Mechanics of Next Token Prediction with Self-Attention

Yingcong Li[*1] **Yixiao Huang**[*1] **M. Emrullah Ildiz**[1] **Ankit Singh Rawat**[2] **Samet Oymak**[1]
University of Michigan, Ann Arbor[1] Google Research NYC[2]

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 led to revolutionary advances in natural language processing. Underlying this success is the self-attention mechanism. In this work, we ask: *What does a single self-attention layer learn from next-token prediction?* We show that training self-attention with gradient descent learns an automaton which generates the next token in two distinct steps: **(1) Hard retrieval:** Given input sequence, self-attention precisely selects the *high-priority input tokens* associated with the last input token. **(2) Soft composition:** It then creates a convex combination of the high-priority tokens from which the next token can be sampled. Under suitable conditions, we rigorously characterize these mechanics through a directed graph over tokens extracted from the training data. We prove that gradient descent implicitly discovers the strongly-connected components (SCC) of this graph and self-attention learns to retrieve the tokens that belong to the highest-priority SCC available in the context window. Our theory relies on decomposing the model weights into a directional component and a finite component that correspond to hard retrieval and soft composition steps respectively. This also formalizes a related implicit bias formula conjectured in [Tarzanagh et al. 2023]. We hope that these findings shed light on how self-attention processes sequential data and pave the path toward demystifying more complex architectures.

[*]Equal contribution. Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

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) have not only led to breakthroughs 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 tasks 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

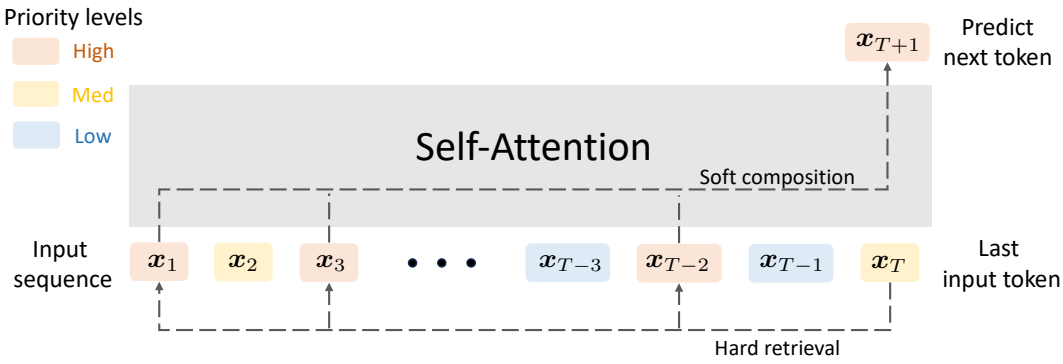


Figure 1: Overview of our result on next-token prediction. We study the implicit bias of gradient descent where a 1-layer self-attention model is trained until convergence. We prove that, during test-time, this model implements a hard retrieval to precisely select the high-priority tokens and then outputs a convex combination of these as the output from which the next token can be sampled. The notion of *high-priority* is formalized through the strongly-connected components of a directed graph associated to the last input token.

with Transformers to the celebrated support vector machines (SVMs) (Tarzanagh et al., 2023b,a).

As illustrated in Figure 1, given training data as a collection of (input sequence, next token) pairs, self-attention model learns to (1) retrieve the high-priority tokens (highlighted with red color) to the last input token; and then (2) build a convex combination of these high-priority tokens. The notion of *high-priority* is dictated by a directed graph learned from training data. SGD training accomplishes this by learning hard and soft components of the attention weights $\boldsymbol{W}$ to execute (1) and (2) respectively. Concretely, the following theorem dictates the evolution of attention weights during gradient descent.

Theorem 1 (*informal*) *Consider training a single-layer self-attention model with gradient descent. The combined attention weights $\boldsymbol{W} := \boldsymbol{W}_K \boldsymbol{W}_Q^\top$ evolve as*

$$\boldsymbol{W}_{GD} \approx C \cdot \boldsymbol{W}_{hard} + \boldsymbol{W}_{soft},$$

where $C \cdot \boldsymbol{W}_{hard}$ is the hard retrieval component selecting the high-priority tokens when $C \to \infty$; and $\boldsymbol{W}_{soft}$ is the soft composition component allocating nonzero softmax probabilities over selected tokens.

To capture the priority order among different tokens as observed in Figure 1, we construct directed graphs among the tokens in the vocabulary, namely *token-priority graphs* (TPGs). An illustration is provided in Figure 2, where a *strongly connected component* (SCC, highlighted as dashed black rectangles) in a TPG corresponds to the tokens that are reachable from each other, indicating the absence of a strict priority among those tokens. The hard retrieval component $\boldsymbol{W}_{\text{hard}}$ captures the topological order of different SCCs (orange arrows) whereas the soft composition component $\boldsymbol{W}_{\text{soft}}$ captures the relationships of different tokens within each SCC (black arrows). These TPGs will geometrically capture the learning dynamics of self-attention. Specifically, we propose the SVM problem (Graph-SVM), solution of which describes the direction gradient descent converges to. This way, SGD asymptotically enforces the topological order between SCCs i.e. the $C \cdot \boldsymbol{W}_{\text{hard}}$ term in Theorem 1 as $C \to \infty$. In practice, this implies that self-attention model favors suppressing lower priority tokens in favour of sampling higher priority tokens.

A conjecture on the decomposition in Theorem 1 was first proposed in (Tarzanagh et al., 2023a)[1]. This decomposition is also related to the implicit bias of logistic regression on non-separable data (Ji & Telgarsky, 2019a). Our theory fully formalizes this decomposition under the next-token prediction setting and reveals fundamental connections to graphical structure in data (e.g. through SCCs, TPGs).

Overall, we carefully study the gradient descent and regularization path algorithms for attention-based next-token prediction and make the following contributions:

1. We study the optimization landscape of self-attention with log-loss and show that the problem is convex under suitable assumptions. We then establish a global convergence result to fully formalize Theorem 1 in terms of a directional component (Graph-SVM) and a finite component (see Sec 3). Notably, results apply to arbitrary datasets as we don't require distributional assumptions.

2. Our theory reveals insightful connections between continuous and discrete optimization, namely: *Self-attention implicitly discovers the strongly-connected components of the TPGs during training.*

[1]Their conjecture aims to characterize the impact of the MLP layer that follow self-attention in a binary classification setting. However, the high-level claim is same.

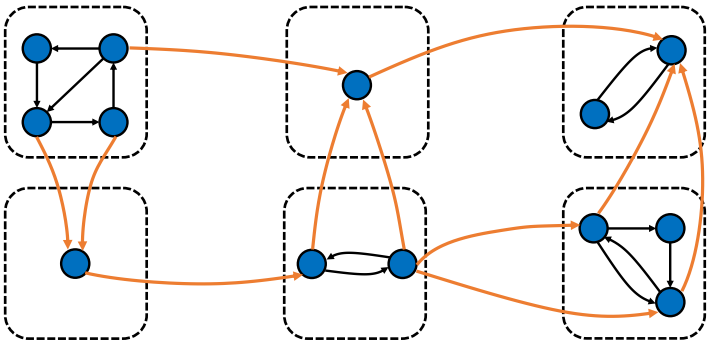


Figure 2: A token-priority graph (TPG) is a directed graph derived from training data (see Sec 2.1 for definition). The edges in TPG capture the input-output relationships between different tokens. A TPG can be partitioned into several SCCs depicted as dashed black squares. In light of Theorem 1, black intra-SCC edges within each SCC induce the soft-composition component of the attention weights whereas the orange edges induce the hard-retrieval component enforcing the priority orders among various SCCs.

3. Under a general setting, we establish the implicit bias of the solution obtained by vanishing regularization (Rosset et al., 2003; Suggala et al., 2018; Ji et al., 2020). This yields a result similar to the gradient descent theory (see Sec 4). We also show that, in general, gradient descent can exhibit local directional convergence rather than global. We characterize these local directions through the SVM solutions of *pseudo TPGs* (see Sec 5).

2 PROBLEM SETUP

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

Next-token prediction problem. Let K be the vocabulary size with $\boldsymbol{E} = \begin{bmatrix}\boldsymbol{e}_1 & \dots & \boldsymbol{e}_K\end{bmatrix}^\top \in \mathbb{R}^{K\times d}$ 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 $\boldsymbol{X} = [\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_T]^\top \in \mathbb{R}^{T\times d}$, where $\boldsymbol{x}_t \in \boldsymbol{E}$ for all $t \in [T]$.

Suppose that we have a training dataset $\texttt{DSET} = \{(\boldsymbol{X}_i, y_i) \in \mathbb{R}^{T_i\times d} \times [K]\}_{i=1}^n$ 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 $\boldsymbol{X}_i$, i.e., $\boldsymbol{x}_{it} = \boldsymbol{e}_{x_{it}}$.

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 linear combination of the input token embeddings in $\boldsymbol{X}$.

Empirical risk minimization (ERM) problem. Let $\ell : \mathbb{R} \to \mathbb{R}$ be the loss function. Given training dataset $\texttt{DSET}$, we consider the ERM problem with the following objective:

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

Throughout this paper, we fix the linear classification head $\boldsymbol{c}_k$[2] and assume $\|\boldsymbol{c}_k\|$ is bounded for all $k \in [K]$. Note that even though the classification head is linear, the problem of learning attention parameters $\boldsymbol{W}$ via ERM is not necessarily convex due to the softmax operator. In this work, we focus on this exact problem and consider the following two algorithms to optimze for $\boldsymbol{W}$:

[2]Specifically, we assume well-pretrained head $\boldsymbol{c}_1, \cdots, \boldsymbol{c}_K$ such that $\ell(\boldsymbol{c}_y^\top \boldsymbol{e}_k)$ returns the minimal risk when $k = y$.

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

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

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

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

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

2.1 Token-priority Graph of the Dataset

A token-priority graph (TPG) 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 TPGs $\{\mathcal{G}^{(k)}\}_{k=1}^K$, 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 $\texttt{DSET}$ 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 (x, y) 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. 3, where we construct two TPGs ($\mathcal{G}^{(1)}$ 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. 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, 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 be learnt to suppress x and 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 TPGs $\{\mathcal{G}^{(k)}\}_{k=1}^K$ associated with the dataset $\texttt{DSET}$, 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. 3, each of the dashed grey rectangle represents an SCC. $\mathcal{G}^{(1)}$(left) contains four SCCs and therefore, all tokens within the graph have strict priority orders. In contrast, $\mathcal{G}^{(2)}$(right) 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 TPGs 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 both two nodes (i, j) are in the same SCC of $\mathcal{G}$.

From the construction, for any two distinct nodes i, j in the same TPG, they either satisfy $(i \Rightarrow j)/(j \Rightarrow i)$ or $(i \asymp j)$.

2.2 SVM Bias of Self-attention Learning

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

$$\boldsymbol{W}^{\text{svm}} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \quad \text{(Graph-SVM)}$$

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

Fix last token $\boldsymbol{e}_k$, and consider any token IDs $i, j \in [K]$, $i \neq j$. When $(i \Rightarrow j)$, token ID i has a higher *priority*

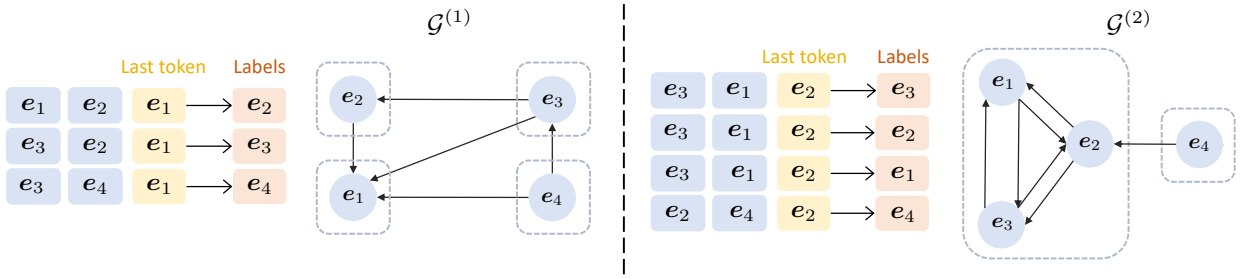


Figure 3: Illustration of token-priority graph (TPG). Given the input sequences and labels (next tokens), we construct the TPGs $\{\mathcal{G}^{(k)}\}_{k=1}^{K}$ according to the last token. Two TPGs $\mathcal{G}^{(1)}$ (left) and $\mathcal{G}^{(2)}$ (right) are constructed using the samples with $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ as the last tokens, respectively. In each graph, directed edges (label token→input token) are added between tokens/nodes. Based on these directed edges, each graph can be partitioned into its strongly-connected components (SCCs, highlighted as dashed grey rectangles). Each SCC is a set of tokens where each token is reachable from every other token within that SCC. Further details are deferred to Section 2.1.

than j and hence, the SVM problem (Graph-SVM) aims to find a $\boldsymbol{W}$ such that $\boldsymbol{W}\boldsymbol{e}_k$ achieves strictly higher correlation to token embedding $\boldsymbol{e}_i$ than $\boldsymbol{e}_j$, that is, $\boldsymbol{e}_i^\top \boldsymbol{W}\boldsymbol{e}_k \geq \boldsymbol{e}_j^\top \boldsymbol{W}\boldsymbol{e}_k + 1$, and then softmax operation will assign higher probability to the token i. While if $(i \asymp j)$, 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 $\boldsymbol{W}^{\text{svm}}$ 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 $\boldsymbol{W}^{\text{svm}} = 0$.

Lemma 1 *Suppose that the embedding matrix $\boldsymbol{E}$ is full row rank. Then,* (Graph-SVM) *is feasible.*

Next focusing on the nodes i, j, with $(i \asymp j)$, we introduce the following subspace definition.

Definition 1 (Cyclic subspace) *Define cyclic subspace $\mathcal{S}_{fin}$ as 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 $k \in [K]$.*

Note that since $\boldsymbol{W}^{\text{svm}}$ satisfies all the "$= 0$" constraints in (Graph-SVM), if (Graph-SVM) is feasible and $\boldsymbol{W}^{\text{svm}} \neq 0$, $\boldsymbol{W}^{\text{svm}} \perp \mathcal{S}_{\text{fin}}$.

2.3 Technical Assumptions

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

Assumption 1 *For $\forall y, k \in [K]$, $k \neq y$, $\boldsymbol{c}_y^\top \boldsymbol{e}_y = 1$ and $\boldsymbol{c}_y^\top \boldsymbol{e}_k = 0$.*

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

This assumption essentially enforces that the rows of the prediction head $\boldsymbol{C}$ are aligned with the corresponding vocabulary embeddings in $\boldsymbol{E}$. This is a variation of the *weight tying* strategy which is commonly employed in language models (Press & Wolf, 2017; Vaswani et al., 2017). It should be noted that Assumption 1 implies $K \leq d$, which further establishes the feasibility of (Graph-SVM) as demonstrated by Lemma 1. Given objective (ERM) and a decreasing loss function ℓ, our ideal goal is for the attention (1) to output $\boldsymbol{e}_y$ which minimizes the training risk. 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 *For any $(\boldsymbol{X}, y) \in \texttt{DSET}$, the token $\boldsymbol{e}_y$ is contained in the input sequence $\boldsymbol{X}$.*

In the scenario where $(\boldsymbol{X}, y)$ 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 Assumption 1 holds, 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. $\nabla_{\boldsymbol{W}} \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)) = 0$.

3 GLOBAL CONVERGENCE OF GRADIENT DESCENT

In this section, we assume the log-loss function, i.e., $\ell(u) = -\log(u)$, and establish the gradient descent 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 2 *Suppose Assumptions 1 and 2 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 $\mathcal{S}_{fin}$.*

Let $(\boldsymbol{X}, y) \in \texttt{DSET}$ be any sample and set $\gamma_t = \boldsymbol{c}_y^\top \boldsymbol{x}_t$. Assumption 1 guarantees that $\gamma_t = 1$ when $x_t = y$, otherwise $\gamma_t = 0$. Consider the attention output $\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}})$ (cf. (ERM)) 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 $\boldsymbol{X}$ is $\ell(\bar{\gamma})$ where $\bar{\gamma} = \sum_t \gamma_t s_t = \sum_{x_t = y} s_t$. Given log-loss, 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 $s_t \not\to 0$ for all t satisfying $x_t = y$. Given different label y for different input and recalling the SCC definition in Section 2.1, attention selects all $\boldsymbol{x}_t$'s within the same SCC as y. The following result characterizes the global directional convergence of the GD iterates to the solution of (Graph-SVM).

Theorem 2 *Consider a dataset $\texttt{DSET}$ and suppose Assumptions 1 and 2 hold. Set loss function as $\ell(u) = -\log(u)$. Let $\boldsymbol{W}^{svm} \in \mathcal{S}_{fin}^{\perp}$ be the solution of (Graph-SVM). Starting from any $\boldsymbol{W}(0)$ with constant step size η, the algorithm Algo-GD satisfies $\lim_{\tau\to\infty} \|\boldsymbol{W}(\tau)\|_F = \infty$ and*

$$\lim_{\tau\to\infty} \boldsymbol{\Pi}_{\mathcal{S}_{fin}}(\boldsymbol{W}(\tau)) = \boldsymbol{W}^{fin}. \tag{2}$$

Here $\boldsymbol{W}^{fin}$ is the unique finite minima of the loss $\tilde{\mathcal{L}}(\boldsymbol{W}) := \lim_{R\to\infty} \mathcal{L}(\boldsymbol{W} + R \cdot \boldsymbol{W}^{svm})$ over $\mathcal{S}_{fin}$. Additionally, if $\boldsymbol{W}^{svm} \neq 0$,

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

Otherwise, $\boldsymbol{\Pi}_{\mathcal{S}_{fin}^{\perp}}(\boldsymbol{W}(\tau))$ remains unchanged throughout the optimization.

This theorem demonstrates the directional convergence of attention weight $\boldsymbol{W}$, and the limits imply the decomposition $\boldsymbol{W}(\tau) \approx C(\tau)\cdot\boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^{\text{fin}}$ for an appropriate $C(\tau) > 0$ with $\lim_{\tau\to\infty} C(\tau) = \infty$. Note that, for directional convergence to happen, we need $\boldsymbol{W}^{\text{svm}} \neq 0$ which happens if and only if (Graph-SVM) has "≥ 1" constraints. That is, the token graph contains a strict priority order. This is consistent with our Theorem 1 where $\boldsymbol{W}^{\text{svm}}$ corresponds to the hard retrieval component ($\boldsymbol{W}_{\text{hard}}$) that selects the high-priority tokens and $\boldsymbol{W}^{\text{fin}}$ is the soft composition component ($\boldsymbol{W}_{\text{soft}}$) that determines the softmax probability assignments among these selected tokens. Importantly, this is a global convergence result thanks to the convexity of the optimization problem, which is enabled by the log likelihood optimization combined with Assumption 1. In Section 5, we will demonstrate that global convergence of gradient descent does not hold in general.

To illustrate Theorem 2, we conduct experiments with results presented in Figure 4. We create embedding tables with $K = 6$, $d = 8$ and randomly generate dataset with $n = 6$, $T = 4$. Here we choose step size $\eta = 0.01$ and perform the normalized gradient descent method to accelerate the increase in 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 $\langle \boldsymbol{W}(\tau), \boldsymbol{W}^{\text{svm}} \rangle / (\|\boldsymbol{W}(\tau)\|_F \|\boldsymbol{W}^{\text{svm}}\|_F)$, and results averaged over 100 random instances are displayed in Fig. 4a, which end in correlation ≈ 0.987 after training with 4000 iterations. In addition to the directional convergence of $\boldsymbol{W}(\tau)$, we also verify the convergence of finite component by tracking the matrix distance $\left\|\widehat{\boldsymbol{W}}^{\text{fin}} - \boldsymbol{W}^{\text{fin}}\right\|_F$ where $\widehat{\boldsymbol{W}}^{\text{fin}} = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau))$, and results are displayed in Fig. 4b, which obtains < 0.01 final distance. Both results validate our Theorem 2.

4 IMPLICIT BIAS OF SELF-ATTENTION

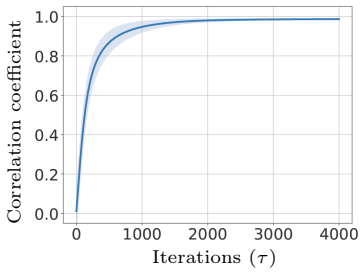
In Section 3, we have discussed that gradient descent with log loss guarantees the global convergence. To proceed, in this section, we discuss the implicit bias of attention via analysis of regularization path (RP) as employed in Algo-RP and identify the implicit bias of self-attention on more general next-token prediction problems.

Theorem 3 *Consider any dataset $\texttt{DSET}$ and suppose Assumptions 1 and 2 hold. Additionally, assume loss $\ell : \mathbb{R} \to \mathbb{R}$ is strictly decreasing and $|\ell'|$ is bounded. Let $\boldsymbol{W}^{svm} \in \mathcal{S}_{fin}^{\perp}$ be the solution of (Graph-SVM) and suppose $\boldsymbol{W}^{svm} \neq 0$. Then the solution of regularization path Algo-RP obeys*

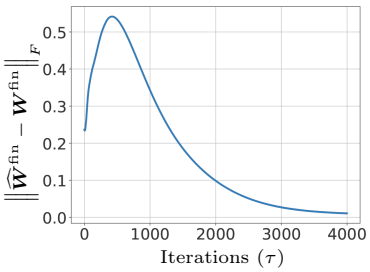
$$\lim_{R\to\infty} \frac{\bar{\boldsymbol{W}}_R}{R} = \frac{\boldsymbol{W}^{svm}}{\|\boldsymbol{W}^{svm}\|_F} \quad \text{and} \quad \lim_{R\to\infty} \boldsymbol{\Pi}_{\mathcal{S}_{fin}}(\bar{\boldsymbol{W}}_R) \in \mathcal{W}^{fin}.$$

Here $\mathcal{W}^{fin} = \arg\min_{\boldsymbol{W}\in\mathcal{S}_{fin}} \lim_{R\to\infty} \mathcal{L}(\boldsymbol{W} + R\cdot\boldsymbol{W}^{svm})$ and we assume that $\mathcal{W}^{fin}$ is a bounded set.

Here, we allow more general loss function ℓ which is different from the log-loss employed in Section 3, and the ERM problem (cf. (ERM)) is not guaranteed to be convex.



(a) Evolution of $\frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F} \to \frac{\boldsymbol{W}^{\mathrm{svm}}}{\|\boldsymbol{W}^{\mathrm{svm}}\|_F}$



(b) Evolution of $\boldsymbol{\Pi}_{\mathcal{S}_{\mathrm{fin}}}(\boldsymbol{W}(\tau)) \to \boldsymbol{W}^{\mathrm{fin}}$

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

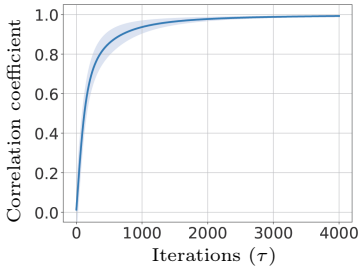


Figure 5: GD convergence of attention weight $\boldsymbol{W}$ when training with acyclic dataset (Def. 2). Correlation coefficient between $\boldsymbol{W}(\tau)$ and $\boldsymbol{W}^{\mathrm{svm}}$ are presented.

4.1 Acyclic Dataset

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 TPGs are directed acyclic graphs.*

For an acyclic dataset, there are not $i, j \in [K]$ satisfying $(i \asymp j) \in \mathcal{G}^{(k)}$, for all $k \in [K]$. Thus, SVM formulation (Graph-SVM) reduces to the following simpler form:

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

We next make the following assumption on the linear head. Notably, Assumption 1 is a special case of Assumption 3.

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

Lemma 3 *Consider acyclic dataset* `DSET` *per Def. 2 and suppose Assumptions 2 and 3 hold. Additionally, assume loss function ℓ is strictly decreasing and $|\ell'|$ is bounded. 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, then for any $\boldsymbol{W}$ that satisfies the constraints in (Acyc-SVM), $\lim_{R\to\infty} \mathcal{L}(R\cdot\boldsymbol{W}) = \mathcal{L}_\star$.*

Assumption 3 and Lemma 3 ensure that the best 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 towards the solution of (Acyc-SVM).

Theorem 4 *Suppose* `DSET` *is acyclic per Definition 2 and Assumptions 2 and 3 hold. Additionally, assume loss $\ell : \mathbb{R} \to \mathbb{R}$ is strictly decreasing and $|\ell'|$ is bounded. Suppose (Acyc-SVM) is feasible with $\boldsymbol{W}^{svm}$ denoting its solution. Then, Algorithm Algo-RP satisfies*

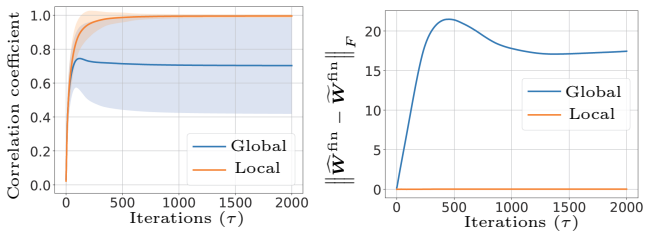
$$\lim_{R\to\infty} \frac{\bar{\boldsymbol{W}}_R}{R} = \frac{\boldsymbol{W}^{svm}}{\|\boldsymbol{W}^{svm}\|_F}.$$

Recall that a dataset being acyclic implies that there is strict priority order among all tokens in each TPG. Theorem 4 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 $\boldsymbol{W}^{\mathrm{svm}}$.

Following the same implementation setting as in Section 3, in Fig. 5, we again conduct 100 trials but with randomly generated acyclic dataset `DSET` under the setting of $K = d = 8, n = 4$ and $T = 6$. The results averaged over 100 random instances are presented in Fig. 5 with correlation coefficient exceeds 0.99 after training with 4000 iterations. Note that in these experiments, log-loss is employed as loss function and Assumption 1 is satisfied which guarantee the convexity of $\mathcal{L}(\boldsymbol{W})$ following Lemma 2 and hence, the connection between Algo-GD and Algo-RP is built by Ji et al. (2020).

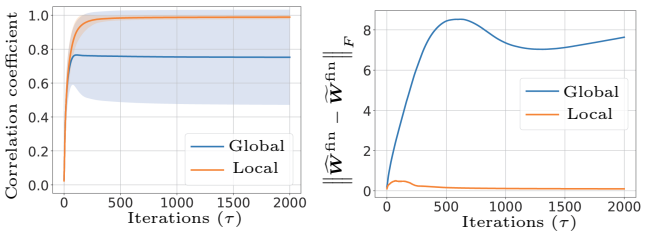
5 FURTHER INVESTIGATION ON LOCAL CONVERGENCE

So far, we have proved the global GD convergence of attention weight when one employs the log-loss (Section 3) and studied the implicit bias of self-attention over next-token prediction problem using RP analysis (Section 4). In this section, we investigate further on the convergence performance of GD and ask:



(a) $\frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F} \to \frac{\widetilde{\boldsymbol{W}}^{\text{svm}}}{\|\widetilde{\boldsymbol{W}}^{\text{svm}}\|_F}$ (b) $\boldsymbol{\Pi}_{\widetilde{\mathcal{S}}_{\text{fin}}}(\boldsymbol{W}(\tau)) \to \widetilde{\boldsymbol{W}}^{\text{fin}}$

Figure 6: Squared loss with general classifier



(a) $\frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F} \to \frac{\widetilde{\boldsymbol{W}}^{\text{svm}}}{\|\widetilde{\boldsymbol{W}}^{\text{svm}}\|_F}$ (b) $\boldsymbol{\Pi}_{\widetilde{\mathcal{S}}_{\text{fin}}}(\boldsymbol{W}(\tau)) \to \widetilde{\boldsymbol{W}}^{\text{fin}}$

Figure 7: Cross-entropy loss with general classifier

When does the GD exhibit local convergence rather than global? Can we characterize 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* TPGs for obtaining *locally-optimal* SVM equivalence $\widetilde{\boldsymbol{W}}^{\text{svm}}$ and cyclic component $\widetilde{\boldsymbol{W}}^{\text{fin}}$ as follows:

1. Given any dataset DSET, consider GD solution $\boldsymbol{W}^{\text{GD}}$. For each training example $(\boldsymbol{X}, y) \in$ DSET, let $\boldsymbol{s} = \mathbb{S}(\boldsymbol{X}\boldsymbol{W}^{\text{GD}}\bar{\boldsymbol{x}})$.
2. Construct TPGs 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 TPGs defined in Section 2.1 which is uniquely determined by the dataset and the ground truth labels, pseudo-TPGs build edges based on the tokens selected by GD solution $\boldsymbol{W}^{\text{GD}}$. 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 3, 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 (Figures 6 and 7).

Extended linear head. In this work, GD experiments are conducted under Assumptions 1 and 2, which implies the convex training loss $\mathcal{L}(\boldsymbol{W})$ and global convergence performance. Now consider a more general linear head (i.e., Assumption 3). As have also been observed and discussed in Tarzanagh et al. (2023b,a), Algo-GD can converge to a locally-optimal solution.

Figures 6 and 7 display our local convergence results where Fig. 6 employs squared loss, i.e. $\ell(u) = (1-u)^2$ and Fig. 7 utilizes cross-entropy loss. Both apply general head following Assumption 3. Similar to Fig. 4, we present (directional) convergence performance of $\boldsymbol{W}$ towards $\boldsymbol{W}^{\text{svm}}$ and $\boldsymbol{W}^{\text{fin}}$. Results indicate that instead of converging to the global solution (blue curves), attention weights trained via GD align more closely with the locally-optimal SVM solution defined via the pseudo TPGs constructed by $\boldsymbol{W}^{\text{GD}}$ (orange curves). In Fig. 6b, the norm difference to $\widetilde{\boldsymbol{W}}^{\text{fin}}$ remains zero, indicating that all SCCs in the pseudo TPGs are singleton and GD optimizes attention weights towards selecting one token per sequence. While in Fig. 7, multiple tokens can be selected by $\boldsymbol{W}^{\text{GD}}$. Note that in Fig. 7b, the norm of difference does not end with zero value on average. The potential explanations can be: Due to the non-convexity of training loss, training $\boldsymbol{W}$ with GD may not fully capture its RP solution $\widetilde{\boldsymbol{W}}^{\text{fin}}$ over the cyclic subspace, 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 $\widetilde{\boldsymbol{W}}^{\text{fin}}$ within its cyclic subspace. We defer a rigorous definition of local $\widetilde{\boldsymbol{W}}^{\text{fin}}$ and guarantees related to gradient descent for future exploration. Experimental details are deferred to the appendix.

6 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); Bombari & Mondelli (2024) 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 gen-

eralization 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., 2023c; Huang et al., 2023; Li et al., 2023b; Collins et al., 2024; Jeon et al., 2024; Chen et al., 2024; Li et al., 2024).

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. The recent work Thrampoulidis (2024) also explores the next-token prediction problem under a classification-like setting, employing a related SVM formulation. Since we study transformers, the main messages are fairly different, e.g., our theory relies on graph-theoretic concepts such as SCCs and token-priority graphs to capture the generative process learned by SGD. Notably, several recent efforts (Jelassi et al., 2022; Li et al., 2023a,d; Oymak et al., 2023; Deora et al., 2023; Chen & Li, 2024) 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. Related work by Tian et al. (2023) studies the training dynamics of next-token prediction. Compared to us, their analysis is restricted to a specific statistical data model, including the requirement of long input sequences ($T \to \infty$). Ildiz et al. (2024); Makkuva et al. (2024) build connections between self-attention and Markov chains. 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. (2020); 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.

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 token-priority graphs 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 softmax probabilities among all equal priority tokens as modeled by the strongly-connected components of the next-token graph. Ultimately, our results comprehensively capture the automaton implemented by a 1-layer self-attention under realistic assumptions.

A natural future direction is relaxing our assumptions in SGD analysis and providing a comprehensive characterization of the training dynamics, accounting for non-convexities. It would also 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 dynamics and aid in the aforementioned token selection and composition mechanisms during next-token prediction.

Acknowledgements

This work was supported in part by the NSF grants CCF-2046816, CCF-2212426, CNS-1932254, UMich's MIDAS PODS program, a Google Research Scholar award, and an Adobe Data Science Research award.

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 neu-

ral 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.

Simone Bombari and Marco Mondelli. Towards understanding the word sensitivity of attention layers: A study via random features. *arXiv preprint arXiv:2402.02969*, 2024.

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.

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

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

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.

Liam Collins, Advait Parulekar, Aryan Mokhtari, Sujay Sanghavi, and Sanjay Shakkottai. In-context learning with transformers: Softmax attention adapts to function lipschitzness. *arXiv preprint arXiv:2402.11639*, 2024.

Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*, 2023.

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.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.

M Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: Unveiling the dynamics of generative transformers. *arXiv preprint arXiv:2402.13512*, 2024.

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.

Hong Jun Jeon, Jason D Lee, Qi Lei, and Benjamin Van Roy. An information-theoretic analysis of in-context learning. *arXiv preprint arXiv:2401.15530*, 2024.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on learning theory*, pp. 1772–1798. PMLR, 2019a.

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. 2019b.

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 *Conference on Learning Theory*, pp. 2109–2136. PMLR, 2020.

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.

Hongkang Li, Meng Wang, Songtao Lu, Hui Wan, Xiaodong Cui, and Pin-Yu Chen. Transformers as multi-task feature selectors: Generalization analysis of in-context learning. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023b.

Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. Training nonlinear transformers for efficient in-context learning: A theoretical learning and generalization analysis. *arXiv preprint arXiv:2402.15607*, 2024.

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 2023c. 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.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*, pp. 19689–19729. PMLR, 2023d.

Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with markov: A framework for principled analysis of transformers via markov chains. *arXiv preprint arXiv:2402.04161*, 2024.

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 preprint-arXiv: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.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 2017.

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.

Christos Thrampoulidis. Implicit bias of next-token prediction. *arXiv preprint arXiv:2402.18551*, 2024.

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 Confer-*

ence 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.

Mechanics of Next Token Prediction with Transformers Supplementary Materials

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

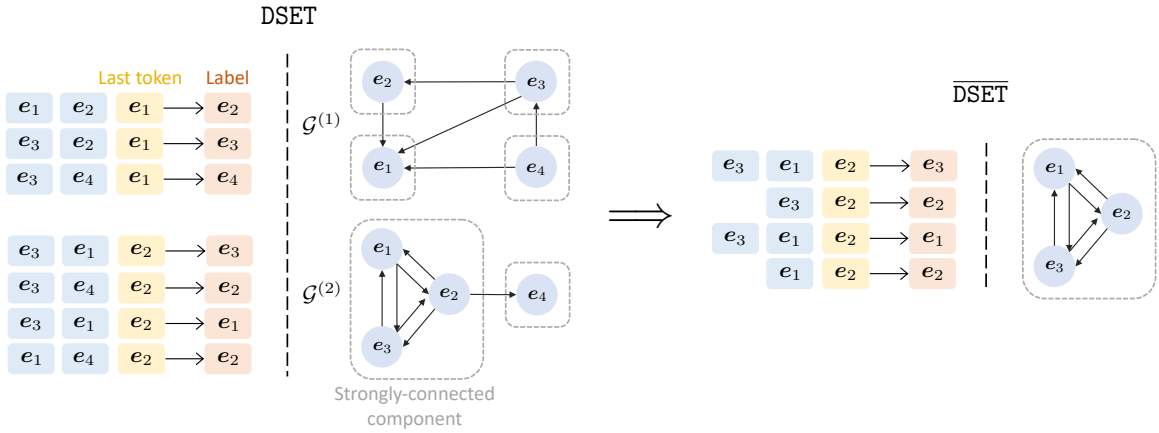


Figure 8: Illustration of token-priority graph (TPG). Given the input sequences and labels (next tokens), we construct the TPGs $\{\mathcal{G}^{(k)}\}_{k=1}^{K}$ according to the last token. **Left hand side:** Two TPGs $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ are constructed using the samples with $\boldsymbol{e}_1$ 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, highlighted as dashed grey rectangles). **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 3). Then, $\overline{\texttt{DSET}}$ contains non-singleton SCCs only as shown on the right.

Contents

A AUXILIARY RESULTS

A.1 Soft-composition Component

In Section 3, we have theoretically shown that when training a single-layer self-attention model with gradient descent and log-loss function, once $\boldsymbol{W}^{\text{svm}} \neq 0$, the composed attention weight $\boldsymbol{W}(\tau)$ will diverge in Frobenius norm and $\boldsymbol{W}(\tau)$ converges towards direction $\boldsymbol{W}^{\text{svm}}/\left\|\boldsymbol{W}^{\text{svm}}\right\|_F$; while in the subspace of $\mathcal{S}_{\text{fin}}$, $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau))$ converges to a finite $\boldsymbol{W}^{\text{fin}}$ which is the unique solution minimizing the training loss over subspace $\mathcal{S}_{\text{fin}}$ as described in Theorem 2. Here, $\boldsymbol{W}^{\text{svm}}$ follows the solution of (Graph-SVM) and plays a role in separating tokens from different SCCs within the same TPG. Specifically, nodes satisfy $(i \Rightarrow j) \in \mathcal{G}^{(k)}$.

As for the nodes contained within the same SCC (e.g., $(i \asymp j)$), 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, $\boldsymbol{W}^{\text{svm}}$ does not truly capture the essence of the ERM solution. In the following, we introduce cyclic subdataset and the so-called *cyclic-component*, and an equivalence between the cyclic term and $\boldsymbol{W}^{\text{fin}}$ can be established under mild assumptions.

Definition 3 (Cyclic subdataset) *Given any training sample $(\boldsymbol{X}, y) \in \texttt{DSET}$, we obtain the corresponding sample $(\boldsymbol{X}', y) \in \overline{\texttt{DSET}}$ by removing all tokens in $\boldsymbol{X}$ that satisfy $(y \Rightarrow x)$ in the corresponding TPG.*

In short, cyclic subdataset focuses on the input tokens that are part of the same SCC as the label token. Fig. 8(Right) presents the cyclic subdataset $\overline{\texttt{DSET}}$ of DSET given in Fig. 8(Left), which is the same as Fig. 3. In $\mathcal{G}^{(1)}$, all nodes are separated into different SCCs, and therefore, none of them is present in $\overline{\texttt{DSET}}$; 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 $\overline{\texttt{DSET}}$ while $\boldsymbol{e}_4$ is removed from the dataset. Note that $\overline{\texttt{DSET}}$ provides a self-contained sub-problem that solely focuses on intra-SCC edges.

Definition 4 (Cyclic component) *$\mathcal{W}^{fin}$ is obtained as the solution set of the ERM problem over the cyclic subdataset $\overline{\texttt{DSET}}$ per Definition 3. Concretely,*

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

Lemma 4 *Consider a dataset $\texttt{DSET}$ and let $\boldsymbol{W}^{svm}$ be the corresponding SVM solution of (Graph-SVM) with $\boldsymbol{W}^{svm} \neq 0$. Then we have $\boldsymbol{W}^{svm} \perp \mathcal{S}_{fin}$, and for any $\bar{\boldsymbol{W}}^{fin} \neq 0 \in \mathcal{W}^{fin}$, $\bar{\boldsymbol{W}}^{fin}$ and $\boldsymbol{W}^{svm}$ are orthogonal.*

Lemma 5 *Let $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ be an arbitrary matrix, then we have $\bar{\mathcal{L}}(\boldsymbol{W}) = \bar{\mathcal{L}}(\boldsymbol{\Pi}_{\mathcal{S}_{fin}}(\boldsymbol{W}))$.*

Lemma 6 *Suppose Assumptions 1 and 2 hold, and loss function $\ell(u) = -\log(u)$, then for any finite $\boldsymbol{W}$, $\tilde{\mathcal{L}}(\boldsymbol{W}) = \bar{\mathcal{L}}(\boldsymbol{W})$ where $\tilde{\mathcal{L}}(\boldsymbol{W})$ and $\bar{\mathcal{L}}(\boldsymbol{W})$ are defined in Theorem 2 and Definition 4, respectively.*

A.2 Useful Notations

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

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

$$\mathcal{O}_i := \left\{ t \;\middle|\; x_{it} = y_i, \forall t \in [T_i] \right\} \quad \text{and} \quad \bar{\mathcal{O}}_i = [T_i] - \mathcal{O}_i, \tag{3a}$$
$$\mathcal{R}_i := \mathcal{O}_i \bigcup \left\{ t \;\middle|\; (x_{it} \asymp y_i) \in \mathcal{G}^{(\bar{x}_i)}, \forall t \in [T_i] \right\} \quad \text{and} \quad \bar{\mathcal{R}}_i = [T_i] - \mathcal{R}_i \tag{3b}$$

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 $\mathcal{G}^{(\bar{x}_i)}$ is the corresponding token-priority graph (TPG) associated with the last/query token of $\boldsymbol{X}_i$. Concretely, $\mathcal{O}_i$ returns the token indices of i-th input that have the same token ID as label y_i, while $\mathcal{R}_i$ returns the token indices of i-th input that are included in the same strongly-connected component (SCC) as label y_i in the corresponding TPG. Then for any

$t \in \bar{\mathcal{R}}_i$, we have $(y_i \Rightarrow x_{it}) \in \mathcal{G}^{(\bar{x}_i)}$. Take the last input sequence in Figure 8(left) as an example, where $\mathcal{O} = \{3\}$ and $\mathcal{R} = \{1, 3\}$.

• **Datasets $\texttt{DSET}$, $\overline{\texttt{DSET}}$ and sample index set $\mathcal{I}$, $\bar{\mathcal{I}}$.** Recap the training dataset $\texttt{DSET} = (\boldsymbol{X}_i, y_i)_{i=1}^n$. Based on the relationships between input tokens and label token, following instructions in Section 2.1 we can construct the TPGs 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 TPG. Or equivalently,

$$\mathcal{I} = \left\{ i \;\middle|\; \mathcal{R}_i - \mathcal{O}_i \neq \emptyset, i \in [n] \right\} \quad \text{and} \quad \bar{\mathcal{I}} = [n] - \mathcal{I}. \tag{4}$$

Then the cyclic subset defined in Definition 3 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 $\boldsymbol{x}_{it}$, $t \in \bar{\mathcal{R}}_i$. Hence, 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., $(y_i \Rightarrow x_{it}) \in \mathcal{G}^{(\bar{x}_i)}$ for $t \in \bar{\mathcal{O}}_i$. In Figure 8(left), we have $\mathcal{I} = \{4, 5, 6, 7\}$ and $\bar{\mathcal{I}} = \{1, 2, 3\}$.

• **Token scores $\boldsymbol{\gamma}_i, i \in [n]$ and loss $\mathcal{L}(\boldsymbol{W})$ under Assumption 1.** Let $\boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$ be the token score vectors. Then under Assumption 1, we have

$$\gamma_{it} = \begin{cases} 1, & t \in \mathcal{O}_i \\ 0, & t \in \bar{\mathcal{O}}_i \end{cases} \qquad \text{for all } i \in [n]. \tag{6}$$

Additionally, letting $s_i^{\boldsymbol{W}} = \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)$, we can rewrite the training risk as follows:

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell \left(\sum_{t \in \mathcal{O}_i} s_{it}^{\boldsymbol{W}} \right). \tag{7}$$

A.3 Proof of Lemma 1

Proof. Recap the constraints in (Graph-SVM) problem:

$$(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W} \boldsymbol{e}_k \begin{cases} = 0 & \forall (i \asymp j) \in \mathcal{G}^{(k)} \\ \geq 1 & \forall (i \Rightarrow j) \in \mathcal{G}^{(k)} \end{cases} \quad \text{for all} \quad k \in [K]. \tag{8}$$

Since $\boldsymbol{E} = [\boldsymbol{e}_1 \; \boldsymbol{e}_2 \; \cdots \; \boldsymbol{e}_K]^\top \in \mathbb{R}^{K \times d}$ is full row rank, then $K \leq d$ and $\boldsymbol{e}_k, k \in [K]$ are linearly independent. Let $\bar{\boldsymbol{E}} \in \mathbb{R}^{K \times d}$ satisfying $\bar{\boldsymbol{E}} \boldsymbol{E}^\top = \boldsymbol{I}$. Then for any $\bar{\boldsymbol{W}} \in \mathbb{R}^{K \times K}$, we get

$$(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \bar{\boldsymbol{E}}^\top \bar{\boldsymbol{W}} \bar{\boldsymbol{E}} \boldsymbol{e}_k \begin{cases} = 0 & \forall (i \asymp j) \in \mathcal{G}^{(k)} \\ \geq 1 & \forall (i \Rightarrow j) \in \mathcal{G}^{(k)} \end{cases} \quad \text{for all} \quad k \in [K] \tag{9}$$

and feasibility of (9) implies $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ in (8) is feasible. Since we can set $\boldsymbol{W} = \bar{\boldsymbol{E}}^\top \bar{\boldsymbol{W}} \bar{\boldsymbol{E}}$. Next let $\boldsymbol{u}_i = \bar{\boldsymbol{E}} \boldsymbol{e}_i$ $i \in [K]$ be K-dimensional one-hot vectors. Then it remains to show that there exists $\bar{\boldsymbol{W}} \in \mathbb{R}^{K \times K}$ such that

$$(\boldsymbol{u}_i - \boldsymbol{u}_j)^\top \bar{\boldsymbol{W}} \boldsymbol{u}_k \begin{cases} = 0 & \forall (i \asymp j) \in \mathcal{G}^{(k)} \\ \geq 1 & \forall (i \Rightarrow j) \in \mathcal{G}^{(k)} \end{cases} \quad \text{for all} \quad k \in [K] \tag{10}$$

is feasible. Additionally, it is equivalent with showing that for any $k \in [K]$, there exists $\boldsymbol{w} \in \mathbb{R}^K$, such that

$$(\boldsymbol{u}_i - \boldsymbol{u}_j)^\top \boldsymbol{w} \begin{cases} = 0 & \forall (i \asymp j) \in \mathcal{G}^{(k)} \\ \geq 1 & \forall (i \Rightarrow j) \in \mathcal{G}^{(k)}. \end{cases} \tag{11}$$

To start with, we first derive the priority order of each graph (referring to the topological sorting of directed graph). Let M_i be the order of $\boldsymbol{u}_i$ where M_i's $i \in [K]$ are positive integers. Then if $(i \asymp j)$, $M_i = M_j$; if $(i \Rightarrow j)$, $M_i > M_j$. Then let $\boldsymbol{w} = \sum_{i\in[K]} M_i \boldsymbol{u}_i$. We obtain that for any $k \in [K]$,

$$\forall (i \asymp j) \in \mathcal{G}^{(k)},\ (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top \boldsymbol{w} = (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top (M_i \boldsymbol{u}_i + M_j \boldsymbol{u}_j) = M_i - M_j = 0$$
$$\forall (i \Rightarrow j) \in \mathcal{G}^{(k)},\ (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top \boldsymbol{w} = (\boldsymbol{u}_i - \boldsymbol{u}_j)^\top (M_i \boldsymbol{u}_i + M_j \boldsymbol{u}_j) = M_i - M_j \geq 1$$

which indicates that (11) is feasible for any $k \in [K]$ and it completes the proof. ■

A.4 Proof of Lemma 4

Proof. Recall from Definition 1 that $\mathcal{S}_{\text{fin}}$ is 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 $k \in [K]$. Then for any matrix $\boldsymbol{W} \in \mathcal{S}_{\text{fin}}$, there exist a_{ijk}'s satisfying

$$\boldsymbol{W} = \sum_{i,j,k} a_{ijk} (\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top$$

where $(i \asymp j) \in \mathcal{G}^{(k)}$ and $k \in [K]$. Since $\boldsymbol{W}^{\text{svm}}$ is the solution of (Graph-SVM) that satisfies the all "$= 0$" constraints, for any matrix $\boldsymbol{W} \in \mathcal{S}_{\text{fin}}$, we have

$$\langle \boldsymbol{W}^{\text{svm}}, \boldsymbol{W} \rangle = \sum_{i,j,k} a_{ijk} \left\langle \boldsymbol{W}^{\text{svm}}, (\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top \right\rangle = 0.$$

Therefore, $\boldsymbol{W}^{\text{svm}} \perp \mathcal{S}_{\text{fin}}$. ■

A.5 Proof of Lemma 5

Proof. Recap the definition of $\bar{\mathcal{L}}(\boldsymbol{W})$ from Def. 4 and $\mathcal{I}$, $\bar{\boldsymbol{X}}_i$ from (4), (5). Then

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

Let $\boldsymbol{W}^\perp = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^\perp}(\boldsymbol{W})$ and $\boldsymbol{W}^\parallel = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W})$. Then it remains to show that for any $(\bar{\boldsymbol{X}}, y) \in \overline{\texttt{DSET}}$, $\mathbb{S}(\bar{\boldsymbol{X}}\boldsymbol{W}\bar{\boldsymbol{x}}) = \mathbb{S}(\bar{\boldsymbol{X}}\boldsymbol{W}^\parallel\bar{\boldsymbol{x}})$.

For simplification, let $\bar{\boldsymbol{x}} = \boldsymbol{e}_k$, and following the definition of TPG, SCC and $\overline{\texttt{DSET}}$, we have that all tokens $\boldsymbol{x} \in \bar{\boldsymbol{X}}$ are in the same SCC and denote the token set as $\mathcal{C}^{(k)}$. Then $\mathcal{S}_{\text{fin}}$ spans the matrices $(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top$ for $i, j \in \mathcal{C}^{(k)}$. For any $i \in \mathcal{C}^{(k)}$, we get

$$\boldsymbol{e}_i^\top \boldsymbol{W} \boldsymbol{e}_k = \boldsymbol{e}_i^\top \boldsymbol{W}^\parallel \boldsymbol{e}_k + \boldsymbol{e}_i^\top \boldsymbol{W}^\perp \boldsymbol{e}_k.$$

Next, let $a_{ik} = \boldsymbol{e}_i^\top \boldsymbol{W}^\perp \boldsymbol{e}_k$, $i \in \mathcal{C}^{(k)}$. Since $\boldsymbol{W}^\perp \perp \mathcal{S}_{\text{fin}}$, and $(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top \in \mathcal{S}_{\text{fin}}$, we obtain

$$\begin{aligned} & (\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W}^\perp \boldsymbol{e}_k = 0 \\ \Longrightarrow\ & \boldsymbol{e}_i^\top \boldsymbol{W}^\perp \boldsymbol{e}_k - \boldsymbol{e}_j^\top \boldsymbol{W}^\perp \boldsymbol{e}_k = 0 \\ \Longrightarrow\ & a_{ik} - a_{jk} = 0 \\ \Longrightarrow\ & a_{ik} = a_{jk} =: \bar{a}_k. \end{aligned} \tag{12}$$

Then we have that for any $\boldsymbol{x} \in \bar{\boldsymbol{X}}$, $\boldsymbol{x}^\top \boldsymbol{W}^\perp \bar{\boldsymbol{x}} = \bar{a}_k$ where $\bar{a}_k$ is associated with the last/query token $\bar{\boldsymbol{x}}$ and hence

$$\bar{\boldsymbol{X}}\boldsymbol{W}\bar{\boldsymbol{x}} = \bar{\boldsymbol{X}}\boldsymbol{W}^\parallel\bar{\boldsymbol{x}} + \bar{\boldsymbol{X}}\boldsymbol{W}^\perp\bar{\boldsymbol{x}} = \bar{\boldsymbol{X}}\boldsymbol{W}^\parallel\bar{\boldsymbol{x}} + \bar{a}_k\mathbf{1}$$
$$\mathbb{S}(\bar{\boldsymbol{X}}\boldsymbol{W}\bar{\boldsymbol{x}}) = \mathbb{S}(\bar{\boldsymbol{X}}\boldsymbol{W}^\parallel\bar{\boldsymbol{x}} + \bar{a}_k\mathbf{1}) = \mathbb{S}(\bar{\boldsymbol{X}}\boldsymbol{W}^\parallel\bar{\boldsymbol{x}}),$$

which completes the proof. ■

A.6 Proof of Lemma 6

In the following, we present an additional lemma that incorporates Lemma 6.

Lemma 7 *Suppose Assumptions 1 and 2 hold and loss function $\ell : \mathbb{R} \to \mathbb{R}$ is strictly decreasing. For any finite $\boldsymbol{W}$, once $\overline{\texttt{DSET}} \neq \texttt{DSET}$, we have that*

$$\mathcal{L}(\boldsymbol{W}) > \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \bar{\mathcal{L}}(\boldsymbol{W}). \tag{13}$$

Additionally, we have

$$\min_{\boldsymbol{W}' \in \mathcal{S}_{\text{fin}}^{\perp}} \mathcal{L}(\boldsymbol{W}' + \boldsymbol{W}) = \tilde{\mathcal{L}}(\boldsymbol{W}) = \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \bar{\mathcal{L}}(\boldsymbol{W}), \tag{14a}$$

$$\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) = \min_{\boldsymbol{W}} \tilde{\mathcal{L}}(\boldsymbol{W}) = \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \min_{\boldsymbol{W}} \bar{\mathcal{L}}(\boldsymbol{W}). \tag{14b}$$

Proof. We start with proving that for any finite $\boldsymbol{W}$, $\mathcal{L}(\boldsymbol{W}) \geq \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \bar{\mathcal{L}}(\boldsymbol{W})$. Let $\boldsymbol{W}^{\perp} = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^{\perp}}(\boldsymbol{W})$, $\boldsymbol{W}^{\parallel} = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W})$ where we have $\boldsymbol{W} = \boldsymbol{W}^{\perp} + \boldsymbol{W}^{\parallel}$. Let $\boldsymbol{a}_i = \boldsymbol{X}_i\boldsymbol{W}^{\perp}\bar{\boldsymbol{x}}_i$, $\boldsymbol{b}_i = \boldsymbol{X}_i\boldsymbol{W}^{\parallel}\bar{\boldsymbol{x}}_i$ and $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}\bar{\boldsymbol{x}}_i) = \mathbb{S}(\boldsymbol{a}_i + \boldsymbol{b}_i)$. Following (7) and the definition of $\bar{\mathcal{L}}(\boldsymbol{W})$, training losses obey

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\sum_{t\in\mathcal{O}_i} s_{it}\right) \quad \text{and} \quad \bar{\mathcal{L}}(\boldsymbol{W}) = \frac{1}{n}\sum_{i\in\bar{\mathcal{I}}} \ell\left(\frac{\sum_{t\in\mathcal{O}_i} s_{it}}{\sum_{t\in\mathcal{R}_i} s_{it}}\right).$$

From proof of Lemma 5 (more specifically (12)), we have that for any $t \in \mathcal{R}_i$,

$$\boldsymbol{x}_{it}^{\top}\boldsymbol{W}^{\perp}\bar{\boldsymbol{x}}_i = \bar{a}_i \implies a_{it} = \bar{a}_i,\ \forall t \in \mathcal{R}_i$$

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

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

$$\frac{\sum_{t\in\mathcal{O}_i} s_{it}}{\sum_{t\in\mathcal{R}_i} s_{it}} = \frac{\sum_{t\in\mathcal{O}_i} e^{\bar{a}_i + b_{it}}}{\sum_{t\in\mathcal{R}_i} e^{\bar{a}_i + b_{it}}} = \frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}}}.$$

Next following (4) we have that for $i \in \bar{\mathcal{I}}$, $\mathcal{R}_i = \mathcal{O}_i$ and therefore

$$\frac{\sum_{t\in\mathcal{O}_i} s_{it}}{\sum_{t\in\mathcal{R}_i} s_{it}} = 1 \quad \text{for all} \quad i \in \bar{\mathcal{I}}.$$

Note that since a_{it}, b_{it} are finite and $\overline{\texttt{DSET}} \neq \texttt{DSET}$, there exists $i \in [n]$ such that $\sum_{t\in\mathcal{O}_i} s_{it} < \frac{\sum_{t\in\mathcal{O}_i} s_{it}}{\sum_{t\in\mathcal{R}_i} s_{it}}$. Given strictly decreasing loss function ℓ and any finite $\boldsymbol{W}$, the training risks obey

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\sum_{i\in\mathcal{O}_i} s_{it}\right) > \frac{1}{n}\sum_{i\in\bar{\mathcal{I}}} \ell(1) + \frac{1}{n}\sum_{i\in\mathcal{I}} \ell\left(\frac{\sum_{t\in\mathcal{O}_i} s_{it}}{\sum_{t\in\mathcal{R}_i} s_{it}}\right) = \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \bar{\mathcal{L}}(\boldsymbol{W}).$$

It completes the proof of (13).

We next show that

$$\min_{\boldsymbol{W}' \in \mathcal{S}_{\text{fin}}^{\perp}} \mathcal{L}(\boldsymbol{W}' + \boldsymbol{W}) = \tilde{\mathcal{L}}(\boldsymbol{W}) = \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \bar{\mathcal{L}}(\boldsymbol{W}).$$

Recap from Theorem 2 that $\tilde{\mathcal{L}}(\boldsymbol{W}) = \lim_{R\to\infty} \mathcal{L}(\boldsymbol{W} + R\cdot\boldsymbol{W}^{\text{svm}})$. Let $\boldsymbol{W}^{\perp} = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^{\perp}}(\boldsymbol{W})$, $\boldsymbol{W}^{\parallel} = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W})$ where we have $\boldsymbol{W} = \boldsymbol{W}^{\perp} + \boldsymbol{W}^{\parallel}$. Let $\boldsymbol{a}_i = \boldsymbol{X}_i\boldsymbol{W}^{\perp}\bar{\boldsymbol{x}}_i$, $\boldsymbol{b}_i = \boldsymbol{X}_i\boldsymbol{W}^{\parallel}\bar{\boldsymbol{x}}_i$, $\boldsymbol{c}_i = \boldsymbol{X}_i\boldsymbol{W}^{\text{svm}}\bar{\boldsymbol{x}}_i$ and $\boldsymbol{s}_i^R = \mathbb{S}(\boldsymbol{X}_i(R\cdot\boldsymbol{W}^{\text{svm}} + \boldsymbol{W})\bar{\boldsymbol{x}}_i) = \mathbb{S}(\boldsymbol{a}_i + \boldsymbol{b}_i + R\cdot\boldsymbol{c}_i)$. Similarly, for any $t \in \mathcal{R}_i$,

$$\boldsymbol{x}_{it}^{\top}\boldsymbol{W}^{\perp}\bar{\boldsymbol{x}}_i = \bar{a}_i \implies a_{it} = \bar{a}_i,\ \forall t \in \mathcal{R}_i$$

where $\bar{a}_i$ is some constant associated with $\boldsymbol{W}^\perp$. Additionally, since $\boldsymbol{W}^{\text{svm}}$ follows (Graph-SVM), we have

$$(\boldsymbol{x}_{i\tau} - \boldsymbol{x}_{it})^\top \boldsymbol{W}^{\text{svm}} \bar{\boldsymbol{x}}_i \begin{cases} = 0 & \forall t \in \mathcal{R}_i \\ \geq 1 & \forall t \in \bar{\mathcal{R}}_i \end{cases}, \quad \text{for all } \tau \in \mathcal{R}_i \quad \Longrightarrow \quad \begin{cases} c_{it} = \bar{c}_i, & \forall t \in \mathcal{R}_i, \\ c_{it} \leq \bar{c}_i - 1, & \forall t \in \bar{\mathcal{R}}_i. \end{cases} \tag{15}$$

Then for any $i \in [n]$, we get

$$\sum_{t\in\mathcal{O}_i} s_{it}^R = \frac{\sum_{t\in\mathcal{O}_i} e^{\bar{a}_i + b_{it} + R\bar{c}_i}}{\sum_{t\in\mathcal{R}_i} e^{\bar{a}_i + b_{it} + R\bar{c}_i} + \sum_{t\in\bar{\mathcal{R}}_i} e^{a_{it} + b_{it} + Rc_{it}}} = \frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}} + \sum_{t\in\bar{\mathcal{R}}_i} e^{b_{it} + a_{it} - \bar{a}_i + R(c_{it} - \bar{c}_i)}} \leq \frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}}}.$$

Case 1: $\boldsymbol{W}^{\text{svm}} = 0$. Then for all $i \in [n]$, $\bar{\mathcal{R}}_i = \emptyset$ and the equality holds for all $i \in [n]$.

Case 2: $\boldsymbol{W}^{\text{svm}} \neq 0$. Since a_{it}, b_{it} are finite and $c_{it} - \bar{c}_i \leq -1$ for $t \in \bar{\mathcal{R}}_i$ following (15), the equality holds when $R \to \infty$, and therefore we have for any $i \in [n]$,

$$\lim_{R\to\infty} \sum_{t\in\mathcal{O}_i} s_{it}^R = \frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}}}$$

and

$$\begin{aligned} \tilde{\mathcal{L}}(\boldsymbol{W}) = \lim_{R\to\infty} \mathcal{L}(R \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}) &= \lim_{R\to\infty} \frac{1}{n} \sum_{i=1}^n \ell\left(\sum_{i\in\mathcal{O}_i} s_{it}^R\right) \\ &= \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \frac{1}{n}\sum_{i\in\mathcal{I}} \ell\left(\frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}}}\right) \\ &= \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \bar{\mathcal{L}}(\boldsymbol{W}). \end{aligned} \tag{16}$$

Additionally, we have for any $\boldsymbol{W}' \in \mathcal{S}_{\text{fin}}^\perp$,

$$\mathcal{L}(\boldsymbol{W}' + \boldsymbol{W}) \geq \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \bar{\mathcal{L}}(\boldsymbol{W}' + \boldsymbol{W}) = \frac{|\bar{\mathcal{I}}|}{n}\ell(1) + \bar{\mathcal{L}}(\boldsymbol{W})$$

where the inequality uses (13) and the equality comes from Lemma 5. Since bound is achievable (by choosing $\boldsymbol{W}' = \lim_{R\to\infty} R \cdot \boldsymbol{W}^{\text{svm}}$ as in (16)), then combining it with (16) completes the proof of (14a). (14b) is directly obtained from (14a).

■

B GLOBAL CONVERGENCE OF GRADIENT DESCENT

B.1 Supporting Results under the Setting of Theorem 2

In this section, we introduce results useful for the main proof. Recap the setting of Theorem 2 where $\ell(u) = -\log(u)$. Therefore loss defined in (7) is

$$\mathcal{L}(\boldsymbol{W}) = -\frac{1}{n}\sum_{i=1}^n \log\left(\sum_{t\in\mathcal{O}_i} s_{it}^{\boldsymbol{W}}\right) \tag{17}$$

where $\boldsymbol{s}_i^{\boldsymbol{W}} = \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)$ and $\mathcal{O}_i$'s follow (3).

• **$\nabla\mathcal{L}(\boldsymbol{W})$ under the setting of Theorem 2.** For any $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, let $\boldsymbol{h}_i = \boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i$, $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{h}_i)$, $\boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$.

$$\begin{aligned} \nabla\mathcal{L}(\boldsymbol{W}) &= \frac{1}{n}\sum_{i=1}^n \ell'(\boldsymbol{\gamma}_i^\top \boldsymbol{s}_i) \boldsymbol{X}_i^\top \mathbb{S}'(\boldsymbol{h}_i) \boldsymbol{\gamma}_i \bar{\boldsymbol{x}}_i^\top \\ &= \frac{1}{n}\sum_{i=1}^n -\frac{1}{\boldsymbol{\gamma}_i^\top \boldsymbol{s}_i} \boldsymbol{X}_i^\top (\operatorname{diag}(\boldsymbol{s}_i) - \boldsymbol{s}_i \boldsymbol{s}_i^\top) \boldsymbol{\gamma}_i \bar{\boldsymbol{x}}_i^\top \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{t\in\bar{\mathcal{O}}_i} s_{it}(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i}) \bar{\boldsymbol{x}}_i^\top \end{aligned} \tag{18}$$

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

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

Here, the second equality comes from (6).

• **Lipschitzness of $\nabla\mathcal{L}(\boldsymbol{W})$ in (18).** For any $\boldsymbol{W}, \dot{\boldsymbol{W}} \in \mathbb{R}^{d\times d}$, let $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}\bar{\boldsymbol{x}}_i)$ and $\dot{\boldsymbol{s}}_i = \mathbb{S}(\boldsymbol{X}_i\dot{\boldsymbol{W}}\bar{\boldsymbol{x}}_i)$. Consider bounded tokens and let $M := \max_{k\in[K]} \|\boldsymbol{e}_k\|$. Following (18), we have:

$$\begin{aligned}\left\|\nabla\mathcal{L}(\boldsymbol{W}) - \nabla\mathcal{L}(\dot{\boldsymbol{W}})\right\|_F &= \left\|\frac{1}{n}\sum_{i=1}^n \sum_{t\in\bar{\mathcal{O}}_i} (s_{it} - \dot{s}_{it})(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i})\bar{\boldsymbol{x}}_i^\top\right\|_F \\ &\le \frac{1}{n}\sum_{i=1}^n \sum_{t\in\bar{\mathcal{O}}_i} |s_{it} - \dot{s}_{it}| \ \left\|(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i})\bar{\boldsymbol{x}}_i^\top\right\|_F \\ &\le \frac{2M^2}{n}\sum_{i=1}^n \sum_{t\in\bar{\mathcal{O}}_i} |s_{it} - \dot{s}_{it}| \\ &\le \frac{2M^2}{n}\sum_{i=1}^n \|\boldsymbol{s}_i - \dot{\boldsymbol{s}}_i\|_1 \\ &\le \frac{2M^2}{n}\sum_{i=1}^n \sqrt{T_i}\cdot\|\boldsymbol{s}_i - \dot{\boldsymbol{s}}_i\|.\end{aligned} \tag{19}$$

Next for any $\boldsymbol{s}, \dot{\boldsymbol{s}}$, we get

$$\begin{aligned}\|\boldsymbol{s} - \dot{\boldsymbol{s}}\| &= \|\mathbb{S}(\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}}) - \mathbb{S}(\boldsymbol{X}\dot{\boldsymbol{W}}\bar{\boldsymbol{x}})\| \\ &\le \|\boldsymbol{X}\boldsymbol{W}\bar{\boldsymbol{x}} - \boldsymbol{X}\dot{\boldsymbol{W}}\bar{\boldsymbol{x}}\| \\ &\le M^2\left\|\boldsymbol{W} - \dot{\boldsymbol{W}}\right\|_F.\end{aligned} \tag{20}$$

Combining results in that

$$\left\|\nabla\mathcal{L}(\boldsymbol{W}) - \nabla\mathcal{L}(\dot{\boldsymbol{W}})\right\|_F \le 2M^4\sqrt{T_{\max}}\cdot\left\|\boldsymbol{W} - \dot{\boldsymbol{W}}\right\|_F \tag{21}$$

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

$$L := 2M^4\sqrt{T_{\max}} \tag{22}$$

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

B.2 Proof of Lemma 2

In this subsection, we provide and prove a general version of Lemma 2. To to that, we first introduce the following new subspaces.

Definition 5 *Define the subspace $\mathcal{S}_{active}$ as the span of all matrices $(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top$ for all $(i \to j) \in \mathcal{G}^{(k)}$ and $k \in [K]$.*

Definition 6 (Cyclic subspace (Restated)) *Define cyclic subspace $\mathcal{S}_{fin}$ as 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 $k \in [K]$.*

Observe that $\mathcal{S}_{\text{fin}}$ is a subspace of $\mathcal{S}_{\text{active}}$. This is because if two nodes $i, j \in \mathcal{G}^{(k)}$ and $i \asymp j$, this also implies that $i \to j$ and $j \to i$.

Definition 7 (SVM subspace) *Define svm subspace $\mathcal{S}_{svm}$ as the orthogonal complement of the subspace $\mathcal{S}_{fin}$ inside the subspace $\mathcal{S}_{active}$.*

Lemma 8 *We have the following:*

(i) Recall the definition of $\boldsymbol{W}^{svm}$ in (Graph-SVM). $\boldsymbol{W}^{svm} \in \mathcal{S}_{svm}$.

(ii) Let $\mathcal{S}_{active}^{\perp}$ be the orthogonal complement of $\mathcal{S}_{active}$ inside $\mathbb{R}^{d\times d}$. Then,

$$\left\|\boldsymbol{\Pi}_{\mathcal{S}_{active}^{\perp}}(\nabla\mathcal{L}(\boldsymbol{W}))\right\|_F = 0, \quad \forall \boldsymbol{W} \in \mathbb{R}^{d\times d}.$$

Proof.

• **(i):** Recall the definition of $\boldsymbol{W}^{\text{svm}}$:

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

Assume that the statement is not correct. Then, either $\|\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}^{\text{svm}})\|_F > 0$ or $\left\|\boldsymbol{\Pi}_{\mathcal{S}_{\text{active}}^{\perp}}(\boldsymbol{W}^{\text{svm}})\right\|_F > 0$.

By definition, $\|\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}^{\text{svm}})\|_F = 0$ since for all $(i \asymp j) \in \mathcal{G}^{(k)}$, $(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W}^{\text{svm}} \boldsymbol{e}_k = 0$.

On the other hand, if $\left\|\boldsymbol{\Pi}_{\mathcal{S}_{\text{active}}^{\perp}}(\boldsymbol{W}^{\text{svm}})\right\|_F > 0$, then $\boldsymbol{W}^{\text{svm}} - \boldsymbol{\Pi}_{\mathcal{S}_{\text{active}}^{\perp}}(\boldsymbol{W}^{\text{svm}})$ also satisfies all of the constraints of (Graph-SVM), and

$$\left\|\boldsymbol{W}^{\text{svm}} - \boldsymbol{\Pi}_{\mathcal{S}_{\text{active}}^{\perp}}(\boldsymbol{W}^{\text{svm}})\right\|_F < \|\boldsymbol{W}^{\text{svm}}\|_F,$$

which is a contradiction. Therefore, $\boldsymbol{W}^{\text{svm}} \in \mathcal{S}_{\text{svm}}$.

• **(ii):** From (18), we know that

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

where $\bar{\mathcal{O}}_i$ is given by (3). By definition of $\mathcal{S}_{\text{active}}$, $\left\|\boldsymbol{\Pi}_{\mathcal{S}_{\text{active}}^{\perp}}(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i})\bar{\boldsymbol{x}}_i^\top\right\|_F = 0$ for any $i \in [n]$ and $t \in \bar{\mathcal{O}}_i$. As $\nabla\mathcal{L}(\boldsymbol{W})$ is the summation of these terms, the advertised result is proved. ■

Now, we are ready to prove a stronger version of Lemma 2.

Lemma 9 (Stronger version of Lemma 2) *Suppose Assumptions 1 and 2 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 $\mathcal{S}_{active}$.*

Proof. Let $\mathcal{S}_K$ be the span of all $\boldsymbol{e}_i\boldsymbol{e}_j^\top$ where $i, j \in [K]$.

• **First Case: $\boldsymbol{W} \in \mathcal{S}_K$.** Let $g : \mathcal{S}_K \to \mathbb{R}^{K\times K}$ such that $g(\boldsymbol{W}) = \boldsymbol{E}\boldsymbol{W}\boldsymbol{E}^\top$. By definition, this function is linear. In addition to that, this function g is invertible by Assumption 1 and the domain of the function is $\mathcal{S}_K$. Note that Assumption 1 ensures $\text{rank}(\boldsymbol{E}) = K$.

Let $\boldsymbol{E}' = \boldsymbol{C}' = \boldsymbol{I}_k$, $(\boldsymbol{X}_i', y_i')$ be a DSET such that $y_i' = y_i$ and $\boldsymbol{X}_i' = \boldsymbol{X}_i\boldsymbol{E}^\dagger$. Then, for any $\boldsymbol{W}' \in \mathbb{R}^{K\times K}$, we have the following:

$$\mathcal{L} \circ g^{-1}(\boldsymbol{W}') = \frac{1}{n}\sum_{i=1}^{n} \ell\left((\boldsymbol{c}_{y_i}')^\top (\boldsymbol{X}_i')^\top \mathbb{S}(\boldsymbol{X}_i'\boldsymbol{W}'\bar{\boldsymbol{x}}_i')\right)$$

Using Lemma 10 and 11, we know that $\mathcal{L} \circ g^{-1}(\boldsymbol{W}')$ is convex on $\mathbb{R}^{K\times K}$ and strictly convex on $g(\mathcal{S}_{\text{active}})$. Using these two facts and Lemma 10, we have $\mathcal{L}(\boldsymbol{W})$ is convex on $\mathcal{S}_K$ and strictly convex on $\mathcal{S}_{\text{active}} \cap \mathcal{S}_K = \mathcal{S}_{\text{active}}$.

• **Second Case: $\boldsymbol{W} \notin \mathcal{S}_K$.** By definition of loss function in (ERM), we have

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{\Pi}_{\mathcal{S}_K}(\boldsymbol{W}) \bar{\boldsymbol{x}}_i)) = \mathcal{L}(\boldsymbol{\Pi}_{\mathcal{S}_K}(\boldsymbol{W})) \tag{23}$$

Let $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{d\times d}$ be arbitrary variables. For any $0 < \lambda < 1$, we have the following:

$$\mathcal{L}(\lambda \boldsymbol{W}_1 + (1-\lambda)\boldsymbol{W}_2) = \mathcal{L}(\lambda \boldsymbol{\Pi}_{\mathcal{S}_K}(\boldsymbol{W}_1) + \lambda \boldsymbol{\Pi}_{\mathcal{S}_K}(\boldsymbol{W}_2)) \tag{24}$$

Then, using (23) and (24), we have the following:

$$\begin{aligned}\lambda\mathcal{L}(\boldsymbol{W}_1) + (1-\lambda)\mathcal{L}(\boldsymbol{W}_2) &= \lambda\mathcal{L}(\boldsymbol{\Pi}_{\mathcal{S}_K}(\boldsymbol{W}_1)) + (1-\lambda)\mathcal{L}(\boldsymbol{\Pi}_{\mathcal{S}_K}(\boldsymbol{W}_2)) \\ &\overset{(a)}{\geq} \mathcal{L}(\lambda \boldsymbol{\Pi}_{\mathcal{S}_K}(\boldsymbol{W}_1) + \lambda \boldsymbol{\Pi}_{\mathcal{S}_K}(\boldsymbol{W}_2)) = \mathcal{L}(\lambda \boldsymbol{W}_1 + (1-\lambda)\boldsymbol{W}_2)\end{aligned}$$

where (a) follows from the convexity of $\mathcal{L}(\boldsymbol{W})$ inside $\mathcal{S}_K$. This implies that $\mathcal{L}(\boldsymbol{W})$ is convex when $\boldsymbol{W} \notin \mathcal{S}_K$. Note that $\mathcal{S}_{\text{active}} \subset \mathcal{S}_K$, therefore we do not look at the strict convexity in this case. ■

Lemma 10 *Let $T : \mathcal{X} \to \mathcal{Y}$ be an invertible linear map. If a function $f : \mathcal{Y} \to \mathbb{R}$ is convex/strictly convex on $\mathcal{Y}$, then $f \circ T(x)$ is a convex/strictly convex function on $\mathcal{X}$.*

Proof. Let $x_1 \neq x_2 \in \mathcal{X}$ be arbitrary variables. Let $y_1 = T(x_1)$ and $y_2 = T(x_2)$. Since T is an invertible map, $y_1 \neq y_2$. Since T is a linear map, $T(\lambda x_1 + (1-\lambda)x_2) = \lambda y_1 + (1-\lambda)y_2$ for $0 < \lambda < 1$. Then, we obtain the following

$$\begin{aligned}\lambda(f \circ T(x_1)) + (1-\lambda)(f \circ T(x_2)) &= \lambda f(y_1) + (1-\lambda) f(y_2) \\ &\overset{(a)}{>} f(\lambda y_1 + (1-\lambda)y_2) \\ &= f \circ T(\lambda x_1 + (1-\lambda)x_2)\end{aligned}$$

where (a) follows from the strict convexity of the function f. This implies that $f \circ T(x)$ is a strictly convex function on $\mathcal{X}$. Note that if $y_1 = y_2$, then we cannot achieve (a). Additionally, if f is convex instead of strictly convex, then $>$ in (a) is changed to $\geq$, and $f \circ T(x)$ is convex. ■

Lemma 11 *Suppose that Assumption 2 holds and $\boldsymbol{E} = \boldsymbol{I}_d$. Let $f : \mathbb{R}^{d\times d} \to \mathbb{R}^{d^2}$ be a linear transformation defined as $f(\boldsymbol{W}) = \boldsymbol{v}$ where $v_{i\times d+j} = \boldsymbol{e}_i^\top \boldsymbol{W} \boldsymbol{e}_j$. Then, $\mathcal{L} \circ f^{-1}(\boldsymbol{v})$ is convex. Furthermore, $\mathcal{L} \circ f^{-1}(\boldsymbol{v})$ is strictly convex on $f(\mathcal{S}_{active})$, where $\mathcal{S}_{active}$ is defined in Definition 5.*

Proof. • **We first prove that $\mathcal{L} \circ f^{-1}(\boldsymbol{v})$ is convex.** Let $\bar{\ell} : \mathbb{R}^{d^2} \times \mathbb{R}^{T\times d} \times \mathbb{R} \to \mathbb{R}$ be defined as follows:

$$\bar{\ell}(\boldsymbol{v}, \boldsymbol{X}, y) := \ell\left(\boldsymbol{c}_y^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\left(f^{-1}(\boldsymbol{v})\right)\bar{\boldsymbol{x}})\right).$$

Then, using (ERM), we have the following:

$$\mathcal{L} \circ f^{-1}(\boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\left(f^{-1}(\boldsymbol{v})\right)\bar{\boldsymbol{x}}_i)\right) = \frac{1}{n}\sum_{i=1}^{n} \bar{\ell}(\boldsymbol{v}, \boldsymbol{X}_i, y_i). \tag{25}$$

Note that the summation of convex functions is convex. Therefore, it is sufficient to prove the convexity of $\mathcal{L} \circ f^{-1}(\boldsymbol{v})$ by proving the convexity of $\bar{\ell}(\boldsymbol{v}, \boldsymbol{X}, y)$ for an arbitrary pair of input sequence and label $(\boldsymbol{X}, y)$. For the simplicity of notation, we use $\bar{\ell}(\boldsymbol{v})$ instead of $\bar{\ell}(\boldsymbol{v}, \boldsymbol{X}, y)$. Let m_j be the number of token ID j inside input sequence $\boldsymbol{X}$ for $j \in [K]$. Let k be the last token of $\boldsymbol{X}$. By Assumption 1 and log-loss, we know that

$$\bar{\ell}(\boldsymbol{v}) := \bar{\ell}(\boldsymbol{v}, \boldsymbol{X}, y) = -\log\left(\frac{m_y \cdot e^{\boldsymbol{v}_{y\times d+k}}}{\sum_{j\in[K]} m_j \cdot e^{\boldsymbol{v}_{j\times d+k}}}\right) = \log\left(\sum_{j\in[K]} m_j \cdot e^{\boldsymbol{v}_{j\times d+k}}\right) - \log(m_y \cdot e^{\boldsymbol{v}_{y\times d+k}}).$$

Let $\boldsymbol{z} \in \mathbb{R}^{d^2}$ be a vector such that the $(j \times d + k)^{\text{th}}$ element of $\boldsymbol{z}$ is $z_{j\times d+k} = m_j \cdot e^{v_{j\times d+k}}$ for $k \in [K]$, otherwise $z_i = 0$. Then, the Hessian matrix of $\bar{\ell}(\boldsymbol{v})$ is

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

For any $\boldsymbol{u} \in \mathbb{R}^{d^2}$, we obtain that

$$\boldsymbol{u}^\top \nabla^2 \bar{\ell}(\boldsymbol{v}) \boldsymbol{u} = \frac{1}{(\mathbf{1}^\top \boldsymbol{z})^2} \left(\left(\sum_{j=1}^{d^2} z_j \right) \left(\sum_{j=1}^{d^2} u_j^2 z_j \right) - \left(\sum_{j=1}^{d^2} u_j z_j \right)^2 \right) \geq 0. \tag{26}$$

Since $\boldsymbol{z}_i \geq 0$, $i \in [d^2]$, (26) follows from the Cauchy-Schwarz inequality $(\boldsymbol{\alpha}^\top \boldsymbol{\alpha})(\boldsymbol{\beta}^\top \boldsymbol{\beta}) \geq (\boldsymbol{\alpha}^\top \boldsymbol{\beta})^2$ 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 $\bar{\ell}(\boldsymbol{v})$ is convex.

• **Next, we will show that $\mathcal{L} \circ f^{-1}(\boldsymbol{v})$ is strictly convex on $f(\mathcal{S}_{\text{active}})$.** Assume that $\mathcal{L} \circ f^{-1}(\boldsymbol{v})$ is not strictly convex on $f(\mathcal{S}_{\text{active}})$. Using the convexity of $\mathcal{L} \circ f^{-1}(\boldsymbol{v})$, this implies that there exist $\boldsymbol{u}, \boldsymbol{v} \in f(\mathcal{S}_{\text{active}})$, $\|\boldsymbol{u}\|_2 > 0$ such that

$$\boldsymbol{u}^\top \left(\nabla^2 \mathcal{L} \circ f^{-1}(\boldsymbol{v}) \right) \boldsymbol{u} = 0$$

Using the convexity of $\bar{\ell}(\boldsymbol{v})$ and (25), we have the following:

$$\boldsymbol{u}^\top \left(\nabla^2 \bar{\ell}(\boldsymbol{v}, \boldsymbol{X}_i, \boldsymbol{y}_i) \right) \boldsymbol{u} = 0 \qquad \forall i \in [n] \tag{27}$$

Now, we are going to prove that $\|\boldsymbol{u}\|_2 = 0$ if (27) holds. As $\boldsymbol{u} \in f(\mathcal{S}_{\text{active}})$, there exists $\boldsymbol{W} \in \mathcal{S}_{\text{active}}$ such that $f(\boldsymbol{W}) = \boldsymbol{u}$. As the function f preserves the norm, $\|\boldsymbol{W}\|_F > 0$. By definition of $\mathcal{S}_{\text{active}}$, there exist $\bar{i}, \bar{j}, \bar{k} \in [K]$ and $(\boldsymbol{X}_{\bar{n}}, y_{\bar{n}}) \in \texttt{DSET}$ such that $\langle (\boldsymbol{e}_{\bar{i}} - \boldsymbol{e}_{\bar{j}}) \boldsymbol{e}_{\bar{k}}^\top, \boldsymbol{W} \rangle > 0$, $\boldsymbol{X}_{\bar{n}}$ includes the $\bar{j}^{\text{th}}$ token, the last token of $\boldsymbol{X}_{\bar{n}}$ is the $\bar{k}^{\text{th}}$ token, and $y_{\bar{n}} = \bar{i}$. On the other hand, by Assumption 2, $z_{\bar{i}\times d+\bar{k}}$ and $z_{\bar{j}\times d+\bar{k}}$ in (26) are non-zero for this input sequence $\boldsymbol{X}_{\bar{n}}$. Using the equality condition of Cauchy-Schwartz Inequality in (26), we obtain that $u_{\bar{i}\times d+\bar{k}} - u_{\bar{j}\times d+\bar{k}} = 0$. This implies that

$$\begin{aligned} 0 &= u_{\bar{i}\times d+\bar{k}} - u_{\bar{j}\times d+\bar{k}} \\ &= \boldsymbol{e}_{\bar{i}}^\top \boldsymbol{W} \boldsymbol{e}_{\bar{k}} - \boldsymbol{e}_{\bar{j}}^\top \boldsymbol{W} \boldsymbol{e}_{\bar{k}} \\ &= (\boldsymbol{e}_{\bar{i}} - \boldsymbol{e}_{\bar{j}})^\top \boldsymbol{W} \boldsymbol{e}_{\bar{k}} = \langle (\boldsymbol{e}_{\bar{i}} - \boldsymbol{e}_{\bar{j}}) \boldsymbol{e}_{\bar{k}}^\top, \boldsymbol{W} \rangle \end{aligned}$$

which contradicts with the fact that $\|\boldsymbol{u}\|_2 > 0$. This completes the proof. ■

B.3 Divergence of $\|\boldsymbol{\Pi}_{\mathcal{S}_{\text{svm}}}(\mathbf{W}(\tau))\|_F$

We first introduce the following lemmas establishing the descent property of gradient descent for $\mathcal{L}(\boldsymbol{W})$ (Lemma 12) and the correlation between $\nabla\mathcal{L}(\boldsymbol{W})$ and the solution of (Graph-SVM) $\boldsymbol{W}^{\text{svm}}$ (Lemma 13) under the setting of Theorem 2. The proofs in this section follow Appendix B.1 of Tarzanagh et al. (2023a).

Lemma 12 (Descent Lemma) *Consider the loss in (17) and choose step size $\eta \leq 1/L$ where L is the Lipschitzness of $\nabla\mathcal{L}(\boldsymbol{W})$ defined in (22). 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 with L defined in (22), we get

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

The inequality above also indicates that

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

■

Lemma 13 *Let $\boldsymbol{W}^{svm}$ be the SVM solution of (Graph-SVM) and suppose $\boldsymbol{W}^{svm} \neq 0$. For any $\boldsymbol{W}$ with $\|\boldsymbol{W}\|_F < \infty$, the training loss $\mathcal{L}(\boldsymbol{W})$ in (17) obeys $\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{svm} \rangle < 0$. Equivalently, $\langle \boldsymbol{\Pi}_{\mathcal{S}_{svm}}(\nabla \mathcal{L}(\boldsymbol{W})), \boldsymbol{W}^{svm} \rangle < 0$.*

Proof. Recap $\mathcal{O}_i, \bar{\mathcal{O}}_i, \mathcal{R}_i, \bar{\mathcal{R}}_i, i \in [n]$ in (3). From (18), 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

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

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

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

Using the fact that $\boldsymbol{W}^{\text{svm}} \in \mathcal{S}_{\text{svm}}$ (Lemma 4) completes the proof. ■

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

Theorem 5 *Consider the same setting as in Theorem 2, 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 (22) and any starting point $\boldsymbol{W}(0)$ satisfies $\lim_{\tau \to \infty} \|\boldsymbol{\Pi}_{\mathcal{S}_{svm}}(\boldsymbol{W}(\tau))\|_F = \infty$.*

Proof. Following Lemma 12, 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 $\langle \boldsymbol{\Pi}_{\mathcal{S}_{\text{svm}}}(\nabla \mathcal{L}(\boldsymbol{W})), \boldsymbol{W}^{\text{svm}} \rangle = 0$. This leads to a contradiction with Lemma 13 which says that for any finite $\boldsymbol{W}$, $\langle \boldsymbol{\Pi}_{\mathcal{S}_{\text{svm}}}(\nabla \mathcal{L}(\boldsymbol{W})), \boldsymbol{W}^{\text{svm}} \rangle < 0$. This implies that $\|\boldsymbol{\Pi}_{\mathcal{S}_{\text{svm}}}(\boldsymbol{W}(\tau))\|_F \to \infty$. ■

B.4 Uniqueness and Finiteness of $\mathbf{W}^{\text{fin}}$

Lemma 14 *Consider the setting of Theorem 2. $\boldsymbol{W}^{fin}$ defined in Theorem 2 is unique and finite.*

Proof. Following Lemma 6, it is equivalent to show $\mathcal{W}^{\text{fin}}$ defined in Def. 3 has unique element $\bar{\boldsymbol{W}}^{\text{fin}}$ and $\bar{\boldsymbol{W}}^{\text{fin}} = \boldsymbol{W}^{\text{fin}}$ is unique and finite. To start with, recap the definition of $\overline{\texttt{DSET}}$ (Definition 3). Denote $\mathcal{I} \subset [n]$ as in (5), and let $\boldsymbol{s}_i = \mathbb{S}(\bar{\boldsymbol{X}}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)$ where $(\bar{\boldsymbol{X}}_i, y_i) \in \overline{\texttt{DSET}}$. What's more, recap the ERM loss from (17) and loss function $\ell(u) = -\log(u)$. Then we have

$$\bar{\boldsymbol{W}}^{\text{fin}} = \arg\min_{\boldsymbol{W} \in \mathcal{S}_{\text{fin}}} \bar{\mathcal{L}}(\boldsymbol{W}) \quad \text{where} \quad \bar{\mathcal{L}}(\boldsymbol{W}) = \frac{1}{n} \sum_{i \in \mathcal{I}} -\log \left(\sum_{t \in \mathcal{O}_i} s_{it} \right). \tag{28}$$

Different from (3), $\mathcal{O}_i$ for dataset $\overline{\texttt{DSET}}$ is defined as follows:

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

where $\bar{T}_i$ is the number of tokens – tokens that are in the same SCC as label token y_i within their corresponding TPG – in $\bar{\boldsymbol{X}}_i$ and if recap the notation of $\mathcal{R}_i$ in (3), here we have $|\mathcal{R}_i| = \bar{T}_i$.

We will first prove that $\bar{\boldsymbol{W}}^{\text{fin}}$ is finite by contradiction. Specifically, we will show that for any $\boldsymbol{W} \in \mathcal{S}_{\text{fin}}$ with $\|\boldsymbol{W}\|_F \neq 0$, $\lim_{R \to \infty} \bar{\mathcal{L}}(R \cdot \boldsymbol{W}) = \infty$ which implies that the optimal solution $\bar{\boldsymbol{W}}^{\text{fin}}$ has to be finite.

Let $\boldsymbol{W} \in \mathcal{S}_{\text{fin}}$ be arbitrary attention weight. Following the definition of $\mathcal{S}_{\text{fin}}$ as in Def. 1, we have that $(\boldsymbol{e}_i - \boldsymbol{e}_j)\boldsymbol{e}_k^\top \in \mathcal{S}_{\text{fin}}$ for all $(i \asymp j) \in \mathcal{G}^{(k)}$ and $k \in [K]$. For any $\bar{\boldsymbol{X}}$, let $\mathcal{O}, \bar{\mathcal{O}}$ correspond to the token index sets. Then we have

$$\sum_{t \in \mathcal{O}} s_t = \frac{|\mathcal{O}| e^{\boldsymbol{e}_y^\top (R \cdot \boldsymbol{W}) \boldsymbol{e}_k}}{|\mathcal{O}| e^{\boldsymbol{e}_y^\top (R \cdot \boldsymbol{W}) \boldsymbol{e}_k} + \sum_{t \in \bar{\mathcal{O}}} e^{\boldsymbol{e}_t^\top (R \cdot \boldsymbol{W}) \boldsymbol{e}_k}} = \frac{1}{1 + \sum_{t \in \bar{\mathcal{O}}} e^{(\boldsymbol{e}_t - \boldsymbol{e}_y)^\top (R \cdot \boldsymbol{W}) \boldsymbol{e}_k} / |\mathcal{O}|}.$$

Given the sample loss $\ell = -\log\left(\sum_{t \in \mathcal{O}} s_t\right)$ and to prevent it from divergence as $R \to \infty$, that is, $\sum_{t \in \mathcal{O}} s_t \not\to 0$, we have

$$\sum_{t \in \bar{\mathcal{O}}} e^{(\boldsymbol{e}_t - \boldsymbol{e}_y)^\top (R \cdot \boldsymbol{W}) \boldsymbol{e}_k} \not\to \infty \quad \Longrightarrow \quad (\boldsymbol{e}_y - \boldsymbol{e}_t)^\top \boldsymbol{W} \boldsymbol{e}_k \geq 0 \text{ for all } t \in \bar{\mathcal{O}}$$

where $\boldsymbol{e}_y$ is the label token and $\boldsymbol{e}_t$ is any other token in $\bar{\mathcal{O}}$. Recap from the construction of TPG in Section 2.1, the directed edge $y \to t$ exists in the graph $\mathcal{G}^{(k)}$. Since SCC is bidirectionally reachable, which means there exists route from t to y, e.g., $t \to p_1 \to p_2 \to \cdots p_m \to y$, similarly we have

$$(\boldsymbol{e}_t - \boldsymbol{e}_{p_1})^\top \boldsymbol{W} \boldsymbol{e}_k,\ (\boldsymbol{e}_{p_1} - \boldsymbol{e}_{p_2})^\top \boldsymbol{W} \boldsymbol{e}_k,\ \cdots, (\boldsymbol{e}_{p_m} - \boldsymbol{e}_y)^\top \boldsymbol{W} \boldsymbol{e}_k \geq 0 \Longrightarrow (\boldsymbol{e}_t - \boldsymbol{e}_y)^\top \boldsymbol{W} \boldsymbol{e}_k \geq 0.$$

Combining results in that $(\boldsymbol{e}_t - \boldsymbol{e}_y)^\top \boldsymbol{W} \boldsymbol{e}_k = 0$. This implies that for all $(i \asymp j) \in \mathcal{G}^{(k)}$ and $R \to \infty$, to ensure the training loss $\bar{\mathcal{L}}(R \cdot \boldsymbol{W})$ finite, $(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{W} \boldsymbol{e}_k = 0$, which contradicts the facts that $\boldsymbol{W} \in \mathcal{S}_{\text{fin}}$ and $\boldsymbol{W} \neq 0$.

Next, we prove that there is at most one local minimum for $\bar{\mathcal{L}}(\boldsymbol{W})$ based on Lemma 2. Suppose to the contrary that we have two optimal solutions satisfying $\min_{\boldsymbol{W}} \bar{\mathcal{L}}(\boldsymbol{W}) = \bar{\mathcal{L}}(\boldsymbol{W}_1^{\text{fin}}) = \bar{\mathcal{L}}(\boldsymbol{W}_2^{\text{fin}}), \boldsymbol{W}_1^{\text{fin}} \neq \boldsymbol{W}_2^{\text{fin}}$. From Lemma 2, since $\mathcal{L}(\boldsymbol{W})$ is strictly convex over subspace $\mathcal{S}_{\text{fin}}$, for any $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathcal{S}_{\text{fin}}, \lambda \in (0,1)$, if $\boldsymbol{W}_1 \neq \boldsymbol{W}_2$, we have

$$\bar{\mathcal{L}}((1-\lambda)\boldsymbol{W}_1 + \lambda \boldsymbol{W}_2) < (1-\lambda)\bar{\mathcal{L}}(\boldsymbol{W}_1) + \lambda \bar{\mathcal{L}}(\boldsymbol{W}_2) \tag{29}$$

Substitute $\boldsymbol{W}_1 = \boldsymbol{W}_1^{\text{fin}}, \boldsymbol{W}_2 = \boldsymbol{W}_2^{\text{fin}}$, we get

$$\bar{\mathcal{L}}((1-\lambda)\boldsymbol{W}_1^{\text{fin}} + \lambda \boldsymbol{W}_2^{\text{fin}}) < (1-\lambda)\bar{\mathcal{L}}(\boldsymbol{W}_1^{\text{fin}}) + \lambda \bar{\mathcal{L}}(\boldsymbol{W}_2^{\text{fin}}) = \min_{\boldsymbol{W}} \bar{\mathcal{L}}(\boldsymbol{W}) \tag{30}$$

which leads to a contradiction to the assumption that $\boldsymbol{W}_1^{\text{fin}}$ and $\boldsymbol{W}_2^{\text{fin}}$ are both optimal solutions. Combining this with the fact that $\boldsymbol{W}^{\text{fin}}$ is not attained at infinity, there exists one unique and finite solution with $\bar{\boldsymbol{W}}^{\text{fin}} = \arg\min_{\boldsymbol{W} \in \mathcal{S}_{\text{fin}}} \bar{\mathcal{L}}(\boldsymbol{W})$. ∎

B.5 Proof of Theorem 2

Lemma 15 *Consider the same setting of Theorem 2. Given any $\pi > 0$, there exists $R_\pi > 0$ such that for any $\boldsymbol{W}$ with $\left\|\boldsymbol{\Pi}_{\mathcal{S}_{fin}}(\boldsymbol{W})\right\|_F < \infty$ and $\left\|\boldsymbol{\Pi}_{\mathcal{S}_{fin}^\perp}(\boldsymbol{W})\right\|_F > R_\pi$,*

$$\mathcal{L}(\boldsymbol{W}) \geq \mathcal{L}\left((1+\pi)\frac{\left\|\boldsymbol{\Pi}_{\mathcal{S}_{fin}^\perp}(\boldsymbol{W})\right\|_F}{\|\boldsymbol{W}^{svm}\|_F}\boldsymbol{W}^{svm} + \boldsymbol{\Pi}_{\mathcal{S}_{fin}}(\boldsymbol{W})\right).$$

Proof. Recap $\mathcal{O}_i, \bar{\mathcal{O}}_i, \mathcal{R}_i, \bar{\mathcal{R}}_i$, $i \in [n]$ from (3) and recap from (17), we get that for any $\boldsymbol{W}$,

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

where $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}\bar{\boldsymbol{x}}_i)$.

To obtain the result, we establish a refined softmax probability control by studying the distance to $\bar{\mathcal{L}}(\boldsymbol{W})$ as defined in Definition 4. Let $\boldsymbol{W}^\parallel = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W})$, $\boldsymbol{W}^\perp = \boldsymbol{W} - \boldsymbol{W}^\parallel$, $\left\|\boldsymbol{W}^\perp\right\|_F = R$, and $\Theta = 1/\left\|\boldsymbol{W}^{\text{svm}}\right\|_F$. Let $\boldsymbol{b}_i = \boldsymbol{X}_i\boldsymbol{W}^\parallel\bar{\boldsymbol{x}}_i$, $\boldsymbol{a}_i^\star = \boldsymbol{X}_i((1+\pi)R\Theta\cdot\boldsymbol{W}^{\text{svm}})\bar{\boldsymbol{x}}_i$, and $\boldsymbol{s}_i^\star = \mathbb{S}(\boldsymbol{X}_i((1+\pi)R\Theta\cdot\boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel)\bar{\boldsymbol{x}}_i) = \mathbb{S}(\boldsymbol{a}_i^\star + \boldsymbol{b}_i)$. Additionally, let $\boldsymbol{a}_i = \boldsymbol{X}_i\boldsymbol{W}^\perp\bar{\boldsymbol{x}}_i$, $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}\bar{\boldsymbol{x}}_i) = \mathbb{S}(\boldsymbol{a}_i + \boldsymbol{b}_i)$, $\gamma_i^\star := \boldsymbol{c}_{y_i}^\top\boldsymbol{X}_i^\top\boldsymbol{s}_i^\star$, and $\gamma_i := \boldsymbol{c}_{y_i}^\top\boldsymbol{X}_i^\top\boldsymbol{s}_i$.

From proof of Lemma 5 (more specifically (12)), we get for all $t, t' \in \mathcal{R}_i$

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

Additionally, since $\frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_F} \neq \Theta\boldsymbol{W}^{\text{svm}}$, there exist $i \in [n], t \in \mathcal{O}_i, t' \in \bar{\mathcal{R}}_i$ such that $(\boldsymbol{x}_{it} - \boldsymbol{x}_{it'})^\top\boldsymbol{W}\bar{\boldsymbol{x}}_i < R\Theta$. Then,

$$\sum_{t\in\mathcal{O}_i} s_{it} = \frac{\sum_{t\in\mathcal{O}_i} e^{a_{it}+b_{it}}}{\sum_{t\in[T_i]} e^{a_{it}+b_{it}}} \leq \frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}} + e^{-R\Theta+b_{it'}}} \leq \frac{c_i}{d_i + e^{-R\Theta-\bar{b}}}, \quad \exists i \in [n]$$

$$\sum_{t\in\mathcal{O}_i} s_{it}^\star = \frac{\sum_{t\in\mathcal{O}_i} e^{a_{it}^\star+b_{it}}}{\sum_{t\in[T_i]} e^{a_{it}^\star+b_{it}}} \geq \frac{\sum_{t\in\mathcal{O}_i} e^{b_{it}}}{\sum_{t\in\mathcal{R}_i} e^{b_{it}} + \sum_{t\in\bar{\mathcal{R}}_i} e^{-(1+\pi)R\Theta+b_{it}}} \geq \frac{c_i}{d_i + Te^{-(1+\pi)R\Theta+\bar{b}}}, \quad \forall i \in [n],$$

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

$$\bar{\mathcal{L}}(\boldsymbol{W}) = \bar{\mathcal{L}}(\boldsymbol{W}^\parallel) = -\frac{1}{n}\sum_{i\in\mathcal{I}} \log\left(\frac{c_i}{d_i}\right).$$

Then we obtain

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}) - \bar{\mathcal{L}}(\boldsymbol{W}) &\geq -\frac{1}{n}\left(\log\left(\frac{c_i}{d_i + e^{-R\Theta-\bar{b}}}\right) - \log\left(\frac{c_i}{d_i}\right)\right)\\ &\geq \frac{1}{n}\log\left(1 + e^{-R\Theta-\bar{b}}/d_i\right)\end{aligned}$$

and let $j := \arg\max_{i\in[n]}\left(-\log\left(\sum_{t\in\mathcal{O}_i} s_{it}^\star\right) + \log\left(\frac{c_i}{d_i}\right)\right)$. We can upper-bound the loss difference for $(1+\pi)R\Theta\cdot\boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel$ as follows:

$$\begin{aligned}\mathcal{L}((1+\pi)R\Theta\cdot\boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel) - \bar{\mathcal{L}}(\boldsymbol{W}) &\leq \max_{i\in[n]}\left(-\log\left(\sum_{t\in\mathcal{O}_i} s_{it}^\star\right) + \log\left(\frac{c_i}{d_i}\right)\right)\\ &= \log\left(1 + Te^{-(1+\pi)R\Theta+\bar{b}}/d_j\right).\end{aligned}$$

Combining them together results in that, $\mathcal{L}(\boldsymbol{W}) > \mathcal{L}((1+\pi)R\Theta\cdot\boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel)$ whenever

$$\frac{1}{n}\log\left(1 + e^{-R\Theta-\bar{b}}/d_i\right) \geq \log\left(1 + Te^{-(1+\pi)R\Theta+\bar{b}}/d_j\right).$$

Given that $x/2 \le \log(1+x)$ for any $0 \le x \le 1$, we get $\mathcal{L}(\boldsymbol{W}) > \mathcal{L}((1+\pi)R\Theta \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^{\parallel})$ whenever

$$\frac{e^{-R\Theta-\bar{b}}}{2nd_i} \ge \frac{Te^{-(1+\pi)R\Theta+\bar{b}}}{d_j} \quad \text{when} \quad R \ge -\frac{\log(d_j)+\bar{b}}{\Theta}$$
$$\implies R > R_\pi := \max\left\{\frac{1}{\pi\Theta}\log\left(\frac{2nTd_i}{d_j}\right) + \frac{2\bar{b}}{\pi\Theta}, -\frac{\log(d_j)+\bar{b}}{\Theta}\right\}. \tag{31}$$

Here, since $\left\|\boldsymbol{W}^{\parallel}\right\|_F < \infty$, $d_i, d_j, \bar{b} < \infty$. ■

Proof of Theorem 2. Now gathering all the results so far, we are ready to prove the gradient descent convergence. The divergence of $\|\boldsymbol{W}(\tau)\|_F$ as $\tau \to \infty$ has been proven by Theorem 5.

• **We first show that $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau)) \to \boldsymbol{W}^{\text{fin}}$.** Lemma 6 has established the equivalence between $\bar{\boldsymbol{W}}^{\text{fin}}$ and $\boldsymbol{W}^{\text{fin}}$. Since $\mathcal{L}(\boldsymbol{W})$ is convex following Lemma 2, we have that $\mathcal{L}(\boldsymbol{W}(\tau)) \to \mathcal{L}_\star := \min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W})$. Additionally, Lemma 2 shows that $\mathcal{L}(\boldsymbol{W})$ is strictly convex on $\mathcal{S}_{\text{fin}}$ and Lemma 14 shows that $\boldsymbol{W}^{\text{fin}}$ is the unique finite solution. Suppose $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau)) \not\to \boldsymbol{W}^{\text{fin}}$. Let $\boldsymbol{W}$ be any matrix with $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}) \neq \boldsymbol{W}^{\text{fin}}$. Then Lemma 7 and Lemma 5 give that $\mathcal{L}(\boldsymbol{W}) \ge \bar{\mathcal{L}}(\boldsymbol{W}) = \bar{\mathcal{L}}(\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W})) > \mathcal{L}_\star$ where $\mathcal{L}_\star = \min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}) = \min_{\boldsymbol{W}} \bar{\mathcal{L}}(\boldsymbol{W})$. Given that $\mathcal{L}_\star$ is achievable, the proof is done by contradiction and we have that $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau)) \to \boldsymbol{W}^{\text{fin}}$.

• **We next prove that when $\boldsymbol{W}^{\text{svm}} = 0$, $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^\perp}(\boldsymbol{W}(\tau)) = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^\perp}(\boldsymbol{W}(0))$.** Recap the gradient in (18) where

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

Since $\boldsymbol{W}^{\text{svm}} = 0$ implies that for all $i \in [n]$, $\bar{\mathcal{R}}_i = \emptyset$, then $\bar{\mathcal{O}}_i \subseteq \mathcal{R}_i$. Additionally, since following definition of $\mathcal{S}_{\text{fin}}$ from Def. 1, for any $t \in \mathcal{R}_i$, $(\boldsymbol{x}_{it} - \boldsymbol{e}_{y_i})\bar{\boldsymbol{x}}_i^\top \in \mathcal{S}_{\text{fin}}$. Then we obtain

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

Therefore, $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^\perp}(\nabla\mathcal{L}(\boldsymbol{W})) = 0$ for any $\boldsymbol{W}$, which completes the proof.

• **Last, we show that when $\boldsymbol{W}^{\text{svm}} \neq 0$, $\boldsymbol{W}(\tau)/\|\boldsymbol{W}(\tau)\|_F \to \boldsymbol{W}^{\text{svm}}/\|\boldsymbol{W}^{\text{svm}}\|_F$.**

Consider any $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, and let $\boldsymbol{W}^\perp = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^\perp}(\boldsymbol{W})$, $\boldsymbol{W}^\parallel = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W})$, $R = \left\|\boldsymbol{W}^\perp\right\|_F$, and $\Theta = 1/\|\boldsymbol{W}^{\text{svm}}\|_F$. Next, from Lemma 12, for any $\tau \ge 0$, $\mathcal{L}(\boldsymbol{W}(\tau+1)) \le \mathcal{L}(\boldsymbol{W}(\tau))$. Let $\boldsymbol{W}^\perp(\tau) = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^\perp}(\boldsymbol{W}(\tau))$ and $\boldsymbol{W}^\parallel(\tau) = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau))$. Following Lemma 7, since loss satisfies $\ell(1) = -\log(1) = 0$, we obtain

$$\mathcal{L}(\boldsymbol{W}(\tau)) \ge \bar{\mathcal{L}}(\boldsymbol{W}^\parallel(\tau)).$$

Since following Lemma 14, the training risk $\bar{\mathcal{L}}(\boldsymbol{W}^\parallel(\tau))$ is infinite if $\left\|\boldsymbol{W}^\parallel(\tau)\right\|_F \to \infty$, which implies $\left\|\boldsymbol{W}^\parallel(\tau)\right\|_F < \infty$ for any $\tau \ge 0$. Additionally, Theorem 5 proves the divergence of $\boldsymbol{W}(\tau)$ as $\tau \to \infty$, hence we have $\left\|\boldsymbol{W}^\perp(\tau)\right\|_F \to \infty$.

Applying Lemma 15, as well as the fact that $\left\|\boldsymbol{W}^\parallel(\tau)\right\|_F < \infty$ and $\left\|\boldsymbol{W}^\perp(\tau)\right\|_F \to \infty$, there exists sufficiently large R_π as defined in (31) such that once $\left\|\boldsymbol{W}^\perp(\tau)\right\|_F = R > R_\pi$, $\mathcal{L}(\boldsymbol{W}(\tau)) - \mathcal{L}((1+\pi)R\Theta \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel(\tau)) \ge 0$. Since $\mathcal{L}(\boldsymbol{W})$ is convex following Lemma 2, we have that

$$\begin{aligned}
\mathcal{L}(\boldsymbol{W}) &\le \mathcal{L}((1+\pi)R\Theta \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel) + \left\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W} - \left((1+\pi)R\Theta \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel\right)\right\rangle \\
&= \mathcal{L}((1+\pi)R\Theta \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel) + \left\langle \nabla\mathcal{L}(\boldsymbol{W}), \left(\boldsymbol{W}^\perp - (1+\pi)R\Theta \cdot \boldsymbol{W}^{\text{svm}}\right)\right\rangle \\
&= \mathcal{L}((1+\pi)R\Theta \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel) + \left\langle \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^\perp}(\nabla\mathcal{L}(\boldsymbol{W})), \left(\boldsymbol{W}^\perp - (1+\pi)R\Theta \cdot \boldsymbol{W}^{\text{svm}}\right)\right\rangle.
\end{aligned} \tag{32}$$

Here, the first inequality uses the convexity of $\mathcal{L}(\boldsymbol{W})$ and last equation is obtained from the fact that $\boldsymbol{W}^{\text{svm}} \perp \mathcal{S}_{\text{fin}}$. It implies that once $\left\|\boldsymbol{W}^\perp(\tau)\right\|_F = R > R_\pi$,

$$\left\langle \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^\perp}(\nabla\mathcal{L}(\boldsymbol{W})), \left(\boldsymbol{W}^\perp - (1+\pi)R\Theta \cdot \boldsymbol{W}^{\text{svm}}\right)\right\rangle \ge 0.$$

Now we choose τ_0 such that for all $\tau \geq \tau_0$, $\left\|\boldsymbol{W}^{\perp}(\tau)\right\|_F > R_\pi$. Then for $\tau > \tau_0$, we get

$$\left\langle \boldsymbol{W}^{\perp}(\tau+1) - \boldsymbol{W}^{\perp}(\tau), \frac{\boldsymbol{W}^{\text{svm}}}{\|\boldsymbol{W}^{\text{svm}}\|_F} \right\rangle \geq \frac{1}{1+\pi} \left\langle \boldsymbol{W}^{\perp}(\tau+1) - \boldsymbol{W}^{\perp}(\tau), \frac{\boldsymbol{W}^{\perp}(\tau)}{\|\boldsymbol{W}^{\perp}(\tau)\|_F} \right\rangle \tag{33}$$

where

$$\begin{aligned}
&\left\langle \boldsymbol{W}^{\perp}(\tau+1) - \boldsymbol{W}^{\perp}(\tau), \frac{\boldsymbol{W}^{\perp}(\tau)}{\|\boldsymbol{W}^{\perp}(\tau)\|_F} \right\rangle \\
&= \frac{1}{2\|\boldsymbol{W}^{\perp}(\tau)\|_F} \left(\left\|\boldsymbol{W}^{\perp}(\tau+1)\right\|_F^2 - \left\|\boldsymbol{W}^{\perp}(\tau)\right\|_F^2 - \left\|\boldsymbol{W}^{\perp}(\tau+1) - \boldsymbol{W}^{\perp}(\tau)\right\|_F^2 \right) \\
&\geq \frac{\left\|\boldsymbol{W}^{\perp}(\tau+1)\right\|_F^2 - \left\|\boldsymbol{W}^{\perp}(\tau)\right\|_F^2}{2\|\boldsymbol{W}^{\perp}(\tau)\|_F} - \left\|\boldsymbol{W}^{\perp}(\tau+1) - \boldsymbol{W}^{\perp}(\tau)\right\|_F^2 &(34)\\
&\geq \left\|\boldsymbol{W}^{\perp}(\tau+1)\right\|_F - \left\|\boldsymbol{W}^{\perp}(\tau)\right\|_F - \left\|\boldsymbol{W}^{\perp}(\tau+1) - \boldsymbol{W}^{\perp}(\tau)\right\|_F^2 &(35)\\
&\geq \left\|\boldsymbol{W}^{\perp}(\tau+1)\right\|_F - \left\|\boldsymbol{W}^{\perp}(\tau)\right\|_F - \|\boldsymbol{W}(\tau+1) - \boldsymbol{W}(\tau)\|_F^2 &(36)\\
&\geq \left\|\boldsymbol{W}^{\perp}(\tau+1)\right\|_F - \left\|\boldsymbol{W}^{\perp}(\tau)\right\|_F + 2\eta\left(\mathcal{L}(\boldsymbol{W}(\tau+1)) - \mathcal{L}(\boldsymbol{W}(\tau))\right). &(37)
\end{aligned}$$

Here, (33) is obtained from (32) and holds for all $\tau > \tau_0$; (34) comes from the fact that $\left\|\boldsymbol{W}^{\perp}(\tau)\right\|_F > 0.5$; (35) follows that for any $a, b > 0$, $(a^2 - b^2)/2b > a - b$; (36) follows the projection property that $\|\boldsymbol{W}(\tau+1) - \boldsymbol{W}(\tau)\|_F^2 = \left\|\boldsymbol{W}^{\perp}(\tau+1) - \boldsymbol{W}^{\perp}(\tau)\right\|_F^2 + \left\|\boldsymbol{W}^{\parallel}(\tau+1) - \boldsymbol{W}^{\parallel}(\tau)\right\|_F^2$; and (37) is obtained via Lemma 12.

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

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

$$\implies \left\langle \frac{\boldsymbol{W}^{\perp}(\tau)}{\|\boldsymbol{W}^{\perp}(\tau)\|_F}, \frac{\boldsymbol{W}^{\text{svm}}}{\|\boldsymbol{W}^{\text{svm}}\|_F} \right\rangle \geq \frac{1}{1+\pi} \left(1 + \frac{C}{\|\boldsymbol{W}^{\perp}(\tau)\|_F} \right)$$

where

$$C := \left\langle \boldsymbol{W}^{\perp}(\tau_0), \frac{\boldsymbol{W}^{\text{svm}}}{\|\boldsymbol{W}^{\text{svm}}\|_F} \right\rangle - \left\|\boldsymbol{W}^{\perp}(\tau_0)\right\|_F + 2\eta\left(\mathcal{L}(\boldsymbol{W}(\tau)) - \mathcal{L}(\boldsymbol{W}(\tau_0))\right).$$

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

$$\lim_{\tau\to\infty} \left\langle \frac{\boldsymbol{W}^{\perp}(\tau)}{\|\boldsymbol{W}^{\perp}(\tau)\|_F}, \frac{\boldsymbol{W}^{\text{svm}}}{\|\boldsymbol{W}^{\text{svm}}\|_F} \right\rangle \geq \frac{1}{1+\pi}. \tag{38}$$

Choosing $\pi \to 0$ and combining (38) with the fact that $\lim_{\tau\to\infty} \left\|\boldsymbol{W}^{\parallel}(\tau)\right\|_F < \infty$ completes the proof.

■

C GLOBAL CONVERGENCE OF REGULARIZATION PATH

C.1 Proof of Theorem 3

Lemma 16 *Suppose Assumptions 1 and 2 hold. Additionally, assume loss $\ell : \mathbb{R} \to \mathbb{R}$ is strictly decreasing and $|\ell'|$ is bounded. Define $\bar{\boldsymbol{W}}_R^{\perp} := \bar{\boldsymbol{W}}_R^{\perp}(\boldsymbol{W}^{\parallel}) \in \mathcal{S}_{fin}^{\perp}$ by*

$$\bar{\boldsymbol{W}}_R^{\perp} := \arg \min_{\boldsymbol{W} \in \mathcal{S}_{fin}^{\perp}, \|\boldsymbol{W}\|_F \leq R} \mathcal{L}(\boldsymbol{W} + \boldsymbol{W}^{\parallel}). \tag{39}$$

Let $\boldsymbol{W}^{svm} \neq 0$ denote the solution of (Graph-SVM). Then we have that for any $\boldsymbol{W}^{\parallel} \in \mathcal{S}_{fin}$ with $\left\|\boldsymbol{W}^{\parallel}\right\|_F < \infty$

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

Proof. Recap $\mathcal{O}_i, \bar{\mathcal{O}}_i, \mathcal{R}_i, \bar{\mathcal{R}}_i$, $i \in [n]$ from (3). Let $\boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$ where following Assumption 1 we have

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

Recap $\mathcal{I}, \bar{\mathcal{I}}$ from (4). To proceed, define $\gamma_i^{\max}$ as follows:

- Consider $i \in \bar{\mathcal{I}}$. Given $\min_{\boldsymbol{W}} \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \bar{\boldsymbol{x}}_i)) \geq \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{e}_{y_i})$, 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 1 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 4, for any $\boldsymbol{W}^\perp \in \mathcal{S}_{\text{fin}}^\perp$ 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

$$\frac{s_{it}}{\sum_{t' \in \mathcal{R}_i} s_{it'}} = \frac{e^{b_{it} + \bar{a}_i}}{\sum_{t' \in \mathcal{R}_i} e^{b_{it'} + \bar{a}_i}} = \frac{e^{b_{it}}}{\sum_{t' \in \mathcal{R}_i} e^{b_{it'}}}, \tag{40}$$

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

  Combining both, we define the maximal score as follows:

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

  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 $\gamma_i^{\max} = \boldsymbol{c}_{y_i}^\top \bar{\boldsymbol{X}}_i^\top \mathbb{S}(\bar{\boldsymbol{X}}_i \boldsymbol{W}^\parallel \bar{\boldsymbol{x}}_i)$.

Hence, given $\boldsymbol{b}_i = \boldsymbol{X}_i \boldsymbol{W}^\parallel \bar{\boldsymbol{x}}_i$, we obtain

$$\gamma_i^{\max} = \begin{cases} 1, & i \in \bar{\mathcal{I}} \\ \frac{|\mathcal{O}_i| e^{\boldsymbol{e}_{y_i}^\top \boldsymbol{W}^\parallel \bar{\boldsymbol{x}}_i}}{\sum_{t' \in \mathcal{R}_i} e^{b_{it'}}}, & i \in \mathcal{I}. \end{cases}$$

Then we define the optimal risk of (39) 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

$$\gamma_i^{\max} = \boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \boldsymbol{s}_i^{\max} = \sum_{t \in \mathcal{O}_i} s_{it}^{\max} = \frac{\sum_{t \in \mathcal{O}_i} e^{b_{it}}}{\sum_{t \in \mathcal{R}_i} e^{b_{it}}} \quad \text{and} \quad \mathcal{L}_\star^{\boldsymbol{W}^\parallel} = \frac{1}{n} \sum_{i \in \bar{\mathcal{I}}} \ell(1) + \bar{\mathcal{L}}(\boldsymbol{W}^\parallel) \tag{41}$$

where $\bar{\mathcal{L}}(\boldsymbol{W}^\parallel)$ is the empirical risk over cyclic subdataset defined in Definition 4.

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

Step 1: We first show that $\lim_{R \to \infty} \mathcal{L}(R \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel) = \mathcal{L}_\star^{\boldsymbol{W}^\parallel}$. It can be easily proven using Lemma 7 and (41) by showing that for any $\boldsymbol{W}^\parallel$ with $\left\|\boldsymbol{W}^\parallel\right\|_F < \infty$,

$$\lim_{R \to \infty} \mathcal{L}(R \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^\parallel) = \min_{\boldsymbol{W} \in \mathcal{S}_{\text{fin}}^\perp} \mathcal{L}(\boldsymbol{W} + \boldsymbol{W}^\parallel) = \frac{|\bar{\mathcal{I}}|}{n} \ell(1) + \bar{\mathcal{L}}(\boldsymbol{W}^\parallel) = \mathcal{L}_\star^{\boldsymbol{W}^\parallel}.$$

Step 2: Next, we will prove that for any $\boldsymbol{W}^{\|} \in \mathcal{S}_{\text{fin}}$ with $\left\|\boldsymbol{W}^{\|}\right\|_F < \infty$, $\bar{\boldsymbol{W}}_R^{\perp}$ 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 $\boldsymbol{W} \in \mathcal{S}_{\text{fin}}^{\perp}$ with $\|\boldsymbol{W}\|_F < \infty$. Let $\boldsymbol{s}_i := \mathbb{S}(\boldsymbol{X}_i(\boldsymbol{W} + \boldsymbol{W}^{\|})\bar{\boldsymbol{x}}_i)$ 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 $\boldsymbol{W}^{\text{svm}} \neq 0$, then there exists some $i \in [n]$ such that $\gamma_i^{\boldsymbol{W}} < \gamma_i^{\max}$. Therefore, for any finite $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{W} \in \mathcal{S}_{\text{fin}}^{\perp}$, since loss function is strictly decreasing

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

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

$$\left\|\bar{\boldsymbol{W}}_R^{\perp} \cdot \left\|\boldsymbol{W}^{\text{svm}}\right\|_F / R - \boldsymbol{W}^{\text{svm}}\right\|_F \ge \delta.$$

Let $\boldsymbol{W}' = \bar{\boldsymbol{W}}_R^{\perp} \cdot \left\|\boldsymbol{W}^{\text{svm}}\right\|_F / R$ where we have $\|\boldsymbol{W}'\|_F \le \left\|\boldsymbol{W}^{\text{svm}}\right\|_F$ and $\boldsymbol{W}' \neq \boldsymbol{W}^{\text{svm}}$. Following Definition 1, we obtain

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

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 $\mathcal{L}(R \cdot \boldsymbol{W}^{\text{svm}}/\left\|\boldsymbol{W}^{\text{svm}}\right\|_F + \boldsymbol{W}^{\|}) < \mathcal{L}(\bar{\boldsymbol{W}}_R + \boldsymbol{W}^{\|})$ for sufficiently large R. Let $\Theta = 1/\left\|\boldsymbol{W}^{\text{svm}}\right\|_F$ and we will show that $\mathcal{L}(R\Theta \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^{\|}) < \mathcal{L}(R\Theta \cdot \boldsymbol{W}' + \boldsymbol{W}^{\|})$ for sufficiently large R.

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

Following Definition 1, we get for all $t, t' \in \mathcal{R}_i$

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

Then

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

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

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 $|\ell'|$ is bounded, let $c_{\text{dn}} \le -\ell' \le c_{\text{up}}$ for some constants $c_{\text{dn}}, c_{\text{up}} > 0$. Note that $c_{\text{dn}}, c_{\text{up}}$ are data-dependent. Then we have

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

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

$$\begin{aligned}
\mathcal{L}(R\Theta \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^{\|}) - \mathcal{L}_\star^{\boldsymbol{W}^{\|}} &\le \max_{i\in[n]} (\ell(\gamma_i^\star) - \ell(\gamma_i^{\max})) \le c_{\text{up}} \left(\gamma_j^{\max} - \gamma_j^\star\right) \\
&= c_{\text{up}} \left(\gamma_j^{\max} - \sum_{t\in\mathcal{O}_j} s_{it}^\star\right) \\
&\le c_{\text{up}} \gamma_j^{\max} \left(1 - \frac{1}{1 + Te^{-R\Theta+\bar{b}}/d_j}\right) \\
&\le \frac{c_{\text{up}} \gamma_j^{\max} T}{d_j} e^{-R\Theta+\bar{b}}.
\end{aligned}$$

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

$$\begin{aligned}
&\frac{c_{\text{dn}} \gamma_i^{\max}}{n} \left(1 - \frac{1}{1 + e^{-(1-\epsilon)R\Theta-\bar{b}}/d_i}\right) > \frac{c_{\text{up}} \gamma_j^{\max} T}{d_j} e^{-R\Theta+\bar{b}} \\
&\implies R > R_\epsilon := \frac{1}{\Theta \cdot \min(\epsilon, 1)} \log\left(\frac{2nTc_{\text{up}} \gamma_j^{\max} \cdot \max(d_i, 1)}{c_{\text{dn}} \gamma_i^{\max} d_j}\right) + \frac{2\bar{b}}{\Theta \cdot \min(\epsilon, 1)}.
\end{aligned} \tag{42}$$

Note that since $\boldsymbol{W}^{\|}$ is finite, b_{it} for all $i \in [n], t \in [T_i]$ are bounded and fixed, and therefore, $0 < d_i < \infty$, for all $i \in [n]$ and $\bar{b} < \infty$. (42) completes the proof by contradiction. ∎

Now, gathering all the results we have obtained so far, we are ready to prove Theorem 3.

Proof of Theorem 3. Recap the dataset $\texttt{DSET} = (\boldsymbol{X}_i, y_i)_{i=1}^n$ and index sets $\mathcal{O}_i, \bar{\mathcal{O}}_i, \mathcal{R}_i, \bar{\mathcal{R}}_i$, $i \in [n]$ from (3). Let $\boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$ denote the score vector of i-th input. Since Assumption 1 holds, then

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

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

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

Additionally, let $\boldsymbol{W}_R^\perp = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}^\perp}(\bar{\boldsymbol{W}}_R)$ and $\boldsymbol{W}_R^{\|} = \boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\bar{\boldsymbol{W}}_R)$. Lemma 16 has shown that for any finite $\lim_{R\to\infty} \boldsymbol{W}_R^{\|}$,

$$\lim_{R\to\infty} \frac{\bar{\boldsymbol{W}}_R}{R} = \lim_{R\to\infty} \frac{\boldsymbol{W}_R^\perp}{\sqrt{R^2 - \left\|\boldsymbol{W}_R^{\|}\right\|_F^2}} = \lim_{R\to\infty} \frac{\boldsymbol{W}_R^\perp}{\left\|\boldsymbol{W}_R^\perp\right\|_F} = \frac{\boldsymbol{W}^{\text{svm}}}{\|\boldsymbol{W}^{\text{svm}}\|_F}.$$

Therefore it remains to prove that $\lim_{R\to\infty} \boldsymbol{W}_R^{\|} \in \mathcal{W}^{\text{fin}}$.

Suppose $\lim_{R\to\infty} \boldsymbol{W}_R^{\|} := \boldsymbol{W}' \notin \mathcal{W}^{\text{fin}}$. Then for any $\boldsymbol{W}^{\|} \in \mathcal{W}^{\text{fin}}$, applying Lemma 7, we obtain

$$\min_{\boldsymbol{W}^\perp \in \mathcal{S}_{\text{fin}}^\perp} \mathcal{L}(\boldsymbol{W}^\perp + \boldsymbol{W}') > \min_{\boldsymbol{W}^\perp \in \mathcal{S}_{\text{fin}}^\perp} \mathcal{L}(\boldsymbol{W}^\perp + \boldsymbol{W}^{\|}) = \lim_{R\to\infty} \mathcal{L}(R \cdot \boldsymbol{W}^{\text{svm}} + \boldsymbol{W}^{\|}).$$

Therefore, $\boldsymbol{W}'$ does not achieve the minimal loss as $R \to \infty$. ∎

C.2 Proof of Lemma 3

Proof. Let $\boldsymbol{\gamma}_i = \boldsymbol{X}_i \boldsymbol{c}_{y_i}$ denote the score vector of i-th 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 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 $i \in [n]$

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

Since loss function ℓ is strictly decreasing, we get

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

Let $\boldsymbol{W}$ be any attention weight satisfying all the "≥ 1" constraints in (Acyc-SVM). We next prove that $\lim_{R\to\infty} \mathcal{L}(R\cdot\boldsymbol{W}) = \mathcal{L}_\star$. Recap $\mathcal{O}_i$ and $\bar{\mathcal{O}}_i$ from (3). 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, $\boldsymbol{W}$ separates tokens $\boldsymbol{e}_{y_i}$ from the rest of the tokens within $\boldsymbol{X}_i$. Then $\lim_{R\to\infty} \mathbb{S}(\boldsymbol{X}_i(R\cdot\boldsymbol{W})\bar{\boldsymbol{x}}_i)$ will output $1/|\mathcal{O}_i|$ for $t \in \mathcal{O}_i$ and zero for the left. Specifically, let $\boldsymbol{s}_i^R := \mathbb{S}(\boldsymbol{X}_i(R\cdot\boldsymbol{W})\bar{\boldsymbol{x}}_i)$, and following the SVM objective (Acyc-SVM) for any $i \in [n]$, we get

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

$$\text{and then,} \quad \sum_{t\in\mathcal{O}_i} s_{it}^R = \frac{|\mathcal{O}_i|}{|\mathcal{O}_i| + \sum_{t\in\bar{\mathcal{O}}_i} e^{(\boldsymbol{x}_{it}-\boldsymbol{e}_{y_i})^\top(R\cdot\boldsymbol{W})\bar{\boldsymbol{x}}_i}} \geq \frac{1}{1+e^{-R}}.$$

Then $\lim_{R\to\infty}\sum_{t\in\mathcal{O}_i} s_{it}^R = 1$ and therefore,

$$\lim_{R\to\infty} s_{it}^R = 1/|\mathcal{O}_i| \quad \text{for } t \in \mathcal{O}_i, \qquad \text{and} \qquad \lim_{R\to\infty} s_{it}^R = 0 \quad \text{for } t \in \bar{\mathcal{O}}_i.$$

Hence we have

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

Since $|\ell'|$ is bounded, then $\lim_{R\to\infty} \mathcal{L}(R\cdot\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{c}_{y_i}^\top \boldsymbol{e}_{y_i}) = \frac{1}{n}\sum_{i=1}^n \ell(\gamma_i^{\max}) = \mathcal{L}_\star$. ■

C.3 Proof of Theorem 4

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 3, 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 3, we have that for any finite $\boldsymbol{W}$, $\mathcal{L}(\boldsymbol{W}) < \lim_{R\to\infty}\mathcal{L}(R\cdot\boldsymbol{W}^{\text{svm}}) = \mathcal{L}^\star$. 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 $\boldsymbol{W}^{\text{svm}}$.

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

$$\left\|\bar{\boldsymbol{W}}_R \cdot \|\boldsymbol{W}^{\text{svm}}\|_F/R - \boldsymbol{W}^{\text{svm}}\right\|_F \geq \delta.$$

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

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

Now, we will argue that this leads to a contradiction by proving $\mathcal{L}(R \cdot \boldsymbol{W}^{\mathrm{svm}} / \|\boldsymbol{W}^{\mathrm{svm}}\|_F) < \mathcal{L}(\bar{\boldsymbol{W}}_R)$ for sufficiently large R. Let $\Theta = 1/\|\boldsymbol{W}^{\mathrm{svm}}\|_F$ and we will show that $\mathcal{L}(R\Theta \cdot \boldsymbol{W}^{\mathrm{svm}}) < \mathcal{L}(R\Theta \cdot \boldsymbol{W}')$ for sufficiently large R.

To obtain the result, we establish a refined softmax probability control as in the proof of Theorem 3 by studying the distance to $\mathcal{L}_\star$. Let $\boldsymbol{a}_i^\star = \boldsymbol{X}_i(R\Theta \cdot \boldsymbol{W}^{\mathrm{svm}})\bar{\boldsymbol{x}}_i$, $\boldsymbol{a}_i^R = \boldsymbol{X}_i(R\Theta \cdot \boldsymbol{W}')\bar{\boldsymbol{x}}_i$, $\boldsymbol{s}_i^\star := \mathbb{S}(\boldsymbol{a}_i^\star)$, $\boldsymbol{s}_i^R := \mathbb{S}(\boldsymbol{a}_i^R)$, $\gamma_i^\star := \boldsymbol{c}_{y_i}^\top \boldsymbol{X}_i^\top \boldsymbol{s}_i^\star$, 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

$$\sum_{t \in \mathcal{O}_i} s_{it}^R = \frac{\sum_{t \in \mathcal{O}_i} e^{a_{it}^R}}{\sum_{t \in [T_i]} e^{a_{it}^R}} \le \frac{|\mathcal{O}_i|}{|\mathcal{O}_i| + e^{-(1-\epsilon)R\Theta}} \le \frac{1}{1 + e^{-(1-\epsilon)R\Theta}/T}, \quad \exists i \in [n] \tag{43}$$

$$\sum_{t \in \mathcal{O}_i} s_{it}^\star = \frac{\sum_{t \in \mathcal{O}_i} e^{a_{it}^\star}}{\sum_{t \in [T_i]} e^{a_{it}^\star}} \ge \frac{|\mathcal{O}_i|}{|\mathcal{O}_i| + (T - |\mathcal{O}_i|)e^{-R\Theta}} \ge \frac{1}{1 + Te^{-R\Theta}}, \quad \forall i \in [n]. \tag{44}$$

Since ℓ is strictly decreasing and $|\ell'|$ is bounded, let $c_{\mathrm{dn}} \le -\ell' \le c_{\mathrm{up}}$ for some constants $c_{\mathrm{dn}}, c_{\mathrm{up}} > 0$. Note that $c_{\mathrm{dn}}, c_{\mathrm{up}}$ are data-dependent. Additionally, define the score minimal/maximal score gaps as

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

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

$$\begin{aligned} \mathcal{L}(R\Theta \cdot \boldsymbol{W}') - \mathcal{L}_\star &\ge \frac{1}{n}\left(\ell(\gamma_i^R) - \ell(\gamma_i^{\max})\right) \ge \frac{c_{\mathrm{dn}}}{n}\left(\gamma_i^{\max} - \gamma_i^R\right) \\ &\ge \frac{c_{\mathrm{dn}}}{n} c_{\min}\left(1 - \sum_{t \in \mathcal{O}_i} s_{it}^R\right) \ge \frac{c_{\mathrm{dn}} c_{\min}}{n} \frac{1}{1 + Te^{(1-\epsilon)R\Theta}} \end{aligned} \tag{45}$$

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 $R\Theta \cdot \boldsymbol{W}^{\mathrm{svm}}$ as follows:

$$\begin{aligned} \mathcal{L}(R\Theta \cdot \boldsymbol{W}^{\mathrm{svm}}) - \mathcal{L}_\star &\le \max_{i \in [n]} \left(\ell(\gamma_i^\star) - \ell(\gamma_i^{\max})\right) \le c_{\mathrm{up}}\left(\gamma_j^{\max} - \gamma_j^\star\right) \\ &\le c_{\mathrm{up}} c_{\max}\left(1 - \sum_{t \in \mathcal{O}_i} s_{it}^\star\right) \le c_{\mathrm{up}} c_{\max} \frac{1}{1 + e^{R\Theta}/T} \le c_{\mathrm{up}} c_{\max} T e^{-R\Theta}. \end{aligned} \tag{46}$$

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

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

This completes the proof by contradiction. ∎

D IMPLEMENTATION DETAILS AND ADDITIONAL EXPERIMENTS

D.1 Implementation Details

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 TPGs, 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, $\boldsymbol{e}_{y_i}$ are also randomly sampled from $\boldsymbol{X}_i$.

3. Construct TPGs and apply Tarjan's algorithm Tarjan (1972) to find SCCs of each TPG. For global convergence experiments (Section 3), TPGs are created based on the token relations between $\boldsymbol{x}_{it}$'s and $\boldsymbol{e}_{y_i}$'s in the dataset $\texttt{DSET}$; while for local convergence analysis (Section 5), we instead establish the token relations between $\boldsymbol{x}_{it}$'s and $\hat{\boldsymbol{e}}_{y_i}$'s following the instruction in Section 5, where $\hat{\boldsymbol{e}}_{y_i}$ is determined by the GD solution.

4. $\overline{\texttt{DSET}}$ is created following Definition 3 based on the SCCs of the corresponding TPGs.

Here, we set the sequence length to be the same for all the samples in $\texttt{DSET}$, and we emphasize that though $\texttt{DSET}$ 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 TPGs and SCCs, and $\overline{\texttt{DSET}}$ may contain inputs with varying sequence lengths (see Figure 3).

• **Generating $\boldsymbol{W}^{\text{fin}}$ and $\widetilde{\boldsymbol{W}}^{\text{fin}}$.** Inspired by the convexity and finiteness of $\bar{\mathcal{L}}(\boldsymbol{W})$ per Definition 4 under the setting of Theorem 2, we can derive $\boldsymbol{W}^{\text{fin}}$ via gradient descent. Hence, to obtain $\boldsymbol{W}^{\text{fin}}$, 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 5, we follow the same method as generating $\boldsymbol{W}^{\text{fin}}$. However, we emphasize that under the local convergence setting, there is no guarantee that gradient descent will converge to the $\widetilde{\boldsymbol{W}}^{\text{fin}}$ 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 6 and 7 indicate that $\widetilde{\boldsymbol{W}}^{\text{fin}}$ can predict the GD convergence performance better than $\boldsymbol{W}^{\text{fin}}$ which is drawn from the dataset-based TPGs. We defer a rigorous definition of local $\widetilde{\boldsymbol{W}}^{\text{fin}}$ and guarantees related to gradient descent for future exploration.

• **Local convergence experiments (Figures 6 and 7).** 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 6 and cross-entropy loss in Figure 7. 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. 6a and 7a and the distance of $\left\|\boldsymbol{\Pi}_{\widetilde{\mathcal{S}}_{\text{fin}}}(\boldsymbol{W}(\tau)) - \widetilde{\boldsymbol{W}}^{\text{fin}}\right\|_F$ are presented in the orange curves in Figs. 6b and 7b. In both experiments, correlations between $\frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F}$ and $\frac{\widetilde{\boldsymbol{W}}^{\text{svm}}}{\|\widetilde{\boldsymbol{W}}^{\text{svm}}\|_F}$ end with > 0.99 values. Fig. 6b 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. 7b, 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 $\boldsymbol{\Pi}_{\widetilde{\mathcal{S}}_{\text{fin}}}(\boldsymbol{W}(\tau))$ is much closer to $\widetilde{\boldsymbol{W}}^{\text{fin}}$ compared to the distance between $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau))$ and $\boldsymbol{W}^{\text{fin}}$.

D.2 Additional Experiments

Global convergence experiments on large K (Figures 9 and 10). Assumption 1 in our work requires $K \leq d$, which helps make the optimization landscape more benign such that global convergence of GD is guaranteed. Unlike previous work Tarzanagh et al. (2023b,a) that relies on strong equal score conditions to induce global convergence, our assumption is much less strict. Empirically, we argue that this constraint is not necessary as we can apply a mask $\boldsymbol{M} \in \mathbb{R}^{n\times K\times T}$ to directly collect the attention probability for each distinct token from the attention map without explicitly calculating the linear head. Therefore, we can still impose Assumption 1 when $K > d$, which aligns more closely with the real-world setting. In Figs. 9 and 10, we repeat the global convergence experiments by setting $n = 16, T = 64, d = 128$ and $K = 10000$. Results are averaged over 100 random instances. The averaged correlation is ≈ 0.987 and the soft component error reaches 0.025. The results again validate Theorem 2.

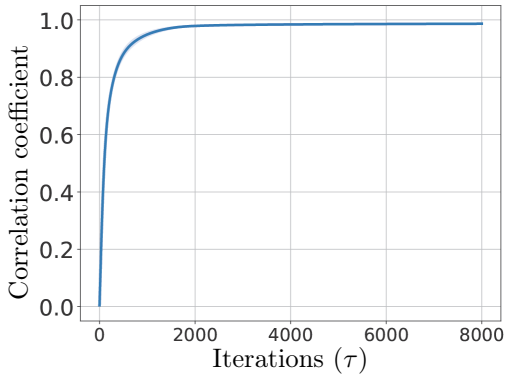


Figure 9: $\frac{\boldsymbol{W}(\tau)}{\|\boldsymbol{W}(\tau)\|_F} \to \frac{\boldsymbol{W}^{\text{svm}}}{\|\boldsymbol{W}^{\text{svm}}\|_F}$

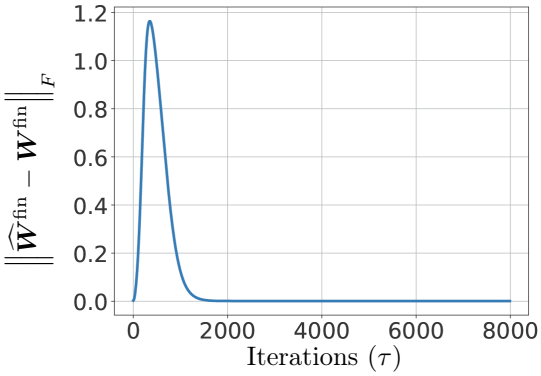


Figure 10: $\boldsymbol{\Pi}_{\mathcal{S}_{\text{fin}}}(\boldsymbol{W}(\tau)) \to \boldsymbol{W}^{\text{fin}}$

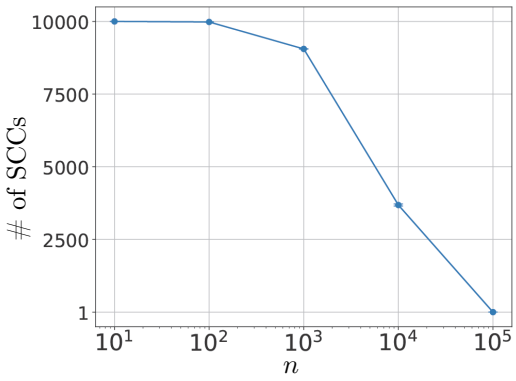


Figure 11: Number of of SCCs vs n

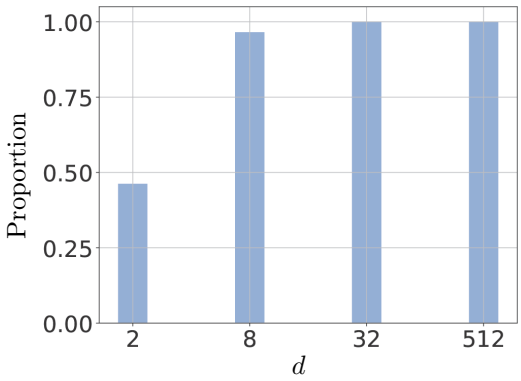


Figure 12: Feasibility of $\boldsymbol{W}^{\text{svm}}$

SCC Structure on large n (Figure 11). When we increase the sample size n and fix others, more edges and cycles will be added to the graphs. Different SCCs can then be merged into one. As illustrated in Fig. 11, the graph eventually collapses to a single SCC.

Feasible condition of (Graph-SVM) (Figure 12). To verify that (Graph-SVM) is feasible when $d \geq K$ (Lemma 1), in Fig. 12, we run experiments with fixed $n = 16, T = 128, K = 512$ and varying d from 2 to 512. Define $\mathcal{C}_y$ as the SCC that the label token belongs to. We calculate the proportion of selected tokens that are in $\mathcal{C}_y$ to the size of $\mathcal{C}_y$, and (Graph-SVM) is feasible when the value reaches 1. The interpretation is that: When d is small, the problem focuses on separating an optimally feasible subset of training data from the others and the empirical SVM bias is captured by a relaxed Graph-SVM solution with constraints based on the subset. As d grows, the exact Graph-SVM becomes feasible. This is similar to the findings in Ji & Telgarsky (2019b).