Margin Maximization in Attention Mechanism

Anonymous Author(s)
Affiliation
Address
email

Abstract

Attention mechanism is a central component of the transformer architecture which led to the phenomenal success of large language models. However, the theoretical principles underlying the attention mechanism are poorly understood, especially its nonconvex optimization dynamics. In this work, we explore the seminal softmax-attention model $f(\boldsymbol{X}) = \boldsymbol{v}^\top \boldsymbol{X}^\top \texttt{softmax}(\boldsymbol{X}\boldsymbol{W}^\top \boldsymbol{p})$, where, $\boldsymbol{X}$ is the tokenized input, $\boldsymbol{v}$ is the value weights, $\boldsymbol{W}$ is the key-query weights, and $\boldsymbol{p}$ is a tunable token/prompt. We prove that running gradient descent on $\boldsymbol{p}$, or equivalently $\boldsymbol{W}$, converges to a max-margin solution that separates locally-optimal tokens from non-optimal ones. We also develop regularization path analysis that generalizes these findings to nonlinear classifier head – rather than linear $\boldsymbol{v}$. When optimizing $\boldsymbol{v}$ and $\boldsymbol{p}$ simultaneously with logistic loss, we identify conditions under which the regularization paths converge to their respective max-margin solutions where $\boldsymbol{v}$ separates the input features based on their labels. Finally, we verify our results through numerical insights.

1 Introduction

Since its introduction in the seminal work [1], attention mechanism has played an influential role in the advances in natural language processing, and more recently, large language models [2, 3, 4, 5]. Attention is initially introduced for encoder-decoder RNN architectures in order to allow the decoder to use the most relevant parts of the input sequence, rather than relying on a fixed-length hidden state. Attention mechanism has taken the center stage in the transformers [6] where the self-attention layer – which calculates softmax similarities between input tokens – forms the backbone of the architecture. Since their inception, transformers have revolutionized natural language processing (from BERT to ChatGPT [7, 8]) and they have also become the architecture of choice for foundation models [9] to address diverse challenges in generative modeling [3, 10], computer vision [11, 12], and reinforcement learning [13, 14, 15].

The prominence of the attention mechanism motivate a fundamental theoretical understanding of its role in optimization and learning. While it is well-known that attention enables the model to focus on the relevant parts of the input sequence, the precise mechanism by which this is achieved is far from clear. To this end, we ask

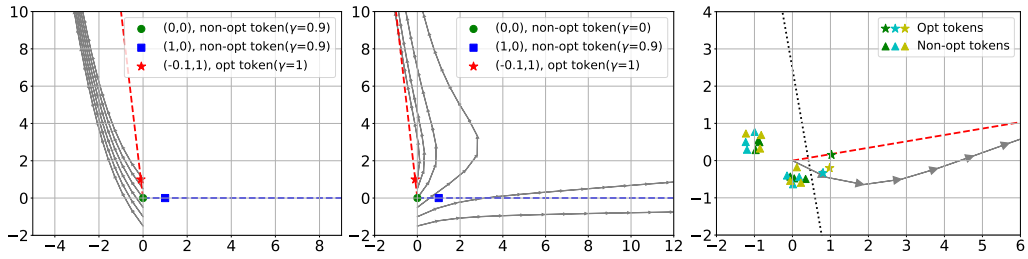
Q: What are the optimization dynamics and inductive biases of the attention mechanism?

We study this question using the fundamental attention model $f(\boldsymbol{X}) = \boldsymbol{v}^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}^\top \boldsymbol{p})$. Here, $\boldsymbol{X}$ is the sequence of input tokens, $\boldsymbol{v}$ is the classifier head, $\boldsymbol{W}$ is the trainable key-query weights, and $\mathbb{S}$ denotes the softmax nonlinearity. For transformers, $\boldsymbol{p}$ corresponds to the `[CLS]` token or tunable prompt [16] whereas for RNN architectures [1], $\boldsymbol{p}$ corresponds to the hidden state.

Given training data $(Y_i, \boldsymbol{X}_i)_{i=1}^n$ with labels $Y_i \in \{-1, 1\}$ and inputs $\boldsymbol{X}_i \in \mathbb{R}^{T\times d}$, we consider the empirical risk minimization with a decreasing loss function $\ell(\cdot) : \mathbb{R} \to \mathbb{R}$,

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{p}, \boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i \cdot f(\boldsymbol{X}_i)).$$

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.



(a) Global convergence (b) Global and local optimal directions (c) Multiple inputs

Figure 1: The convergence behavior of the gradient descent on the attention weights $\boldsymbol{p}$ using the logistic loss in (ERM). Here, (- - -) and (- - -) denote the global and local max-margin solutions. γ denotes the *score* of a token per Definition 1. Discussion is provided under Theorem 1.

We operate under the assumption that the most relevant tokens within each input are separable from the rest through softmax nonlinearity. Our main contributions are as follows:

• **Optimize $\boldsymbol{p}$ or $\boldsymbol{W}$ for fixed $\boldsymbol{v}$ (Sec. 2):** We first prove that gradient iterations of $\boldsymbol{p}$ and $\boldsymbol{W}$ admit a one-to-one mapping, thus we focus on optimizing $\boldsymbol{p}$ without losing generality. We prove that gradient iterates of $\boldsymbol{p}$ converges to a max-margin solution (namely (ATT-SVM)) that separates locally-optimal tokens from non-optimal ones. The notion of *relevant tokens* is clearly quantified in terms of scores $\boldsymbol{\gamma}_t = Y \cdot \boldsymbol{v}^\top \boldsymbol{x}_t$ where $\boldsymbol{x}_t$ is the t'th token of input $\boldsymbol{X}$. The *locally-optimal* tokens are those with higher scores than their nearest neighbors determined by the SVM solution. These are illustrated in Figure 1.

• **Optimize $(\boldsymbol{v}, \boldsymbol{p})$ jointly (Sec. 3):** We study the joint problem under logistic loss function. We use regularization path analysis where (ERM) is solved under ridge constraints and we study the solution trajectory as the constraints are relaxed. Since the problem is linear in $\boldsymbol{v}$, if the attention features $\boldsymbol{x}_i^{att} = \mathbb{S}(\boldsymbol{X}_i \boldsymbol{W}^\top \boldsymbol{p})$ are separable based on their labels Y_i, $\boldsymbol{v}$ would implement a max-margin classifier. Building on this, we prove that $\boldsymbol{p}$ and $\boldsymbol{v}$ converges to their respective max-margin solutions under certain margin conditions. Relaxing these conditions, we obtain a more general solution where margin constraints on $\boldsymbol{p}$ are relaxed on the inputs whose attention features are not support vectors of $\boldsymbol{v}$. Figure 2 illustrates these outcomes.

In Sec. 4, we extend the ideas in Sec. 2 to the more general model $f(\boldsymbol{X}) = \psi(\boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}^\top \boldsymbol{p}))$ with nonlinear head ψ. Overall, our results clearly formalize the role of the attention mechanism as a token-selection/context-discovery mechanism and lay the groundwork for future research by connecting it to the implicit bias literature and max-margin SVM formulation.

Next section introduces preliminaries, Section 5 discusses related literature, and Section 6 provides a discussion of limitations and future work.

1.1 Preliminaries

Notations. For any integer $N \geq 1$, let $[N] = \{1, \ldots, N\}$. We use lower-case and upper-case bold letters (e.g. $\boldsymbol{a}$ and $\boldsymbol{A}$) to represent vectors and matrices, respectively. The entries of $\boldsymbol{a}$ are denoted as $\boldsymbol{a}_i$. We use $\sigma_{\max}(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A})$ to denote the maximum and minimum singular values of $\boldsymbol{A}$, respectively. We denote the minimum of two numbers a, b as $a \wedge b$, and the maximum $a \vee b$. We use the standard big-Oh notation $O(\cdot)$ to hide universal constants.

Optimization. Given a function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ and an ℓ_2-norm bound R, the regularized solution is defined as

$$\bar{\boldsymbol{p}}(R) := \arg\min_{\|\boldsymbol{p}\| \leq R} \mathcal{L}(\boldsymbol{p}). \tag{1}$$

Note that $\bar{\boldsymbol{p}}(R)$ is not unique in general. For gradient descent, we assume the objective $\mathcal{L}(\boldsymbol{p})$ is smooth and describe the gradient descent process as

$$\boldsymbol{p}(t+1) = \boldsymbol{p}(t) - \eta(t) \nabla \mathcal{L}(\boldsymbol{p}(t)), \tag{2}$$

where $\eta(t)$ is the stepsize at time t and $\nabla \mathcal{L}(\boldsymbol{p}(t))$ is the gradient of $\mathcal{L}$ at $\boldsymbol{p}(t)$.

Attention in Transformers. We now describe how our model relates to the attention mechanism in transformers. Our exposition follows the recent work [17] which focuses on the theoretical properties of prompt-tuning.

• **Self-attention** is the core building block of transformers [6]. Given an input consisting of T tokens $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]^\top \in \mathbb{R}^{T\times d}$, self-attention with key-query matrix $\boldsymbol{W} \in \mathbb{R}^{d\times d}$, and value matrix $\boldsymbol{V} \in \mathbb{R}^{d\times v}$, the self-attention model is defined as follows:

$$f_{sa}(\boldsymbol{X}) = \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{V}. \tag{3}$$

Here, $\mathbb{S}(\cdot)$ is the softmax nonlinearity that applies row-wise on the similarity matrix $\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top$.

• **Tunable tokens: `[CLS]` and prompt-tuning.** In practice, we append additional tokens to the raw input features $\boldsymbol{X}$: For instance, a `[CLS]` token is used for classification purposes [7] and prompt vectors can be appended for adapting a pretrained model to new tasks [16, 18]. Let $\boldsymbol{p} \in \mathbb{R}^d$ be the tunable token (`[CLS]` or prompt vector) and concatenate it to $\boldsymbol{X}$ to obtain $\boldsymbol{X}_{\boldsymbol{p}} := [\boldsymbol{p}\ \boldsymbol{X}^\top]^\top \in \mathbb{R}^{(T+1)\times d}$. Consider the cross-attention features obtained from $\boldsymbol{X}_{\boldsymbol{p}}$ and $\boldsymbol{X}$ given by

$$\begin{bmatrix} f_{cls}^\top(\boldsymbol{X}) \\ f_{sa}(\boldsymbol{X}) \end{bmatrix} = \mathbb{S}(\boldsymbol{X}_{\boldsymbol{p}}\boldsymbol{W}\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{V} = \begin{bmatrix} \mathbb{S}(\boldsymbol{p}^\top\boldsymbol{W}\boldsymbol{X}^\top) \\ \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top) \end{bmatrix}\boldsymbol{X}\boldsymbol{V}.$$

The beauty of cross-attention is that it isolates the contribution of $\boldsymbol{p}$ under the upper term $f_{cls}(\boldsymbol{X}) = \boldsymbol{V}^\top\boldsymbol{X}^\top\mathbb{S}(\boldsymbol{X}\boldsymbol{W}^\top\boldsymbol{p}) \in \mathbb{R}^v$. In this work, we use the value weights for classification, thus we set $v = 1$, and denote $\boldsymbol{v} = \boldsymbol{V} \in \mathbb{R}^d$. This brings us to our attention model of interest:

$$f(\boldsymbol{X}, \Theta) = \boldsymbol{v}^\top\boldsymbol{X}^\top\mathbb{S}(\boldsymbol{K}\boldsymbol{p}) \quad \text{where} \quad \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^\top. \tag{4}$$

Above $\Theta = (\boldsymbol{v}, \boldsymbol{W}, \boldsymbol{p})$ are the tunable model parameters and $\boldsymbol{K}$ is the key embeddings. Note that $\boldsymbol{W}$ and $\boldsymbol{p}$ are playing the same role within softmax, thus, it is intuitive that they exhibit similar optimization dynamics. Confirming this, the next lemma shows that gradient iterations of $\boldsymbol{p}$ (after setting $\boldsymbol{W} \leftarrow$ Identity) and $\boldsymbol{W}$ admit a one-to-one mapping.

Lemma 1 *Fix $\boldsymbol{a} \in \mathbb{R}^d$. Let $\psi : \mathbb{R}^d \to \mathbb{R}$ and $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be differentiable functions. On the same training data, define $\mathcal{L}(\boldsymbol{p}) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \psi(\boldsymbol{X}_i^\top\mathbb{S}(\boldsymbol{X}_i\boldsymbol{p})))$ and $\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \psi(\boldsymbol{X}_i^\top\mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}^\top\boldsymbol{a})))$. Starting from $\boldsymbol{p}(0)$ and $\boldsymbol{W}(0) = \boldsymbol{a}\boldsymbol{p}(0)^\top/\|\boldsymbol{a}\|^2$ consider the gradient descent iterations with stepsize η:*

$$\boldsymbol{p}(t+1) = \boldsymbol{p}(t) - \eta\nabla\mathcal{L}(\boldsymbol{p}_t),$$
$$\boldsymbol{W}(t+1) = \boldsymbol{W}(t) - \eta\|\boldsymbol{a}\|^{-2}\nabla\mathcal{L}(\boldsymbol{W}(t)).$$

We have that $\boldsymbol{W}_t = \boldsymbol{a}\boldsymbol{p}_t^\top/\|\boldsymbol{a}\|^2$ for all $t \geq 0$.

Thanks to this lemma, $\boldsymbol{W}$'s optimization dynamics is directly characterized by $\boldsymbol{p}$'s dynamics, since we can always reconstruct $\boldsymbol{W}$ from $\boldsymbol{p}$ using the relationship between their gradient iterations. Hence, in what follows, we fix $\boldsymbol{W}$, and focus on optimizing $\boldsymbol{p}$ in Sec 2 and joint optimization of $(\boldsymbol{v}, \boldsymbol{p})$ in Sec 3.

Problem definition: Throughout, $(Y_i, \boldsymbol{X}_i)_{i=1}^n$ denotes our training dataset where $Y_i \in \{-1, 1\}$ and $\boldsymbol{X}_i \in \mathbb{R}^{T\times d}$. We denote the key embeddings of $\boldsymbol{X}_i$ via $\boldsymbol{K}_i = \boldsymbol{X}_i\boldsymbol{W}^\top$ and explore the training risk

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

Importantly, our results apply to general tuples $(Y_i, \boldsymbol{X}_i, \boldsymbol{K}_i)$ and do not assume that $(\boldsymbol{X}_i, \boldsymbol{K}_i)$ are tied via $\boldsymbol{W}$. Finally, the t^{th} tokens of $\boldsymbol{X}_i, \boldsymbol{K}_i$ are denoted by $\boldsymbol{x}_{it}, \boldsymbol{k}_{it} \in \mathbb{R}^d$ respectively for $t \in [T]$.

2 Global and Local Margin Maximization with Attention

In this section, we establish the main results of this paper (Theorems 1 and 3) by characterizing the implicit bias of gradient descent on learning $\boldsymbol{p} \in \mathbb{R}^d$ for a fixed $\boldsymbol{v} \in \mathbb{R}^d$ choice. A notable feature of our results is that they apply to general decreasing loss functions. The underlying reason is that margin maximization arises from the exponentially-tailed nature of the softmax within attention rather than ℓ. Throughout, we make the following assumption on the loss function:

Assumption A (Well-Behaved Losses) *Over any bounded set in $\mathbb{R}$, $\ell : \mathbb{R} \to \mathbb{R}$ obeys*

A1. *ℓ is strictly decreasing and bounded from below.*

A2. *ℓ' is M_0-Lipschitz continuous and $|\ell'(u)| \leq M_1$.*

Assumption A includes many common loss functions, including the logistic loss $\ell(u) = \log(1 + e^{-u})$, exponential loss $\ell(u) = e^{-u}$, and correlation loss $\ell(u) = -u$. Assumption A implies that $\mathcal{L}(\boldsymbol{p})$ is a $O\left((M_0 + M_1)\bar{\sigma}_{\max}^4\right)$-smooth function (see supplementary), where $\bar{\sigma}_{\max} := 1/n \sum_{i=1}^n \sigma_{\max}(\boldsymbol{X}_i)$. The central feature of this assumption is that *we do not require convexity for the loss function*.

We now introduce a convex hard-margin SVM problem that separates one token of the input sequence from the rest, jointly solved over all inputs. We will show that this problem captures the optimization properties of softmax-attention. Fix indices $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ and consider

$$\boldsymbol{p}^{mm}(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{p}} \|\boldsymbol{p}\| \quad \text{subject to} \quad \min_{t \neq \alpha_i} \boldsymbol{p}^\top (\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it}) \geq 1 \quad \text{for all} \quad 1 \leq i \leq n. \qquad \text{(ATT-SVM)}$$

We are ready to introduce our main results by characterizing global and local convergence of the attention weights $\boldsymbol{p}$ in the direction of (ATT-SVM) solutions.

2.1 Global Convergence of the Attention Weights p

We first identify the conditions that guarantees the global convergence of gradient descent for $\boldsymbol{p}$. The intuition is that, in order for attention to exhibit implicit bias, the softmax nonlinearity should be forced to select the *optimal token within each input sequence*. Fortunately, the optimal tokens that achieve the smallest training objective under decreasing loss function $\ell(\cdot)$ have a clear definition.

Definition 1 (Token Score and Optimality) *The score of 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 tokens for input $\boldsymbol{X}_i$ are those tokens with highest scores given by*

$$\texttt{opt}_i \in \arg\max_{t \in [T]} \boldsymbol{\gamma}_{it}.$$

We denote the solution of (ATT-SVM) with optimal indices $(\texttt{opt}_i)_{i=1}^n$ by $\boldsymbol{p}^{mm\star}$. Note that multiple tokens within an $\boldsymbol{X}_i$ might attain same score, thus $\texttt{opt}_i$ and $\boldsymbol{p}^{mm\star}$ may not be unique.

To proceed with our global convergence analysis, we need to make the assumption that all non-optimal tokens have equal scores. In other words, if a potential solution includes tokens that do not appear in the final optimal solution, all of these tokens are assumed to have the same score value.

Assumption B *For all $i \in [n]$ and $t, \tau \neq \texttt{opt}_i$, the scores per Def. 1 obey $\boldsymbol{\gamma}_{it} = \boldsymbol{\gamma}_{i\tau} < \boldsymbol{\gamma}_{i\texttt{opt}_i}$.*

Theorem 1 (Global Convergence of Gradient Descent) *Suppose Assumption A on the loss function ℓ and Assumption B on the tokens' score hold. Then the gradient descent iterates $\boldsymbol{p}(t+1) = \boldsymbol{p}(t) - \eta \nabla \mathcal{L}_{\boldsymbol{p}}(\boldsymbol{p}(t))$ on (ERM), with the step size $\eta \leq O\left(\bar{\sigma}_{\max}^{-4}/(M_0 + M_1)\right)$ and any starting point $p(0)$ satisfies $\lim_{t\to\infty} \boldsymbol{p}(t)/\|\boldsymbol{p}(t)\| = \boldsymbol{p}^{mm\star}/\|\boldsymbol{p}^{mm\star}\|$.*

Theorem 1 shows that gradient descent dynamics of the normalized predictor $\boldsymbol{p}(t)/\|\boldsymbol{p}(t)\|$ converges towards $\boldsymbol{p}^{mm\star}/\|\boldsymbol{p}^{mm\star}\|$, effectively separating globally optimal tokens from non-optimal ones. To illustrate this theorem, we have conducted synthetic experiments. Let us first explain the setup used in Figure 1 and 2(a). We set $d = 3$ with each token having three entries $\boldsymbol{x} = [x_1, x_2, x_3]$. We reserve the first two coordinates as key embeddings $\boldsymbol{k} = [x_1, x_2, 0]$ by setting $\boldsymbol{W} = \text{diag}([1, 1, 0])$. This is what we display in our figures as token positions. Finally, in order to assign scores to the tokens we use the last coordinate by setting $\boldsymbol{v} = [0, 0, 1]$. This way score becomes $Y \cdot \boldsymbol{v}^\top \boldsymbol{x} = Y \cdot x_3$, allowing us to assign any score (regardless of key embedding).

In Figure 1(a), the gray paths represent gradient descent trajectories initiated from different points, while the points $(0, 0)$ and $(1, 0)$ correspond to non-optimal tokens, and $(-0.1, 1)$ represents the optimal token. Notably, gradient descent iterates with various starting points converge towards the direction of the max-margin solution $\boldsymbol{p}^{mm\star}$ (depicted by - - -). Moreover, as the iteration count t increases, the inner product $\langle \boldsymbol{p}(t)/\|\boldsymbol{p}(t)\|, \boldsymbol{p}^{mm\star}/\|\boldsymbol{p}^{mm\star}\| \rangle$ consistently increases. Figure 1(c) also illustrates the result of Theorem 1 on multiple inputs (gray dot line is the separating hyperplane). These observations emphasize the gradual alignment between the evolving predictor and the max-margin solution throughout optimization.

Transient optimization dynamics and the role of loss function. While asymptotic direction of gradient descent is determined by $\boldsymbol{p}^{mm}$, intuitively transient dynamics can exhibit bias towards tokens with extreme scores. We aim to capture this intuition in Figure 2(a) which depicts the gradient

trajectories for different scores and loss functions. We have two optimal tokens ($\star$) with scores $\boldsymbol{\gamma}_1 = 1, \boldsymbol{\gamma}_2 = C$ for varying C and we consider correlation loss $\ell(x) = -x$ and exponential loss $\ell(x) = e^{-x}$. In a nutshell, as C grows, it can be seen that $\ell(x) = -x$ is biased towards token with high-score whereas $\ell(x) = e^{-x}$ is biased towards the low-score token. The underlying reason can be seen from the gradient of individual inputs: $\nabla \mathcal{L}_i(\boldsymbol{p}) = \ell_i' \cdot \boldsymbol{K}_i^\top \mathbb{S}'(\boldsymbol{X}\boldsymbol{p})\boldsymbol{X}\boldsymbol{v}$ where $\mathbb{S}'(\cdot)$ is the softmax derivative and $\ell_i' := \ell'(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{p}))$. Assuming $\boldsymbol{p}$ (approximately) selects the optimal tokens, this would simplify to $\ell_i' \approx \ell'(\boldsymbol{\gamma}_i)$ and $\|\nabla \mathcal{L}_i(\boldsymbol{p})\| \propto |\ell'(\boldsymbol{\gamma}_i)| \cdot \boldsymbol{\gamma}_i$. Now, with correlation loss $|\ell'| = 1$, thus, $\|\nabla \mathcal{L}_i(\boldsymbol{p})\| \propto \boldsymbol{\gamma}_i$ and larger score induces larger gradient. Whereas with exponential loss $|\ell'| = e^{-u}$, thus, $\|\nabla \mathcal{L}_i(\boldsymbol{p})\| \propto \boldsymbol{\gamma}_i e^{-\boldsymbol{\gamma}_i}$ and smaller score induces larger gradient explaining the empirical behavior.

We next provide the regularization path analysis that requires relaxed assumptions on both loss function and tokens' score.

Theorem 2 (Regularization Path) *Suppose Assumption A on the loss function holds, and for all $i \in [n]$ and $t \neq \texttt{opt}_i$, scores obey $\boldsymbol{\gamma}_{it} < \boldsymbol{\gamma}_{i\texttt{opt}_i}$. Then the regularization path $\bar{\boldsymbol{p}}(R) = \arg\min_{\|\boldsymbol{p}\| \leq R} \mathcal{L}(\boldsymbol{p})$ satisfies $\lim_{R\to\infty} \bar{\boldsymbol{p}}(R)/R = \boldsymbol{p}^{mm\star}/\|\boldsymbol{p}^{mm\star}\|$.*

Theorem 2 reveals that as we loosen the regularization strength R to achieve ridgeless optimization with $\min_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{p})$, the optimal direction $\bar{\boldsymbol{p}}(R)$ gradually aligns with the max-margin solution $\boldsymbol{p}^{mm\star}$. A central feature of this theorem is its ability to handle non-optimal tokens that possess different arbitrary scores. Thus, it demonstrates that max-margin convergence is a global feature of attention mechanism. As we shall see in the next section, due to nonconvex landscape and nonlinearity of softmax, convergence of regularization path without Assumption B does not imply that Theorem 1 can avoid this condition.

2.2 Local convergence of the attention weights p

Theorem 1 on the global convergence of gradient descent serves as a prelude to the general behavior of the optimization. Once we relax Assumption B by allowing for arbitrary token scores, we will show that $\boldsymbol{p}$ can converge (in direction) to a locally-optimal solution. However, this locally-optimal solution is still characterized in terms of (ATT-SVM) which separates *locally-optimal* tokens from the rest. Our theory builds on two new concepts: locally-optimal tokens and neighbors of these tokens.

Definition 2 (SVM-Neighbor and Locally-Optimal Tokens) *Fix token indices $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$. Solve (ATT-SVM) to obtain $\boldsymbol{p}^{mm} = \boldsymbol{p}^{mm}(\boldsymbol{\alpha})$. Consider tokens $\mathcal{T}_i \subset [T]$ such that $(\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it})^\top \boldsymbol{p}^{mm} = 1$ for all $t \in \mathcal{T}_i$. We refer to $\mathcal{T}_i$ as SVM-neighbors of $\boldsymbol{k}_{i\alpha_i}$. Additionally, tokens $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ are called locally-optimal if for all $i \in [n]$, $t \in \mathcal{T}_i$ scores per Def. 1 obey $\boldsymbol{\gamma}_{i\alpha_i} > \boldsymbol{\gamma}_{it}$.*

To provide a basis for discussing local convergence, we establish a cone centered around $\boldsymbol{p}^{mm}$ using the following construction. Let μ be a positive scalar, and define the cone as:

$$\texttt{cone}_\mu(\boldsymbol{p}^{mm}) := \left\{ \boldsymbol{p} \in \mathbb{R}^d \;\middle|\; \left\langle \frac{\boldsymbol{p}}{\|\boldsymbol{p}\|}, \frac{\boldsymbol{p}^{mm}}{\|\boldsymbol{p}^{mm}\|} \right\rangle \geq 1 - \mu \right\}. \tag{5}$$

In the subsequent theorem, we demonstrate the existence of a scalar $\mu = \mu(\boldsymbol{\alpha}) > 0$ and a radius R such that when R is sufficiently large, there are no stationary points within the intersection of $\texttt{cone}_\mu(\boldsymbol{p}^{mm})$ and the set $\{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R\}$. Further, the gradient descent initialized within this intersection converges in direction to $\boldsymbol{p}^{mm}/\|\boldsymbol{p}^{mm}\|$.

Theorem 3 (Local Convergence of Gradient Descent) *Suppose Assumption A on the loss function ℓ holds and assume $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ are locally-optimal tokens per Definition 2. Then, there exists a scalar $\mu = \mu(\boldsymbol{\alpha}) \in (0, 1)$ and a radius $R > 0$ such that $\texttt{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R\}$ does not contain any stationary points. Further, the gradient descent iterates $\boldsymbol{p}(t+1) = \boldsymbol{p}(t) - \eta \nabla \mathcal{L}(\boldsymbol{p}(t))$ on (ERM) with*

$$\eta \leq O\left(\min\left(\frac{1}{(M_0 + M_1)\bar{\sigma}_{\max}^4}, \frac{\mu - \epsilon}{(1-\mu)}\right)\right), \tag{6}$$

for any $\epsilon \in (0, \min(\mu, 1))$, and any starting point $p(0) \in \texttt{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R\}$ satisfies $\lim_{t\to\infty} \boldsymbol{p}(t)/\|\boldsymbol{p}(t)\| = \boldsymbol{p}^{mm}/\|\boldsymbol{p}^{mm}\|$.

Proof sketch. We provide the proof in four steps:

Step 1. We begin by proving that there are no stationary points within $\texttt{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \left\{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R_\mu\right\}$

for a specific radius R_μ. Let $(\mathcal{T}_i)_{i=1}^n$ denote the set of SVM-neighbors as defined in Definition 2. We define $\bar{\mathcal{T}}_i = [T] - \mathcal{T}_i - \alpha_i$ as the tokens that are non-SVM neighbors. Additionally, let

$$\delta := \frac{1}{2} \min_{i\in[n]} \min_{t\in\mathcal{T}_i, \tau\in\bar{\mathcal{T}}_i} (\boldsymbol{k}_{it} - \boldsymbol{k}_{i\tau})^\top \boldsymbol{p}^{mm}, \qquad A := \max_{i\in[n], t\in[T]} \|\boldsymbol{k}_{it}\| \cdot \|\boldsymbol{p}^{mm}\|, \qquad \mu := \frac{1}{8}\left(\frac{\min(0.5, \delta)}{A}\right)^2.$$

For all $\boldsymbol{q}, \boldsymbol{p} \in \mathsf{cone}_\mu(\boldsymbol{p}^{mm})$ with $\|\boldsymbol{q}\| = \|\boldsymbol{p}^{mm}\|$, we establish the existence of R_μ such that $-\boldsymbol{q}^\top \nabla \mathcal{L}(\boldsymbol{p})$ is strictly positive for $\|\boldsymbol{p}\| \geq R_\mu$. Specifically, we show the existence of positive constants C and c satisfying:

$$C \cdot \max_{i\in[n]} q_i \geq -\langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{q} \rangle \geq c \cdot \min_{i\in[n]} q_i > 0.$$

Here, $q_i = 1 - \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p})_{\alpha_i}$ and $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ are locally-optimal tokens per Definition 2.

Step 2. We demonstrate that for any $\epsilon \in (0, \min(\mu, 1))$, there exists R_ϵ such that all $\boldsymbol{p} \in \mathsf{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R_\epsilon\}$ satisfy

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

Step 3. By leveraging the results from **Step 1** and **Step 2**, we can demonstrate that the gradient iterates, with an appropriate step size, starting from $p(0) \in \mathsf{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R\}$, remain within this cone. Specifically, if $\boldsymbol{p}(t) \in \mathsf{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R\}$, then $\|\boldsymbol{p}(t+1)\| \geq \|\boldsymbol{p}(t)\|$, and

$$\left\langle \frac{\boldsymbol{p}(t+1)}{\|\boldsymbol{p}(t+1)\|}, \frac{\boldsymbol{p}^{mm}}{\|\boldsymbol{p}^{mm}\|} \right\rangle \geq 1 - \mu + O\Big(\eta(\mu - \epsilon) - \eta^2(1-\mu)\Big),$$

which implies that, with the step size η satisfying (6), $\boldsymbol{p}(t+1) \in \mathsf{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R\}$.

Step 4. The remaining part of the proof follows the same reasoning as the proof of Theorem 1 and is provided in the supplementary material. ■

To further illustrate Theorem 3, we can consider Figure 1(b) where $n = 1$ and $T = 3$. In this figure, the point $(0, 0)$ represents the non-optimal tokens, while $(1, 0)$ represents the locally optimal token. Additionally, the gray paths represent the trajectories of gradient descent initiated from different points. By observing the figure, we can see that gradient descent, when properly initialized, converges towards the direction of $\boldsymbol{p}^{mm}$ (depicted by - - -). This direction of convergence effectively separates the locally optimal tokens $(1, 0)$ from the non-optimal token $(0, 0)$.

2.3 Tightness of the locally-optimal token definition

An important question is whether our definition of locally-optimal tokens (Def. 2) covers all token configurations $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ that can be selected by the attention mechanism asymptotically (as $\|\boldsymbol{p}\| \to \infty$). The following theorem essentially establishes the tightness of our definition: It shows that, given $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$, if any of the α_i's have an SVM-neighbor with a higher score, then regularization path will not prefer the $\boldsymbol{p}^{mm}(\boldsymbol{\alpha})$ direction.

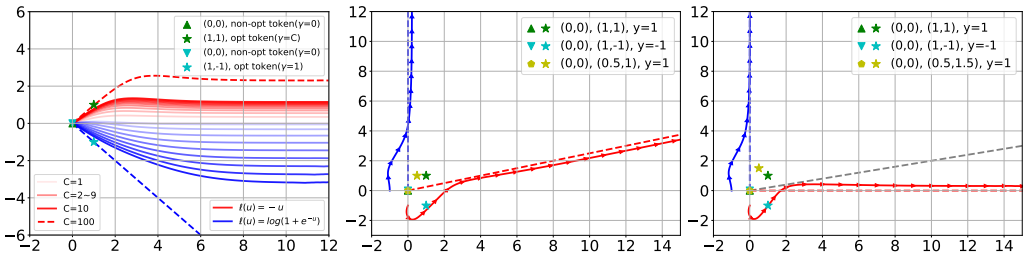
Theorem 4 *Fix indices $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ with SVM-neighbors $(\mathcal{T}_i)_{i=1}^n$. Set $\boldsymbol{p}^{mm} := \boldsymbol{p}^{mm}(\boldsymbol{\alpha})$. Suppose that:*

- *For some $j \in [n]$, there exists $\beta \in \mathcal{T}_j$ with higher score than α_j, i.e., $Y_j \cdot \boldsymbol{v}^\top \boldsymbol{x}_{j\beta} > Y_j \cdot \boldsymbol{v}^\top \boldsymbol{x}_{j\alpha_j}$.*
- *For all $i \in [n]$ and $t \in \mathcal{T}_i$, the vectors $\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it}$ are linearly independent (We note that this holds for almost all datasets).*

For any $\varepsilon > 0$, there exists $R_\varepsilon > 0$ as follows: Consider the neighborhood of $\boldsymbol{p}^{mm}$: $C_\varepsilon = \mathsf{cone}_\varepsilon(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R_\varepsilon\}$. Define the local path $\bar{\boldsymbol{p}}(R) = \min_{\boldsymbol{p}\in C_\varepsilon, \|\boldsymbol{p}\| \leq R} \mathcal{L}(\boldsymbol{p})$. Then $\lim_{R\to\infty} \frac{\bar{\boldsymbol{p}}(R)}{\|\bar{\boldsymbol{p}}(R)\|} \neq \frac{\boldsymbol{p}^{mm}}{\|\boldsymbol{p}^{mm}\|}$.

3 Joint Convergence of Head $\boldsymbol{v}$ and Attention Weights $\boldsymbol{p}$

In this section, we extend the preceding results to the general case of joint optimization of head $\boldsymbol{v}$ and attention weights $\boldsymbol{p}$ using a logistic loss function. To this aim, we focus on regularization path analysis, which involves solving (ERM) under ridge constraints and examining the solution trajectory as the constraints are relaxed.



(a) The role of loss function (b) All inputs are support vectors (c) (0.5,1.5) is not a support vector

Figure 2: **(a)** Global convergence of $\boldsymbol{p}$ with different loss functions and scores. **(b)&(c)** Joint convergence of attention weights $\boldsymbol{p}$ and classifier head $\boldsymbol{v}$ to max-margin directions.

High-level intuition. Since the prediction is linear as a function of $\boldsymbol{v}$, logistic regression in $\boldsymbol{v}$ can exhibit its own implicit bias to a max-margin solution. Concretely, define the attention features $\boldsymbol{x}_i^{\boldsymbol{p}} = \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p})$ and define the dataset $\mathcal{S}^{\boldsymbol{p}} = (Y_i, \boldsymbol{x}_i^{\boldsymbol{p}})$. If the dataset $\mathcal{S}^{\boldsymbol{p}}$ is separable by $\boldsymbol{v}$, then optimizing only $\boldsymbol{v}$ will converge in the direction of the max-margin classifier by setting $\boldsymbol{r}_i \leftarrow \boldsymbol{x}_i^{\boldsymbol{p}}$:

$$\boldsymbol{v}^{mm} = \arg\min_{\boldsymbol{v}\in\mathbb{R}^d} \|\boldsymbol{v}\| \quad \text{subject to} \quad Y_i \cdot \boldsymbol{v}^\top \boldsymbol{r}_i \geq 1. \tag{CLS-SVM}$$

This motivates a clear question: *Under what conditions, optimizing $\boldsymbol{v}, \boldsymbol{p}$ jointly will converge to their respective max-margin solutions?* We study this question in two steps. Loosely speaking, we will first assume that when solving (CLS-SVM), all inputs $i \in [n]$ are also the support vectors. We will then relax this condition to uncover a more general implicit bias for $\boldsymbol{p}$. Throughout we assume that the joint problem is separable and there exists $(\boldsymbol{v}, \boldsymbol{p})$ asymptotically achieving zero training loss.

3.1 When all attention features are support vectors

In (CLS-SVM), define *label margin* to be $1/\|\boldsymbol{v}^{mm}\|$. Our first insight in quantifying joint implicit bias is that, optimal tokens admit a natural definition: Those that maximize the downstream label margin when selected. This is formalized below where we assume that: (1) selecting the token indices $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ from each input data achieves the largest label margin. (2) The optimality of the $\boldsymbol{\alpha}$ choice is strict in the sense that mixing other tokens will shrink the label margin in (CLS-SVM).

Assumption C *Let $\Gamma > 0$ be the label margin when solving (CLS-SVM) with $\boldsymbol{r}_i \leftarrow \boldsymbol{x}_{i\alpha_i}$. There exists $\nu > 0$ such that for all $\boldsymbol{p}$, solving (CLS-SVM) with $\boldsymbol{r}_i \leftarrow \boldsymbol{x}_i^{\boldsymbol{p}}$ results in a label margin of at most $\Gamma - \nu \cdot \max_{i\in[n]}(1 - \boldsymbol{s}_{i\alpha_i})$ where $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p})$.*

Example: To gain intuition, let us fix $\boldsymbol{v}_\star \in \mathbb{R}^d$ and consider the dataset obeying $\boldsymbol{x}_{i1} = Y_i \cdot \boldsymbol{v}_\star$ and $\|\boldsymbol{x}_{it}\| < \|\boldsymbol{v}_\star\|$ for all $t \geq 2$ and all $i \in [n]$. For this dataset, we can choose $\alpha_i = 1$, $\Gamma = \|\boldsymbol{v}_\star\|$ and $\nu = \|\boldsymbol{v}_\star\| - \sup_{i\in[n],t\geq 2} \|\boldsymbol{x}_{it}\|$.

Theorem 5 *Consider the ridge-constrained solutions $(\boldsymbol{v}_r, \boldsymbol{p}_R)$ of (ERM) defined as*

$$\boldsymbol{v}_r, \boldsymbol{p}_R = \arg\min_{\|\boldsymbol{v}\|\leq r, \|\boldsymbol{p}\|\leq R} \mathcal{L}(\boldsymbol{v}, \boldsymbol{p}).$$

Suppose Assumption C holds for some $\Gamma, \nu > 0$. As $r, R \to \infty$, the joint regularization path $(\boldsymbol{v}_r, \boldsymbol{p}_R)$ converges as follows: $\frac{\boldsymbol{p}_R}{R} \to \frac{\boldsymbol{p}^{mm}}{\|\boldsymbol{p}^{mm}\|}$ where $\boldsymbol{p}^{mm}$ is the solution of (ATT-SVM). $\frac{\boldsymbol{v}_r}{r} \to \frac{\boldsymbol{v}^{mm}}{\|\boldsymbol{v}^{mm}\|}$ where $\boldsymbol{v}^{mm}$ is the solution of (CLS-SVM) with $\boldsymbol{r}_i = \boldsymbol{x}_{i1}$.

As further discussion, consider Figure 2(b) where we set $n = 3, T = d = 2$ and $\boldsymbol{W} =$ Identity. All three inputs share the point $(0, 0)$ which corresponds to their non-optimal tokens. The optimal tokens (denoted by $\star$) are all support vectors of the (CLS-SVM) since $\boldsymbol{v}^{mm} = [0, 1]$ is the optimal classifier direction (blue color). Because of this, $\boldsymbol{p}^{mm}$ will separate optimal tokens from $(0, 0)$ coordinate via (ATT-SVM) which results in the red direction (yellow and teal $\star$ are the support tokens).

3.2 General solution when selecting one token per input

Can we relax Assumption C, and if so, what is the resulting behavior? Consider the scenario where the optimal $\boldsymbol{p}$ diverges to ∞ and ends up selecting one token per input. Suppose this $\boldsymbol{p}$ selects some

coordinates $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$. Let $\mathcal{N} \subset [n]$ be the set of indices where the associated token $\boldsymbol{x}_{i\alpha_i}$ is **not** a support vector when solving (CLS-SVM). Our intuition is as follows: Even if we slightly perturb this $\boldsymbol{p}$ choice and mix other tokens $t \neq \alpha_i$ over the input set $\mathcal{N} \subset [n]$, since $\mathcal{N}$ is not support vector for (CLS-SVM), we can preserve the label margin (by only preserving the support vectors $[n] - \mathcal{N}$). This means that $\boldsymbol{p}$ may not have to enforce *max-margin* constraint over inputs $i \in \mathcal{N}$, instead, it suffices to just select these tokens (asymptotically). This results in the following **relaxed SVM** problem:

$$\boldsymbol{p}^{\text{relax}} = \min_{\boldsymbol{p}} \|\boldsymbol{p}\| \quad \text{such that} \quad \boldsymbol{p}^\top(\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it}) \geq \begin{cases} 1 & \text{for all} \quad t \neq \alpha_i,\ i \in [n] - \mathcal{N} \\ 0 & \text{for all} \quad t \neq \alpha_i,\ i \in \mathcal{N} \end{cases}. \tag{7}$$

Here, $\boldsymbol{p}^\top \boldsymbol{x}_{i\alpha_i} \geq 0$ corresponds to the *selection* idea. Building on this intuition, the following theorem captures the generalized behavior of the joint regularization path.

Theorem 6 *Consider the path of $(\boldsymbol{v}_r, \boldsymbol{p}_R)$ as $r, R \to \infty$ as in Theorem 5. Suppose $\mathbb{S}(\boldsymbol{K}_i \boldsymbol{p}_R)_{\alpha_i} \to 1$, i.e., the tokens $(\alpha_i)_{i=1}^n$ are asymptotically selected. Let $\boldsymbol{v}^{\text{mm}}$ be the solution of (CLS-SVM) with $\boldsymbol{r}_i = \boldsymbol{x}_{i\alpha_i}$ and $\mathcal{N}$ be its set of non-support indices. Suppose Assumption C holds over the support vectors $[n] - \mathcal{N}$. Then, $\frac{\boldsymbol{v}_r}{r} \to \frac{\boldsymbol{v}^{\text{mm}}}{\|\boldsymbol{v}^{\text{mm}}\|}$ and $\frac{\boldsymbol{p}_R}{R} \to \frac{\boldsymbol{p}^{\text{relax}}}{\|\boldsymbol{p}^{\text{relax}}\|}$ where $\boldsymbol{p}^{\text{relax}}$ is the solution of (7) with α_i choices.*

To illustrate this numerically, consider Figure 2(c) which modifies Figure 2(b) by pushing the yellow $\star$ to the northern position $(0.5, 1.5)$. We still have $\boldsymbol{v}^{\text{mm}} = [0, 1]$ however the yellow $\star$ is no longer a support vector of (CLS-SVM). Thus, $\boldsymbol{p}$ solves the relaxed problem which separates green and teal $\star$'s by enforcing the max-margin constraint on $\boldsymbol{p}$ (which is the red direction). Instead, yellow $\star$ only needs to achieve positive correlation with $\boldsymbol{p}$ (unlike Figure 2(c) where it dictates the direction).

4 Regularization Path of Attention with Nonlinear Head

So far our discussion has focused on the attention model with linear head. However, the conceptual ideas on optimal token selection via margin maximization also extends to a general nonlinear model under mild assumptions. The aim of this section is showcasing this generalization. Specifically, we consider the prediction model $f(\boldsymbol{X}) = \psi(\boldsymbol{X}^\top \mathbb{S}(\boldsymbol{K}\boldsymbol{p}))$ where $\psi(\cdot) : \mathbb{R}^d \to \mathbb{R}$ generalizes the linear head $\boldsymbol{v}$ of our attention model. For instance, following exposition in Section 1.1, $\psi(\cdot)$ can represent a multilayer transformer with $\boldsymbol{p}$ being a tunable prompt at the input layer. Recall that $\mathcal{S} = (\boldsymbol{X}_i, \boldsymbol{K}_i, Y_i)_{i=1}^n$ is the dataset of the input-key-label tuples. We consider the training risk

$$\mathcal{L}(\boldsymbol{p}) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, \psi(\boldsymbol{X}_i^\top \boldsymbol{s}_i^{\boldsymbol{p}})) \quad \text{where} \quad \boldsymbol{s}_i^{\boldsymbol{p}} = \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p}) \in \mathbb{R}^T. \tag{8}$$

The challenge with nonlinear $\psi(\cdot)$ is that, we lack a clear score function (Def. 1) unlike the previous sections. The assumption below introduces a generic condition that splits the tokens of each $\boldsymbol{X}_i$ into an *optimal* set $\mathcal{O}_i$ and *non-optimal* set $\bar{\mathcal{O}}_i = [T] - \mathcal{O}_i$. In words, non-optimal tokens are those that strictly increase the training risk $\mathcal{L}(\boldsymbol{p})$ if they are not fully suppressed by attention probabilities $\boldsymbol{s}_i^{\boldsymbol{p}}$.

Assumption D (Mixing non-optimal tokens hurt) *There exists sets $(\mathcal{O}_i)_{i=1}^n \subset [T]$ as follows. Let $q_i^{\boldsymbol{p}} = \sum_{t \in \bar{\mathcal{O}}_i} \boldsymbol{s}_{it}^{\boldsymbol{p}}$ be the sum of softmax similarities over the non-optimal set for $\boldsymbol{p}$. Set $q_{\max}^{\boldsymbol{p}} = \max_{i \in [n]} q_i^{\boldsymbol{p}}$. For any $\Delta > 0$, there exists $\rho < 0$ such that:*

$$\text{For all } \boldsymbol{p}, \boldsymbol{p}' \in \mathbb{R}^d, \text{ if } \log(q_{\max}^{\boldsymbol{p}}) \leq (1 + \Delta) \log(q_{\max}^{\boldsymbol{p}'}) \wedge \rho, \text{ then } \mathcal{L}(\boldsymbol{p}) < \mathcal{L}(\boldsymbol{p}').$$

This assumption is titled *mixing hurts* because the attention output $\boldsymbol{X}_i^\top \boldsymbol{s}_i^{\boldsymbol{p}}$ is mixing the tokens of $\boldsymbol{X}_i$ and our condition is that, to achieve optimal risk, this mixture should not contain any non-optimal tokens. In particular, we require that, a model $\boldsymbol{p}$ that contains *exponentially less non-optimality* (quantified via $\log(q_{\max})$) compared to $\boldsymbol{p}'$ is strictly preferable. As we discuss in the supplementary material, Theorem 2 is in fact a concrete instance (with linear head $\boldsymbol{v}$) satisfying this condition.

Before stating our generic theorem, we need to introduce the max-margin separator towards which regularization path of attention will converge. This is a slightly general version of Section 2's (ATT-SVM) problem where we allow for a set of optimal tokens $\mathcal{O}_i$ for each input.

$$\boldsymbol{p}^{\text{mm}} = \arg\min_{\boldsymbol{p}} \|\boldsymbol{p}\| \quad \text{subject to} \quad \max_{\alpha \in \mathcal{O}_i} \min_{\beta \in \bar{\mathcal{O}}_i} \boldsymbol{p}^\top(\boldsymbol{k}_{i\alpha} - \boldsymbol{k}_{i\beta}) \geq 1 \quad \text{for all} \quad i \in [n]. \tag{ATT-SVM'}$$

Unlike (ATT-SVM), this problem is not necessarily convex when the optimal set $\mathcal{O}_i$ is not a singleton. To see this, imagine $n = d = 1$ and $T = 3$: Set the two optimal tokens as $\boldsymbol{k}_1 = 1$ and $\boldsymbol{k}_2 = -1$ and the

non-optimal token as $\boldsymbol{k}_3 = 0$. The solution set of (ATT-SVM') is $\boldsymbol{p}^{mm} \in \{-1, 1\}$ whereas their convex combination $\boldsymbol{p} = 0$ violates the constraints. To proceed, our final result establishes the convergence of regularization path to the solution set of (ATT-SVM') under Assumption D.

Theorem 7 *Let $\mathcal{P}^{mm}$ be the set of global minima of* (ATT-SVM'). *Suppose its objective $\Gamma := \|\boldsymbol{p}^{mm}\|$ is finite and Assumption D holds. Let $\texttt{dist}(\cdot,\cdot)$ denote the ℓ_2-distance between a vector and a set. Following* (8), *define $\bar{\boldsymbol{p}}(R) = \arg\min_{\|\boldsymbol{p}\|\le R} \mathcal{L}(\boldsymbol{p})$. We have that $\lim_{R\to\infty} \texttt{dist}\left(\Gamma\frac{\bar{\boldsymbol{p}}(R)}{R}, \mathcal{P}^{mm}\right) = 0$.*

We note that Theorem 2 is a corollary of this result where the set $\mathcal{P}^{mm}$ is a singleton.

5 Related Work

Implicit Regularization. The implicit bias of gradient descent in classification tasks involving separable data has been extensively examined by [19, 20, 21, 22, 23, 24]. These works typically use logistic loss or, more generally, exponentially-tailed losses to make connections to margin maximization. These results are also extended to non-separable data by [25, 26, 27]. Furthermore, there have been notable investigations into the implicit bias in regression problems/losses utilizing techniques such as mirror descent [28, 20, 29, 30, 31, 32]. In addition, several papers have explored the implicit bias of stochastic gradient descent [33, 34, 35, 36, 37, 38], as well as adaptive and momentum-based methods [39, 40, 41, 42]. Although there are similarities between our optimization approach for $\boldsymbol{v}$ and existing works, the optimization of $\boldsymbol{p}$ stands out as significantly different. Firstly, our optimization problem is nonconvex, introducing new challenges and complexities. Secondly, it necessitates the introduction of novel concepts such as locally-optimal tokens and requires a fresh analysis specifically tailored to the cones surrounding them.

Attention Mechanism. Transformers, introduced by [6], revolutionized the field of NLP and machine translation, with earlier works on self-attention by [43, 44, 45, 46]. Self-attention differs from traditional models like MLPs and CNNs by leveraging global interactions for feature representations, showing exceptional empirical performance. However, the underlying mechanisms and learning processes of the attention layer remain unknown. Recent studies such as [47, 48, 49, 50, 51] have focused on specific aspects like representing sparse functions, convex-relaxations, and expressive power. [52, 53] have developed initial results to characterize the optimization and generalization dynamics of attention. [17] is another closely related work where the authors analyze the same attention model (ERM) as us. However, all of these works make stringent assumptions on the data, namely, tokens are tightly clusterable or can be clearly split into clear relevant and irrelevant sets. Additionally [53] requires assumptions on initialization and [52] considers a simplified attention structure where the attention matrix is not directly parameterized with respect to the input. Our work offers a comprehensive optimization-theoretic analysis of the attention model by establishing a formal connection to max-margin problems. Notably, our work presents the first theoretical understanding of the implicit bias exhibited by gradient descent methods in the context of the attention model.

6 Discussion

We have provided a thorough optimization-theoretic characterization of the fundamental attention model $f(\boldsymbol{X}) = \boldsymbol{v}^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{p})$ by formally connecting it to max-margin problems. We first established the convergence of gradient descent on $\boldsymbol{p}$ (or equivalently $\boldsymbol{W}$) in isolation. We also explored joint convergence of $(\boldsymbol{v}, \boldsymbol{p})$ via regularization path which revealed surprising implicit biases such as (7). These findings motivate several exciting avenues for future research. An immediate open problem is characterizing the (local) convergence of gradient descent for joint optimization of $(\boldsymbol{v}, \boldsymbol{p})$. Another major direction is to extend similar analysis to study self-attention layer (3) or to allow for multiple tunable tokens (where $\boldsymbol{p}$ becomes a matrix). Either setting will enrich the problem by allowing the attention to discover multiple hyperplanes to separate tokens. While we assumed the tokens to be separable, it would be interesting to relax this assumption by leveraging results developed for logistic regression analysis [26, 19]. Ideas from these results can also be useful for characterizing the non-asymptotic behavior of how gradient descent aligns with the max-margin direction.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *The International Conference on Learning Representations*,

2015.

[2] 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.

[3] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and et al. Language models are unsupervised multitask learners. *arXiv preprint arXiv:1911.05786*, 2019.

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

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

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

[9] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[10] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[14] 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.

[15] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[16] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

[17] Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

[18] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

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

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

[27] 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.

[28] 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.

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

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

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

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

[33] 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.

[34] 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.

[35] 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.

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

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

[38] 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.

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

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

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

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

[51] 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.

[52] 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.

[53] 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.

[54] Olvi L Mangasarian. *Nonlinear programming*. SIAM, 1994.

[55] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

[56] Peter Bartlett. For valid generalization the size of the weights is more important than the size of the network. *Advances in neural information processing systems*, 9, 1996.

[57] Albert B Novikoff. On convergence proofs for perceptrons. Technical report, STANFORD RESEARCH INST MENLO PARK CA, 1963.

[58] Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E Schapire. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686, 1998.

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

[60] Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR, 2013.

[61] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

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

[63] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[64] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

[65] Spencer Frei, Gal Vardi, Peter L Bartlett, and Nathan Srebro. Benign overfitting in linear classifiers and leaky relu networks from kkt conditions for margin maximization. *arXiv e-prints*, pages arXiv–2303, 2023.

[66] Gal Vardi, Ohad Shamir, and Nati Srebro. On margin maximization in linear and relu networks. *Advances in Neural Information Processing Systems*, 35:37024–37036, 2022.

[67] Gal Vardi and Ohad Shamir. Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, pages 4224–4258. PMLR, 2021.

[68] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[69] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1059–1068, 2018.

[70] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. Behavior sequence transformer for e-commerce recommendation in alibaba. In *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*, pages 1–4, 2019.

[71] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1441–1450, 2019.

[72] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[73] Michael Janner, Qiyang Li, and Sergey Levine. Reinforcement learning as one big sequence modeling problem. In *ICML 2021 Workshop on Unsupervised Reinforcement Learning*, 2021.

[74] Qinqing Zheng, Amy Zhang, and Aditya Grover. Online decision transformer. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[75] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[76] 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.

[77] Zi-Hang Jiang, Qibin Hou, Li Yuan, Daquan Zhou, Yujun Shi, Xiaojie Jin, Anran Wang, and Jiashi Feng. All tokens matter: Token labeling for training better vision transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 18590–18602, 2021.

[78] Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.

[79] Zizheng Pan, Bohan Zhuang, Jing Liu, Haoyu He, and Jianfei Cai. Scalable vision transformers with hierarchical pooling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 377–386, 2021.

[80] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, volume 34, pages 13937–13949, 2021.

[81] Youwei Liang, GE Chongjian, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

[82] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12165–12174, 2022.

[83] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10809–10818, 2022.

[84] Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pages 5562–5571. PMLR, 2021.

[85] Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: Nngp and ntk for deep attention networks. In *International Conference on Machine Learning*, pages 4376–4386. PMLR, 2020.

[86] Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

[87] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2018.

[88] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2019.

[89] Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. *arXiv preprint arXiv:2009.11264*, 2020.

[90] Valerii Likhosherstov, Krzysztof Choromanski, and Adrian Weller. On the expressive power of self-attention matrices. *arXiv preprint arXiv:2106.03764*, 2021.

[91] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2019.

[92] Yoav Levine, Noam Wies, Or Sharir, Hofit Bata, and Amnon Shashua. Limits to depth efficiencies of self-attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 22640–22651, 2020.

[93] Charlie Snell, Ruiqi Zhong, Dan Klein, and Jacob Steinhardt. Approximating how single head attention learns. *arXiv preprint arXiv:2103.07601*, 2021.

[94] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[95] 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.

Roadmap. The appendix is organized as follows: Section A provides basic facts about the training risk. Section B presents the proof of local and global gradient descent and regularized path for learning $\boldsymbol{p} \in \mathbb{R}^d$ with a fixed $\boldsymbol{v} \in \mathbb{R}^d$ choice. Section C provides the proof of regularized path applied to the general case of joint optimization of head $\boldsymbol{v}$ and attention weights $\boldsymbol{p}$ using a logistic loss function. Section D presents the proof for the regularized path applied to a more general model $f(\boldsymbol{X}) = \psi(\boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}^\top \boldsymbol{p}))$ with a nonlinear head ψ. Section E provides implementation details. Finally, Section F discusses additional related work on implicit bias and self-attention.

Corrections and Refinements. We have made the following changes to the main submission.

- In the first bullet point of Theorem 4, we corrected indices i into j. This was a typo.
- In the statement of Theorem 4, we now include the norm lower bound R_ε over the conic neighborhood. Note that, this is consistent with the setting of success guarantee Theorem 3 and the main message on the tightness of local optimality remains intact.
- In Theorem 6, we corrected the statement from $\boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p}_R) \to \boldsymbol{x}_{i\alpha_i}$ to $\mathbb{S}(\boldsymbol{K}_i \boldsymbol{p}_R)_{\alpha_i} \to 1$. Note that, the former statement does not actually imply token index α_i is selected because combination of other tokens can still add up to $\boldsymbol{x}_{i\alpha_i}$. Instead, the new statement says softmax probability fully concentrates over α_i.

Table of Contents

A Addendum to Section 1

A.1 Preliminaries on the Training Risk

Recall the objective

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{p}, \boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i \cdot f(\boldsymbol{X}_i)). \tag{9}$$

with the generic prediction model $f(\boldsymbol{X}) = \psi(\boldsymbol{X}^\top \mathbb{S}(\boldsymbol{K}\boldsymbol{p}))$ and $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^\top$. Here, we write down the gradients of $\boldsymbol{W}$ and $\boldsymbol{p}$ in (9) to highlight the connection. Set $\boldsymbol{q} := \boldsymbol{W}^\top \boldsymbol{p}$, $z\{\boldsymbol{X}\} := \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{K}\boldsymbol{p})$ and $\boldsymbol{a}\{\boldsymbol{X}\} := \boldsymbol{K}\boldsymbol{p}$. Given $\boldsymbol{X}$ and using $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}^\top$, we have that

$$\nabla_{\boldsymbol{q}} f_\psi(\Theta) = \boldsymbol{X}^\top \mathbb{S}'(\boldsymbol{a}\{\boldsymbol{X}\})\boldsymbol{X} \cdot \nabla\psi(z\{\boldsymbol{X}\}) \tag{10a}$$

$$\nabla_{\boldsymbol{p}} f_\psi(\Theta) = \boldsymbol{W}\nabla_{\boldsymbol{q}} f_\psi(\Theta), \tag{10b}$$

$$\nabla_{\boldsymbol{W}} f_\psi(\Theta) = \boldsymbol{p}\nabla_{\boldsymbol{q}}^\top f_\psi(\Theta). \tag{10c}$$

Setting $\psi(\boldsymbol{z}) = \boldsymbol{v}^\top \boldsymbol{z}$ and recalling the score definition $\boldsymbol{\gamma} = \boldsymbol{X}\boldsymbol{v}$, for linear head, we obtain

$$\nabla_{\boldsymbol{q}} f_\psi(\Theta) = \boldsymbol{X}^\top \mathbb{S}'(\boldsymbol{a}\{\boldsymbol{X}\})\boldsymbol{\gamma} \tag{11a}$$

$$\nabla_{\boldsymbol{p}} f_\psi(\Theta) = \boldsymbol{W}\nabla_{\boldsymbol{q}} f_\psi(\Theta) = \boldsymbol{K}^\top \mathbb{S}'(\boldsymbol{a}\{\boldsymbol{X}\})\boldsymbol{\gamma}, \tag{11b}$$

$$\nabla_{\boldsymbol{W}} f_\psi(\Theta) = \boldsymbol{p}\nabla_{\boldsymbol{q}}^\top f_\psi(\Theta) = \boldsymbol{p}\boldsymbol{\gamma}^\top \mathbb{S}'(\boldsymbol{a}\{\boldsymbol{X}\})\boldsymbol{X}. \tag{11c}$$

Note that the gradient of $\boldsymbol{W}$ is rank-1 with fixed left singular direction. The proof of Lemma 1 below shows that solutions induced by matrix $\boldsymbol{W}$ and vectors $\boldsymbol{q}, \boldsymbol{p}$ can be mapped to each other exactly.

A.2 Proof of Lemma 1

Proof. Let us prove the result for a general step size sequence $(\eta_t)_{t\geq 0}$. By our assumption $\psi : \mathbb{R}^d \to \mathbb{R}$ and $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ are differentiable functions. Recall $\mathcal{L}(\boldsymbol{p}) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \psi(\boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{p})))$ and $\mathcal{L}(\boldsymbol{W}) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \psi(\boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}^\top\boldsymbol{a})))$ for fixed $\boldsymbol{a}$. Suppose claim is true till iteration t. For iteration $t+1$, using $\boldsymbol{W}_t^\top \boldsymbol{a} = \boldsymbol{p}_t$, define and observe that

$$\boldsymbol{S}_i^t = \mathbb{S}'(\boldsymbol{X}_i\boldsymbol{W}_t^\top\boldsymbol{a}) = \mathbb{S}'(\boldsymbol{X}_i\boldsymbol{p}_t) \tag{12}$$

$$\boldsymbol{s}_i^t = \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}_t^\top\boldsymbol{a}) = \mathbb{S}(\boldsymbol{X}_i\boldsymbol{p}_t) \tag{13}$$

$$z\{\boldsymbol{X}_i\} := \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{p}_t) = \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}_t^\top\boldsymbol{a}) \tag{14}$$

for all $i \in [n]$. Thus, recalling (10a) and (10c), and defining $\ell_i' = \ell'(Y_i, \psi(z\{\boldsymbol{X}_i\}))$ we have that

$$\nabla_{\boldsymbol{p}} \ell(Y_i, \psi(\boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{p}_t))) = \ell_i' \cdot \boldsymbol{X}_i^\top \boldsymbol{S}_i^t \boldsymbol{X}_i \cdot \nabla\psi(z\{\boldsymbol{X}_i\}), \tag{15}$$

$$\nabla_{\boldsymbol{W}} \ell(Y_i, \psi(\boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{W}_t^\top\boldsymbol{a}))) = \boldsymbol{a}\left(\ell_i' \cdot \boldsymbol{X}_i^\top \boldsymbol{S}_i^t \boldsymbol{X}_i \cdot \nabla\psi(z\{\boldsymbol{X}_i\})\right)^\top. \tag{16}$$

Consequently, we found that gradient is rank-1 with left singular space equal given by $\boldsymbol{a}$

$$\nabla_{\boldsymbol{W}} \mathcal{L}_W(\boldsymbol{W}_t) = \boldsymbol{a}\nabla_{\boldsymbol{p}}^\top \mathcal{L}_q(\boldsymbol{p}_t).$$

Since $\boldsymbol{W}_t$'s left singular space is guaranteed to be in $\boldsymbol{a}$ (including $\boldsymbol{W}_0$ by initialization), we only need to study the right singular vector. Using the induction till t, this yields

$$\boldsymbol{W}_{t+1}^\top \boldsymbol{a} = \boldsymbol{W}_t^\top \boldsymbol{a} - \eta_t \|\boldsymbol{a}\|^{-2} \nabla_{\boldsymbol{W}}^\top \mathcal{L}_W(\boldsymbol{W}_t)\boldsymbol{a} \tag{17}$$

$$= \boldsymbol{p}_t - \eta_t \|\boldsymbol{a}\|^{-2} \boldsymbol{a}^\top\boldsymbol{a} \nabla_{\boldsymbol{p}} \mathcal{L}_q(\boldsymbol{p}_t) \tag{18}$$

$$= \boldsymbol{q}_{t+1}. \tag{19}$$

This concludes the induction. ∎

B Addendum to Section 2

B.1 Local Gradient Condition

Lemma 2 (Key lemma) *Let $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^d$, $\boldsymbol{a} = \boldsymbol{K}\boldsymbol{q}$, $\boldsymbol{s} = \mathbb{S}(\boldsymbol{K}\boldsymbol{p})$, $\boldsymbol{\gamma} = \boldsymbol{X}\boldsymbol{v}$. Set $\Gamma = \sup_{t,\tau\in[T]} |\boldsymbol{\gamma}_t - \boldsymbol{\gamma}_\tau|$ and $A = \sup_{t\in[T]} \|\boldsymbol{k}_t\| \cdot \|\boldsymbol{q}\|$. We have that*

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

Proof. Set $\bar{\gamma} = \sum_{t=1}^{T} \boldsymbol{\gamma}_t \boldsymbol{s}_t$. $\boldsymbol{\gamma}_1 - \bar{\gamma} = \sum_{t\geq 2}^{T} (\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)\boldsymbol{s}_t$. Also note that

$$|\bar{\gamma} - \boldsymbol{\gamma}_1| \leq \Gamma(1-\boldsymbol{s}_1).$$

Proceeding,

$$\begin{aligned}
\boldsymbol{a}^\top \mathrm{diag}(\boldsymbol{s})\boldsymbol{\gamma} - \boldsymbol{a}^\top \boldsymbol{s}\boldsymbol{s}^\top \boldsymbol{\gamma} &= \sum_{t=1}^{T} \boldsymbol{a}_t \boldsymbol{\gamma}_t \boldsymbol{s}_t - \sum_{t=1}^{T} \boldsymbol{a}_t \boldsymbol{s}_t \sum_{t=1}^{T} \boldsymbol{\gamma}_t \boldsymbol{s}_t \\
&= \boldsymbol{a}_1 \boldsymbol{s}_1 (\boldsymbol{\gamma}_1 - \bar{\gamma}) - \sum_{t\geq 2} \boldsymbol{a}_t \boldsymbol{s}_t (\bar{\gamma} - \boldsymbol{\gamma}_t). \qquad (20)
\end{aligned}$$

Now using $|\sum_{t\geq 2} \boldsymbol{a}_t \boldsymbol{s}_t(\bar{\gamma} - \boldsymbol{\gamma}_t) - \sum_{t\geq 2} \boldsymbol{a}_t \boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)| \leq A\Gamma(1-\boldsymbol{s}_1)^2$, we obtain[1]

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

Above, in the last inequality (i.e., $\pm$ on the right handside), we used the fact that

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

■

This lemma will play a key role in the following lemma.

Lemma 3 (Local Gradient Condition) *Let $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ be locally-optimal tokens per Definition 2. Define $\mathsf{cone}_\mu(\boldsymbol{p}^{mm})$ to be the set of vectors obeying $\mathsf{corr}(\boldsymbol{p}, \boldsymbol{p}^{mm}) \geq 1-\mu$. There exists a scalar $\mu = \mu(\boldsymbol{\alpha}) > 0$ such that for sufficiently large $R = R_\mu$:*

- *There is no stationary point within $\mathsf{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R\}$.*
- *Let $q_i = 1 - \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p})_{\alpha_i}$ and $\ell_i' = \ell'(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p})) < 0$, $\gamma_i^{gap} = \min_{t\in\mathcal{T}_i} Y_i \cdot (\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{it})^\top \boldsymbol{v}$, $\bar{\gamma}_i^{gap} = \max_{t\in\mathcal{T}_i} Y_i \cdot (\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{it})^\top \boldsymbol{v}$. For all $\boldsymbol{q}, \boldsymbol{p} \in \mathsf{cone}_\mu(\boldsymbol{p}^{mm})$ with $\|\boldsymbol{q}\| = \|\boldsymbol{p}^{mm}\|$, we have*

$$\frac{2}{n} \sum_{i\in[n]} \ell_i' \cdot q_i \cdot \bar{\gamma}_i^{gap} \geq \langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{q} \rangle \geq \frac{1}{8n} \sum_{i\in[n]} \ell_i' \cdot q_i \cdot \gamma_i^{gap}. \qquad (21)$$

  Note that above $-\ell_i'$ and $\gamma_i^{gap}, \bar{\gamma}_i^{gap}$ are upper/lower bounded by positive dataset-dependent constants. The only term that can vanish (as $\|\boldsymbol{p}\| \to \infty$) is q_i. Consequently, there exists constants $C, c > 0$ such that,

$$C \cdot \max_{i\in[n]} q_i \geq -\langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{q} \rangle \geq c \cdot \min_{i\in[n]} q_i > 0. \qquad (22)$$

  Note that, the identical bound holds by setting $\boldsymbol{q} = \boldsymbol{p}^{mm}$ or $\boldsymbol{q} = \|\boldsymbol{p}^{mm}\|\boldsymbol{p}/\|\boldsymbol{p}\|$.

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

- *Denote $\bar{\boldsymbol{p}} = \|\boldsymbol{p}^{mm}\|\boldsymbol{p}/\|\boldsymbol{p}\|$. For any $\pi > 0$, there exists $R := R_\pi$ such that all $\boldsymbol{p} \in \mathsf{cone}_\mu(\boldsymbol{p}^{mm})$ with $\|\boldsymbol{p}\| \geq R$ obeys*

$$\langle \nabla \mathcal{L}(\boldsymbol{p}), \bar{\boldsymbol{p}} \rangle \geq (1+\pi)\langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{p}^{mm} \rangle,$$

Proof. Let $\boldsymbol{p}^{mm} = \boldsymbol{p}^{mm}(\boldsymbol{\alpha})$ be the solution of (ATT-SVM). Define $\mathsf{cone}_{\mu,R}(\boldsymbol{p}^{mm}) = \{\boldsymbol{p} \in \mathbb{R}^d \;\big|\; \mathtt{corr}(\boldsymbol{p}, \boldsymbol{p}^{mm}) \geq 1-\mu,\ \|\boldsymbol{p}\| \geq R\}$. Let $(\mathcal{T}_i)_{i=1}^n$ be the set of all SVM-neighbors per Definition 2. Let $\bar{\mathcal{T}}_i = [T] - \mathcal{T}_i - \{\alpha_i\}$ be the non-SVM-neighbor tokens. Introduce the notation

$$\Theta = 1/\|\boldsymbol{p}^{mm}\|, \tag{23}$$

$$\delta = 0.5 \min_{i\in[n]} \min_{t\in\mathcal{T}_i, \tau\in\bar{\mathcal{T}}_i} (\boldsymbol{k}_{it} - \boldsymbol{k}_{i\tau})^\top \boldsymbol{p}^{mm}, \tag{24}$$

$$A = \max_{i\in[n], t\in[T]} \|\boldsymbol{k}_{it}\|/\Theta, \tag{25}$$

$$\mu = \mu(\delta) = \frac{1}{8}\left(\frac{\min(0.5,\delta)}{A}\right)^2. \tag{26}$$

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

$$(\boldsymbol{k}_{it} - \boldsymbol{k}_{i\tau})^\top \boldsymbol{p} \geq \delta > 0, \tag{27}$$

$$(\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{i\tau})^\top \boldsymbol{p} \geq 1 + \delta, \tag{28}$$

$$3/2 \geq (\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it})^\top \boldsymbol{p} \geq 1/2. \tag{29}$$

Above we used $\|\boldsymbol{p} - \boldsymbol{p}^{mm}\|^2/\|\boldsymbol{p}^{mm}\|^2 \leq 2\mu$ which implies $\|\boldsymbol{p} - \boldsymbol{p}^{mm}\| \leq \sqrt{2\mu}/\Theta$.

Proving Steps 1 and 2: No stationary point and $-\boldsymbol{q}^\top \nabla \mathcal{L}(\boldsymbol{p}) > 0$ within cone. Now that the choice of local cone is determined, we need to prove the main claims. We will lower bound $-\boldsymbol{q}^\top \nabla \mathcal{L}(\boldsymbol{p})$ and establish its strict positivity for $\|\boldsymbol{p}\| \geq R$. This will show that there is no stationary point as a by product. Given any $\boldsymbol{p} \in \mathsf{cone}_{\mu,R}(\boldsymbol{p}^{mm})$, denote $\bar{\boldsymbol{p}} = (\|\boldsymbol{p}^{mm}\|/\|\boldsymbol{p}\|)\boldsymbol{p}$ and recall $\|\boldsymbol{q}\| = \|\boldsymbol{p}^{mm}\|$. To proceed, we write the gradient correlation following (44) and (46)

$$\langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{q} \rangle = \frac{1}{n}\sum_{i=1}^n \ell_i' \cdot \langle \boldsymbol{a}_i, \mathbb{S}'(\boldsymbol{a}_i')\boldsymbol{\gamma}_i \rangle. \tag{30}$$

where we denoted $\ell_i' = \ell'(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}))$, $\boldsymbol{a}_i = \boldsymbol{K}_i\boldsymbol{q}$, $\boldsymbol{a}_i' = \boldsymbol{K}_i\boldsymbol{p}$, $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p})$. Using (27), for all $t \in \mathcal{T}_i, \tau \in \bar{\mathcal{T}}_i$, for all $\boldsymbol{p} \in \mathsf{cone}_{\mu,R}(\boldsymbol{p}^{mm})$, we have that

$$\boldsymbol{a}_{i\alpha_i}' - \boldsymbol{a}_{i\tau}' \geq R\Theta(1+\delta),\ \boldsymbol{a}_{it}' - \boldsymbol{a}_{i\tau}' \geq R\Theta\delta$$

Consequently, we can bound the softmax probabilities $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p})$ over non-neighbors as follows: For all $i \in [n]$ and any $t_i \in \mathcal{T}_i$

$$S_i := \sum_{\tau\in\bar{\mathcal{T}}_i} \boldsymbol{s}_{i\tau} \leq \sum_{\tau\neq\alpha_i} \boldsymbol{s}_{i\tau} \leq Te^{-R\Theta/2}\boldsymbol{s}_{i\alpha_i} \leq Te^{-R\Theta/2}, \tag{31}$$

$$Q_i := \sum_{\tau\in\bar{\mathcal{T}}_i} \boldsymbol{s}_{i\tau} \leq Te^{-R\Theta\delta}\boldsymbol{s}_{it_i} \leq Te^{-R\Theta\delta}S_i. \tag{32}$$

Recall scores $\boldsymbol{\gamma}_{it} = Y_i \cdot \boldsymbol{v}^\top \boldsymbol{x}_{it}$. Define the score gaps over neighbors: $\gamma_i^{gap} = \boldsymbol{\gamma}_{i\alpha_i} - \max_{t\in\mathcal{T}_i}\boldsymbol{\gamma}_{it}$, $\bar{\gamma}_i^{gap} = \boldsymbol{\gamma}_{i\alpha_i} - \min_{t\in\mathcal{T}_i}\boldsymbol{\gamma}_{it}$. Recall that $A := \max_{i\in[n],t\in[T]}\|\boldsymbol{k}_{it}\|/\Theta \geq \max_{i,t\in[T]}\|\boldsymbol{a}_{it}\| = \|\boldsymbol{k}_{it}\boldsymbol{q}\|$. Define the $\boldsymbol{\alpha}$-dependent global scalar $\Gamma = \sup_{i\in[n],t,\tau\in[T]}|\boldsymbol{\gamma}_{it} - \boldsymbol{\gamma}_{i\tau}|$.

Let us focus on a fixed datapoint $i \in [n]$, assume (without losing generality) $\alpha := \alpha_i = 1$, and drop subscripts i, that is, $\alpha := \alpha_i$, $\boldsymbol{X} := \boldsymbol{X}_i$, $Y := Y_i$, $\boldsymbol{K} := \boldsymbol{K}_i$, $\boldsymbol{a}' = \boldsymbol{K}\boldsymbol{p}$, $\boldsymbol{a} = \boldsymbol{K}\boldsymbol{q}$, $\boldsymbol{s} = \mathbb{S}(\boldsymbol{K}\boldsymbol{p})$, $\boldsymbol{\gamma} = Y \cdot \boldsymbol{X}\boldsymbol{v}$, $\gamma^{gap} := \gamma_i^{gap}$. Directly applying Lemma 2, we obtain

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

To proceed, let us decouple the non-neighbors within $\sum_{t\geq 2}^T (\boldsymbol{a}_1 - \boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)$ via

$$|\sum_{t\in\bar{\mathcal{T}}} (\boldsymbol{a}_1 - \boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)| \leq 2Q\Gamma A.$$

Aggregating these, we found

$$\left|\boldsymbol{a}^\top \mathrm{diag}(\boldsymbol{s})\boldsymbol{\gamma} - \boldsymbol{a}^\top \boldsymbol{s}\boldsymbol{s}^\top \boldsymbol{\gamma} - \sum_{t\in\mathcal{T}} (\boldsymbol{a}_1 - \boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)\right| \le 2\Gamma A((1-\boldsymbol{s}_1)^2 + Q). \tag{33}$$

To proceed, let us upper/lower bound the gradient correlation. Since $1.5 \ge \boldsymbol{a}_1 - \boldsymbol{a}_t \ge 0.5$, we find

$$1.5 \cdot S \cdot \bar{\gamma}^{gap} \sum_{t\in\mathcal{T}} (\boldsymbol{a}_1 - \boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \ge 0.5 \cdot S \cdot \gamma^{gap}.$$

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

$$S \cdot \gamma^{gap}/4 \ge 4\Gamma A \max((1-\boldsymbol{s}_1)^2, Q) \iff S \ge 16\frac{\Gamma A}{\gamma^{gap}} \max((1-\boldsymbol{s}_1)^2, Q). \tag{34}$$

Now choose $R \ge \delta^{-1}\log(T)/\Theta$ to ensure $Q \le S$ since $Q \le Te^{-R\Theta\delta}S$. Consequently

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

Combining these, what we wish is ensured by guaranteeing

$$S \ge 16\frac{\Gamma A}{\gamma^{gap}} \max(4STe^{-R\Theta/2}, Te^{-R\Theta\delta}S). \tag{35}$$

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

$$R = \frac{\max(2, \delta^{-1})}{\Theta} \log\left(\frac{64T\Gamma A}{\gamma^{gap}_{\min}}\right), \tag{36}$$

where $\gamma^{gap}_{\min} = \sup_{i\in[n]} \gamma^{gap}_i$ is the global scalar which is the worst case score gap over all inputs. With the above choice of R, we guaranteed

$$2(1-\boldsymbol{s}_1)\cdot\bar{\gamma}^{gap} \ge 2\cdot S\cdot\bar{\gamma}^{gap} \ge \sum_{t\in\mathcal{T}} (\boldsymbol{a}_1 - \boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \ge \frac{S\cdot\gamma^{gap}}{4} \ge \frac{(1-\boldsymbol{s}_1)\gamma^{gap}}{8}.$$

via (34) and (39). Since this holds over all inputs, going back to the gradient correlation (30) 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)

$$\frac{2}{n}\sum_{i\in[n]} \ell'_i \cdot q_i \cdot \bar{\gamma}^{gap}_i \ge \langle \nabla\mathcal{L}(\boldsymbol{p}), \boldsymbol{q}\rangle \ge \frac{1}{8n}\sum_{i\in[n]} \ell'_i \cdot q_i \cdot \gamma^{gap}_i. \tag{37}$$

Proving Step 3: Establishing gradient correlation. Our final goal is establishing gradient comparison between $\boldsymbol{p}, \boldsymbol{p}^{mm}$ for the same choice of $\mu > 0$ provided in (23). Define $\bar{\boldsymbol{p}} = \|\boldsymbol{p}^{mm}\|\boldsymbol{p}/\|\boldsymbol{p}\|$ to be the normalized vector. Set notations $\boldsymbol{a}_i = \boldsymbol{K}_i\bar{\boldsymbol{p}}$, $\bar{\boldsymbol{a}}_i = \boldsymbol{K}_i\boldsymbol{p}^{mm}$, and $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p})$. To establish the result, we will prove that, for sufficiently large $R = R_\pi$, for any $\boldsymbol{p} \in \mathsf{cone}_{\mu,R}(\boldsymbol{p}^{mm})$ and for any $i \in [n]$,

$$\langle \boldsymbol{a}_i, \mathbb{S}'(\boldsymbol{a}_i)\boldsymbol{\gamma}_i\rangle \le (1+\pi)\langle \bar{\boldsymbol{a}}_i, \mathbb{S}'(\boldsymbol{a}_i)\boldsymbol{\gamma}_i\rangle. \tag{38}$$

Once (38) holds for all i, the same conclusion will hold for the gradient correlations via (30). 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 (39), for all $\boldsymbol{q} \in \mathsf{cone}_\mu$ with $\|\boldsymbol{q}\| = \|\boldsymbol{p}^{mm}\|$ and $\boldsymbol{a}' = \boldsymbol{K}\boldsymbol{q}$, we have found

$$\left|\boldsymbol{a}'^\top \mathrm{diag}(\boldsymbol{s})\boldsymbol{\gamma} - \boldsymbol{a}'^\top \boldsymbol{s}\boldsymbol{s}^\top \boldsymbol{\gamma} - \sum_{t\in\mathcal{T}} (\boldsymbol{a}'_1 - \boldsymbol{a}'_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)\right| \le 2\Gamma A((1-\boldsymbol{s}_1)^2 + Q). \tag{39}$$

Plugging in $\boldsymbol{a}, \bar{\boldsymbol{a}}$ in the bound above and assuming $\pi \le 1$ (w.l.o.g.), (38) is implied by the following stronger inequality

$$6\Gamma A((1-\boldsymbol{s}_1)^2 + Q) + \sum_{t\in\mathcal{T}} (\boldsymbol{a}_1 - \boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \le (1+\pi)\sum_{t\in\mathcal{T}} (\bar{\boldsymbol{a}}_1 - \bar{\boldsymbol{a}}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) = (1+\pi)\sum_{t\in\mathcal{T}} \boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t)$$

First, we claim that $0.5\pi\sum_{t\in\mathcal{T}} \boldsymbol{s}_t(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_t) \ge 6\Gamma A((1-\boldsymbol{s}_1)^2 + Q)$. The proof of this claim directly follows the earlier argument, namely, following (34), (36) and (35) which leads to the choice $R_\pi = \frac{\max(2,\delta^{-1})}{\Theta}\log\left(\frac{C\cdot T\Gamma A}{\pi\gamma^{gap}_{\min}}\right)$ for some constant $C > 0$.

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

$$\sum_{t\in\mathcal{T}}(\boldsymbol{a}_1-\boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)\leq(1+0.5\pi)\sum_{t\in\mathcal{T}}\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t) \tag{40}$$

To proceed, we split the problem into two scenarios.

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

$$|\boldsymbol{a}_t-\bar{\boldsymbol{a}}_t|=|\boldsymbol{k}_t^\top(\bar{\boldsymbol{p}}-\boldsymbol{p}^{mm})|\leq A\Theta\varepsilon=\pi/4.$$

Consequently, we obtain

$$\boldsymbol{a}_1-\boldsymbol{a}_t\leq\bar{\boldsymbol{a}}_1-\bar{\boldsymbol{a}}_t+2A\Theta\varepsilon\leq 1+0.5\pi.$$

Similarly, $\boldsymbol{a}_1-\boldsymbol{a}_t\geq 1-0.5\pi\geq 0.5$. Since all terms $\boldsymbol{a}_1-\boldsymbol{a}_t,\boldsymbol{s}_t,\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t$ in (40) are nonnegative and $(\boldsymbol{a}_1-\boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)\leq(1+0.5\pi)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)$, above implies the desired result (40).

Scenario 2: $\|\bar{\boldsymbol{p}}-\boldsymbol{p}^{mm}\|\geq\varepsilon=\frac{\pi}{4A\Theta}$. Since $\bar{\boldsymbol{p}}$ is not (locally) max-margin, in this scenario, for some $\nu=\nu(\varepsilon)>0$ and $\tau\in\mathcal{T}$, we have that $\bar{\boldsymbol{p}}^\top(\boldsymbol{k}_1-\boldsymbol{k}_\tau)=\boldsymbol{a}_1-\boldsymbol{a}_\tau\leq 1-2\nu$. Here $\tau=\arg\max_{\tau\in\mathcal{T}}\bar{\boldsymbol{p}}^\top\boldsymbol{k}_\tau$ denotes the nearest point to $\boldsymbol{k}_1$ (along the $\bar{\boldsymbol{p}}$ direction). Note that a non-neighbor $\tau\in\bar{\mathcal{T}}$ cannot be nearest because $\boldsymbol{p}\in C_\mu$ and (27) holds. Recall that $\boldsymbol{s}=\mathbb{S}(\bar{R}\boldsymbol{a})$ where $\bar{R}=R\Theta$. To proceed, split the tokens into two groups: Let $\mathcal{N}$ be the group of tokens obeying $\boldsymbol{p}^\top(\boldsymbol{k}_1-\boldsymbol{k}_\tau)\leq 1-\nu$ and $\mathcal{T}-\mathcal{N}$ be the rest of the neighbors. 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 $|\boldsymbol{a}_1-\boldsymbol{a}_t|\leq 2A$ and recalling the definition of γ^{gap}, observe that

$$\sum_{t\in\mathcal{T}-\mathcal{N}}(\boldsymbol{a}_1-\boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)\leq\frac{2\Gamma ATe^{-\bar{R}\nu}}{\gamma^{gap}}\sum_{t\in\mathcal{N}}\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t).$$

Plugging this into (40), we obtain

$$\begin{aligned}\sum_{t\in\mathcal{T}}(\boldsymbol{a}_1-\boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)&=\sum_{t\in\mathcal{N}}(\boldsymbol{a}_1-\boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)+\sum_{t\in\mathcal{T}-\mathcal{N}}(\boldsymbol{a}_1-\boldsymbol{a}_t)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)\\&\leq\sum_{t\in\mathcal{N}}(1-\nu)\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t)+\sum_{t\in\mathcal{T}-\mathcal{N}}2A\Gamma Te^{-\bar{R}\nu}\\&\leq(1-\nu+\frac{2\Gamma ATe^{-\bar{R}\nu}}{\gamma^{gap}})\sum_{t\in\mathcal{T}}\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t) \qquad (41)\\&\leq(1+\frac{2\Gamma ATe^{-\bar{R}\nu}}{\gamma^{gap}})\sum_{t\in\mathcal{T}}\boldsymbol{s}_t(\boldsymbol{\gamma}_1-\boldsymbol{\gamma}_t). \qquad (42)\end{aligned}$$

(43)

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

$$R\geq R_\pi=\frac{1}{\nu\Theta}\log(\frac{4\Gamma A}{\gamma^{gap}_{\min}\pi}),$$

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

Combining this with the prior R_π choice (by taking maximum), we conclude with the statement. ■

B.2 Descent and Gradient Correlation Conditions

The lemma below identifies conditions under which $\boldsymbol{p}^{mm\star}$ is a global descent direction for $\mathcal{L}(\boldsymbol{p})$.

Lemma 4 (Global descent conditions) *Suppose $\ell(\cdot)$ is a strictly decreasing loss function and either of the following two conditions holds*

- ***Scores of non-optimal tokens are same:*** *For all* $i \in [n]$ *and* $t_1, t_2 \neq \texttt{opt}_i$, $\boldsymbol{v}^\top \boldsymbol{x}_{it_1} = \boldsymbol{v}^\top \boldsymbol{x}_{it_1}$.
- ***All tokens are support vectors:*** *Consider* (ATT-SVM) *with optimal indices* $(\texttt{opt}_i)_{i=1}^n$. $(\boldsymbol{k}_i^{\texttt{opt}} - \boldsymbol{k}_{it})^\top \boldsymbol{p}^{mm\star} = 1$ *for all* $t \neq \texttt{opt}_i, i \in [n]$.

Define

- $\boldsymbol{a}_{gap}^i := 1 = \inf_{t \neq \texttt{opt}_i} (\boldsymbol{k}_i^{\texttt{opt}} - \boldsymbol{k}_{it})^\top \boldsymbol{p}^{mm\star}$,
- $\boldsymbol{\gamma}_{gap}^i = \inf_{t \neq \texttt{opt}_i} Y_i \cdot (\boldsymbol{x}_i^{\texttt{opt}} - \boldsymbol{x}_{it})^\top \boldsymbol{v}$,
- $\texttt{lgt}_i' = p_{\texttt{opt}}(1 - p_{\texttt{opt}})$ *where* $p_{\texttt{opt}} = \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p})_{\texttt{opt}_i}$,
- $\ell_i' = \ell'(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p})) < 0$.

Then, for all $\boldsymbol{p} \in \mathbb{R}^d$, *the training loss* (ERM) *obeys*

$$-\left\langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{p}^{mm\star} \right\rangle \geq \min_{i \in [n]} \left\{ -\ell_i' \cdot \texttt{lgt}_i' \cdot \boldsymbol{a}_{gap}^i \cdot \boldsymbol{\gamma}_{gap}^i \right\} > 0.$$

Proof. Set $\bar{\boldsymbol{a}}_i = \boldsymbol{K}_i \boldsymbol{p}^{mm\star}$ to obtain In order to show this result, let us recall the gradient evaluated at $\boldsymbol{p}$ which is given by

$$\nabla \mathcal{L}(\boldsymbol{p}) = \frac{1}{n} \sum_{i=1}^n \ell_i' \cdot \boldsymbol{K}_i^\top \mathbb{S}'(\boldsymbol{a}_i) \boldsymbol{\gamma}_i. \tag{44}$$

Here $\boldsymbol{\gamma}_i = Y_i \cdot X_i \boldsymbol{v}$, $\boldsymbol{a}_i = \boldsymbol{K}_i \boldsymbol{p}$, and $\ell_i' = \ell'(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p}))$. This implies that

$$\left\langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{p}^{mm\star} \right\rangle = \frac{1}{n} \sum_{i=1}^n \ell_i' \cdot \langle \bar{\boldsymbol{a}}_i, \mathbb{S}'(\boldsymbol{a}_i) \boldsymbol{\gamma}_i \rangle .$$

To proceed, we will prove that individual summands are all strictly negative. To show that, without losing generality, let us focus on the first input and drop the subscript i for cleaner notation. This yields

$$\langle \bar{\boldsymbol{a}}, \mathbb{S}'(\boldsymbol{a}) \boldsymbol{\gamma} \rangle = \bar{\boldsymbol{a}}^\top \mathrm{diag}(\mathbb{S}(\boldsymbol{a})) \boldsymbol{\gamma} - \bar{\boldsymbol{a}}^\top \mathbb{S}(\boldsymbol{a}) \mathbb{S}(\boldsymbol{a})^\top \boldsymbol{\gamma}. \tag{45}$$

Without losing generality, assume optimal token is the first one. The lemma has two scenarios. In the first scenario (same non-optimal scores), $\boldsymbol{\gamma}_t$ is a constant for all $t \geq 2$. In the second scenario (all tokens are support), $\bar{\boldsymbol{a}}_t = \boldsymbol{k}_t \boldsymbol{p}^{mm\star}$ is constant for all $t \geq 2$. Since $\bar{\boldsymbol{a}}, \boldsymbol{\gamma}$ vectors are represented symmetrically in the gradient correlation, verifying these two conditions are equivalent.

To proceed, we will prove the following (focusing on the first condition): Suppose $\gamma = \boldsymbol{\gamma}_{t \geq 2}$ is constant, $\boldsymbol{\gamma}_1, \bar{\boldsymbol{a}}_1$ are the largest indices of $\boldsymbol{\gamma}, \bar{\boldsymbol{a}}$. Then, for any $\boldsymbol{s}$ obeying $\sum_{t \in [T]} \boldsymbol{s}_t = 1, \boldsymbol{s}_t \geq 0$, we have that $\bar{\boldsymbol{a}}^\top \mathrm{diag}(\boldsymbol{s}) \boldsymbol{\gamma} - \bar{\boldsymbol{a}}^\top \boldsymbol{s} \boldsymbol{s}^\top \boldsymbol{\gamma} > 0$. To see this, we write

$$\begin{aligned}
\bar{\boldsymbol{a}}^\top \mathrm{diag}(\boldsymbol{s}) \boldsymbol{\gamma} - \bar{\boldsymbol{a}}^\top \boldsymbol{s} \boldsymbol{s}^\top \boldsymbol{\gamma} &= \sum_{t=1}^T \bar{\boldsymbol{a}}_t \boldsymbol{\gamma}_t \boldsymbol{s}_t - \sum_{t=1}^T \bar{\boldsymbol{a}}_t \boldsymbol{s}_t \sum_{t=1}^T \boldsymbol{\gamma}_t \boldsymbol{s}_t && (46) \\
&= (\bar{\boldsymbol{a}}_1 \boldsymbol{\gamma}_1 \boldsymbol{s}_1 + \gamma \sum_{t \geq 2} \bar{\boldsymbol{a}}_t \boldsymbol{s}_t) - (\boldsymbol{\gamma}_1 \boldsymbol{s}_1 + \gamma(1 - \boldsymbol{s}_1))(\bar{\boldsymbol{a}}_1 \boldsymbol{s}_1 + \sum_{t \geq 2}^T \bar{\boldsymbol{a}}_t \boldsymbol{s}_t) && (47) \\
&= \bar{\boldsymbol{a}}_1 (\boldsymbol{\gamma}_1 - \gamma) \boldsymbol{s}_1 (1 - \boldsymbol{s}_1) + (\gamma - (\boldsymbol{\gamma}_1 \boldsymbol{s}_1 + \gamma(1 - \boldsymbol{s}_1))) \sum_{t \geq 2}^T \bar{\boldsymbol{a}}_t \boldsymbol{s}_t && (48) \\
&= \bar{\boldsymbol{a}}_1 (\boldsymbol{\gamma}_1 - \gamma) \boldsymbol{s}_1 (1 - \boldsymbol{s}_1) - (\boldsymbol{\gamma}_1 - \gamma) \boldsymbol{s}_1 \sum_{t \geq 2}^T \bar{\boldsymbol{a}}_t \boldsymbol{s}_t && (49) \\
&= (\boldsymbol{\gamma}_1 - \gamma)(1 - \boldsymbol{s}_1) \boldsymbol{s}_1 [\bar{\boldsymbol{a}}_1 - \frac{\sum_{t \geq 2}^T \bar{\boldsymbol{a}}_t \boldsymbol{s}_t}{\sum_{t \geq 2} \boldsymbol{s}_t}]. && (50)
\end{aligned}$$

To proceed, recall the definitions $\boldsymbol{\gamma}_{gap} = \boldsymbol{\gamma}_1 - \max_{t \geq 2} \boldsymbol{\gamma}_t$ and $\boldsymbol{a}_{gap} = \bar{\boldsymbol{a}}_1 - \max_{t \geq 2} \boldsymbol{a}_t$. With these, we obtain

$$\bar{\boldsymbol{a}}^\top \mathrm{diag}(\boldsymbol{s}) \boldsymbol{\gamma} - \bar{\boldsymbol{a}}^\top \boldsymbol{s} \boldsymbol{s}^\top \boldsymbol{\gamma} \geq \boldsymbol{a}_{gap} \boldsymbol{\gamma}_{gap} \boldsymbol{s}_1 (1 - \boldsymbol{s}_1),$$

which is the advertised result after noticing $\boldsymbol{s}_1(1 - \boldsymbol{s}_1)$ is the logistic derivative and infimum'ing over all inputs and multiplying by ℓ_i'. ■

Lemma 5 (Gradient correlation conditions) *Fix indices $\boldsymbol{\alpha} = (\alpha_{i=1}^n)$ and let $\boldsymbol{p}^{mm} = \boldsymbol{p}^{mm}(\boldsymbol{\alpha})$ be the SVM solution separating α_i from remaining tokens of input $\boldsymbol{X}_i$ for $i \in [n]$. Suppose for all $i \in [n]$ and $t_1, t_2 \neq \alpha_i$, $\boldsymbol{v}^\top \boldsymbol{x}_{it_1} = \boldsymbol{v}^\top \boldsymbol{x}_{it_2} < \boldsymbol{v}^\top \boldsymbol{x}_{i\alpha_i}$ and $\ell(\cdot)$ is strictly decreasing. Let $\bar{\boldsymbol{p}} = \|\boldsymbol{p}^{mm}\|\boldsymbol{p}/\|\boldsymbol{p}\|$. $M = \sup_{i,t} \|\boldsymbol{k}_i\|$ and $\Xi = 1/\|\boldsymbol{p}^{mm}\|$. For any choice of $\pi > 0$, there exists $R := R_\pi$ such that, for any $\boldsymbol{p}$ with $\|\boldsymbol{p}\| \geq R$, we have*

$$\langle \nabla \mathcal{L}(\boldsymbol{p}), \bar{\boldsymbol{p}} \rangle \geq (1+\pi) \langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{p}^{mm} \rangle .$$

Above, observe that as $R \to \infty$, we eventually get to set $\pi = 0$.

Proof. The proof is similar to Lemma 4 at a high-level. However, we also need to account for the impact of $\boldsymbol{p}$ besides $\boldsymbol{p}^{mm}$ in the gradient correlation. The main goal is showing that $\boldsymbol{p}^{mm}$ is the near-optimal descent direction, thus, $\boldsymbol{p}$ cannot significantly outperform it.

To proceed, set $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p})$, $\boldsymbol{a}_i = \boldsymbol{K}_i \bar{\boldsymbol{p}}$, $\bar{\boldsymbol{a}}_i = \boldsymbol{K}_i \boldsymbol{p}^{mm}$. Without losing generality assume $\alpha_i = 1$ for all $i \in [n]$. Set $\texttt{lgt}'_i = \boldsymbol{s}_{i1}/(1 - \boldsymbol{s}_{i1})$. Repeating the proof of Lemma 4 yields

$$\langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{p}^{mm} \rangle = \frac{1}{n} \sum_{i=1}^n \ell'_i \cdot \texttt{lgt}'_i \cdot (\boldsymbol{\gamma}_{i1} - \gamma_i) \left[\bar{\boldsymbol{a}}_{i1} - \frac{\sum_{t\geq 2}^T \bar{\boldsymbol{a}}_{it} \boldsymbol{s}_{it}}{\sum_{t\geq 2} \boldsymbol{s}_{it}} \right] \tag{51}$$

$$\langle \nabla \mathcal{L}(\boldsymbol{p}), \bar{\boldsymbol{p}} \rangle = \frac{1}{n} \sum_{i=1}^n \ell'_i \cdot \texttt{lgt}'_i \cdot (\boldsymbol{\gamma}_{i1} - \gamma_i) \left[\boldsymbol{a}_{i1} - \frac{\sum_{t\geq 2}^T \boldsymbol{a}_{it} \boldsymbol{s}_{it}}{\sum_{t\geq 2} \boldsymbol{s}_{it}} \right] \tag{52}$$

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

$$\left[\boldsymbol{a}_1 - \frac{\sum_{t\geq 2}^T \boldsymbol{a}_t \boldsymbol{s}_t}{\sum_{t\geq 2} \boldsymbol{s}_t} \right] \leq (1+\pi) \cdot \left[\bar{\boldsymbol{a}}_1 - \frac{\sum_{t\geq 2}^T \bar{\boldsymbol{a}}_t \boldsymbol{s}_t}{\sum_{t\geq 2} \boldsymbol{s}_t} \right]. \tag{53}$$

We consider two scenarios. Let $M = \max_{i\in[n], t\in[T]} \|\boldsymbol{k}_{it}\|$.

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

$$|\boldsymbol{a}_t - \bar{\boldsymbol{a}}_t| = |\boldsymbol{k}_t^\top (\bar{\boldsymbol{p}} - \boldsymbol{p}^{mm})| \leq M \|\bar{\boldsymbol{p}} - \boldsymbol{p}^{mm}\| \leq M\varepsilon.$$

Consequently, we obtain

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

Also noticing $\bar{\boldsymbol{a}}_1 - \frac{\sum_{t\geq 2}^T \bar{\boldsymbol{a}}_t \boldsymbol{s}_t}{\sum_{t\geq 2} \boldsymbol{s}_t} \geq 1$ (thanks to $\boldsymbol{p}^{mm}$ satisfying ≥ 1 margin), this implies (53).

Scenario 2: $\|\bar{\boldsymbol{p}} - \boldsymbol{p}^{mm}\| \geq \varepsilon := \pi/2M$. In this scenario, for some $\delta = \delta(\varepsilon)$ and $\tau \geq 2$, we have that $\boldsymbol{p}^\top(\boldsymbol{k}_1 - \boldsymbol{k}_\tau) = \boldsymbol{a}_1 - \boldsymbol{a}_\tau \leq 1 - 2\delta$. Here $\tau = \arg\max_{t\geq 2} \boldsymbol{p}^\top \boldsymbol{k}_t$ denotes the nearest point to $\boldsymbol{k}_1$. Recall that $\boldsymbol{s} = \mathbb{S}(\bar{R}\boldsymbol{a})$ where $\bar{R} = R\Xi = R/\|\boldsymbol{p}^{mm}\|$. To proceed, split the tokens into two groups: Let $\mathcal{N}$ be the group of tokens obeying $\boldsymbol{p}^\top(\boldsymbol{k}_1 - \boldsymbol{k}_\tau) \geq 1 - \delta$ and $[T] - \mathcal{N}$ be the rest. Observe that

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

Set $\bar{M} = M/\Xi$ and note that $\|\boldsymbol{a}_t\| \leq \|\boldsymbol{p}^{mm}\| \cdot \|\boldsymbol{k}_t\| \leq \bar{M}$. Using $\boldsymbol{p}^\top(\boldsymbol{k}_1 - \boldsymbol{k}_\tau) < 1 - \delta$ over $\tau \in [T] - \mathcal{N}$ and plugging in the above bound, we obtain

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

Using the fact that $\bar{\boldsymbol{a}}_1 - \frac{\sum_{t\geq 2}^T \bar{\boldsymbol{a}}_t \boldsymbol{s}_t}{\sum_{t\geq 2} \boldsymbol{s}_t} \geq 1$, the above implies (53) with $\pi' = (1-\delta) + 2\bar{M}Te^{-\bar{R}\delta}$. To proceed, choose $R_\pi = \delta^{-1}\Xi^{-1} \log(2\bar{M}T/\pi)$ to ensure $\pi' \leq \pi$. ∎

The following lemma states the descent property of gradient descent for $\mathcal{L}(\boldsymbol{p})$ under Assumption A. It is important to note that although the infimum of the optimization problem is $\mathcal{L}^*$, it is not achieved at any finite $\boldsymbol{p}$. Additionally, there are no finite critical points $\boldsymbol{p}$.

Lemma 6 *Under Assumption A, the objective $\mathcal{L}(\boldsymbol{p})$ is L_p-smooth, where*

$$L_p := \frac{1}{n}\sum_{i=1}^{n}\left(M_0\|\mathbf{v}\|^2\|\mathbf{W}\|^2 + M_1\|\mathbf{v}\|\|\mathbf{W}\|^3\right)\|\mathbf{X}_i\|^4. \tag{54}$$

Further, if $\eta \le 2/L_p$, then, for any initialization $\boldsymbol{p}(0)$, with the GD sequence $\boldsymbol{p}(t+1) = \boldsymbol{p}(t) - \eta\nabla\mathcal{L}(\boldsymbol{p}(t))$, we have

$$\mathcal{L}(\boldsymbol{p}(t+1)) - \mathcal{L}(\boldsymbol{p}(t)) \le -\frac{\eta}{2}\|\nabla\mathcal{L}(\boldsymbol{p}(t))\|^2, \tag{55}$$

for all $t \ge 0$, $\sum_{t=0}^{\infty}\|\nabla\mathcal{L}(\boldsymbol{p}(t))\|^2 < \infty$ and $\lim_{t\to\infty}\|\nabla\mathcal{L}(\boldsymbol{p}(t))\|^2 = 0$.

Proof. Recall that we defined $\boldsymbol{\gamma}_i = Y_i \cdot \mathbf{X}_i\boldsymbol{v}$, $\boldsymbol{a}_i = \boldsymbol{K}_i\boldsymbol{p}$, and $\ell_i' = \ell'(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}))$. The gradient evaluated at $\boldsymbol{p}$ is given by

$$\nabla\mathcal{L}(\boldsymbol{p}) = \frac{1}{n}\sum_{i=1}^{n}\ell_i' \cdot \boldsymbol{K}_i^\top \mathbb{S}'(\boldsymbol{a}_i)\boldsymbol{\gamma}_i.$$

Now, for any $\boldsymbol{p}, \dot{\boldsymbol{p}} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\|\nabla\mathcal{L}(\boldsymbol{p}) - \nabla\mathcal{L}(\dot{\boldsymbol{p}})\| &\le \frac{1}{n}\sum_{i=1}^{n}\left\|\ell'(\boldsymbol{\gamma}_i^\top\mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}))\cdot\boldsymbol{K}_i^\top\mathbb{S}'(\boldsymbol{K}_i\boldsymbol{p})\boldsymbol{\gamma}_i - \ell'(\boldsymbol{\gamma}_i^\top\mathbb{S}(\boldsymbol{K}_i\dot{\boldsymbol{p}}))\cdot\boldsymbol{K}_i^\top\mathbb{S}'(\boldsymbol{K}_i\dot{\boldsymbol{p}})\boldsymbol{\gamma}_i\right\| \\
&\le \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{K}_i^\top\mathbb{S}'(\boldsymbol{K}_i\dot{\boldsymbol{p}})\boldsymbol{\gamma}_i\|\left\|\ell'(\boldsymbol{\gamma}_i^\top\mathbb{S}(\boldsymbol{K}_i\boldsymbol{p})) - \ell'(\boldsymbol{\gamma}_i^\top\mathbb{S}(\boldsymbol{K}_i\dot{\boldsymbol{p}}))\right\| \\
&+ \frac{1}{n}\sum_{i=1}^{n}\|\ell'(\boldsymbol{\gamma}_i^\top\mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}))\|\left\|\boldsymbol{K}_i^\top\mathbb{S}'(\boldsymbol{K}_i\boldsymbol{p})\boldsymbol{\gamma}_i - \boldsymbol{K}_i^\top\mathbb{S}'(\boldsymbol{K}_i\dot{\boldsymbol{p}})\boldsymbol{\gamma}_i\right\| \\
&\le \frac{1}{n}\sum_{i=1}^{n}M_0\|\boldsymbol{\gamma}_i\|^2\|\boldsymbol{K}_i\|\,\|\mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}) - \mathbb{S}(\boldsymbol{K}_i\dot{\boldsymbol{p}})\| + M_1\|\boldsymbol{K}_i\|\|\boldsymbol{\gamma}_i\|\left\|\mathbb{S}'(\boldsymbol{K}_i\boldsymbol{p}) - \mathbb{S}'(\boldsymbol{K}_i\dot{\boldsymbol{p}})\right\|,
\end{aligned}$$

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

Note also that for any $\boldsymbol{p}, \dot{\boldsymbol{p}} \in \mathbb{R}^d$,

$$\|\mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}) - \mathbb{S}(\boldsymbol{K}_i\dot{\boldsymbol{p}})\| \le \|\boldsymbol{K}_i\|\|\boldsymbol{p} - \dot{\boldsymbol{p}}\| \quad\text{and}\quad \left\|\mathbb{S}'(\boldsymbol{K}_i\boldsymbol{p}) - \mathbb{S}'(\boldsymbol{K}_i\boldsymbol{p})\right\| \le \|\boldsymbol{K}_i\|^2\|\boldsymbol{p} - \dot{\boldsymbol{p}}\|.$$

Hence,

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

where L_p is defined in (54).

The reminder of proof is similar to the proof of [19, Lemma 10]. Since $\mathcal{L}(\boldsymbol{p})$ is L_p-smooth, we get

$$\begin{aligned}
\mathcal{L}(\boldsymbol{p}(t+1)) &\le \mathcal{L}(\boldsymbol{p}(t)) + \nabla\mathcal{L}(\boldsymbol{p}(t))^\top(\boldsymbol{p}(t+1) - \boldsymbol{p}(t)) + \frac{L_p}{2}\|\boldsymbol{p}(t+1) - \boldsymbol{p}(t)\|^2 \\
&= \mathcal{L}(\boldsymbol{p}(t)) - \eta\|\nabla\mathcal{L}(\boldsymbol{p}(t))\|^2 + \frac{L_p\eta^2}{2}\|\nabla\mathcal{L}(\boldsymbol{p}(t))\|^2 \\
&= \mathcal{L}(\boldsymbol{p}(t)) - \eta\left(1 - \frac{L_p\eta}{2}\right)\|\nabla\mathcal{L}(\boldsymbol{p}(t))\|^2 \\
&= \mathcal{L}(\boldsymbol{p}(t)) - \frac{\eta}{2}\|\nabla\mathcal{L}(\boldsymbol{p}(t))\|^2,
\end{aligned}$$

where the last inequality follows from our assumption on the stepsize.

The above inequality implies that

$$\sum_{t=0}^{\infty} \|\nabla \mathcal{L}(\boldsymbol{p}(t))\|^2 \leq \frac{2}{\eta}\left(\mathcal{L}(\boldsymbol{p}(0)) - \mathcal{L}^*\right).$$

Here, the right hand side is upper bounded by a finite constant, since by Assumption A, $\mathcal{L}(\boldsymbol{p}(0)) < \infty$ and $\mathcal{L}^* \leq \mathcal{L}(\boldsymbol{p}(t))$. This implies $\sum_{t=0}^{\infty} \|\nabla \mathcal{L}(\boldsymbol{p}(t))\|^2 < \infty$ and therefore $\|\nabla \mathcal{L}(\boldsymbol{p}(t))\|^2 \to 0$. ■

B.3 Proof of Theorem 1

Proof. We first show that $\lim_{t\to\infty} \|\boldsymbol{p}(t)\| = \infty$. From Lemma 4, we have

$$\left\langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{p}^{mm\star} \right\rangle = \frac{1}{n} \sum_{i=1}^{n} \ell_i' \cdot \left\langle \boldsymbol{K}_i \boldsymbol{p}^{mm\star}, \mathbb{S}'(\boldsymbol{a}_i)\boldsymbol{\gamma}_i \right\rangle,$$

where $\boldsymbol{\gamma}_i = Y_i \cdot \mathbf{X}_i \boldsymbol{v}$, $\boldsymbol{a}_i = \boldsymbol{K}_i \boldsymbol{p}$, and $\ell_i' = \ell'(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p}))$.

It follows from Lemma 4 that $\langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{p}^{mm\star} \rangle < 0$ for all $\boldsymbol{p} \in \mathbb{R}^d$. Hence, for any finite $\mathbf{p}$, $\langle \nabla \mathcal{L}(\boldsymbol{p}), \boldsymbol{p}^{mm\star} \rangle$ cannot be equal to zero, as a sum of negative terms. Therefore, there are no finite critical points $\mathbf{p}$, for which $\nabla \mathcal{L}(\boldsymbol{p}) = 0$ which contradicts Lemma 6. This implies that $\|\boldsymbol{p}(t)\| \to \infty$.

Now, given any $\epsilon \in (0, 1)$, let $\pi = \epsilon/(1-\epsilon)$. Since $\lim_{t\to\infty} \|\boldsymbol{p}(t)\| = \infty$, we can choose t_0 such that for any $t \geq t_0$, it holds that $\|\boldsymbol{p}(t)\| > R_\epsilon \vee 1/2$ for some radius R_ϵ. Now for any $t \geq t_0$, it follows from Lemma 5 that

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

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

$$\begin{aligned}
\left\langle \boldsymbol{p}(t+1) - \boldsymbol{p}(t), \frac{\boldsymbol{p}^{mm\star}}{\|\boldsymbol{p}^{mm\star}\|} \right\rangle &\geq (1-\epsilon) \left\langle \boldsymbol{p}(t+1) - \boldsymbol{p}(t), \frac{\boldsymbol{p}(t)}{\|\boldsymbol{p}(t)\|} \right\rangle \\
&= \frac{(1-\epsilon)}{2\|\boldsymbol{p}(t)\|} \left(\|\boldsymbol{p}(t+1)\|^2 - \|\boldsymbol{p}(t)\|^2 - \|\boldsymbol{p}(t+1) - \boldsymbol{p}(t)\|^2 \right) \\
&\geq (1-\epsilon) \left(\|\boldsymbol{p}(t+1)\|^2 - \|\boldsymbol{p}(t)\|^2 - \|\boldsymbol{p}(t+1) - \boldsymbol{p}(t)\|^2 \right) \\
&\geq (1-\epsilon) \left(\|\boldsymbol{p}(t+1)\| - \|\boldsymbol{p}(t)\| - \|\boldsymbol{p}(t+1) - \boldsymbol{p}(t)\|^2 \right) \\
&\geq (1-\epsilon) \Big(\|\boldsymbol{p}(t+1)\| - \|\boldsymbol{p}(t)\| - 2\eta \left(\mathcal{L}(\boldsymbol{p}(t)) - \mathcal{L}(\boldsymbol{p}(t+1)) \right) \Big).
\end{aligned} \tag{56}$$

Here, the last inequality uses Lemma 6.

Summing the above inequality over $t \geq t_0$ gives

$$\left\langle \frac{\boldsymbol{p}(t)}{\|\boldsymbol{p}(t)\|}, \frac{\boldsymbol{p}^{mm\star}}{\|\boldsymbol{p}^{mm\star}\|} \right\rangle \geq 1 - \epsilon + \frac{C(\epsilon, \eta)}{\|\boldsymbol{p}(t)\|},$$

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

$$C(\epsilon, \eta) := \left\langle \boldsymbol{p}(t_0), \frac{\boldsymbol{p}^{mm\star}}{\|\boldsymbol{p}^{mm\star}\|} \right\rangle - (1-\epsilon)\|\boldsymbol{p}(t_0)\| - 2\eta(1-\epsilon)(\mathcal{L}(\boldsymbol{p}(t_0)) - \mathcal{L}^*), \tag{57}$$

where $\mathcal{L}^* \leq \mathcal{L}(\boldsymbol{p}(t))$ for all $t \geq 0$.

Since $\|\boldsymbol{p}(t)\| \to \infty$, we get

$$\liminf_{t\to\infty} \left\langle \frac{\boldsymbol{p}(t)}{\|\boldsymbol{p}(t)\|}, \frac{\boldsymbol{p}^{mm\star}}{\|\boldsymbol{p}^{mm\star}\|} \right\rangle \geq 1 - \epsilon.$$

Given that ϵ is arbitrary, we can consider the limit as ϵ approaches zero. Thus, we have: $\boldsymbol{p}(t)/\|\boldsymbol{p}(t)\| \to \boldsymbol{p}^{mm\star}/\|\boldsymbol{p}^{mm\star}\|$. ■

B.4 Proof of Theorem 2

This proof is a direct corollary of Lemma 9 which itself is a special case of the nonlinear head Theorem 7. Let us verify that $f(\boldsymbol{X}) = \boldsymbol{v}^\top \boldsymbol{X}^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{p})$ satisfies the assumptions of Lemma 9 where we replace the nonlinear head with linear $\boldsymbol{v}$. To see this, set the optimal sets to be the singletons $O_i = \{\texttt{opt}_i\}$. given $(\boldsymbol{X}_i, Y_i)$ and defining $\boldsymbol{s}_i = \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p})$ and $q_i := q_i^{\boldsymbol{p}} = \sum_{t\neq \texttt{opt}_i} \boldsymbol{s}_{it}$. Recalling score definition $\boldsymbol{\gamma}_i = Y_i \cdot \boldsymbol{X}_i\boldsymbol{v}$ and setting $\nu_i := \boldsymbol{\gamma}_{i\texttt{opt}_i}$ and $Z_i := \sum_{t\neq \texttt{opt}_i} \boldsymbol{\gamma}_{it}\boldsymbol{s}_{it}$, a particular prediction can be written as

$$Y_i \cdot \boldsymbol{v}^\top \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{X}_i\boldsymbol{p}) = \boldsymbol{\gamma}_i^\top \boldsymbol{s}_i = \boldsymbol{\gamma}_{i\texttt{opt}_i}(1-q_i) + \sum_{t\neq \texttt{opt}_i} \boldsymbol{\gamma}_{it}\boldsymbol{s}_{it} \tag{58}$$

$$= \nu_i(1-q_i) + Z_i. \tag{59}$$

To proceed, we demonstrate the choices for $C, \varepsilon > 0$. Let $C := -\min_{i\in[n],t\in[T]} \boldsymbol{\gamma}_{it} \wedge 0$ and $q_{\max} = \max_{i\in[n]} q_i$. Note that $Z_i \geq \sum_{t\neq \texttt{opt}_i} \boldsymbol{\gamma}_{it}\boldsymbol{s}_{it} \geq q_i\boldsymbol{\gamma}_{\min} \geq -Cq_{\max}$. Now, using strict score optimality of $\texttt{opt}_i$'s for all $i \in [n]$, we set

$$\varepsilon := 1 - \sup_{i\in[n]} \frac{\sum_{t\neq \texttt{opt}_i} \boldsymbol{\gamma}_{it}\boldsymbol{s}_{it}}{\nu_i q_i} \geq 1 - \sup_{i\in[n]} \frac{\sup_{t\neq \texttt{opt}_i} \boldsymbol{\gamma}_{it}}{\boldsymbol{\gamma}_{i\texttt{opt}_i}} > 0.$$

We conclude by observing $Z_i \leq \nu_i q_i \frac{\sum_{t\neq \texttt{opt}_i} \boldsymbol{\gamma}_{it}\boldsymbol{s}_{it}}{\nu_i q_i} \leq \nu_i q_i \varepsilon$ as desired.

B.5 Proof of Theorem 3

Proof. We provide the proof in four steps:

Step 1: There are no stationary points within the cone. We begin by proving that there are no stationary points within $\texttt{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \left\{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R_\mu\right\}$ for a specific radius R_μ. Let $(\mathcal{T}_i)_{i=1}^n$ denote the set of SVM-neighbors as defined in Definition 2. We define $\bar{\mathcal{T}}_i = [T] - \mathcal{T}_i - \alpha_i$ as the tokens that are non-SVM neighbors. Additionally, let μ be defined as in (23). For all $\boldsymbol{q}, \boldsymbol{p} \in \texttt{cone}_\mu(\boldsymbol{p}^{mm})$ with $\|\boldsymbol{q}\| = \|\boldsymbol{p}^{mm}\|$, it follows from Lemma 3 that there exists R_μ such that $-\boldsymbol{q}^\top \nabla \mathcal{L}(\boldsymbol{p})$ is strictly positive for $\|\boldsymbol{p}\| \geq R_\mu$.

Step 2: Let $\epsilon \in (0, \min(\mu, 1))$, $1/(1+\pi) = 1-\epsilon$. It follows from Lemma 5 that, there exists R_ϵ such that all $\boldsymbol{p} \in \texttt{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R_\epsilon\}$ satisfy

$$\left\langle -\nabla \mathcal{L}(\boldsymbol{p}), \frac{\boldsymbol{p}^{mm}(\alpha)}{\|\boldsymbol{p}^{mm}(\alpha)\|} \right\rangle \geq (1-\epsilon) \left\langle -\nabla \mathcal{L}(\boldsymbol{p}), \frac{\boldsymbol{p}}{\|\boldsymbol{p}\|} \right\rangle. \tag{60}$$

Step 3: Updates remain inside the cone. By leveraging the results from **Step 1** and **Step 2**, we show that that the gradient iterates, with an appropriate step size, starting from $p(0) \in \texttt{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R\}$, remain within this cone.

We proceed by induction. Suppose that the claim holds up to iteration $t \geq 0$. This implies that $\boldsymbol{p}(t) \in \texttt{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R\}$. Hence, there exists scalar $\mu = \mu(\alpha) \in (0, 1]$ and R_μ such that $\texttt{corr}(\boldsymbol{p}(t), \boldsymbol{p}^{mm}(\alpha)) \geq 1-\mu$ and $\|\boldsymbol{p}(t)\| \geq R_\mu$. Let $\rho := -(1/(1-\epsilon)) \left\langle \nabla \mathcal{L}(\boldsymbol{p}(t)), \frac{\boldsymbol{p}^{mm}(\alpha)}{\|\boldsymbol{p}^{mm}(\alpha)\|} \right\rangle > 0$. We have

$$\begin{aligned}
\left\langle \frac{\boldsymbol{p}(t+1)}{\|\boldsymbol{p}(t)\|}, \frac{\boldsymbol{p}^{mm}(\alpha)}{\|\boldsymbol{p}^{mm}(\alpha)\|} \right\rangle &= \left\langle \frac{\boldsymbol{p}(t)}{\|\boldsymbol{p}(t)\|} - \frac{\eta}{\|\boldsymbol{p}(t)\|} \nabla \mathcal{L}(\boldsymbol{p}(t)), \frac{\boldsymbol{p}^{mm}(\alpha)}{\|\boldsymbol{p}^{mm}(\alpha)\|} \right\rangle \\
&\geq 1 - \mu - \frac{\eta}{\|\boldsymbol{p}(t)\|} \left\langle \nabla \mathcal{L}(\boldsymbol{p}(t)), \frac{\boldsymbol{p}^{mm}(\alpha)}{\|\boldsymbol{p}^{mm}(\alpha)\|} \right\rangle \qquad (61a) \\
&\geq 1 - \mu + \frac{\eta\rho(1-\epsilon)}{\|\boldsymbol{p}(t)\|}.
\end{aligned}$$

Note that from Lemma 3, we have $\langle \nabla f(\boldsymbol{p}(t)), \boldsymbol{p}(t) \rangle < 0$ which implies that $\|\boldsymbol{p}(t+1)\| \leq \|\boldsymbol{p}(t)\| - \frac{\eta}{\|\boldsymbol{p}(t)\|} \langle \nabla f(\boldsymbol{p}(t)), \boldsymbol{p}(t) \rangle + \eta^2 \|\nabla f(\boldsymbol{p}(t))\|^2$. Hence, $\|\boldsymbol{p}(t+1)\| \geq \|\boldsymbol{p}(t)\|$, and

$$\begin{aligned}
\frac{\boldsymbol{p}(t+1)}{\|\boldsymbol{p}(t)\|} &\leq 1 - \eta \left\langle \nabla f(\boldsymbol{p}(t)), \frac{\boldsymbol{p}(t)}{\|\boldsymbol{p}(t)\|} \right\rangle + \eta^2 \frac{\|\nabla \mathcal{L}(\boldsymbol{p}(t))\|^2}{\|\boldsymbol{p}(t)\|} \\
&\leq 1 - \frac{\eta}{1-\epsilon} \left\langle \nabla \mathcal{L}(\boldsymbol{p}(t)), \frac{\boldsymbol{p}^{mm}(\alpha)}{\|\boldsymbol{p}^{mm}(\alpha)\|} \right\rangle + \eta^2 \frac{\|\nabla \mathcal{L}(\boldsymbol{p}(t))\|^2}{\|\boldsymbol{p}(t)\|} \qquad (61b) \\
&\leq 1 + \frac{\eta\rho}{\|\mathbf{p}(t)\|} + \frac{\eta^2 \|\nabla \mathcal{L}(\boldsymbol{p}(t))\|^2}{\|\boldsymbol{p}(t)\|} =: C(\eta, \rho).
\end{aligned}$$

Here, the second inequality follows from (60).

Now, it follows from (61a) and (61b) that

$$\begin{aligned}\left\langle \frac{\boldsymbol{p}(t+1)}{\|\boldsymbol{p}(t+1)\|}, \frac{\boldsymbol{p}^{mm}(\alpha)}{\|\boldsymbol{p}^{mm}(\alpha)\|} \right\rangle &\geq \frac{1}{C(\eta,\rho)}\left(1-\mu+\frac{\eta\rho(1-\epsilon)}{\|\boldsymbol{p}(t)\|}\right) \\ &\geq \frac{1}{C(\eta,\rho)}\left(1-\mu+\frac{\eta\rho(1-\epsilon)}{\|\boldsymbol{p}(t)\|}\right) \\ &\geq 1-\mu+\frac{\eta}{C(\eta,\rho)}\left(\frac{\rho(\mu-\epsilon)}{\|\boldsymbol{p}(t)\|}-\eta(1-\mu)\frac{\|\nabla\mathcal{L}(\boldsymbol{p}(t))\|^2}{\|\boldsymbol{p}(t)\|}\right) \\ &\geq 1-\mu,\end{aligned} \tag{62}$$

where the last inequality uses $\eta \leq \frac{(\mu-\epsilon)\rho}{1-\mu}\frac{1}{\|\nabla f(\boldsymbol{p}(t))\|^2}$.

Hence, $\boldsymbol{p}(t+1) \in \mathsf{cone}_\mu(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R_\mu\}$.

Step 4: The correlation of $\boldsymbol{p}(t)$ and $\boldsymbol{p}^{mm}(\alpha)$ increases over t. The reminder is similar to the proof of Theorem 1. Note that it follows from Lemma 4 that $\langle \nabla\mathcal{L}(\boldsymbol{p}), \boldsymbol{p}^{mm}(\alpha)/\|\boldsymbol{p}^{mm}(\alpha)\|\rangle < 0$, for any finite $\mathbf{p}$. Hence, there are no finite critical points $\mathbf{p}$, for which $\nabla\mathcal{L}(\mathbf{p}) = 0$ which contradicts Lemma 6. This implies that $\|\mathbf{p}(t)\| \to \infty$. Hence, we can choose t_0 such that for any $t \geq t_0$, it holds that $\|\boldsymbol{p}(t)\| > R$ for some $R \geq R_\mu \vee R_\epsilon \vee 1/2$. Now, following similar steps in (56) and (57), we obtain

$$\left\langle \frac{\boldsymbol{p}(t)}{\|\boldsymbol{p}(t)\|}, \frac{\boldsymbol{p}^{mm}(\alpha)}{\|\boldsymbol{p}^{mm}(\alpha)\|} \right\rangle \geq 1-\epsilon+\frac{C(\epsilon,\eta)}{\|\boldsymbol{p}(t)\|},$$

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

Consequently,

$$\liminf_{t\to\infty}\left\langle \frac{\boldsymbol{p}(t)}{\|\boldsymbol{p}(t)\|}, \frac{\boldsymbol{p}^{mm}(\alpha)}{\|\boldsymbol{p}^{mm}(\alpha)\|} \right\rangle \geq 1-\epsilon.$$

Since $\epsilon \in (0, \min(\mu, 1))$ is arbitrary, we get $\boldsymbol{p}(t)/\|\boldsymbol{p}(t)\| \to \boldsymbol{p}^{mm}(\alpha)/\|\boldsymbol{p}^{mm}(\alpha)\|$. ■

B.6 Proof of Theorem 4: Regularization Path Fails for Non-Locally-Optimal Tokens

The theorem below is essentially a restatement of Theorem 4 and shows that regularization path does not converge to the max-margin solution if token indices $\boldsymbol{\alpha}$ does not satisfy Definition 2. The only difference is that, Theorem 4 replaces the second condition below with a cleaner statement which assumes the linear-independence of the support vectors.

Theorem 8 (Failure of Local Regularization Path) *Fix token indices $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^n$ with SVM-neighbors $(\mathcal{T}_i)_{i=1}^n$. Suppose for some $j \in [n]$, there exists an SVM-neighbor $\beta \in \mathcal{T}_j$ satisfying the following:*

- *$\boldsymbol{x}_{j\beta}$ has a higher score than $\boldsymbol{x}_{j\alpha_j}$: $Y_j \cdot \boldsymbol{v}^\top \boldsymbol{x}_{j\beta} > Y_j \cdot \boldsymbol{v}^\top \boldsymbol{x}_{j\alpha_j}$.*
- *Recall $\boldsymbol{p}^{mm} = \boldsymbol{p}^{mm}(\boldsymbol{\alpha})$ be the solution of (ATT-SVM) and let $\boldsymbol{p}^\beta$ be the solution of (ATT-SVM) where the constraint $(\boldsymbol{k}_{j\alpha_j} - \boldsymbol{k}_{j\beta})^\top \boldsymbol{p} \geq 1$ is not enforced. β is an* active *SVM-neighbor in the sense that $\boldsymbol{p}^\beta$ violates the constraint i.e. $(\boldsymbol{k}_{j\alpha_j} - \boldsymbol{k}_{j\beta})^\top \boldsymbol{p}^\beta < 1$.*

For any $\varepsilon > 0$, there exists $R_\varepsilon > 0$ as follows: Consider the neighborhood of $\boldsymbol{p}^{mm}$: $C_\varepsilon = \mathsf{cone}_\varepsilon(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \geq R_\varepsilon\}$. Define the local path $\bar{\boldsymbol{p}}(R) = \min_{\boldsymbol{p}\in C_\varepsilon, \|\boldsymbol{p}\|\leq R} \mathcal{L}(\boldsymbol{p})$. Then $\lim_{R\to\infty} \frac{\bar{\boldsymbol{p}}(R)}{\|\bar{\boldsymbol{p}}(R)\|} \neq \frac{\boldsymbol{p}^{mm}}{\|\boldsymbol{p}^{mm}\|}$.

Proof of Theorem 4: Using the above theorem we can now conclude the proof of Theorem 4 by showing that, second bullet of Theorem 4 implies the second bullet of Theorem 8. We are given solution $\boldsymbol{p}^{mm}$ and $\boldsymbol{p}^\beta$. Suppose that $\boldsymbol{p}^\beta$ in Theorem 4 does not violate the constraint $(\boldsymbol{k}_{j\alpha_j} - \boldsymbol{k}_{j\beta})^\top \boldsymbol{p}^\beta \geq 1$. Then, it would imply that $\boldsymbol{p}^\beta = \boldsymbol{p}^{mm}$ because $\boldsymbol{p}^\beta$ satisfies all margin constraints and $\|\boldsymbol{p}^\beta\| \leq \|\boldsymbol{p}^{mm}\|$ (because it solves the problem with less constraints), thus, if $\boldsymbol{p}^\beta \neq \boldsymbol{p}^{mm}$, it would contradict with the optimality of $\boldsymbol{p}^{mm}$. Since the active constraints are linearly independent, their Lagrange multipliers are unique. Since $\boldsymbol{p}^\beta$ is missing a linearly independent constraint, the solution $\boldsymbol{p}^\beta$ expressed in terms of Lagrange-weighted constraints cannot equate to the solution $\boldsymbol{p}^{mm}$ expressed in terms of its own Lagrange-weighted constraints that also include the constraint induced by $\boldsymbol{k}_{j\alpha_j} - \boldsymbol{k}_{j\beta}$.

B.6.1 Proof of Theorem 8

Proof strategy: Without losing generality, let us prove the result for 2ε (to simplify the downstream notation). To accomplish the proof, we will follow the following strategy. Fix $\bar{\boldsymbol{p}}^{mm}_{\varepsilon} = \frac{\varepsilon \boldsymbol{p}^{\beta} + (1-\varepsilon)\boldsymbol{p}^{mm}}{\|\varepsilon \boldsymbol{p}^{\beta} + (1-\varepsilon)\boldsymbol{p}^{mm}\|}$ and $\bar{\boldsymbol{p}}^{mm} = \boldsymbol{p}^{mm}/\|\boldsymbol{p}^{mm}\|$. Using $\|\boldsymbol{p}^{\beta}\| \le \|\boldsymbol{p}^{mm}\|$, we observe that $\bar{\boldsymbol{p}}^{mm}_{\varepsilon}$ obeys the correlation inequality

$$(\bar{\boldsymbol{p}}^{mm}_{\varepsilon})^{\top}\bar{\boldsymbol{p}}^{mm} \ge \frac{(1-\varepsilon)\|\boldsymbol{p}^{mm}\|^2 - \varepsilon\|\boldsymbol{p}^{mm}\|\|\boldsymbol{p}^{\beta}\|}{\|\boldsymbol{p}^{mm}\|((1-\varepsilon)\|\boldsymbol{p}^{mm}\| + \varepsilon\|\boldsymbol{p}^{\beta}\|)} \ge \frac{(1-2\varepsilon)\|\boldsymbol{p}^{mm}\|^2}{\|\boldsymbol{p}^{mm}\|^2} \ge 1 - 2\varepsilon.$$

This establishes that $r \cdot \bar{\boldsymbol{p}}^{mm}_{\varepsilon} \in \mathsf{cone}_{2\varepsilon}(\boldsymbol{p}^{mm})$. Thus, we will use $\bar{\boldsymbol{p}}^{mm}_{\varepsilon}$ to show that it is a superior direction to $\boldsymbol{p}^{mm}$. Concretely, for all $R \ge R_{\varepsilon}$, suppose that, there exists $\delta = \delta(\varepsilon)$ such that,

$$\mathcal{L}(R \cdot \bar{\boldsymbol{p}}^{mm}_{\varepsilon}) < \inf_{\|\boldsymbol{p}\|=R, \boldsymbol{p} \in \mathsf{cone}_{\delta}(\boldsymbol{p}^{mm})} \mathcal{L}(\boldsymbol{p}). \tag{63}$$

In words, suppose that $R \cdot \bar{\boldsymbol{p}}^{mm}_{\varepsilon}$ achieves strictly better loss than all points of ℓ_2-norm R within $\mathsf{cone}_{\delta}(\boldsymbol{p}^{mm})$. Establishing this would imply the desired result $\lim_{R\to\infty} \frac{\tilde{\boldsymbol{p}}(R)}{\|\tilde{\boldsymbol{p}}(R)\|} \neq \frac{\boldsymbol{p}^{mm}}{\|\boldsymbol{p}^{mm}\|}$. Since for any choice of $R \ge R_{\varepsilon}$, (63) implies that the optimal direction $\frac{\tilde{\boldsymbol{p}}(R)}{\|\tilde{\boldsymbol{p}}(R)\|}$ is at least δ bounded away from $\frac{\boldsymbol{p}^{mm}}{\|\boldsymbol{p}^{mm}\|}$. In what follows, we will prove this by establishing (63).

First, let us establish the critical properties of $\boldsymbol{p}^{\beta}$. Set $K = n(T-1)$ and gather the set of margin equalities $\boldsymbol{p}^{mm} := \boldsymbol{p}^{mm}(\boldsymbol{\alpha})$ satisfies: These inequalities are given by vectors $(\boldsymbol{v}_{k=1}^{K})$ where $\boldsymbol{v}_k$ is the form $\boldsymbol{k}_{\alpha_i} - \boldsymbol{k}_t$ for $t \neq \alpha_i$. Also let $\boldsymbol{v}_1 = \boldsymbol{k}_{j\alpha_j} - \boldsymbol{k}_{j\beta}$ be the active constraint described in the theorem.

Note that $\|\boldsymbol{p}^{\beta}\| \le \|\boldsymbol{p}^{mm}\|$ since $\boldsymbol{p}^{\beta}$ is solving a max-margin problem with strictly less constraints (over $k \ge 2$). Secondly, we claim that $\boldsymbol{p}^{\beta}$ achieves a strictly larger margin compared to $\boldsymbol{p}^{mm}$ over $k \ge 2$, namely setting $\Gamma = \|\boldsymbol{p}^{mm}\|$ and $\Gamma_{\beta} = \|\boldsymbol{p}^{\beta}\|$

$$\min_{k\ge 2} \boldsymbol{v}_k^{\top}\boldsymbol{p}^{\beta}/\|\boldsymbol{p}^{\beta}\| = 1/\Gamma_{\beta} > \min_{k\ge 2} \boldsymbol{v}_k^{\top}\boldsymbol{p}^{mm}/\|\boldsymbol{p}^{mm}\| = 1/\Gamma.$$

If not, it would imply that $\|\boldsymbol{p}^{\beta}\| = \|\boldsymbol{p}^{mm}\|$ and that $\min_{k\ge 2} \boldsymbol{v}_k^{\top}\boldsymbol{p}^{mm} = \min_{k\ge 2} \boldsymbol{v}_k^{\top}\boldsymbol{p}^{\beta}$. Since theorem's statement guarantees $\boldsymbol{p}^{mm} \neq \boldsymbol{p}^{\beta}$, this contradicts with the unique optimality of $\boldsymbol{p}^{\beta}$ when satisfying constraints $k \ge 2$ as $\boldsymbol{p}^{mm}$ would achieve the same objective.

Finally, using same argument, we also note that, $\boldsymbol{p}^{\beta}$ achieves strictly less margin over $\boldsymbol{v}_1$, namely

$$\boldsymbol{v}_1^{\top}\boldsymbol{p}^{\beta}/\|\boldsymbol{p}^{\beta}\| < \boldsymbol{v}_1^{\top}\boldsymbol{p}^{mm}/\|\boldsymbol{p}^{mm}\| = 1/\Gamma. \tag{64}$$

If not, it would imply that $\boldsymbol{p}^{\beta}$ achieves a better or equal margin at all constraints which would contradict with the optimality of $\boldsymbol{p}^{mm}$ over constraints $k \ge 1$.

Now, let us define $\boldsymbol{p}^{mm}_{\varepsilon} = \varepsilon\boldsymbol{p}^{\beta} + (1-\varepsilon)\boldsymbol{p}^{mm}$ and observe that $\boldsymbol{p}^{mm}_{\varepsilon}$ also satisfies the discussion above. Namely, using $\|\boldsymbol{p}^{mm}_{\varepsilon}\| \le \varepsilon\|\boldsymbol{p}^{\beta}\| + (1-\varepsilon)\|\boldsymbol{p}^{mm}\| < \Gamma$ we find

$$\frac{\min_{k\ge 2} \boldsymbol{v}_k^{\top}\boldsymbol{p}^{mm}_{\varepsilon}}{\|\boldsymbol{p}^{mm}_{\varepsilon}\|} \ge \frac{1}{\varepsilon\|\boldsymbol{p}^{\beta}\| + (1-\varepsilon)\|\boldsymbol{p}^{mm}\|} > \frac{1}{\Gamma}. \tag{65}$$

Similarly, on constraint $\boldsymbol{v}_1$, we have that

$$\boldsymbol{v}_1^{\top}\boldsymbol{p}^{mm}_{\varepsilon}/\|\boldsymbol{p}^{mm}_{\varepsilon}\| := 1/\Gamma_{\varepsilon} < 1/\Gamma. \tag{66}$$

If not, it would imply that $\boldsymbol{p}^{mm}_{\varepsilon}$ achieves a better or equal margin on all constraints which would contradict with the unique optimality of $\boldsymbol{p}^{mm}$ over constraints $k \ge 1$.

We will use $\Gamma_{\varepsilon} > \Gamma$, (66), and (65) to conclude that $\boldsymbol{p}^{mm}_{\varepsilon}$ is a strictly better direction compared to a $\delta = \delta(\varepsilon)$ conic neighborhood of $\bar{\boldsymbol{p}}^{mm} := \boldsymbol{p}^{mm}/\|\boldsymbol{p}^{mm}\|$. Pick the δ neighborhood of $\boldsymbol{p}^{mm}$ such that, all $\boldsymbol{p}^{mm}_{\delta}$ it satisfies

$$\boldsymbol{v}_k^{\top}\boldsymbol{p}^{mm}_{\delta}/\|\boldsymbol{p}^{mm}_{\delta}\| \ge 1/\Gamma_{\delta} = 0.5(1/\Gamma_{\varepsilon} + 1/\Gamma) > 1/\Gamma_{\varepsilon} \quad \text{for all} \quad k \in [K]. \tag{67}$$

In words, we choose a neighborhood with correlation profile dominated by $\boldsymbol{p}^{mm}_{\varepsilon}$ (on $k = 1$ and $k \ge 2$). We now lower bound the loss function over δ-neighborhood and upper bound over $\boldsymbol{p}^{mm}_{\varepsilon}$. Specifically, we will compare a $\boldsymbol{p}^{mm}_{\delta}$ within the δ neighborhood of $\boldsymbol{p}^{mm}$ with $\|\boldsymbol{p}^{mm}_{\delta}\| = R$ and $\tilde{\boldsymbol{p}}^{mm}_{\varepsilon} := R \cdot \bar{\boldsymbol{p}}^{mm}_{\varepsilon}$. To proceed, define:

$$q_i^{\star} = 1 - \mathbb{S}(\boldsymbol{K}_i\tilde{\boldsymbol{p}}^{mm}_{\varepsilon})_{\alpha_i}, \quad \hat{q}_i = 1 - \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}^{mm}_{\delta})_{\alpha_i}$$

Also define $q^\beta = \mathbb{S}(\boldsymbol{K}_j \tilde{\boldsymbol{p}}_\varepsilon^{mm})_\beta$ which is the j'th softmax likelihood at token β. We will use the fact that margin at β is small for j'th example to lower bound q^β carefully. We next bound these as follows based on (66), (65), (67) (e.g. following derivation of (109))[2]

$$\log(q_i^\star) \le -(R/\Gamma) + \log T \quad \text{for all} \quad i \neq j \tag{68}$$

$$\log(q^\beta) \ge -(R/\Gamma_\varepsilon) - \log T \quad \text{for all} \quad i \neq j \tag{69}$$

$$\log(q_j^\star - q^\beta) \le -(R/\Gamma) + \log T \quad \text{for all} \quad i \neq j \tag{70}$$

$$\log(\hat{q}_i) \le -(R/\Gamma_\delta) + \log T \quad \text{for all} \quad i \in [n]. \tag{71}$$

• **Lower bounding $\mathcal{L}(\boldsymbol{p}_\delta^{mm})$:** Using the last inequality, on $\boldsymbol{p}_\delta^{mm}$ (within the δ neighborhood of $\boldsymbol{p}^{mm}$), we have the following lower bound: Set $\boldsymbol{x}_i^\delta = \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i \boldsymbol{p}_\delta^{mm})$ and $M := \sup_{i\in[n],t,\tau\in[T]} \|\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau}\|$ and note that $\|\boldsymbol{x}_i^\delta - \boldsymbol{x}_{\alpha_i}\| \le M\hat{q}_i$. Also let B and A be the lower and upper bound of $-\ell'$ over $[-M\|\boldsymbol{v}\|, M\|\boldsymbol{v}\|]$ interval. Finally, define $\mathcal{L}_\star = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{v}^\top \boldsymbol{x}_{i\alpha_i})$. We find

$$|\mathcal{L}(\boldsymbol{p}_\delta^{mm}) - \mathcal{L}_\star| = \frac{1}{n}\sum_{i=1}^n |\ell(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{x}_i^\delta) - \ell(Y_i \cdot \boldsymbol{v}^\top \boldsymbol{x}_{i\alpha_i})| \tag{72}$$

$$\le B\hat{q}_{\max} M\|\boldsymbol{v}\| \tag{73}$$

$$\le TBM\|\boldsymbol{v}\| e^{-R/\Gamma_\delta}. \tag{74}$$

This implies $\mathcal{L}(\boldsymbol{p}_\delta^{mm}) \ge \mathcal{L}_\star - TBM\|\boldsymbol{v}\| e^{-R/\Gamma_\delta}$. Note that this holds for all $\boldsymbol{p}_\delta^{mm}$ within the conic neighborhood $\mathsf{cone}_\delta(\boldsymbol{p}^{mm}) \cap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| = R\}$ (defined above (63)).

• **Upper bounding $\mathcal{L}(\tilde{\boldsymbol{p}}_\varepsilon^{mm})$:** On $\boldsymbol{p}_\varepsilon^{mm}$, we upper bound the loss as follows. Define the loss $\mathcal{L}^{-j}(\boldsymbol{p}) = \frac{1}{n}\sum_{i\neq j} \ell(Y_i \cdot f(\boldsymbol{X}_i))$ i.e. loss over all training data except the j'th one. Repeating the argument identical to (72), we find that

$$|\mathcal{L}^{-j}(\tilde{\boldsymbol{p}}_\varepsilon^{mm}) - \mathcal{L}_\star^{-j}| \le TBM\|\boldsymbol{v}\| e^{-R/\Gamma}.$$

The critical term is the j'th loss which we need to upper bound as follows. Set $\boldsymbol{x}_j^\varepsilon = \boldsymbol{X}_j^\top \mathbb{S}(\boldsymbol{K}_j \tilde{\boldsymbol{p}}_\varepsilon^{mm})$ and define the score improvement by β to be $\boldsymbol{\gamma}_{gap} = Y_j \cdot \boldsymbol{v}^\top(\boldsymbol{x}_{i\beta} - \boldsymbol{x}_{j\alpha_j}) > 0$. We note that

$$Y_j \cdot \boldsymbol{v}^\top \boldsymbol{x}_j^\varepsilon - Y_j \cdot \boldsymbol{v}^\top \boldsymbol{x}_{\alpha_j} = q^\beta Y_j \cdot \boldsymbol{v}^\top(\boldsymbol{x}_{i\beta} - \boldsymbol{x}_{\alpha_j}) + \sum_{t\notin\{\alpha_j,\beta\}} \mathbb{S}(\boldsymbol{K}_j \tilde{\boldsymbol{p}}_\varepsilon^{mm})_t Y_j \cdot \boldsymbol{v}^\top(\boldsymbol{x}_{it} - \boldsymbol{x}_{\alpha_j}) \tag{75}$$

$$\ge q^\beta \boldsymbol{\gamma}_{gap} - (q_j^\star - q^\beta) M\|\boldsymbol{v}\| \tag{76}$$

$$\ge T^{-1}\boldsymbol{\gamma}_{gap} e^{-R/\Gamma_\varepsilon} - TBM\|\boldsymbol{v}\| e^{-R/\Gamma}. \tag{77}$$

Combining these into $\mathcal{L}(\tilde{\boldsymbol{p}}_\varepsilon^{mm}) = \mathcal{L}^{-j}(\tilde{\boldsymbol{p}}_\varepsilon^{mm}) + n^{-1}\ell(Y_j \cdot \boldsymbol{v}^\top \boldsymbol{x}_j^\varepsilon)$ and using $A \le -\ell' \le B$, we obtain the lower bound

$$\mathcal{L}(\tilde{\boldsymbol{p}}_\varepsilon^{mm}) - \mathcal{L}_\star \le 2TBM\|\boldsymbol{v}\| e^{-R/\Gamma} - An^{-1}T^{-1}\boldsymbol{\gamma}_{gap} e^{-R/\Gamma_\varepsilon}. \tag{78}$$

In conclusion, we find that $\mathcal{L}(\tilde{\boldsymbol{p}}_\varepsilon^{mm}) > \mathcal{L}(\boldsymbol{p}_\delta^{mm})$ whenever

$$An^{-1}T^{-1}\boldsymbol{\gamma}_{gap} e^{-R/\Gamma_\varepsilon} > 2TBM\|\boldsymbol{v}\|(e^{-R/\Gamma} + e^{-R/\Gamma_\delta}).$$

Using the relationship $\Gamma_\varepsilon \ge \Gamma_\delta \ge \Gamma$ and noticing $1/\Gamma_\delta - 1/\Gamma_\varepsilon = (1/\Gamma - 1/\Gamma_\varepsilon)/2$, this is implied by

$$e^{R(1/\Gamma_\delta - 1/\Gamma_\varepsilon)} > 4T^2 nBMA^{-1}\|\boldsymbol{v}\| \iff R > \frac{2\Gamma_\varepsilon\Gamma}{\Gamma_\varepsilon - \Gamma}\log(4T^2 nBMA^{-1}\|\boldsymbol{v}\|).$$

Thus, as advertised, we found that, for any $\varepsilon > 0$, there exists R_ε such that, over the set $\mathsf{cone}_\varepsilon(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| \ge R_\varepsilon\}$, $\tilde{\boldsymbol{p}}_\varepsilon^{mm}$ with $\|\tilde{\boldsymbol{p}}_\varepsilon^{mm}\| = R > R_\varepsilon$ achieves smaller loss compared to $\mathcal{L}(\boldsymbol{p}_\delta^{mm})$ for all $\boldsymbol{p}_\delta^{mm} \in \mathsf{cone}_\delta(\boldsymbol{p}^{mm}) \bigcap \{\boldsymbol{p} \mid \|\boldsymbol{p}\| = R\}$. This in turn implies (63) for all $R > R_\varepsilon$ concluding the proof.

[2] We are essentially following identical arguments developed in the proofs of Theorem 7 or Theorem 5.

C Addendum to Section 3

C.1 Proof of Theorem 5

Proof. Suppose the claim is incorrect and either $\boldsymbol{p}_R/R$ or $\boldsymbol{v}_r/r$ fails to converge as R, r grows. Set $\Xi = 1/\|\boldsymbol{p}^{mm}\|$, $\tilde{\boldsymbol{p}}^{mm} = R\Xi\boldsymbol{p}^{mm}$ and $\tilde{\boldsymbol{v}}^{mm} = r\Gamma\boldsymbol{v}^{mm}$. The proof strategy is obtaining a contradiction by proving that $(\tilde{\boldsymbol{v}}^{mm}, \tilde{\boldsymbol{p}}^{mm})$ is a strictly better solution compared to $(\boldsymbol{v}_r, \boldsymbol{p}_R)$ for large R, r. Without losing generality, we will set $\alpha_i = 1$ for all $i \in [n]$ as the problem is invariant to tokens' permutation. Define $q_i^{\boldsymbol{p}} = 1 - \boldsymbol{s}_{i1}^{\boldsymbol{p}}$ to be the amount of non-optimality (cumulative probability of non-first tokens) where $\boldsymbol{s}_i^{\boldsymbol{p}} = \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p})$ is the softmax probabilities.

• **Case 1: $\boldsymbol{p}_R/R$ does not converge.** Under this scenario there exists $\delta, \gamma = \gamma(\delta) > 0$ such that we can find arbitrarily large R with $\|\boldsymbol{p}_R/R - \tilde{\boldsymbol{p}}^{mm}/R\| \geq \delta$ and margin induced by $\boldsymbol{p}_R/R$ is at most $\Xi(1-\gamma)$ (from strong convexity of (ATT-SVM)). Following $q_i^{\boldsymbol{p}}$ definition above, set $\hat{q}_{\max} = \sup_{i\in[n]} q_i^{\boldsymbol{p}_R}$ to be worst non-optimality in $\boldsymbol{p}_R$ and $q^\star_{\max} = \sup_{i\in[n]} q_i^{\tilde{\boldsymbol{p}}^{mm}}$ to be the same for $\tilde{\boldsymbol{p}}^{mm}$. Repeating the identical argument in Theorem 7 (specifically (109)), we can bound the non-optimality amount $q_i^\star$ of $\boldsymbol{p}_R^\star$ as

$$q_i^\star = \frac{\sum_{t\neq\alpha_i}\exp(\boldsymbol{k}_{it}^\top\boldsymbol{p}_R^\star)}{\sum_{t\in[T]}\exp(\boldsymbol{k}_{it}^\top\boldsymbol{p}_R^\star)} \leq \frac{\sum_{t\neq\alpha_i}\exp(\boldsymbol{k}_{it}^\top\boldsymbol{p}_R^\star)}{\exp(\boldsymbol{k}_{i\alpha_i}^\top\boldsymbol{p}_R^\star)} \leq T\exp(-R\Xi). \tag{79}$$

Thus, $q^\star_{\max} = \max_{i\in[n]} q_i^\star \leq T\exp(-R\Xi)$. Next without losing generality, assume first margin constraint is γ-violated by $\boldsymbol{p}_R$ and $\min_{t\neq\alpha_1}(\boldsymbol{k}_{1\alpha_1} - \boldsymbol{k}_{1t})^\top\boldsymbol{p}_R \leq \Xi R(1-\gamma)$. Denoting the amount of non-optimality of the first input as $\hat{q}_1$, we find

$$\hat{q}_1 = \frac{\sum_{t\neq\alpha_i}\exp(\boldsymbol{k}_{1t}^\top\boldsymbol{p}_R)}{\sum_{t\in[T]}\exp(\boldsymbol{k}_{1t}^\top\boldsymbol{p}_R)} \geq \frac{1}{T}\frac{\sum_{t\neq\alpha_i}\exp(\boldsymbol{k}_{1t}^\top\boldsymbol{p}_R)}{\exp(\boldsymbol{k}_{1\alpha}^\top\boldsymbol{p}_R)} \geq T^{-1}\exp(-(1-\gamma)R\Xi). \tag{80}$$

We similarly have $q^\star_{\max} \geq T^{-1}\exp(-R\Xi)$ to find that

$$\log(\hat{q}_{\max}) \geq -(1-\gamma)\Xi R - \log T, \tag{81}$$

$$-\Xi R - \log T \leq \log(q^\star_{\max}) \leq -\Xi R + \log T. \tag{82}$$

In words, $\tilde{\boldsymbol{p}}^{mm}$ contains exponentially less non-optimality compared to $\boldsymbol{p}_R$ as R grows. The remainder of the proof differs from Theorem 7 as we need to upper/lower bound the logistic loss of $(\tilde{\boldsymbol{v}}^{mm}, \tilde{\boldsymbol{p}}^{mm})$ and $(\boldsymbol{v}_r, \boldsymbol{p}_R)$ respectively to conclude with the contradiction.

First, let us upper bound the logistic loss of $(\tilde{\boldsymbol{v}}^{mm}, \tilde{\boldsymbol{p}}^{mm})$. Set $\boldsymbol{r}_i = \boldsymbol{X}_i^\top\mathbb{S}(\boldsymbol{K}_i\tilde{\boldsymbol{p}}^{mm})$. Observe that if $\|\boldsymbol{r}_i - \boldsymbol{x}_{i1}\| \leq \varepsilon_i$, we have that $\boldsymbol{v}^{mm}$ satisfies the SVM constraints on $\boldsymbol{r}_i$ with $Y_i\cdot\boldsymbol{r}_i^\top\boldsymbol{v}^{mm} \geq 1 - \varepsilon_i/\Gamma$. Consequently, setting $\varepsilon_{\max} = \sup_{i\in[n]}\varepsilon_i$, $\boldsymbol{v}^{mm}$ achieves a label-margin of $\Gamma - \varepsilon_{\max}$ on the dataset $(Y_i, \boldsymbol{r}_i)_{i\in[n]}$. With this, we upper bound the logistic loss of $(\tilde{\boldsymbol{v}}^{mm}, \tilde{\boldsymbol{p}}^{mm})$ as follows. Let $M = \sup_{i\in[n],t\in[T]}\|\boldsymbol{x}_{it}\|$. In what follows, let us recall the fact (81) that worst-case perturbation is $\varepsilon_{\max} \leq M\exp(-\Xi R + \log T) = MT\exp(-\Xi R)$.

$$\mathcal{L}(\tilde{\boldsymbol{v}}^{mm}, \tilde{\boldsymbol{p}}^{mm}) \leq \max_{i\in[n]}\log(1+\exp(-Y_i\boldsymbol{r}_i^\top\tilde{\boldsymbol{v}}^{mm})). \tag{83}$$

$$\leq \max_{i\in[n]}\exp(-Y_i\boldsymbol{r}_i^\top\tilde{\boldsymbol{v}}^{mm}) \tag{84}$$

$$\leq \exp(-r\Gamma + r\varepsilon_{\max}) \tag{85}$$

$$\leq e^{rMT\exp(-\Xi R)}e^{-r\Gamma}. \tag{86}$$

Conversely, we obtain a lower bound for $(\boldsymbol{v}_r, \boldsymbol{p}_R)$. Set $\boldsymbol{r}_i = \boldsymbol{X}_i^\top\mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}_R)$. Using Assumption C, we find that solving (CLS-SVM) on $(Y_i, r_i)_{i\in[n]}$ achieves at most $\Gamma - \nu e^{-(1-\gamma)\Xi R}/T$ margin. Consequently, we have

$$\mathcal{L}(\boldsymbol{v}_r, \boldsymbol{p}_R) \geq \frac{1}{n}\max_{i\in[n]}\log(1+\exp(-Y_i\boldsymbol{r}_i^\top\boldsymbol{v}_r)) \tag{87}$$

$$\geq \frac{1}{2n}\max_{i\in[n]}\exp(-Y_i\boldsymbol{r}_i^\top\boldsymbol{v}_r) \wedge \log 2 \tag{88}$$

$$\geq \frac{1}{2n}\exp(-r(\Gamma - \nu e^{-(1-\gamma)\Xi R}/T)) \wedge \log 2 \tag{89}$$

$$\geq \frac{1}{2n}e^{r(\nu/T)\exp(-(1-\gamma)\Xi R)}e^{-r\Gamma} \wedge \log 2. \tag{90}$$

Observe that, this lower bound dominates the previous upper bound when R is large, namely, when (ignoring the multiplier $1/2n$ for brevity)

$$(\nu/T)e^{-(1-\gamma)\Xi R} \geq MTe^{-\Xi R} \iff R \geq R_0 := \frac{1}{\gamma\Xi}\log(\frac{MT^2}{\nu}).$$

Thus, we indeed obtain the desired contradiction since such large R is guaranteed to exist when $\boldsymbol{p}_R/R \nrightarrow \boldsymbol{p}^{mm}$.

• **Case 2: $\boldsymbol{v}_r/r$ does not converge.** This is the simpler scenario: There exists $\delta > 0$ such that we can find arbitrarily large r obeying $\|\boldsymbol{v}_r/r - \boldsymbol{v}^{mm}/\|\boldsymbol{v}^{mm}\|\| \geq \delta$. If $\|\boldsymbol{p}_R/R - \Xi\boldsymbol{p}^{mm}\| \nrightarrow 0$, then "Case 1" applies. Otherwise, we have $\|\boldsymbol{p}_R/R - \Xi\boldsymbol{p}^{mm}\| \to 0$, thus we can assume $\|\boldsymbol{p}_R/R - \Xi\boldsymbol{p}^{mm}\| \leq \varepsilon$ for arbitrary choice of $\varepsilon > 0$.

On the other hand, due to the strong convexity of (CLS-SVM), for some $\gamma := \gamma(\delta) > 0$, $\boldsymbol{v}_r$ achieves a margin of at most $(1-\gamma)\Gamma r$ on the dataset $(Y_i, \boldsymbol{x}_{i1})_{i\in[n]}$. Additionally, since $\|\boldsymbol{p}_R/R - \Xi\boldsymbol{p}^{mm}\| \leq \varepsilon$, $\boldsymbol{p}_R$ strictly separates all optimal tokens (for small enough $\varepsilon > 0$) and $\hat{q}_{\max} := f(\varepsilon) \to 0$ as $R \to \infty$. Consequently, setting $\boldsymbol{r}_i = \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}_R)$, for sufficiently large $R > 0$ setting $M = \sup_{i\in[n],t\in[T]}\|\boldsymbol{x}_{it}\|$, we have that

$$\min_{i\in[n]} Y_i\boldsymbol{v}_r^\top\boldsymbol{r}_i \leq \min_{i\in[n]} Y_i\boldsymbol{v}_r^\top\boldsymbol{x}_{i1} + \sup_{i\in[n]}|\boldsymbol{r}_i - \boldsymbol{x}_{i1}|\boldsymbol{v}_r^\top \tag{91}$$

$$\leq (1-\gamma)\Gamma r + Mf(\varepsilon)r \tag{92}$$

$$\leq (1-\gamma/2)\Gamma r. \tag{93}$$

This in turn implies that logistic loss is lower bounded by (following (90)),

$$\mathcal{L}(\boldsymbol{v}_r, \boldsymbol{p}_R) \geq \frac{1}{2n}e^{\gamma\Gamma r/2}e^{-\Gamma r} \wedge \log 2.$$

Going back to (86), this exponentially dominates the upper bound of $(\tilde{\boldsymbol{p}}^{mm}, \tilde{\boldsymbol{v}}^{mm})$ whenever $rMT\exp(-\Xi R) < r\gamma\Gamma/2$, (that is, whenever R, r are sufficiently large), again concluding the proof. ■

C.2 Proof of Theorem 6

We first restate Theorem 6 for ease of reference.

Theorem 9 *Consider the path of $(\boldsymbol{v}_r, \boldsymbol{p}_R)$ as $r, R \to \infty$ as in Theorem 5. Suppose $\mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}_R)_{\alpha_i} \to 1$, i.e., the tokens $(\alpha_i)_{i=1}^n$ are asymptotically selected. Then, $\boldsymbol{v}_r/r \to \boldsymbol{v}^{mm}/\|\boldsymbol{v}^{mm}\|$ where $\boldsymbol{v}^{mm}$ is the solution of (CLS-SVM) with $\boldsymbol{r}_i = \boldsymbol{x}_{i\alpha_i}$, $\mathcal{N}$ is the set of non-support vectors for (CLS-SVM), and $\frac{\boldsymbol{p}_R}{R} \to \frac{\boldsymbol{p}^{relax}}{\|\boldsymbol{p}^{relax}\|}$ where $\boldsymbol{p}^{relax}$ is the solution of (7) with α_i choices.*

We will prove this result in two steps. Our first claim restricts the optimization to the particular quadrant induced by $\min_{t\neq\alpha_i}(\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it})\boldsymbol{p}_R$ under the theorem's condition $\mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}_R) \to \boldsymbol{e}_{\alpha_i}$.

Lemma 7 *Suppose $\mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}_R) \to \boldsymbol{e}_{\alpha_i}$. Then, there exists R_0 such that for all $R \geq R_0$, we have that,*

$$\min_{t\neq\alpha_i}(\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it})\boldsymbol{p}_R \geq 0 \quad \textit{for all} \quad i \in [n]. \tag{94}$$

Proof. Suppose the claim does not hold. Set $\boldsymbol{s}_i^R = \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}_R)$. Fix R_0 such that $\boldsymbol{s}_{i\alpha_i}^R \geq 0.9$ for all $R \geq R_0$. On the other hand, there exists arbitrarily large R for which $(\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it})\boldsymbol{p}_R < 0$ for some $t \neq \alpha_i \in [T]$ and $i \in [n]$. At this (R, i, t) choices, we have that $\boldsymbol{s}_{it}^R \geq \boldsymbol{s}_{i\alpha_i}^R$. Since $\boldsymbol{s}_{it}^R + \boldsymbol{s}_{i\alpha_i}^R \leq 1$, we find $\boldsymbol{s}_{i\alpha_i}^R < 0.5$ which contradicts with $\boldsymbol{s}_{i\alpha_i}^R \geq 0.9$. ■

Let $\mathcal{Q}$ be the set of $\boldsymbol{p}$ satisfying the quadrant constraint (94) – i.e. indices $(\alpha_i)_{i=1}^n$ are selected. Let $\boldsymbol{h}_R$ be the solution of regularization path of $(\boldsymbol{v}, \boldsymbol{p})$ subject to the constraint $\boldsymbol{p} \in \mathcal{Q}$. From Lemma 7, we know that, for some R_0 and all $R \geq R_0$, $\boldsymbol{h}_R = \boldsymbol{p}_R$. Thus, if the limit exists, we have that $\lim_{R\to\infty}\boldsymbol{h}_R/R = \lim_{R\to\infty}\boldsymbol{p}_R/R$.

To proceed, we will prove that $\lim_{R\to\infty}\boldsymbol{h}_R/R$ exists and is equal to $\boldsymbol{p}^{relax}/\|\boldsymbol{p}^{relax}\|$ and simultaneously establish $\boldsymbol{v}_r/r \to \boldsymbol{v}^{mm}/\|\boldsymbol{v}^{mm}\|$.

Lemma 8 $\lim_{R,r\to\infty} \boldsymbol{h}_R/R = \boldsymbol{p}^{relax}/\|\boldsymbol{p}^{relax}\|$ *and* $\lim_{R,r\to\infty} \boldsymbol{v}_r/r = \boldsymbol{v}^{mm}/\|\boldsymbol{v}^{mm}\|$.

Proof. The proof will be similar to that of Theorem 5. As usual, we aim to show that SVM-solutions constitute the most competitive direction. Set $\Xi = 1/\|\boldsymbol{p}^{relax}\|$.

• **Case 1: $\boldsymbol{h}_R/R$ does not converge.** Under this scenario there exists $\delta, \gamma = \gamma(\delta) > 0$ such that we can find arbitrarily large R with $\|\boldsymbol{h}_R/R - \Xi\boldsymbol{p}^{relax}\| \geq \delta$. This implies that margin induced by $\boldsymbol{h}_R/R$ is at most $\Xi(1-\gamma)$ over the support vectors $[n] - \mathcal{N}$ (from strong convexity of (7)). The reason is that, $\boldsymbol{h}_R$ satisfies $\boldsymbol{h}_R^\top(\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it}) \geq 0 \quad$ for all $\quad t \neq \alpha_i$ by construction as $\boldsymbol{h}_R \in \mathcal{Q}$. Thus, a constraint over the support vectors have to be violated (when normalized to the same ℓ_2 norm as $\|\boldsymbol{p}^{relax}\| = 1/\Xi$).

As usual, we will construct a solution strictly superior to $\boldsymbol{h}_R$ and contradicts with its optimality.

Construction of competitor: Rather than using $\boldsymbol{p}^{relax}$ direction, we will choose a slightly deviating direction that ensures the selection of the correct tokens over non-supports $\mathcal{N}$. Specifically, consider the solution of (7) where we tighten the non-support constraints by arbitrarily small $\varepsilon > 0$.

$$\boldsymbol{p}^{\varepsilon\text{-}rlx} = \min_{\boldsymbol{p}} \|\boldsymbol{p}\| \quad \text{such that} \quad \boldsymbol{p}^\top(\boldsymbol{k}_{i\alpha_i} - \boldsymbol{k}_{it}) \geq \begin{cases} 1 & \text{for all} \quad t \neq \alpha_i,\ i \in [n] - \mathcal{N} \\ \varepsilon & \text{for all} \quad t \neq \alpha_i,\ i \in \mathcal{N} \end{cases}. \tag{95}$$

Let $\boldsymbol{p}^{mm}$ be the solution of (ATT-SVM) with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_i)_{i=1}^n$ (which was assumed to be separable). Observe that $\boldsymbol{p}_\varepsilon^{mm} = \varepsilon\boldsymbol{p}^{mm} + (1-\varepsilon)\boldsymbol{p}^{relax}$ satisfies the constraints of (95). Additionally, $\boldsymbol{p}_\varepsilon^{mm}$ would achieve a margin of $\frac{1}{(1-\varepsilon)/\Xi + \varepsilon/\Delta} = \frac{\Delta\Xi}{\Delta + \varepsilon(\Xi - \Delta)}$ where $\Delta = 1/\|\boldsymbol{p}^{mm}\|$. Using optimality of $\boldsymbol{p}^{\varepsilon\text{-}rlx}$, this implies that the reduced margin $\Xi_\varepsilon = 1/\|\boldsymbol{p}^{\varepsilon\text{-}rlx}\|$ (by enforcing ε over non-support) over the support vectors is a Lipschitz function of ε. That is $\Xi_\varepsilon \geq \Xi - \varepsilon M$ for some $M \geq 0$. To proceed, choose an $\varepsilon > 0$ such that, it is strictly superior to margin induced by $\boldsymbol{h}_R$, that is,

$$\Xi_\varepsilon \geq \Xi(1 - \frac{\gamma}{2}).$$

To proceed, set $\tilde{\boldsymbol{p}}^{\varepsilon\text{-}rlx} = R\Xi_\varepsilon\boldsymbol{p}^{\varepsilon\text{-}rlx}$. Let us recall the following notation from the proof of Theorem 5: $\boldsymbol{s}_i^{\boldsymbol{p}} = \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p})$ and $q_i^{\boldsymbol{p}} = 1 - \boldsymbol{s}_{i\alpha_i}$. Set $\hat{q}_{\max} = \max_i \hat{q}_{i\in[n]-\mathcal{N}}$ to be worst non-optimality of $\boldsymbol{h}_R$ over **support set**. Similarly, define $q^\star_{\max} = \max_{i\in[n]-\mathcal{N}} q_i^\star$ to be the same for $\tilde{\boldsymbol{p}}^{\varepsilon\text{-}rlx}$. Repeating the identical arguments to (79), (80), (81), and using the fact that $\boldsymbol{p}^{\varepsilon\text{-}rlx}$ achieves a margin $\Xi(1-\frac{\gamma}{2}) \leq \Xi_\varepsilon \leq \Xi$, we end up with the lines

$$\log(\hat{q}_{\max}) \geq -(1-\gamma)\Xi R - \log T, \tag{96a}$$

$$-\Xi R - \log T \leq \log(q^\star_{\max}) \leq -\Xi(1 - 0.5\gamma)R + \log T. \tag{96b}$$

In what follows, we will prove that $\tilde{\boldsymbol{p}}^{\varepsilon\text{-}rlx}$ achieves a strictly smaller logistic loss contradicting with the optimality of $\boldsymbol{p}_R$ (whenever $\|\boldsymbol{h}_R/R - \Xi\boldsymbol{p}^{relax}\| \geq \delta$).

Upper bounding logistic loss. Let us now upper bound the logistic loss of $(\tilde{\boldsymbol{v}}^{mm}, \tilde{\boldsymbol{p}}^{\varepsilon\text{-}rlx})$ where $\tilde{\boldsymbol{v}}^{mm} = r\Gamma\boldsymbol{v}^{mm}$ with $\boldsymbol{v}^{mm}$ being the solution of (CLS-SVM) with $\boldsymbol{r}_i \leftarrow \boldsymbol{x}_{i\alpha_i}$ and $\Gamma = 1/\|\boldsymbol{v}^{mm}\|$. Set $\boldsymbol{r}_i = \boldsymbol{X}_i^\top\mathbb{S}(\boldsymbol{K}_i\tilde{\boldsymbol{p}}^{\varepsilon\text{-}rlx})$. Set $\upsilon = \min_{i\in\mathcal{N}} Y_i \cdot \boldsymbol{x}_{i\alpha_i}^\top\boldsymbol{v}^{mm} - 1$ to be the additional margin buffer that non-support vectors have access to. Also set $M = \sup_{i\in[n],t,\tau\in[T]} \|\boldsymbol{x}_{it} - \boldsymbol{x}_{i\tau}\|$. Observe that we can write

$$\boldsymbol{x}_{i\alpha_i} - \boldsymbol{r}_i = \sum_{t\neq\alpha_i} \boldsymbol{s}_{it}(\boldsymbol{x}_{i\alpha_i} - \boldsymbol{x}_{it}) \implies \|\boldsymbol{x}_{i\alpha_i} - \boldsymbol{r}_i\| \leq q_i M.$$

Non-supports achieve strong label-margin: Using above and (95) for all $i \in \mathcal{N}$ and $t \neq \alpha_i$, we have that $\boldsymbol{s}_{it} \leq e^{-\varepsilon\Xi_\varepsilon R}\boldsymbol{s}_{i\alpha_i} \leq e^{-\varepsilon\Xi(1-\gamma/2)R}\boldsymbol{s}_{i\alpha_i}$. Consequently, whenever $R \geq \bar{R}_0 := (\varepsilon\Xi(1-\gamma/2))^{-1}\log(\frac{TM}{\Gamma\upsilon})$,

$$q_i^\star \leq \frac{\sum_{t\neq\alpha_i}\boldsymbol{s}_{it}}{\boldsymbol{s}_{i\alpha_i}} \leq Te^{-\varepsilon\Xi(1-\gamma/2)R} \leq \frac{\Gamma\upsilon}{M}.$$

This implies that, on $i \in \mathcal{N}$

$$Y_i \cdot \boldsymbol{r}_i^\top\boldsymbol{v}^{mm} \geq 1 + \upsilon + Y_i \cdot (\boldsymbol{r}_i - \boldsymbol{x}_{i\alpha_i})^\top\boldsymbol{v}^{mm} \geq 1 + \upsilon - q_i M\|\boldsymbol{v}^{mm}\| \geq 1. \tag{97}$$

In words: Above a fixed $\bar{R}_0$ that only depends on $\gamma = \gamma(\delta)$, features $\boldsymbol{r}_i$ induced by all non-support indices $i \in \mathcal{N}$ achieve margin at least 1. What remains is analyzing the margin shrinkage over the support vectors as in Theorem 5.

Controlling support margin and combining bounds: Over $[n] - \mathcal{N}$, suppose $\boldsymbol{v}^{mm}$ satisfies the SVM constraints on $\boldsymbol{r}_i$ with $Y_i \cdot \boldsymbol{r}_i^\top \boldsymbol{v}^{mm} \geq 1 - \varepsilon_i/\Gamma$. Consequently, setting $\varepsilon_{\max} = \sup_{i\in[n]} \varepsilon_i$, $\boldsymbol{v}^{mm}$ achieves a label-margin of $\Gamma - \varepsilon_{\max}$ on the dataset $(Y_i, \boldsymbol{r}_i)_{i\in[n]}$. Next, we recall the fact (96b) that worst-case perturbation is $\varepsilon_{\max} \leq M \exp(-\Xi(1-0.5\gamma)R + \log T) = MT \exp(-\Xi(1-0.5\gamma)R)$. With this and (97), we upper bound the logistic loss of $(\tilde{\boldsymbol{v}}^{mm}, \tilde{\boldsymbol{p}}^{\varepsilon\text{-rlx}})$ as follows.

$$\mathcal{L}(\tilde{\boldsymbol{v}}^{mm}, \tilde{\boldsymbol{p}}^{mm}) \leq \max_{i\in[n]} \log(1 + \exp(-Y_i \boldsymbol{r}_i^\top \tilde{\boldsymbol{v}}^{mm})). \tag{98}$$

$$\leq \max_{i\in[n]} \exp(-Y_i \boldsymbol{r}_i^\top \tilde{\boldsymbol{v}}^{mm}) \tag{99}$$

$$\leq \exp(-r\Gamma + r\varepsilon_{\max}) \tag{100}$$

$$\leq e^{rMT \exp(-\Xi(1-0.5\gamma)R)} e^{-r\Gamma}. \tag{101}$$

Conversely, we obtain a lower bound for $(\boldsymbol{v}_r, \boldsymbol{h}_R)$. Set $\boldsymbol{r}_i = \boldsymbol{X}_i^\top \mathbb{S}(\boldsymbol{K}_i \boldsymbol{h}_R)$. Recall the lower bound (96a) over the support vector set $[n] - \mathcal{N}$. Combining this with our Assumption C over the support vectors of (CLS-SVM) implies that, solving (CLS-SVM) on $(Y_i, r_i)_{i\in[n]}$ achieves at most $\Gamma - \nu e^{-(1-\gamma)\Xi R}/T$ margin. Consequently, we have

$$\mathcal{L}(\boldsymbol{v}_r, \boldsymbol{h}_R) \geq \frac{1}{n} \max_{i\in[n]} \log(1 + \exp(-Y_i \boldsymbol{r}_i^\top \boldsymbol{v}_r)) \tag{102}$$

$$\geq \frac{1}{2n} \max_{i\in[n]} \exp(-Y_i \boldsymbol{r}_i^\top \boldsymbol{v}_r) \wedge \log 2 \tag{103}$$

$$\geq \frac{1}{2n} \exp(-r(\Gamma - \nu e^{-(1-\gamma)\Xi R}/T)) \wedge \log 2 \tag{104}$$

$$\geq \frac{1}{2n} e^{r(\nu/T)\exp(-(1-\gamma)\Xi R)} e^{-r\Gamma} \wedge \log 2. \tag{105}$$

Observe that, this lower bound dominates the previous upper bound when R is large, namely, when (ignoring the multiplier $1/2n$ for brevity)

$$(\nu/T) e^{-(1-\gamma)\Xi R} \geq MT e^{-\Xi(1-0.5\gamma)R} \iff R \geq R_0 := \frac{2}{\gamma\Xi} \log(\frac{MT^2}{\nu}).$$

Thus, we obtain the desired contradiction since $\tilde{\boldsymbol{p}}^{\varepsilon\text{-rlx}}$ is a strictly better solution compared to $\boldsymbol{p}_R = \boldsymbol{h}_R$ (once R is sufficiently large).

• **Case 2: $\boldsymbol{v}_r/r$ does not converge.** This is the simpler scenario: There exists $\delta > 0$ such that we can find arbitrarily large r obeying $\|\boldsymbol{v}_r/r - \boldsymbol{v}^{mm}/\|\boldsymbol{v}^{mm}\|\| \geq \delta$. First, note that, due to the strong convexity of (CLS-SVM), for some $\gamma := \gamma(\delta) > 0$, $\boldsymbol{v}_r$ achieves a margin of at most $(\Gamma - \gamma)r$ on the dataset $(Y_i, \boldsymbol{x}_{i1})_{i\in[n]}$. By theorem's condition, we are provided that $\mathbb{S}(\boldsymbol{K}_i \boldsymbol{p}_R)_{\alpha_i} \to 1$. This immediately implies that, for any choice of $\varepsilon = \gamma/3 > 0$, above some sufficiently large (r_0, R_0), we have that $\|\boldsymbol{x}_i^{\boldsymbol{p}_R} - \boldsymbol{r}_i\| \leq \varepsilon$. Following (101), this implies that, choosing $\tilde{\boldsymbol{v}}^{mm} = r\boldsymbol{v}^{mm}/\|\boldsymbol{v}^{mm}\|$ achieves a logistic loss of at most $e^{r\gamma/3} e^{-r\Gamma}$. Again using $\|\boldsymbol{x}_i^{\boldsymbol{p}_R} - \boldsymbol{r}_i\| \leq \varepsilon$, for sufficiently large (r, R) we have that

$$\min_{i\in[n]} Y_i \boldsymbol{v}_r^\top \boldsymbol{r}_i \leq \min_{i\in[n]} Y_i \boldsymbol{v}_r^\top \boldsymbol{x}_{i1} + \sup_{i\in[n]} |\boldsymbol{r}_i - \boldsymbol{x}_{i1}| \boldsymbol{v}_r^\top \tag{106}$$

$$\leq (\Gamma - \gamma)r + \varepsilon r \tag{107}$$

$$\leq (\Gamma - 2\gamma/3)r. \tag{108}$$

This in turn implies that logistic loss is lower bounded by (following (105)),

$$\mathcal{L}(\boldsymbol{v}_r, \boldsymbol{p}_R) \geq \frac{1}{2n} e^{2\gamma r/3} e^{-r\Gamma} \wedge \log 2.$$

This dominates the above upper bound $e^{r\gamma/3} e^{-r\Gamma}$ of $\tilde{\boldsymbol{v}}^{mm}$ whenever $\frac{1}{2n} e^{\gamma r/3} > 1 \iff r > \frac{3}{\gamma} \log(2n)$, (that is, when r is sufficiently large), again concluding the proof. ■

D Addendum to Section 4

D.1 Proof of Theorem 7

Proof. The key idea is showing that, thanks to the exponential tail of softmax-attention, (harmful) contribution of the non-optimal token with the minimum margin can dominate the contribution of

all other tokens as $R \to \infty$. This high-level approach is similar to earlier works on implicit bias of gradient descent with logistic loss.

Pick $\boldsymbol{p}^{mm} \in \mathcal{P}^{mm}$ and set $\boldsymbol{p}_R^\star = R\frac{\boldsymbol{p}^{mm}}{\|\boldsymbol{p}^{mm}\|}$. This will be the baseline model that $\boldsymbol{p}_R$ has to compete against. Also let $\bar{\boldsymbol{p}}_R = \Gamma\frac{\boldsymbol{p}_R}{R}$. Now suppose $\texttt{dist}\,(\bar{\boldsymbol{p}}_R, \mathcal{P}^{mm}) \nrightarrow 0$ as $R \to \infty$. Then, there exists $\delta > 0$ such that, we can always find arbitrarily large R obeying $\texttt{dist}\,(\bar{\boldsymbol{p}}_R, \mathcal{P}^{mm}) \geq \delta$.

Since $\bar{\boldsymbol{p}}_R$ is $\delta > 0$ bounded away from $\mathcal{P}^{mm}$, $\bar{\boldsymbol{p}}_R$ and $\|\bar{\boldsymbol{p}}_R\| = \|\boldsymbol{p}^{mm}\|$, $\bar{\boldsymbol{p}}_R$ strictly violates at least one of the inequality constraints in (ATT-SVM'). Otherwise, we would have $\bar{\boldsymbol{p}}_R \in \mathcal{P}^{mm}$. Without losing generality, suppose $\bar{\boldsymbol{p}}_R$ violates the first margin constraint, that is, for some $\gamma := \gamma(\delta) > 0$, $\max_{\alpha\in O_1} \min_{\beta\in\bar{O}_1} \boldsymbol{p}^\top(\boldsymbol{k}_{1\alpha} - \boldsymbol{k}_{1\beta}) \leq 1 - \gamma$. Now, we will argue that this will lead to a contradiction as $R \to \infty$ since we will show that $\mathcal{L}(\boldsymbol{p}_R^\star) < \mathcal{L}(\boldsymbol{p}_R)$ for sufficiently large R.

First, let us control $\mathcal{L}(\boldsymbol{p}_R^\star)$. We study $\boldsymbol{s}_i^\star = \mathbb{S}(\boldsymbol{K}_i\boldsymbol{p}_R^\star)$ and let $\alpha_i \in O_i$ be the index α in (ATT-SVM') for which $\text{margin}_i = \max_{\alpha\in O_i} \min_{\beta\in\bar{O}_i}(\boldsymbol{k}_{i\alpha} - \boldsymbol{k}_{i\beta})^\top\boldsymbol{p}^{mm} \geq 1$ is attained. Then, we bound the non-optimality amount $q_i^\star$ of $\boldsymbol{p}_R^\star$ as

$$q_i^\star = \frac{\sum_{t\in\bar{O}_i}\exp(\boldsymbol{k}_{it}^\top\boldsymbol{p}_R^\star)}{\sum_{t\in[T]}\exp(\boldsymbol{k}_{it}^\top\boldsymbol{p}_R^\star)} \leq \frac{\sum_{t\in\bar{O}_i}\exp(\boldsymbol{k}_{it}^\top\boldsymbol{p}_R^\star)}{\exp(\boldsymbol{k}_{i\alpha_i}^\top\boldsymbol{p}_R^\star)} \leq T\exp(-R/\Gamma).$$

Thus, $q_{\max}^\star = \max_{i\in[n]} q_i^\star \leq T\exp(-R/\Gamma)$. Secondly, we wish to control $\mathcal{L}(\boldsymbol{p}_R)$ by lower bounding the non-optimality in $\boldsymbol{p}_R$. Focusing on the first margin constraint, let $\alpha \in O_1$ be the index in (ATT-SVM') for which $\text{margin}_1 \leq 1 - \gamma$ is attained. Denoting the amount of non-optimality of the first input as $\hat{q}_1$, we find[3]

$$\hat{q}_1 = \frac{\sum_{t\in\bar{O}_1}\exp(\boldsymbol{k}_{1t}^\top\boldsymbol{p}_R)}{\sum_{t\in[T]}\exp(\boldsymbol{k}_{1t}^\top\boldsymbol{p}_R)} \geq \frac{1}{T}\frac{\sum_{t\in\bar{O}_1}\exp(\boldsymbol{k}_{1t}^\top\boldsymbol{p}_R)}{\exp(\boldsymbol{k}_{1\alpha}^\top\boldsymbol{p}_R)} \geq T^{-1}\exp(-R(1-\gamma)/\Gamma).$$

We similarly have $q_{\max}^\star \geq T^{-1}\exp(-R/\Gamma)$. In conclusion, for $\boldsymbol{p}_R, \boldsymbol{p}_R^\star$, denoting maximum non-optimality by $\hat{q}_{\max} \geq \hat{q}_1$ and $q_{\max}^\star$, we respectively obtained

$$\log(\hat{q}_{\max}) \geq -(1-\gamma)(R/\Gamma) - \log T, \tag{109}$$

$$-(R/\Gamma) - \log T \leq \log(q_{\max}^\star) \leq -(R/\Gamma) + \log T. \tag{110}$$

The above inequalities satisfy Assumption D as follows where $\boldsymbol{p} \leftarrow \boldsymbol{p}_R^\star$ and $\boldsymbol{p}' \leftarrow \boldsymbol{p}_R$: Set $R_0 = 3\gamma^{-1}\Gamma\log T$ so that $\log T \leq \frac{\gamma R_0}{3\Gamma}$. Secondly, set $\rho_0 = -(R_0/\Gamma) - \log T$. This way, $\rho_0 \geq \log(q_{\max}^\star)$ implies $R \geq R_0$ and $\log T \leq \frac{\gamma R}{3\Gamma}$. Using the latter inequality, we bound the $\log T$ terms to obtain

- $\log(\hat{q}_{\max}) \geq -(1 - 2\gamma/3)(R/\Gamma)$.
- $\log(q_{\max}^\star) \leq -(1 - \gamma/3)(R/\Gamma)$.

To proceed, we pick $1 + \Delta = \frac{1-\gamma/3}{1-2\gamma/3}$ implying $\Delta := \frac{\gamma}{3-2\gamma}$. Finally, for this Δ, there exists $\rho(\Delta)$ which we need to ensure $\log(\hat{q}_{\max}) \leq \rho(\Delta)$. This can be guaranteed by picking sufficiently large R that ensures $\log(q_{\max}^\star) \leq -(1 - \gamma/3)(R/\Gamma) \leq \rho(\Delta)$ to satisfy all conditions of Assumption D. Since such large R exists by initial assumption $\texttt{dist}\,(\bar{\boldsymbol{p}}_R, \mathcal{P}^{mm}) \nrightarrow 0$, Assumption D in turn implies that $\mathcal{L}(\boldsymbol{p}_R^\star) < \mathcal{L}(\boldsymbol{p}_R)$ contradicting with the optimality of $\boldsymbol{p}_R$ in (8). ■

D.2 Application to Linearly-mixed Labels

The following example shows that if non-optimal tokens result in reduced score (in terms of the alignment of prediction and label), Assumption D holds. The high-level idea behind this lemma is that, if the optimal risk is achieved by setting $q_{\max}^{\boldsymbol{p}} = 0$, then, Assumption D will hold.

Lemma 9 (Linear label mixing) *Recall $q_i^{\boldsymbol{p}} = \sum_{t\in\bar{O}_i} \boldsymbol{s}_{it}^{\boldsymbol{p}}$ from Assumption D. Suppose $Y_i \in \{-1, 1\}$ and*

$$Y_i \cdot \psi(\boldsymbol{X}_i^\top\boldsymbol{s}_i^{\boldsymbol{p}}) = \nu_i(1 - q_i^{\boldsymbol{p}}) + Z_i,$$

[3] Here, we assumed margin is non-negative i.e. $\boldsymbol{k}_{1\alpha}^\top\boldsymbol{p}_R \geq \sup_{t\in\bar{O}_1}\boldsymbol{k}_{1t}^\top\boldsymbol{p}_R$. Otherwise, $\sup_{t\in[T]}\boldsymbol{k}_{1t}^\top\boldsymbol{p}_R$ is attained in $\bar{O}_1$ which implies $\hat{q}_1 \geq T^{-1}$. Thus, we can still use the identical inequality (109) with the choice $\gamma = 1$.

for some $(\nu_i)_{i=1}^n > 0$. *Here* $Z_i = Z_i(\boldsymbol{p})$ *is the contribution of non-optimal tokens to prediction. For some* $C, \varepsilon > 0$ *and for all* $\boldsymbol{p} \in \mathbb{R}^d$, *assume*

$$-C q_{\max}^{\boldsymbol{p}} \le Z_i \le (1-\varepsilon)\nu_i q_i^{\boldsymbol{p}}. \tag{111}$$

Then, Assumption D holds for $\mathcal{L}(\boldsymbol{p}) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i \cdot \psi(\boldsymbol{X}_i^\top \boldsymbol{s}_i^{\boldsymbol{p}}))$ *when* $\ell(\cdot)$ *is a strictly decreasing loss function with continuous derivative.*

Proof. Recall the assumption $Y_i \cdot \psi(\boldsymbol{X}_i^\top \boldsymbol{s}_i^{\boldsymbol{p}}) = \nu_i(1 - q_i^{\boldsymbol{p}}) + Z_i$ with Z_i obeying (111). Let us also write the loss function

$$\mathcal{L}(\boldsymbol{p}) = \frac{1}{n}\sum_{i=1}^n \ell(\nu_i(1 - s_i^{\boldsymbol{p}}) + Z_i).$$

Define $q_{\max}^{\boldsymbol{p}} = \sup_{i\le[n]} q_i^{\boldsymbol{p}}$. Let M be the maximum absolute value of score over tokens. Let $B = \max_{|x|\le M} -\ell'(x) \ge A = \min_{|x|\le M} -\ell'(x) > 0$. Through Taylor's Theorem (integral remainder), we have that

$$B(q_i^{\boldsymbol{p}}\nu_i - Z_i) \ge \ell(\nu_i(1-q_i^{\boldsymbol{p}}) + Z_i) - \ell(\nu_i) \ge A(q_i^{\boldsymbol{p}}\nu_i - Z_i) \ge \varepsilon A \nu_i q_i^{\boldsymbol{p}}.$$

Set $\mathcal{L}_\star = \frac{1}{n}\sum_{i=1}^n \ell(\nu_i)$. Set $C_+ = B(C + \max_{i\in[n]}\nu_i)$ and $C_- = n^{-1}A\varepsilon \min_{i\in[n]}\nu_i$. This also implies

$$C_+ q_{\max}^{\boldsymbol{p}} \ge \frac{1}{n}\sum_{i\in[n]} B(q_i^{\boldsymbol{p}}\nu_i - Z_i) \ge \mathcal{L}(\boldsymbol{p}) - \mathcal{L}_\star \ge \frac{1}{n}\sum_{i\in[n]} A(q_i^{\boldsymbol{p}}\nu_i - Z_i) \ge \frac{1}{n}\sum_{i\in[n]} \varepsilon A \nu_i q_i^{\boldsymbol{p}} \ge C_- q_{\max}^{\boldsymbol{p}}.$$

Thus, to prove $\mathcal{L}(\boldsymbol{p}') > \mathcal{L}(\boldsymbol{p})$, we simply need to establish the stronger statement $C_- q_{\max}^{\boldsymbol{p}'} > C_+ q_{\max}^{\boldsymbol{p}}$.

Going back to the condition of Assumption D, any $\log(q_{\max}^{\boldsymbol{p}}) \le (1+\Delta)\log(q_{\max}^{\boldsymbol{p}'})$ obeys $q_{\max}^{\boldsymbol{p}} \le (q_{\max}^{\boldsymbol{p}'})^{1+\Delta}$ i.e. $q_{\max}^{\boldsymbol{p}'} \ge (q_{\max}^{\boldsymbol{p}})^{(1+\Delta)^{-1}}$. Following above, we wish to ensure $q_{\max}^{\boldsymbol{p}'} > \Theta q_{\max}^{\boldsymbol{p}}$ for such $(\boldsymbol{p}, \boldsymbol{p}')$ pairs where $\Theta = C_+/C_- > 1$. This is guaranteed by

$$(q_{\max}^{\boldsymbol{p}})^{(1+\Delta)^{-1}-1} > \Theta \iff \frac{\Delta}{1+\Delta}\log(q_{\max}^{\boldsymbol{p}}) < -\log(\Theta).$$

The above is satisfied by choosing a $\rho(\Delta) := -2(1+\Delta^{-1})\log(\Theta)$ in Assumption D. Thus, all $\boldsymbol{p}, \boldsymbol{p}'$ with $\log(q_{\max}^{\boldsymbol{p}}) \le \rho = \rho(\Delta)$ satisfies the condition of Assumption D finishing the proof. ∎

E Experimental Details

In this section, we provide additional implementation details for the experiments.

1. We build one attention layer using `PyTorch`, and set input and embedding dimensions to be 3. During training, we use `SGD` optimizer with learning rate 1 in Figure 1 and 0.1 in Figure 2 and train the model for 1000 iterations. To better visualize the generalization path, we normalize the gradient of $\boldsymbol{p}$ (and $\boldsymbol{v}$) at each iteration.
2. Next, given the solution $\hat{\boldsymbol{p}}$, we determine locally-optimal indices to be those with the highest softmax scores. Using these optimal indices, we utilize python package `cvxopt` to build and solve (ATT-SVM), and then get solution $\boldsymbol{p}^{mm}$. After obtaining $\boldsymbol{p}^{mm}$, we also verify that these indices satisfy our local-optimal definition. The examples we use in the paper are all trivial to verify (by construction).
3. In Fig. 2(b) and Fig. 2(c), $\boldsymbol{v}^{mm}$ is solved using python package `sklearn.svm` based on the given label information, and $\boldsymbol{p}^{mm}$ is the solution of (7) instead.

F Addendum to Section 5

We provide an overview of the current literature on implicit regularization and attention mechanism.

F.1 Related Work on Implicit Regularization

The introduction of Support Vector Machines (SVM), which utilize explicit regularization to choose maximum margin classifiers, represents one of the earliest relevant literature in this field [55]. The concept of maximizing the margin was later connected to generalization performance [56]. From a practical perspective, exponential losses with decaying regularization exhibit asymptotic behavior similar to SVMs, as demonstrated in [19]. While the analysis of the perceptron [57] originally introduced the concept of margins, the method itself does not possess an inherent bias as it terminates with zero classification error. However, establishing a meaningful lower bound for the attained margin is not possible. Initial empirical investigations highlighting the implicit bias of descent methods focused on ℓ_1-regularization, revealing that coordinate descent, when combined with the exponential loss, exhibits an inherent inclination towards ℓ_1-regularized solutions [58].

This work draws extensively from the literature on implicit bias and regularization, which has provided valuable techniques and inspiration. A common observation in these studies is the convergence to a specific optimal solution over the training set. This phenomenon has been observed in various approaches, including coordinate descent [59, 60], gradient descent [25, 19], deep linear networks [61, 62], ReLU networks [63, 64, 24, 65, 66, 67], mirror descent [20], and many others. The implicit bias of gradient descent in classification tasks involving separable data has been extensively examined by [19, 20, 21, 22, 23, 24]. These works typically utilize logistic loss or exponentially-tailed losses to establish connections to margin maximization. The results have also been extended to non-separable data by [25, 26, 27]. Furthermore, there have been notable investigations into the implicit bias in regression problems and losses, utilizing techniques such as mirror descent [28, 20, 29, 30, 31, 32]. Additionally, several papers have explored the implicit bias of stochastic gradient descent [33, 34, 35, 36, 37, 38], as well as adaptive and momentum-based methods [39, 40, 41, 42].

While there are some similarities between our optimization approach for $\boldsymbol{v}$ and existing works, the optimization of $\boldsymbol{p}$ presents notable differences. Firstly, our optimization problem is nonconvex and involves a composition of loss and softmax, which introduces new challenges and complexities. The presence of softmax adds a nonlinearity to the problem, requiring specialized techniques for analysis and optimization. Secondly, our analysis introduces the concept of locally-optimal tokens, which refers to tokens that achieve locally optimal solutions in their respective attention cones. This concept is crucial for understanding the behavior of the attention mechanism and its convergence properties. By focusing on the cones surrounding locally-optimal tokens, we provide a tailored analysis that captures the unique characteristics of the attention model. Overall, our work offers novel insights into the optimization of attention-based models and sheds light on the behavior of the attention mechanism during training.

F.2 Related Work on Attention Mechanism

As the backbone of Transformers [6], the self-attention mechanism [68] plays a crucial role in computing feature representations by globally modeling long-range interactions within the input. Transformers have achieved remarkable empirical success in various domains, including natural language processing [4, 2], recommendation systems [69, 70, 71], and reinforcement learning [72, 73, 74]. With the introduction of Vision Transformer (ViT) [75], Transformer-based models [76, 77, 78] have gradually replaced convolutional neural network (CNN) architectures and become prevalent in vision tasks. To train ViT efficiently, several techniques have been developed, among which token sparsification [79, 80, 81, 82, 83] remove redundant tokens (image patches) from the data, improving computational complexity while maintaining comparable learning performance.

However, the theoretical foundation of Transformers and self-attention mechanisms has remained largely unexplored. Some studies have established important results, including the Lipschitz constant of self-attention [84], properties of the neural tangent kernel [85, 86], and the expressive power and Turing-completeness of Transformers [87, 88, 89, 47, 51, 90, 91, 92] with statistical guarantees [93, 94].

Focusing on the self-attention component, Edelman et al. [47] theoretically proved that a single self-attention head can represent a sparse function of the input with a sample complexity for the generalization gap between the training loss and the test loss. However, they did not delve into the algorithmic aspects of training Transformers to achieve desirable loss. Sahiner et al. [48] and Ergen et al. [49] further explored the analysis of convex relaxations for self-attention, investigating

potential optimization techniques and properties. In terms of expressive ability, Baldi and Vershynin [50] investigated the capacity of attention layers to capture complex patterns and information, while Dong et al. [51] provided additional insights into the expressive power of attention layers in various contexts. Likhosherstov et al. [90] studied the model complexity for function approximation of the self-attention module, and Cordonnier et al. [91] provided sufficient and necessary conditions for multi-head self-attention structures to simulate convolution layers.

Recent works have made progress in characterizing the optimization and generalization dynamics of attention. Jelassi et al. [52] studied gradient-based methods from random initialization and provided a theoretical analysis of the empirical finding that Vision Transformers learn position embeddings that recapitulate the spatial structure of the training data, even though this spatial structure is no longer explicitly represented after the image is split into patches. Li et al. [53] provided theoretical results on training three-layer ViTs for classification tasks. They quantified the importance of self-attention in terms of sample complexity for achieving zero generalization error, as well as the sparsity of attention maps when trained by stochastic gradient descent (SGD). In another related work, Nguyen et al. [95] proposed a primal-dual optimization framework that focuses on deriving attention as the dual expansion of a primal neural network layer. By solving a support vector regression problem, they gained a deeper understanding and explanation of various attention mechanisms. This framework also enables the creation of novel attention mechanisms, offering flexibility and customization in designing attention-based models. In another closely related work, Oymak et al. [17] analyzed the same attention model as ours, denoted by (ERM). However, it is important to note that all of these works make certain assumptions about the data. Specifically, they assume that tokens are tightly clusterable or can be clearly split into relevant and irrelevant sets. Additionally, Li et al. [53] require specific assumptions on the initialization of the model, while Jelassi et al. [52] consider a simplified attention structure where the attention matrix is not directly parameterized with respect to the input.

In contrast, our work offers a comprehensive optimization-theoretic analysis of the attention model, establishing a formal connection to max-margin problems. This analysis allows us to gain a deeper understanding of the attention mechanism and its behavior during the training process. Notably, our work presents the first theoretical understanding of the implicit bias exhibited by gradient descent methods in the context of the attention model. By uncovering the underlying optimization principles, we provide valuable insights into the dynamics and generalization properties of attention-based models.