Finite-Time Analysis of Gradient Descent for Shallow Transformers

Enes Arda
The Ohio State University
arda.2@osu.edu

Semih Cayci
RWTH Aachen University
cayci@mathc.rwth-aachen.de

Atilla Eryilmaz
The Ohio State University
eryilmaz.2@osu.edu

Abstract

Understanding why Transformers perform so well remains challenging due to their nonconvex optimization landscape. In this work, we analyze a shallow Transformer with m independent heads trained by projected gradient descent in the kernel regime. Our analysis reveals two main findings: (i) the width required for nonasymptotic guarantees scales only *logarithmically* with the sample size n, and (ii) the optimization error is *independent* of the sequence length T. This contrasts sharply with recurrent architectures, where the optimization error can grow exponentially with T. The trade-off is memory: to keep the full context, the Transformer's memory requirement grows with the sequence length. We validate our theoretical results numerically in a teacher–student setting and compare Transformers with recurrent architectures on an autoregressive task.

1 INTRODUCTION

Transformers have reshaped modern machine learning, achieving state-of-the-art results in language, vision, and multimodal tasks (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020; Dosovitskiy et al., 2021). Despite this empirical success, a sharp nonasymptotic understanding of their training dynamics remains elusive. Two obstacles stand out: (i) self-attention is highly nonlinear—each output coordinate depends on all inputs through a softmax reweighting—so standard arguments for fully-connected networks or CNNs do not apply directly; and (ii) the training landscape is nonconvex, complicating guarantees for gradient-based methods beyond infinite-width limits.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

A growing theory literature has therefore analyzed simplified settings. To make comparisons precise, let us recall the single-head self-attention layer. For a length-T input $X \in \mathbb{R}^{d\times T}$ and weights $W_Q, W_K, W_V \in \mathbb{R}^{m\times d}$, the attention layer outputs

$$a(X) \;=\; W_V X\, \sigma_s\big(\beta X^\top W_K^\top W_Q X\big), \qquad (1)$$

where σ_s denotes the column-wise softmax, and β is the softmax scaling. In the original Transformer, one takes $\beta = m^{-1/2}$ to stabilize the dot-product variance (Vaswani et al., 2017). However, to make the analysis of (1) tractable, prior work often studies variants that take one of the following directions: (i) avoid σ_s entirely to obtain linear attention (Katharopoulos et al., 2020; Choromanski et al., 2021; Lu et al., 2021; Zheng et al., 2022; Zhang et al., 2024); (ii) replace σ_s with a pointwise activation such as ReLU (Wortsman et al., 2023; Hron et al., 2020); or (iii) take infinite-width limits under $\beta = m^{-1}$, in which the softmax weights approach uniform distribution and the attention layer reduces to mean-pooling (Wu et al., 2023; Noci et al., 2023; Bordelon et al., 2024). While analytically convenient, these modifications limit the scope of the resulting guarantees.

More recently, Wu et al. (2023) established global convergence for shallow Transformers at finite width under the realistic $\beta = m^{-1/2}$ scaling, showing that *quadratic* overparameterization in the sample size n suffices for gradient descent to converge at a *linear* rate. Their proof follows the standard recipe of enforcing strict positive definiteness of a Gram matrix to drive convergence (Du et al., 2019a,b; Allen-Zhu et al., 2019; Zou and Gu, 2019; Nguyen, 2021; Nguyen et al., 2021). In particular, Wu et al. (2023) control the minimum singular value of a feature matrix built from the last hidden layer, which relies on the regime $d \geq n$, thereby limiting its practicality.

Parallel to these developments for Transformers, recent work has sharpened finite-time analyses for recurrent architectures (Cayci and Eryilmaz, 2024a,b). In particular, Cayci and Eryilmaz (2024a) prove nonasymptotic convergence of gradient descent for in-

dependently recurrent neural networks (IndRNNs, Li et al., 2018) with width scaling only *logarithmically* in the sample size n, using transportation mappings to represent targets in the associated kernel function space. Their bounds also quantify the cost of *long-term dependencies*: beyond an activation-controlled cutoff, both the required width and iteration complexity can grow exponentially with the sequence length T.

In this paper, building on techniques from Cayci and Eryilmaz (2024a), we analyze a shallow multi-head Transformer trained by gradient descent in the near-initialization regime while preserving the nonlinearity of the attention layer. The main contributions of this work can be highlighted as follows:

- **Preserving attention nonlinearity with multiple heads.** Unlike the approaches discussed above that remove or linearize softmax, or that enforce degeneracy via the scaling $\beta = m^{-1}$, our analysis retains the *genuine attention nonlinearity* and allows independent heads.

- **Bypassing positive definiteness of the NTK/Gram.** Unlike the original works in the NTK regime, our analysis *does not* require strict positive definiteness of the NTK/Gram matrix.

- **Logarithmic overparameterization.** We show that, in our setting, the width required for nonasymptotic training guarantees scales only *logarithmically* with the sample size n. To the best of our knowledge, this has not been shown for Transformers.

- **Optimization error independent of input sequence length T.** Our convergence bound does *not* degrade with the sequence length T. This contrasts sharply with RNN analyses, where the dependence on T can be exponential under long-term dependencies. The trade-off is memory: retaining full context implies that the Transformer's memory footprint grows with T due to attention over all tokens.

- **Experimental validation.** With controlled experiments in teacher–student settings, we validate the predicted scaling laws for our Transformer model and contrast Transformers with IndRNNs in their handling of long-term dependencies and memory complexity.

Technical challenges and ideas. (i) We preserve the genuine softmax attention nonlinearity while avoiding any uniformly positive definite Gram/NTK assumption by working in a near-initialization regime, where the linearization error can be controlled under an explicit width requirement (Lemma 3). (ii) Within this NTK viewpoint, transportation mappings (Section 5.4) provide a convenient representation of targets in the Transformer-NTK RKHS and yield an approximation bound with only logarithmic dependence on the sample size (Lemma 5). (iii) To obtain bounds independent of the sequence length T, we exploit the covariance structure induced by the softmax Jacobian to bound attention gradient/Hessian norms uniformly in T (Remark 3); this feeds into T-independent local Lipschitz/smoothness constants (Lemma 2) and ultimately yields the T-independent optimization error (Theorem 1).

All proofs of technical results are deferred to the appendix, which also provides additional details.

2 RELATED WORK

Neural Tangent Kernel (NTK) analysis. The NTK framework linearizes networks around initialization and connects gradient descent to kernel gradient flow, yielding optimization and generalization guarantees in overparameterized regimes (Jacot et al., 2018; Lee et al., 2019; Du et al., 2019b; Chizat et al., 2019). For CNNs, NTK limits capture locality and weight sharing (Garriga-Alonso et al., 2019; Novak et al., 2019), and the convolutional NTK admits exact computation (Arora et al., 2019). For RNNs, the Recurrent NTK defines sequence kernels that aggregate contributions across time (Alemohammad et al., 2021).

Asymptotic theory for Transformers. From a representation-theoretic perspective, Transformers are universal approximators of permutation-equivariant sequence-to-sequence maps, and, when equipped with trainable positional encodings, of general continuous sequence-to-sequence functions on compact domains (Yun et al., 2020). At infinite width, attention admits NTK descriptions under suitable scalings: Hron et al. (2020) derive attention NTK limits, showing that under the common $\beta = m^{-1/2}$, letting the number of heads grow results in a deterministic NTK. Under $\beta = m^{-1}$, deterministic kernels can exist but may degenerate toward mean-pooling or require $W_Q = W_K$ (Yang, 2020). Wu et al. (2023) complement their finite-width convergence proof with an NTK analysis for shallow Transformers with the degenerate scaling $\beta = m^{-1}$.

Nonasymptotic theory for Transformers. Wu et al. (2023) prove global convergence of a shallow Transformer while preserving the softmax attention nonlinearity and carefully diagnosing scaling/initialization choices. Deora et al. (2024) derive finite-time training and generalization bounds for gradient descent on a single-layer multi-head self-attention model in binary classification, where they obtain polylogarithmic head requirements under realizability and initialization conditions. Complemen-

tary results capture in-context learning dynamics with true softmax: Huang et al. (2024) give finite-time guarantees and a phase-wise view of attention; Song et al. (2024) show convergence under suitable initialization and reveal that softmax can introduce spurious local minima whereas a Gaussian attention smooths the landscape. Simplified settings using linear attention establish convergence for linear predictors learned in-context (Zhang et al., 2024). Beyond linear tasks, Shen et al. (2024) obtain linear-rate convergence to the Bayes classifier for in-context classification with a shallow multi-head Transformer. Further optimization views sharpen what attention learns: Tarzanagh et al. (2023) show that gradient descent on a single attention layer aligns with a max-margin (SVM-like) token-selection rule. Vasudeva et al. (2025) study implicit bias and convergence rates for single-layer self-attention in binary classification. Nichani et al. (2024) analyze a two-layer attention-only Transformer and prove that gradient descent recovers the latent causal structure of the in-context learning problem.

Comparison to Wu et al. (2023). Our work is closest in spirit to Wu et al. (2023), which also studies shallow Transformers with genuine softmax attention under a regression setting. The main differences are: **(i) Regime:** Wu et al. (2023) prove global convergence via a well-conditioned last-layer feature/Gram matrix, whereas we work in a projected near-initialization NTK regime and do not assume a positive definite Gram matrix. **(ii) Rates/width:** their result achieves a linear rate with polynomial overparameterization in n (Wu et al., 2023, Proposition 1), while our bounds have the canonical $1/\sqrt{\tau}$ optimization rate and require only logarithmic dependence on n in the width (Theorem 1). **(iii) Sequence length:** our optimization constants are explicitly independent of T, while their conditions involve T-dependent quantities. **(iv) NTK features:** they complement their analysis with an NTK study under $1/m$-type scaling, which yields a vanishing attention component (Wu et al., 2023, Lemma 1); whereas our NTK decomposition retains a nontrivial attention block (Lemma 4).

Notation. For $k \in \mathbb{N}$ we write $[k] = \{1, 2, \ldots, k\}$. For matrices A, B we write $\langle A, B\rangle_F = \mathrm{tr}(A^\top B)$ and $\|A\|_F$ for the Frobenius norm. The probability simplex in $\mathbb{R}^T$ is $\Delta_T = \{\, \alpha \in \mathbb{R}^T : \alpha_t \geq 0,\ \sum_{t=1}^T \alpha_t = 1 \,\}$ and $\mathbb{R}_+ = (0, \infty)$. For a vector x, $\mathrm{diag}(x)$ denotes the diagonal matrix with diagonal x. For matrices $A_1, \ldots, A_r$, $\mathrm{blkdiag}(A_1, \ldots, A_r)$ denotes the block-diagonal matrix with diagonal blocks $A_1, \ldots, A_r$ and zero elsewhere. For $A \in \mathbb{R}^{m\times n}$, $\mathrm{vec}(A) \in \mathbb{R}^{mn}$ denotes the column-wise vectorization of A. The notation $\lesssim_{a,b}$ omits the constants depending on the parameters a, b. For vectors x, y, $x \leq y$ denotes element-wise comparison. $\mathrm{Unif}(V)$ denotes the uniform distribution over the set V, and $\mathrm{Rad}(\alpha)$ denotes $\mathrm{Unif}(\{-\alpha, \alpha\})$.

3 PROBLEM SETUP

We start by specifying the objective, introduce the shallow multi-head Transformer studied in this work, and formalize the optimization algorithms analyzed throughout the paper.

3.1 Data

Each sample $X = (X_1, \ldots, X_T) \in \mathbb{R}^{d\times T}$ is a sequence of d-dimensional tokens coming from a compact set

$$\mathcal{X} \;\coloneqq\; \big\{\, X \in \mathbb{R}^{d\times T} \;:\; \max_{t\in[T]} \|X_t\|_2 \leq 1 \,\big\}.$$

The label is a scalar $y \in \mathbb{R}$. We observe samples from a dataset

$$\mathcal{D} = \big\{(X^{(j)}, y^{(j)})\big\}_{j=1}^n \;\subset\; \mathcal{X} \times \mathbb{R}$$

that were generated by a function $f^\star$. In other words, $f^\star(X^{(j)}) = y^{(j)}$ for every $j \in [n]$. Our goal will be to approximate the function $f^\star$ with the Transformer model defined in the next section.

3.2 Transformer Architecture

Fix $X \in \mathcal{X}$. Since each training sample has a scalar target, we employ pooled self-attention driven by a single query $q_X \in \mathbb{R}^d$ with $\|q_X\|_2 \leq 1$. The query q_X is not trainable: it may be a fixed vector (`[CLS]` as in BERT (Devlin et al., 2019), or ViT's class token (Dosovitskiy et al., 2021)), a deterministic function of the input (e.g., last-token pooling $q_X = X_T$ for next-token prediction (Radford et al., 2019)), or a pre-specified seed query used to summarize high-dimensional inputs (Jaegle et al., 2021).

Let $m \in \mathbb{N}$ be the width of our Transformer. For $i \in [m]$ and trainable $W_i \in \mathbb{R}^{d\times d}$, $U_i \in \mathbb{R}^d$, $c_i \in \mathbb{R}$, we consider the following model.

- **Attention layer.**

$$a(X; W_i) \coloneqq X\, \sigma_s\big(X^\top W_i q_X\big) \in \mathbb{R}^d, \tag{2}$$

where $\sigma_s : \mathbb{R}^T \to \Delta_T$ is the softmax function: $(\sigma_s(z))_t = \frac{\exp(z_t)}{\sum_{s=1}^T \exp(z_s)}$.

- **Feed-forward layer.**

$$h(X; \theta_i) \coloneqq \sigma\big(U_i^\top a(X; W_i)\big) \in \mathbb{R}, \tag{3}$$

where $\theta_i \coloneqq (U_i, \mathrm{vec}(W_i))$ and the activation $\sigma : \mathbb{R} \to \mathbb{R}$ satisfies uniform bounds

$$|\sigma(u)| \leq \sigma_0,\ |\sigma'(u)| \leq \sigma_1,\ |\sigma''(u)| \leq \sigma_2 \quad \forall u \in \mathbb{R}.$$

We collect these bounds as $\bar{\sigma} = (\sigma_0, \sigma_1, \sigma_2) \in \mathbb{R}^3_+$.

- **Linear layer.**

$$f(X;\varphi) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} c_i\, h(X;\theta_i) \in \mathbb{R}, \tag{4}$$

where $\varphi_i := (c_i, U_i, \mathrm{vec}(W_i))$ and $\varphi := (\varphi_1, \ldots, \varphi_m)$.

Remark 1. Following prior analyses (Wu et al., 2023; Deora et al., 2024; Nichani et al., 2024; Oymak et al., 2023), we simplify the standard (W_Q, W_K, W_V) parameterization in (1). For pooled attention with a single query, the usual logits can be written as $\beta\,(W_K X)^\top (W_Q q_X) \in \mathbb{R}^T$ and hence $\beta\,(W_K X)^\top (W_Q q_X) = X^\top (\beta W_K^\top W_Q) q_X$. Thus, collapsing W_Q and W_K into a single matrix in (2) amounts to setting $W_i := \beta W_{K,i}^\top W_{Q,i}$. By initializing W_i with bounded-variance entries, the scaling $\beta = 1/\sqrt{m}$ is absorbed into the initialization and does not appear explicitly in (2). In contrast, $1/m$ scaling would drive logits toward zero and make the softmax weights nearly uniform, degenerating attention toward mean pooling (cf. Section 1). Finally, in this one-hidden-layer setting we absorb the value/output matrices into the subsequent linear map U_i.

Remark 2. In standard multi-head attention one concatenates h attention heads into $A(X) = (a(X;W_1), \ldots, a(X;W_h)) \in \mathbb{R}^{hd}$ and applies a feed-forward layer with a weight matrix $V \in \mathbb{R}^{m \times hd}$, resulting in $\sigma(VA(X))$. Our architecture is equivalent to this with a block-diagonal weight matrix with one neuron per head, i.e., $h = m$ and $V = \mathrm{blkdiag}(U_1^\top, \ldots, U_m^\top)$. This independent head constraint acts as a structural regularizer: it preserves per-head interpretability, reduces cross-head interference, and restores a deterministic NTK, thereby simplifying the analysis without requiring any of the architectural changes mentioned in Section 1.

3.3 Algorithms for Empirical Risk Minimization

For a given dataset $\mathcal{D}$, parameters φ and the predictor $f(\cdot;\varphi)$, we analyze the empirical mean-squared error (MSE) loss, which is expressed as

$$\widehat{\mathcal{L}}_n(\varphi) := \frac{1}{n} \sum_{j=1}^{n} \big(f(X^{(j)};\varphi) - y^{(j)}\big)^2.$$

We optimize $\widehat{\mathcal{L}}_n$ by projected methods onto the feasible set Ω. Let $\ell_j(\varphi) := \big(f(X^{(j)};\varphi) - y^{(j)}\big)^2$ and, for any index set $B \subseteq [n]$, $\widehat{\mathcal{L}}_B(\varphi) := \frac{1}{|B|}\sum_{j\in B} \ell_j(\varphi)$. Given a step size η and mini-batches $\{B_s\}_{s\geq 0}$, the iterate updates are, for $s = 0, 1, 2, \ldots$:

$$\varphi^{(s+1)} := \Pi_\Omega \left(\varphi^{(s)} - \eta\, \nabla_\varphi \widehat{\mathcal{L}}_{B_s}(\varphi^{(s)})\right),$$

where Π_Ω denotes projection onto the set Ω. We analyze (i) ProjGD with $B_s = [n]$ for all s, and (ii) ProjSGD with $B_s = \{J_s\}$ with $J_s \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([n])$.

Why projection? Our analysis is local around the random initialization: all bounds are proved on the neighborhood Ω_ρ (Definition 2), where (i) the NTK linearization error is uniformly controlled (Lemma 3) and (ii) the change-of-feature terms in the optimization argument can be bounded using local Lipschitz/smoothness constants (Lemma 2). Projecting the iterates onto Ω_ρ is therefore a simple analytical device that enforces this near-initialization regime. In practice, other mechanisms that keep parameters close to initialization (e.g., weight decay, explicit norm regularization, or early stopping) can play a similar role, and in Section 6 we also report experiments trained without projection.

4 OVERVIEW OF MAIN RESULTS

Below we present our main results informally. Formal statements are given in Section 5.5, with detailed proofs deferred to Appendix A.

(Informal) Theorem 1. *Let $f^\star \in \mathcal{F}_{\bar{\nu}}$ be the target function (formalized in Section 5.4). After τ steps of* ProjGD *with an appropriate step size and projection radius, the average iterate $\bar{\varphi}^{(\tau)} := \frac{1}{\tau}\sum_{s=0}^{\tau-1} \varphi^{(s)}$ satisfies, with probability at least $1-\delta-\delta'$ over the random initialization (introduced in Section 5.2),*

$$\widehat{\mathcal{L}}_n\left(\bar{\varphi}^{(\tau)}\right) \lesssim \underbrace{\frac{D^2}{\sqrt{\tau}}}_{\substack{\textit{optimization}\\ \textit{error}}} + \underbrace{\sqrt{\frac{D\log(n/\delta)}{m}}}_{\substack{\textit{approximation}\\ \textit{error}}} + \underbrace{\sqrt{\frac{D^3}{m}}}_{\substack{\textit{linearization}\\ \textit{error}}}$$

where $D := d + \log(m/\delta')$.

For ProjSGD, the same bound holds for $\mathbb{E}[\widehat{\mathcal{L}}_n(\bar{\varphi}^{(\tau)}) \,|\, \varphi^{(0)}]$ with the same rates (cf. Proposition 1).

We highlight the following features:

- **Non-asymptotic in time and width.** The training loss admits a finite-time optimization–approximation–linearization decomposition, with the canonical $1/\sqrt{\tau}$ optimization rate and $1/\sqrt{m}$ width dependence. This decomposition holds for any $\tau, m \geq 1$, with no hidden asymptotic regime.

- **Logarithmic in sample size.** The sample size enters only through the approximation error (Lemma 5). To make this term at most ε it suffices to take $m \gtrsim \varepsilon^{-2}\log(n/\delta)$; in particular, m needs to grow only *logarithmically* with n.

• **No explicit dependence on sequence length.** Thanks to the attention gradients being uniformly bounded in T via a softmax-covariance structure (Remark 3), we obtain an optimization error independent of the sequence length T.

5 TRANSFORMER NTK ANALYSIS

In this section, we analyze the gradients of the Transformer architecture from Section 3.2, specify the initialization, derive the associated NTK, and introduce transportation mappings. These components define the function class used to establish our main results in Section 5.5.

5.1 Transformer Gradient Analysis

Lemma 1 (Gradients). *Fix $i \in [m]$ and write $a_i := a(X; W_i)$. Then*

$$\frac{\partial f(X;\varphi)}{\partial c_i} = m^{-1/2}\, h(X;\theta_i), \tag{5}$$

$$\nabla_{U_i} f(X;\varphi) = m^{-1/2}\, c_i\, \sigma'\big(U_i^\top a_i\big)\, a_i, \tag{6}$$

$$\nabla_{W_i} f(X;\varphi) = m^{-1/2}\, c_i\, \sigma'\big(U_i^\top a_i\big)\, \big(M_i U_i\big)\, q_X^\top, \tag{7}$$

where

$$M_i \;:=\; M(X;W_i) \;=\; X\, J_s\big(X^\top W_i q_X\big)\, X^\top, \tag{8}$$

and $J_s(z) = \operatorname{diag}\big(\sigma_s(z)\big) - \sigma_s(z)\sigma_s(z)^\top$ is the Jacobian matrix of softmax.

Remark 3. With $\alpha = \sigma_s(X^\top W_i q_X) \in \Delta_T$ and $\mu = \sum_t \alpha_t X_t$, notice that $M(X;W_i) = \sum_{t=1}^T \alpha_t X_t X_t^\top - \mu\mu^\top$ is a softmax-weighted covariance matrix. This is the only term in Lemma 1 that involves a T-dependence. Since $\operatorname{tr} M(X;W_i) \le 1$ for all $X \in \mathcal{X}$, we obtain $\|M(X;W_i)\|_F \le 1$ *uniformly in T*. As a consequence of this key property, we observe that $\|\nabla_{W_i} f(X;\varphi)\|_F$ is independent of T. In contrast, for RNNs the analogous gradient contains an explicit sum over T recurrent steps, leading to constants that can grow with T (Cayci and Eryilmaz, 2024a, Prop. 3.1). This structural difference underlies our T-independent optimization bounds in Section 5.5. See Appendix A.1 for the detailed derivation.

Remark 4. Similarly, we have $a(X;W_i) = \sum_{t=1}^T \alpha_t X_t$, which is a convex combination of the tokens $\{X_t\}_{t=1}^T$. Hence, by the triangle inequality, $\|a(X;W_i)\|_2 \le \sum_{t=1}^T \alpha_t \|X_t\|_2 \le \sum_{t=1}^T \alpha_t = 1$. This allows us to bound $\|\nabla_{U_i} f(X;\varphi)\|_2$, and together with the previous remark, these bounds yield the continuity and smoothness results stated in Lemma 2.

5.2 Initialization

As noted in previous work (Wu et al., 2023; Cayci and Eryilmaz, 2024a; Lee et al., 2019; Jacot et al., 2018; Chizat et al., 2019), initialization plays an important role in the performance of gradient descent in the kernel regime. Following (Bai and Lee, 2020; Cayci and Eryilmaz, 2024a; Chizat et al., 2019), we adopt a symmetric random initialization that centers the model at zero output, which will be important in the convergence analysis. Assume m is even. For $i \in [m/2]$ initialize

$$W_i^{(0)} \in \mathbb{R}^{d\times d} \text{ with i.i.d. entries } \mathcal{N}(0,1),$$
$$U_i^{(0)} \sim \mathcal{N}(0, I_d), \qquad c_i^{(0)} \sim \mathrm{Rad}(1).$$

and set

$$W_{i+m/2}^{(0)} = W_i^{(0)}, \quad U_{i+m/2}^{(0)} = U_i^{(0)}, \quad c_{i+m/2}^{(0)} = -\,c_i^{(0)},$$

with all draws mutually independent. We will denote this random initialization by $\varphi^{(0)}$.

Remark 5. Writing $h_i(X) = h(X;\theta_i^{(0)})$, note that the pairing above gives $h_{i+m/2}(X) = h_i(X)$, and therefore $f\big(X;\varphi^{(0)}\big) = 0$ for all $X \in \mathcal{X}$. This fact will be critical to achieve the error bounds in Lemmas 3 and 5.

Because $U_i^{(0)}$ is a sequence of sub-Gaussian vectors, the running maximum of its norm stays bounded with high probability, which allows us to bound the norm of the attention weight gradient in (7) with high probability. For this, we define the following event.

Definition 1. For $\delta' \in (0,1)$, define

$$\mathcal{E}_{U,m}(\delta') := \Big\{ \max_{i\in[m]} \|U_i^{(0)}\|_2 \;\le\; \sqrt{d} + \sqrt{2\log(m/2\delta')} \Big\}.$$

By Gaussian concentration for the 1-Lipschitz map $u \mapsto \|u\|_2$ (Boucheron et al., 2013, Thm. 5.6) and a union bound over $i \in [m/2]$, we obtain $\mathbb{P}\big(\mathcal{E}_{U,m}(\delta')\big) \ge 1 - \delta'$.

Throughout the paper, we will work in the following parameter set centered around the initialization $\varphi^{(0)}$.

Definition 2. Let $\rho = (\rho_c, \rho_u, \rho_w) \in \mathbb{R}_+^3$. For $i \in [m]$ define

$$\mathcal{C}_{\rho_c,i} := \{\, c \in \mathbb{R} : |c - c_i^{(0)}| \le \rho_c/\sqrt{m} \,\},$$
$$\mathcal{U}_{\rho_u,i} := \{\, U \in \mathbb{R}^d : \|U - U_i^{(0)}\|_2 \le \rho_u/\sqrt{m} \,\},$$
$$\mathcal{W}_{\rho_w,i} := \{\, W \in \mathbb{R}^{d\times d} : \|W - W_i^{(0)}\|_F \le \rho_w/\sqrt{m} \,\},$$

and the product set

$$\Omega_{\rho,i} \;:=\; \mathcal{C}_{\rho_c,i} \times \mathcal{U}_{\rho_u,i} \times \mathcal{W}_{\rho_w,i}, \quad \Omega_\rho \;:=\; \prod_{i=1}^m \Omega_{\rho,i}.$$

Notice that on $\mathcal{E}_{U,m}(\delta')$ we also have for all $i \in [m]$ and all $U_i \in \mathcal{U}_{\rho_u,i}$

$$\|U_i\|_2 \le \sqrt{d} + \sqrt{2\log(m/2\delta')} + \frac{\rho_u}{\sqrt{m}} =: B_{U,m}(\delta'). \quad (9)$$

This allows us to establish local Lipschitz continuity and smoothness, as formalized in the following lemma, which in turn enables control of the optimization error in Theorem 1.

Lemma 2. *On $\mathcal{E}_{U,m}(\delta')$, the mapping $\theta_i \mapsto h(X;\theta_i)$ is $L_{1,m}(\delta')$-Lipschitz and $L_{2,m}(\delta')$-smooth in Ω_ρ, where $L_{1,m}(\delta') := \sigma_1\sqrt{1+B_{U,m}(\delta')^2}$, and*

$$L_{2,m}(\delta') := \sigma_2(1+B_{U,m}(\delta')^2) + 8\sigma_1\sqrt{1+B_{U,m}(\delta')^2}.$$

Next, we define the linear model, which will be useful in the analysis to characterize the linearization error.

Definition 3 (Linear model)**.** Define the function

$$f_{\text{lin}}(X;\varphi) := f(X;\varphi^{(0)}) \;+\; \big\langle \nabla_\varphi f(X;\varphi^{(0)}),\, \varphi - \varphi^{(0)}\big\rangle,$$

which is the first-order approximation of f at initialization. Thanks to symmetric initialization, we get

$$f_{\text{lin}}(X;\varphi) = \big\langle \nabla_\varphi f(X;\varphi^{(0)}),\, \varphi - \varphi^{(0)}\big\rangle.$$

Near initialization, higher-order errors vanish, so the linear model closely approximates the original model. Indeed, inside the neighborhood Ω_ρ, we get the following uniform linearization error bound.

Lemma 3 (Linearization error)**.** *Assume $\varphi \in \Omega_\rho$ and define $\varepsilon_{\text{lin}} := \sup_{j\in[n]} \big|f(X^{(j)};\varphi) - f_{\text{lin}}(X^{(j)};\varphi)\big|$. Conditioned on the event $\mathcal{E}_{U,m}(\delta')$, we have*

$$\varepsilon_{\text{lin}} \le \tfrac{1}{\sqrt{m}}\Big(L_{1,m}(\delta')\rho_c\sqrt{\rho_w^2+\rho_u^2} + L_{2,m}(\delta')(\rho_w^2+\rho_u^2)\Big),$$

where the upper bound is denoted as $B_{\text{lin},m}(\delta')$.

5.3 Transformer NTK

For our convergence analysis, we work in the NTK regime, where training is well-approximated by linearization around initialization. Concretely, for a given random initialization $\varphi^{(0)}$, the Neural Tangent Kernel (NTK) is defined as

$$K(X,X') := \lim_{m\to\infty} \Big\langle \nabla_\varphi f\big(X;\varphi^{(0)}\big),\, \nabla_\varphi f\big(X';\varphi^{(0)}\big)\Big\rangle.$$

Viewing $\nabla_\varphi f(X;\varphi^{(0)})$ as a feature map for the input X, K is the kernel induced by these features in the infinite-width limit. In this regime, gradient descent on f linearizes around $\varphi^{(0)}$ and corresponds to kernel regression with kernel K.

Since $\nabla_\varphi f(X;\varphi^{(0)})$ is block-separable across (c,U,W), we investigate the features coming from these parameters separately. Using the gradient identities in Lemma 1, the Transformer NTK decomposes into the following blocks.

Lemma 4 (NTK decomposition)**.** *Let $a = a(X;W)$, $a' = a(X';W)$, $M = X\,J_s(X^\top W q_X)\,X^\top$, and $M' = X'\,J_s(X'^\top W q_{X'})\,X'^\top$. Then, with expectation taken over the random initialization $(c,U,\text{vec}(W)) \sim \varphi_1^{(0)}$,*

$$\begin{aligned}
K_c(X,X') &= \mathbb{E}\big[\,\sigma(U^\top a)\,\sigma(U^\top a')\,\big],\\
K_u(X,X') &= \mathbb{E}\big[\,\sigma'(U^\top a)\,\sigma'(U^\top a')\,\langle a,a'\rangle\,\big],\\
K_w(X,X') &= \langle q_X, q_{X'}\rangle \mathbb{E}\big[\sigma'(U^\top a)\sigma'(U^\top a')\,U^\top M M' U\big]
\end{aligned}$$

and the NTK decomposes as $K = K_c + K_u + K_w$.

Remark 6. Unlike the NTK analysis in (Wu et al., 2023, Lemma 1), here the attention-weight component K_w is *nonzero*. This is a consequence of not using the degenerate m^{-1} scaling and keeping a random attention weight in (2), which preserves attention features in the limit. Moreover, with multiple independent heads, the empirical NTK converges to the deterministic kernel K above, solving the limit kernel problem of a single-head attention mentioned in (Hron et al., 2020).

5.4 Transportation Mappings

Let $\varphi_0 := (c,U,\text{vec}(W)) \sim \varphi_1^{(0)}$. It can be seen that the NTK in Lemma 4 is a Mercer kernel constructed with the following random feature maps:

$$\begin{aligned}
\phi_c(X) &:= \sigma\big(U^\top a(X;W)\big) \in \mathbb{R},\\
\phi_u(X) &:= \sigma'\big(U^\top a(X;W)\big)\,a(X;W) \in \mathbb{R}^d,\\
\phi_w(X) &:= \sigma'\big(U^\top a(X;W)\big)\,(M(X;W)U)\,q_X^\top \in \mathbb{R}^{d\times d}.
\end{aligned}$$

In order to characterize the Reproducing Kernel Hilbert Space (RKHS) associated with the NTK, we will use the concept of transportation mappings (Ji et al., 2020).

Let $v : \mathbb{R}^{d^2+d+1} \to \mathbb{R}\times\mathbb{R}^d\times\mathbb{R}^{d\times d}$ be a measurable function with $v(\varphi_0) = (v_c(\varphi_0), v_u(\varphi_0), v_w(\varphi_0))$ such that $\mathbb{E}\Big[\,|v_c(\varphi_0)|^2 + \|v_u(\varphi_0)\|_2^2 + \|v_w(\varphi_0)\|_F^2\,\Big] < \infty$. Define the function class

$$\mathcal{F} := \Big\{\tilde f(\cdot;v) : \mathbb{E}|v_c(\varphi_0)|^2 < \infty,\ \mathbb{E}\|v_u(\varphi_0)\|_2^2 < \infty,\ \mathbb{E}\|v_w(\varphi_0)\|_F^2 < \infty\Big\},$$

where

$$\tilde f(X;v) := \mathbb{E}\big[\phi_c(X)\,v_c(\varphi_0)\big] + \mathbb{E}\big[\langle \phi_u(X), v_u(\varphi_0)\rangle\big] + \mathbb{E}\big[\langle \phi_w(X), v_w(\varphi_0)\rangle_F\big] \quad (10)$$

and equip $\mathcal{F}$ with the inner product

$$\langle \tilde{f}(\cdot; v), \tilde{f}(\cdot; v')\rangle_{\mathcal{F}} := \mathbb{E}\Big[v_c\, v_c' + \langle v_u, v_u'\rangle + \langle v_w, v_w'\rangle_F\Big].$$

Notice that for a fixed $X' \in \mathcal{X}$, one can set $v'(\varphi_0) := (\phi_c(X'), \phi_u(X'), \phi_w(X'))$ to get

$$\tilde{f}(X; v') = \mathbb{E}\big[\phi_c(X)\phi_c(X')\big] + \mathbb{E}\big[\langle \phi_u(X), \phi_u(X')\rangle\big] + \mathbb{E}\big[\langle \phi_w(X), \phi_w(X')\rangle_F\big] = K(X, X').$$

In other words, we have $K(\cdot, X') \in \mathcal{F}$ for every $X' \in \mathcal{X}$, and $\langle \tilde{f}(\cdot; v), K(\cdot, X')\rangle_{\mathcal{F}} = \tilde{f}(X'; v)$. Hence, K has the *reproducing property* on $\mathcal{F}$, which implies that the completion $\overline{\mathcal{F}}$ is the RKHS associated with the Transformer NTK K.

We will consider the norm-constrained functions from this class. In particular, for a given $\bar{\nu} = (\bar{\nu}_c, \bar{\nu}_u, \bar{\nu}_w) \in \mathbb{R}_+^3$, the function class of interest is defined as

$$\mathcal{F}_{\bar{\nu}} := \big\{\tilde{f}(\cdot; v) : \ v \in \mathcal{V}_{\bar{\nu}}\big\},$$

where

$$\mathcal{V}_{\bar{\nu}} := \Big\{v : \ \sup_{\varphi} |v_c(\varphi)| \le \bar{\nu}_c, \ \sup_{\varphi} \|v_u(\varphi)\|_2 \le \bar{\nu}_u, \ \sup_{\varphi} \|v_w(\varphi)\|_F \le \bar{\nu}_w\Big\}.$$

Remark 7. Note that for any $\tilde{f} \in \mathcal{F}_{\bar{\nu}}$, we have $\|\tilde{f}\|_{\mathcal{F}_{\bar{\nu}}} \le \|\bar{\nu}\|_2$. Thus, $\|\bar{\nu}\|_2$ quantifies the complexity of the function class $\mathcal{F}_{\bar{\nu}}$ and serves as a natural measure of the hardness of the regression problem. It will play a central role in controlling the approximation error (Lemma 5) and the convergence of ProjGD (Theorem 1).

For any $\tilde{f} \in \mathcal{F}_{\bar{\nu}}$, the transportation mapping v associated with $\tilde{f}$ "transports" the initial weights to a point at which the model f more closely approximates $\tilde{f}$. Indeed, for $i \in [m]$, define the following parameters

$$\begin{aligned} \tilde{c}_i &:= c_i^{(0)} \ + \ m^{-1/2}\, v_c(\varphi_i^{(0)}), \\ \tilde{U}_i &:= U_i^{(0)} \ + \ c_i^{(0)} m^{-1/2}\, v_u(\varphi_i^{(0)}), \\ \tilde{W}_i &:= W_i^{(0)} \ + \ c_i^{(0)} m^{-1/2}\, v_w(\varphi_i^{(0)}), \end{aligned} \tag{11}$$

and $\tilde{\varphi} := (\tilde{c}_i, \tilde{U}_i, \mathrm{vec}(\tilde{W}_i))_{i=1}^m$. Then we have the following approximation error bound.

Lemma 5 (Approximation error)**.** *Let $\tilde{f}(\cdot; v) \in \mathcal{F}_{\bar{\nu}}$ and $\tilde{\varphi}$ be as in (11). Define $\varepsilon_{\mathrm{app}} := \sup_{j\in[n]} \big|f_{\mathrm{lin}}(X^{(j)}; \tilde{\varphi}) - \tilde{f}(X^{(j)}; v)\big|$. Then for any $\delta \in (0, 1)$, we have*

$$\varepsilon_{\mathrm{app}} \le 4(\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + \sigma_1\bar{\nu}_w)\sqrt{\tfrac{\log(2n/\delta)}{m}} =: B_{\mathrm{app},m}(\delta),$$

with probability at least $1 - \delta$. We denote this event with $\mathcal{E}_{\mathrm{app}}(\delta)$.

Remark 8. In contrast to the RNN counterpart (Cayci and Eryilmaz, 2024a, Proposition 3.5), where gradient norms may grow with the sequence length, the softmax-covariance structure of attention yields an approximation bound independent of T.

5.5 Main Results

We now present the main theorem, which formally establishes the result stated in Section 4.

Theorem 1 (ProjGD Convergence)**.** *Assume $f^\star \in \mathcal{F}_{\bar{\nu}}$ and the projection radius satisfies $\rho \ge \bar{\nu}$. Run* ProjGD *onto Ω_ρ for τ steps with the step size $\eta = 1/\sqrt{\tau}$. Then, conditioned on the event $\mathcal{E}_{U,m}(\delta') \cap \mathcal{E}_{\mathrm{app}}(\delta)$, we have*

$$\min_{s<\tau} \ \widehat{\mathcal{L}}_n\big(\varphi^{(s)}\big) \ \lesssim_{\bar{\sigma},\bar{\nu},\rho} \ \frac{L_{1,m}(\delta')^4}{\sqrt{\tau}} + L_{1,m}(\delta')(B_{\mathrm{app},m}(\delta) + B_{\mathrm{lin},m}(\delta') + \varepsilon_{\mathrm{CoF}}),$$

and for the average iterate $\bar{\varphi}^{(\tau)} := \frac{1}{\tau}\sum_{s=0}^{\tau-1} \varphi^{(s)}$,

$$\widehat{\mathcal{L}}_n\big(\bar{\varphi}^{(\tau)}\big) \ \lesssim_{\bar{\sigma},\bar{\nu},\rho} \ \frac{L_{1,m}(\delta')^4}{\sqrt{\tau}} + L_{1,m}(\delta')\big(B_{\mathrm{app},m}(\delta) + B_{\mathrm{lin},m}(\delta') + \varepsilon_{\mathrm{CoF}}\big) + B_{\mathrm{lin},m}(\delta')^2,$$

where $B_{\mathrm{lin},m}(\delta')$ and $B_{\mathrm{app},m}(\delta)$ are defined as in Lemmas 3 and 5, and

$$\varepsilon_{\mathrm{CoF}} \ \lesssim_{\bar{\sigma},\bar{\nu},\rho} \ \frac{L_{1,m}(\delta') + L_{2,m}(\delta')}{\sqrt{m}}.$$

Remark 9. The realizability condition $f^\star \in \mathcal{F}_{\bar{\nu}}$ is used only in the approximation step (Lemma 5) to control how well $f^\star$ can be represented in the Transformer-NTK RKHS. If $f^\star$ is not realizable by $\mathcal{F}_{\bar{\nu}}$, the optimization and linearization arguments (Lemmas 2–3) and the Lyapunov analysis in Theorem 1 go through unchanged, and the final bound acquires the additional term

$$\inf_{\tilde{f}\in\mathcal{F}_{\bar{\nu}}} \ \max_{j\in[n]} \big|f^\star(X^{(j)}) - \tilde{f}(X^{(j)})\big|.$$

We leave a detailed study of the approximation of broader target functions by the Transformer-NTK RKHS for future work.

Remark 10. The condition $\rho \ge \bar{\nu}$ ensures that the transported parameters $\tilde{\varphi}$ constructed from $v \in \mathcal{V}_{\bar{\nu}}$ (cf. (11)) lie in the feasible set Ω_ρ. Larger $\bar{\nu}$ corresponds to a richer target class (larger RKHS norm budget cf. Remark 7) and requires a larger projection radius, which in turn increases the hidden constants in Theorem 1. See Appendix A.7 for explicit dependencies.

We can state a similar result for the convergence of ProjSGD as well.

Proposition 1 (ProjSGD Convergence). *Running* ProjSGD *under the same setup as Theorem 1 gives, on the event* $\mathcal{E}_{U,m}(\delta') \cap \mathcal{E}_{\mathrm{app}}(\delta)$,

$$\mathbb{E}\left[\min_{s<\tau} \widehat{\mathcal{L}}_n(\varphi^{(s)}) \,|\, \varphi^{(0)}\right] \lesssim_{\bar{\sigma},\bar{\nu},\rho} \frac{L_{1,m}(\delta')^4}{\sqrt{\tau}} + L_{1,m}(\delta')(B_{\mathrm{app},m}(\delta) + B_{\mathrm{lin},m}(\delta') + \varepsilon_{\mathrm{CoF}}),$$

and for the average iterate,

$$\mathbb{E}\left[\widehat{\mathcal{L}}_n(\bar{\varphi}^{(\tau)}) \,|\, \varphi^{(0)}\right] \lesssim_{\bar{\sigma},\bar{\nu},\rho} \frac{L_{1,m}(\delta')^4}{\sqrt{\tau}} + L_{1,m}(\delta')(B_{\mathrm{app},m}(\delta) + B_{\mathrm{lin},m}(\delta') + \varepsilon_{\mathrm{CoF}}) + B_{\mathrm{lin},m}(\delta')^2.$$

Comparison of results. Theorem 1 gives a high-probability finite-time bound for ProjGD, stated for the minimum iterate and the averaged iterate. Proposition 1 shows that ProjSGD enjoys the *same* error decomposition and the *same* rates in (τ, m, n), with the guarantee stated in *conditional expectation* over the sampling randomness.

6 EXPERIMENTAL RESULTS

We complement the theory with two experiments. Section 6.1 validates the scaling laws predicted by Lemma 3, Lemma 5, and Theorem 1 in a realizable teacher–student setting aligned with $f^\star \in \mathcal{F}_{\bar{\nu}}$ and near-initialization assumptions. Section 6.2 then probes behavior beyond this regime by comparing IndRNNs and our Transformer on an autoregressive AR(L) task trained without projection, highlighting long-term dependency behavior and the associated memory trade-off.

6.1 Teacher–Student Scaling-Law Validation

We assess the three components of our theory—linearization, approximation, and optimization—with controlled teacher–student experiments using the same shallow Transformer as in Section 3.2 with $q_X = X_T$. Throughout, inputs are i.i.d. $X_t \sim \mathcal{N}(0, I_d)$ with token-wise clipping to keep $\max_t \|X_t\|_2 \le 1$.

Teacher. To realize a ground-truth $f^\star \in \mathcal{F}_{\bar{\nu}}$ as in (10) we define a sup-norm-bounded transportation mapping $v \in \mathcal{V}_{\bar{\nu}}$ as $v(\varphi) = \frac{1}{R}\sum_{r=1}^{R} v^{(r)}(\varphi)$, where

$$v_c^{(r)}(\varphi) = \tfrac{\nu_c}{\sigma_0}\,\phi_c(X_R^{(r)}), \quad v_u^{(r)}(\varphi) = \tfrac{\nu_u}{\sigma_1}\,\phi_u(X_R^{(r)}),$$
$$v_w^{(r)}(\varphi) = \mathrm{clip}_F\big(\phi_w(X_R^{(r)}),\, \nu_w\big),$$

and the anchor samples $X_R^{(r)}$ are drawn from the same distribution as inputs once and kept fixed. For a large $\tilde{m} \in \mathbb{N}$, we sample an i.i.d. pool $\{\varphi_i^{(0)}\}_{i=1}^{\tilde{m}}$ and generate the labels via a Monte Carlo approximation of (10), i.e.,

$$\tilde{y}^{(j)} = \frac{1}{\tilde{m}}\sum_{i=1}^{\tilde{m}}\big[\phi_c(X^{(j)})\, v_c(\varphi_i^{(0)}) + \langle \phi_u(X^{(j)}), v_u(\varphi_i^{(0)})\rangle + \langle \phi_w(X^{(j)}), v_w(\varphi_i^{(0)})\rangle_F\big],$$

so the teacher lies in the RKHS induced by the model's NTK, matching the realizability assumption used in the proofs. All experiments are conducted with $\tilde{m} = 8192$, $d = 8$, $T = 16$, $n = 5000$, $\bar{\nu} = (3, 3, 3)$, and tanh activation.

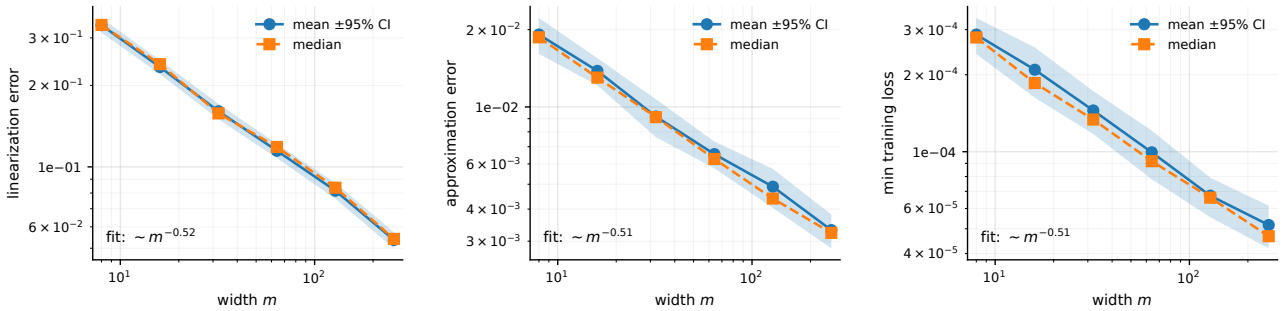
Students and metrics. Students are initialized with the symmetric initialization from Section 5.2 so that $f(\cdot; \varphi^{(0)}) \equiv 0$. For each width $m \in \{8, 16, 32, 64, 128, 256\}$, we record:

- **Linearization error.** We select parameters on the boundary of the neighborhood Ω_ρ and compute $\sup_{j\in[n]} \big|f(X^{(j)}; \varphi) - f_{\mathrm{lin}}(X^{(j)}; \varphi)\big|$, which should scale like $m^{-1/2}$ per Lemma 3.
- **Approximation error.** We set the parameters $\tilde{\varphi}$ as in (11) and measure $\sup_{j\in[n]} \big|f_{\mathrm{lin}}(X^{(j)}; \tilde{\varphi}) - \tilde{y}^{(j)}\big|$, which is expected to scale like $m^{-1/2}$ from Lemma 5.
- **Minimum training loss.** We train a student Transformer with $\tau = 4000$ steps, a constant step size $\eta = 1/\sqrt{\tau}$, and projection radii $\rho = \bar{\nu}$. Minimum training loss $\min_{t<\tau} \widehat{\mathcal{L}}_n(\varphi^{(t)})$ versus m is supposed to scale like $m^{-1/2}$ from Theorem 1.

We average curves over 10 seeds, show ±95% confidence interval (CI) around the mean, and fit a straight line to the log–log curves to estimate the scaling exponent. Figure 1 shows that all three quantities—linearization error, approximation error, and minimum training loss—decrease monotonically with width and exhibit a clear power-law dependence on m. The fitted slopes are in close agreement with the predicted $m^{-1/2}$ rate. Medians closely track the mean and the CIs remain small across widths, indicating that the observed scaling is robust to random initialization.

6.2 IndRNN vs. Transformer Architectures

To compare how IndRNNs (Cayci and Eryilmaz, 2024a) and our Transformer cope with long-term dependencies and training stability, we use a controlled autoregressive AR(L) task where the lag L sets the dependency horizon. Theory for diagonal RNNs predicts that, depending on a Lipschitz cutoff, gradients can grow exponentially due to repeated multiplication over a horizon of T steps (Cayci and Eryilmaz, 2024a), whereas our Transformer's bound has no explicit T-dependence (cf. Lemma 2).



(a) Linearization error (b) Approximation error (c) Min training loss

Figure 1: Width scaling across three metrics (log–log). Shaded bands show mean ±95% CI across seeds. Fitted slopes are consistent with the predicted $m^{-1/2}$ decay.

Each input sequence $X \in \mathbb{R}^{1\times T}$ is constructed as a window from a univariate process

$$X_t = \alpha\, X_{t-L} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

and the label is $y = X_{T+1}$. We append fixed positional channels (d_{pos}=8 sinusoidal pairs) to the input so that the Transformer has positional information. We fix $n = 5{,}000$, $\alpha = 0.9$, $\sigma_\varepsilon^2 = 0.1$, and vary L. For the Transformer we set $T = L + 1$ so that the lag falls within the context window, yielding a memory footprint $\mathcal{O}(L)$ compared to IndRNN's $\mathcal{O}(1)$. Both models use symmetric initialization and are trained with gradient descent for $\tau = 2{,}000$ steps with step size $\eta = 1/\sqrt{\tau}$ and no projection to observe gradient-norm dynamics. We set $m = 64$ and use tanh as the activation. Results are averaged over 20 seeds.

Metrics. We track training stability via the maximum Jacobian norm over training, computed with respect to the diagonal recurrent weights (IndRNN) or the attention weights (Transformer). Predictive performance is summarized by the minimum validation loss achieved over training.

Figure 2 summarizes the comparison across lags L. The top panel shows that IndRNN exhibits larger and more variable norms, consistent with the potential gradient growth predicted in Cayci and Eryilmaz (2024a), whereas the Transformer remains stable. The bottom panel shows the best validation loss achieved over training: IndRNN performs well for small lags but degrades as L increases, while the Transformer maintains roughly constant performance as L grows, at the cost of an $\mathcal{O}(L)$ context window. Overall, the experiment illustrates the performance-memory trade-off between the two architectures anticipated by the theory. Although this task lies outside our realizable/projection regime, the observed stability of attention is consistent with the softmax-covariance control highlighted in Remark 3.

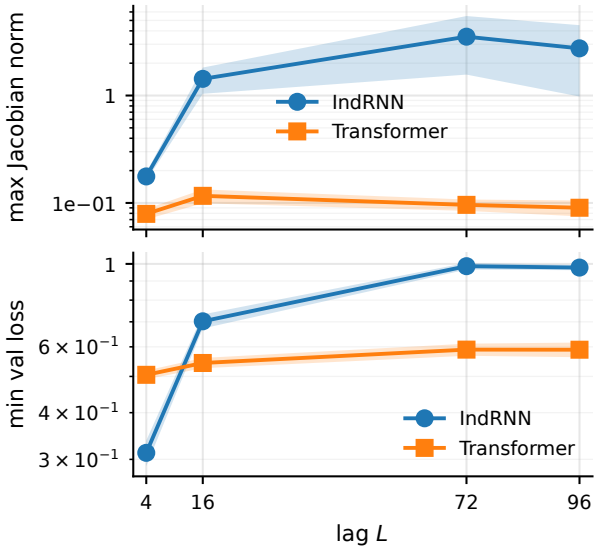


Figure 2: IndRNN vs. Transformer on the AR(L) task with $T = L + 1$. Top: maximum Jacobian norm over training (stability); bottom: minimum validation loss achieved over training (performance). Results are averaged over 20 seeds.

7 CONCLUSION

We performed a finite-time NTK analysis of projected gradient methods for multi-head shallow Transformers. We established nonasymptotic convergence bounds with explicit dependencies on the training period, width, and the number of samples. To the best of our knowledge, this is the first work that establishes near-optimality of first-order methods without massive overparameterization for Transformers. Interesting future directions that can build on this work include the study of deep Transformer architectures and the extension of the setting to sequence-to-sequence learning.

Acknowledgements

We thank the anonymous reviewers for their thoughtful comments and helpful suggestions, which improved the clarity and presentation of the paper. This work was supported in part by the NSF grants: NSF AI Institute (AI-EDGE) 2112471, CNS-NeTS-2106679, the ONR Grant N00014-19-1-2621; and the ARO Grant W911NF-24-1-0103.

References

Alemohammad, S., Wang, Z., Balestriero, R., and Baraniuk, R. G. (2021). The recurrent neural tangent kernel. *arXiv preprint arXiv:2006.10246*.

Allen-Zhu, Z., Li, Y., and Song, Z. (2019). A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. (2019). On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8139–8148.

Bai, Y. and Lee, J. D. (2020). Beyond linearization: On quadratic and higher-order approximation of wide neural networks. In *International Conference on Learning Representations*.

Bordelon, B., Chaudhry, H. T., and Pehlevan, C. (2024). Infinite limits of multi-head transformer dynamics. In *Advances in Neural Information Processing Systems*, volume 37.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Series in Probability and Statistics. Oxford University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., and et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Cayci, S. and Eryilmaz, A. (2024a). Convergence of gradient descent for recurrent neural networks: A nonasymptotic analysis. *arXiv preprint arXiv:2402.12241*.

Cayci, S. and Eryilmaz, A. (2024b). Recurrent natural policy gradient for pomdps. *arXiv preprint arXiv:2405.18221*.

Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc.

Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlós, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, Ł., Belanger, D. B., Colwell, L. J., and Weller, A. (2021). Rethinking attention with performers. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Deora, P., Ghaderi, R., Taheri, H., and Thrampoulidis, C. (2024). On the optimization and generalization of multi-head attention.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., and et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. (2019a). Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR.

Du, S. S., Zhai, X., Póczos, B., and Singh, A. (2019b). Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations (ICLR)*.

Garriga-Alonso, A., Aitchison, L., and Rasmussen, C. E. (2019). Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations (ICLR)*.

Hron, J., Bahri, Y., Sohl-Dickstein, J., and Novak, R. (2020). Infinite attention: Nngp and ntk for deep attention networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Huang, Y., Cheng, Y., and Liang, Y. (2024). In-context convergence of transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Touvron, H., et al. (2021). Perceiver io: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*.

Ji, Z., Telgarsky, M., and Xian, R. (2020). Neural tangent kernels, transportation mappings, and universal approximation. In *International Conference on Learning Representations*.

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. (2020). Transformers are rnns: Fast autoregressive transformers with linear attention. In Daumé III, H. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Li, S., Li, W., Cook, C., Zhu, C., and Gao, Y. (2018). Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5457–5466, Salt Lake City, UT. IEEE.

Lu, J., Yao, J., Zhang, J., Zhu, X., Xu, H., Gao, W., Xu, C., Xiang, T., and Zhang, L. (2021). SOFT: Softmax-free transformer with linear complexity. In *Advances in Neural Information Processing Systems*, volume 34, pages 21297–21309.

Nguyen, Q. (2021). On the proof of global convergence of gradient descent for deep relu networks with linear widths. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR.

Nguyen, Q., Mondelli, M., and Montúfar, G. F. (2021). Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR.

Nichani, E., Damian, A., and Lee, J. D. (2024). How Transformers Learn Causal Structure with Gradient Descent. arXiv:2402.14735 [cs, math, stat].

Noci, L., Li, C., Li, M. B., He, B., Hofmann, T., Maddison, C. J., and Roy, D. M. (2023). The shaped transformer: Attention models in the infinite depth-and-width limit. In *Advances in Neural Information Processing Systems*, volume 36.

Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A., Sohl-Dickstein, J., and Schoenholz, S. (2019). Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations (ICLR)*.

Oymak, S., Rawat, A. S., Soltanolkotabi, M., and Thrampoulidis, C. (2023). On the role of attention in prompt-tuning. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, pages 26724–26768, Honolulu, HI, USA. arXiv preprint arXiv:2306.03435.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI. OpenAI Technical Report.

Shen, W., Zhou, R., Yang, J., and Shen, C. (2024). On the training convergence of transformers for in-context classification of gaussian mixtures. *arXiv preprint arXiv:2410.11778*.

Song, B., Han, B., Zhang, S., Ding, J., and Hong, M. (2024). Unraveling the gradient descent dynamics of transformers. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Tarzanagh, D. A., Li, Y., Zhang, X., and Oymak, S. (2023). Max-Margin Token Selection in Attention Mechanism. arXiv:2306.13596 [cs, math].

Vasudeva, B., Deora, P., and Thrampoulidis, C. (2025). Implicit bias and fast convergence rates for self-attention.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wortsman, M., Lee, J., Gilmer, J., and Kornblith, S. (2023). Replacing softmax with relu in vision transformers. *arXiv preprint arXiv:2309.08586*.

Wu, Y., Liu, F., Chrysos, G. G., and Cevher, V. (2023). On the convergence of encoder-only shallow transformers. In *Advances in Neural Information Processing Systems*.

Yang, G. (2020). Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., and Kumar, S. (2020). Are Transformers universal approximators of sequence-to-sequence functions? arXiv:1912.10077 [cs, stat].

Zhang, R., Frei, S., and Bartlett, P. L. (2024). Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25:1–55.

Zheng, L., Wang, C., and Kong, L. (2022). Linear complexity randomized self-attention mechanism. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27011–27041. PMLR.

Zou, D. and Gu, Q. (2019). An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] **Justification:** The setup, the model, and the algorithm are described in Section 3. The function that is assumed to generate the underlying data is described in Section 5.4.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] **Justification:** The analysis of the algorithms is informally made in Section 4 and formally made in Section 5.5.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] **Justification:** The source code is provided at `https://github.com/enesarda22/nonasymptotic-transformer`.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] **Justification:** Every stated theoretical result contains the assumptions if any.
   (b) Complete proofs of all theoretical results. [Yes] **Justification:** Complete proofs are provided in the Supplementary Material.
   (c) Clear explanations of any assumptions. [Yes] **Justification:** The only assumption we make is on the function generating the underlying data, which is explained in Section 5.4.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] **Justification:** The code and instructions are provided in a Github repo in the Supplementary Material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] **Justification:** Experimental details are provided in Section 6 and additional details are provided in the Supplementary Material.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] **Justification:** Experimental details are provided in Section 6 and additional details are provided in the Supplementary Material.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No] **Justification:** Conducted experiments are not resource-heavy, so we do not describe the computing infrastructure.

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

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

Supplementary Material

Organization. In Appendix A, we restate and prove the results in the main text, and in Appendix B, we give further experimental details along with the code to reproduce the experiments.

A PROOFS

A.1 Useful facts

Fact 1 (Jacobian of softmax)**.** *For $\sigma_s : \mathbb{R}^T \to \Delta_T$ given by $(\sigma_s(z))_t = \exp(z_t)/\sum_{s=1}^T \exp(z_s)$, the Jacobian equals*

$$J_s(z) \;=\; \operatorname{diag}\big(\sigma_s(z)\big) \;-\; \sigma_s(z)\,\sigma_s(z)^\top.$$

It is symmetric, satisfies $J_s(z)\mathbf{1} = 0$, and obeys the operator bound $\|J_s(z)\|_{\infty\to 1} \le 2$. Here, $\mathbf{1}$ is the all-ones vector.

Proof. $\partial(\sigma_s)_i/\partial z_j = \sigma_{s,i}(\delta_{ij} - \sigma_{s,j})$ gives the formula and symmetry. Row sums vanish since $\sum_j \partial(\sigma_s)_i/\partial z_j = 0$, hence $J_s(z)\mathbf{1} = 0$. For the $\infty \to 1$ bound, if $\|u\|_\infty \le 1$ then $\|J_s(z)u\|_1 = \|\operatorname{diag}(\alpha)u - \alpha(\alpha^\top u)\|_1 \le \sum_i \alpha_i|u_i| + |\alpha^\top u|\,\|\alpha\|_1 \le 1 + 1 = 2$ with $\alpha = \sigma_s(z)$. □

Fact 2 (Bound for $a(X;W)$)**.** *Fix $X \in \mathcal{X}$. With $\alpha = \sigma_s(X^\top W q_X) \in \Delta_T$,*

$$a(X;W) \;=\; X\alpha \;=\; \sum_{t=1}^T \alpha_t\, X_t \quad\Rightarrow\quad \|a(X;W)\|_2 \;\le\; \sum_{t=1}^T \alpha_t \|X_t\|_2 \;\le\; \sum_{t=1}^T \alpha_t \;=\; 1.$$

Fact 3 (Bound for $M(X;W)$)**.** *Fix $X \in \mathcal{X}$. Let $\alpha = \sigma_s(X^\top W q_X)$ and $\mu = X\alpha$. Then*

$$M(X;W) \;=\; X\big(\operatorname{diag}(\alpha) - \alpha\alpha^\top\big)X^\top \;=\; \sum_{t=1}^T \alpha_t\,(X_t - \mu)(X_t - \mu)^\top \;\succeq\; 0.$$

Consequently,

$$M \;\preceq\; \sum_{t=1}^T \alpha_t\, X_t X_t^\top \quad \textit{and} \quad \|M\|_F \;\le\; \operatorname{tr}(M) \;\le\; \sum_{t=1}^T \alpha_t \|X_t\|_2^2 \;\le\; 1.$$

Fact 4 (Gaussian norm concentration)**.** *If $G \sim \mathcal{N}(0, I_d)$, then*

$$\mathbb{P}\big(\big|\|G\|_2 - \mathbb{E}\|G\|_2\big| \ge t\big) \;\le\; 2e^{-t^2/2} \quad \textit{for all } t > 0, \qquad \mathbb{E}\|G\|_2 \;\le\; \sqrt{d}.$$

In particular, for independent $G_1, \ldots, G_m \sim \mathcal{N}(0, I_d)$,

$$\mathbb{P}\left(\max_{1\le i\le m} \|G_i\|_2 \;\le\; \sqrt{d} + \sqrt{2\log(m/\delta)}\right) \;\ge\; 1 - \delta.$$

Proof. By Lipschitz Gaussian concentration (Boucheron et al., 2013, Theorem 5.6) applied to $u \mapsto \|u\|_2$, we get the tail bound; the expectation bound follows from $\mathbb{E}\|G\|_2 \le \sqrt{\mathbb{E}\|G\|_2^2} = \sqrt{d}$. The union bound over m copies yields the shown high-probability bound. □

A.2 Proof of Lemma 1

Lemma 1. *Fix $i \in [m]$ and write $a_i := a(X; W_i)$. Then*

$$\frac{\partial f(X;\varphi)}{\partial c_i} = m^{-1/2}\, h(X;\theta_i), \tag{12}$$

$$\nabla_{U_i} f(X;\varphi) = m^{-1/2}\, c_i\, \sigma'\big(U_i^\top a_i\big)\, a_i, \tag{13}$$

$$\nabla_{W_i} f(X;\varphi) = m^{-1/2}\, c_i\, \sigma'\big(U_i^\top a_i\big)\, \big(M_i U_i\big)\, q_X^\top, \tag{14}$$

where

$$M_i \;:=\; M(X;W_i) \;=\; X\, J_s\big(X^\top W_i q_X\big)\, X^\top, \tag{15}$$

and $J_s(z) \;=\; \mathrm{diag}\big(\sigma_s(z)\big) - \sigma_s(z)\sigma_s(z)^\top$ is the Jacobian matrix of softmax.

Proof. Write $\alpha_i := \sigma_s\big(X^\top W_i q_X\big) \in \Delta_T$, $a_i = X\alpha_i \in \mathbb{R}^d$, and $z_i := X^\top W_i q_X \in \mathbb{R}^T$.

Derivative w.r.t. c_i. From $f(X;\varphi) = \frac{1}{\sqrt{m}} \sum_{j=1}^m c_j h(X;\theta_j)$, $\partial f/\partial c_i = m^{-1/2}\, h(X;\theta_i)$.

Gradient w.r.t. U_i. Since $h(X;\theta_i) = \sigma(U_i^\top a_i)$,

$$\nabla_{U_i} h(X;\theta_i) \;=\; \sigma'\big(U_i^\top a_i\big)\, a_i, \qquad \nabla_{U_i} f \;=\; m^{-1/2}\, c_i\, \nabla_{U_i} h \;=\; m^{-1/2}\, c_i\, \sigma'\big(U_i^\top a_i\big)\, a_i.$$

Gradient w.r.t. W_i. We compute via differentials. First, by the chain rule for softmax,

$$d\alpha_i \;=\; J_s(z_i)\, dz_i, \qquad dz_i \;=\; X^\top\, dW_i\, q_X,$$

and hence

$$da_i \;=\; X\, d\alpha_i \;=\; X\, J_s(z_i)\, X^\top\, dW_i\, q_X \;=\; M_i\, dW_i\, q_X.$$

Next,

$$dh(X;\theta_i) \;=\; \sigma'\big(U_i^\top a_i\big)\, U_i^\top\, da_i \;=\; \sigma'\big(U_i^\top a_i\big)\, U_i^\top\, M_i\, dW_i\, q_X.$$

Therefore

$$df \;=\; m^{-1/2}\, c_i\, dh \;=\; m^{-1/2}\, c_i\, \sigma'\big(U_i^\top a_i\big)\, U_i^\top M_i\, dW_i\, q_X.$$

Using $\mathrm{tr}(A^\top B) = \langle A, B\rangle_F$ and cyclicity of trace,

$$U_i^\top M_i\, dW_i\, q_X \;=\; \mathrm{tr}\big(q_X\, U_i^\top M_i\, dW_i\big) \;=\; \mathrm{tr}\big((M_i U_i q_X^\top)^\top dW_i\big),$$

where we used that M_i is symmetric. Identifying the coefficient of dW_i in the Frobenius inner product yields

$$\nabla_{W_i} f(X;\varphi) \;=\; m^{-1/2}\, c_i\, \sigma'\big(U_i^\top a_i\big)\, (M_i U_i)\, q_X^\top,$$

as claimed. □

A.3 Proof of Lemma 2

Lemma 2. *On $\mathcal{E}_{U,m}(\delta')$, the mapping $\theta_i \mapsto h(X;\theta_i)$ is $L_{1,m}(\delta')$-Lipschitz and $L_{2,m}(\delta')$-smooth in Ω_ρ, where*

$$L_{1,m}(\delta') := \sigma_1 \sqrt{1 + B_{U,m}(\delta')^2},$$

and

$$L_{2,m}(\delta') := \sigma_2(1 + B_{U,m}(\delta')^2) + 8\sigma_1 \sqrt{1 + B_{U,m}(\delta')^2},$$

where $B_{U,m}(\delta')$ is defined as in (9).

Proof. Throughout, $D[\cdot]$ denotes the Fréchet derivative. For a scalar map $g : \mathbb{R}^{d\times d} \to \mathbb{R}$ the Fréchet derivative is $Dg(W)[\Delta] = \frac{d}{dt} g(W + t\Delta)\big|_{t=0}$. By the Riesz representation theorem, there is a unique $\nabla_W g(W)$ with $Dg(W)[\Delta] = \langle \nabla_W g(W), \Delta\rangle_F$, hence $\|\nabla_W g(W)\|_F = \sup_{\|\Delta\|_F=1} |Dg(W)[\Delta]|$. For a matrix-valued operator G, we measure $DG(W)$ by the induced norm $\|DG(W)\|_{F\to F} := \sup_{\|\Delta\|_F=1} \|DG(W)[\Delta]\|_F$.

Fix $i \in [m]$ and write

$$\alpha_i \;=\; \sigma_s\big(X^\top W_i q_X\big) \in \Delta_T, \qquad a_i \;=\; a(X; W_i) \;=\; X\alpha_i, \qquad z_i \;=\; U_i^\top a_i,$$

and

$$M_i \;=\; X\, J_s\big(X^\top W_i q_X\big)\, X^\top,$$

where J_s is the softmax Jacobian.

(a) Local Lipschitz continuity. From Lemma 1,

$$\nabla_{U_i} h(X;\theta_i) = \sigma'(z_i)\, a_i, \qquad \nabla_{W_i} h(X;\theta_i) = \sigma'(z_i)\,(M_i U_i)\, q_X^\top.$$

Therefore

$$\big\|\nabla_{U_i} h(X;\theta_i)\big\|_2 \le \sigma_1 \|a_i\|_2 \le \sigma_1, \qquad \big\|\nabla_{W_i} h(X;\theta_i)\big\|_F \le \sigma_1 \|M_i\|_F \|U_i\|_2 \|q_X\|_2 \le \sigma_1 \|U_i\|_2.$$

Collecting the two blocks,

$$\big\|\nabla_{\theta_i} h(X;\theta_i)\big\|_2^2 = \big\|\nabla_{U_i} h\big\|_2^2 + \big\|\nabla_{W_i} h\big\|_F^2 \;\le\; \sigma_1^2\,(1 + \|U_i\|_2^2).$$

Hence, on $\mathcal{E}_{U,m}(\delta')$, the mapping $\theta_i \mapsto h(X;\theta_i)$ is $L_{1,m}(\delta')$-Lipschitz in Ω_ρ with $L_{1,m}(\delta') = \sigma_1\sqrt{1 + B_{U,m}(\delta')^2}$.

(b) Local smoothness. Write $h_i := h(X;\theta_i) = \sigma(z_i)$. By the chain rule,

$$\nabla_{\theta_i} h_i = \sigma'(z_i)\, \nabla_{\theta_i} z_i, \qquad \nabla^2_{\theta_i} h_i = \sigma''(z_i)\,(\nabla_{\theta_i} z_i)(\nabla_{\theta_i} z_i)^\top + \sigma'(z_i)\, \nabla^2_{\theta_i} z_i,$$

and thus

$$\big\|\nabla^2_{\theta_i} h_i\big\|_{\mathrm{op}} \;\le\; \sigma_2 \underbrace{\big\|\nabla_{\theta_i} z_i\big\|_2^2}_{(i)} + \sigma_1 \underbrace{\big\|\nabla^2_{\theta_i} z_i\big\|_{\mathrm{op}}}_{(ii)}. \tag{16}$$

Bounding (i). We have $\nabla_{U_i} z_i = a_i$ and, for any direction $\Delta \in \mathbb{R}^{d\times d}$,

$$Dz_i[\Delta] \;=\; U_i^\top M_i\, \Delta\, q_X, \quad \Rightarrow \quad \|\nabla_{W_i} z_i\|_F \le \|M_i\|_F\, \|U_i\|_2\, \|q_X\|_2 \le \|U_i\|_2.$$

Therefore

$$\big\|\nabla_{\theta_i} z_i\big\|_2^2 = \|\nabla_{U_i} z_i\|_2^2 + \|\nabla_{W_i} z_i\|_F^2 \;\le\; 1 + \|U_i\|_2^2. \tag{17}$$

Bounding (ii). Note $\nabla^2_{U_i} z_i = 0$ and the W_i–U_i block of the Hessian equals the Jacobian of a_i w.r.t. W_i:

$$D[\nabla_{U_i} z_i]\,[\Delta] \;=\; D[a_i][\Delta] \;=\; M_i\, \Delta\, q_X, \quad \Rightarrow \quad \big\|\nabla^2_{U_i,\mathrm{vec}(W_i)} z_i\big\|_{\mathrm{op}} \le \big\|\nabla^2_{U_i,\mathrm{vec}(W_i)} z_i\big\|_F \;\le\; \|M_i\|_F\, \|q_X\|_2 \;\le\; 1.$$

For the W_i–W_i block, differentiate $\nabla_{W_i} z_i = M_i U_i q_X^\top$ once more in the direction Δ:

$$D[\nabla_{W_i} z_i]\,[\Delta] \;=\; D[M_i][\Delta]\, U_i\, q_X^\top.$$

From the covariance form of M_i, $D[M_i][\Delta] = \sum_{t=1}^T (D[\alpha_i][\Delta])_t\, (X_t - \mu_i)(X_t - \mu_i)^\top$, with $\mu_i := \sum_{t=1}^T (\alpha_i)_t X_t = a_i$ and $D[\alpha_i][\Delta] = J_s(X^\top W_i q_X)\, X^\top \Delta\, q_X$. Hence

$$\|D[M_i][\Delta]\|_F \;\le\; \sum_{t=1}^T |(D\alpha_i[\Delta])_t|\, \|X_t - \mu_i\|_2^2 \;\le\; 4\, \|D\alpha_i[\Delta]\|_1,$$

and using $\|\operatorname{diag}(p) - pp^\top\|_{\infty\to 1} \le 2$ (Fact 1),

$$\|D\alpha_i[\Delta]\|_1 = \big\|(\operatorname{diag}(\alpha_i) - \alpha_i\alpha_i^\top)\, X^\top \Delta\, q_X\big\|_1 \;\le\; 2\,\|X^\top \Delta\, q_X\|_\infty \;\le\; 2\,\|\Delta\|_F,$$

where we used $|X_t^\top \Delta q_X| \le \|X_t\|_2\,\|\Delta\|_F\,\|q_X\|_2 \le \|\Delta\|_F$. Thus $\|D[M_i][\Delta]\|_F \le 8\|\Delta\|_F$, and therefore

$$\|D[\nabla_{W_i} z_i]\,[\Delta]\|_F \le \|D[M_i][\Delta]\|_F \|U_i\|_2 \|q_X\|_2 \le 8\|\Delta\|_F \|U_i\|_2,$$

which implies

$$\big\|\nabla^2_{\mathrm{vec}(W_i)} z_i\big\|_{\mathrm{op}} \;\le\; 8\,\|U_i\|_2.$$

Let $B_i := \nabla^2_{U_i,\mathrm{vec}(W_i)} z_i$ and $C_i := \nabla^2_{\mathrm{vec}(W_i)} z_i$. Then

$$\nabla^2_{\theta_i} z_i = \begin{pmatrix} 0 & B_i^\top \\ B_i & C_i \end{pmatrix}.$$

For any (u, v) with $\|u\|_2^2 + \|v\|_2^2 = 1$,

$$\left\| \begin{pmatrix} 0 & B_i^\top \\ B_i & C_i \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \right\|_2^2 \le \|B_i\|_{\mathrm{op}}^2 \|v\|_2^2 + \big(\|B_i\|_{\mathrm{op}}\|u\|_2 + \|C_i\|_{\mathrm{op}}\|v\|_2\big)^2 \le 2\|B_i\|_{\mathrm{op}}^2 + \|C_i\|_{\mathrm{op}}^2.$$

Therefore

$$\big\|\nabla^2_{\theta_i} z_i\big\|_{\mathrm{op}} \;\le\; \sqrt{2\big\|\nabla^2_{U_i,\mathrm{vec}(W_i)} z_i\big\|_{\mathrm{op}}^2 + \big\|\nabla^2_{\mathrm{vec}(W_i)} z_i\big\|_{\mathrm{op}}^2} \;\le\; \sqrt{2 + 64\,\|U_i\|_2^2} \;\le\; 8\sqrt{1 + \|U_i\|_2^2}. \tag{18}$$

Finally, plugging (17) and (18) into (16) yields

$$\big\|\nabla^2_{\theta_i} h(X;\theta_i)\big\|_{\mathrm{op}} \;\le\; \sigma_2\,(1 + \|U_i\|_2^2) + 8\sigma_1\,\sqrt{1 + \|U_i\|_2^2}.$$

Therefore, on $\mathcal{E}_{U,m}(\delta')$ and for all $U_i \in \mathcal{U}_{\rho_u,i}$,

$$\big\|\nabla_{\theta_i} h(X;\theta_i)\big\|_2 \le \sigma_1\sqrt{1 + B_{U,m}(\delta')^2} \quad \text{and} \quad \big\|\nabla^2_{\theta_i} h(X;\theta_i)\big\|_{\mathrm{op}} \le \sigma_2\big(1 + B_{U,m}(\delta')^2\big) + 8\sigma_1\sqrt{1 + B_{U,m}(\delta')^2}.$$

□

A.4 Proof of Lemma 3

Lemma 3. *Assume $\varphi \in \Omega_\rho$ and define $\varepsilon_{\mathrm{lin}} := \sup_{j\in[n]} \big|f(X^{(j)};\varphi) - f_{\mathrm{lin}}(X^{(j)};\varphi)\big|$. Conditioned on the event $\mathcal{E}_{U,m}(\delta')$, we have*

$$\varepsilon_{\mathrm{lin}} \le \tfrac{1}{\sqrt{m}}\Big(L_{1,m}(\delta')\rho_c\sqrt{\rho_w^2 + \rho_u^2} + L_{2,m}(\delta')(\rho_w^2 + \rho_u^2)\Big),$$

where the upper bound is denoted as $B_{\mathrm{lin},m}(\delta')$.

Proof. Fix any $j \in [n]$ and write $X = X^{(j)}$. For $i \in [m]$ set $\delta c_i := c_i - c_i^{(0)}$, $\delta U_i := U_i - U_i^{(0)}$, $\delta W_i := W_i - W_i^{(0)}$, and

$$\delta\theta_i \;:=\; (\delta U_i, \mathrm{vec}(\delta W_i)), \qquad \|\delta\theta_i\|_2^2 \;=\; \|\delta U_i\|_2^2 + \|\delta W_i\|_F^2.$$

We can write $f_{\mathrm{lin}}(X;\varphi)$ as

$$\begin{aligned} f_{\mathrm{lin}}(X;\varphi) &= f\big(X;\varphi^{(0)}\big) \;+\; \sum_{i=1}^m \Big\langle \nabla_{c_i} f\big(X;\varphi^{(0)}\big),\, \delta c_i \Big\rangle \;+\; \sum_{i=1}^m \Big\langle \nabla_{\theta_i} f\big(X;\varphi^{(0)}\big),\, \delta\theta_i \Big\rangle \\ &= f\big(X;\varphi^{(0)}\big) \;+\; \frac{1}{\sqrt{m}} \sum_{i=1}^m \Big[\, h\big(X;\theta_i^{(0)}\big)\,\delta c_i + c_i^{(0)}\,\big\langle \nabla_{\theta_i} h\big(X;\theta_i^{(0)}\big),\, \delta\theta_i \big\rangle \Big]. \end{aligned}$$

A direct expansion gives the per-sample linearization error

$$\begin{aligned} e_{\text{lin}}(X) &:= f(X;\varphi) - f_{\text{lin}}(X;\varphi) \\ &= \frac{1}{\sqrt{m}}\sum_{i=1}^{m} \delta c_i \Big(h(X;\theta_i) - h\big(X;\theta_i^{(0)}\big)\Big) \\ &\quad + \frac{1}{\sqrt{m}}\sum_{i=1}^{m} c_i^{(0)} \Big(h(X;\theta_i) - h\big(X;\theta_i^{(0)}\big) - \big\langle \nabla_{\theta_i} h\big(X;\theta_i^{(0)}\big),\, \delta\theta_i\big\rangle\Big). \end{aligned} \tag{19}$$

where we use the fact that $f(X;\varphi^{(0)}) = 0$. Since $\varphi \in \Omega_\rho$, we have

$$|\delta c_i| \le \frac{\rho_c}{\sqrt{m}}, \qquad \|\delta\theta_i\|_2 \le \frac{\sqrt{\rho_u^2+\rho_w^2}}{\sqrt{m}}, \qquad \|\delta\theta_i\|_2^2 \le \frac{\rho_u^2+\rho_w^2}{m}.$$

Using Lemma 2 and the fact that $|c_i^{(0)}| = 1$ at initialization, on the event $\mathcal{E}_{U,m}(\delta')$ we obtain

$$\begin{aligned} |e_{\text{lin}}(X)| &\le \frac{1}{\sqrt{m}}\sum_{i=1}^{m} \frac{\rho_c}{\sqrt{m}} \cdot L_{1,m}(\delta') \cdot \frac{\sqrt{\rho_u^2+\rho_w^2}}{\sqrt{m}} \;+\; \frac{1}{\sqrt{m}}\sum_{i=1}^{m} 1 \cdot \frac{L_{2,m}(\delta')}{2} \cdot \frac{\rho_u^2+\rho_w^2}{m} \\ &= \frac{1}{\sqrt{m}}\, L_{1,m}(\delta')\, \rho_c \sqrt{\rho_u^2+\rho_w^2} \;+\; \frac{1}{\sqrt{m}}\, \frac{L_{2,m}(\delta')}{2}\, (\rho_u^2+\rho_w^2) \\ &\le \frac{1}{\sqrt{m}}\Big(L_{1,m}(\delta')\, \rho_c \sqrt{\rho_u^2+\rho_w^2} + L_{2,m}(\delta')\, (\rho_u^2+\rho_w^2)\Big). \end{aligned}$$

This bound holds for every $X = X^{(j)}$, hence after taking a supremum over $j \in [n]$ we get

$$\varepsilon_{\text{lin}} \;=\; \sup_{j\in[n]} |e_{\text{lin}}(X^{(j)})| \;\le\; \frac{1}{\sqrt{m}}\Big(L_{1,m}(\delta')\, \rho_c \sqrt{\rho_u^2+\rho_w^2} + L_{2,m}(\delta')\, (\rho_u^2+\rho_w^2)\Big) \;=:\; B_{\text{lin},m}(\delta').$$

□

A.5 Proof of Lemma 4

Lemma 4. *Let $a = a(X;W)$, $a' = a(X';W)$, $M = X\, J_s(X^\top W q_X)\, X^\top$, and $M' = X'\, J_s(X'^\top W q_{X'})\, X'^\top$. Then, with expectation taken over the random initialization $(c, U, \text{vec}(W)) \sim \varphi_1^{(0)}$,*

$$\begin{aligned} K_c(X,X') &= \mathbb{E}\big[\, \sigma(U^\top a)\, \sigma(U^\top a')\, \big]\,, \\ K_u(X,X') &= \mathbb{E}\big[\, \sigma'(U^\top a)\, \sigma'(U^\top a')\, \langle a, a'\rangle\, \big]\,, \\ K_w(X,X') &= \langle q_X, q_{X'}\rangle \mathbb{E}\big[\sigma'(U^\top a)\sigma'(U^\top a') U^\top M M' U\big] \end{aligned}$$

and the NTK decomposes as $K = K_c + K_u + K_w$.

Proof. Define the width-m neural tangent kernel at initialization as

$$K_m(X,X') \;:=\; \big\langle \nabla_\varphi f\big(X;\varphi^{(0)}\big),\, \nabla_\varphi f\big(X';\varphi^{(0)}\big)\big\rangle.$$

By block structure, this decomposes as a sum over units and parameter blocks:

$$\begin{aligned} K_m(X,X') \;=\; \sum_{i=1}^{m}\Bigg[&\underbrace{\frac{\partial f}{\partial c_i}(X;\varphi^{(0)})\, \frac{\partial f}{\partial c_i}(X';\varphi^{(0)})}_{c\text{–block}} \;+\; \underbrace{\big\langle \nabla_{U_i} f(X;\varphi^{(0)}),\, \nabla_{U_i} f(X';\varphi^{(0)})\big\rangle}_{U\text{–block}} \\ &+\; \underbrace{\big\langle \nabla_{W_i} f(X;\varphi^{(0)}),\, \nabla_{W_i} f(X';\varphi^{(0)})\big\rangle_F}_{W\text{–block}}\Bigg]. \end{aligned}$$

We remove the superscript and write $(c_i^{(0)}, U_i^{(0)}, W_i^{(0)}) = (c_i, U_i, W_i)$

c–**component.**

$$\sum_{i=1}^{m} \frac{\partial f}{\partial c_i}(X;\varphi^{(0)})\,\frac{\partial f}{\partial c_i}(X';\varphi^{(0)}) = \frac{1}{m}\sum_{i=1}^{m} \sigma\big({U_i}^\top a(X;W_i)\big)\,\sigma\big(U_i^\top a(X';W_i)\big)$$
$$= \frac{2}{m}\sum_{i=1}^{m/2} \sigma\big({U_i}^\top a(X;W_i)\big)\,\sigma\big(U_i^\top a(X';W_i)\big).$$

Since the pairs $(c_i, U_i, W_i)_{i=1}^{m/2}$ are i.i.d. at initialization, by the strong law of large numbers (SLLN),

$$K_c(X,X') \;:=\; \lim_{m\to\infty} \frac{2}{m}\sum_{i=1}^{m/2} \sigma\big(U_i^\top a(X;W_i)\big)\,\sigma\big(U_i^\top a(X';W_i)\big) = \mathbb{E}\big[\,\sigma(U^\top a)\,\sigma(U^\top a')\,\big]\,.$$

U–**component.** Writing $a_i = a(X;W_i)$ and $a_i' = a(X';W_i)$,

$$\sum_{i=1}^{m} \big\langle \nabla_{U_i} f(X;\varphi^{(0)}),\, \nabla_{U_i} f(X';\varphi^{(0)})\big\rangle = \frac{1}{m}\sum_{i=1}^{m} c_i^2\,\sigma'\big(U_i^\top a_i\big)\,\sigma'\big(U_i^\top a_i'\big)\;\langle a_i, a_i'\rangle$$
$$= \frac{2}{m}\sum_{i=1}^{m/2} \sigma'\big(U_i^\top a_i\big)\,\sigma'\big(U_i^\top a_i'\big)\;\langle a_i, a_i'\rangle$$

where we used the fact that $c_i^2 = 1$ surely at initialization. Then SLLN yields

$$K_u(X,X') \;:=\; \lim_{m\to\infty} \frac{2}{m}\sum_{i=1}^{m/2} \sigma'\big(U_i^\top a_i\big)\,\sigma'\big(U_i^\top a_i'\big)\;\langle a_i, a_i'\rangle = \mathbb{E}\big[\,\sigma'(U^\top a)\,\sigma'(U^\top a')\,\langle a, a'\rangle\,\big]\,.$$

W–**component.** Writing $M_i = X\,J_s(X^\top W_i q_X)\,X^\top$ and $M_i' = X'\,J_s(X'^\top W_i q_{X'})\,X'^\top$, using the Frobenius product identity $\langle xy^\top, x'y'^\top\rangle_F = \langle x, x'\rangle\,\langle y, y'\rangle$ and that M_i and M_i' are symmetric,

$$\sum_{i=1}^{m} \big\langle \nabla_{W_i} f(X),\, \nabla_{W_i} f(X')\big\rangle_F = \frac{1}{m}\sum_{i=1}^{m} \sigma'\big(U_i^\top a_i\big)\sigma'\big(U_i^\top a_i'\big)\big\langle (M_iU_i)q_X^\top,\; (M_i'U_i)q_{X'}^\top\big\rangle_F$$
$$= \frac{2}{m}\sum_{i=1}^{m/2} \sigma'\big(U_i^\top a_i\big)\sigma'\big(U_i^\top a_i'\big)\;\langle q_X, q_{X'}\rangle\; U_i^\top M_i M_i' U_i.$$

Again using SLLN,

$$K_w(X,X') := \lim_{m\to\infty} \frac{2}{m}\sum_{i=1}^{m/2} \sigma'\big(U_i^\top a_i\big)\sigma'\big(U_i^\top a_i'\big)\,\langle q_X, q_{X'}\rangle\, U_i^\top M_i M_i' U_i$$
$$= \langle q_X, q_{X'}\rangle\,\mathbb{E}\big[\sigma'(U^\top a)\sigma'(U^\top a')\,U^\top M M' U\big]\,.$$

Putting the three blocks together and taking $m \to \infty$,

$$K(X,X') \;=\; \lim_{m\to\infty} K_m(X,X') \;=\; K_c(X,X') + K_u(X,X') + K_w(X,X').$$

□

A.6 Proof of Lemma 5

Lemma 5. *Let $\tilde{f}(\cdot;v) \in \mathcal{F}_{\bar{\nu}}$ and $\tilde{\varphi}$ be as in* (11)*. Define*

$$\varepsilon_{\text{app}} := \sup_{j\in[n]} \big|f_{\text{lin}}(X^{(j)};\tilde{\varphi}) - \tilde{f}(X^{(j)};v)\big|.$$

Then for any $\delta \in (0,1)$, we have

$$\varepsilon_{\text{app}} \;\le\; 4(\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + \sigma_1\bar{\nu}_w)\sqrt{\tfrac{\log(2n/\delta)}{m}} =: B_{\text{app},m}(\delta),$$

with probability at least $1-\delta$. We denote this event with $\mathcal{E}_{\text{app}}(\delta)$.

Proof. Fix $X \in \mathcal{X}$, and write $a := a(X;W)$, $M := M(X;W)$. Define, for each $i \in [m]$,

$$g_i(X) := g_i^{(c)}(X) + g_i^{(u)}(X) + g_i^{(w)}(X),$$

where

$$\begin{aligned} g_i^{(c)}(X) &:= \phi_c\big(X;\varphi_i^{(0)}\big)\, v_c\big(\varphi_i^{(0)}\big), \\ g_i^{(u)}(X) &:= \Big\langle \phi_u\big(X;\varphi_i^{(0)}\big),\, v_u\big(\varphi_i^{(0)}\big)\Big\rangle, \\ g_i^{(w)}(X) &:= \Big\langle \phi_w\big(X;\varphi_i^{(0)}\big),\, v_w\big(\varphi_i^{(0)}\big)\Big\rangle_F, \end{aligned}$$

with

$$\phi_c(X;\varphi_0) = \sigma(U^\top a(X;W)), \qquad \phi_u(X;\varphi_0) = \sigma'(U^\top a(X;W))\, a(X;W),$$

and

$$\phi_w(X;\varphi_0) = \sigma'(U^\top a(X;W))\,(M(X;W)U)\, q_X^\top.$$

By the definition of f_{lin} and $\tilde{f}(\cdot;v)$,

$$f_{\text{lin}}(X;\tilde{\varphi}) = \frac{1}{m}\sum_{i=1}^{m} g_i(X), \qquad \tilde{f}(X;v) = \mathbb{E}[g_1(X)].$$

Hence

$$f_{\text{lin}}(X;\tilde{\varphi}) - \tilde{f}(X;v) = \frac{1}{m}\sum_{i=1}^{m}\big(g_i(X) - \mathbb{E}[g_i(X)]\big).$$

Using the paired initialization, define

$$\bar{g}_i^{(\ell)}(X) := \frac{1}{2}\Big(g_i^{(\ell)}(X) + g_{i+m/2}^{(\ell)}(X)\Big), \qquad \ell \in \{c,u,w\}, \quad i \in [m/2].$$

Since the pairs $\big(\varphi_i^{(0)}, \varphi_{i+m/2}^{(0)}\big)$ are i.i.d. across $i \in [m/2]$, the random variables

$$Z_i(X) := \sum_{\ell\in\{c,u,w\}} \Big(\bar{g}_i^{(\ell)}(X) - \mathbb{E}[\bar{g}_1^{(\ell)}(X)]\Big), \qquad i \in [m/2],$$

are i.i.d. and mean zero, and

$$f_{\text{lin}}(X;\tilde{\varphi}) - \tilde{f}(X;v) = \frac{2}{m}\sum_{i=1}^{m/2} Z_i(X).$$

c-component. Since $|\phi_c| \le \sigma_0$ and $|v_c| \le \bar{\nu}_c$, we have

$$|g_i^{(c)}(X)| \le \sigma_0\bar{\nu}_c, \qquad |\bar{g}_i^{(c)}(X)| \le \sigma_0\bar{\nu}_c.$$

Therefore, by Hoeffding's lemma,

$$\mathbb{E}\exp\Big(\lambda\big(\bar{g}_i^{(c)}(X) - \mathbb{E}\bar{g}_i^{(c)}(X)\big)\Big) \le \exp\Big(\frac{\lambda^2(\sigma_0\bar{\nu}_c)^2}{2}\Big), \qquad \forall\lambda \in \mathbb{R}.$$

u-component. By Fact 2, $\|a(X;W)\|_2 \le 1$. Since $|\sigma'| \le \sigma_1$ and $\|v_u\|_2 \le \bar{\nu}_u$,

$$|g_i^{(u)}(X)| \le \|\phi_u(X;\varphi_i^{(0)})\|_2 \, \|v_u(\varphi_i^{(0)})\|_2 \le \sigma_1 \bar{\nu}_u,$$

and hence $|\bar{g}_i^{(u)}(X)| \le \sigma_1 \bar{\nu}_u$. Another application of Hoeffding's lemma gives

$$\mathbb{E}\exp\Big(\lambda\big(\bar{g}_i^{(u)}(X) - \mathbb{E}\bar{g}_i^{(u)}(X)\big)\Big) \le \exp\Big(\frac{\lambda^2(\sigma_1\bar{\nu}_u)^2}{2}\Big), \qquad \forall \lambda \in \mathbb{R}.$$

w-component. Let $Y_i := \bar{g}_i^{(w)}(X)$. We have $|Y_i| \le \sigma_1 \bar{\nu}_w \|M(X;W_i^{(0)})U_i^{(0)}\|_2$. We claim that

$$\mathbb{E}\exp\Big(\lambda(Y_i - \mathbb{E}Y_i)\Big) \le \exp(2\lambda^2(\sigma_1\bar{\nu}_w)^2), \qquad \forall\, \lambda \in \mathbb{R}.$$

By standard symmetrization trick, it suffices to show

$$\mathbb{E}|Y_i|^{2k} \le (\sigma_1\bar{\nu}_w)^{2k} \cdot \frac{(2k)!}{2^k k!}, \qquad k \ge 1. \tag{20}$$

Conditional on $W_i^{(0)}$, let $s_1, \dots, s_d$ be the singular values of $M(X;W_i^{(0)})$. Since $U_i^{(0)} \sim \mathcal{N}(0, I_d)$, we have $\|MU_i^{(0)}\|_2^2 \overset{d}{=} \sum_{r=1}^d s_r^2 G_r^2$ with $G_r \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. By Jensen's inequality (viewing $\{s_r^2/\sum s_r^2\}$ as weights and using convexity of $x \mapsto x^k$),

$$\Big(\sum_r s_r^2 G_r^2\Big)^k \le \Big(\sum_r s_r^2\Big)^{k-1} \sum_r s_r^2 G_r^{2k}.$$

Taking conditional expectation and using $\sum_r s_r^2 = \operatorname{tr}(M^\top M) \le (\operatorname{tr} M)^2 \le 1$ (Fact 3), we get

$$\mathbb{E}\big[\|MU_i^{(0)}\|_2^{2k} \mid W_i^{(0)}\big] \le \mathbb{E}|G_1|^{2k} = \frac{(2k)!}{2^k k!},$$

which gives (20) after multiplying by $(\sigma_1\bar{\nu}_w)^{2k}$ and taking full expectation.

Substituting (20) into the symmetrized MGF series yields

$$\mathbb{E}\, e^{\lambda(Y_i - \mathbb{E}Y_i)} \le \sum_{k=0}^{\infty} \frac{(2\lambda\sigma_1\bar{\nu}_w)^{2k}}{(2k)!} \cdot \frac{(2k)!}{2^k k!} = \sum_{k=0}^{\infty} \frac{(2\lambda^2(\sigma_1\bar{\nu}_w)^2)^k}{k!} = e^{2\lambda^2(\sigma_1\bar{\nu}_w)^2},$$

proving the claim.

Combining the bounds. Write

$$Z_i^{(c)} := \bar{g}_i^{(c)} - \mathbb{E}\bar{g}_i^{(c)}, \qquad Z_i^{(u)} := \bar{g}_i^{(u)} - \mathbb{E}\bar{g}_i^{(u)}, \qquad Z_i^{(w)} := \bar{g}_i^{(w)} - \mathbb{E}\bar{g}_i^{(w)}.$$

Then $Z_i = Z_i^{(c)} + Z_i^{(u)} + Z_i^{(w)}$. For any $\lambda \in \mathbb{R}$, applying Hölder's inequality with exponents $1/p_c + 1/p_u + 1/p_w = 1$ we get

$$\begin{aligned}
\mathbb{E}e^{\lambda Z_i} &\le \Big(\mathbb{E}e^{p_c\lambda Z_i^{(c)}}\Big)^{1/p_c} \Big(\mathbb{E}e^{p_u\lambda Z_i^{(u)}}\Big)^{1/p_u} \Big(\mathbb{E}e^{p_w\lambda Z_i^{(w)}}\Big)^{1/p_w} \\
&\le \exp\Big(\frac{p_c\lambda^2(\sigma_0\bar{\nu}_c)^2}{2}\Big) \exp\Big(\frac{p_u\lambda^2(\sigma_1\bar{\nu}_u)^2}{2}\Big) \exp\Big(2p_w\lambda^2(\sigma_1\bar{\nu}_w)^2\Big) \\
&= \exp\Big(\frac{\lambda^2}{2}(\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + 2\sigma_1\bar{\nu}_w)^2\Big),
\end{aligned}$$

where in the last equality we set

$$p_c = \frac{\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + 2\sigma_1\bar{\nu}_w}{\sigma_0\bar{\nu}_c}, \qquad p_u = \frac{\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + 2\sigma_1\bar{\nu}_w}{\sigma_1\bar{\nu}_u}, \qquad p_w = \frac{\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + 2\sigma_1\bar{\nu}_w}{2\sigma_1\bar{\nu}_w}.$$

Thus each Z_i is centered sub-Gaussian with proxy $\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + 2\sigma_1\bar{\nu}_w$. Since $Z_1, \ldots, Z_N$ are i.i.d., for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}\exp\Big(\lambda \sum_{i=1}^{m/2} Z_i\Big) \le \exp\Big(\frac{m\lambda^2}{4}(\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + 2\sigma_1\bar{\nu}_w)^2\Big).$$

By Chernoff's bound,

$$\mathbb{P}\left(\left|\frac{2}{m}\sum_{i=1}^{m/2} Z_i\right| \ge t\right) \le 2\exp\left(-\frac{mt^2}{4(\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + 2\sigma_1\bar{\nu}_w)^2}\right).$$

Choosing

$$t = \sqrt{2}\,(\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + 2\sigma_1\bar{\nu}_w)\sqrt{\frac{2\log(2n/\delta)}{m}},$$

and a union bound over $X^{(1)}, \ldots, X^{(n)}$ yields, with probability at least $1-\delta$,

$$\varepsilon_{\text{app}} = \sup_{j\in[n]} \left| f_{\text{lin}}(X^{(j)};\tilde{\varphi}) - \tilde{f}(X^{(j)};v)\right| \le \sqrt{2}\,(\sigma_0\bar{\nu}_c + \sigma_1\bar{\nu}_u + 2\sigma_1\bar{\nu}_w)\sqrt{\frac{2\log(2n/\delta)}{m}}.$$

This proves the lemma. □

A.7 Proof of Theorem 1

The following lemma will be useful.

Lemma 6 (Uniform envelope bounds)**.** *On the event $\mathcal{E}_{U,m}(\delta')$ and for any $\varphi \in \Omega_\rho$,*

$$\sup_{X\in\mathcal{X}} |f(X;\varphi)| \;\le\; L_{1,m}(\delta')\,\sqrt{\rho_u^2+\rho_w^2} \;+\; \sigma_0\,\rho_c =: f^{\max}.$$

Moreover, for any $\tilde{f} \in \mathcal{F}_{\bar{\nu}}$,

$$\sup_{X\in\mathcal{X}} |\tilde{f}(X;v)| \;\le\; \sigma_0\,\bar{\nu}_c \;+\; \sigma_1\,\bar{\nu}_u \;+\; \sigma_1\sqrt{d}\,\bar{\nu}_w =: y^{\max}.$$

Proof. **Bounding $|f(X;\varphi)|$.** Fix $X \in \mathcal{X}$ and write $\delta c_i = c_i - c_i^{(0)}$, $\delta\theta_i = (U_i - U_i^{(0)}, \text{vec}(W_i - W_i^{(0)}))$. Thanks to symmetric initialization, $f(X;\varphi^{(0)}) = 0$, hence

$$|f(X;\varphi)| = \underbrace{\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left|c_i^{(0)}\big(h(X;\theta_i) - h(X;\theta_i^{(0)})\big)\right|}_{(i)} + \underbrace{\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left|\delta c_i\, h(X;\theta_i)\right|}_{(ii)}.$$

On $\mathcal{E}_{U,m}(\delta')$, Lemma 2 gives $\left|h(X;\theta_i) - h(X;\theta_i^{(0)})\right| \le L_{1,m}(\delta')\|\delta\theta_i\|_2$ and we also have $|c_i^{(0)}| = 1$ surely. Using Cauchy–Schwarz and $\varphi \in \Omega_\rho$, we get

$$(i) \le \frac{1}{\sqrt{m}}\sum_{i=1}^{m} L_{1,m}(\delta')\|\delta\theta_i\|_2 \;\le\; L_{1,m}(\delta')\sqrt{\rho_u^2+\rho_w^2}.$$

Similarly, since we have $|h(X;\theta_i)| \le \sigma_0$, we get

$$(ii) \;\le\; \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\frac{\rho_c}{\sqrt{m}}\sigma_0 \;=\; \sigma_0\,\rho_c.$$

Therefore, $|f(X;\varphi)| \le L_{1,m}(\delta')\sqrt{\rho_u^2+\rho_w^2} + \sigma_0\rho_c$, and taking the supremum over $X \in \mathcal{X}$ gives the stated bound.

Bounding $|\tilde{f}(X;v)|$. Recall

$$\tilde{f}(X;v) = \mathbb{E}[\phi_c(X)\,v_c(\varphi_0)] + \mathbb{E}[\langle\phi_u(X), v_u(\varphi_0)\rangle] + \mathbb{E}[\langle\phi_w(X), v_w(\varphi_0)\rangle_F]\,,$$

with $\phi_c(X) = \sigma(U^\top a)$, $\phi_u(X) = \sigma'(U^\top a)\, a$, $\phi_w(X) = \sigma'(U^\top a)\,(MU)\, q_X^\top$. Using $|\sigma| \le \sigma_0$, $|\sigma'| \le \sigma_1$, $\|a\|_2 \le 1$, and $\|M\|_F \le 1$, we obtain

$$\begin{aligned}
\big|\mathbb{E}[\phi_c(X)\, v_c(\varphi_0)]\big| &\le \sigma_0\, \bar{\nu}_c, \\
\big|\mathbb{E}[\langle \phi_u(X), v_u(\varphi_0)\rangle]\big| &\le \sigma_1\, \bar{\nu}_u, \\
\big|\mathbb{E}[\langle \phi_w(X), v_w(\varphi_0)\rangle_F]\big| &\le \sigma_1\, \mathbb{E}\big[\|(MU)q_X^\top\|_F\big]\, \bar{\nu}_w \;\le\; \sigma_1\, \mathbb{E}[\|U\|_2]\, \bar{\nu}_w \;\le\; \sigma_1 \sqrt{d}\, \bar{\nu}_w,
\end{aligned}$$

where the last inequality uses $\mathbb{E}\|U\|_2 \le \sqrt{d}$ for $U \sim \mathcal{N}(0, I_d)$. Summing the three bounds and taking the supremum over $X \in \mathcal{X}$ proves the claim.

□

Theorem 1. *Assume $f^\star \in \mathcal{F}_{\bar{\nu}}$ and the projection radius satisfies $\rho \ge \bar{\nu}$. Run* ProjGD *onto Ω_ρ for τ steps with the step size $\eta = 1/\sqrt{\tau}$. Then, conditioned on the event $\mathcal{E}_{U,m}(\delta') \cap \mathcal{E}_{\mathrm{app}}(\delta)$, we have*

$$\min_{s<\tau} \widehat{\mathcal{L}}_n\big(\varphi^{(s)}\big) \;\lesssim_{\bar{\sigma},\bar{\nu},\rho}\; \frac{L_{1,m}(\delta')^4}{\sqrt{\tau}} + L_{1,m}(\delta')(B_{\mathrm{app},m}(\delta) + B_{\mathrm{lin},m}(\delta') + \varepsilon_{\mathrm{CoF}}),$$

and for the average iterate $\bar{\varphi}^{(\tau)} := \frac{1}{\tau}\sum_{s=0}^{\tau-1} \varphi^{(s)}$,

$$\widehat{\mathcal{L}}_n\big(\bar{\varphi}^{(\tau)}\big) \;\lesssim_{\bar{\sigma},\bar{\nu},\rho}\; \frac{L_{1,m}(\delta')^4}{\sqrt{\tau}} + L_{1,m}(\delta')(B_{\mathrm{app},m}(\delta) + B_{\mathrm{lin},m}(\delta') + \varepsilon_{\mathrm{CoF}}) + B_{\mathrm{lin},m}(\delta')^2,$$

where $B_{\mathrm{lin},m}(\delta')$ and $B_{\mathrm{app},m}(\delta)$ are defined as in Lemmas 3 and 5, and

$$\varepsilon_{\mathrm{CoF}} \;\lesssim_{\bar{\sigma},\bar{\nu},\rho}\; \frac{L_{1,m}(\delta') + L_{2,m}(\delta')}{\sqrt{m}}.$$

Proof. Choose v with $f^\star = \tilde{f}(\cdot; v)$ and set the Lyapunov function $\mathcal{V}(\varphi) = \|\varphi - \tilde{\varphi}\|_2^2$, where $\tilde{\varphi}$ is from (11). Write the pre-projection point $\varphi'^{(s)} = \varphi^{(s)} - \eta\, \nabla\widehat{\mathcal{L}}_n(\varphi^{(s)})$. By non-expansivity of Euclidean projection onto a convex set and the Pythagorean identity,

$$\mathcal{V}(\varphi^{(s+1)}) \;\le\; \mathcal{V}(\varphi'^{(s)}) \;=\; \mathcal{V}(\varphi^{(s)}) + 2\eta\, \underbrace{\big\langle \nabla\widehat{\mathcal{L}}_n(\varphi^{(s)}),\, \tilde{\varphi} - \varphi^{(s)}\big\rangle}_{\text{(i)}} + \eta^2\, \underbrace{\big\|\nabla\widehat{\mathcal{L}}_n(\varphi^{(s)})\big\|_2^2}_{\text{(ii)}}. \tag{21}$$

Controlling (i). For shorthand write $f_j^{(s)} := f(X^{(j)}; \varphi^{(s)})$, $f_j^\star := f^\star(X^{(j)})$ and $g_j^{(s)} := \nabla_\varphi f(X^{(j)}; \varphi^{(s)})$, $g_j^{(0)} := \nabla_\varphi f(X^{(j)}; \varphi^{(0)})$. Then

$$\big\langle \nabla\widehat{\mathcal{L}}_n(\varphi^{(s)}),\, \tilde{\varphi} - \varphi^{(s)}\big\rangle = \frac{2}{n}\sum_{j=1}^{n} \big(f_j^{(s)} - f_j^\star\big)\, \big\langle g_j^{(s)},\, \tilde{\varphi} - \varphi^{(s)}\big\rangle. \tag{22}$$

Introduce the change-of-feature error

$$\varepsilon_j^{\mathrm{CoF}}(\varphi^{(s)}) := \big\langle g_j^{(s)} - g_j^{(0)},\, \tilde{\varphi} - \varphi^{(s)}\big\rangle, \tag{23}$$

and note that

$$\begin{aligned}
\big\langle g_j^{(s)},\, \tilde{\varphi} - \varphi^{(s)}\big\rangle &= \big\langle g_j^{(0)},\, \tilde{\varphi} - \varphi^{(s)}\big\rangle + \varepsilon_j^{\mathrm{CoF}}(\varphi^{(s)}) \\
&= f_{\mathrm{lin}}(X^{(j)}; \tilde{\varphi}) - f_{\mathrm{lin}}(X^{(j)}; \varphi^{(s)}) + \varepsilon_j^{\mathrm{CoF}}(\varphi^{(s)}).
\end{aligned} \tag{24}$$

Now add and subtract $f_j^\star - f_j^{(s)}$ in (24) to obtain

$$\big\langle g_j^{(s)},\, \tilde{\varphi} - \varphi^{(s)}\big\rangle = \underbrace{f_{\mathrm{lin}}(X^{(j)}; \tilde{\varphi}) - f_j^\star}_{\varepsilon_j^{\mathrm{app}}} + \big(f_j^\star - f_j^{(s)}\big) + \underbrace{f_j^{(s)} - f_{\mathrm{lin}}(X^{(j)}; \varphi^{(s)})}_{\varepsilon_j^{\mathrm{lin}}(\varphi^{(s)})} + \varepsilon_j^{\mathrm{CoF}}(\varphi^{(s)}).$$

From Lemmas 3 and 5 we have,

$$\left|\varepsilon_j^{\text{lin}}(\varphi^{(s)})\right| \le B_{\text{lin},m}(\delta'), \tag{25}$$

$$\left|\varepsilon_j^{\text{app}}\right| \le B_{\text{app},m}(\delta), \tag{26}$$

To bound $\left|\varepsilon_j^{\text{CoF}}(\varphi^{(s)})\right|$, we start from (23) and use the triangle and Cauchy–Schwarz inequalities,

$$\left|\varepsilon_j^{\text{CoF}}(\varphi^{(s)})\right| \;\le\; \sum_{i=1}^{m}\left|\left\langle \nabla_{\varphi_i} f(X^{(j)};\varphi^{(s)}) - \nabla_{\varphi_i} f(X^{(j)};\varphi^{(0)}),\; \tilde{\varphi}_i - \varphi_i^{(s)}\right\rangle\right| \tag{27}$$

$$\le\; \sum_{i=1}^{m}\left\|\nabla_{\varphi_i} f(X^{(j)};\varphi^{(s)}) - \nabla_{\varphi_i} f(X^{(j)};\varphi^{(0)})\right\|_2 \left\|\tilde{\varphi}_i - \varphi_i^{(s)}\right\|_2. \tag{28}$$

By construction of $\tilde{\varphi}$ and Ω_ρ,

$$\left\|\tilde{\varphi}_i - \varphi_i^{(s)}\right\|_2 \;\le\; \frac{\|\bar{\nu}\|_2 + \|\rho\|_2}{\sqrt{m}}. \tag{29}$$

Set $\delta c_i := c_i^{(s)} - c_i^{(0)}$, $\delta\theta_i := \theta_i^{(s)} - \theta_i^{(0)}$. From Lemma 1,

$$\nabla_{\theta_i} f(X;\varphi) \;=\; m^{-1/2}\, c_i\, \nabla_{\theta_i} h(X;\theta_i), \qquad \nabla_{c_i} f(X;\varphi) \;=\; m^{-1/2}\, h(X;\theta_i).$$

Hence

$$\nabla_{\theta_i} f\big(X^{(j)};\varphi^{(s)}\big) - \nabla_{\theta_i} f\big(X^{(j)};\varphi^{(0)}\big) = \frac{c_i^{(0)}}{\sqrt{m}}\Big(\nabla_{\theta_i} h\big(X^{(j)};\theta_i^{(s)}\big) - \nabla_{\theta_i} h\big(X^{(j)};\theta_i^{(0)}\big)\Big) + \frac{\delta c_i}{\sqrt{m}}\,\nabla_{\theta_i} h\big(X^{(j)};\theta_i^{(s)}\big), \tag{30}$$

$$\nabla_{c_i} f\big(X^{(j)};\varphi^{(s)}\big) - \nabla_{c_i} f\big(X^{(j)};\varphi^{(0)}\big) = \frac{1}{\sqrt{m}}\Big(h\big(X^{(j)};\theta_i^{(s)}\big) - h\big(X^{(j)};\theta_i^{(0)}\big)\Big). \tag{31}$$

Invoking Lemma 2, we have for all i:

$$\left\|\nabla_{\theta_i} h(X;\theta_i^{(s)}) - \nabla_{\theta_i} h(X;\theta_i^{(0)})\right\|_2 \le L_{2,m}(\delta')\,\|\delta\theta_i\|_2, \qquad \left|h(X;\theta_i^{(s)}) - h(X;\theta_i^{(0)})\right| \le L_{1,m}(\delta')\,\|\delta\theta_i\|_2,$$

and $\|\nabla_{\theta_i} h(X;\theta_i^{(s)})\|_2 \le L_{1,m}(\delta')$. With $|c_i^{(0)}| = 1$ at initialization, $|\delta c_i| \le \rho_c/\sqrt{m}$, and $\|\delta\theta_i\|_2 \le \sqrt{\rho_u^2+\rho_w^2}/\sqrt{m}$, the bounds (30)–(31) yield

$$\begin{aligned}\left\|\nabla_{\theta_i} f(X^{(j)};\varphi^{(s)}) - \nabla_{\theta_i} f(X^{(j)};\varphi^{(0)})\right\|_2 &\le \frac{L_{2,m}(\delta')}{\sqrt{m}}\,\frac{\sqrt{\rho_u^2+\rho_w^2}}{\sqrt{m}} \;+\; \frac{\rho_c}{\sqrt{m}}\,\frac{L_{1,m}(\delta')}{\sqrt{m}} \\ &= \frac{L_{2,m}(\delta')\sqrt{\rho_u^2+\rho_w^2} + \rho_c L_{1,m}(\delta')}{m},\end{aligned} \tag{32}$$

$$\left\|\nabla_{c_i} f(X^{(j)};\varphi^{(s)}) - \nabla_{c_i} f(X^{(j)};\varphi^{(0)})\right\|_2 \le \frac{L_{1,m}(\delta')}{\sqrt{m}}\,\frac{\sqrt{\rho_u^2+\rho_w^2}}{\sqrt{m}} = \frac{L_{1,m}(\delta')\sqrt{\rho_u^2+\rho_w^2}}{m}. \tag{33}$$

Combining the c- and θ-blocks via $\|(x,y)\|_2 \le \|x\|_2 + \|y\|_2$,

$$\left\|\nabla_{\varphi_i} f(X^{(j)};\varphi^{(s)}) - \nabla_{\varphi_i} f(X^{(j)};\varphi^{(0)})\right\|_2 \;\le\; \frac{(L_{1,m}(\delta') + L_{2,m}(\delta'))\sqrt{\rho_u^2+\rho_w^2} + \rho_c L_{1,m}(\delta')}{m}. \tag{34}$$

Plugging (29) and (34) into (27), we obtain, for every j,

$$\left|\varepsilon_j^{\text{CoF}}(\varphi^{(s)})\right| \;\le\; \Big((L_{1,m}(\delta') + L_{2,m}(\delta'))\sqrt{\rho_u^2+\rho_w^2} + \rho_c L_{1,m}(\delta')\Big)\,\frac{\|\bar{\nu}\|_2 + \|\rho\|_2}{\sqrt{m}} \;=:\; B_{\text{CoF},m}(\delta'). \tag{35}$$

Putting (24) in (22) and using the bounds (25), (26) and (35), we have, for all $s \ge 0$,

$$\begin{aligned}\left\langle \nabla\widehat{\mathcal{L}}_n(\varphi^{(s)}),\, \tilde{\varphi} - \varphi^{(s)}\right\rangle &= -\,2\,\widehat{\mathcal{L}}_n(\varphi^{(s)}) + \frac{2}{n}\sum_{j=1}^{n}\big(f_j^{(s)} - f_j^{\star}\big)\Big(\varepsilon_j^{\text{app}} + \varepsilon_j^{\text{lin}}(\varphi^{(s)}) + \varepsilon_j^{\text{CoF}}(\varphi^{(s)})\Big) \\ &\le -\,2\,\widehat{\mathcal{L}}_n(\varphi^{(s)}) + 2\,(f^{\max} + y^{\max})\Big(B_{\text{app},m}(\delta) + B_{\text{lin},m}(\delta') + B_{\text{CoF},m}(\delta')\Big),\end{aligned} \tag{36}$$

where $f^{\max}$ and $y^{\max}$ are defined as in Lemma 6.

Controlling (ii). For any $j \in [n]$ and head $i \in [m]$,

$$\left\|\nabla_{\varphi_i} f\big(X^{(j)};\varphi^{(s)}\big)\right\|_2 \;\le\; \frac{|h(X^{(j)};\theta_i^{(s)})|}{\sqrt{m}} + \frac{|c_i^{(s)}|}{\sqrt{m}}\left\|\nabla_{\theta_i} h\big(X^{(j)};\theta_i^{(s)}\big)\right\|_2 \;\le\; \frac{1}{\sqrt{m}}\Big(\sigma_0 + |c_i^{(s)}|\, L_{1,m}(\delta')\Big),$$

and since $|c_i^{(s)}| \le 1 + \rho_c/\sqrt{m}$ on Ω_ρ,

$$\left\|\nabla_{\varphi_i} f\big(X^{(j)};\varphi^{(s)}\big)\right\|_2 \;\le\; \frac{1}{\sqrt{m}}\Big(\sigma_0 + (1+\tfrac{\rho_c}{\sqrt{m}})\, L_{1,m}(\delta')\Big). \tag{37}$$

Summing the squares over $i = 1,\dots,m$ gives

$$\left\|\nabla_{\varphi} f\big(X^{(j)};\varphi^{(s)}\big)\right\|_2 \;\le\; A_m(\delta') \qquad \text{with} \qquad A_m(\delta') := \sigma_0 + \Big(1+\tfrac{\rho_c}{\sqrt{m}}\Big) L_{1,m}(\delta'). \tag{38}$$

Therefore,

$$\begin{aligned}\left\|\nabla \widehat{\mathcal{L}}_n(\varphi^{(s)})\right\|_2 &= \left\|\frac{2}{n}\sum_{j=1}^{n}\big(f_j^{(s)} - f_j^\star\big)\,\nabla_\varphi f\big(X^{(j)};\varphi^{(s)}\big)\right\|_2 \\ &\le \frac{2}{n}\sum_{j=1}^{n}|f_j^{(s)} - f_j^\star|\,\left\|\nabla_\varphi f\big(X^{(j)};\varphi^{(s)}\big)\right\|_2 \;\le\; 2\,(f^{\max} + y^{\max})\,A_m(\delta'),\end{aligned}$$

and hence

$$\left\|\nabla \widehat{\mathcal{L}}_n(\varphi^{(s)})\right\|_2^2 \;\le\; 4\,(f^{\max} + y^{\max})^2\, A_m(\delta')^2 \;=:\; B_{\mathrm{grad},m}(\delta'). \tag{39}$$

Lyapunov drift. Plugging (36) and (39) into (21) yields

$$\mathcal{V}(\varphi^{(s+1)}) - \mathcal{V}(\varphi^{(s)}) \;\le\; -4\eta\,\widehat{\mathcal{L}}_n(\varphi^{(s)}) + 4\eta\,(f^{\max} + y^{\max})\big(B_{\mathrm{app},m}(\delta) + B_{\mathrm{lin},m}(\delta') + B_{\mathrm{CoF},m}(\delta')\big) + \eta^2 B_{\mathrm{grad},m}(\delta'). \tag{40}$$

Summing (40) over $s = 0,\dots,\tau-1$ and using $\sum_{s<\tau}(\mathcal{V}(\varphi^{(s+1)}) - \mathcal{V}(\varphi^{(s)})) = \mathcal{V}(\varphi^{(\tau)}) - \mathcal{V}(\varphi^{(0)})$, we obtain after rearranging the terms

$$\frac{1}{\tau}\sum_{s=0}^{\tau-1}\widehat{\mathcal{L}}_n(\varphi^{(s)}) \;\le\; \frac{\mathcal{V}(\varphi^{(0)})}{4\eta\tau} + (f^{\max} + y^{\max})\big(B_{\mathrm{app},m}(\delta) + B_{\mathrm{lin},m}(\delta') + B_{\mathrm{CoF},m}(\delta')\big) + \frac{\eta}{4}\,B_{\mathrm{grad},m}(\delta'). \tag{41}$$

Choosing $\eta = \tau^{-1/2}$ and using $\mathcal{V}(\varphi^{(0)}) \le \|\bar{\nu}\|_2^2$ gives

$$\min_{0\le s<\tau}\widehat{\mathcal{L}}_n\big(\varphi^{(s)}\big) \;\le\; \frac{\|\bar{\nu}\|_2^2 + B_{\mathrm{grad},m}(\delta')}{4\sqrt{\tau}} + (f^{\max} + y^{\max})\big(B_{\mathrm{app},m}(\delta) + B_{\mathrm{lin},m}(\delta') + B_{\mathrm{CoF},m}(\delta')\big). \tag{42}$$

Average iterate. Define $\widehat{\mathcal{L}}_n^{\mathrm{lin}}(\varphi) := \frac{1}{n}\sum_{j=1}^{n}(f_{\mathrm{lin}}(X^{(j)};\varphi) - y^{(j)})^2$. From the uniform linearization bound in Lemma 3, for all $\varphi \in \Omega_\rho$,

$$\widehat{\mathcal{L}}_n(\varphi) \le 2\,\widehat{\mathcal{L}}_n^{\mathrm{lin}}(\varphi) + 2\,B_{\mathrm{lin},m}(\delta')^2, \qquad \widehat{\mathcal{L}}_n^{\mathrm{lin}}(\varphi) \le 2\,\widehat{\mathcal{L}}_n(\varphi) + 2\,B_{\mathrm{lin},m}(\delta')^2.$$

By convexity of $\widehat{\mathcal{L}}_n^{\mathrm{lin}}$ and Jensen,

$$\widehat{\mathcal{L}}_n^{\mathrm{lin}}\big(\bar{\varphi}^{(\tau)}\big) \;\le\; \frac{1}{\tau}\sum_{s=0}^{\tau-1}\widehat{\mathcal{L}}_n^{\mathrm{lin}}\big(\varphi^{(s)}\big) \;\le\; \frac{2}{\tau}\sum_{s=0}^{\tau-1}\widehat{\mathcal{L}}_n\big(\varphi^{(s)}\big) + 2\,B_{\mathrm{lin},m}(\delta')^2.$$

Combining the two inequalities yields the "sandwich" inequality

$$\widehat{\mathcal{L}}_n\big(\bar{\varphi}^{(\tau)}\big) \;\le\; \frac{4}{\tau}\sum_{s=0}^{\tau-1}\widehat{\mathcal{L}}_n\big(\varphi^{(s)}\big) + 6\,B_{\mathrm{lin},m}(\delta')^2. \tag{43}$$

Substituting (41) into (43) gives the desired inequality. □

A.8 Proof of Proposition 1

Proposition 1. *Under the same conditions as Theorem 1, we have, on the event $\mathcal{E}_{U,m}(\delta') \cap \mathcal{E}_{\text{app}}(\delta)$,*

$$\mathbb{E}\left[\min_{s<\tau} \widehat{\mathcal{L}}_n(\varphi^{(s)}) \,|\, \varphi^{(0)}\right] \lesssim_{\bar{\sigma},\bar{\nu},\rho} \frac{L_{1,m}(\delta')^4}{\sqrt{\tau}} + L_{1,m}(\delta')(B_{\text{app},m}(\delta) + B_{\text{lin},m}(\delta') + \varepsilon_{\text{CoF}}),$$

and for the average iterate,

$$\mathbb{E}\left[\widehat{\mathcal{L}}_n\big(\bar{\varphi}^{(\tau)}\big) \,|\, \varphi^{(0)}\right] \lesssim_{\bar{\sigma},\bar{\nu},\rho} \frac{L_{1,m}(\delta')^4}{\sqrt{\tau}} + L_{1,m}(\delta')(B_{\text{app},m}(\delta) + B_{\text{lin},m}(\delta') + \varepsilon_{\text{CoF}}) + B_{\text{lin},m}(\delta')^2.$$

Proof. Choose v with $f^\star = \tilde{f}(\cdot; v)$ and define the Lyapunov function $\mathcal{V}(\varphi) = \|\varphi - \tilde{\varphi}\|_2^2$, where $\tilde{\varphi}$ is from (11). Write the pre-projection iterate $\varphi'^{(s)} = \varphi^{(s)} - \eta\, \nabla_\varphi \ell_{J_s}(\varphi^{(s)})$. By non-expansivity of projection and the Pythagorean identity,

$$\mathcal{V}(\varphi^{(s+1)}) \le \mathcal{V}(\varphi^{(s)}) + 2\eta \underbrace{\langle \nabla_\varphi \ell_{J_s}(\varphi^{(s)}),\, \tilde{\varphi} - \varphi^{(s)} \rangle}_{\text{(i)}} + \eta^2 \underbrace{\|\nabla_\varphi \ell_{J_s}(\varphi^{(s)})\|_2^2}_{\text{(ii)}}. \tag{44}$$

Controlling (i). Let $E_s[\cdot] := \mathbb{E}_{J_s}[\cdot \,|\, J_{<s}, \varphi^{(0)}]$ denote expectation w.r.t. the fresh index J_s, keeping the past $J_{<s}$ and $\varphi^{(0)}$ fixed (so $\varphi^{(s)}$ is fixed). Since $E_s\big[\nabla_\varphi \ell_{J_s}(\varphi^{(s)})\big] = \nabla \widehat{\mathcal{L}}_n(\varphi^{(s)})$, the same "Controlling (i)" computation as in the proof of Theorem 1 gives

$$E_s\big[\langle \nabla_\varphi \ell_{J_s}(\varphi^{(s)}),\, \tilde{\varphi} - \varphi^{(s)} \rangle\big] \le -2\, \widehat{\mathcal{L}}_n(\varphi^{(s)}) + 2\,(f^{\max} + y^{\max})\Big(B_{\text{app},m}(\delta) + B_{\text{lin},m}(\delta') + B_{\text{CoF},m}(\delta')\Big).$$

Controlling (ii). Similarly, using the same arguments leading to (39), we have $\|\nabla_\varphi \ell_{J_s}(\varphi^{(s)})\|_2^2 \le B_{\text{grad},m}(\delta')$ almost surely and therefore

$$E_s\big[\|\nabla_\varphi \ell_{J_s}(\varphi^{(s)})\|_2^2\big] \le B_{\text{grad},m}(\delta').$$

Lyapunov drift. Taking $E_s[\cdot]$ in (44) and combining the previous two steps yields

$$E_s\big[\mathcal{V}(\varphi^{(s+1)}) - \mathcal{V}(\varphi^{(s)})\big] \le -4\eta\, \widehat{\mathcal{L}}_n(\varphi^{(s)}) + 4\eta\,(f^{\max} + y^{\max})\Big(B_{\text{app},m}(\delta) + B_{\text{lin},m}(\delta') + B_{\text{CoF},m}(\delta')\Big) + \eta^2 B_{\text{grad},m}(\delta').$$

Taking expectation over $J_0, \ldots, J_{\tau-1}$, summing $s = 0, \ldots, \tau-1$, using $\sum_{s<\tau}\big(\mathcal{V}(\varphi^{(s+1)}) - \mathcal{V}(\varphi^{(s)})\big) = \mathcal{V}(\varphi^{(\tau)}) - \mathcal{V}(\varphi^{(0)})$, and $\mathcal{V}(\varphi^{(0)}) \le \|\bar{\nu}\|_2^2$, we obtain after rearranging the terms

$$\frac{1}{\tau}\sum_{s=0}^{\tau-1} \mathbb{E}\big[\widehat{\mathcal{L}}_n(\varphi^{(s)}) | \varphi^{(0)}\big] \le \frac{\|\bar{\nu}\|_2^2}{4\eta\tau} + (f^{\max} + y^{\max})\Big(B_{\text{app},m}(\delta) + B_{\text{lin},m}(\delta') + B_{\text{CoF},m}(\delta')\Big) + \frac{\eta}{4} B_{\text{grad},m}(\delta'). \tag{45}$$

With $\eta = \tau^{-1/2}$ and $\min_{0\le s<\tau} \widehat{\mathcal{L}}_n(\varphi^{(s)}) \le \frac{1}{\tau}\sum_{s<\tau} \widehat{\mathcal{L}}_n(\varphi^{(s)})$, we get

$$\mathbb{E}\left[\min_{0\le s<\tau} \widehat{\mathcal{L}}_n\big(\varphi^{(s)}\big) | \varphi^{(0)}\right] \;\le\; \frac{\|\bar{\nu}\|_2^2 + B_{\text{grad},m}(\delta')}{4\sqrt{\tau}} + (f^{\max} + y^{\max})\Big(B_{\text{app},m}(\delta) + B_{\text{lin},m}(\delta') + B_{\text{CoF},m}(\delta')\Big). \tag{46}$$

Average iterate. Taking conditional expectation given $\varphi^{(0)}$ on both sides of (43), we get

$$\mathbb{E}\Big[\widehat{\mathcal{L}}_n\big(\bar{\varphi}^{(\tau)}\big) \,\Big|\, \varphi^{(0)}\Big] \;\le\; \frac{4}{\tau}\sum_{s=0}^{\tau-1} \mathbb{E}\Big[\widehat{\mathcal{L}}_n\big(\varphi^{(s)}\big) \,\Big|\, \varphi^{(0)}\Big] + 6\, B_{\text{lin},m}(\delta')^2. \tag{47}$$

Substituting (45) into (47) gives the desired inequality. □

B EXPERIMENTAL DETAILS

Code and Reproducibility. All code to reproduce our experiments is available at the following repository:

`https://github.com/enesarda22/nonasymptotic-transformer`

B.1 IndRNN Architecture

Our recurrent model is an Independent RNN (IndRNN) studied by Cayci and Eryilmaz (2024a) with one key difference: because the prediction target in our case is a scalar, we use only the final hidden state and apply a linear layer, rather than producing a sequence of outputs. Concretely, for an input sequence $X = [X_1, \ldots, X_T] \in \mathbb{R}^{d\times T}$, width m, activation σ, hidden matrix $U \in \mathbb{R}^{m\times d}$, diagonal recurrent weights $w \in \mathbb{R}^m$, and linear layer weights $c \in \mathbb{R}^m$, we define

$$h_0 = 0 \in \mathbb{R}^m, \tag{48}$$

$$h_t = \sigma\big(U\, X_t + w \odot h_{t-1}\big) \in \mathbb{R}^m, \qquad t = 1, \ldots, T, \tag{49}$$

$$f_{\mathrm{rnn}}(X; \varphi_{\mathrm{rnn}}) := \frac{1}{\sqrt{m}}\, c^\top h_T \in \mathbb{R}, \tag{50}$$

where $\odot$ denotes elementwise multiplication and $\varphi_{\mathrm{rnn}} = (\varphi_{\mathrm{rnn},i})_{i=1}^m$ with $\varphi_{\mathrm{rnn},i} := (U_i, w_i, c_i)$.

Symmetric initialization. Similar to Section 5.2, for IndRNN, we use a symmetric initialization: assume m is even and draw the first half independently as

$$U_i^{(0)} \sim \mathcal{N}(0, I_d), \qquad w_i^{(0)} \sim \mathrm{Rad}(\gamma), \quad c_i^{(0)} \sim \mathrm{Rad}(1), \qquad i = 1, \ldots, \tfrac{m}{2}. \tag{51}$$

The second half is set as

$$U_{i+m/2}^{(0)} = U_i^{(0)}, \quad w_{i+m/2}^{(0)} = w_i^{(0)}, \quad c_{i+m/2}^{(0)} = -\,c_i^{(0)}.$$

This pairing ensures that the initial predictor is identically zero for any input $X \in \mathcal{X}$.

Parameter set. Similar to Definition 2, for IndRNN, we define the neighborhood for projection to be

$$\Omega_{\mathrm{rnn},\rho} := \Big\{(U, w, c) : \ \|U - U^{(0)}\|_{2,\mathrm{row}} \le \tfrac{\rho_u}{\sqrt{m}},\ \|w - w^{(0)}\|_\infty \le \tfrac{\rho_w}{\sqrt{m}},\ \|c - c^{(0)}\|_\infty \le \tfrac{\rho_c}{\sqrt{m}}\Big\}.$$

Training error for IndRNN. For completeness, we state the training loss result for the IndRNN architecture with scalar outputs, which follows as a corollary of (Cayci and Eryilmaz, 2024a, Thm. 4.4).

Proposition 2. *Assume $\tilde{f} \in \mathcal{F}_{\mathrm{rnn},\bar{\nu}}$ (see (Cayci and Eryilmaz, 2024a, Eq. (3.5))), and that the targets are generated as $y^{(j)} = \tilde{f}(X^{(j)})_T$ for $j = 1, \ldots, n$, with projection radius $\rho \ge \bar{\nu}$. Run* ProjGD *on f_{rnn} with projection onto $\Omega_{\mathrm{rnn},\rho}$ for τ steps with step size $\eta = 1/\sqrt{\tau}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\min_{s<\tau}\ \widehat{\mathcal{L}}_n\big(\varphi_{\mathrm{rnn}}^{(s)}\big) \ \lesssim_{\bar{\sigma},\bar{\nu},\rho,\mu_T}\ \frac{1}{\sqrt{\tau}} \ + \ \sqrt{\frac{\log(2n/\delta)}{m}},$$

where $\mu_T = \mathcal{O}(1)$ if $\gamma + \frac{\rho_w}{\sqrt{m}} \le \frac{1}{\sigma_1}$, and $\mu_T = \exp(\Omega(T))$ otherwise. Here $\gamma > 0$ is the Rademacher scale in the symmetric initialization of the diagonal recurrent weights (see (51)*).*

Max Jacobian norms in training. For Transformer and IndRNN we calculate

$$\mathcal{G}_{\mathrm{rnn}}(\varphi_{\mathrm{rnn}}) := \frac{1}{n}\sum_{j=1}^n \nabla_w f_{\mathrm{rnn}}(X^{(j)}; \varphi_{\mathrm{rnn}}), \qquad \mathcal{G}_{\mathrm{tf}}(\varphi) := \frac{1}{n}\sum_{j=1}^n \nabla_W f(X^{(j)}; \varphi).$$

In Figure 2, we plot $\max_{s<\tau} \|\mathcal{G}_{\mathrm{rnn}}(\varphi_{\mathrm{rnn}}^{(s)})\|_2$ and $\max_{s<\tau} \|\mathcal{G}_{\mathrm{tf}}(\varphi^{(s)})\|_F$.