Mechanics of Next Token Prediction with Self-Attention

Yingcong Li[*]
University of California, Riverside
yli692@ucr.edu

Yixiao Huang[*]
University of Michigan
yixiao.huang@my.cityu.edu.hk

M. Emrullah Ildiz
University of California, Riverside
mildi001@ucr.edu

Ankit Singh Rawat
Google Research
ankitsrawat@google.com

Samet Oymak
University of Michigan
oymak@umich.edu

Abstract

Transformer-based language models are trained on large datasets to predict the next token given an input sequence. Despite this simple training objective, they have revolutionized natural language processing within a short timeframe. Underlying this success is the self-attention mechanism. In this work, we ask: *What does 1-layer self-attention learn from next-token prediction?* We show that training self-attention with gradient descent learns an automaton which generates the next token based on the last input token and a data-induced *token hierarchy* as follows: **(1)** Given input sequence, self-attention implements a *hard retrieval* to precisely select the *highest priority input tokens*. **(2)** It then creates a *soft convex composition* of these tokens to obtain the next token. Under suitable conditions, we precisely characterize this hierarchy through extracted from the training data. The priority of a token is governed by the *strongly-connected components (SCC)* of the which induces cyclic and acyclic subgraphs. Self-attention implicitly learns the acyclic subgraph through a max-margin bias to enforce a strict priority order among SCCs. Cyclic subgraph is responsible for the soft composition by distributing softmax probabilities to the tokens within a SCC. We hope these findings shed light on self-attention's distinct mechanics of processing sequential data and pave the path towards demystifying more complex transformer architectures.

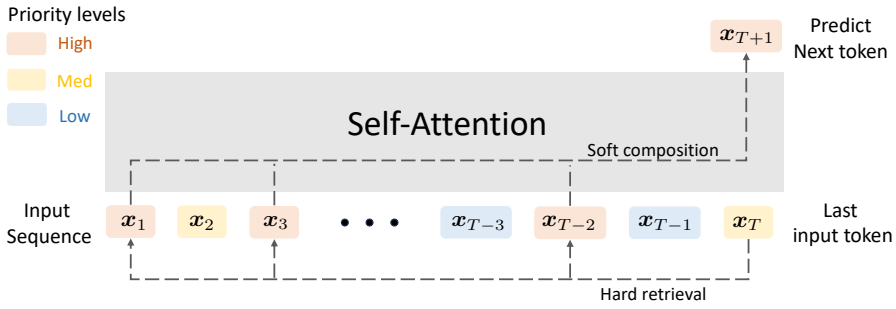


Figure 1: Overview of next-token prediction. The last input token is used as a query token and fed into a single-layer self-attention together with the input sequence to predict the next token. We show that self-attention first implements a hard retrieval to select the highest priority tokens and then outputs a soft convex composition of these tokens as the next token.

1 Introduction

Language modeling as enabled by Transformer architecture Vaswani et al. (2017) and seemingly simple training objectives such as next-token prediction Radford et al. (2018, 2019) has not only led to breakthroughs

[*]Equal contribution.

in the field of natural language processing (NLP) Brown et al. (2020); Chowdhery et al. (2022); OpenAI (2023); Touvron et al. (2023), but rather straightforward adaptations of this symbiosis between Transformers and next-token prediction task have also realized remarkable performance in other domains, including vision Chen et al. (2020), speech Chung & Glass (2020), reinforcement learning Chen et al. (2021), and even protein design Ferruz et al. (2022); Nijkamp et al. (2022). This widespread empirical success is often attributed to the (self-)attention mechanism of Transformers that produces high-quality contextual representations needed to realize excellent prediction performance in a wide range of domains. However, a rigorous understanding of how Transformers can learn such high-quality representations by solving next-token prediction task via natural algorithms such as gradient descent is largely missing from the literature.

This work aims to bridge this gap between the empirical success and principled understanding of Transformer-based language modeling by shedding light on the optimization landscape and key implicit biases faced by the self-attention mechanism in solving the next-token prediction task. In particular, focusing on a *single-layer* self-attention model with linear classification head, and solving the next-token prediction task, we consider the following questions:

- *What relationships in the training data are captured by the single-layer self-attention model?*
- *How exactly do these relationships dictate the optimization geometry of natural algorithms such as gradient descent?*

We show that the answers to both of these questions are intertwined which we achieve by significantly expanding the recently proposed framework that connects learning with Transformers to the celebrated support vector machines (SVMs) Tarzanagh et al. (2023b,a).

Given training data in the form of a collection of (input sequence, next-token) pairs, we construct directed graphs among the tokens in the vocabulary, namely , capturing the priority order or hierarchy among different tokens as observed in the training data. A *strongly connected component* (SCC) in a corresponds to the tokens that can potentially follow each other, indicating the absence of a priority order among those tokens. Notably, label token has higher (or equal) priority compared to the other tokens in the sequence, as illustrated in Figure 2. Based on these , we propose an SVM formulation, namely (Graph-SVM), for self-attention learning with the objective of suppressing lower priority tokens in favour of higher priority tokens while predicting the next token for an input sequence (cf. Section 3.2).

Subsequently, we focus on the regularization path and gradient descent algorithms to learn the self-attention model for next-token prediction via empirical risk minimization (ERM). Our main contributions are as follows:

1. We rigorously show the implicit bias of the solution obtained by regularization path Rosset et al. (2003); Suggala et al. (2018); Ji et al. (2020b) towards (Graph-SVM) solution, and introduce *cyclic correction* component to capture the behaviour on tokens with equal priority (cf. Section 4).
2. As for the optimization landscape of the gradient descent algorithm, we consider the setting where log-loss defines the ERM objective and establish a global convergence result similar to the regularization path algorithm (cf. Section 5).
3. Finally, we consider more general settings and characterize local directional convergence while highlighting their implicit bias towards SVM solution as defined based on the *pseudo* TPGs constructed by gradient descent solution (cf. Section 6).

2 Related Work

Inspired by the increasing popularity of Transformer-based models, a large number of research efforts have focused on developing theoretical understanding of various aspects of such models. Yun et al. (2020a); Fu et al. (2023) studied the expressive power of Transformers and showed that they are universal approximators for sequence-to-sequence functions. A similar result for efficient variants of Transformers based on sparse attention was presented in Yun et al. (2020b). Edelman et al. (2022) studied bias of single attention layer towards representing sparse functions of input sequence with favourable generalization behaviour. Interestingly, Baldi & Vershynin (2023) explored key building blocks of attention mechanism beyond modern neural networks and studied the functional capacity of the resulting attention-based models. Other lines of theoretical efforts have focused on explaining various properties of Transformer-based models, including rank collapse Dong et al. (2021) and realization of in-context learning Xie et al. (2022); Garg et al. (2022); Akyürek et al. (2023); Von Oswald et al. (2023); Li et al. (2023b).

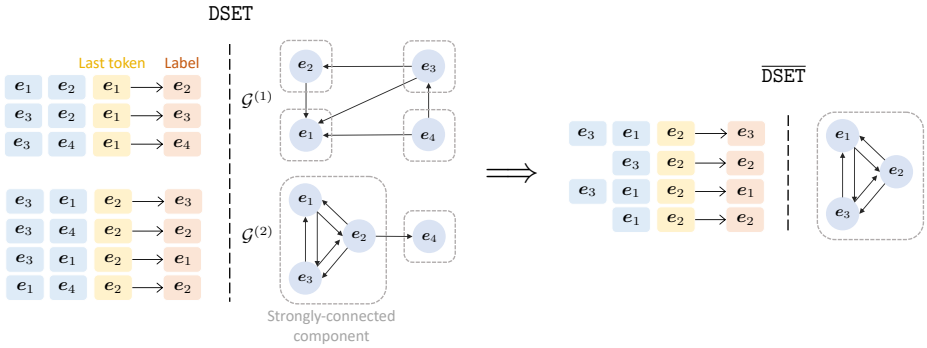


Figure 2: Illustration of token-priority graph (TPG). Given the input sequences and labels (next tokens), we construct the TPGs $\{\mathcal{G}^{(k)}\}_{k=1}^{K}$ according to the last token. **Left hand side:** Two TPGs $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ are constructed using the samples with $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ as the last tokens, respectively. In each graph, directed edges (label→input token) are added, and based on the token relations, each graph can be partitioned into different strongly-connected components (SCCs, presented in dashed circles). **Right hand side:** We consider a cyclic subdataset $\overline{\texttt{DSET}}$ by removing all the singleton SCCs, as well as their corresponding edges (see Definition 1). Then, $\overline{\texttt{DSET}}$ contains non-singleton SCCs only as shown on the right. More details are deferred to Section 3.

Unlike these prior work, we focus on optimization-theoretic analysis of attention-based models for the next-token prediction objective. Our work sheds light on the implicit bias of underlying optimization problem towards SVM formulations, which builds on the recent research efforts Tarzanagh et al. (2023b,a). However, different from these prior efforts that deal with traditional (supervised) classification tasks, we focus on next-token prediction task – the main workhorse of Transformer-based language modeling. We emphasize that our study requires us to introduce multiple novel concepts such as to capture token hierarchy and cyclic correction component to characterize the impact of equal priority tokens. Notably, several recent efforts Jelassi et al. (2022); Li et al. (2023a); Oymak et al. (2023) have also analyzed optimization and generalization dynamics of attention-based models. However, these works again only focus on traditional classification tasks and consider simplifications of the attention mechanism Jelassi et al. (2022) or work with strict statistical data assumptions Jelassi et al. (2022); Li et al. (2023a); Oymak et al. (2023). In contrast, we provide a detailed optimization-theoretic treatment of the original (non-linear input dependent) attention mechanism without any statistical assumption on the underlying data. In a closely related recent work, Tian et al. (2023) studied training dynamics of next-token prediction with analysis restricted to a specific statistical data model and requires working with very long input sequences ($T \to \infty$). In contrast, we characterize the implicit bias of self-attention learning to novel SVM formulations without any such assumptions on the data model or sequence lengths.

We would also like to note the rich literature on studying implicit bias of gradient-based optimization methods (see, e.g., Soudry et al. (2018); Gunasekar et al. (2018); Ji et al. (2020a); Ji & Telgarsky (2021); Kini et al. (2021); Li et al. (2019); Blanc et al. (2020); Qian & Qian (2019); Wang et al. (2021) and references therein). However, this prior work does not focus on the optimization landscape of learning Transformer-based models and thus, does not provide specific insights into their inner-workings, which is the main objective of our work.

3 Problem Setup

Notation. Let $[n]$ denote the set $\{1, \cdots, n\}$. For a subspace $\mathcal{S}$, let $\mathcal{S}^{\perp}$ denote the space orthogonal to $\mathcal{S}$ and $\boldsymbol{\Pi}_{\mathcal{S}}$ denote the subspace projection on $\mathcal{S}$ with respect to Euclidean distance.

Next-token prediction problem. Let K be the vocabulary size with $\boldsymbol{E} = [\boldsymbol{e}_1 \ \ldots \ \boldsymbol{e}_K]^\top \in \mathbb{R}^{K\times d}$ denoting the embedding matrix consisting of d-dimensional token embeddings for the K tokens in the vocabulary. The next-token prediction is a multi-class classification problem and the goal is to predict the ID $y \in [K]$ of the next token given an input sequence $\boldsymbol{X} = [\boldsymbol{x}_1 \ \ldots \ \boldsymbol{x}_T]^\top \in \mathbb{R}^{T\times d}$, where $\boldsymbol{x}_t \in \boldsymbol{E}, t \in [T]$.

Suppose that we have a training dataset $\texttt{DSET} = \big\{(\boldsymbol{X}_i, y_i) \in \mathbb{R}^{T_i\times d} \times [K]\big\}_{i=1}^{n}$ consisting of n sequences where we allow the sequences to have different lengths T_i, $i \in [n]$. Throughout this paper, we use $x_{it} \in [K]$ to denote the scalar token ID corresponding to the t-th token $\boldsymbol{x}_{it} \in \mathbb{R}^d$ of the input sequence $\boldsymbol{X}_i$. Note that the i-th sample in the dataset $(\boldsymbol{X}_i, y_i)$ essentially corresponds to an extended (T_i+1)-length sequence $[\boldsymbol{e}_{x_{i1}} \ \ldots \ \boldsymbol{e}_{x_{iT_i}} \ \boldsymbol{e}_{y_i}]^\top \in \mathbb{R}^{(T_i+1)\times d}$. Here, all $T_i + 1$ elements of the extended sequence, including the input

tokens and the label token, belong to the same discrete set corresponding to the rows of the embedding matrix $\boldsymbol{E}$. Thus, the next-token prediction task enabling language modeling differs from many *traditional* multi-class classification setups where input features and labels belong to different spaces.

Self-attention model. We consider a *single-layer* self-attention model when making a prediction on a given input sequence $\boldsymbol{X} \in \mathbb{R}^{T\times d}$. Following the previous work Tarzanagh et al. (2023a), we denote the combined key-query weights by a trainable $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ matrix, and assume identity value matrix. Let $\bar{\boldsymbol{x}} := \boldsymbol{x}_T$ be the last token of the input sequence $\boldsymbol{X}$. Then, the (single-layer) self-attention outputs the following embedding to predict the next-token ID y:

$$f_{\boldsymbol{W}}(\boldsymbol{X}) = \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}}), \tag{1}$$

where $\mathbb{S}(\cdot)$ denotes the softmax operation which facilitates weighing tokens of $\boldsymbol{X}$ based on the data-dependent probabilities $\mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}})$. Note that the output embedding $f_{\boldsymbol{W}}(\boldsymbol{X}) \in \mathbb{R}^d$ in (1) is a weighted combination of the input token embeddings in $\boldsymbol{X}$.

As for the underlying next-token prediction task, we rely on a linear classifier defined by $\boldsymbol{C} \in \mathbb{R}^{K\times d}$ that provides per-class predictions $\hat{\boldsymbol{y}} \in \mathbb{R}^K$ based on the output of the self-attention model (cf. (1)).

Empirical risk minimization (ERM) problem. Let $\ell(y, \hat{\boldsymbol{y}})$ be the loss function that assesses the mismatch between prediction $\hat{\boldsymbol{y}}$ and true label $y \in [K]$. Given training dataset DSET, we consider the empirical risk minimization problem with the following objective:

$$\mathcal{L}(\boldsymbol{C}, \boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \boldsymbol{C}\boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}\bar{\boldsymbol{x}}_i)).$$

Throughout this paper, we fix the linear classification head $\boldsymbol{C}$[2] and assume $\|\boldsymbol{C}\|_F < \infty$. Accordingly, we define $\mathcal{L}(\boldsymbol{W}) := \mathcal{L}(\boldsymbol{C}, \boldsymbol{W})$. Note that even though the classification head is fixed and linear, the problem of learning attention parameters $\boldsymbol{W}$ via ERM is non-linear due to the softmax operator. In this work, we focus on this exact problem and consider the following two algorithms to optimze for $\boldsymbol{W}$ with a fixed $\boldsymbol{C}$:

1. Regularization path: Given $R > 0$, $\boldsymbol{W} \in \mathbb{R}^{d\times d}$

$$\bar{\boldsymbol{W}}_R = \arg\min_{\|\boldsymbol{W}\|_F \le R} \mathcal{L}(\boldsymbol{W}). \tag{Algo-RP}$$

2. Gradient descent: Given starting point $\boldsymbol{W}(0) \in \mathbb{R}^{d\times d}$ and step size $\eta > 0$, for $\tau \ge 0$,

$$\boldsymbol{W}(\tau+1) = \boldsymbol{W}(\tau) - \eta\nabla\mathcal{L}(\boldsymbol{W}(\tau)). \tag{Algo-GD}$$

The next-token prediction task aims to capture various patterns present in the underlying dataset. Towards this, we introduce that summarizes the sequential priority orders presented in the training data. As we will see later that TPGs play a crucial role in characterizing the optimization geometry for both (Algo-RP) and (Algo-GD) above.

3.1 Token-priority Graph of Dataset

A is a directed graph with at most K nodes corresponding to the elements in the vocabulary. We associate the dataset $\texttt{DSET} = \{(\boldsymbol{X}_i, y_i)\}_{i=1}^n$ with multiple $\{\mathcal{G}^{(k)}\}_{k=1}^K$, with each TPG focusing on a subset of the dataset comprising of those input sequences that agree on the last token $\bar{\boldsymbol{x}}$. Concretely, we construct $\mathcal{G}^{(k)}$'s as follows:

1. Split DSET into K subsets $\{\texttt{DSET}^{(k)}\}_{k=1}^K$ with $\texttt{DSET}^{(k)}$ containing all input sequences that end with the same last token $\bar{\boldsymbol{x}} = \boldsymbol{e}_k$.
2. For each $(\boldsymbol{X}, y) \in \texttt{DSET}^{(k)}$ and for all (y, x) pairs in $(\boldsymbol{X}, y)$ where x is the corresponding token ID of $\boldsymbol{x} \in \boldsymbol{X}$, add a directed edge $(y \to x)$ to $\mathcal{G}^{(k)}$.

An illustration is provided in Fig. 2(Left). Here, we construct two ($\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$) based on the last tokens (depicted in yellow), and the directed edges are presented as arrows starting from labels (orange) to input tokens (blue/yellow) within each sequence in the figure. Note that nodes of each $\mathcal{G}^{(k)}$ constitute a subset of the indices $[K]$. The edges in $\mathcal{G}^{(k)}$ capture the priorities across the tokens in an extended data sequence,

[2]Specifically, we assume well-pretrained head $\boldsymbol{C}$ such that $\ell(y, \boldsymbol{e}_k)$ returns the minimal risk when $k = y$.

conditioned on the last token of the input being $\bar{\boldsymbol{x}} = \boldsymbol{e}_k$. We will see that if there is a cycle, i.e., $y \to x$ and $x \to y$ are both directionally reachable in the graph, then the self-attention learnt via next-token prediction task can assign comparable priorities to the tokens x and y. In contrast, if y always *dominates* x, i.e., $y \to x$ is reachable but $x \not\to y$, then, when x and y are both present in an input sequence, self-attention will suppress x and only select y through an SVM mechanism along the line of Tarzanagh et al. (2023a).

Strongly-connected components in TPGs. To formalize the aforementioned SVM mechanism, we need the notion of *strongly-connected components* (SCCs). A directed graph is strongly connected if every node in the graph is reachable from every other node. SCCs of a directed graph form a partition into subgraphs that are themselves strongly connected. Given the $\{\mathcal{G}^{(k)}\}_{k=1}^{K}$ associated with the dataset DSET, we can split the directed graph $\mathcal{G}^{(k)}$ into its SCCs, denoted by $\{\mathcal{C}_i^{(k)}\}_{i=1}^{N_k}$. Note that the number of SCCs in $\mathcal{G}^{(k)}$, as denoted by N_k, is at most the number of nodes in $\mathcal{G}^{(k)}$, which is upper bounded by the vocabulary size K. Furthermore, by definition, different SCCs within a graph consist of distinct nodes, i.e., $\mathcal{C}_i^{(k)} \bigcap \mathcal{C}_j^{(k)} = \emptyset$, for $i \neq j$. Now, returning to Fig. 2, each of the dashes circles represents an SCC. $\mathcal{G}^{(1)}$(upper) contains four SCCs and therefore, all tokens within the graph have strict priority orders. In contrast, $\mathcal{G}^{(2)}$(lower) consists of two SCCs, with one containing three nodes. Following the arrows, we can see that all the tokens/nodes within this specific SCC are directional reachable.

Before formally connecting and their SCCs to the SVM mechanism that enables next-token prediction, we introduce some necessary graph-related notation. Given a directed graph $\mathcal{G}$, for $i, j \in [K]$ such that $i \neq j$:

- $i \in \mathcal{G}$ denotes that the node i belongs to $\mathcal{G}$.
- $(i \Rightarrow j) \in \mathcal{G}$ denotes that the directed edge $(i \to j)$ is present in $\mathcal{G}$ but $j \to i$ is not reachable.
- $(i \asymp j) \in \mathcal{G}$ means that two nodes (i, j) are in the same SCC of $\mathcal{G}$.

For $i \neq j \in [K]$, $\mathcal{C}_i \bigcap \mathcal{C}_j = \emptyset$ implies that $(i \Rightarrow j)$ and $(i \asymp j)$ cannot both present in the same graph $\mathcal{G}$.

3.2 SVM Bias of Self-attention Learning

The main contribution of this paper is to establish that capture the optimization geometry of the next-token prediction problem. We will show that the self-attention model learnt via either (Algo-RP) or (Algo-GD) converges to the solution of an SVM defined by the of the underlying dataset DSET. In particular, given $(\mathcal{G}_k^{(k)})_{k=1}^{K}$, we introduces the following SVM formulation:

$$\boldsymbol{W}^{\mathrm{mm}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \qquad \text{(Graph-SVM)}$$

$$\text{s.t. } (\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W} \boldsymbol{e}_k \begin{cases} = 0 & \forall (i \asymp j) \in \mathcal{G}^{(k)} \\ \geq 1 & \forall (i \Rightarrow j) \in \mathcal{G}^{(k)} \end{cases} \quad k \in [K].$$

When $(i \Rightarrow j)$, token ID i has a higher *priority* than j and hence, the SVM problem (Graph-SVM) aims to find a $\boldsymbol{W}$ such that $\boldsymbol{W}\boldsymbol{e}_k$ achieves higher correlation to token embedding $\boldsymbol{e}_i$ than $\boldsymbol{e}_j$, that is, $\boldsymbol{e}_i^\top \boldsymbol{W} \boldsymbol{e}_k > \boldsymbol{e}_j^\top \boldsymbol{W} \boldsymbol{e}_k$, and then softmax will assign higher probability to the token i. While for $(i \asymp j)$, there is not strict priority order between i and j, and hence we set the correlation difference equal to zero to prevent the SVM solution $\boldsymbol{W}$ from distinguishing them. The existence of the solution $\boldsymbol{W}^{\mathrm{mm}}$ ensures the separability of tokens i's from the j's for all pairs $(i \Rightarrow j) \in \mathcal{G}^{(k)}$. Additionally, if for all $k \in [K]$, the number of SCCs[3] $N_k \leq 1$, then $\boldsymbol{W}^{\mathrm{mm}} = 0$.

Given DSET, we define a sub-dataset induced by the interactions within SCCs.

Definition 1 (Cyclic subdataset) *Given any training sample $(\boldsymbol{X}, y) \in$ DSET, we obtain the corresponding sample $(\boldsymbol{X}', y) \in \overline{\text{DSET}}$ by removing all tokens in $\boldsymbol{X}$ that satisfy $(y \Rightarrow x)$ in the corresponding .*

As illustrated in Fig. 2, the right handside presents the cyclic subdataset $\overline{\text{DSET}}$ of the left handside DSET. In $\mathcal{G}^{(1)}$, all nodes are separated into different SCCs, and therefore, none of them is present in $\overline{\text{DSET}}$; while in $\mathcal{G}^{(2)}$, token $\boldsymbol{e}_1, \boldsymbol{e}_2$ and $\boldsymbol{e}_3$ are reachable from each other, and then are utilized to construct $\overline{\text{DSET}}$ while $\boldsymbol{e}_4$ is removed from the dataset. Note that cyclic subdataset $\overline{\text{DSET}}$ provides a self-contained sub-problem that solely

[3]Note that $N_k = 0$ implies that within DSET, there are not training samples whose input sequence has $\boldsymbol{e}_k$ as its last token; equivalently, $\text{DSET}^{(k)} = \emptyset$.

focuses on intra-SCC edges. Below, in contrast, we introduce the concept of acyclic dataset which implies that the next-token prediction task always encounters a strict priority order among tokens within each TPG. This corresponds to the setting where all SCCs $((\mathcal{C}_i^{(k)})_{i=1}^{N_k})_{k=1}^{K}$ are all singletons; or equivalently, $\overline{\texttt{DSET}} = \emptyset$.

Definition 2 (Acyclic dataset) *We call* `DSET` *acyclic if all of its are directed acyclic graphs.*

For an acyclic dataset, we have no i and j, such that $i \asymp j \in \mathcal{G}^{(k)}$, for all $i, j, k \in [K]$. Thus, SVM formulation (Graph-SVM) reduces to the following simpler form.

$$\boldsymbol{W}^{\mathrm{mm}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \qquad \text{(Acyc-SVM)}$$
$$\text{s.t.} \quad (\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W} \boldsymbol{e}_k \geq 1 \quad \forall (i \Rightarrow j) \in \mathcal{G}^{(k)},\ k \in [K].$$

3.3 Technical Assumptions

In what follows, we work with a few assumptions that will make the optimization landscape of the underlying learning problem more benign. Below, we introduce these assumptions along with their justifications.

Assumption 1 (Loss function) *For any $\boldsymbol{v} \in \mathbb{R}^d$, $y \in [K]$, we have $\ell(y, \boldsymbol{v}) = \ell(\boldsymbol{c}_y^\top \boldsymbol{v})$ where $\boldsymbol{c}_y \in \mathbb{R}^d$ is the y-th row of linear head $\boldsymbol{C}$. Additionally, $\ell : \mathbb{R} \to \mathbb{R}$ is strictly decreasing and $|\ell'|$ is bounded.*

Under Assumption 1 above, we can rewrite the ERM problem in a simplified form as follows:

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)). \qquad \text{(ERM)}$$

Recall that, single-layer self-attention outputs a convex combination of the input tokens (cf. (1)). If tokens are linearly independent, the only way model can output the embedding $\boldsymbol{e}_y$ corresponding to the target label y would be if $\boldsymbol{e}_y$ was among the input sequence. This motivates the following realizability assumption.

Assumption 2 (Realizable labels) `DSET` *is called realizable if for any $(\boldsymbol{X}, y) \in$* `DSET`*, the token $\boldsymbol{e}_y$ is contained in the input sequence $\boldsymbol{X}$.*

Note that if $(\boldsymbol{X}, y)$ is not realizable, self-attention would select $\boldsymbol{e}_{\hat{y}} \neq \boldsymbol{e}_y$ and the SVM formula would be established via separating $\hat{y}$ from the other tokens in the sequence instead of the true label y. Additionally, when head satisfies $\boldsymbol{C}\boldsymbol{E}^\top = \boldsymbol{I}$ and the model can only make a random prediction over the output labels (since any output of the self-attention model would result in the same training risk); consequently, such non-realizable examples will not play roles in optimizing $\boldsymbol{W}$, i.e. $\nabla_{\boldsymbol{W}} \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)) = 0$.

4 Implicit Bias of Self-Attention

To proceed, we wish to identify the implicit bias of self-attention on general next-token prediction problems. This necessitates SVM formulations (Acyc-SVM) and (Graph-SVM) as well as additional ideas.

4.1 Acyclic Dataset

We start by establishing the global convergence of algorithm Algo-RP under the scenario where dataset `DSET` is acyclic (see Definition 2).

Since $\boldsymbol{C}$ choice is under our control and $\boldsymbol{E}$ is fixed, we choose $\boldsymbol{C}$ such that y-th row of $\boldsymbol{C}$ picks up token y and suppresses other tokens, where $y \in [K]$ is the corresponding true label of each sample. Concretely, we enforce the following criteria.

Assumption 3 *For $y \in [K]$, $\arg\max_{k \in [K]} \boldsymbol{c}_y^\top \boldsymbol{e}_k = y$.*

Lemma 1 *Consider acyclic dataset* `DSET` *per Def. 2. Suppose Assumptions 1,2 and 3 hold, then for any finite $\boldsymbol{W} \in \mathbb{R}^{d \times d}$, training risk obeys $\mathcal{L}(\boldsymbol{W}) > \mathcal{L}_\star := \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{e}_{y_i})$. Additionally, if (Acyc-SVM) is feasible and denote its solution as $\boldsymbol{W}^{mm}$, $\lim_{R \to \infty} \mathcal{L}(R \cdot \boldsymbol{W}^{mm}) = \mathcal{L}_\star$.*

Assumption 3 and Lemma 1 ensure that a reduced ERM training risk (see (ERM)) is indicative of enhanced prediction accuracy, and the only way for attention to make a correct prediction on class k is to output the vector $\boldsymbol{e}_k$, i.e., $f_{\boldsymbol{W}}(\boldsymbol{X}) = \boldsymbol{e}_k$. The next theorem states the directional bias of self-attention on the acyclic dataset (per Definition 2) towards the solution of (Acyc-SVM).

Theorem 1 *Suppose* DSET *is acyclic per Definition 2 and Assumptions 1,2, and 3 hold. Suppose* (Acyc-SVM) *is feasible with $\boldsymbol{W}^{mm}$ denoting its solution. Then, Algorithm Algo-RP satisfies* $\lim_{R\to\infty} \frac{\bar{\boldsymbol{W}}_R}{R} \to \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F}$.

Recall that a dataset being acyclic implies that there is strict priority order among tokens in each . Theorem 1 establishes the implicit bias of self-attention model for next-token prediction problem in the presence of such strict priority order. It demonstrates that once the SVM problem (Acyc-SVM) is feasible, the regularized path of optimizing (ERM) converges directionally toward it solution $\boldsymbol{W}^{\text{mm}}$.

4.2 General Dataset

Now we focus on the general data setting and establish connections between algorithm Algo-RP and SVM solution of (Graph-SVM). Consider nodes i and j, where $i \neq j$ and both belong to the same SCC. To ensure that i and j will not suppress each other, (Graph-SVM) solves the SVM problem with the constraint $(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W}\boldsymbol{e}_k = 0$. This essentially disregards the influence of distinct tokens within the same SCC. Consequently, (Graph-SVM) does not truly capture the essence of the ERM solution. To this end, we will introduce the so-called *cyclic-correction* term, which was first proposed in Tarzanagh et al. (2023a). We remark that Tarzanagh et al. (2023a)'s proposal is highly informal and is developed for understanding the MLP layer following self-attention. In contrast, we will develop clear formulas and theory that shed light on the mechanics of next-token prediction. Notably, our theory critically relies on SCCs of the underlying . We begin with cyclic correction component.

Definition 3 (Cyclic correction $\boldsymbol{W}^{\text{cyc}}$) *$\boldsymbol{W}^{cyc}$ is obtained as the minimal norm solution of the ERM problem over the cyclic subset $\overline{DSET}$ per Definition 1. Concretely, $\boldsymbol{W}^{cyc} = \arg\min_{\boldsymbol{W}\in\mathcal{W}^{cyc}} \|\boldsymbol{W}\|_F$, where $\mathcal{W}^{cyc}$ denotes the set of ERM solutions defined as*

$$\mathcal{W}^{cyc} = \arg\min_{\boldsymbol{W}} \bar{\mathcal{L}}(\boldsymbol{W}) \quad \textit{where}$$
$$\bar{\mathcal{L}}(\boldsymbol{W}) = \frac{1}{|\overline{DSET}|} \sum_{(\boldsymbol{X},y)\in\overline{DSET}} \ell(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}})).$$

We also define the subspace $\mathcal{S}_{cyc}$ associated to $\boldsymbol{W}^{cyc}$ as follows. For any non-singleton SCC $((\mathcal{C}_i^{(k)})_{i=1}^{N_k})_{k=1}^{K}$, define its mean embedding $\bar{\boldsymbol{e}}_i^{(k)} = \frac{1}{|\mathcal{C}_i^{(k)}|}\sum_{j\in\mathcal{C}_i^{(k)}} \boldsymbol{e}_j$. Let $\mathcal{S}_{cyc}$ be the span of all matrices $(\boldsymbol{e}_j - \bar{\boldsymbol{e}}_i^{(k)})\boldsymbol{e}_k^\top$ for all $j \in \mathcal{C}_i^{(k)}$, $i \in [N_k]$, $k \in [K]$.

Lemma 2 *Consider $\boldsymbol{W}^{cyc}$, $\mathcal{S}_{cyc}$ as in Definition 3. Let $\boldsymbol{W}^{mm}$ be the corresponding SVM solution of* (Graph-SVM) *and suppose $\boldsymbol{W}^{cyc}, \boldsymbol{W}^{mm} \neq 0$. We have that $\boldsymbol{W}^{cyc} \in \mathcal{S}_{cyc}$ and $\boldsymbol{W}^{mm} \perp \mathcal{S}_{cyc}$. Thus, $\boldsymbol{W}^{cyc}$ and $\boldsymbol{W}^{mm}$ are orthogonal i.e. $\langle \boldsymbol{W}^{cyc}, \boldsymbol{W}^{mm}\rangle = 0$.*

We next make the following assumption on the linear head. Notably, Assumption 4 is a special case of Assumption 3.

Assumption 4 *Classifier $\boldsymbol{C}$ and embeddings $\boldsymbol{E}$ satisfy $\boldsymbol{C}\boldsymbol{E}^\top = \boldsymbol{I}_K$.*

Under Assumption 4, the k-th entry of attention output $\hat{\boldsymbol{y}} = \boldsymbol{C}\boldsymbol{X}^\top\mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}})$ will return the likelihood of token k. The reason is that, $\mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}})$ returns a probability distribution over tokens in the sequence and $\boldsymbol{C}\boldsymbol{X}^\top$ aggregates these probabilities into the token alphabet $[K]$. For instance, if a token ID k shows up twice or more, $\boldsymbol{C}\boldsymbol{X}^\top$ will aggregate multiple occurrences into the k-th entry of $\hat{\boldsymbol{y}}$. As a result, similar to cross-entropy loss, we assume that loss function only relies on $\hat{\boldsymbol{y}}_y$ where y is the true label. Note that Assumption 4 results in $\boldsymbol{c}_y^\top \boldsymbol{e}_k = 0$ for all $k \neq y$, ensuring that none of the tokens except $\boldsymbol{e}_y$ can contribute to minimizing the training risk. Therefore, only tokens within the same SCC as label y will be selected by softmax. In Section 6, we will introduce the failure of global convergence without assuming $\boldsymbol{C}\boldsymbol{E}^\top = \boldsymbol{I}$. Next we have the following results regarding cyclic correction $\boldsymbol{W}^{\text{cyc}}$ and SVM-induced direction $\boldsymbol{W}^{\text{mm}}$.

Theorem 2 *Consider any dataset* DSET *and suppose Assumption 1, 2, and 4 hold. Suppose* (Graph-SVM) *is feasible with* $\boldsymbol{W}^{mm}$ *denoting its solution and assume* $\boldsymbol{W}^{mm} \neq 0$. *If* $\|\boldsymbol{W}^{cyc}\|_F < \infty$, *then the regularization path solution Algo-RP obeys* $\lim_{R\to\infty} \frac{\bar{\boldsymbol{W}}_R}{R} = \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F}$, *and* $\lim_{R\to\infty} \boldsymbol{\Pi}_{\mathcal{S}_{cyc}}(\bar{\boldsymbol{W}}_R) = \boldsymbol{W}^{cyc}$.

Together, the limits above imply the decomposition $\bar{\boldsymbol{W}}_R \approx C_R \cdot \boldsymbol{W}^{\text{mm}} + \boldsymbol{W}^{\text{cyc}}$ for an appropriate $C_R > 0$. Here, we assume $\boldsymbol{W}^{\text{mm}} \neq 0$ to guarantee the directional convergence of attention weights. Specifically, we assume that there exists $i, j, k \in [K]$ such that $(i \Rightarrow j) \in \mathcal{G}^{(k)}$, and (Graph-SVM) has ≥ 1 constraints.

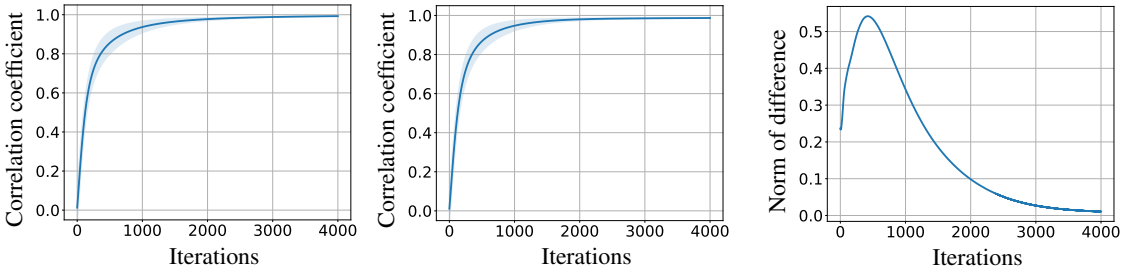


Figure 3: GD convergence of attention weight $\boldsymbol{W}$ when training with acyclic dataset. Correlation coefficient between $\boldsymbol{W}(\tau)$ and $\boldsymbol{W}^{\text{mm}}$ are presented.

(a) Evolution of $\frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F} \to \frac{\boldsymbol{W}^{\text{mm}}}{\|\boldsymbol{W}^{\text{mm}}\|_F}$ (b) Evolution of $\boldsymbol{\Pi}_{\mathcal{S}_{\text{cyc}}}(\boldsymbol{W}(\tau)) \to \boldsymbol{W}^{\text{cyc}}$

Figure 4: GD convergence of attention weight $\boldsymbol{W}$ when training with general/cyclic dataset. **(a)** shows the directional convergence of $\boldsymbol{W}(\tau)$ (cf. (2)); while **(b)** presents the convergence of $\boldsymbol{\Pi}_{\mathcal{S}_{\text{cyc}}}(\boldsymbol{W}(\tau))$ (cf. (3)).

5 Global Convergence of Gradient Descent

In Section 4, we discuss the implicit bias of attention via analysis of regularization path (RP) as employed in Algo-RP. In this section, we focus on the gradient descent (GD) following Algo-GD. Previous work Ji et al. (2020b); Rosset et al. (2003); Suggala et al. (2018) has established the connection between RP and GD for general convex and strictly decreasing loss. Inspired by their results, we assume the log-loss function, i.e., $\ell(u) = -\log(u)$, and establish the GD convergence of attention weight $\boldsymbol{W}$ via the convexity of $\mathcal{L}(\boldsymbol{W})$. Note that although loss function ℓ is convex and the classification head is linear, due to the non-convexity of softmax, the convexity of $\mathcal{L}(\boldsymbol{W})$ is not immediately clear. Towards this, we introduce the following lemma:

Lemma 3 *Suppose Assumptions 2 and 4 hold and consider the log-loss* $\ell(u) = -\log(u)$, *then* $\mathcal{L}(\boldsymbol{W})$ *is convex. Furthermore,* $\mathcal{L}(\boldsymbol{W})$ *is strictly convex on* $\mathcal{S}_{cyc}$.

Now consider log-loss ℓ and let $(\boldsymbol{X}, y) \in$ DSET be any sample. Set $\gamma_t = \boldsymbol{c}_y^\top \boldsymbol{x}_t$. Assumption 4 guarantees that $\gamma_t = 1$ when $\boldsymbol{x}_t = \boldsymbol{e}_y$, otherwise $\gamma_t = 0$. Consider the attention output $\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}})$ and let softmax probabilities be $s_t = \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}})_t$, where $\sum_t s_t = 1$. Then, the loss of this single sample is $\ell(\bar{\gamma})$ where $\bar{\gamma} = \sum_t \gamma_t s_t > 0$. Note that when $\bar{\gamma} \to 0^+$, $-\log(\bar{\gamma})$ results in the infinite loss, which suggests that, once attention weight $\boldsymbol{W}$ diverges to saturate the softmax probability, the finite training risk is achievable only when attention selects all $\boldsymbol{x}_t$'s within the same SCC as y, and then $\boldsymbol{W}^{\text{cyc}}$ following Definition 3 is finite. The following result characterizes the global directional convergence of the GD iterates to the solution of (Graph-SVM).

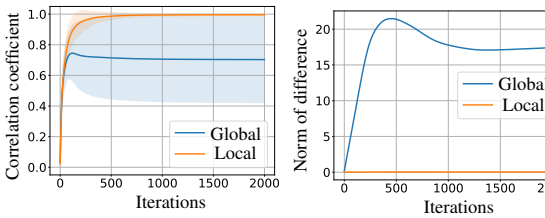
Theorem 3 *Consider any dataset* DSET *and suppose Assumptions 2 and 4 hold. Let* ℓ *be the log-loss. Finally, suppose that* (Graph-SVM) *is feasible with* $\boldsymbol{W}^{mm}$ *denoting its solution such that* $\boldsymbol{W}^{mm} \neq 0$. *Starting from any* $\boldsymbol{W}(0)$ *and with a constant step size* η, *algorithm Algo-GD satisfies*

$$\lim_{\tau\to\infty} \frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F} = \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F}. \tag{2}$$

Additionally, let $\boldsymbol{W}^{cyc}$ *be defined as in Def. 3, then*

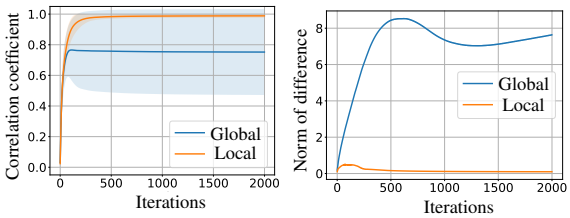
$$\lim_{\tau\to\infty} \boldsymbol{\Pi}_{\mathcal{S}_{cyc}}(\boldsymbol{W}(\tau)) = \boldsymbol{W}^{cyc}. \tag{3}$$

This theorem demonstrates the directional convergence of attention weight $\boldsymbol{W}$, and $\|\boldsymbol{W}(\tau)\|_F \to \infty$ as $\tau \to \infty$.



(a) $\frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F} \to \frac{\widetilde{\boldsymbol{W}}^{\mathrm{mm}}}{\|\widetilde{\boldsymbol{W}}^{\mathrm{mm}}\|_F}$ (b) $\boldsymbol{\Pi}_{\widetilde{\mathcal{S}}_{\mathrm{cyc}}}(\boldsymbol{W}(\tau)) \to \widetilde{\boldsymbol{W}}^{\mathrm{cyc}}$

Figure 5: Squared loss with general classifier



(a) $\frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F} \to \frac{\widetilde{\boldsymbol{W}}^{\mathrm{mm}}}{\|\widetilde{\boldsymbol{W}}^{\mathrm{mm}}\|_F}$ (b) $\boldsymbol{\Pi}_{\widetilde{\mathcal{S}}_{\mathrm{cyc}}}(\boldsymbol{W}(\tau)) \to \widetilde{\boldsymbol{W}}^{\mathrm{cyc}}$

Figure 6: Cross-entropy loss with general classifier

To illustrate Theorem 3, we conduct experiments for both acyclic and general data setting and results are presented in Fig. 3 and 4, respectively. In Fig. 3, we create embedding tables with $K = d = 8$ and generate acyclic dataset with $n = 4, T = 6$. Here we choose step size $\eta = 0.01$ and conduct normalized gradient descent method to accelerate the increasing of the norm of attention weight, so that softmax can easily saturate. Specifically, we update attention weight $\boldsymbol{W}$ via $\boldsymbol{W}(\tau+1) = \boldsymbol{W}(\tau) - \eta \frac{\nabla \mathcal{L}(\boldsymbol{W}(\tau))}{\|\nabla \mathcal{L}(\boldsymbol{W}(\tau))\|_F}$. At each iteration τ, correlation coefficient is computed by $\langle \boldsymbol{W}(\tau), \boldsymbol{W}^{\mathrm{mm}} \rangle / (\|\boldsymbol{W}(\tau)\|_F \|\boldsymbol{W}^{\mathrm{mm}}\|_F)$, and results averaged over 100 random instances are presented in Fig. 3, which end in correlation exceeds 0.99 after training with 4000 iterations.

Moreover, in Fig. 4, we again conduct 100 trails but with randomly generated cyclic dataset DSET under the setting of $K = 6, d = 8, n = 6$ and $T = 4$. The averaged results are presented in Fig. 4a with correlation coefficient ≈ 0.987. In addition to the directional convergence of $\boldsymbol{W}(\tau)$, we also verify the convergence of cyclic component by tracking the matrix distance $\|\boldsymbol{\Pi}_{\mathcal{S}_{\mathrm{cyc}}}(\boldsymbol{W}(\tau)) - \boldsymbol{W}^{\mathrm{cyc}}\|_F$, and results are displayed in Fig. 4b, which obtains < 0.01 final distance. Both Fig. 3 and 4 validate our Theorem 3.

6 Further Investigation on Local Convergence

So far, we have studied the implicit bias of self-attention over next-token prediction problem using RP analysis (Section 4) and proved the global GD convergence of attention weight when one employs the log-loss (Section 5). In this section, we investigate further on the convergence performance of GD and ask: *When will the GD converge to locally-optimal direction and what is its implicit bias?*

Convergence performance of learning 1-layer attention has been analyzed in the previous work Tarzanagh et al. (2023b,a), and they have observed the local convergence phenomenon, and also provided the theoretical explanation and empirical evidence. Inspired by their work, we define the *pseudo* for obtaining locally-optimal SVM equivalence $\widetilde{\boldsymbol{W}}^{\mathrm{mm}}$ and cyclic component $\widetilde{\boldsymbol{W}}^{\mathrm{cyc}}$ as follows:

1. Given any dataset DSET, consider GD solution $\boldsymbol{W}^{\mathrm{GD}}$. For each training example $(\boldsymbol{X}, y) \in$ DSET, let $\boldsymbol{s} = \mathbb{S}(\boldsymbol{X}\boldsymbol{W}^{\mathrm{GD}}\bar{\boldsymbol{x}})$.
2. Construct based on $\boldsymbol{s}$ by adding directed edge $(x_{t_1} \to x_{t_2})$ to $\mathcal{G}^{(k)}$ if $\boldsymbol{s}_{t_1} > 0$, where k, x_t are the token IDs of last token and $\boldsymbol{x}_t$, respectively.

Different from the defined in Section 3.1 which is uniquely determined by the dataset and the ground truth labels, build edges based on the tokens selected by GD solution $\boldsymbol{W}^{\mathrm{GD}}$. To further investigate under which scenarios local convergence phenomenon exists, we consider the following cases and provide experimental evidence.

General loss function ℓ. In Section 5, we analyze the convergence performance of gradient descent when employing log-loss. As we have discussed, such loss guarantees the convexity of the problem and therefore, GD of attention weight (directional) converges to its global minima. Here, we investigate the performance of more general loss function, i.e., squared loss, and find empirical evidence of local convergence.

Extended linear head $\boldsymbol{C}$. In this work, we assume $\boldsymbol{C}\boldsymbol{E}^\top = \boldsymbol{I}$. Let $\gamma_k = \boldsymbol{c}_y^\top \boldsymbol{e}_k$. Then $\gamma_{k_1} = \gamma_{k_2} < \gamma_y$ where $k_1 \neq k_2 \neq y$, and hence token y outperforms all the rest and is guaranteed to be chosen which implies the global convergence. Now consider a general head $\boldsymbol{C}$ (i.e., Assumption 3). Then, as also observed and discussed in Tarzanagh et al. (2023b,a), Algo-GD can converge to a locally-optimal solution.

Fig. 5 and 6 display our local convergence results where Fig. 5 employs squared loss, i.e. $\ell(u) = (1-u)^2$ and Fig. 6 utilizes cross-entropy loss. Both apply general head following Assumption 3. Similar to Fig. 4, we

present (directional) convergence of $\boldsymbol{W}$ towards $\boldsymbol{W}^{\text{mm}}$ and $\boldsymbol{W}^{\text{cyc}}$. Results indicate that instead of converging to the global solution (blue curves), attention weight trained via GD aligns more closely with the locally-optimal SVM solution defined via the pseudo constructed by $\boldsymbol{W}^{\text{GD}}$ (orange curves). In Fig. 5b, the norm difference to $\widetilde{\boldsymbol{W}}^{\text{cyc}}$ remains zero, indicating that all SCCs in the pseudo are singleton and GD optimizes attention weights towards selecting one token per sequence. While in Fig. 6, multiple tokens can be selected by $\boldsymbol{W}^{\text{GD}}$. Note that in Fig. 6b, the norm of difference does not end with zero value on average. The potential explanations can be: Training $\boldsymbol{W}$ and $\widetilde{\boldsymbol{W}}^{\text{cyc}}$ with GD may not fully capture its RP solution, and general classification head induces correlation among tokens, leading the attention mechanism to generate more intricate composed tokens. Nevertheless, our empirical results indicate that $\boldsymbol{W}$ more closely aligns with the local $\widetilde{\boldsymbol{W}}^{\text{cyc}}$ within its cyclic subspace. We defer a rigorous definition of local $\widetilde{\boldsymbol{W}}^{\text{cyc}}$ and guarantees related to gradient descent for future exploration. Experimental details are deferred to the appendix.

7 Discussion

In this work we set out to demystify Transformer-based language modeling via next-token prediction task. We established that single-layer self-attention learning has implicit bias towards the solution of a support vector machine (SVM) formulation based on which encode the priority order among the tokens as per the training data. Our analysis shows that a self-attention model learned via next-token prediction objective implements a selection mechanism to suppress the lower priority tokens in order to predict the higher priority tokens as the next-token for an input sequence. At the same time, such an attention model would distribute its attention weight among all equal priority tokens as modeled by the strongly-connected components of the next-token graph.

A natural future direction is to obtain rigorous implicit bias associated with gradient descent and local convergence. Also, it would be interesting to extend our analysis to multi-layer multi-head self-attention models or explore how feed-forward layers (a.k.a. MLP layers) in Transformers affect the optimization bias to the SVM solution and aid in the aforementioned selection mechanism for the next-token prediction.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Pierre Baldi and Roman Vershynin. The quarks of attention: Structure and capacity of neural attention building blocks. *Artificial Intelligence*, 319:103901, 2023. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2023.103901. URL https://www.sciencedirect.com/science/article/pii/S0004370223000474.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pp. 483–513. PMLR, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International*

Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20s.html.

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

Yu-An Chung and James Glass. Generative pre-training for speech with autoregressive predictive coding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3497–3501. IEEE, 2020.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/dong21a.html.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5793–5831. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/edelman22a.html.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? a study through the random features lens. *arXiv preprint arXiv:2307.11353*, 2023.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c529dba08a146ea8d6cf715ae8930cbe-Paper-Conference.pdf.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

Samy Jelassi, Michael Eli Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=eMW9AkXaREI.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 772–804. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/ji21a.html.

Ziwei Ji, Miroslav Dudík, Robert E. Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2109–2136. PMLR, 09–12 Jul 2020a. URL https://proceedings.mlr.press/v125/ji20a.html.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pp. 2109–2136. PMLR, 2020b.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34:18970–18983, 2021.

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

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/li23l.html.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Erik Nijkamp, Jeffrey Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. Progen2: exploring the boundaries of protein language models. *arXiv preprint arXiv:2206.13517*, 2022.

OpenAI. Gpt-4 technical report. *arXiv preprintarXiv:2303.08774*, 2023.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26724–26768. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/oymak23a.html.

Qian Qian and Xiaoyuan Qian. The implicit bias of adagrad on separable data. *Advances in Neural Information Processing Systems*, 32, 2019.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Margin maximizing loss functions. *Advances in neural information processing systems*, 16, 2003.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018. URL http://jmlr.org/papers/v19/18-188.html.

Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. *Advances in Neural Information Processing Systems*, 31, 2018.

Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972. doi: 10.1137/0201010. URL https://doi.org/10.1137/0201010.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023a.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Margin maximization in attention mechanism. *arXiv preprint arXiv:2306.13596*, 2023b.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10849–10858. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21q.html.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=ByxRM0Ntvr.

Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. O (n) connections are expressive enough: Universal approximability of sparse transformers. *Advances in Neural Information Processing Systems*, 33:13783–13794, 2020b.

Contents

A Useful Notations

In this section, we will introduce additional notations used in the subsequent proofs.

• **Token index sets $\mathcal{O}_i$ and $\mathcal{R}_i$, $i \in [n]$.** Consider dataset $\texttt{DSET}$. Throughout, for any sample $(\boldsymbol{X}_i, y_i) \in \texttt{DSET}$, $i \in [n]$, we define

$$\mathcal{O}_i := \Big\{t \;\Big|\; x_{it} = y_i, t \in [T_i]\Big\} \quad \text{and} \quad \mathcal{R}_i := \mathcal{O}_i \bigcup \Big\{t \;\Big|\; x_{it} = j, (j \asymp y_i) \in \mathcal{G}^{(k)}, \forall j \in [K], t \in [T_i]\Big\} \tag{4}$$

where x_{it} is the token ID of $\boldsymbol{x}_{it}$, T_i is the number of tokens in the input sequence $\boldsymbol{X}_i$ and $\mathcal{G}^{(k)}$ is the corresponding next-token graph (NTG) associated with the last token of $\boldsymbol{X}_i$ (k is the token ID of the last token). Throughout, let $\bar{\mathcal{O}}_i = [T_i] - \mathcal{O}_i$ and $\bar{\mathcal{R}}_i = [T_i] - \mathcal{R}_i$. Concretely, $\mathcal{O}_i$ returns the token indices of ith input that have the same token ID as label y_i, while $\mathcal{R}_i$ returns the token indices of ith input that are included in the same strongly-connected component (SCC) as label y_i, and for any $t \in \bar{\mathcal{R}}_i$, we have $(y_i \Rightarrow x_{it}) \in \mathcal{G}^{(k)}$. Take the last input sequence in Figure 2(left) as an example, then $\mathcal{O} = \{3\}$ and $\mathcal{R} = \{1, 3\}$.

• **Datasets $\texttt{DSET}$, $\overline{\texttt{DSET}}$ and sample index set $\mathcal{I}$.** Recap the training dataset $\texttt{DSET} = (\boldsymbol{X}_i, y_i)_{i=1}^n$. Based on the token relations between input tokens and label, following instructions in Section 3.1 we can construct the NTGs of dataset $\texttt{DSET}$. Then, let $\mathcal{I} \subseteq [n]$ be the sample index set such that for any $i \in \mathcal{I}$, $\boldsymbol{X}_i$ contains distinct tokens from the same SCC as label y_i in their corresponding NTG. Then the cyclic subset defined in Definition 1 can be written by

$$\overline{\texttt{DSET}} = (\bar{\boldsymbol{X}}_i, y_i)_{i \in \mathcal{I}}, \tag{5}$$

where $\bar{\boldsymbol{X}}_i$ is obtained by removing all input tokens of $\boldsymbol{X}_i$ that are in the different SCCs from the label token y_i, or equivalently, removing $\boldsymbol{x}_{it}$, $t \in \bar{\mathcal{R}}_i$. Consequently, let $\bar{\mathcal{I}} := [n] - \mathcal{I}$ and for all $i \in \bar{\mathcal{I}}$, $\boldsymbol{X}_i$ only contains input tokens (ignoring the ones with the same token ID as label) that have strictly lower priority than its label token, i.e., $(y_i \Rightarrow x_{it}) \in \mathcal{G}^{(k)}$ for $t \in [T_i]$ and $x_{it} \neq y_i$ and $\mathcal{G}^{(k)}$ is the corresponding NTG associated with the last token. In Figure 2(left), we have $\mathcal{I} = \{4, 5, 6, 7\}$.

B Global Convergence of Regularization Path

B.1 Proof of Lemma 1

Proof. Let $\boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$ denote the score vector of ith input and $\gamma_{it} = \boldsymbol{x}_{it}^\top \boldsymbol{c}_{y_i}$. Let $\gamma_i^{\max} = \boldsymbol{e}_{y_i}^\top \boldsymbol{c}_{y_i} = \max_{t \in [T_i]} \gamma_{it}$ following Assumptions 2 and 3. What's more, since Assumption 1 ensures that loss ℓ is strictly

decreasing, we define the optimal loss as follows:

$$\mathcal{L}_\star := \frac{1}{n}\sum_{i=1}^{n} \ell(\gamma_i^{\max}).$$

For any $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, let $\boldsymbol{s}_i = \boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i$, $i \in [n]$. If $\|\boldsymbol{W}\|_F < \infty$, then $\min_{t\in[T_i], i\in[n]} s_{it} > 0$ and for any $i \in [n]$

$$\boldsymbol{s}_i^\top \boldsymbol{\gamma}_i = \sum_{t=1}^{T_i} s_{it}\gamma_{it} < \gamma_i^{\max}.$$

Using Assumption 1 that loss function ℓ is strictly decreasing, we get

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{s}_i^\top \boldsymbol{\gamma}_i) > \frac{1}{n}\sum_{i=1}^{n} \ell(\gamma_i^{\max}) = \mathcal{L}_\star.$$

We next prove that $\lim_{R\to\infty} \mathcal{L}(R\cdot \boldsymbol{W}^{\mathrm{mm}}) = \mathcal{L}_\star$. Recap $\mathcal{O}_i$ and $\bar{\mathcal{O}}_i$ from (4). Since token $\boldsymbol{e}_{y_i}$ is always contained in $\boldsymbol{X}_i$ following Assumption 2, we have $|\mathcal{O}_i| \geq 1, i \in [n]$, and $\boldsymbol{X}_i$ contains $|\mathcal{O}_i|$ optimal tokens $\boldsymbol{e}_{y_i}$. Note that under acyclic data setting, $\boldsymbol{W}^{\mathrm{mm}}$ separates tokens $\boldsymbol{e}_{y_i}$ from the rest of the tokens for each $i \in [n]$. Then $\lim_{R\to\infty} \mathbb{S}(\boldsymbol{X}_i(R\cdot \boldsymbol{W}^{\mathrm{mm}})\bar{\boldsymbol{x}}_i)$ will output $1/|\mathcal{O}_i|$ for $t \in \mathcal{O}_i$ and zero for the left. Specifically, let $\boldsymbol{s}_i^R := \mathbb{S}(\boldsymbol{X}_i(R\cdot \boldsymbol{W}^{\mathrm{mm}})\bar{\boldsymbol{x}}_i)$, and following the SVM objective (Acyc-SVM) we get

$$s_{it}^R = \frac{e^{\boldsymbol{x}_{it}^\top (R\cdot \boldsymbol{W}^{\mathrm{mm}})\bar{\boldsymbol{x}}_i}}{\sum_{t\in[T_i]} e^{\boldsymbol{x}_{it}^\top (R\cdot \boldsymbol{W}^{\mathrm{mm}})\bar{\boldsymbol{x}}_i}} = \frac{1}{|\mathcal{O}_i| + \sum_{t\in\bar{\mathcal{O}}_i} e^{(\boldsymbol{x}_{it}-\boldsymbol{e}_{y_i})^\top (R\cdot \boldsymbol{W}^{\mathrm{mm}})\bar{\boldsymbol{x}}_i}} \geq \frac{1}{|\mathcal{O}_i| + e^{-R}} \quad \text{for all } t \in \mathcal{O}_i.$$

Then $\lim_{R\to\infty} s_{it}^R = 1/|\mathcal{O}_i|$ for $t \in \mathcal{O}_i$ and $\lim_{R\to\infty} s_{it}^R = 0$ for $t \in \bar{\mathcal{O}}_i$. Hence we have

$$\boldsymbol{X}_i^\top \boldsymbol{s}_i^R = \sum_{t\in\mathcal{O}_i} \frac{1}{|\mathcal{O}_i|}\boldsymbol{e}_{y_i} = \boldsymbol{e}_{y_i}$$

and $\lim_{R\to\infty} \mathcal{L}(R\cdot \boldsymbol{W}^{\mathrm{mm}}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{e}_{y_i}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\gamma_i^{\max}) = \mathcal{L}_\star$. ■

B.2 Proof of Theorem 1

Proof. Recap the dataset $\texttt{DSET} = (\boldsymbol{X}_i, y_i)_{i=1}^n$. The regularization path solution of the ERM problem (per Algo-RP and (ERM)) is defined as follows:

$$\bar{\boldsymbol{W}}_R = \arg\min_{\|\boldsymbol{W}\|_F \leq R} \mathcal{L}(\boldsymbol{W}) \quad \text{where} \quad \mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)).$$

The proof is similar to the proof of Theorem 2 in Tarzanagh et al. (2023a) by choosing $\texttt{opt}_i = y_i$. However in our work, we allow each sequence contains more than one optimal tokens, while Tarzanagh et al. (2023a) forces that the optimal token is unique.

Following the proof in Lemma 1, let $\boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$, $\gamma_i^{\max} = \boldsymbol{e}_{y_i}^\top \boldsymbol{c}_{y_i} = \max_{t\in[T_i]} \gamma_{it}$, and the optimal training risk

$$\mathcal{L}_\star := \frac{1}{n}\sum_{i=1}^{n} \ell(\gamma_i^{\max}).$$

From Lemma 1, we have that for any finite $\boldsymbol{W}$, $\mathcal{L}(\boldsymbol{W}) < \lim_{R\to\infty} \mathcal{L}(R\cdot \boldsymbol{W}^{\mathrm{mm}}) = \mathcal{L}^\star$. Then the optimal risk $\mathcal{L}_\star$ is achievable and to achieve the limit, R has to be infinite. Then it remains to prove that $\bar{\boldsymbol{W}}_R$ converges in direction to $\boldsymbol{W}^{\mathrm{mm}}$.

Suppose convergence fails. We will obtain a contradiction by showing that $R\cdot \boldsymbol{W}^{\mathrm{mm}}/\|\boldsymbol{W}^{\mathrm{mm}}\|_F$ achieves a strictly superior loss compared to $\bar{\boldsymbol{W}}_R$. Since $\bar{\boldsymbol{W}}_R$ fails to converge to $\boldsymbol{W}^{\mathrm{mm}}$, for some $\delta > 0$, there exists arbitrarily large $R > 0$ such that

$$\|\bar{\boldsymbol{W}}_R \cdot \|\boldsymbol{W}^{\mathrm{mm}}\|_F/R - \boldsymbol{W}^{\mathrm{mm}}\|_F \geq \delta.$$

Let $\boldsymbol{W}' = \bar{\boldsymbol{W}}_R \cdot \|\boldsymbol{W}^{\mathrm{mm}}\|_F/R$ where we have $\|\boldsymbol{W}'\|_F \leq \|\boldsymbol{W}^{\mathrm{mm}}\|_F$ and $\boldsymbol{W}' \neq \boldsymbol{W}^{\mathrm{mm}}$. Since $\boldsymbol{W}^{\mathrm{mm}}$ is the min-norm solution of (Acyc-SVM), then for some $\epsilon := \epsilon(\delta)$, there exists i, j, k such that

$$(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W}' \boldsymbol{e}_k \leq 1 - \epsilon \quad \text{where} \quad (i \Rightarrow j) \in \mathcal{G}^{(k)}.$$

Now, we will argue that this leads to a contradiction by proving $\mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}}/\|\boldsymbol{W}^{\text{mm}}\|_F) < \mathcal{L}(\bar{\boldsymbol{W}}_R)$ for sufficiently large R. For simplification, we update $R \leftarrow R/\|\boldsymbol{W}^{\text{mm}}\|_F$, and it is equivalent to prove that $\mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}}) < \mathcal{L}(R \cdot \boldsymbol{W}')$ for sufficiently large R.

To obtain the result, we establish a refined softmax probability control by studying the distance to $\mathcal{L}_\star$. Let $\boldsymbol{a}_i^\star = \boldsymbol{X}_i(R \cdot \boldsymbol{W}^{\text{mm}})\bar{\boldsymbol{x}}_i$, $\boldsymbol{a}_i^R = \boldsymbol{X}_i(R \cdot \boldsymbol{W}')\bar{\boldsymbol{x}}_i$, $\boldsymbol{s}_i^\star := \mathbb{S}(\boldsymbol{a}_i^\star)$, $\boldsymbol{s}_i^R := \mathbb{S}(\boldsymbol{a}_i^R)$, $\gamma_i^\star := \boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \boldsymbol{s}_i^\star$, and $\gamma_i^R := \boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \boldsymbol{s}_i^R$. Recap that $\gamma_i^{\max} = \boldsymbol{e}_{y_i}^\top \boldsymbol{c}_{y_i}$. Then

$$\sum_{t\in\mathcal{O}_i} s_{it}^R = \frac{\sum_{t\in\mathcal{O}_i} e^{a_{it}^R}}{\sum_{t\in[T_i]} e^{a_{it}^R}} \le \frac{|\mathcal{O}_i|}{|\mathcal{O}_i| + e^{-(1-\epsilon)R}} \le \frac{1}{1+e^{-(1-\epsilon)R}/T}, \quad \exists i \in [n] \tag{6}$$

$$\sum_{t\in\mathcal{O}_i} s_{it}^\star = \frac{\sum_{t\in\mathcal{O}_i} e^{a_{it}^\star}}{\sum_{t\in[T_i]} e^{a_{it}^\star}} \ge \frac{|\mathcal{O}_i|}{|\mathcal{O}_i| + (T-|\mathcal{O}_i|)e^{-R}} \ge \frac{1}{1+Te^{-R}}, \quad \forall i \in [n]. \tag{7}$$

Since ℓ is strictly decreasing and ℓ' is bounded following Assumption 1, let $c_{\text{dn}} \le -\ell' \le c_{\text{up}}$ for some constants $c_{\text{dn}}, c_{\text{up}} > 0$. Additionally, define the score minimal/maximal score gaps as

$$c_{\min} = \min_{i,k\in[K],i\neq k} (\boldsymbol{e}_k - \boldsymbol{e}_i)^\top \boldsymbol{c}_k, \quad c_{\max} = \max_{i,k\in[K],i\neq k} (\boldsymbol{e}_k - \boldsymbol{e}_i)^\top \boldsymbol{c}_k$$

where $c_{\max} \ge c_{\min} > 0$. Then we have that there exists $i \in [n]$,

$$\begin{aligned}\mathcal{L}(R \cdot \boldsymbol{W}') - \mathcal{L}_\star &\ge \frac{1}{n}\left(\ell(\gamma_i^R) - \ell(\gamma_i^{\max})\right) \ge \frac{c_{\text{dn}}}{n}\left(\gamma_i^{\max} - \gamma_i^R\right) \\ &\ge \frac{c_{\text{dn}}}{n} c_{\min}\left(1 - \sum_{t\in\mathcal{O}_i} s_{it}^R\right) \ge \frac{c_{\text{dn}} c_{\min}}{n} \frac{1}{1+Te^{(1-\epsilon)R}}\end{aligned} \tag{8}$$

and letting $j := \arg\max_{i\in[n]} \left(\ell(\gamma_i^\star) - \ell(\gamma_i^{\max})\right)$, we can upper-bound the loss difference for $R \cdot \boldsymbol{W}^{\text{mm}}$ as follows:

$$\begin{aligned}\mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}}) - \mathcal{L}_\star &\le \max_{i\in[n]} \left(\ell(\gamma_i^\star) - \ell(\gamma_i^{\max})\right) \le c_{\text{up}}\left(\gamma_j^{\max} - \gamma_j^\star\right) \\ &\le c_{\text{up}} c_{\max}\left(1 - \sum_{t\in\mathcal{O}_i} s_{it}^\star\right) \le c_{\text{up}} c_{\max} \frac{1}{1+e^R/T} \le c_{\text{up}} c_{\max} T e^{-R}.\end{aligned} \tag{9}$$

Combining them together results in that, $\mathcal{L}(R \cdot \boldsymbol{W}') > \mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}})$ whenever

$$\frac{c_{\text{dn}} c_{\min}}{n} \frac{1}{1+Te^{(1-\epsilon)R}} > c_{\text{up}} c_{\max} T e^{-R} \Longrightarrow R > \frac{1}{\epsilon} \log\left(\frac{2nT^2 c_{\text{up}} c_{\max}}{c_{\text{dn}} c_{\min}}\right).$$

This completes the proof by contradiction. ■

B.3 Proof of Lemma 2

Proof. Recap the $\boldsymbol{W}^{\text{cyc}}$ and $\mathcal{S}_{\text{cyc}}$ definitions in Definition 3 and (5) that

$$\boldsymbol{W}^{\text{cyc}} = \arg\min_{\boldsymbol{W}} \bar{\mathcal{L}}(\boldsymbol{W}) \quad \text{where} \quad \bar{\mathcal{L}}(\boldsymbol{W}) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \ell(\boldsymbol{c}_{y_i}^\top \bar{\boldsymbol{X}}_i^\top \mathbb{S}(\bar{\boldsymbol{X}}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)),$$

and $\mathcal{S}_{\text{cyc}}$ is the span of all matrices $(\boldsymbol{e}_j - \bar{\boldsymbol{e}}_i^{(k)})\boldsymbol{e}_k^\top$ for all $j \in \mathcal{C}_i^{(k)}$, $i \in [N_k]$, $k \in [K]$.

• **We first prove that $\boldsymbol{W}^{\text{cyc}} \in \mathcal{S}_{\text{cyc}}$.** Let $\boldsymbol{W}^\perp$ be any matrix such that $\boldsymbol{W}^\perp \perp \mathcal{S}_{\text{cyc}}$. We will show that for any $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, $\bar{\mathcal{L}}(\boldsymbol{W} + \boldsymbol{W}^\perp) = \bar{\mathcal{L}}(\boldsymbol{W})$. Then since $\boldsymbol{W}^{\text{cyc}}$ is the min-norm solution, it will imply $\boldsymbol{W}^{\text{cyc}} \in \mathcal{S}_{\text{cyc}}$.

$$\bar{\mathcal{L}}(\boldsymbol{W} + \boldsymbol{W}^\perp) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \ell(\boldsymbol{c}_{y_i}^\top \bar{\boldsymbol{X}}_i^\top \mathbb{S}(\bar{\boldsymbol{X}}_i(\boldsymbol{W} + \boldsymbol{W}^\perp)\bar{\boldsymbol{x}}_i)).$$

Then it remains to show that for any $(\bar{\boldsymbol{X}}, y) \in \overline{\texttt{DSET}}$, $\mathbb{S}(\bar{\boldsymbol{X}}(\boldsymbol{W} + \boldsymbol{W}^\perp)\bar{\boldsymbol{x}}) = \mathbb{S}(\bar{\boldsymbol{X}}\boldsymbol{W}\bar{\boldsymbol{x}})$. For simplification, let $\bar{\boldsymbol{x}} = \boldsymbol{e}_k$, and following the definition of NTG, SCC and $\overline{\texttt{DSET}}$, we have that all tokens $\boldsymbol{x} \in \bar{\boldsymbol{X}}$ are in the same

SCC and denote it as $\mathcal{C}^{(k)}$. Then $\mathcal{S}_{\text{cyc}}$ spans the matrices $(\boldsymbol{e}_j - \bar{\boldsymbol{e}})\boldsymbol{e}_k^\top$ for $j \in \mathcal{C}^{(k)}$ where $\bar{\boldsymbol{e}} := \frac{1}{|\mathcal{C}^{(k)}|}\sum_{j\in\mathcal{C}^{(k)}} \boldsymbol{e}_j$. For any $j \in \mathcal{C}^{(k)}$, we get

$$\boldsymbol{e}_j^\top(\boldsymbol{W} + \boldsymbol{W}^\perp)\boldsymbol{e}_k = \boldsymbol{e}_j^\top \boldsymbol{W}\boldsymbol{e}_k + \boldsymbol{e}_j^\top \boldsymbol{W}^\perp \boldsymbol{e}_k.$$

Next, let $a_j = \boldsymbol{\Pi}_{\boldsymbol{e}_j\boldsymbol{e}_k^\top}(\boldsymbol{W}^\perp)$, $j \in \mathcal{C}^{(k)}$. Since $\boldsymbol{W}^\perp \perp \mathcal{S}_{\text{cyc}}$, and $(\boldsymbol{e}_j - \bar{\boldsymbol{e}})\boldsymbol{e}_k^\top \in \mathcal{S}_{\text{cyc}}$, we obtain

$$\begin{aligned}
&(\boldsymbol{e}_j - \bar{\boldsymbol{e}})^\top \boldsymbol{W}^\perp \boldsymbol{e}_k = 0 \qquad (10)\\
\implies &\boldsymbol{e}_j \boldsymbol{W}^\perp \boldsymbol{e}_k - \frac{1}{|\mathcal{C}^{(k)}|}\sum_{j'\in\mathcal{C}^{(k)}} \boldsymbol{e}_{j'}^\top \boldsymbol{W}^\perp \boldsymbol{e}_k = 0\\
\implies &a_j - \frac{1}{|\mathcal{C}^{(k)}|}\sum_{j'\in\mathcal{C}^{(k)}} a_{j'} = 0\\
\implies &a_j = \bar{a} \quad \text{where} \quad \bar{a} = \frac{1}{|\mathcal{C}^{(k)}|}\sum_{j\in\mathcal{C}^{(k)}} a_j.
\end{aligned}$$

Here (10) uses the fact that $\boldsymbol{W}^\perp \perp \mathcal{S}_{\text{cyc}}$. Then for any $\boldsymbol{x} \in \bar{\boldsymbol{X}}$, $\boldsymbol{x}^\top \boldsymbol{W}^\perp \bar{\boldsymbol{x}} = \bar{a}$ and hence

$$\begin{aligned}
\bar{\boldsymbol{X}}(\boldsymbol{W} + \boldsymbol{W}^\perp)\bar{\boldsymbol{x}} &= \bar{\boldsymbol{X}}\boldsymbol{W}\bar{\boldsymbol{x}} + \bar{\boldsymbol{X}}\boldsymbol{W}^\perp\bar{\boldsymbol{x}} = \bar{\boldsymbol{X}}\boldsymbol{W}\bar{\boldsymbol{x}} + \bar{a}\mathbf{1}\\
\mathbb{S}(\bar{\boldsymbol{X}}(\boldsymbol{W} + \boldsymbol{W}^\perp)\bar{\boldsymbol{x}}) &= \mathbb{S}(\bar{\boldsymbol{X}}\boldsymbol{W}\bar{\boldsymbol{x}} + \bar{a}\mathbf{1}) = \mathbb{S}(\bar{\boldsymbol{X}}\boldsymbol{W}\bar{\boldsymbol{x}}),
\end{aligned}$$

which completes the proof.

• **We next prove that $\boldsymbol{W}^{\text{mm}} \perp \mathcal{S}_{\text{cyc}}$.** Define $\mathcal{S}$ be the span of all matrices $(\boldsymbol{e}_j - \boldsymbol{e}_{j'})\boldsymbol{e}_k^\top$ for all $(j \asymp j') \in \mathcal{G}^{(k)}$ and $k \in [K]$. From the formulation of (Graph-SVM), we know that $\boldsymbol{W}^{\text{mm}} \perp \mathcal{S}$. Also $(j \asymp j') \in \mathcal{G}^{(k)}$ implies that j, j' are in the same SCC, i.e., $j, j' \in \mathcal{C}_i^{(k)}$. Recap that $\mathcal{S}_{\text{cyc}}$ is span of all matrices $(\boldsymbol{e}_j - \bar{\boldsymbol{e}}_i^{(k)})\boldsymbol{e}_k^\top$ for all $j \in \mathcal{C}_i^{(k)}, i \in [N_k], k \in [K]$. Below we show that any matrix $(\boldsymbol{e}_j - \bar{\boldsymbol{e}}_i^{(k)})\boldsymbol{e}_k^\top \in \mathcal{S}$.

$$(\boldsymbol{e}_j - \bar{\boldsymbol{e}}_i^{(k)})\boldsymbol{e}_k^\top = \left(\boldsymbol{e}_j - \frac{1}{|\mathcal{C}_i^{(k)}|}\sum_{j'\in\mathcal{C}_i^{(k)}} \boldsymbol{e}_{j'}\right)\boldsymbol{e}_k^\top = \frac{1}{|\mathcal{C}_i^{(k)}|}\sum_{j'\in\mathcal{C}_i^{(k)}} (\boldsymbol{e}_j - \boldsymbol{e}_{j'})\boldsymbol{e}_k^\top,$$

where $(\boldsymbol{e}_j - \boldsymbol{e}_{j'})\boldsymbol{e}_k^\top \in \mathcal{S}$, and therefore, $(\boldsymbol{e}_j - \bar{\boldsymbol{e}}_i^{(k)})\boldsymbol{e}_k^\top \in \mathcal{S}$ and $\mathcal{S}_{\text{cyc}} \subseteq \mathcal{S}$. Since $\boldsymbol{W}^{\text{mm}} \perp \mathcal{S}$, then it implies that $\boldsymbol{W}^{\text{mm}} \perp \mathcal{S}_{\text{cyc}}$, which completes the proof.

Additionally, we can also get that any matrix $(\boldsymbol{e}_j - \boldsymbol{e}_{j'})\boldsymbol{e}_k^\top \in \mathcal{S}_{\text{cyc}}$ for all $j, j' \in \mathcal{C}_i^{(k)}$.

$$(\boldsymbol{e}_j - \boldsymbol{e}_{j'})\boldsymbol{e}_k^\top = (\boldsymbol{e}_j - \bar{\boldsymbol{e}})\boldsymbol{e}_k^\top - (\boldsymbol{e}_{j'} - \bar{\boldsymbol{e}})\boldsymbol{e}_k^\top \quad \text{where} \quad \bar{\boldsymbol{e}} = \frac{1}{|\mathcal{C}_i^{(k)}|}\sum_{j\in\mathcal{C}_i^{(k)}} \boldsymbol{e}_j.$$

Since $(\boldsymbol{e}_j - \bar{\boldsymbol{e}})\boldsymbol{e}_k^\top, (\boldsymbol{e}_{j'} - \bar{\boldsymbol{e}})\boldsymbol{e}_k^\top \in \mathcal{S}_{\text{cyc}}$, then $\mathcal{S} \subseteq \mathcal{S}_{\text{cyc}}$. Combining them together results in that $\mathcal{S} = \mathcal{S}_{\text{cyc}}$. ■

B.4 Proof of Theorem 2

Lemma 4 *Consider the setting of Theorem 2. For any $\boldsymbol{W}^\parallel \in \mathcal{S}_{cyc}$ with $\|\boldsymbol{W}^\parallel\|_F < \infty$, consider the following objective*

$$\bar{\boldsymbol{W}}_R^\perp := \arg\min_{\boldsymbol{W}^\perp \in \mathcal{S}_{cyc}^\perp, \|\boldsymbol{W}^\perp\|_F \le R} \mathcal{L}(\boldsymbol{W}^\perp + \boldsymbol{W}^\parallel). \qquad (11)$$

Then we have that

$$\lim_{R\to\infty} \frac{\bar{\boldsymbol{W}}_R^\perp}{R} = \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F}.$$

Proof. Recap $\mathcal{O}_i, \bar{\mathcal{O}}_i, \mathcal{R}_i, \bar{\mathcal{R}}_i, i \in [n]$ from (4). Let $\boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$ and

$$\gamma_{it} = \boldsymbol{x}_{it}^\top \boldsymbol{c}_{y_i} = \begin{cases} 1, & t \in \mathcal{O}_i \\ 0, & t \in \bar{\mathcal{O}}_i. \end{cases}$$

To proceed, define $\gamma_i^{\max}$ as follows:

- Consider $i \in \bar{\mathcal{I}}$. Then the optimal risk of input sequence $\boldsymbol{X}_i$ is obtained by selecting token ID y_i and we define the maximal score $\gamma_i^{\max} := \boldsymbol{c}_{y_i}^\top \boldsymbol{e}_{y_i} = 1$.
- Consider $i \in \mathcal{I}$.
  1. Assumption 4 ensures that all tokens, excluding the ones with token ID $x_{it} = y_i$, return zero score, that is, $\boldsymbol{c}_{y_i}^\top \boldsymbol{e}_k = 0$ for $k \neq y_i$
  2. From proof of Lemma 2, for any $\boldsymbol{W}^\perp \in \mathcal{S}_{\text{cyc}}^\perp$ and $t \in \mathcal{R}_i$, $\boldsymbol{x}_{it}^\top (\boldsymbol{W}^\perp + \boldsymbol{W}^\parallel)\bar{\boldsymbol{x}}_i = \boldsymbol{x}_{it}^\top \boldsymbol{W}^\parallel \bar{\boldsymbol{x}}_i + \bar{a}_i$, where $\bar{a}_i$ is some constant associated with $\boldsymbol{W}^\perp$ and remains the same value within the same SCC. Let $\boldsymbol{b}_i = \boldsymbol{X}_i \boldsymbol{W}^\parallel \bar{\boldsymbol{x}}_i$. Then the probabilities for $t \in \mathcal{R}_i$ (if denoted by s_{it}) obey

$$\frac{s_{it}}{\sum_{t' \in \mathcal{R}_i} s_{it'}} = \frac{e^{b_{it} + \bar{a}_i}}{\sum_{t' \in \mathcal{R}_i} e^{b_{it'} + \bar{a}_i}} = \frac{e^{b_{it}}}{\sum_{t' \in \mathcal{R}_i} e^{b_{it'}}}, \tag{12}$$

     which means that the probability distribution over set $\mathcal{R}_i$ remains the same with varying $\boldsymbol{W}^\perp$.

  Combining both, we can see that the optimal risk is achieved by assigning zero probabilities for tokens in $\bar{\mathcal{R}}_i$ since for all $t \in \bar{\mathcal{R}}_i$, $\gamma_{it} = 0$ and it never contributes to reducing risk. What's more, (12) indicates that for any $\boldsymbol{W}^\perp \perp \mathcal{S}_{\text{cyc}}$ with fixed $\boldsymbol{W}^\parallel \in \mathcal{S}_{\text{cyc}}$, the probability assignment of tokens over $\mathcal{R}_i$ is fixed (if assuming zero probabilities for $t \in \bar{\mathcal{R}}_i$). Then we can define

$$\gamma_i^{\max} := |\mathcal{O}_i| \cdot \bar{s}_i \quad \text{where} \quad \bar{s}_i = \frac{e^{\boldsymbol{e}_{y_i}^\top \boldsymbol{W}^\parallel \bar{\boldsymbol{x}}_i}}{\sum_{t' \in \mathcal{R}_i} e^{b_{it'}}}.$$

  Note that if consider the cyclic subdataset $\overline{\texttt{DSET}}$ as in (5). Let $(\bar{\boldsymbol{X}}_i, y_i) \in \overline{\texttt{DSET}}$ where $\bar{\boldsymbol{X}}_i$ is the corresponding sequence by removing the tokens in $\bar{\mathcal{R}}_i$. Then we have $\gamma_i^{\max} = \boldsymbol{c}_{y_i}^\top \bar{\boldsymbol{X}}_i^\top \mathbb{S}(\bar{\boldsymbol{X}}_i \boldsymbol{W}^\parallel \bar{\boldsymbol{x}}_i)$.

Then we define the optimal risk of (11) and its corresponding softmax probabilities $\boldsymbol{s}_i^{\max}$, $i \in [n]$ as follows:

$$\mathcal{L}_\star^{\boldsymbol{W}^\parallel} := \frac{1}{n} \sum_{i=1}^n \ell(\gamma_i^{\max}), \quad \text{and} \quad s_{it}^{\max} = \begin{cases} 0, & t \in \bar{\mathcal{R}}_i \\ \frac{e^{b_{it}}}{\sum_{t' \in \mathcal{R}_i} e^{b_{it'}}}, & t \in \mathcal{R}_i \end{cases} \quad \text{for all } i \in [n].$$

Note that we also have

$$\gamma_i^{\max} = \boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \boldsymbol{s}_i^{\max} = \sum_{t \in \mathcal{O}_i} s_{it}^{\max} = \frac{\sum_{t \in \mathcal{O}_i} e^{b_{it}}}{\sum_{t \in \mathcal{R}_i} e^{b_{it}}}.$$

In the following, we will complete the proof in three steps.

Step 1: We first show that $\lim_{R \to \infty} \mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}} + \boldsymbol{W}^\parallel) = \mathcal{L}_\star^{\boldsymbol{W}^\parallel}$. Since from the formulation of (Graph-SVM), $\boldsymbol{W}^{\text{mm}}$ returns higher correlation to the tokens in $\mathcal{R}_i$ than the tokens in $\bar{\mathcal{R}}_i$, then with $R \to \infty$, $R \cdot \boldsymbol{W}^{\text{mm}}$ ensures that tokens in $\bar{\mathcal{R}}_i$ output zero softmax probabilities. Specifically, let $\boldsymbol{a}_i^\star = \boldsymbol{X}_i (R \cdot \boldsymbol{W}^{\text{mm}}) \bar{\boldsymbol{x}}_i$, and recap $\boldsymbol{b}_i = \boldsymbol{X}_i \boldsymbol{W}^\parallel \bar{\boldsymbol{x}}_i$. Then let $\boldsymbol{s}_i^\star = \mathbb{S}(\boldsymbol{X}_i (R \cdot \boldsymbol{W}^{\text{mm}} + \boldsymbol{W}^\parallel)) = \mathbb{S}(\boldsymbol{a}_i + \boldsymbol{b}_i)$. For $t \in \mathcal{R}_i$, we have

$$s_{it}^\star = \frac{e^{a_{it}^\star + b_{it}}}{\sum_{t' \in [T_i]} e^{a_{it'}^\star + b_{it'}}} \geq \frac{e^{b_{it}}}{\sum_{t' \in \mathcal{R}_i} e^{b_{it'}} + \sum_{t' \in \bar{\mathcal{R}}_i} e^{b_{it'} - R}}, \quad t \in \mathcal{R}_i$$

where the inequality comes from the constraints of (Graph-SVM). Then

$$\lim_{R \to \infty} s_{it}^\star = \begin{cases} 0, & t \in \bar{\mathcal{R}}_i \\ \frac{e^{b_{it}}}{\sum_{t' \in \mathcal{R}_i} e^{b_{it'}}}, & t \in \mathcal{R}_i \end{cases} \quad \text{for all } i \in [n].$$

Therefore, $\lim_{R \to \infty} \boldsymbol{s}_i^\star = \boldsymbol{s}_i^{\max}$ and the fact that $\gamma_i^{\max} = \boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \boldsymbol{s}_i^{\max}$ implies

$$\lim_{R \to \infty} \mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}} + \boldsymbol{W}^\parallel) = \mathcal{L}_\star^{\boldsymbol{W}^\parallel}.$$

Step 2: Next, we will prove that for any $\boldsymbol{W}^\parallel \in \mathcal{S}_{\text{cyc}}$ with $\|\boldsymbol{W}^\parallel\|_F < \infty$, $\bar{\boldsymbol{W}}_R^\perp$ achieves the optimal risk as $R \to \infty$ – rather than problem having finite optima. It is to show that there is no finite R can achieve optimal

risk. Consider any $\boldsymbol{W} \in \mathcal{S}_{\text{cyc}}^{\perp}$ with $\|\boldsymbol{W}\|_F < \infty$. Let $\boldsymbol{s}_i := \mathbb{S}(\boldsymbol{X}_i(\boldsymbol{W} + \boldsymbol{W}^{\parallel})\bar{\boldsymbol{x}}_i)$ and $\gamma_i^{\boldsymbol{W}} := \boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \boldsymbol{s}_i$. Then we have that

$$\gamma_i^{\boldsymbol{W}} = \sum_{t \in \mathcal{O}_i} s_{it} = \sum_{t \in \mathcal{O}_i} \frac{\sum_{t' \in \mathcal{R}_i} s_{it'}}{\sum_{t' \in [T_i]} s_{it'}} s_{it}^{\max} \le \sum_{t \in \mathcal{O}_i} s_{it}^{\max} = \gamma_i^{\max}$$

where the equality holds when $\mathcal{R}_i = [T_i]$. Since $\boldsymbol{W}^{\text{mm}} \neq 0$, then there exists some $i \in [n]$ that $\gamma_i^{\boldsymbol{W}} < \gamma_i^{\max}$. Therefore, for any finite $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{W} \in \mathcal{S}_{\text{cyc}}^{\perp}$, since loss function is strictly decreasing,

$$\mathcal{L}(\boldsymbol{W} + \boldsymbol{W}^{\parallel}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\gamma_i^{\boldsymbol{W}}) > \frac{1}{n} \sum_{i=1}^{n} \ell(\gamma^{\max}) = \mathcal{L}_\star^{\boldsymbol{W}^{\parallel}}.$$

Step 3: Now, it remains to show that $\bar{\boldsymbol{W}}_R^{\perp}$ converges in direction to $\boldsymbol{W}^{\text{mm}}$. Suppose convergence fails. We will obtain a contradiction by showing that $R \cdot \boldsymbol{W}^{\text{mm}}/\|\boldsymbol{W}^{\text{mm}}\|_F$ achieves a strictly superior loss compared to $\bar{\boldsymbol{W}}_R^{\perp}$. Since $\bar{\boldsymbol{W}}_R^{\perp}$ fails to converge to $\boldsymbol{W}^{\text{mm}}$, for some $\delta > 0$, there exists arbitrarily large $R > 0$ such that

$$\|\bar{\boldsymbol{W}}_R^{\perp} \cdot \|\boldsymbol{W}^{\text{mm}}\|_F / R - \boldsymbol{W}^{\text{mm}}\|_F \ge \delta$$

Let $\boldsymbol{W}' = \bar{\boldsymbol{W}}_R^{\perp} \cdot \|\boldsymbol{W}^{\text{mm}}\|_F / R$ where we have $\|\boldsymbol{W}'\|_F \le \|\boldsymbol{W}^{\text{mm}}\|_F$ and $\boldsymbol{W}' \neq \boldsymbol{W}^{\text{mm}}$. Following proof of Lemma 2, the subspace $\mathcal{S}$ (the span of all matrices $(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top$ for all $(i \asymp j) \in \mathcal{G}^{(k)}$ and $k \in [K]$) is equal to $\mathcal{S}_{\text{cyc}}$. Then we have

$$(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W}' \boldsymbol{e}_k = 0 \quad \text{where} \quad (i \asymp j) \in \mathcal{G}^{(k)}.$$

Then for some $\epsilon := \epsilon(\delta)$, there exists i, j, k such that

$$(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W}' \boldsymbol{e}_k \le 1 - \epsilon \quad \text{where} \quad (i \Rightarrow j) \in \mathcal{G}^{(k)}.$$

Now, we will argue that this leads to a contradiction by proving $\mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}}/\|\boldsymbol{W}^{\text{mm}}\|_F + \boldsymbol{W}^{\parallel}) < \mathcal{L}(\bar{\boldsymbol{W}}_R + \boldsymbol{W}^{\parallel})$ for sufficiently large R. For simplification, we update $R \leftarrow R/\|\boldsymbol{W}^{\text{mm}}\|_F$, and it is equivalent to prove that $\mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}} + \boldsymbol{W}^{\parallel}) < \mathcal{L}(R \cdot \boldsymbol{W}' + \boldsymbol{W}^{\parallel})$ for sufficiently large R.

To obtain the result, we establish a refined softmax probability control as in the proof of Theorem 1 by studying the distance to $\mathcal{L}_\star^{\boldsymbol{W}^{\parallel}}$. Recap the definitions of $\gamma_i^{\max}$ and $\boldsymbol{s}_i^{\max}$, and recap that $\boldsymbol{b}_i = \boldsymbol{X}_i \boldsymbol{W}^{\parallel} \bar{\boldsymbol{x}}_i$, $\boldsymbol{a}_i^\star = \boldsymbol{X}_i(R \cdot \boldsymbol{W}^{\text{mm}})\bar{\boldsymbol{x}}_i$, and $\boldsymbol{s}_i^\star = \mathbb{S}(\boldsymbol{X}_i(R \cdot \boldsymbol{W}^{\text{mm}} + \boldsymbol{W}^{\parallel})\bar{\boldsymbol{x}}_i) = \mathbb{S}(\boldsymbol{a}_i^\star + \boldsymbol{b}_i)$. Additionally, let $\boldsymbol{a}_i^R = \boldsymbol{X}_i(R \cdot \boldsymbol{W}')\bar{\boldsymbol{x}}_i$, $\boldsymbol{s}_i^R = \mathbb{S}(\boldsymbol{X}_i(R \cdot \boldsymbol{W}' + \boldsymbol{W}^{\parallel})\bar{\boldsymbol{x}}_i) = \mathbb{S}(\boldsymbol{a}_i^R + \boldsymbol{b}_i)$, $\gamma_i^\star := \boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \boldsymbol{s}_i^\star$, and $\gamma_i^R := \boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \boldsymbol{s}_i^R$.

From the proof of Lemma 2, subspaces $\mathcal{S} = \mathcal{S}_{\text{cyc}}$. Then we get for all $t, t' \in \mathcal{R}_i$

$$(\boldsymbol{x}_{it} - \boldsymbol{x}_{it'})^\top \boldsymbol{W} \bar{\boldsymbol{x}}_i = 0 \quad \text{for any } \boldsymbol{W} \perp \mathcal{S}_{\text{cyc}} \Longrightarrow a_{it}^\star - a_{it'}^\star = a_{it}^R - a_{it'}^R = 0.$$

Then

$$\sum_{t \in \mathcal{O}_i} s_{it}^R = \frac{\sum_{t \in \mathcal{O}_i} e^{a_{it}^R + b_{it}}}{\sum_{t \in [T_i]} e^{a_{it}^R + b_{it}}} \le \frac{\sum_{t \in \mathcal{O}_i} e^{b_{it}}}{\sum_{t \in \mathcal{R}_i} e^{b_{it}} + \sum_{t \in \bar{\mathcal{R}}_i} e^{-(1-\epsilon)R + b_{it}}} \le \frac{c_i}{d_i + e^{-(1-\epsilon)R - \bar{b}}}, \quad \exists i \in [n]$$

$$\sum_{t \in \mathcal{O}_i} s_{it}^\star = \frac{\sum_{t \in \mathcal{O}_i} e^{a_{it}^\star + b_{it}}}{\sum_{t \in [T_i]} e^{a_{it}^\star + b_{it}}} \ge \frac{\sum_{t \in \mathcal{O}_i} e^{b_{it}}}{\sum_{t \in \mathcal{R}_i} e^{b_{it}} + \sum_{t \in \bar{\mathcal{R}}_i} e^{-R + b_{it}}} \ge \frac{c_i}{d_i + T e^{-R + \bar{b}}}, \quad \forall i \in [n],$$

where $c_i = \sum_{t \in \mathcal{O}_i} e^{b_{it}}$, $d_i = \sum_{t \in \mathcal{R}_i} e^{b_{it}}$, and $\bar{b} := \max_{t \in \mathcal{R}_i, i \in [n]} |b_{it}|$, and we have $\gamma_i^{\max} = c_i/d_i$.

Since ℓ is strictly decreasing and ℓ' is bounded, let $c_{\text{dn}} \le -\ell' \le c_{\text{up}}$ for some constants $c_{\text{dn}}, c_{\text{up}} > 0$. Then we have

$$\begin{aligned}
\mathcal{L}(R \cdot \boldsymbol{W}' + \boldsymbol{W}^{\parallel}) - \mathcal{L}_\star^{\boldsymbol{W}^{\parallel}} &\ge \frac{1}{n} \left(\ell(\gamma_i^R) - \ell(\gamma_i^{\max})\right) \ge \frac{c_{\text{dn}}}{n} \left(\gamma_i^{\max} - \gamma_i^R\right) \\
&= \frac{c_{\text{dn}}}{n} (\boldsymbol{s}_i^{\max} - \boldsymbol{s}_i^R)^\top \boldsymbol{\gamma}_i \\
&\ge \frac{c_{\text{dn}}}{n} \left(\gamma_i^{\max} - \sum_{t \in \mathcal{O}_i} s_{it}^R\right) \\
&\ge \frac{c_{\text{dn}} c_i}{n d_i} \left(1 - \frac{1}{1 + e^{-(1-\epsilon)R - \bar{b}}/d_i}\right)
\end{aligned}$$

and let $j := \max_{i\in[n]} \left(\ell(\gamma_i^\star) - \ell(\gamma_i^{\max})\right)$. We can upper-bound the loss difference for $R \cdot \boldsymbol{W}^{\mathrm{mm}} + \boldsymbol{W}^{\parallel}$ as follows:

$$\begin{aligned}
\mathcal{L}(R \cdot \boldsymbol{W}^{\mathrm{mm}} + \boldsymbol{W}^{\parallel}) - \mathcal{L}_\star^{\boldsymbol{W}^{\parallel}} &\le \max_{i\in[n]} \left(\ell(\gamma_i^\star) - \ell(\gamma_i^{\max})\right) \le c_{\mathrm{up}} \left(\gamma_j^{\max} - \gamma_j^\star\right) \\
&= c_{\mathrm{up}} (\boldsymbol{s}_j^{\max} - \boldsymbol{s}_j^\star)^\top \boldsymbol{\gamma}_j \\
&\le c_{\mathrm{up}} \left(\gamma_j^{\max} - \sum_{t\in\mathcal{O}_j} s_{it}^\star\right) \\
&\le c_{\mathrm{up}} \frac{c_j}{d_j} \left(1 - \frac{1}{1 + Te^{-R+\bar{b}}/d_j}\right) \\
&\le c_{\mathrm{up}} \frac{c_j}{d_j^2} Te^{-R+\bar{b}}.
\end{aligned}$$

Combining them together results in that, $\mathcal{L}(R \cdot \boldsymbol{W}' + \boldsymbol{W}^{\parallel}) > \mathcal{L}(R \cdot \boldsymbol{W}^{\mathrm{mm}} + \boldsymbol{W}^{\parallel})$ whenever

$$\frac{c_{\mathrm{dn}} c_i}{n d_i} \left(1 - \frac{1}{1 + e^{-(1-\epsilon)R-\bar{b}}/d_i}\right) > c_{\mathrm{up}} \frac{c_j}{d_j^2} Te^{-R+\bar{b}} \Longrightarrow R > R_\epsilon := \frac{1}{\epsilon} \log\left(\frac{2nTc_{\mathrm{up}} c_j d_i^2}{c_{\mathrm{dn}} c_i d_j^2}\right) + \frac{2\bar{b}}{\epsilon}. \tag{13}$$

Note that since $\boldsymbol{W}^{\parallel}$ is finite, b_{it} for all $i \in [n], t \in [T_i]$ are bounded, and therefore, $0 < c_i \le d_i < \infty$, $i \in [n]$ and $\bar{b} < \infty$. (13) completes the proof by contradiction. ∎

Now, gathering all the results we have obtained so far, we are ready to prove Theorem 2.

Proof of Theorem 2. Recap the dataset $\mathtt{DSET} = (\boldsymbol{X}_i, y_i)_{i=1}^n$ and index sets $\mathcal{O}_i, \bar{\mathcal{O}}_i, \mathcal{R}_i, \bar{\mathcal{R}}_i$, $i \in [n]$ from (4). Let $\boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$ denote the score vector of ith input and

$$\gamma_{it} = \boldsymbol{x}_{it}^\top \boldsymbol{c}_{y_i} = \begin{cases} 1, & t \in \mathcal{O}_i \\ 0, & t \in \bar{\mathcal{O}}_i. \end{cases}$$

Let $\boldsymbol{s}_i^{\boldsymbol{W}} = \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)$, then the regularization path solution of the ERM problem is defined as follows:

$$\bar{\boldsymbol{W}}_R = \arg\min_{\|\boldsymbol{W}\|_F \le R} \mathcal{L}(\boldsymbol{W}) \quad \text{where} \quad \mathcal{L}(\boldsymbol{W}) = \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)) = \frac{1}{n} \sum_{i=1}^n \ell\left(\sum_{t\in\mathcal{O}_i} s_{it}^{\boldsymbol{W}}\right).$$

Let $\boldsymbol{W}_R^\perp = \boldsymbol{\Pi}_{\mathcal{S}_{\mathrm{cyc}}^\perp}(\bar{\boldsymbol{W}}_R)$ and $\boldsymbol{W}_R^{\parallel} = \boldsymbol{\Pi}_{\mathcal{S}_{\mathrm{cyc}}}(\bar{\boldsymbol{W}}_R)$.

Step 1: We start with showing that $\lim_{R\to\infty} \|\boldsymbol{W}_R^\perp\|_F = \infty$.

Similar to the analysis in the proof of Lemma 4, let $\boldsymbol{a}_i^R = \boldsymbol{X}_i \boldsymbol{W}_R^\perp \bar{\boldsymbol{x}}_i$, $\boldsymbol{b}_i^R = \boldsymbol{X}_i \boldsymbol{W}_R^{\parallel} \bar{\boldsymbol{x}}_i$, and $\boldsymbol{s}_i^R = \mathbb{S}(\boldsymbol{X}_i \bar{\boldsymbol{W}}_R \bar{\boldsymbol{x}}_i) = \mathbb{S}(\boldsymbol{a}_i^R + \boldsymbol{b}_i^R)$. From proof of Lemma 2, for any $t \in \mathcal{R}_i$,

$$\boldsymbol{x}_{it}^\top (\boldsymbol{W}_R^\perp + \boldsymbol{W}_R^{\parallel}) \bar{\boldsymbol{x}}_i = \boldsymbol{x}_{it}^\top \boldsymbol{W}^{\parallel} \bar{\boldsymbol{x}}_i + \bar{a}_i \implies a_{it}^R = \bar{a}_i$$

where $\bar{a}_i$ is some constant associated with $\boldsymbol{W}_R^\perp$. Then for any $i \in [n]$, we get

$$\sum_{t\in\mathcal{O}_i} s_{it}^R = \frac{\sum_{t\in\mathcal{O}_i} e^{a_{it}^R + b_{it}^R}}{\sum_{t\in[T_i]} e^{a_{it}^R + b_{it}^R}} = \frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}^R}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}^R} + \sum_{t\in\bar{\mathcal{R}}_i} e^{b_{it}^R + a_{it}^R - \bar{a}_i}} \le \frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}^R}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}^R}},$$

and given strictly decreasing loss function ℓ, the training risk obeys

$$\mathcal{L}(\bar{\boldsymbol{W}}_R) = \frac{1}{n} \sum_{i=1}^n \ell\left(\sum_{i\in\mathcal{O}_i} s_{it}^R\right) \ge \frac{1}{n} \sum_{i=1}^n \ell\left(\frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}^R}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}^R}}\right),$$

where the equality holds once that for all $t \in \bar{\mathcal{R}}_i$, $a_{it}^R - \bar{a} \to -\infty$ as $R \to \infty$, which implies that $\|\boldsymbol{W}_R^\perp\|_F \to \infty$. Note that $\boldsymbol{W}_R^\perp := \|\boldsymbol{W}_R^\perp\|_F \boldsymbol{W}^{\mathrm{mm}} / \|\boldsymbol{W}^{\mathrm{mm}}\|_F$ reaches the bound, which ensures that

$a_{it}^R - \bar{a} \leq -\|\boldsymbol{W}_R^\perp\|_F / \|\boldsymbol{W}^{\text{mm}}\|_F \to -\infty$. Then the bound is reachable. Given that $\bar{\boldsymbol{W}}_R$ returns the optimal risk, we have

$$\lim_{R\to\infty} \mathcal{L}(\bar{\boldsymbol{W}}_R) = \frac{1}{n}\sum_{i=1}^n \ell\left(\frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}^R}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}^R}}\right). \tag{14}$$

Step 2: Next, we prove that for any $\boldsymbol{W}^{\text{cyc}}$ with $\|\boldsymbol{W}^{\text{cyc}}\|_F < \infty$, $\lim_{R\to\infty} \boldsymbol{W}_R^{\parallel} = \boldsymbol{W}^{\text{cyc}}$.

Consider the ERM problem defined in Definition 3. Recap the definitions of $\mathcal{I}$ and $\bar{\mathcal{I}}$ in (5). We obtain

$$\bar{\mathcal{L}}(\boldsymbol{W}_R^{\parallel}) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \ell(\boldsymbol{c}_{y_i}^\top \bar{\boldsymbol{X}}_i^\top \mathbb{S}(\bar{\boldsymbol{X}}_i \boldsymbol{W}_R^{\parallel} \bar{\boldsymbol{x}}_i)) = \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \ell\left(\frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}^R}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}^R}}\right). \tag{15}$$

Combining (14) and (15) results in that

$$\lim_{R\to\infty} \mathcal{L}(\bar{\boldsymbol{W}}_R) = \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \frac{|\mathcal{I}|}{n}\lim_{R\to\infty} \bar{\mathcal{L}}(\boldsymbol{W}_R^{\parallel}), \tag{16}$$

where for $i \in \bar{\mathcal{I}}$, $\mathcal{O}_i = \mathcal{R}_i$. Following Definition 3 and Lemma 2, where $\boldsymbol{W}^{\text{cyc}}$ is the optimal solution and $\boldsymbol{W}^{\text{cyc}} \in \mathcal{S}_{\text{cyc}}$, and since $\bar{\boldsymbol{W}}_R$ returns the min-norm optimal solution, we have $\lim_{R\to\infty} \boldsymbol{W}_R^{\parallel} = \boldsymbol{W}^{\text{cyc}}$.

Step 3: It remains to prove that $\lim_{R\to\infty} \frac{\bar{\boldsymbol{W}}_R}{R} = \frac{\boldsymbol{W}^{\text{mm}}}{\|\boldsymbol{W}^{\text{mm}}\|_F}$. Since in Step 2, $\lim_{R\to\infty} \boldsymbol{W}_R^{\parallel} = \boldsymbol{W}^{\text{cyc}}$ and $\lim_{R\to\infty} \|\boldsymbol{W}_R^{\parallel}\|_F = \|\boldsymbol{W}^{\text{cyc}}\|_F < \infty$, then the proof is done by Lemma 4. ■

C Global Convergence of Gradient Descent

C.1 Supporting results under the setting of Theorem 3

In this section, we introduce results useful for the main proof. Recap the setting of Theorem 3, and the index sets of $\mathcal{O}_i, \bar{\mathcal{O}}_i, \mathcal{R}_i, \bar{\mathcal{R}}_i, i \in [n]$ in (4). Let $\boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$. Then under Assumption 4, we have

$$\gamma_{it} = \begin{cases} 1, & t \in \mathcal{O}_i \\ 0, & t \in \bar{\mathcal{O}}_i \end{cases} \qquad \text{for all } i \in [n]. \tag{17}$$

Additionally, given loss function $\ell(u) = -\log(u)$ and letting $\boldsymbol{s}_i^{\boldsymbol{W}} = \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)$, we can write the training risk as follows:

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n -\log\left(\sum_{t\in\mathcal{O}_i} s_{it}^{\boldsymbol{W}}\right). \tag{18}$$

• $\nabla\mathcal{L}(\boldsymbol{W})$ **under the setting of Theorem 3.** For any $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, let $\boldsymbol{h}_i = \boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i, \boldsymbol{s}_i = \mathbb{S}(\boldsymbol{h}_i), \boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$.

$$\begin{aligned} \nabla\mathcal{L}(\boldsymbol{W}) &= \frac{1}{n}\sum_{i=1}^n \ell'(\boldsymbol{\gamma}_i^\top \boldsymbol{s}_i) \boldsymbol{X}_i^\top \mathbb{S}'(\boldsymbol{h}) \boldsymbol{\gamma}_i \bar{\boldsymbol{x}}_i^\top \\ &= \frac{1}{n}\sum_{i=1}^n -\frac{1}{\boldsymbol{\gamma}_i^\top \boldsymbol{s}_i} \boldsymbol{X}_i^\top (\text{diag}(\boldsymbol{s}_i) - \boldsymbol{s}_i \boldsymbol{s}_i^\top) \boldsymbol{\gamma}_i \bar{\boldsymbol{x}}_i^\top \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{t\in\bar{\mathcal{O}}_i} s_{it}(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i}) \bar{\boldsymbol{x}}_i^\top \end{aligned} \tag{19}$$

where the last equation uses the fact that for any example $(\boldsymbol{X}_i, y_i) \in \texttt{DSET}$, $i \in [n]$,

$$\begin{aligned} \frac{\boldsymbol{X}_i^\top (\text{diag}(\boldsymbol{s}_i) - \boldsymbol{s}_i \boldsymbol{s}_i^\top) \boldsymbol{\gamma}_i}{\boldsymbol{\gamma}_i^\top \boldsymbol{s}_i} &= \frac{\boldsymbol{X}_i^\top \text{diag}(\boldsymbol{s}_i) \boldsymbol{\gamma}_i}{\boldsymbol{\gamma}_i^\top \boldsymbol{s}_i} - \boldsymbol{X}_i^\top \boldsymbol{s}_i = \frac{\sum_{t\in\mathcal{O}_i} s_{it} \boldsymbol{e}_{y_i}}{\sum_{t\in\mathcal{O}_i} s_{it}} - \boldsymbol{X}_i^\top \boldsymbol{s}_i \\ &= \boldsymbol{e}_{y_i} - \boldsymbol{X}_i^\top \boldsymbol{s}_i = \sum_{t\in\bar{\mathcal{O}}_i} s_{it}(\boldsymbol{e}_{y_i} - \boldsymbol{x}_{it}). \end{aligned}$$

• **Lipschitzness of $\nabla\mathcal{L}(\boldsymbol{W})$.** For any $\boldsymbol{W}, \dot{\boldsymbol{W}} \in \mathbb{R}^{d\times d}$, letting $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}\bar{\boldsymbol{x}}_i)$ and $\dot{\boldsymbol{s}}_i = \mathbb{S}(\boldsymbol{X}_i\dot{\boldsymbol{W}}\bar{\boldsymbol{x}}_i)$, and following (19), we have:

$$\begin{aligned}
\|\nabla\mathcal{L}(\boldsymbol{W}) - \nabla\mathcal{L}(\dot{\boldsymbol{W}})\|_F &\le \frac{1}{n}\sum_{i=1}^{n}\sum_{t\in\bar{\mathcal{O}}_i}(s_{it}-\dot{s}_{it})\|(\boldsymbol{x}_{it}-\boldsymbol{e}_{y_i})\bar{\boldsymbol{x}}_i^\top\|_F \\
&\le \frac{2\|\boldsymbol{E}\|^2}{n}\sum_{i=1}^{n}\sum_{t\in\bar{\mathcal{O}}_i}(s_{it}-\dot{s}_{it}) \\
&\le \frac{2\|\boldsymbol{E}\|^2}{n}\sum_{i=1}^{n}\sum_{t\in\bar{\mathcal{O}}_i}|s_{it}-\dot{s}_{it}| \\
&\le \frac{2\|\boldsymbol{E}\|^2}{n}\sum_{i=1}^{n}\|\boldsymbol{s}_i-\dot{\boldsymbol{s}}_i\|_1 \\
&\le \frac{2\|\boldsymbol{E}\|^2}{n}\sum_{i=1}^{n}\sqrt{T_i}\cdot\|\boldsymbol{s}_i-\dot{\boldsymbol{s}}_i\|.
\end{aligned} \tag{20}$$

It remains to bound $\|\boldsymbol{s}_i - \dot{\boldsymbol{s}}_i\|$. For any $\boldsymbol{s}, \dot{\boldsymbol{s}}$,

$$\begin{aligned}
\|\boldsymbol{s}-\dot{\boldsymbol{s}}\| &= \|\mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}}) - \mathbb{S}(\boldsymbol{X}\dot{\boldsymbol{W}}\bar{\boldsymbol{x}})\| \\
&\le \|\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}} - \boldsymbol{X}\dot{\boldsymbol{W}}\bar{\boldsymbol{x}}\| \\
&\le \|\boldsymbol{E}\|^2\|\boldsymbol{W}-\dot{\boldsymbol{W}}\|_F.
\end{aligned} \tag{21}$$

Combining results in that

$$\|\nabla\mathcal{L}(\boldsymbol{W}) - \nabla\mathcal{L}(\dot{\boldsymbol{W}})\|_F \le 2\sqrt{T_{\max}}\cdot\|\boldsymbol{E}\|^4\|\boldsymbol{W}-\dot{\boldsymbol{W}}\|_F \tag{22}$$

where $T_{\max} := \max_{i\in[n]} T_i$. Then let

$$L := 2\sqrt{T_{\max}}\cdot\|\boldsymbol{E}\|^4 \tag{23}$$

and $\nabla\mathcal{L}(\boldsymbol{W})$ is L-Lipschitz continuous.

C.2 Proof of Lemma 3

Proof.

• **We first prove that $\mathcal{L}(\boldsymbol{W})$ is convex.** Let $(\boldsymbol{X}, y)$ be an arbitrary pair of input sequence and label. Let $\boldsymbol{v}(\boldsymbol{W}) := \boldsymbol{E}\boldsymbol{W}\bar{\boldsymbol{x}}$ and $v_k = \boldsymbol{e}_k^\top\boldsymbol{W}\bar{\boldsymbol{x}}$ for $k\in[K]$ and let m_k be the number of token ID k inside input sequence $\boldsymbol{X}$. By Assumption 4 and log-loss, we know that

$$\ell(\boldsymbol{v}) := \ell(\boldsymbol{c}_y^\top\boldsymbol{X}^\top\mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}})) = -\log\left(\frac{m_y\cdot e^{v_y}}{\sum_{k\in[K]} m_k\cdot e^{v_k}}\right) = \log\left(\sum_{k\in[K]} m_k\cdot e^{v_k}\right) - \log(m_y\cdot e^{v_y}).$$

Since $\mathcal{L}(\boldsymbol{W})$ is the summation of $\ell(\boldsymbol{v})$ and $\boldsymbol{v}$ is a linear transformation of $\boldsymbol{W}$, we will prove $\mathcal{L}(\boldsymbol{W})$ is convex by showing that $\ell(\boldsymbol{v})$ is convex with respect to $\boldsymbol{v}$.

Let $\boldsymbol{z}\in\mathbb{R}^K$ be a vector such that the kth element of $\boldsymbol{z}$ is $z_k = m_k\cdot e^{v_k}$. Then, the Hessian matrix of $\ell(\boldsymbol{v})$ is

$$\nabla^2\ell(\boldsymbol{v}) = \frac{1}{(\boldsymbol{1}^\top\boldsymbol{z})^2}\left((\boldsymbol{1}^\top\boldsymbol{z})\mathrm{diag}(\boldsymbol{z}) - \boldsymbol{z}\boldsymbol{z}^\top\right)$$

For any $\boldsymbol{u}\in\mathbb{R}^K$, we obtain that

$$\boldsymbol{u}^\top\nabla^2\ell(\boldsymbol{v})\boldsymbol{u} = \frac{1}{(\boldsymbol{1}^\top\boldsymbol{z})^2}\left(\left(\sum_{k=1}^K z_k\right)\left(\sum_{k=1}^K u_k^2 z_k\right) - \left(\sum_{k=1}^K u_k z_k\right)^2\right) \ge 0. \tag{24}$$

Since $\boldsymbol{z}_k > 0$, $k\in[K]$, (24) follows from the Cauchy-Schwarz inequality $(\boldsymbol{\alpha}^\top\boldsymbol{\alpha})(\boldsymbol{\beta}^\top\boldsymbol{\beta}) \ge (\boldsymbol{\alpha}^\top\boldsymbol{\beta})^2$ applied to the vectors with $\alpha_i = u_i\sqrt{z_i}$ and $\beta_i = \sqrt{z_i}$. The equality condition holds $k\boldsymbol{\alpha} = \boldsymbol{\beta}$ for $k\neq 0$. This means that $\ell(\boldsymbol{v})$ is convex.

• **Next, we will show that $\mathcal{L}(\boldsymbol{W})$ is strictly convex on subspace $\mathcal{S}_{\text{cyc}}$.** Following Definition 3, let us define the subspace $\mathcal{S}_{\text{cyc}}^{ik}$ for $k \in [K], i \in [N_k]$ such that it is equal to the span of all matrices $(\boldsymbol{e}_j - \bar{\boldsymbol{e}}_i^{(k)})\boldsymbol{e}_k^\top$ for $j \in \mathcal{C}_i^{(k)}$. Assume that $\mathcal{L}(\boldsymbol{W})$ is not strictly convex on $\mathcal{S}_{\text{cyc}}$. This means that there exist $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathcal{S}_{\text{cyc}}$, $\|\boldsymbol{W}_2\|_F > 0$ and $0 < \lambda < 1$ such that

$$\mathcal{L}((1-\lambda)\boldsymbol{W}_1 + \lambda(\boldsymbol{W}_1 + \boldsymbol{W}_2)) = (1-\lambda)\mathcal{L}(\boldsymbol{W}_1) + \lambda\mathcal{L}(\boldsymbol{W}_1 + \boldsymbol{W}_2). \tag{25}$$

We will show that $\|\Pi_{\mathcal{S}_{\text{cyc}}^{ik}}(\boldsymbol{W}_2)\|_F = 0$ for all i, k. Let us first assume that there exist $\bar{i}, \bar{k}$ such that $\|\Pi_{\mathcal{S}_{\text{cyc}}^{\bar{i}\bar{k}}}(\boldsymbol{W}_2)\|_F > 0$ and let $\boldsymbol{W}_1^{\bar{i}\bar{k}} = \Pi_{\mathcal{S}_{\text{cyc}}^{\bar{i}\bar{k}}}(\boldsymbol{W}_1)$ and $\boldsymbol{W}_2^{\bar{i}\bar{k}} = \Pi_{\mathcal{S}_{\text{cyc}}^{\bar{i}\bar{k}}}(\boldsymbol{W}_2)$. Let $(\boldsymbol{X}_i, y_i)_{i=1}^m$ be all input sequence label pairs of $\overline{\texttt{DSET}}$ that induces an edge in $\mathcal{C}_{\bar{i}}^{(\bar{k})}$. For $i \in [m]$, let $\boldsymbol{v}_i = \boldsymbol{X}_i\boldsymbol{W}_1^{\bar{i}\bar{k}}\boldsymbol{e}_{\bar{k}}$, $\ell(\boldsymbol{v}_i) = \ell(\boldsymbol{c}_{y_i}\boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}_1^{\bar{i}\bar{k}}\boldsymbol{e}_{\bar{k}}))$, and $\boldsymbol{u} \in \mathbb{R}^K$ such that $u_k = \boldsymbol{e}_k^\top \boldsymbol{W}_2^{\bar{i}\bar{k}}\boldsymbol{e}_{\bar{k}}$ for $k \in [K]$. Combining (25) and the fact that $\ell(\boldsymbol{v})$ is convex, we obtain

$$\boldsymbol{u}^\top \nabla^2 \ell(\boldsymbol{v}_i)\boldsymbol{u} = 0 \qquad \forall i \in [m]. \tag{26}$$

Combining the equality condition of Cauchy-Schwarz inequality and the fact that there exists a way to go between two nodes if they are in the same SCC, we obtain that $u_i = u_j$ for each $i, j \in \mathcal{C}_{\bar{i}}^{(\bar{k})}$ (Otherwise $\sum_i \boldsymbol{u}^\top \nabla^2 \ell(\boldsymbol{v}_i)\boldsymbol{u} > 0$). This implies that $(\mathrm{e}_i - \mathrm{e}_j)^\top \boldsymbol{W}_2^{\bar{i}\bar{k}}\mathrm{e}_k = 0$. Therefore, we have the following:

$$0 = \frac{1}{|\mathcal{C}_i^{(k)}|}\sum_{j' \in \mathcal{C}_i^{(k)}} (\boldsymbol{e}_j - \boldsymbol{e}_{j'})\boldsymbol{W}_2^{\bar{i}\bar{k}}\boldsymbol{e}_k^\top = \left(\boldsymbol{e}_j - \frac{1}{|\mathcal{C}_i^{(k)}|}\sum_{j' \in \mathcal{C}_i^{(k)}} \boldsymbol{e}_{j'}\right)\boldsymbol{W}_2^{\bar{i}\bar{k}}\boldsymbol{e}_k^\top = (\boldsymbol{e}_j - \bar{\boldsymbol{e}}_i^{(k)})\boldsymbol{W}_2^{\bar{i}\bar{k}}\boldsymbol{e}_k^\top$$

This means that $\boldsymbol{W}_2^{\bar{i}\bar{k}} \perp \mathcal{S}_{\text{cyc}}^{\bar{i}\bar{k}}$, which is contradiction. Therefore, $\mathcal{L}(\boldsymbol{W})$ is strictly convex on $\mathcal{S}_{\text{cyc}}$. ■

C.3 Divergence of $\|\mathbf{W}(\tau)\|_F$

We first propose the following lemmas showing the descent property of gradient descent for $\mathcal{L}(\boldsymbol{W})$ (Lemma 5) and the correlation between $\nabla\mathcal{L}(\boldsymbol{W})$ and the solution of (Graph-SVM) $\boldsymbol{W}^{\text{mm}}$ (Lemma 6) under the setting of Theorem 3. The proofs in this section follow Appendix B.1 of Tarzanagh et al. (2023a).

Lemma 5 (Descent Lemma) *Consider the loss in (18) and choose step size $\eta \leq 1/L$ where L is the Lipschitzness of $\nabla\mathcal{L}(\boldsymbol{W})$ defined in (23). Then from any initialization $\boldsymbol{W}(0)$, Algorithm Algo-GD satisfies:*

$$\mathcal{L}(\boldsymbol{W}(\tau+1)) - \mathcal{L}(\boldsymbol{W}(\tau)) \leq -\frac{\eta}{2}\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F^2$$

for all $\tau \geq 0$. Additionally, it holds that $\sum_{\tau=0}^{\infty}\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F^2 < \infty$, and $\lim_{\tau\to\infty}\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F^2 = 0$

Proof. From (Algo-GD), for $\tau \geq 0$, we have that $\boldsymbol{W}(\tau+1) = \boldsymbol{W}(\tau) - \eta\nabla\mathcal{L}(\boldsymbol{W}(\tau))$. Since $\mathcal{L}(\boldsymbol{W})$ is L-smooth following (23), we get

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}(\tau+1)) &\leq \mathcal{L}(\boldsymbol{W}(\tau)) + \langle\nabla\mathcal{L}(\boldsymbol{W}(\tau)), \boldsymbol{W}(\tau+1) - \boldsymbol{W}(\tau)\rangle + \frac{L}{2}\|\boldsymbol{W}(\tau+1) - \boldsymbol{W}(\tau)\|_F^2 \\
&= \mathcal{L}(\boldsymbol{W}(\tau)) - \eta\cdot\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F^2 + \frac{L\eta^2}{2}\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F^2 \\
&= \mathcal{L}(\boldsymbol{W}(\tau)) - \eta\left(1 - \frac{L\eta}{2}\right)\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F^2 \\
&\leq \mathcal{L}(\boldsymbol{W}(\tau)) - \frac{\eta}{2}\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F^2.
\end{aligned}$$

The inequality above also indicates that

$$\sum_{\tau=0}^{\infty}\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F^2 \leq \frac{2}{\eta}(\mathcal{L}(\boldsymbol{W}(0)) - \mathcal{L}^*) < \infty, \quad \text{and} \quad \lim_{\tau\to\infty}\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F^2 = 0.$$

■

Lemma 6 *Let $\boldsymbol{W}^{mm}$ be the SVM solution of (Graph-SVM) and suppose $\boldsymbol{W}^{mm} \neq 0$. For all $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ with $\|\boldsymbol{W}\|_F < \infty$, the training loss $\mathcal{L}(\boldsymbol{W})$ in (18) obeys $\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{mm}\rangle < 0$.*

Proof. Recap $\mathcal{O}_i, \bar{\mathcal{O}}_i, \mathcal{R}_i, \bar{\mathcal{R}}_i, i \in [n]$ in (4). From (19), for any $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, we obtain the gradient

$$\nabla\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t\in\bar{\mathcal{O}}_i} s_{it}(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i})\bar{\boldsymbol{x}}_i^\top.$$

Then

$$\begin{aligned}
\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{\mathrm{mm}}\rangle &= \frac{1}{n}\sum_{i\in[n]}\sum_{t\in\bar{\mathcal{O}}_i} \left\langle s_{it}(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i})\bar{\boldsymbol{x}}_i^\top, \boldsymbol{W}^{\mathrm{mm}}\right\rangle \\
&= \frac{1}{n}\sum_{i\in[n]}\sum_{t\in\bar{\mathcal{O}}_i} s_{it}\cdot \mathrm{trace}\left((\boldsymbol{W}^{\mathrm{mm}})^\top(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i})\bar{\boldsymbol{x}}_i^\top\right) \\
&= \frac{1}{n}\sum_{i\in[n]}\sum_{t\in\bar{\mathcal{O}}_i} s_{it}\cdot \bar{\boldsymbol{x}}_i(\boldsymbol{W}^{\mathrm{mm}})^\top(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i}).
\end{aligned}$$

From the (Graph-SVM) formulation, we have that $\bar{\boldsymbol{x}}_i(\boldsymbol{W}^{\mathrm{mm}})^\top(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i}) = 0$ for $t \in \mathcal{R}_i$ and $\bar{\boldsymbol{x}}_i(\boldsymbol{W}^{\mathrm{mm}})^\top(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i}) = -1$ for $t \in \bar{\mathcal{R}}_i$. Then $\boldsymbol{W}^{\mathrm{mm}} \neq 0$ ensures that there exists $i \in [n]$ such that $\bar{\mathcal{R}}_i \neq \emptyset$, which implies that

$$\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{\mathrm{mm}}\rangle < 0.$$

It completes the proof. ∎

The next theorem proves the divergence of norm of the iterates $\boldsymbol{W}(\tau)$.

Theorem 4 *Consider the same setting as in Theorem 3, then there is no finite $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ satisfying $\nabla\mathcal{L}(\boldsymbol{W}) = 0$. Furthermore, Algorithm Algo-GD with the step size $\eta \leq 1/L$ where L is the Lipschitzness of $\nabla\mathcal{L}(\boldsymbol{W})$ defined in (23) and any starting point $\boldsymbol{W}(0)$ satisfies $\lim_{\tau\to\infty}\|\boldsymbol{W}(\tau)\|_F = \infty$.*

Proof. Following Lemma 5, when using log-loss $\ell(u) = -\log(u)$, for any starting point $\boldsymbol{W}(0)$, the Algorithm Algo-GD satisfies $\lim_{\tau\to\infty}\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F^2 = 0$. Moreover, assume that the first claim is wrong and that there is a finite critical point $\boldsymbol{W}$ that satisfies $\nabla\mathcal{L}(\boldsymbol{W}) = 0$. We then have $\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{\mathrm{mm}}\rangle = 0$. This leads to a contradiction with Lemma 6 which says that for any finite $\boldsymbol{W}$, $\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{\mathrm{mm}}\rangle < 0$. This implies that $\|\boldsymbol{W}(\tau)\|_F \to \infty$. ∎

C.4 Proof of Theorem 3

Lemma 7 *Consider the setting of Theorem 3. $\boldsymbol{W}^{cyc}$ defined in Definition 3 is unique and finite.*

Proof. To start with, recap the definition of $\overline{\mathtt{DSET}}$ (Definition 1) and $\boldsymbol{W}^{\mathrm{cyc}}$ (Definition 3). Denote $\mathcal{I}$ following (5), and let $\boldsymbol{s}_i = \mathbb{S}(\bar{\boldsymbol{X}}_i\boldsymbol{W}\bar{\boldsymbol{x}}_i)$ where $(\bar{\boldsymbol{X}}_i, y_i) \in \overline{\mathtt{DSET}}$. What's more, recap the ERM loss under Assumption 4 from (18) and loss function $\ell(u) = -\log(u)$. Then we have

$$\boldsymbol{W}^{\mathrm{cyc}} = \arg\min_{\boldsymbol{W}} \bar{\mathcal{L}}(\boldsymbol{W}) \quad \text{where} \quad \bar{\mathcal{L}}(\boldsymbol{W}) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} -\log\left(\sum_{t\in\mathcal{O}_i} s_{it}\right). \tag{27}$$

Here the input sequence $\bar{\boldsymbol{X}}_i \neq \boldsymbol{X}_i$, since it removes all the tokens that are associated with singleton SCCs within their corresponding NTG. Therefore, different from (4), $\mathcal{O}_i$ in (27) is defined as follows:

$$\mathcal{O}_i := \left\{t \mid \boldsymbol{x}_{it} = \boldsymbol{e}_{y_i},\ t \in [\bar{T}_i]\right\},$$

where $\bar{T}_i$ is the number of tokens – tokens that are in the same SCC as label token $\boldsymbol{e}_{y_i}$ within their corresponding NTG – in $\bar{\boldsymbol{X}}_i$ and if recap the notation of $\mathcal{R}_i$ in (4), we have $\bar{T}_i = |\mathcal{R}_i|$.

From Lemma 2, we have known that $\boldsymbol{W}^{\mathrm{cyc}} \in \mathcal{S}_{\mathrm{cyc}}$. What's more for any $\boldsymbol{W}$ with $\|\boldsymbol{W}\|_F < \infty$, we have $\bar{\mathcal{L}}(\boldsymbol{W}) < \infty$. It can be easily seen since $\mathbb{S}(\bar{\boldsymbol{X}}_i\boldsymbol{W}\bar{\boldsymbol{x}}_i)$ will not have zero entries (thanks to the finite $\|\boldsymbol{W}\|_F$) and then $-\log\left(\sum_{t\in\mathcal{O}_i} s_{it}\right)$ in (27) cannot reach infinite number. Next we will prove that $\boldsymbol{W}^{\mathrm{cyc}}$ is finite by

contradiction. Specifically, we will show that for any $\boldsymbol{W} \in \mathcal{S}_{\text{cyc}}$, $\|\boldsymbol{W}\|_F \to \infty$, $\bar{\mathcal{L}}(\boldsymbol{W}) = \infty$ which implies that the optimal solution $\boldsymbol{W}^{\text{cyc}}$ has to be finite.

Let $\boldsymbol{W} \in \mathcal{S}_{\text{cyc}}$ be any attention weight with $\|\boldsymbol{W}\|_F = \infty$. Consider any SCC of dataset $\overline{\text{DSET}}$, and denote it as $\mathcal{C}_i^{(k)}, i \in [N_k], k \in [K]$. Then following the definition of $\mathcal{S}_{\text{cyc}}$ as in Def. 3, we have that $(\boldsymbol{e}_j - \bar{\boldsymbol{e}})\boldsymbol{e}_k^\top \in \mathcal{S}_{\text{cyc}}$ for all $j \in \mathcal{C}_i^{(k)}$ and $k \in [K]$. Let $a_{ijk} = (\boldsymbol{e}_j - \bar{\boldsymbol{e}})^\top \boldsymbol{W} \boldsymbol{e}_k$. Since $\|\boldsymbol{W}\|_F = \infty$, there exist $j' \in \mathcal{C}_{i'}^{(k')}, i' \in [N_{k'}], k' \in [K]$ such that $a_{i'j'k'} = \infty$. Let $\bar{\boldsymbol{e}}' = \frac{1}{|\mathcal{C}_{i'}^{(k')}|} \sum_{j \in \mathcal{C}_{i'}^{(k')}} \boldsymbol{e}_j$. Then for any $\bar{\boldsymbol{X}}$ with last token $\bar{\boldsymbol{x}} = \boldsymbol{e}_{k'}$, and any $t \in [\bar{T}]$, softmax outputs

$$s_t = \frac{e^{\boldsymbol{x}_t^\top \boldsymbol{W} \boldsymbol{e}_{k'}}}{\sum_{t' \in [\bar{T}]} e^{\boldsymbol{x}_t^\top \boldsymbol{W} \boldsymbol{e}_{k'}}} = \frac{e^{(\boldsymbol{x}_t - \bar{\boldsymbol{e}}')^\top \boldsymbol{W} \boldsymbol{e}_{k'}}}{\sum_{t' \in [\bar{T}]} e^{(\boldsymbol{x}_t - \bar{\boldsymbol{e}}')^\top \boldsymbol{W} \boldsymbol{e}_{k'}}}$$

where $\boldsymbol{x}_t$ has label $x_t \in \mathcal{C}_{i'}^{(k')}$. Therefore to prevent softmax outputing zero probability, for all $j \in \mathcal{C}_{i'}^{(k')}$, $(\boldsymbol{e}_j - \bar{\boldsymbol{e}}')^\top \boldsymbol{W} \boldsymbol{e}_{k'} = \infty$. Then we have the sum

$$\sum_{j \in \mathcal{C}_{i'}^{(k')}} (\boldsymbol{e}_j - \bar{\boldsymbol{e}}')^\top \boldsymbol{W} \boldsymbol{e}_{k'} = \infty \neq \left(\left(\sum_{j \in \mathcal{C}_{i'}^{(k')}} \boldsymbol{e}_j \right) - |\mathcal{C}_{i'}^{(k')}| \bar{\boldsymbol{e}}' \right)^\top \boldsymbol{W} \boldsymbol{e}_{k'} = 0,$$

which shows that: If attention weights $\boldsymbol{W} \in \mathcal{S}_{\text{cyc}}$ and $\|\boldsymbol{W}\|_F \to \infty$, there must exist at least one training example that outputs zero softmax probability on the set $\mathcal{O}_i$, which then results in infinite training risk. Therefore, the optimal solution $\boldsymbol{W}^{\text{cyc}}$ cannot have infinite norm.

Next, the strict convexity of $\bar{\mathcal{L}}(\boldsymbol{W})$, as substantiated by a proof similar to that presented in Lemma 3, confirms the uniqueness of the solution $\boldsymbol{W}^{\text{cyc}}$. ■

Proof of Theorem 3.

Now gathering all the results so far, we are ready to prove the gradient descent convergence.

• **We start with showing that $\boldsymbol{W}(\tau)/\|\boldsymbol{W}(\tau)\|_F \to \boldsymbol{W}^{\text{mm}}/\|\boldsymbol{W}^{\text{mm}}\|_F$.**

Consider any $\boldsymbol{W} \in \mathbb{R}^{d \times d}$, and let $\boldsymbol{W}^\perp = \boldsymbol{\Pi}_{\mathcal{S}_{\text{cyc}}^\perp}(\boldsymbol{W})$, $\boldsymbol{W}^\parallel = \boldsymbol{\Pi}_{\mathcal{S}_{\text{cyc}}}(\boldsymbol{W})$, and $R = \|\boldsymbol{W}^\perp\|_F$. Since $\mathcal{L}(\boldsymbol{W})$ is convex following Lemma 3, we have that

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}) &\leq \mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}} + \boldsymbol{W}^\parallel) + \left\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{W} - \left(R \cdot \boldsymbol{W}^{\text{mm}} + \boldsymbol{W}^\parallel \right) \right\rangle \\
&= \mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}} + \boldsymbol{W}^\parallel) + \left\langle \nabla \mathcal{L}(\boldsymbol{W}), \left(\boldsymbol{W}^\perp - R \cdot \boldsymbol{W}^{\text{mm}} \right) \right\rangle \\
&= \mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}} + \boldsymbol{W}^\parallel) + \left\langle \boldsymbol{\Pi}_{\mathcal{S}_{\text{cyc}}^\perp}(\nabla \mathcal{L}(\boldsymbol{W})), \left(\boldsymbol{W}^\perp - R \cdot \boldsymbol{W}^{\text{mm}} \right) \right\rangle,
\end{aligned} \tag{28}$$

Here, the first inequality uses the convexity of $\mathcal{L}(\boldsymbol{W})$ and last equation is obtained from the fact that $\boldsymbol{W}^\perp$, $\boldsymbol{W}^{\text{mm}} \perp \mathcal{S}_{\text{cyc}}$.

Next, from Lemma 5, for any $\tau \geq 0$, $\mathcal{L}(\boldsymbol{W}(\tau + 1)) \leq \mathcal{L}(\boldsymbol{W}(\tau))$. Let $\boldsymbol{W}^\perp(\tau) = \boldsymbol{\Pi}_{\mathcal{S}_{\text{cyc}}^\perp}(\boldsymbol{W}(\tau))$ and $\boldsymbol{W}^\parallel(\tau) = \boldsymbol{\Pi}_{\mathcal{S}_{\text{cyc}}}(\boldsymbol{W}(\tau))$. Recap (16) in the proof of Theorem 2. We obtain

$$\mathcal{L}(\boldsymbol{W}(\tau)) \geq \frac{|\bar{\mathcal{I}}|}{n} \ell(1) + \frac{|\mathcal{I}|}{n} \bar{\mathcal{L}}(\boldsymbol{W}^\parallel(\tau)).$$

Since following Lemma 7, the training risk $\bar{\mathcal{L}}(\boldsymbol{W}^\parallel(\tau))$ is infinite if $\|\boldsymbol{W}^\parallel(\tau)\|_F \to \infty$, which contradicts with Lemma 5. Hence for any $\tau \geq 0$, $\|\boldsymbol{W}^\parallel(\tau)\|_F < \infty$. Since additionally Theorem 4 proves the divergence of $\boldsymbol{W}(\tau)$ as $\tau \to \infty$, we have $\|\boldsymbol{W}^\perp(\tau)\|_F \to \infty$.

Applying Lemma 4, as well as the fact that $\|\boldsymbol{W}^\parallel(\tau)\|_F < \infty$ and $\|\boldsymbol{W}^\perp(\tau)\|_F \to \infty$, there exists sufficiently large R_ϵ as defined in (13) such that once $\|\boldsymbol{W}^\perp(\tau)\|_F = R > \max\{R_\epsilon, 0.5\}$, $\mathcal{L}(\boldsymbol{W}(\tau)) - \mathcal{L}(R \cdot \boldsymbol{W}^{\text{mm}}(\tau) + \boldsymbol{W}^\parallel(\tau)) > 0$. Note that although R_ϵ defined in (13) is associated with the projection component $\boldsymbol{W}^\parallel(\tau)$, since $\boldsymbol{W}^\parallel(\tau)$ is bounded for all $\tau \geq 0$, R_ϵ is also bounded by some worst-case choices of $c_i, d_i, c_j, d_j, \bar{b}$, which are all finite. Now we choose τ_0 such that for all $\tau \geq \tau_0$, $\|\boldsymbol{W}^\perp(\tau)\|_F > \max\{R_\epsilon, 0.5\}$. Then for

$\tau > \tau_0$, we get

$$\left\langle \boldsymbol{W}^\perp(\tau+1) - \boldsymbol{W}^\perp(\tau), \frac{\boldsymbol{W}^{\rm mm}}{\|\boldsymbol{W}^{\rm mm}\|_F} \right\rangle \geq \left\langle \boldsymbol{W}^\perp(\tau+1) - \boldsymbol{W}^\perp(\tau), \frac{\boldsymbol{W}^\perp(\tau)}{\|\boldsymbol{W}^\perp(\tau)\|_F} \right\rangle \tag{29}$$

$$= \frac{1}{2\|\boldsymbol{W}^\perp(\tau)\|_F} \left(\|\boldsymbol{W}^\perp(\tau+1)\|_F^2 - \|\boldsymbol{W}^\perp(\tau)\|_F^2 - \|\boldsymbol{W}^\perp(\tau+1) - \boldsymbol{W}^\perp(\tau)\|_F^2 \right)$$

$$\geq \frac{\|\boldsymbol{W}^\perp(\tau+1)\|_F^2 - \|\boldsymbol{W}^\perp(\tau)\|_F^2}{2\|\boldsymbol{W}^\perp(\tau)\|_F} - \|\boldsymbol{W}^\perp(\tau+1) - \boldsymbol{W}^\perp(\tau)\|_F^2 \tag{30}$$

$$\geq \|\boldsymbol{W}^\perp(\tau+1)\|_F - \|\boldsymbol{W}^\perp(\tau)\|_F - \|\boldsymbol{W}^\perp(\tau+1) - \boldsymbol{W}^\perp(\tau)\|_F^2 \tag{31}$$

$$\geq \|\boldsymbol{W}^\perp(\tau+1)\|_F - \|\boldsymbol{W}^\perp(\tau)\|_F - \|\boldsymbol{W}(\tau+1) - \boldsymbol{W}(\tau)\|_F^2 \tag{32}$$

$$\geq \|\boldsymbol{W}^\perp(\tau+1)\|_F - \|\boldsymbol{W}^\perp(\tau)\|_F + 2\eta \left(\mathcal{L}(\boldsymbol{W}(\tau+1)) - \mathcal{L}(\boldsymbol{W}(\tau)) \right). \tag{33}$$

Here, (29) is obtained from (28) and holds for all $\tau > \tau_0$; (30) comes from the fact that $\|\boldsymbol{W}^\perp(\tau)\|_F > 0.5$; (31) follows that for any $a, b > 0$, $(a^2 - b^2)/2b > a - b$; (32) follows the projection property that $\|\boldsymbol{W}(\tau+1) - \boldsymbol{W}(\tau)\|_F^2 = \|\boldsymbol{W}^\perp(\tau+1) - \boldsymbol{W}^\perp(\tau)\|_F^2 + \|\boldsymbol{W}^\parallel(\tau+1) - \boldsymbol{W}^\parallel(\tau)\|_F^2$; and (33) is obtained via Lemma 5.

Summing the above inequality over $\tau \geq \tau_0$ obtains

$$\left\langle \boldsymbol{W}^\perp(\tau) - \boldsymbol{W}^\perp(\tau_0), \frac{\boldsymbol{W}^{\rm mm}}{\|\boldsymbol{W}^{\rm mm}\|_F} \right\rangle \geq \|\boldsymbol{W}^\perp(\tau)\|_F - \|\boldsymbol{W}^\perp(\tau_0)\|_F + 2\eta \left(\mathcal{L}(\boldsymbol{W}(\tau)) - \mathcal{L}(\boldsymbol{W}(\tau_0)) \right)$$

$$\implies \left\langle \frac{\boldsymbol{W}^\perp(\tau)}{\|\boldsymbol{W}^\perp(\tau)\|_F}, \frac{\boldsymbol{W}^{\rm mm}}{\|\boldsymbol{W}^{\rm mm}\|_F} \right\rangle \geq 1 + \frac{\left\langle \boldsymbol{W}^\perp(\tau_0), \frac{\boldsymbol{W}^{\rm mm}}{\|\boldsymbol{W}^{\rm mm}\|_F} \right\rangle - \|\boldsymbol{W}^\perp(\tau_0)\|_F + 2\eta \left(\mathcal{L}(\boldsymbol{W}(\tau)) - \mathcal{L}(\boldsymbol{W}(\tau_0)) \right)}{\|\boldsymbol{W}^\perp(\tau)\|_F}.$$

Since $\|\boldsymbol{W}^\perp(\tau)\|_F \to \infty$ and $0 < \mathcal{L}(\boldsymbol{W}(\tau)) \leq \mathcal{L}(\boldsymbol{W}(0)) < \infty$, we get

$$\lim_{\tau\to\infty} \left\langle \frac{\boldsymbol{W}^\perp(\tau)}{\|\boldsymbol{W}^\perp(\tau)\|_F}, \frac{\boldsymbol{W}^{\rm mm}}{\|\boldsymbol{W}^{\rm mm}\|_F} \right\rangle = 1. \tag{34}$$

Combining (34) with the fact that $\lim_{\tau\to\infty} \|\boldsymbol{W}^\parallel(\tau)\|_F < \infty$ completes the proof.

• **We next show that $\boldsymbol{\Pi}_{\mathcal{S}_{\rm cyc}}(\boldsymbol{W}(\tau)) \to \boldsymbol{W}^{\rm cyc}$.** From Lemma 5 we have that $\lim_{\tau\to\infty} \|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F = 0$. Then we will prove it by contradiction, that is to show $\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F \not\to 0$ if $\boldsymbol{\Pi}_{\mathcal{S}_{\rm cyc}}(\boldsymbol{W}(\tau)) \not\to \boldsymbol{W}^{\rm cyc}$.

Consider any $\boldsymbol{W} := \boldsymbol{W}(\tau) \in \mathbb{R}^{d\times d}$. Let $\boldsymbol{W} = R \cdot \boldsymbol{W}^{\rm mm} + \boldsymbol{W}^\parallel + \boldsymbol{W}'$ where $\boldsymbol{W}^\parallel = \boldsymbol{\Pi}_{\mathcal{S}_{\rm cyc}}(\boldsymbol{W})$, $R \cdot \boldsymbol{W}^{\rm mm} + \boldsymbol{W}' = \boldsymbol{\Pi}_{\mathcal{S}_{\rm cyc}^\perp}(\boldsymbol{W})$ and $\langle \boldsymbol{W}^{\rm mm}, \boldsymbol{W}' \rangle = 0$. From (34), we get

$$\lim_{\tau\to\infty} \left\langle \frac{R \cdot \boldsymbol{W}^{\rm mm} + \boldsymbol{W}'}{\|R \cdot \boldsymbol{W}^{\rm mm} + \boldsymbol{W}'\|_F}, \frac{\boldsymbol{W}^{\rm mm}}{\|\boldsymbol{W}^{\rm mm}\|_F} \right\rangle = \lim_{\tau\to\infty} \frac{R \cdot \|\boldsymbol{W}^{\rm mm}\|_F}{\sqrt{R^2 \cdot \|\boldsymbol{W}^{\rm mm}\|_F^2 + \|\boldsymbol{W}'\|_F^2}} = 1,$$

which implies that as $\tau \to \infty$, $\|\boldsymbol{W}'\|_F / (R \cdot \|\boldsymbol{W}^{\rm mm}\|_F) \to 0$, and since $\boldsymbol{\Pi}_{\mathcal{S}_{\rm cyc}^\perp}(\boldsymbol{W}(\tau)) \to \infty$, $R \to \infty$.

Since $\mathcal{L}(\boldsymbol{W})$ is convex (obtained via Lemma 3), we have that

$$\mathcal{L}(\boldsymbol{W}) \leq \mathcal{L}(R \cdot \boldsymbol{W}^{\rm mm} + \boldsymbol{W}^{\rm cyc} + \boldsymbol{W}') + \left\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^\parallel - \boldsymbol{W}^{\rm cyc} \right\rangle. \tag{35}$$

Since $\boldsymbol{W}^\parallel \neq \boldsymbol{W}^{\rm cyc}$, then $\mathcal{L}(\boldsymbol{W}) - \mathcal{L}(R \cdot \boldsymbol{W}^{\rm mm} + \boldsymbol{W}^{\rm cyc} + \boldsymbol{W}') > 0$ will result in $\|\nabla\mathcal{L}(\boldsymbol{W})\|_F > 0$. Therefore, we next show that $\mathcal{L}(\boldsymbol{W}) > \mathcal{L}(R \cdot \boldsymbol{W}^{\rm mm} + \boldsymbol{W}^{\rm cyc} + \boldsymbol{W}')$ for $R \to \infty$.

Consider dataset $\texttt{DSET} = (\boldsymbol{X}_i, y_i)_{i=1}^n$. Let $\boldsymbol{a}_i = \boldsymbol{X}_i(R \cdot \boldsymbol{W}^{\rm mm})\bar{\boldsymbol{x}}_i$, $\boldsymbol{b}_i = \boldsymbol{X}_i\boldsymbol{W}^\parallel\bar{\boldsymbol{x}}_i$, $\boldsymbol{b}_i^\star = \boldsymbol{X}_i\boldsymbol{W}^{\rm cyc}\bar{\boldsymbol{x}}_i$, and $\boldsymbol{c}_i = \boldsymbol{X}_i\boldsymbol{W}'\bar{\boldsymbol{x}}_i$. What's more, let $\bar{c}_i = \boldsymbol{e}_{y_i}^\top \boldsymbol{W}'\bar{\boldsymbol{x}}_i$. Recall the $\mathcal{O}_i, \mathcal{R}_i, i \in [n]$ defined in (4). From the same analysis as in the proof of Lemma 2 and the facts that $\boldsymbol{W}^{\rm mm}$ is the solution of (Graph-SVM) and $\boldsymbol{W}' \perp \mathcal{S}_{\rm cyc}$, we have that

$$(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i})^\top (R \cdot \boldsymbol{W}^{\rm mm})\bar{\boldsymbol{x}}_i \begin{cases} = 0, & t \in \mathcal{R}_i \\ \leq -R, & t \in \bar{\mathcal{R}}_i \end{cases} \quad \text{and} \quad (\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i})^\top \boldsymbol{W}'\bar{\boldsymbol{x}}_i = \begin{cases} 0, & t \in \mathcal{R}_i \\ c_{it} - \bar{c}_i, & t \in \bar{\mathcal{R}}_i. \end{cases}$$

Recap the training risk from (18), and let $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i) = \mathbb{S}(\boldsymbol{a}_i + \boldsymbol{b}_i + \boldsymbol{c}_i)$ and $\boldsymbol{s}_i^\star = \mathbb{S}(\boldsymbol{X}_i(R \cdot \boldsymbol{W}^{\mathrm{mm}} + \boldsymbol{W}^{\mathrm{cyc}} + \boldsymbol{W}')\bar{\boldsymbol{x}}_i) = \mathbb{S}(\boldsymbol{a}_i + \boldsymbol{b}_i^\star + \boldsymbol{c}_i)$. Then for any $i \in [n]$, we get

$$\lim_{R\to\infty} \sum_{t\in\mathcal{O}_i} s_{it} = \lim_{R\to\infty} \frac{\sum_{t\in\mathcal{O}_i} e^{a_{it}+b_{it}+c_{it}}}{\sum_{t\in[T_i]} e^{a_{it}+b_{it}+c_{it}}} = \frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}}} =: d_i$$

$$\lim_{R\to\infty} \sum_{t\in\mathcal{O}_i} s_{it}^\star = \lim_{R\to\infty} \frac{\sum_{t\in\mathcal{O}_i} e^{a_{it}+b_{it}^\star+c_{it}}}{\sum_{t\in[T_i]} e^{a_{it}+b_{it}^\star+c_{it}}} = \frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}^\star}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}^\star}} =: d_i^\star.$$

Here $R \to \infty$ ensures that all tokens in $\bar{\mathcal{R}}_i$ are assigned with zero probability. Then following the similar analysis in the proof of Theorem 2 and (16), the ERM objective is related to the problem defined in Definition 1, where we have

$$\lim_{R\to\infty} \mathcal{L}(R \cdot \boldsymbol{W}^{\mathrm{mm}} + \boldsymbol{W}^{\|} + \boldsymbol{W}') = \frac{|\mathcal{I}|}{n} \bar{\mathcal{L}}(\boldsymbol{W}^{\|}).$$

Here employing log-loss, $\ell(1) = 0$. Since $\boldsymbol{W}^{\mathrm{cyc}}$ returns the optimal solution of $\bar{\mathcal{L}}(\boldsymbol{W})$ and $\mathcal{L}(\boldsymbol{W})$ is strictly convex on $\mathcal{S}_{\mathrm{cyc}}$, we have that

$$\lim_{R\to\infty} \mathcal{L}(R \cdot \boldsymbol{W}^{\mathrm{mm}} + \boldsymbol{W}^{\|} + \boldsymbol{W}') = \bar{\mathcal{L}}(\boldsymbol{W}^{\|}) > \bar{\mathcal{L}}(\boldsymbol{W}^{\mathrm{cyc}}) = \lim_{R\to\infty} \mathcal{L}(R \cdot \boldsymbol{W}^{\mathrm{mm}} + \boldsymbol{W}^{\mathrm{cyc}} + \boldsymbol{W}').$$

It combining with (35) results in that if $\lim_{\tau\to\infty} \boldsymbol{W}^{\|}(\tau) \neq \boldsymbol{W}^{\mathrm{cyc}}$, $\lim_{\tau\to\infty} \nabla\mathcal{L}(\boldsymbol{W}(\tau)) \neq 0$, which completes the proof by contradiction. ■

D Experimental Details

In this section, we provide implementation details of the experiments.

In all the experiments, we train single-layer self-attention layer models using PyTorch and SGD optimizer. We conduct normalized gradient descent method to enhance the increasing of the norm of attention weight, so that softmax can easily saturate. Specifically, at each iteration τ, we update attention weight $\boldsymbol{W}$ via

$$\boldsymbol{W}(\tau+1) = \boldsymbol{W}(\tau) - \eta \frac{\nabla\mathcal{L}(\boldsymbol{W}(\tau))}{\|\nabla\mathcal{L}(\boldsymbol{W}(\tau))\|_F}.$$

All the results are averaged over 100 random trails and in each trail, we create the dataset and its corresponding NTGs, SCCs as follows:

1. Given dimension d and vocabulary size K, generate random embedding table $\boldsymbol{E} = [\boldsymbol{e}_1 \ \cdots \ \boldsymbol{e}_K]^\top \in \mathbb{R}^{K\times d}$ such that each $\boldsymbol{e} \in \boldsymbol{E}$ is randomly sampled from unit sphere.
2. Given sample size n and sequence length T, create dataset $\texttt{DSET} = (\boldsymbol{X}_i, y_i)_{i=1}^n$ and $\boldsymbol{X}_i = [\boldsymbol{x}_{it} \ \cdots \ \boldsymbol{x}_{iT}]^\top \in \mathbb{R}^{T\times d}$ where $\boldsymbol{x}_{it}$ are randomly sampled from $\boldsymbol{E}$. For acyclic setting, label $\boldsymbol{e}_{y_i}$ is determined by the token in the $\boldsymbol{X}_i$ that has the highest priority order; while for general cyclic setting, same as tokens $\boldsymbol{x}_{it}$, $\boldsymbol{e}_{y_i}$ are also randomly sampled from $\boldsymbol{E}$.
3. Construct NTGs and SCCs of each NTG applying Tarjan's algorithm Tarjan (1972). For global convergence experiments (Section 5), NTGs are created based on the token relations between (x_{it}, y_i)s in the dataset DSET; while for local convergence analysis (Section 6), we instead establish the token relations $(x_{it}, \hat{y}_i)$ following the instruction in Section 6, where $\hat{y}_i$ is not unique and is determined by the GD solution.
4. $\overline{\texttt{DSET}}$ is created following Definition 1 based on the SCCs of the corresponding NTGs.

Here, we set the sequence length to be the same for all the samples in DSET, and we emphasize that though DSET contains inputs with same number of tokens, the randomness in sampling $\boldsymbol{x}_{it}$ and $\boldsymbol{e}_{y_i}$ will still result in a variety of NTGs and SCCs, and $\overline{\texttt{DSET}}$ may contain inputs with varying sequence lengths (see Figure 2).

• **Generating $\boldsymbol{W}^{\mathrm{cyc}}$ and $\widetilde{\boldsymbol{W}}^{\mathrm{cyc}}$.** Inspired by the convexity and finiteness of $\bar{\mathcal{L}}(\boldsymbol{W})$ per Definition 3 under the setting of Theorem 3, we can derive $\boldsymbol{W}^{\mathrm{cyc}}$ via gradient descent. Hence, to obtain $\boldsymbol{W}^{\mathrm{cyc}}$, we train separate models but with the same architecture from zero initialization on the sub-dataset $\overline{\texttt{DSET}}$. As for the experiments shown in Section 6, we follow the same method as generating $\boldsymbol{W}^{\mathrm{cyc}}$. However, we emphasize that under the local convergence setting, there is no guarantee that gradient descent will converge to the $\widetilde{\boldsymbol{W}}^{\mathrm{cyc}}$ solution as

problem is more general, i.e., with nonconvex head, and dataset $\overline{\texttt{DSET}}$ might not be enough to capture the performance of tokens within the same SCCs. Though, our results in Figures 5 and 6 indicate that $\widetilde{\boldsymbol{W}}^{\text{cyc}}$ can predict the GD convergence performance better than $\boldsymbol{W}^{\text{cyc}}$ which is drawn from the dataset-based NTGs. We defer a rigorous definition of local $\widetilde{\boldsymbol{W}}^{\text{cyc}}$ and guarantees related to gradient descent for future exploration.

• **Local convergence experiments (Figures 5 and 6).** To evaluate our local convergence conjecture, we conduct random experiments with more general head (satisfying Assumption 3) and, and consider squared loss $\ell(u) = (1-u)^2$ in Figure 5 and cross-entropy loss in Figure 6. In both experiment, we create embedding tabels with $K = 8, d = 8$ and datasets with $n = 4, T = 6$. We choose step size $\eta = 0.1$ and also conduct normalized gradient descent. Correlations are reported in Figs. 5a and 6a and the distance of $\|\boldsymbol{\Pi}_{\widetilde{\mathcal{S}}_{\text{cyc}}}(\boldsymbol{W}(\tau)) - \widetilde{\boldsymbol{W}}^{\text{cyc}}\|_F$ are presented in the orange curves in Figs. 5b and 6b. In both experiments, correlations between $\frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F}$ and $\frac{\widetilde{\boldsymbol{W}}^{\text{mm}}}{\|\widetilde{\boldsymbol{W}}^{\text{mm}}\|_F}$ end with > 0.99 values. Fig. 5b achieves 0 distance error since employing squared loss, attention is inclined to select tokens that appear mostly frequently in the labels of the dataset, resulting in $\mathcal{R}_i = \mathcal{O}_i$ for $i \in [n]$ and $\overline{\texttt{DSET}} = \emptyset$. While in Fig. 6b, the global and local norm of difference is around 9.59 and 0.09 respectively, where $\overline{\texttt{DSET}} \neq \emptyset$. This implies that the distance of $\boldsymbol{\Pi}_{\widetilde{\mathcal{S}}_{\text{cyc}}}(\boldsymbol{W}(\tau))$ is much closer to $\widetilde{\boldsymbol{W}}^{\text{cyc}}$ compared to the distance between $\boldsymbol{\Pi}_{\mathcal{S}_{\text{cyc}}}(\boldsymbol{W}(\tau))$ and $\boldsymbol{W}^{\text{cyc}}$.