On the Optimization and Generalization of Multi-head Attention

Anonymous authors
Paper under double-blind review

Abstract

The training and generalization dynamics of the Transformer's core mechanism, namely the Attention mechanism, remain under-explored. Besides, existing analyses primarily focus on single-head attention. Inspired by the demonstrated benefits of overparameterization when training fully-connected networks, we investigate the potential optimization and generalization advantages of using multiple attention heads. Towards this goal, we derive convergence and generalization guarantees for gradient-descent training of a single-layer multi-head self-attention model, under a suitable realizability condition on the data. We then establish primitive conditions on the initialization that ensure realizability holds. Finally, we demonstrate that these conditions are satisfied for a simple tokenized-mixture model. We expect the analysis can be extended to various data-model and architecture variations.

1 Introduction

Transformers have emerged as a promising paradigm in deep learning, primarily attributable to their distinctive self-attention mechanism. Motivated by the model's state-of-the-art performance in natural language processing (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020) and computer vision (Dosovitskiy et al., 2021; Radford et al., 2021; Touvron et al., 2021), the theoretical study of the attention mechanism has seen a notable surge in interest recently. Numerous studies have already explored the expressivity of Attention, e.g. (Baldi & Vershynin, 2022; Dong et al., 2021; Yun et al., 2020a;b; Sanford et al., 2023; Bietti et al., 2023), and initial findings regarding memory capacity have been very recently studied in (Baldi & Vershynin, 2022; Dong et al., 2021; Yun et al., 2020a;b; Mahdavi et al., 2023). In an attempt to comprehend optimization aspects of training attention models, Sahiner et al. (2022); Ergen et al. (2022) have investigated convex-relaxations, while Tarzanagh et al. (2023a) investigates the model's implicit bias. Additionally, Edelman et al. (2021) have presented capacity and Rademacher complexity-based generalization bounds for Self-Attention. However, the exploration of the *finite-time* optimization and generalization dynamics of gradient-descent (GD) for training attention models largely remains an open question.

Recent contributions in this direction, which serve as motivation for our work, include the studies by (Jelassi et al., 2022; Li et al., 2023a; Oymak et al., 2023). These works concentrate on single-layer attention models with a *single attention head*. Furthermore, despite necessary simplifying assumptions made for the data, the analyses in all three cases are rather intricate and appear highly specialized on the individual attention and data model. These direct and highly specialized analyses present certain challenges. First, it remains uncertain whether they can be encompassed within a broader framework that can potentially be extended to more complex attention architectures and diverse data models. Second, they appear disconnected from existing frameworks that have been flourishing in recent years for conventional architectures like fully-connected and convolutional neural networks e.g., (Jacot et al., 2018; Ji & Telgarsky, 2020; Richards & Rabbat, 2021; Banerjee et al., 2022; Taheri & Thrampoulidis, 2023). Consequently, it is also unclear how the introduction of attention alters the analysis landscape.

In this work, we study the optimization and generalization properties of multi-head attention mechanism trained by gradient methods. Our approach specifically leverages the use of *multiple attention heads*. Despite the operational differences between attention heads in an attention model and hidden nodes in an MLP,

we demonstrate, from an analysis perspective, that this parallelism enables the exploitation of frameworks developed for the latter to study the former. Particularly for the generalization analysis, we leverage recent advancements in the application of the algorithmic-stability framework to overparameterized MLPs (Richards & Kuzborskij, 2021; Taheri & Thrampoulidis, 2023).

Contributions. We study training and generalization of gradient descent optimization for a multi-head attention (MHA) layer with H heads in a binary classification setting. For this setting, detailed in Section 2, we analyze training with logistic loss both the attention weights (parameterizing the softmax logits), as well as, the linear decoder that turns output tokens to label prediction.

In Section 3, we characterize key properties of the empirical loss $\widehat{L}$, specifically establishing that it is self-bounded and satisfies a key self-bounded weak-convexity property, i.e. $\lambda_{\min}(\nabla^2\widehat{L}(\boldsymbol{\theta})) \gtrsim -\frac{\kappa}{\sqrt{H}}\widehat{L}(\boldsymbol{\theta})$ for a parameter κ that depends only mildly on the parameter vector $\boldsymbol{\theta}$. Establishing these properties (and also quantifying κ) involves carefully computing and bounding the gradient and Hessian of the MHA layer, calculations that can be useful beyond the context of our paper.

In Sections 4.1-4.2, we present our training and generalization bounds in their most general form. The bounds are given in terms of the empirical loss $\widehat{L}(\boldsymbol{\theta})$ and the distance $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|$ to initialization $\boldsymbol{\theta}_0$ of an appropriately chosen target vector $\boldsymbol{\theta}$. The distance to initialization also controls the minimum number of heads $H \gtrsim \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^6$ required for the bounds to hold. The choice of an appropriate parameter $\boldsymbol{\theta}$ that makes the bounds tight is generically specific to the data setting and the chosen initialization. To guide such a choice, in Section 4.3, we formalize primitive and straightforward-to-check conditions on the initialization $\boldsymbol{\theta}_0$ that ensure it is possible to find an appropriate $\boldsymbol{\theta}$. In short, provided the model output at initialization is logarithmic on the train-set size n and the data are separable with respect to the neural-tangent kernel (NTK) features of the MHA model with constant margin γ, then Corollary 2 shows that with step-size $\eta = \widetilde{O}(1)$ and $\Theta(n)$ gradient descent steps, the train loss and generalization gap is bounded by $\widetilde{\mathcal{O}}(1/n)$ provided only a polylogarithmic number of heads $H = \Omega(\log^6(n))$. We remark that the aforementioned NTK separability assumption, although related to, differs from the standard NTK analysis. Besides, while this assumption is sufficient to apply our general bounds, it is not a necessary condition.

In Section 5, we investigate a tokenized mixture data model with label-(ir)relevant tokens. We show that after one randomized gradient step from zero initialization, the NTK features of the MHA model separate the data with margin $\gamma_\star$. Thus, applying our general analysis from Section 4.1, we establish training and generalization bounds as described above, for a logarithmic number of heads. Towards assessing the optimality of these bounds, we demonstrate that MHA is expressive enough to achieve margin γ_{attn} that is superior to $\gamma_\star$. The mechanism to reach γ_{attn} involves selecting key-query weights of sufficiently large norm, which saturates the softmax nonlinearity by suppressing label-irrelevant tokens. We identify the large-norm requirement as a potential bottleneck in selecting those weights as target parameters in our theory framework and discuss open questions regarding extending the analytical framework into this specific regime.

The remaining parts are organised as follows. Proof sketches of our main training/generalization bounds are given in Section 6. The paper concludes in Section 7 with remarks on our findings' implications and open questions. Detailed proofs are in the appendix, where we also present synthetic numerical experiments.

Related work. We give a brief overview of the most relevant works on understanding optimization/generalization of self-Attention or its variants. Please see Section H for more detailed exposition. Oymak et al. (2023) diverges from traditional self-Attention by focusing on a variant called prompt-Attention, aiming to gain understanding of prompt-tuning. Jelassi et al. (2022) shed light on how ViTs learn spatially localized patterns using gradient-based methods. Li et al. (2023a) provides sample complexity bounds for achieving zero generalization error on training three-layer ViTs for classification tasks for a similar tokenized mixture data model as ours. Contemporaneous work Tian et al. (2023) presents SGD-dynamics of single-layer attention for next-token prediction by re-parameterizing the original problem in terms of the softmax and classification logit matrices, while Tarzanagh et al. (2023b;a) study the implicit bias of training the softmax weights $\boldsymbol{W}$ with a fixed decoder $\boldsymbol{U}$. All these works focus on a single attention head; instead, we leverage the use of multiple heads to establish connections to the literature on GD training of overparameterized MLPs. Conceptually, Hron et al. (2020) drew similar connections, linking multi-head attention to a Gaussian process

in the limit as the number of heads approaches infinity. In contrast, we study the more practical regime of finite heads and obtain *finite-time* optimization and generalization bounds.

Among the extensive studies on training/generalization of overparameterized MLPs, our work closely aligns with Nitanda et al. (2019); Ji & Telgarsky (2020); Cao & Gu (2019); Chen et al. (2020); Telgarsky (2022); Taheri & Thrampoulidis (2023) focusing on classification with logistic loss. Conceptually, our findings extend this research to attention models. The use of algorithmic-stability tools towards order-optimal generalization bounds for overparameterized MLPs has been exploited recently by Richards & Kuzborskij (2021); Richards & Rabbat (2021); Taheri & Thrampoulidis (2023); Lei et al. (2022). To adapt these tools to the MHA layer, we critically utilize the smoothness of the softmax function and derive bounds on the growth of the model's gradient/Hessian, which establish a self-bounded weak convexity property for the empirical risk (see Corollary 1). Our approach also involves training both the classifier and attention weights, necessitating several technical adjustments detailed in Section 6 and Appendix B.1.

2 Preliminaries

Notation. $\boldsymbol{\varphi}(\cdot):\mathbb{R}^T\to\mathbb{R}^T$ denotes the softmax map and $\boldsymbol{\varphi}'(\boldsymbol{v}) \coloneqq \nabla\boldsymbol{\varphi}(\boldsymbol{v}) = \mathrm{diag}(\boldsymbol{\varphi}(\boldsymbol{v})) - \boldsymbol{v}\boldsymbol{v}^\top$ its gradient at $\boldsymbol{v}\in\mathbb{R}^T$. For $t\in[T]$, $\boldsymbol{\varphi}_t(\boldsymbol{v})$ is the t-th entry of $\boldsymbol{\varphi}(\boldsymbol{v})\in\mathbb{R}^T$. For $\boldsymbol{A}\in\mathbb{R}^{n\times m}$, $\boldsymbol{A}_{i,:}$ is its i-th row and $\boldsymbol{A}_{:,j}$ is its j-th column. Recall the induced matrix norm $\|\boldsymbol{A}\|_{p,q} = \max_{\|\boldsymbol{v}\|_p=1}\|\boldsymbol{A}\boldsymbol{v}\|_q$ and particularly the following: $\|\boldsymbol{A}\|_{2,\infty} = \max_{i\in[n]}\|\boldsymbol{A}_{i,:}\|$, $\|\boldsymbol{A}\|_{1,2} = \max_{j\in[m]}\|\boldsymbol{A}_{:,j}\|$, and $\|\boldsymbol{A}\|_{1,\infty} = \max_{j\in[m]}\|\boldsymbol{A}_{:,j}\|_\infty$. For simplicity, $\|\boldsymbol{A}\|, \|\boldsymbol{v}\|$ denote Euclidean norms and $\lambda_{\min}(\boldsymbol{A})$ the minimum eigenvalue. We let $a\wedge b = \min\{a,b\}$ and $a\vee b = \max\{a,b\}$. concat denotes vector concatenation. All logarithms are natural logarithms (base e). We represent the line segment between $\boldsymbol{w}_1,\boldsymbol{w}_2\in\mathbb{R}^{d'}$ as $[\boldsymbol{w}_1,\boldsymbol{w}_2] = \{\boldsymbol{w}:\boldsymbol{w}=\alpha\boldsymbol{w}_1+(1-\alpha)\boldsymbol{w}_2, \alpha\in[0,1]\}$. Finally, to simplify the exposition we use "$\gtrsim$" or "$\lesssim$" notation to hides absolute constants. We also occasionally use standard notations $\mathcal{O},\Omega$ and $\widetilde{\mathcal{O}},\widetilde{\Omega}$ to hide poly-logarithmic factors. Unless otherwise stated these order-wise notations are with respect to the training-set size n. Whenever used, exact constants are specified in the appendix.

Single-head Self-attention. A single-layer self-attention head $\mathrm{ATTN}:\mathbb{R}^{T\times d}\to\mathbb{R}^{T\times d}$ with context size T and dimension d parameterized by key, query and value matrices $\boldsymbol{W}_Q,\boldsymbol{W}_K\in\mathbb{R}^{d\times d_h}, \boldsymbol{W}_V\in\mathbb{R}^{d\times d_v}$ is given by:

$$\mathrm{ATTN}(\boldsymbol{X};\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V) \coloneqq \boldsymbol{\varphi}(\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\boldsymbol{X}\boldsymbol{W}_V\,.$$

Here, $\boldsymbol{X} = [\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_T]^\top\in\mathbb{R}^{T\times d}$ is the input token matrix and $\boldsymbol{\varphi}(\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top)\in\mathbb{R}^{T\times T}$ is the attention matrix. (Softmax applied to a matrix acts row-wise.) To turn the Attention output in a prediction label, we compose ATTN with a linear projection head (aka decoder). Thus, the model's output is[*]

$$\Phi(\boldsymbol{X};\boldsymbol{W},\boldsymbol{U}) \coloneqq \langle\boldsymbol{U},\boldsymbol{\varphi}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top)\boldsymbol{X}\rangle\,. \tag{1}$$

Note that we absorb the value weight matrix $\boldsymbol{W}_V$ into the projector $\boldsymbol{U} = [\boldsymbol{u}_1,\ldots,\boldsymbol{u}_T]^\top\in\mathbb{R}^{T\times d}$. Also, we parameterize throughout the key-query product matrix as $\boldsymbol{W}\coloneqq\boldsymbol{W}_Q\boldsymbol{W}_K^\top$.

Multi-head Self-attention. Our focus is on the multi-head attention (MHA) model with H heads:

$$\sum_{h\in[H]}\mathrm{ATTN}(\boldsymbol{X};\boldsymbol{W}_{Q_h},\boldsymbol{W}_{Kh},\boldsymbol{W}_{Vh})\boldsymbol{W}_{Oh},$$

for output matrices $\boldsymbol{W}_{Oh}\in\mathbb{R}^{d_v\times d}$. Absorbing $\boldsymbol{W}_{Vh}\boldsymbol{W}_{Oh}$ into a projection layer (similar to the single-head attention) and parameterizing $\boldsymbol{W}_h\coloneqq\boldsymbol{W}_{Q_h}\boldsymbol{W}_{Kh}^\top$,[†] we arrive at the following MHA model:

$$\widetilde{\Phi}(\boldsymbol{X};\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{U}}) \coloneqq \frac{1}{\sqrt{H}}\sum\nolimits_{h\in[H]}\Phi(\boldsymbol{X};\boldsymbol{W}_h,\boldsymbol{U}_h) = \frac{1}{\sqrt{H}}\sum\nolimits_{h\in[H]}\langle\boldsymbol{U}_h,\boldsymbol{\varphi}(\boldsymbol{X}\boldsymbol{W}_h\boldsymbol{X}^\top)\boldsymbol{X}\rangle\,, \tag{2}$$

parameterized by $\widetilde{\boldsymbol{W}}\coloneqq\mathrm{concat}\big(\{\boldsymbol{W}_h\}_{h\in[H]}\big)$ and $\widetilde{\boldsymbol{U}}\coloneqq\mathrm{concat}\big(\{\boldsymbol{U}_h\}_{h\in[H]}\big)$. The $1/\sqrt{H}$ scaling is analogous to the normalization in MLP literature e.g. (Du et al., 2019; Ji & Telgarsky, 2021; Richards & Kuzborskij, 2021), ensuring the model variance is of constant order when $\boldsymbol{U}_h$ is initialized $\mathcal{O}_H(1)$.

[*]While we focus on (i) Full-projection: trainable matrix $\boldsymbol{U}\in\mathbb{R}^{T\times d}$, our results also apply to (ii) Pooling: $\boldsymbol{U}=\boldsymbol{u}\mathbb{1}_T^\top$ with trainable $\boldsymbol{u}\in\mathbb{R}^d$, and (iii) Last-token output: $\boldsymbol{U} = \begin{bmatrix}0_{d\times(T-1)} & \boldsymbol{u}\end{bmatrix}^\top$ with trainable $\boldsymbol{u}\in\mathbb{R}^d$.

[†]These relaxations sacrifice some generality since it is common practice to set values for d_h and d_v such that $d_v = d/H < d$, thus imposing low-rank restrictions on the product matrices $\boldsymbol{W}_{Q_h}\boldsymbol{W}_{K_h}^\top$, $\boldsymbol{W}_{Vh}\boldsymbol{W}_{Oh}$. We defer a treatment of these to future work.

Throughout, we will use $\boldsymbol{\theta}_h := \mathrm{concat}(\boldsymbol{U}_h, \boldsymbol{W}_h) \in \mathbb{R}^{dT+d^2}$, to denote the trainable parameters of the h-attention head and $\widetilde{\boldsymbol{\theta}} := \mathrm{concat}(\{\boldsymbol{\theta}_h\}_{h\in[H]}) \in \mathbb{R}^{H(dT+d^2)}$ for the trainable parameters of the overall model. More generally, we use the convention of applying " $\widetilde{\cdot}$ " notation for quantities relating to the multi-head model. Finally, with some slight abuse of notation, we define: $\|\widetilde{\boldsymbol{\theta}}\|_{2,\infty} := \max_{h\in[H]} \|\boldsymbol{\theta}_h\|$.

Training. Given training set $(\boldsymbol{X}_i, y_i)_{i\in[n]}$, with n IID samples, we minimize logistic-loss based empirical risk

$$\widehat{L}(\widetilde{\boldsymbol{\theta}}) := \frac{1}{n}\sum_{i\in[n]} \ell(y_i \widetilde{\Phi}(\boldsymbol{X}_i; \widetilde{\boldsymbol{\theta}})) := \frac{1}{n}\sum_{i\in[n]} \log(1 + e^{-y_i \widetilde{\Phi}(\boldsymbol{X}_i;\widetilde{\boldsymbol{\theta}})}) .$$

Our analysis extends to any convex, smooth, Lipschitz and self-bounded loss.[‡] The empirical risk is minimized as an approximation of the *test loss* defined as $L(\widetilde{\boldsymbol{\theta}}) := \mathbb{E}_{(\boldsymbol{X},y)}[\ell(y\widetilde{\Phi}(\boldsymbol{X};\widetilde{\boldsymbol{\theta}}))]$. We consider standard gradient-descent (GD) applied to empirical risk $\widehat{L}$. Formally, initialized at $\widetilde{\boldsymbol{\theta}}^{(0)}$ and equipped with step-size $\eta > 0$, at each iteration $k \geq 0$, GD performs the following update:

$$\widetilde{\boldsymbol{\theta}}^{(k+1)} = \widetilde{\boldsymbol{\theta}}^{(k)} - \eta \nabla \widehat{L}(\widetilde{\boldsymbol{\theta}}^{(k)}) .$$

3 Gradient and Hessian bounds of soft-max attention

This section establishes bounds on the gradient and Hessian of the logistic empirical risk $\widehat{L}(.)$ evaluated on the multi-head attention model. To do this, we first derive bounds on the Euclidean norm and spectral-norm for the gradient and Hessian of the self-attention model. In order to simplify notations, we state here the bounds for the single-head model (see App. A.1 for multi-head model): $\Phi(\boldsymbol{X};\boldsymbol{\theta}) := \Phi(\boldsymbol{X};\boldsymbol{W},\boldsymbol{U}) = \langle \boldsymbol{U}, \boldsymbol{\varphi}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top)\boldsymbol{X}\rangle$.

Lemma 1 (Gradient/Hessian formulas)**.** *For all $\boldsymbol{a} \in \mathbb{R}^T$, $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^d$ the model's gradients satisfy:*

- $\nabla_{\boldsymbol{U}} \Phi(\boldsymbol{X};\boldsymbol{\theta}) = \boldsymbol{\varphi}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top)\boldsymbol{X}$, *and* $\nabla_{\boldsymbol{W}} \Phi(\boldsymbol{X};\boldsymbol{\theta}) = \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{u}_t^\top \boldsymbol{X}^\top \boldsymbol{\varphi}'(\boldsymbol{X}\boldsymbol{W}^\top \boldsymbol{x}_t)\boldsymbol{X}$.
- $\nabla_{\boldsymbol{W}} \langle \boldsymbol{a}, \nabla_{\boldsymbol{U}} \Phi(\boldsymbol{X};\boldsymbol{\theta})\, \boldsymbol{b}\rangle = \sum_{t=1}^{T} \boldsymbol{x}_t a_t \boldsymbol{b}^\top \boldsymbol{X}^\top \boldsymbol{\varphi}'(\boldsymbol{X}\boldsymbol{W}^\top \boldsymbol{x}_t)\boldsymbol{X}$, *and*

$$\nabla_{\boldsymbol{W}} \langle \boldsymbol{c}, \nabla_{\boldsymbol{W}} \Phi(\boldsymbol{X};\boldsymbol{\theta})\, \boldsymbol{b}\rangle = \sum_{t=1}^{T} (\boldsymbol{c}^\top \boldsymbol{x}_t)\, \boldsymbol{x}_t \mathbf{d}^\top \boldsymbol{\varphi}'(\boldsymbol{X}\boldsymbol{W}^\top \boldsymbol{x}_t)\boldsymbol{X}$$

where $\mathbf{d} := \mathrm{diag}(\boldsymbol{X}\boldsymbol{b})\boldsymbol{X}\boldsymbol{u}_t - \boldsymbol{X}\boldsymbol{u}_t\boldsymbol{b}^\top\boldsymbol{X}^\top\boldsymbol{\varphi}(\boldsymbol{X}\boldsymbol{W}^\top\boldsymbol{x}_t) - \boldsymbol{X}\boldsymbol{b}\boldsymbol{u}_t^\top\boldsymbol{X}^\top\boldsymbol{\varphi}(\boldsymbol{X}\boldsymbol{W}^\top\boldsymbol{x}_t)$.

These calculations imply the following useful bounds.

Proposition 1 (Model Gradient/Hessian bounds)**.** *The Euclidean norm of the gradient and the spectral norm of the Hessian of the single-head Attention model (1) are bounded as follows:*

- $\|\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{X};\boldsymbol{\theta})\| \leq 2\|\boldsymbol{X}\|_{2,\infty}^2 \sum_{t=1}^{T} \|\boldsymbol{X}\boldsymbol{u}_t\|_\infty + \sqrt{T}\,\|\boldsymbol{X}\|_{2,\infty}$.
- $\|\nabla^2_{\boldsymbol{\theta}} \Phi(\boldsymbol{X};\boldsymbol{\theta})\| \leq 6d\,\|\boldsymbol{X}\|_{2,\infty}^2 \|\boldsymbol{X}\|_{1,\infty}^2 \sum_{t=1}^{T} \|\boldsymbol{X}\boldsymbol{u}_t\|_\infty + 2\sqrt{Td}\,\|\boldsymbol{X}\|_{2,\infty}^2 \|\boldsymbol{X}\|_{1,\infty}$.

Next, we focus on the empirical loss $\widehat{L}$. To derive bounds on its gradient and Hessian, we leverage the model's bounds from Proposition 1 and the fact that logistic loss is self-bounded, i.e., $|\ell'(t)| \leq \ell(t)$. To provide concrete statements, we introduce first a mild boundedness assumption.

Assumption 1 (Bounded data)**.** *Data $(\boldsymbol{X}, y) \in \mathbb{R}^{T\times d} \times \mathbb{R}$ satisfy the following conditions almost surely: $y \in \{\pm 1\}$, and for some $R \geq 1$, it holds for all $t \in [T]$ that $\|\boldsymbol{x}_t\| \leq R$.*

Corollary 1 (Loss properties)**.** *Under Assumption 1, the objective's gradient and Hessian satisfy the bounds:*[§]

[‡] A function $\ell : \mathbb{R} \to \mathbb{R}$ is self-bounded if $\exists\, C > 0$ such that $|\ell'(t)| \leq C\ell(t)$.

[§] In all the bounds in this paper involving $\|\widetilde{\boldsymbol{\theta}}\|_{2,\infty}$, it is possible to substitute this term with $\max_{h\in[H]} \|\boldsymbol{U}_h\|$. However, for the sake of notation simplicity, we opt for a slightly looser bound $\max_{h\in[H]} \|\boldsymbol{U}_h\| \leq \max_{h\in[H]} \|\boldsymbol{\theta}_h\| =: \|\widetilde{\boldsymbol{\theta}}\|_{2,\infty}$.

(1) $\|\nabla \widehat{L}(\widetilde{\boldsymbol{\theta}})\| \le \beta_1(\widetilde{\boldsymbol{\theta}})\,\widehat{L}(\widetilde{\boldsymbol{\theta}})\,,$ $\qquad \beta_1(\widetilde{\boldsymbol{\theta}}) \coloneqq \sqrt{T}\,R\,\left(2\,R^2\,\|\widetilde{\boldsymbol{\theta}}\|_{2,\infty}+1\right).$

(2) $\|\nabla^2 \widehat{L}(\widetilde{\boldsymbol{\theta}})\| \le \beta_2(\widetilde{\boldsymbol{\theta}})\,,$ $\qquad \beta_2(\widetilde{\boldsymbol{\theta}}) \coloneqq \frac{1}{\sqrt{H}}\,\beta_3(\widetilde{\boldsymbol{\theta}}) + \frac{1}{4}\,\beta_1(\widetilde{\boldsymbol{\theta}})^2\,.$

(3) $\lambda_{\min}(\nabla^2 \widehat{L}(\widetilde{\boldsymbol{\theta}})) \ge -\frac{\beta_3(\widetilde{\boldsymbol{\theta}})}{\sqrt{H}}\widehat{L}(\widetilde{\boldsymbol{\theta}})$ $\qquad \beta_3(\widetilde{\boldsymbol{\theta}}) \coloneqq 2\sqrt{T\,d}\,R^3\,\left(3\sqrt{d}\,R^2\,\|\widetilde{\boldsymbol{\theta}}\|_{2,\infty}+1\right).$

The loss properties above are crucial for the training and generalization analysis. Property (1) establishes self-boundedness of the empirical loss, which is used to analyze stability of GD updates for generalization. Property (2) is used to establish descent of gradient updates for appropriate choice of step-size η. Note that the smoothness upper bound is $\widetilde{\boldsymbol{\theta}}$-dependent, hence to show descent we need to also guarantee boundedness of the updates. Finally, property (3) establishes a self-bounded weak-convexity property of the loss, which is crucial to both the training and generalization analysis. Specifically, as the number of heads H increases, the minimum eigenvalue becomes less positive, indicating an approach towards convex-like behavior.

4 Main results

In this section, we present our training and generalization bounds for multi-head attention.

4.1 Training bounds

We state our main result on train loss convergence in the following theorem. See App. B for exact constants and the detailed proofs.

Theorem 1 (Training loss)**.** *Fix iteration horizon $K \ge 1$ and any $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^{H(dT+d^2)}$ and H satisfying*

$$\sqrt{H} \gtrsim d\,T^{1/2}R^5\|\widetilde{\boldsymbol{\theta}}\|_{2,\infty}\|\widetilde{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}^{(0)}\|^3. \tag{3}$$

Fix step-size $\eta \le 1 \wedge 1/\rho(\widetilde{\boldsymbol{\theta}}) \wedge \frac{\|\widetilde{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}^{(0)}\|^2}{K\,\widehat{L}(\widetilde{\boldsymbol{\theta}})} \wedge \frac{\|\widetilde{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}^{(0)}\|^2}{\widehat{L}(\widetilde{\boldsymbol{\theta}}^{(0)})}$, with $\rho(\widetilde{\boldsymbol{\theta}}) \lesssim d^{3/2}\,T^{3/2}\,R^{13}\,\|\widetilde{\boldsymbol{\theta}}\|_{2,\infty}^2\,\|\widetilde{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}^{(0)}\|^2$. Then, the following bounds hold for the training loss and the weights' norm at iteration K of GD:

$$\begin{aligned}&\widehat{L}(\widetilde{\boldsymbol{\theta}}^{(K)}) \le \frac{1}{K}\sum_{k=1}^{K}\widehat{L}(\widetilde{\boldsymbol{\theta}}_k) \le 2\widehat{L}(\widetilde{\boldsymbol{\theta}}) + \frac{5\|\widetilde{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}^{(0)}\|^2}{4\eta K},\\ &\|\widetilde{\boldsymbol{\theta}}^{(K)}-\widetilde{\boldsymbol{\theta}}^{(0)}\| \le 4\|\widetilde{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}^{(0)}\|.\end{aligned} \tag{4}$$

Yielding a concrete train loss bound requires an appropriate set of target parameters $\widetilde{\boldsymbol{\theta}}$ in the sense of minimizing the bound in (4). Hence, $\widetilde{\boldsymbol{\theta}}$ should simultaneously attain small loss ($\widehat{L}(\widetilde{\boldsymbol{\theta}})$) and distance to initialization ($\|\widetilde{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}^{(0)}\|$). This desiderata is formalized in Assumption 2 below. The distance to initialization, as well as $\|\widetilde{\boldsymbol{\theta}}\|_{2,\infty}$, determine how many heads are required for our bounds to hold. Also, in view of the bound in (4), it is reasonable that an appropriate choice for $\widetilde{\boldsymbol{\theta}}$ attains $\widehat{L}(\widetilde{\boldsymbol{\theta}})$ of same order as $\|\widetilde{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}^{(0)}\|^2/K$. Hence, the theorem's restriction on the step-size is governed by the inverse local-smoothness of the loss: $\eta \lesssim 1/\rho(\widetilde{\boldsymbol{\theta}})$.

4.2 Generalization bounds

Next we bound the expected generalization gap. Expectations are with respect to (w.r.t) randomness of the train set. See App. C for the detailed proof, which is based on algorithmic-stability.

Theorem 2 (Generalization loss)**.** *Fix any $K \ge 1$, any $\widetilde{\boldsymbol{\theta}}$ and H satisfying (3), and any step-size η satisfying the conditions of Thm. 1. Then the expected generalization gap at iteration K satisfies,*

$$\mathbb{E}\big[L(\widetilde{\boldsymbol{\theta}}^{(K)}) - \widehat{L}(\widetilde{\boldsymbol{\theta}}^{(K)})\big] \le \frac{4}{n}\,\mathbb{E}\big[2\,K\,\widehat{L}(\widetilde{\boldsymbol{\theta}}) + \frac{9\|\widetilde{\boldsymbol{\theta}}-\widetilde{\boldsymbol{\theta}}^{(0)}\|^2}{4\eta}\big]. \tag{5}$$

The condition on the number of heads is same up to constants to the corresponding condition in Theorem 1. Also, the generalization-gap bound translates to test-loss bound by combining with Thm. 1. Finally, similar

to Thm. 1, we can get concrete bounds under the realizability assumption; see Cor. 4 in App. C.2. For the generalization analysis, we require that the realizability assumption holds almost surely over all training sets sampled from the data distribution.

The bounds on optimization and generalization are up to constants same as analogous bounds for logistic regression (Soudry et al., 2018; Ji & Telgarsky, 2018; Shamir, 2021). Yet, for these bounds to be valid, we require sufficiently large number of heads as well as the existence of an appropriate set of target parameters $\widetilde{\boldsymbol{\theta}}$, as stated in the conditions of theorem. Namely, these conditions are related to the realizability condition, which guarantees small training error near initialization. The next assumption formalizes these conditions.

Assumption 2 (Realizability). *There exist non-increasing functions $g : \mathbb{R}_+ \to \mathbb{R}_+$ and $g_0 : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\forall \epsilon > 0$, there exists model parameters $\widetilde{\boldsymbol{\theta}}_{(\epsilon)} \in \mathbb{R}^{H(dT+d^2)}$ for which: (i) the empirical loss over n data samples satisfies $\widehat{L}(\widetilde{\boldsymbol{\theta}}_{(\varepsilon)}) \le \varepsilon$, (ii) $\|\widetilde{\boldsymbol{\theta}}_{(\varepsilon)} - \widetilde{\boldsymbol{\theta}}^{(0)}\| \le g_0(\varepsilon)$, and, (iii) $\|\widetilde{\boldsymbol{\theta}}_{(\varepsilon)}\|_{2,\infty} \le g(\varepsilon)$.*

With this assumption, we can specialize the result of Thms. above to specific data settings; see Cor. 3 and 4 in App. B.5 and C.2. In the next section we will further show how the realizability assumption is satisfied.

4.3 Primitive conditions for checking realizability

Here, we introduce a set of more primitive and straightforward-to-check conditions on the data and initialization that ensure the realizability Assumption 2 holds.

Definition 1 (Good initialization). *We say $\widetilde{\boldsymbol{\theta}}^{(0)} = \mathrm{concat}(\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_H^{(0)})$ is a `good` initialization with respect to training data $(\boldsymbol{X}_i, y_i)_{i\in[n]}$ provided the following three properties hold.*

P1. ***Parameter $L_{2,\infty}$-bound:*** *There exists parameter $B_2 \ge 1$ such that $\forall h \in [H]$ it holds $\|\boldsymbol{\theta}_h^{(0)}\|_2 \le B_2$.*

P2. ***Model bound:*** *There exists parameter $B_\Phi \ge 1$ such that $\forall i \in [n]$ it holds $\left|\widetilde{\Phi}(\boldsymbol{X}_i; \widetilde{\boldsymbol{\theta}}^{(0)})\right| \le B_\Phi$.*

P3. ***NTK separability:*** *There exists $\widetilde{\boldsymbol{\theta}}_\star \in \mathbb{R}^{H(dT+d^2)}$ and $\gamma > 0$ such that $\|\widetilde{\boldsymbol{\theta}}_\star\| = \sqrt{2}$ and $\forall i \in [n]$, it holds $y_i \left\langle \nabla \widetilde{\Phi}\left(\boldsymbol{X}_i; \widetilde{\boldsymbol{\theta}}^{(0)}\right), \widetilde{\boldsymbol{\theta}}_\star \right\rangle \ge \gamma$.*

Prop. 7 in the appendix shows that starting from a `good` initialization we can always find $\widetilde{\boldsymbol{\theta}}_{(\epsilon)}$ satisfying the realizability Assumption 2 provided large enough number of heads. Thus, given `good` initialization, we can immediately apply Theorems 1 and 2 to get the following concrete bounds.

Corollary 2 (General bounds under `good` initialization). *Suppose `good` initialization $\widetilde{\boldsymbol{\theta}}^{(0)}$ and let*

$$\sqrt{H} \gtrsim d\, T^{1/2}\, R^5\, B_2^2 \left(g_0(1/K)\right)^3, \qquad \text{where} \quad g_0\Big(\frac{1}{K}\Big) = \frac{2B_\Phi + \log(K)}{\gamma}.$$

Further fix step-size $\eta \le 1 \wedge 1/\rho(K) \wedge \frac{4B_\Phi^2}{\gamma^2 \log(1+e^{B_\Phi})}$ with $\rho(K) \gtrsim d^{3/2}\, T^{3/2}\, R^{13}\, g_0(\frac{1}{K})^4$. Then, it holds that

$$\widehat{L}(\widetilde{\boldsymbol{\theta}}^{(K)}) \le \frac{2}{K} + \frac{5\left(2B_\Phi + \log(K)\right)^2}{4\gamma^2\, \eta\, K}, \quad \text{and} \quad \mathbb{E}\big[L(\widetilde{\boldsymbol{\theta}}^{(K)}) - \widehat{L}(\widetilde{\boldsymbol{\theta}}^{(K)})\big] \le \frac{17\left(2B_\Phi + \log(K)\right)^2}{\gamma^2\, \eta\, n}.$$

Consider training loss after K GD steps: Assuming $B_\Phi = \widetilde{\mathcal{O}}_K(1)$ and $\gamma = \mathcal{O}_K(1)$, then choosing $\eta = \widetilde{\mathcal{O}}_K(1)$, the corollary guarantees train loss is $\widetilde{\mathcal{O}}_K(\frac{1}{K})$ provided polylogarithmic number of heads $H = \Omega(\log^6(K))$. Moreover, after $K \approx n$ GD steps the expected test loss is $\mathcal{O}(1/n)$.

Remark 1. *The last two conditions (P2 and P3) for `good` initialization are similar to the conditions needed in (Taheri & Thrampoulidis, 2023; Ji & Telgarsky, 2020; Nitanda et al., 2019) for analysis of two-layer MLPs. Compared to (Ji & Telgarsky, 2020; Nitanda et al., 2019) which assume random Gaussian initialization $\widetilde{\boldsymbol{\theta}}^{(0)}$, and similar to (Taheri & Thrampoulidis, 2023) the NTK separability assumption (P3) can potentially accommodate deterministic $\boldsymbol{\theta}^{(0)}$. Condition (P1) appears because we allow training both layers of the model. Specifically the $L_{2,\infty}$ norm originates from the Hessian bounds in Corollary 1.*

5 Application to tokenized-mixture model

We now demonstrate through an example how our results apply to specific data models.

Data model: An example. Consider $M+2$ distinct patterns $\{\boldsymbol{\mu}_+, \boldsymbol{\mu}_-, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, ..., \boldsymbol{\nu}_M\}$, where discriminative patterns $\boldsymbol{\mu}_\pm$ correspond to labels $y = \pm 1$. The tokens are split into (i) a label-relevant set ($\mathcal{R}$) and (ii) a label-irrelevant set ($\mathcal{R}^c := [T] \backslash \mathcal{R}$). Conditioned on the label and $\mathcal{R}$, the tokens $\boldsymbol{x}_t, t \in [T]$ are IID as follows

$$\boldsymbol{x}_t | y = \begin{cases} \boldsymbol{\mu}_y & , t \in \mathcal{R} \\ \boldsymbol{\nu}_{j_t} + \boldsymbol{z}_t & , j_t \sim \text{Unif}(1, ..., M) \text{ and } t \in \mathcal{R}^c, \end{cases} \tag{DM1}$$

where $\boldsymbol{z}_t$ are noise vectors. Let $\mathcal{D}$ denote the joint distribution induced by the described $(\mathbf{X}, y)$ pairs.

Assumption 3. *The labels are equi-probable and we further assume the following:*

- ***Orthogonal, equal-energy means:*** *All patters are orthogonal to each other, i.e.* $\boldsymbol{\mu}_+ \perp \boldsymbol{\mu}_- \perp \boldsymbol{\nu}_\ell \perp \boldsymbol{\nu}_{\ell'}$, $\forall\, \ell, \ell' \in [M]$. *Also, for all* $y \in \{\pm 1\}, \ell \in [M]$ *that* $\|\boldsymbol{\mu}_y\| = \|\boldsymbol{\nu}_\ell\| = S$, *where* S *denotes the* signal strength.
- ***Sparsity level:*** *The number of label-relevant tokens is* $|\mathcal{R}| = \zeta T$; *for sparsity level* $\zeta \in (0, 1)$.
- ***Noise distribution:*** *The noise tokens* $\boldsymbol{z}_t$ *are sampled from a distribution* $\mathcal{D}_z$, *such that it holds almost surely for* $\boldsymbol{z}_t \sim \mathcal{D}_z$ *that* $|\langle \boldsymbol{z}_t, \boldsymbol{\mu}_y \rangle| \le Z_\mu$, $y \in \{\pm 1\}$ *and* $|\langle \boldsymbol{z}_t, \boldsymbol{\nu}_\ell \rangle| \le Z_\nu / M$, $\forall\, \ell \in [M]$. *Moreover,* $\|\boldsymbol{z}_t\| \le Z$. *Overall, Assumption 1 is satisfied with* $R = \sqrt{S^2 + Z^2 + 2Z_\nu / M}$.

The above assumptions can be relaxed, but without contributing new insights. We have chosen to present a model that is representative and transparent in its analysis.

5.1 Finding a good initialization

To apply the general Corollary 2 to the specific data model DM1, it suffices to find good initialization. While we cannot directly show that $\widetilde{\boldsymbol{\theta}}^{(0)} = 0$ is good, we can show this is the case for first step of gradient descent $\widetilde{\boldsymbol{\theta}}^{(1)}$. Thus, we consider training in two phases as follows.

First phase: One step of GD as initialization. We use n_1 training samples to update the model parameters by running one-step of gradient descent starting from zero initialization. Specifically,

$$(\boldsymbol{U}_h^{(1)}, \boldsymbol{W}_h^{(1)}) = \boldsymbol{\theta}_h^{(1)} = \boldsymbol{\theta}_h^{(0)} - \alpha_h \sqrt{H} \cdot \nabla_{\boldsymbol{\theta}_h} \widehat{L}_{n_1}(\boldsymbol{\theta}_h^{(0)}), \text{ where } \quad \boldsymbol{\theta}_h^{(0)} = \mathbf{0} \;\forall\; h \in [H].$$

Here, α_h denotes the step-size for head $h \in [H]$ and the scaling by $\sqrt{H}$ guarantees the update of each head is $\mathcal{O}(1)$. The lemma below shows that at the end of this phase, we have $\|\boldsymbol{U}_h^{(1)} - \frac{\zeta \alpha_h}{2} \mathbb{1}_T \boldsymbol{u}_\star^\top\|_F = O(1/\sqrt{n_1})$, where $\boldsymbol{u}_\star$ is the oracle classifier $\boldsymbol{u}_\star = \boldsymbol{\mu}_+ - \boldsymbol{\mu}_-$. On the other hand, the attention weight-matrix does *not* get updated; the interesting aspect of the training lies in the second phase, which involves updating $\boldsymbol{W}$.

Lemma 2 (First phase)**.** *After the first-gradient step as described above, we have* $\boldsymbol{U}_h^{(1)} = \alpha_h \mathbb{1}_T \left(\frac{\zeta}{2} \boldsymbol{u}_\star^\top + \boldsymbol{p}^\top\right)$ *and* $\boldsymbol{W}_h^{(1)} = \mathbf{0}$. *where with probability at least* $1 - \delta \in (0, 1)$ *over the randomness of labels there exists positive universal constant* $C > 0$ *such that*

$$\|\boldsymbol{p}\| \le C\,(2S + Z)\left(\sqrt{\frac{d}{n_1}} + \sqrt{\frac{\log(1/\delta)}{n_1}}\right) =: P\,. \tag{6}$$

Second phase: GD with constant step size. During the second phase, K gradient steps are performed on n new samples (distinct from those used in the first phase). Concretely, $\widetilde{\boldsymbol{\theta}}^{(k+1)} = \widetilde{\boldsymbol{\theta}}^{(k)} - \eta \cdot \nabla_{\widetilde{\boldsymbol{\theta}}} \widehat{L}_n(\widetilde{\boldsymbol{\theta}}^{(k)})$, $\quad k = 1, \ldots, K$, with $\widetilde{\boldsymbol{\theta}}^{(1)} = \text{concat}\left(\{\boldsymbol{\theta}_h^{(1)}\}_{h \in [H]}\right)$ the step obtained by the first-phase update and η the step-size of the second phase. In order to analyze the second phase, during which both $\widetilde{\boldsymbol{W}}$ and $\widetilde{\boldsymbol{U}}$ get updated, we employ the general results of Section 4. To do so, we show that $\widetilde{\boldsymbol{\theta}}^{(1)}$ serves as good initialization as per Definition 1.

Proposition 2. *Consider the first-phase iterate* $\{\boldsymbol{\theta}_h^{(1)}\}_{h \in [H]}$ *and condition on the event* $\|\boldsymbol{p}\| \le P$ *(depending only on the data randomness in the first phase) of Lemma 2. Suppose the step-size of the first phase is chosen IID* $\alpha_h \sim \text{Unif}(\pm 1)$, $h \in [H]$. *Then, the initialization* $\widetilde{\boldsymbol{\theta}}^{(1)} = \text{concat}\left(\boldsymbol{\theta}_1^{(1)}, \ldots, \boldsymbol{\theta}_H^{(1)}\right)$ *is* ***good*** *with respect to data sampled from DM1 and satisfying Assumption 3. Specifically, the three desired properties hold as follows.*

- *Almost surely,* ***P1*** *holds with* $B_2 = \sqrt{T}(S + P)$.

- *With probability* $1-\delta \in (0,1)$, ***P2*** *holds with* $B_\Phi = TR(S+P)\sqrt{2\log(n/\delta)}$.
- *Suppose* $\sqrt{H} \gtrsim \frac{R^4 T(S+P)}{\gamma_\star} \cdot \sqrt{2\log(n/\delta)}$. *Then, with probability* $1-\delta \in (0,1)$, ***P3*** *holds with* $\gamma = \gamma_\star/2$ *where*

$$\gamma_\star := \frac{T(1-\zeta)\zeta\left(\zeta S^4 - 7\bar{Z}S^2 - 12\bar{Z}^2 - 16\frac{\bar{Z}^3}{S^2}\right)}{4\sqrt{2(M+1)}} - PT^{5/2}(S+Z)^3 + \frac{S\sqrt{T}\left(\zeta - 2(1-\zeta)\frac{Z_\mu}{S^2}\right)}{\sqrt{2}}, \tag{7}$$

and $\bar{Z} := Z_\mu \vee Z_\nu$. *The randomness is with respect to the sampling of* α_h, $h \in [H]$.

The parameter $\gamma_\star$ in (7) represents the NTK margin of the model at initialization $\widetilde{\boldsymbol{\theta}}^{(1)}$. By Corollary 2, larger margin translates to better train/generalization bounds and smaller requirements on the number of heads. For a concrete example, suppose $T \vee M = \mathcal{O}(1)$ and $Z \vee \bar{Z} = \mathcal{O}(S)$. Then, provided first-phase sample size $n_1 \gtrsim S^2 d$ so that $P = \mathcal{O}(1)$, it holds $\gamma_\star = \gamma_{\text{lin}} + \Omega(\zeta^2(1-\zeta)S^4)$, where $\gamma_{\text{lin}} = \Omega(\zeta S)$ is the margin of a linear model for the same dataset (see App. F). Overall, applying Cor. 2 for $K = n$ and a polylogarithmic polylog(n) number of heads leads to $\tilde{\mathcal{O}}\left(\frac{1}{\eta\gamma_\star^2 n}\right)$ train loss and expected generalization gap.

5.2 Proof sketch of P3: NTK separability

It is instructive to see how **P3** follows as it sheds light on the choice of an appropriate target parameter $\widetilde{\boldsymbol{\theta}}$ as per Thms. 1 and 2. We choose

$$\boldsymbol{W}_\star = \boldsymbol{\mu}_+\boldsymbol{\mu}_+^\top + \boldsymbol{\mu}_-\boldsymbol{\mu}_-^\top + \sum_{\ell\in[M]} \boldsymbol{\nu}_\ell(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)^\top \quad \text{and} \quad \boldsymbol{U}_\star = \mathbb{1}_T \boldsymbol{u}_\star^\top = \mathbb{1}_T(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^\top,$$

and normalize parameters such that $\boldsymbol{\theta}_\star := (\overline{\boldsymbol{U}}_\star = \frac{1}{\|\boldsymbol{U}_\star\|_F}\boldsymbol{U}_\star,\ \texttt{sign}(\alpha)\overline{\boldsymbol{W}}_\star = \texttt{sign}(\alpha)\frac{1}{\|\boldsymbol{W}_\star\|_F}\boldsymbol{W}_\star)$. It is easy to see that $\boldsymbol{U}_\star$ is the optimal classifier for the label-relevant tokens. To gain intuition on the choice of $\boldsymbol{W}_\star$, note that $\boldsymbol{W}_\star = \boldsymbol{W}_{K,\star}\boldsymbol{W}_{Q,\star}^\top$, with key-query matrices chosen as $\boldsymbol{W}_{K,\star} = \begin{bmatrix}\boldsymbol{\mu}_+ & \boldsymbol{\mu}_- & \boldsymbol{\nu}_1 & \cdots & \boldsymbol{\nu}_M\end{bmatrix} \in \mathbb{R}^{d\times(M+2)}$ and $\boldsymbol{W}_{Q,\star} = \begin{bmatrix}\boldsymbol{\mu}_+ & \boldsymbol{\mu}_- & \boldsymbol{\mu}_+ + \boldsymbol{\mu}_- & \cdots & \boldsymbol{\mu}_+ + \boldsymbol{\mu}_-\end{bmatrix} \in \mathbb{R}^{d\times(M+2)}$. With these choices, the relevance scores (aka softmax logits) of relevant tokens turn out to be strictly larger compared to the irrelevant tokens. Concretely, we show in App. D.2.3 that the t-th row $\boldsymbol{r}_t(\boldsymbol{X};\boldsymbol{W}_\star) = \boldsymbol{X}\boldsymbol{W}_\star^\top\boldsymbol{x}_t$ of the softmax-logit matrix satisfies the following:

$$\forall t : [\boldsymbol{r}_t]_{t'} = \begin{cases}\mathcal{O}(S^4) & , t' \in \mathcal{R}, \\ \mathcal{O}(S^2) & , t' \in \mathcal{R}^c. \end{cases} \tag{8}$$

Thus, under this parameter choice, softmax can attend to label-relevant tokens and supresses noisy irrelevant tokens. In turn, this increases the signal-to-noise ratio for classification using $\boldsymbol{U}_\star$.

We now show how to compute $\mathbb{E}_{\boldsymbol{\theta}^{(1)}} y\langle\nabla_{\boldsymbol{\theta}}\Phi(\boldsymbol{X};\boldsymbol{\theta}^{(1)}), \boldsymbol{\theta}_\star\rangle$ for a single head. Recall $\boldsymbol{\theta}_\star$ consists of $\overline{\boldsymbol{U}}_\star$, $\overline{\boldsymbol{W}}_\star$. First, since $\boldsymbol{W}^{(1)} = \boldsymbol{0}$, using Assumption 3, a simple calculation shows $y\langle\nabla_{\boldsymbol{U}}\Phi(\boldsymbol{X};\boldsymbol{\theta}^{(1)}), \overline{\boldsymbol{U}}_\star\rangle \geq \frac{S\sqrt{T}}{\sqrt{2}}\left(\zeta - 2(1-\zeta)\frac{Z_\mu}{S^2}\right)$. Second, to compute $\mathbb{E}_{\alpha\sim\text{Unif}(\pm1)} y\langle\nabla_{\boldsymbol{W}}\Phi\left(\boldsymbol{X};\boldsymbol{\theta}^{(1)}\right), \texttt{sign}(\alpha)\overline{\boldsymbol{W}}_\star\rangle$ it follows from Lemma 1 that

$$\nabla_{\boldsymbol{W}}\Phi(\boldsymbol{X};\boldsymbol{\theta}^{(1)}) = \frac{\alpha\zeta}{2}\sum_{t\in[T]} \boldsymbol{x}_t\boldsymbol{u}_\star^\top\boldsymbol{X}^\top\boldsymbol{\varphi}'(\boldsymbol{0})\boldsymbol{X} + \alpha\sum_{t\in[T]} \boldsymbol{x}_t\boldsymbol{p}^\top\boldsymbol{X}^\top\boldsymbol{\varphi}'(\boldsymbol{0})\boldsymbol{X}.$$

Note the first term is dominant here since the second term can be controlled by making $\|\boldsymbol{p}\|_2$ small as per Lemma 2. Thus, ignoring here the second term (see Appendix D.2.3 for full calculation) $y\langle\nabla_{\boldsymbol{W}}\Phi(\boldsymbol{X};\boldsymbol{\theta}^{(1)}), \boldsymbol{W}_\star\rangle$ is governed by the following term: $\frac{\alpha\zeta}{2}\sum_{t\in[T]} y\,\boldsymbol{u}_\star^\top\boldsymbol{X}^\top\boldsymbol{\varphi}'(\boldsymbol{0})\boldsymbol{X}\boldsymbol{W}_\star^\top\boldsymbol{x}_t = \frac{\alpha\zeta}{2}\sum_{t\in[T]} y\,\boldsymbol{u}_\star^\top\boldsymbol{X}^\top\boldsymbol{\varphi}'(\boldsymbol{0})\,\boldsymbol{r}_t$. Note that $\boldsymbol{\varphi}'(\boldsymbol{0}) = \mathbf{I} - \frac{1}{T}\mathbb{1}_T\mathbb{1}_T^\top$. To simplify the exposition here, let us focus on the identity component and leave treatment of the the rank-one term to the detailed proof. The corresponding term then becomes

$$\frac{\alpha\zeta}{2}\sum_{t\in[T]}\sum_{t'\in[T]} \underbrace{\left(y\,\boldsymbol{u}_\star^\top\boldsymbol{x}_{t'}\right)}_{\text{class. logits}} \cdot \underbrace{\left([\boldsymbol{r}_t]_{t'}\right)}_{\text{softmax logits}},$$

which involves for each output token t, the sum of products over all tokens $t' \in [T]$ of softmax logits (i.e. relevant scores $[\boldsymbol{r}_t]_{t'}$) and corresponding classification logits (i.e. $y\,\boldsymbol{u}_\star^\top \boldsymbol{x}_{t'}$). Note that by choice of $\boldsymbol{u}_\star$ and $\boldsymbol{W}_\star$, both the classification and softmax logits are large from label-relevant tokens, while being small for noise tokens. Intuitively, this allows for a positive margin $\gamma_\star$ as stated in Proposition 2. We defer the detailed calculations to Appendix D.2.3.

In the appendix, we also detail how to yield the computation for the MHA, which builds on the calculations for the single-head attention model above. In short, we simply choose multi-head parameter $\widetilde{\boldsymbol{\theta}}_\star$ as $\widetilde{\boldsymbol{\theta}}_\star = \frac{1}{\sqrt{H}}\operatorname{concat}\left(\boldsymbol{\theta}_\star(\boldsymbol{\theta}_1^{(1)}), \ldots, \boldsymbol{\theta}_\star(\boldsymbol{\theta}_H^{(1)})\right)$. This guarantees that $\|\widetilde{\boldsymbol{\theta}}_\star\| = \sqrt{2}$ and maintains the multi-head NTK margin be at least $\gamma_\star$ in expectation. To complete the proof, it remains to get a high-probability version of this bound. To do this, notice that $\boldsymbol{\theta}_h^{(1)}$ are IID, hence we can apply Hoeffding's inequality, which finally gives the desired bound on the NTK margin provided sufficient number of heads H, which controls the degree of concentration when applying Hoeffding's inequality. See Lemmas 14 and 15 for details.

5.3 Is the NTK margin optimal?

Below, we discuss the optimality of the NTK margin $\gamma_\star$. First, define set of parameters $\boldsymbol{\theta}_{\text{opt}} := (\boldsymbol{U}_{\text{opt}}, \boldsymbol{W}_{\text{opt}})$:

$$\boldsymbol{U}_{\text{opt}} := \frac{1}{S\sqrt{2T}}\boldsymbol{U}_\star \qquad \text{and} \qquad \boldsymbol{W}_{\text{opt}} := \frac{1}{S^2\sqrt{2(M+1)}}\boldsymbol{W}_\star\,, \tag{9}$$

normalized so that $\|\boldsymbol{\theta}_{\text{opt}}\|_F = \sqrt{2}$. Recall here the definitions of $\boldsymbol{U}_\star, \boldsymbol{W}_\star$ in the section above. As we already explained above, this choice of parameters guarantees that relevant tokens are assigned larger relevance and classification scores compared to irrelevant ones. Specifically about $\boldsymbol{W}_\star$, we saw in Eq. (8) that it ensures a gap of $\mathcal{O}(S^2)$ between relevance scores of label-relevant and label-irrelevant tokens. Thanks to this gap, it is possible for softmax to fully attend to the label-relevant tokens by saturating the softmax. To do this, it suffices to scale-up $\boldsymbol{W}_\star$ by an amount $\propto 1/S^2$. This is formalized in the proposition below.

Proposition 3 (Attention expressivity for tokenized mixture model)**.** *Consider single-head attention model. Suppose the noise level is such that $Z_\mu = Z_\nu \le S^2/8$. For any $\epsilon > 0$, consider Γ_ϵ satisfying $\Gamma_\epsilon \ge \frac{8\sqrt{2(M+1)}}{3S^2}\log\left(\frac{\zeta^{-1}-1}{\epsilon}\right)$. Then, the attention scores corresponding to weights $\Gamma_\epsilon \cdot \boldsymbol{W}_{opt}$ satisfy*

$$\forall t \in [T] : 0 \le 1 - \sum_{t' \in \mathcal{R}} \boldsymbol{\varphi}_{t'}(\boldsymbol{x}_t^\top \Gamma_\epsilon \boldsymbol{W}_{opt} \boldsymbol{X}^T) = \sum_{t' \in \mathcal{R}^c} \boldsymbol{\varphi}_{t'}(\boldsymbol{x}_t^\top \Gamma_\epsilon \boldsymbol{W}_{opt} \boldsymbol{X}^T) \le \epsilon\,. \tag{10}$$

Thus, almost surely over data $(\boldsymbol{X}, y)$ generated from data model DM1 the margin of single-head attention with parameters $(\boldsymbol{U}_{opt}, \Gamma_\epsilon \cdot \boldsymbol{W}_{opt})$ satisfies

$$y\Phi(\boldsymbol{X}; \boldsymbol{U}_{opt}, \Gamma_\epsilon \cdot \boldsymbol{W}_{opt}) \ge \gamma_{attn} := \gamma_{attn}(\epsilon) := \frac{\sqrt{T}}{\sqrt{2}S}\left(S^2(1-\epsilon) - 2\epsilon Z_\mu\right)\,. \tag{11}$$

From Eq. (10), note that as $\epsilon \to 0$ and $\Gamma_\epsilon \to \infty$, the softmax map saturates, i.e. it approaches a hard-max map that attends only to the label-relevant tokens ($\mathcal{R}$) and suppress the rest ($\mathcal{R}^c$). As a consequence of this, Eq. (11) shows that the achieved margin approaches $\gamma_{\text{attn}} := S\sqrt{T}/\sqrt{2}$. Note this is independent of the sparsity level ζ. In particular, $\gamma_{\text{attn}} \ge \gamma_\star \ge \gamma_{\text{lin}}$ and the gap increases with decreasing sparsity. See appendix for experiments and discussion regarding the margin achieved by GD for data model DM1.

Following Proposition 3, a natural question arises: Is it possible to choose "good" parameters $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{U}}, \widetilde{\boldsymbol{W}})$ based on the set of optimal parameters $\boldsymbol{\theta}_{\text{opt}}$? This would then yield train-loss and expected generalization-gap bounds $\tilde{\mathcal{O}}\left(1/(\eta\gamma_{\text{attn}}^2 n)\right)$ after $\Theta(n)$ steps of GD starting at $\widetilde{\boldsymbol{\theta}}^{(0)} = \boldsymbol{0}$. To investigate this question, define the following parameters for each head, aligning with the aforementioned "good" directions of Proposition 3:

$$\boldsymbol{U}_h := \frac{\log(n)}{\gamma_{\text{attn}}}\frac{1}{H^{1/2}}\boldsymbol{U}_{\text{opt}}, \quad \boldsymbol{W}_h := \frac{C}{H^p}\boldsymbol{W}_{\text{opt}}\,,$$

for some $C > 0$, $p > 0$, and $\forall h \in [H]$. To yield the margin γ_{attn} of (10), we need that each $\boldsymbol{W}_h$ has norm at least $\Gamma_\epsilon \propto 1/S^2$. Thus, we need $\|\boldsymbol{W}_h\| \gtrsim \frac{1}{S^2} \implies S^2 \gtrsim \frac{1}{C}\cdot H^p$. Now, in order to apply Thms. 1 and 2, the

requirement on the number of heads H in terms of distance of $\widetilde{\boldsymbol{\theta}}$ to $\widetilde{\boldsymbol{\theta}}^{(0)} = \mathbf{0}$ yields the following condition:

$$H^{1/2} \gtrsim S^5 \|\widetilde{\boldsymbol{\theta}}\|^3 . \tag{12}$$

Note that $\|\widetilde{\boldsymbol{U}}\| = \frac{\log(n)}{\gamma_{\text{attn}}} \approx \frac{\log(n)}{S}$, $\|\widetilde{\boldsymbol{W}}\| = C \cdot H^{1/2-p}$. Hence, in computing $\|\widetilde{\boldsymbol{\theta}}\|$, we distinguish two cases. First, assume that $\|\widetilde{\boldsymbol{W}}\| \geq \|\widetilde{\boldsymbol{U}}\|$ which implies that $S \gtrsim \frac{\log(n)}{C} \cdot H^{p-1/2}$ and $\|\widetilde{\boldsymbol{\theta}}\| \gtrsim \|\widetilde{\boldsymbol{W}}\| \vee \|\widetilde{\boldsymbol{U}}\| = C \cdot H^{1/2-p}$. Since

$$S \gtrsim \frac{1}{C^{1/2}} \cdot H^{p/2} \vee \frac{\log(n)}{C} \cdot H^{p-1/2} ,$$

by using Eq. (12), we get the following conditions on H:

$$H^{1/2} \gtrsim S^5 \cdot C^3 \cdot H^{3/2-3p} \gtrsim C^{1/2} \cdot H^{3/2-p/2} \vee \frac{\log^5(n)}{C^2} \cdot H^{2p-1} \implies H^{p-2} \gtrsim C \;\text{ and }\; H^{p-1/4} \lesssim \frac{C}{\log^{5/2}(n)} .$$

Combining these two gives $C \lesssim \frac{C}{\log^{5/2}(n)} \implies \log(n) \lesssim 1$, a contradiction since $n > 1$. Thus, there are no possible choices for p and C that satisfy both conditions. The case $\|\widetilde{\boldsymbol{W}}\| \leq \|\widetilde{\boldsymbol{U}}\|$ can be treated similarly leading to the same conclusion; thus, is omitted for brevity.

Intuitively, this contradiction arises because of the large $\|\boldsymbol{W}_h\|$ requirement to achieve margin γ_{attn}. Finally, one can ask if it is possible to resolve the contradiction by changing the scaling of normalization with respect to H in the MHA model Eq. (2), from $1/H^{1/2}$ to $1/H^c$ for $c > 0$. It can be shown via the same argument that no such value of c exists for which $\widetilde{\boldsymbol{\theta}}$ constructed above satisfies the overparameterization requirement $H^c \gtrsim S^5 \|\widetilde{\boldsymbol{\theta}}\|^3$. We thus conclude that the construction of weights in Proposition 3 does not yield a target parameter that simultaneously achieves low empirical loss and allows choosing H large enough as per (3). This triggers interesting questions for future research: Does GD converge to weights attaining margin γ_{attn} as in Proposition 3? If so, under what conditions on initialization? See also the remarks in Section 7.

6 Proof Sketch of Section 4

Throughout this section we drop the $\widetilde{\cdot}$ in $\widetilde{\boldsymbol{\theta}}$ and $\widetilde{\Phi}(\boldsymbol{X}_i; \widetilde{\boldsymbol{\theta}})$ as everything refers to the full model. Moreover, $\widetilde{\boldsymbol{\theta}}^{(K)}$ and $\widetilde{\boldsymbol{\theta}}^{(0)}$ are denoted by $\boldsymbol{\theta}_K$ and $\boldsymbol{\theta}_0$.

6.1 Training analysis

The proof begins by showing step-wise descent for any iteration $k \geq 0$ of GD (see Lemma 7), where step-size at each iteration $\eta_k \leq \frac{1}{\rho_k}$ depends on the objective's local smoothness parameters $\rho_k = \beta_2(\boldsymbol{\theta}_k) \vee \beta_2(\boldsymbol{\theta}_{k+1})$:

$$\widehat{L}(\boldsymbol{\theta}_{k+1}) \leq \widehat{L}(\boldsymbol{\theta}_k) - \frac{\eta_k}{2} \left\|\nabla \widehat{L}(\boldsymbol{\theta}_k)\right\|^2 . \tag{13}$$

Now, using Taylor's theorem we can link $\widehat{L}(\boldsymbol{\theta}_k)$ to $\widehat{L}(\boldsymbol{\theta})$ for any $\boldsymbol{\theta}$ as follows:

$$\widehat{L}(\boldsymbol{\theta}) \leq \widehat{L}(\boldsymbol{\theta}_k) + \langle \nabla \widehat{L}(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_k \rangle + \frac{1}{2} \min_{\boldsymbol{\theta}_{k_\alpha}} \lambda_{\min}\left(\nabla^2 \widehat{L}(\boldsymbol{\theta}_{k_\alpha})\right) \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|^2, \tag{14}$$

where $\boldsymbol{\theta}_{k_\alpha} := \alpha \boldsymbol{\theta}_k + (1-\alpha)\boldsymbol{\theta},\ \alpha \in [0,1]$. We can plug this into (13) to relate the loss at iterates $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k+1}$. To continue, we need to lower bound $\min_{\boldsymbol{\theta}_{k_\alpha}} \lambda_{\min}\left(\nabla^2 \widehat{L}(\boldsymbol{\theta}_{k_\alpha})\right)$. For this, we use the following property of the loss objective from Corollary 1: $\forall \boldsymbol{\theta} :\ \lambda_{\min}\left(\nabla^2 \widehat{L}(\boldsymbol{\theta})\right) \geq -\kappa(\boldsymbol{\theta}) \cdot \widehat{L}(\boldsymbol{\theta})$, where $\kappa(\boldsymbol{\theta}) := \frac{\beta_3(\boldsymbol{\theta})}{\sqrt{H}}$. Note from the definition of $\beta_3(\cdot)$ that $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 :\ \max_{\boldsymbol{\theta} \in [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]} \beta_3(\boldsymbol{\theta}) = \beta_3(\boldsymbol{\theta}_1) \vee \beta_3(\boldsymbol{\theta}_2)$. Thus, the above property of the loss implies the following *local self-bounded weak convexity* property on the line $[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]$ for arbitrary points $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$:

$$\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 :\ \min_{\boldsymbol{\theta} \in [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]} \lambda_{\min}\left(\nabla^2 \widehat{L}(\boldsymbol{\theta})\right) \geq -\frac{\beta_3(\boldsymbol{\theta}_1) \vee \beta_3(\boldsymbol{\theta}_2)}{\sqrt{H}} \cdot \max_{\boldsymbol{\theta} \in [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]} \widehat{L}(\boldsymbol{\theta}) . \tag{15}$$

Therefore, using Eq. (15) in Eq. (14), we can get:

$$\widehat{L}(\boldsymbol{\theta}) \geq \widehat{L}(\boldsymbol{\theta}_k) + \langle \nabla \widehat{L}(\boldsymbol{\theta}_k), \boldsymbol{\theta} - \boldsymbol{\theta}_k \rangle - \frac{1}{2} \frac{\beta_3(\boldsymbol{\theta}_1) \vee \beta_3(\boldsymbol{\theta}_2)}{\sqrt{H}} \cdot \max_{\alpha \in [0,1]} \widehat{L}(\boldsymbol{\theta}_{k_\alpha}) \, \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|^2 \,. \tag{16}$$

To apply the Descent Lemma in (13), we need to fix a step-size such that satisfies the condition of the Lemma at each iteration $\eta \leq \eta_k$ for all $k < K$. Then, combining with Eq. (16) and applying standard telescope summation, we arrive at the following:

$$\frac{1}{K} \sum_{k=1}^{K} \widehat{L}(\boldsymbol{\theta}_k) \leq \widehat{L}(\boldsymbol{\theta}) + \frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2}{2\eta K} + \frac{1}{2K} \sum_{k=0}^{K-1} \frac{\beta_3(\boldsymbol{\theta}) \vee \beta_3(\boldsymbol{\theta}_k)}{\sqrt{H}} \cdot \max_{\alpha \in [0,1]} \widehat{L}(\boldsymbol{\theta}_{k_\alpha}) \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|^2 \,. \tag{17}$$

Next, we use the following generalized local quasi-convexity (GLQC) of the loss function.

Proposition 4 (GLQC property: Slight variation of Prop. 8 of Taheri & Thrampoulidis (2023))**.** *Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be two points that are sufficiently close to each other, such that*

$$2 \left(\beta_3(\boldsymbol{\theta}_1) \vee \beta_3(\boldsymbol{\theta}_2)\right) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 \leq \sqrt{H} \,. \tag{18}$$

Then, $\max_{\boldsymbol{\theta} \in [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2]} \widehat{L}(\boldsymbol{\theta}) \leq \frac{4}{3} \left(\widehat{L}(\boldsymbol{\theta}_1) \vee \widehat{L}(\boldsymbol{\theta}_2)\right)$.

Using Proposition 4 in Eq. (17) and assuming sufficiently large heads H such that $\sqrt{H} \geq 2 \left(\beta_3(\boldsymbol{\theta}) \vee \beta_3(\boldsymbol{\theta}_k)\right) \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|^2$, we can get the advertised regret bound in (4).

In order to remove the dependence of H on iteration k, by an induction argument we can show bounded iterates-norm i.e. $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}\| \leq 3\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$ (see Lemma 10). Using this and the definition of $\beta_3(\cdot)$ we can control $\beta_3(\boldsymbol{\theta}) \vee \beta_3(\boldsymbol{\theta}_k)$ as $\left(\beta_3(\boldsymbol{\theta}) \vee \beta_3(\boldsymbol{\theta}_k)\right) \lesssim \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \|\boldsymbol{\theta}\|_{2,\infty}$ to get the desired requirement of heads $\sqrt{H} \gtrsim \|\boldsymbol{\theta}\|_{2,\infty} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^3$ stated in Eq. (3).

The remaining piece to guarantee descent at each step is establishing a $\rho(\boldsymbol{\theta})$ such that $\rho_k \leq \rho(\boldsymbol{\theta})$ for all $k < K$. To do this, we recall that $\rho_k = \beta_2(\boldsymbol{\theta}_k) \vee \beta_2(\boldsymbol{\theta}_{k+1})$. By definition of $\beta_2(\cdot)$ in Corollary 1, we can control $\beta_2(\boldsymbol{\theta}_k) \vee \beta_2(\boldsymbol{\theta}_{k+1})$ with controlling $\|\boldsymbol{\theta}_k\|_{2,\infty} \vee \|\boldsymbol{\theta}_{k+1}\|_{2,\infty}$ as $\left(\|\boldsymbol{\theta}_k\|_{2,\infty} \vee \|\boldsymbol{\theta}_{k+1}\|_{2,\infty}\right) \lesssim \|\boldsymbol{\theta} - \boldsymbol{\theta}_k\| + \|\boldsymbol{\theta}\|_{2,\infty} + 1$. Using iterates-norm bound and setting $\rho(\boldsymbol{\theta}) = \left(\frac{2\sqrt{T\,d}\,R^3}{\sqrt{H}} + \frac{T\,R^2}{4}\right) \alpha(\boldsymbol{\theta})^2$ with $\alpha(\boldsymbol{\theta}) := 3\sqrt{d}\,R^2 \left[3\sqrt{T}\,R^3 \left(3\,\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| + \|\boldsymbol{\theta}\|_{2,\infty}\right) + 2\sqrt{T}\,R\right]$, satisfies the desired condition for the Descent Lemma completing the proof.

6.2 Generalization analysis

In order to bound the expected generalization gap, we leverage the algorithmic stability framework. To begin, consider the leave-one-out (loo) training loss $\widehat{L}^{\neg i}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{j \neq i} \ell_j(\boldsymbol{\theta})$ for $i \in [n]$, where $\ell_j(\boldsymbol{\theta}) := \ell(y_j \Phi(\boldsymbol{X}_j; \boldsymbol{\theta}))$ denotes the j-th sample loss. With these, define the loo model updates of GD on the loo loss for $\eta > 0$:

$$\boldsymbol{\theta}_{k+1}^{\neg i} := \boldsymbol{\theta}_k^{\neg i} - \eta \nabla \widehat{L}^{\neg i}(\boldsymbol{\theta}_k^{\neg i}), \;\; k \geq 0, \qquad \boldsymbol{\theta}_0^{\neg i} = \boldsymbol{\theta}_0 \,.$$

The following lemma relates expected generalization loss to average model stability for any G-Lipschitz loss.

Lemma 3 (Lei & Ying (2020), Thm. 2)**.** *For G-Lipschitz loss and for all iterates K, it holds that* $\mathbb{E}\Big[L(\boldsymbol{\theta}_K) - \widehat{L}(\boldsymbol{\theta}_K)\Big] \leq 2G \cdot \mathbb{E}\Big[\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{\theta}_K - \boldsymbol{\theta}_K^{\neg i}\|\Big]$.

To bound the average model-stability on the r.h.s of the lemma's inequality, we use GD expansiveness. Specifically applying (Taheri & Thrampoulidis, 2023, Lemma B.1.) to our setting, fives $\forall \boldsymbol{\theta}, \boldsymbol{\theta}'$:

$$\|(\boldsymbol{\theta} - \eta \nabla \widehat{L}(\boldsymbol{\theta})) - (\boldsymbol{\theta}' - \eta \nabla \widehat{L}(\boldsymbol{\theta}'))\| \leq \max_{\alpha \in [0,1]} \left\{\left(1 + \frac{\eta \beta_3(\boldsymbol{\theta}_\alpha)}{\sqrt{H}} \widehat{L}(\boldsymbol{\theta}_\alpha)\right) \vee \eta \beta_2(\boldsymbol{\theta}_\alpha)\right\} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,, \tag{19}$$

where, $\boldsymbol{\theta}_\alpha = \alpha \boldsymbol{\theta} + (1 - \alpha)\boldsymbol{\theta}'$, $\alpha \in [0, 1]$. Using this and gradient self-boundedness from Corollary 1, we get:

$$\left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_{k+1}^{\neg i}\right\| \leq \max_{\alpha \in [0,1]} \left\{\left(1 + \frac{\eta \beta_3(\boldsymbol{\theta}_{k_\alpha}^{\neg i})}{\sqrt{H}} \widehat{L}^{\neg i}(\boldsymbol{\theta}_{k_\alpha}^{\neg i})\right) \vee \eta \beta_2(\boldsymbol{\theta}_{k_\alpha}^{\neg i})\right\} \cdot \left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^{\neg i}\right\| + \frac{\eta \beta_1(\boldsymbol{\theta}_k)}{n} \ell_i(\boldsymbol{\theta}_k) \,, \tag{20}$$

where $\boldsymbol{\theta}_{k_\alpha}^{\neg i} := \alpha\boldsymbol{\theta}_k + (1-\alpha)\boldsymbol{\theta}_k^{\neg i}$ for $\alpha \in [0,1]$. Further using the bounded iterates-norm property from the training analysis, we can control $\beta_2(\boldsymbol{\theta}_{k_\alpha}^{\neg i}) \le \tilde{\beta}_2(\boldsymbol{\theta})$ and $\beta_3(\boldsymbol{\theta}_{k_\alpha}^{\neg i}) \le \tilde{\beta}_3(\boldsymbol{\theta})$ making them independent of k (See Lemma 11 for the definitions of $\tilde{\beta}_2(\cdot), \tilde{\beta}_3(\cdot)$). In order to invoke the Descent Lemma, we set the step-size same as in the training analysis. Thus, (20) becomes:

$$\left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_{k+1}^{\neg i}\right\| \le \left((1 + \frac{\eta\tilde{\beta}_3(\boldsymbol{\theta})}{\sqrt{H}}) \max_{\alpha\in[0,1]} \widehat{L}^{\neg i}(\boldsymbol{\theta}_{k_\alpha}^{\neg i})\right)\left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^{\neg i}\right\| + \frac{\eta\beta_1(\boldsymbol{\theta}_k)}{n}\ell_i(\boldsymbol{\theta}_k)\,. \tag{21}$$

As in the training analysis, we can control the loo empirical loss $\widehat{L}^{\neg i}$ for any point on the line $[\boldsymbol{\theta}_k, \boldsymbol{\theta}_k^{\neg i}]$ of two sufficiently close points satisfying $\sqrt{H} \ge 2\left(\beta_3(\boldsymbol{\theta}_k) \vee \beta_3(\boldsymbol{\theta}_k^{\neg i})\right)\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^{\neg i}\|^2$. Using Prop. 4, Eq. (21) becomes

$$\left\|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_{k+1}^{\neg i}\right\| \le (1+\alpha_{k,i})\left\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_k^{\neg i}\right\| + \frac{\eta\tilde{\beta}_1(\boldsymbol{\theta})}{n}\ell_i(\boldsymbol{\theta}_k)\,, \tag{22}$$

where $\alpha_{k,i} := \frac{4\eta\tilde{\beta}_3(\boldsymbol{\theta})}{3\sqrt{H}}\left(\widehat{L}^{\neg i}(\boldsymbol{\theta}_k) + \widehat{L}^{\neg i}(\boldsymbol{\theta}_k^{\neg i})\right)$ and $\beta_1(\boldsymbol{\theta}_k) \le \tilde{\beta}_1(\boldsymbol{\theta})$ similar to $\beta_2(\cdot), \beta_3(\cdot)$ using bounded iterates-norm. Unrolling the iterates in (22), summing over $i \in [n]$ and using training regret bounds, we have the following average model stability bound for any iterate K: $\frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{\theta}_K - \boldsymbol{\theta}_K^{\neg i}\right\| \le \frac{2\eta\tilde{\beta}_1(\boldsymbol{\theta})}{n}\left(2K\widehat{L}(\boldsymbol{\theta}) + \frac{9\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2}{4\eta}\right)$, Combining this with an application of Lemma 3 for our objective, which is $G \le \tilde{\beta}_1(\boldsymbol{\theta})$-Lipschitz from Corollary 1, and using $\eta \le \frac{1}{\rho(\boldsymbol{\theta})} \le \frac{1}{(\tilde{\beta}_1(\boldsymbol{\theta}))^2}$, we get the desired generalization gap stated in Thm. 2.

7 Concluding remarks

We studied convergence and generalization of GD for training a multi-head attention layer in a classification task. Our training and generalization bounds hold under an appropriate realizability condition asking for the existence of an a target model $\widetilde{\boldsymbol{\theta}}$ achieving good train loss while being sufficiently close to initialization. In particular, from the condition on the number of heads H in (3), we need $\widetilde{\boldsymbol{\theta}}$ is at most $\tilde{\mathcal{O}}\left(d^{-1/3}T^{-1/6}R^{-5/3}H^{1/6}\right)$ far from initialization (provided $\|\widetilde{\boldsymbol{\theta}}\|_{2,\infty} = \mathcal{O}(1)$). In Sec. 4.3 we showed that such a model exists if the initialization is chosen appropriately. Specifically it suffices that $\|\widetilde{\boldsymbol{\theta}}^{(0)}\|_{2,\infty} = \mathcal{O}(1)$, the model output at initialization is $\tilde{\mathcal{O}}(1)$-bounded and that the data are linearly separable with margin γ with respect to the NTK features of the model at initialization. Then, $\mathcal{O}\left(d^2TR^{10}\,\mathrm{polylog}(n)/\gamma^6\right)$ number of heads guarantee that $\Theta(n)$ GD steps result in train and test loss bounds $\tilde{\mathcal{O}}\left(1/(\eta\gamma^2 n)\right)$. In Sec. 5 we applied our results to a tokenized-mixture model. We showed that after one randomized gradient step from $\mathbf{0}$, the model satisfies the above conditions for `good` intialization. For this initialization, we computed the NTK margin $\gamma_\star$ which in turn governs the guaranteed rate of convergence and generalization based on our general bounds. This opens several interesting questions for future work.

First, does random initialization of attention weights satisfy NTK separability, and if so, what is the corresponding margin? Second, are there other initialization strategies that guarantee the realizability conditions are satisfied? Here, note that our conditions for `good` initialization are only shown to be sufficient for realizability leaving room for improvements. Third, how suboptimal is the best NTK margin (among other potential natural initializations) compared to the model's global margin $\arg\max_{\|\widetilde{\boldsymbol{\theta}}\|=\sqrt{2}} \min_{i\in[n]} y_i\widetilde{\Phi}(\boldsymbol{X}_i;\widetilde{\boldsymbol{\theta}})$? In Proposition 3 we showed for the data model DM1 that there exists single-head attention model $\boldsymbol{\theta}_{\mathrm{opt}} = (\boldsymbol{U}_{\mathrm{opt}}, \boldsymbol{W}_{\mathrm{opt}})$ with $\|\boldsymbol{\theta}_{\mathrm{opt}}\| = \sqrt{2}$ such that $y\Phi(\boldsymbol{X};\boldsymbol{U}_{\mathrm{opt}},\Gamma_\epsilon\cdot\boldsymbol{W}_{\mathrm{opt}}) = \frac{S\sqrt{T}}{\sqrt{2}}\left((1-\epsilon) - 2\epsilon Z_\mu/S\right)$ for all $\Gamma_\epsilon \gtrsim \frac{\log((\zeta^{-1}-1)/\epsilon)}{S}$ and any $\epsilon \in (0,1)$ (see App. E). In particular, as $\epsilon \to 0$ and $\Gamma_\epsilon \to \infty$ (for which the softmax map gets saturated and attends to tokens with highest relevance score) the achieved margin approaches $\gamma_{\mathrm{attn}} := S\sqrt{T}/\sqrt{2}$, which is independent of the sparsity level ζ. In particular, $\gamma_{\mathrm{attn}} \ge \gamma_\star \ge \gamma_{\mathrm{lin}}$ and the gap increases with decreasing sparsity. Is it possible to establish finite-time convergence bounds to models with margin $\approx \gamma_{\mathrm{attn}}$ under appropriate initialization? How is the answer affected by the fact that the optimal attention weights in this case are diverging in norm ($\Gamma_\epsilon \to \infty$)? Using our approach, we argued in Sec. 5.3 that the key challenge is the saturation of norm of $\boldsymbol{W}_{\mathrm{opt}}$ (Γ_ϵ), which does not allow the appropriate realizability condition to hold (at least for $\mathbf{0}$ initialization). Finally, it is interesting to consider other data models for which multiple heads are necessary to interpolate the data.

References

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

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332. PMLR, 2019.

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

Arindam Banerjee, Pedro Cisneros-Velarde, Libin Zhu, and Mikhail Belkin. Restricted strong convexity of deep learning models with smooth activations. *arXiv preprint arXiv:2209.15106*, 2022.

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

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

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.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep relu networks? In *International Conference on Learning Representations*, 2020.

Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 551–561. Association for Computational Linguistics, November 2016. doi: 10.18653/v1/D16-1053. URL https://aclanthology.org/D16-1053.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

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.

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.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pp. 1675–1685. PMLR, 2019.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. *arXiv preprint arXiv:2110.10090*, 2021.

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

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

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

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

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

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. In *International Conference on Learning Representations*, 2020.

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

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pp. 5809–5819. PMLR, 2020.

Yunwen Lei, Rong Jin, and Yiming Ying. Stability and generalization analysis of gradient methods for shallow neural networks. In *Advances in Neural Information Processing Systems*, 2022.

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*, 2023a.

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

Valerii Likhosherstov, Krzysztof Choromanski, and Adrian Weller. On the expressive power of self-attention matrices, 2021.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJC_jUqxe.

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

Quynh Nguyen, Marco Mondelli, and Guido F Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8119–8129. PMLR, 18–24 Jul 2021.

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.

Konstantinos E Nikolakakis, Farzin Haddadpour, Amin Karbasi, and Dionysios S Kalogerias. Beyond Lipschitz: Sharp generalization and excess risk bounds for full-batch gd. *arXiv preprint arXiv:2204.12446*, 2022.

Atsushi Nitanda, Geoffrey Chinot, and Taiji Suzuki. Gradient descent can learn less over-parameterized two-layer neural networks on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

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

OpenAI. Gpt-4 technical report, 2023.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1 (1):84–105, 2020.

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.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1244. URL https://aclanthology.org/D16-1244.

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.

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.

Dominic Richards and Ilja Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. *Advances in Neural Information Processing Systems*, 34: 8609–8621, 2021.

Dominic Richards and Mike Rabbat. Learning with gradient descent and weakly convex losses. In *International Conference on Artificial Intelligence and Statistics*, pp. 1990–1998. PMLR, 2021.

Itay M Safran, Gilad Yehudai, and Ohad Shamir. The effects of mild over-parameterization on the optimization landscape of shallow relu neural networks. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 3889–3934. PMLR, 15–19 Aug 2021.

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.

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

Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3380–3394. PMLR, 02–05 Jul 2022.

Ohad Shamir. Gradient methods never overfit on separable data. *Journal of Machine Learning Research*, 22 (85):1–20, 2021.

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.

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

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines, 2023a.

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

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

Matus Telgarsky. Feature selection and low test error in shallow low-rotation relu networks. In *The Eleventh International Conference on Learning Representations*, 2022.

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

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Johannes von Oswald, Eyvind Niklasson, E. Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. *ArXiv*, abs/2212.07677, 2022.

Weihang Xu and Simon Du. Over-parameterization exponentially slows down gradient descent for learning a single neuron. In Gergely Neu and Lorenzo Rosasco (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1155–1198. PMLR, 2023.

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

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.

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

Mo Zhou, Rong Ge, and Chi Jin. A local convergence theory for mildly over-parameterized two-layer neural network. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4577–4632. PMLR, 15–19 Aug 2021.

Zhenyu Zhu, Fanghui Liu, Grigorios Chrysos, Francesco Locatello, and Volkan Cevher. Benign overfitting in deep neural networks under lazy training. In *International Conference on Machine Learning*, pp. 43105–43128. PMLR, 2023.