direct parent state; subsequently, the value matrix executes a one-step probability transition to predict the location of the next state based on this parent state. We also show that certain edge cases not covered by our theory are indeed failure cases, demonstrating that our theoretical conditions are tight. By investigating these success and failure cases, it is revealed that gradient descent with small initialization may fail or struggle to converge to a good solution in certain simple tasks even beyond random walks. Experiments are conducted to support our theoretical findings.

1 Introduction

In recent years, transformers [31] have revolutionized many fields such as natural language processing [2, 23, 31], computer vision [8, 24], reinforcement learning [5, 13, 15], and have rapidly emerged as a key component in state-of-the-art deep learning models due to their ability to capture complex dependencies in data. While transformers exhibit remarkable practical performance, the underlying mechanisms of transformers are still not well understood due to their complex architecture.

In order to theoretically understand transformers, a number of recent works have investigated their capability in learning from sequential data that follows certain classic statistical models. Specifically, [4, 22] studied sequential data with an underlying causal graph, and theoretically showed how transformers can encode the causal structure with the self-attention mechanism for in-context learning. [6, 9] considered the task of in-context learning of Markov chains, and investigated how two-layer transformers can make predictions for Markov chains according to the context. [12] considered Markov chain forecasting tasks and reveals a connection between a context-conditioned Markov chain and the self-attention mechanism. [21] characterized the loss landscape of a one-layer transformer and demonstrated the existence of global minima and bad local minima in learning Markovian data with vocabularies of size two. Although these works provided valuable insights, there remain many open questions that require further exploration. Notably, the Markov chain data studied in [21] can be essentially understood as a random walk over a space of only two states. Therefore, given the results in [21], a natural question is when and how transformers can learn more general random walks over a larger state space.

In this paper, we consider a classic statistical model for random sequences, namely random walks on circles, and study the capability of one-layer transformers to learn from such data and make predictions. We consider a general setting that allows a large number of nodes (possible locations of the walker) on the circle, which is an extension to the two-state setting studied in [21]. We also consider the general setting where a walker moves on the circle clockwise with probability p , and counter-clockwise with probability $1 - p$, and we build theories that cover the full range of $p \in [0, 1]$. For such a classic, yet relatively general class of random sequences, our goal is to theoretically study the performance of a one-layer transformer trained by gradient descent in predicting the next location, and reveal the interpretability of the obtained transformer model.

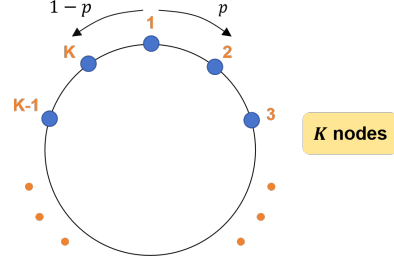


Figure 1: Illustration of random walks on circles with K nodes and transition probability p .

The main contributions of this paper are as follows:

- We theoretically demonstrate that a one-layer transformer can be trained to optimally predict the next location of a random walk with $p \in (0, 1)$. Despite non-convexity, we prove that the trained model converges to the optimal prediction function with a rate $\mathcal{O}(T^{-1/2})$, where T is the iteration number of gradient descent. Furthermore, we show that the trained transformer achieves optimal prediction accuracy $\max\{p, 1 - p\}$ after a constant number of iterations.
- Our analysis reveals that the trained transformer model is interpretable, as we precisely delineate the role of each component in the model. First, the trained softmax attention can select the “direct parent” token by assigning it a near-one score. Second, the trained value matrix serves as a one-step transition model that recovers the ground-truth probability transition matrix, which is applied to the “direct parent” token to make an optimal prediction.
- We also identify failure cases when $p = 0$ or 1 . In these cases, we show that starting from zero initialization, the training of the one-layer transformer model with any loss function and any learning rate will always fail, resulting in a transformer model whose performance is no better than a random guess. This negative result is complementary to our positive guarantees for learning random walks with $p \in (0, 1)$, and they together give a comprehensive characterization of the capability of transformers in learning random walks with $p \in [0, 1]$.
- We provide intuitive explanations that the failure cases with $p = 0$ or 1 are optimization failures caused by zero initialization. Notably, similar optimization failures may also happen beyond the cases of random walks, as we can construct simple question answering tasks that also suffer from similar issues in optimization. We also empirically demonstrate that although training may still take longer, these failure cases can be resolved to a certain extent with random initialization.

2 Problem Setup

In this section, we present our problem formulations, including the random walk prediction task, the one-layer transformer model, and the training algorithm.

We study random walks on circles. Specifically, consider K nodes (possible locations) that are arranged on a circle so that each node has two neighbors. Without loss of generality, we suppose that the nodes are assigned with node IDs $1, 2, \dots, K$ in a clockwise manner. A “walk” on the circle refers to the process where a “walker” moves step-by-step among the nodes of the circle. We suppose that, starting from a random initial location, at each step, the walker moves either clockwise with probability p or counterclockwise with probability $1 - p$, to a neighboring node of its current position, where $p \in (0, 1)$ is a fixed probability. In this way, a random walk of length N generates a sequence of “states” $s_1, \dots, s_N$, where $s_i \in [K]$ denotes the location (node ID) of the walker at the i -th step. We aim to address the problem of predicting the walker’s next location s_N based on the historical locations $s_1, \dots, s_{N-1}$.

To better formulate this random walk prediction task, we map $s_1, \dots, s_{N-1}$ to embeddings $\mathbf{x}_1, \dots, \mathbf{x}_{N-1} \in \mathbb{R}^K$. Our goal is then to train a model to predict the target $y = s_N$ based on $\mathbf{x}_1, \dots, \mathbf{x}_{N-1}$. In the following, we give a detailed definition and discuss some basic properties.

Random walk on circles. Suppose that there are K nodes on a circle and the transition probability is p . $\mathbf{x}_1, \dots, \mathbf{x}_{N-1}, y$ are generated as follows:

1. Draw $s_1 \sim \text{Unif}([K])$.
2. For $i = 2, \dots, N$, sample either $s_i = \langle s_{i-1} + 1 \rangle_K$ with probability p or $s_i = \langle s_{i-1} - 1 \rangle_K$ with probability $1 - p$.
3. Set $\mathbf{x}_i = \mathbf{e}_{s_i}$, $i \in [N - 1]$, and $y = s_N$.

Here, $\langle s \rangle_K$ is defined as the integer satisfying $\langle s \rangle_K \in [K]$ and $\langle s \rangle_K \equiv s \pmod{K}$. It is clear that the sequence $\mathbf{x}_1, \dots, \mathbf{x}_{N-1}, \mathbf{e}_y$ form a Markov chain, and $\mathbb{P}(y|\mathbf{x}_1, \dots, \mathbf{x}_{N-1}) = \mathbb{P}(y|\mathbf{x}_{N-1})$. Moreover, the transition matrix of the Markov model is $\mathbf{\Pi} = (\pi_{i,j})_{K \times K}$, where $\pi_{i,j} = p \cdot \mathbb{1}\{i \equiv j - 1 \pmod{K}\} + (1 - p) \cdot \mathbb{1}\{i \equiv j + 1 \pmod{K}\}$. A visualization of $\mathbf{\Pi}$ is given in Figure 2. The Markov property indicates that the optimal predictor of y is given by

$$f^{\text{OPT}}(\mathbf{x}_1, \dots, \mathbf{x}_{N-1}) = \mathbf{\Pi}^\top \mathbf{x}_{N-1},$$

and the optimal prediction accuracy any predictor can achieve is $\text{OPT} = \max\{p, 1 - p\}$. Based on the formulation of $f^{\text{OPT}}(\cdot)$, it is clear that a simple autoregressive model $f(\mathbf{X}) = \mathbf{V} \mathbf{x}_{N-1}$ can already solve this random walk prediction task. However, the goal of this work is not to identify the optimal model for solving random walk tasks. Instead, we aim to understand and analyze the capability and interpretability of transformers when learning such classic statistical tasks from scratch. Therefore, in the following, we introduce a simple transformer model.

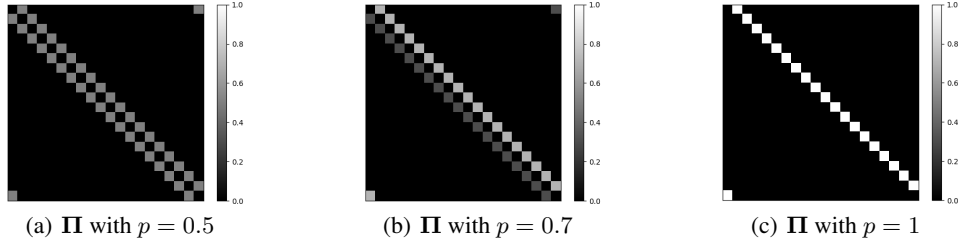


Figure 2: Visualization of the transition matrices $\mathbf{\Pi}$ with $p = 0.5, 0.7, 1.0$ respectively.

We consider learning the random walk prediction task with a simple one-layer transformer model. By naturally treating the one-hot vectors $\mathbf{x}_1, \dots, \mathbf{x}_{N-1}$ as token embeddings, the task to predict the next position y is a problem of next token prediction. Therefore, we define the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N-1}, \mathbf{0}] \in \mathbb{R}^{K \times N}$. We also employ a positional embedding matrix $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N] \in \mathbb{R}^{M \times N}$, where M is the embedding dimension with $M = \Omega(N^{3/2})$ and $\mathbf{p}_i \in \mathbb{R}^M$ is defined as

$$\mathbf{p}_i = \left[\sin\left(\frac{i\pi}{M+1}\right), \sin\left(\frac{2i\pi}{M+1}\right), \dots, \sin\left(\frac{Mi\pi}{M+1}\right) \right]^\top$$

for $i = 1, 2, \dots, N$. The positional embeddings above are inspired by the fact that $\langle \mathbf{p}_i, \mathbf{p}_j \rangle = 0$ for all $i \neq j$, which significantly helps to simplify our theoretical analysis (see Lemma E.5 in the appendix). Additionally, orthogonal positional embeddings are commonly considered in existing theoretical studies [3, 22, 32, 33]. Then, we define the matrix $\widetilde{\mathbf{X}}$ by concatenating the input matrix $\mathbf{X}$ and the positional matrix $\mathbf{P}$ as

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{N-1} & \mathbf{0} \\ \mathbf{p}_1 & \mathbf{p}_2 & \cdots & \mathbf{p}_{N-1} & \mathbf{p}_N \end{bmatrix} := [\widetilde{\mathbf{x}}_1, \widetilde{\mathbf{x}}_2, \dots, \widetilde{\mathbf{x}}_N] \in \mathbb{R}^{(K+M) \times N}.$$

We consider a one-layer transformer model to make a prediction on a given input matrix $\mathbf{X}$. The transformer is defined as follows:

$$f_\theta(\mathbf{X}) = \mathbf{V} \mathbf{X} \mathcal{S}(\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{x}}_N), \quad (2.1)$$

where $\mathbf{V} \in \mathbb{R}^{K \times K}$, $\mathbf{W} \in \mathbb{R}^{(K+M) \times (K+M)}$ are the trainable parameter matrices, $\mathcal{S}: \mathbb{R}^N \rightarrow \mathbb{R}^N$ is the softmax function defined by $[\mathcal{S}(\mathbf{z})]_i = \frac{\exp(z_i)}{\sum_{j=1}^N \exp(z_j)}$, and $\theta = (\mathbf{V}, \mathbf{W})$ denotes the collection of all the trainable parameters. In this definition, we consider a reparameterization where we

use a single matrix $\mathbf{W}$ to denote the product of the key and query parameter matrices in self-attention. Such kind of reparameterizations are widely considered in theoretical studies of transformer models [11, 12, 14, 16, 22, 29, 32, 34]. In addition, we omit the softmax scaling factor, which is mainly for simplicity. Such omission has also been considered in most of the theoretical studies [7, 17, 18, 22, 25, 26].

Note that by (2.1), given any input matrix $\mathbf{X}$, the transformer model outputs a K -dimensional vector. This follows the standard practice of K -class classification: for $i \in [K]$, $[f_\theta(\mathbf{X})]_i$ can be treated as a predicted “score” of the i -th class. More specifically, we define the prediction rule as follows.

Definition 2.1. For any predictor $f(\mathbf{X}) : \mathbb{R}^{K \times N} \rightarrow \mathbb{R}^K$, the predicted label is given as

$$\text{Pred}(f(\mathbf{X})) := \min \left\{ j \in [K] : [f(\mathbf{X})]_j = \max_{i \in [K]} \{ [f(\mathbf{X})]_i \} \right\}.$$

The definition above matches the common practice to predict the label that corresponds to the entry in $f(\mathbf{X})$ with the maximum function value. It also gives a naive way to handle ties – when $f(\mathbf{X})$ contains multiple dimensions with the same (and maximum) function value, we always predict the dimension corresponding to the smallest label. We remark that this definition to handle ties is just to exclude ambiguity, and the detailed rule to handle ties is not essential. Our result works for all reasonable methods to handle ties.

We train the transformer model defined in (2.1) by gradient descent, minimizing the loss function

$$L(\theta) = \mathbb{E}_{(\mathbf{X}, y)} \left[\ell \left(\mathbf{e}_y^\top f_\theta(\mathbf{X}) \right) \right], \quad (2.2)$$

where $\ell(\cdot)$ is a loss function. In terms of the specific choice of $\ell(\cdot)$, our analysis will cover learning random walks on a circle by minimizing the log-loss $\ell(z) = -\log(z + \epsilon)$, which has been considered in a series of recent works [12, 16, 21, 28].

We consider gradient descent with zero initialization $\mathbf{V}^{(0)} = \mathbf{0}^{K \times K}$, $\mathbf{W}^{(0)} = \mathbf{0}^{(K+M) \times (K+M)}$ to train the model. The update rule for the parameter matrices $\mathbf{W}, \mathbf{V}$ are as follows:

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \nabla_{\mathbf{W}} L(\theta^{(t)}), \quad \mathbf{V}^{(t+1)} = \mathbf{V}^{(t)} - \eta \nabla_{\mathbf{V}} L(\theta^{(t)}), \quad (2.3)$$

where $\eta > 0$ is the learning rate and $t \geq 0$ is the iteration number. Note that the log-loss $\ell(z) = -\log(z + \epsilon)$ is well-defined and does not blow up at zero initialization due to the stability constant $\epsilon > 0$. Gradient descent with appropriate learning rate further ensures that it does not blow up during training.

3 Main Results

In this section, we present our main theoretical results on learning random walks with a self-attention layer. In our result, we can choose any $T^* = \text{poly}(\eta, \epsilon^{-1}, K, N, M)$ as the maximum admissible number of iterations, and only consider the training period $0 \leq t \leq T^*$. This technical assumption regarding a polynomially large maximum admissible number of iterations prevents training from becoming exponentially long and is a mild assumption since exponentially long training is impractical.

Our main results are given in the following theorem.

Theorem 3.1. Suppose that $0 < p < 1$, $K > 4$, and $N \geq C_p \cdot \text{poly}(K)$ for some constant C_p that only depends on p . Further suppose that the transformer is trained by gradient descent (2.3) to minimize the loss (2.2) with $\ell(z) = -\log(z + \epsilon)$, and $\eta, \epsilon = \Theta(1)$. Then there exists $T_0 = \Theta(1)$, such that for all $T_0 \leq T \leq T^*$, it holds that:

1. The trained transformer achieves optimal prediction accuracy:

$$\mathbb{P}[\text{Pred}(f_{\theta^{(T)}}(\mathbf{X})) = y] = \text{OPT} = \max\{p, 1 - p\}.$$

2. The transformer converges to the optimal predictor:

$$\left\| \frac{f_{\theta^{(T)}}(\mathbf{X})}{\|f_{\theta^{(T)}}(\mathbf{X})\|_2} - f^{\text{OPT}}(\mathbf{X}) \right\|_2 = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

3. The value matrix converges to the true transition matrix in direction:

$$\left\| \frac{\mathbf{V}^{(T)}}{\|\mathbf{V}^{(T)}\|_F} - \frac{\mathbf{\Pi}^\top}{\|\mathbf{\Pi}^\top\|_F} \right\|_F = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

4. Softmax attention selects the “direct parent” of y :

$$\begin{aligned} [\mathcal{S}(\widehat{\mathbf{X}}^\top \mathbf{W}^{(T)} \tilde{\mathbf{x}}_N)]_{N-1} &\geq 1 - \exp(-\Omega(N)), \\ [\mathcal{S}(\widehat{\mathbf{X}}^\top \mathbf{W}^{(T)} \tilde{\mathbf{x}}_N)]_j &\leq \exp(-\Omega(N)) \text{ for all } j \neq N-1. \end{aligned}$$

The first result in Theorem 3.1 states that the transformer trained by gradient descent for a constant number of iterations can achieve a prediction accuracy $\max\{p, 1-p\}$, which matches the optimal accuracy OPT. The second result in Theorem 3.1 further gives a more detailed characterization of the trained transformer, and demonstrates that the normalized model converges to the optimal prediction model $f^{\text{OPT}}(\mathbf{X}) = \mathbf{\Pi}^\top \mathbf{x}_{N-1}$.

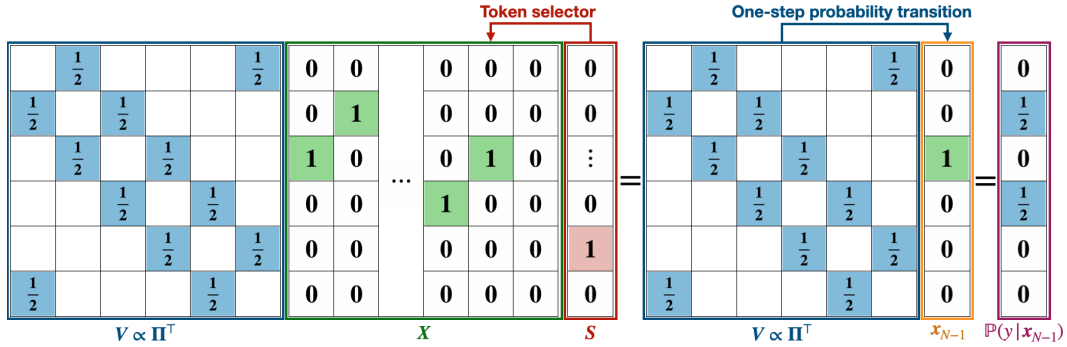


Figure 3: Illustration of how the trained one-layer transformer performs optimal prediction in an example with $p = 1/2$, $K = 6$. Here we denote $\mathbf{S} = \mathcal{S}(\widehat{\mathbf{X}}^\top \mathbf{W}^{(T)} \tilde{\mathbf{x}}_N)$.

The third and fourth results in Theorem 3.1 further back up the first two results by a precise characterization on how the self-attention mechanism works in predicting random walks. Specifically, the third result demonstrates that in direction, **the value matrix $\mathbf{V}^{(T)}$ serves as a one-step transition model** by recovering the ground-truth one-step transition matrix, and the last result indicates that **softmax attention performs optimal token selection** by assigning a near-one score to the $(N-1)$ -th token (the direct parent of y). These two results show that, when learning to predict random walks, the one-layer transformer obtained through gradient descent is **interpretable**: the trained one-layer transformer model makes predictions by (i) selecting the correct parent token $\mathbf{x}_{N-1}$ of y by assigning a softmax weighting close to 1 to it, and (ii) predicting y by applying a one-step transition model to $\mathbf{x}_{N-1}$ through the linear mapping defined by $\mathbf{V}$. An illustration is given in Figure 3.

A recent related work [21] studied how one-layer transformers can learn Markov chains with a state space of size two, and analyzed the loss landscape by identifying global minima and bad local minima. We remark that we consider a different parameterization of one-layer transformers, which makes our results not rigorously comparable. However, by studying random walks over state spaces of arbitrary sizes and establishing theoretical guarantees directly on transformers trained by gradient descent, our work explores a setting that is not covered in [21].

Notably, Theorem 3.1 is based on the assumption that the transition probability satisfies $0 < p < 1$, excluding the edge cases when $p = 0, 1$. In Section 4, we will show a negative result showing that when $p = 0, 1$, gradient descent with zero initialization fails to achieve good prediction accuracy. Therefore, the assumption $0 < p < 1$ in Theorem 3.1 is inevitable.

Here we informally explain how Theorem 3.1 is proved. The idea of the proof can be shown by investigating the first several gradient descent steps:

Step 1. After the first gradient descent step, due to the zero initialization and the Toeplitz property of the transition matrix $\mathbf{\Pi}$, it can be shown that $\mathbf{V}^{(1)}$ is also a Toeplitz matrix whose largest entries appear exactly on the locations of the largest entries of $\mathbf{\Pi}^\top$. $\mathbf{W}^{(1)}$ is still a zero matrix due to the fact that $\mathbf{V}^{(0)} = \mathbf{0}$.

Step 2. With the same analysis, we can also show that $\mathbf{V}^{(2)}$ is a Toeplitz matrix whose largest entries appear exactly on the locations of the largest entries of $\mathbf{\Pi}^\top$. Moreover, the locations of the largest entries in $\mathbf{V}^{(1)}$ encourage $\mathbf{W}^{(2)}$ to be updated so that higher softmax weightings are put upon the “direct parent” token $\mathbf{x}_{N-1}$ (see Lemma C.5 in the appendix).

Step 3. The higher weighting on $\mathbf{x}_{N-1}$ given by $\mathbf{W}^{(2)}$ further encourages $\mathbf{V}^{(3)}$ to be updated towards $\mathbf{\Pi}^\top$ in direction. And $\mathbf{V}^{(2)}$ obtained in Step 2 continues to encourage $\mathbf{W}^{(3)}$ to place a even higher weighting on $\mathbf{x}_{N-1}$ (see Lemma C.7 in the appendix).

From the three gradient descent steps discussed above, it is clear that $\mathbf{V}^{(t)}$ will converge to the direction of $\mathbf{\Pi}^\top$, and $\mathbf{W}^{(t)}$ will consistently place a high weighting on the direct parent token $\mathbf{x}_{N-1}$. This is our key intuition for proving Theorem 3.1, and in our formal proof (given in the appendix), we use an induction to characterize the whole training procedure.

4 “Deterministic Walks” with $p = 0, 1$

In Section 3, we demonstrate that a one-layer transformer can be trained to optimally predict random walks under the assumption $0 < p < 1$. In this section, we justify the necessity of this assumption and analyze the edge cases $p = 0, 1$, which lead to a failure for learning random walks. Random walks with $p = 0, 1$ are essentially “deterministic walks”, and for them we have the following theorem.

Theorem 4.1. Suppose that $N = rK + 1$ with $r \geq 1$, and $p = 0$ or 1 . Further suppose that the transformer is trained by gradient descent (2.3) to minimize the loss (2.2). Then for any loss function $\ell(\cdot)$, any learning rate $\eta > 0$, and any $T \geq 0$, it holds that

$$\mathbb{P}(\text{Pred}(f_{\theta(T)}(\mathbf{X})) = y) = \frac{1}{K}.$$

Moreover, with probability 1, for all $T \geq 0$, it holds that

$$\mathbf{V}^{(T)} \propto \mathbf{1}_{K \times K}, \quad \left[\mathcal{S}(\widetilde{\mathbf{X}}^\top \mathbf{W}^{(T)} \widetilde{\mathbf{x}}_N) \right]_1 = \dots = \left[\mathcal{S}(\widetilde{\mathbf{X}}^\top \mathbf{W}^{(T)} \widetilde{\mathbf{x}}_N) \right]_{N-1}.$$

Theorem 4.1 shows that the prediction accuracy of the trained transformer on deterministic walks is $1/K$, equal to the accuracy of a random guess. Moreover, the characterizations of the softmax scores and the value matrix $\mathbf{V}^{(T)}$ further demonstrate that the transformer takes average over all tokens, and then gives the same prediction scores for all possible values of y . Notably, these results hold for any choice of the loss function and any learning rate, indicating that this failure case of the transformer cannot be resolved by simply adjusting these training setups.

Theorems 3.1 and 4.1 together provide a comprehensive characterization of the transformer’s performance when learning random walks with transition probabilities $p \in [0, 1]$. Based on Theorem 4.1, it is clear that the assumption in Theorem 3.1 that $0 < p < 1$ is necessary and cannot be further relaxed. Theorems 3.1 and 4.1 also point out an interesting fact, that compared to random walks with $0 < p < 1$, the seemingly easier task of predicting deterministic walks with $p = 0, 1$ may be more challenging for a transformer to learn. Here, we would like to remark that the failure of learning deterministic walks is an optimization failure, and is highly due to zero initialization. To see the reason, we can consider the two initial gradient descent steps:

Step 1. Since the initial softmax weightings on all tokens are the same, $\mathbf{V}^{(1)}$ is essentially trained based on the averaged token $\bar{\mathbf{x}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathbf{x}_i$. Importantly, we can see that

$$\bar{\mathbf{x}} \text{ is a constant vector that does not depend on } y.$$

This means that $\bar{\mathbf{x}}$ does not provide any helpful information in predicting y , and is therefore “uninformative”. As a result, it can be shown that all entries in $\mathbf{V}^{(1)}$ are equal (see Lemma D.1 in the appendix). Additionally, $\mathbf{W}^{(1)}$ is still a zero matrix since $\mathbf{V}^{(0)} = \mathbf{0}$.

Step 2. With the same analysis as Step 1, we can show that all entries in $\mathbf{V}^{(2)}$ are equal. Moreover, due to the fact that $\mathbf{V}^{(1)}$ is proportional to the all-one matrix, it can be further shown that $\mathbf{W}^{(2)}$ is updated so that the softmax weightings on all tokens $\mathbf{x}_1, \dots, \mathbf{x}_{N-1}$ remain equal.

In our formal proof, we inductively show that throughout training, the value matrix $\mathbf{V}^{(t)}$ is always proportional to the all-one matrix, and the softmax weights on all tokens are always the same.

From the discussion above, we can observe that transformers struggle to learn deterministic walks due to the unbreakable symmetry in the training dynamics, arising from the “uninformative” token average $\bar{x}$ given by the zero initialization. We remark that if random initialization is used, the symmetry will be broken, and transformers may succeed in learning the optimal predictor. However, if the random initialization is too small in scale, we can still expect that the optimization for learning deterministic walks to be more challenging compared to that for random walks. While our theoretical analysis does not cover random initialization, we present empirical studies in Sections 5 and 6 to verify this claim. We believe that theoretically analyzing the impact of random initializations can be an important future work direction.

5 Experiment

In this section, we present experimental results to support our theoretical analysis. We consider two cases: the first one is the zero initialization case that aligns with the setting used in our theoretical analysis, and the second one is the random initialization case, which helps verify our discussion about the optimization failure caused by zero initialization. In all experiments introduced in this section, we set the number of nodes $K = 6$ and the length of each sequence $N = 97$. We utilize the transformer model introduced in Section 2 and utilize the gradient method to train the model. The prediction accuracy is calculated based on 1000 test data.

Zero initialization. In this case, we set the length of the positional embedding $M = 1000$, the initialization $\mathbf{V}^{(0)} = \mathbf{0}_{K \times K}$, $\mathbf{W}^{(0)} = \mathbf{0}_{(K+M) \times (K+M)}$, and the learning rate $\eta = 1$. The constant ϵ in the log-loss is set as $\epsilon = 0.1$. For both tasks, we generate 1000 sequences to train the model.

Figure 4 and Figure 5 illustrate the results of the experiment for $p = 0.5$ and $p = 1$ respectively: Figure 4(a) and Figure 5(a) present the prediction accuracy; Figure 4(b) and Figure 5(b) visualize the value matrix $\mathbf{V}^{(T)}$ after 50 iterations; Figure 4(c) and Figure 5(c) display the attention scores attached to each token after 50 iterations. To clearly observe the results, we also provide Figure 4(d) that represents the part of Figure 4(c).

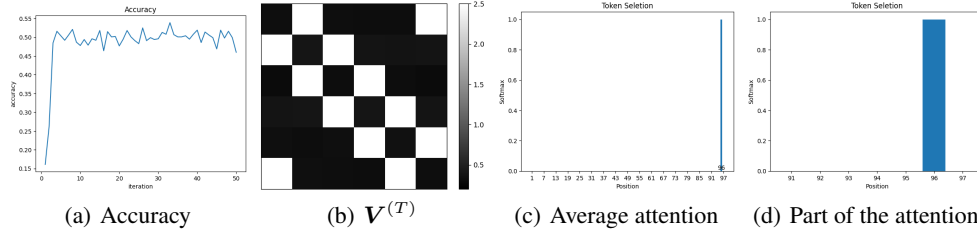


Figure 4: The results of the experiments for $p = 0.5$ with zero initialization: (a) is the test accuracy; (b) is the visualization of $\mathbf{V}$; (c) and (d) present the average attention of the test data with x -axis representing the position of the token and y -axis representing the attention score.

We can observe that these experimental results for $p = 0.5$ provide strong support for Theorem 3.1. Figure 4(a) shows that the prediction accuracy is close to the optimal accuracy (50%) within constant iterations. Figure 4(b) indicates that $\mathbf{V}^{(T)}$ can recover the transition matrix $\mathbf{\Pi}^\top$ as shown in Figure 2(a). Figure 4(c) presents that the $(N - 1)$ -th attention score is the highest and close to 1, indicating that the self-attention layer is able to select the true parent token. All of these experimental results demonstrate the performance of transformers in learning random walks.

In addition, we can find that the experimental outcomes for $p = 1$ match the theoretical results stated in Theorem 4.1. We obtain an accuracy close to 0.167 from Figure 5(a), which suggests that the prediction accuracy for learning deterministic walks is approximately equal to $1/K$, far away from the optimal accuracy (100%) and no better than a random guess. Figure 5(b) indicates that $\mathbf{V}^{(T)}$ is approximately proportional to $\mathbf{1}_{K \times K}$. Figure 5(c) shows that the attention scores attached to all tokens are identical, which proves that the self-attention layer cannot select any of the tokens when learning deterministic walks. These experimental results demonstrate that the self-attention mechanism struggles in learning deterministic walks with $p = 0, 1$.

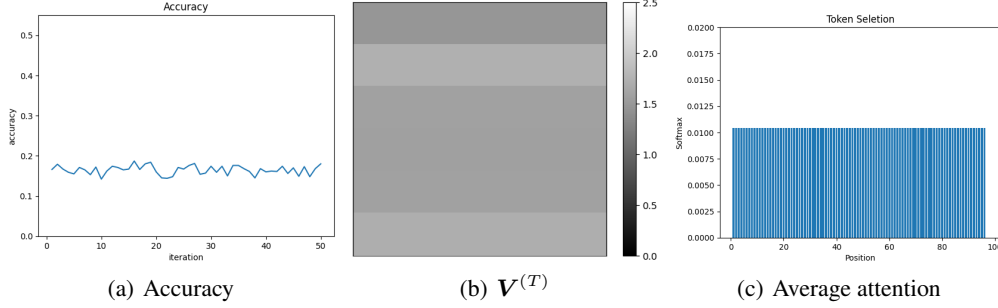


Figure 5: Experiments for $p = 1$ with zero initialization. (a) gives the prediction accuracy along training. (b) is the visualization of $\mathbf{V}$. (c) is the average attention of the test data with x -axis representing the position of the token and y -axis representing the attention score.

Random initialization. In this case, we set the length of the positional embedding $M = 1000$, the initialization $\mathbf{V}_{ij}^{(0)}, \mathbf{W}_{ij}^{(0)} \sim N(0, \sigma^2)$ with $\sigma = 0.01$, and the learning rate $\eta = 0.01$. The constant ϵ in the log-loss is set as $\epsilon = 0.1$. For both tasks, we generate 1000 sequences to train the model. To ensure numerical stability, we normalize $\tilde{\mathbf{x}}_i$'s to unit length in the softmax attention.

Figure 6 illustrates the results of the experiment for $p = 0.5$ and $p = 1$. Figure 6(a) and Figure 6(c) show the prediction accuracy within 1000 iterations, respectively. In Figure 6(b) and Figure 6(d), we first normalize the output of the trained transformer model to get a K -dimensional vector, which can be regarded as the prediction distribution of K locations. The KL-divergence between this prediction distribution and the true distribution of $y|x_{N-1}$ is illustrated in Figure 6(b) and Figure 6(d).

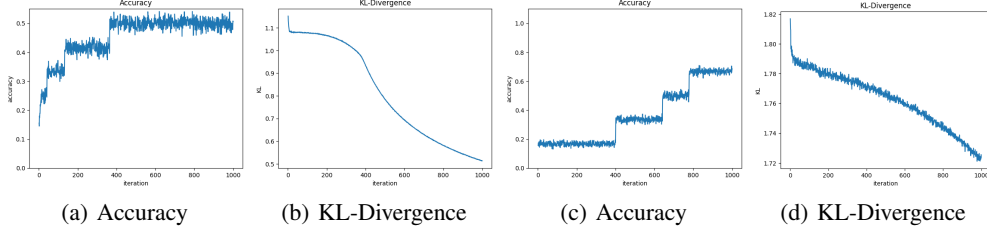


Figure 6: The results of the synthetic experiment with random initialization: (a) and (b) correspond to the experiment for $p = 0.5$; (c) and (d) correspond to the experiment for $p = 1$. (a) and (c) present the prediction accuracy. In (b) and (d), we first normalize the output of the trained transformer model to get a K -dimensional vector, representing the prediction distribution of K locations. Then, we display the KL-divergence between this prediction distribution and the true distribution of $y|x_{N-1}$.

Figure 6(a) clearly shows that in the experiment for $p = 0.5$, the accuracy is close to the optimal accuracy (50%) after around 400 iterations. However, as shown in Figure 6(c), for $p = 1$, the prediction accuracy cannot reach the optimal accuracy (100%) within 1000 iterations. Based on the plots of KL-divergence, we can also see that the transformer learns the true prediction distribution of random walks much faster than learning that of deterministic walks. Note that these results are for training with random initialization, and hence the results do not perfectly match our theory for zero initialization in Section 3. However, the experiment results still show that the optimization task is more challenging, even with small random initialization.

6 Beyond Random Walks

In Section 4, we intuitively explain that one-layer transformers may suffer from optimization issues when learning deterministic walks with $p = 0, 1$ due to the fact that zero initialization produces a token average which is “uninformative”. In this section, we briefly discuss other tasks beyond random walks where transformers with zero/small initialization may face similar optimization challenges.

We construct two question answering tasks. The detailed descriptions are given as follows.

Task 1. We consider a simple question answering task. Possible input questions are of the form: *Based on the list ‘apple, orange, apple, apple, orange’, which type of fruit appears most frequently?*

Here, the list stated in the question can be any combination of ‘apple’ and ‘orange’ with a fixed length of 5. Therefore, there are a total of 32 possible questions the model may see, and each of these questions occurs with probability $1/32$. Ignoring punctuation marks, each input sample is assumed to be 16 words involving the list and other words in the inquiry sentence. The correct response (the ‘label’ for classification) is the fruit that appears most frequently in the list. For example, for the question “Based on the list ‘apple, orange, apple, apple, orange’, which type of fruit appears most frequently?”, the correct response is *apple*.

Task 2. We again consider a simple question answering task with only two possible questions:

Based on the sentence ‘I prefer an apple to an orange’, which type of fruit do I prefer?
Based on the sentence ‘I prefer an orange to an apple’, which type of fruit do I prefer?

Here, each of the two questions above occurs with probability $1/2$. Similar to Task 1, we ignore the punctuation marks, and the input is the 18 words in the sentence. The correct response (the ‘label’ for classification) is *apple* for the first question above, and *orange* for the second question above.

Combining all the words appearing in two tasks, we attain a vocabulary with a length of 19: {‘apple’, ‘orange’, ‘Based’, ‘on’, ‘the’, ‘which’, ‘type’, ‘of’, ‘fruit’, ‘list’, ‘appears’, ‘most’, ‘frequently’, ‘sentence’, ‘I’, ‘prefer’, ‘an’, ‘to’, ‘do’}. We embed this sequence as a matrix $\mathbf{E} = [e_1, e_2, \dots, e_{19}] \in \mathbb{R}^{19 \times 19}$. Then, we know that the length of the vocabulary K and the length of each input sequence N are set as $(K, N) = (19, 17), (19, 19)$ for Task 1 and Task 2 respectively.

In the experiments, we consider a similar transformer model as we introduced in our theoretical analysis. To train the model, we consider Gaussian random initialization $\mathbf{V}_{ij}^{(0)}, \mathbf{W}_{ij}^{(0)} \sim N(0, \sigma^2)$ with $\sigma = 0.01$, and we use gradient descent with learning rate $\eta = 0.1$ to train the model. The constant ϵ in the log-loss is set as $\epsilon = 0.1$. Both the training and test datasets contain 1000 samples.

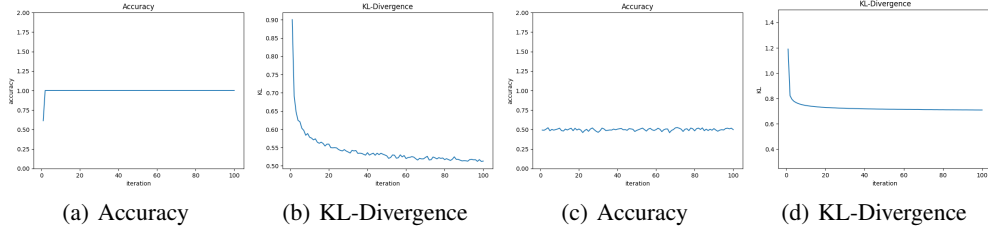


Figure 7: The results of the experiment for Task 1 and Task 2: (a) and (b) correspond to the experiment for Task 1; (c) and (d) correspond to the experiment for Task 2.

Figure 7 shows the experiment results for Task 1 and Task 2. Figure 7(a) and Figure 7(c) present the test accuracy. In Figure 7(b) and Figure 7(d), we first normalize the output of the trained transformer model to get a K -dimensional vector, representing the prediction distribution of K words. Then, we report the KL-divergence between this prediction distribution and the true distribution of $y|x_1, x_2, \dots, x_{N-1}$ in Figure 7(b) and Figure 7(d). The experiment results show a clear difference between the performances of the transformer model in the two tasks. In Task 1, the trained transformer model can successfully approach the optimal accuracy (100%) within 100 iterations. However, in Task 2, the test accuracy remains around 50% within 100 iterations.

The results can be explained following the discussion in Section 4. Specifically, in Task 1, the average of the word embeddings $\bar{x}$ in a question can help the model to find the correct response, while in Task 2, the two questions give the *same* average of word embeddings $\bar{x}$, and therefore, it causes inefficient optimization.

The results for these two tasks demonstrate that our theories and explanations for random walks can also guide the construction of various other learning tasks and predict the performance of a transformer model in these tasks. This confirms the validity of our theories and explanations and highlights the insights provided by our study.

7 Conclusion

This paper investigates the capability of a one-layer transformer model to learn random walks on circles. We demonstrate that transformers successfully learn such walks for the transition probability $0 < p < 1$, achieving the optimal prediction accuracy. We also show that the trained model is

interpretable: the softmax attention mechanism effectively selects the correct “parent” token, while the value matrix recovers a one-step transition model that applies to the selected “parent” token for optimal prediction. In addition, we identify that the edge cases ($p = 0, 1$) are failure cases, thereby proving the necessity of the assumption $0 < p < 1$. Motivated by the analysis of success and failure in learning random walks, we design simple question answering tasks that exhibit similar optimization challenges, showing the broader applicability of our analysis to other tasks beyond random walks. We also provide experimental results to validate our theoretical findings.

In future works, it is important to theoretically study the impact of random initialization in learning random walks. Moreover, an interesting future work direction is to extend the results and study the performance of deeper transformer architectures, which may require more advanced theoretical tools. Moreover, extending the finding to more complicated learning tasks, such as random sequences generated by general Markov chains or Bayesian networks, is also an important future work direction.

Acknowledgments

We thank the anonymous reviewers and area chairs for their valuable and constructive comments. Yuan Cao is partially supported by NSFC 12301657 and Hong Kong ECS award 27308624.

References

- [1] ABBE, E., BENGIO, S., BOIX-ADSERA, E., LITWIN, E. and SUSSKIND, J. (2024). Transformers learn through gradual rank increase. *Advances in Neural Information Processing Systems* **36**.
- [2] ACHIAM, J., ADLER, S., AGARWAL, S., AHMAD, L., AKKAYA, I., ALEMAN, F. L., ALMEIDA, D., ALTENSCHMIDT, J., ALTMAN, S., ANADKAT, S. ET AL. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- [3] BIETTI, A., CABANNES, V., BOUCHACOURT, D., JEGOU, H. and BOTTOU, L. (2024). Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems* **36**.
- [4] CAO, Y., HE, Y., WU, D., CHEN, H.-Y., FAN, J. and LIU, H. (2025). Transformers simulate mle for sequence generation in bayesian networks. *arXiv preprint arXiv:2501.02547*.
- [5] CHEN, L., LU, K., RAJESWARAN, A., LEE, K., GROVER, A., LASKIN, M., ABBEEL, P., SRINIVAS, A. and MORDATCH, I. (2021). Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* **34** 15084–15097.
- [6] CHEN, S., SHEEN, H., WANG, T. and YANG, Z. (2024). Unveiling induction heads: Provable training dynamics and feature learning in transformers. *Advances in Neural Information Processing Systems* **37** 66479–66567.
- [7] D’ANGELO, F., CROCE, F. and FLAMMARION, N. (2025). Selective induction heads: How transformers select causal structures in context. In *The Thirteenth International Conference on Learning Representations*.
- [8] DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEGHANI, M., MINDERER, M., HEIGOLD, G., GELLY, S. ET AL. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- [9] EDELMAN, E., TSILIVIS, N., EDELMAN, B., MALACH, E. and GOEL, S. (2024). The evolution of statistical induction heads: In-context learning markov chains. *Advances in Neural Information Processing Systems* **37** 64273–64311.
- [10] HE, H. and SU, W. J. (2025). A law of next-token prediction in large language models. *Physical Review E* **112** 035317.
- [11] HUANG, Y., CHENG, Y. and LIANG, Y. (2024). In-context convergence of transformers. In *Forty-first International Conference on Machine Learning*.

- [12] ILDIZ, M. E., HUANG, Y., LI, Y., RAWAT, A. S. and OYMAK, S. (2024). From self-attention to markov models: Unveiling the dynamics of generative transformers. In *Forty-first International Conference on Machine Learning*.
- [13] JANNER, M., LI, Q. and LEVINE, S. (2021). Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems* **34** 1273–1286.
- [14] JELASSI, S., SANDER, M. and LI, Y. (2022). Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems* **35** 37822–37836.
- [15] JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A., POTAPENKO, A. ET AL. (2021). Highly accurate protein structure prediction with alphafold. *nature* **596** 583–589.
- [16] LI, Y., HUANG, Y., ILDIZ, M. E., RAWAT, A. S. and OYMAK, S. (2024). Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*. PMLR.
- [17] LI, Y., LI, Y. and RISTESKI, A. (2023). How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*. PMLR.
- [18] LI, Z., CAO, Y., GAO, C., HE, Y., LIU, H., JASON, K., FAN, J. and WANG, M. (2024). One-layer transformer provably learns one-nearest neighbor in context. In *Advances in Neural Information Processing Systems*.
- [19] LU, C., SHI, R., LIU, Y., HU, K., DU, S. S. and XU, H. (2024). Rethinking transformers in solving pomdps. In *Forty-first International Conference on Machine Learning*.
- [20] MAHANKALI, A. V., HASHIMOTO, T. and MA, T. (2024). One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*.
- [21] MAKKUVA, A. V., BONDASCHI, M., GIRISH, A., NAGLE, A., JAGGI, M., KIM, H. and GASTPAR, M. (2025). Attention with markov: A curious case of single-layer transformers. In *The Thirteenth International Conference on Learning Representations*.
- [22] NICHANI, E., DAMIAN, A. and LEE, J. D. (2024). How transformers learn causal structure with gradient descent. In *Forty-first International Conference on Machine Learning*.
- [23] RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., SUTSKEVER, I. ET AL. (2019). Language models are unsupervised multitask learners. *OpenAI blog* **1** 9.
- [24] RAO, Y., ZHAO, W., LIU, B., LU, J., ZHOU, J. and HSIEH, C.-J. (2021). Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems* **34** 13937–13949.
- [25] SVETE, A. and COTTERELL, R. (2024). Transformers can represent n-gram language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.
- [26] TARZANAGH, D. A., LI, Y., THRAMPOULIDIS, C. and OYMAK, S. (2023). Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*.
- [27] TARZANAGH, D. A., LI, Y., ZHANG, X. and OYMAK, S. (2023). Max-margin token selection in attention mechanism. *Advances in Neural Information Processing Systems* **36** 48314–48362.
- [28] THRAMPOULIDIS, C. (2024). Implicit optimization bias of next-token prediction in linear models. *Advances in Neural Information Processing Systems* **37** 22624–22656.
- [29] TIAN, Y., WANG, Y., CHEN, B. and DU, S. S. (2023). Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems* **36** 71911–71947.

- [30] TIAN, Y., WANG, Y., ZHANG, Z., CHEN, B. and DU, S. S. (2024). Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. In *The Twelfth International Conference on Learning Representations*.
- [31] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. and POLOSUKHIN, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* .
- [32] WANG, Z., WEI, S., HSU, D. and LEE, J. D. (2024). Transformers provably learn sparse token selection while fully-connected nets cannot. In *Forty-first International Conference on Machine Learning*.
- [33] ZHANG, C., MENG, X. and CAO, Y. (2024). Transformer learns optimal variable selection in group-sparse classification. In *The Thirteenth International Conference on Learning Representations*.
- [34] ZHANG, R., FREI, S. and BARTLETT, P. L. (2024). Trained transformers learn linear models in-context. *Journal of Machine Learning Research* **25** 1–55.

A Additional Related Work

In this section, we give an overview of some additional related works.

Token selection. Our work reveals that a one-layer transformer can learn to perform the optimal token selection by focusing on the direct parent in a random walk. A line of recent works has studied the token selection of the self-attention mechanism from different perspectives. [26, 27] propose an equivalence between the optimization dynamics of one self-attention layer and an SVM problem and prove the global convergence under certain assumptions. [16] shows that when training a self-attention layer, the priority in token selection is determined by a directed graph extracted from the training data. [32] demonstrates that transformer models can learn the sparse token selection task effectively while fully connected networks fail in the worst case. [18] shows that a self-attention layer can be trained to perform proper token selection so that the model acts as a one-nearest neighbor classifier in context.

Next-token prediction. [28] explores the implicit bias of next-token prediction employing a related SVM formulation. [19] demonstrates that transformers fail to solve the Partially Observable Markov Decision Processes problem (POMDP) even with sufficient data. [10] observes a phenomenon of next-token prediction in LLM that each layer contributes equally to enhancing the prediction accuracy. [29] studies the SGD training dynamics of a transformer with one self-attention layer and one decoder layer for next-token prediction, restricted to some specific assumptions like no positional encoding, long input sequences, and the fact that the decoder layer learns faster than the self-attention layer.

Training dynamics of transformers. [20, 34] investigate the training dynamics of in-context learning in transformers with a single self-attention layer trained through gradient flow on linear regression tasks. [11] solves in-context linear regression with the orthogonal input data by gradient descent on a single softmax attention layer. [14] demonstrates that the position-position block of a single attention layer in a vision transformer can encode spatial structure by dealing with a binary classification task. [30] delves into the training process of transformers with multi-layers by analyzing the dynamics of the MLP layers. [3] analyzes a synthetic in-context learning task and emphasizes the significance of weight matrices as associative memories. [1] shows incremental learning dynamics in transformers with diagonal attention matrices.

B Gradient Calculation

Recall that the population loss is

$$L(\theta) = \mathbb{E}[\ell(\theta)] = \mathbb{E}[-\log(e_y^\top f_\theta(\mathbf{X}) + \epsilon)] = \mathbb{E}[-\log(e_y^\top \mathbf{V} \mathbf{X} \mathcal{S}(\tilde{\mathbf{X}}^\top \mathbf{W} \tilde{\mathbf{x}}_N) + \epsilon)].$$

The following lemma calculates the gradients of the population function.

Lemma B.1. The gradients regarding $\mathbf{V}$ and $\mathbf{W}$ are

$$\begin{aligned} \nabla_{\mathbf{V}} \ell(\theta) &= \ell' \cdot e_y \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top = -\frac{1}{e_y^\top \mathbf{V} \mathbf{X} \mathcal{S} + \epsilon} \cdot e_y \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top, \\ \nabla_{\mathbf{W}} \ell(\theta) &= \ell' \cdot \begin{bmatrix} \mathbf{0} & \left(\sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i \mathbf{x}_i^\top - \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top \right) \mathbf{V}^\top e_y \mathbf{p}_N^\top \\ \mathbf{0} & \left(\sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{p}_i \mathbf{x}_i^\top - \sum_{i=1}^N \mathcal{S}_i \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top \right) \mathbf{V}^\top e_y \mathbf{p}_N^\top \end{bmatrix} \\ &= -\frac{1}{e_y^\top \mathbf{V} \mathbf{X} \mathcal{S} + \epsilon} \cdot \begin{bmatrix} \mathbf{0} & \left(\sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i \mathbf{x}_i^\top - \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top \right) \mathbf{V}^\top e_y \mathbf{p}_N^\top \\ \mathbf{0} & \left(\sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{p}_i \mathbf{x}_i^\top - \sum_{i=1}^N \mathcal{S}_i \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top \right) \mathbf{V}^\top e_y \mathbf{p}_N^\top \end{bmatrix}, \end{aligned}$$

where $\mathcal{S} = \mathcal{S}(\tilde{\mathbf{X}}^\top \mathbf{W} \tilde{\mathbf{x}}_N)$, and $\mathcal{S}_i$ is the i -th element of $\mathcal{S}$.

Proof of Lemma B.1. For $\mathbf{V}$, we have

$$\begin{aligned} \nabla_{\mathbf{V}} \ell(\theta) &= -\frac{1}{e_y^\top f_\theta(\mathbf{X}) + \epsilon} \cdot \frac{\partial e_y^\top \mathbf{V} \mathbf{X} \mathcal{S}}{\partial \mathbf{V}} \\ &= -\frac{1}{e_y^\top \mathbf{V} \mathbf{X} \mathcal{S} + \epsilon} \cdot e_y \mathcal{S}^\top \mathbf{X}^\top \end{aligned}$$

$$= -\frac{1}{e_y^\top \mathbf{V} \mathbf{X} \mathbf{S} + \epsilon} \cdot e_y \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top.$$

For $\mathbf{W}$, we have

$$\begin{aligned} \nabla_{\mathbf{W}} \ell(\theta) &= -\frac{1}{e_y^\top f_\theta(\mathbf{X}) + \epsilon} \cdot \frac{\partial e_y^\top \mathbf{V} \mathbf{X} \mathbf{S} (\widetilde{\mathbf{X}}^\top \mathbf{W} \tilde{\mathbf{x}}_N)}{\partial \mathbf{W}} \\ &= -\frac{1}{e_y^\top \mathbf{V} \mathbf{X} \mathbf{S} + \epsilon} \cdot \widetilde{\mathbf{X}} \mathbf{S}'(\widetilde{\mathbf{X}}^\top \mathbf{W} \tilde{\mathbf{x}}_N) \mathbf{X}^\top \mathbf{V}^\top e_y \tilde{\mathbf{x}}_N^\top \\ &= -\frac{1}{e_y^\top \mathbf{V} \mathbf{X} \mathbf{S} + \epsilon} \cdot \widetilde{\mathbf{X}} [\text{diag}(\mathbf{S}) - \mathbf{S} \mathbf{S}^\top] \mathbf{X}^\top \mathbf{V}^\top e_y \tilde{\mathbf{x}}_N^\top \\ &= -\frac{1}{e_y^\top \mathbf{V} \mathbf{X} \mathbf{S} + \epsilon} \cdot \begin{bmatrix} \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i \mathbf{x}_i^\top - \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top \\ \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{p}_i \mathbf{x}_i^\top - \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top \end{bmatrix} \cdot \begin{bmatrix} \mathbf{0} & \mathbf{V}^\top e_y \mathbf{p}_N^\top \end{bmatrix} \\ &= -\frac{1}{e_y^\top \mathbf{V} \mathbf{X} \mathbf{S} + \epsilon} \cdot \begin{bmatrix} \mathbf{0} & (\sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i \mathbf{x}_i^\top - \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top) \mathbf{V}^\top e_y \mathbf{p}_N^\top \\ \mathbf{0} & (\sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{p}_i \mathbf{x}_i^\top - \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top) \mathbf{V}^\top e_y \mathbf{p}_N^\top \end{bmatrix}, \end{aligned}$$

where we use the fact that $\mathbf{S}'(\widetilde{\mathbf{X}}^\top \mathbf{W} \tilde{\mathbf{x}}_N) = [\text{diag}(\mathbf{S}) - \mathbf{S} \mathbf{S}^\top]$ and

$$\text{diag}(\mathbf{S}) := \begin{bmatrix} \mathcal{S}_1 & & & \\ & \mathcal{S}_2 & & \\ & & \ddots & \\ & & & \mathcal{S}_N \end{bmatrix}.$$

□

To simplify the notation, we denote

$$\begin{aligned} \mathbf{A} &:= \left(\sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i \mathbf{x}_i^\top - \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top \right) \mathbf{V}^\top e_y \mathbf{p}_N^\top, \\ \mathbf{B} &:= \left(\sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{p}_i \mathbf{x}_i^\top - \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i \mathbf{x}_i^\top \right) \mathbf{V}^\top e_y \mathbf{p}_N^\top. \end{aligned}$$

With this notations, by Lemma B.1, we have

$$\nabla_{\mathbf{W}} \ell(\theta) = -\frac{1}{e_y^\top \mathbf{V} \mathbf{X} \mathbf{S} + \epsilon} \cdot \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}. \quad (\text{B.1})$$

We also denote $\mathbf{A}^{(t)}, \mathbf{B}^{(t)}$ the corresponding matrices at the t -th iteration of gradient descent.

Moreover, we can observe that $\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix}$, where $\mathbf{W}_{11} \in \mathbb{R}^{K \times K}$, $\mathbf{W}_{12} \in \mathbb{R}^{K \times M}$,

$\mathbf{W}_{21} \in \mathbb{R}^{M \times K}$, and $\mathbf{W}_{22} \in \mathbb{R}^{M \times M}$. By (B.1), we know that $\mathbf{W}_{11}^{(t)} = \mathbf{0}_{K \times K}$ and $\mathbf{W}_{21}^{(t)} = \mathbf{0}_{M \times K}$ for all $t \geq 1$.

By the definition of the random walks on circles, we can write the transition matrix $\mathbf{\Pi}$ as $\mathbf{\Pi} = p\mathbf{\Pi}_0^\top + (1-p)\mathbf{\Pi}_0$, where

$$\mathbf{\Pi}_0 = \begin{bmatrix} 0 & & & 1 \\ 1 & 0 & & \\ & 1 & 0 & \\ & & \ddots & \ddots \\ & & & 1 & 0 \end{bmatrix}.$$

C Proof of Theorem 3.1

In this section, we analyze random walks with the transition probability p satisfying $0 < p < 1$. Without loss of generality, we assume $\frac{1}{2} \leq p < 1$. We consider gradient descent starting from zero initialization. The following lemma presents the result of the first iteration.

Lemma C.1. Under the same conditions as Theorem 3.1, it holds that

$$\mathbf{V}^{(1)} = \frac{\eta}{\epsilon N K} \sum_{i=1}^{N-1} (\mathbf{\Pi}^\top)^{N-i} \text{ and } \mathbf{W}^{(1)} = \mathbf{0}_{(K+M) \times (K+M)}.$$

Proof of Lemma C.1. By Lemma B.1, we have

$$\begin{aligned} \mathbb{E}[\nabla_{\mathbf{V}} \ell(\theta^{(0)})] &= -\frac{1}{\epsilon N} \sum_{i=1}^{N-1} \mathbb{E}[e_y \mathbf{x}_i^\top] \\ &= -\frac{1}{\epsilon N} \sum_{i=1}^{N-1} \mathbb{E}[(\mathbf{\Pi}^\top)^{N-i} \mathbf{x}_i \mathbf{x}_i^\top] \\ &= -\frac{1}{\epsilon N K} \sum_{i=1}^{N-1} (\mathbf{\Pi}^\top)^{N-i} \end{aligned}$$

where the first equation is by the initialization of $\mathbf{V}^{(0)}$ and $\mathbf{W}^{(0)}$, the second equation is by the sampling method, and the third equation is by $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top] = \frac{1}{K} \mathbf{I}_K$ for $i \in [N-1]$ since $\mathbf{x}_i$ is uniformly distributed in $\mathbf{E}$. Thus, by the update, we can get

$$\begin{aligned} \mathbf{V}^{(1)} &= \mathbf{V}^{(0)} - \eta \mathbb{E}[\nabla_{\mathbf{V}} \ell(\theta^{(0)})] \\ &= \frac{\eta}{\epsilon N K} \sum_{i=1}^{N-1} (\mathbf{\Pi}^\top)^{N-i}. \end{aligned}$$

Since $\mathbf{V}^{(0)} = \mathbf{0}_{K \times K}$ and $\mathbf{W}^{(0)} = \mathbf{0}_{(K+M) \times (K+M)}$, we can get $\mathbb{E}[\nabla_{\mathbf{W}} \ell(\theta^{(0)})] = \mathbf{0}_{(K+M) \times (K+M)}$. Thus,

$$\mathbf{W}^{(1)} = \mathbf{W}^{(0)} - \eta \mathbb{E}[\nabla_{\mathbf{W}} \ell(\theta^{(0)})] = \mathbf{0}_{(K+M) \times (K+M)}.$$

□

Lemma C.1 gives explicit calculations of $\mathbf{V}^{(1)}$. Based on these, we can further derive some properties of $\mathbf{V}^{(1)}$, given in Lemma C.2 below. In this lemma, for all matrix indices, we consider simplified notations following the rule that if an index i is not in the set $\{1, \dots, K\}$, we treat it as $i' \in \{1, \dots, K\}$ with $i' \equiv i \pmod{K}$.

Lemma C.2. Under the same conditions as Theorem 3.1, it holds that $[\mathbf{V}^{(1)}]_{i_1, j_2} = [\mathbf{V}^{(1)}]_{i_2, j_2}$ for $i_1 - j_1 \equiv i_2 - j_2 \pmod{K}$.

Proof of Lemma C.2. We use induction to prove that for any $R \in \mathbb{N}$, $[\mathbf{\Pi}^R]_{i_1, j_2} = [\mathbf{\Pi}^R]_{i_2, j_2}$ for $i_1 - j_1 \equiv i_2 - j_2 \pmod{K}$. The result is obvious at $R = 1$. Suppose that the results hold for $\mathbf{\Pi}^R$. We aim to prove the results hold for $\mathbf{\Pi}^{R+1}$. By the definition of $\mathbf{\Pi}$, for any i, j , we obtain that

$$[\mathbf{\Pi}^{R+1}]_{i, j} = p \cdot [\mathbf{\Pi}^R]_{i, j-1} + (1-p) \cdot [\mathbf{\Pi}^R]_{i, j+1}. \quad (\text{C.1})$$

By induction, for any i_1, j_1, i_2, j_2 satisfying $i_1 - j_1 \equiv i_2 - j_2 \pmod{K}$, we have $[\mathbf{\Pi}^R]_{i_1, j_1-1} = [\mathbf{\Pi}^R]_{i_2, j_2-1}$ and $[\mathbf{\Pi}^R]_{i_1, j_1+1} = [\mathbf{\Pi}^R]_{i_2, j_2+1}$. Thus, by (C.1), we can easily get $[\mathbf{\Pi}^{R+1}]_{i_1, j_1} = [\mathbf{\Pi}^{R+1}]_{i_2, j_2}$, which completes the induction.

From Lemma C.1, we know $(\mathbf{V}^{(1)})^\top = \frac{\eta}{\epsilon N K} \sum_{i=1}^{N-1} \mathbf{\Pi}^{N-i}$ holds the result. Therefore, $\mathbf{V}^{(1)}$ also has the same property. □

Lemma C.3. Under the same conditions as Theorem 3.1, it holds that $[\mathbf{V}^{(1)}]_{2,1} = \|\mathbf{V}^{(1)}\|_{\max}$.

Proof of Lemma C.3. Without loss of generality, we assume $\frac{1}{2} \leq p < 1$. From Lemma C.1, we know $\mathbf{V}^{(1)} = \frac{\eta}{\epsilon N K} \sum_{i=1}^{N-1} (\mathbf{\Pi}^\top)^{N-i}$. We can observe that since $p < 1$, with increasing R , $(\mathbf{\Pi}^\top)^R$ will be closer to $\frac{1}{K} \mathbf{1} \mathbf{1}^\top$, as stated in Lemma E.2 and E.3. Thus, the location of the largest entries in $\mathbf{V}^{(1)}$ is mainly determined by the first several terms. In $\mathbf{\Pi}^\top$, there are two locations with non-zero

values $p, 1 - p$. In $(\mathbf{\Pi}^\top)^2$, there are three locations with non-zero values $p^2, 2p(1 - p), (1 - p)^2$. We can easily check that p is larger than p^2 and $2p(1 - p)$. Thus, we know the location of the value p in $\mathbf{\Pi}^\top$ is also the largest in $\mathbf{V}^{(1)}$, i.e. $[\mathbf{V}^{(1)}]_{2,1}$. In detail, we can first get that

$$\sum_{i=0}^{N-1} (\mathbf{\Pi}^\top)^{N-i} = (\mathbf{\Pi}^\top - (\mathbf{\Pi}^\top)^N)(\mathbf{I} - \mathbf{\Pi}^\top)^{-1}$$

matrix of interest:

$$\begin{aligned} \mathbf{e}_1^\top \sum_{i=0}^{N-2} (\mathbf{\Pi}^\top)^i (\mathbf{\Pi}^\top \mathbf{\Pi} \mathbf{e}_1). \\ \mathbf{e}_1^\top \sum_{i=0}^{N-2} (\mathbf{\Pi}^\top)^i (\mathbf{\Pi}^\top \mathbf{\Pi}^j \mathbf{e}_1). \end{aligned}$$

Denote

$$\mathbf{\Gamma}(N) = \sum_{i=0}^{N-2} (\mathbf{\Pi}^\top)^i.$$

Claim: for all N , the diagonal entries in $\mathbf{\Gamma}(N)$ are larger than or equal to the other entries.

$$\mathbf{\Gamma}(N) = \mathbf{\Pi}^\top \mathbf{\Gamma}(N-1) + \mathbf{I}.$$

Induction hypothesis:

$$\mathbf{\Gamma}(N)_{1,1} \geq \mathbf{\Gamma}(N)_{i,j} + c_p.$$

Suppose the induction hypothesis holds for $\mathbf{\Gamma}(N-1)$.

For i, j with $|i - j| \geq 1$,

$$\mathbf{\Gamma}(N)_{ij} = p \cdot \mathbf{\Gamma}(N-1)_{i,j-1} + (1-p) \cdot (\mathbf{\Gamma}(N-1))_{i,j+1} \leq (\mathbf{\Gamma}(N-1))_{1,1} + c_p.$$

$$\mathbf{\Gamma}(N)_{1,1} \geq \mathbf{\Gamma}(N-1)_{1,1}$$

$\sum_{i=1}^{N-1} (\mathbf{\Pi}^\top)^{N-i} = (\mathbf{\Pi}^\top - (\mathbf{\Pi}^\top)^N) C_p \sum_{k=0}^{\infty} \left(\frac{1-p}{p}\right)^k (\mathbf{\Pi}_0^\top)^k$, where C_p is a positive constant regarding p . Considering the terms $\mathbf{\Pi}^\top (\mathbf{\Pi}_0^\top)^k$ with $0 \leq k \leq K-1$, we can easily observe that the coefficient of $\mathbf{\Pi}^\top (\mathbf{\Pi}_0^\top)^0$ is much larger than that of other terms. Since $\mathbf{\Pi}_0^\top$ is a cyclic shift matrix and the entries in $(\mathbf{\Pi}^\top)^N$ are almost same (Lemma E.2 and E.3), the location of the largest entry in $\mathbf{\Pi}^\top$ is also that in $\mathbf{V}^{(1)}$. Therefore, it holds that $[\mathbf{V}^{(1)}]_{2,1} = \|\mathbf{V}^{(1)}\|_{\max}$. $\square$

Further, the following two lemmas provide some properties of the weights for the second iteration.

Lemma C.4. Under the same conditions as Theorem 3.1, it holds that $\|\mathbf{V}^{(2)}\|_{\max} \leq \frac{\eta}{\epsilon K} + 2\epsilon K^2$.

Proof of Lemma C.4. First, we have

$$\mathbf{V}^{(2)} = \mathbf{V}^{(1)} - \eta \mathbb{E}[\nabla_{\mathbf{V}} \ell(\theta^{(1)})] = \mathbf{V}^{(1)} + \eta \mathbb{E} \left[\frac{\mathbf{e}_y \sum_{i=1}^{N-1} \mathcal{S}_i^{(1)} \mathbf{x}_i^\top}{\mathbf{e}_y^\top \mathbf{V}^{(1)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(1)} \mathbf{x}_i + \epsilon} \right].$$

Thus,

$$\|\mathbf{V}^{(2)}\|_{\max} \leq \|\mathbf{V}^{(1)}\|_{\max} + \left\| \eta \mathbb{E} \left[\frac{\frac{1}{N} \mathbf{e}_y \sum_{i=1}^{N-1} \mathbf{x}_i^\top}{\frac{1}{N} \mathbf{e}_y^\top \mathbf{V}^{(1)} \sum_{i=1}^{N-1} \mathbf{x}_i + \epsilon} \right] \right\|_{\max}$$

$$\begin{aligned}
&\leq \left\| \mathbf{V}^{(1)} \right\|_{\max} + \frac{\eta}{\frac{N-1}{N} \min_{i,j} [\mathbf{V}^{(1)}]_{i,j}} \cdot \left\| \mathbb{E} \left[\frac{1}{N} \mathbf{e}_y \sum_{i=1}^{N-1} \mathbf{x}_i^\top \right] \right\|_{\max} \\
&\leq \left\| \frac{\eta}{\epsilon N K} \sum_{i=1}^{N-1} \mathbf{\Pi}^{N-i} \right\|_{\max} + \frac{\eta}{\frac{N-1}{N} \min_{i,j} \left[\frac{\eta}{\epsilon N K} \sum_{i=1}^{N-1} \mathbf{\Pi}^{N-i} \right]_{i,j}} \cdot \frac{N-1}{N} \\
&\leq \frac{\eta}{\epsilon K} + 2\epsilon K^2,
\end{aligned}$$

where the second inequality is by $\mathbf{e}_y^\top \mathbf{V}^{(1)} \mathbf{x}_i \geq \min_{i,j} [\mathbf{V}^{(1)}]_{i,j}$, and the last inequality is by Lemma E.2 and E.3. $\square$

Lemma C.5. Under the same conditions as Theorem 3.1, it holds that $\mathcal{S}_{N-1}^{(2)} \geq \mathcal{S}_j^{(2)} \exp(\Omega(N))$ for $j \neq N-1$. Further, $\mathcal{S}_{N-1}^{(2)} \geq 1 - \exp(-\Omega(N))$ and $\mathcal{S}_j^{(2)} \leq \exp(-\Omega(N))$ for $j \neq N-1$.

Proof of Lemma C.5. By Lemma B.1, we have

$$\begin{aligned}
&\mathbb{E}[\mathbf{A}^{(1)}] \\
&= \mathbb{E} \left[\left(\sum_{i=1}^{N-1} \mathcal{S}_i^{(1)} \mathbf{x}_i \mathbf{x}_i^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y - \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \mathcal{S}_{i_1}^{(1)} \mathcal{S}_{i_2}^{(1)} \mathbf{x}_{i_1} \mathbf{x}_{i_2}^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y \right) \mathbf{p}_N^\top \right] \\
&= \mathbb{E} \left[\left(\frac{\eta}{\epsilon N^2 K} \sum_{i=1}^{N-1} \mathbf{x}_i \mathbf{x}_i^\top \sum_{i'=1}^{N-1} \mathbf{\Pi}^{N-i'} \mathbf{e}_y - \frac{\eta}{\epsilon N^3 K} \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \mathbf{x}_{i_1} \mathbf{x}_{i_2}^\top \sum_{i'=1}^{N-1} \mathbf{\Pi}^{N-i'} \mathbf{e}_y \right) \mathbf{p}_N^\top \right] \\
&= \mathbb{E} \left[\frac{\eta}{\epsilon N^2 K} \sum_{i=1}^{N-1} \mathbf{x}_i \mathbf{x}_i^\top \sum_{i'=1}^{N-1} \mathbf{\Pi}^{N-i'} (\mathbf{\Pi}^\top)^{N-i} \mathbf{x}_i \cdot \mathbf{p}_N^\top \right] \\
&\quad - \mathbb{E} \left[\frac{\eta}{\epsilon N^3 K} \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \mathbf{x}_{i_1} \mathbf{x}_{i_1}^\top \mathbf{\Pi}^{i_2-i_1} \sum_{i'=1}^{N-1} \mathbf{\Pi}^{N-i'} (\mathbf{\Pi}^\top)^{N-i_1} \mathbf{x}_{i_1} \cdot \mathbf{p}_N^\top \right] \\
&= \mathbb{E} \left[\frac{\eta}{\epsilon N^2 K} \sum_{i=1}^{N-1} \sum_{i'=1}^{N-1} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{\Pi}^{N-i'} (\mathbf{\Pi}^\top)^{N-i} \mathbf{x}_i \right] \mathbf{p}_N^\top \\
&\quad - \mathbb{E} \left[\frac{\eta}{\epsilon N^3 K} \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \sum_{i'=1}^{N-1} \mathbf{x}_{i_1} \mathbf{x}_{i_1}^\top \mathbf{\Pi}^{N-i'+i_2-i_1} (\mathbf{\Pi}^\top)^{N-i_1} \mathbf{x}_{i_1} \cdot \mathbf{p}_N^\top \right] \\
&= \frac{\eta}{\epsilon N^2 K^2} \sum_{i=1}^{N-1} \sum_{i'=1}^{N-1} \text{tr} \left(\mathbf{\Pi}^{N-i'} (\mathbf{\Pi}^\top)^{N-i} \right) \mathbf{1}_K \mathbf{p}_N^\top \\
&\quad - \frac{\eta}{\epsilon N^3 K^2} \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \sum_{i'=1}^{N-1} \text{tr} \left(\mathbf{\Pi}^{N-i'+i_2-i_1} (\mathbf{\Pi}^\top)^{N-i_1} \right) \mathbf{1}_K \mathbf{p}_N^\top,
\end{aligned}$$

where the second equation is by Lemma C.1, the third equation is by the sampling method, and the fifth equation is by the fact that all the $\mathbf{x}_i$ is uniformly distributed in $\mathbf{E}$ for $i \in [N-1]$. Then, $\mathbf{W}_{12}^{(2)} = \mathbf{W}_{12}^{(1)} - \eta \mathbb{E}[\nabla_{\mathbf{W}} \ell(\theta^{(1)})]_{12} \propto \mathbf{1}_K \mathbf{p}_N^\top$. Thus, We also have

$$\begin{aligned}
&\mathbb{E}[\mathbf{B}^{(1)}] \\
&= \mathbb{E} \left[\left(\sum_{i=1}^{N-1} \mathcal{S}_i^{(1)} \mathbf{p}_i \mathbf{x}_i^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y - \sum_{i=1}^N \mathcal{S}_i^{(1)} \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i^{(1)} \mathbf{x}_i^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y \right) \mathbf{p}_N^\top \right] \\
&= \mathbb{E} \left[\left(\frac{\eta}{\epsilon N^2 K} \sum_{i=1}^{N-1} \mathbf{p}_i \mathbf{x}_i^\top \sum_{i'=1}^{N-1} \mathbf{\Pi}^{N-i'} \mathbf{e}_y - \frac{\eta}{\epsilon N^3 K} \sum_{i=1}^N \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathbf{x}_i^\top \sum_{i'=1}^{N-1} \mathbf{\Pi}^{N-i'} \mathbf{e}_y \right) \mathbf{p}_N^\top \right] \\
&= \mathbb{E} \left[\frac{\eta}{\epsilon N^2 K} \sum_{i=1}^{N-1} \mathbf{p}_i \mathbf{x}_i^\top \sum_{i'=1}^{N-1} \mathbf{\Pi}^{N-i'} (\mathbf{\Pi}^\top)^{N-i} \mathbf{x}_i \cdot \mathbf{p}_N^\top \right]
\end{aligned}$$

$$\begin{aligned}
& - \mathbb{E} \left[\frac{\eta}{\epsilon N^3 K} \sum_{i=1}^N \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathbf{x}_i^\top \sum_{i'=1}^{N-1} \mathbf{\Pi}^{N-i'} (\mathbf{\Pi}^\top)^{N-i} \mathbf{x}_i \cdot \mathbf{p}_N^\top \right] \\
& = \frac{\eta}{\epsilon N^2 K^2} \sum_{i=1}^{N-1} \sum_{i'=1}^{N-1} \mathbf{p}_i \operatorname{tr} \left(\mathbf{\Pi}^{N-i'} (\mathbf{\Pi}^\top)^{N-i} \right) \mathbf{p}_N^\top \\
& \quad - \frac{\eta}{\epsilon N^3 K^2} \sum_{i=1}^N \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \sum_{i'=1}^{N-1} \operatorname{tr} \left(\mathbf{\Pi}^{N-i'} (\mathbf{\Pi}^\top)^{N-i} \right) \mathbf{p}_N^\top,
\end{aligned}$$

where the second equation is by Lemma C.1, the third equation is by the sampling method, and the last equation is by the fact that all the $\mathbf{x}_i$ is uniformly distributed in $\mathbf{E}$ for $i \in [N-1]$. Since $[\widetilde{\mathbf{X}} \mathbf{W}^{(2)} \widetilde{\mathbf{x}}_N]_N = \mathbf{p}_N^\top \mathbf{W}_{22}^{(2)} \mathbf{p}_N$ and $[\widetilde{\mathbf{X}} \mathbf{W}^{(2)} \widetilde{\mathbf{x}}_N]_j = \mathbf{x}_j^\top \mathbf{W}_{12}^{(2)} \mathbf{p}_N + \mathbf{p}_j^\top \mathbf{W}_{22}^{(2)} \mathbf{p}_N$ for $j \in \{1, 2, \dots, N-1\}$, we can obtain that

$$\begin{aligned}
[\widetilde{\mathbf{X}} \mathbf{W}^{(2)} \widetilde{\mathbf{x}}_N]_N & = \mathbf{p}_N^\top \mathbf{W}_{22}^{(2)} \mathbf{p}_N \\
& = \mathbb{E} \left[\mathbf{p}_N^\top \frac{\eta}{\mathbf{e}_y^\top \mathbf{V} \mathbf{X} \mathbf{S} + \epsilon} \mathbf{B}^{(1)} \mathbf{p}_N \right] \\
& = \mathbb{E} \left[\frac{\eta}{\mathbf{e}_y^\top \mathbf{V} \mathbf{X} \mathbf{S} + \epsilon} \left(-\mathcal{S}_N^{(1)} \mathbf{p}_N^\top \mathbf{p}_N \cdot \sum_{i=1}^{N-1} \mathcal{S}_i^{(1)} \mathbf{x}_i^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y \right) \mathbf{p}_N^\top \mathbf{p}_N \right] \\
& < 0,
\end{aligned}$$

where the third equation is by $\mathbf{p}_i^\top \mathbf{p}_j = 0$ for $i \neq j$. And for $j \in \{1, 2, \dots, N-2\}$, we can get

$$\begin{aligned}
& [\widetilde{\mathbf{X}} \mathbf{W}^{(2)} \widetilde{\mathbf{x}}_N]_{N-1} - [\widetilde{\mathbf{X}} \mathbf{W}^{(2)} \widetilde{\mathbf{x}}_N]_j \\
& = \mathbf{x}_{N-1}^\top \mathbf{W}_{12}^{(2)} \mathbf{p}_N + \mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(2)} \mathbf{p}_N - \mathbf{x}_j^\top \mathbf{W}_{12}^{(2)} \mathbf{p}_N - \mathbf{p}_j^\top \mathbf{W}_{22}^{(2)} \mathbf{p}_N \\
& \stackrel{(i)}{=} \mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(2)} \mathbf{p}_N - \mathbf{p}_j^\top \mathbf{W}_{22}^{(2)} \mathbf{p}_N \\
& \stackrel{(ii)}{=} \mathbb{E} \left[\frac{\eta}{\mathbf{e}_y^\top \mathbf{V}^{(1)} \mathbf{X} \mathbf{S}^{(1)} + \epsilon} \cdot \frac{1}{N} \left(\mathbf{p}_{N-1}^\top \mathbf{p}_{N-1} \mathbf{x}_{N-1}^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y - \mathbf{p}_j^\top \mathbf{p}_j \mathbf{x}_j^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y \right) \mathbf{p}_N^\top \mathbf{p}_N \right] \\
& \stackrel{(iii)}{=} \mathbb{E} \left[\frac{\eta}{\mathbf{e}_y^\top \mathbf{V}^{(1)} \mathbf{X} \mathbf{S}^{(1)} + \epsilon} \cdot \frac{1}{N} \left(\mathbf{p}_{N-1}^\top \mathbf{p}_{N-1} \mathbf{x}_{N-1}^\top (\mathbf{V}^{(1)})^\top \mathbf{\Pi}^\top \mathbf{x}_{N-1} - \mathbf{p}_j^\top \mathbf{p}_j \mathbf{x}_j^\top (\mathbf{V}^{(1)})^\top (\mathbf{\Pi}^\top)^{N-j} \mathbf{x}_j \right) \mathbf{p}_N^\top \mathbf{p}_N \right] \\
& \stackrel{(iv)}{=} \mathbb{E} \left[\frac{\eta}{\mathbf{e}_y^\top \mathbf{V}^{(1)} \mathbf{X} \mathbf{S}^{(1)} + \epsilon} \cdot \frac{(\mathbf{p}_N^\top \mathbf{p}_N)^2}{N} \left(\left[(\mathbf{V}^{(1)})^\top \mathbf{\Pi}^\top \right]_{1,1} - \left[(\mathbf{V}^{(1)})^\top (\mathbf{\Pi}^\top)^{N-j} \right]_{1,1} \right) \right] \\
& \stackrel{(v)}{\geq} \frac{\eta}{\max_{i,j} [\mathbf{V}^{(1)}]_{i,j} + \epsilon} \cdot \frac{(\mathbf{p}_N^\top \mathbf{p}_N)^2}{N} \mathbb{E} \left[\left[(\mathbf{V}^{(1)})^\top \mathbf{\Pi}^\top \right]_{1,1} - \left[(\mathbf{V}^{(1)})^\top (\mathbf{\Pi}^\top)^{N-j} \right]_{1,1} \right] \\
& \stackrel{(vi)}{\geq} \frac{\eta}{\frac{\eta}{\epsilon K} - \epsilon} \cdot \frac{\eta (\mathbf{p}_N^\top \mathbf{p}_N)^2}{\epsilon N^2 K} C_p \\
& \geq \Omega \left(\frac{\eta M^2}{N^2} \right) \\
& \geq \Omega(N),
\end{aligned}$$

where (i) is by $\mathbf{W}_{12}^{(2)} \propto \mathbf{1}_K \mathbf{p}_N^\top$, (ii) is by $\mathbf{p}_i^\top \mathbf{p}_{i'} = 0$ for $i \neq i'$, (iii) is by the sampling methods, (iv) is by the fact that all the $\mathbf{x}_i$ is uniformly distributed in $\mathbf{E}$ for $i \in [N-1]$, (v) and (vi) are by Lemma E.4. Therefore, we have $\mathcal{S}_{N-1}^{(2)} / \mathcal{S}_j^{(2)} = \exp \left([\widetilde{\mathbf{X}} \mathbf{W}^{(2)} \widetilde{\mathbf{x}}_N]_{N-1} - [\widetilde{\mathbf{X}} \mathbf{W}^{(2)} \widetilde{\mathbf{x}}_N]_j \right) \geq \exp(\Omega(N))$ for $j \neq N-1$. Further,

$$\mathcal{S}_{N-1}^{(2)} = 1 - \sum_{j \neq N-1} \mathcal{S}_j^{(2)} \geq 1 - (N-1) \exp(-\Omega(N)) \mathcal{S}_{N-1}^{(2)},$$

which implies that

$$\mathcal{S}_{N-1}^{(2)} \geq \frac{1}{1 + (N-1)\exp(-\Omega(N))} = 1 - \frac{N-1}{\exp(\Omega(N)) + N-1} = 1 - \exp(-\Omega(N)).$$

Then, we have $\mathcal{S}_j^{(2)} \leq 1 - \mathcal{S}_{N-1}^{(2)} \leq \exp(-\Omega(N))$ for $j \neq N-1$. $\square$

Then, we can derive the bounds of entries in $\mathbf{V}^{(t)}$.

Lemma C.6. Under the same conditions as Theorem 3.1, it holds for $t \geq 3$ that

$$\min_{i,j} [\mathbf{V}^{(t)}]_{i,j} \geq \frac{\eta}{2\epsilon K^2} \text{ and } \|\mathbf{V}^{(t)}\|_{\max} \leq \frac{\eta}{\epsilon K} + (t-2) \cdot 2\epsilon K^2.$$

Proof of Lemma C.6. First, we have

$$\begin{aligned} \min_{i,j} [\mathbf{V}^{(t)}]_{i,j} &\geq \min_{i,j} [\mathbf{V}^{(1)}]_{i,j} \\ &\geq \min_{i,j} \left[\frac{\eta}{\epsilon N K} \sum_{i'=1}^{N-1} (\mathbf{\Pi}^\top)^{N-i'} \right]_{i,j} \\ &\geq \frac{\eta}{\epsilon N K} \cdot \frac{N}{2K} \\ &= \frac{\eta}{2\epsilon K^2}, \end{aligned}$$

where the third inequality is by Lemma E.2 and E.3. Then, we can get that

$$\begin{aligned} \|\mathbf{V}^{(t)}\|_{\max} &\leq \|\mathbf{V}^{(t-1)}\|_{\max} + \left\| \mathbb{E} \left[\frac{\eta \mathbf{e}_y \sum_{i=1}^{N-1} \mathcal{S}_i^{(t-1)} \mathbf{x}_i^\top}{\mathbf{e}_y^\top \mathbf{V}^{(t-1)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t-1)} \mathbf{x}_i + \epsilon} \right] \right\|_{\max} \\ &\leq \|\mathbf{V}^{(t-1)}\|_{\max} + \mathbb{E} \left[\frac{\eta \left\| \mathbf{e}_y \sum_{i=1}^{N-1} \mathcal{S}_i^{(t-1)} \mathbf{x}_i^\top \right\|_{\max}}{\min \left[\mathbf{e}_y^\top \mathbf{V}^{(t-1)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t-1)} \mathbf{x}_i \right]} \right] \\ &\leq \|\mathbf{V}^{(t-1)}\|_{\max} + \frac{\eta}{\min_{i,j} [\mathbf{V}^{(t-1)}]_{i,j}} \\ &\leq \|\mathbf{V}^{(t-1)}\|_{\max} + 2\epsilon K^2 \\ &\leq \|\mathbf{V}^{(2)}\|_{\max} + (t-3) \cdot 2\epsilon K^2 \\ &\leq \frac{\eta}{\epsilon K} + (t-2) \cdot 2\epsilon K^2, \end{aligned}$$

where the third inequality is by $\mathbf{e}_y^\top \mathbf{V} \mathbf{x}_i \geq \min_{i,j} [\mathbf{V}]_{i,j}$, and the last inequality is by Lemma C.4. $\square$

Next, we analyze the training dynamics over multiple iterations.

Lemma C.7. Assume the same conditions as Theorem 3.1. For $2 \leq t \leq T^*$, it holds that $\mathcal{S}_{N-1}^{(t)} \geq 1 - \exp(-\Omega(N))$ and $\mathbf{V}^{(t)} = \beta^{(t)} \mathbf{\Pi}^\top + \tilde{\mathbf{V}}^{(t)}$ where $\|\tilde{\mathbf{V}}^{(t)}\|_{\max} \leq \gamma^{(t)}$. Here, $\beta^{(t)} \geq \sqrt{\eta t} - \frac{2\eta}{\epsilon K}$ and $\gamma^{(t)} \leq \frac{2\eta}{\epsilon K} + 2(t-1)\epsilon K^2 N \exp(-\Omega(N))$.

Proof of Lemma C.7. We use induction to prove the results that

$$\begin{aligned} \beta^{(t)} &\geq \sqrt{\eta t} - \frac{2\eta}{\epsilon K}, \\ \gamma^{(t)} &\leq \frac{2\eta}{\epsilon K} + 2(t-1)\epsilon K^2 N \exp(-\Omega(N)), \\ \left[\tilde{\mathbf{X}} \mathbf{W}^{(t)} \tilde{\mathbf{x}}_N \right]_{N-1} - \left[\tilde{\mathbf{X}} \mathbf{W}^{(t)} \tilde{\mathbf{x}}_N \right]_j &\geq \Omega(N), \\ \mathcal{S}_{N-1}^{(t)} &\geq 1 - \exp(-\Omega(N)). \end{aligned}$$

It can be easily checked that the results hold for $t = 2$. Suppose that the results hold for the t -th iteration. We aim to prove that the results hold for $t + 1$.

For $\mathbf{V}^{(t+1)}$, we can get

$$\begin{aligned}
\mathbf{V}^{(t+1)} &= \mathbf{V}^{(t)} - \eta \mathbb{E}[\nabla_{\mathbf{V}} \ell(\theta^{(t)})] \\
&= \mathbf{V}^{(t)} + \eta \mathbb{E} \left[\frac{\mathbf{e}_y \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i^\top}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\
&= \mathbf{V}^{(t)} + \mathbb{E} \left[\frac{\eta \mathcal{S}_{N-1}^{(t)} \mathbf{e}_y \mathbf{x}_{N-1}^\top}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] + \mathbb{E} \left[\frac{\eta \mathbf{e}_y \sum_{i=1}^{N-2} \mathcal{S}_i^{(t)} \mathbf{x}_i^\top}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\
&= \mathbf{V}^{(t)} + \mathbb{E} \left[\frac{\eta \mathcal{S}_{N-1}^{(t)} \mathbf{x}_{N-1} \mathbf{x}_{N-1}^\top}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \mathbf{\Pi}^\top + \mathbb{E} \left[\frac{\eta \mathbf{e}_y \sum_{i=1}^{N-2} \mathcal{S}_i^{(t)} \mathbf{x}_i^\top}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right]
\end{aligned}$$

Then, we have

$$\begin{aligned}
\left[\mathbb{E} \left[\frac{\eta \mathcal{S}_{N-1}^{(t)} \mathbf{x}_{N-1} \mathbf{x}_{N-1}^\top}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \right]_{1,1} &\geq \frac{\eta \mathcal{S}_{N-1}^{(t)}}{\|\mathbf{V}^{(t)}\|_{\max} + \epsilon} \\
&\geq \frac{\eta [1 - \exp(-\Omega(N))]}{\beta^{(t)} + \gamma^{(t)} + \epsilon} \\
&\geq \frac{\eta}{2 [\beta^{(t)} + \frac{\eta}{\epsilon K} + 2\epsilon K^2 + 2t\epsilon K^2 N \exp(-\Omega(N)) + \epsilon]} \\
&\geq \frac{\eta}{2 (\beta^{(t)} + \frac{2\eta}{\epsilon K})}, \tag{C.2}
\end{aligned}$$

where the first inequality is by $\mathbf{e}_y^\top \mathbf{V}^{(t)} \mathbf{x}_i \leq \|\mathbf{V}^{(t)}\|_{\max}$, the second inequality is by induction, and the third inequality is by the assumption of ϵ . And, we have

$$\begin{aligned}
\left\| \mathbb{E} \left[\frac{\eta \mathbf{e}_y \sum_{i=1}^{N-2} \mathcal{S}_i^{(t)} \mathbf{x}_i^\top}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \right\|_{\max} &\leq \frac{\eta \exp(-\Omega(N))}{\min_{i,j} [\mathbf{V}^{(t)}]_{i,j}} \cdot \left\| \mathbb{E} \left[\sum_{i=1}^{N-2} \mathbf{e}_y \mathbf{x}_i^\top \right] \right\|_{\max} \\
&\leq 2\epsilon K^2 \exp(-\Omega(N)) \cdot N, \tag{C.3}
\end{aligned}$$

where the first inequality is by induction and $\mathbf{e}_y^\top \mathbf{V}^{(t)} \mathbf{x}_i \geq \min_{i,j} [\mathbf{V}^{(t)}]_{i,j}$, and the second inequality is by Lemma C.6. Thus, we can get that

$$\begin{aligned}
\beta^{(t+1)} &\geq \beta^{(t)} + \left[\mathbb{E} \left[\frac{\eta \mathcal{S}_{N-1}^{(t)} \mathbf{x}_{N-1} \mathbf{x}_{N-1}^\top}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \right]_{1,1} \\
&\geq \beta^{(t)} + \frac{\eta}{2\beta^{(t)} + \frac{4\eta}{\epsilon K}} \\
&\geq \sqrt{\eta t} - \frac{2\eta}{\epsilon K} + \frac{\eta}{2\sqrt{\eta t}} \\
&\geq \sqrt{\eta t} - \frac{2\eta}{\epsilon K} + \frac{\sqrt{\eta}}{\sqrt{t+1} + \sqrt{t}} \\
&= \sqrt{\eta(t+1)} - \frac{2\eta}{\epsilon K}
\end{aligned}$$

where the second inequality is by (C.2), and the third inequality is by induction and the fact that $x + \frac{\eta}{2x + \frac{4\eta}{\epsilon K}}$ is monotonically increasing for $x \geq \frac{\sqrt{\eta}}{\sqrt{2}} - \frac{2\eta}{\epsilon K}$. And, we can get

$$\begin{aligned}
\gamma^{(t+1)} &\leq \gamma^{(t)} + \left\| \mathbb{E} \left[\frac{\eta \mathbf{e}_y \sum_{i=1}^{N-2} \mathcal{S}_i^{(t)} \mathbf{x}_i^\top}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \right\|_{\max} \\
&\leq \gamma^{(t)} + 2\epsilon K^2 N \exp(-\Omega(N))
\end{aligned}$$

$$\leq \frac{2\eta}{\epsilon K} + 2t\epsilon K^2 N \exp(-\Omega(N)),$$

where the second inequality is by (C.3), and the third inequality is by induction.

Next, we consider $\mathcal{S}^{(t+1)}$. Recall that

$$\mathbf{W}_{12}^{(t+1)} = \mathbf{W}_{12}^{(t)} + \eta \mathbb{E} \left[\frac{\mathbf{A}^{(t)}}{\mathbf{e}_y^\top \mathbf{V} \mathbf{X} \mathcal{S} + \epsilon} \right] \text{ and } \mathbf{W}_{22}^{(t+1)} = \mathbf{W}_{22}^{(t)} + \eta \mathbb{E} \left[\frac{\mathbf{B}^{(t)}}{\mathbf{e}_y^\top \mathbf{V} \mathbf{X} \mathcal{S} + \epsilon} \right],$$

where

$$\begin{aligned} \mathbf{A}^{(t)} &= \left(\sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i \mathbf{x}_i^\top (\mathbf{V}^{(t)})^\top \mathbf{e}_y - \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \mathcal{S}_{i_1}^{(t)} \mathcal{S}_{i_2}^{(t)} \mathbf{x}_{i_1} \mathbf{x}_{i_2}^\top (\mathbf{V}^{(t)})^\top \mathbf{e}_y \right) \mathbf{p}_N^\top, \\ \mathbf{B}^{(t)} &= \left(\sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{p}_i \mathbf{x}_i^\top (\mathbf{V}^{(t)})^\top \mathbf{e}_y - \sum_{i=1}^N \mathcal{S}_i^{(t)} \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i^\top (\mathbf{V}^{(t)})^\top \mathbf{e}_y \right) \mathbf{p}_N^\top. \end{aligned}$$

We also have $\left[\widetilde{\mathbf{X}} \mathbf{W}^{(t)} \widetilde{\mathbf{x}}_N \right]_N = \mathbf{p}_N^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N$ and $\left[\widetilde{\mathbf{X}} \mathbf{W}^{(t)} \widetilde{\mathbf{x}}_N \right]_j = \mathbf{x}_j^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N + \mathbf{p}_j^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N$ for $j \in \{1, 2, \dots, N-1\}$. Then, for $j = N$, we have

$$\begin{aligned} \mathbf{p}_N^\top \mathbf{W}_{22}^{(t+1)} \mathbf{p}_N &= \mathbf{p}_N^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N + \eta \mathbb{E} \left[\frac{\mathbf{p}_N^\top \mathbf{B}^{(t)} \mathbf{p}_N}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\ &= \mathbf{p}_N^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N - \eta M \mathcal{S}_N^{(t)} \mathbb{E} \left[\frac{\sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i^\top \mathbf{V}^{(t)} \mathbf{e}_y}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\ &\leq \mathbf{p}_N^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N. \end{aligned}$$

For $j \in \{1, 2, \dots, N-2\}$, we have

$$\begin{aligned} \mathbf{x}_j^\top \mathbf{W}_{12}^{(t+1)} \mathbf{p}_N &= \mathbf{x}_j^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N + \eta \mathbb{E} \left[\frac{\mathbf{x}_j^\top \mathbf{A}^{(t)} \mathbf{p}_N}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\ &= \mathbf{x}_j^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N + \eta \sqrt{M} \mathcal{S}_j^{(t)} \mathbb{E} \left[\frac{\mathbf{x}_j^\top \mathbf{V}^{(t)} \mathbf{e}_y - \sum_{i_2=1}^{N-1} \mathcal{S}_{i_2}^{(t)} \mathbf{x}_{i_2}^\top \mathbf{V}^{(t)} \mathbf{e}_y}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\ &\leq \mathbf{x}_j^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N + \eta \sqrt{M} \mathcal{S}_j^{(t)} \frac{\|\mathbf{V}^{(t)}\|_{\max}}{\min_{i,j} [\mathbf{V}^{(t)}]_{i,j}} \\ &\leq \mathbf{x}_j^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N + \eta \sqrt{M} \exp(-\Omega(N)) \left(2K + \frac{4(t-2)\epsilon^2 K^4}{\eta} \right), \end{aligned}$$

where the first inequality is by $\mathbf{e}_y^\top \mathbf{V}^{(t)} \mathbf{x}_{N-1} = \|\mathbf{V}^{(t)}\|_{\max}$, and the second inequality is by induction and Lemma C.6. And,

$$\begin{aligned} \mathbf{p}_j^\top \mathbf{W}_{22}^{(t+1)} \mathbf{p}_N &= \mathbf{p}_j^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N + \eta \mathbb{E} \left[\frac{\mathbf{p}_j^\top \mathbf{B}^{(t)} \mathbf{p}_N}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\ &= \mathbf{p}_j^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N + \eta M \mathcal{S}_j^{(t)} \mathbb{E} \left[\frac{\mathbf{x}_j^\top \mathbf{V}^{(t)} \mathbf{e}_y - \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i^\top \mathbf{V}^{(t)} \mathbf{e}_y}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\ &\leq \mathbf{p}_j^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N + \eta M \mathcal{S}_j^{(t)} \frac{\|\mathbf{V}^{(t)}\|_{\max}}{\min_{i,j} [\mathbf{V}^{(t)}]_{i,j}} \\ &\leq \mathbf{p}_j^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N + \eta M \exp(-\Omega(N)) \left(2K + \frac{4(t-2)\epsilon^2 K^4}{\eta} \right), \end{aligned}$$

where the first inequality is by $\mathbf{e}_y^\top \mathbf{V}^{(t)} \mathbf{x}_{N-1} = \|\mathbf{V}^{(t)}\|_{\max}$, and the second inequality is by induction and Lemma C.6. For $j = N-1$, we have

$$\mathbf{x}_{N-1}^\top \mathbf{W}_{12}^{(t+1)} \mathbf{p}_N = \mathbf{x}_{N-1}^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N + \eta \mathbb{E} \left[\frac{\mathbf{x}_{N-1}^\top \mathbf{A}^{(t)} \mathbf{p}_N}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right]$$

$$\begin{aligned}
&= \mathbf{x}_{N-1}^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N + \eta \sqrt{M} \mathcal{S}_{N-1}^{(t)} \mathbb{E} \left[\frac{\mathbf{x}_{N-1}^\top \mathbf{V}^{(t)} \mathbf{e}_y - \sum_{i_2=1}^{N-1} \mathcal{S}_{i_2}^{(t)} \mathbf{x}_{i_2}^\top \mathbf{V}^{(t)} \mathbf{e}_y}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\
&\geq \mathbf{x}_{N-1}^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N + \eta \sqrt{M} \mathbb{E} \left[\frac{-\sum_{i_2=1}^{N-2} \mathcal{S}_{i_2}^{(t)} \mathbf{x}_{i_2}^\top \mathbf{V}^{(t)} \mathbf{e}_y}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\
&\geq \mathbf{x}_{N-1}^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N - \eta \sqrt{M} \frac{\sum_{i_2=1}^{N-2} \mathcal{S}_{i_2}^{(t)} \|\mathbf{V}^{(t)}\|_{\max}}{\min_{i,j} [\mathbf{V}^{(t)}]_{i,j}} \\
&\geq \mathbf{x}_{N-1}^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N - \eta \sqrt{M} N \exp(-\Omega(N)) \left(2K + \frac{4(t-2)\epsilon^2 K^4}{\eta} \right),
\end{aligned}$$

where the second inequality is by $\mathbf{e}_y^\top \mathbf{V}^{(t)} \mathbf{x}_{N-1} = \|\mathbf{V}^{(t)}\|_{\max}$, and the third inequality is by induction and Lemma C.6. And,

$$\begin{aligned}
\mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(t+1)} \mathbf{p}_N &= \mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N + \eta \mathbb{E} \left[\frac{\mathbf{p}_{N-1}^\top \mathbf{B}^{(t)} \mathbf{p}_N}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\
&= \mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N + \eta M \mathcal{S}_{N-1}^{(t)} \mathbb{E} \left[\frac{\mathbf{x}_{N-1}^\top \mathbf{V}^{(t)} \mathbf{e}_y - \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i^\top \mathbf{V}^{(t)} \mathbf{e}_y}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\
&\geq \mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N + \eta M \mathbb{E} \left[\frac{-\sum_{i=1}^{N-2} \mathcal{S}_i^{(t)} \mathbf{x}_i^\top \mathbf{V}^{(t)} \mathbf{e}_y}{\mathbf{e}_y^\top \mathbf{V}^{(t)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(t)} \mathbf{x}_i + \epsilon} \right] \\
&\geq \mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N - \eta M \frac{\sum_{i_2=1}^{N-2} \mathcal{S}_{i_2}^{(t)} \|\mathbf{V}^{(t)}\|_{\max}}{\min_{i,j} [\mathbf{V}^{(t)}]_{i,j}} \\
&\geq \mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N - \eta M N \exp(-\Omega(N)) \left(2K + \frac{4(t-2)\epsilon^2 K^4}{\eta} \right),
\end{aligned}$$

where the second inequality is by $\mathbf{e}_y^\top \mathbf{V}^{(t)} \mathbf{x}_{N-1} = \|\mathbf{V}^{(t)}\|_{\max}$, and the third inequality is by induction and Lemma C.6. Therefore, we can get that for $j \neq N-1$,

$$\begin{aligned}
&\left[\widetilde{\mathbf{X}} \mathbf{W}^{(t+1)} \widetilde{\mathbf{x}}_N \right]_{N-1} - \left[\widetilde{\mathbf{X}} \mathbf{W}^{(t+1)} \widetilde{\mathbf{x}}_N \right]_j \\
&= \mathbf{x}_{N-1}^\top \mathbf{W}_{12}^{(t+1)} \mathbf{p}_N + \mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(t+1)} \mathbf{p}_N - \mathbf{x}_j^\top \mathbf{W}_{12}^{(t+1)} \mathbf{p}_N - \mathbf{p}_j^\top \mathbf{W}_{22}^{(t+1)} \mathbf{p}_N \\
&\geq \mathbf{x}_{N-1}^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N + \mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N - \mathbf{x}_j^\top \mathbf{W}_{12}^{(t)} \mathbf{p}_N - \mathbf{p}_j^\top \mathbf{W}_{22}^{(t)} \mathbf{p}_N \\
&\quad - 4\eta M N \exp(-\Omega(N)) \left(2K + \frac{4(t-2)\epsilon^2 K^4}{\eta} \right) \\
&= \left[\widetilde{\mathbf{X}} \mathbf{W}^{(t)} \widetilde{\mathbf{x}}_N \right]_{N-1} - \left[\widetilde{\mathbf{X}} \mathbf{W}^{(t)} \widetilde{\mathbf{x}}_N \right]_j - \exp(-\Omega(N)) \\
&\geq \Omega(N),
\end{aligned}$$

where the last inequality is by induction. Thus, $\mathcal{S}_{N-1}^{(t+1)}/\mathcal{S}_j^{(t+1)} \geq \exp(\Omega(N))$ for $j \neq N-1$, which implies that $\mathcal{S}_{N-1}^{(t+1)} \geq 1 - \exp(-\Omega(N))$. Therefore, we prove that the results hold for $t+1$, which completes the proof. $\square$

The next two lemmas show the convergence rates of $\mathbf{V}^{(T)}/\|\mathbf{V}^{(T)}\|_F$ and $f_{\theta^T}(\mathbf{X})/\|f_{\theta^T}(\mathbf{X})\|_2$.

Lemma C.8. Assume the same conditions as Theorem 3.1. For $\Omega(\eta\epsilon^{-2}K^{-2}) \leq T \leq T^*$, it holds that

$$\left\| \frac{\mathbf{V}^{(T)}}{\|\mathbf{V}^{(T)}\|_F} - \frac{\mathbf{\Pi}^\top}{\|\mathbf{\Pi}^\top\|_F} \right\|_F \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Proof of Lemma C.8. By Lemma C.7, we can get that

$$\left\| \frac{\mathbf{V}^{(T)}}{\|\mathbf{V}^{(T)}\|_F} - \frac{\mathbf{\Pi}^\top}{\|\mathbf{\Pi}^\top\|_F} \right\|_F = \left\| \frac{\beta^{(T)} \mathbf{\Pi}^\top + \widetilde{\mathbf{V}}^{(T)}}{\|\mathbf{V}^{(T)}\|_F} - \frac{\mathbf{\Pi}^\top}{\|\mathbf{\Pi}^\top\|_F} \right\|_F$$

$$\leq \left\| \left(\frac{\beta^{(T)}}{\|\mathbf{V}^{(T)}\|_F} - \frac{1}{\|\mathbf{\Pi}^\top\|_F} \right) \mathbf{\Pi}^\top \right\|_F + \left\| \frac{\tilde{\mathbf{V}}^{(T)}}{\|\mathbf{V}^{(T)}\|_F} \right\|_F.$$

For the first part, we have

$$\begin{aligned} \left\| \left(\frac{\beta^{(T)}}{\|\mathbf{V}^{(T)}\|_F} - \frac{1}{\|\mathbf{\Pi}^\top\|_F} \right) \mathbf{\Pi}^\top \right\|_F &= \left| \frac{\beta^{(T)} \|\mathbf{\Pi}^\top\|_F}{\|\beta^{(T)} \mathbf{\Pi}^\top + \tilde{\mathbf{V}}^{(T)}\|_F} - 1 \right| \\ &\leq 1 - \frac{\beta^{(T)} \|\mathbf{\Pi}^\top\|_F}{\beta^{(T)} \|\mathbf{\Pi}^\top\|_F + \|\tilde{\mathbf{V}}^{(T)}\|_F} \\ &\stackrel{(i)}{=} \frac{\|\tilde{\mathbf{V}}^{(T)}\|_F}{\sqrt{K(p^2 + (1-p)^2)\beta^{(T)} + \|\tilde{\mathbf{V}}^{(T)}\|_F}} \\ &\leq \frac{K\gamma^{(T)}}{\sqrt{K(p^2 + (1-p)^2)\beta^{(T)} + K\gamma^{(T)}}} \\ &\stackrel{(ii)}{\leq} \frac{\frac{2\eta}{\epsilon} + 2T\epsilon K^3 N \exp(-\Omega(N))}{\frac{\sqrt{2K}}{2} (\sqrt{\eta T} - \frac{2\eta}{\epsilon K}) + \frac{2\eta}{\epsilon} + 2T\epsilon K^3 N \exp(-\Omega(N))} \\ &\leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \end{aligned}$$

where (i) is by $\|\mathbf{\Pi}^\top\|_F = \sqrt{K(p^2 + (1-p)^2)}$, and (ii) is by Lemma C.7. For the second part, we have

$$\begin{aligned} \left\| \frac{\tilde{\mathbf{V}}^{(T)}}{\|\mathbf{V}^{(T)}\|_F} \right\|_F &= \frac{\|\tilde{\mathbf{V}}^{(T)}\|_F}{\|\beta^{(T)} \mathbf{\Pi}^\top + \tilde{\mathbf{V}}^{(T)}\|_F} \\ &\leq \frac{\|\tilde{\mathbf{V}}^{(T)}\|_F}{\|\beta^{(T)} \mathbf{\Pi}^\top\|_F + \|\tilde{\mathbf{V}}^{(T)}\|_F} \\ &\stackrel{(i)}{\leq} \frac{K\gamma^{(T)}}{\frac{\sqrt{2K}}{2} \beta^{(T)} + K\gamma^{(T)}} \\ &\stackrel{(ii)}{\leq} \frac{\frac{2\eta}{\epsilon} + 2T\epsilon K^3 N \exp(-\Omega(N))}{\frac{\sqrt{2K}}{2} (\sqrt{\eta T} - \frac{2\eta}{\epsilon K}) + \frac{2\eta}{\epsilon} + 2T\epsilon K^3 N \exp(-\Omega(N))} \\ &\leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \end{aligned}$$

where (i) is by $\|\mathbf{\Pi}^\top\|_F = \sqrt{K(p^2 + (1-p)^2)} \geq \sqrt{2K}/2$, and (ii) is by Lemma C.7. Therefore, we can obtain that

$$\left\| \frac{\mathbf{V}^{(T)}}{\|\mathbf{V}^{(T)}\|_F} - \frac{\mathbf{\Pi}^\top}{\|\mathbf{\Pi}^\top\|_F} \right\|_F \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

□

Lemma C.9. Assume the same conditions as Theorem 3.1. For $\Omega(\eta\epsilon^{-2}K^{-2}) \leq T \leq T^*$, it holds that

$$\left\| \frac{f_{\theta^{(T)}}(\mathbf{X})}{\|f_{\theta^{(T)}}(\mathbf{X})\|_2} - \mathbf{\Pi}^\top \mathbf{x}_{N-1} \right\|_2 \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Proof of Lemma C.9. The output with $\theta = \theta^{(T)}$ is $f_{\theta^{(T)}}(\mathbf{X}) = \mathbf{V}^{(T)} \mathbf{X} \mathcal{S}(\tilde{\mathbf{X}}^\top \mathbf{W}^{(T)} \tilde{\mathbf{x}}_N) = \mathbf{V}^{(T)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i$. Then, we can get that

$$\left\| \frac{f_{\theta^{(T)}}(\mathbf{X})}{\|f_{\theta^{(T)}}(\mathbf{X})\|_2} - \mathbf{\Pi}^\top \mathbf{x}_{N-1} \right\|_2$$

$$\begin{aligned}
&= \left\| \frac{\left(\beta^{(T)} \mathbf{\Pi}^\top + \tilde{\mathbf{V}}^{(T)} \right) \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i}{\left\| \mathbf{V}^{(T)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i \right\|_2} - \mathbf{\Pi}^\top \mathbf{x}_{N-1} \right\|_2 \\
&\leq \left\| \left(\frac{\beta^{(T)} \mathcal{S}_{N-1}^{(T)}}{\left\| \mathbf{V}^{(T)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i \right\|_2} - 1 \right) \mathbf{\Pi}^\top \mathbf{x}_{N-1} \right\|_2 \\
&\quad + \left\| \frac{\beta^{(T)} \mathbf{\Pi}^\top \sum_{i=1}^{N-2} \mathcal{S}_i^{(T)} \mathbf{x}_i + \tilde{\mathbf{V}}^{(T)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i}{\left\| \mathbf{V}^{(T)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i \right\|_2} \right\|_2.
\end{aligned}$$

For the first part, we have

$$\begin{aligned}
&\left\| \left(\frac{\beta^{(T)} \mathcal{S}_{N-1}^{(T)}}{\left\| \mathbf{V}^{(T)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i \right\|_2} - 1 \right) \mathbf{\Pi}^\top \mathbf{x}_{N-1} \right\|_2 \\
&\leq \left\| \left(1 - \frac{\beta^{(T)} \mathcal{S}_{N-1}^{(T)}}{\left\| \beta^{(T)} \mathbf{\Pi}^\top \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i \right\|_2 + \left\| \tilde{\mathbf{V}}^{(T)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i \right\|_2} \right) \mathbf{\Pi}^\top \mathbf{x}_{N-1} \right\|_2 \\
&\leq \left\| \left(1 - \frac{\beta^{(T)} [1 - \exp(-\Omega(N))]}{\frac{\sqrt{2}}{2} \beta^{(T)} + \sqrt{K} \gamma^{(T)}} \right) \mathbf{\Pi}^\top \mathbf{x}_{N-1} \right\|_2 \\
&\leq \left(1 - \frac{(\sqrt{\eta T} - \frac{2\eta}{\epsilon K}) [1 - \exp(-\Omega(N))]}{(\sqrt{\eta T} - \frac{2\eta}{\epsilon K}) + \sqrt{K} (\frac{2\eta}{\epsilon K} + 2T\epsilon K^2 N \exp(-\Omega(N)))} \right) \cdot \sqrt{p^2 + (1-p)^2} \\
&\leq \mathcal{O} \left(\frac{\sqrt{K} (\frac{2\eta}{\epsilon K} + 2T\epsilon K^2 N \exp(-\Omega(N)))}{\sqrt{\eta T} + 2T\epsilon K^2 N \exp(-\Omega(N))} \right) \\
&\leq \mathcal{O} \left(\frac{1}{\sqrt{T}} \right),
\end{aligned}$$

where the first inequality is by Lemma C.7, the second inequality is by Lemma C.7 and $\|\mathbf{\Pi}^\top \mathbf{x}_i\|_2 = \sqrt{p^2 + (1-p)^2}$, and the third inequality is by Lemma C.7. For the second part, we have

$$\begin{aligned}
&\left\| \frac{\beta^{(T)} \mathbf{\Pi}^\top \sum_{i=1}^{N-2} \mathcal{S}_i^{(T)} \mathbf{x}_i + \tilde{\mathbf{V}}^{(T)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i}{\left\| \mathbf{V}^{(T)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i \right\|_2} \right\|_2 \\
&\leq \frac{\left\| \beta^{(T)} \mathbf{\Pi}^\top \sum_{i=1}^{N-2} \mathcal{S}_i^{(T)} \mathbf{x}_i \right\|_2 + \left\| \tilde{\mathbf{V}}^{(T)} \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i \right\|_2}{\left\| \beta^{(T)} \mathbf{\Pi}^\top \sum_{i=1}^{N-1} \mathcal{S}_i^{(T)} \mathbf{x}_i \right\|_2} \\
&\leq \frac{\exp(-\Omega(N)) \|\mathbf{V}^{(T)}\|_{\max} + \sqrt{K} \gamma^{(T)}}{\sqrt{p^2 + (1-p)^2} \beta^{(T)}} \\
&\leq \frac{\exp(-\Omega(N)) \left(\frac{\eta}{\epsilon K} + 2T\epsilon K^2 \right) + \sqrt{K} \left(\frac{2\eta}{\epsilon K} + 2T\epsilon K^2 N \exp(-\Omega(N)) \right)}{\sqrt{p^2 + (1-p)^2} \cdot (\sqrt{\eta T} - \frac{2\eta}{\epsilon K})} \\
&\leq \mathcal{O} \left(\frac{1}{\sqrt{T}} \right)
\end{aligned}$$

where the second inequality is by Lemma C.7 and $\|\mathbf{\Pi}^\top \mathbf{x}_i\|_2 = \sqrt{p^2 + (1-p)^2}$, and the third inequality is by Lemma C.7. Therefore, we can obtain that

$$\left\| \frac{f_{\theta^{(T)}}(\mathbf{X})}{\|f_{\theta^{(T)}}(\mathbf{X})\|_2} - \mathbf{\Pi}^\top \mathbf{x}_{N-1} \right\|_2 \leq \mathcal{O} \left(\frac{1}{\sqrt{T}} \right).$$

□

D Proof of Theorem 4.1

In this section, we analyze “deterministic walks” with $p = 0, 1$. Without loss of generality, we set $p = 1$, which means $\mathbf{\Pi} = \mathbf{\Pi}_0^\top$. The following lemma shows the results of the first iteration.

Lemma D.1. Under the same condition as Theorem 4.1, for any loss function $\ell(\cdot)$, it holds that

$$\mathbf{V}^{(1)} = -\ell'(\theta^{(0)}) \cdot \frac{\eta r}{NK} \mathbf{1}_{K \times K} \text{ and } \mathbf{W}^{(1)} = \mathbf{0}_{(K+M) \times (K+M)}.$$

Proof of Lemma D.1. By Lemma B.1, we have

$$\begin{aligned} \mathbb{E}[\nabla_{\mathbf{V}} \ell(\theta^{(0)})] &= \ell'(\theta^{(0)}) \cdot \frac{1}{N} \sum_{i=1}^{N-1} \mathbb{E}[\mathbf{e}_y \mathbf{x}_i^\top] \\ &= \ell'(\theta^{(0)}) \cdot \frac{1}{N} \sum_{i=1}^{N-1} \mathbb{E}[(\mathbf{\Pi}^\top)^{N-i} \mathbf{x}_i \mathbf{x}_i^\top] \\ &= \ell'(\theta^{(0)}) \cdot \frac{1}{NK} \sum_{i=1}^{N-1} (\mathbf{\Pi}_0)^{N-i} \\ &= \ell'(\theta^{(0)}) \cdot \frac{r}{NK} \mathbf{1}_{K \times K}, \end{aligned}$$

where the first equation is by the initialization of $\mathbf{V}^{(0)}$ and $\mathbf{W}^{(0)}$, the second equation is by the sampling method, the third equation is by $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top] = \frac{1}{K} \mathbf{I}_K$ for $i \in [N-1]$, and the last equation is by Lemma E.1. Thus, by the update, we can get

$$\mathbf{V}^{(1)} = \mathbf{V}^{(0)} - \eta \mathbb{E}[\nabla_{\mathbf{V}} \ell(\theta^{(0)})] = -\ell'(\theta^{(0)}) \cdot \frac{\eta r}{NK} \mathbf{1}_{K \times K}.$$

Since $\mathbf{V}^{(0)} = \mathbf{0}_{K \times K}$ and $\mathbf{W}^{(0)} = \mathbf{0}_{(K+M) \times (K+M)}$, we can get $\mathbb{E}[\nabla_{\mathbf{W}} \ell(\theta^{(0)})] = \mathbf{0}_{(K+M) \times (K+M)}$. Thus,

$$\mathbf{W}^{(1)} = \mathbf{W}^{(0)} - \eta \mathbb{E}[\nabla_{\mathbf{W}} \ell(\theta^{(0)})] = \mathbf{0}_{(K+M) \times (K+M)}.$$

□

The following lemma states the results of the second iteration.

Lemma D.2. If $\mathbf{\Pi} = \mathbf{\Pi}_2$, then it holds that

$$\begin{aligned} \mathbf{V}^{(2)} &= -(\ell'(\theta^{(0)}) + \ell'(\theta^{(1)})) \cdot \frac{\eta r}{NK} \mathbf{1}_{K \times K}, \\ \mathbf{W}_{12}^{(2)} &= \ell'(\theta^{(1)}) \ell'(\theta^{(0)}) \frac{\eta^2 r^2}{N^3 K} \mathbf{1}_K \mathbf{p}_N^\top, \\ \mathbf{W}_{22}^{(2)} &= \ell'(\theta^{(1)}) \ell'(\theta^{(0)}) \left(\frac{\eta^2 r}{N^3 K} \sum_{i=1}^{N-1} \mathbf{p}_i - \frac{\eta^2 r^2}{N^3} \mathbf{p}_N \right) \mathbf{p}_N^\top. \end{aligned}$$

Proof of Lemma D.2. By Lemma B.1, we have

$$\begin{aligned} \mathbb{E}[\nabla_{\mathbf{V}} \ell(\theta^{(1)})] &= \ell'(\theta^{(1)}) \cdot \frac{1}{N} \sum_{i=1}^{N-1} \mathbb{E}[\mathbf{e}_y \mathbf{x}_i^\top] \\ &= \ell'(\theta^{(1)}) \cdot \frac{1}{N} \sum_{i=1}^{N-1} \mathbb{E}[(\mathbf{\Pi}^\top)^{N-i} \mathbf{x}_i \mathbf{x}_i^\top] \\ &= \ell'(\theta^{(1)}) \cdot \frac{1}{NK} \sum_{i=1}^{N-1} (\mathbf{\Pi}_0)^{N-i} \\ &= \ell'(\theta^{(1)}) \cdot \frac{r}{NK} \mathbf{1}_{K \times K}, \end{aligned}$$

where the second equation is by the sampling method, the third equation is by $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\top] = \frac{1}{K} \mathbf{I}_K$ for $i \in [N-1]$, and the last equation is by Lemma E.1. Thus, we can get

$$\begin{aligned} \mathbf{V}^{(2)} &= \mathbf{V}^{(1)} - \eta \mathbb{E}[\nabla_{\mathbf{V}} \ell(\theta^{(1)})] \\ &= -\ell'(\theta^{(0)}) \cdot \frac{\eta r}{NK} \mathbf{1}_{K \times K} - \ell'(\theta^{(1)}) \cdot \frac{\eta r}{NK} \mathbf{1}_{K \times K}. \end{aligned}$$

By Lemma B.1, we have

$$\begin{aligned} \mathbb{E}[\mathbf{A}^{(1)}] &= \mathbb{E} \left[\left(\sum_{i=1}^{N-1} \mathcal{S}_i^{(1)} \mathbf{x}_i \mathbf{x}_i^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y - \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \mathcal{S}_{i_1}^{(1)} \mathcal{S}_{i_2}^{(1)} \mathbf{x}_{i_1} \mathbf{x}_{i_2}^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y \right) \mathbf{p}_N^\top \right] \\ &= \mathbb{E} \left[-\ell'(\theta^{(0)}) \left(\frac{\eta r}{N^2 K} \sum_{i=1}^{N-1} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{1}_K - \frac{\eta r}{N^3 K} \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \mathbf{x}_{i_1} \mathbf{x}_{i_2}^\top \mathbf{1}_K \right) \mathbf{p}_N^\top \right] \\ &= \mathbb{E} \left[-\ell'(\theta^{(0)}) \left(\frac{\eta r}{N^2 K} \sum_{i=1}^{N-1} \mathbf{x}_i - \frac{\eta r}{N^3 K} \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \mathbf{x}_{i_1} \right) \mathbf{p}_N^\top \right] \\ &= -\ell'(\theta^{(0)}) \left(\frac{\eta r^2}{N^2 K} \mathbf{1}_K - \frac{\eta r^2 (N-1)}{N^3 K} \mathbf{1}_K \right) \mathbf{p}_N^\top \\ &= -\ell'(\theta^{(0)}) \frac{\eta r^2}{N^3 K} \mathbf{1}_K \mathbf{p}_N^\top, \end{aligned}$$

where the second equation is by Lemma D.1, and the fourth equation is by the fact that all the $\mathbf{x}_i$ are uniformly distributed in $\mathbf{E}$. We also have

$$\begin{aligned} \mathbb{E}[\mathbf{B}^{(1)}] &= \mathbb{E} \left[\left(\sum_{i=1}^{N-1} \mathcal{S}_i^{(1)} \mathbf{p}_i \mathbf{x}_i^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y - \sum_{i=1}^N \mathcal{S}_i^{(1)} \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i^{(1)} \mathbf{x}_i^\top (\mathbf{V}^{(1)})^\top \mathbf{e}_y \right) \mathbf{p}_N^\top \right] \\ &= \mathbb{E} \left[-\ell'(\theta^{(0)}) \left(\frac{\eta r}{N^2 K} \sum_{i=1}^{N-1} \mathbf{p}_i \mathbf{x}_i^\top \mathbf{1}_K - \frac{\eta r}{N^3 K} \sum_{i=1}^N \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathbf{x}_i^\top \mathbf{1}_K \right) \mathbf{p}_N^\top \right] \\ &= -\ell'(\theta^{(0)}) \left(\frac{\eta r}{N^2 K} \sum_{i=1}^{N-1} \mathbf{p}_i - \frac{\eta r (N-1)}{N^3 K} \sum_{i=1}^N \mathbf{p}_i \right) \mathbf{p}_N^\top \\ &= -\ell'(\theta^{(0)}) \left(\frac{\eta r}{N^3 K} \sum_{i=1}^{N-1} \mathbf{p}_i - \frac{\eta r^2}{N^3} \mathbf{p}_N \right) \mathbf{p}_N^\top, \end{aligned}$$

where the second equation is by Lemma D.1. Thus, we can get that

$$\begin{aligned} \mathbf{W}_{12}^{(2)} &= \mathbf{W}_{12}^{(1)} - \eta \mathbb{E}[\nabla_{\mathbf{W}} \ell(\theta^{(1)})]_{12} \\ &= -\mathbb{E} \left[\eta \ell'(\theta^{(1)}) \cdot \mathbf{A}^{(1)} \right] \\ &= \ell'(\theta^{(1)}) \ell'(\theta^{(0)}) \frac{\eta^2 r^2}{N^3 K} \mathbf{1}_K \mathbf{p}_N^\top, \end{aligned}$$

and

$$\begin{aligned} \mathbf{W}_{22}^{(2)} &= \mathbf{W}_{22}^{(1)} - \eta \mathbb{E}[\nabla_{\mathbf{W}} \ell(\theta^{(1)})]_{22} \\ &= -\mathbb{E} \left[\eta \ell'(\theta^{(1)}) \cdot \mathbf{B}^{(1)} \right] \\ &= \ell'(\theta^{(1)}) \ell'(\theta^{(0)}) \left(\frac{\eta^2 r}{N^3 K} \sum_{i=1}^{N-1} \mathbf{p}_i - \frac{\eta^2 r^2}{N^3} \mathbf{p}_N \right) \mathbf{p}_N^\top. \end{aligned}$$

□

Next, we can analyze the gradient descent dynamics over multiple iterations.

Lemma D.3. If $\Pi = \Pi_2$, then for any $t \geq 0$ and any sequence of learning rates $\{\eta_t\}$, it holds that

$$\mathbf{V}^{(t)} \propto \mathbf{1}_{K \times K}, \text{ and } \mathcal{S}_1^{(t)} = \mathcal{S}_2^{(t)} = \dots = \mathcal{S}_{N-1}^{(t)}.$$

Proof of Lemma D.3. We use induction to prove that for some scalar $\alpha_1^{(t)}, \alpha_2^{(t)}, \alpha_3^{(t)}, \alpha_4^{(t)}$, it holds that for $t \geq 2$, $\mathbf{V}^{(t)} = \alpha_1^{(t)} \mathbf{1}_{K \times K}$, $\mathbf{W}_{12}^{(t)} = \alpha_2^{(t)} \mathbf{1}_K \mathbf{p}_N^\top$, and $\mathbf{W}_{22}^{(t)} = \left(\alpha_3^{(t)} \sum_{i=1}^{N-1} \mathbf{p}_i - \alpha_4^{(t)} \mathbf{p}_N \right) \mathbf{p}_N^\top$. By Lemma D.2, we know that the hypothesis holds for $t = 2$. Suppose that the hypothesis holds for $t = t'$. We aim to prove that the hypothesis holds for $t = t' + 1$. We have

$$\begin{aligned} \widetilde{\mathbf{X}} \mathbf{W}^{(t')} \widetilde{\mathbf{x}}_N &= \begin{bmatrix} \mathbf{x}_1^\top \mathbf{W}_{12}^{(t')} \mathbf{p}_N + \mathbf{p}_1^\top \mathbf{W}_{22}^{(t')} \mathbf{p}_N \\ \mathbf{x}_2^\top \mathbf{W}_{12}^{(t')} \mathbf{p}_N + \mathbf{p}_2^\top \mathbf{W}_{22}^{(t')} \mathbf{p}_N \\ \vdots \\ \mathbf{x}_{N-1}^\top \mathbf{W}_{12}^{(t')} \mathbf{p}_N + \mathbf{p}_{N-1}^\top \mathbf{W}_{22}^{(t')} \mathbf{p}_N \\ \mathbf{p}_N^\top \mathbf{W}_{22}^{(t')} \mathbf{p}_N \end{bmatrix} \\ &= \begin{bmatrix} \alpha_2^{(t')} \mathbf{p}_N^\top \mathbf{p}_N + \alpha_3^{(t')} \mathbf{p}_1^\top \mathbf{p}_1 \mathbf{p}_N^\top \mathbf{p}_N \\ \alpha_2^{(t')} \mathbf{p}_N^\top \mathbf{p}_N + \alpha_3^{(t')} \mathbf{p}_2^\top \mathbf{p}_2 \mathbf{p}_N^\top \mathbf{p}_N \\ \vdots \\ \alpha_2^{(t')} \mathbf{p}_N^\top \mathbf{p}_N + \alpha_3^{(t')} \mathbf{p}_{N-1}^\top \mathbf{p}_{N-1} \mathbf{p}_N^\top \mathbf{p}_N \\ -\alpha_4^{(t')} (\mathbf{p}_N^\top \mathbf{p}_N)^2 \end{bmatrix}. \end{aligned} \quad (\text{D.1})$$

Since $\mathbf{p}_1^\top \mathbf{p}_1 = \mathbf{p}_2^\top \mathbf{p}_2 = \dots = \mathbf{p}_N^\top \mathbf{p}_N$, we have $[\widetilde{\mathbf{X}} \mathbf{W}^{(t')} \widetilde{\mathbf{x}}_N]_1 = [\widetilde{\mathbf{X}} \mathbf{W}^{(t')} \widetilde{\mathbf{x}}_N]_2 = \dots = [\widetilde{\mathbf{X}} \mathbf{W}^{(t')} \widetilde{\mathbf{x}}_N]_{N-1}$. Thus, we can get that $\mathcal{S}_1^{(t')} = \mathcal{S}_2^{(t')} = \dots = \mathcal{S}_{N-1}^{(t')} := s^{(t')}$. Then, we have

$$\begin{aligned} \mathbb{E}[\nabla_{\mathbf{V}} \ell(\theta^{(t')})] &= \mathbb{E} \left[\ell'(\theta^{(t')}) \cdot \mathbf{e}_y \sum_{i=1}^{N-1} \mathcal{S}_i^{(t')} \mathbf{x}_i^\top \right] \\ &= \ell'(\theta^{(t')}) s^{(t')} \sum_{i=1}^{N-1} \mathbb{E}[\mathbf{e}_y \mathbf{x}_i^\top] \\ &= \ell'(\theta^{(t')}) s^{(t')} \sum_{i=1}^{N-1} \mathbb{E}[(\Pi^\top)^{N-i} \mathbf{x}_i \mathbf{x}_i^\top] \\ &= \ell'(\theta^{(t')}) \frac{s^{(t')}}{K} \sum_{i=1}^{N-1} (\Pi_0)^{N-i} \\ &= \ell'(\theta^{(t')}) \frac{s^{(t')} r}{K} \mathbf{1}_{K \times K}, \end{aligned}$$

where the second equation is by the induction, the third equation is by the sampling method, the fourth equation is by the fact that $\mathbf{x}_i$ is uniformly distributed in $\mathbf{E}$, and the last equation is by Lemma E.1. Thus, we can get $\mathbf{V}^{(t'+1)} = \mathbf{V}^{(t')} - \eta^{(t')} \mathbb{E}[\nabla_{\mathbf{V}} \ell(\theta^{(t')})] \propto \mathbf{1}_{K \times K}$. We also have

$$\begin{aligned} \mathbb{E}[\mathbf{A}^{(t')}] &= \mathbb{E} \left[\left(\sum_{i=1}^{N-1} \mathcal{S}_i^{(t')} \mathbf{x}_i \mathbf{x}_i^\top (\mathbf{V}^{(t')})^\top \mathbf{e}_y - \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \mathcal{S}_{i_1}^{(t')} \mathcal{S}_{i_2}^{(t')} \mathbf{x}_{i_1} \mathbf{x}_{i_2}^\top (\mathbf{V}^{(t')})^\top \mathbf{e}_y \right) \mathbf{p}_N^\top \right] \\ &= \mathbb{E} \left[\left(\alpha_1^{(t')} s^{(t')} \sum_{i=1}^{N-1} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{1}_K - \alpha_1^{(t')} (s^{(t')})^2 \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \mathbf{x}_{i_1} \mathbf{x}_{i_2}^\top \mathbf{1}_K \right) \mathbf{p}_N^\top \right] \\ &= \mathbb{E} \left[\left(\alpha_1^{(t')} s^{(t')} \sum_{i=1}^{N-1} \mathbf{x}_i - \alpha_1^{(t')} (s^{(t')})^2 \sum_{i_1=1}^{N-1} \sum_{i_2=1}^{N-1} \mathbf{x}_{i_1} \right) \mathbf{p}_N^\top \right] \\ &= \left(\frac{\alpha_1^{(t')} s^{(t')} (N-1)}{K} \mathbf{1}_K - \frac{\alpha_1^{(t')} (s^{(t')})^2 (N-1)^2}{K} \mathbf{1}_K \right) \mathbf{p}_N^\top, \end{aligned}$$

where the second equation is by the induction, and the fourth equation is by the fact that all the $\mathbf{x}_i$ is uniformly distributed in $\mathbf{E}$. And,

$$\begin{aligned}\mathbb{E}[\mathbf{B}^{(t')}] &= \mathbb{E}\left[\left(\sum_{i=1}^{N-1} \mathcal{S}_i^{(t')} \mathbf{p}_i \mathbf{x}_i^\top (\mathbf{V}^{(t')})^\top \mathbf{e}_y - \sum_{i=1}^N \mathcal{S}_i^{(t')} \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathcal{S}_i^{(t')} \mathbf{x}_i^\top (\mathbf{V}^{(t')})^\top \mathbf{e}_y\right) \mathbf{p}_N^\top\right] \\ &= \mathbb{E}\left[\left(\alpha_1^{(t')} s^{(t')} \sum_{i=1}^{N-1} \mathbf{p}_i \mathbf{x}_i^\top \mathbf{1}_K - \alpha_1^{(t')} (s^{(t')})^2 \sum_{i=1}^N \mathbf{p}_i \cdot \sum_{i=1}^{N-1} \mathbf{x}_i^\top \mathbf{1}_K\right) \mathbf{p}_N^\top\right] \\ &= \left(\alpha_1^{(t')} s^{(t')} \sum_{i=1}^{N-1} \mathbf{p}_i - \alpha_1^{(t')} (s^{(t')})^2 (N-1) \sum_{i=1}^N \mathbf{p}_i\right) \mathbf{p}_N^\top,\end{aligned}$$

where the second equation is by the induction. Therefore, we can get

$$\begin{aligned}\mathbf{W}_{12}^{(t'+1)} &= \mathbf{W}_{12}^{(t')} - \eta \mathbb{E}[\nabla_{\mathbf{W}} \ell(\theta^{(t')})]_{12} \\ &= \alpha_2^{(t')} \mathbf{1}_K \mathbf{p}_N^\top - \eta \ell'(\theta^{(t')}) \mathbb{E}[\mathbf{A}^{(t')}] \\ &= \alpha_2^{(t')} \mathbf{1}_K \mathbf{p}_N^\top - \eta \ell'(\theta^{(t')}) \left(\frac{\alpha_1^{(t')} s^{(t')} (N-1)}{K} - \frac{\alpha_1^{(t')} (s^{(t')})^2 (N-1)^2}{K}\right) \mathbf{1}_K \mathbf{p}_N^\top \\ &:= \alpha_2^{(t'+1)} \mathbf{1}_K \mathbf{p}_N^\top,\end{aligned}$$

and

$$\begin{aligned}\mathbf{W}_{22}^{(t'+1)} &= \mathbf{W}_{22}^{(t')} - \eta \mathbb{E}[\nabla_{\mathbf{W}} \ell(\theta^{(t')})]_{22} \\ &= \left(\alpha_3^{(t')} \sum_{i=1}^{N-1} \mathbf{p}_i - \alpha_4^{(t')} \mathbf{p}_N\right) \mathbf{p}_N^\top - \eta \ell'(\theta^{(t')}) \mathbb{E}[\mathbf{B}^{(t')}] \\ &= \left(\alpha_3^{(t')} \sum_{i=1}^{N-1} \mathbf{p}_i - \alpha_4^{(t')} \mathbf{p}_N\right) \mathbf{p}_N^\top \\ &\quad - \eta \ell'(\theta^{(t')}) \left(\alpha_1^{(t')} s^{(t')} \sum_{i=1}^{N-1} \mathbf{p}_i - \alpha_1^{(t')} (s^{(t')})^2 (N-1) \sum_{i=1}^N \mathbf{p}_i\right) \mathbf{p}_N^\top \\ &:= \left(\alpha_3^{(t'+1)} \sum_{i=1}^{N-1} \mathbf{p}_i - \alpha_4^{(t'+1)} \mathbf{p}_N\right) \mathbf{p}_N^\top.\end{aligned}$$

Therefore, by induction, we can conclude that for all $t \geq 2$, $\mathbf{V}^{(t)} = \alpha_1^{(t)} \mathbf{1}_{K \times K}$, $\mathbf{W}_{12}^{(t)} = \alpha_2^{(t)} \mathbf{1}_K \mathbf{p}_N^\top$, and $\mathbf{W}_{22}^{(t)} = \left(\alpha_3^{(t)} \sum_{i=1}^{N-1} \mathbf{p}_i - \alpha_4^{(t)} \mathbf{p}_N\right) \mathbf{p}_N^\top$. Similar to (D.1), we have $[\tilde{\mathbf{X}} \mathbf{W}^{(t)} \tilde{\mathbf{x}}_N]_1 = [\tilde{\mathbf{X}} \mathbf{W}^{(t)} \tilde{\mathbf{x}}_N]_2 = \dots = [\tilde{\mathbf{X}} \mathbf{W}^{(t)} \tilde{\mathbf{x}}_N]_{N-1}$, which implies that $\mathcal{S}_1^{(t)} = \mathcal{S}_2^{(t)} = \dots = \mathcal{S}_{N-1}^{(t)}$. $\square$

E Auxiliary Lemmas

In this section, we present some auxiliary lemmas. The following lemma states the properties of $\mathbf{\Pi}_0$.

Lemma E.1. By the definition of $\mathbf{\Pi}_0$, it holds that $\mathbf{\Pi}_0^K = \mathbf{I}_K$, $\mathbf{\Pi}_0 \mathbf{\Pi}_0^\top = \mathbf{I}_K$, and $\sum_{k=1}^K \mathbf{\Pi}_0^k = \mathbf{1}_{K \times K}$.

Proof of Lemma E.1. In this proof, the index i larger than K represents $i - K$. For $\mathbf{\Pi}_0$, only $[\mathbf{\Pi}_0]_{i+1,i} = 1$ for $i \in [K]$ and other elements are 0. We can get that for $\mathbf{\Pi}_0^k$, only $[\mathbf{\Pi}_0^k]_{i+k,i} = 1$ for $i \in [K]$ and other elements are 0. By this observation, we can derive that $\mathbf{\Pi}_0^K = \mathbf{I}_K$ and $\sum_{k=1}^K \mathbf{\Pi}_0^k = \mathbf{1}_{K \times K}$. Also, we have $\mathbf{\Pi}_0^\top = \mathbf{\Pi}_0^{K-1}$, so we can get $\mathbf{\Pi}_0 \mathbf{\Pi}_0^\top = \mathbf{\Pi}_0^K = \mathbf{I}_K$. $\square$

Further, the following three lemmas show some properties of $\mathbf{\Pi}$ with transition probability p satisfying $0 < p < 1$.

Lemma E.2. Assume that K is odd. It holds that

$$\left| [\mathbf{\Pi}^R]_{i,j} - \frac{1}{K} \right| \leq \exp \left(-\frac{8p(1-p)R}{K^2} \right).$$

Proof of Lemma E.2. $\mathbf{\Pi}$ has eigenvalues $\lambda_0, \dots, \lambda_{K-1}$, where

$$\lambda_k = pe^{-\frac{2\pi ik}{K}} + (1-p)e^{\frac{2\pi ik}{K}} = \cos \left(\frac{2\pi k}{K} \right) + i(1-2p) \sin \left(\frac{2\pi k}{K} \right)$$

with

$$|\lambda_k| = \sqrt{1 - 4p(1-p) \sin^2 \left(\frac{2\pi k}{K} \right)} \leq \sqrt{1 - \frac{16p(1-p)}{K^2}} \leq 1 - \frac{8p(1-p)}{K^2}$$

for $k \neq 0$, where the first inequality is by $\sin(2\pi k/K) \geq \sin(\pi/K) \geq 2/K$. The eigendecomposition of each entry in $\mathbf{\Pi}$ can be written as

$$[\mathbf{\Pi}]_{i,j} = \sum_{k=0}^{K-1} c_{k,i,j} \lambda_k = \frac{1}{K} \lambda_0 + \sum_{k \neq 0} c_{k,i,j} \lambda_k,$$

where $c_{k,i,j} = \frac{1}{\sqrt{K}} e^{2\pi i(i-1)k/K} \cdot \frac{1}{\sqrt{K}} e^{2\pi i(j-1)k/K} = \frac{1}{K} e^{2\pi i(i+j-2)k/K}$. Then, we can get that

$$[\mathbf{\Pi}^R]_{i,j} = \sum_{k=0}^{K-1} c_{k,i,j} \lambda_k^R = \frac{1}{K} + \sum_{k \neq 0} c_{k,i,j} \lambda_k^R$$

Thus,

$$\begin{aligned} \left| [\mathbf{\Pi}^R]_{i,j} - \frac{1}{K} \right| &= \left| \sum_{k \neq 0} c_{k,i,j} \lambda_k^R \right| \\ &\leq \sum_{k \neq 0} |c_{k,i,j}| |\lambda_k|^R \\ &\leq \left(1 - \frac{8p(1-p)}{K^2} \right)^R \sum_{k \neq 0} |c_{k,i,j}| \\ &\leq \exp \left(-\frac{8p(1-p)R}{K^2} \right), \end{aligned}$$

where the second inequality is by the bound of the absolute values of the eigenvalues, and the last inequality is by $(1-t)^R \leq e^{-Rt}$ for any $0 < t < 1$ and $|c_{k,i,j}| = \frac{1}{K}$ for all k, i, j . $\square$

Lemma E.3. Assume that K is even. For the case that R is even,

$$\begin{aligned} [\mathbf{\Pi}^R]_{i,j} &= 0 \text{ for odd } (j-i); \\ \left| [\mathbf{\Pi}^R]_{i,j} - \frac{2}{K} \right| &\leq \exp \left(-\frac{8p(1-p)R}{K^2} \right) \text{ for even } (j-i). \end{aligned}$$

For the case that R is odd,

$$\begin{aligned} \left| [\mathbf{\Pi}^R]_{i,j} - \frac{2}{K} \right| &\leq \exp \left(-\frac{8p(1-p)R}{K^2} \right) \text{ for odd } (j-i); \\ [\mathbf{\Pi}^R]_{i,j} &= 0 \text{ for even } (j-i). \end{aligned}$$

Proof of Lemma E.3. $\mathbf{\Pi}$ has eigenvalues $\lambda_0, \dots, \lambda_{K-1}$, where

$$\lambda_k = pe^{-\frac{2\pi ik}{K}} + (1-p)e^{\frac{2\pi ik}{K}} = \cos \left(\frac{2\pi k}{K} \right) + i(1-2p) \sin \left(\frac{2\pi k}{K} \right)$$

with

$$|\lambda_k| = \sqrt{1 - 4p(1-p) \sin^2 \left(\frac{2\pi k}{K} \right)} \leq \sqrt{1 - \frac{16p(1-p)}{K^2}} \leq 1 - \frac{8p(1-p)}{K^2}.$$

for $k \neq 0, K/2$, where the first inequality is by $\sin(2\pi k/K) \geq \sin(\pi/K) \geq 2/K$. The eigendecomposition of each entry in Π can be written as

$$[\Pi]_{i,j} = \sum_{k=0}^{K-1} c_{k,i,j} \lambda_k = \frac{1}{K} \lambda_0 + \frac{(-1)^{i+j}}{K} \lambda_{K/2} + \sum_{k \neq 0, K/2} c_{k,i,j} \lambda_k,$$

where $c_{k,i,j} = \frac{1}{\sqrt{K}} e^{2\pi i(i-1)k/K} \cdot \frac{1}{\sqrt{K}} e^{2\pi i(j-1)k/K} = \frac{1}{K} e^{2\pi i(i+j-2)k/K}$. Then, we can get that

$$\begin{aligned} [\Pi^R]_{i,j} &= \sum_{k=0}^{K-1} c_{k,i,j} \lambda_k^R = \frac{1}{K} \lambda_0^R + \frac{(-1)^{i+j}}{K} \lambda_{K/2}^R + \sum_{k \neq 0, K/2} c_{k,i,j} \lambda_k^R \\ &= \frac{1}{K} + \frac{(-1)^{i+j+R}}{K} + \sum_{k \neq 0, K/2} c_{k,i,j} \lambda_k^R \end{aligned}$$

When $i+j+R$ is odd, it is easy to obtain that

$$\begin{aligned} c_{t,i,j} \lambda_t^R + c_{t+\frac{K}{2},i,j} \lambda_{t+\frac{K}{2}}^R &= \frac{1}{K} e^{2\pi i(i+j-2)t/K} \lambda_t^R + \frac{1}{K} (-1)^{i+j-2} e^{2\pi i(i+j-2)t/K} (-\lambda_t)^R \\ &= \frac{1}{K} e^{2\pi i(i+j-2)t/K} \lambda_t^R + (-1)^{i+j+R} \cdot \frac{1}{K} e^{2\pi i(i+j-2)t/K} \lambda_t^R \\ &= 0, \end{aligned}$$

which indicates that $[\Pi^R]_{i,j} = 0$.

When $i+j+R$ is even,

$$\begin{aligned} \left| [\Pi^R]_{i,j} - \frac{2}{K} \right| &= \left| \sum_{k \neq 0, K/2} c_{k,i,j} \lambda_k^R \right| \\ &\leq \sum_{k \neq 0, K/2} |c_{k,i,j}| |\lambda_k|^R \\ &\leq \left(1 - \frac{8p(1-p)}{K^2} \right)^R \sum_{k \neq 0, K/2} |c_{k,i,j}| \\ &\leq \exp \left(-\frac{8p(1-p)R}{K^2} \right), \end{aligned}$$

where the second inequality is by the bound of the absolute values of the eigenvalues, and the last inequality is by $(1-t)^R \leq e^{-Rt}$ for any $0 < t < 1$ and $|c_{k,i,j}| = \frac{1}{K}$ for all k, i, j . $\square$

Lemma E.4. For Π with $0 < p < 1$, $N = \omega(p^{-1}(1-p)^{-1})$ and $k \in \{2, \dots, N-1\}$, it holds that

$$\left[\sum_{i=1}^{N-1} \Pi^i \cdot \Pi^\top \right]_{1,1} \geq \left[\sum_{i=1}^{N-1} \Pi^i \cdot (\Pi^\top)^k \right]_{1,1} + C_p,$$

where C_p is a positive constant only depending on p .

Proof of Lemma E.4. Without loss of generality, we assume $\frac{1}{2} \leq p < 1$. We observe that

$$\left[\sum_{i=1}^{N-1} \Pi^i \Pi^\top \right]_{1,1} = \mathbf{e}_1^\top \sum_{i=1}^{N-1} \Pi^i \Pi^\top \mathbf{e}_1 = \mathbf{e}_1^\top \sum_{i=0}^{N-2} \Pi^i \cdot \Pi \Pi^\top \mathbf{e}_1,$$

$$\left[\sum_{i=1}^{N-1} \mathbf{\Pi}^i (\mathbf{\Pi}^\top)^k \right]_{1,1} = \mathbf{e}_1^\top \sum_{i=1}^{N-1} \mathbf{\Pi}^i (\mathbf{\Pi}^\top)^k \mathbf{e}_1 = \mathbf{e}_1^\top \sum_{i=0}^{N-2} \mathbf{\Pi}^i \cdot \mathbf{\Pi} (\mathbf{\Pi}^\top)^k \mathbf{e}_1.$$

Denote $\mathbf{\Gamma}(N) = \sum_{i=0}^{N-2} \mathbf{\Pi}^i$. We use induction to prove that

$$\mathbf{\Gamma}(N)_{1,1} \geq \mathbf{\Gamma}(N)_{i,j} + (1-p)^{N-2} \text{ for } i \neq j.$$

It can be easily obtained that for $N = 2$, $\mathbf{\Gamma}(2)_{1,1} = \mathbf{\Gamma}(2)_{i,j} + 1$ for any $i \neq j$. Suppose that the induction hypothesis holds for $\mathbf{\Gamma}(N-1)$. Then, for $\mathbf{\Gamma}(N)$, by the definition of $\mathbf{\Gamma}(N)$, we can get $\mathbf{\Gamma}(N) = \mathbf{\Pi} \cdot \mathbf{\Gamma}(N-1) + \mathbf{I}$. Thus, based on the definition of $\mathbf{\Pi}$, we obtain that for $i \neq j$,

$$\mathbf{\Gamma}(N)_{i,j} = p \cdot \mathbf{\Gamma}(N-1)_{i,j-1} + (1-p) \cdot \mathbf{\Gamma}(N-1)_{i,j+1}.$$

When $|i-j| \geq 2$, by induction, we have

$$\begin{aligned} \mathbf{\Gamma}(N)_{i,j} &= p \cdot \mathbf{\Gamma}(N-1)_{i,j-1} + (1-p) \cdot \mathbf{\Gamma}(N-1)_{i,j+1} \\ &\leq p(\mathbf{\Gamma}(N-1)_{1,1} - (1-p)^{N-3}) + (1-p)(\mathbf{\Gamma}(N-1)_{1,1} - (1-p)^{N-3}) \\ &= \mathbf{\Gamma}(N-1)_{1,1} - (1-p)^{N-3} \\ &\leq \mathbf{\Gamma}(N-1)_{1,1} - (1-p)^{N-2}. \end{aligned}$$

When $i = j+1$, we have

$$\begin{aligned} \mathbf{\Gamma}(N)_{i,j} &= p \cdot \mathbf{\Gamma}(N-1)_{i,j-1} + (1-p) \cdot \mathbf{\Gamma}(N-1)_{i,j+1} \\ &\leq p(\mathbf{\Gamma}(N-1)_{1,1} - (1-p)^{N-3}) + (1-p)\mathbf{\Gamma}(N-1)_{1,1} \\ &= \mathbf{\Gamma}(N-1)_{1,1} - p(1-p)^{N-3} \\ &\leq \mathbf{\Gamma}(N-1)_{1,1} - (1-p)^{N-2}. \end{aligned}$$

When $i = j-1$, we have

$$\begin{aligned} \mathbf{\Gamma}(N)_{i,j} &= p \cdot \mathbf{\Gamma}(N-1)_{i,j-1} + (1-p) \cdot \mathbf{\Gamma}(N-1)_{i,j+1} \\ &\leq p\mathbf{\Gamma}(N-1)_{1,1} + (1-p)(\mathbf{\Gamma}(N-1)_{1,1} - (1-p)^{N-3}) \\ &= \mathbf{\Gamma}(N-1)_{1,1} - (1-p)^{N-2}. \end{aligned}$$

Obviously, $\mathbf{\Gamma}(N-1)_{1,1} \leq \mathbf{\Gamma}(N)_{1,1}$ as $\mathbf{\Gamma}(N) = \mathbf{\Gamma}(N-1) + \mathbf{\Pi}^{N-2}$. Therefore, we prove that the induction hypothesis holds for $\mathbf{\Gamma}(N)$, which completes the induction. Thus, for all $N \geq 2$, the diagonal element is the largest entry in $\mathbf{\Gamma}(N)$. We analyze this result from the perspective of random walks. Given that $\mathbf{\Pi}^k$ is the k -step transition matrix in the random walk task, the entry $\mathbf{\Gamma}(N)_{i,j} = \sum_{l=0}^{N-2} \mathbf{\Pi}^l$ can be regarded as the expected number of visits to state j within $N-1$ steps, starting from state i . Since the largest entry in $\mathbf{\Gamma}(N)$ is found at the diagonal, we can conclude that the expected number of visits back to state i within $N-1$ steps, starting from state i , is the largest. From Lemma E.2 and E.3, we know that for any i, j ,

$$\left| [\mathbf{\Pi}^R]_{i,j} + [\mathbf{\Pi}^{R+1}]_{i,j} - \frac{2}{K} \right| \leq \exp\left(-\frac{8p(1-p)R}{K^2}\right) + \exp\left(-\frac{8p(1-p)(R+1)}{K^2}\right).$$

Hence, we get that for any i, j ,

$$|[\mathbf{\Pi}^R]_{1,1} + [\mathbf{\Pi}^{R+1}]_{1,1} - [\mathbf{\Pi}^R]_{i,j} - [\mathbf{\Pi}^{R+1}]_{i,j}| \leq 2 \left[\exp\left(-\frac{8p(1-p)R}{K^2}\right) + \exp\left(-\frac{8p(1-p)(R+1)}{K^2}\right) \right].$$

Thus, for a constant integer $N_0 = \omega(p^{-1}(1-p)^{-1})$ and $N > N_0$, we can get for $i \neq j$,

$$\mathbf{\Gamma}(N)_{1,1} - \mathbf{\Gamma}(N)_{i,j} \geq (1-p)^{N_0-2} - c_{p0} \exp(-N_0),$$

where c_{p0} are constants related to p . In addition, we can get $|\mathbf{\Gamma}(N)_{1,j_1} - \mathbf{\Gamma}(N)_{1,j_2}| \leq c_{p0} \exp(-N_0)$ for any $j_1 \neq 1, j_2 \neq 1$.

Next, we focus on $\mathbf{\Pi}(\mathbf{\Pi}^\top)^k$ with $k = 1, 2, \dots, N-1$. Denote $\mathbf{\Gamma}_2(k) = \mathbf{\Pi}(\mathbf{\Pi}^\top)^k$. Similar to the analysis of induction above, we can use induction to prove that for $i \neq j$,

$$\mathbf{\Gamma}_2(2k-1)_{1,1} + \mathbf{\Gamma}_2(2k)_{1,1} \geq \mathbf{\Gamma}_2(2k-1)_{i,j} + \mathbf{\Gamma}_2(2k)_{i,j} - (1-p^2)^k.$$

The proof can be directly extended from the proof for $\Gamma(N)$ above. Then, we use this result to prove

$$[\Gamma_2(1)]_{1,1} + [\Gamma_2(2)]_{1,1} \geq [\Gamma_2(2k-1)]_{1,1} + [\Gamma_2(2k)]_{1,1}$$

for any $k \geq 2$. We also use induction to prove this. It is easily checked that $[\Gamma_2(1)]_{1,1} + [\Gamma_2(2)]_{1,1} = p^2 + (1-p)^2 \geq 3p^3(1-p) + 3p(1-p)^3 = [\Gamma_2(3)]_{1,1} + [\Gamma_2(4)]_{1,1}$. Suppose that $[\Gamma_2(2k-1)]_{1,1} + [\Gamma_2(2k)]_{1,1} \geq [\Gamma_2(2k+1)]_{1,1} + [\Gamma_2(2k+2)]_{1,1}$. Then, by $\Gamma_2(2k+1) + \Gamma_2(2k+2) = (\Gamma_2(2k-1) + \Gamma_2(2k))(\Pi^\top)^2$, we can get

$$\begin{aligned} \Gamma_2(2k+1)_{1,1} + \Gamma_2(2k+2)_{1,1} &= p^2[\Gamma_2(2k-1) + \Gamma_2(2k)]_{1,K-1} \\ &\quad + 2p(1-p)[\Gamma_2(2k-1) + \Gamma_2(2k)]_{1,1} \\ &\quad + (1-p)^2[\Gamma_2(2k-1) + \Gamma_2(2k)]_{1,3} \\ &\leq \Gamma_2(2k-1)_{1,1} + \Gamma_2(2k)_{1,1}, \end{aligned}$$

where the inequality is by the result demonstrated before. Therefore, we complete the induction and get that for $k \geq 1$,

$$\begin{aligned} \Gamma_2(1)_{1,1} + \Gamma_2(2)_{1,1} - \Gamma_2(2k+1)_{1,1} - \Gamma_2(2k+2)_{1,1} &\geq \Gamma_2(1)_{1,1} + \Gamma_2(2)_{1,1} - \Gamma_2(3)_{1,1} - \Gamma_2(4)_{1,1} \\ &= (p^2 + (1-p)^2)(1 - 3p(1-p)). \end{aligned}$$

Since $\Gamma_2(2)_{1,1} = 0$, we obtain for $k \geq 2$,

$$\Gamma_2(1)_{1,1} - \Gamma_2(k)_{1,1} \geq (p^2 + (1-p)^2)(1 - 3p(1-p)) := c_{p1}.$$

We provide an intuitive explanation from the perspective of random walks. $\Pi(\Pi^\top)^k$ represents first taking a step according to the transition Π , followed by k steps according to the transition $\Pi^\top$. With increasing k , the distribution of the possible state after $k+1$ steps is closer to uniform across all states, regardless of the starting state chosen, which can also be shown by Lemma E.2 and E.3. Thus, $\Pi(\Pi^\top)^k \mathbf{e}_1$ will converge to a vector corresponding uniform distribution, and $\Pi\Pi^\top \mathbf{e}_1$ represents the most sparse case with the largest value concentrating on the first entry.

Combining all the results obtained above, we can conclude that

$$\begin{aligned} &\left[\sum_{i=1}^{N-1} \Pi^i \cdot \Pi^\top \right] - \left[\sum_{i=1}^{N-1} \Pi^i \cdot (\Pi^\top)^k \right] \\ &= \mathbf{e}_1^\top \Gamma(N) \cdot \Gamma_2(1) \mathbf{e}_1 - \mathbf{e}_1^\top \Gamma(N) \cdot \Gamma_2(k) \mathbf{e}_1 \\ &= \Gamma(N)_{1,1} \Gamma_2(1)_{1,1} + \sum_{i=2}^K \Gamma(N)_{1,i} \Gamma_2(1)_{i,1} - \Gamma(N)_{1,1} \Gamma_2(k)_{1,1} - \sum_{i=2}^K \Gamma(N)_{1,i} \Gamma_2(k)_{i,1} \\ &\geq \Gamma(N)_{1,1} (\Gamma_2(1)_{1,1} - \Gamma_2(k)_{1,1}) + \min_{2 \leq i \leq K} \{ \Gamma(N)_{1,i} \} (1 - \Gamma_2(1)_{1,1}) - \max_{2 \leq i \leq K} \{ \Gamma(N)_{1,i} \} (1 - \Gamma_2(k)_{1,1}) \\ &= \Gamma_2(1)_{1,1} (\Gamma(N)_{1,1} - \min_{2 \leq i \leq K} \{ \Gamma(N)_{1,i} \}) - \Gamma_2(k)_{1,1} (\Gamma(N)_{1,1} - \max_{2 \leq i \leq K} \{ \Gamma(N)_{1,i} \}) \\ &\quad + \min_{2 \leq i \leq K} \{ \Gamma(N)_{1,i} \} - \max_{2 \leq i \leq K} \{ \Gamma(N)_{1,i} \} \\ &\geq (\Gamma_2(1)_{1,1} - \Gamma_2(k)_{1,1}) (\Gamma(N)_{1,1} - \min_{2 \leq i \leq K} \{ \Gamma(N)_{1,i} \}) + \min_{2 \leq i \leq K} \{ \Gamma(N)_{1,i} \} - \max_{2 \leq i \leq K} \{ \Gamma(N)_{1,i} \} \\ &\geq c_{p1} ((1-p)^{N_0-2} - c_{p0} \exp(-N_0)) - c_{p0} \exp(-N_0) \\ &\geq C_p, \end{aligned}$$

where C_p is a positive constant related to p . □

The following lemma shows the basic property of the positional embedding.

Lemma E.5. Assume that

$$\mathbf{p}_i = \left[\sin\left(\frac{i\pi}{M+1}\right), \sin\left(\frac{2i\pi}{M+1}\right), \dots, \sin\left(\frac{Mi\pi}{M+1}\right) \right]^\top$$

for $i \in [M]$. It holds that

$$\mathbf{p}_{i_1}^\top \mathbf{p}_{i_2} = \begin{cases} \frac{M+1}{2} & \text{for } i_1 = i_2; \\ 0 & \text{for } i_1 \neq i_2. \end{cases}$$

Proof of Lemma E.5. When $i_1 \neq i_2$ and $i_1 + i_2$ are even, we have

$$\begin{aligned}
\mathbf{p}_{i_1}^\top \mathbf{p}_{i_2} &= \sum_{j=1}^M \sin\left(\frac{j i_1 \pi}{M+1}\right) \sin\left(\frac{j i_2 \pi}{M+1}\right) \\
&= \sum_{j=0}^M \sin\left(\frac{j i_1 \pi}{M+1}\right) \sin\left(\frac{j i_2 \pi}{M+1}\right) \\
&= \frac{1}{4} \sum_{j=0}^M \left[\exp\left(i\pi \frac{i_1 - i_2}{M+1} j\right) + \exp\left(-i\pi \frac{i_1 - i_2}{M+1} j\right) \right. \\
&\quad \left. - \exp\left(i\pi \frac{i_1 + i_2}{M+1} j\right) - \exp\left(-i\pi \frac{i_1 + i_2}{M+1} j\right) \right] \\
&= \frac{1}{4} \cdot \frac{\exp(i\pi(i_1 - i_2)) - 1}{\exp\left(i\pi \frac{i_1 - i_2}{M+1}\right) - 1} + \frac{1}{4} \cdot \frac{\exp(-i\pi(i_1 - i_2)) - 1}{\exp\left(-i\pi \frac{i_1 - i_2}{M+1}\right) - 1} \\
&\quad - \frac{1}{4} \cdot \frac{\exp(i\pi(i_1 + i_2)) - 1}{\exp\left(i\pi \frac{i_1 + i_2}{M+1}\right) - 1} - \frac{1}{4} \cdot \frac{\exp(-i\pi(i_1 + i_2)) - 1}{\exp\left(-i\pi \frac{i_1 + i_2}{M+1}\right) - 1} \\
&= 0,
\end{aligned}$$

where the third equation is by $\sin(x) = \frac{\exp(ix) - \exp(-ix)}{2i}$, and the last inequality is by $\exp(i\pi k) = 1$ for even k . When $i_1 \neq i_2$ and $i_1 + i_2$ are odd, we have

$$\begin{aligned}
\mathbf{p}_{i_1}^\top \mathbf{p}_{i_2} &= \sum_{j=1}^M \sin\left(\frac{j i_1 \pi}{M+1}\right) \sin\left(\frac{j i_2 \pi}{M+1}\right) \\
&= \sum_{j=0}^M \sin\left(\frac{j i_1 \pi}{M+1}\right) \sin\left(\frac{j i_2 \pi}{M+1}\right) \\
&= \frac{1}{4} \sum_{j=0}^M \left[\exp\left(i\pi \frac{i_1 - i_2}{M+1} j\right) + \exp\left(-i\pi \frac{i_1 - i_2}{M+1} j\right) \right. \\
&\quad \left. - \exp\left(i\pi \frac{i_1 + i_2}{M+1} j\right) - \exp\left(-i\pi \frac{i_1 + i_2}{M+1} j\right) \right] \\
&= \frac{1}{4} \cdot \frac{\exp(i\pi(i_1 - i_2)) - 1}{\exp\left(i\pi \frac{i_1 - i_2}{M+1}\right) - 1} + \frac{1}{4} \cdot \frac{\exp(-i\pi(i_1 - i_2)) - 1}{\exp\left(-i\pi \frac{i_1 - i_2}{M+1}\right) - 1} \\
&\quad - \frac{1}{4} \cdot \frac{\exp(i\pi(i_1 + i_2)) - 1}{\exp\left(i\pi \frac{i_1 + i_2}{M+1}\right) - 1} - \frac{1}{4} \cdot \frac{\exp(-i\pi(i_1 + i_2)) - 1}{\exp\left(-i\pi \frac{i_1 + i_2}{M+1}\right) - 1} \\
&= -\frac{1}{2} \left(\frac{1}{\exp\left(i\pi \frac{i_1 - i_2}{M+1}\right) - 1} + \frac{1}{\exp\left(-i\pi \frac{i_1 - i_2}{M+1}\right) - 1} \right) \\
&\quad + \frac{1}{2} \left(\frac{1}{\exp\left(i\pi \frac{i_1 + i_2}{M+1}\right) - 1} + \frac{1}{\exp\left(-i\pi \frac{i_1 + i_2}{M+1}\right) - 1} \right) \\
&= 0,
\end{aligned}$$

where the third equation is by $\sin(x) = \frac{\exp(ix) - \exp(-ix)}{2i}$, the fifth inequality is by $\exp(i\pi k) = -1$ for odd k , and the last equation is by $\frac{1}{\exp(x)-1} + \frac{1}{\exp(-x)-1} = -1$. When $i_1 = i_2$, we have

$$\mathbf{p}_{i_1}^\top \mathbf{p}_{i_2} = \sum_{j=1}^M \sin\left(\frac{j i_1 \pi}{M+1}\right) \sin\left(\frac{j i_2 \pi}{M+1}\right)$$

$$\begin{aligned}
&= \sum_{j=0}^M \sin\left(\frac{ji_1\pi}{M+1}\right) \sin\left(\frac{ji_2\pi}{M+1}\right) \\
&= \frac{1}{4} \sum_{j=0}^M \left[\exp\left(i\pi \frac{i_1-i_2}{M+1}j\right) + \exp\left(-i\pi \frac{i_1-i_2}{M+1}j\right) \right. \\
&\quad \left. - \exp\left(i\pi \frac{i_1+i_2}{M+1}j\right) - \exp\left(-i\pi \frac{i_1+i_2}{M+1}j\right) \right] \\
&= \frac{M+1}{2} - \frac{1}{4} \sum_{j=0}^M \left[\exp\left(i\pi \frac{i_1+i_2}{M+1}j\right) + \exp\left(-i\pi \frac{i_1+i_2}{M+1}j\right) \right] \\
&= \frac{M+1}{2} - \frac{1}{4} \cdot \frac{\exp(i\pi(i_1+i_2)) - 1}{\exp\left(i\pi \frac{i_1+i_2}{M+1}\right) - 1} - \frac{1}{4} \cdot \frac{\exp(-i\pi(i_1+i_2)) - 1}{\exp\left(-i\pi \frac{i_1+i_2}{M+1}\right) - 1} \\
&= \frac{M+1}{2},
\end{aligned}$$

where the third equation is by $\sin(x) = \frac{\exp(ix) - \exp(-ix)}{2i}$, and the last inequality is by $\exp(i\pi k) = 1$ for even k . $\square$

F Additional Experiments

F.1 Additional Experiments on Random/Deterministic Walks

In this subsection, we provide additional experiments on synthetic data with $(K, N) = (20, 101)$. We consider the transformer model introduced in Section 2 with the length of the position embedding $M = 1000$. To train the model, we utilize gradient descent starting with zero initialization, where the learning rate $\eta = 1$ and the constant ϵ in the log-loss is set as $\epsilon = 0.1$. And, we run the gradient descent algorithm for $T = 50$ training epochs. Figure 8 and Figure 9 illustrate the experiments for $p = 0.5$ and $p = 1$ respectively. These experimental results match Theorem 3.1 and Theorem 4.1, which also strongly supports our theoretical results.

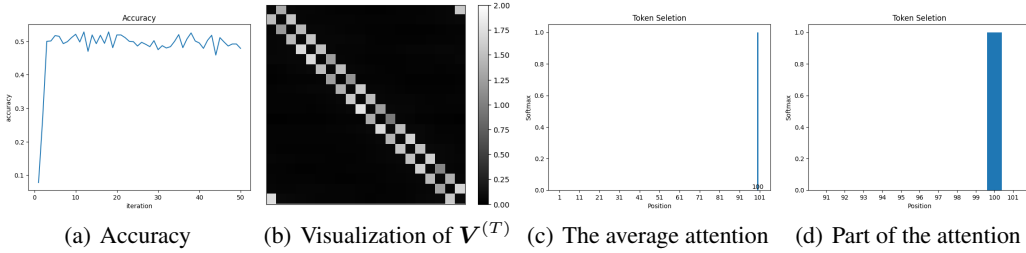


Figure 8: The results of the experiment for $p = 0.5$ with $(K, N) = (20, 101)$: (a) is the test accuracy; (b) is the visualization of $V^{(T)}$; (c) and (d) present the average attention of the test data with x-axis representing the position of the token and y-axis representing the attention score.

F.2 Additional Experiments on the Question Answering Tasks in Section 6

In this section, we conduct some additional experiments for Task 1 and Task 2 discussed in Section 6. We conduct some additional experiments extending the one-layer transformer model to a more complicated model by adding a fully connected layer with ReLU activation to the transformer model. The new model has the form

$$f_{\theta}(X) = A \cdot \text{ReLU}\left(VX \text{Softmax}\left(\widetilde{X}^{\top} W \widetilde{x}_N\right)\right), \quad (\text{F.1})$$

where $A \in \mathbb{R}^{K \times m}$, $V \in \mathbb{R}^{m \times K}$, $W \in \mathbb{R}^{(K+M) \times (K+M)}$ are the trainable parameter matrices, and m is the number of neurons in the fully connected layer. For Task 1 and Task 2, the length of

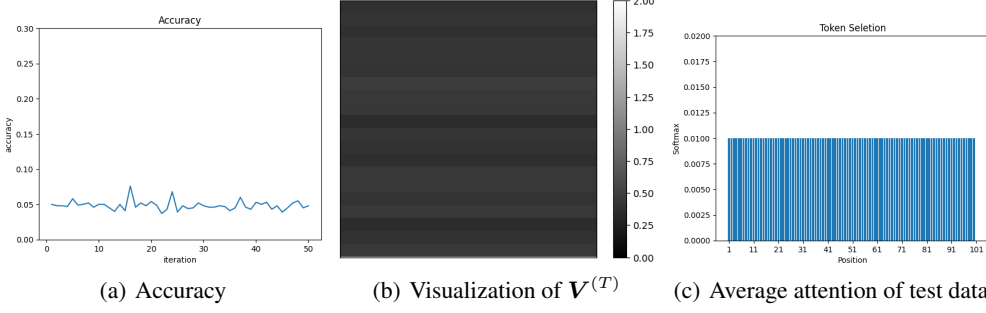


Figure 9: The results of the experiment for $p = 1$ with $K = 20, N = 101$. (a) is the prediction accuracy with x -axis representing the iteration and y -axis representing the accuracy. (b) is the visualization of V . (c) is the average attention of the test data with x -axis representing the position of the token and y -axis representing the attention score.

the vocabulary K and the length of each input sequence N are set as $(K, N) = (19, 17), (19, 19)$ respectively. In addition, we set the positional embedding $M = 1000$ and the number of neurons $m = 19$. To train the model, we consider the Gaussian random initialization $A_{ij}^{(0)}, V_{ij}^{(0)}, W_{ij}^{(0)} \sim N(0, \sigma^2)$ with $\sigma = 0.01$, and use gradient descent with learning rate $\eta = 0.1$. The constant ϵ in the log-loss is set as $\epsilon = 0.1$. Both experiments are conducted on 1024 training data and 1024 test data. Here, most of the settings remain the same as in the previous experiments in Section 6.

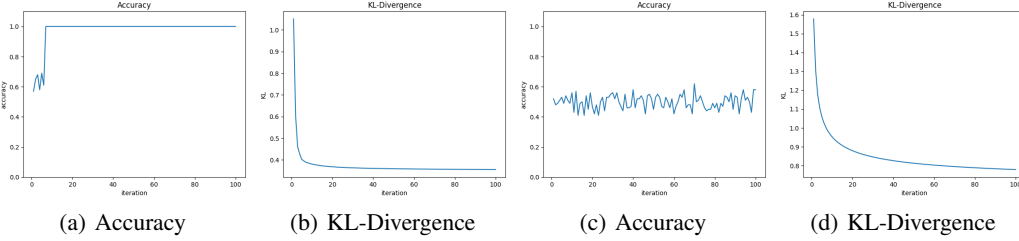


Figure 10: The results of the experiment conducted using a more complicated transformer for Task 1 and Task 2: (a) and (b) correspond to the experiment for Task 1; (c) and (d) correspond to the experiment for Task 2.

Figure 10 shows the experiment results using the more complicated transformer in (F.1) to learn Task 1 and Task 2. In Figure 10(a) and Figure 10(c), we present the test accuracy achieved by the transformer model in learning Task 1 and Task 2 respectively. In Figure 10(b) and Figure 10(d), we first normalize the output of the trained transformer model to get a K -dimensional vector, representing the prediction distribution of K words. Then, we report the KL-divergence between this prediction distribution and the true distribution of $y|x_1, x_2, \dots, x_{N-1}$. The experiment results show a clear difference between the performances of the transformer model in the two tasks. In Task 1, the trained transformer model can successfully approach the optimal accuracy (100%) within 100 iterations. However, in Task 2, the test accuracy always remains around 50%, which is the accuracy of a random guess.

Despite using a more complicated transformer model with an additional feedforward layer of non-linearities compared to the one considered in our theoretical analysis and previous experiments, the experimental results are still similar to those reported in Section 6. These results demonstrate that more complex transformer models may still struggle with the relatively 'simple' Task 2 but excel at the relatively 'difficult' Task 1. This indicates that our findings can be applied to cases involving additional nonlinearities, implying their applicability to more complex and general conditions.

F.3 Visualizations of the value matrix and Softmax scores corresponding to Figure 6

In this section, we present visualizations of the value matrix and softmax scores corresponding to Figure 6.

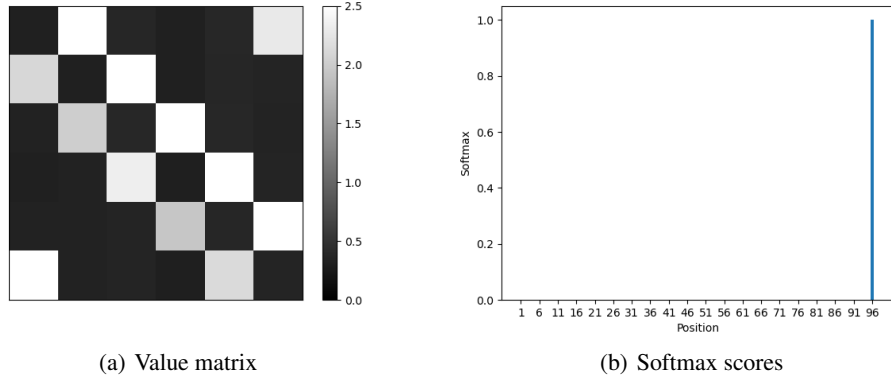


Figure 11: Visualizations of the trained value matrix and the softmax scores corresponding to Figure 6(a) and Figure 6(b).

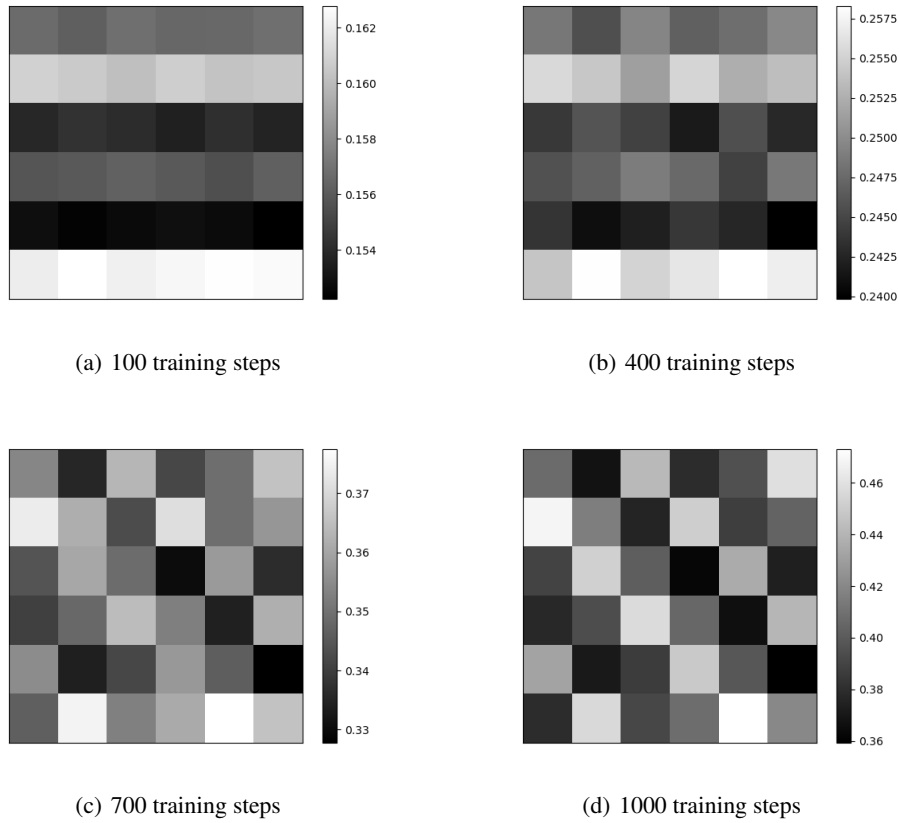


Figure 12: Visualizations of the value matrix corresponding to Figure 6(c) and Figure 6(d) at 100, 400, 700, and 1000 training steps, respectively.

Figure 11 visualizes the trained value matrix and the softmax scores attached to all tokens, corresponding to Figure 6(a) and Figure 6(b). It shows that $\mathbf{V}$ can recover the transition matrix $\mathbf{\Pi}^\top$, and the softmax score attached to the direct parent token is the largest and close to 1. These findings demonstrate that our positive results for learning random walks in Theorem 3.1 are fairly robust to random initialization.

Figure 12 and Figure 13 display the value matrix and the weighted average token $\mathbf{X} \cdot \mathcal{S}(\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{x}}_N)$ corresponding to Figure 6(c) and Figure 6(d) at 100, 400, 700, and 1000 training steps, respectively. In Figure 13, the weighted average token embedding is calculated based on one sequence $\mathbf{X}$ as an example, while the variance of each dimension of the weighted average token is calculated across 1000 sequences. We can observe that when considering the case for $p = 1$ with random initialization,

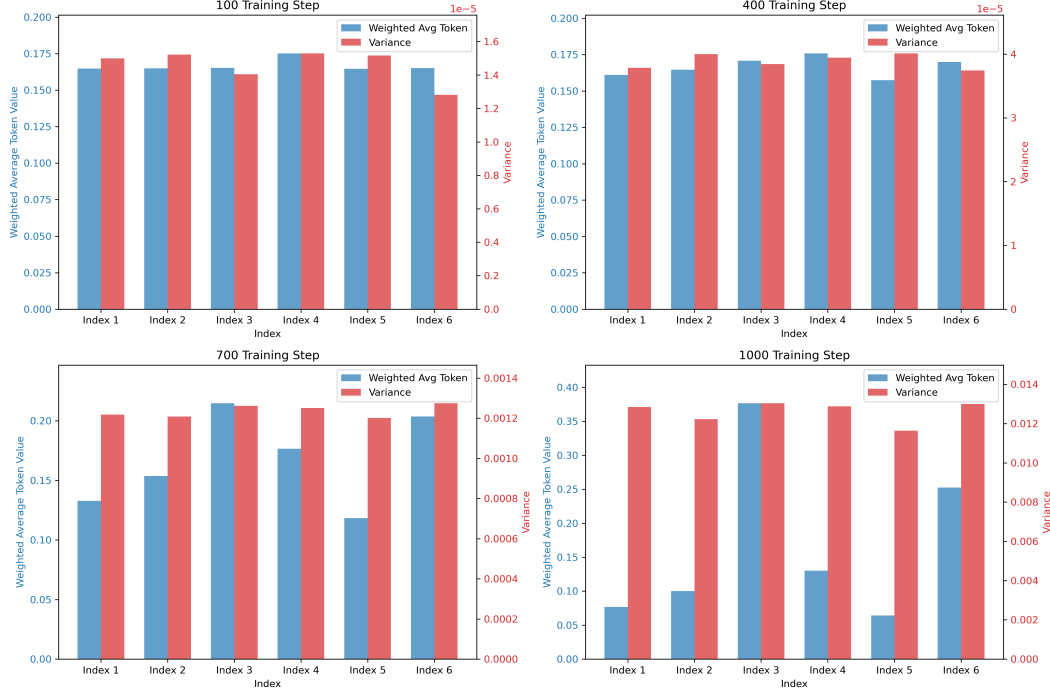


Figure 13: Visualizations of the weighted average token $\mathbf{X} \cdot \mathcal{S}(\tilde{\mathbf{X}}^\top \mathbf{W} \tilde{\mathbf{x}}_N)$ corresponding to Figure 6(c) and Figure 6(d) at 100, 400, 700, and 1000 training steps, respectively.

the value matrix and the weighted average token embedding remain approximately proportional to the all-one matrix and the all-one vector within 700 iterations, aligning with the results stated in Theorem 4.1. Additionally, the variances of all dimensions are close to 0, which shows that this result is general for different sequences. However, after 1000 iterations, $\mathbf{V}$ is no longer proportional to an all-one matrix, and the softmax score for one arbitrary token becomes much larger than the others. As discussed in Section 4, the failure case for $p = 0, 1$ is indeed due to an unbreakable symmetry caused by zero initialization. The random initialization may break this symmetry, enabling the transformer to successfully learn the optimal predictor.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: This paper mainly focus on theory, and our main results are stated in the abstract and introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We have discussed limitations on the studies of the impact of random initialization and the relatively simple setting of one-layer transformers.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We have clearly stated all theoretical assumptions in the problem setting section and in the formal theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Experiments in this paper are mainly to back up the theoretical results and are not too complicated. We have provided detailed descriptions on the experimental setups.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: Experiments in this paper are on synthetic data and are relatively simple. The purpose is mainly to back up our theory.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experiments are relatively simple and the purpose is mainly to back up our theory. We have clearly explained all the experiment setups.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The experiments in this paper are to verify our theory, and there is no need to analyze statistical significance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: The experiments in this paper are very simple and are mainly to verify our theory. There is no need to specify compute resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We confirm that this research conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper mainly focuses on theoretical analyses on simple one-layer transformer models. We do not see any potential negative social impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper mainly focuses on theoretical analyses on simple one-layer transformer models. We do not see any potential risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This paper mainly focuses on theoretical analyses on simple one-layer transformer models. We have provided appropriate citations and there is no other relevant assets to credit.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: We have only used LLMs for editing purposes.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.