Transformers as Support Vector Machines

Anonymous Author(s)
Affiliation
Address
email

Abstract

The transformer architecture has led to revolutionary advancements in NLP. The attention layer within the transformer admits a sequence of input tokens $\boldsymbol{X}$ and makes them interact through pairwise similarities computed as $\texttt{softmax}(\boldsymbol{X}\boldsymbol{Q}\boldsymbol{K}^\top\boldsymbol{X}^\top)$, where $(\boldsymbol{K}, \boldsymbol{Q})$ are the trainable key-query parameters. In this work, we establish a formal equivalence between the optimization geometry of self-attention and a hard-margin SVM problem that separates optimal input tokens from non-optimal tokens using linear constraints on the outer-products of token pairs. This formalism allows us to characterize the implicit bias of 1-layer transformers optimized with gradient descent: **(1)** Optimizing the attention layer, parameterized by $(\boldsymbol{K}, \boldsymbol{Q})$, with vanishing regularization, converges in direction to an SVM solution minimizing the nuclear norm of the combined parameter $\boldsymbol{W} := \boldsymbol{K}\boldsymbol{Q}^\top$. Instead, directly parameterizing by $\boldsymbol{W}$ minimizes a Frobenius norm SVM objective. **(2)** Complementing this, for $\boldsymbol{W}$-parameterization, we prove the local/global directional convergence of gradient descent under suitable geometric conditions, and propose a more general SVM equivalence that predicts the implicit bias of attention with nonlinear heads/MLPs.

1 Introduction

Self-attention, the central component of the transformer architecture, has revolutionized NLP [VSP+17]. This mechanism has proven highly effective in capturing long-range dependencies, which is essential for applications arising in NLP [KT19, BMR+20, RSR+20], computer vision [FXM+21, LLC+21, TCD+21, CSL+23], and reinforcement learning [JLL21, CLR+21, WWX+22]. Remarkable success of the self-attention mechanism and transformers has paved the way for the development of LLMs such as GPT4 [Ope23], Bard [Goo23], LLaMA [TLI+23], and ChatGPT [Ope22].

Q: Can we characterize the optimization landscape and implicit bias of transformers?

We address this question by rigorously connecting the optimization geometry of the attention layer and a hard max-margin SVM problem, namely (Att-SVM), that separates and selects the optimal tokens from each input sequence. This formalism follows [TLZO23], which sheds light on the intricacies of self-attention. Throughout, given input sequences $\boldsymbol{X}, \boldsymbol{Z} \in \mathbb{R}^{T\times d}$ with length T and embedding dimension d, we study the core cross-attention and self-attention models:

$$f_{\texttt{cross}}(\boldsymbol{X}, \boldsymbol{Z}) := \mathbb{S}(\boldsymbol{Z}\boldsymbol{Q}\boldsymbol{K}^\top\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{V}, \quad f_{\texttt{self}}(\boldsymbol{X}) := \mathbb{S}(\boldsymbol{X}\boldsymbol{Q}\boldsymbol{K}^\top\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{V}.$$

Here, $\boldsymbol{K}, \boldsymbol{Q} \in \mathbb{R}^{d\times m}$, $\boldsymbol{V} \in \mathbb{R}^{d\times v}$ are the trainable key, query, value matrices respectively; $\mathbb{S}(\cdot)$ denotes the softmax nonlinearity. Note that self-attention is a special instance of the cross-attention by setting $\boldsymbol{Z} \leftarrow \boldsymbol{X}$. To expose our main results, suppose the first token of $\boldsymbol{Z}$, denoted by $\boldsymbol{z}$, is used for prediction. Concretely, given a dataset $(Y_i, \boldsymbol{X}_i, \boldsymbol{z}_i)_{i=1}^n$ with labels $Y_i \in \{-1, 1\}$ and inputs $\boldsymbol{X}_i \in \mathbb{R}^{T\times d}$, $\boldsymbol{z}_i \in \mathbb{R}^d$, we consider the empirical risk minimization with a loss $\ell(\cdot) : \mathbb{R} \to \mathbb{R}$, defined as follows:

$$\mathcal{L}(\boldsymbol{K}, \boldsymbol{Q}) = \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i \cdot f(\boldsymbol{X}_i, \boldsymbol{z}_i)\right), \quad \text{where } f(\boldsymbol{X}_i, \boldsymbol{z}_i) = h\left(\boldsymbol{X}_i^\top \mathbb{S}\left(\boldsymbol{X}_i\boldsymbol{K}\boldsymbol{Q}^\top\boldsymbol{z}_i\right)\right). \tag{1}$$

Submitted to the Mathematics of Modern Machine Learning Workshop at NeurIPS 2023. Do not distribute.

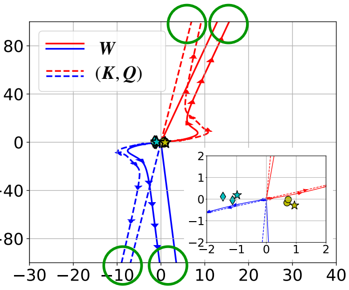


Figure 1: GD convergence of attention weights. Markers represent tokens; lines depict attention-SVM directions mapped to z; arrows illustrate GD paths converging towards these SVM directions. Green circles denote GD $\leftrightarrow$ SVM pairings.

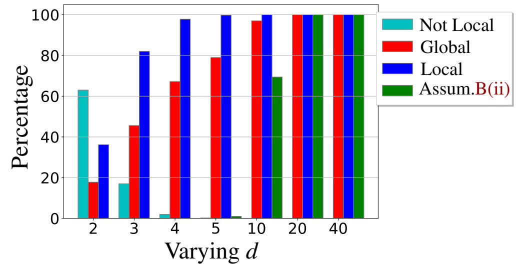


Figure 2: Percentage of different convergence types when training $\boldsymbol{W}$. Red and blue bars represent the percentages of convergence to globally and locally-optimal SVM solutions; teal are complements of the blue; green depict Assum. B(ii).

Here, $h(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x}$ is the linear prediction head and $f(\cdot)$ precisely represents a one-layer transformer. The softmax operation, due to its nonlinear nature, poses a significant challenge when optimizing (1). In this study, we focus on optimizing the attention weights ($\boldsymbol{K}, \boldsymbol{Q}$ or $\boldsymbol{W}$) and overcome such challenges to establish a fundamental SVM equivalence. The paper's main contributions are as follows:

- **Implicit bias of the attention layer (Sec. 2).** Optimizing the attention parameters $\boldsymbol{W}$ or $(\boldsymbol{K}, \boldsymbol{Q})$ with vanishing regularization converges in direction towards a solution of (Att-SVM) or (Att-SVM$_\star$) with the Frobenius norm or the nuclear norm objective, respectively. To our knowledge, this is the first result to formally distinguish the optimization dynamics of $\boldsymbol{W}$ vs $(\boldsymbol{K}, \boldsymbol{Q})$ parameterizations, revealing the low-rank bias of the latter.
- **Convergence of gradient descent (Sec. 3).** We prove the local/global directional convergence of gradient descent for optimizing the attention layer parameterized by $\boldsymbol{W}$ under suitable geometric conditions. Beyond these, we propose a more general SVM equivalence with nonlinear head, which predicts the implicit bias of attention trained by gradient descent.

1.1 Preliminaries

Optimization algorithms. Given a parameter $R > 0$, we define the regularized path solution as (W-RP) and (KQ-RP). For GD, with appropriate $\eta > 0$, we describe the optimization process as (W-GD) and (KQ-GD). Here for (W-RP) and (W-GD), $\mathcal{L}(\boldsymbol{Q}, \boldsymbol{K})$ is replaced with $\mathcal{L}(\boldsymbol{W})$ with $\boldsymbol{W} := \boldsymbol{K}\boldsymbol{Q}^\top$.

Given $\boldsymbol{W}(0) \in \mathbb{R}^{d\times d}$, $\eta > 0$, for $k \geq 0$ do:

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

Given $\boldsymbol{Q}(0), \boldsymbol{K}(0) \in \mathbb{R}^{d\times m}$, $\eta > 0$, for $k \geq 0$ do:

$$\begin{bmatrix} \boldsymbol{K}(k+1) \\ \boldsymbol{Q}(k+1) \end{bmatrix} = \begin{bmatrix} \boldsymbol{K}(k) \\ \boldsymbol{Q}(k) \end{bmatrix} - \eta \begin{bmatrix} \nabla_{\boldsymbol{K}} \mathcal{L}(\boldsymbol{K}(k), \boldsymbol{Q}(k)) \\ \nabla_{\boldsymbol{Q}} \mathcal{L}(\boldsymbol{K}(k), \boldsymbol{Q}(k)) \end{bmatrix}. \quad \text{(KQ-GD)}$$

Given $R > 0$, find $d \times d$ matrix:

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

Given $R > 0$, find $d \times m$ matrices:

$$(\bar{\boldsymbol{K}}_R, \bar{\boldsymbol{Q}}_R) = \underset{\|\boldsymbol{K}\|_F^2 + \|\boldsymbol{Q}\|_F^2 \leq 2R}{\arg\min}\ \mathcal{L}(\boldsymbol{K}, \boldsymbol{Q}). \quad \text{(KQ-RP)}$$

Definition 1 (Token Score and Optimality) *Given a prediction head $\boldsymbol{v} \in \mathbb{R}^d$, the score of a token $\boldsymbol{x}_{it}$ of input $\boldsymbol{X}_i$ is defined as $\boldsymbol{\gamma}_{it} = Y_i \cdot \boldsymbol{v}^\top \boldsymbol{x}_{it}$. The optimal token for each input $\boldsymbol{X}_i$ is given by the index $\texttt{opt}_i \in \arg\max_{t\in[T]} \boldsymbol{\gamma}_{it}$ for all $i \in [n]$.*

By introducing token scores and identifying optimal tokens, we can better understand the importance of individual tokens and their impact on the overall objective. Next, we present SVM problems.

- **Hard-margin SVM for $\boldsymbol{W}$-parameterization.** Equipped with the set of optimal indices $(\texttt{opt}_i)_{i=1}^n$ as per Definition 1, we introduce the following SVM formulation associated to $\boldsymbol{W}$-parameterization:

$$\boldsymbol{W}^{mm} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \quad \text{s.t.} \quad (\boldsymbol{x}_{i\texttt{opt}_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W} \boldsymbol{z}_i \geq 1 \quad \text{for all } t \neq \texttt{opt}_i,\ i \in [n]. \quad \text{(Att-SVM)}$$

Throughout, we assume the SVM problems are feasible. We also note that GD can provably converge to an SVM solution over locally-optimal tokens, as detailed in Section 3.2.

• **SVM problem for $(\boldsymbol{K},\boldsymbol{Q})$-parameterization.** The objective function has an extra layer of nonconvexity as $(\boldsymbol{K},\boldsymbol{Q})$ corresponds to a matrix factorization of $\boldsymbol{W}$. Fortunately, our experiments reveal that GD is indeed biased towards the global minima. This yields the following $\boldsymbol{W}$-parameterized SVM with nuclear norm objective:

$$\boxed{\boldsymbol{W}_\star^{mm} \in \underset{\mathrm{rank}(\boldsymbol{W})\le m}{\arg\min} \|\boldsymbol{W}\|_\star \quad \text{s.t.} \quad (\boldsymbol{x}_{i\mathrm{opt}_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W} z_i \ge 1 \quad \text{for all } t \ne \mathtt{opt}_i,\ i \in [n]. \quad \text{(Att-SVM}_\star)}$$

Above, the nonconvex rank constraint arises from the fact that the rank of $\boldsymbol{W} = \boldsymbol{K}\boldsymbol{Q}^\top$ is at most m. Lemma 1, presented below, demonstrates that this guarantee holds whenever $n \le m$.

Lemma 1 *Any optimal solution of* (Att-SVM) *or* (Att-SVM$_\star$) *is at most rank n. More precisely, the row space of* $\boldsymbol{W}^{mm}$ *or* $\boldsymbol{W}_\star^{mm}$ *lies within* $\mathtt{span}(\{z_i\}_{i=1}^n)$.

2 Understanding Implicit Bias of Self-Attention

We start by establishing the global convergence of regulrized paths.

Assumption A *Over any bounded interval* $[a,b]$*: (i)* $\ell : \mathbb{R} \to \mathbb{R}$ *is strictly decreasing; (ii) The derivative* ℓ' *is bounded as* $|\ell'(u)| \le M_1$*; (iii)* ℓ' *is* M_0*-Lipschitz continuous.*

Theorem 1 *Suppose Assumption A holds, optimal indices* $(\mathtt{opt}_i)_{i=1}^n$ *are unique. Let* $\boldsymbol{W}^{mm}$ *be the unique solution of* (Att-SVM)*, and let* $\mathcal{W}_\star^{mm}$ *be the solution set of* (Att-SVM$_\star$) *with nuclear norm achieving objective* $C_\star$*. Then, Algorithms W-RP and KQ-RP, respectively, satisfy:*

- $\boldsymbol{W}$*-parameterization has Frobenius norm bias:* $\lim_{R\to\infty} \frac{\bar{\boldsymbol{W}}_R}{R} = \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F}$.
- $(\boldsymbol{K},\boldsymbol{Q})$*-parameterization has nuclear norm bias:* $\lim_{R\to\infty} \mathtt{dist}\left(\frac{\bar{\boldsymbol{K}}_R\bar{\boldsymbol{Q}}_R^\top}{R}, \frac{\mathcal{W}_\star^{mm}}{C_\star}\right) = 0$.

Theorem 1 shows that the RP of the $\boldsymbol{W}$ and $(\boldsymbol{K},\boldsymbol{Q})$-parameterization converge to the max-margin solutions of (Att-SVM) and (Att-SVM$_\star$) with Frobenius and nuclear norm objectives, respectively. This result is the first to distinguish the optimization dynamics of $\boldsymbol{W}$ and $(\boldsymbol{K},\boldsymbol{Q})$ parameterizations, revealing the low-rank bias of the latter. To study the RP theory predictivity of the implicit bias exhibited by GD, we examine the GD paths in Figure 1, where $n = d = 2$, $T = 3$. The teal and yellow markers correspond to tokens from $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and the stars indicate the optimal tokens. We illustrate the iterations of the attention weight in the form of $\boldsymbol{W}z_i$ and $\boldsymbol{K}\boldsymbol{Q}^\top z_i$, $i = 1,2$. The red/blue solid lines delineate the directions of $\boldsymbol{W}^{mm}z_1/\boldsymbol{W}^{mm}z_2$; the red/blue dashed lines show the directions of $\boldsymbol{W}_\star^{mm}z_1/\boldsymbol{W}_\star^{mm}z_2$; the arrows denote the corresponding directions of gradient evolution. Figure 1 provides a clear depiction of the incremental alignment of $\boldsymbol{W}(k)$ and $\boldsymbol{K}(k)\boldsymbol{Q}(k)^\top$ with their respective attention SVM solutions as k increases. This strongly supports the assertions of Theorem 1.

3 Convergence and Implicit Bias of Gradient Descent

3.1 Global convergence

In this section, we will establish conditions that guarantee the global convergence of GD.

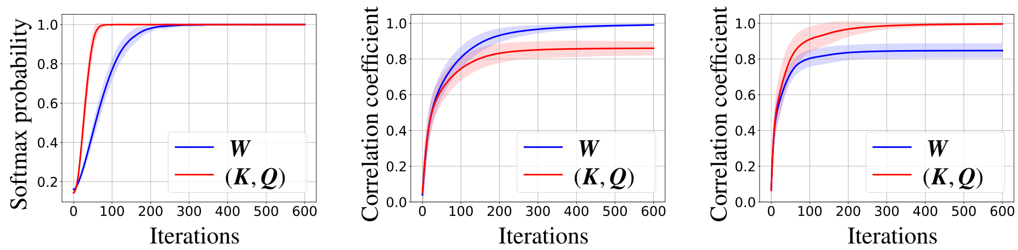
Lemma 2 *Under Assumption A,* $\nabla\mathcal{L}(\boldsymbol{W})$ *is* $L_{\boldsymbol{W}}$*-Lipschitz continuous , where* $L_{\boldsymbol{W}} := \frac{1}{n}\sum_{i=1}^n a_i b_i$*, and* $a_i = \|\boldsymbol{v}\|\,\|z_i\|^2\|\boldsymbol{X}_i\|^3$, $b_i = M_0\|\boldsymbol{v}\|\,\|\boldsymbol{X}_i\| + 3M_1$ *for all* $i \in [n]$.

Assumption B *Optimal tokens' indices* $(\mathtt{opt}_i)_{i=1}^n$ *are unique and one of the following conditions on the tokens holds: For all* $t \ne \mathtt{opt}_i$ *and* $i \in [n]$*, (i) the tokens' scores, as defined in Def. 1, satisfy* $\gamma_{it} = \gamma_{i\tau} < \gamma_{i\mathtt{opt}_i}$*. (ii) all tokens are support vectors, i.e.,* $(\boldsymbol{x}_{i\mathtt{opt}_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W}^{mm} z_i = 1$*;*

Here, we provide conditions for achieving global convergence towards the max-margin direction $\boldsymbol{W}^{mm}$ based on token score constraints and over-parameterization. For the former, we provide precise theoretical guarantees. For the latter, we provide strong empirical evidence.

(I) Global convergence under score constraints. Our next result establishes the global convergence of GD to the max-margin direction $\boldsymbol{W}^{mm}$ under Assumption B(i) that non-optimal tokens have identical scores but lower than the score of the optimal token.

Theorem 2 *Suppose Assumption A on the loss* ℓ *and Assumption B(i) on the tokens' score hold. Then, Algorithm W-GD with* $\eta \le 1/L_{\boldsymbol{W}}$ *and any starting point* $\boldsymbol{W}(0)$ *satisfies* $\lim_{k\to\infty} \frac{\boldsymbol{W}(k)}{\|\boldsymbol{W}(k)\|_F} = \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F}$.



(a) Evolution of softmax prob. (b) Corr. coeff. of GD and W^{mm}_{α} (c) Corr. coeff. of GD and $W^{mm}_{\star,\alpha}$

Figure 3: Local convergence behaviour of GD when training $\boldsymbol{W}$ or $(\boldsymbol{K}, \boldsymbol{Q})$ with random data.

(II) Global convergence via overparameterization. Considering that Assumption B(ii) is anticipated to hold as the dimension d increases, the norm of the GD solution is bound to diverge to infinity. This satisfies a prerequisite for converging towards the globally-optimal SVM direction $\boldsymbol{W}^{mm}$. The trend depicted in Figure 2, where the percentage of global convergence (red bars) approaches 100% and Assumption B(ii) holds with higher probability (green bars) as d grows, reinforces this insight.

3.2 Local convergence

Definition 2 (Local Optimality) *Fix token indices* $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$. *Solve* (Att-SVM) *with* $(\texttt{opt}_i)_{i=1}^n$ *replaced with* $\boldsymbol{\alpha}$ *to obtain* $\boldsymbol{W}^{mm}_{\alpha}$. *Consider the set* $\mathcal{T}_i \subset [T]$ *such that* $(\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W}^{mm}_{\alpha} \boldsymbol{z}_i = 1$. *If for all* $i \in [n]$ *and* $t \in \mathcal{T}_i$ *scores per Def. 1 obey* $\boldsymbol{\gamma}_{i\alpha_i} > \boldsymbol{\gamma}_{it}$, $\boldsymbol{W}^{mm}_{\alpha}$ *is called a* locally-optimal direction.

To provide a basis for discussing local convergence of GD, we establish a cone centered around $\boldsymbol{W}^{mm}_{\alpha}$: For $\mu \in (0,1)$ and $R > 0$, we define $C_{\mu,R}(\boldsymbol{W}^{mm}_{\alpha}) := \left\{ \|\boldsymbol{W}\|_F \geq R \;\middle|\; \langle \boldsymbol{W}/\|\boldsymbol{W}\|_F, \boldsymbol{W}^{mm}_{\alpha}/\|\boldsymbol{W}^{mm}_{\alpha}\|_F \rangle \geq 1 - \mu \right\}$.

Theorem 3 *Suppose Assumption A holds, and let* $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ *be locally optimal tokens and* $\boldsymbol{W}^{mm}_{\alpha}$ *be a* locally-optimal direction *according to Def. 2. Then, Algorithm W-GD with* $\eta \leq 1/L_{\boldsymbol{W}}$ *and any* $\boldsymbol{W}(0) \in C_{\mu,R}(\boldsymbol{W}^{mm}_{\alpha})$ *satisfies* $\lim_{k\to\infty} \|\boldsymbol{W}(k)\|_F = \infty$ *and* $\lim_{k\to\infty} \frac{\boldsymbol{W}(k)}{\|\boldsymbol{W}(k)\|_F} = \frac{\boldsymbol{W}^{mm}_{\alpha}}{\|\boldsymbol{W}^{mm}_{\alpha}\|_F}$.

This theorem indicates that if GD is initiated within $C_{\mu,R}(\boldsymbol{W}^{mm}_{\alpha})$, it will converge in the direction of $\boldsymbol{W}^{mm}_{\alpha}/\|\boldsymbol{W}^{mm}_{\alpha}\|_F$. Importantly, Theorem 3 does not make any assumptions on the tokens as opposed to Theorem 2. In Figure 3 we consider setting where $n = 6$, $T = 8$, and $d = 10$. In Fig. 3(a) we calculate the softmax probabilities, which result in probability 1, indicating that attention weights succeed in selecting one token per input. Following Def. 2 let $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ be the token indices selected by GD and denote $\boldsymbol{W}^{mm}_{\star,\alpha}$ as the corresponding SVM solution of (Att-SVM$_\star$). Figs. 3(b) and 3(c) illustrate the correlation coefficients of attention weights with respect to $\boldsymbol{W}^{mm}_{\alpha}$ and $\boldsymbol{W}^{mm}_{\star,\alpha}$. The results demonstrate that $\boldsymbol{W}$ $(\boldsymbol{K}\boldsymbol{Q}^\top)$ ultimately reaches a 1 correlation with $\boldsymbol{W}^{mm}_{\alpha}$ $(\boldsymbol{W}^{mm}_{\star,\alpha})$, which validates Theorem 3.

3.3 Implicit bias under MLP nonlinearity

So far, we focus on the setting that $h(\cdot)$ is linear and attention selects a single token per sequence. In this section, we analyze the scenarios where $h(\cdot)$ is nonlinear and nonconvex, and GD solution is composed by multiple tokens. Suppose optimal solution outputs softmax probability of $\boldsymbol{s}^\star_i, i \in [n]$. Intuitively, $\boldsymbol{W}(k)$ should be decomposed into two components via

$$\boldsymbol{W}(k) \approx \boldsymbol{W}^{fin} + \|\boldsymbol{W}(k)\|_F \cdot \bar{\boldsymbol{W}}^{mm}. \tag{2}$$

where $\boldsymbol{W}^{fin}$ is the finite component and $\bar{\boldsymbol{W}}^{mm}$ is the directional component with $\|\bar{\boldsymbol{W}}^{mm}\|_F = 1$. Define the selected set $\mathcal{O}_i \subseteq [T]$ to be the indices $s^\star_{it} \neq 0$ and the masked set as $\bar{\mathcal{O}}_i = [T] - \mathcal{O}_i$.

Finite component ($\boldsymbol{W}^{fin}$): The job of $\boldsymbol{W}^{fin}$ is to assign nonzero softmax probabilities within each $\boldsymbol{s}^\star_i$. Then, $\boldsymbol{W}^{fin}$ should satisfy the linear constraints:

$$(\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W}^{fin} \boldsymbol{z}_i = \log(s^\star_{it}/s^\star_{i\tau}) \quad \text{for all} \quad t, \tau \in \mathcal{O}_i,\ i \in [n]. \tag{3}$$

Directional component ($\bar{\boldsymbol{W}}^{mm}$): While $\boldsymbol{W}^{fin}$ creates the composition by allocating the nonzero softmax probabilities, it does not explain sparsity of attention map. This is the role of $\bar{\boldsymbol{W}}^{mm}$, and we obtain the following convex generalized SVM formulation

$$\boldsymbol{W}^{mm} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \quad \text{subj. to} \quad \begin{cases} \forall\, t \in \mathcal{O}_i, \tau \in \bar{\mathcal{O}}_i : & (\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W} \boldsymbol{z}_i \geq 1, \\ \forall\, t, \tau \in \mathcal{O}_i : & (\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W} \boldsymbol{z}_i = 0, \end{cases} \quad \forall 1 \leq i \leq n, \tag{4}$$

and $\bar{\boldsymbol{W}}^{mm} = \boldsymbol{W}^{mm}/\|\boldsymbol{W}^{mm}\|_F$. It is important to note that (4) offers a substantial generalization beyond the scope of the previous sections. Remarkably, in Appendix B, we empirically demonstrate that this general form indeed seems to predict the implicit bias of gradient descent with MLPs.

References

[ACDS23] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

[ACHL19] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

[ALH21] Navid Azizan, Sahin Lale, and Babak Hassibi. Stochastic mirror descent on overparameterized nonlinear models. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7717–7727, 2021.

[ASA+22] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv:2211.15661*, 2022.

[AW20a] Ehsan Amid and Manfred K Warmuth. Winnowing with gradient descent. In *Conference on Learning Theory*, pages 163–182. PMLR, 2020.

[AW20b] Ehsan Amid and Manfred KK Warmuth. Reparameterizing mirror descent as gradient descent. *Advances in Neural Information Processing Systems*, 33:8430–8439, 2020.

[BALA+23] Enric Boix-Adsera, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua Susskind. Transformers learn through gradual rank increase. *arXiv preprint arXiv:2306.07042*, 2023.

[BCW+23] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

[BGVV20] 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*, pages 483–513. PMLR, 2020.

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901, 2020.

[BV22] Pierre Baldi and Roman Vershynin. The quarks of attention. *arXiv preprint arXiv:2202.08371*, 2022.

[Car21] Marcus Carlsson. von neumann's trace inequality for hilbert–schmidt operators. *Expositiones Mathematicae*, 39(1):149–157, 2021.

[CDL16] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas, November 2016. Association for Computational Linguistics.

[CDS01] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[CLR+21] 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 *Advances in Neural Information Processing Systems*, volume 34, pages 15084–15097, 2021.

[CRT06] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[CSL+23] Yingyi Chen, Xi Shen, Yahui Liu, Qinghua Tao, and Johan AK Suykens. Jigsaw-vit: Learning jigsaw puzzles in vision transformer. *Pattern Recognition Letters*, 166:53–60, 2023.

[DCL21] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021.

[DML21] Alex Damian, Tengyu Ma, and Jason Lee. Label noise sgd provably prefers flat global minimizers. *arXiv preprint arXiv:2106.06530*, 2021.

[Don06] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[EGKZ22] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

[ENM22] Tolga Ergen, Behnam Neyshabur, and Harsh Mehta. Convexifying transformers: Improving optimization and understanding of transformer networks. *arXiv:2211.11052*, 2022.

[FGBM23] 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.

[FXM+21] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[GLSS18] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[Goo23] Google. Try bard, an ai experiment by google. *https://bard.google.com*, 2023.

[GRS+23] Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. *arXiv:2301.13196*, 2023.

[GWB+17] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

[Hah20] Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

[HWLM20] Jeff Z HaoChen, Colin Wei, Jason D Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. *arXiv preprint arXiv:2006.08680*, 2020.

[JDST20] Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

[JLL21] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.

[JSL22] 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, editors, *Advances in Neural Information Processing Systems*, 2022.

[JST21] Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.

[JT18] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[JT19] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

[JT20] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17176–17186. Curran Associates, Inc., 2020.

[JT21] Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

[KPOT21] 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.

[KT19] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[LFS+17] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *International Conference on Learning Representations*, 2017.

[LIPO23] Yingcong Li, M Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, 2023.

[LLC+21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[LMZ18] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in overparameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

[LR20] TENGYUAN LIANG and ALEXANDER RAKHLIN. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

[LWA22] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022.

[LWLC23] Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. *arXiv preprint arXiv:2302.06015*, 2023.

[LWM19] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *arXiv preprint arXiv:1907.04595*, 2019.

[MRG+20] William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah Smith. Effects of parameter norm growth during transformer training: Inductive bias from gradient descent. *arXiv preprint arXiv:2010.09697*, 2020.

[MWG+20] Edward Moroshko, Blake E Woodworth, Suriya Gunasekar, Jason D Lee, Nati Srebro, and Daniel Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. *Advances in neural information processing systems*, 33:22182–22193, 2020.

[NLG+19] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[NLL+23] Lorenzo Noci, Chuning Li, Mufan Bill Li, Bobby He, Thomas Hofmann, Chris Maddison, and Daniel M Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *arXiv preprint arXiv:2306.17759*, 2023.

[NNH+23] Tan Minh Nguyen, Tam Minh Nguyen, Nhat Ho, Andrea L Bertozzi, Richard Baraniuk, and Stanley Osher. A primal-dual framework for transformers and neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

[OH10] Samet Oymak and Babak Hassibi. New null space results and recovery thresholds for matrix rank minimization. *arXiv preprint arXiv:1011.6326*, 2010.

[OMFH11] Samet Oymak, Karthik Mohan, Maryam Fazel, and Babak Hassibi. A simplified approach to recovery conditions for low rank matrices. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2318–2322. IEEE, 2011.

[Ope22] OpenAI. OpenAI: Introducing ChatGPT. https://openai.com/blog/chatgpt, 2022.

[Ope23] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[ORST23] Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, 2023.

[PHD20] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[PTDU16] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics.

[PXS18] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*, 2018.

[QQ19] Qian Qian and Xiaoyuan Qian. The implicit bias of adagrad on separable data. *Advances in Neural Information Processing Systems*, 32, 2019.

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[RXH11] Benjamin Recht, Weiyu Xu, and Babak Hassibi. Null space conditions and thresholds for rank minimization. *Mathematical programming*, 127:175–202, 2011.

[RZH03] Saharon Rosset, Ji Zhu, and Trevor Hastie. Margin maximizing loss functions. *Advances in neural information processing systems*, 16, 2003.

[SATA22] Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. *Advances in Neural Information Processing Systems*, 35:31089–31101, 2022.

[SEO+22] Arda Sahiner, Tolga Ergen, Batu Ozturkler, John Pauly, Morteza Mardani, and Mert Pilanci. Unraveling attention via convex duality: Analysis and interpretations of vision transformers. In *International Conference on Machine Learning*, pages 19050–19088. PMLR, 2022.

[SHN+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[SRJ04] Nathan Srebro, Jason Rennie, and Tommi Jaakkola. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17, 2004.

[SS21] Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843, 2021.

[TBS+16] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

[TCD+21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[TG07] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[TKVB22] Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

[TLI+23] 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.

[TLZO23] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Margin maximization in attention mechanism. *arXiv preprint arXiv:2306.13596*, 2023.

[TVS23] Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. In *International Conference on Algorithmic Learning Theory*, pages 1429–1459. PMLR, 2023.

[TWCD23] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *arXiv:2305.16380*, 2023.

[VKR19] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32:2972–2983, 2019.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[WGL+20] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

[WMCL21] Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pages 10849–10858. PMLR, 2021.

[WMZ+21] Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, and Zhi-Ming Ma. Momentum doesn't change the implicit bias. *arXiv preprint arXiv:2110.03891*, 2021.

[WWX+22] Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. In *International Conference on Machine Learning*, pages 24226–24242, 2022.

[YKM20] Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. *arXiv preprint arXiv:2010.02501*, 2020.

[ZFB23] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

[ZWB+21] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P Foster, and Sham Kakade. The benefits of implicit regularization from sgd in least squares problems. *Advances in Neural Information Processing Systems*, 34:5456–5468, 2021.

[ZY05] Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, page 1538, 2005.

Roadmap. The appendix is organized as follows:

- Appendix A provides related work.
- Appendix B provides detailed discussion and experimental evaluation about Section 3.3.
- Appendix C provides auxiliary lemmas.
- Appendix D presents the proof for the global regularization path analysis (Section 2).
- Appendix E presents the proofs for the gradient descent convergences (Section 3).
- Appendix F provides additional experiments and their discussion.
- Appendix G discusses potential further directions.

Contents

A Related work

A.1 Implicit Regularization, Matrix Factorization, Sparsity

Extensive research has delved into gradient descent's implicit bias in separable classification tasks, often using logistic or exponentially-tailed losses for margin maximization [SHN+18, GLSS18, NLG+19, JT21, KPOT21, MWG+20, JT20]. The findings have also been extended to non-separable data using gradient-based techniques [JT18, JT19, JDST20]. Implicit bias in regression problems and losses has been investigated, utilizing methods like mirror descent [WGL+20, GLSS18, YKM20, VKR19, AW20a, AW20b, ALH21, SATA22]. Stochastic gradient descent has also been a subject of interest regarding its implicit bias [LWM19, BGVV20, LR20, HWLM20, LWA22, DML21, ZWB+21]. This extends to the implicit bias of adaptive and momentum-based methods [QQ19, WMZ+21, WMCL21, JST21].

In linear classification, GD iterations on logistic loss and separable datasets converge to the hard margin SVM solution [SHN+18, RZH03, ZY05]. The attention layer's softmax nonlinearity behaves similarly, potentially favoring margin-maximizing solutions. Yet, the layer operates on tokens in input sequences, not for direct classification. Its bias leans toward an (Att-SVM), selecting relevant tokens while suppressing others. However, formalizing this intuition presents significant challenges: Firstly, our problem is nonconvex (even in terms of the $\boldsymbol{W}$-parameterization), introducing new challenges and complexities. Secondly, it requires the introduction of novel concepts such as locally-optimal tokens, demanding a tailored analysis focused on the cones surrounding them. Our findings on the implicit bias of $(\boldsymbol{K},\boldsymbol{Q})$-parameterization share conceptual similarities with [SRJ04], which proposes and analyzes a max-margin matrix factorization problem. Similar problems have also been studied more recently in the context of neural-collapse phenomena [PHD20] through an analysis of the implicit bias and regularization path of the unconstrained features model with cross-entropy loss [TKVB22]. However, a fundamental distinction from these works lies in the fact that attention solves a different max-margin problem that separate tokens. Moreover, our results on $(\boldsymbol{K},\boldsymbol{Q})$-parameterization are inherently connected to the rich literature on low-rank factorization [GWB+17, ACHL19, TVS23, TBS+16, SS21], stimulating further research. [TLZO23] is the first work to establish the connection between attention and SVM, which is closest to our work. Here, we augment their framework, initially developed for a simpler attention model, to transformers by providing the first guarantees for self/cross-attention layers, nonlinear prediction heads, and realistic global convergence guarantees. While our Assumption (i) and local-convergence analysis align with [TLZO23], our contributions in global convergence analysis, benefits of overparameterization, and the generalized SVM-equivalence in Section B are unique to this work.

It is well-known that attention map (i.e. softmax outputs) act as a feature selection mechanism and reveal the tokens that are relevant to classification. On the other hand, sparsity and lasso regression (i.e. ℓ_1 penalization) [Don06, Tib96, TG07, CDS01, CRT06] have been pivotal tools in the statistics literature for feature selection. Softmax and lasso regression exhibit interesting parallels: The Softmax output $\boldsymbol{s} = \mathbb{S}(\boldsymbol{XWz})$ obeys $\|\boldsymbol{s}\|_{\ell_1} = 1$ by design. Softmax is also highly receptive to being sparse because decreasing the temperature (i.e. scaling up the weights $\boldsymbol{W}$) eventually leads to a one-hot vector unless all logits are equal. We (also, [TLZO23]) have used these intuitions to formalize attention as a *token selection mechanism*. This aspect is clearly visible in our primary SVM formulation (Att-SVM) which selects precisely one token from each input sequence (i.e. hard attention). Section B has also demonstrated how (Gen-SVM) can explain more general sparsity patterns by precisely selecting desired tokens and suppressing others. We hope that this SVM-based token-selection viewpoint will motivate future work and deeper connections to the broader feature-selection and compressed sensing literature.

A.2 Attention Mechanism and Transformers

Transformers, as highlighted by [VSP+17], revolutionized the domains of NLP and machine translation. Prior work on self-attention [CDL16, PTDU16, PXS18, LFS+17] laid the foundation for this transformative paradigm. In contrast to conventional models like MLPs and CNNs, self-attention models employ global interactions to capture feature representations, resulting in exceptional empirical performance.

Despite their achievements, the mechanisms and learning processes of attention layers remain enigmatic. Recent investigations [EGKZ22, SEO+22, ENM22, BV22, DCL21] have concentrated on specific aspects such as sparse function representation, convex relaxations, and expressive power. Expressivity discussions concerning hard-attention [Hah20] or attention-only architectures [DCL21] are connected to our findings when $h(\cdot)$ is linear. In fact, our work reveals how linear h results in attention's optimization dynamics to collapse on a single token whereas nonlinear h provably requires attention to select and compose multiple tokens. This supports the benefits of the MLP layer for expressivity of transformers. There is also a growing body of research aimed at a theoretical comprehension of in-context learning and the role played by the attention mechanism [ASA+22, LIPO23, ACDS23, ZFB23, BCW+23, GRS+23]. [SEO+22] investigate self-attention with linear activation instead of softmax, while [ENM22] approximate softmax using a linear operation with unit simplex constraints. Their primary goal is to derive convex reformulations for training problems grounded in empirical risk minimization (ERM). In contrast, our methodologies, detailed in equations (W-ERM) and (KQ-ERM), delve into the nonconvex domain.

[MRG+20, BALA+23] offer insights into the implicit bias of optimizing transformers. Specifically, [MRG+20] provide empirical evidence that an increase in attention weights results in a sparser softmax, which aligns with our theoretical framework. [BALA+23] study incremental learning and furnish both theory and numerical evidence that increments of the softmax attention weights ($\boldsymbol{K}\boldsymbol{Q}^\top$) are low-rank. Our theory aligns with this concept, as the SVM formulation of $(\boldsymbol{K}, \boldsymbol{Q})$ parameterization inherently exhibits low-rank properties through the nuclear norm objective, rank-m constraint, and implicit constraint induced by Lemma 1.

Several recent works [JSL22, LWLC23, TWCD23, NLL+23, ORST23, NNH+23, FGBM23] aim to delineate the optimization and generalization dynamics of transformers. However, their findings usually apply under strict statistical assumptions about the data, while our study offers a comprehensive optimization-theoretic analysis of the attention model, establishing a formal linkage to max-margin problems and SVM geometry. This allows our findings to encompass the problem geometry and apply to diverse datasets. Overall, the max-margin equivalence provides a fundamental comprehension of the optimization geometry of transformers, offering a framework for prospective research endeavors, as outlined in the subsequent section.

B Understanding Multi-token Compositions: Toward A More General Max-Margin and Directional Convergence Theory

So far, our theory has focused on the setting where the attention layer selects a single optimal token within each sequence. As we have discussed, this is theoretically well-justified under linear head assumption and certain nonlinear generalizations. On the other hand, for arbitrary nonconvex $h(\cdot)$ or multilayer transformer architectures, it is expected that attention will select multiple tokens per sequence. This motivates us to ask:

> **Q:** What is the implicit bias and the form of $\boldsymbol{W}(k)$ when the GD solution is composed by multiple tokens?

In this section, our goal is to derive and verify the generalized behavior of GD. Let $\boldsymbol{o}_i = \boldsymbol{X}_i^\top \boldsymbol{s}_i^{\boldsymbol{W}}$ denote the token generated by the attention layer where $\boldsymbol{s}_i^{\boldsymbol{W}} = \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{z}_i)$ are the softmax probabilities.

Suppose GD trajectory converges to achieve the risk $\mathcal{L}_\star = \min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W})$. Suppose the eventual token composition achieving $\mathcal{L}_\star$ is given by

$$\boldsymbol{o}_i^\star = \boldsymbol{X}_i^\top \boldsymbol{s}_i^\star,$$

where $\boldsymbol{s}_i^\star$ are the eventual softmax probability vectors that dictate the token composition. Since attention maps are sparse in practice, we are interested in the scenario where $\boldsymbol{s}_i^\star$ is sparse i.e. it contains some zero entries. This can only be accomplished by letting $\|\boldsymbol{W}\|_F \to \infty$. However, unlike the earlier sections, we wish to allow for arbitrary $\boldsymbol{s}_i^\star$ rather than a one-hot vector which selects a single token.

To proceed, we aim to understand the form of GD solution $\boldsymbol{W}(k)$ responsible for composing $\boldsymbol{o}_i^\star$ via the softmax map $\boldsymbol{s}_i^\star$ as $R \to \infty$. Intuitively, $\boldsymbol{W}(k)$ should be decomposed into two components via

$$\boldsymbol{W}(k) \approx \boldsymbol{W}^{fin} + \|\boldsymbol{W}(k)\|_F \cdot \bar{\boldsymbol{W}}^{mm}. \tag{5}$$

where $\boldsymbol{W}^{fin}$ is the finite component and $\bar{\boldsymbol{W}}^{mm}$ is the directional component with $\|\bar{\boldsymbol{W}}^{mm}\|_F = 1$. Define the selected set $\mathcal{O}_i \subseteq [T]$ to be the indices $s_{it}^\star \neq 0$ and the masked (i.e. suppressed) set as $\bar{\mathcal{O}}_i = [T] - \mathcal{O}_i$ where softmax entries are zero. In the context of earlier sections, we could also call these the *optimal set* and the *non-optimal set*, respectively.

- **Finite component:** The job of $\boldsymbol{W}^{fin}$ is to assign nonzero softmax probabilities within each $\boldsymbol{s}_i^\star$.

This is accomplished by ensuring that, $\boldsymbol{W}^{fin}$ induces the probabilities of $\boldsymbol{s}_i^\star$ over $\mathcal{O}_i$ by satisfying the softmax equations

$$\frac{e^{\boldsymbol{x}_{it}^\top \boldsymbol{W}^{fin} \boldsymbol{z}_i}}{e^{\boldsymbol{x}_{i\tau}^\top \boldsymbol{W}^{fin} \boldsymbol{z}_i}} = e^{(\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W}^{fin} \boldsymbol{z}_i} = s_{it}^\star / s_{i\tau}^\star.$$

for $t, \tau \in \mathcal{O}_i$. Consequently, this $\boldsymbol{W}^{fin}$ should satisfy the following linear constraints

$$(\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W}^{fin} \boldsymbol{z}_i = \log(s_{it}^\star / s_{i\tau}^\star) \quad \text{for all} \quad t, \tau \in \mathcal{O}_i,\ i \in [n]. \tag{6}$$

- **Directional component:** While $\boldsymbol{W}^{fin}$ creates the composition by allocating the nonzero softmax probabilities, it does not explain sparsity of attention map. This is the role of $\bar{\boldsymbol{W}}^{mm}$, which is responsible for selecting the selected tokens $\mathcal{O}_i$ and suppressing the masked ones $\bar{\mathcal{O}}_i$ by assigning zero softmax probability to them. To predict direction component, we build on the theory developed in earlier sections. Concretely, there are two constraints $\bar{\boldsymbol{W}}^{mm}$ should satisfy

1. **Equal similarity over selected tokens:** For all $t, \tau \in \mathcal{O}_i$, we have that $(\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W} \boldsymbol{z}_i = 0$. This way, softmax scores assigned by $\boldsymbol{W}^{fin}$ are not disturbed by the directional component and $\boldsymbol{W}^{fin} + R\bar{\boldsymbol{W}}^{mm}$ will still satisfy the softmax equations (6).
2. **Max-margin against masked tokens:** For all $t \in \mathcal{O}_i, \tau \in \bar{\mathcal{O}}_i$, enforce the margin constraint $(\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W} \boldsymbol{z}_i \geq 1$ subject to minimum norm $\|\boldsymbol{W}\|_F$.

Combining these, we obtain the following convex generalized SVM formulation

$$\boldsymbol{W}^{mm} = \arg\min_{\boldsymbol{W}} \|\boldsymbol{W}\|_F \quad \text{subj. to} \quad \begin{cases} \forall\, t \in \mathcal{O}_i, \tau \in \bar{\mathcal{O}}_i: & (\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W} \boldsymbol{z}_i \geq 1, \\ \forall\, t, \tau \in \mathcal{O}_i: & (\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W} \boldsymbol{z}_i = 0, \end{cases} \quad \forall 1 \leq i \leq n. \tag{Gen-SVM}$$

and set the normalized direction in (5) to $\bar{\boldsymbol{W}}^{mm} = \boldsymbol{W}^{mm} / \|\boldsymbol{W}^{mm}\|_F$.

It is important to note that (Gen-SVM) offers a substantial generalization beyond the scope of the previous sections, where the focus was on selecting a single token from each sequence, as described in the main formulation (Att-SVM). This broader solution class introduces a more flexible approach to the problem.

We present experiments showcasing the predictive power of the (Gen-SVM) equivalence in nonlinear scenarios. We conducted these experiments on random instances using an MLP denoted as $h(\cdot)$, which takes the form of $\mathbf{1}^\top \texttt{ReLU}(\boldsymbol{x})$. We begin by detailing the preprocessing step and our setup. For the attention SVM equivalence analytical prediction, clear definitions of the selected and masked sets are crucial. These sets include token indices with nonzero and zero softmax outputs, respectively. However, practically, reaching a precisely zero output is not feasible. Hence, we define the selected set as tokens with softmax outputs exceeding 10^{-3}, and the masked set as tokens with softmax outputs

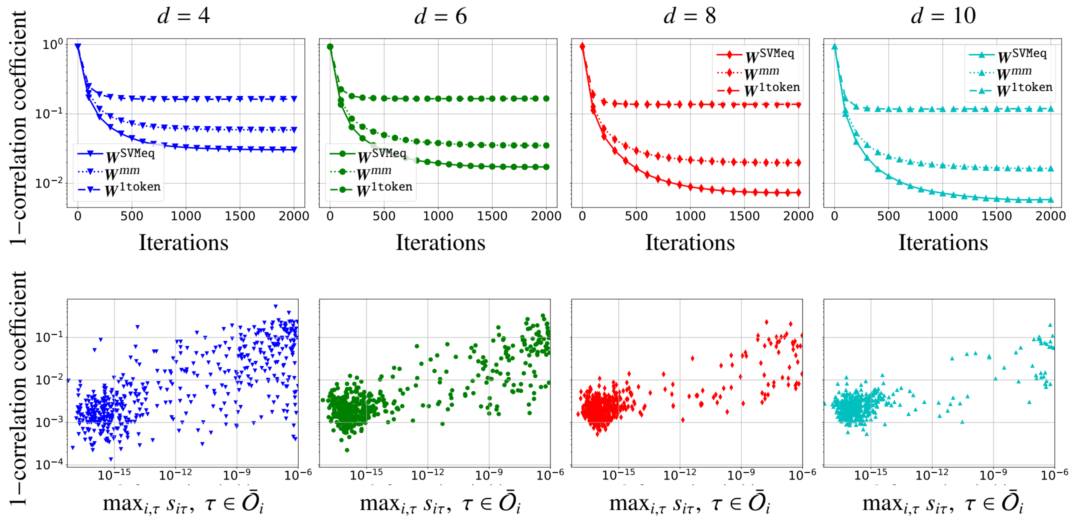
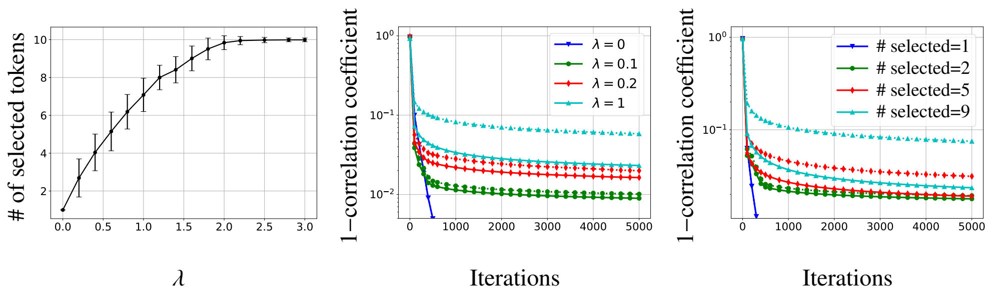


Figure 4: Behavior of GD with nonlinear nonconvex prediction head and multi-token compositions. **Upper:** The correlation between GD solution and three distinct baselines: $(\cdots)$ $\boldsymbol{W}^{mm}$ obtained from (Gen-SVM); (**—**) $\boldsymbol{W}^{\texttt{SVMeq}}$ obtained by calculating $\boldsymbol{W}^{fin}$ and determining the best linear combination $\boldsymbol{W}^{fin} + \gamma\bar{\boldsymbol{W}}^{mm}$ that maximizes correlation with the GD solution; and (**- -**) $\boldsymbol{W}^{\texttt{1token}}$ obtained by solving (Att-SVM) and selecting the highest probability token from the GD solution. **Lower:** Scatterplot of the largest softmax probability over masked tokens (per our $s_{i\tau} \leq 10^{-6}$ criteria) vs correlation coefficient.

below 10^{-6}. We also excluded instances with softmax outputs falling between 10^{-6} and 10^{-3} to distinctly separate the concepts of *selected* and *masked* sets, thereby enhancing the predictive accuracy of the attention SVM equivalence. In addition to the filtering process, we focus on scenarios where the label $Y = -1$ exists to enforce *non-convexity* of prediction head $Y_i \cdot h(\cdot)$. It is worth mentioning that when all labels are 1, due to the convexity of $Y_i \cdot h(\cdot)$, GD tends to select one token per input, and Equations (Gen-SVM) and (Att-SVM) yield the same solutions. The results are displayed in Figure 4, where $n = 3$, $T = 4$, and d varies within $4, 6, 8, 10$. We conduct 500 random trials for different choices of d, each involving $\boldsymbol{x}_{it}$, $\boldsymbol{z}_i$, and $\boldsymbol{v}$ randomly sampled from the unit sphere. We apply normalized GD with a step size $\eta = 0.1$ and run 2000 iterations for each trial.

- Figure 4 (upper) illustrates the correlation evolution between the GD solution and three distinctive baselines: $(\cdots)$ $\boldsymbol{W}^{mm}$ obtained from (Gen-SVM); (**—**) $\boldsymbol{W}^{\texttt{SVMeq}}$ obtained by calculating $\boldsymbol{W}^{fin}$ and determining the best linear combination $\boldsymbol{W}^{fin}+\gamma\bar{\boldsymbol{W}}^{mm}$ that maximizes correlation with the GD solution; and (**- -**) $\boldsymbol{W}^{\texttt{1token}}$ obtained by solving (Att-SVM) and selecting the highest probability token from the GD solution. For clearer visualization, the logarithmic scale of correlation misalignment is presented in Figure 4. In essence, our findings show that $\boldsymbol{W}^{\texttt{1token}}$ yields unsatisfactory outcomes, whereas $\boldsymbol{W}^{mm}$ attains a significant correlation coefficient in alignment with our expectations. Ultimately, our comprehensive SVM-equivalence $\boldsymbol{W}^{\texttt{SVMeq}}$ further enhances correlation, lending support to our analytical formulas. It's noteworthy that SVM-equivalence displays higher predictability in a larger d regime (with an average correlation exceeding 0.99). This phenomenon might be attributed to more frequent directional convergence in higher dimensions, with overparameterization contributing to a smoother loss landscape, thereby expediting optimization.

- Figure 4 (lower) offers a scatterplot overview of the 500 random problem instances that were solved. The x-axis represents the largest softmax probability over the masked set, denoted as $\max_{i,\tau} s_{i\tau}$ where $\tau \in \bar{\mathcal{O}}_i$. Meanwhile, the y-axis indicates the predictivity of the SVM-equivalence, quantified as $1 - \texttt{corr\_coef}(\boldsymbol{W}, \boldsymbol{W}^{\texttt{SVMeq}})$. From this analysis, two significant observations arise. Primarily, there exists an inverse correlation between softmax probability and SVM-predictivity. This correlation is intuitive, as higher softmax probabilities signify a stronger divergence from our desired *masked set* state (ideally set to 0). Secondly, as dimensionality (d) increases, softmax probabilities over the masked set tend to converge towards the range of 10^{-15} (effectively zero). Simultaneously, attention SVM-predictivity improves, creating a noteworthy correlation.



(a) λ vs # selected tokens (b) λ vs correlation coefficient (c) # selected tokens vs correlation

Figure 5: Behavior of GD when selecting multiple tokens. **(a)** The number of selected tokens increases with λ. **(b)** Predictivity of attention SVM solutions for varying λ; Dotted curves depict the correlation corresponding to $\boldsymbol{W}^{mm}$ calculated via (Gen-SVM) and solid curves represent the correlation to $\boldsymbol{W}^{\texttt{SVMeq}}$, which incorporates the $\boldsymbol{W}^{fin}$ correction. **(c)** Similar to (b), but evaluating correlations over different numbers of selected tokens.

B.1 When does attention select multiple tokens?

In this section, we provide a concrete example where the optimal solution indeed requires combining multiple tokens in a nontrivial fashion. Here, by nontrivial we mean that, we select more than 1 tokens from an input sequence but we don't select all of its tokens. Recall that, for linear prediction head, attention will ideally select the single token with largest score for almost all datasets. Perhaps not surprisingly, this behavior will not persist for nonlinear prediction heads. For instance in Figure 4, the GD output $\boldsymbol{W}$ aligned better in direction with $\boldsymbol{W}^{mm}$ than $\boldsymbol{W}^{\texttt{1token}}$. Specifically, here we prove that if we make the function $h_Y(\boldsymbol{x}) := Y \cdot h(\boldsymbol{x})$ concave, then optimal softmax map can select multiple tokens in a controllable fashion. $h_Y(\boldsymbol{x})$ can be viewed as generalization of the linear score function $Y \cdot \boldsymbol{v}^\top \boldsymbol{x}$. In the example below, we induce concavity by incorporating a small $-\lambda\|\boldsymbol{x}\|^2$ term within a linear prediction head and setting $h(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x} - \lambda\|\boldsymbol{x}\|^2$ with $Y = 1$.

Lemma 3 *Given $\boldsymbol{v} \in \mathbb{R}^d$, recall the score vector $\boldsymbol{\gamma} = \boldsymbol{X}\boldsymbol{v}$. Without losing generality, assume $\boldsymbol{\gamma}$ is non-increasing. Define the vector of score gaps $\boldsymbol{\gamma}^{gap} \in \mathbb{R}^{T-1}$ with entries $\boldsymbol{\gamma}_t^{gap} = \boldsymbol{\gamma}_t - \boldsymbol{\gamma}_{t+1}$. Suppose all tokens within the input sequence are orthonormal and for some $\tau \geq 2$, we have that*

$$\tau \boldsymbol{\gamma}_\tau^{gap}/2 > \boldsymbol{\gamma}_1^{gap}. \tag{7}$$

Set $h(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x} - \lambda\|\boldsymbol{x}\|^2$ where $\tau\boldsymbol{\gamma}_\tau^{gap}/2 > \lambda > \boldsymbol{\gamma}_1^{gap}$, $\ell(x) = -x$, and $Y = 1$. Let $\boldsymbol{\Delta}_T$ denote the T-dimensional simplex. Define the unconstrained softmax optimization associated to the objective h where we make $\boldsymbol{s} := \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{z})$ a free variable, namely,

$$\min_{\boldsymbol{s} \in \boldsymbol{\Delta}_T} \ell(h(\boldsymbol{X}\boldsymbol{s})) = \min_{\boldsymbol{s} \in \boldsymbol{\Delta}_T} \lambda\|\boldsymbol{X}^\top \boldsymbol{s}\|^2 - \boldsymbol{v}^\top \boldsymbol{X}^\top \boldsymbol{s}. \tag{8}$$

Then, the optimal solution $\boldsymbol{s}^\star$ contains at least 2 and at most τ nonzero entries.

Figure 5 presents experimental findings concerning Lemma 3 across random problem instances. For this experiment, we set $n = 1$, $T = 10$, and $d = 10$. The results are averaged over 100 random trials, with each trial involving the generation of randomly orthonormal vectors $\boldsymbol{x}_{1t}$ and the random sampling of vector $\boldsymbol{v}$ from the unit sphere. Similar to the processing step in Figure 4, and following Figure 4 (lower) which illustrates that smaller softmax outputs over masked sets correspond to higher correlation coefficients, we define the selected and masked token sets. Specifically, tokens with softmax outputs $> 10^{-3}$ are considered selected, while tokens with softmax outputs $< 10^{-8}$ are masked. Instances with softmax outputs between 10^{-8} and 10^{-3} are filtered out.

Figure 5(a) shows that the number of selected tokens grows alongside λ, a prediction consistent with Lemma 3. When $\lambda = 0$, the head $h(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x}$ is linear, resulting in the selection of only one token per input. Conversely, as λ exceeds a certain threshold (e.g., $\lambda > 2.0$ based on our criteria), the optimization consistently selects all tokens. Figure 5(b) and 5(c) delve into the predictivity of attention SVM solutions for varying λ and different numbers of selected tokens. The dotted curves in both figures represent $1 - \texttt{corr\_coef}(\boldsymbol{W}, \boldsymbol{W}^{mm})$, while solid curves indicate $1 - \texttt{corr\_coef}(\boldsymbol{W}, \boldsymbol{W}^{\texttt{SVMeq}})$, where $\boldsymbol{W}$ denotes the GD solution. Overall, the SVM-equivalence demonstrates a strong correlation with the GD solution (consistently above 0.95). However, selecting more tokens (aligned with larger λ values) leads to reduced predictivity.

To sum up, we have showcased the predictive capacity of the generalized SVM equivalence regarding the inductive bias of 1-layer transformers with nonlinear heads. Nevertheless, it's important to acknowledge that this section represents an initial approach to a complex problem, with certain caveats requiring further investigation (e.g., the use of filtering in Figures 4 and 5, and the presence of imperfect correlations). We aspire to conduct a more comprehensive investigation, both theoretically and empirically, in forthcoming work.

B.2 Proof of Lemma 3

Suppose τ described by (7) exists and set λ accordingly. Let $\mathcal{S} \subset [T]$ denote the top τ indices of $\boldsymbol{\gamma}$ with largest scores. Denote $\boldsymbol{X}^1 \in \mathbb{R}^{\tau\times d}$ to be the sequence corresponding to $\mathcal{S}$ and $\boldsymbol{X}^2 \in \mathbb{R}^{(T-\tau)\times d}$ to be the sequence corresponding to $[T]-\mathcal{S}$. Similarly, denote the subvectors $\boldsymbol{\gamma}_1, \boldsymbol{s}^{(1)} \in \mathbb{R}^{\tau}$ and $\boldsymbol{\gamma}_2, \boldsymbol{s}^{(2)} \in \mathbb{R}^{T-\tau}$ and define the probability over $\mathcal{S}$ as $S_1 = \sum_{i\in\mathcal{S}} s_i$. The orthogonality and unit norm assumption on the tokens imply

$$1 \geq \|\boldsymbol{X}^\top \boldsymbol{s}\|^2 = \sum_{i=1}^{T} s_i^2 \geq S_1^2/\tau + (1-S_1)^2/(T-\tau).$$

Also note that $\boldsymbol{v}^\top \boldsymbol{X}\boldsymbol{s} = \boldsymbol{\gamma}_1^\top \boldsymbol{s}^{(1)} + \boldsymbol{\gamma}_2^\top \boldsymbol{s}^{(2)}$. With these, we can write the objective $\mathcal{L}(\boldsymbol{s}) := \ell(h(\boldsymbol{X}\boldsymbol{s}))$ as follows

$$\mathcal{L}(\boldsymbol{s}) = \lambda \sum_{i=1}^{T} s_i^2 - \boldsymbol{\gamma}_1^\top \boldsymbol{s}^{(1)} - \boldsymbol{\gamma}_2^\top \boldsymbol{s}^{(2)}.$$

Note that, for fixed $\boldsymbol{\gamma}$ and over all permutations of entries of $\boldsymbol{s}$, $\boldsymbol{\gamma}^\top \boldsymbol{s}$ is maximized when $\boldsymbol{s}$ and $\boldsymbol{\gamma}$ are aligned namely, when the entries of $\boldsymbol{s}$ are sorted according to the entries of $\boldsymbol{\gamma}$. Otherwise, we could swap two unsorted entries of $\boldsymbol{s}$ (i.e. with unaligned $\boldsymbol{\gamma}$ entries) to a sorted position to obtain a strictly better optimal (where we also used the fact that $\boldsymbol{s}$ has nonnegative entries). Thus, we can assume the entries of $\boldsymbol{s}^\star$ are sorted according to $\boldsymbol{\gamma}$. Specifically, the largest τ entries of $\boldsymbol{s}^\star$ lie on the set $\mathcal{S}$.

• **We first show that $\boldsymbol{s} := \boldsymbol{s}^\star$ cannot have more than τ entries.** To prove this, we compare $\boldsymbol{s}$ against the baseline $\bar{\boldsymbol{s}}$ where $\bar{\boldsymbol{s}}^1 = \boldsymbol{s}^{(1)}/S_1$ and $\bar{\boldsymbol{s}}^2 = 0$ so that $\bar{\boldsymbol{s}}$ is τ-sparse. In this scenario, $\bar{\boldsymbol{s}}$ yields the objective

$$\mathcal{L}(\bar{\boldsymbol{s}}) = \frac{\lambda}{S_1^2} \sum_{i\in\mathcal{S}} s_i^2 - \frac{1}{S_1}\boldsymbol{\gamma}_1^\top \boldsymbol{s}^{(1)}.$$

We claim that $\mathcal{L}(\bar{\boldsymbol{s}}) < \mathcal{L}(\boldsymbol{s})$. To see this, we first observe that $\boldsymbol{\gamma}_1^\top \boldsymbol{s}^{(1)}/S_1 \geq \boldsymbol{\gamma}_2^\top \boldsymbol{s}^{(2)}/(1-S_1) + \boldsymbol{\gamma}_\tau^{gap}$. This implies

$$(1/S_1 - 1)\boldsymbol{\gamma}_1^\top \boldsymbol{s}^{(1)} - \boldsymbol{\gamma}_2^\top \boldsymbol{s}^{(2)} \geq (1-S_1)\boldsymbol{\gamma}_\tau^{gap}.$$

Recalling $\sum_{i\in\mathcal{S}} s_i^2 \leq S_1^2/\tau$, we can now utilize the following chain of implications

$$\begin{aligned}
&\mathcal{L}(\bar{\boldsymbol{s}}) < \mathcal{L}(\boldsymbol{s})\\
\iff &\frac{\lambda}{S_1^2} \sum_{i\in\mathcal{S}} s_i^2 - \frac{1}{S_1}\boldsymbol{\gamma}_1^\top \boldsymbol{s}^{(1)} < \lambda \sum_{i=1}^{T} s_i^2 - \boldsymbol{\gamma}_1^\top \boldsymbol{s}^{(1)} - \boldsymbol{\gamma}_2^\top \boldsymbol{s}^{(2)}\\
\Longleftarrow\ &\lambda(1/S_1^2 - 1)\sum_{i\in\mathcal{S}} s_i^2 < (1/S_1 - 1)\boldsymbol{\gamma}_1^\top \boldsymbol{s}^{(1)} - \boldsymbol{\gamma}_2^\top \boldsymbol{s}^{(2)}\\
\Longleftarrow\ &\lambda(1/S_1^2 - 1)\sum_{i\in\mathcal{S}} s_i^2 < (1-S_1)\boldsymbol{\gamma}_\tau^{gap}\\
\Longleftarrow\ &\lambda(1-S_1^2)/\tau < (1-S_1)\boldsymbol{\gamma}_\tau^{gap}\\
\Longleftarrow\ &\lambda(1+S_1)/\tau < \boldsymbol{\gamma}_\tau^{gap}\\
\Longleftarrow\ &2\lambda/\tau < \boldsymbol{\gamma}_\tau^{gap}\\
\Longleftarrow\ &\lambda < \tau\boldsymbol{\gamma}_\tau^{gap}/2.
\end{aligned}$$

• **We next prove that there are at least two nonzeros in the optimal solution.** Denote the largest and second largest entry of $\boldsymbol{\gamma}$ by $\bar{\gamma}_1$ and $\bar{\gamma}_2$ respectively. For $\boldsymbol{s}^{\text{one}} \in \boldsymbol{\Delta}_T$ containing a single nonzero (i.e. one-hot vector), the best achievable risk is given by

$$\mathcal{L}(\boldsymbol{s}^{\text{one}}) = \lambda - \bar{\gamma}_1.$$

On the other hand consider the 2-sparse reference solution $\boldsymbol{s}^{\texttt{ref}}$ which assigns equal likelihood over the top two entries. This achieves

$$\mathcal{L}(\boldsymbol{s}^{\texttt{ref}}) = \frac{\lambda}{2} - \boldsymbol{\gamma}^\top \boldsymbol{s}^{\texttt{ref}} \leq \frac{\lambda}{2} - \frac{\bar{\gamma}_1 + \bar{\gamma}_2}{2}.$$

The latter is superior as soon as

$$\frac{\lambda}{2} - \frac{\bar{\gamma}_1 + \bar{\gamma}_2}{2} < \lambda - \bar{\gamma}_1 \iff \lambda > \boldsymbol{\gamma}_1^{gap}.$$

Thus, we conclude with the statement by selecting $\tau\boldsymbol{\gamma}_\tau^{gap}/2 > \lambda > \gamma_1^{gap}$. ■

C Auxiliary Lemmas

C.1 Proof of Lemma 1

Suppose the claim is wrong and row space of $\boldsymbol{W}_\diamond^{mm}$ does not lie within $\mathcal{S} = \texttt{span}(\{z_i\}_{i=1}^n)$. Let $\boldsymbol{W} = \Pi_{\mathcal{S}}(\boldsymbol{W}_\diamond^{mm})$ denote the matrix obtained by projecting the rows of $\boldsymbol{W}_\diamond^{mm}$ on $\mathcal{S}$. Observe that $\boldsymbol{W}$ satisfies all SVM constraints since $\boldsymbol{W}\boldsymbol{z}_i = \boldsymbol{W}_\diamond^{mm}\boldsymbol{z}_i$ for all $i \in [n]$. For Frobenius norm, using $\boldsymbol{W}_\diamond^{mm} \neq \boldsymbol{W}$, we obtain a contradiction via $\|\boldsymbol{W}_\diamond^{mm}\|_F^2 = \|\boldsymbol{W}\|_F^2 + \|\boldsymbol{W}_\diamond^{mm} - \boldsymbol{W}\|_F^2 > \|\boldsymbol{W}\|_F^2$. For nuclear norm, we can write $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ with $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ where r is dimension of $\mathcal{S}$ and $\texttt{column\_span}(\boldsymbol{V}) = \mathcal{S}$.

To proceed, we split the problem into two scenarios.

Scenario 1: Let $\boldsymbol{U}_\perp, \boldsymbol{V}_\perp$ be orthogonal complements of $\boldsymbol{U}, \boldsymbol{V}$ – viewing matrices with orthonormal columns as subspaces. Suppose $\boldsymbol{U}_\perp^\top \boldsymbol{W}_\diamond^{mm} \boldsymbol{V}_\perp \neq 0$. Then, singular value inequalities (which were also used in earlier works on nuclear norm analysis [RXH11, OH10, OMFH11]) guarantee that $\|\boldsymbol{W}_\diamond^{mm}\|_\star \geq \|\boldsymbol{U}^\top \boldsymbol{W}_\diamond^{mm}\boldsymbol{V}\|_\star + \|\boldsymbol{U}_\perp^\top \boldsymbol{W}_\diamond^{mm}\boldsymbol{V}_\perp\|_\star > \|\boldsymbol{W}\|_\star$.

Scenario 2: Now suppose $\boldsymbol{U}_\perp^\top \boldsymbol{W}_\diamond^{mm} \boldsymbol{V}_\perp = 0$. Since $\boldsymbol{W}_\diamond^{mm}\boldsymbol{V}_\perp \neq 0$, this implies $\boldsymbol{U}^\top \boldsymbol{W}_\diamond^{mm}\boldsymbol{V}_\perp \neq 0$. Let $\boldsymbol{W}' = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{W}_\diamond^{mm}$ which is a rank-r matrix. Since $\boldsymbol{W}'$ is a subspace projection, we have $\|\boldsymbol{W}'\|_\star \leq \|\boldsymbol{W}_\diamond^{mm}\|_\star$. Next, observe that $\|\boldsymbol{W}\|_\star = \texttt{trace}(\boldsymbol{U}^\top\boldsymbol{W}\boldsymbol{V}) = \texttt{trace}(\boldsymbol{U}^\top\boldsymbol{W}'\boldsymbol{V})$. On the other hand, $\texttt{trace}(\boldsymbol{U}^\top\boldsymbol{W}'\boldsymbol{V}) < \|\boldsymbol{W}'\|_\star$ because the equality in *von Neumann's trace inequality* happens if and only if the two matrices we are inner-producting, namely $(\boldsymbol{W}', \boldsymbol{U}\boldsymbol{V}^\top)$, share a joint set of singular vectors [Car21]. However, this is not true as the row space of $\boldsymbol{W}_\diamond^{mm}$ does not lie within $\mathcal{S}$. Thus, we obtain $\|\boldsymbol{W}\|_\star < \|\boldsymbol{W}'\|_\star \leq \|\boldsymbol{W}_\diamond^{mm}\|_\star$ concluding the proof via contradiction. ■

C.2 Proof of Lemma 2

Lemma 4 (Lemma 2 restated) *Under Assumption A, $\nabla\mathcal{L}(\boldsymbol{W})$, $\nabla_{\boldsymbol{K}}\mathcal{L}(\boldsymbol{K},\boldsymbol{Q})$, and $\nabla_{\boldsymbol{Q}}\mathcal{L}(\boldsymbol{K},\boldsymbol{Q})$ are $L_{\boldsymbol{W}}$, $L_{\boldsymbol{K}}$, $L_{\boldsymbol{Q}}$–Lipschitz continuous, respectively, where $a_i = \|\boldsymbol{v}\|\,\|z_i\|^2\|\boldsymbol{X}_i\|^3$, $b_i = M_0\|\boldsymbol{v}\|\,\|\boldsymbol{X}_i\| + 3M_1$ for all $i \in [n]$,*

$$L_{\boldsymbol{W}} := \frac{1}{n}\sum_{i=1}^n a_i b_i, \qquad L_{\boldsymbol{K}} := \|\boldsymbol{Q}\| L_{\boldsymbol{W}}, \qquad \text{and} \qquad L_{\boldsymbol{Q}} := \|\boldsymbol{K}\| L_{\boldsymbol{W}}. \tag{9}$$

Proof. Let

$$\boldsymbol{\gamma}_i = Y_i \cdot \boldsymbol{X}_i\boldsymbol{v}, \quad \boldsymbol{h}_i = \boldsymbol{X}_i\boldsymbol{W}z_i. \tag{10}$$

From Assumption A, $\ell : \mathbb{R} \to \mathbb{R}$ is differentiable. Hence, the gradient evaluated at $\boldsymbol{W}$ is given by

$$\nabla\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n \ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\boldsymbol{h}_i)\right) \cdot \boldsymbol{X}_i^\top \mathbb{S}'(\boldsymbol{h}_i)\boldsymbol{\gamma}_i z_i^\top, \tag{11}$$

where

$$\mathbb{S}'(\boldsymbol{h}) = \text{diag}\left(\mathbb{S}(\boldsymbol{h})\right) - \mathbb{S}(\boldsymbol{h})\mathbb{S}(\boldsymbol{h})^\top \in \mathbb{R}^{T\times T}. \tag{12}$$

Note that

$$\|\mathbb{S}'(\boldsymbol{h})\| \leq \|\mathbb{S}'(\boldsymbol{h})\|_F \leq 1. \tag{13}$$

Hence, for any $\boldsymbol{W}, \dot{\boldsymbol{W}} \in \mathbb{R}^{d\times d}$, $i \in [n]$, we have

$$\left\|\mathbb{S}(\boldsymbol{h}_i) - \mathbb{S}(\dot{\boldsymbol{h}}_i)\right\| \leq \left\|\boldsymbol{h}_i - \dot{\boldsymbol{h}}_i\right\| \leq \|\boldsymbol{X}_i\| \, \|z_i\| \, \left\|\boldsymbol{W} - \dot{\boldsymbol{W}}\right\|_F, \tag{14a}$$

where $\dot{\boldsymbol{h}}_i = \boldsymbol{X}_i \dot{\boldsymbol{W}} z_i$.

Similarly,

$$\begin{aligned}\left\|\mathbb{S}'(\boldsymbol{h}_i) - \mathbb{S}'(\dot{\boldsymbol{h}}_i)\right\|_F &\leq \left\|\mathbb{S}(\boldsymbol{h}_i) - \mathbb{S}(\dot{\boldsymbol{h}}_i)\right\| + \left\|\mathbb{S}(\boldsymbol{h}_i)\mathbb{S}(\boldsymbol{h}_i)^\top - \mathbb{S}(\dot{\boldsymbol{h}}_i)\mathbb{S}(\dot{\boldsymbol{h}}_i)^\top\right\|_F \\ &\leq 3\|\boldsymbol{X}_i\| \, \|z_i\| \, \left\|\boldsymbol{W} - \dot{\boldsymbol{W}}\right\|_F. \end{aligned} \tag{14b}$$

Next, for any $\boldsymbol{W}, \dot{\boldsymbol{W}} \in \mathbb{R}^{d\times d}$, we get

$$\begin{aligned}\left\|\nabla\mathcal{L}(\boldsymbol{W}) - \nabla\mathcal{L}(\dot{\boldsymbol{W}})\right\|_F &\leq \frac{1}{n}\sum_{i=1}^n \left\|\ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\boldsymbol{h}_i)\right) \cdot z_i \boldsymbol{\gamma}_i^\top \mathbb{S}'(\boldsymbol{h}_i)\boldsymbol{X}_i - \ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\dot{\boldsymbol{h}}_i)\right) \cdot z_i \boldsymbol{\gamma}_i^\top \mathbb{S}'(\dot{\boldsymbol{h}}_i)\boldsymbol{X}_i\right\|_F \\
&\leq \frac{1}{n}\sum_{i=1}^n \left\|z_i \boldsymbol{\gamma}_i^\top \mathbb{S}'(\dot{\boldsymbol{h}}_i)\boldsymbol{X}_i\right\|_F \, \left|\ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\boldsymbol{h}_i)\right) - \ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\dot{\boldsymbol{h}}_i)\right)\right| \\
&+ \frac{1}{n}\sum_{i=1}^n \left|\ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\boldsymbol{h}_i)\right)\right| \, \left\|z_i \boldsymbol{\gamma}_i^\top \mathbb{S}'(\boldsymbol{h}_i)\boldsymbol{X}_i - z_i \boldsymbol{\gamma}_i^\top \mathbb{S}'(\dot{\boldsymbol{h}}_i)\boldsymbol{X}_i\right\|_F \\
&\leq \frac{1}{n}\sum_{i=1}^n M_0 \, \|\boldsymbol{\gamma}_i\|^2 \, \|z_i\| \, \|\boldsymbol{X}_i\| \, \left\|\mathbb{S}(\boldsymbol{h}_i) - \mathbb{S}(\dot{\boldsymbol{h}}_i)\right\| \\
&+ \frac{1}{n}\sum_{i=1}^n M_1 \, \|\boldsymbol{\gamma}_i\| \, \|z_i\| \, \|\boldsymbol{X}_i\| \, \left\|\mathbb{S}'(\boldsymbol{h}_i) - \mathbb{S}'(\dot{\boldsymbol{h}}_i)\right\|_F, \end{aligned} \tag{15}$$

where the second inequality follows from the fact that $|ab - cd| \leq |d||a - c| + |a||b - d|$ and the third inequality uses Assumption A and (13).

Substituting (14a) and (14b) into (15), we get

$$\begin{aligned}\left\|\nabla\mathcal{L}(\boldsymbol{W}) - \nabla\mathcal{L}(\dot{\boldsymbol{W}})\right\|_F &\leq \frac{1}{n}\sum_{i=1}^n \left(M_0 \, \|\boldsymbol{\gamma}_i\|^2\|z_i\|^2\|\boldsymbol{X}_i\|^2 + 3M_1\|\boldsymbol{\gamma}_i\| \, \|z_i\|^2 \, \|\boldsymbol{X}_i\|^2\right)\|\boldsymbol{W} - \dot{\boldsymbol{W}}\|_F \\
&\leq \frac{1}{n}\sum_{i=1}^n \left(M_0 \, \|\boldsymbol{v}\|^2\|z_i\|^2\|\boldsymbol{X}_i\|^4 + 3M_1\|\boldsymbol{v}\| \, \|z_i\|^2 \, \|\boldsymbol{X}_i\|^3\right)\|\boldsymbol{W} - \dot{\boldsymbol{W}}\|_F \\
&\leq L_{\boldsymbol{W}} \, \|\boldsymbol{W} - \dot{\boldsymbol{W}}\|_F, \end{aligned}$$

where $L_{\boldsymbol{W}}$ is defined in (9).

Let $\boldsymbol{g}_i = \boldsymbol{X}_i \boldsymbol{K}\boldsymbol{Q}^\top z_i$. We have

$$\nabla_{\boldsymbol{K}}\mathcal{L}(\boldsymbol{K}, \boldsymbol{Q}) = \frac{1}{n}\sum_{i=1}^n \ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\boldsymbol{g}_i)\right) \cdot z_i \boldsymbol{\gamma}_i^\top \mathbb{S}'(\boldsymbol{g}_i)\boldsymbol{X}_i\boldsymbol{Q}, \tag{16a}$$

$$\nabla_{\boldsymbol{Q}}\mathcal{L}(\boldsymbol{K}, \boldsymbol{Q}) = \frac{1}{n}\sum_{i=1}^n \ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\boldsymbol{g}_i)\right) \cdot \boldsymbol{X}_i^\top \mathbb{S}'(\boldsymbol{g}_i)\boldsymbol{\gamma}_i z_i^\top \boldsymbol{K}. \tag{16b}$$

By the similar argument as in (15), for any $\boldsymbol{Q}$ and $\dot{\boldsymbol{Q}} \in \mathbb{R}^{d\times m}$, we have

$$\begin{aligned}\left\|\nabla_{\boldsymbol{Q}}\mathcal{L}(\boldsymbol{K}, \boldsymbol{Q}) - \nabla_{\boldsymbol{Q}}\mathcal{L}(\boldsymbol{K}, \dot{\boldsymbol{Q}})\right\|_F &\leq \frac{\|\boldsymbol{K}\|}{n}\sum_{i=1}^n \left\|\ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\boldsymbol{h}_i)\right) \cdot z_i \boldsymbol{\gamma}_i^\top \mathbb{S}'(\boldsymbol{h}_i)\boldsymbol{X}_i - \ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\dot{\boldsymbol{h}}_i)\right) \cdot z_i \boldsymbol{\gamma}_i^\top \mathbb{S}'(\dot{\boldsymbol{h}}_i)\boldsymbol{X}_i\right\|_F \\
&\leq L_{\boldsymbol{W}}\|\boldsymbol{K}\| \, \|\boldsymbol{Q} - \dot{\boldsymbol{Q}}\|_F. \end{aligned} \tag{17}$$

Similarly, for any $\boldsymbol{K}, \dot{\boldsymbol{K}} \in \mathbb{R}^{d\times m}$, we get

$$\left\|\nabla_{\boldsymbol{K}}\mathcal{L}(\boldsymbol{K}, \boldsymbol{Q}) - \nabla_{\boldsymbol{K}}\mathcal{L}(\dot{\boldsymbol{K}}, \boldsymbol{Q})\right\|_F \leq L_{\boldsymbol{W}}\|\boldsymbol{Q}\| \, \|\boldsymbol{K} - \dot{\boldsymbol{K}}\|_F.$$

■

C.3 Useful Lemmas

Lemma 5 (Optimal Tokens Minimize Training Loss) *Suppose Assumption A (i)-(ii) hold, and not all tokens are optimal per Definition 1. Then, training risk obeys* $\mathcal{L}(\boldsymbol{W}) > \mathcal{L}_\star := \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\gamma}_{i\mathrm{opt}_i})$. *Additionally, suppose there are optimal indices* $(\mathrm{opt}_i)_{i=1}^n$ *for which (Att-SVM) is feasible, i.e. there exists a* $\boldsymbol{W}$ *separating optimal tokens. This* $\boldsymbol{W}$ *choice obeys* $\lim_{R\to\infty} \mathcal{L}(R\cdot\boldsymbol{W}) = \mathcal{L}_\star$.

The result presented in Lemma 5 originates from the observation that the output tokens of the attention layer constitute a convex combination of the input tokens. Consequently, when subjected to a strictly decreasing loss function, attention optimization inherently leans towards the selection of a singular token, specifically, the optimal token $(\mathrm{opt}_i)_{i=1}^n$.

Proof. The token at the output of the attention layer is given by $\boldsymbol{a}_i = \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}\boldsymbol{z}_i)$. Here, $\boldsymbol{a}_i$ can be written as $\boldsymbol{a}_i = \sum_{t\in[T]} c_{it}\boldsymbol{x}_{it}$ where $c_{it} \ge 0$ and $\sum_{t\in[T]} c_{it} = 1$. Note that, for any finite $\boldsymbol{W}$, c_{it} as softmax probabilities are strictly positive. To proceed, using the linearity of $h(\boldsymbol{x}) = \boldsymbol{v}^\top\boldsymbol{x}$ and strictly-decreasing nature of the loss ℓ, we find that

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i\cdot h(\boldsymbol{a}_i)) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i\cdot \sum_{t\in[T]} c_{it} h(\boldsymbol{x}_{it})) \ge \frac{1}{n}\sum_{i=1}^n \ell(Y_i\cdot h(\boldsymbol{x}_{i\mathrm{opt}_i})) = \mathcal{L}_\star,$$

which implies that $\mathcal{L}(\boldsymbol{W}) \ge \mathcal{L}_\star$ for any $\boldsymbol{W}$.

On the other hand, since not all tokens are optimal, there exists a token index (i,t) for which $Y_i\cdot h(\boldsymbol{x}_{it}) < Y_i\cdot h(\boldsymbol{x}_{i\mathrm{opt}_i})$. Since all softmax entries obey $c_{it} > 0$ for finite $\boldsymbol{W}$, this implies the strict inequality $\ell(Y_i\cdot h(\boldsymbol{a}_i)) > \ell(Y_i\cdot h(\boldsymbol{x}_{i\mathrm{opt}_i}))$. This leads to the desired conclusion $\mathcal{L}(\boldsymbol{W}) > \mathcal{L}_\star$.

Secondly, suppose (Att-SVM) is feasible i.e. there exists a $\boldsymbol{W}$ separating some optimal indices $(\mathrm{opt}_i)_{i=1}^n$ from the other tokens. Note that, this does not exclude the existence of other optimal indices. This implies that, letting $\lim_{R\to\infty}\mathbb{S}(\boldsymbol{X}_i(R\cdot\boldsymbol{W})\boldsymbol{z}_i)$ saturates the softmax and will be equal to the indicator function at opt_i for all inputs $i\in[n]$. Thus, $c_{it}\to 0$ for $t\neq \mathrm{opt}_i$ and $c_{it}\to 1$ for $t = \mathrm{opt}_i$. Using M_1-Lipschitzness of ℓ, we can write

$$\left|\ell(Y_i\cdot h(\boldsymbol{x}_{i\mathrm{opt}_i})) - \ell(Y_i\cdot h(\boldsymbol{a}_i))\right| \le M_1\left|h(\boldsymbol{a}_i) - h(\boldsymbol{x}_{i\mathrm{opt}_i})\right|.$$

Since h is linear, it is $\|\boldsymbol{v}\|$-Lipschitz implying

$$\left|\ell(Y_i\cdot h(\boldsymbol{x}_{i\mathrm{opt}_i})) - \ell(Y_i\cdot h(\boldsymbol{a}_i))\right| \le M_1\|\boldsymbol{v}\|\cdot\|\boldsymbol{a}_i - \boldsymbol{x}_{i\mathrm{opt}_i}\|.$$

Since $\boldsymbol{a}_i \to \boldsymbol{x}_{i\mathrm{opt}_i}$ as $R\to\infty$, we conclude with the advertised result. ■ ■

Lemma 6 *For any* $\boldsymbol{X}\in\mathbb{R}^{T\times d}$, $\boldsymbol{W},\boldsymbol{V}\in\mathbb{R}^{d\times d}$ *and* $\boldsymbol{z},\boldsymbol{v}\in\mathbb{R}^d$, *let* $\boldsymbol{a} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{z}$, $\boldsymbol{s} = \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{z})$, *and* $\boldsymbol{\gamma} = \boldsymbol{X}\boldsymbol{v}$. *Set*

$$\Gamma = \sup_{t,\tau\in[T]} |\boldsymbol{\gamma}_t - \boldsymbol{\gamma}_\tau| \quad\text{and}\quad A = \sup_{t\in[T]} \|\boldsymbol{a}_t\|.$$

We have that

$$\left|\boldsymbol{a}^\top\mathrm{diag}(\boldsymbol{s})\boldsymbol{\gamma} - \boldsymbol{a}^\top\boldsymbol{s}\boldsymbol{s}^\top\boldsymbol{\gamma} - \sum_{t\ge 2}^T(\boldsymbol{a}_1 - \boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)\right| \le 2\Gamma A(1-\boldsymbol{s}_1)^2.$$

Proof. The proof is similar to [TLZO23, Lemma 4], but for the sake of completeness, we provide it here. Set $\bar{\gamma} = \sum_{t=1}^T \boldsymbol{\gamma}_t\boldsymbol{s}_t$. We have

$$\boldsymbol{\gamma}_1 - \bar{\gamma} = \sum_{t\ge 2}^T(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)\boldsymbol{s}_t,\ \ \text{and}\ \ |\boldsymbol{\gamma}_1 - \bar{\gamma}| \le \Gamma(1-\boldsymbol{s}_1).$$

Then,

$$\begin{aligned}\boldsymbol{a}^\top\mathrm{diag}(\boldsymbol{s})\boldsymbol{\gamma} - \boldsymbol{a}^\top\boldsymbol{s}\boldsymbol{s}^\top\boldsymbol{\gamma} &= \sum_{t=1}^T \boldsymbol{a}_t\boldsymbol{\gamma}_t\boldsymbol{s}_t - \sum_{t=1}^T\boldsymbol{a}_t\boldsymbol{s}_t\sum_{t=1}^T\boldsymbol{\gamma}_t\boldsymbol{s}_t\\ &= \boldsymbol{a}_1\boldsymbol{s}_1(\boldsymbol{\gamma}_1 - \bar{\gamma}) - \sum_{t\ge 2}^T \boldsymbol{a}_t\boldsymbol{s}_t(\bar{\gamma} - \boldsymbol{\gamma}_t). \qquad (18)\end{aligned}$$

Since

$$\left|\sum_{t\geq 2}^{T} \boldsymbol{a}_t \boldsymbol{s}_t(\bar{\gamma}-\boldsymbol{\gamma}_t) - \sum_{t\geq 2}^{T} \boldsymbol{a}_t \boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)\right| \leq A\Gamma(1-\boldsymbol{s}_1)^2,$$

we obtain[1]

$$\begin{aligned}
\boldsymbol{a}^\top \mathrm{diag}(\boldsymbol{s})\boldsymbol{\gamma} - \boldsymbol{a}^\top \boldsymbol{s}\boldsymbol{s}^\top \boldsymbol{\gamma} &= \boldsymbol{a}_1 \boldsymbol{s}_1(\boldsymbol{\gamma}_1 - \bar{\gamma}) - \sum_{t\geq 2}^{T} \boldsymbol{a}_t \boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t) \pm A\Gamma(1-\boldsymbol{s}_1)^2 \\
&= \boldsymbol{a}_1 \boldsymbol{s}_1 \sum_{t\geq 2}^{T} (\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)\boldsymbol{s}_t - \sum_{t\geq 2}^{T} \boldsymbol{a}_t \boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t) \pm A\Gamma(1-\boldsymbol{s}_1)^2 \\
&= \sum_{t\geq 2}^{T} (\boldsymbol{a}_1\boldsymbol{s}_1 - \boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t) \pm A\Gamma(1-\boldsymbol{s}_1)^2 \\
&= \sum_{t\geq 2}^{T} (\boldsymbol{a}_1 - \boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t) \pm 2A\Gamma(1-\boldsymbol{s}_1)^2.
\end{aligned}$$

Here, $\pm$ on the right handside uses the fact that

$$\left|\sum_{t\geq 2}^{T} (\boldsymbol{a}_1\boldsymbol{s}_1 - \boldsymbol{a}_1)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)\right| \leq (1-\boldsymbol{s}_1)A\Gamma\sum_{t\geq 2}^{T} \boldsymbol{s}_t = (1-\boldsymbol{s}_1)^2 A\Gamma.$$

■

D Global Regularization Path

D.1 Proof of Theorem 1

Throughout $\diamond$ denotes either Frobenius norm or nuclear norm. We will prove that $\bar{\boldsymbol{W}}(R)$ asymptotically aligns with the set of globally-optimal directions and also $\|\bar{\boldsymbol{W}}(R)\|_\diamond \to \infty$. $\mathcal{R}_m \subseteq \mathbb{R}^{d\times d}$ denote the manifold of rank $\leq m$ matrices.

Step 1: Let us first prove that $\bar{\boldsymbol{W}}(R)$ achieves the optimal risk as $R \to \infty$ – rather than problem having finite optima. Define $\Xi_\diamond = 1/\|\boldsymbol{W}^{mm}\|_\diamond$ and norm-normalized $\bar{\boldsymbol{W}}^{mm} = \Xi_\diamond \boldsymbol{W}^{mm}$. Note that $\boldsymbol{W}^{mm}$ separates tokens $\texttt{opt}$ from rest of the tokens for each $i \in [n]$. Thus, we have that

$$\lim_{R\to\infty} \mathcal{L}(\bar{\boldsymbol{W}}(R)) \leq \lim_{R\to\infty} \mathcal{L}(R\cdot \bar{\boldsymbol{W}}^{mm}) := \mathcal{L}_\star = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\gamma}_i^{\texttt{opt}}). \tag{19}$$

On the other hand, for any $\boldsymbol{W} \in \mathcal{R}_m$, define the softmax probabilities $\boldsymbol{s}^{(i)} = \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}\boldsymbol{z}_i)$ and attention features $\boldsymbol{x}_i^{\boldsymbol{W}} = \sum_{t=1}^{T} \boldsymbol{s}_t^{(i)}\boldsymbol{x}_t$. Decompose $\boldsymbol{x}_i^{\boldsymbol{W}}$ as $\boldsymbol{x}_i^{\boldsymbol{W}} = \boldsymbol{s}_{\texttt{opt}_i}^{(i)}\boldsymbol{x}_{i\texttt{opt}_i} + \sum_{t\neq \texttt{opt}_i} \boldsymbol{s}_t^{(i)}\boldsymbol{x}_{it}$. Set $\gamma_{it}^{gap} = \boldsymbol{\gamma}_i^{\texttt{opt}} - \boldsymbol{\gamma}_{it} = Y_i\cdot \boldsymbol{v}^\top(\boldsymbol{x}_{i\texttt{opt}_i} - \boldsymbol{x}_{it}) > 0$, and define

$$B := \max_{i\in[n]} \max_{t,\tau\in[T]} \|\boldsymbol{v}\|\cdot\|\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau}\| \geq \gamma_{it}^{gap}. \tag{20}$$

Define $c_{\texttt{opt}} = \min_{i\in[n],t\neq\texttt{opt}_i} \gamma_{it}^{gap} > 0$ and $\boldsymbol{\gamma}_i^{\boldsymbol{W}} = Y_i\cdot\boldsymbol{v}^\top \boldsymbol{x}_i^{\boldsymbol{W}}$. We obtain the following score inequalities

$$\begin{aligned}
&\boldsymbol{\gamma}_i^{\boldsymbol{W}} \leq \boldsymbol{\gamma}_i^{\texttt{opt}} - c_{\texttt{opt}}(1-\boldsymbol{s}_{\texttt{opt}_i}^{(i)}) < \boldsymbol{\gamma}_i^{\texttt{opt}}, \\
&|\boldsymbol{\gamma}_i^{\boldsymbol{W}} - \boldsymbol{\gamma}_i^{\texttt{opt}}| \leq \|\boldsymbol{v}\|\cdot\|\boldsymbol{x}_i^{\boldsymbol{W}} - \boldsymbol{x}_i^{\alpha}\| \leq \|\boldsymbol{v}\| \sum_{t\neq\texttt{opt}_i} \boldsymbol{s}_t^{(i)}\|\boldsymbol{x}_{it} - \boldsymbol{x}_i^{\alpha}\| \leq B(1-\boldsymbol{s}_{\texttt{opt}_i}^{(i)}).
\end{aligned} \tag{21}$$

We will use the $\boldsymbol{\gamma}_i^{\boldsymbol{W}} - \boldsymbol{\gamma}_i^{\texttt{opt}}$ term in (21) to evaluate $\boldsymbol{W}$ against the reference loss $\mathcal{L}_\star$ of (19). Using the strictly-decreasing nature of ℓ, we conclude with the fact that for all (finite) $\boldsymbol{W} \in \mathcal{R}_m$,

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\gamma}_i^{\boldsymbol{W}}) > \mathcal{L}_\star = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\gamma}_i^{\texttt{opt}}),$$

[1]For simplicity, we use $\pm$ on the right hand side to denote the upper and lower bounds.

which implies $\|\bar{W}(R)\|_\diamond \to \infty$ together with (19).

Step 2: To proceed, we show that $\bar{W}(R)$ converges in direction to $\mathcal{W}^{mm}$, which denotes the set of SVM minima. Suppose this is not the case and convergence fails. We will obtain a contradiction by showing that $\bar{W}_R^{mm} = R \cdot \bar{W}^{mm}$ achieves a strictly superior loss compared to $\bar{W}(R)$. Let us introduce the normalized parameters $\bar{W}_0(R) = \frac{\bar{W}(R)}{R\Xi_\diamond}$ and $W' = \frac{\bar{W}(R)}{\|\bar{W}(R)\|_\diamond \Xi_\diamond}$. Note that $\bar{W}_0(R)$ is obtained by scaling down W' since $\|\bar{W}(R)\|_\diamond \le R$ and W' obeys $\|W'\|_\diamond = \|W^{mm}\|_\diamond$. Since $\bar{W}_0(R)$ fails to converge to $\mathcal{W}^{mm}$, for some $\delta > 0$, there exists arbitrarily large $R > 0$ such that $\texttt{dist}\left(\bar{W}_0(R), \mathcal{W}^{mm}\right) \ge \delta$. This translates to the suboptimality in terms of the margin constraints as follows: First, since nuclear norm dominates Frobenius, distance with respect to the $\diamond$-norm obeys $\texttt{dist}_\diamond\left(\bar{W}_0(R), \mathcal{W}^{mm}\right) \ge \delta$. Secondly, using triangle inequality,

$$\text{this implies that either} \quad \|\bar{W}_0(R)\|_\diamond \le \|W^{mm}\|_\diamond - \delta/2 \quad \text{or} \quad \texttt{dist}_\diamond(W', \mathcal{W}^{mm}) \ge \delta/2.$$

In either scenario, $\bar{W}_0(R)$ strictly violates one of the margin constraints of (Att-SVM) ($\diamond = F$) or (Att-SVM$_\star$) ($\diamond = \star$): If $\|\bar{W}_0(R)\|_\diamond \le \|W^{mm}\|_\diamond - \delta/2$, then, since the optimal SVM objective is $\|W^{mm}\|_\diamond$, there exists a constraint $i, t \neq \texttt{opt}_i$ for which $\left\langle (x_i^{\texttt{opt}} - x_{it}) z_i^\top, \bar{W}_0(R)\right\rangle \le 1 - \frac{\delta}{2\|W^{mm}\|_\diamond}$. If $\texttt{dist}_\diamond(W', \mathcal{W}^{mm}) \ge \delta/2$, then, W' has the same SVM objective but it is strictly bounded away from the solution set. Thus, for some $\epsilon := \epsilon(\delta) > 0$, W' and its scaled down version $\bar{W}_0(R)$ strictly violate an SVM constraint achieving margin $\le 1 - \epsilon$. Without losing generality, suppose $\bar{W}_0(R)$ violates the first constraint $i = 1$. Thus, for a properly updated $\delta > 0$ (that is function of the initial $\delta > 0$) and for $i = 1$ and some support index $\tau \in \mathcal{T}_1$,

$$\left\langle (x_1^{\texttt{opt}} - x_{1t}) z_1^\top, \bar{W}_0(R)\right\rangle \le 1 - \delta. \tag{22}$$

Now, we will argue that this leads to a contradiction by proving $\mathcal{L}(\bar{W}_R^{mm}) < \mathcal{L}(\bar{W}(R))$ for sufficiently large R.

To obtain the result, we establish a refined softmax probability control as in Step 1 by studying distance to $\mathcal{L}_\star$. Following (21), denote the score function at $\bar{W}(R)$ via $\gamma_i^R := \gamma_i^{\bar{W}(R)}$. Similarly, let $s_i^R = \mathbb{S}(a_i^R)$ with $a_i^R = X_i \bar{W}(R) z_i$. Set the corresponding notation for the reference parameter $\bar{W}_R^{mm}$ as $\gamma_i^\star, s_i^\star, a_i^\star$. Recall that $R \ge \|\bar{W}(R)\|_\diamond$ and $\Xi_\diamond := 1/\|W^{mm}\|_\diamond$. We note the following softmax inequalities

$$\begin{aligned} s_{i\texttt{opt}_i}^\star &\ge \frac{1}{1 + Te^{-R\Xi_\diamond}} \ge 1 - Te^{-R\Xi_\diamond} \quad \text{for all} \quad i \in [n], \\ s_{i\texttt{opt}_i}^R &\le \frac{1}{1 + e^{-(1-\delta)\|\bar{W}(R)\|_\diamond \Xi_\diamond}} \le \frac{1}{1 + e^{-(1-\delta)R\Xi_\diamond}} \quad \text{for} \quad i = 1. \end{aligned} \tag{23}$$

The former inequality is thanks to W^{mm} achieving $\ge$1 margins on all tokens $[T] - \texttt{opt}_i$ and the latter arises from the δ-margin violation of $\bar{W}(R)$ at $i = 1$ i.e. Eq. (22). Since ℓ is strictly decreasing with Lipschitz derivative and the scores are upper/lower bounded by an absolute constant (as tokens are bounded and fixed), we have that $c_{\text{up}} \ge -\ell'(\gamma_i^W) \ge c_{\text{dn}}$ for some constants $c_{\text{up}} > c_{\text{dn}} > 0$. Thus, following Eq. (20), the score decomposition (21), and (23) we can write

$$\begin{aligned} \mathcal{L}(\bar{W}(R)) - \mathcal{L}_\star &\ge \frac{1}{n}[\ell(\gamma_1^{\bar{W}(R)}) - \ell(\gamma_1^{\texttt{opt}})] \ge \frac{c_{\text{dn}}}{n}(\gamma_1^{\texttt{opt}} - \gamma_1^{\bar{W}(R)}) \\ &\ge \frac{c_{\text{dn}}}{n} c_{\texttt{opt}} (1 - s_{1\texttt{opt}_1}^R). \\ &\ge \frac{c_{\text{dn}} c_{\texttt{opt}}}{n} \frac{1}{1 + e^{(1-\delta)R\Xi_\diamond}}. \end{aligned} \tag{24}$$

Conversely, we upper bound the difference between $\mathcal{L}(\bar{W}_R^{mm})$ and $\mathcal{L}_\star$ as follows. Define the worst-case loss difference for $\bar{W}(R)$ as $j = \arg\max_{i \in [n]}[\ell(\gamma_i^\star) - \ell(\gamma_i^{\texttt{opt}})]$. Using (21)&(23), we write

$$\begin{aligned} \mathcal{L}(\bar{W}_R^{mm}) - \mathcal{L}_\star &\le \max_{i \in [n]}[\ell(\gamma_i^\star) - \ell(\gamma_i^{\texttt{opt}})] \le c_{\text{up}} \cdot (\gamma_j^{\texttt{opt}} - \gamma_j^\star) \\ &\le c_{\text{up}} \cdot (1 - s_{j\texttt{opt}_j}^\star) B \\ &\le c_{\text{up}} \cdot Te^{-R\Xi_\diamond} B. \end{aligned}$$

Combining the last inequality and (24), we conclude that $\mathcal{L}(\bar{\boldsymbol{W}}_R^{mm}) < \mathcal{L}(\bar{\boldsymbol{W}}(R))$ whenever

$$c_{\text{up}}T \cdot e^{-R\Xi_\diamond}B < \frac{c_{\text{dn}} \cdot c_{\text{opt}}}{n}\frac{1}{1+e^{(1-\delta)R\Xi_\diamond}} \iff \frac{e^{R\Xi_\diamond}}{1+e^{(1-\delta)R\Xi_\diamond}} > \frac{c_{\text{up}}TnB}{c_{\text{dn}}c_{\text{opt}}}.$$

The left hand-side inequality holds for all sufficiently large R: Specifically, as soon as R obeys $R > \frac{1}{\delta\Xi_\diamond}\log(\frac{2c_{\text{up}}TnB}{c_{\text{dn}}c_{\text{opt}}})$. This completes the proof of the theorem by contradiction since we obtained $\mathcal{L}(\bar{\boldsymbol{W}}(R)) > \mathcal{L}(\bar{\boldsymbol{W}}_R^{mm})$.

E Convergence of Gradient Descent

Optimization problem definition. Recap the problem, where we use a linear head $h(\boldsymbol{x}) = \boldsymbol{v}^\top \boldsymbol{x}$ for most of our theoretical exposition. Given dataset $(Y_i, \boldsymbol{X}_i, \boldsymbol{z}_i)_{i=1}^n$, we minimize the empirical risk of an 1-layer transformer using combined weights $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ or individual weights $\boldsymbol{K}, \boldsymbol{Q} \in \mathbb{R}^{d\times m}$ for a fixed head and decreasing loss function:

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{z}_i)\right), \tag{W-ERM}$$

$$\mathcal{L}(\boldsymbol{K}, \boldsymbol{Q}) = \frac{1}{n}\sum_{i=1}^n \ell\left(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i \boldsymbol{K}\boldsymbol{Q}^\top \boldsymbol{z}_i)\right). \tag{KQ-ERM}$$

We can recover the self-attention model by setting $\boldsymbol{z}_i$ to be the first token of $\boldsymbol{X}_i$, i.e., $\boldsymbol{z}_i \leftarrow \boldsymbol{x}_{i1}$.

E.1 Divergence of norm of the iterates $\mathbf{W}(k)$

The next lemma establishes the descent property of gradient descent for $\mathcal{L}(\boldsymbol{W})$ under Assumption A.

Lemma 7 (Descent Lemma) *Under Assumption A, if $\eta \le 1/L_{\boldsymbol{W}}$, then for any initialization $\boldsymbol{W}(0)$, Algorithm W-GD satisfies:*

$$\mathcal{L}(\boldsymbol{W}(k+1)) - \mathcal{L}(\boldsymbol{W}(k)) \le -\frac{\eta}{2}\|\nabla\mathcal{L}(\boldsymbol{W}(k))\|_F^2, \tag{25}$$

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

Proof. The proof is similar to [TLZO23, Lemma 5]. ∎

The lemma below reveals that the correlation between the training loss's gradient at any arbitrary matrix $\boldsymbol{W}$ and the attention SVM solution $\boldsymbol{W}^{mm}$ is negative. Consequently, for any finite $\boldsymbol{W}$, $\langle\nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{mm}\rangle$ cannot be equal to zero.

Lemma 8 *Let $\boldsymbol{W}^{mm}$ be the SVM solution of (Att-SVM). Suppose Assumptions A and B hold. Then, for all $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, the training loss (W-ERM) obeys $\langle\nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{mm}\rangle \le -c < 0$, for some constant $c > 0$ (see (34)) depending on the data, the head $\mathbf{v}$, and a loss derivative bound.*

Proof. Let

$$\bar{\boldsymbol{h}}_i = \boldsymbol{X}_i \boldsymbol{W}^{mm} \boldsymbol{z}_i, \quad \boldsymbol{\gamma}_i = Y_i \cdot \boldsymbol{X}_i \boldsymbol{v}, \quad \text{and} \quad \boldsymbol{h}_i = \boldsymbol{X}_i \boldsymbol{W} \boldsymbol{z}_i. \tag{26}$$

Let us recall the gradient evaluated at $\boldsymbol{W}$ which is given by

$$\nabla\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n \ell'\left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\boldsymbol{h}_i)\right) \cdot \boldsymbol{X}_i^\top \mathbb{S}'(\boldsymbol{h}_i)\boldsymbol{\gamma}_i \boldsymbol{z}_i^\top, \tag{27}$$

which implies that

$$
\begin{aligned}
\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{mm} \rangle &= \frac{1}{n} \sum_{i=1}^{n} \ell' \left(\boldsymbol{\gamma}_i^\top \mathbb{S}(\boldsymbol{h}_i) \right) \cdot \left\langle \boldsymbol{X}_i^\top \mathbb{S}'(\boldsymbol{h}_i) \boldsymbol{\gamma}_i \boldsymbol{z}_i^\top, \boldsymbol{W}^{mm} \right\rangle \\
&= \frac{1}{n} \sum_{i=1}^{n} \ell_i' \cdot \operatorname{trace}\left((\boldsymbol{W}^{mm})^\top \boldsymbol{X}_i^\top \mathbb{S}'(\boldsymbol{h}_i) \boldsymbol{\gamma}_i \boldsymbol{z}_i^\top \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \ell_i' \cdot \bar{\boldsymbol{h}}_i^\top \mathbb{S}'(\boldsymbol{h}_i) \boldsymbol{\gamma}_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \ell_i' \cdot \left(\bar{\boldsymbol{h}}_i^\top \operatorname{diag}(\boldsymbol{s}_i) \boldsymbol{\gamma}_i - \bar{\boldsymbol{h}}_i^\top \boldsymbol{s}_i \boldsymbol{s}_i^\top \boldsymbol{\gamma}_i \right).
\end{aligned}
\tag{28}
$$

Here, let $\ell_i' := \ell'(\boldsymbol{\gamma}_i^\top \mathbb{S}(\boldsymbol{h}_i))$, $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{h}_i)$ and the third equality uses $\operatorname{trace}(\boldsymbol{b}\boldsymbol{a}^\top) = \boldsymbol{a}^\top \boldsymbol{b}$.

In order to move forward, we will establish the following result, with a focus on the equal score condition (Assumption (i)): Let $\gamma = \boldsymbol{\gamma}_{t\geq 2}$ be a constant, and let $\boldsymbol{\gamma}_1$ and $\bar{\boldsymbol{h}}_1$ represent the largest indices of vectors $\boldsymbol{\gamma}$ and $\bar{\boldsymbol{h}}$ respectively. For any vector $\boldsymbol{s}$ that satisfies $\sum_{t\in[T]} \boldsymbol{s}_t = 1$ and $\boldsymbol{s}_t > 0$, we aim to prove that $\bar{\boldsymbol{h}}^\top \operatorname{diag}(\boldsymbol{s})\boldsymbol{\gamma} - \bar{\boldsymbol{h}}^\top \boldsymbol{s}\boldsymbol{s}^\top \boldsymbol{\gamma} > 0$. To demonstrate this, we proceed by writing the following:

$$
\begin{aligned}
\bar{\boldsymbol{h}}^\top \operatorname{diag}(\boldsymbol{s})\boldsymbol{\gamma} - \bar{\boldsymbol{h}}^\top \boldsymbol{s}\boldsymbol{s}^\top \boldsymbol{\gamma} &= \sum_{t=1}^{T} \bar{\boldsymbol{h}}_t \boldsymbol{\gamma}_t \boldsymbol{s}_t - \sum_{t=1}^{T} \bar{\boldsymbol{h}}_t \boldsymbol{s}_t \sum_{t=1}^{T} \boldsymbol{\gamma}_t \boldsymbol{s}_t \\
&= \left(\bar{\boldsymbol{h}}_1 \boldsymbol{\gamma}_1 \boldsymbol{s}_1 + \gamma \sum_{t\geq 2}^{T} \bar{\boldsymbol{h}}_t \boldsymbol{s}_t \right) - \left(\boldsymbol{\gamma}_1 \boldsymbol{s}_1 + \gamma(1 - \boldsymbol{s}_1) \right) \left(\bar{\boldsymbol{h}}_1 \boldsymbol{s}_1 + \sum_{t\geq 2}^{T} \bar{\boldsymbol{h}}_t \boldsymbol{s}_t \right) \\
&= \bar{\boldsymbol{h}}_1 (\boldsymbol{\gamma}_1 - \gamma) \boldsymbol{s}_1 (1 - \boldsymbol{s}_1) - (\boldsymbol{\gamma}_1 - \gamma) \boldsymbol{s}_1 \sum_{t\geq 2}^{T} \bar{\boldsymbol{h}}_t \boldsymbol{s}_t \\
&= (\boldsymbol{\gamma}_1 - \gamma)(1 - \boldsymbol{s}_1)\boldsymbol{s}_1 \left[\bar{\boldsymbol{h}}_1 - \frac{\sum_{t\geq 2}^{T} \bar{\boldsymbol{h}}_t \boldsymbol{s}_t}{\sum_{t\geq 2}^{T} \boldsymbol{s}_t} \right] \\
&\geq (\boldsymbol{\gamma}_1 - \gamma)(1 - \boldsymbol{s}_1)\boldsymbol{s}_1 (\bar{\boldsymbol{h}}_1 - \max_{t\geq 2} \bar{\boldsymbol{h}}_t).
\end{aligned}
\tag{29}
$$

To proceed, define

$$
\gamma_{gap}^i = \boldsymbol{\gamma}_{i\mathrm{opt}_i} - \max_{t\neq \mathrm{opt}_i} \boldsymbol{\gamma}_{it} \quad \text{and} \quad \bar{h}_{gap}^i = \bar{\boldsymbol{h}}_{i\mathrm{opt}_i} - \max_{t\neq \mathrm{opt}_i} \bar{\boldsymbol{h}}_{it}.
$$

With these, we obtain

$$
\bar{\boldsymbol{h}}_i^\top \operatorname{diag}(\boldsymbol{s}_i)\boldsymbol{\gamma}_i - \bar{\boldsymbol{h}}_i^\top \boldsymbol{s}_i \boldsymbol{s}_i^\top \boldsymbol{\gamma}_i \geq \gamma_{gap}^i \bar{h}_{gap}^i (1 - \boldsymbol{s}_{i\mathrm{opt}_i}) \boldsymbol{s}_{i\mathrm{opt}_i}. \tag{30}
$$

Note that

$$
\begin{aligned}
\bar{h}_{gap}^i &= \min_{t\neq \mathrm{opt}_i} (\boldsymbol{x}_{i\mathrm{opt}_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W}^{mm} \boldsymbol{z}_i \geq 1, \\
\gamma_{gap}^i &= \min_{t\neq \mathrm{opt}_i} \boldsymbol{\gamma}_{i\mathrm{opt}_i} - \boldsymbol{\gamma}_{it} > 0, \\
\boldsymbol{s}_{i\mathrm{opt}_i} & (1 - \boldsymbol{s}_{i\mathrm{opt}_i}) > 0.
\end{aligned}
$$

Hence,

$$
c_0 := \min_{i\in[n]} \left\{ \left(\min_{t\neq \mathrm{opt}_i} (\boldsymbol{x}_{i\mathrm{opt}_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W}^{mm} \boldsymbol{z}_i \right) \cdot \left(\min_{t\neq \mathrm{opt}_i} \boldsymbol{\gamma}_{i\mathrm{opt}_i} - \boldsymbol{\gamma}_{it} \right) \cdot \boldsymbol{s}_{i\mathrm{opt}_i} (1 - \boldsymbol{s}_{i\mathrm{opt}_i}) \right\} > 0. \tag{31}
$$

It follows from (30) and (31) that

$$
\min_{i\in[n]} \left\{ \bar{\boldsymbol{h}}_i^\top \operatorname{diag}(\boldsymbol{s}_i)\boldsymbol{\gamma}_i - \bar{\boldsymbol{h}}_i^\top \boldsymbol{s}_i \boldsymbol{s}_i^\top \boldsymbol{\gamma}_i \right\} \geq c_0 > 0. \tag{32}
$$

Further, by our assumption $\ell_i' < 0$. Since by Assumption A, ℓ' is continuous and the domain is bounded, the maximum is attained and negative, and thus

$$
-c_1 = \max_x \ell'(x), \qquad \text{for some} \quad c_1 > 0. \tag{33}
$$

Hence, using (32) and (33) in (28), we obtain

$$\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{mm} \rangle \leq -c < 0, \quad \text{where} \quad c = c_1 \cdot c_0. \tag{34}$$

In the scenario that Assumption B(ii) holds (all tokens are support), $\bar{\boldsymbol{h}}_t = \boldsymbol{x}_{it}^\top \boldsymbol{W}^{mm} \boldsymbol{z}_i$ is constant for all $t \geq 2$. Hence, following similar steps as in (29) completes the proof. ■

Theorem 4 *Suppose Assumption A on the loss function ℓ and Assumption B on the tokens hold. Then,*

- *There is no $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ satisfying $\nabla \mathcal{L}(\boldsymbol{W}) = 0$.*
- *Algorithm W-GD with the step size $\eta \leq 1/L_{\boldsymbol{W}}$ and any starting point $\boldsymbol{W}(0)$ satisfies $\lim_{k\to\infty} \|\boldsymbol{W}(k)\|_F = \infty$.*

Proof. It follows from Lemma 7 that under Assumption A, $\eta \leq 1/L_{\boldsymbol{W}}$, and for any initialization $\boldsymbol{W}(0)$, the gradient descent sequence $\boldsymbol{W}(k+1) = \boldsymbol{W}(k) - \eta \nabla \mathcal{L}(\boldsymbol{W}(k))$ satisfies $\lim_{k\to\infty} \|\nabla \mathcal{L}(\boldsymbol{W}(k))\|_F^2 = 0$.

Further, it follows from Lemma 8 that $\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{mm} \rangle < 0$ for all $\boldsymbol{W} \in \mathbb{R}^{d\times d}$. Hence, for any finite $\boldsymbol{W}$, $\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{mm} \rangle$ cannot be equal to zero. Therefore, there are no finite critical points $\boldsymbol{W}$, for which $\nabla \mathcal{L}(\boldsymbol{W}) = 0$ which contradicts Lemma 7. This implies that $\|\boldsymbol{W}(k)\| \to \infty$. ■

E.2 Global Convergence of Gradient Descent

The following lemma illustrates that when non-optimal tokens within an input share the same scores, the negative gradient of the loss function at $\boldsymbol{W}$ becomes more correlated with the max-margin solution ($\boldsymbol{W}^{mm}$) than with $\boldsymbol{W}$ itself.

Lemma 9 *Let $\boldsymbol{W}^{mm}$ be the SVM solution of* (Att-SVM). *Suppose Assummption (i) on the tokens' score hold and $\ell(\cdot)$ is strictly decreasing and differentiable. For any choice of $\pi > 0$, there exists $R := R_\pi$ such that, for any $\boldsymbol{W}$ with $\|\boldsymbol{W}\|_F \geq R$, we have*

$$\left\langle \nabla \mathcal{L}(\boldsymbol{W}), \frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_F} \right\rangle \geq (1+\pi) \left\langle \nabla \mathcal{L}(\boldsymbol{W}), \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle.$$

Proof. Let $\bar{\boldsymbol{W}} = \|\boldsymbol{W}^{mm}\|_F \boldsymbol{W}/\|\boldsymbol{W}\|_F$, $M = \sup_{i,t} \|\boldsymbol{x}_{it} \boldsymbol{z}_i^\top\|$, $\Theta = 1/\|\boldsymbol{W}^{mm}\|_F$, $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{z}_i)$, $\boldsymbol{h}_i = \boldsymbol{X}_i \bar{\boldsymbol{W}} \boldsymbol{z}_i$, $\bar{\boldsymbol{h}}_i = \boldsymbol{X}_i \boldsymbol{W}^{mm} \boldsymbol{z}_i$, and $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_{i,t\geq 2}$. Without losing generality assume $\alpha_i = \texttt{opt}_i = 1$ for all $i \in [n]$. Repeating the proof of Lemma 8 yields

$$\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{W}^{mm} \rangle = \frac{1}{n} \sum_{i=1}^{n} \ell_i' \cdot (\boldsymbol{\gamma}_{i1} - \boldsymbol{\gamma}_i)(1 - \boldsymbol{s}_{i1}) \boldsymbol{s}_{i1} \left[\bar{\boldsymbol{h}}_{i1} - \frac{\sum_{t\geq 2}^{T} \bar{\boldsymbol{h}}_{it} \boldsymbol{s}_{it}}{\sum_{t\geq 2}^{T} \boldsymbol{s}_{it}} \right],$$

$$\left\langle \nabla \mathcal{L}(\boldsymbol{W}), \bar{\boldsymbol{W}} \right\rangle = \frac{1}{n} \sum_{i=1}^{n} \ell_i' \cdot (\boldsymbol{\gamma}_{i1} - \boldsymbol{\gamma}_i)(1 - \boldsymbol{s}_{i1}) \boldsymbol{s}_{i1} \left[\boldsymbol{h}_{i1} - \frac{\sum_{t\geq 2}^{T} \boldsymbol{h}_{it} \boldsymbol{s}_{it}}{\sum_{t\geq 2}^{T} \boldsymbol{s}_{it}} \right].$$

Focusing on a single example $i \in [n]$ with $\boldsymbol{s}, \boldsymbol{h}, \bar{\boldsymbol{h}}$ vectors (dropping subscript i), given π, for sufficiently large R, we wish to show that

$$\left[\boldsymbol{h}_1 - \frac{\sum_{t\geq 2}^{T} \boldsymbol{h}_t \boldsymbol{s}_t}{\sum_{t\geq 2}^{T} \boldsymbol{s}_t} \right] \leq (1+\pi) \cdot \left[\bar{\boldsymbol{h}}_1 - \frac{\sum_{t\geq 2}^{T} \bar{\boldsymbol{h}}_t \boldsymbol{s}_t}{\sum_{t\geq 2}^{T} \boldsymbol{s}_t} \right]. \tag{35}$$

We consider two scenarios.

Scenario 1: $\|\bar{\boldsymbol{W}} - \boldsymbol{W}^{mm}\|_F \leq \epsilon := \pi/(2M)$. In this scenario, for any token, we find that

$$|\boldsymbol{h}_t - \bar{\boldsymbol{h}}_t| = |\boldsymbol{x}_t^\top (\bar{\boldsymbol{W}} - \boldsymbol{W}^{mm}) \boldsymbol{z}_t| \leq M \|\bar{\boldsymbol{W}} - \boldsymbol{W}^{mm}\|_F \leq M\epsilon.$$

Consequently, we obtain

$$\bar{\boldsymbol{h}}_1 - \frac{\sum_{t\geq 2}^{T} \bar{\boldsymbol{h}}_t \boldsymbol{s}_t}{\sum_{t\geq 2}^{T} \boldsymbol{s}_t} \geq \boldsymbol{h}_1 - \frac{\sum_{t\geq 2}^{T} \boldsymbol{h}_t \boldsymbol{s}_t}{\sum_{t\geq 2}^{T} \boldsymbol{s}_t} - 2M\epsilon = \boldsymbol{h}_1 - \frac{\sum_{t\geq 2}^{T} \boldsymbol{h}_t \boldsymbol{s}_t}{\sum_{t\geq 2}^{T} \boldsymbol{s}_t} - \pi.$$

Also noticing $\bar{\boldsymbol{h}}_1 - \frac{\sum_{t\geq 2}^T \bar{\boldsymbol{h}}_t \boldsymbol{s}_t}{\sum_{t\geq 2}^T \boldsymbol{s}_t} \geq 1$ (thanks to $\boldsymbol{W}^{mm}$ satisfying ≥ 1 margin), this implies (35).

Scenario 2: $\|\bar{\boldsymbol{W}} - \boldsymbol{W}^{mm}\|_F \geq \epsilon := \pi/(2M)$. In this scenario, for some $\delta = \delta(\epsilon)$ and $\tau \geq 2$, we have that

$$\boldsymbol{h}_1 - \boldsymbol{h}_\tau \leq 1 - 2\delta.$$

Recall that $\boldsymbol{s} = \mathbb{S}(\bar{R}\boldsymbol{h})$ where $\bar{R} = \|\boldsymbol{W}\|_F / \|\boldsymbol{W}^{mm}\|_F$. To proceed, split the tokens into two groups: Let $\mathcal{N}$ be the group of tokens obeying $(\boldsymbol{x}_1 - \boldsymbol{x}_t)^\top \bar{\boldsymbol{W}} \boldsymbol{z} \geq 1 - \delta$ for $t \in \mathcal{N}$ and $[T] - \mathcal{N}$ be the rest. Observe that

$$\frac{\sum_{t\in\mathcal{N}} \boldsymbol{s}_t}{\sum_{t\geq 2}^T \boldsymbol{s}_t} \leq \frac{\sum_{t\in\mathcal{N}} \boldsymbol{s}_t}{\boldsymbol{s}_\tau} \leq T \frac{e^{\delta\bar{R}}}{e^{2\delta\bar{R}}} = T e^{-\bar{R}\delta}.$$

Set $\bar{M} = M/\Theta$ and note that $\|\boldsymbol{h}_t\| \leq \|\boldsymbol{W}^{mm}\|_F \cdot \|\boldsymbol{x}_t \boldsymbol{z}^\top\| \leq \bar{M}$. Using $(\boldsymbol{x}_1 - \boldsymbol{x}_t)^\top \bar{\boldsymbol{W}} \boldsymbol{z} < 1 - \delta$ over $t \in [T] - \mathcal{N}$ and plugging in the above bound, we obtain

$$\begin{aligned} \frac{\sum_{t\geq 2}^T (\boldsymbol{h}_1 - \boldsymbol{h}_t)\boldsymbol{s}_t}{\sum_{t\geq 2}^T \boldsymbol{s}_t} &= \frac{\sum_{t\in[T]-\mathcal{N}} (\boldsymbol{h}_1 - \boldsymbol{h}_t)\boldsymbol{s}_t}{\sum_{t\geq 2}^T \boldsymbol{s}_t} + \frac{\sum_{t\in\mathcal{N}} (\boldsymbol{h}_1 - \boldsymbol{h}_t)\boldsymbol{s}_t}{\sum_{t\geq 2}^T \boldsymbol{s}_t} \\ &\leq (1-\delta) + 2\bar{M}T e^{-\bar{R}\delta}. \end{aligned}$$

Using the fact that $\bar{\boldsymbol{h}}_1 - \frac{\sum_{t\geq 2}^T \bar{\boldsymbol{h}}_t \boldsymbol{s}_t}{\sum_{t\geq 2}^T \boldsymbol{s}_t} \geq 1$, the above implies (35) with $\pi' = 2\bar{M}T e^{-\bar{R}\delta} - \delta$. To proceed, choose

$$R_\pi = \delta^{-1}\Theta^{-1} \log(\frac{2\bar{M}T}{\pi}) \quad \text{to ensure} \quad \pi' \leq \pi. \tag{36}$$

■

E.2.1 Proof of Theorem 2.

The proof is similar to [TLZO23, Theorem 2]. Given any $\epsilon \in (0,1)$, let $\pi = \epsilon/(1-\epsilon)$. It follows from Theorem 4 that $\lim_{k\to\infty} \|\boldsymbol{W}(k)\|_F = \infty$. Hence, we can choose k_ϵ such that for any $k \geq k_\epsilon$, it holds that $\|\boldsymbol{W}(k)\|_F > R_\epsilon \vee 1/2$ for some parameter R_ϵ. Now for any $k \geq k_\epsilon$, it follows from Lemma 9 that

$$\left\langle -\nabla\mathcal{L}(\boldsymbol{W}(k)), \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle \geq (1-\epsilon) \left\langle -\nabla\mathcal{L}(\boldsymbol{W}(k)), \frac{\boldsymbol{W}(k)}{\|\boldsymbol{W}(k)\|_F} \right\rangle.$$

Multiplying both sides by the stepsize η and using the gradient descent update, we get

$$\begin{aligned} \left\langle \boldsymbol{W}(k+1) - \boldsymbol{W}(k), \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle &\geq (1-\epsilon) \left\langle \boldsymbol{W}(k+1) - \boldsymbol{W}(k), \frac{\boldsymbol{W}(k)}{\|\boldsymbol{W}(k)\|_F} \right\rangle \\ &= \frac{(1-\epsilon)}{2\|\boldsymbol{W}(k)\|_F} \left(\|\boldsymbol{W}(k+1)\|_F^2 - \|\boldsymbol{W}(k)\|_F^2 - \|\boldsymbol{W}(k+1) - \boldsymbol{W}(k)\|_F^2 \right) \\ &\geq (1-\epsilon) \left(\frac{1}{2\|\boldsymbol{W}(k)\|_F} \left(\|\boldsymbol{W}(k+1)\|_F^2 - \|\boldsymbol{W}(k)\|_F^2 \right) - \|\boldsymbol{W}(k+1) - \boldsymbol{W}(k)\|_F^2 \right) \\ &\geq (1-\epsilon) \left(\|\boldsymbol{W}(k+1)\|_F - \|\boldsymbol{W}(k)\|_F - \|\boldsymbol{W}(k+1) - \boldsymbol{W}(k)\|_F^2 \right) \\ &\geq (1-\epsilon) \Big(\|\boldsymbol{W}(k+1)\|_F - \|\boldsymbol{W}(k)\|_F - 2\eta \left(\mathcal{L}(\boldsymbol{W}(k)) - \mathcal{L}(\boldsymbol{W}(k+1)) \right) \Big). \end{aligned} \tag{37}$$

Here, the second inequality is obtained from $\|\boldsymbol{W}(k)\|_F \geq 1/2$; the third inequality follows since for any $a, b > 0$, we have $(a^2 - b^2)/(2b) - (a-b) \geq 0$; and the last inequality uses Lemma 7.

Summing the above inequality over $k \geq k_\epsilon$ gives

$$\left\langle \frac{\boldsymbol{W}(k)}{\|\boldsymbol{W}(k)\|_F}, \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle \geq 1 - \epsilon + \frac{C(\epsilon,\eta)}{\|\boldsymbol{W}(k)\|_F},$$

for some finite constant $C(\epsilon, \eta)$ defined as

$$C(\epsilon,\eta) := \left\langle \boldsymbol{W}(k_\epsilon), \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle - (1-\epsilon)\|\boldsymbol{W}(k_\epsilon)\|_F - 2\eta(1-\epsilon)(\mathcal{L}(\boldsymbol{W}(k_\epsilon)) - \mathcal{L}_\star), \tag{38}$$

where $\mathcal{L}_\star \le \mathcal{L}(\boldsymbol{W}(k))$ for all $k \ge 0$.

Since $\|\boldsymbol{W}(k)\| \to \infty$, we get

$$\liminf_{k\to\infty} \left\langle \frac{\boldsymbol{W}(k)}{\|\boldsymbol{W}(k)\|_F}, \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle \ge 1 - \epsilon. \tag{39}$$

Given that ϵ is arbitrary, we can consider the limit as ϵ approaches zero. Thus, $\boldsymbol{W}(k)/\|\boldsymbol{W}(k)\|_F \to \boldsymbol{W}^{mm}/\|\boldsymbol{W}^{mm}\|_F$. ∎

E.3 Local Convergence of Gradient Descent

To provide a basis for discussing local convergence of GD, we establish a cone centered around $\boldsymbol{W}_{\boldsymbol{\alpha}}^{mm}$ using the following construction. For parameters $\mu \in (0,1)$ and $R > 0$, we define $C_{\mu,R}(\boldsymbol{W}_{\boldsymbol{\alpha}}^{mm})$ as the set of matrices $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ such that $\|\boldsymbol{W}\|_F \ge R$ and the correlation coefficient between $\boldsymbol{W}$ and $\boldsymbol{W}_{\boldsymbol{\alpha}}^{mm}$ is at least $1-\mu$:

$$\mathcal{S}_\mu(\boldsymbol{W}_{\boldsymbol{\alpha}}^{mm}) := \left\{ \boldsymbol{W} \in \mathbb{R}^{d\times d} \ : \ \left\langle \frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_F}, \frac{\boldsymbol{W}_{\boldsymbol{\alpha}}^{mm}}{\|\boldsymbol{W}_{\boldsymbol{\alpha}}^{mm}\|_F} \right\rangle \ge 1-\mu \right\}, \tag{40a}$$

$$C_{\mu,R}(\boldsymbol{W}_{\boldsymbol{\alpha}}^{mm}) := \mathcal{S}_\mu(\boldsymbol{W}_{\boldsymbol{\alpha}}^{mm}) \cap \left\{ \boldsymbol{W} \in \mathbb{R}^{d\times d} \ : \ \|\boldsymbol{W}\|_F \ge R \right\}. \tag{40b}$$

Lemma 10 *Suppose Assumption A on the loss function ℓ holds, and let $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ be locally optimal tokens according to Definition 2. Let $\boldsymbol{W}^{mm} = \boldsymbol{W}_{\boldsymbol{\alpha}}^{mm}$ denote the SVM solution obtained via (Att-SVM) by applying the Frobenius norm and replacing $(\texttt{opt}_i)_{i=1}^n$ with $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$. To provide a basis for discussing the local convergence of gradient descent, we establish a cone centered around $\boldsymbol{W}^{mm}$ using the following construction. There exists a scalar $\mu = \mu(\boldsymbol{\alpha}) > 0$ such that for sufficiently large $\bar{R}_\mu$:*

L1. *There is no stationary point within $C_{\mu,\bar{R}_\mu}(\boldsymbol{W}^{mm})$.*

L2. *For all $\boldsymbol{V} \in \mathcal{S}_\mu(\boldsymbol{W}^{mm})$ with $\|\boldsymbol{V}\|_F = \|\boldsymbol{W}^{mm}\|_F$ and $\boldsymbol{W} \in C_{\mu,\bar{R}_\mu}(\boldsymbol{W}^{mm})$, there exist dataset dependent constants $C, c > 0$ such that*

$$C \cdot \frac{1}{n}\sum_{i=1}^n (1 - \boldsymbol{s}_{i\alpha_i}) \ge -\left\langle \nabla\mathcal{L}(\boldsymbol{W}), \boldsymbol{V} \right\rangle \ge c \cdot \frac{1}{n}\sum_{i=1}^n (1 - \boldsymbol{s}_{i\alpha_i}) > 0, \tag{41a}$$

$$\|\nabla\mathcal{L}(\boldsymbol{W})\|_F \le \bar{A}C \cdot \frac{1}{n}\sum_{i=1}^n (1 - \boldsymbol{s}_{i\alpha_i}), \tag{41b}$$

$$-\left\langle \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|_F}, \frac{\nabla\mathcal{L}(\boldsymbol{W})}{\|\nabla\mathcal{L}(\boldsymbol{W})\|_F} \right\rangle \ge \frac{c}{C} \cdot \frac{\Theta}{\bar{A}} > 0. \tag{41c}$$

Here, $\boldsymbol{s}_{i\alpha_i} = (\mathbb{S}(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{z}_i))_{\alpha_i}$, $\bar{A} = \max_{i\in[n],t,\tau\in[T]} \|(\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})\| \, \|\boldsymbol{z}_i\|$, and $\Theta = 1/\|\boldsymbol{W}^{mm}\|_F$.

Proof. Let $R = \bar{R}_\mu$, $(\mathcal{T}_i)_{i=1}^n$ be the set of all support indices per Definition 2. Let $\bar{\mathcal{T}}_i = [T] - \mathcal{T}_i - \{\alpha_i\}$ be the non-support indices. Let

$$\begin{aligned} &\Theta = 1/\|\boldsymbol{W}^{mm}\|_F, \\ &\delta = \frac{1}{2} \min_{i\in[n]} \min_{t\in\mathcal{T}_i, \tau\in\bar{\mathcal{T}}_i} (\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W}^{mm} \boldsymbol{z}_i, \\ &A = \max_{i\in[n], t\in[T]} \frac{\|\boldsymbol{x}_{it}\boldsymbol{z}_i^\top\|_F}{\Theta}, \\ &\mu \le \mu(\delta) = \frac{1}{8}\left(\frac{\min(0.5,\delta)}{A}\right)^2. \end{aligned} \tag{42}$$

Since $\boldsymbol{W}^{mm}$ is the max-margin model ensuring $(\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W}^{mm} \boldsymbol{z}_i \ge 1$, the following inequalities hold for all $\boldsymbol{W} \in \mathcal{S}_\mu(\boldsymbol{W}^{mm})$, $\|\boldsymbol{W}\|_F = \|\boldsymbol{W}^{mm}\|_F$ and all $i \in [n], t \in \mathcal{T}_i, \tau \in \bar{\mathcal{T}}_i$:

$$\begin{aligned} (\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W} \boldsymbol{z}_i &\ge \delta > 0, \\ (\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{i\tau})^\top \boldsymbol{W} \boldsymbol{z}_i &\ge 1 + \delta, \\ \frac{3}{2} \ge (\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W} \boldsymbol{z}_i &\ge \frac{1}{2}. \end{aligned} \tag{43}$$

Here, we used $\|W - W^{mm}\|_F^2/\|W^{mm}\|_F^2 \le 2\mu$ which implies $\|W - W^{mm}\|_F \le \sqrt{2\mu}/\Theta$.

To proceed, we write the gradient correlation following (11) and (29)

$$\langle \nabla \mathcal{L}(W), V \rangle = \frac{1}{n}\sum_{i=1}^{n} \ell_i' \cdot h_i^\top \mathbb{S}'(\tilde{h}_i)\gamma_i, \tag{44}$$

where we denoted $\ell_i' = \ell'(Y_i \cdot v^\top X_i^\top \mathbb{S}(\tilde{h}_i))$, $h_i = X_i V z_i$, $\tilde{h}_i = X_i W z_i$, $s_i = \mathbb{S}(\tilde{h}_i)$.

Using (43), for all $t \in \mathcal{T}_i, \tau \in \bar{\mathcal{T}}_i$, for all $W \in C_{\mu,R}(W^{mm})$, we have that

$$\begin{aligned}
\tilde{h}_{it} - \tilde{h}_{i\tau} &\ge R\Theta\delta, \\
\tilde{h}_{i\alpha_i} - \tilde{h}_{i\tau} &\ge R\Theta(1+\delta), \\
\tilde{h}_{i\alpha_i} - \tilde{h}_{it} &\ge R\Theta/2.
\end{aligned}$$

Consequently, we can bound the softmax probabilities $s_i = \mathbb{S}(\tilde{h}_i)$ over non-support indices as follows: For all $i \in [n]$ and any $t_i \in \mathcal{T}_i$

$$S_i := \sum_{\tau \in \mathcal{T}_i} s_{i\tau} \le T e^{-R\Theta/2} s_{i\alpha_i} \le T e^{-R\Theta/2}, \tag{45a}$$

$$Q_i := \sum_{\tau \in \bar{\mathcal{T}}_i} s_{i\tau} \le T e^{-R\Theta\delta} s_{it_i} \le T e^{-R\Theta\delta} S_i. \tag{45b}$$

Recall scores $\gamma_{it} = Y_i \cdot v^\top x_{it}$. Define the score gaps over support indices:

$$\gamma_i^{gap} = \gamma_{i\alpha_i} - \max_{t \in \mathcal{T}_i} \gamma_{it} \quad \text{and} \quad \bar{\gamma}_i^{gap} = \gamma_{i\alpha_i} - \min_{t \in \mathcal{T}_i} \gamma_{it}.$$

It follows from (42) that

$$A = \max_{i \in [n], t \in [T]} \frac{\|x_{it} z_i^\top\|_F}{\Theta} \ge \max_{i \in [n], t \in [T]} \|h_{it}\|.$$

Define the α-dependent global scalar $\Gamma = \sup_{i \in [n], t, \tau \in [T]} |\gamma_{it} - \gamma_{i\tau}|$.

Let us focus on a fixed datapoint $i \in [n]$, assume (without losing generality) $\alpha_i = 1$, and drop subscripts i. Directly applying Lemma 6, we obtain

$$\left| h^\top \mathrm{diag}(s)\gamma - h^\top s s^\top \gamma - \sum_{t \ge 2}^{T} (h_1 - h_t) s_t (\gamma_1 - \gamma_t) \right| \le 2\Gamma A (1 - s_1)^2.$$

To proceed, let us decouple the non-support indices within $\sum_{t \ge 2}^{T} (h_1 - h_t) s_t (\gamma_1 - \gamma_t)$ via

$$\left| \sum_{t \in \bar{\mathcal{T}}} (h_1 - h_t) s_t (\gamma_1 - \gamma_t) \right| \le 2Q\Gamma A.$$

Aggregating these, we found

$$\left| h^\top \mathrm{diag}(s)\gamma - h^\top s s^\top \gamma - \sum_{t \in \mathcal{T}} (h_1 - h_t) s_t (\gamma_1 - \gamma_t) \right| \le 2\Gamma A((1 - s_1)^2 + Q). \tag{46}$$

To proceed, let us upper/lower bound the gradient correlation. We use two bounds depending on $V \in \mathcal{S}_\mu(W^{mm})$ (**Case 1**) or general $V \in \mathbb{R}^{d \times d}$ (**Case 2**).

• **Case 1: $V \in \mathcal{S}_\mu(W^{mm})$.** Since $1.5 \ge h_1 - h_t \ge 0.5$ following (43), we find

$$1.5 \cdot S \cdot \bar{\gamma}^{gap} \ge \sum_{t \in \mathcal{T}} (h_1 - h_t) s_t (\gamma_1 - \gamma_t) \ge 0.5 \cdot S \cdot \gamma^{gap},$$

where recall the definition of S (having dropped subscripts) in (45a).

• **Case 2: $V \in \mathbb{R}^{d \times d}$ and $\|V\|_F = \|W^{mm}\|_F$.** Define $\bar{A} = \max_{i \in [n], t, \tau \in [T]} \|x_{it} - x_{i\tau}\| \, \|z_i\|$. For any $\|V\|_F = \|W^{mm}\|$, we use the fact that

$$\|h_1 - h_t\| \le \|(x_{it} - x_{i\tau}) z_i^\top\|_F \cdot \|V\|_F \le \frac{\bar{A}}{\Theta}.$$

Note that by definition $\frac{\bar{A}}{\Theta} \geq 1$. To proceed, we can upper bound

$$\frac{\bar{A}}{\Theta} \cdot S \cdot \bar{\gamma}^{gap} \geq \sum_{t \in \mathcal{T}} (\boldsymbol{h}_1 - \boldsymbol{h}_t) \boldsymbol{s}_t (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t). \tag{47}$$

Next we claim that for both cases, S dominates $((1 - \boldsymbol{s}_1)^2 + Q)$ for large R. Specifically, we wish for

$$\frac{S \cdot \gamma^{gap}}{4} \geq 4\Gamma A \max((1 - \boldsymbol{s}_1)^2, Q) \iff S \geq 16 \frac{\Gamma A}{\gamma^{gap}} \max((1 - \boldsymbol{s}_1)^2, Q). \tag{48}$$

Now choose $R \geq \delta^{-1} \log(T)/\Theta$ to ensure $Q \leq S$ since $Q \leq T e^{-R\Theta\delta} S$ from (45a). Consequently

$$(1 - \boldsymbol{s}_1)^2 = (Q + S)^2 \leq 4S^2 \leq 4STe^{-R\Theta/2}.$$

Combining these, what we wish is ensured by guaranteeing

$$S \geq 16 \frac{\Gamma A}{\gamma^{gap}} \max(4STe^{-R\Theta/2}, Te^{-R\Theta\delta} S). \tag{49}$$

This in turn is ensured for all inputs $i \in [n]$ by choosing

$$R \geq \frac{\max(2, \delta^{-1})}{\Theta} \log\left(\frac{64T\Gamma A}{\gamma^{gap}_{\min}}\right), \tag{50}$$

where $\gamma^{gap}_{\min} = \min_{i \in [n]} \gamma^{gap}_i$ is the global scalar which is the worst case score gap over all inputs.

• **Case 1:** $\boldsymbol{V} \in \mathcal{S}_\mu(\boldsymbol{W^{mm}})$. With the above choice of R, we guaranteed

$$2(1 - \boldsymbol{s}_1) \cdot \bar{\gamma}^{gap} \geq 2 \cdot S \cdot \bar{\gamma}^{gap} \geq \boldsymbol{h}^\top \mathrm{diag}(\boldsymbol{s}) \boldsymbol{\gamma} - \boldsymbol{h}^\top \boldsymbol{s}\boldsymbol{s}^\top \boldsymbol{\gamma} \geq \frac{S \cdot \gamma^{gap}}{4} \geq \frac{(1 - \boldsymbol{s}_1)\gamma^{gap}}{8}.$$

via (48) and (46).

Since this holds over all inputs, going back to the gradient correlation (44) and averaging above over all inputs $i \in [n]$ and plugging back the indices i, we obtain the advertised bound by setting $q_i = 1 - \boldsymbol{s}_{i\alpha_i}$ (where we set $\alpha_i = 1$ above without losing generality)

$$\frac{2}{n} \sum_{i \in [n]} -\ell'_i \cdot q_i \cdot \bar{\gamma}^{gap}_i \geq -\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{V} \rangle \geq \frac{1}{8n} \sum_{i \in [n]} -\ell'_i \cdot q_i \cdot \gamma^{gap}_i. \tag{51}$$

Let $-\ell'_{\min / \max}$ be the min/max values negative loss derivative admits over the ball $[-A, A]$ and note that $\max_{i \in [n]} \bar{\gamma}^{gap}_i > 0$ and $\min_{i \in [n]} \gamma^{gap}_i > 0$ are dataset dependent constants. Then, we declare the constants $C = -2\ell'_{\max} \cdot \max_{i \in [n]} \bar{\gamma}^{gap}_i > 0, c = -(1/8)\ell'_{\min} \cdot \min_{i \in [n]} \gamma^{gap}_i > 0$ to obtain the bound (41a).

• **Case 2:** $\boldsymbol{V} \in \mathbb{R}^{d \times d}$ **and** $\|\boldsymbol{V}\|_F = \|\boldsymbol{W^{mm}}\|_F$**.** Next, we show (41b) and (41c). For any $\boldsymbol{V} \in \mathbb{R}^{d \times d}$ satisfying $\|\boldsymbol{V}\|_F = \|\boldsymbol{W}^{mm}\|_F$, using (47) and the choice of R in (50) similarly guarantees

$$\frac{2\bar{A}}{\Theta} (1 - \boldsymbol{s}_1) \bar{\gamma}^{gap} \geq \boldsymbol{h}^\top \mathrm{diag}(\boldsymbol{s}) \boldsymbol{\gamma} - \boldsymbol{h}^\top \boldsymbol{s}\boldsymbol{s}^\top \boldsymbol{\gamma},$$

for fixed input. Going back to the gradient correlation (44) and averaging above over all inputs $i \in [n]$, with the same definition of $C > 0$, we obtain

$$\frac{\bar{A}C}{\Theta n} \sum_{i \in [n]} q_i \geq -\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{V} \rangle. \tag{52}$$

To proceed, since (52) holds for any $\boldsymbol{V} \in \mathbb{R}^{d \times d}$, we observe that when setting $\boldsymbol{V} = \frac{\|\boldsymbol{W}^{mm}\|_F}{\|\nabla \mathcal{L}(\boldsymbol{W})\|_F} \cdot \nabla \mathcal{L}(\boldsymbol{W})$, this implies that

$$\langle \nabla \mathcal{L}(\boldsymbol{W}), \boldsymbol{V} \rangle = \|\nabla \mathcal{L}(\boldsymbol{W})\|_F \cdot \|\boldsymbol{W}^{mm}\|_F \leq \frac{\bar{A}C}{\Theta n} \sum_{i \in [n]} q_i.$$

Simplifying $\Theta = 1/\|\boldsymbol{W}^{mm}\|_F$ on both sides gives (41b).
Combining the above inequality with (51), we obtain that for all $\boldsymbol{V}, \boldsymbol{W} \in \mathcal{S}_\mu(\boldsymbol{W}^{mm})$

$$-\left\langle \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|_F}, \frac{\nabla \mathcal{L}(\boldsymbol{W})}{\|\nabla \mathcal{L}(\boldsymbol{W})\|_F} \right\rangle \geq \frac{c\Theta}{C\bar{A}},$$

which gives (41c).

■

Lemma 11 *Suppose Assumption A on the loss function ℓ holds, and let $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ be locally optimal tokens according to Definition 2. Let $\boldsymbol{W}^{mm} = \boldsymbol{W}^{mm}_{\boldsymbol{\alpha}}$ denote the SVM solution obtained via (Att-SVM) by replacing $(\texttt{opt}_i)_{i=1}^n$ with $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$. Let $\mu = \mu(\boldsymbol{\alpha}) > 0$ and $\bar{R}_\mu$ be defined as in Lemma 10. For any choice of $\pi > 0$, there exists $R_\pi \geq \bar{R}_\mu$ such that, for any $\boldsymbol{W} \in C_{\mu,R_\pi}(\boldsymbol{W}^{mm})$, we have*

$$\left\langle \nabla\mathcal{L}(\boldsymbol{W}), \frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_F} \right\rangle \geq (1+\pi) \left\langle \nabla\mathcal{L}(\boldsymbol{W}), \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle.$$

Proof. Let $R = R_\pi$, $\bar{\boldsymbol{W}} = \|\boldsymbol{W}^{mm}\|_F \boldsymbol{W}/\|\boldsymbol{W}\|_F$, $\boldsymbol{h}_i = \boldsymbol{X}_i \bar{\boldsymbol{W}} \boldsymbol{z}_i$, and $\bar{\boldsymbol{h}}_i = \boldsymbol{X}_i \boldsymbol{W}^{mm} \boldsymbol{z}_i$. To establish the result, we will prove that, for sufficiently large R, for any $\boldsymbol{W} \in C_{\mu,R}(\boldsymbol{W}^{mm})$ and for any $i \in [n]$,

$$\langle \boldsymbol{h}_i, \mathbb{S}'(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{z}_i)\boldsymbol{\gamma}_i \rangle \leq (1+\pi) \left\langle \bar{\boldsymbol{h}}_i, \mathbb{S}'(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{z}_i)\boldsymbol{\gamma}_i \right\rangle. \tag{53}$$

Once (53) holds for all i, the same conclusion will hold for the gradient correlations via (44). Moving forward, we shall again focus on a single point $i \in [n]$ and drop all subscripts i. Also, assume $\alpha = \alpha_i = 1$ without losing generality (same as above).

Following (46), for all $\boldsymbol{W} \in \mathcal{S}_\mu(\boldsymbol{W}^{mm})$ with $\|\boldsymbol{W}\|_F = \|\boldsymbol{W}^{mm}\|_F$ and $\tilde{\boldsymbol{h}} = \boldsymbol{X}\boldsymbol{W}\boldsymbol{z}$, and $\boldsymbol{s} = \mathbb{S}(\tilde{\boldsymbol{h}})$, we have found

$$\left|\tilde{\boldsymbol{h}}^\top \operatorname{diag}(\boldsymbol{s})\boldsymbol{\gamma} - \tilde{\boldsymbol{h}}^\top \boldsymbol{s}\boldsymbol{s}^\top \boldsymbol{\gamma} - \sum_{t\in\mathcal{T}} (\tilde{\boldsymbol{h}}_1 - \tilde{\boldsymbol{h}}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)\right| \leq 2\Gamma A((1-\boldsymbol{s}_1)^2 + Q), \tag{54}$$

where $\mathcal{T}$ is the set of support indices. Plugging in $\boldsymbol{h}, \bar{\boldsymbol{h}}$ in the bound above and assuming $\pi \leq 1$ (w.l.o.g.), (53) is implied by the following stronger inequality

$$\begin{aligned} 6\Gamma A((1-\boldsymbol{s}_1)^2 + Q) + \sum_{t\in\mathcal{T}} (\boldsymbol{h}_1 - \boldsymbol{h}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) &\leq (1+\pi) \sum_{t\in\mathcal{T}} (\bar{\boldsymbol{h}}_1 - \bar{\boldsymbol{h}}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \\ &= (1+\pi) \sum_{t\in\mathcal{T}} \boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t). \end{aligned}$$

First, we claim that $0.5\pi \sum_{t\in\mathcal{T}} \boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \geq 6\Gamma A((1-\boldsymbol{s}_1)^2 + Q)$. The proof of this claim directly follows the earlier argument, namely, following (48), (50) and (49) which leads to the choice

$$R \geq \frac{\max(2, \delta^{-1})}{\Theta} \log\left(\frac{C_0 \cdot T\Gamma A}{\pi \gamma^{gap}_{\min}}\right), \tag{55}$$

for some constant $C_0 > 0$. Using (50), we choose $C_0 \geq 64\pi$ to guarantee $R = R_\pi \geq \bar{R}_\mu$.

Following this control over the perturbation term $6\Gamma A((1-\boldsymbol{s}_1)^2 + Q)$, to conclude with the result, what remains is proving the comparison

$$\sum_{t\in\mathcal{T}} (\boldsymbol{h}_1 - \boldsymbol{h}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \leq (1+0.5\pi) \sum_{t\in\mathcal{T}} \boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t). \tag{56}$$

To proceed, we split the problem into two scenarios.

Scenario 1: $\|\bar{\boldsymbol{W}} - \boldsymbol{W}^{mm}\|_F \leq \epsilon = \frac{\pi}{4A\Theta}$ for some $\epsilon > 0$. In this scenario, for any token, we find that

$$|\boldsymbol{h}_t - \bar{\boldsymbol{h}}_t| \leq A\Theta\epsilon = \pi/4.$$

Consequently, we obtain

$$\boldsymbol{h}_1 - \boldsymbol{h}_t \leq \bar{\boldsymbol{h}}_1 - \bar{\boldsymbol{h}}_t + 2A\Theta\epsilon = 1 + 0.5\pi.$$

Similarly, $\boldsymbol{h}_1 - \boldsymbol{h}_t \geq 1 - 0.5\pi \geq 0.5$. Since all terms $\boldsymbol{h}_1 - \boldsymbol{h}_t, \boldsymbol{s}_t, \boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t$ in (56) are nonnegative and $(\boldsymbol{h}_1 - \boldsymbol{h}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \leq (1+0.5\pi)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)$, the above implies the desired result (56).

Scenario 2: $\|\bar{\boldsymbol{W}} - \boldsymbol{W}^{mm}\|_F \geq \epsilon = \frac{\pi}{4A\Theta}$. Since $\bar{\boldsymbol{W}}$ is not (locally) max-margin, in this scenario, for some $\nu = \nu(\epsilon) > 0$ and $\tau \in \mathcal{T}$, we have that $\boldsymbol{h}_1 - \boldsymbol{h}_\tau \leq 1 - 2\nu$. Here $\tau = \arg\max_{\tau\in\mathcal{T}} \boldsymbol{x}_\tau \bar{\boldsymbol{W}} \boldsymbol{z}$ denotes the nearest point to $\boldsymbol{h}_1$ (along the $\bar{\boldsymbol{W}}$ direction). Note that a non-support index $\tau \in \bar{\mathcal{T}}$ cannot be closest because $\boldsymbol{W} \in C_\mu$ and (43) holds. Recall that $\boldsymbol{s} = \mathbb{S}(\bar{R}\boldsymbol{h})$ where $\bar{R} = \|\boldsymbol{W}\|_F \Theta \geq R\Theta$. To proceed, split the tokens into two groups: Let $\mathcal{N}$ be the group of tokens obeying $(\boldsymbol{x}_1 - \boldsymbol{x}_\tau)\boldsymbol{W}\boldsymbol{z} \leq 1 - \nu$ and $\mathcal{T} - \mathcal{N}$ be the rest of the support indices. Observe that

$$\frac{\sum_{t\in\mathcal{T}-\mathcal{N}} \boldsymbol{s}_t}{\sum_{t\in\mathcal{T}} \boldsymbol{s}_t} \leq \frac{\sum_{t\in\mathcal{T}-\mathcal{N}} \boldsymbol{s}_t}{\sum_{t=\tau} \boldsymbol{s}_t} \leq T \frac{e^{\nu\bar{R}}}{e^{2\nu\bar{R}}} = Te^{-\bar{R}\nu}.$$

Thus, using $|\boldsymbol{h}_1 - \boldsymbol{h}_t| \le 2A$ and recalling the definition of γ^{gap}, observe that

$$\sum_{t\in\mathcal{T}-\mathcal{N}} (\boldsymbol{h}_1 - \boldsymbol{h}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \le \frac{2\Gamma A T e^{-\bar{R}\nu}}{\gamma^{gap}} \sum_{t\in\mathcal{T}} \boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t).$$

Plugging this into (56), we obtain

$$\begin{aligned}
\sum_{t\in\mathcal{T}}(\boldsymbol{h}_1 - \boldsymbol{h}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) &= \sum_{t\in\mathcal{N}}(\boldsymbol{h}_1 - \boldsymbol{h}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) + \sum_{t\in\mathcal{T}-\mathcal{N}}(\boldsymbol{h}_1 - \boldsymbol{h}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \\
&\le \sum_{t\in\mathcal{N}}(1-\nu)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) + \sum_{t\in\mathcal{T}-\mathcal{N}} 2A\Gamma T e^{-\bar{R}\nu} \\
&\le \left(1 - \nu + \frac{2\Gamma A T e^{-\bar{R}\nu}}{\gamma^{gap}}\right)\sum_{t\in\mathcal{T}} \boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \\
&\le \left(1 + \frac{2\Gamma A T e^{-\bar{R}\nu}}{\gamma^{gap}}\right)\sum_{t\in\mathcal{T}} \boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t).
\end{aligned}$$

Consequently, the proof boils down to ensuring the perturbation term $\frac{2\Gamma A T e^{-\bar{R}\nu}}{\gamma^{gap}} \le 0.5\pi$. Recalling $\bar{R} \ge R\Theta$, this is guaranteed for all inputs $i \in [n]$ by recalling $\gamma^{gap}_{\min} = \min_{i\in[n]} \gamma^{gap}_i$ and choosing

$$R \ge \frac{1}{\nu\Theta}\log\left(\frac{4\Gamma A T}{\gamma^{gap}_{\min}\pi}\right),$$

where $\nu = \nu(\frac{\pi}{4A\Theta})$ depends only on π and global problem variables.

Combining this with the prior R lower bound of (55) (by taking maximum), we conclude with the statement. ■

E.3.1 Proof of Theorem 3

Theorem 5 (Theorem 3 restated) *Suppose Assumption A on the loss ℓ holds, and let $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ be locally optimal tokens according to Definition 2. Let $\boldsymbol{W}^{mm}_{\boldsymbol{\alpha}}$ denote the SVM solution obtained via (Att-SVM) by replacing $(\texttt{opt}_i)_{i=1}^n$ with $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$. Then,*

- *There exist parameters $\mu = \mu(\boldsymbol{\alpha}) \in (0,1)$ and $R > 0$ such that $C_{\mu,R}(\boldsymbol{W}^{mm}_{\boldsymbol{\alpha}})$ does not contain any stationary points.*
- *Algorithm W-GD with $\eta \le 1/L_{\boldsymbol{W}}$ and any $\boldsymbol{W}(0) \in C_{\mu,R}(\boldsymbol{W}^{mm}_{\boldsymbol{\alpha}})$ satisfies $\lim_{k\to\infty} \|\boldsymbol{W}(k)\|_F = \infty$ and $\lim_{k\to\infty} \frac{\boldsymbol{W}(k)}{\|\boldsymbol{W}(k)\|_F} = \frac{\boldsymbol{W}^{mm}_{\boldsymbol{\alpha}}}{\|\boldsymbol{W}^{mm}_{\boldsymbol{\alpha}}\|_F}$.*

The proof of this theorem follows the proof of [TLZO23, Theorem 3]. Let us denote the initialization lower bound as $R^0_\mu := R$, where R is given in the Theorem 3's statement. Consider an arbitrary value of $\epsilon \in (0, \mu/2)$ and let $1/(1+\pi) = 1-\epsilon$. We additionally denote $R_\epsilon \leftarrow R_\pi \vee 1/2$ where R_π was defined in Lemma 11. At initialization $\boldsymbol{W}(0)$, we set $\epsilon = \mu/2$ to obtain $R^0_\mu = R_{\mu/2}$, and provide the proof in four steps:

Step 1: There are no stationary points within $C_{\mu,R^0_\mu}(\boldsymbol{W}^{mm})$. We begin by proving that there are no stationary points within $C_{\mu,R^0_\mu}(\boldsymbol{W}^{mm})$. Let $(\mathcal{T}_i)_{i=1}^n$ denote the sets of support indices as defined in Definition 2. We define $\bar{\mathcal{T}}_i = [T] - \mathcal{T}_i - \{\alpha_i\}$ as the tokens that are non-support indices. Additionally, let μ be defined as in (42). Then, since $R^0_\mu \ge \bar{R}_\mu$ per Lemma 11, we can apply Lemma 10 to find that: For all $\boldsymbol{V}, \boldsymbol{W} \in \mathcal{S}_\mu(\boldsymbol{W}^{mm})$ with $\|\boldsymbol{W}\|_F \ne 0$ and $\|\boldsymbol{W}\|_F \ge R^0_\mu$, we have that $-\langle \boldsymbol{V}, \nabla\mathcal{L}(\boldsymbol{W})\rangle$ is strictly positive.

Step 2: It follows from Lemma 11 that, there exists $R_\epsilon \ge \bar{R}_\mu \vee 1/2$ such that all $\boldsymbol{W} \in C_{\mu,R_\epsilon}(\boldsymbol{W}^{mm})$ satisfy

$$\left\langle -\nabla\mathcal{L}(\boldsymbol{W}), \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F}\right\rangle \ge (1-\epsilon)\left\langle -\nabla\mathcal{L}(\boldsymbol{W}), \frac{\boldsymbol{W}}{\|\boldsymbol{W}\|_F}\right\rangle. \tag{57}$$

The argument below applies to a general $\epsilon \in (0, \mu/2)$. However, at initialization $\boldsymbol{W}(0)$, we set $\epsilon = \mu/2$ and, recalling above, initialization lower bound was defined as $R_\mu^0 := R_{\mu/2}$. To proceed, for any $\epsilon \in (0, \mu/2)$, we will show that after gradient descent enters the conic set $C_{\mu,R_\epsilon}(\boldsymbol{W}^{mm})$ for the first time, it will never leave the set. Let t_ϵ be the first time gradient descent enters $C_{\mu,R_\epsilon}(\boldsymbol{W}^{mm})$. In **Step 4**, we will prove that such t_ϵ is guaranteed to exist. Additionally, for $\epsilon \leftarrow \mu/2$, note that $t_\epsilon = 0$ i.e. the point of initialization.

Step 3: Updates remain inside the cone $C_{\mu,R_\epsilon}(\boldsymbol{W}^{mm})$. By leveraging the results from **Step 1** and **Step 2**, we demonstrate that the gradient iterates, with an appropriate constant step size, starting from $\boldsymbol{W}(k_\epsilon) \in C_{\mu,R_\epsilon}(\boldsymbol{W}^{mm})$, remain within this cone.

We proceed by induction. Suppose that the claim holds up to iteration $k \geq k_\epsilon$. This implies that $\boldsymbol{W}(k) \in C_{\mu,R_\epsilon}(\boldsymbol{W}^{mm})$. Hence, recalling cone definition, there exists scalar $\mu = \mu(\boldsymbol{\alpha}) \in (0,1)$ and R such that $\|\boldsymbol{W}(k)\|_F \geq R$, and

$$
\left\langle \frac{\boldsymbol{W}(k)}{\|\boldsymbol{W}(k)\|_F}, \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle \geq 1 - \mu.
$$

For all $k \geq 1$, let

$$
\rho(k) := -\frac{1}{1-\epsilon} \left\langle \nabla \mathcal{L}(\boldsymbol{W}(k)), \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle. \tag{58}
$$

Note that $\rho(k) > 0$ due to **Step 1**. This together with the gradient descent update rule gives

$$
\begin{aligned}
\left\langle \frac{\boldsymbol{W}(k+1)}{\|\boldsymbol{W}(k)\|_F}, \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle &= \left\langle \frac{\boldsymbol{W}(k)}{\|\boldsymbol{W}(k)\|_F} - \frac{\eta}{\|\boldsymbol{W}(k)\|_F} \nabla \mathcal{L}(\boldsymbol{W}(k)), \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle \\
&\geq 1 - \mu - \frac{\eta}{\|\boldsymbol{W}(k)\|_F} \left\langle \nabla \mathcal{L}(\boldsymbol{W}(k)), \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle \\
&\geq 1 - \mu + \frac{\eta \rho(k)(1-\epsilon)}{\|\boldsymbol{W}(k)\|_F}.
\end{aligned} \tag{59a}
$$

Note that from Lemma 10, we have $\langle \nabla \mathcal{L}(\boldsymbol{W}(k)), \boldsymbol{W}(k) \rangle < 0$ which implies that $\|\boldsymbol{W}(k+1)\|_F \geq \|\boldsymbol{W}(k)\|_F$. This together with R_ϵ definition and $\|\boldsymbol{W}(k)\|_F \geq 1/2$ implies that

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

which gives

$$
\begin{aligned}
\frac{\|\boldsymbol{W}(k+1)\|_F}{\|\boldsymbol{W}(k)\|_F} &\leq 1 - \frac{\eta}{\|\boldsymbol{W}(k)\|_F} \left\langle \nabla \mathcal{L}(\boldsymbol{W}(k)), \frac{\boldsymbol{W}(k)}{\|\boldsymbol{W}(k)\|_F} \right\rangle + \eta^2 \frac{\|\nabla \mathcal{L}(\boldsymbol{W}(k))\|^2}{\|\boldsymbol{W}(k)\|_F} \\
&\leq 1 - \frac{\eta}{(1-\epsilon)\|\boldsymbol{W}(k)\|_F} \left\langle \nabla \mathcal{L}(\boldsymbol{W}(k)), \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F} \right\rangle + \eta^2 \frac{\|\nabla \mathcal{L}(\boldsymbol{W}(k))\|^2}{\|\boldsymbol{W}(k)\|_F} \\
&\leq 1 + \frac{\eta \rho(k)}{\|\boldsymbol{W}(k)\|_F} + \frac{\eta^2 \|\nabla \mathcal{L}(\boldsymbol{W}(k))\|^2}{\|\boldsymbol{W}(k)\|_F} =: C_1(\rho(k), \eta).
\end{aligned} \tag{59b}
$$

Here, the second inequality follows from (57) and (58).

Now, it follows from (59a) and (59b) that

$$\begin{aligned}\left\langle \frac{\boldsymbol{W}(k+1)}{\|\boldsymbol{W}(k+1)\|}, \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|} \right\rangle &\geq \frac{1}{C_1(\rho(k),\eta)} \left(1-\mu+\frac{\eta\rho(k)(1-\epsilon)}{\|\boldsymbol{W}(k)\|_F}\right) \\ &= 1-\mu+\frac{1}{C_1(\rho(k),\eta)}\left((1-\mu)(1-C_1(\rho(k),\eta))+\frac{\eta\rho(k)(1-\epsilon)}{\|\boldsymbol{W}(k)\|_F}\right) \\ &= 1-\mu+\frac{\eta}{C_1(\rho(k),\eta)}\left((\mu-1)(\frac{\rho(k)}{\|\boldsymbol{W}(k)\|_F}+\frac{\eta\|\nabla\mathcal{L}(\boldsymbol{W}(k))\|^2}{\|\boldsymbol{W}(k)\|_F})+\frac{\rho(k)(1-\epsilon)}{\|\boldsymbol{W}(k)\|_F}\right) \\ &= 1-\mu+\frac{\eta}{C_1(\rho(k),\eta)}\left(\frac{\rho(k)(\mu-\epsilon)}{\|\boldsymbol{W}(k)\|_F}-\eta(1-\mu)\frac{\|\nabla\mathcal{L}(\boldsymbol{W}(k))\|^2}{\|\boldsymbol{W}(k)\|_F}\right) \\ &\geq 1-\mu, \end{aligned} \tag{60}$$

where the last inequality uses our choice of stepsize $\eta \leq 1/L_W$ in Theorem 3's statement. Specifically, we need η to be small to ensure the last inequality. We will guarantee this by choosing a proper R_ϵ in Lemma 11. Specifically, Lemma 11 leaves the choice of C_0 in R_ϵ lower bound of (55) open (it can always be chosen larger). Here, by choosing $C_0 \gtrsim 1/L_W$ will ensure $\eta \leq 1/L_W$ works well.

$$\begin{aligned}\eta &\leq \frac{\mu}{2(1-\mu)(1-\frac{\mu}{2})}\frac{c}{C}\frac{\Theta}{\bar{A}}\frac{1}{\bar{A}CT}e^{R_\mu^0\Theta/2} \\ &\leq \frac{\mu-\epsilon}{1-\mu}\cdot\frac{1}{1-\epsilon}\cdot\frac{c}{C}\cdot\frac{\Theta}{\bar{A}}\cdot\frac{1}{\bar{A}CT}e^{R_\mu^0\Theta/2} \leq \frac{(\mu-\epsilon)}{1-\mu}\frac{\rho(k)}{\|\nabla\mathcal{L}(\boldsymbol{W}(k))\|_F^2}.\end{aligned} \tag{61}$$

Here, the first inequality uses our choice of $\epsilon \in (0, \mu/2)$ (see **Step 2**), and the last inequality is obtained from Lemma 10 since

$$\begin{aligned}\frac{\rho(k)}{\|\nabla\mathcal{L}(\boldsymbol{W}(k))\|_F} &= -\frac{1}{1-\epsilon}\left\langle\frac{\nabla\mathcal{L}(\boldsymbol{W}(k))}{\|\nabla\mathcal{L}(\boldsymbol{W}(k))\|_F}, \frac{\boldsymbol{W}^{mm}}{\|\boldsymbol{W}^{mm}\|_F}\right\rangle \geq \frac{1}{1-\epsilon}\cdot\frac{c}{C}\cdot\frac{\Theta}{\bar{A}}, \\ \frac{1}{\|\nabla\mathcal{L}(\boldsymbol{W}(k))\|_F} &\geq \frac{1}{\bar{A}C\cdot\frac{1}{n}\sum_{i=1}^n(1-\boldsymbol{s}_{i\alpha_i})} \geq \frac{1}{\bar{A}CTe^{-R_\mu^0\Theta/2}}\end{aligned}$$

for some data dependent constrants c and C, $\bar{A} = \max_{i\in[n],t,\tau\in[T]}\|(\boldsymbol{x}_{it}-\boldsymbol{x}_{i\tau})\|\,\|z_i\|$, and $\Theta = 1/\|\boldsymbol{W}^{mm}\|_F$.

Next, we will demonstrate that the choice of η in (61) does indeed meet our step size condition as stated in the theorem, i.e., $\eta \leq 1/L_W$. Recall that $1/(1+\pi) = 1-\epsilon$, which implies that $\pi = \epsilon/(1-\epsilon)$. Combining this with (55), we obtain:

$$R_\pi \geq \frac{\max(2,\delta^{-1})}{\Theta}\log\left(\frac{C_0T\Gamma A}{\pi\gamma_{\min}^{gap}}\right), \quad \text{where} \quad C_0 \geq 64\pi. \tag{62}$$

$$\Rightarrow R_\epsilon \geq \frac{\max(2,\delta^{-1})}{\Theta}\log\left(\frac{(1-\epsilon)C_0T\Gamma A}{\epsilon\gamma_{\min}^{gap}}\right), \quad \text{where} \quad C_0 \geq 64\frac{\epsilon}{1-\epsilon}. \tag{63}$$

On the other hand, at the initialization, we have $\epsilon = \mu/2$ which implies that

$$R_\mu^0 \geq \frac{\max(2,\delta^{-1})}{\Theta}\log\left(\frac{(2-\mu)C_0T\Gamma A}{\mu\gamma_{\min}^{gap}}\right), \quad \text{where} \quad C_0 \geq 64\frac{\mu}{2(1-\frac{\mu}{2})}. \tag{64}$$

In the following, we will determine a lower bound on C_0 such that our step size condition in Theorem 3's statement, i.e., $\eta \leq 1/L_W$, is satisfied. Note that for the choice of η in (61) to meet the condition $\eta \leq 1/L_W$, the following condition must hold:

$$\frac{1}{L_W} \leq \frac{\mu}{(2-\mu)}\frac{1}{C_2T}e^{R_\mu^0\Theta/2} \Rightarrow R_\mu^0 \geq \frac{2}{\Theta}\log\left(\frac{1}{L_W}\frac{2-\mu}{\mu}C_2T\right). \tag{65}$$

where $C_2 = (1-\mu)\frac{\bar{A}^2C^2}{\Theta c}$.

This together with (64) implies that

$$\frac{C_0\Gamma A}{\gamma_{\min}^{gap}} \geq (1-\mu)\frac{C_2}{L_W} \Rightarrow C_0 \geq \max\left(\frac{(1-\mu)C_2}{L_W}\frac{\gamma_{\min}^{gap}}{\Gamma A}, \frac{64\mu}{2-\mu}\right). \tag{66}$$

Therefore, with this lower bound on C_0, the step size bound in (61) is sufficiently large to ensure that $\eta \leq 1/L_W$ guarantees (60).

Hence, it follows from (60) that $W(k+1) \in C_{\mu,R_\epsilon}(W^{mm})$.

Step 4: The correlation of $W(k)$ and W^{mm} increases over k. The remainder is similar to the proof of Theorem 2. From Step 3, we have that all iterates remain within the initial conic set i.e. $W(k) \in C_{\mu,R_\mu^0}(W^{mm})$ for all $k \geq 0$. Note that it follows from Lemma 10 that $\langle \nabla\mathcal{L}(W), W^{mm}/\|W^{mm}\|_F \rangle < 0$, for any finite $W \in C_{\mu,R_\mu^0}(W^{mm})$. Hence, there are no finite critical points $W \in C_{\mu,R_\mu^0}(W^{mm})$, for which $\nabla\mathcal{L}(W) = 0$. Now, based on Lemma 7, which guarantees that $\nabla\mathcal{L}(W(k)) \to 0$, this implies that $\|W(t)\|_F \to \infty$. Consequently, for any choice of $\epsilon \in (0, \mu/2)$ there is an iteration k_ϵ such that, for all $k \geq k_\epsilon$, $W(k) \in C_{\mu,R_\epsilon}(W^{mm})$. Once within $C_{\mu,R_\epsilon}(W^{mm})$, following similar steps in (37) and (38), for any $k \geq k_\epsilon$,

$$\left\langle \frac{W(k)}{\|W(k)\|_F}, \frac{W^{mm}}{\|W^{mm}\|_F} \right\rangle \geq 1 - \epsilon + \frac{C(\epsilon,\eta)}{\|W(k)\|_F}, \qquad W(k) \in C_{\mu,R_\epsilon}(W^{mm}),$$

for some finite constant $C(\epsilon,\eta)$ (that depends only on $\eta, \epsilon, \|W(k_\epsilon)\|_F$).

Consequently, as $k \to \infty$

$$\liminf_{k\to\infty} \left\langle \frac{W(k)}{\|W(k)\|_F}, \frac{W^{mm}}{\|W^{mm}\|_F} \right\rangle \geq 1 - \epsilon, \qquad W(k) \in C_{\mu,R_\epsilon}(W^{mm}).$$

Since $\epsilon \in (0, \mu/2)$ is arbitrary, we get $W(k)/\|W(k)\|_F \to W^{mm}/\|W^{mm}\|_F$. ■

F Supporting Experiments

In this section, we introduce implementation details and additional experiments. We create a 1-layer self-attention using `PyTorch`, training it with the SGD optimizer and a learning rate of $\eta = 0.1$. We apply normalized gradient descent to ensure divergence of attention weights. The attention weight W is then updated through

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

In the setting of (K, Q)-parameterization, we noted that with extended training iterations, the norm of the combined parameter $KQ^\top$ consistently rises, despite the gradient being treated as zero due to computational limitations. To tackle this issue, we introduce a minor regularization penalty to the loss function, ensuring that the norms of K and Q remain within reasonable bounds. This adjustment involves

$$\widetilde{\mathcal{L}}(K, Q) = \mathcal{L}(K, Q) + \lambda(\|K\|_F^2 + \|Q\|_F^2).$$

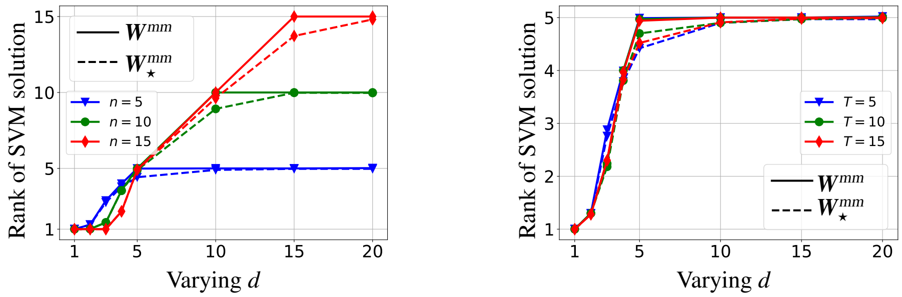
Here, we set λ to be the the smallest representable number, e.g. computed as $1 + \lambda = 1$ in `Python`, which is around 2.22×10^{-16}. Therefore, K, Q parameters are updated as follows.

$$K(k+1) = K(k) - \eta \frac{\nabla\widetilde{\mathcal{L}}_K(K(k), Q(k))}{\|\nabla\widetilde{\mathcal{L}}_K(K(k), Q(k))\|_F}, \qquad Q(k+1) = Q(k) - \eta \frac{\nabla\widetilde{\mathcal{L}}_Q(K(k), Q(k))}{\|\nabla\widetilde{\mathcal{L}}_Q(K(k), Q(k))\|_F}.$$

• As observed in previous work [TLZO23], and due to the exponential expression of softmax nonlinearity and computation limitation, `PyTorch` has no guarantee to select optimal tokens when the score gap is too small. Therefore in Figures 2, 9 and 10, we generate random tokens making sure that $\min_{i\in[n], t\neq \mathrm{opt}_i} \gamma_{i\mathrm{opt}_i} - \gamma_{it} \geq \underline{\gamma}$ and we choose $\underline{\gamma} = 0.1$ in our experiments.

Rank sensitivity of (K, Q)-parameterization (Figures 6&7). In Lemma 1, we have theoretically established that the rank of the SVM solution, denoted as W^{mm} in (Att-SVM) or $W_\star^{mm}$ in (Att-SVM$_\star$), is at most rank $\max(n, d)$. To further verify it, Figure 6 illustrates rank range of W^{mm} and $W_\star^{mm}$, solved using optimal tokens $(\mathrm{opt}_i)_{i=1}^n$ and setting $m = d$ (the rank constraint is eliminated). Each result is averaged over 100 trials, and for each trial, x_{it}, z_i, and linear head v are randomly sampled from the unit sphere. In Fig. 6(a), we fix $T = 5$ and vary n across $\{5, 10, 15\}$. Conversely, in Fig. 6(b), we keep $n = 5$ constant and alter T across $\{5, 10, 15\}$. Both figures confirm rank of W^{mm} and $W_\star^{mm}$ are bounded by $\max(n, d)$, validating Lemma 1.

Now, moving to Figure 7, we delve into GD performance across various dimensions of $K, Q \in \mathbb{R}^{d\times m}$ while keeping $d = 20$ fixed and varying m from 1 to 10. In the upper subfigure, we maintain a constant



(a) Rank of SVM solutions with fixed $T = 5$ (b) Rank of SVM solutions with fixed $n = 5$

Figure 6: Rank range of solutions for (Att-SVM) and (Att-SVM$_\star$), denoted as W^{mm} and $W^{mm}_\star$, solved using optimal tokens $(\texttt{opt}_i)_{i=1}^n$ and setting $m = d$ (the rank constraint is eliminated). Both figures confirm ranks of W^{mm} and $W^{mm}_\star$ are bounded by $\max(n, d)$, validating Lemma 1.

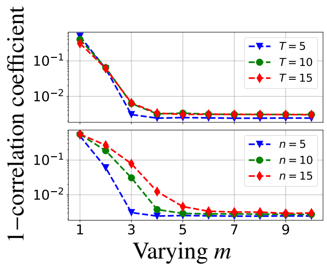
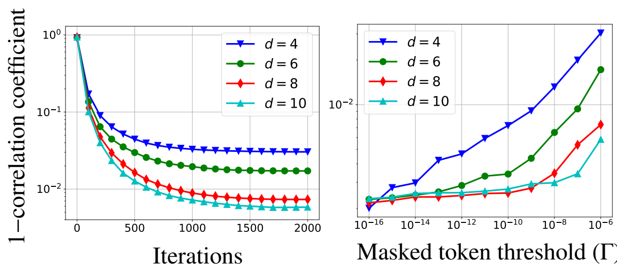


Figure 7: Convergence behavior of GD when training $(\boldsymbol{K}, \boldsymbol{Q}) \in \mathbb{R}^{d\times m}$ with varying m. The misalignment, $1-\texttt{corr\_coef}(W^{mm}_{\star,\alpha}, \boldsymbol{K}\boldsymbol{Q}^\top)$, is studied, where $W^{mm}_{\star,\alpha}$ is from (Att-SVM$_\star$) with $\texttt{opt}$ replaced by α and $m = d$. Subfigures with fixed $n = 5$ (upper) and $T = 5$ (lower) show that as m approaches or exceeds n, $\boldsymbol{K}\boldsymbol{Q}^\top$ aligns more with $W^{mm}_{\star,\alpha}$.

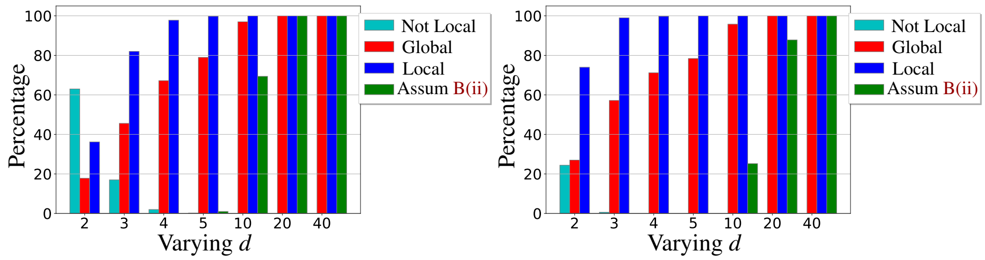


(a) Evolution of correlation (b) Γ vs correlation coefficient

Figure 8: Behavior of GD with nonlinear nonconvex prediction head and multi-token compositions. **(a)**: Blue, green, red and teal curves represent the evolution of $1 - \texttt{corr\_coef}(W, W^{\texttt{SVMeq}})$ for $d = 4, 6, 8$ and 10 respectively, which have been displayed in Figure 4(upper). **(b)**: Over the 500 random instances as discussed in Figure 4, we filter different instances by constructing masked set with tokens whose softmax output $< \Gamma$ and vary Γ from 10^{-16} to 10^{-6}. The corresponding results of $1 - \texttt{corr\_coef}(W, W^{\texttt{SVMeq}})$ are displayed in blue, green, red and teal curves.

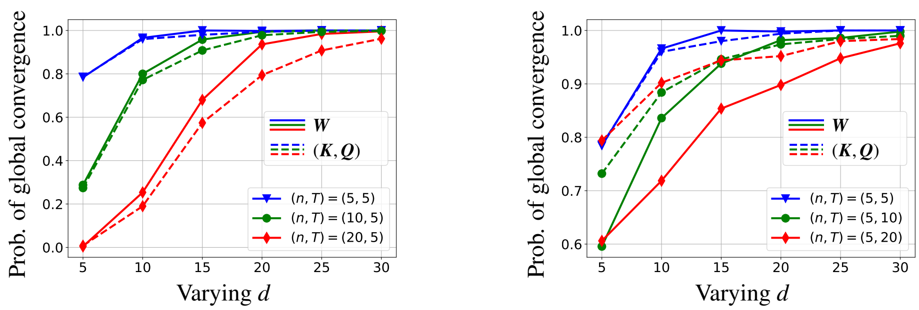
$n = 5$ and vary T within $\{5, 10, 15\}$, while in the lower subfigure, T is fixed at 5 and n changes within $\{5, 10, 15\}$. Results are depicted using blue, green, and red dashed curves, with both y-axes representing $1 - \texttt{corr\_coef}(W, W^{mm}_{\star,\alpha})$, where W represents the GD solution and $W^{mm}_{\star,\alpha}$ is obtained from (Att-SVM$_\star$) by employing token indices α selected via GD and setting the rank limit to $m = d$. Observing both subfigures, we note that a larger n necessitates a larger m for attention weights $\boldsymbol{K}\boldsymbol{Q}^\top$ to accurately converge to the SVM solution (Figure 7(lower)). Meanwhile, performances remain consistent across varying T values (Figure 7(upper)). This observation further validates Lemma 1. Furthermore, the results demonstrate that W converges directionally towards $W^{mm}_{\star,\alpha}$ as long as $m \gtrsim n$.

Global Convergence via overparameterization (Figures 9&10). The trend depicted in Figure 9, where the percentage of global convergence (red bars) approaches 100% and Assumption B(ii) holds with higher probability (green bars) as d grows, reinforces this insight. Specifically, Fig. 9(a) is same as Figure 2, and Fig. 9(b) displays the same evaluation over $(\boldsymbol{K}, \boldsymbol{Q})$-parameterization setting. In both experiments, and for each chosen d value, a total of 500 random instances are conducted under the conditions of $n = T = 5$. The outcomes are reported in terms of the percentages of Not Local, Local, and Global convergence, represented by the teal, blue, and red bars, respectively. We validate Assumption B(ii) as follows: Given a problem instance, we compute the average margin over all non-optimal tokens of all inputs and declare that problem satisfies Assumption B(ii), if the average



(a) $\boldsymbol{W}$-parameterization (b) $(\boldsymbol{K},\boldsymbol{Q})$-parameterization

Figure 9: Percentage of different convergence types of GD when training cross-attention weights **(a):** $\boldsymbol{W}$ or **(b):** $(\boldsymbol{K},\boldsymbol{Q})$ with varying d. In both figures, red, blue, and teal bars represent the percentages of Global, Local (including Global), and Not Local convergence, respectively. The green bar corresponds to Assumption B(ii) where all tokens act as support vectors. Larger overparameterization (d) relates to a higher percentage of globally-optimal SVM convergence.



(a) Global convergence for varying n, d (b) Global convergence for varying T, d

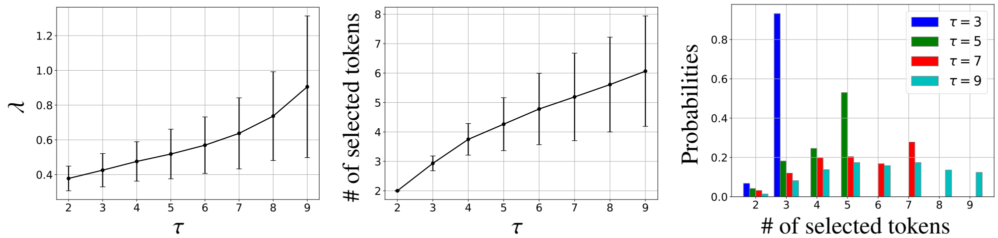
Figure 10: Global convergence behavior of GD when training cross-attention weights $\boldsymbol{W}$ (solid) or $(\boldsymbol{K},\boldsymbol{Q})$ (dashed) with random data. The blue, green, and red curves represent the probabilities of global convergence for **(a)**: fixing $T=5$ and varying $n \in \{5,10,20\}$ and **(b)**: fixing $n=5$ and varying $T \in \{5,10,20\}$. Results demonstrate that for both attention models, as d increases (due to over-parameterization), attention weights tend to select optimal tokens $(\texttt{opt}_i)_{i=1}^n$.

margin is below 1.1 (where 1 is the minimum). Here, recall that margin of a non-optimal token is defined as $(\boldsymbol{x}_{i\texttt{opt}_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W}^{mm} \boldsymbol{z}_i$ or $(\boldsymbol{x}_{i\texttt{opt}_i} - \boldsymbol{x}_{it})^\top \boldsymbol{W}^{mm}_\star \boldsymbol{z}_i$ for $t \neq \texttt{opt}_i$.

Furthermore, the observations in Figure 10 regarding the percentages of achieving global convergence reaching 100 with larger d reaffirm that overparameterization leads the attention weights to converge directionally towards the optimal max-margin direction outlined by (Att-SVM) and (Att-SVM$_\star$).

Behavior of GD with nonlinear nonconvex prediction head and multi-token compositions (Figure 8). To better investigate how correlation changes with data dimension d, we collect the solid curves in Figure 4(upper) and construct as Figure 8(a). Moreover, Figure 8(b) displays the average correlation of instances (refer to scatters in Figure 4 (lower)), considering masked tokens with softmax probability $< \Gamma$. Both findings highlight that higher d enhances alignment. For $d \geq 8$ or $\Gamma \leq 10^{-9}$, the GD solution $\boldsymbol{W}$ achieves a correlation of > 0.99 with the SVM-equivalence $\boldsymbol{W}^{\texttt{SVMeq}}$, defined in Section B.

Investigation of Lemma 3 over different τ selections (Figure 11). Consider the setting of Section B.1 and Lemma 3. Figure 5 explores the influence of λ on the count of tokens selected by GD-derived attention weights. As λ increases, the likelihood of selecting more tokens also increases. Shifting focus to Figure 11, we examine the effect of τ. For each outcome, we generate random λ values, retaining pairs $(\lambda, \boldsymbol{X})$ satisfying τ constraints, with averages derived from 100 successful trials. The results indicate a positive correlation among τ, λ, and the number of selected tokens.



(a) τ vs λ parameters (b) τ vs # of selected tokens (c) Distribution of # selected tokens

Figure 11: Behavior of GD when selecting multiple tokens.

Moreover, Figure 11(c) provides a precise distribution of selected token counts across various τ values (specifically $\tau \in \{3, 5, 7, 9\}$). The findings confirm that the number of selected tokens remains within the limit of τ, thus validating the assertion made in Lemma 3.

G Discussion, Future Directions, and Open Problems

Our optimization-theoretic characterization of the self-attention model provides a comprehensive understanding of its underlying principles. The developed framework, along with the research presented in [TLZO23], introduces new avenues for studying transformers and language models. The key findings include:

✓ The optimization geometry of self-attention exhibits a fascinating connection to hard-margin SVM problems. By leveraging linear constraints formed through outer products of token pairs, optimal input tokens can be effectively separated from non-optimal ones.

✓ When gradient descent is employed without early-stopping, implicit regularization and convergence of self-attention naturally occur. This convergence leads to the maximum margin solution when minimizing specific requirements using logistic loss, exp-loss, or other smooth decreasing loss functions. Moreover, this implicit bias is unaffected by the step size, as long as it is sufficiently small for convergence, and remains independent of the initialization process.

The fact that gradient descent leads to a maximum margin solution may not be surprising to those who are familiar with the relationship between regularization path and gradient descent in linear and nonlinear neural networks [SHN+18, GLSS18, NLG+19, JT21, MWG+20, JT20]. However, there is a lack of prior research or discussion regarding this connection to the attention mechanism. Moreover, there has been no rigorous analysis or investigation into the exactness and independence of this bias with respect to the initialization and step size. Thus, we believe our findings and insights deepen our understanding of transformers and language models, paving the way for further research in this domain. Below, we discuss some notable directions and highlight open problems that are not resolved by the existing theory.

- **Convergence Rates**: The current paper establishes asymptotic convergence of gradient descent; nonetheless, there is room for further exploration to characterize non-asymptotic convergence rates. Indeed, such an exploration can also provide valuable insights into the choice of learning rate, initialization, and the optimization method.
- **Gradient descent on $(\boldsymbol{K}, \boldsymbol{Q})$ parameterization:** We find it remarkable that regularization path analysis was able to predict the implicit bias of gradient descent. Complete analysis of gradient descent is inherently connected to the fundamental question of low-rank factorization [GWB+17, LMZ18]. We believe formalizing the implicit bias of gradient descent under margin constraints presents an exciting open research direction for further research.
- **Generalization analysis:** An important direction is the generalization guarantees for gradient-based algorithms. The established connection to hard-margin SVM can facilitate this because the SVM problem is amenable to statistical analysis. This would be akin to how kernel/NTK analysis for deep nets enabled a rich literature on generalization analysis for traditional deep learning.

- **Realistic architectures:** Naturally, we wish to explore whether max-margin equivalence can be extended to more realistic settings: Can the theory be expanded to handle multi-head attention, multi-layer architectures, and MLP nonlinearities? We believe the results in Section B take an important step towards this direction by including analytical formulae for the implicit bias of the attention layer under nonlinear prediction heads.
- **Jointly optimizing attention and prediction head:** It would be interesting to study the joint optimization dynamics of attention weights and prediction head $h(\cdot)$. This problem can be viewed as a novel low-rank factorization type problem where $h(\cdot)$ and $\boldsymbol{W}$ are factors of the optimization problem, only, here, $\boldsymbol{W}$ passes through the softmax nonlinearity. To this aim, [TLZO23] provides a preliminary geometric characterization of the implicit bias for a simpler attention model using regularization path analysis. Such findings can potentially be generalized to the analysis of gradient methods and full transformer block.