Implicit Bias and Fast Convergence Rates for Self-attention

Bhavya Vasudeva[*] *bvasudev@usc.edu*
University of Southern California

Puneesh Deora[*] *puneeshdeora@ece.ubc.ca*
University of British Columbia

Christos Thrampoulidis *cthrampo@ece.ubc.ca*
University of British Columbia

Reviewed on OpenReview: *https://openreview.net/forum?id=pKilnjQsb0*

Abstract

We study the fundamental optimization principles of self-attention, the defining mechanism of transformers, by analyzing the implicit bias of gradient-based optimizers in training a self-attention layer with a linear decoder in binary classification. Building on prior studies in linear logistic regression, recent findings demonstrate that the key-query matrix $\boldsymbol{W}_t$ from gradient-descent (GD) converges in direction towards $\boldsymbol{W}_{\text{mm}}$, which maximizes the margin between optimal and non-optimal tokens across sequences. However, this convergence is local, dependent on initial conditions, only holds asymptotically as the number of iterations increases, and leaves questions about the potential benefits of adaptive step-size rules unaddressed. To bridge this gap, we first establish scenarios for which convergence is provably *global*. We then analyze two adaptive step-size strategies: normalized GD and Polyak step-size, demonstrating *finite-time* convergence rates for $\boldsymbol{W}_t$ to $\boldsymbol{W}_{\text{mm}}$, and quantifying the sparsification rate of the attention map. These findings not only show that these strategies can accelerate parameter convergence over standard GD in a non-convex setting but also deepen the understanding of the implicit bias in self-attention, linking it more closely to the phenomena observed in linear logistic regression despite its intricate non-convex nature.

1 Introduction

Self-attention serves as the fundamental building block of transformers, distinguishing them from traditional neural networks (Vaswani et al., 2017) and driving their outstanding performance across various applications, including natural language processing and generation (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020), as well as computer vision (Dosovitskiy et al., 2021; Radford et al., 2021; Touvron et al., 2021). With transformers establishing themselves as the de-facto deep-learning architecture, driving advancements in applications seamlessly integrated into society's daily life at an unprecedented pace (OpenAI, 2022), there has been a surge of recent interest in the mathematical study of the fundamental optimization and statistical principles of the self-attention mechanism; see Section 6 on related work for an overview.

In pursuit of this objective, Tarzanagh et al. (2023b;a) have initiated an investigation into the *implicit bias* of gradient descent (GD) in training a self-attention layer with fixed linear decoder in a binary classification task. Concretely, the study paradigm of implicit bias seeks to characterize structural properties of the weights learned by GD when the training objective has multiple solutions. The prototypical instance of this paradigm is GD training of linear logistic regression on separable data: among infinitely many possible solutions to logistic-loss minimization (each linear separator defines one such solution), GD learns weights that converge in direction to the (unique) max-margin class separator (Soudry et al., 2018; Ji & Telgarsky, 2018). Notably, convergence is global, holding irrespective of the initial weights' direction, and comes with explicit rates that

[*] Equal Contribution.

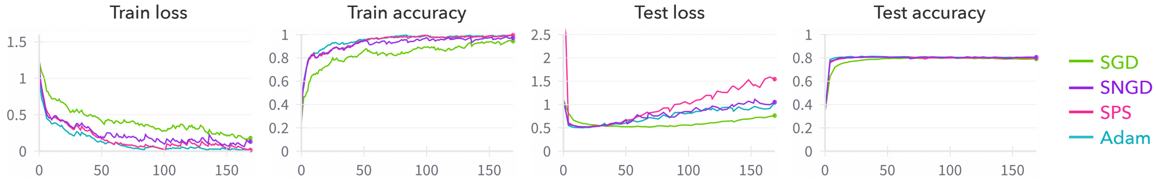

Figure 1: Comparison of train and test dynamics of various optimizers—SGD, stochastic normalized GD (SNGD), stochastic Polyak step (SPS), and Adam—while fine-tuning a pre-trained BERT model on the MNLI dataset; see App. C for details. SNGD and SPS, employing adaptive step-size rules, demonstrate significantly faster training, closely resembling the performance of Adam. Motivated in part by this observation, our work establishes fast convergence rates for NGD and PS for single-layer self-attention.

characterize its speed with respect to the number of iterations. Drawing an analogy to this prototypical instance, when training self-attention with linear decoder in a binary classification task, Tarzanagh et al. (2023a) defines a hard-margin SVM problem (W-SVM) that separates, with maximal margin, *optimal* input tokens from non-optimal ones based on their respective softmax logits. For this, they show that the key-query weights $\boldsymbol{W}_t$ found by GD converge locally to the solution $\boldsymbol{W}_{\text{mm}}$ of (W-SVM) as the number of iterations t grows to infinity.

Despite the intriguing analogy between the two settings, the findings of Tarzanagh et al. (2023a) are highly non-trivial not only because the nature of the max-margin solution differs, but also because of the intricate non-convex optimization landscape introduced by the presence of self-attention. The non-convexity, induced by the softmax operator in the self-attention layer, complicates the analysis and is the reason why the convergence result of Tarzanagh et al. (2023a) is: (i) *local*, holding only for an appropriate initialization direction, and (ii) *asymptotic*, applicable only as iterations t approach infinity.

Identifying these limitations, this work is motivated by the following questions:

Q1: *Are there settings under which GD iterates converge* globally *to the solution $\boldsymbol{W}_{mm}$ of* (W-SVM)*?*

Q2: *Is it possible to obtain* finite-time *rates of convergence to $\boldsymbol{W}_{mm}$?*

Additionally, motivated by the practical benefit of using adaptive learning rates in transformer optimization (see for example Fig. 1), we pose an additional open question:

Q3: *Can using* adaptive learning rates *in self-attention optimization accelerate the convergence to $\boldsymbol{W}_{mm}$?*

Contributions. Our work addresses the above open questions, thus contributing fundamental insights on the optimization properties of the self-attention mechanism.

Concretely, we study a single-layer self-attention model $\Phi(\boldsymbol{X},\boldsymbol{\theta}) = \boldsymbol{u}^\top \boldsymbol{X}^\top \varphi(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_1)$, where, $\varphi(\cdot)$ is the softmax nonlinearity, $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{\mathrm{T}}]$ is the sequence of input tokens, $\boldsymbol{W}$ is the key-query matrix, and $\boldsymbol{u}$ is a linear decoder. Following the setup of Tarzanagh et al. (2023a), for each sequence $\boldsymbol{X}$, we associate each token $\boldsymbol{x}_\tau$ with a score $\gamma_\tau := y\boldsymbol{u}^\top \boldsymbol{x}_\tau$ and let $\texttt{opt} \in [\mathrm{T}]$ be the index of the token with the largest score. Given a training set of n sequences $\boldsymbol{X}_i$, this defines a hard-margin SVM problem with solution $\boldsymbol{W}_{\text{mm}}$ that separates the optimal tokens with maximum margin; see Section 2 for details.

Motivated by question Q3, we study here the optimization properties of training $\Phi(\boldsymbol{X};\boldsymbol{\theta})$ with exponential loss using *normalized GD*, which sets the learning rate η_t at iteration t adaptively as $\eta_t = \frac{\eta}{\|\nabla_{\boldsymbol{\theta}} \widehat{\mathcal{L}}(\boldsymbol{\theta}_t)\|}$, for some constant $\eta > 0$. Our results also extend to Polyak step-size, another adaptive step-size rule.

Our first set of results (Section 3), for fixed decoder $\boldsymbol{u}$ (similar to Tarzanagh et al. (2023a)), answer questions Q1-Q3 as follows. In response to Q1, we begin by identifying sufficient conditions on the input data under which GD converges *globally*, in direction, to $\boldsymbol{W}_{\text{mm}}$. Then, simultaneously addressing Q2 and Q3, we establish fast finite-time rates of convergence for normalized GD by proving that the iterates $\boldsymbol{W}_t$, at any time t, satisfy $\left\| \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|} - \frac{\boldsymbol{W}_{\text{mm}}}{\|\boldsymbol{W}_{\text{mm}}\|} \right\| \leq \mathcal{O}\left(t^{-1/2}\right)$. We identify two key ingredients towards establishing these results. First, we show that the Euclidean norm $\|\boldsymbol{W}_t\|$ of the iterates grows at a rate $\Theta(t)$. Second, and more intricate, we demonstrate that even if the iterate at any time t violates the constraints of (W-SVM) corresponding to any training sample, the softmax score of the optimal token for that sample must be non-decreasing. In turn, this establishes that the iterates, initialized in any direction, remain inside a cone around $\boldsymbol{W}_{\text{mm}}$. Our convergence results also imply an explicit rate at which softmax-attention gets sparsified as softmax scores of optimal tokens converge to 1 at an exponential rate $\mathcal{O}(\exp(-\eta t))$.

Our second set of results (Section 4), raises the limitation of fixed linear decoder of prior work and applies to joint training of $\boldsymbol{u}$ and of the attention weights $\boldsymbol{W}$. In response to Q1, we construct a representative data model with Gaussian-distributed tokens and prove that GD converges *globally*, in direction, to $\boldsymbol{W}_{\text{mm}}$. To address Q2, we show that for normalized GD with an aggressive step-size, the iterates $\boldsymbol{W}_t$, at any time t, satisfy $\left\|\frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|} - \frac{\boldsymbol{W}_{\text{mm}}}{\|\boldsymbol{W}_{\text{mm}}\|}\right\| \leq \mathcal{O}\left(1/\log t\right)$.[*] Further, we prove that the linear decoder $\boldsymbol{u}$ converges, in direction, to $\boldsymbol{u}_{\text{mm}}$, the solution of the hard-margin SVM problem (u-SVM) that separates the examples with maximal margin, using only the *optimal* input tokens at a $\mathcal{O}(t^{-\eta})$ rate. Finally, to completely characterize the training dynamics, we show fast train loss convergence at a $\mathcal{O}(\exp(-t^{1/3}))$ rate. We highlight three key technical contributions towards proving these results: (i) a growing token score gap $\gamma_{\texttt{opt},t} - \gamma_{\tau,t} > 0$, between optimal $\texttt{opt}$ and non-optimal $\tau \neq \texttt{opt}$ tokens, for any time $t > 0$; (ii) non-decreasing softmax scores of the optimal tokens with time; and (iii) a constant loss ratio across all training sequences valid for any time. These properties pave way for a PL-like inequality which is crucial to show the loss convergence rates.

Throughout, we validate our findings on both synthetic and real datasets (see Section 5). We end with Section 7 discussing the implications of our results and the open questions they raise.

2 Preliminaries

Lowercase and uppercase bold letters represent vectors and matrices, respectively. $\|\boldsymbol{a}\|$ and $\|\boldsymbol{A}\|$ denote Euclidean norms. $\bar{\boldsymbol{a}}$ and $\bar{\boldsymbol{A}}$ denote $\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}$ and $\frac{\boldsymbol{A}}{\|\boldsymbol{A}\|}$, respectively. $[\boldsymbol{a}_1, \boldsymbol{a}_2] = \{\boldsymbol{a} : \boldsymbol{a} = \beta \boldsymbol{a}_1 + (1-\beta)\boldsymbol{a}_2, \beta \in [0,1]\}$ denotes the line segment between $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, and concat$(\cdot,\cdot)$ denotes the concatenation operation. $a \wedge b$ denotes the minimum of numbers a and b. $a \vee b$ denotes their maximum. We use standard notation $\mathcal{O}$, Ω to hide absolute constants, and $\widetilde{\mathcal{O}}$, $\widetilde{\Omega}$ to hide poly-logarithmic factors. All logarithms are natural logarithms.

The output of a single-head self-attention layer, parameterized by key, query and value matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K \in \mathbb{R}^{d \times d_1}$, $\boldsymbol{W}_V \in \mathbb{R}^{d \times d_2}$, is given by $\boldsymbol{\varphi}(\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{W}_V$, where $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{\mathrm{T}}] \in \mathbb{R}^{\mathrm{T} \times d}$ is the input sequence, and the softmax map $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^{\mathrm{T}} \to \mathbb{R}^{\mathrm{T}}$ is applied row-wise. We can compose the output of the attention layer with a linear projection head to obtain the final prediction as

$$\Phi(\boldsymbol{X};\boldsymbol{\theta}) := \boldsymbol{u}^\top \boldsymbol{X}^\top \boldsymbol{\varphi}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_1), \tag{1}$$

where $\boldsymbol{\theta} := \text{concat}(\boldsymbol{u}, \boldsymbol{W})$ denotes the set of trainable parameters. Here, similar to Tarzanagh et al. (2023a); Deora et al. (2023); Tian et al. (2023a); Oymak et al. (2023) we reparameterize the key-query matrix as $\boldsymbol{W} := \boldsymbol{W}_Q\boldsymbol{W}_K^\top \in \mathbb{R}^{d \times d}$, use the first token for prediction[†] and subsume the value weights $\boldsymbol{W}_V$ within the prediction head $\boldsymbol{u} \in \mathbb{R}^d$. Let $\boldsymbol{a}(\boldsymbol{W}) := \boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_1$ denote the vector of softmax logits, then $\boldsymbol{\varphi}'(\boldsymbol{a}(\boldsymbol{W})) \in \mathbb{R}^{\mathrm{T} \times \mathrm{T}}$ denotes the Jacobian at $\boldsymbol{a}(\boldsymbol{W})$.

Given training data $\{(\boldsymbol{X}_i, y_i)\}_{i=1}^n$ with labels $y_i \in \{\pm 1\}$ and $\boldsymbol{X}_i := [\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \ldots, \boldsymbol{x}_{i,\mathrm{T}}]$, and decreasing loss $\ell : \mathbb{R} \to \mathbb{R}_+$, the empirical risk is $\widehat{L}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i \in [n]} \ell(y_i \Phi(\boldsymbol{X}_i;\boldsymbol{\theta}))$. We focus on GD optimization with adaptive time-dependent step-size η_t. Concretely, we study two variants: normalized gradient descent (NGD) (Hazan et al., 2015; Nacson et al., 2019), and Polyak step-size (PS) (Loizou et al., 2021), which set

$$\eta_t \propto 1/\|\nabla_{\boldsymbol{\theta}}\widehat{L}(\boldsymbol{\theta}_t)\|, \quad \text{and} \quad \eta_t = \big(\widehat{L}(\boldsymbol{\theta}_t) - \widehat{L}^*\big)/\big(2\|\nabla_{\boldsymbol{\theta}}\widehat{L}(\boldsymbol{\theta}_t)\|^2\big) \text{ with } \widehat{L}^* = \min_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}),$$

respectively. Our results are applicable to both update rules. For specificity, we present them for NGD and include remarks for PS.

[*] Note the rate here is slower compared to the fixed-decoder case. As detailed in Section 4, this is due to the additional $(t+1)^{-1}$ factor in the step-size, which results in a slow down of the rate of growth of $\|\boldsymbol{W}_t\|$. Our proof requires this additional factor to account for the non-smooth objective.

[†] This is without loss of generality as our results hold for any token $\tau \in [\mathrm{T}]$.

We follow Tarzanagh et al. (2023b;a) in defining token scores as follows.

Definition 1 (Token scores and Optimality). *Given a fixed prediction head $\boldsymbol{u}_* \in \mathbb{R}^d$, the token score vector for a sample $(y, \boldsymbol{X})$ is given by $\boldsymbol{\gamma} = y\boldsymbol{X}\boldsymbol{u}_*$. The optimal token index*[‡] *is $\boldsymbol{opt} = \operatorname{argmax}_{\tau\in[\mathrm{T}]} \gamma_\tau$, where $\gamma_\tau = y\boldsymbol{x}_\tau^\top \boldsymbol{u}_*$ denotes the token score for the token $\boldsymbol{x}_\tau$.*

Similarly, $\texttt{opt}_i$ denotes the optimal token index for a sample $(y_i, \boldsymbol{X}_i)$, $i \in [n]$. Intuitively, these are the tokens that minimize the training loss upon selection (Lemma 2 in Tarzanagh et al. (2023a)). Given a set of optimal token indices $\texttt{OPT} := \{\texttt{opt}_i\}_{i=1}^n$, define the following hard-margin SVM problem, which separates, with maximal margin, optimal tokens from the other tokens for every input sequence:

$$\boldsymbol{W}_{\mathrm{mm}} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \, \|\boldsymbol{W}\| \quad \text{s.t. } (\boldsymbol{x}_{i,\texttt{opt}_i} - \boldsymbol{x}_{i,\tau})^\top \boldsymbol{W} \boldsymbol{x}_{i,1} \geq 1 \;\; \forall i \in [n], \tau \in [\mathrm{T}] \setminus \{\texttt{opt}_i\}. \tag{W-SVM}$$

Throughout, we assume that (W-SVM) is feasible,[§] *i.e.* softmax logits $\boldsymbol{x}_{i,\texttt{opt}_i}^\top \boldsymbol{W} \boldsymbol{x}_{i,1}$ of optimal tokens can be separated from logits $\boldsymbol{x}_{i,\tau}^\top \boldsymbol{W}\boldsymbol{x}_{i,1}$ of the other tokens $\tau \neq \texttt{opt}_i$.

3 Training Dynamics of $\boldsymbol{W}$

Here, let fixed prediction head $\boldsymbol{u} = \boldsymbol{u}_*$. This allows focusing first on the dynamics of token selection induced by training key-query weight matrix $\boldsymbol{W}$, which is the only trainable parameter:

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \frac{\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t)}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t)\|}, \, t > 0. \tag{2}$$

We start by setting up (mild) assumptions on the data and initialization, and then present our main results.

3.1 Setup

For convenience, first define $\Lambda := \|\boldsymbol{W}_{\mathrm{mm}}\|$, $B := \max_{i,\tau} \|\boldsymbol{x}_{i,\tau}\|$, $\gamma_i := \min_{\tau\neq\texttt{opt}_i} \gamma_{i,\tau}$, $\kappa_+ := \max_i \exp(\gamma_{i,\texttt{opt}_i} - \gamma_i)$, $\kappa_- := \min_i \exp(\gamma_{i,\texttt{opt}_i} - \gamma_i)$, and $\Upsilon = \kappa_+ \cdot \frac{\log(\kappa_+)}{\log(\kappa_-)}$. Note $1/\Lambda$ is the margin achieved by $\boldsymbol{W}_{\mathrm{mm}}$ separating optimal tokens from the rest, and B is a uniform bound on the tokens' magnitudes. The parameters $\kappa_\pm$ represent the largest and smallest degradation factors across sequences for each sequence's individual loss term when suboptimal tokens are selected. Respectively, Υ can be interpreted as a conditioning parameter for the problem, measuring the variability in token gaps across different sequences.

Our first technical assumption towards ensuring global convergence requires that tokens are nearly orthogonal.[¶] This is often the case in high-dimensional settings; see Example 1 and the references below.

Assumption 1 (Nearly-orthogonal Tokens). *For any $i, j, k \in [n]$, and for any $\tau, \tau' \in [\mathrm{T}]$,*

$$\|\boldsymbol{x}_{i,\tau}\|^2 \geq 4n\Upsilon\mathrm{T}|\langle \boldsymbol{x}_{i,\tau}, \boldsymbol{x}_{j,\tau'}\rangle|, \quad j \neq i, \tau' \neq \boldsymbol{opt}_j,$$
$$\langle \boldsymbol{x}_{i,\boldsymbol{opt}_i}, \boldsymbol{x}_{j,\boldsymbol{opt}_j}\rangle \geq 4n\Upsilon\mathrm{T}|\langle \boldsymbol{x}_{i,\tau}, \boldsymbol{x}_{k,\tau'}\rangle|, \quad y_i = y_j \neq y_k.$$

We will use Ass. 1 to prove that softmax scores of optimal tokens are lower bounded by a constant throughout the optimization trajectory. Note that Ass. 1 itself makes *no* further guarantees on softmax scores of optimal tokens approaching 1 or even increasing during training, which is essential for global convergence and we prove separately. Furthermore, we impose no assumptions on the direction of initialization $\boldsymbol{W}_0$. In particular, our main results hold even when $\boldsymbol{W}_0$ is aligned with "bad" stationary directions $\overline{\boldsymbol{W}}$ that could asymptotically saturate non-optimal tokens, i.e. $\nabla_{\boldsymbol{W}} \widehat{L}(\alpha\overline{\boldsymbol{W}}) \to \boldsymbol{0}$ and $\varphi_{i,\texttt{opt}_i}(\alpha\overline{\boldsymbol{W}}) \to 0$ as $\alpha \to \infty$. Thus, the only mild requirement on the initialization $\boldsymbol{W}_0$ regards its scale, rather than its direction, and is formalized in Lem. 2 in the App. See Fig. 8 in the App. for numerical validation of global convergence despite initializing in a "bad" stationary direction. The following example illustrates the above assumption.

[‡] Similar to Tarzanagh et al. (2023a), we assume unique optimal token which holds for almost all datasets.

[§] This is analogous to linear separability assumptions in linear models, which are fundamental to understand the implicit bias in GD-based methods. For linear models, feasibility of hard-margin SVM is guaranteed when the ambient dimension exceeds the number of samples. Similarly, for our self-attention model, (W-SVM) feasibility is guaranteed when the ambient dimension exceeds both the number of samples and the context window size (as shown in Theorem 1 of Tarzanagh et al. (2023a).

[¶] Our results also extend to cases where $\texttt{OPT}$ tokens are antipodal and the remaining tokens are nearly orthogonal.

Example 1. *Let $\boldsymbol{\mu}_\pm \in \mathbb{R}^d$, $\boldsymbol{\mu}_+ \perp \boldsymbol{\mu}_-$, $\|\boldsymbol{\mu}_+\| = \|\boldsymbol{\mu}_-\| = U$, and data generated as follows:*

$$y \sim \text{Unif}(\{\pm1\}), \quad \boldsymbol{opt} \sim \text{Unif}([\mathrm{T}]), \quad \boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}),$$
$$\boldsymbol{x}_{opt} = \boldsymbol{\nu} + \begin{cases}\boldsymbol{\mu}_+ & , y = 1\\ \boldsymbol{\mu}_- & , y = -1\end{cases}, \quad \boldsymbol{x}_\tau \sim \mathcal{N}(\mathbf{0}, \rho^2\boldsymbol{\Sigma}),\ \forall\tau \in [\mathrm{T}] \setminus \boldsymbol{opt},$$

where $\boldsymbol{\Sigma} := \boldsymbol{I}_d - U^{-2}\boldsymbol{\mu}_+\boldsymbol{\mu}_+^\top - U^{-2}\boldsymbol{\mu}_-\boldsymbol{\mu}_-^\top$. It is easy to show that when $d = \widetilde{\Omega}(n^2\mathrm{T}^2)$, the tokens are nearly orthogonal with high probability. Further let high enough signal-to-noise ratio $U^2 \geq \widetilde{\Omega}((\sigma \vee \rho)^2 n\mathrm{T}\sqrt{d})$ and appropriate $\boldsymbol{u}_*$ such that the token scores satisfy $\Upsilon = \mathcal{O}(1)$, then, the data model satisfies Ass. 1 (see Lem. 6 in the App. for details). We remark that similar data models have been considered in prior works on the analysis of linear overparameterized models (Muthukumar et al., 2019; Chatterji & Long, 2021; Wang & Thrampoulidis, 2022; Cao & Gu, 2019; Wang et al., 2021), NNs with Leaky ReLU activation (Frei et al., 2022a; 2023), CNNs (Cao et al., 2022; Kou et al., 2023a), and self-attention models (Deora et al., 2023; Li et al., 2023a).

Our second technical assumption is similar to Tarzanagh et al. (2023a) with two key distinctions: Firstly, it applies to self-attention rather than their simplified attention model. Secondly, and most importantly, under this same assumption, we will prove a stronger results for global convergence and finite-time rates.

Assumption 2. *For any $i \in [n]$, and any $\tau \neq \boldsymbol{opt}_i \in [\mathrm{T}]$, the token scores satisfy $\gamma_{i,\tau} = \gamma_i$.*

Note that in Ex. 1, Ass. 2 is satisfied by choosing an appropriate $\boldsymbol{u}_*$ (see Lem. 6 in the App. for details). It has been recently shown that the optimization landscape of self-attention can lead GD to converge to local directions that are different from $\boldsymbol{W}_{\text{mm}}$ (Tarzanagh et al., 2023a). Thus, global convergence necessitates additional conditions. Assumptions 1 and 2, as outlined above, serve this purpose. While being only sufficient conditions, they lead to the first-known global convergence result with finite-time guarantees, contrasting with the local and asymptotic convergence in Tarzanagh et al. (2023a).

3.2 Main Results

We now present our main results. For ease of exposition, we use exponential loss, but our results directly extend to continuously differentiable decreasing loss functions. First, we establish the rate of NGD's directional convergence to $\overline{\boldsymbol{W}}_{\text{mm}}$. A proof sketch is given in Sec. 3.4.

Theorem 1 (IB rate)**.** *Under small initialization scale (Lem. 2 in the App.) and Ass. 1-2, using the NGD updates in Eq. (2) with $\eta = \widetilde{\mathcal{O}}(B^{-2})$, it holds for any $t \geq t_0 = \text{poly}(\eta, B, \Lambda, \mathrm{T}, n\Upsilon)$,*

$$\left\langle \overline{\boldsymbol{W}}_t, \overline{\boldsymbol{W}}_{mm} \right\rangle \geq 1 - C\,\frac{(\log t)^2}{t},$$

where $C := C(\eta, B, \Lambda, t_0) = \text{poly}(\eta, B, \Lambda, t_0)$. In particular, assuming $B = \mathcal{O}(1), T = \mathcal{O}(1)$ and $\Upsilon = \mathcal{O}(1)$, we have $\Lambda = \mathcal{O}(1)$. Thus, for any $t \geq t_0 = \Omega(n)$, it holds $\left\|\overline{\boldsymbol{W}}_t - \overline{\boldsymbol{W}}_{mm}\right\| \leq \tilde{\mathcal{O}}\left(t^{-1/2}\right)$.

Thm. 1 establishes that normalized iterates $\overline{\boldsymbol{W}}_t$ of NGD converge *globally* to $\overline{\boldsymbol{W}}_{\text{mm}}$, starting from *any* initialization direction. It also provides a lower bound on the rate of $\overline{\boldsymbol{W}}_t$ approaching $\overline{\boldsymbol{W}}_{\text{mm}}$. In Sec. 3.3, we establish an upper bound for the rate of convergence of GD, demonstrating that NGD's adaptive step-size provably accelerates convergence. This is the first global convergence result and finite-time convergence rate for self-attention using either GD or NGD. To the best of our knowledge, this is also the first demonstration of directional convergence of NGD and its superiority over GD in a non-convex setting (Sec. 3.3 for comparisons with existing results in convex linear settings).

We now discuss an implication of Thm. 1 concerning the evolution of the attention map during NGD training. Specifically, we demonstrate that as training progresses, the attention map increasingly becomes sparse, selecting optimal tokens at an exponential rate.

Lemma 1 (Softmax score rate)**.** *Under the setting of Thm. 1, it holds for any $i \in [n]$ and any $t \geq \text{poly}(\eta, B, \Lambda, t_0)$ that $\varphi^t_{i,opt_i} \geq \left(1 + (T-1)e^{-\eta(8B^2\Lambda^2)^{-1}t}\right)^{-1}$.*

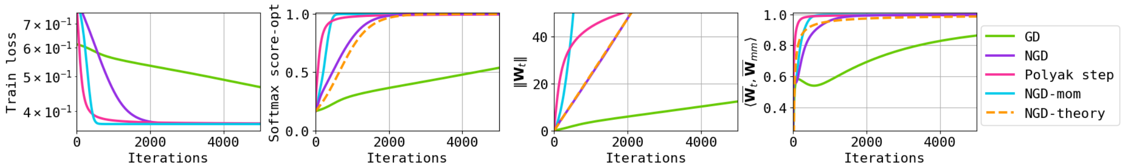

Figure 2: Training dynamics of a single-head self-attention model (Eq. (1)) when optimizing only $\boldsymbol{W}$ on synthetic data with nearly orthogonal tokens (Example 1 with $\sigma = 0$). The observed softmax score saturation, norm growth of $\boldsymbol{W}$ and directional alignment with $\boldsymbol{W}_{\mathrm{mm}}$ closely match with our theoretical results (Lemma 1, Eq. (9) and Theorem 1, respectively).

To understand why the attention map becomes sparser during NGD training, recall from Thm. 1 that for sufficiently large iterations, $\boldsymbol{W}_t \approx \frac{\|\boldsymbol{W}_t\|}{\|\boldsymbol{W}_{\mathrm{mm}}\|}\boldsymbol{W}_{\mathrm{mm}}$, and that $\boldsymbol{W}_{\mathrm{mm}}$ separates optimal tokens from others. In proving Thm. 1, we also show a norm-growth rate $\|\boldsymbol{W}_t\| = \Theta(t)$. Combining these yields that softmax scores of optimal tokens $\varphi^t_{i,\texttt{opt}_i}$ approach 1, while the rest approach 0. Concretely, given $\boldsymbol{W}_t \approx \frac{\|\boldsymbol{W}_t\|}{\|\boldsymbol{W}_{\mathrm{mm}}\|}\boldsymbol{W}_{\mathrm{mm}}$, the $\texttt{opt}$ softmax score $\varphi^t_{i,\texttt{opt}_i} \geq \frac{1}{1+\sum_{\tau\neq\texttt{opt}_i}\exp(-\frac{\|\boldsymbol{W}_t\|}{\|\boldsymbol{W}_{\mathrm{mm}}\|})}$ approaches 1 as $\|\boldsymbol{W}_t\|$ grows. Lem. 1 formalizes this intuition and pairs it with an explicit sparsification rate.

Finally, we validate the predictions of the above results using synthetic data generated as per Ex. 1 with $\sigma = 0$ (see App. C for details). In Fig. 2, we plot the train loss, the norm growth of $\boldsymbol{W}_t$, the softmax score $\varphi_{i,\texttt{opt}_i}$ for the $\texttt{opt}$ token (averaged over $i \in [n]$), and the alignment of $\boldsymbol{W}_t$ with $\boldsymbol{W}_{\mathrm{mm}}$. Note that since $\boldsymbol{u}_*$ is fixed, the train loss does not converge to 0. Observe that the directional alignment for NGD is closely predicted by Thm. 1. Similarly, the softmax-score saturation rate in Lem. 1 closely aligns with the empirical rate in Fig. 2. Moreover, note that NGD is significantly faster than standard GD, which we formally prove below.

3.3 Remarks

NGD is faster than GD. We demonstrate that there exists a dataset, satisfying the assumptions of Thm. 1, for which the correlation of vanilla GD parameters to $\boldsymbol{W}_{\mathrm{mm}}$, $\langle \overline{\boldsymbol{W}}_t, \overline{\boldsymbol{W}}_{\mathrm{mm}} \rangle$, is upper bounded by $\leq 1 - \Omega((\log t)^{-2})$. In contrast, according to Thm. 1, the rate for NGD is $\geq 1 - \mathcal{O}((\log t)^2/t)$, thereby highlighting a significant gap between the two.

To construct the dataset, let $n = 1$, $\mathrm{T} = 2$, $y = 1$, $\boldsymbol{x}_1 = [1,0]^\top$, $\boldsymbol{x}_2 = [0,0]^\top$ and $\boldsymbol{u}_* = [1,0]^\top$. Clearly, Ass. 2 holds since there is only two tokens; specifically, $\texttt{opt} = 1$, $\gamma_{\texttt{opt}} - \gamma = 1$. Also, Ass. 1 holds trivially since $n = 1$. For this dataset, the max-margin classifier can be easily expressed in closed form as $\boldsymbol{W}_{\mathrm{mm}} = (\boldsymbol{x}_1 - \boldsymbol{x}_2)\boldsymbol{x}_1^\top$. Using this, we observe that $\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t) = -\widehat{L}(\boldsymbol{W}_t)\varphi^t_{\texttt{opt}}(1-\varphi^t_{\texttt{opt}})\boldsymbol{W}_{\mathrm{mm}}$, which implies

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_0 + \eta G_t \boldsymbol{W}_{\mathrm{mm}}\,, \quad \text{where we set } \; G_t := \sum_{t'=0}^{t} \|\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_{t'})\|\,.$$

Leveraging this, we now show that $\|\boldsymbol{W}_t\| \leq \mathcal{O}(\log t)$: Note that $\|\boldsymbol{W}_t\| \to \infty$ as $t \to \infty$ which leads to softmax saturation $\varphi^t_{\texttt{opt}} \to 1$ and hence, loss minimization. This implies that G_t grows with t, and there exists some $t_0 > 0$, such that for any $t \geq t_0$, it holds that $\|\boldsymbol{W}_{t+1}\| \approx \eta G_t + W_0^{11}$, where W_0^{11} is the (1,1) entry of the initialization matrix. Using this, we can show for $t \geq t_0$ that $1 - \varphi^{t+1}_{\texttt{opt}} = \frac{1}{1+\exp((\boldsymbol{x}_1-\boldsymbol{x}_2)^\top \boldsymbol{W}_{t+1}\boldsymbol{x}_1)} \leq \exp(-(\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \boldsymbol{W}_{t+1}\boldsymbol{x}_1) = \exp(-\eta G_t - W_0^{11}) \approx \exp\left(-\|\boldsymbol{W}_{t+1}\|\right)$. Thus,

$$\|\boldsymbol{W}_{t+1}\| \leq \|\boldsymbol{W}_0\| + \eta \sum_{t'=0}^{t} \widehat{L}(\boldsymbol{W}_{t'}) \exp\left(-\|\boldsymbol{W}_{t'}\|\right) \leq \mathcal{O}\Big(\sum_{t'=0}^{t} \exp\left(-\|\boldsymbol{W}_{t'}\|\right)\Big),$$

where we use $\widehat{L}(\boldsymbol{W}_t) \leq \mathcal{O}(1)$, since the decoder is fixed. This proves that $\|\boldsymbol{W}_t\| \leq \mathcal{O}(\log t)$, which we use to upper bound the correlation as follows:

$$\langle \overline{\boldsymbol{W}}_t, \overline{\boldsymbol{W}}_{\mathrm{mm}} \rangle = \frac{\eta G_t + W_0^{11}}{\sqrt{(\eta G_t + W_0^{11})^2 + \|\boldsymbol{W}_0\|^2 - (W_0^{11})^2}} \approx 1 - \frac{\|\boldsymbol{W}_0\|^2 - (W_0^{11})^2}{2(\eta G_t + W_0^{11})^2} \approx 1 - \frac{\|\boldsymbol{W}_0\|^2 - (W_0^{11})^2}{2\|\boldsymbol{W}_t\|^2}. \tag{3}$$

Thus, $\langle \overline{\boldsymbol{W}}_t, \overline{\boldsymbol{W}}_{\text{mm}} \rangle \leq 1 - \Omega((\log t)^{-2})$, or equivalently $\|\overline{\boldsymbol{W}}_t - \overline{\boldsymbol{W}}_{\text{mm}}\| \geq \Omega((\log t)^{-1})$. This shows that using an adaptive step-size accelerates convergence to $\boldsymbol{W}_{\text{mm}}$, compared to standard GD.

Extension to Polyak step-size. In Thm. 7 in the App., we establish an analogue of Thm. 1 specifically for the Polyak step-size rule. This extension required us to further establish a lower bound on the gradient norm. This is because PS has an additional $\|\nabla_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}_t)\|^{-1}$ factor compared to NGD. To accomplish this, we used the variational form of the Euclidean norm for a "good" direction $\boldsymbol{W}_{\text{mm}}$, yielding $\|\nabla_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}_t)\| \geq \text{poly}\left(\Lambda^{-1}\right) \widehat{L}(\boldsymbol{\theta}_t)$.

As seen in Fig. 2, PS leads to even faster convergence than NGD across all metrics, including the train loss. In our implementation, we use a hyperparameter $\eta_{\max}$ as a threshold on the step-size value, which has been shown to stabilize training (Loizou et al., 2021). Further, since the selection of `opt` tokens minimizes the loss, $\widehat{L}^*$ is calculated by setting $\varphi_{i,\texttt{opt}_i} = 1$ for all $i \in [n]$. Note that after some time, when $\eta_t \geq \eta_{\max}$, the PS updates used in practice are identical to standard GD with step-size $\eta_{\max}$. Hence, the rates are much faster in the beginning, but slow down as training progresses.

3.4 Proof Sketch of Theorem 1

The first key intermediate result we need to show is that after sufficient iterations, the negative gradient of the loss at $\boldsymbol{W}_t$ is more correlated with $\boldsymbol{W}_{\text{mm}}$ compared to $\boldsymbol{W}_t$ itself. Formally, we show that for all $\epsilon > 0$, there exists $R_\epsilon := 2\Lambda\epsilon^{-1}\log(C\epsilon^{-1})$ with $C = \text{poly}\left(\mathrm{T}, B^2\Lambda, n\Upsilon\right)$, such that for every t for which $\|\boldsymbol{W}_t\| \geq R_\epsilon$, it holds that

$$\left\langle -\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t), \overline{\boldsymbol{W}}_{\text{mm}} \right\rangle \geq (1-\epsilon) \left\langle -\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t), \overline{\boldsymbol{W}}_t \right\rangle. \tag{4}$$

A proof sketch of this follows; see Lem. 7 in the App. for details. First, under Ass. 2, for any $\boldsymbol{W}$:

$$\langle -\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t), \boldsymbol{W} \rangle = \frac{1}{n}\sum_{i=1}^{n} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)\varphi^t_{i,\texttt{opt}_i}(1 - \varphi^t_{i,\texttt{opt}_i})h_{t,i}, \tag{5}$$

where $h_{t,i} = a_{i,\texttt{opt}_i} - \frac{\sum_{\tau \neq \texttt{opt}_i} \varphi^t_{i,\tau} a_{i,\tau}}{\sum_{\tau \neq \texttt{opt}_i} \varphi^t_{i,\tau}}$, $\boldsymbol{a}_i = \boldsymbol{X}_i \boldsymbol{W} \boldsymbol{x}_{i,1}$. Analogously, define $\tilde{h}$, h^*, using $\tilde{\boldsymbol{a}} = \Lambda \boldsymbol{X} \overline{\boldsymbol{W}}_t \boldsymbol{x}_1$, $\boldsymbol{a}^* = \boldsymbol{X}\boldsymbol{W}_{\text{mm}}\boldsymbol{x}_1$, respectively. Then, to prove Eq. (4), it suffices that

$$\tilde{h}_{t,i} \leq (1+\nu)h^*_{t,i}, \quad \text{where } \nu = \epsilon(1-\epsilon)^{-1}. \tag{6}$$

Suppose that $\|\overline{\boldsymbol{W}}_t - \overline{\boldsymbol{W}}_{\text{mm}}\| > \frac{\nu}{2B^2\Lambda}$, since otherwise the claim follows easily. We categorize the training samples into three subsets based on the satisfaction of the constraints in (W-SVM). Concretely, define the "minimum SVM-gap per sample" $i \in [n]$ as $\delta_i^{\min} := \min_{\tau \notin \{\texttt{opt}_i\}} \tilde{a}_{i,\texttt{opt}_i} - \tilde{a}_{i,\tau} - 1$ and based on this consider three sets of samples as follows:

- $\mathcal{I}_1 := \{i \in [n] \mid \delta_i^{\min} \leq 0.3\nu\}$, for samples where the constraints are violated or weakly satisfied,
- $\mathcal{I}_2 := \{i \in [n] \mid 0.3\nu < \delta_i^{\min} \leq 0.8\nu\}$, for samples where constraints are satisfied with small gap,
- $\mathcal{I}_3 := \{i \in [n] \mid \delta_i^{\min} \geq 0.8\nu\}$, for samples where the constraints are well-satisfied.

Since $\|\overline{\boldsymbol{W}}_t - \overline{\boldsymbol{W}}_{\text{mm}}\|$ is large, there is at least one point $i \in [n]$ that violates the constraints, i.e. $\mathcal{I}_1$ is non-empty. We first show the following for examples in $\mathcal{I}_1$ and $\mathcal{I}_2$

$$\tilde{h}_{t,i} \leq (1+0.5\nu)h^*_{t,i},\ i \in \mathcal{I}_1, \quad \text{and} \quad \tilde{h}_{t,i} \leq (1+\nu)h^*_{t,i},\ i \in \mathcal{I}_2\,. \tag{7}$$

Note that we prove a tighter inequality for $\mathcal{I}_1$ than required by (6) in order to use this residual to get the desired inequality for examples in $\mathcal{I}_3$. Specifically, when R_ϵ is large enough, we wish to show (following Eq. (5)) that

$$\sum_{i \in \mathcal{I}_3} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)\varphi^t_{i,\texttt{opt}_i}(1 - \varphi^t_{i,\texttt{opt}_i}) \leq \frac{0.5\nu}{2B^2\Lambda} \sum_{i \in \mathcal{I}_1} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)\varphi^t_{i,\texttt{opt}_i}(1 - \varphi^t_{i,\texttt{opt}_i}). \tag{8}$$

Intuitively, for the LHS, any sample $i \in \mathcal{I}_3$ satisfies (W-SVM) constraints with large gap giving $\varphi_{i,\texttt{opt}_i}$ close to 1 and growing with $\|\boldsymbol{W}_t\|$. Similarly, for the RHS, we expect $\varphi_{i,\texttt{opt}_i}$ to be smaller. Thus, for a sufficiently large $\|\boldsymbol{W}_t\|$ and controlling $\varphi_{i,\texttt{opt}_i}$ for any $i \in \mathcal{I}_1$ and $t > 0$ would allow us to show the above inequality. This is achieved by establishing a non-trivial lower bound on the softmax scores for $i \in \mathcal{I}_1$ that holds throughout training (Lem. 3 in the App.). Finally, since $h^* \geq 1$ and $\tilde{h} \leq 2B^2\Lambda$, we combine Eqs. (7) and (8) using the formulation in (5) and complete the proof of Eq. (4).

The next key ingredient in the proof of Thm. 1 is showing that the iterate norm grows at rate $\Theta(t)$. Specifically, we prove in Lem. 5 in the App. that

$$2\eta t \geq \|\boldsymbol{W}_t\| \geq \eta(4B^2\Lambda)^{-1}t. \tag{9}$$

The upper bound follows by applying Cauchy-Schwarz. For the lower bound, we write the iterate $\boldsymbol{W}_t$ using its gradient updates, use the dual norm characterization of the ℓ_2-norm, and select the $\boldsymbol{W}_{\text{mm}}$ direction to get

$$\|\boldsymbol{W}_t\| \geq \left\langle \boldsymbol{W}_0, \overline{\boldsymbol{W}}_{\text{mm}} \right\rangle + \sum_{t'=0}^{t-1} \eta \left\langle -\frac{\nabla_{\boldsymbol{W}} \widehat{\mathcal{L}}(\boldsymbol{W}_t)}{\|\nabla_{\boldsymbol{W}} \widehat{\mathcal{L}}(\boldsymbol{W}_t)\|}, \frac{\boldsymbol{W}_{\text{mm}}}{\|\boldsymbol{W}_{\text{mm}}\|} \right\rangle,$$

where we use the key property that for any $\boldsymbol{W}$ the gradient correlates positively with $\boldsymbol{W}_{\text{mm}}$ (formalized in Lem. 4 in the App.), i.e. $\left\langle -\frac{\nabla_{\boldsymbol{W}} \widehat{\mathcal{L}}(\boldsymbol{W})}{\|\nabla_{\boldsymbol{W}} \widehat{\mathcal{L}}(\boldsymbol{W})\|}, \frac{\boldsymbol{W}_{\text{mm}}}{\|\boldsymbol{W}_{\text{mm}}\|} \right\rangle \geq (2B^2\Lambda)^{-1} > 0$.

Next, using Eq. (2) to substitute the gradient update in Eq. (4), we can show that

$$\begin{aligned}\left\langle \boldsymbol{W}_{t+1} - \boldsymbol{W}_t, \frac{\boldsymbol{W}_{\text{mm}}}{\|\boldsymbol{W}_{\text{mm}}\|} \right\rangle \geq (1-\epsilon_t) \left\langle \boldsymbol{W}_{t+1} - \boldsymbol{W}_t, \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|} \right\rangle &= \frac{\|\boldsymbol{W}_{t+1}\|^2 - \|\boldsymbol{W}_t\|^2 - \|\boldsymbol{W}_{t+1} - \boldsymbol{W}_t\|^2}{2\|\boldsymbol{W}_t\|} - \epsilon_t \left\langle \boldsymbol{W}_{t+1} - \boldsymbol{W}_t, \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|} \right\rangle \\ &\geq \|\boldsymbol{W}_{t+1}\| - \|\boldsymbol{W}_t\| - \frac{\eta^2}{2\|\boldsymbol{W}_t\|} - \epsilon_t\eta.\end{aligned}$$

We then select $\epsilon_t = (80B^2\Lambda^2 \log t)(\eta t)^{-1}$, and use Eq. (9) and Lem. 7 to obtain a $\mathcal{O}((\log t)^2 \|\boldsymbol{W}_t\|^{-1})$ rate. Using Eq. (9) then finishes the proof.

4 Training Dynamics for Joint Optimization

We now study training dynamics when jointly optimizing prediction head $\boldsymbol{u}$ and attention weights $\boldsymbol{W}$. Compared to Sec. 3, the key challenge is that the token scores $\boldsymbol{\gamma}^t$ evolve with time, driven by the changing nature of $\boldsymbol{u}_t$. Additionally, the objective function becomes non-smooth in this context, given the dynamic changes in $\boldsymbol{u}_t$. Addressing these challenges necessitates additional technical considerations compared to Sec. 3, which we also discuss in this section.

4.1 Setup

We consider the following updates for $\boldsymbol{u}_t$ and $\boldsymbol{W}_t$.

$$\boldsymbol{u}_{t+1} = \boldsymbol{u}_t - \eta_t^u \nabla_{\boldsymbol{u}} \widehat{\mathcal{L}}(\boldsymbol{\theta}_t),\ \eta_t^u = \frac{\eta(t+1)^{-2/3}}{\|\nabla_{\boldsymbol{u}} \widehat{\mathcal{L}}(\boldsymbol{\theta}_t)\|},\ \boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta_t^W \nabla_{\boldsymbol{W}} \widehat{\mathcal{L}}(\boldsymbol{\theta}_t),\ \eta_t^W = \frac{\eta(t+1)^{-1}}{\|\nabla_{\boldsymbol{W}} \widehat{\mathcal{L}}(\boldsymbol{\theta}_t)\|}. \tag{10}$$

Here, we focus on exponential loss. Note that in η_t^u, the factor $(t+1)^{-2/3}$ can be replaced with $(t+1)^{-p}$, where $p \in [2/3, 1)$. In our results, we present findings using $p = 2/3$, as it yields the fastest rate. We also remark that common adaptive learning rates like Adam (Kingma & Ba, 2014), AdaGrad (Duchi et al., 2011), etc. similarly set per-parameter learning rates that vary dynamically.

Similar to Def. 1, given a trainable prediction head $\boldsymbol{u}_t$ at any iteration $t > 0$, the token score vector for a sample $(y, \boldsymbol{X})$ is given by $\boldsymbol{\gamma}^t = y\boldsymbol{X}\boldsymbol{u}_t$.

Data Model. We study the following data model. Within each example $\boldsymbol{X}$, a single $\texttt{opt} \sim \text{Unif}([\text{T}])$ token and an additional $T-1$ tokens are sampled i.i.d. Gaussian as follows:

$$\boldsymbol{x}_{\texttt{opt}} \sim \mathcal{N}(\boldsymbol{0}, \alpha^2\rho^2\boldsymbol{I}_d), \quad \boldsymbol{x}_\tau \sim \mathcal{N}(\boldsymbol{0}, \rho^2\boldsymbol{I}_d),\ \forall\, \tau \in \mathcal{S}, \quad y = \texttt{sign}(\boldsymbol{u}_*^\top \boldsymbol{x}_{\texttt{opt}}), \tag{DM}$$

where $\mathcal{S}:=[\mathrm{T}]\setminus\{\texttt{opt}\}$, $\boldsymbol{u}_*\in\mathbb{R}^d$ is a fixed vector that generates the label $y\in\{\pm1\}$. Here, ρ controls the token norms, and $\alpha>1$ separates the $\texttt{opt}$ token from the other tokens. We consider the following mild conditions on the initialization, token norms, overparameterization and step-size.

Condition 1. *Let $B=\alpha\rho\sqrt{1.5d}\geq 1$, and $C_0>1$, $C_1,C_2>0$ be some absolute constants. We consider the following conditions for any $\delta\in(0,1)$: i) Zero initialization: $\|\boldsymbol{u}_0\|=0, \|\boldsymbol{W}_0\|=0$; ii) $\texttt{opt}$ token has larger norm: $\alpha\geq C_0$; iii) Sufficient overparameterization: $d\geq C_1\alpha^4 n^4\log\left(\frac{10n^2}{\delta}\right)$; iv) Small step-size: $\eta\leq(C_2\sqrt{n}B^4(B\vee d))^{-1}$.*

The condition on the initialization is for ease of exposition; our analysis and results extend to random initialization with small scale (similar to the conditions of Lem. 2 in Sec. 3). We also remark that under Cond. 1 and data model DM, the tokens are nearly orthogonal, similar to Ass. 1 in Sec. 3. We formalize this in Lem. 9 in the Appendix.

4.2 Main Results

We now present our main results for joint training. For ease of exposition, we consider $\mathrm{T}=2$, but our results directly extend to $\mathrm{T}>2$ assuming token scores $\boldsymbol{\gamma}^t$ satisfy Ass. 2 at every iteration t. It is also possible to extend our results to the case where Ass. 2 is not exactly satisfied, but the non-$\texttt{opt}$ token scores show small deviation.

Our first result shows that train loss goes to 0 at $\mathcal{O}(\exp(-t^{1/3}))$ rate; see Sec. 4.3 for a proof sketch.

Theorem 2 (Train loss convergence)**.** *Under Cond. 1, data model DM, using the updates in Eq. (10), it holds for any $t>0$ that $\widehat{L}(\boldsymbol{\theta}_{t+1})\leq\mathcal{O}\left(\exp\left(-\frac{\eta B C_0}{\sqrt{n}}(t+1)^{1/3}\right)\right)$.*

The next theorem illustrates that $\boldsymbol{W}_t$ converges to $\boldsymbol{W}_{\mathrm{mm}}$ in direction at an $\mathcal{O}(1/\log t)$ rate.

Theorem 3 (IB rate of $\boldsymbol{W}$)**.** *Under Cond. 1 and the data model DM, using the updates in Eq. (10), for any $t\geq t_\epsilon=\exp(\mathrm{poly}\left(\eta,B,\Lambda,n,d,\log\delta^{-1},\epsilon^{-1}\right))$,*

$$\left\langle\overline{\boldsymbol{W}}_t,\overline{\boldsymbol{W}}_{mm}\right\rangle\geq 1-\epsilon-\mathrm{poly}\left(\eta,B,\Lambda,\epsilon\right)\frac{1}{\log t}.$$

In particular, assuming $\rho=\mathcal{O}\left(\eta^{1/2}d^{-3/4}(n\Lambda)^{-1}\right)$, for any $t\geq t_\epsilon=\Omega(\exp(\epsilon^{-4/3}))$, $\left\|\overline{\boldsymbol{W}}_t-\overline{\boldsymbol{W}}_{\mathrm{mm}}\right\|\leq\mathcal{O}\left(\epsilon\vee(\log t)^{-1}\right)$. Intuitively, since the updates for $\boldsymbol{W}_t$ have an additional $(t+1)^{-1}$ factor due to the smaller step-size η_t, the iterate norm $\|\boldsymbol{W}_t\|$ grows as $\Theta(\log t)$ in contrast to the $\Theta(t)$ rate in Sec. 3, where we were only optimizing $\boldsymbol{W}_t$. Consequently, the convergence to implicit bias is slower — $\mathcal{O}(1/\log t)$ instead of $\tilde{\mathcal{O}}(t^{-1})$. The additional $(t+1)^{-1}$ factor in the updates for $\boldsymbol{W}_t$ helps to account for the non-smoothness of the objective, which will become apparent in Sec. 4.3.

We now discuss the implicit bias and convergence of $\boldsymbol{u}_t$. From prior work (Soudry et al., 2018; Ji & Telgarsky, 2019), one could guess that $\boldsymbol{u}_t$ converges to the max-margin predictor separating the set of samples $\{(\varphi^t_{i,\texttt{opt}_t}\boldsymbol{x}_{i,\texttt{opt}_i}+(1-\varphi^t_{i,\texttt{opt}_i})\boldsymbol{x}_{i,\tau},y_i)_{i=1}^n\}$. But as shown above, when $t\to\infty$, $\varphi_{\texttt{opt}}\to 1$. This motivates the following hard-margin SVM problem,

$$\boldsymbol{u}_{\mathrm{mm}}=\underset{\boldsymbol{u}}{\mathrm{argmin}}\|\boldsymbol{u}\| \text{ s.t. } y_i\boldsymbol{u}^\top\boldsymbol{x}_{i,\texttt{opt}_i}\geq 1\ \forall i\in[n]. \tag{u-SVM}$$

The following result confirms that the learned model attains margin that is asymptotically a constant factor $(1/4)$ of the maximum margin $\overline{\gamma}:=\max_{\boldsymbol{u}:\|\boldsymbol{u}\|\leq 1}\min_i y_i\boldsymbol{u}^\top\boldsymbol{x}_{i,\texttt{opt}_i}$.

Theorem 4 (IB rate of $\boldsymbol{u}$)**.** *Under Cond. 1 and the data model DM, using the updates in Eq. (10), for any $t\geq t_\epsilon\vee\exp(C(\eta,B,\Lambda,\epsilon)(\epsilon^{-1}\vee(8B^2\Lambda)^4))$, it holds that*

$$\min_i y_i\overline{\boldsymbol{u}_t}^\top\boldsymbol{x}_{i,opt_i}\geq\frac{\overline{\gamma}}{4}-\frac{1}{1+\exp(\eta(8B^2\Lambda^2)^{-1}\log t)}.$$

4.3 Proof Sketch of Theorem 2

The main challenge in establishing the loss convergence rate in Thm. 2 is the non-smoothness of the objective function. To overcome this, we show three key properties that hold throughout training: i) constant loss ratio for any two samples, ii) growing token score gap between `opt` and non-`opt` tokens for every sample, and iii) a non-decreasing `opt`-token softmax score. We formalize these in Lem. 11 in the App. Property (ii) is crucial in proving the implicit bias rate in Thm. 3. Property (iii) is crucial in obtaining a PL-like inequality which is the key challenging step in proving Thm. 2. Specifically, using properties (i) and (iii) we demonstrate that the training loss satisfies the following (see Lem. 13 and Rem. 3 in the App. for details):

1. PL-inequality-like result: $\|\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t)\| \geq \Omega(Bn^{-1/2})\widehat{L}(\boldsymbol{\theta}_t)$,
2. controlled loss between current and next iterate: $\max_{\boldsymbol{\theta}'\in[\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}]}\widehat{L}(\boldsymbol{\theta}') \leq 8\widehat{L}(\boldsymbol{\theta}_t)$,
3. second-order self-boundedness: $\max\limits_{\boldsymbol{\theta}'\in[\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}]}\|\nabla^2_{\boldsymbol{\theta}}\widehat{L}(\boldsymbol{\theta}')\| \leq 8(\omega(\boldsymbol{\theta}_t)\vee\omega(\boldsymbol{\theta}_{t+1}))\widehat{L}(\boldsymbol{\theta}_t)$,

where $\omega(\boldsymbol{\theta}_t) \coloneqq 13B^5(B\vee d)(\|\boldsymbol{u}_t\|\vee 1)^2$.

To prove the first point, we use the dual norm characterization of the ℓ_2-norm,

$$\|\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t)\| = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1}\left\langle\frac{1}{n}\sum_{i=1}^{n}|\ell'_{i,t}|y_i\nabla_{\boldsymbol{u}}\Phi(\boldsymbol{u}_t,\boldsymbol{X}_i),\boldsymbol{v}\right\rangle \geq \widehat{L}(\boldsymbol{\theta}_t)\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1}\min_i y_i(\varphi^t_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i}+(1-\varphi^t_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau})^\top\boldsymbol{v}.$$

To proceed, we lower bound the softmax scores for the `opt` tokens for every $i\in[n]$ and $t>0$ as

$$\varphi^t_{i,\mathtt{opt}_i} \geq \tfrac{1}{2}. \tag{11}$$

Using this, we show that the set of samples $\{(0.5(\boldsymbol{x}_{i,\mathtt{opt}_i}+\boldsymbol{x}_{i,\tau}),y_i)\}_{i=1}^n$, where $\tau\neq\mathtt{opt}_i$, are separable using $\tilde{\boldsymbol{u}}=\sum_{i\in[n]}y_i\boldsymbol{x}_{i,\mathtt{opt}_i}$ (formalized in Lem. 12 in the App.). Then, selecting $\boldsymbol{v}=\tilde{\boldsymbol{u}}$ proves the first point.

We now briefly discuss the process to prove Eq. (11). We can show this by induction. Since $\boldsymbol{W}_0=\boldsymbol{0}$, then for any $i\in[n]$, $\varphi^0_{i,\mathtt{opt}_i}=1/2$ at initialization. To show the inductive step, note that for any $i\in[n]$, if

$$(\boldsymbol{x}_{i,\mathtt{opt}_i}-\boldsymbol{x}_{i,\tau})^\top(-\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{\theta}_t))\boldsymbol{x}_{i,1}>0, \tag{12}$$

$$\implies \frac{\varphi^{t+1}_{j,\mathtt{opt}_j}/\varphi^{t+1}_{j,\tau}}{\varphi^{t}_{j,\mathtt{opt}_j}/\varphi^{t}_{j,\tau}}=\exp\big((\boldsymbol{x}_{j,\mathtt{opt}_j}-\boldsymbol{x}_{j,\tau})^\top(\boldsymbol{W}_{t+1}-\boldsymbol{W}_t)\boldsymbol{x}_{j,1}\big)\geq 1.$$

We consider two cases. If $\varphi^t_{i,\mathtt{opt}_i}\geq 3/4$, Eq. (12) follows as η is small. Specifically, for any $j\in[n]$, $\tau\neq\mathtt{opt}_j$,

$$\varphi^t_{j,\mathtt{opt}_j}=\frac{1}{1+\exp((\boldsymbol{x}_{i,\tau}-\boldsymbol{x}_{j,\mathtt{opt}_j})^\top\boldsymbol{W}_t\boldsymbol{x}_{j,1})}>3/4 \qquad\implies\qquad (\boldsymbol{x}_{j,\tau}-\boldsymbol{x}_{j,\mathtt{opt}_j})^\top\boldsymbol{W}_{t-1}\boldsymbol{x}_{j,1}\leq-\log(3).$$

After the gradient step, since $\eta\leq(2B^2)^{-1}\log(3)$, we have

$$(\boldsymbol{x}_{j,\tau}-\boldsymbol{x}_{j,\mathtt{opt}_j})^\top\boldsymbol{W}_{t+1}\boldsymbol{x}_{j,1}=(\boldsymbol{x}_{j,\tau}-\boldsymbol{x}_{j,\mathtt{opt}_j})^\top\boldsymbol{W}_t\boldsymbol{x}_{j,1}-\eta_{t-1}(\boldsymbol{x}_{j,\mathtt{opt}_j}-\boldsymbol{x}_{j,\tau})^\top(-\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t))\boldsymbol{x}_{j,1}\leq 2\eta B^2-\log(3)\leq 0.$$

If $\varphi^t_{i,\mathtt{opt}_i}\leq 3/4$, we proceed as follows. Consider the expansion of the LHS of Eq. (12),

$$\begin{aligned}(\boldsymbol{x}_{i,\mathtt{opt}_i}-\boldsymbol{x}_{i,\tau})^\top(-\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{\theta}_t))\boldsymbol{x}_{i,1}&=\frac{\ell_{t,i}}{n}(\gamma^t_{i,\mathtt{opt}_i}-\gamma^t_i)\varphi^t_{i,\mathtt{opt}_i}(1-\varphi^t_{i,\mathtt{opt}_i})\|\boldsymbol{x}_{i,1}\|^2\|\boldsymbol{x}_{i,\mathtt{opt}_i}-\boldsymbol{x}_{i,\tau}\|^2\\&\quad-\frac{1}{n}\sum_{j\neq i}\ell_{t,j}(\gamma^t_{j,\mathtt{opt}_j}-\gamma^t_j)\varphi^t_{j,\mathtt{opt}_j}(1-\varphi^t_{j,\mathtt{opt}_j})\boldsymbol{x}_{i,1}^\top\boldsymbol{x}_{j,1}(\boldsymbol{x}_{j,\mathtt{opt}_j}-\boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{i,\mathtt{opt}_i}-\boldsymbol{x}_{i,\tau'}).\end{aligned}$$

To use this to prove Eq. (12), we first show that the token scores $\boldsymbol{\gamma}^t$ satisfy Defn. 1 ($\gamma^t_{\mathtt{opt}_i}>\gamma^t_i$ for any $i\in[n]$), and that the loss ratio $\max_{i,j\in[n]}\frac{\ell_{t,i}}{\ell_{t,j}}$ is bounded by a constant throughout training, *i.e.*, for any $i\in[n]$,

$$\gamma^t_{i,\mathtt{opt}_i}-\gamma^t_i\geq\Omega(t^{1/3})\quad\text{and}\quad\max_{i,j\in[n]}\frac{\ell_{t,i}}{\ell_{t,j}}\leq C, \tag{13}$$

where $C>0$ is a universal constant. This is formalized in Lem. 11 in the App. We then prove Eqs. (11) and (13) jointly by induction.

For the second point, we first show that for any $\boldsymbol{\theta}, \boldsymbol{\theta}'$,

$$\begin{aligned}|y\Phi(\boldsymbol{\theta}, \boldsymbol{X}) - y\Phi(\boldsymbol{\theta}', \boldsymbol{X})| &= |\boldsymbol{u}^\top \boldsymbol{X}^\top \varphi(\boldsymbol{X}\boldsymbol{W}^\top \boldsymbol{x}_1) - \boldsymbol{u}'^\top \boldsymbol{X}^\top \varphi(\boldsymbol{X}\boldsymbol{W}'^\top \boldsymbol{x}_1)| \\ &\leq \|\boldsymbol{X}\|_{2,\infty}\|\boldsymbol{u} - \boldsymbol{u}'\| + 2\|\boldsymbol{X}\|_{2,\infty}^3\|\boldsymbol{u}'\|\|\boldsymbol{W} - \boldsymbol{W}'\|.\end{aligned}$$

Since we are using exponential loss, we use this to get,

$$\max_{\lambda\in[0,1]} \frac{\ell(y\Phi(\boldsymbol{\theta}_t + \lambda(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t), \boldsymbol{X}))}{\ell(y\Phi(\boldsymbol{\theta}_t, \boldsymbol{X}))} \leq \max_{\lambda\in[0,1]} \exp\left(2\lambda\|\boldsymbol{X}\|_{2,\infty}^3\|\boldsymbol{u}_t\|\|\boldsymbol{W}_{t+1} - \boldsymbol{W}_t\| + \lambda\|\boldsymbol{X}\|_{2,\infty}\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\|\right).$$

Further, we show that the iterate norm $\|\boldsymbol{u}_t\|$ grows as $\mathcal{O}(t^{1/3})$ (formalized in Lem. 10 in the App.). Then, using the updates in Eq. (10), we get

$$\max_{i\in[n]} \max_{\lambda\in[0,1]} \frac{\ell(y_i\Phi(\boldsymbol{\theta}_t + \lambda(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t), \boldsymbol{X}_i))}{\ell(y_i\Phi(\boldsymbol{\theta}_t, \boldsymbol{X}_i))} \leq \exp\left(6\eta^2 B^3 + \eta B\right).$$

Then, using Cond. 1 for η finishes the proof for the second point. For the third point, we first upper bound $\|\nabla^2_{\boldsymbol{\theta}}\Phi\|$ as $\|\nabla^2_{\boldsymbol{\theta}}\Phi(\boldsymbol{\theta}, \boldsymbol{X})\| \leq 6dB^5\|\boldsymbol{u}\| + 2\sqrt{d}B^3$. Using this and the upper bound on $\|\nabla_{\boldsymbol{\theta}}\Phi\|$, we show that for any $\boldsymbol{\theta}' \in [\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]$,

$$\|\nabla^2_{\boldsymbol{\theta}}\widehat{L}(\boldsymbol{\theta}')\| \leq \max_{i\in[n]}(\|\nabla_{\boldsymbol{\theta}}\Phi(\boldsymbol{\theta}_t, \boldsymbol{X}_i)\|^2 + \|\nabla^2_{\boldsymbol{\theta}}\Phi(\boldsymbol{\theta}_t, \boldsymbol{X}_i)\|)\widehat{L}(\boldsymbol{\theta}') \leq \omega(\boldsymbol{\theta}')\widehat{L}(\boldsymbol{\theta}').$$

Then, we leverage the second point to finish proving the third point.

With these ingredients, we prove Thm. 2 by working with the second-order Taylor expansion of $\widehat{L}(\boldsymbol{\theta}_{t+1})$. Specifically, using the updates in Eq. (10) and the three key properties of the training loss established above, we show that

$$\widehat{L}(\boldsymbol{\theta}_{t+1}) \leq \widehat{L}(\boldsymbol{\theta}_t) - \frac{\eta B}{10\sqrt{n}(t+1)^{2/3}}\widehat{L}(\boldsymbol{\theta}_t) + \frac{8\eta^2}{(t+1)^{4/3}}(\omega(\boldsymbol{\theta}_t) \vee \omega(\boldsymbol{\theta}_{t+1}))\widehat{L}(\boldsymbol{\theta}_t).$$

Further, since $\|\boldsymbol{u}_t\| \leq \mathcal{O}(t^{1/3})$, we show that $\omega(\boldsymbol{\theta}_t) \leq \mathcal{O}(t^{2/3})$. Using a small enough η, and telescoping, gives the advertised result.

5 Experimental Results

To complement our theory, we present experiments on synthetic/ real-world data demonstrating that (S)NGD-based training leads to faster convergence for various metrics compared to vanilla (S)GD.

Synthetic Data. We first consider the self-attention model in Eq. (1) and data generated as per data model DM. Fig. 3 shows the train loss, iterate norms $\|\boldsymbol{u}_t\|$ and $\|\boldsymbol{W}_t\|$, average softmax score for the `opt` tokens, and alignment of iterates $\boldsymbol{u}_t$ and $\boldsymbol{W}_t$ to the respective max-margin solutions $\boldsymbol{W}_{\text{mm}}$ and $\boldsymbol{u}_{\text{mm}}$. We compare standard GD, NGD updates in Eq. (10) (without the additional $t^{-2/3}, t^{-1}$ factors in the step-size), PS updates, and joint NGD updates. We observe that all other algorithms are faster than GD, with PS being the fastest. We also see that the train loss converges at a similar rate for NGD and joint-NGD, while for the other metrics, NGD is slightly faster than joint-NGD.

Language and Vision Datasets. Fig. 1 compares the train and test dynamics of various optimizers while fine-tuning a pre-trained BERT (Devlin et al., 2019) model on the MNLI (Williams et al., 2018) dataset. As evident, both gradient normalized step-sizes SNGD and SPS train significantly faster than SGD. Fig. 4 in the App. shows similar results for the CivilComments dataset (Borkan et al., 2019). In App. C we present additional results on vision datasets. Interestingly, in this setting, while SPS still outperforms SGD, SNGD trains slower, which could be explored further in future work. It is also worth noting that in this setting we see no significant gap between Adam and SGD; this is consistent with the observations reported in Xie et al. (2023); Kunstner et al. (2023).

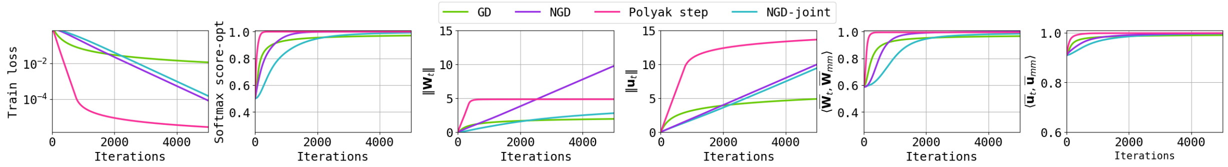

Figure 3: Training dynamics of a self-attention model (Eq. (1)) with data generated using model DM.

6 Related Work

Implicit bias of NNs. There has been extensive work on the implicit bias of GD for both linear predictors (see Sec. 1) and NNs. Despite non-convexity, our results are more closely related to those for linear predictors since we give an explicit formulation of the solution in the convergence limit (compare to implicit formulations in terms of KKT points for nonlinear NNs; see App. D for a review).

Thm. 1 provides a non-trivial extension to self-attention of corresponding results for linear logistic regression. Specifically, Ji & Telgarsky (2021) demonstrate that NGD on logistic loss under separable data converges globally, in direction, at rate $\widetilde{\mathcal{O}}(1/t)$ to the max-margin separator of data points belonging to distinct classes. While the convergence rate of Thm. 1 for self-attention is slower, it holds under a more intricate non-convex landscape induced by the softmax nonlinearity; see App. A.4 for a discussion on the tightness of our results. As we have seen, in addition to NGD, our convergence results also apply to Polyak step-size, which empirically outperforms NGD (see Fig. 2). We also present results for NGD with momentum in Fig. 2, which results in a performance boost. In the simpler setting of linear logistic regression, this method is proven to have a faster implicit bias rate of $\widetilde{\mathcal{O}}(1/t^2)$ (Ji et al., 2021). Extending our theory to incorporate momentum in the self-attention setting is an interesting direction for future work. The loss convergence rate in Thm. 2 is analogous to the $\mathcal{O}(\exp(-t))$ loss rate for linear predictors on separable data with $\eta_t = \eta/\|\nabla\widehat{L}(\boldsymbol{\theta}_t)\|$ (Ji & Telgarsky, 2019). The difference $\mathcal{O}(\exp(-t^{1/3}))$ vs. $\mathcal{O}(\exp(-t))$ shows up as we use smaller step-size. Nacson et al. (2019) similarly show a slower rate of $\mathcal{O}(\exp(-t^{1/2}))$ for linear predictors with $\eta_t = \eta(t+1)^{-1/2}/\|\nabla\widehat{L}(\boldsymbol{\theta}_t)\|$. Contrary to all these works, to the best of our knowledge, we are the first to prove parameter convergence of NGD and PS in non-convex settings. See App. D for a discussion.

Transformers theory. To understand the optimization and generalization dynamics, (Jelassi et al., 2022) shows that Vision Transformers (ViTs) learn spatially localized patterns in a binary classification task using gradient-based methods, while Li et al. (2023b) shows that attention maps sparsify as SGD training progresses. Oymak et al. (2023) studies the initial trajectory of GD for the closely related prompt-attention. Additionally, Tian et al. (2023a) studies SGD-dynamics for the next-token prediction task in a one-layer transformer with a linear decoder. More recently, Tian et al. (2023b) extends this analysis to the joint training of multi-layer transformer with an MLP. As detailed in Sec. 1, our work is most closely related and improves upon Tarzanagh et al. (2023b;a), which adopt an implicit-bias view to attention training. After completion of this work, we became aware of contemporaneous work (Sheen et al., 2024) that also studies implicit bias in the same problem setting (single-layer self-attention, fixed decoder, Ass. 2). Unlike ours, their results hold only for gradient flow (i.e. GD with infinitesimal step-size), and are asymptotic. On the other hand, they study separate optimization of key-query matrices, albeit with appropriate initialization assumptions. See App. D for a full review.

7 Conclusion and Future Work

We studied implicit optimization bias of GD for self-attention in binary classification. For both fixed and trainable decoder settings, we identified data conditions under which global convergence holds and characterized, for the first time, the convergence rates. Our convergence rates hold for NGD and PS, which are shown both theoretically and empirically to outperform vanilla GD. In future work, we aim to further relax Ass. 2 and identify and investigate other sufficient conditions for global convergence, such as overparameterization. Additionally, extending our theory to incorporate momentum and explain its fast convergence is also an interesting future direction. Finally, motivated by our experiments, we want to further investigate why NGD trains faster on language datasets, yet appears to be slower on vision datasets. This may involve extending our findings to the next-token prediction setting, which is also of interest.

Acknowledgements

PD and CT were supported by NSERC Discovery Grant No. 2021-03677, the Alliance Grant ALLRP 581098-22 and a CIFAR AI Catalyst grant. BV was supported by NSF CAREER Award CCF-2239265. The authors acknowledge the use of the Discovery cluster by USC CARC.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *Int. Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I. 46

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 46

Pierre Baldi and Roman Vershynin. The quarks of attention. *arXiv preprint arXiv:2202.08371*, 2022. 46

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *arXiv preprint arXiv:2306.00802*, 2023. 46

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019. URL http://arxiv.org/abs/1903.04561. 11, 42, 45

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019. 5

Yuan Cao, Zixiang Chen, Mikhail Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *arXiv preprint arXiv:2202.06526*, 2022. 5

Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021. URL http://jmlr.org/papers/v22/20-974.html. 5

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*, pp. 1305–1338. PMLR, 2020. 46

Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention, 2023. 3, 5, 37, 39, 46

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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423. 1, 11, 42, 45

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*, pp. 2793–2803. PMLR, 2021. 46

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, 2021. 1

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html. 8

Tolga Ergen, Behnam Neyshabur, and Harsh Mehta. Convexifying transformers: Improving optimization and understanding of transformer networks. *arXiv preprint arXiv:2211.11052*, 2022. 46

Spencer Frei and Quanquan Gu. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=NVpGLJUuPx5. 37

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pp. 2668–2703. PMLR, 2022a. 5

Spencer Frei, Gal Vardi, Peter L Bartlett, Nathan Srebro, and Wei Hu. Implicit bias in leaky relu networks trained on high-dimensional data. *arXiv preprint arXiv:2210.07082*, 2022b. 46

Spencer Frei, Gal Vardi, Peter Bartlett, and Nathan Srebro. Benign overfitting in linear classifiers and leaky relu networks from kkt conditions for margin maximization. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3173–3228. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/frei23a.html. 5

Elad Hazan, Kfir Y. Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 1594–1602, Cambridge, MA, USA, 2015. MIT Press. 3

M. Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: Unveiling the dynamics of generative transformers, 2024. 46

Samy Jelassi, Michael Eli Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=eMW9AkXaREI. 12, 46

Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018. 1

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *International Conference on Algorithmic Learning Theory*, 2019. URL https://api.semanticscholar.org/CorpusID:226245936. 9, 12, 46

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020. 46

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pp. 772–804. PMLR, 2021. 12, 46

Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2021. 12, 46

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021. 42

Yiwen Kou, Zi-Yuan Chen, Yuanzhou Chen, and Quanquan Gu. Benign overfitting in two-layer relu convolutional neural networks. In *International Conference on Machine Learning*, 2023a. URL https://api.semanticscholar.org/CorpusID:260871074. 5

Yiwen Kou, Zixiang Chen, and Quanquan Gu. Implicit bias of gradient descent for two-layer relu and leaky relu networks on nearly-orthogonal data, 2023b. 46

Alex Krizhevsky. Learning multiple layers of features from tiny images. pp. 32–33, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. 44

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=a65YK0cqH8g. 11, 44, 46

Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 2005. 43

Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets, 2021. 43

Hongkang Li, M. Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. *ArXiv*, abs/2302.06015, 2023a. 5

Hongkang Li, Meng Weng, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *International Conference on Learning Representations*, 2023b. 12, 46

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning, 2023c. 46

Yingcong Li, Yixiao Huang, Muhammed E Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 685–693. PMLR, 02–04 May 2024. 46

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning, 2021. 46

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022. 37

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1306–1314. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/loizou21a.html. 3, 7, 46

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020. 46

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Memorization capacity of multi-head attention in transformers. *arXiv preprint arXiv:2306.02010*, 2023. 46

Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with markov: A framework for principled analysis of transformers via markov chains, 2024. 46

Vidya Muthukumar, Kailas Vodrahalli, and Anant Sahai. Harmless interpolation of noisy data in regression. *arXiv preprint arXiv:1903.09139*, 2019. 5

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*, pp. 3420–3428. PMLR, 2019. 3, 12, 30, 37, 46

Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33:11961–11972, 2020. 37

OpenAI. Openai: Introducing chatgpt, 2022. URL https://openai.com/blog/chatgpt,2022. 1

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. 3, 12, 46

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. 42

Mary Phuong and Christoph H Lampert. The inductive bias of re{lu} networks on orthogonally separable data. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=krz7T0xU9Z_. 46

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html. 1

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. 21(1), 2020. ISSN 1532-4435. 1

Arda Sahiner, Tolga Ergen, Batu Ozturkler, John Pauly, Morteza Mardani, and Mert Pilanci. Unraveling attention via convex duality: Analysis and interpretations of vision transformers. *International Conference on Machine Learning*, 2022. 46

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. *arXiv preprint arXiv:2306.02896*, 2023. 46

Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, and Robert M. Gower. Momo: Momentum models for adaptive learning rates, 2024. URL https://arxiv.org/abs/2305.07583. 46

Heejune Sheen, Siyu Chen, Tianhao Wang, and Harrison H. Zhou. Implicit regularization of gradient flow on one-layer softmax attention, 2024. 12

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. 1, 9, 46

Hossein Taheri and Christos Thrampoulidis. Fast convergence in learning two-layer neural networks with separable data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9944–9952, Jun. 2023a. doi: 10.1609/aaai.v37i8.26186. URL https://ojs.aaai.org/index.php/AAAI/article/view/26186. 37

Hossein Taheri and Christos Thrampoulidis. Generalization and stability of interpolating neural networks with minimal width. *arXiv preprint arXiv:2302.09235*, 2023b. 37

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines, 2023a. 1, 2, 3, 4, 5, 12, 24, 27, 29, 41

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism, 2023b. 1, 4, 12, 24, 27

Christos Thrampoulidis. Implicit bias of next-token prediction, 2024. 46

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer, 2023a. 3, 12, 46

Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention, 2023b. 12, 46

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/touvron21a.html. 1

Gal Vardi. On the implicit bias in deep-learning algorithms, 2022. 46

Gal Vardi and Ohad Shamir. Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, pp. 4224–4258. PMLR, 2021. 46

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. *arXiv preprint arXiv:2212.07677*, 2022. 46

Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science*, 4(1): 260–284, 2022. doi: 10.1137/21M1415121. 5

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *Advances in Neural Information Processing Systems*, 34:24164–24179, 2021. 5

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101. 11, 42, 45

Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng YAN. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models, 2023. URL https://openreview.net/forum?id=mPzpPv0geS2. 11, 44

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions?, 2020a. 46

Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. $o(n)$ connections are expressive enough: Universal approximability of sparse transformers, 2020b. 46

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context, 2023. 46

Appendix

A Training only $\boldsymbol{W}$

A.1 Preliminaries

$\|\boldsymbol{A}\|_2$ denotes the spectral norm and $\lambda_{\min}(\boldsymbol{A})$ denotes the minimum eigenvalue. Also, $\|\boldsymbol{A}\|_{2,\infty} = \max_i \|\boldsymbol{a}_i\|$, where $\boldsymbol{a}_i$ is the i^{th} row of $\boldsymbol{A}$. We list the useful notations used in the Appendix in Table 1 for convenience. Further, let $\boldsymbol{\varphi}'(\boldsymbol{v}) := \nabla\boldsymbol{\varphi}(\boldsymbol{v}) = \text{diag}(\boldsymbol{\varphi}(\boldsymbol{v})) - \boldsymbol{\varphi}(\boldsymbol{v})\boldsymbol{\varphi}(\boldsymbol{v})^\top$ denote the Jacobian of the softmax vector $\boldsymbol{\varphi}(\boldsymbol{v})$ at

$\boldsymbol{x}_{i,\tau}$	τ^{th} token in the i^{th} sample
$\boldsymbol{a}(\boldsymbol{W})$	$\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_1$
γ	$y\boldsymbol{X}\boldsymbol{u}_*$
$\boldsymbol{\varphi}(\cdot)$	softmax vector
$\varphi_{i,\tau}^t$	$\boldsymbol{\varphi}(\cdot)\|_\tau$, $\tau \in [\mathrm{T}]$ for sample i at time t
Λ	$\|\boldsymbol{W}_{\text{mm}}\|$
B	$\max_i \|\boldsymbol{X}_i\|_{2,\infty}$
κ	$\frac{\max_i(\gamma_{i,\texttt{opt}_i} - \gamma_i)}{\min_i(\gamma_{i,\texttt{opt}_i} - \gamma_i)}$
Υ	$\kappa\frac{\min_i \exp(\gamma_{i,\texttt{opt}_i})}{\max_i \exp(\gamma_i)}$
ζ	$\frac{\max_i(\gamma_{i,\texttt{opt}_i} - \gamma_i)}{\max_i \exp(\gamma_i)}$

Table 1: Useful Notation.

$\boldsymbol{v} \in \mathbb{R}^{\mathrm{T}}$. We note that under Assumption 2, for any $\boldsymbol{W}$, we have

$$\begin{aligned}\langle -\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t), \boldsymbol{W} \rangle &= -\frac{1}{n}\sum_{i=1}^{n} \ell'_{t,i} \langle y_i \nabla_{\boldsymbol{W}} \Phi(\boldsymbol{W}_t, \boldsymbol{X}_i), \boldsymbol{W} \rangle \\ &= \frac{1}{n}\sum_{i=1}^{n} \ell_{t,i} (\gamma_{i,\mathtt{opt}_i} - \gamma_i) \varphi^t_{i,\mathtt{opt}_i} (1 - \varphi^t_{i,\mathtt{opt}_i}) \left[a_{i,\mathtt{opt}_i} - \frac{\sum_{\tau \neq \mathtt{opt}_i} \varphi^t_{i,\tau} a_{i,\tau}}{\sum_{\tau \neq \mathtt{opt}_i} \varphi^t_{i,\tau}} \right], \end{aligned} \tag{14}$$

where $\boldsymbol{a}_i = \boldsymbol{X}_i \boldsymbol{W} \boldsymbol{x}_{i,1}$.

Intuition for the Constraints in Equation (W-SVM). The intuition behind the constraints in Equation (W-SVM) is as follows. Consider the loss for the i^{th} sample:

$$\ell_i = \ell(y_i \boldsymbol{u}_*^\top \boldsymbol{X}_i^\top \boldsymbol{\varphi}(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{x}_{i,1})) = \ell(\boldsymbol{\gamma}_i^\top \boldsymbol{\varphi}(\boldsymbol{X}_i \boldsymbol{W} \boldsymbol{x}_{i,1})) = \ell(\gamma_{i,\mathtt{opt}_i} \varphi_{i,\mathtt{opt}_i} + \gamma_i (1 - \varphi_{i,\mathtt{opt}_i})).$$

Since ℓ is a decreasing function, it is minimized when $\varphi_{i,\mathtt{opt}_i} = 1$.

Next, consider the softmax weight for the optimal token:

$$\frac{\exp(\boldsymbol{x}_{\text{opt}}^\top \boldsymbol{W}_t \boldsymbol{x}_1)}{\sum_\tau \exp(\boldsymbol{x}_\tau^\top \boldsymbol{W}_t \boldsymbol{x}_1)} = \frac{1}{1 + \sum_{\tau \neq opt} \exp((\boldsymbol{x}_\tau - \boldsymbol{x}_{\text{opt}})^\top \boldsymbol{W}_t \boldsymbol{x}_1)}.$$

For loss minimization, this score should tend to 1 as the iterate norm $\|\boldsymbol{W}_t\|$ tends to infinity. This can only happen when the exponent is negative, that is, $\boldsymbol{W}_t$ satisfies $(\boldsymbol{x}_\tau - \boldsymbol{x}_{\mathtt{opt}})^\top \boldsymbol{W}_t \boldsymbol{x}_1 < 0$. Then, analogous to translating the constraints to the SVM for training linear models on linearly separable data, we can obtain Equation (W-SVM) for self-attention.

A.2 Key Lemmas

We first prove some key Lemmas that are useful in the proof of Theorem 1.

Lemma 2. *Let $w^0_{u,v} \sim \mathcal{N}(0, \sigma_0^2)$ for all $u, v \in [d]$, with the initialization scale*

$$\sigma_0 \le B^{-2} d^{-1/2} \left(C'^{-1} \log\left(\frac{k\mathrm{T} - 1}{\mathrm{T} - 1} \right) \wedge \eta d^{-1/2} (8\Lambda)^{-1} \right),$$

for some $k > 1$, and absolute constant $C' > 0$. Further, if $d \ge \log(2/\delta)$ for any $\delta \in (0,1)$, then with probability at least $1 - \delta$, we have $\varphi^0_{i,\mathtt{opt}_i} \ge \frac{1}{k\mathrm{T}}$ for every $i \in [n]$.

Proof. Let $w^0_{u,v} \sim \mathcal{N}(0, \sigma_0^2)$, for any $u, v \in [d]$. Then for any $i \in [n]$ and $\tau \in [T]$, for any $\delta \in (0,1)$, with probability at least $1 - \delta$ we have

$$\begin{aligned}(\boldsymbol{x}_{i,\tau} - \boldsymbol{x}_{i,\mathtt{opt}_i})^\top \boldsymbol{W}_0 \boldsymbol{x}_{i,1} &\le \left| (\boldsymbol{x}_{i,\tau} - \boldsymbol{x}_{i,\mathtt{opt}_i})^\top \boldsymbol{W} \boldsymbol{x}_{i,1} \right| \\ &\le \|\boldsymbol{x}_{i,\tau} - \boldsymbol{x}_{i,\mathtt{opt}_i}\| \|\boldsymbol{W}_0\|_2 \|\boldsymbol{x}_{i,1}\| \\ &\le 2B^2 (\sigma_0 C (2\sqrt{d} + \sqrt{\log(2/\delta)})) \\ &\le 6CB^2 \sigma_0 \sqrt{d} =: B^2 \sigma_0 \sqrt{d} C'. \end{aligned}$$

Using this, we have

$$\begin{aligned}\varphi^0_{i,\mathtt{opt}_i} &= \frac{1}{1 + \sum\limits_{\tau \neq \mathtt{opt}_i} \exp((\boldsymbol{x}_{i,\tau} - \boldsymbol{x}_{i,\mathtt{opt}_i})^\top \boldsymbol{W}_0 \boldsymbol{x}_{i,1})} \\ &\ge \frac{1}{1 + (T-1)\exp(B^2 \sigma_0 \sqrt{d} C')} \\ &\ge \frac{1}{kT} && \left(\text{using } \sigma_0 \le C'^{-1} B^{-2} d^{-1/2} \log\left(\frac{k\mathrm{T} - 1}{\mathrm{T} - 1}\right)\right) \end{aligned}$$

Also, $\|\boldsymbol{W}_0\| \le 2\sigma_0 d \le \eta (4B^2\Lambda)^{-1}$. □

Lemma 3. *Under the conditions of Lemma 2, Assumptions 1 and 2, using the updates in Eq. (2) with $\eta \leq (2B^2)^{-1}\log(2\mathrm{T}-1)$, at any time $t \geq 0$, for any sample $j \in [n]$, the softmax weight for the $\texttt{opt}_j$ token satisfies*

$$\varphi^t_{j,\texttt{opt}_j} \geq \varphi^0_{j,\texttt{opt}_j} \wedge \frac{1}{\mathrm{T}}. \tag{15}$$

Proof. We prove this by induction. Clearly, Eq. (15) is true for $t = 0$. Next, we assume that it is true at time $t-1$. Combining this with the conditions of Lemma 2, we have $\varphi^{t-1}_{j,\texttt{opt}_j} \geq \frac{1}{2\mathrm{T}}$.

Further, for any $j \in [n]$ we can have one of the following two scenarios:

Scenario 1: At time $t-1$, $\varphi^{t-1}_{j,\texttt{opt}_j} > 1 - \frac{1}{2\mathrm{T}}$. This implies

$$\begin{aligned}
&\varphi^{t-1}_{j,\texttt{opt}_j} = \frac{1}{1 + \sum_{\tau \neq \texttt{opt}_i} \exp((\boldsymbol{x}_{i,\tau} - \boldsymbol{x}_{j,\texttt{opt}_j})^\top \boldsymbol{W}_{t-1}\boldsymbol{x}_{j,1})} > 1 - \frac{1}{2\mathrm{T}} \\
\implies \quad &\sum_{\tau \neq \texttt{opt}_j} \exp((\boldsymbol{x}_{j,\tau} - \boldsymbol{x}_{j,\texttt{opt}_j})^\top \boldsymbol{W}_{t-1}\boldsymbol{x}_{j,1}) < \frac{1}{2\mathrm{T}-1} \\
\implies \quad &(\boldsymbol{x}_{j,\tau} - \boldsymbol{x}_{j,\texttt{opt}_j})^\top \boldsymbol{W}_{t-1}\boldsymbol{x}_{j,1} \leq -\log(2\mathrm{T}-1), \qquad \forall \tau \neq \texttt{opt}_j.
\end{aligned}$$

After the gradient step at time $t-1$, for any $\tau \neq \texttt{opt}_j$, we have,

$$\begin{aligned}
(\boldsymbol{x}_{j,\tau} - \boldsymbol{x}_{j,\texttt{opt}_j})^\top \boldsymbol{W}_t \boldsymbol{x}_{j,1} &= (\boldsymbol{x}_{j,\tau} - \boldsymbol{x}_{j,\texttt{opt}_j})^\top \boldsymbol{W}_{t-1}\boldsymbol{x}_{j,1} - \eta_{t-1}(\boldsymbol{x}_{j,\texttt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(-\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_{t-1}))\boldsymbol{x}_{j,1} \\
&\leq -\log(2\mathrm{T}-1) + \eta\|\boldsymbol{x}_{j,\tau} - \boldsymbol{x}_{j,\texttt{opt}_j}\|\|\boldsymbol{x}_{j,1}\| \\
&\leq -\log(2\mathrm{T}-1) + 2\eta B^2 \\
&\leq 0,
\end{aligned}$$

where the last step follows since $\eta \leq (2B^2)^{-1}\log(2\mathrm{T}-1)$. This gives

$$\begin{aligned}
&\sum_{\tau \neq \texttt{opt}_j} \exp((\boldsymbol{x}_{j,\tau} - \boldsymbol{x}_{j,\texttt{opt}_j})^\top \boldsymbol{W}_t \boldsymbol{x}_{j,1}) \leq \mathrm{T} - 1 \\
&\qquad\qquad \implies \varphi^t_{j,\texttt{opt}_j} \geq \frac{1}{\mathrm{T}}.
\end{aligned}$$

Scenario 2: At time $t-1$, $\varphi^{t-1}_{j,\texttt{opt}_j} \leq 1 - \frac{1}{2\mathrm{T}}$. This gives

$$\varphi^{t-1}_{j,\texttt{opt}_j}(1 - \varphi^{t-1}_{j,\texttt{opt}_j}) \geq \frac{2\mathrm{T}-1}{4\mathrm{T}^2} \geq \frac{1}{4\mathrm{T}}. \tag{16}$$

We will use this to show that

$$(\boldsymbol{x}_{j,\texttt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(-\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t))\boldsymbol{x}_{j,1} \geq 0. \tag{17}$$

Let $\boldsymbol{a}' = \boldsymbol{a}((\boldsymbol{x}_{j,\texttt{opt}_j} - \boldsymbol{x}_{j,\tau})\boldsymbol{x}_{j,1}^\top) = \boldsymbol{X}(\boldsymbol{x}_{j,\texttt{opt}_j} - \boldsymbol{x}_{j,\tau})\boldsymbol{x}_{j,1}^\top \boldsymbol{x}_1$. Using similar calculations as Eq. (14), we get

$$\begin{aligned}
&(\boldsymbol{x}_{j,\texttt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(-\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_{t-1}))\boldsymbol{x}_{j,1} \\
&\qquad = \frac{1}{n}\sum_{i=1}^n \ell_{t-1,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)\varphi^{t-1}_{i,\texttt{opt}_i}(1 - \varphi^{t-1}_{i,\texttt{opt}_i})\left[a'_{i,\texttt{opt}_i} - \frac{\sum_{\tau \neq \texttt{opt}_i}\varphi^{t-1}_{i,\tau}a'_{i,\tau}}{\sum_{\tau \neq \texttt{opt}_i}\varphi^{t-1}_{i,\tau}}\right].
\end{aligned} \tag{18}$$

There are two cases:

Case 1: $\mathtt{opt}_j \neq 1$. In this case, by splitting the sum over $i \in [n]$ in Eq. (18) into $i = j$ and $i \neq j$, and using non-negativity of the softmax weights, we get

$$\begin{aligned}
&(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top (-\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_{t-1}))\boldsymbol{x}_{j,1} \\
&\geq \frac{\ell_{t-1,j}}{n}(\gamma_{j,\mathtt{opt}_j} - \gamma_j)\varphi^{t-1}_{j,\mathtt{opt}_j}(1 - \varphi^{t-1}_{j,\mathtt{opt}_j})\|\boldsymbol{x}_{j,1}\|^2 \min_{\tau' \neq \mathtt{opt}_j}(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau'}) \\
&- \max_{i \neq j} \ell_{t-1,i}(\gamma_{i,\mathtt{opt}_i} - \gamma_i)\varphi^{t-1}_{i,\mathtt{opt}_i}(1 - \varphi^{t-1}_{i,\mathtt{opt}_i})|\boldsymbol{x}_{i,1}^\top \boldsymbol{x}_{j,1}| \max_{\tau' \neq \mathtt{opt}_i} |(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{i,\mathtt{opt}_i} - \boldsymbol{x}_{i,\tau'})|.
\end{aligned} \tag{19}$$

Next, we use Assumption 1 and that $n\Upsilon\mathrm{T} \geq 1$ to get

$$\begin{aligned}
\max_{i \neq j} \max_{\tau' \neq \mathtt{opt}_i} (\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{i,\mathtt{opt}_i} - \boldsymbol{x}_{i,\tau'}) &\leq \max_{i \neq j} |\boldsymbol{x}_{j,\mathtt{opt}_j}^\top \boldsymbol{x}_{i,\mathtt{opt}_i}| + 3 \max_{\substack{i,j: i \neq j, \tau \in [\mathrm{T}], \\ \tau' \in [\mathrm{T}] \setminus \{\mathtt{opt}_j\}}} |\boldsymbol{x}_{i,\tau}^\top \boldsymbol{x}_{j,\tau'}| \\
&\leq 4 \frac{\|\boldsymbol{x}_{j,\mathtt{opt}_j}\|^2}{4} = \|\boldsymbol{x}_{j,\mathtt{opt}_j}\|^2.
\end{aligned} \tag{20}$$

Similarly,

$$\begin{aligned}
\min_{\tau' \neq \mathtt{opt}_j} (\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau'}) &\geq \|\boldsymbol{x}_{j,\mathtt{opt}_j}\|^2 - 3 \max_{\substack{\tau \in [\mathrm{T}], \\ \tau' \in [\mathrm{T}] \setminus \{\mathtt{opt}_j\}}} |\boldsymbol{x}_{j,\tau}^\top \boldsymbol{x}_{j,\tau'}| \\
&\geq \|\boldsymbol{x}_{j,\mathtt{opt}_j}\|^2 - 3\frac{\|\boldsymbol{x}_{j,\mathtt{opt}_j}\|^2}{4} = \frac{1}{4}\|\boldsymbol{x}_{j,\mathtt{opt}_j}\|^2.
\end{aligned} \tag{21}$$

Combining Eqs. (16), (20) and (21), we have

$$\begin{aligned}
\zeta &:= \frac{n \max_{i \neq j} \ell_{t,i}(\gamma_{i,\mathtt{opt}_i} - \gamma_i)\varphi^{t-1}_{i,\mathtt{opt}_i}(1 - \varphi^{t-1}_{i,\mathtt{opt}_i})}{\ell_{t,j}(\gamma_{j,\mathtt{opt}_j} - \gamma_j)\varphi^{t-1}_{j,\mathtt{opt}_j}(1 - \varphi^{t-1}_{j,\mathtt{opt}_j})} \frac{\max_{i,\tau' \neq \mathtt{opt}_i}(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{i,\mathtt{opt}_i} - \boldsymbol{x}_{i,\tau'})}{\min_{\tau' \neq \mathtt{opt}_j}(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau'})} \\
&\leq n \frac{\max_i \ell_{t-1,i}(\gamma_{i,\mathtt{opt}_i} - \gamma_i)}{\min_i \ell_{t-1,i}(\gamma_{i,\mathtt{opt}_i} - \gamma_i)} \frac{1/4}{1/(4\mathrm{T})} \frac{\|\boldsymbol{x}_{j,\mathtt{opt}_j}\|^2}{\|\boldsymbol{x}_{j,\mathtt{opt}_j}\|^2/4} \\
&\leq 4n\mathrm{T}\kappa \frac{\max_i \ell_{t-1,i}}{\min_i \ell_{t-1,i}} \\
&\leq 4n\mathrm{T}\kappa \frac{\max_i \exp(-(\varphi^{t-1}_{i,\mathtt{opt}_i}\gamma_{i,\mathtt{opt}_i} + (1 - \varphi^{t-1}_{i,\mathtt{opt}_i})\gamma_i))}{\min_i \exp(-(\varphi^{t-1}_{i,\mathtt{opt}_i}\gamma_{i,\mathtt{opt}_i} + (1 - \varphi^{t-1}_{i,\mathtt{opt}_i})\gamma_i))} \\
&\leq 4n\mathrm{T}\kappa \frac{\max_i \exp(-\gamma_i)}{\min_i \exp(-\gamma_{i,\mathtt{opt}_i})} = 4n\Upsilon\mathrm{T}.
\end{aligned} \tag{22}$$

Using Assumption 1 and Eq. (22), we have

$$\|\boldsymbol{x}_{j,1}\|^2 \geq \zeta \max_{i \neq j} |\langle \boldsymbol{x}_{i,1}, \boldsymbol{x}_{j,1} \rangle|,$$

which implies that Eq. (19) is non-negative.

Case 2: $\mathtt{opt}_j = 1$. In this case, let $\mathcal{I} := \{i \in [n] : y_i = y_j \wedge \mathtt{opt}_i = 1\}$. Using Eq. (18), we get

$$\begin{aligned}
&(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top (-\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_{t-1}))\boldsymbol{x}_{j,1} \\
&\geq \frac{\ell_{t-1,j}}{n}(\gamma_{j,\mathtt{opt}_j} - \gamma_j)\varphi^{t-1}_{j,\mathtt{opt}_j}(1 - \varphi^{t-1}_{j,\mathtt{opt}_j})\|\boldsymbol{x}_{j,1}\|^2 \min_{\tau' \neq \mathtt{opt}_j}(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau'}) \\
&+ \min_{i \neq j, i \in \mathcal{I}} \ell_{t-1,i}(\gamma_{i,\mathtt{opt}_i} - \gamma_i)\varphi^{t-1}_{i,\mathtt{opt}_i}(1 - \varphi^{t-1}_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,1}^\top \boldsymbol{x}_{j,1} \min_{\tau' \neq \mathtt{opt}_i}(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{i,\mathtt{opt}_i} - \boldsymbol{x}_{i,\tau'}) \\
&- \max_{i \neq j, i \notin \mathcal{I}} \ell_{t-1,i}(\gamma_{i,\mathtt{opt}_i} - \gamma_i)\varphi^{t-1}_{i,\mathtt{opt}_i}(1 - \varphi^{t-1}_{i,\mathtt{opt}_i})|\boldsymbol{x}_{i,1}^\top \boldsymbol{x}_{j,1}| \max_{\tau' \neq \mathtt{opt}_i} |(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{i,\mathtt{opt}_i} - \boldsymbol{x}_{i,\tau'})|.
\end{aligned} \tag{23}$$

Using Assumption 1, the second term is non-negative. For the remaining two terms, we can proceed in a similar way as the previous case to show that using Assumption 1, these terms are also non-negative.

Using Eq. (17), we get

$$(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top (\boldsymbol{W}_t - \boldsymbol{W}_{t-1})\boldsymbol{x}_{j,1} = \eta \frac{(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top (-\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_{t-1}))\boldsymbol{x}_{j,1}}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_{t-1})\|} \geq 0.$$

Then, for any $\tau \neq \mathtt{opt}_j$, we have

$$\frac{\varphi^t_{j,\mathtt{opt}_j}/\varphi^t_{j,\tau}}{\varphi^{t-1}_{j,\mathtt{opt}_j}/\varphi^{t-1}_{j,\tau}} = \exp\left((\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top (\boldsymbol{W}_t - \boldsymbol{W}_{t-1})\boldsymbol{x}_{j,1}\right) \geq 1.$$

By telescoping, we get $\frac{\varphi^t_{j,\mathtt{opt}_j}}{\varphi^t_{j,\tau}} \geq \frac{\varphi^0_{j,\mathtt{opt}_j}}{\varphi^0_{j,\tau}}$, which implies that $\varphi^t_{j,\mathtt{opt}_j} \geq \varphi^0_{j,\mathtt{opt}_j}$.

□

Lemma 4. *Under Assumption 2, for $\boldsymbol{W}_{mm}$ as defined in (W-SVM) and any $\boldsymbol{W}$,*

$$\left\langle -\frac{\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W})}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W})\|_F}, \frac{\boldsymbol{W}_{mm}}{\|\boldsymbol{W}_{mm}\|} \right\rangle \geq (2B^2\Lambda)^{-1} > 0.$$

Proof. First, using Eq. (14) for $\boldsymbol{W}_{\mathrm{mm}}$, we have

$$\begin{aligned} \langle -\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}), \boldsymbol{W}_{\mathrm{mm}} \rangle &\geq \min_i (a^*_{i,\mathtt{opt}_i} - \max_{\tau \neq \mathtt{opt}_i} a^*_{i,\tau}) \frac{1}{n} \sum_{i=1}^n \ell_i (\gamma_{i,\mathtt{opt}_i} - \gamma_i)\varphi_{i,\mathtt{opt}_i}(1 - \varphi_{i,\mathtt{opt}_i}) \\ &\geq \frac{1}{n} \sum_{i=1}^n \ell_i (\gamma_{i,\mathtt{opt}_i} - \gamma_i)\varphi_{i,\mathtt{opt}_i}(1 - \varphi_{i,\mathtt{opt}_i}), \end{aligned} \tag{24}$$

where $\boldsymbol{a}^*_i = \boldsymbol{X}_i \boldsymbol{W}_{\mathrm{mm}} \boldsymbol{x}_{i,1}$, and the second inequality follows by the definition of $\boldsymbol{W}_{\mathrm{mm}}$.

Let $\hat{\boldsymbol{W}} := -\frac{\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W})}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W})\|_F}$, and $\hat{\boldsymbol{a}}_i = \boldsymbol{X}_i \hat{\boldsymbol{W}} \boldsymbol{x}_{i,1}$, then by similar calculations as Eq. (14), we have

$$\begin{aligned} \|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W})\|_F &= \langle -\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}), \hat{\boldsymbol{W}} \rangle \\ &= \frac{1}{n} \sum_{i=1}^n \ell_i (\gamma_{i,\mathtt{opt}_i} - \gamma_i)\varphi_{i,\mathtt{opt}_i}(1 - \varphi_{i,\mathtt{opt}_i}) \left[\hat{a}_{i,\mathtt{opt}_i} - \frac{\sum_{\tau \neq \mathtt{opt}_i} \varphi_{i,\tau} \hat{a}_{i,\tau}}{\sum_{\tau \neq \mathtt{opt}_i} \varphi_{i,\tau}} \right] \\ &\leq \max_i (\hat{a}_{i,\mathtt{opt}_i} - \min_{\tau \neq \mathtt{opt}_i} \hat{a}_{i,\tau}) \frac{1}{n} \sum_{i=1}^n \ell_i (\gamma_{i,\mathtt{opt}_i} - \gamma_i)\varphi_{i,\mathtt{opt}_i}(1 - \varphi_{i,\mathtt{opt}_i}) \\ &\leq 2B^2 \frac{1}{n} \sum_{i=1}^n \ell_i (\gamma_{i,\mathtt{opt}_i} - \gamma_i)\varphi_{i,\mathtt{opt}_i}(1 - \varphi_{i,\mathtt{opt}_i}), \end{aligned} \tag{25}$$

where for the last step we use

$$\begin{aligned} \max_i (\hat{a}_{i,\mathtt{opt}_i} - \min_{\tau \neq \mathtt{opt}_i} \hat{a}_{i,\tau}) \leq 2 \max_{i,\tau} |\hat{a}_{i,\tau}| &= 2 \max_{i,\tau} |\boldsymbol{x}^\top_{i,\tau} \hat{\boldsymbol{W}} \boldsymbol{x}_{i,1}| \\ &\leq 2 \max_i \|\boldsymbol{X}_i\|_{2,\infty} \|\boldsymbol{x}_{i,1}\| \|\hat{\boldsymbol{W}}\| \leq 2B^2. \end{aligned}$$

Using Eqs. (24) and (25), we get

$$\left\langle -\frac{\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W})}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W})\|_F}, \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|} \right\rangle \geq (2B^2\Lambda)^{-1}.$$

□

Lemma 5 (Iterate Norm)**.** *Using the updates in Eq.* (2)*, under the conditions of Lemma 2 and Assumption 2, at any time $t > 0$, we have*

$$2\eta((4B^2\Lambda)^{-1} \vee 1)t \geq \|\boldsymbol{W}_t\|_F \geq \eta(4B^2\Lambda)^{-1}t. \tag{26}$$

Proof. Using Eq. (2) and Lemma 4 with $\boldsymbol{W} = \boldsymbol{W}_t$, we have

$$\begin{aligned}
\|\boldsymbol{W}_t\| &= \left\| \boldsymbol{W}_0 - \sum_{t'=0}^{t-1} \eta_{t'} \nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t) \right\| \\
&\geq \left\langle \boldsymbol{W}_0, \overline{\boldsymbol{W}}_{\mathrm{mm}} \right\rangle + \sum_{t'=0}^{t-1} \eta \left\langle -\frac{\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t)}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t)\|_F}, \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|} \right\rangle \\
&\geq \eta(2B^2\Lambda)^{-1}t - \left| \left\langle \boldsymbol{W}_0, \overline{\boldsymbol{W}}_{\mathrm{mm}} \right\rangle \right|.
\end{aligned}$$

Under the conditions of Lemma 2,

$$\left| \left\langle \boldsymbol{W}_0, \overline{\boldsymbol{W}}_{\mathrm{mm}} \right\rangle \right| \leq \|\boldsymbol{W}_0\|_F \leq \eta(4B^2\Lambda)^{-1},$$

which gives the required result for $t > 0$.

Similarly, for the upper bound we have

$$\begin{aligned}
\|\boldsymbol{W}_t\|_F &= \left\| \boldsymbol{W}_0 - \sum_{t'=0}^{t-1} \eta_{t'} \nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t) \right\|_F \\
&\leq \|\boldsymbol{W}_0\| + \sum_{t'=0}^{t-1} \eta \\
&\leq \eta(4B^2\Lambda)^{-1} + \eta t.
\end{aligned}$$

□

Lemma 6 (Data Example 1)**.** *Using $\boldsymbol{u}_* = \frac{1}{U^2}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$ and $d \gtrsim \Omega(n^2 \mathrm{T}^2 \log\left(\frac{4n}{\delta}\right))$, data generated as per Example 1 satisfies Ass. 1 and 2 with probability at least $1 - \delta$, for any $\delta \in (0, 1)$.*

Proof. Using $\boldsymbol{u}_* = \frac{1}{U^2}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$, for all $i \in [n]$ and $\tau \neq \texttt{opt}_i$, we have

$$\gamma_{i,\texttt{opt}_i} = y\boldsymbol{u}_*^\top \boldsymbol{x}_{i,\texttt{opt}_i} = 1, \quad \gamma_{i,\tau} = y\boldsymbol{u}_*^\top \boldsymbol{x}_{i,\tau} = 0.$$

Using this, we have

$$\Upsilon = \frac{\max_i(\gamma_{i,\texttt{opt}_i} - \gamma_i)}{\min_i(\gamma_{i,\texttt{opt}_i} - \gamma_i)} \frac{\min_i \exp(\gamma_{i,\texttt{opt}_i})}{\max_i \exp(\gamma_i)} = e.$$

Next, by using Bernstein inequality, we know with probability at least $1 - \frac{\delta}{2n}$, for every $i \in [n]$ and $\tau \neq \texttt{opt}_i$ we have

$$\begin{aligned}
\left| \|\boldsymbol{x}_{i,\texttt{opt}_i}\|^2 - \sigma^2 d - U^2 \right| &\leq c\sigma^2 \sqrt{d \log\left(\frac{4n}{\delta}\right)} \leq \frac{\sigma^2 d}{2}, \\
\left| \|\boldsymbol{x}_{i,\tau}\|^2 - \rho^2 d \right| &\leq c\rho^2 \sqrt{d \log\left(\frac{4n}{\delta}\right)} \leq \frac{\rho^2 d}{2}.
\end{aligned}$$

where $c > 0$ is an absolute constant and we use $d \geq 4c^2 \log\left(\frac{4n}{\delta}\right)$.

Further, for any $i, j \in [n], i \neq j$ and any $\tau, \tau' \in [\mathrm{T}]$. By applying Bernstein's inequality, each of the following is true with probability at least $1 - \frac{\delta}{2n^2}$,

$$\begin{aligned}
|\langle \boldsymbol{x}_{i,\texttt{opt}_i}, \boldsymbol{x}_{j,\texttt{opt}_j} \rangle - \langle \boldsymbol{\mu}_{y_i}, \boldsymbol{\mu}_{y_j} \rangle| \leq 2\sigma^2 \sqrt{d \log\left(\frac{4n}{\delta}\right)}, \quad |\langle \boldsymbol{x}_{i,\tau'}, \boldsymbol{x}_{j,\tau} \rangle| \leq 2\rho^2 \sqrt{d \log\left(\frac{4n}{\delta}\right)}, \\
|\langle \boldsymbol{x}_{i,\texttt{opt}_i}, \boldsymbol{x}_{j,\tau} \rangle| \leq 2\sigma\rho \sqrt{d \log\left(\frac{4n}{\delta}\right)},
\end{aligned}$$

where $\tau' \neq \texttt{opt}_i, \tau \neq \texttt{opt}_j$.

In order to satisfy Assumption 1, ignoring the log factors, we require

$$\begin{aligned}(U^2 + \sigma^2 d) \wedge \rho^2 d \ &\gtrsim\ nT(\rho^2\sqrt{d} \vee \sigma\rho\sqrt{d})\\ U^2 + \sigma^2\sqrt{d} \ &\gtrsim\ nT(\sigma^2\sqrt{d} \vee \rho^2\sqrt{d} \vee \sigma\rho\sqrt{d})\end{aligned}$$

The above inequalities hold if we have $d = \widetilde{\Omega}(n^2\mathrm{T}^2)$ and $U^2 \geq \widetilde{\Omega}((\sigma \vee \rho)^2 n\mathrm{T}\sqrt{d})$. The proof finishes by the application of a union bound. □

We now state the key Lemma used in the proof of Thm. 1 below, with the expression for R_ϵ.

Lemma 7. *Under the conditions of Lemma 2 and Assumptions 1 and 2, for any $\epsilon \in (0,1)$, there exists*

$$R_\epsilon := 2\Lambda\epsilon^{-1}(\log(4n(B^2\Lambda)^{-1}\Upsilon\mathrm{T}^3\epsilon^{-1}) \vee 5\log(20\mathrm{T}B^2\Lambda\epsilon^{-1})),$$

such that for every t where $\|\boldsymbol{W}_t\| \geq R_\epsilon$,

$$\left\langle -\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t), \frac{\boldsymbol{W}_{mm}}{\|\boldsymbol{W}_{mm}\|}\right\rangle \geq (1-\epsilon)\left\langle -\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t), \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}\right\rangle. \tag{27}$$

Remark 1. *Lem. 7 improves Lem. 10 of (Tarzanagh et al., 2023a), which only holds for $n=1$, i.e., for a single training sample (same restriction as (Tarzanagh et al., 2023b)). Our key idea to extend the result to $n \geq 1$ is to divide the samples into three sets, based on whether the constraints in* (W-SVM) *are violated ($\mathcal{I}_1$), satisfied with a small margin ($\mathcal{I}_2$) or well satisfied ($\mathcal{I}_3$). We prove a tighter version for samples in $\mathcal{I}_1$ and the residual allows us to prove the result for the sum over all samples.*

Proof. Let $\tilde{\boldsymbol{a}}_i := \boldsymbol{X}_i\widetilde{\boldsymbol{W}}_t\boldsymbol{x}_{i,1}$, $\boldsymbol{a}_i^* := \boldsymbol{X}_i\boldsymbol{W}_{\mathrm{mm}}\boldsymbol{x}_{i,1}$. We consider two scenarios based on whether $\|\widetilde{\boldsymbol{W}}_t - \boldsymbol{W}_{\mathrm{mm}}\| \leq \frac{\nu}{2B^2}$ (Scenario 1) or not (Scenario 2), where

$$\widetilde{\boldsymbol{W}}_t := \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}\|\boldsymbol{W}_{\mathrm{mm}}\|,\ \ \nu := \frac{\epsilon}{(1-\epsilon)}. \tag{28}$$

Scenario 1 of the proof is the same as (Tarzanagh et al., 2023a). Below, we consider Scenario 2.

Using the definition of ν and similar calculations to Eq. (14), Eq. (27) translates to

$$\frac{1}{n}\sum_{i=1}^{n}\ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)\varphi^t_{i,\texttt{opt}_i}(1-\varphi^t_{i,\texttt{opt}_i})\tilde{h}_{t,i} \leq (1+\nu)\frac{1}{n}\sum_{i=1}^{n}\ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)\varphi^t_{i,\texttt{opt}_i}(1-\varphi^t_{i,\texttt{opt}_i})h^*_{t,i}, \tag{29}$$

where $\tilde{h}_{t,i} := \tilde{a}_{i,\texttt{opt}_i} - \frac{\sum_{\tau\neq\texttt{opt}_i}\varphi^t_{i,\tau}\tilde{a}_{i,\tau}}{\sum_{\tau\neq\texttt{opt}_i}\varphi^t_{i,\tau}}$ and $h^*_{t,i}$ is defined similarly.

For any sample $i \in [n]$, let

$$\begin{aligned}\delta_i^{\max}(\epsilon,\omega) &:= \max_{\tau\notin\{\texttt{opt}_i\}} \tilde{a}_{i,\texttt{opt}_i} - \tilde{a}_{i,\tau} - 1,\\ \delta_i^{\min}(\epsilon,\omega) &:= \min_{\tau\notin\{\texttt{opt}_i\}} \tilde{a}_{i,\texttt{opt}_i} - \tilde{a}_{i,\tau} - 1.\end{aligned}$$

Define the following sets

$$\begin{aligned}\mathcal{I}_1 &:= \{i \in [n] \mid \delta_i^{\min} \leq 0.3\nu\},\\ \mathcal{I}_2 &:= \{i \in [n] \mid 0.3\nu < \delta_i^{\min} \leq 0.8\nu\},\\ \mathcal{I}_3 &:= \{i \in [n] \mid \delta_i^{\min} \geq 0.8\nu\}.\end{aligned}$$

Intuitively, $\mathcal{I}_1$ is the set of samples for which some (W-SVM) constraints are either violated or all constraints are barely satisfied. Similarly, $\mathcal{I}_3$ contains samples where all the (W-SVM) constraints are satisfied. Finally,

$\mathcal{I}_2$ makes up the rest of the samples.

First, we will show that for any $i \in \mathcal{I}_1$,

$$\left[\tilde{a}_{i,\mathtt{opt}_i} - \frac{\sum_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}\tilde{a}_{i,\tau}}{\sum_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}}\right] \le (1+0.5\nu)\left[a^*_{i,\mathtt{opt}_i} - \frac{\sum_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}a^*_{i,\tau}}{\sum_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}}\right]. \tag{30}$$

Next, we will show that for any $i \in \mathcal{I}_2$, we have $\tilde{h}_{t,i} \le (1+\nu)h^*_{t,i}$, i.e.

$$\left[\tilde{a}_{i,\mathtt{opt}_i} - \frac{\sum_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}\tilde{a}_{i,\tau}}{\sum_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}}\right] \le (1+\nu)\left[a^*_{i,\mathtt{opt}_i} - \frac{\sum_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}a^*_{i,\tau}}{\sum_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}}\right]. \tag{31}$$

Finally, notice that in order to complete the proof for Eq. (29), using Eqs. (30) and (31) it suffices to show that

$$\begin{aligned}&\sum_{i\in\mathcal{I}_3}\ell_{t,i}(\gamma_{i,\mathtt{opt}_i}-\gamma_i)\varphi^t_{i,\mathtt{opt}_i}(1-\varphi^t_{i,\mathtt{opt}_i})(\tilde{h}_{t,i}-(1+\nu)h^*_{t,i})\\ &\qquad\le 0.5\nu\sum_{i\in\mathcal{I}_1}\ell_{t,i}(\gamma_{i,\mathtt{opt}_i}-\gamma_i)\varphi^t_{i,\mathtt{opt}_i}(1-\varphi^t_{i,\mathtt{opt}_i})h^*_{t,i}.\end{aligned} \tag{32}$$

Using $h^*_{t,i} \ge 1$, Eq. (32) follows when

$$\sum_{i\in\mathcal{I}_3}\ell_{t,i}(\gamma_{i,\mathtt{opt}_i}-\gamma_i)\varphi^t_{i,\mathtt{opt}_i}(1-\varphi^t_{i,\mathtt{opt}_i})(\tilde{h}_{t,i}-1-\nu) \le 0.5\nu\sum_{i\in\mathcal{I}_1}\ell_{t,i}(\gamma_{i,\mathtt{opt}_i}-\gamma_i)\varphi^t_{i,\mathtt{opt}_i}(1-\varphi^t_{i,\mathtt{opt}_i}). \tag{33}$$

Case 1: Sample $i \in \mathcal{I}_1$, i.e. $\delta_i^{\min} \le 0.3\nu$.

If $\delta_i^{\max} < 0.5\nu$, (30) follows directly. For the rest of the samples i where $\delta_i^{\max} \ge 0.5\nu$, let

$$\mathcal{N} := \{\tau \in [\mathrm{T}] : a_{i,\mathtt{opt}_i} - a_{i,\tau} \le 1 + 0.4\nu\}.$$

Let $R' := \frac{\|\boldsymbol{W}_t\|}{\|\boldsymbol{W}_{\mathrm{mm}}\|}$, then by definition of $\mathcal{N}, \delta_{\min}$, we have

$$\begin{aligned}\frac{\sum\limits_{\tau\neq\mathtt{opt}_i,\tau\notin\mathcal{N}}\varphi^t_{i,\tau}}{\sum\limits_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}} &\le \frac{\mathrm{T}\max_{\tau\neq\mathtt{opt}_i,\tau\notin\mathcal{N}}\varphi^t_{i,\tau}}{\varphi^t_{i,\tau}} \le \frac{\mathrm{T}\exp(-R'(1+0.4\nu))}{\exp(-R'(1+\delta_i^{\min}))}\\ &\le \mathrm{T}\exp(-0.1R'\nu).\end{aligned}$$

Combining these, we have

$$\begin{aligned}\frac{\sum\limits_{\tau\neq\mathtt{opt}_i}(\tilde{a}_{i,\mathtt{opt}_i}-\tilde{a}_{i,\tau})\varphi^t_{i,\tau}}{\sum\limits_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}} &\le (1+0.4\nu)\frac{\sum\limits_{\tau\in\mathcal{N}}\varphi^t_{i,\tau}}{\sum\limits_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}} + (1+\delta_i^{\max})\frac{\sum\limits_{\tau\neq\mathtt{opt}_i,\tau\notin\mathcal{N}}\varphi^t_{i,\tau}}{\sum\limits_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}}\\ &\le (1+0.4\nu)\frac{\sum\limits_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau} - \sum\limits_{\tau\neq\mathtt{opt}_i\notin\mathcal{N}}\varphi^t_{i,\tau}}{\sum\limits_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}} + (1+\delta_i^{\max})\frac{\sum\limits_{\tau\neq\mathtt{opt}_i,\tau\notin\mathcal{N}}\varphi^t_{i,\tau}}{\sum\limits_{\tau\neq\mathtt{opt}_i}\varphi^t_{i,\tau}}\\ &\le 1 + 0.4\nu + (\delta_i^{\max} - 0.4\nu)\mathrm{T}\exp(-0.1R'\nu).\end{aligned}$$

To satisfy (30), we need $(\delta_i^{\max} - 0.4\nu)\mathrm{T}\exp(-0.1R'\nu) \le 0.1\nu$, which is true when

$$R' \ge (0.1\nu)^{-1}\log(\mathrm{T}(\delta_i^{\max}/(0.1\nu) - 4)).$$

This is true since $R' \ge 10\epsilon^{-1}\log(10\mathrm{T}\max_{i\in[n]}\delta_i^{\max}/\epsilon)$, where we use the definition of ν from Eq. (28).

Case 2: Sample $i \in \mathcal{I}_2$, i.e. $0.3\nu < \delta^{\min} \le 0.8\nu$.

If $\delta_i^{\max} < \nu$, (31) follows directly. For the rest of the samples where $\delta_i^{\max} \ge \nu$, we can proceed in a similar way as Case 1. We use a threshold of 0.9ν to split the tokens, and obtain

$$\frac{\sum_{\tau \neq \texttt{opt}_i} (a_{i,\texttt{opt}_i} - a_{i,\tau})\varphi_{t,i,\tau}}{\sum_{\tau \neq \texttt{opt}_i} \varphi_{t,i,\tau}} \le 1 + \nu,$$

when $R' \ge 10\epsilon^{-1} \log(10\mathrm{T} \max_{i \in [n]} \delta_i^{\max}/\epsilon)$.

Case 3: Sample $i \in \mathcal{I}_3$, i.e. $\delta_i^{\min} \ge 0.8\nu$.

As $\|\widetilde{\boldsymbol{W}}_t - \boldsymbol{W}_{\mathrm{mm}}\| \ge \frac{\nu}{2B^2}$, at least one sample $i \in [n]$ violates the (W-SVM) constraints, i.e. $|\mathcal{I}_1| \ge 1$. We have

$$\begin{aligned}\sum_{i \in \mathcal{I}_1} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)\varphi^t_{i,\texttt{opt}_i}(1 - \varphi^t_{i,\texttt{opt}_i}) &\ge \min_{i \in \mathcal{I}_1} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i) \min_t (\varphi^t_{i,\texttt{opt}_i})^2 \sum_{\tau \neq \texttt{opt}_i} \varphi^t_{i,\tau}/\varphi^t_{i,\texttt{opt}_i} \\ &\ge \min_{i \in \mathcal{I}_1} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)(4\mathrm{T})^{-2} \exp(-(1 + 0.3\nu)R'), \end{aligned} \tag{34}$$

where the last inequality follows by using Lemmas 2 and 3, and that $\delta_i^{\min} \le 0.3\nu$ for any $i \in \mathcal{I}_1$. Also,

$$\begin{aligned}\sum_{i \in \mathcal{I}_3} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)\varphi^t_{i,\texttt{opt}_i}(1 - \varphi^t_{i,\texttt{opt}_i}) &\le (n-1) \max_{i \in \mathcal{I}_3} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i) \max_{i \in \mathcal{I}_3} \sum_{\tau \neq \texttt{opt}_i} \varphi^t_{i,\tau}/\varphi^t_{i,\texttt{opt}_i} \\ &\le (n-1) \max_{i \in \mathcal{I}_3} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)(\mathrm{T} - 1) \exp(-(1 + 0.8\nu)R'), \end{aligned} \tag{35}$$

where the last inequality follows because $0.8\nu \le \min_{i \in \mathcal{I}_3} \delta_i^{\min}$. Combining Eqs. (34) and (35) gives us (33) when

$$\exp(0.5\nu R') \ge 0.5\nu(2B^2\Lambda)^{-1}(n-1)(\mathrm{T}-1)(4\mathrm{T})^2 \frac{\max_{i \in \mathcal{I}_3} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)}{\min_{i \in \mathcal{I}_1} \ell_{t,i}(\gamma_{i,\texttt{opt}_i} - \gamma_i)}, \tag{36}$$

which is true when

$$\exp(0.5\nu R') \ge 4n(B^2\Lambda)^{-1}\Upsilon\mathrm{T}^3\epsilon^{-1},$$

which is true when $R' \ge 2\epsilon^{-1}\log(4n(B^2\Lambda)^{-1}\Upsilon\mathrm{T}^3\epsilon^{-1})$.

Combining all conditions on R', and using $\delta_i^{\max} \le 2B^2\Lambda$ for all $i \in [n]$, we get the desired result since

$$R_\epsilon \ge \Lambda^{-1} = 2\Lambda\epsilon^{-1}(\log(4n(B^2\Lambda)^{-1}\Upsilon\mathrm{T}^3\epsilon^{-1}) \vee 5\log(20\mathrm{T}B^2\Lambda\epsilon^{-1})).$$

□

A.3 Proof of Theorem 1

We first restate Theorem 1, this time with the exact constants.

Theorem 5 (IB Rate)**.** *Under the conditions of Lemma 2 and Assumptions 1 and 2, using the updates in Eq.* (2) *with* $\eta \le (2B^2)^{-1}\log(2\mathrm{T}-1)$*, for any* $t \ge t_0 := \left(\frac{10B\Lambda}{\sqrt{\eta}}\right)^3 \vee \log\left(\frac{\eta\mathrm{T}(n\Upsilon(B^2\Lambda)^{-2}\mathrm{T}^2 \vee 5)}{20\Lambda}\right)$,

$$\left\langle \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}, \frac{\boldsymbol{W}_{mm}}{\|\boldsymbol{W}_{mm}\|} \right\rangle \ge 1 - \frac{C(\eta, B, \Lambda, t_0)(\log t)^2}{t},$$

where $C(\eta, B, \Lambda, t_0) = 4\eta^{-1}B^2\Lambda\left(\|\boldsymbol{W}_{t_0}\| - \langle \boldsymbol{W}_{t_0}, \overline{\boldsymbol{W}}_{mm}\rangle + 2B^2\Lambda(40\Lambda + \eta)\right)$.

Remark 2 (Comparison to (Tarzanagh et al., 2023a).)**.** *Thm. 1 establishes* finite-time *convergence of NGD to* $\boldsymbol{W}_{mm}$ *starting from* any *initialization direction. This marks a clear improvement over prior work (Tarzanagh et al., 2023a), which demonstrates that GD converges* asymptotically *($t \to \infty$) to* $\boldsymbol{W}_{mm}$ *only under an appropriate* local *initialization direction. The only previously known global convergence result in Tarzanagh et al. (2023b;a) requires Ass. 2 and* $n = 1$*, limiting its applicability to training with a single sample. Thm. 1 addresses this limitation and additionally provides convergence rates. For completeness, we demonstrate in Thm. 6 in the Appendix that GD (with constant step size) also converges globally to* $\boldsymbol{W}_{mm}$ *under the conditions specified in Thm. 1, without requiring* $n = 1$.

Proof. First, we use Lemma 5 to show that Lemma 7 is true for any $t \geq t_0$, by selecting $\epsilon = \epsilon_t := (80B^2\Lambda^2 \log t)(\eta t)^{-1}$. Since $\log t < 1.25t^{1/3}$ for $t > 0$, $\epsilon_t \leq 1$ for $t \geq (10B\Lambda\eta^{-1/2})^3$. We also have

$$\begin{aligned}
&2\Lambda\epsilon^{-1}(\log(4n\Upsilon(B^2\Lambda)^{-1}\mathrm{T}^3\epsilon^{-1}) \vee 5\log(20\mathrm{T}B^2\Lambda\epsilon^{-1})) \leq 10\Lambda\epsilon^{-1}\log((4n\Upsilon(B^2\Lambda)^{-1}\mathrm{T}^3 \vee 20\mathrm{T}B^2\Lambda)\epsilon^{-1})\\
&\quad = 10\Lambda\frac{\eta t}{80B^2\Lambda^2\log t}\log\left((4n\Upsilon(B^2\Lambda)^{-1}\mathrm{T}^3 \vee 20\mathrm{T}B^2\Lambda)\frac{\eta t}{80B^2\Lambda^2\log t}\right)\\
&\quad \leq \frac{\eta(4B^2\Lambda)^{-1}t}{\log t}\left(\log t \vee \log\left(\frac{\eta\mathrm{T}(n\Upsilon(B^2\Lambda)^{-2}\mathrm{T}^2 \vee 5)}{20\Lambda}\right)\right)\\
&\quad \leq \eta(4B^2\Lambda)^{-1}t.
\end{aligned}$$

Thus, $\|\boldsymbol{W}_t\| \geq R$ for $t \geq t_0$ and we can use Lemma 7 with $\epsilon = \epsilon_t$, *i.e.*,

$$\left\langle -\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t), \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|}\right\rangle \geq (1-\epsilon_t)\left\langle -\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t), \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}\right\rangle.$$

Using the update Eq. (2), we have

$$\begin{aligned}
&\left\langle \boldsymbol{W}_{t+1} - \boldsymbol{W}_t, \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|}\right\rangle \geq (1-\epsilon_t)\left\langle \boldsymbol{W}_{t+1} - \boldsymbol{W}_t, \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}\right\rangle\\
&= \frac{1}{2\|\boldsymbol{W}_t\|}(\|\boldsymbol{W}_{t+1}\|^2 - \|\boldsymbol{W}_t\|^2 - \|\boldsymbol{W}_{t+1} - \boldsymbol{W}_t\|^2) - \epsilon_t\left\langle \boldsymbol{W}_{t+1} - \boldsymbol{W}_t, \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}\right\rangle\\
&= \frac{1}{2\|\boldsymbol{W}_t\|}(\|\boldsymbol{W}_{t+1}\|^2 - \|\boldsymbol{W}_t\|^2) - \frac{\eta^2}{2\|\boldsymbol{W}_t\|} + \epsilon_t\eta\left\langle \nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t), \overline{\boldsymbol{W}}_t\right\rangle && \text{(using Eq. (2))}\\
&\geq \|\boldsymbol{W}_{t+1}\| - \|\boldsymbol{W}_t\| - \frac{\eta^2}{2\|\boldsymbol{W}_t\|} - \epsilon_t\eta && \left(\text{since } \frac{a^2-b^2}{2b} \geq a - b \ \forall\ a, b > 0\right)\\
&\geq \|\boldsymbol{W}_{t+1}\| - \|\boldsymbol{W}_t\| - \frac{80B^2\Lambda^2\log t}{t} - 2\eta B^2\Lambda(t+1)^{-1} && \text{(substituting } \epsilon_t\text{, using Lemma 5).}
\end{aligned}$$

Summing over $t \geq t_0$, we get

$$\left\langle \boldsymbol{W}_t, \overline{\boldsymbol{W}}_{\mathrm{mm}}\right\rangle \geq \|\boldsymbol{W}_t\| - \|\boldsymbol{W}_{t_0}\| + \left\langle \boldsymbol{W}_{t_0}, \overline{\boldsymbol{W}}_{\mathrm{mm}}\right\rangle - 80B^2\Lambda^2\sum_{t'=t_0}^{t}(\log t')(t')^{-1} - 2\eta B^2\Lambda\sum_{t'=t_0}^{t}(t'+1)^{-1}. \tag{37}$$

We bound the last two terms as follows

$$\begin{aligned}
\sum_{t'=t_0}^{t}(\log t')(t')^{-1} &\leq \log t\sum_{t'=t_0}^{t}(t')^{-1} \leq (\log t)^2,\\
\sum_{t'=t_0}^{t}(t'+1)^{-1} &\leq \log t.
\end{aligned}$$

Using these in Eq. (37) and dividing by $\|\boldsymbol{W}_t\|$ throughout, we get

$$\begin{aligned}
\left\langle \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}, \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|}\right\rangle &\geq 1 - \frac{1}{\|\boldsymbol{W}_t\|}\left(\|\boldsymbol{W}_{t_0}\| - \left\langle \boldsymbol{W}_{t_0}, \overline{\boldsymbol{W}}_{\mathrm{mm}}\right\rangle + 2B^2\Lambda(40\Lambda + \eta)(\log t)^2\right)\\
&\geq 1 - \frac{C(\eta, B, \Lambda, t_0)(\log t)^2}{t},
\end{aligned}$$

where the last step follows by Lemma 5. □

A.4 Optimal Rate for NGD

In Th. 1, we showed an upper bound on the correlation $\langle \overline{\boldsymbol{W}}_t, \overline{\boldsymbol{W}}_{\mathrm{mm}} \rangle \le 1 - \tilde{\Omega}(t^{-1})$ for NGD. In this section, we consider the example used in Sec. 3.3 to derive a lower bound on the correlation.

We restate the example here for convenience. Let $n = 1$, $\mathrm{T} = 2$, $y = 1$, $\boldsymbol{x}_1 = [1, 0]^\top$, $\boldsymbol{x}_2 = [0, 0]^\top$ and $\boldsymbol{u}_* = [1, 0]^\top$. As we saw in Sec. 3.3, $\texttt{opt} = 1$, $\gamma_{\texttt{opt}} - \gamma = 1$, $\boldsymbol{W}_{\mathrm{mm}} = (\boldsymbol{x}_1 - \boldsymbol{x}_2)\boldsymbol{x}_1^\top = \boldsymbol{X}$ and $\Lambda = 1$. Using this, we can write

$$\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t) = -\widehat{L}(\boldsymbol{W}_t)\varphi^t_{\texttt{opt}}(1 - \varphi^t_{\texttt{opt}})\boldsymbol{W}_{\mathrm{mm}}, \text{ and } \boldsymbol{W}_{t+1} = \boldsymbol{W}_0 + \eta G_t \boldsymbol{W}_{\mathrm{mm}},$$

where $G_t \boldsymbol{W}_{\mathrm{mm}} \coloneqq \sum_{t'=0}^{t} \frac{\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_{t'})}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_{t'})\|} = (t+1)\boldsymbol{W}_{\mathrm{mm}}$.

Clearly, $\|\boldsymbol{W}_t\| \le \eta t + \|\boldsymbol{W}_0\| \le \mathcal{O}(t)$. Using Eq. (3), we have

$$\langle \overline{\boldsymbol{W}}_t, \overline{\boldsymbol{W}}_{\mathrm{mm}} \rangle \approx 1 - \frac{\|\boldsymbol{W}_0\|^2 - (W_0^{11})^2}{2\|\boldsymbol{W}_t\|^2} \le 1 - \Omega(t^{-2}).$$

Comparing this with the upper bound, we see that the rate in Th. 1 may not be the optimal rate for NGD. However, it is the first finite-time convergence rate for self-attention. Improving this rate is an interesting direction for future work.

A.5 Proofs for Additional Results in Section 3.2

Lemma 8 (Softmax score rate)**.** *Under the conditions of Lemma 2 and Assumptions 1 and 2, using the updates in Eq. (2), for any* $i \in [n]$ *and any* $t \ge 2^{11}(B^2\Lambda)^2 C(\eta, B, \Lambda, t_0) \vee t_0$*, where* $C(\eta, B, \Lambda, t_0) \coloneqq 4\eta^{-1}B^2\Lambda\left(\|\boldsymbol{W}_{t_0}\| - \langle \boldsymbol{W}_{t_0}, \overline{\boldsymbol{W}}_{mm}\rangle + 2B^2\Lambda(40\Lambda + \eta)\right)$*,* $t_0 \coloneqq \left(\frac{10B\Lambda}{\sqrt{\eta}}\right)^3 \vee \log\left(\frac{\eta\mathrm{T}(n\Upsilon(B^2\Lambda)^{-2}\mathrm{T}^2 \vee 5)}{20\Lambda}\right)$*,*

$$\varphi^t_{i,opt_i} \ge \frac{1}{1 + (T-1)\exp(-\eta(8B^2\Lambda^2)^{-1}t)}.$$

Proof. Using Theorem 1 and Lemma 5, we have

$$\begin{aligned}
(\boldsymbol{x}_{i,\tau} - \boldsymbol{x}_{i,\texttt{opt}_i})^\top(\boldsymbol{W}_t - \overline{\boldsymbol{W}}_{\mathrm{mm}}\|\boldsymbol{W}_t\| + \overline{\boldsymbol{W}}_{\mathrm{mm}}\|\boldsymbol{W}_t\|)\boldsymbol{x}_{i,1} &\le 2B^2\|\boldsymbol{W}_t\|\|\overline{\boldsymbol{W}_t} - \overline{\boldsymbol{W}}_{\mathrm{mm}}\| - \frac{1}{\|\boldsymbol{W}_{\mathrm{mm}}\|}\|\boldsymbol{W}_t\| \\
&\le 2\sqrt{2}B^2\frac{\sqrt{C(\eta, B, \Lambda, t_0)}}{t^{1/2}}(2\eta t) - \eta(4B^2\Lambda^2)^{-1}t \\
&= 4\eta B^2\sqrt{2C(\eta, B, \Lambda, t_0)}\sqrt{t} - \eta(4B^2\Lambda^2)^{-1}t \\
&\le -\eta(8B^2\Lambda^2)^{-1}t, \qquad (38)
\end{aligned}$$

since $t \ge 2^{11}(B^2\Lambda)^2 C(\eta, B, \Lambda, t_0)$.

We use this to find the softmax rate as follows

$$\begin{aligned}
\varphi^t_{i,\texttt{opt}_i} &= \frac{1}{1 + \sum_{\tau \ne \texttt{opt}_i}\exp((\boldsymbol{x}_{i,\tau} - \boldsymbol{x}_{i,\texttt{opt}_i})^\top \boldsymbol{W}_t \boldsymbol{x}_{i,1})} \\
&= \frac{1}{1 + \sum_{\tau \ne \texttt{opt}_i}\exp((\boldsymbol{x}_{i,\tau} - \boldsymbol{x}_{i,\texttt{opt}_i})^\top(\boldsymbol{W}_t - \overline{\boldsymbol{W}}_{\mathrm{mm}}\|\boldsymbol{W}_t\| + \overline{\boldsymbol{W}}_{\mathrm{mm}}\|\boldsymbol{W}_t\|)\boldsymbol{x}_{i,1})} \\
&\ge \frac{1}{1 + (T-1)\exp(-\eta(8B^2\Lambda^2)^{-1}t)}.
\end{aligned}$$

□

Theorem 6 (Asymptotic Convergence of GD)**.** *Under the conditions of Lemma 2 and Assumptions 1 and 2, using the standard GD updates with $\eta \le \frac{\log(2\mathrm{T}-1)}{B^4\zeta}$, for any $t \ge t_\epsilon$ such that $\|\boldsymbol{W}_t\| \ge R_\epsilon \vee 1/2$,*

$$\left\langle \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}, \frac{\boldsymbol{W}_{mm}}{\|\boldsymbol{W}_{mm}\|} \right\rangle \ge 1 - \epsilon - \frac{C(\eta, B, \Lambda, \epsilon)}{\|\boldsymbol{W}_t\|},$$

where $C(\eta, B, \Lambda, \epsilon) = \frac{2B^2\Lambda}{\eta}(1-\epsilon)\|\boldsymbol{W}_{t_\epsilon}\| \left(1 - (1-\epsilon)^{-1}\left\langle \overline{\boldsymbol{W}}_{t_\epsilon}, \overline{\boldsymbol{W}}_{mm}\right\rangle - 2\eta\|\boldsymbol{W}_{t_\epsilon}\|^{-1}\widehat{L}(\boldsymbol{\theta}_{t_\epsilon})\right)$.

Proof. First, we show that Lemma 3 holds for standard GD updates, *i.e.*, for any $t \ge 0$ and any $j \in [n]$,

$$\varphi^t_{j,\mathtt{opt}_j} \ge \varphi^0_{j,\mathtt{opt}_j} \wedge \frac{1}{\mathrm{T}}.$$

Following similar steps as the proof of Lemma 3, we consider two scenarios,
Scenario 1: At time $t-1$, $\varphi^{t-1}_{j,\mathtt{opt}_j} > 1 - \frac{1}{2\mathrm{T}}$. This implies

$$\varphi^{t-1}_{j,\mathtt{opt}_j} \le -\log(2\mathrm{T}-1), \qquad \forall \tau \ne \mathtt{opt}_j.$$

After the gradient step at time $t-1$, for any $\tau \ne \mathtt{opt}_j$, we have,

$$\begin{aligned}
(\boldsymbol{x}_{j,\tau} - \boldsymbol{x}_{j,\mathtt{opt}_j})^\top \boldsymbol{W}_t \boldsymbol{x}_{j,1} &= (\boldsymbol{x}_{j,\tau} - \boldsymbol{x}_{j,\mathtt{opt}_j})^\top \boldsymbol{W}_{t-1}\boldsymbol{x}_{j,1} - \eta(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(-\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_{t-1}))\boldsymbol{x}_{j,1} \\
&\le -\log(2\mathrm{T}-1) + \eta\|\boldsymbol{x}_{j,\tau} - \boldsymbol{x}_{j,\mathtt{opt}_j}\|\|\boldsymbol{x}_{j,1}\|\|\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_{t-1})\| \\
&\le -\log(2\mathrm{T}-1) + \eta(2B^2)(B^2\widehat{L}(\boldsymbol{W}_{t-1})\max_i(\gamma_{i,\mathtt{opt}_i} - \gamma_i)\varphi^t_{i,\mathtt{opt}_i}(1-\varphi^t_{i,\mathtt{opt}_i})) \\
&\le -\log(2\mathrm{T}-1) + \eta B^4 \max_i \exp(-\gamma_i)\max_i(\gamma_{i,\mathtt{opt}_i} - \gamma_i) \\
&\le 0,
\end{aligned}$$

where the last step follows since $\eta \le \frac{\log(2\mathrm{T}-1)}{B^4\zeta}$. This gives $\varphi^t_{j,\mathtt{opt}_j} \ge \frac{1}{\mathrm{T}}$.
Scenario 2: At time $t-1$, $\varphi^{t-1}_{j,\mathtt{opt}_j} \le 1 - \frac{1}{2\mathrm{T}}$. The proof for this part is the same as Lemma 3.

Using this, we can easily show that Lemma 7 is true for GD updates, *i.e.*, for any $\epsilon \in (0,1)$, there exists

$$R_\epsilon := 2\Lambda\epsilon^{-1}(\log(4n(B^2\Lambda)^{-1}\Upsilon\mathrm{T}^3\epsilon^{-1}) \vee 5\log(20\mathrm{T}B^2\Lambda\epsilon^{-1})),$$

such that for every t where $\|\boldsymbol{W}_t\| \ge R_\epsilon$,

$$\left\langle -\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t), \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|} \right\rangle \ge (1-\epsilon)\left\langle -\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t), \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|} \right\rangle.$$

Once we have this, the remaining proof is the same as Theorem 4 in (Tarzanagh et al., 2023a). □

Theorem 7 (IB rate for Polyak-step)**.** *Under the conditions of Lemma 2 and Assumptions 1 and 2, using GD updates with Polyak-step, with $\eta \le (2B^2)^{-1}\omega_1\log(2\mathrm{T}-1)$, where $\omega_1 = \beta(1-\beta)\Lambda^{-1}\min_i(\gamma_{i,opt_i} - \gamma_i)$, for any $T_0 \ge t \ge t_0 := \omega_2\nu\left(\left(\frac{10B\Lambda}{\sqrt{\eta}}\right)^3 \vee \log\left(\frac{\eta\mathrm{T}(n\Upsilon(B^2\Lambda)^{-2}\mathrm{T}^2\vee 5)}{20\Lambda}\right)\right)$, where $\omega_2 = 0.5B^2\max_i(\gamma_{i,opt_i} - \gamma_i)$, such that for any $i \in [n]$, $\varphi^t_{i,opt_i} \le \beta$,*

$$\left\langle \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}, \frac{\boldsymbol{W}_{mm}}{\|\boldsymbol{W}_{mm}\|} \right\rangle \ge 1 - \frac{C(\eta, B, \Lambda, t_0)(\log t)^2}{t},$$

where $C(\eta, B, \Lambda, t_0) = 4\eta^{-1}B^2\Lambda\omega_2\nu\left(\|\boldsymbol{W}_{t_0}\| - \left\langle \boldsymbol{W}_{t_0}, \overline{\boldsymbol{W}}_{mm}\right\rangle + 2B^2\Lambda\omega_1^{-1}(40\Lambda + \eta\upsilon\omega_1^{-1}\omega_2)\right)$.

Proof. To extend our analysis and results for NGD updates to the Polyak-step, we mainly use the following lower bound on the gradient norm,

$$\begin{aligned}
\|\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t)\| &\geq \widehat{L}(\boldsymbol{W}_t)\min_i \left\langle y_i\nabla_{\boldsymbol{W}}\Phi(\boldsymbol{\theta}_t,\boldsymbol{X}_i), \frac{\boldsymbol{W}_{\text{mm}}}{\|\boldsymbol{W}_{\text{mm}}\|}\right\rangle \\
&= \Lambda^{-1}\widehat{L}(\boldsymbol{W}_t)\min_i(\gamma_{i,\texttt{opt}_i}-\gamma_i)\varphi^t_{i,\texttt{opt}_i}(1-\varphi^t_{i,\texttt{opt}_i})\left[a^*_{i,\texttt{opt}_i}-\frac{\sum_{\tau\neq\texttt{opt}_i}\varphi^t_{i,\tau}a^*_{i,\tau}}{\sum_{\tau\neq\texttt{opt}_i}\varphi^t_{i,\tau}}\right] \\
&\geq \Lambda^{-1}\widehat{L}(\boldsymbol{W}_t)\min_i(\gamma_{i,\texttt{opt}_i}-\gamma_i)\min_{i,t'\leq t}\varphi^{t'}_{i,\texttt{opt}_i}(1-\varphi^{t'}_{i,\texttt{opt}_i}) \\
&\geq \beta(1-\beta)\Lambda^{-1}\min_i(\gamma_{i,\texttt{opt}_i}-\gamma_i)\widehat{L}(\boldsymbol{W}_t) = \omega_1\widehat{L}(\boldsymbol{W}_t).
\end{aligned}$$

Note that similar lower bounds have appeared when deriving loss convergence rates of linear predictors trained with NGD on separable data (Nacson et al., 2019).

We also have the following upper bound,

$$\|\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t)\| \leq \widehat{L}(\boldsymbol{W}_t)(1/4)\max_i(\gamma_{i,\texttt{opt}_i}-\gamma_i)(2B^2) = \omega_2\widehat{L}(\boldsymbol{W}_t).$$

First, we show that Lemma 3 holds for Polyak-step updates, *i.e.*, for any $t \geq 0$ and any $j \in [n]$,

$$\varphi^t_{j,\texttt{opt}_j} \geq \varphi^0_{j,\texttt{opt}_j} \wedge \frac{1}{\mathrm{T}}.$$

Following similar steps as the proof of Lemma 3, we consider two scenarios,
Scenario 1: At time $t-1$, $\varphi^{t-1}_{j,\texttt{opt}_j} > 1-\frac{1}{2\mathrm{T}}$. This implies

$$\varphi^{t-1}_{j,\texttt{opt}_j} \leq -\log(2\mathrm{T}-1), \quad \forall\tau\neq\texttt{opt}_j.$$

After the gradient step at time $t-1$, for any $\tau\neq\texttt{opt}_j$, we have,

$$\begin{aligned}
(\boldsymbol{x}_{j,\tau}-\boldsymbol{x}_{j,\texttt{opt}_j})^\top\boldsymbol{W}_t\boldsymbol{x}_{j,1} &= (\boldsymbol{x}_{j,\tau}-\boldsymbol{x}_{j,\texttt{opt}_j})^\top\boldsymbol{W}_{t-1}\boldsymbol{x}_{j,1} - \eta_{t-1}(\boldsymbol{x}_{j,\texttt{opt}_j}-\boldsymbol{x}_{j,\tau})^\top(-\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_{t-1}))\boldsymbol{x}_{j,1} \\
&\leq -\log(2\mathrm{T}-1)+\eta\frac{\widehat{L}(\boldsymbol{W}_{t-1})-\widehat{L}^*}{\|\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_{t-1})\|^2}\|\boldsymbol{x}_{j,\tau}-\boldsymbol{x}_{j,\texttt{opt}_j}\|\|\boldsymbol{x}_{j,1}\|\|\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_{t-1})\| \\
&\leq -\log(2\mathrm{T}-1)+\eta\frac{\widehat{L}(\boldsymbol{W}_{t-1})}{\omega_1\widehat{L}(\boldsymbol{W}_{t-1})}(2B^2) \\
&= -\log(2\mathrm{T}-1)+\eta\frac{2B^2}{\omega_1} \\
&\leq 0,
\end{aligned}$$

where the last step follows since $\eta \leq \frac{\omega_1\log(2\mathrm{T}-1)}{2B^2}$. This gives $\varphi^t_{j,\texttt{opt}_j} \geq \frac{1}{\mathrm{T}}$.

Scenario 2: At time $t-1$, $\varphi^{t-1}_{j,\texttt{opt}_j} \leq 1-\frac{1}{2\mathrm{T}}$. The proof for this part is the same as Lemma 3.

Using this, we can easily show that Lemma 7 is true for Polyak-step updates, *i.e.*, for any $\epsilon \in (0,1)$, there exists

$$R_\epsilon := 2\Lambda\epsilon^{-1}(\log(4n(B^2\Lambda)^{-1}\Upsilon\mathrm{T}^3\epsilon^{-1}) \vee 5\log(20\mathrm{T}B^2\Lambda\epsilon^{-1})),$$

such that for every t where $\|\boldsymbol{W}_t\| \geq R_\epsilon$,

$$\left\langle -\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t), \frac{\boldsymbol{W}_{\text{mm}}}{\|\boldsymbol{W}_{\text{mm}}\|}\right\rangle \geq (1-\epsilon)\left\langle -\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{W}_t), \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}\right\rangle. \tag{39}$$

Next, following similar steps as the proof of Lemma 5, the iterate norm growth can be characterized as follows. Using Lemma 4 with $\boldsymbol{W} = \boldsymbol{W}_t$, for any $t \leq T_0$,

$$\begin{aligned}
\|\boldsymbol{W}_t\| &= \left\| \boldsymbol{W}_0 - \sum_{t'=0}^{t-1} \eta_{t'} \nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t) \right\| \\
&\geq \left\langle \boldsymbol{W}_0, \overline{\boldsymbol{W}}_{\mathrm{mm}} \right\rangle + \sum_{t'=0}^{t-1} \eta \frac{\widehat{L}(\boldsymbol{W}_t) - \widehat{L}^*}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t)\|} \left\langle - \frac{\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t)}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t)\|}, \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|} \right\rangle \\
&\geq \eta (2B^2 \Lambda \omega_2 \upsilon)^{-1} t - \left| \left\langle \boldsymbol{W}_0, \overline{\boldsymbol{W}}_{\mathrm{mm}} \right\rangle \right| \\
&\geq \eta (4B^2 \Lambda \omega_2 \upsilon)^{-1} t,
\end{aligned}$$

since $\omega_2 > 1$. Using this, and following similar steps as the proof of Theorem 1, we can show that Eq. (39) is true for any $t \geq t_0$. Using $\epsilon_t = \epsilon_t := (80B^2\Lambda^2 \log t)(\eta t)^{-1}$, we have,

$$\begin{aligned}
&\left\langle \boldsymbol{W}_{t+1} - \boldsymbol{W}_t, \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|} \right\rangle \geq (1 - \epsilon_t) \left\langle \boldsymbol{W}_{t+1} - \boldsymbol{W}_t, \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|} \right\rangle \\
&= \frac{1}{2\|\boldsymbol{W}_t\|} (\|\boldsymbol{W}_{t+1}\|^2 - \|\boldsymbol{W}_t\|^2) - \frac{\eta^2}{2\|\boldsymbol{W}_t\|} \frac{(\widehat{L}(\boldsymbol{W}_t) - \widehat{L}^*)^2}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t)\|^2} + \epsilon_t \eta \frac{\widehat{L}(\boldsymbol{W}_t) - \widehat{L}^*}{\|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t)\|} \left\langle \overline{\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{W}_t)}, \overline{\boldsymbol{W}_t} \right\rangle \\
&\geq \|\boldsymbol{W}_{t+1}\| - \|\boldsymbol{W}_t\| - \frac{\eta^2}{2\omega_1^2 \|\boldsymbol{W}_t\|} - \frac{\epsilon_t \eta}{\omega_1} \\
&\geq \|\boldsymbol{W}_{t+1}\| - \|\boldsymbol{W}_t\| - 80B^2\Lambda^2\omega_1^{-1}(\log t)t^{-1} - 2\eta\upsilon B^2\Lambda\omega_1^{-2}\omega_2(t+1)^{-1}.
\end{aligned}$$

Then, telescoping over $t \geq t_0$, we get

$$\begin{aligned}
\left\langle \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}, \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|} \right\rangle &\geq 1 - \frac{1}{\|\boldsymbol{W}_t\|} \left(\|\boldsymbol{W}_{t_0}\| - \left\langle \boldsymbol{W}_{t_0}, \overline{\boldsymbol{W}}_{\mathrm{mm}} \right\rangle + 2B^2\Lambda\omega_1^{-1}(40\Lambda + \eta\upsilon\omega_1^{-1}\omega_2)(\log t)^2 \right) \\
&\geq 1 - \frac{C(\eta, B, \Lambda, t_0)(\log t)^2}{t}.
\end{aligned}$$

□

B Joint Optimization

B.1 Preliminaries

We first state the complete version of Condition 1 below.

Condition 2 (Complete version of Condition 1)**.** *Let $B = \alpha\rho\sqrt{1.5d}$. We consider the following conditions for any $\delta \in (0,1)$,*

- *Zero initialization: $\|\boldsymbol{u}_0\| = 0, \|\boldsymbol{W}_0\| = 0$,*
- *$\texttt{opt}$ token has larger norm: $\alpha \geq 6$,*
- *Sufficient overparameterization: $d \geq C_0 \alpha^4 n^4 \log\left(\frac{10n^2}{\delta}\right)$, where C_0 is an absolute constant,*
- *Small step-size: $\eta \leq \frac{1}{18\alpha^2\rho^2 d} \wedge \frac{\alpha\rho}{160n} \wedge \frac{\log(2)}{3B} \wedge \frac{B}{128\omega_0\sqrt{1.5n}}$, where $\omega_0 = 13(B \vee 1)^5 (B \vee d)$.*

Lemma 9. *Under the data model DM and Condition 1, the following events are true with probability at least $1 - \delta$,*

- $E_1 := \{\frac{\alpha^2\rho^2 d}{2} \leq \|\boldsymbol{x}_{i,\texttt{opt}_i}\|^2 \leq \frac{3\alpha^2\rho^2 d}{2}, \forall i \in [n]\}$,
- $E_2 := \{\frac{\rho^2 d}{2} \leq \|\boldsymbol{x}_{i,\tau}\|^2 \leq \frac{3\rho^2 d}{2}, \forall i \in [n], \tau \neq \texttt{opt}_i\}$,

- $E_3 := \{|\langle \boldsymbol{x}_{i,\boldsymbol{opt}_i}, \boldsymbol{x}_{j,\boldsymbol{opt}_j}\rangle| \leq 2\alpha^2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)},\ \forall i \neq j \in [n]\}$,
- $E_4 := \{|\langle \boldsymbol{x}_{i,\tau'}, \boldsymbol{x}_{j,\tau}\rangle| \leq 2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)},\ \forall i \neq j \in [n], \tau' \neq \boldsymbol{opt}_i, \tau \neq \boldsymbol{opt}_j\}$,
- $E_5 := \{|\langle \boldsymbol{x}_{i,\boldsymbol{opt}_i}, \boldsymbol{x}_{j,\tau}\rangle| \leq 2\alpha\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)},\ \forall i \neq j \in [n], \tau \neq \boldsymbol{opt}_j\}$,

Proof. First, by applying Bernstein's inequality, with probability at least $1 - \frac{\delta}{5n}$, for every $i \in [n]$,

$$\left|\|\boldsymbol{x}_{i,\mathtt{opt}_i}\|^2 - \alpha^2\rho^2 d\right| \leq c\alpha^2\rho^2\sqrt{d\log\left(\frac{10n}{\delta}\right)} \leq \frac{\alpha^2\rho^2 d}{2},$$

where $c > 0$ is an absolute constant and we use $d \geq (2c)^2\log\left(\frac{10n}{\delta}\right)$ in the second inequality.

Similarly, since $d \geq (2c)^2\log\left(\frac{10n}{\delta}\right)$, we also have that with probability at least $1 - \frac{\delta}{5n}$, for every $i \in [n]$, $\tau \neq \mathtt{opt}_i$,

$$\left|\|\boldsymbol{x}_{i,\tau}\|^2 - \rho^2 d\right| \leq c\rho^2\sqrt{d\log\left(\frac{10n}{\delta}\right)} \leq \frac{\rho^2 d}{2}.$$

Further, for any $i, j \in [n], i \neq j$ and any $\tau, \tau' \in [\mathrm{T}]$ since $\langle \boldsymbol{x}_{i,\tau}, \boldsymbol{x}_{j,\tau'}\rangle$ is zero-mean. By applying Bernstein's inequality, each of the following is true with probability at least $1 - \frac{\delta}{5n^2}$,

$$|\langle \boldsymbol{x}_{i,\mathtt{opt}_i}, \boldsymbol{x}_{j,\mathtt{opt}_j}\rangle| \leq 2\alpha^2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}, \quad |\langle \boldsymbol{x}_{i,\tau'}, \boldsymbol{x}_{j,\tau}\rangle| \leq 2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)},$$
$$|\langle \boldsymbol{x}_{i,\mathtt{opt}_i}, \boldsymbol{x}_{j,\tau}\rangle| \leq 2\alpha\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)},$$

where $\tau' \neq \mathtt{opt}_i, \tau \neq \mathtt{opt}_j$.

Applying a union bound over all these finishes the proof.

□

B.2 Key Lemmas

We first state some intermediate Lemmas that are useful in the proofs of Theorem 2 and 3.

Lemma 10 (Iterate Norm)**.** *Using the updates in Eq. (10), at any time $t > 0$,*

$$\|\boldsymbol{u}_t\| \leq 3\eta t^{1/3}.$$

Proof. Using Eq. (10) and triangle inequality, we get

$$\|\boldsymbol{u}_t\| \leq \|\boldsymbol{u}_{t-1}\| + \frac{\eta}{t^{2/3}}\frac{\|\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t)\|}{\|\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t)\|} \leq \eta\sum_{t'=0}^{t-1}(1+t')^{-2/3} \leq 3\eta t^{1/3}.$$

□

Lemma 11. *Under Condition 1 and the data model DM, for any $j \in [n]$, at any $t > 0$,*

$$\gamma^t_{j,\boldsymbol{opt}_j} - \gamma^t_j \geq \eta\omega_2\sum_{t'=0}^{t-1}(t+1)^{-2/3}, \quad \varphi^t_{j,\boldsymbol{opt}_j} \geq \frac{1}{2} \quad \textit{and} \quad \max_{i,j\in[n]}\frac{\ell_{t,i}}{\ell_{t,j}} \leq C,$$

where $\omega_2 = \omega_2(\alpha, \rho, n) := \frac{\alpha\rho}{n}\sqrt{\log\left(\frac{10n^2}{\delta}\right)}$ and $C \geq 24$ is a universal constant.

Proof. We will prove this by induction. First, we prove the three parts at $t = 1$ as follows.

Since $\boldsymbol{u}_0 = \mathbf{0}$, for any $j \in [n]$, $\ell_{0,j} = 1$. Using this, We have

$$\begin{aligned}
(\gamma^1_{j,\mathtt{opt}_j} - \gamma^1_j) &= -\eta^u_0(-\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_0))^\top(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau}) \\
&= \frac{\eta^u_0}{n}\sum_{i=1}^{n}(y_i\boldsymbol{X}_i^\top\boldsymbol{\varphi}(\boldsymbol{a}_i(\boldsymbol{W}_t)))^\top(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau}) \\
&\geq \frac{\eta^u_0}{n}(\varphi^t_{j,\mathtt{opt}_j}\boldsymbol{x}_{j,\mathtt{opt}_j} + (1-\varphi^t_{j,\mathtt{opt}_j})\boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau}) \\
&+ \frac{\eta^u_0}{n}\sum_{i\neq j}\ell_{t,i}y_iy_j(\varphi^t_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i} + (1-\varphi^t_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau})^\top(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau}) \\
&\geq \frac{\eta^u_0}{n}\left(\frac{1}{2}\frac{\alpha^2\rho^2 d}{2} - \frac{3\rho^2 d}{2} - \alpha\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)} - 2(n-1)M_t\rho^2\alpha^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}\right) \\
&\geq \frac{\eta^u_0}{n}\frac{\alpha^2\rho^2 d}{16},
\end{aligned}$$

since $\alpha \geq 5$, $d \geq (16n)^2\log\left(\frac{10n^2}{\delta}\right)$. Also, $\|\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_0)\| \leq \max_{i\in[n]}\|\boldsymbol{X}_i\|_{2,\infty} \leq \alpha\rho\sqrt{1.5d}$. Using these, we get

$$\begin{aligned}
(\gamma^1_{j,\mathtt{opt}_j} - \gamma^1_j) &\geq \frac{\frac{\eta}{16n}\alpha^2\rho^2\sqrt{d}16n\sqrt{\log\left(\frac{10n^2}{\delta}\right)}}{\alpha\rho\sqrt{1.5d}} \\
&= \frac{\eta\alpha\rho}{\sqrt{1.5}}\sqrt{\log\left(\frac{10n^2}{\delta}\right)} \geq \eta\omega_2,
\end{aligned}$$

when $n > 1$. Next, we have

$$\begin{aligned}
\frac{\varphi^1_{j,\mathtt{opt}_j}/\varphi^1_{j,\tau}}{\varphi^0_{j,\mathtt{opt}_j}/\varphi^0_{j,\tau}} &= \exp\left((\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{W}_1 - \boldsymbol{W}_0)\boldsymbol{x}_{j,1}\right) \\
&= \exp\left((\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(-\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{\theta}_0))\boldsymbol{x}_{j,1}\right) = 1,
\end{aligned}$$

since $\boldsymbol{u}_0 = \mathbf{0}$. Therefore, $\varphi^1_{j,\mathtt{opt}_j} = \varphi^0_{j,\mathtt{opt}_j} \geq \frac{1}{2}$.

Further, since $\eta \leq \frac{\log(2)}{3\alpha\rho\sqrt{1.5d}}$, we have

$$\begin{aligned}
\max_{i,j\in[n]}\frac{\ell_{1,i}}{\ell_{1,j}} &\leq \max_i \exp(2|\boldsymbol{u}_1^\top\boldsymbol{X}_i\top\boldsymbol{\varphi}_i^1|) \leq \exp(2\|\boldsymbol{u}_1\|\alpha\rho\sqrt{1.5d}) \\
&\leq \exp(6\eta\alpha\rho\sqrt{1.5d}) \leq \exp(2\log(2)) = 4 \leq C.
\end{aligned}$$

Next, we assume that these are true at iteration t. Let $M_t := \max_i \ell_{t,i}$ and $m_t := \min_i \ell_{t,i}$. We will first show that for any $j \in [n]$,

$$\begin{aligned}
\Delta\gamma^{t+1}_j &:= (\gamma^{t+1}_{j,\mathtt{opt}_j} - \gamma^{t+1}_j) - (\gamma^t_{j,\mathtt{opt}_j} - \gamma^t_j) \\
&= -\eta^u_t(-\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t))^\top(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau}) \geq \eta(t+1)^{-1/3}\omega_2. \qquad (40)
\end{aligned}$$

We have

$$\begin{aligned}
n(-\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t))^\top(\boldsymbol{x}_{j,\mathtt{opt}_j}-\boldsymbol{x}_{j,\tau}) &= \sum_{i=1}^{n}(\ell_{t,i}y_i\boldsymbol{X}_i^\top\boldsymbol{\varphi}(\boldsymbol{a}_i(\boldsymbol{W}_t)))^\top(\boldsymbol{x}_{j,\mathtt{opt}_j}-\boldsymbol{x}_{j,\tau})\\
&= \ell_{t,j}(\varphi^t_{j,\mathtt{opt}_j}\boldsymbol{x}_{j,\mathtt{opt}_j}+(1-\varphi^t_{j,\mathtt{opt}_j})\boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{j,\mathtt{opt}_j}-\boldsymbol{x}_{j,\tau})\\
&+\sum_{i\neq j}\ell_{t,i}y_iy_j(\varphi^t_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i}+(1-\varphi^t_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau})^\top(\boldsymbol{x}_{j,\mathtt{opt}_j}-\boldsymbol{x}_{j,\tau})\\
&\geq m_t\left(\frac{1}{2}\frac{\alpha^2\rho^2 d}{2}-\frac{3\rho^2 d}{2}-\alpha\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}\right)-2(n-1)M_t\rho^2\alpha^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}\\
&= 2\rho^2\alpha^2M_t\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}-2n\rho^2\alpha^2m_t\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}\left(\frac{M_t}{m_t}-\frac{\frac{\alpha^2\rho^2 d}{4}-\frac{3\rho^2 d}{2}-\alpha\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}}{2n\rho^2\alpha^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}}\right)\\
&\geq 2\rho^2\alpha^2\widehat{L}(\boldsymbol{\theta}_t)\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}-2n\rho^2\alpha^2m_t\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}\left(\frac{M_t}{m_t}-\frac{1}{16n}\sqrt{\frac{d}{\log\left(\frac{10n^2}{\delta}\right)}}\right)\\
&\geq 2\rho^2\alpha^2\widehat{L}(\boldsymbol{\theta}_t)\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}, \qquad (41)
\end{aligned}$$

where we use the definitions of M_t, m_t, and the first inequality follows by using Lemma 9 and the second part of the IH, the second inequality follows since $M_t \geq \widehat{L}(\boldsymbol{\theta}_t)$, $\alpha \geq 6$ and $d \geq 4\log\left(\frac{10n^2}{\delta}\right)$, and the final inequality follows by using the third part of the IH and that $d \geq (16Cn)^2\log\left(\frac{10n^2}{\delta}\right)$.

Next, we have $\|\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t)\| \leq \max_{i\in[n]}\|\boldsymbol{X}_i\|_{2,\infty}\widehat{L}(\boldsymbol{\theta}_t) \leq \alpha\rho\sqrt{1.5d}\widehat{L}(\boldsymbol{\theta}_t)$. Combining this with Eq. (41), we get

$$\frac{-\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t)^\top(\boldsymbol{x}_{j,\mathtt{opt}_j}-\boldsymbol{x}_{j,\tau})}{\|\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t)\|} \geq \frac{2\rho\alpha}{\sqrt{1.5}n}\sqrt{\log\left(\frac{10n^2}{\delta}\right)} \geq \omega_1,$$

which implies Eq. (40). Using Eq. (40) and the first part of the IH, it follows that

$$\gamma^{t+1}_{j,\mathtt{opt}_j}-\gamma^{t+1}_j \geq \gamma^t_{j,\mathtt{opt}_j}-\gamma^t_j+\eta^u_t\omega_1 \geq \eta\omega_2\sum_{t'=0}^{t}(t'+1)^{-2/3}.$$

Next, we will show that for any $j\in[n]$, $\varphi^{t+1}_{j,\mathtt{opt}_j} \geq \frac{1}{2}$. We consider two cases:

Case 1: $\varphi^t_{j,\mathtt{opt}_j} \geq \frac{3}{4}$. In this case, we can directly use the steps in Scenario 1 of the proof of Lemma 3, and get that $\varphi^{t+1}_{j,\mathtt{opt}_j} \geq \frac{1}{2}$ since $\eta \leq \frac{\log(3)}{3\alpha^2\rho^2 d}$.

Case 2: $\varphi^t_{j,\mathtt{opt}_j} \leq \frac{3}{4}$. Using Lemma 10, we have

$$\gamma^t_{i,\mathtt{opt}_i}-\gamma^t_i \leq 2\|\boldsymbol{u}_t\|\max_i\|\boldsymbol{X}_i\|_{2,\infty} \leq 6\eta\sum_{t'=0}^{t-1}(1+t')^{-2/3}\alpha\rho\sqrt{1.5d}. \qquad (42)$$

Next, using similar calculations as Eqs. (18) and (19), for any $j \in [n]$, we have

$$\begin{aligned}
&(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(-\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{\theta}_t))\boldsymbol{x}_{j,1} \\
&\quad \geq \frac{\ell_{t,j}}{n}(\gamma^t_{j,\mathtt{opt}_j} - \gamma^t_j)\varphi^t_{j,\mathtt{opt}_j}(1 - \varphi^t_{j,\mathtt{opt}_j})\|\boldsymbol{x}_{j,1}\|^2\|\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau}\|^2 \\
&\quad - \frac{(n-1)}{n}\max_{i \neq j}\ell_{t,i}(\gamma^t_{i,\mathtt{opt}_i} - \gamma^t_i)\varphi^t_{i,\mathtt{opt}_i}(1 - \varphi^t_{i,\mathtt{opt}_i})|\boldsymbol{x}_{i,1}^\top\boldsymbol{x}_{j,1}||(\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{x}_{i,\mathtt{opt}_i} - \boldsymbol{x}_{i,\tau'})| \\
&\quad \geq \frac{m_t}{n}(\eta\omega_2\sum_{t'=0}^{t-1}(t'+1)^{-2/3})\frac{3}{16}\frac{\rho^2 d}{2}\left(\frac{\alpha^2\rho^2 d}{2} + \frac{\rho^2 d}{2} - 2\alpha\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}\right) \\
&\quad - \frac{(n-1)}{n}M_t(7.5\eta\alpha\rho\sqrt{d}\sum_{t'=0}^{t-1}(t'+1)^{-2/3})\alpha^2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}(\alpha^2+2\alpha+1)\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)} \\
&\quad \geq \frac{\alpha^3\rho^5 d^{3/2}}{n}(\eta\sum_{t'=0}^{t-1}(t'+1)^{-2/3})\left(\frac{3}{2^6 n}\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}m_t - 32\alpha^2(n-1)M_t\log\left(\frac{10n^2}{\delta}\right)\right) \\
&\quad \geq \frac{32\alpha^5\rho^5 d^{3/2}}{n}(\eta\sum_{t'=0}^{t-1}(t'+1)^{-2/3})\log\left(\frac{10n^2}{\delta}\right)\left(\widehat{L}(\boldsymbol{\theta}_t) + nm_t\left(\frac{1}{2^{10}\alpha^2 n^2}\sqrt{\frac{d}{\log\left(\frac{10n^2}{\delta}\right)}} - \frac{M_t}{m_t}\right)\right) \\
&\quad \geq \frac{32\alpha^5\rho^5 d^{3/2}}{n}(\eta\sum_{t'=0}^{t-1}(t'+1)^{-2/3})\log\left(\frac{10n^2}{\delta}\right)\widehat{L}(\boldsymbol{\theta}_t), \qquad (43)
\end{aligned}$$

where we use the definitions of M_t and m_t, and the second inequality follows by using Lemma 9, the first two parts of the IH, and the final inequality uses the third part of the IH and follows when $d \geq (2^{10}C\alpha^2 n^2)^2\log\left(\frac{10n^2}{\delta}\right)$. This implies that

$$\frac{\varphi^{t+1}_{j,\mathtt{opt}_j}/\varphi^{t+1}_{j,\tau}}{\varphi^t_{j,\mathtt{opt}_j}/\varphi^t_{j,\tau}} = \exp\left((\boldsymbol{x}_{j,\mathtt{opt}_j} - \boldsymbol{x}_{j,\tau})^\top(\boldsymbol{W}_{t+1} - \boldsymbol{W}_t)\boldsymbol{x}_{j,1}\right) \geq 1.$$

Therefore, $\varphi^{t+1}_{j,\mathtt{opt}_j} \geq \varphi^t_{j,\mathtt{opt}_j} \geq \frac{1}{2}$.

Finally, we will show that $A^{\max}_{t+1} := \max_{i,j\in[n]}\frac{\ell_{t+1,i}}{\ell_{t+1,j}} \leq C$. Let $\boldsymbol{\varphi}^t_i := \boldsymbol{\varphi}(\boldsymbol{a}_i(\boldsymbol{W}_t))$. For any $i \in [n]$, we have

$$\begin{aligned}
\ell_{t+1,i} &= \exp(-y_i\boldsymbol{u}_{t+1}^\top\boldsymbol{X}_i^\top\boldsymbol{\varphi}^{t+1}_i) \\
&= \exp(-y_i\boldsymbol{u}_{t+1}^\top\boldsymbol{X}_i^\top(\boldsymbol{\varphi}^{t+1}_i - \boldsymbol{\varphi}^t_i))\exp(-y_i\boldsymbol{u}_t^\top\boldsymbol{X}_i^\top\boldsymbol{\varphi}^t_i)\exp(y_i(\eta^u_t\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t))^\top\boldsymbol{X}_i^\top\boldsymbol{\varphi}^t_i) \\
&= \ell_{t,i}\exp(-y_i\boldsymbol{u}_{t+1}^\top\boldsymbol{X}_i^\top(\boldsymbol{\varphi}^{t+1}_i - \boldsymbol{\varphi}^t_i))\exp\left(-\frac{\eta^u_t}{n}\sum_{j=1}^n\ell_{t,j}(y_j\boldsymbol{X}_j^\top\boldsymbol{\varphi}^t_j)^\top y_i\boldsymbol{X}_i^\top\boldsymbol{\varphi}^t_i\right), \qquad (44)
\end{aligned}$$

where $S_{t,i} := \sum_{j=1}^n \ell_{t,j}(y_j\boldsymbol{X}_j^\top\boldsymbol{\varphi}^t_j)^\top y_i\boldsymbol{X}_i^\top\boldsymbol{\varphi}^t_i$ can be expanded as

$$\begin{aligned}
S_{t,i} &= \ell_{t,i}\|\varphi^t_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i} + (1-\varphi^t_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau}\|^2 \\
&\quad + \sum_{j\neq i}\ell_{t,j}y_iy_j(\varphi^t_{j,\mathtt{opt}_j}\boldsymbol{x}_{j,\mathtt{opt}_j} + (1-\varphi^t_{j,\mathtt{opt}_j})\boldsymbol{x}_{j,\tau'})^\top(\varphi^t_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i} + (1-\varphi^t_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau}). \qquad (45)
\end{aligned}$$

Using Lemma 9 and the second part of the IH, we have

$$\begin{aligned}
&\|\varphi^t_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i} + (1-\varphi^t_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau}\|^2 \leq \frac{3\alpha^2\rho^2 d}{2} \\
&\|\varphi^t_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i} + (1-\varphi^t_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau}\|^2 \geq \frac{1}{2^2}\frac{\alpha^2\rho^2 d}{2} + \frac{1^2}{2^2}\frac{\rho^2 d}{2} - 2\frac{1}{4}\alpha\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)} \geq \frac{\alpha^2\rho^2 d}{8}, \\
&|(\varphi^t_{j,\mathtt{opt}_j}\boldsymbol{x}_{j,\mathtt{opt}_j} + (1-\varphi^t_{j,\mathtt{opt}_j})\boldsymbol{x}_{j,\tau'})^\top(\varphi^t_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i} + (1-\varphi^t_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau})| \leq \alpha^2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}.
\end{aligned}$$

Using these in Eq. (45), we get

$$S_{t,i} \leq \frac{3\alpha^2\rho^2 d}{2}\ell_{t,i} + n\alpha^2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}M_t, \quad S_{t,i} \geq \frac{\alpha^2\rho^2 d}{8}\ell_{t,i} - n\alpha^2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}M_t. \tag{46}$$

Next, for the second term, using similar calculations as Eq. (53) and Lemma 10, we have

$$\begin{aligned} R_{t+1,i} &:= -y_i\boldsymbol{u}_{t+1}^\top\boldsymbol{X}_i^\top(\boldsymbol{\varphi}_i^{t+1} - \boldsymbol{\varphi}_i^t) \leq |\boldsymbol{u}_{t+1}^\top\boldsymbol{X}_i^\top(\boldsymbol{\varphi}_i^{t+1} - \boldsymbol{\varphi}_i^t)| \\ &\leq 2\|\boldsymbol{u}_{t+1}\|\max_i\|\boldsymbol{X}_i\|_{2,\infty}^3\|\boldsymbol{W}_{t+1} - \boldsymbol{W}_t\| \\ &\leq 6\eta(\alpha\rho\sqrt{1.5d})^3(t+1)^{1/3}\eta(t+1)^{-1} = 2\eta^2(\alpha\rho\sqrt{1.5d})^3(t+1)^{-2/3}. \end{aligned} \tag{47}$$

Consider the loss ratio for any two samples $i, j \in [n]$, $A_t := \frac{\ell_{t,i}}{\ell_{t,j}}$. Using Eqs. (44) and (47), we have

$$A_{t+1} = A_t\frac{\exp(R_{t+1,i})}{\exp(R_{t+1,j})}\frac{\exp(-\frac{\eta_t^u}{n}S_{t,i})}{\exp(-\frac{\eta_t^u}{n}S_{t,j})} \leq A_t\exp(12\eta^2(\alpha\rho\sqrt{1.5d})^3(t+1)^{-2/3})\frac{\exp(-\frac{\eta_t^u}{n}S_{t,i})}{\exp(-\frac{\eta_t^u}{n}S_{t,j})} \tag{48}$$

We consider two cases:

Case 1: $A_t < C/2$. In this case, we have

$$\frac{\eta_t^u}{n}S_{t,i} \leq \eta(t+1)^{-2/3}\max_i\|\boldsymbol{X}_i\|_{2,\infty} \leq \eta\alpha\rho\sqrt{1.5d}(t+1)^{-2/3}.$$

Using this in Eq. (48), and that $\eta \leq \frac{\log(2)}{3\alpha\rho\sqrt{1.5d}}$, $\eta \leq \frac{1}{18\alpha^2\rho^2 d}$, we get

$$\begin{aligned} A_{t+1} &\leq A_t\exp(12\eta^2(\alpha\rho\sqrt{1.5d})^3)\exp(2\eta\alpha\rho\sqrt{1.5d}) \\ &\leq A_t\exp\left(\frac{\log(2)}{3\alpha\rho\sqrt{1.5d}}\frac{1}{18\alpha^2\rho^2 d}(18\alpha^3\rho^3 d\sqrt{1.5d})\right)\exp(2\log(2)/3) \\ &= 2A_t \leq C. \end{aligned}$$

Case 2: $A_t \geq C/2$.

In this case, using Eq. (46), we have

$$\begin{aligned} \frac{\exp(-\frac{\eta_t^u}{n}S_{t,i})}{\exp(-\frac{\eta_t^u}{n}S_{t,j})} &\leq \exp\left(-\frac{\eta_t^u}{n}\left(\frac{\alpha^2\rho^2 d}{8}\ell_{t,i} - \frac{3\alpha^2\rho^2 d}{2}\ell_{t,j} - 2n\alpha^2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}M_t\right)\right) \\ &\leq \exp\left(-\frac{\eta_t^u\alpha^2\rho^2\ell_{t,j}}{n}\left(\frac{d}{8}A_t - \frac{3d}{2} - 2n\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}A_t^{\max}\right)\right). \end{aligned}$$

Further,

$$\begin{aligned} \frac{\eta_t^u\alpha^2\rho^2\ell_{t,j}}{n}\left(\frac{d}{8}A_t - \frac{3d}{2} - 2n\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}A_t^{\max}\right) &\geq \frac{\eta_t^u\alpha^2\rho^2 m_t}{n}\left(\frac{d}{8}C - \frac{3d}{2} - 2n\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}C\right) \\ &\geq \frac{\eta_t^u\alpha^2\rho^2 m_t}{n}\frac{C\sqrt{d}}{16}\left(\sqrt{d} - 32n\sqrt{\log\left(\frac{10n^2}{\delta}\right)}\right) \\ \geq \frac{\eta_t^u\alpha^2\rho^2\sqrt{d}M_t}{32n} &\geq \frac{\eta\alpha^2\rho^2\sqrt{d}}{32n}(t+1)^{-2/3}\frac{\widehat{L}(\boldsymbol{\theta}_t)}{\alpha\rho\sqrt{1.5d}\widehat{L}(\boldsymbol{\theta}_t)} \\ &= \frac{\eta\alpha\rho}{40n}(t+1)^{-2/3}, \end{aligned}$$

as $C \geq 24$, $d \geq (32n)^2 \log\left(\frac{10n^2}{\delta}\right)$ and $\|\nabla_{\boldsymbol{u}} \widehat{L}(\boldsymbol{\theta}_t)\| \leq \alpha\rho\sqrt{1.5d}\widehat{L}(\boldsymbol{\theta}_t)$. Further, since $\eta \leq \frac{\alpha\rho}{160n}$, we have

$$\exp(4\eta^2(t+1)^{-2/3}) \exp\left(-\frac{\eta\alpha\rho}{40n}(t+1)^{-2/3}\right) \leq 1,$$

which gives $A_{t+1} \leq A_t \leq C$.

□

Lemma 12. *Let $\tilde{\boldsymbol{u}} \coloneqq \sum\limits_{i\in[n]} y_i \boldsymbol{x}_{i,\mathtt{opt}_i}$. Under Condition 1 and the data model DM, for any $j \in [n]$, $\tau \neq \mathtt{opt}_j$,*

$$\left\langle y_j(\boldsymbol{x}_{j,\mathtt{opt}_j} + \boldsymbol{x}_{j,\tau}), \frac{\tilde{\boldsymbol{u}}}{\|\tilde{\boldsymbol{u}}\|} \right\rangle \geq \frac{\alpha\rho}{4}\sqrt{\frac{d}{n}}.$$

Proof. Using Lemma 9, we have

$$\begin{aligned}
y_j(\boldsymbol{x}_{j,\mathrm{opt}_j} + \boldsymbol{x}_{j,\tau})^\top \left(\sum_{i\in[n]} y_i \boldsymbol{x}_{i,\mathrm{opt}_i}\right) &= \|\boldsymbol{x}_{j,\mathrm{opt}_j}\|^2 + y_j \sum_{i\neq j\in[n]} y_i \boldsymbol{x}_{i,\mathrm{opt}_i}^\top \boldsymbol{x}_{j,\mathrm{opt}_j} + y_j \sum_{i\in[n]} y_i \boldsymbol{x}_{i,\mathrm{opt}_i}^\top \boldsymbol{x}_{j,\tau} \\
&\geq 0.5\alpha^2\rho^2 d - 2(n-1)\alpha^2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)} - 2n\alpha\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)} \\
&\geq 0.5\alpha^2\rho^2 d - 4\alpha^2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)} \\
&\geq \frac{\alpha^2\rho^2 d}{3},
\end{aligned}$$

since $d \geq (12n)^2 \log\left(\frac{10n^2}{\delta}\right)$ Similarly, we also have

$$\begin{aligned}
\|\tilde{\boldsymbol{u}}\| &\leq \sqrt{n \max_i \|\boldsymbol{x}_{i,\mathrm{opt}_i}\|^2 + n^2 \max_{i\neq j} |\langle \boldsymbol{x}_{i,\mathrm{opt}_i}, \boldsymbol{x}_{j,\mathrm{opt}_j}\rangle|} \\
&\leq \sqrt{n(1.5\alpha^2\rho^2 d) + 2n^2\alpha^2\rho^2\sqrt{d\log\left(\frac{10n^2}{\delta}\right)}} \\
&\leq \alpha\rho\sqrt{1.75nd},
\end{aligned}$$

since $d \geq (8n)^2 \log\left(\frac{10n^2}{\delta}\right)$. Combining these and using $3\sqrt{1.75} \leq 4$ finishes the proof. □

Lemma 13. *Under Condition 1 and the data model DM, using the updates in Eq. (10), exponential loss has the following properties:*

- $\|\nabla_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}_t)\| \geq \|\nabla_{\boldsymbol{u}} \widehat{L}(\boldsymbol{\theta}_t)\| \geq \omega_1 \widehat{L}(\boldsymbol{\theta}_t)$,
- $\max\limits_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]} \widehat{L}(\boldsymbol{\theta}') \leq 8\widehat{L}(\boldsymbol{\theta}_t)$,
- $\max\limits_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]} \|\nabla^2_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}')\| \leq 8(\omega(\boldsymbol{\theta}_t) \vee \omega(\boldsymbol{\theta}_{t+1}))\widehat{L}(\boldsymbol{\theta}_t)$,

where $\omega_1 = \omega_1(\alpha, \rho, n, d) = \frac{B}{8\sqrt{1.5n}}$, $\omega(\boldsymbol{\theta}_t) \coloneqq 13(B \vee 1)^5 (B \vee d)(\|\boldsymbol{u}_t\| \vee 1)^2$, $B = \alpha\rho\sqrt{1.5d}$.

Remark 3. *We note that the first property is the most challenging to show in the analysis and it yields a PL-inequality-like form, e.g. for neural networks (Frei & Gu, 2021; Nguyen & Mondelli, 2020; Liu et al., 2022). Analogous properties but for simpler settings, e.g. to obtain loss convergence results for linear predictors using NGD on separable data have been shown in (Nacson et al., 2019). The second point is an analog of the log-Lipschitzness property shown in (Taheri & Thrampoulidis, 2023a) for the analysis of two-layer neural networks trained with NGD and controls the loss on a line between current iterate $\boldsymbol{\theta}_t$ and the next one $\boldsymbol{\theta}_{t+1}$. The third property of second-order self-boundedness has been seen previously in the convergence analysis of multi-head self-attention (Deora et al., 2023) and MLPs (Taheri & Thrampoulidis, 2023b) trained with GD.*

Proof. We first obtain the lower bound on the gradient norm as follows.

$$\begin{aligned}
\|\nabla_{\boldsymbol{\theta}}\widehat{L}(\boldsymbol{\theta}_t)\| &= \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1}\left\langle \frac{1}{n}\sum_{i=1}^{n}|\ell'_{i,t}|y_i\nabla_{\boldsymbol{\theta}}\Phi(\boldsymbol{\theta}_t,\boldsymbol{X}_i),\boldsymbol{v}\right\rangle \\
&\geq \|\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t)\| = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1}\left\langle \frac{1}{n}\sum_{i=1}^{n}|\ell'_{i,t}|y_i\nabla_{\boldsymbol{u}}\Phi(\boldsymbol{\theta}_t,\boldsymbol{X}_i),\boldsymbol{v}\right\rangle \\
&\geq \widehat{L}(\boldsymbol{\theta}_t)\sup_{\boldsymbol{v}:\|\boldsymbol{v}\|=1}\min_i y_i\nabla_{\boldsymbol{u}}\Phi(\boldsymbol{\theta}_t,\boldsymbol{X}_i)^\top\boldsymbol{v} \\
&\geq \widehat{L}(\boldsymbol{\theta}_t)\min_i\left\langle y_i\nabla_{\boldsymbol{u}}\Phi(\boldsymbol{\theta}_t,\boldsymbol{X}_i),\frac{\tilde{\boldsymbol{u}}}{\|\tilde{\boldsymbol{u}}\|}\right\rangle, \qquad (49)
\end{aligned}$$

where $\tilde{\boldsymbol{u}} = \sum_{i\in[n]} y_i\boldsymbol{x}_{i,\mathtt{opt}_i}$. Using Lemma 11 and 12,

$$\begin{aligned}
\min_i\left\langle y_i\nabla_{\boldsymbol{u}}\Phi(\boldsymbol{\theta}_t,\boldsymbol{X}_i),\frac{\tilde{\boldsymbol{u}}}{\|\tilde{\boldsymbol{u}}\|}\right\rangle &= \min_i y_i(\varphi^t_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i} + (1-\varphi^t_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau})^\top\overline{\tilde{\boldsymbol{u}}} \\
&\geq 0.5\min_i y_i(\boldsymbol{x}_{i,\mathtt{opt}_i}+\boldsymbol{x}_{i,\tau})^\top\overline{\tilde{\boldsymbol{u}}} \geq \frac{\alpha\rho}{8}\sqrt{\frac{d}{n}} = \omega_1. \qquad (50)
\end{aligned}$$

Using Eq. (50) in Eq. (49), we get the first bullet point.

We show the second bullet point as follows. For any $\boldsymbol{\theta},\boldsymbol{\theta}'$, we have

$$\begin{aligned}
|y\Phi(\boldsymbol{\theta},\boldsymbol{X}) - y\Phi(\boldsymbol{\theta}',\boldsymbol{X})| &= |\boldsymbol{u}^\top\boldsymbol{X}^\top\varphi(\boldsymbol{X}\boldsymbol{W}^\top\boldsymbol{x}_1) - \boldsymbol{u}'^\top\boldsymbol{X}^\top\varphi(\boldsymbol{X}\boldsymbol{W}'^\top\boldsymbol{x}_1)| \\
&\leq |(\boldsymbol{u}-\boldsymbol{u}')^\top\boldsymbol{X}^\top\varphi(\boldsymbol{X}\boldsymbol{W}^\top\boldsymbol{x}_1)| + |\boldsymbol{u}'^\top\boldsymbol{X}^\top(\varphi(\boldsymbol{X}\boldsymbol{W}^\top\boldsymbol{x}_1) - \varphi(\boldsymbol{X}\boldsymbol{W}'^\top\boldsymbol{x}_1))|. \qquad (51)
\end{aligned}$$

These two terms can be bounded as follows.

$$\begin{aligned}
|(\boldsymbol{u}-\boldsymbol{u}')^\top\boldsymbol{X}^\top\varphi(\boldsymbol{X}\boldsymbol{W}^\top\boldsymbol{x}_1)| &\leq \|\boldsymbol{u}-\boldsymbol{u}'\|\|\boldsymbol{X}^\top\|_{1,2} = \|\boldsymbol{X}\|_{2,\infty}\|\boldsymbol{u}-\boldsymbol{u}'\|, \qquad (52) \\
\|\boldsymbol{X}^\top(\varphi(\boldsymbol{X}\boldsymbol{W}^\top\boldsymbol{x}_1) - \varphi(\boldsymbol{X}\boldsymbol{W}'^\top\boldsymbol{x}_1))\| &\leq \|\boldsymbol{X}^\top\|_{1,2}\|\varphi(\boldsymbol{X}\boldsymbol{W}^\top\boldsymbol{x}_1) - \varphi(\boldsymbol{X}\boldsymbol{W}'^\top\boldsymbol{x}_1)\|_1 \\
&\leq 2\|\boldsymbol{X}\|_{2,\infty}\|\boldsymbol{X}\boldsymbol{W}^\top\boldsymbol{x}_1 - \boldsymbol{X}\boldsymbol{W}'^\top\boldsymbol{x}_1\|_\infty \\
&\leq 2\|\boldsymbol{X}\|^2_{2,\infty}\|(\boldsymbol{W}-\boldsymbol{W}')^\top\boldsymbol{x}_1\| \\
&\leq 2\|\boldsymbol{X}\|^2_{2,\infty}\|\boldsymbol{x}_1\|\|\boldsymbol{W}-\boldsymbol{W}'\| \\
&\leq 2\|\boldsymbol{X}\|^3_{2,\infty}\|\boldsymbol{W}-\boldsymbol{W}'\|. \qquad (53)
\end{aligned}$$

Using Eqs. (52) and (53) in Eq. (51) for $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_t + \lambda(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t)$, we get

$$\begin{aligned}
&\max_{\lambda\in[0,1]}\frac{\ell(y\Phi(\boldsymbol{\theta}_t+\lambda(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t),\boldsymbol{X}))}{\ell(y\Phi(\boldsymbol{\theta}_t,\boldsymbol{X}))} \\
&\quad\leq \max_{\lambda\in[0,1]}\exp(|y\Phi(\boldsymbol{\theta}_t+\lambda(\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t),\boldsymbol{X}) - y\Phi(\boldsymbol{\theta}_t,\boldsymbol{X})|) \\
&\quad\leq \max_{\lambda\in[0,1]}\exp\left(2\lambda\|\boldsymbol{X}\|^3_{2,\infty}\|\boldsymbol{u}_t\|\|\boldsymbol{W}_{t+1}-\boldsymbol{W}_t\| + \lambda\|\boldsymbol{X}\|_{2,\infty}\|\boldsymbol{u}_{t+1}-\boldsymbol{u}_t\|\right) \\
&\quad\leq \exp\left(2\eta^W_t\|\nabla_{\boldsymbol{W}}\widehat{L}(\boldsymbol{\theta}_t)\|\|\boldsymbol{X}\|^3_{2,\infty}\|\boldsymbol{u}_t\| + \eta^u_t\|\nabla_{\boldsymbol{u}}\widehat{L}(\boldsymbol{\theta}_t)\|\|\boldsymbol{X}\|_{2,\infty}\right) \qquad \text{(using Eq. (2))} \\
&\quad\leq \exp\left(6\eta\|\boldsymbol{X}\|^3_{2,\infty}\frac{\eta t^{1/3}}{(1+t)} + \eta\|\boldsymbol{X}\|_{2,\infty}\right) \qquad \text{(using Lemma 10)} \\
&\quad\leq \exp\left(6\eta^2\|\boldsymbol{X}\|^3_{2,\infty} + \eta\|\boldsymbol{X}\|_{2,\infty}\right).
\end{aligned}$$

Since this is true for every $\boldsymbol{X}_i$, and $\eta \le \frac{\log(2)}{3\alpha\rho\sqrt{1.5d}} \wedge \frac{\log(3)}{3\alpha^2\rho^2 d}$, we get

$$\begin{aligned}\max_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]} \widehat{L}(\boldsymbol{\theta}') &\le \max_{i\in[n]} \exp\left(6\eta^2 \|\boldsymbol{X}_i\|_{2,\infty}^3 + \eta \|\boldsymbol{X}_i\|_{2,\infty}\right) \widehat{L}(\boldsymbol{\theta}_t)\\ &= \exp\left(6\eta^2 B^3 + \eta B\right) \widehat{L}(\boldsymbol{\theta}_t) \le \exp\left(6\frac{\log(2)}{3}\frac{\log(3)}{2} + \frac{\log(2)}{3}\right)\widehat{L}(\boldsymbol{\theta}_t)\\ &\le 8\widehat{L}(\boldsymbol{\theta}_t).\end{aligned}$$

Next, we obtain the upper bound on the Hessian norm as follows. Using Prop. 3 from (Deora et al., 2023), we have

$$\begin{aligned}\|\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}_t, \boldsymbol{X})\| &\le \|\boldsymbol{X}\|_{2,\infty} + 2\|\boldsymbol{X}\|_{2,\infty}^2 \|\boldsymbol{X}\boldsymbol{u}_t\|_\infty,\\ \|\nabla^2_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}_t, \boldsymbol{X})\| &\le 6d\|\boldsymbol{X}\|_{2,\infty}^2 \|\boldsymbol{X}\|_{1,\infty}^2 \|\boldsymbol{X}\boldsymbol{u}_t\|_\infty + 2\sqrt{d}\|\boldsymbol{X}\|_{2,\infty}^2\|\boldsymbol{X}\|_{1,\infty}.\end{aligned}$$

Using these, we get

$$\begin{aligned}\|\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}_t, \boldsymbol{X}_i)\| &\le \|\boldsymbol{X}_i\|_{2,\infty} + 2\|\boldsymbol{X}_i\|_{2,\infty}^3 \|\boldsymbol{u}_t\|,\\ \|\nabla^2_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}_t, \boldsymbol{X}_i)\| &\le 6d\|\boldsymbol{X}_i\|_{2,\infty}^5 \|\boldsymbol{u}_t\| + 2\sqrt{d}\|\boldsymbol{X}_i\|_{2,\infty}^3,\\ \implies \|\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}_t, \boldsymbol{X}_i)\|^2 + \|\nabla^2_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}_t, \boldsymbol{X}_i)\| &\le \max_{i\in[n]} 4\|\boldsymbol{X}_i\|_{2,\infty}^6 \|\boldsymbol{u}_t\|^2 + 6d\|\boldsymbol{X}_i\|_{2,\infty}^5\|\boldsymbol{u}_t\| + 2\sqrt{d}\|\boldsymbol{X}_i\|_{2,\infty}^3 + \|\boldsymbol{X}_i\|_{2,\infty}^2\\ &= 4B^6\|\boldsymbol{u}_t\|^2 + 6dB^5\|\boldsymbol{u}_t\| + 2\sqrt{d}B^3 + B^2\\ &\le 13(B\vee 1)^5(B \vee d)(\|\boldsymbol{u}_t\| \vee 1)^2 =: \omega(\boldsymbol{\theta}_t).\end{aligned}$$

Using this, we get

$$\begin{aligned}\max_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]} \|\nabla^2_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}')\| &\le \max_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]} \omega(\boldsymbol{\theta}')\widehat{L}(\boldsymbol{\theta}')\\ &\le 8(\omega(\boldsymbol{\theta}_t) \vee \omega(\boldsymbol{\theta}_{t+1}))\widehat{L}(\boldsymbol{\theta}_t).\end{aligned}$$

□

B.3 Proof of Theorem 2

We first restate Theorem 2, this time with the exact constants.

Theorem 8 (Train loss convergence)**.** *Under Condition 1 and the data model DM, using the updates in Eq. (10), for any $t > 0$,*

$$\widehat{L}(\boldsymbol{\theta}_{t+1}) \le \mathcal{O}\left(\exp\left(-\frac{\eta\omega_1}{2}(t+1)^{1/3}\right)\right),$$

where $\omega_1 = \frac{\alpha\rho}{8}\sqrt{\frac{d}{n}}$.

Proof. First, using Lemma 10, we have

$$\begin{aligned}\omega(\boldsymbol{\theta}_t) \vee \omega(\boldsymbol{\theta}_{t+1}) &\le 13(B\vee 1)^5(B\vee d)(\|\boldsymbol{u}_t\| \vee \|\boldsymbol{u}_{t+1}\| \vee 1)^2\\ &\le 13(B\vee 1)^5(B\vee d)(3\eta(t+1)^{1/3} \vee 1)^2\\ &= 13(B\vee 1)^5(B\vee d)(3\eta \vee (t+1)^{-1/3})^2(t+1)^{2/3}\\ &\le \omega_0(t+1)^{2/3}, \end{aligned} \tag{54}$$

where $\omega_0 = 13(B\vee 1)^5(B\vee d)$, since $\eta \le 1/3$.

Next, for some $\boldsymbol{\theta}' \in [\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]$, using the second order Taylor expansion of $\widehat{L}(\boldsymbol{\theta}_{t+1})$, we have

$$\begin{aligned}
&\widehat{L}(\boldsymbol{\theta}_{t+1}) = \widehat{L}(\boldsymbol{\theta}_t) + \langle \nabla_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}_t), \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t \rangle + \frac{1}{2} (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)^\top \nabla^2_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}')(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) \\
&\leq \widehat{L}(\boldsymbol{\theta}_t) + \langle \nabla_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}_t), \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t \rangle + \frac{1}{2} \|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|^2 \max_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]} \|\nabla^2_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}')\| \\
&\leq \widehat{L}(\boldsymbol{\theta}_t) - \eta_t^u \|\nabla_{\boldsymbol{u}} \widehat{L}(\boldsymbol{\theta}_t)\|^2 - \eta_t^W \|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{\theta}_t)\|^2 \\
&+ \frac{1}{2} ((\eta_t^u)^2 \|\nabla_{\boldsymbol{u}} \widehat{L}(\boldsymbol{\theta}_t)\|^2 + (\eta_t^W)^2 \|\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{\theta}_t)\|^2) \max_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]} \|\nabla^2_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}')\| \\
&\leq \widehat{L}(\boldsymbol{\theta}_t) - \frac{\eta}{(t+1)^{2/3}} \|\nabla_{\boldsymbol{u}} \widehat{L}(\boldsymbol{\theta}_t)\| + \frac{1}{2} \left(\frac{\eta^2}{(t+1)^{4/3}} + \frac{\eta^2}{(t+1)^2} \right) \max_{\boldsymbol{\theta}' \in [\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}]} \|\nabla^2_{\boldsymbol{\theta}} \widehat{L}(\boldsymbol{\theta}')\| && \text{(using Eq. (10))} \\
&\leq \widehat{L}(\boldsymbol{\theta}_t) - \frac{\eta \omega_1}{(t+1)^{2/3}} \widehat{L}(\boldsymbol{\theta}_t) + \frac{8\eta^2}{(t+1)^{4/3}} (\omega(\boldsymbol{\theta}_t) \vee \omega(\boldsymbol{\theta}_{t+1})) \widehat{L}(\boldsymbol{\theta}_t) && \text{(using Lemma 13)} \\
&\leq \left(1 - \frac{\eta \omega_1}{(t+1)^{2/3}} + \frac{8\eta^2 \omega_0 (t+1)^{2/3})}{(t+1)^{4/3})} \right) \widehat{L}(\boldsymbol{\theta}_t) && \text{(using Eq. (54))} \\
&\leq \left(1 - \frac{\eta \omega_1}{2(t+1)^{2/3}} \right) \widehat{L}(\boldsymbol{\theta}_t) \leq \exp\left(-\frac{\eta \omega_1}{2} (t+1)^{-2/3} \right) \widehat{L}(\boldsymbol{\theta}_t) && \text{(using Condition 1)} \\
&\leq \exp\left(-\sum_{t'=0}^{t} \frac{\eta \omega_1}{2} (t+1)^{-2/3} \right) \widehat{L}(\boldsymbol{\theta}_0) && \text{(by telescoping)} \\
&\leq \exp\left(-\frac{\eta \omega_1}{2} (t+1)^{1/3} \right) \widehat{L}(\boldsymbol{\theta}_0).
\end{aligned}$$

□

B.4 Proofs of Theorems 3 and 4

We first restate Theorem 3, this time with the exact constants.

Theorem 9 (IB Rate under Joint Training)**.** *Under Condition 1 and the data model DM, using the updates in Eq. (10), for any* $t \geq t_\epsilon := \exp\left(\frac{(10B\Lambda)^2}{\eta} \left(\frac{Cn^2\sqrt{2d}}{250 \log\left(\frac{10n^2}{\delta}\right)} \vee 10B^2\Lambda \right)^{1/3} \epsilon^{-4/3} \right) \vee \exp\left(\frac{B^2\Lambda}{\eta}\right)$,

$$\left\langle \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}, \frac{\boldsymbol{W}_{mm}}{\|\boldsymbol{W}_{mm}\|} \right\rangle \geq 1 - \epsilon - \frac{C(\eta, B, \Lambda, \epsilon)}{\log t},$$

where $B = \alpha\rho\sqrt{1.5d}$*,* $C \geq 24$ *is an absolute constant, and*

$$C(\eta, B, \Lambda, \epsilon) = \frac{2B^2\Lambda}{\eta} (1-\epsilon) \|\boldsymbol{W}_{t_\epsilon}\| \left(1 - (1-\epsilon)^{-1} \left\langle \overline{\boldsymbol{W}}_{t_\epsilon}, \overline{\boldsymbol{W}}_{mm} \right\rangle - 2\eta \|\boldsymbol{W}_{t_\epsilon}\|^{-1} \widehat{L}(\boldsymbol{\theta}_{t_\epsilon}) \right).$$

Proof. In this case, Lemma 4 follows directly, using Lemma 11. Since $\|\boldsymbol{W}_0\| = 0$, using Eq. 10 and similar calculations as the proof of Lemma 5, for any $t > 0$,

$$\|\boldsymbol{W}_t\| \geq \eta (2B^2\Lambda)^{-1} \log t. \tag{55}$$

Next, using Lemma 11, we will show that for any $\epsilon \in (0,1)$, there exists

$$R_\epsilon := 2\Lambda\epsilon^{-1} \left(\log\left(\frac{4Cn^2\sqrt{2d}}{\log\left(\frac{10n^2}{\delta}\right)} \epsilon^{-1} \right) \vee 5 \log\left(40B^2\Lambda\epsilon^{-1}\right) \right), \tag{56}$$

such that for every t where $\|\boldsymbol{W}_t\| \geq R_\epsilon$,

$$\left\langle -\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{\theta}_t), \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|} \right\rangle \geq (1-\epsilon) \left\langle -\nabla_{\boldsymbol{W}} \widehat{L}(\boldsymbol{\theta}_t), \frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|} \right\rangle. \tag{57}$$

Using similar calculations as the proof of Lemma 7, in the first two cases, since $R' = R\|\boldsymbol{W}_{\mathrm{mm}}\|^{-1} \geq 10\log(40B^2\Lambda\epsilon^{-1})$, Eqs. (30) and (31) follow, respectively. Using Eq. (36), we can show that when

$$\exp(0.5\epsilon(1-\epsilon)^{-1}R') \geq 0.5\nu(n-1)(2)^2\frac{\max_{i\in\mathcal{I}_3}\ell_{t,i}(\gamma_{i,\mathtt{opt}_i}-\gamma_i)}{\min_{i\in\mathcal{I}_1}\ell_{t,i}(\gamma_{i,\mathtt{opt}_i}-\gamma_i)}, \tag{58}$$

Eq. (29) is satisfied. Using Lemma 11, Eq. (58) is true when

$$\exp(0.5\epsilon(1-\epsilon)^{-1}R') \geq \frac{4Cn^2\sqrt{2d}}{\log\left(\frac{10n^2}{\delta}\right)}\epsilon^{-1},$$

which is satisfied by Eq. (56).

Next, using Eqs. (55) and (58), we can show that for any $t \geq t_\epsilon$, $\|\boldsymbol{W}_t\| \geq R_\epsilon \vee 1/2$. We have

$$\begin{aligned}
R_\epsilon &= 2\Lambda\epsilon^{-1}\left(\log\left(\frac{4Cn^2\sqrt{2d}}{\log\left(\frac{10n^2}{\delta}\right)}\epsilon^{-1}\right) \vee 5\log\left(40B^2\Lambda\epsilon^{-1}\right)\right)\\
&\leq 25\Lambda\left(\frac{Cn^2\sqrt{2d}}{250\log\left(\frac{10n^2}{\delta}\right)} \vee 10B^2\Lambda\right)^{1/3}\epsilon^{-4/3}\\
&\leq \eta(4B^2\Lambda)^{-1}\log t \leq \|\boldsymbol{W}_t\|.
\end{aligned}$$

In addition, $\|\boldsymbol{W}_t\| \geq 1/2$ for $t \geq \exp\left(\frac{B^2\Lambda}{\eta}\right)$. Combining these and using similar steps as the proof of Theorem 4 in (Tarzanagh et al., 2023a), we get

$$\begin{aligned}
\left\langle\frac{\boldsymbol{W}_t}{\|\boldsymbol{W}_t\|}, \frac{\boldsymbol{W}_{\mathrm{mm}}}{\|\boldsymbol{W}_{\mathrm{mm}}\|}\right\rangle &\geq 1-\epsilon-\frac{\eta(2B^2\Lambda)^{-1}C(\eta,B,\epsilon)}{\|\boldsymbol{W}_t\|}\\
&\geq 1-\epsilon-\frac{C(\eta,B,\epsilon)}{\log t},
\end{aligned}$$

where the last step follows by using Eq. (55). □

Next, we restate Theorem 4 for convenience.

Theorem 10 (IB Rate of $\boldsymbol{u}$)**.** *Let* $\overline{\gamma} := \max_{\boldsymbol{u}:\|\boldsymbol{u}\|\leq 1}\min_i y_i\boldsymbol{u}^\top\boldsymbol{x}_{i,opt_i}$*. Under Condition 1 and the data model DM, using the updates in Eq.* (10)*, for any* $t \geq t_\epsilon \vee \exp(C(\eta,B,\Lambda,\epsilon)(\epsilon^{-1} \vee (8B^2\Lambda)^4))$*,*

$$\min_i y_i\overline{\boldsymbol{u}_t}^\top\boldsymbol{x}_{i,opt_i} \geq \frac{\overline{\gamma}}{4}-\frac{1}{1+\exp(\eta(8B^2\Lambda^2)^{-1}\log t)}.$$

Proof. First, by definition of $\overline{\gamma}$, we have

$$\|\nabla_{\boldsymbol{\theta}}\widehat{L}(\boldsymbol{\theta}_t)\| \geq \frac{\overline{\gamma}}{3}\widehat{L}(\boldsymbol{\theta}_t).$$

Then, following similar steps as the proof of Theorem 2, we have

$$\widehat{L}(\boldsymbol{\theta}_t) \leq \exp\left(-\frac{\eta\overline{\gamma}}{4}(t+1)^{1/3}\right)\widehat{L}(\boldsymbol{\theta}_0).$$

Using this, we have

$$\begin{aligned}
&\min_i \exp(-y_i(\varphi_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i}+(1-\varphi_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau})^\top\boldsymbol{u}_t) \leq \exp\left(-\frac{\eta\overline{\gamma}}{4}(t+1)^{1/3}\right),\\
&\Longrightarrow \min_i y_i(\varphi_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i}+(1-\varphi_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau})^\top\boldsymbol{u}_t \geq \frac{\eta\overline{\gamma}}{4}(t+1)^{1/3}.
\end{aligned}$$

Then, using the proof of Lemma 10, we get

$$
\begin{aligned}
\min_i y_i(\varphi_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i} + (1-\varphi_{i,\mathtt{opt}_i})\boldsymbol{x}_{i,\tau})^\top \overline{\boldsymbol{u}_t} &\geq \frac{\overline{\gamma}}{4} \\
\implies \min_i y_i \varphi_{i,\mathtt{opt}_i}\boldsymbol{x}_{i,\mathtt{opt}_i}^\top \overline{\boldsymbol{u}_t} &\geq \frac{\overline{\gamma}}{4} - \max_i (1-\varphi_{i,\mathtt{opt}_i})\|\boldsymbol{x}_{i,\tau}\| \\
&\geq \frac{\overline{\gamma}}{4} - \frac{\rho\sqrt{1.5d}}{1+\exp(-\eta(8B^2\Lambda^2)^{-1}\log t)},
\end{aligned}
$$

where for the last inequality, we use the following lower bound on the softmax scores,

$$
\varphi^t_{i,\mathtt{opt}_i} \geq 1 - \frac{1}{1+\exp(\eta(8B^2\Lambda^2)^{-1}\log t)}
$$

which is obtained by using Theorem 3 and following similar steps as the proof of Lemma 1, as

$$
\begin{aligned}
(\boldsymbol{x}_{i,\tau} - \boldsymbol{x}_{i,\mathtt{opt}_i})^\top(\boldsymbol{W}_t - \overline{\boldsymbol{W}}_{\mathrm{mm}}\|\boldsymbol{W}_t\| + \overline{\boldsymbol{W}}_{\mathrm{mm}}\|\boldsymbol{W}_t\|)\boldsymbol{x}_{i,1} &\leq 2B^2\|\boldsymbol{W}_t\|\|\overline{\boldsymbol{W}_t} - \overline{\boldsymbol{W}}_{\mathrm{mm}}\| - \frac{1}{\|\boldsymbol{W}_{\mathrm{mm}}\|}\|\boldsymbol{W}_t\| \\
&\leq 4B^2\sqrt{\epsilon \vee \frac{C(\eta,B,\Lambda,\epsilon)}{\log t}}(2\eta\log t) - \eta(4B^2\Lambda^2)^{-1}\log t \\
&\leq -\eta(8B^2\Lambda^2)^{-1}\log t,
\end{aligned}
$$

since $t \geq \exp(C(\eta,B,\Lambda,\epsilon)(\epsilon^{-1} \vee (8B^2\Lambda)^4))$.

Using $\varphi_{i,\mathtt{opt}_i} \leq 1$ then finishes the proof. □

C Additional Experiments and Settings

This section includes some additional experiments and details about the settings for the results included in this work. We use the PyTorch (Paszke et al., 2019) library for our code, which is licensed under the Modified BSD license.

Fig. 1 In this case, we use the MultiNLI (Williams et al., 2018) dataset, which contains sentence pairs belonging to one of three classes: entailment, neutral, contradiction. The task is to predict whether the second sentence entails, is neutral with, or contradicts the first sentence. It is released under the ODC-By license. We use the Hugging Face `pytorch-transformers` implementation of the BERT `bert-base-uncased` model, with pretrained weights (Devlin et al., 2019), released under Apache License 2.0. We use batch-size 32 to train all models. Learning rates are—Adam: $2e-5$, SGD: $1e-3$, SNGD: 0.01. The η_{max} for SPS and SNGD is set to 0.1.

Fig. 2 In this case, the samples are generated following Example 1, with $U = 1$ and $\rho = 0.05$. We set $n = 20$, $d = 100$ and $\mathrm{T} = 6$. We use a cap on the adaptive step-size, $\eta_{\max} = 100$ for NGD and NGD-mom, and $\eta_{\max} = 10$ for Polyak-step. The learning rate η is set as 0.025 for NGD and NGD-mom, whereas for GD, it is set as 0.25. For NGD-mom, the momentum parameter is set to 0.9.

Fig. 3 In this case, the samples are generated based on the data model DM with $\rho = 0.1$ and $\alpha = 3$. We set $n = 10$, $d = 100$ and $\mathrm{T} = 2$. The learning rate η is set as 0.002 for NGD and NGD-joint, whereas for GD, it is set as 0.02.

Fig. 4 We use the CivilComments dataset (Borkan et al., 2019), which consists of online comments and the task is to classify if the comment is *toxic* or *non-toxic*. It is released under the CC BY-NC 4.0 license. We use the BERT `bert-base-uncased model`, with pretrained weights (Devlin et al., 2019) using the WILDS package (Koh et al., 2021). The learning rate for SGD is 10^{-3}, 10^{-2} for NGD and 10^{-5} for Adam. The parameter η_{max} is set to 10^{-2} for both SNGD and SPS. All models are trained with batch-size 32.

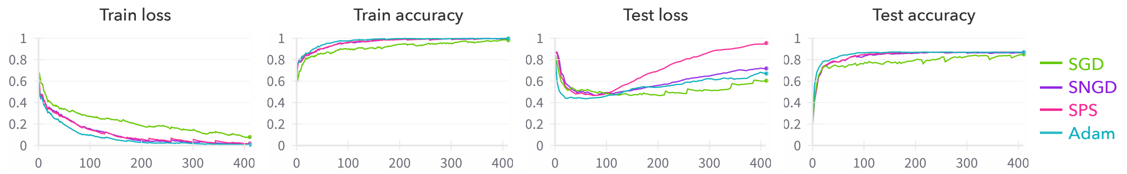

Figure 4: Comparison of train and test dynamics of various optimizers—SGD, SNGD, SPS, and Adam—while fine-tuning a pre-trained BERT model on the CivilComments dataset.

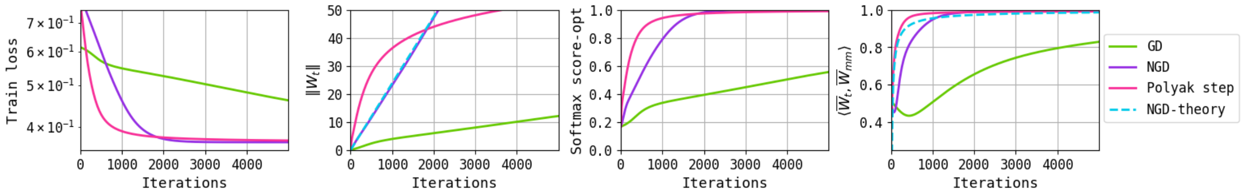

Figure 5: Training dynamics when optimizing only $\boldsymbol{W}$ on synthetic data with antipodal `opt` tokens.

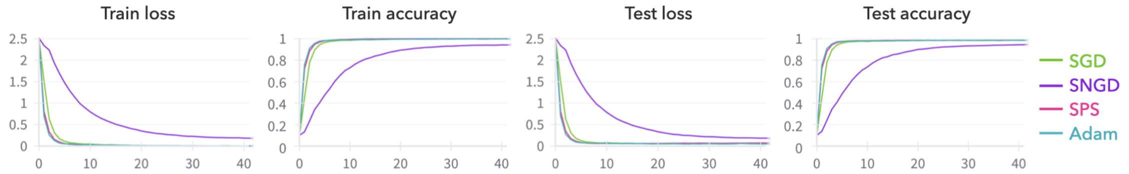

Figure 6: Comparison of train and test dynamics of various optimizers—SGD, SNGD, SPS, and Adam—for a ViT model on the MNIST dataset.

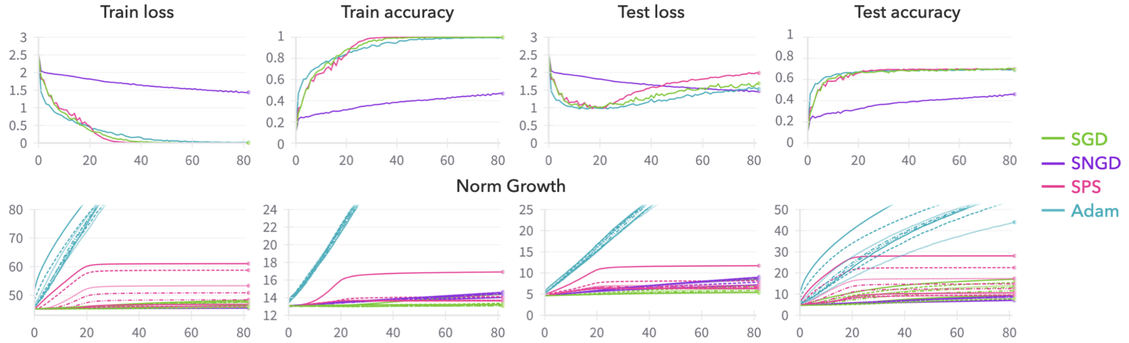

Figure 7: Comparison of train and test dynamics (top row) and parameter norm growth (bottom row) of various optimizers—SGD, SNGD, SPS, and Adam—for a ViT model on the CIFAR-10 dataset.

Fig. 5 Training dynamics of a single-head self-attention model (Eq. (1)) when optimizing only $\boldsymbol{W}$ on synthetic data (Example 1 with $\sigma = 0$). In this case, the `opt` tokens are antipodal instead of orthogonal. The remaining settings are the same as those for Fig. 2, except we set $\eta_{\max} = 5$ for Polyak-step.

Fig. 6 We use the MNIST dataset (LeCun & Cortes, 2005), which contains gray-scale images of handwritten digit 0 – 9. It is released under the CC BY-SA 3.0 license. We use the ViT for small-scale datasets proposed in (Lee et al., 2021), referred to as ViT-small hereon. The implementation is available at `https://github.com/lucidrains/vit-pytorch` under the MIT license. We use patch-size 4, and set depth as 2, number of heads as 8 and MLP width as 128. All models are trained with a batch-size of 100. Learning rates are set as follows. SGD: 0.1, SNGD: 0.001, Adam: 0.001. η_{max} for SPS and SNGD is set to 0.01. We observe that unlike language datasets SNGD trains slower than SGD, and Adam achieves no significant gain

in terms of training speed over SGD (also reported in (Xie et al., 2023)), showcasing behaviour similar to CNNs (Kunstner et al., 2023).

Fig. 7 We consider the CIFAR-10 dataset (Krizhevsky, 2009) which is a benchmark dataset for object recongnition tasks and contains colored images from 10 classes. It is released under the MIT license. We use the ViT-small model with patch-size 4, and set depth as 8, number of heads as 32 and MLP width as 512. All models are trained with batch-size 100. Learning rates are set as follows. SGD: 0.1, SNGD: 0.001, Adam: 0.001. The parameter η_{max} is set to 10 for SPS and 0.1 for SNGD. As evident, SNGD portrays a slower training speed similar to the observations for MNIST. Additionally, we plot the norm growth for various layers of the ViT and see growth in parameters similar to our observations on synthetic datasets.

Compute and Runtimes. The experiments using synthetic data were run on Google Colab. The experiments on vision and language datasets were run on an internal cluster with two NVIDIA V100 GPUs with 32 GB memory each. We train for 40 epochs on MNIST and 80 epochs on CIFAR-10. The runtime for the latter is about 2 hours for each setting. We fine-tune the models on language datasets for 10 epochs, which takes about 32 hours for each run. In Figures 1 and 4, we plot the dynamics for every 200 iterations.

C.1 Initializing in a Bad Stationary Direction

Fig. 8 shows a synthetic setting where we initialize in a bad stationary direction $\boldsymbol{W}_{init}$—one that gives a higher softmax score to the non-opt token. To be precise,

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 0.2 \\ -1 & -0.2 \end{bmatrix}, \qquad \boldsymbol{X}_2 = \begin{bmatrix} -2.5 & 0.5 \\ 2.5 & -0.5 \end{bmatrix},$$

with labels $y_1 = -1, y_2 = 1$, respectively. We set $\boldsymbol{u}_* = [0, 1]^\top$. Using token optimality (Definition 1), this gives $\texttt{opt}_1 = 2, \texttt{opt}_2 = 1$; see Fig. 8 (left) for illustration. The chosen initialization $\boldsymbol{W}_{init}$ violates the (W-SVM) constraints for $i = 2$, that is

$$(\boldsymbol{x}_{2,\texttt{opt}_2} - \boldsymbol{x}_{2,\tau\neq\texttt{opt}_2})^\top \boldsymbol{W}_{init}\boldsymbol{x}_{2,1} < 0.$$

This can be seen in Fig. 8 (left), where the — line correlates more with the non-opt token than the opt one. This consequently forces the softmax score

$$\varphi^0_{2,\texttt{opt}_2} = \frac{1}{1 + \exp((\boldsymbol{x}_{2,\tau\neq\texttt{opt}_2} - \boldsymbol{x}_{2,\texttt{opt}_2})^\top \boldsymbol{W}_{init}\boldsymbol{x}_{2,1})} < \frac{1}{2},$$

which can be seen in Fig. 8 (plot 2). We call $\boldsymbol{W}_{init}$ a "bad" stationary direction as starting with $\alpha\boldsymbol{W}_{init}$ and $\alpha \to \infty$ would result in $\varphi_{2,\texttt{opt}_2}(\alpha\boldsymbol{W}_{init}) \to 0$, and $\nabla_{\boldsymbol{W}}\widehat{L}(\alpha\boldsymbol{W}_{init}) \to \boldsymbol{0}$. On the other hand, for $\boldsymbol{X}_1$, $\boldsymbol{W}_{init}$ behaves the opposite way (see plots 1-2 in Fig. 8).

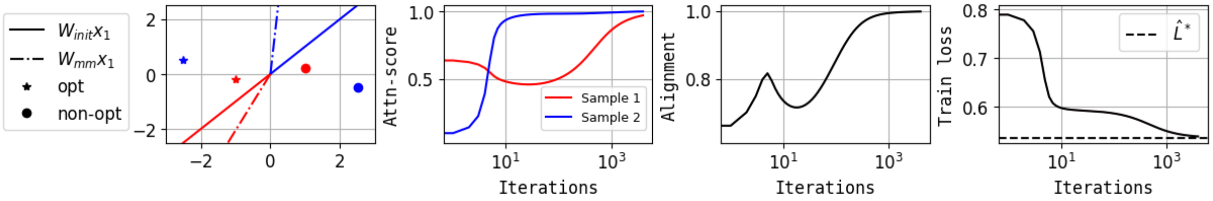

Figure 8: The training dynamics for a synthetic setting with $n = 2, d = 2, T = 2$. Attn-score denotes the opt-token softmax score $\varphi^t_{i,\texttt{opt}_i}$, and alignment shows $\langle \overline{\boldsymbol{W}}_t, \overline{\boldsymbol{W}}_{\text{mm}} \rangle$. Plot 1 shows the initialization $\boldsymbol{W}_{init}$, and the (W-SVM) solution $\boldsymbol{W}_{\text{mm}}$ projected onto the first token, respectively. For sample 2, the $\boldsymbol{W}_{init}$ is more correlated to the non-opt token which consequently gives a small softmax score (plot 2). Plots 2-4 show that despite this "bad" initialization direction GD converges to $\boldsymbol{W}_{\text{mm}}$ and achieves the loss minima $\widehat{L}^*$.

Despite this “bad” starting direction, plots 3-4 in Fig. 8 show that GD still globally convergences to $\boldsymbol{W}_{\text{mm}}$, and minimizes the train loss. This behaviour aligns with our theoretical results (Thm. 1) showing global parametric convergence starting in any direction $\boldsymbol{W}_{init}$, with a small enough norm to avoid the bad stationary directions in the limit as $\|\boldsymbol{W}_{init}\| \to \infty$.

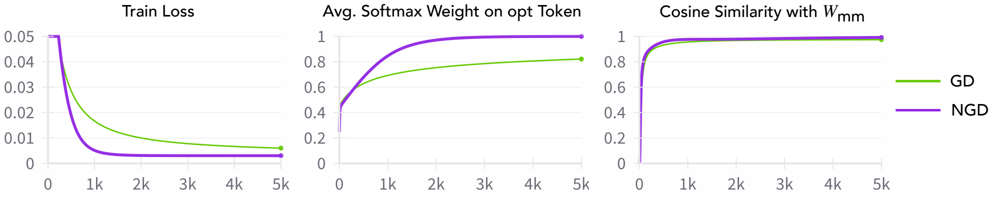

Figure 9: The training dynamics of self-attention parameters $\boldsymbol{W}_t$ for a binary classification on a simplified MNIST. For details on the task see Appendix C.2. Results show parametric convergence to $\boldsymbol{W}_{\text{mm}}$, and fast convergence for NGD aligning with our theoretical findings.

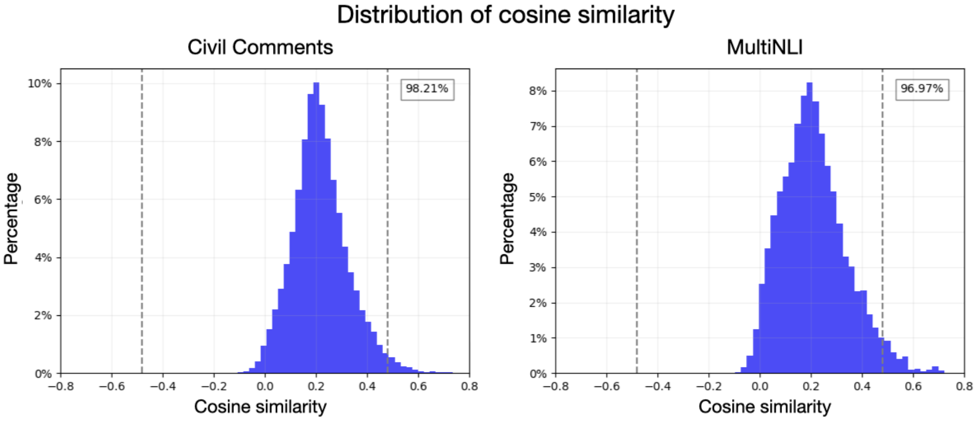

Figure 10: Validation of part 1 of Assumption 1. We compare cosine similarity of token embeddings for tokens within a sentence, and plot the distribution over multiple sentences. See details in Appendix C.3. We observe that a significant portion of the distribution lies around 0, indicating near orthogonality.

C.2 One-layer Self-attention Model Trained on Binary MNIST

In this section, we conduct experiments using a simplified MNIST setting as follows: we created a binary classification task where each input consists of multiple patches, with one patch containing either a 0 or 1 digit (the ’signal’ patch) and the remaining patches containing Gaussian noise. We trained a self-attention module that processes these patches before feeding them through a fixed linear classifier (pre-trained on 0/1 classification). As shown in Figure 9, the results in this setting empirically validate two key theoretical predictions: (1) the self-attention parameters $\boldsymbol{W}$ converge in direction to $\boldsymbol{W}_{\text{mm}}$, and (2) normalized gradient descent achieves faster convergence than standard gradient descent. In this setup, the signal patch naturally corresponds to the token with the highest score, aligning with Ass. 2.

C.3 Validation of Assumption 1

We conduct an experiment to verify the first part of Assumption 1 on two language-based datasets. To do this, we obtain token embeddings from a pre-trained BERT tokenizer (Devlin et al., 2019), and, for every sentence, we compute the cosine similarity between every pair of token embeddings. We consider 50k text examples from the Civil Comments dataset (Borkan et al., 2019) and plot the distribution of these similarity values. We repeat the experiment using 100k samples from the MultiNLI dataset (Williams et al., 2018). The results are shown in Figure 10. Values close to 0 indicate that the tokens are orthogonal.

We observe that a significant fraction of all the cosine similarities fall within a small range around 0, suggesting a nearly orthogonal behavior. Note that while Assumption 1 is only on the *similarity of non-optimal tokens with optimal or other non-optimal tokens*, whereas the experiment considers the similarity between every pair of token embeddings. Nevertheless, we see that the similarities are concentrated in a small region.

D Related Work

Implicit bias of NNs. Since the first few works characterizing the implicit bias of linear predictors on separable data (see Introduction), there has been an abundance of work studying the implicit bias of gradient based methods for both linear predictors and NNs. Nacson et al. (2019); Ji & Telgarsky (2021); Ji et al. (2021) show fast convergence of GD-based methods to the max-margin predictor. For MLPs, early works study the implicit bias of GD/gradient flow using exponentially-tailed classification losses towards the KKT points of the corresponding max-margin problem in finite (Ji & Telgarsky, 2020; Lyu & Li, 2020) and infinite width (Chizat & Bach, 2020). Additionally, works by Phuong & Lampert (2021); Frei et al. (2022b); Kou et al. (2023b) study the implicit bias of GD trained ReLU/Leaky-ReLU networks on orthogonal data. Other works also study the implicit bias towards rank minimization with square loss in regression settings (Vardi & Shamir, 2021; Arora et al., 2019; Li et al., 2021). We encourage the reader to go through these works and a recent survey (Vardi, 2022) for a thorough understanding.

Parametric convergence rates for NGD and PS in non-convex settings. We compare the rates obtained in our results to the corresponding results for linear logistic regression (see Sec. 6). Contrary to these results, we prove the first parametric convergence for NGD and PS in non-convex settings. The work by Loizou et al. (2021) shows loss convergence results for PS. However, the rate of directional convergence might be different from the rate of loss convergence. This has been seen in the series of works analyzing gradient-based methods for linear logistic regression (Soudry et al., 2018; Nacson et al., 2019; Ji & Telgarsky, 2019). If the problem had a unique global minima, then loss convergence implies paramateric convergence. However, the problem of training self-attention does not have a unique minimum. For instance, using some other norm in place of ℓ_2 norm in the objective of (W-SVM) would yield a different solution. While the norm diverges, the convergence is in the direction. In addition to this, there has been some work showing the benefit of adaptive step sizes such as NGD and PS for transformer optimization (Kunstner et al., 2023; Schaipp et al., 2024), which is connected to Q3. We next discuss related work on transformer theory.

Transformers theory. Several studies, such as those by Baldi & Vershynin (2022); Dong et al. (2021); Yun et al. (2020a;b); Sanford et al. (2023); Bietti et al. (2023) have delved into exploring the expressivity of attention, while Baldi & Vershynin (2022); Dong et al. (2021); Yun et al. (2020a;b); Mahdavi et al. (2023) have initiated an investigation of its memory capacity. To gain insights into the optimization aspects of training attention models, Sahiner et al. (2022); Ergen et al. (2022) have explored convex relaxations of attention. Furthermore, a subdomain experiencing growing attention involves the theoretical exploration of in-context learning, as evidenced by recent studies (von Oswald et al., 2022; Akyürek et al., 2023; Zhang et al., 2023; Li et al., 2023c).

In this context, we discuss studies that seek to understand the optimization and generalization dynamics of transformers. Jelassi et al. (2022) show that Vision Transformers (ViTs) learn spatially localized patterns in a binary classification task using gradient-based methods. For a three-layer ViT starting from some structured initialization to mimic a pre-trained network, Li et al. (2023b) show sample complexity bounds to achieve zero generalization and show attention maps sparsify as SGD training proceeds. As a step towards understanding the training dynamics of the closely related prompt-attention, Oymak et al. (2023) study the initial trajectory of GD for one-layer attention. Further, there has also been work to obtain optimization and generalization guarantees of GD in multi-head attention models (Deora et al., 2023). Additionally, recent work by Tian et al. (2023a) attempts to understand SGD-dynamics for the task of next-token prediction for one-layer transformer with a linear decoder showing a similar sparsifying effect in the attention map. More recently, Tian et al. (2023b) extend this by analyzing the joint training dynamics of multi-layer transformer with an MLP. Other than this, there has been recent progress towards a theoretical understanding of next-token prediction dynamics under GD training (Li et al., 2024; Thrampoulidis, 2024). Furthermore, Makkuva et al. (2024) discussed the impact of the transformer architecture and distributional properties on the loss landscape when modeling the data as a Markov source, and Ildiz et al. (2024) linked the dynamics of self-attention to context-conditioned Markov chains.