Transformers Implement Functional Gradient Descent to Learn Non-Linear Functions In Context

Xiang Cheng [1] **Yuxin Chen** [2] **Suvrit Sra** [3]

Abstract

Many neural network architectures are known to be Turing Complete, and can thus, in principle implement arbitrary algorithms. However, Transformers are unique in that they can implement gradient-based learning algorithms *under simple parameter configurations*. This paper provides theoretical and empirical evidence that (non-linear) Transformers naturally learn to implement gradient descent *in function space*, which in turn enable them to learn non-linear functions in context. Our results apply to a broad class of combinations of non-linear architectures and non-linear in-context learning tasks. Additionally, we show that the optimal choice of non-linear activation depends in a natural way on the class of functions that need to be learned.

1. Introduction

Transformers (Vaswani et al., 2017) have been observed to produce the correct output based on contextual demonstrations provided in the prompt alone, a phenomenon commonly known as in-context learning (ICL) (Brown et al., 2020). Understanding ICL and its underlying mechanism may hold the key to explaining the success of the Transformer architecture, and has thus attracted great attention.

A promising conjecture is that Transformers learn in-context by implementing algorithms in their forward pass (Akyürek et al., 2022; von Oswald et al., 2023a; Ahn et al., 2023; Zhang et al., 2023; Mahankali et al., 2023; von Oswald et al., 2023b; Lin et al., 2023; Bai et al., 2023). A subset of this work focuses on ICL for *linear functions* using *linear Transformers* (i.e., the attention module contains no nonlinear activations). In this setting, there exists a *simple parameter configuration* under which the Transformer implements gradient descent. Subsequently, (Ahn et al., 2023; Zhang et al., 2023; Mahankali et al., 2023) verified the local and global optimality of similar parameter configurations, and Zhang et al. (2023) also showed convergence to this parameter configuration when training the Transformer with gradient descent.

The above works provide a convincing explanation of how linear Transformers learn linear functions in context; their theoretical conclusions are also well supported by experiments (von Oswald et al., 2023a; Ahn et al., 2023). An important aspect of this setting is that linear Transformers are *very well suited* to learning linear functions—by simply setting the Query and Key matrices to the identity matrix, a single linear attention can implement one step of gradient descent for the least squares loss.

But in real Transformers, emphnon-linear activations such as softmax are very important; moreover, the training data are more likely generated by *complicated non-linear functions*. It is unclear whether there exists a similarly elegant construction for non-linear Transformers that explains how they could learn non-linear functions in context. These observations raise two central motivating questions:

(Q1) *What learning algorithms are implemented by Transformers with* ***non-linear activations****?*

(Q2) *Can Transformers learn* ***non-linear functions*** *of data in context?*

This paper aims to answer both questions, so as to shed light on the inner workings of Transformers, and in turn explain why Transformers are such powerful learners.

More specifically, we simultaneously consider both ***non-linear architectures***—Attention modules with arbitrary non-linear activations $\tilde{h}$ (e.g., softmax or ReLU) and ***non-linear data***—where labels are sampled from a non-linear process (e.g., a Gaussian Process, or certain more general processes, to be clarified later) conditioned on the covariates. Surprisingly, we show that the answers to questions (Q1) and (Q2) are *deeply intertwined*: there exists ***a simple parameter configuration that makes Transformers implement gradient***

[1]Massachusetts Institute of Technology [2]University of California, Davis [3]Technical University of Munich. Correspondence to: Xiang Cheng <x.cheng@berkeley.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

descent in function space; moreover, we show that this functional gradient descent converges to the ***Bayes optimal predictor if the non-linearity of the attention module matches the underlying data distribution.*** Beyond our construction, we also provide ***theoretical and empirical evidence*** that Transformers do indeed learn to implement functional gradient descent via training. Our analysis applies to a broad range of functions (such as labels generated from two layer ReLU networks, see Ex. 8) and common architectures (such as ReLU and Softmax Transformers, see Ex. 2, 3).

1.1. Summary of Contributions

The main contributions of this work are as follows:

1. In Proposition 3.1, we show that when the non-linearity in the Attention module matches a kernel $\mathcal{K}$, then ***Transformers can implement gradient descent in function space wrt the Reproducing Kernel Hilbert Space (RKHS) metric induced by*** $\mathcal{K}$. In Sections 3.1.1 and 3.1.2, we discuss the connection between the construction in Proposition 3.1 and several common Transformer variants. Our result generalizes the least-squares gradient descent construction from (von Oswald et al., 2023a).
2. In Proposition 3.4, we consider a general setting when the data labels $y^{(i)}$ are generated from a Kernel Gaussian Process. We show that when the non-linear module $\tilde{h}$ matches the generating kernel $\mathcal{K}$, ***the functional gradient descent construction converges to the Bayes optimal predictor as the number of layers increases.*** In Section 3.3, we *verify experimentally* that the highest accuracy is indeed achieved when the ***non-linear module matches the generating kernel***.
3. In Proposition D.1, we present a generalization of Propositions 3.1 and 3.4 to **multi-head** attention. A multi-head Transformer with different activation $\tilde{h}$ per-head can implement the Bayes-optimal functional gradient descent algorithm for any RKHS that is obtainable by composition of the kernels of each individual $\tilde{h}$.
4. We analyze the loss landscape of a Transformer on non-linear data. In Theorem 4.5, we characterize certain stationary points of the in-context loss under a sparsity constraint on the value matrix. When $\tilde{h}$ coincides with a kernel, ***this stationary point is exactly the functional gradient descent construction of Proposition 3.1***. We verify empirically that this stationary point is consistently learned during training.
5. In Theorem 4.6, we characterize stationary points of the in-context loss without the sparsity constraint. Our proposed stationary point implements an algorithm that interleaves steps of covariate transformation with functional gradient descent. Once again, we verify empirically that the stationary point is consistently learned during training.
6. Less importantly, but possibly of independent interest, our experiments in Section 3.3 identify a simple scenario where ReLU Transformers appears to out-perform softmax Transformers (and vice-versa).

In Table 1, we summarize the main theoretical results of this paper, along with their key assumptions. We emphasize that **Theorems 4.5 and 4.6 apply to the commonly used softmax and ReLU attentions.**

1.2. Related Work

Garg et al. (2022) show experimentally that Transformers can learn *simple functions* in context, including linear functions, decision trees, and two layer neural networks. Akyürek et al. (2022); Dai et al. (2022) propose that Transformers learn in-context by implementing learning algorithms. Building upon Akyürek et al. (2022), Lin et al. (2023) propose more efficient constructions for a broader range of learning algorithms. Bai et al. (2023) apply a similar technique to study the in-context reinforcement learning problem. Independent of the ICL motivation, numerous other authors have also studied the algorithmic power of transformers (Pérez et al., 2021; Wei et al., 2022; Giannou et al., 2023; Olsson et al., 2022).

Many of the above papers propose some form of *construction* (i.e., a specific parameter configuration), under which the Transformer implements the desired algorithm. It is however often unclear if the Transformer actually learns these constructions during training. Motivated by the question of *"what do Transformers actually learn,"* a line of recent work turned their attention to linear Transformers (Schlag et al., 2021; von Oswald et al., 2023a).

von Oswald et al. (2023a) devise a *simple weight construction* for the linear Transformer, which can be shown to implement gradient descent (as well as a more sophisticated algorithm known as GD++). Subsequently, Ahn et al. (2023); Zhang et al. (2023); Mahankali et al. (2023) show that for 1-layer Transformers, there exists a global minimum of the in-context loss that closely resembles the construction in (von Oswald et al., 2023a). Zhang et al. (2023) further show that training a 1-layer Transformer with gradient descent converges in polynomial time to the proposed global minimum. For multi-layer Transformers, (Ahn et al., 2023) show the local optimality of preconditioned GD and preconditioned GD++, under different parameter sparsity assumptions. We note that Assumptions 2.1 and 4.4 in this paper closely parallel the parameter sparsity assumptions in (Ahn et al., 2023). Furthermore, Theorems 4.5 and 4.6 can be viewed as generalizations of Theorems 3 and 4 of (Ahn et al., 2023) to non-linear architectures and functions. Finally, Theorem 5 of (Ahn et al., 2023) establishes the global optimality of gradient descent for *one-layer ReLU-activated Transformers*, which was shown in (Wortsman et al., 2023)

Results	$x^{(i)}$	$y^{(i)}$	$\tilde{h}$	Transformer Parameters	Basic Description
Prop. 3.1	None	None	$\tilde{h}$ a kernel	None	Transformer can implement functional GD in RKHS induced by kernel $\tilde{h}$.
Prop. 3.4	None	$\mathcal{K}$-GP (3.3)	$\tilde{h}$ matches $\mathcal{K}$	None	Transformer prediction can be Bayes optimal if $\tilde{h}$ matches $\mathcal{K}$, with sufficient layers.
Thm. 4.5	Asm. 4.1	Asm. 4.2	Asm. 4.3	Asm. 4.4 & $A_\ell = 0$	Functional GD is stationary point of in-context loss under $A_\ell = 0$ constraint.
Thm. 4.6	Asm. 4.1	Asm. 4.2	Asm. 4.3	Asm. 4.4	Characterizing stationary point of in-context loss (unconstrained).

Table 1. Summary of main theoretical results and key assumptions

to perform comparably to softmax transformers on certain tasks. More recently, (von Oswald et al., 2023b) studied the ability of linear Transformers to perform auto-regressive next-token prediction for sequential data. (Wu et al., 2023) study the statistical complexity of ICL with a 1-layer linear Transformer for linear regression. (Huang et al., 2023) uses a similar framework to study 1-layer softmax-activated Transformers, when the covariates are all orthonormal.

Distinct from the above, another relevant line of work views the attention module as a kernel operation, and proposes alternatives to standard attention based on various kernels (Tsai et al., 2019; Choromanski et al., 2020; Ali et al., 2021; Nguyen et al., 2022b;a; Chi et al., 2022). In (Wright & Gonzalez, 2021; Chen et al., 2023), authors consider the connection between Transformers and *asymmetric kernels*.

2. Setup: ICL with non-linear Transformers

Input Data for In-Context Learning

We begin by defining the in-context learning problem. We are given n demonstrations $z^{(i)} := (x^{(i)}, y^{(i)})$, for $i = 1...n$. $x^{(i)} \in \mathbb{R}^d$ are covariates and $y^{(i)} \in \mathbb{R}$ are scalar labels. We are also given a query $x^{(n+1)} \in \mathbb{R}^d$, an the goal is to predict its label $y^{(n+1)}$, which is unobserved. In general, $X := [x^{(1)} ... x^{(n+1)}] \in \mathbb{R}^{d\times(n+1)}$ have joint distribution $\mathcal{P}_X$. We assume that $Y = [y^{(1)}...y^{(n+1)}] \in \mathbb{R}^{1\times n+1}$ have joint distribution $\mathcal{P}_{Y|X}$ conditional on X. An important example of $\mathcal{P}_{Y|X}$ is when $y^{(i)} = \phi(x^{(i)})$ for some unknown function ϕ. For instance, $\phi(x) = \langle\theta, x\rangle$ gives rise to linear regression. We will discuss specific choices of $\mathcal{P}_X$ and $\mathcal{P}_{Y|X}$ in Sections 3.2 and 4.1. Thus the input of the in-context learning problem is given by

$$Z_0 = \begin{bmatrix} z^{(1)} & z^{(2)} & \cdots & z^{(n)} & z^{(n+1)} \end{bmatrix} \quad (1)$$
$$= \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \cdots & y^{(n)} & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)}.$$

Transformers with general non-linear attention

We define the *generalized attention module* as

$$\mathrm{Attn}^{\tilde{h}}_{V,B,C}(Z) := VZM\tilde{h}\left(BX, CX\right), \quad (2)$$

$V \in \mathbb{R}^{(d+1)\times(d+1)}, B \in \mathbb{R}^{d\times d}, C \in \mathbb{R}^{d\times d}$ are the value, query, and key parameter matrices respectively, and they parameterize the attention. $M := \begin{bmatrix} I_{n\times n} & 0 \\ 0 & 0 \end{bmatrix}$ is a mask matrix, and $\tilde{h} : \mathbb{R}^{d\times(n+1)} \times \mathbb{R}^{d\times(n+1)} \to \mathbb{R}^{(n+1)\times(n+1)}$ denotes a matrix-valued function. The matrix X is shorthand for $[x^{(1)} ... x^{(n+1)}] \in \mathbb{R}^{d\times(n+1)}$, the first d rows of Z.

Note. The Attention definition in (2) differs from standard attention, in that X should be replaced by Z. We discuss in Section 2.1 how the common attention modules maps to (2) under a sparsity assumption on the Query, Key matrices (Assumption 2.1).

We construct a k-layer Transformer by stacking k layers of the attention module (with residual). To be precise, let Z_ℓ denote the output of the $(\ell-1)^{th}$ layer of the Transformer. Then $Z_{\ell+1} := Z_\ell + \mathrm{Attn}^{\tilde{h}}_{V_\ell,B_\ell,C_\ell}(Z_\ell)$, or equivalently:

$$Z_{\ell+1} = Z_\ell + V_\ell Z_\ell M\tilde{h}\left(B_\ell X_\ell, C_\ell X_\ell\right) \quad (3)$$

where V_ℓ, B_ℓ, C_ℓ are the value, query and key matrices of the attention module at layer ℓ. $\tilde{h} : \mathbb{R}^{d\times(n+1)} \times \mathbb{R}^{d\times(n+1)} \to \mathbb{R}^{(n+1)\times(n+1)}$ denotes a non-linear activation function.

Consider a k layer Transformer. For the rest of the paper, we let $V := \{V_\ell\}_{\ell=0...k}$, $B := \{B_\ell\}_{\ell=0...k}$, $C := \{C_\ell\}_{\ell=0...k}$ denote collections of the attention parameters across layers. For $\ell = 0 ... k+1$, define

$$\mathsf{TF}_\ell(x; (V, B, C)|z^{(1)} ... z^{(n)}) := [Z_\ell]_{(d+1),(n+1)}, \quad (4)$$

where Z_i evolves as (3), initialized at $Z_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(n)} & x \\ y^{(1)} & y^{(2)} & \cdots & y^{(n)} & 0 \end{bmatrix}$.

We interpret $\mathsf{TF}_\ell(x; (V, B, C)|z^{(1)} ... z^{(n)})$ as *"The predictor for $-y^{(n+1)}$ at layer ℓ, given $x^{(n+1)} = x$, conditioned on demonstrations $z^{(1)} ... z^{(n)}$, parameterized by weight matrices V, B, C"*. This definition is consistent with the setup in (von Oswald et al., 2023a; Ahn et al., 2023).

The In-Context Loss

Given the input Z_0 and the Transformer parameterized by V, B, C, we define the in-context loss as

$$f(V,B,C) \tag{5}$$
$$= \mathbb{E}\left[\left(\mathsf{TF}_{k+1}(x^{(n+1)}; (V,B,C)|z^{(1)} \dots z^{(n)}) + y^{(n+1)}\right)^2\right]$$
$$:= \mathbb{E}\left[\left([Z_{k+1}]_{(d+1),(n+1)} + y^{(n+1)}\right)^2\right],$$

where expectation is taken over Z_0 and $y^{(n+1)}$.

2.1. Examples of Attention Modules

To motivate definition (2), we show in Examples 1, 2, 3 below how the most common variants of the attention module can be realized via specific choices of non-linearity $\tilde{h}$, assuming the following sparsity constraints:

Assumption 2.1 (QK last column and row sparsity)**.** For $Q, K \in \mathbb{R}^{(d+1)\times(d+1)}$, there exist $B, C \in \mathbb{R}^{d\times d}$ such that $Q = \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix}$ and $K = \begin{bmatrix} C & 0 \\ 0 & 0 \end{bmatrix}$.

In words, Assumption 2.1 restricts the last row/column of Q, K to be 0; this restriction was considered in (von Oswald et al., 2023a; Ahn et al., 2023). and is naturally satisfied by the global minimum of 1-layer Linear Transformers in (Ahn et al., 2023; Mahankali et al., 2023; Zhang et al., 2023).

Example 1 (Linear Transformer). The linear attention module of (von Oswald et al., 2023a), with parameters V, Q, K, is given by

$$\mathrm{Attn}^{linear}_{V,Q,K}(Z) := VZMZ^\top Q^\top KZ.$$

Assume Q, K, B, C satisfy Assumption 2.1. By choosing $\tilde{h}(U,W) := U^\top W$, $\mathrm{Attn}^{\tilde{h}}_{VBC}$ from (2) equals $\mathrm{Attn}^{linear}_{VQK}$.

Example 2 (ReLU Transformer). The ReLU attention module, with parameters V, Q, K, is given by

$$\mathrm{Attn}^{relu}_{V,Q,K}(Z) := VZM\mathsf{relu}\left(Z^\top Q^\top KZ\right).$$

In the above, relu denotes element-wise ReLU function. Assume Q, K, B, C satisfy Assumption 2.1. By choosing $\tilde{h}(U,W) = \mathsf{relu}\left(U^\top W\right)$, $\mathrm{Attn}^{\tilde{h}}_{VBC}$ from (2) equals $\mathrm{Attn}^{relu}_{VQK}$.

Example 3 (Softmax Transformer). The softmax attention module, with parameters V, Q, K, is given by

$$\mathrm{Attn}^{softmax}_{V,Q,K}(Z) := VZ\mathsf{softmax}\left(Z^\top Q^\top KZ\right).$$

In the above, $\mathsf{softmax} : \mathbb{R}^{(n+1)\to(n+1)}$ is the masked softmax function:

$$[\mathsf{softmax}(W)]_{ij} = \begin{cases} \frac{\exp(W_{ij})}{\sum_{k=1}^n \exp(W_{kj})} & \text{for } i \neq n+1 \\ 0 & \text{for } i = n+1 \end{cases}.$$

Let Q, K, B, C satisfy Assumption 2.1. Let us define

$$\left[\tilde{h}(U,W)\right]_{ij} := \begin{cases} \frac{\exp\left([U^\top W]_{ij}\right)}{\sum_{k=1}^n \exp\left([U^\top W]_{kj}\right)} & \text{for } i \neq n+1 \\ 0 & \text{for } i = n+1 \end{cases}. \tag{6}$$

Then $\mathrm{Attn}^{\tilde{h}}_{VBC}$ from (2) equals $\mathrm{Attn}^{softmax}_{VQK}$. Note that the mask matrix M from (2) is unnecessary here as $M\tilde{h}(\cdot) = \tilde{h}(\cdot)$.

3. Transformers *can* implement gradient descent in function space.

In this section, we show that under a choice of V, B, C and $\tilde{h}$, the forward pass of the Transformer defined in (3) can implement *Kernel Regression* for a kernel $\mathcal{K}$. We present the necessary background on RKHS in Section G.

We begin by defining *"gradient descent in function space."* Let $\mathbb{H}$ denote a Hilbert space of functions mapping from $\mathbb{R}^d \to \mathbb{R}$, equipped with the metric $\|\cdot\|_{\mathbb{H}}$. Let $L(f) : \mathbb{H} \to \mathbb{R}$ denote some loss. The gradient descent of $L(f)$ with respect to $\|\cdot\|_{\mathbb{H}}$ is defined as the sequence

$$f_{\ell+1} = f_\ell - r_\ell \nabla L(f_\ell), \tag{7}$$

where $\nabla L(f) := \arg\min_{\|g\|_{\mathbb{H}}=1} \frac{d}{dt} L(f+tg)\Big|_{t=0}$, and r_ℓ is a sequence of stepsizes.

3.1. Transformers *can* implement gradient descent in function space.

The first main result of this section is Proposition 3.1 below, which shows that a Transformer can implement the functional descent sequence (7). We highlight that Proposition 3.1 **works for any kernel** $\mathcal{K}$ – as long as the choice of $\tilde{h}$ coincides with $\mathcal{K}$. In Sections 3.1.1 and 3.1.2, we motivate Proposition 3.1 using specific examples.

Proposition 3.1. *Let $\mathcal{K}$ be an arbitrary kernel. Let $\mathbb{H}$ denote the Reproducing Kernel Hilbert space induced by $\mathcal{K}$. Let $z^{(i)} = (x^{(i)}, y^{(i)})$ for $i = 1 \dots n$ be an arbitrary set of in-context examples. Denote the empirical loss functional by $L(f) := \sum_{i=1}^n \left(f(x^{(i)}) - y^{(i)}\right)^2$. Let $f_0 = 0$ and let f_ℓ denote the gradient descent sequence of L wrt $\|\cdot\|_{\mathbb{H}}$, as defined in (7). Then there exist scalars stepsizes $r'_0 \dots r'_k$ such that the following holds:*

Let $\tilde{h}$ be the function defined as $\left[\tilde{h}(U,W)\right]_{i,j} :=$

$\mathcal{K}\left(U^{(i)}, W^{(j)}\right)$, *where* $U^{(i)}$ *and* $W^{(i)}$ *denote the* i^{th} *column of* U *and* W *respectively. Let* $V_\ell = \begin{bmatrix} 0 & 0 \\ 0 & -r'_\ell \end{bmatrix}$, $B_\ell = I_{d\times d}$, $C_\ell = I_{d\times d}$. *Then for any* $x := x^{(n+1)}$, *the Transformer's prediction for* $y^{(n+1)}$ *at each layer* ℓ *matches the prediction of the functional gradient sequence* (7) *at step* ℓ, *i.e. for all* $\ell = 0 \dots k$,

$$\mathsf{TF}_\ell(x; (V, B, C)|z^{(1)} \dots z^{(n)}) = -f_\ell(x). \tag{8}$$

We defer the proof of Proposition 3.1 to Appendix B. As we verify in step (24) in the proof, the functional gradient descent sequence (7) is equivalent to

$$f_{\ell+1}(\cdot) = f_\ell(\cdot) + r'_\ell \sum_{i=1}^{n} \left(y^{(i)} - f_\ell(x^{(i)})\right) \mathcal{K}(\cdot, x^{(i)}) \tag{9}$$

for some stepsizes r'_ℓ.

In Theorem 4.5, we show that above choices of V_ℓ, B_ℓ, C_ℓ form a stationary point of the Transformer training objective. In Appendix H.2, we empirically verify that these parameter choices are consistently learned in experiments. Below, we show how Proposition 3.1 applies to two common settings.

3.1.1. Case Study: Linear Kernel

Consider the simplest setting of the **Euclidean inner product Kernel**, i.e. $\mathcal{K}^{linear}(u, w) := \langle u, w\rangle$. In this setting, the choices of key, value, query matrices in Proposition 3.1 essentially match the constructions in (von Oswald et al., 2023a; Ahn et al., 2023). It is also worth noting that *functional gradient descent in the RKHS induced by* $\mathcal{K}^{linear}$ in fact follows the same trajectory as *(Euclidean) gradient descent of the linear regression parameter*; we provide a short proof in Appendix I.2.

3.1.2. Case Study: Exponential Kernel, and connection to softmax activation

We will now show that the construction in Proposition 3.1, when $\mathcal{K}$ is the exponential kernel, bears remarkable similarity to the softmax Transformer with identity weights.

Let $\mathcal{K}$ denote the exponential kernel with bandwidth σ, i.e. $\mathcal{K}(x, x') := \exp\left(\frac{1}{\sigma^2}\langle x, x'\rangle\right)$. Choosing $\left[\tilde{h}^{exp}(U, W)\right]_{ij} := \mathcal{K}(U_i, W_j)$ (where U_i is the i^{th} column of U), and choose Transformer parameters $V_\ell = \begin{bmatrix} 0 & 0 \\ 0 & -r_\ell \end{bmatrix}, B_\ell = \frac{1}{\sigma} I_{d\times d}, C_\ell = \frac{1}{\sigma} I_{d\times d}$.

Recall that $f_\ell(x^{(n+1)})$ denotes the Transformer's prediction for $y^{(n+1)}$ at layer ℓ. The $\tilde{h}^{exp}$ Transformer's prediction at each layer follows the functional gradient descent sequence from (9).

On the other hand, recall from Example 3 the standard $\tilde{h}^{softmax}$ Transformer with $\tilde{h}^{softmax}$ defined in (6). By choosing parameters $V_\ell = \begin{bmatrix} 0 & 0 \\ 0 & -r_\ell \end{bmatrix}, Q_\ell = \begin{bmatrix} \frac{1}{\sigma} I_{d\times d} & 0 \\ 0 & 0 \end{bmatrix}, K_\ell = \begin{bmatrix} \frac{1}{\sigma} I_{d\times d} & 0 \\ 0 & 0 \end{bmatrix}$, the softmax-activated Transformer implements the update

$$f_{\ell+1}(\cdot) = f_\ell(\cdot) + r'_\ell \tau(\cdot) \sum_{i=1}^{n} \left(y^{(i)} - f_\ell(x^{(i)})\right) \mathcal{K}(\cdot, x^{(i)}), \tag{10}$$

where $\tau(\cdot) = 1/\sum_{j=1}^{n} \mathcal{K}\left(\cdot, x^{(j)}\right)$ is the normalization factor in softmax. Comparing (10) to (9), we see that the algorithms implemented by $\tilde{h}^{exp}$ and $\tilde{h}^{softmax}$ Transformers are very similar, with the only difference being the normalization factor τ.

Remark 3.2. When $\|x^{(i)}\|_2 = 1$ for all $i = 1 \dots n+1$, the exponential kernel is up to scaling equal to the RBF kernel.

3.2. Optimality of $\tilde{h}$ for matching $\mathcal{K}$.

We will show in Proposition 3.4 below, that the functional gradient descent algorithm (7), which is implemented by the Transformer in Proposition 3.1, can in fact lead to a *nearly statistically optimal* prediction when the non-linear activation $\tilde{h}$ matches the data distribution. We begin by defining a general class of data distributions. Let $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ denote a symmetric function. We define a conditional distribution for $Y|X$ as follows:

Definition 3.3 ($\mathcal{K}$ Gaussian Process)**.** Given symmetric $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we define the $\mathcal{K}$ Gaussian Process as the conditional distribution

$$Y|X \sim \mathcal{N}(0, \mathbb{K}_+(X)),$$

where $Y = [y^{(1)} \dots y^{(n+1)}]$, $X = [x^{(1)} \dots x^{(n+1)}]$, and $\mathbb{K}_{ij}(X) := \mathcal{K}\left(x^{(i)}, x^{(j)}\right)$. Let $UDU^\top$ be the Eigenvalue decomposition of $\mathbb{K}(X)$. $\mathbb{K}_+(X) := U|D|U^\top$, where $|D|_{ii} := |D_{ii}|$ is entry-wise the absolute value of D.

Definition 3.3 generalizes the notion of a Gaussian process, to when the "metric" is given by the function $\mathcal{K}$. In Example 7 from Section 4.1, we discuss a few concrete examples of $\mathcal{K}$ Gaussian Processes. Note that Definition 3.3 ***does not assume that*** $\mathcal{K}$ is a kernel (specifically, $\mathcal{K}$ may not be PSD). However, if $\mathcal{K}$ is a kernel, then $\mathbb{K}$ is always positive semidefinite, so that $\mathbb{K}_+ = \mathbb{K}$.

In the following result, we see that when the data labels are generated by a $\mathcal{K}$-Gaussian Process for some kernel $\mathcal{K}$, then the Transformer prediction (4), for the construction from Proposition 3.1, is statistically optimal if $\tilde{h}$ matches $\mathcal{K}$:

Proposition 3.4. *Let* $X = [x^{(1)} \dots x^{(n+1)}]$, $Y = [y^{(1)} \dots y^{(n+1)}]$. *Let* $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel. *Assume that* $Y|X$ *is drawn from the* $\mathcal{K}$ *Gaussian Process. Let the attention activation* $[\tilde{h}(U, W)]_{ij} := \mathcal{K}(U_i, W_j)$. *Consider*

*the functional gradient descent construction in Proposition 3.1. As the layer number $\ell \to \infty$, the Transformer's prediction for $y^{(n+1)}$ at layer ℓ (4) approaches the **Bayes (optimal) estimator** that minimizes the in-context loss (5).*

We defer the proof of Proposition 3.4 to Appendix C. We note a few caveats of Proposition 3.4:

1. The result guarantees optimality of $\tilde{h} = \mathcal{K}$ in the limit of $\ell \to \infty$. For finite ℓ, there may exist a better choice of $\tilde{h}$ that implements a more iteration-efficient algorithm. For example, see discussion of Figure 2(b).
2. The construction in Proposition 3.1 sets the top-left $d \times d$ block of V_ℓ to 0 (i.e. $A_\ell = 0$). In practice, as see in Theorem 4.6 and Figure 6, A_ℓ is often a non-zero multiple of $I_{d\times d}$. For this reason, a choice of $\tilde{h}$ that differs from $\mathcal{K}$ may nonetheless recover the Bayes estimator.

3.3. Experiments for Proposition 3.4

To experimentally verify Proposition 3.4, we compare the performance of different choices of $\tilde{h}$ against different choices of generating kernel $\mathcal{K}$. We present our findings in Figures 1 and 2.

We consider three types of $\mathcal{K}$ Gaussian Processes:

$$\mathcal{K}^{linear}(u,w) := \langle u, w\rangle, \quad \mathcal{K}^{relu}(u,w) := \mathsf{relu}(\langle u, w\rangle),$$
$$\mathcal{K}^{exp}(u,w) := \exp\left(\langle x, y\rangle\right). \tag{11}$$

For each choice of $\mathcal{K}$ above, we try Transformers with four different types of non-linear attention module $\tilde{h}$.

$$[\tilde{h}^{linear}(U,W)]_{ij} := [U^\top W]_{ij},$$
$$[\tilde{h}^{relu}(U,W)]_{ij} := \mathsf{relu}([U^\top W]_{ij}),$$
$$[\tilde{h}^{exp}(U,W)]_{ij} := \exp([U^\top W]_{ij}),$$
$$[\tilde{h}^{softmax}(U,W)]_{ij} := \begin{cases} \frac{\exp\left([U^\top W]_{ij}\right)}{\sum_{k=1}^n \exp\left([U^\top W]_{kj}\right)} & \text{for } i \neq n+1 \\ 0 & \text{for } i = n+1 \end{cases}. \tag{12}$$

Note: Proposition 3.1 does not apply to $\tilde{h}^{\mathrm{relu}}$ and $\tilde{h}^{\mathrm{softmax}}$ as they are not kernels. But we include them in our experiments since they are widely used in practice.

The covariates $x^{(i)}$ are drawn iid from the unit sphere, and the labels $y^{(i)}$ are drawn from one of the three $\mathcal{K}$-Gaussian Processes. In all plots, the loss values are taken after convergence of training loss. Full experiment details are found in Appendix H.1.

In all our Figures, we will show the loss of the **Bayes Estimator** f^{bayes} as a baseline. This represents the information-theoretically optimal loss. Recall from (3.3) that $Y|X \sim \mathcal{N}(0, \mathbb{K}_+(X))$. Let $\mathbb{K}$ and $\mathbb{K}_+$ be as defined in Definition 3.3. Let $\left\{\hat{\mathbb{K}}, \nu, \mu\right\} \in \left\{\mathbb{R}^{d\times d}, \mathbb{R}^d, \mathbb{R}\right\}$ be **defined** as $\begin{bmatrix} \hat{\mathbb{K}} & \nu \\ \nu^\top & \mu \end{bmatrix} := \mathbb{K}_+$. Let $\hat{Y} \in \mathbb{R}^n$ denote the vector of $y^{(1)} \ldots y^{(n)}$. Then the Bayes Estimator is defined as

$$f^{bayes}(x^{(n+1)}) := \nu^\top \hat{\mathbb{K}}^{-1} \hat{Y}. \tag{13}$$

When $\mathcal{K}$ is a PSD kernel, $\mathbb{K}_+ := \mathbb{K}$, and (13) is identical to the Bayes estimator that we derive in (25) in the proof of Proposition 3.4. When $\mathcal{K}$ is not PSD, (13) and (25) are **not equal**, but (13) is nonetheless a well-defined estimator.

Figure 1 plots the in-context loss of a 3-layer Transformer against **number of demonstrations** $n \in \{2, 4, 6, 8, 10, 12\}$, for different combinations of label-generating kernel $\mathcal{K}$ and attention module $\tilde{h}$ (see (12)). Figure 2 plots the in-context loss against **number of layers** $L \in \{1, 2, 3, 4, 5, 6, 7, 8\}$, for $n \in \{14, 6\}$. We show the losses of different combinations of $\mathcal{K}$ and $\tilde{h}$. We summarize key observations of interest below:

1. **Ignoring $\tilde{h}^{softmax}$, the best prediction error is achieved when the attention activation $\tilde{h}$ matches distribution $\mathcal{K}$.** From Figures 1(a), 1(b), 2(a), 2(c), 2(d): the best accuracy is obtained when the attention activation $\tilde{h}$ matches $\mathcal{K}$, as suggested by Proposition 3.4.
2. **The $\tilde{h}^{softmax}$ attention is most accurate for $\mathcal{K}^{exp}$ labels when number of layers (L) is small and context length (n) is large.** From Figure 1(c): for $L = 3$, $\tilde{h}^{softmax}$ is more accurate than $\tilde{h}^{exp}$ on $\mathcal{K}^{exp}$ data for $n \in \{6, 8, 10, 12, 14\}$. From Figure 2(b), when $n = 14$, the gap between $\tilde{h}^{softmax}$ and $\tilde{h}^{exp}$ (for $\mathcal{K}^{exp}$ data) becomes very small for $L \geq 6$. From Figure 2(d), when $n = 6$, $\tilde{h}^{exp}$ is most accurate for $\mathcal{K}^{exp}$ data when $L \geq 5$.
   (a) We conjecture that $\tilde{h}^{softmax}$ implements an algorithm that is *more iteration efficient* but *less statistically efficient* than functional gradient descent. As each layer implements a step of some algorithm, $\tilde{h}^{softmax}$ performs well with few layers (steps), and performs relatively poorly when number of samples is small.
   (b) The relative performance of $\tilde{h}^{softmax}$ and $\tilde{h}^{exp}$ is consistent with Proposition 3.4, which predicts that the $\tilde{h}^{\exp}$ Transformer approaches Bayes-optimal prediction loss **as number of layers increases.** We also note that $\tilde{h}^{softmax}$ is closely related to $\mathcal{K}^{exp}$, as discussed at the end of Example 3.1.2.
3. For each Transformer, the **parameters learned are as predicted in Theorem 4.6**. See experiments in Section H.3 for details.

Finally, we also note that the gap between $\tilde{h}^{relu}$ and the Bayes estimator in Figure 2(a) is quite significant, and does

not seem to decrease with number of layers. This is likely because the Bayes estimator f^{bayes} in (13) uses $\hat{\mathbb{K}}$ which involves flipping eigenvalues on $\mathbb{K}$. Such an operation may not be easily implementable by single head Transformers.

3.4. Composing multiple attention heads

A powerful aspect of RKHS theory is the ability to form complex kernels by composing simple ones via addition and multiplication. Using this idea, we show, both theoretically (Proposition D.1) and empirically (Figure 4), that **multi-head** Transformers with **different activations per-head** can attain much greater representation power; specifically, they can attain optimal prediction loss for a large class of $\mathcal{K}$ Gaussian Processes which are obtained from kernel composition. Due to space constraints, we refer the readers to Appendix D for the detailed discussion.

4. Optimization Landscape Results

In the previous section, we saw that Transformers *can* implement functional gradient descent in its forward pass, and that this implementation can be nearly statistical optimal. However, *does the Transformer learn to implement functional gradient descent when training converges?* To answer this question, we analyze the optimization landscape of the in-context loss, for the Transformer defined in (3).

In Theorem 4.5, we show that the functional gradient descent construction of Proposition 3.1 is a stationary point of the in-context loss when we constrain the top left block of the Value matrix to 0. **In Theorem 4.6**, we characterize stationary points of the in-context loss for general Value matrices. The stationary point implements a sophisticated algorithm that interleaves functional gradient descent steps with transformations of the covariates.

We provide experimental verification of both Theorem 4.5 and 4.6 in Sections H.2 and H.3. We present key assumptions in Sections 4.1 and 4.2. We note that both Theorem 4.5 and 4.6 apply to softmax and ReLU Transformers.

4.1. Distributional Assumptions

We will first state two assumptions on the distribution of covariates X and labels $Y|X$. We motivate these assumptions with Examples 4-8. .

Recall the setup from Section 2. The input is $Z_0 \in \mathbb{R}^{(d+1)\times(n+1)}$. Let $X = \left[x^{(1)} \cdots x^{(n+1)}\right] \in \mathbb{R}^{d\times(n+1)}$ denote the first d rows of Z_0. Let $Y = \left[y^{(1)} \cdots y^{(n+1)}\right] \in \mathbb{R}^{1\times(n+1)}$ denote the row vector of labels $y^{(i)}$'s. Note that the last row of Z_0 has $y^{(n+1)}$ replaced by 0, and thus differs from Y. We will make an assumption each on the distributions of X and Y respectively:

Assumption 4.1 (X distribution assumption)**.** Let $\mathcal{P}_X$ denote the distribution of X, i.e. $\mathcal{P}_X$ is the joint distribution over $x^{(1)} \ldots x^{(n+1)}$. Furthermore, assume that there is a symmetric invertible matrix $\Sigma \in \mathbb{R}^{d\times d}$ such that for any orthogonal matrix U, $\Sigma^{1/2}U\Sigma^{-1/2}X \overset{d}{=} X$.

In Examples 4 and 5 below, we provide two common distributions for $x^{(i)}$ which satisfy Assumption 4.1.

Example 4 ($x^{(i)}$ drawn from rotationally invariant distributions). Assumption 4.1 is satisfied when $x^{(i)} \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$, or when $x^{(i)} = \Sigma^{1/2}\xi^{(i)}$, for ξ drawn uniformly from the unit sphere. This distribution of $x^{(i)}$ has been considered in (Garg et al., 2022; Akyürek et al., 2022; von Oswald et al., 2023a; Ahn et al., 2023; Zhang et al., 2023; Mahankali et al., 2023).

Example 5 ($x^{(i)}$ drawn from Gaussian Mixture Models). More generally, Assumption 4.1 can be satisfied even when $x^{(i)}$ are not iid. Let $\mu \sim \mathcal{N}(0, I)$, and let $x^{(i)} = \mu + \xi^{(i)}$, where $\xi^{(i)} \overset{iid}{\sim} \mathcal{N}(0, I)$. This example can be further generalized to contain two or more cluster means μ_1, μ_2 sampled independently (i.e. mixture of Gaussians).

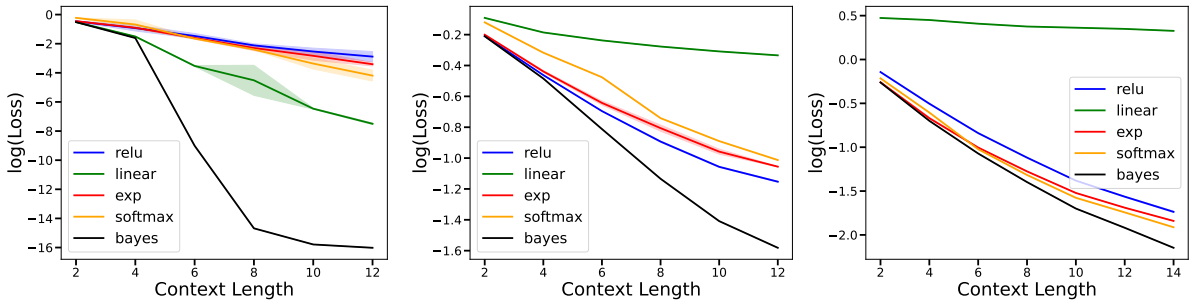
Assumption 4.2 ($Y|X$ distribution assumption)**.** Conditional on $X = \left[x^{(1)} \ldots x^{(n+1)}\right]$, $Y = \left[y^{(1)} \ldots y^{(n+1)}\right] \in \mathbb{R}^{(n+1)}$ has covariance matrix $\mathbb{E}_{Y|X}\left[Y^\top Y\right] =: \mathbb{K}(X)$, where $\mathbb{K}(X) : \mathbb{R}^{d\times(n+1)} \to \mathbb{R}^{(n+1)\times(n+1)}$. Assume that for all orthogonal matrix $U \in \mathbb{R}^{d\times d}$, $\mathbb{K}(\Sigma^{1/2}U\Sigma^{-1/2}X) = \mathbb{K}(X)$, where Σ is the same matrix from Assumption 4.1.

In Examples 6, 7, 8 below, we will discuss a few common label distributions which satisfy Assumptions 4.2. **Example 7 is of particular interest**, as it is quite general, and is the setting for all the experiments presented in Figures 1, 2, 3, 5, 6. Note that Example 6 is a special case of Example 7.

Example 6 ($y^{(i)}$ are linear functions of $x^{(i)}$). One example of Assumption 4.2 is when $\theta \sim \mathcal{N}(0, I)$, $y^{(i)} = \langle \theta, \xi^{(i)} \rangle$, and $x^{(i)} = \Sigma^{1/2}\xi^{(i)}$. We can verify that the covariance matrix $\mathbb{K}(X_0) := \mathbb{E}\left[Y^\top Y\right] = X^\top\Sigma^{-1/2}\,\mathbb{E}\left[\theta\theta^T\right]\Sigma^{-1/2}X = X^\top\Sigma^{-1}X = \mathbb{K}(\Sigma^{1/2}U\Sigma^{-1/2}X)$. This setting was considered in (Ahn et al., 2023; Mahankali et al., 2023).

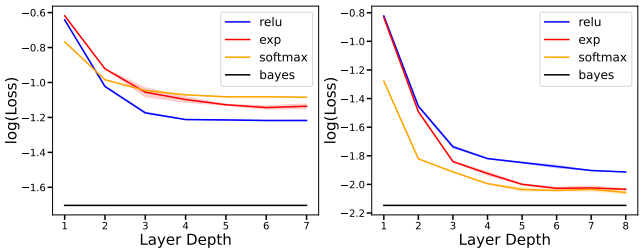
Example 7 (Rotationally Symmetric $\mathcal{K}$ Gaussian Process). Recall the $\mathcal{K}$ Gaussian Process from Definition 3.3. Under this definition, recall that $Y^\top|X \sim \mathcal{N}(0, \mathbb{K}_+(X))$, where $[\mathbb{K}(X)]_{ij} := \mathcal{K}\left(\Sigma^{-1/2}x^{(i)}, \Sigma^{-1/2}x^{(j)}\right)$, and $\mathbb{K}_+(X)$ takes an absolute value on the eigenvalues of $\mathbb{K}(X)$. We verify that Assumption 4.2 holds if, for all orthogonal matrix U, $\mathcal{K}(v, w) = \mathcal{K}(Uv, Uw)$. This is satisfied by the following common choices of $\mathcal{K}$:

$$\mathcal{K}^{linear}(u, w) := \langle u, w \rangle, \quad \mathcal{K}^{relu}(u, w) := \mathsf{relu}(\langle u, w \rangle),$$
$$\mathcal{K}^{exp}_\sigma(u, w) := \exp\left(\langle x, y \rangle / \sigma^2\right). \tag{14}$$

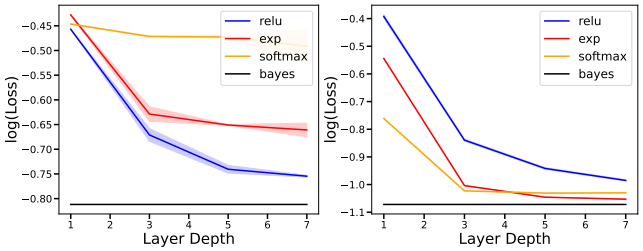


1(a) $\mathcal{K}^{linear}$ 1(b) $\mathcal{K}^{relu}$ 1(c) $\mathcal{K}^{exp}$

Figure 1. Plot of log(test ICL loss) against number of in-context demonstrations. The labels are generated using a $\mathcal{K}$ Gaussian Process (Definition 3.3) Each sub-figure corresponds to one of three choices of $\mathcal{K}$, defined in (11). Each sub-figure contains 4 plots corresponding to 4 choices of $\tilde{h}$, as defined in (12). Black line denotes Bayes Loss.



2(a) $\mathcal{K}^{relu}, n = 14$ 2(b) $\mathcal{K}^{exp}, n = 14$



2(c) $\mathcal{K}^{relu}, n = 6$ 2(d) $\mathcal{K}^{exp}, n = 6$

Figure 2. Plot of log(test ICL loss) against number of layers. The labels are generated using a $\mathcal{K}$ Gaussian Process (Definition 3.3), for $\mathcal{K}^{relu}$ and $\mathcal{K}^{exp}$ as defined in (11). Each sub-figure contains 3 plots corresponding to three choices of $\tilde{h}$, as defined in (12).

Example 8 (Two-layer ReLU network.). Finally, we provide an example where Assumption 4.2 holds for a *random kernel*. Consider the random two-layer ReLU classification function described in (Garg et al., 2022):

$$y^{(i)} = \left\langle \theta_2, \mathsf{relu}\left(\theta_1 x^{(i)}\right)\right\rangle, \tag{15}$$

where $\theta_1 \in \mathbb{R}^{d\times m}, \theta_2 \in \mathbb{R}^m$; assume θ_1, θ_2 are sampled coordinate-wise, independently, from $\mathcal{N}(0,1)$. Thus m is the dimension of the hidden layer. We verify that this satisfies Assumption 4.2 in Appendix I.1.

4.2. Architectural Assumptions

For the rest of this section, we will assume that the non-linear map $\tilde{h}(U,V)$ satisfies the following invariance:

Assumption 4.3. For any $W, V \in \mathbb{R}^{d\times(n+1)}$ and for any matrix $S \in \mathbb{R}^{d\times d}$ with inverse S^{-1}, the function $\tilde{h}(\cdot,\cdot)$ satisfies $\tilde{h}(W,V) = \tilde{h}(S^\top W, S^{-1}V)$.

We verify that the three examples of $\tilde{h}$ from Examples {1, 2, 3} which implement {Linear, ReLU, Softmax}-activated Transformers, all satisfy Assumption 4.3. We also assume that V_ℓ has the following sparsity pattern for $\ell = 0 \dots k$:

Assumption 4.4. For $\ell = 0 \dots k$, the value matrices V_ℓ which parameterize the Transformer layers in (3) satisfy $V_\ell = \begin{bmatrix} A_\ell & 0 \\ 0 & r_\ell \end{bmatrix}$ for some $A_i \in \mathbb{R}^{d\times d}, r_i \in \mathbb{R}$.

The same sparsity pattern was considered in (Ahn et al., 2023) in studying multi-layer linear Transformers.

4.3. Theorem 4.5: Functional gradient descent is a stationary point of (constrained) in-context loss.

We first study the stationary points of the optimization problem, under the constraint that $A_\ell = 0$ in Assumption 4.4. This setting is interesting because of its connection to the *functional gradient descent* construction in Proposition 3.1.

Theorem 4.5 (Informal Statement of Theorem E.1)**.** *Let $\tilde{h}$ satisfy Assumption 4.3, Let $\left(x^{(i)}, y^{(i)}\right)_{i=1\dots n+1}$ have distributions satisfying Assumptions 4.1 and 4.2. Consider the optimization problem $\min_{V,B,C} f(V,B,C)$, for the in-context loss f defined in (5), under the constraint that $V = \{V_\ell\}_{\ell=0\dots k}$ satisfies Assumption 4.4.* ***Additionally constrain $A_\ell = 0$ for $\ell = 0 \dots k$.*** *Then there exist stationary points of the constrained optimization problem where, for all $\ell = 0 \dots k$,*

$$B_\ell = b_\ell \Sigma^{-1/2} \qquad C_\ell = c_\ell \Sigma^{-1/2}, \tag{16}$$

where $b_\ell, c_\ell \in \mathbb{R}$.

The formal version of Theorem 4.5 is stated as Theorem E.1 in Appendix E; its proof is in Appendix E.1. We highlight that the proposed stationary point exactly implements the functional gradient descent construction of Proposition 3.1.

In the simplest case that $\Sigma = I$, we verify that (16) is, **up to**

scaling, identical to the construction in Proposition 3.1.

More generally, when Σ is not identity, but $[\tilde{h}(U,W)]_{ij} = \mathcal{K}(U^{(i)}, W^{(j)})$ for some kernel $\mathcal{K}$ (see Examples 1, 2), (16) implement functional descent with respect to the RKHS induced by $\tilde{\mathcal{K}}(u,w) := \mathcal{K}\left(\Sigma^{-1/2}u, \Sigma^{-1/2}w\right)$. One can view the kernel $\tilde{\mathcal{K}}$ as a rescaled version of $\mathcal{K}$.

Finally, in the case when $\tilde{h}$ **does not coincide with a kernel**, (16) implements the following algorithm:

$$f_{\ell+1}(x^{(n+1)}) = f_\ell(x^{(n+1)}) \qquad (17)$$
$$+ r'_\ell \sum_{i=1}^{n} \left(y^{(i)} - f_\ell(x^{(i)})\right) \left[\tilde{h}\left(\Sigma^{-1/2}X_0, \Sigma^{-1/2}X_0\right)\right]_{i,(n+1)}$$

where $f_{\ell+1}(x^{(n+1)})$ is "the Transformer's prediction for $y^{(n+1)}$ at layer ℓ", for $\ell = 0 \ldots k+1$. It is instructive to compare (17) with (9).

4.4. Theorem 4.6: Characterizing the stationary points of unconstrained in-context loss.

We now study stationary points of the optimization problem under Assumption 4.4 (with arbitrary A_ℓ's). This setting is considerably more general than the setting of Theorem 4.5.

Theorem 4.6 (Informal Statement of Theorem F.1)**.** *Consider the same setup as Theorem 4.5,* ***except do not constrain*** $A_\ell = 0$*. Then there exist stationary points of the constrained optimization problem where, for all* $\ell = 0 \ldots k$*,*

$$A_\ell = a_\ell I, \qquad B_\ell = b_\ell \Sigma^{-1/2}, \qquad C_\ell = c_\ell \Sigma^{-1/2}, \quad (18)$$

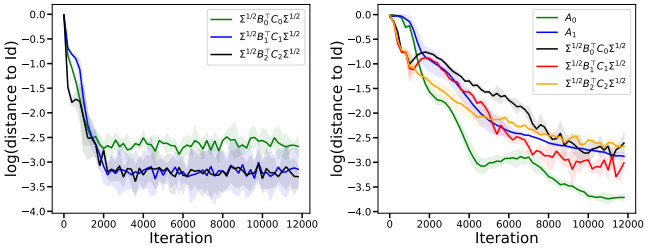
where $a_\ell, b_\ell, c_\ell \in \mathbb{R}$*.*

The formal version of Theorem 4.6 is stated as Theorem F.1 in Appendix F; the proof can be found in Appendix F.1. Unlike in Theorem 4.5, the matrix A_ℓ is not 0; thus the covariates $x^{(\ell)}$ are transformed each layer. If $\tilde{h}$ matches a kernel $\mathcal{K}$, we can verify (using same steps as Proposition 3.1) that the Transformer dynamics 3 implements an algorithm that interleaves functional gradient descent steps with transformations of the covariates. We refer the readers to (58) in the proof of Theorem F.1 for an explicit description of the algorithm implemented by (18).

We leave analysis of this algorithm as future work, but we note that in the special case of $\tilde{h}^{linear}$ (see Example 1), Theorem 4.6 implies Theorem 4 of (Ahn et al., 2023), which is similar to the GD++ construction of (von Oswald et al., 2023a). In the linear case, each covariate transformation step can be shown to improve the condition number of the optimization problem.

4.5. Experiments for Theorems 4.5 and 4.6

To verify experimentally that the stationary points in Theorems 4.5 and 4.6 are indeed learned during training, we plot the difference between each parameter matrix and its predicted value against training time. Figure below 3 illustrates this convergence for a specific setup of $(\mathcal{K}^{exp}, \tilde{h}^{softmax})$. We present more extensive experiments using different combinations of architecture $\tilde{h}$ and label distribution $\mathcal{K}$ in Figures 5 and 6 in Appendices H.2 and H.3 (where we also provide full experiment details).



3(a) Theorem 4.5 3(c) Theorem 4.6

Figure 3. Plots of $\log(\mathsf{dist}(M,I))$ for $M = \{A_0, A_1\} \cup \Sigma^{1/2}\left\{B_0^\top C_0, B_1^\top C_1, B_2^\top C_2\right\}\Sigma^{1/2}$ against number of training iterations, where $\mathsf{dist}(M,I) := \min_\alpha \frac{\|M - \alpha \cdot I\|}{\|M\|_F}$. Data labels are drawn from a $\mathcal{K}^{exp}$-Gaussian Process. The network is a 3-layer $\tilde{h}^{softmax}$ Transformer. Figure 3(a) uses the setting of Theorem 4.5, where we constrain $A_\ell = 0$. Figure 3(b) uses the setting of Theorem 4.6, where A_ℓ's are unconstrained. We only verify $B^\top C$ because for any $\Lambda \in \mathbb{R}^{d\times d}$, (B_ℓ, C_ℓ) gives identical prediction as $(\Lambda^\top B_\ell, \Lambda^{-1} C_\ell)$. (See Remark E.2)

5. Future Directions

Representation Power via Composition: Using the idea of kernel composition, we showed in Proposition D.1 that a multi-headed Transformer can have significantly greater representation power by combining *parallel attention heads*. It will be interesting to also investigate the representation power of composition across layers, i.e. *sequential attention heads*. Another question is whether diverse activations have similar benefits for practical tasks.

Optimal choice of non-linearity: In Section 3.3, we saw how the optimal choice of non-linear activation can depend on the function being learned. This may provide some intuition for how to select the right non-linear activation in practical settings.

Understanding functional gradient descent++ What is the interpretation of the algorithm implemented in Theorem 4.6? In the linear setting, the action of the value matrix improves the condition number of the GD objective (von Oswald et al., 2023a), does a similar statement hold in the non-linear setting?

Stronger Theoretical Guarantees: Can we show global optimality, or even establish convergence guarantees, for the stationary points in Theorem 4.5 and 4.6, in light of our experimental observations?

Acknowledgement

Xiang Cheng acknowledges NSF CCF-2112665 (TILOS AI Research Institute) for their generous support. Suvrit Sra acknowledges NSF CCF-2112665 (TILOS AI Research Institute) and the Alexander von Humboldt foundation for their generous support.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. *International Conference on Learning Representations*, 2022.

Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34: 20014–20027, 2021.

Bai, Y., Chen, F., Wang, H., Xiong, C., and Mei, S. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Neural Information Processing Systems*, 2020.

Chen, Y., Tao, Q., Tonin, F., and Suykens, J. A. Primal-attention: Self-attention through asymmetric kernel svd in primal representation. *arXiv preprint arXiv:2305.19798*, 2023.

Chi, T.-C., Fan, T.-H., Ramadge, P. J., and Rudnicky, A. Kerple: Kernelized relative positional embedding for length extrapolation. *Advances in Neural Information Processing Systems*, 35:8386–8399, 2022.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., and Wei, F. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Giannou, A., Rajput, S., Sohn, J.-y., Lee, K., Lee, J. D., and Papailiopoulos, D. Looped transformers as programmable computers. *arXiv preprint arXiv:2301.13196*, 2023.

Huang, Y., Cheng, Y., and Liang, Y. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.

Lin, L., Bai, Y., and Mei, S. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*, 2023.

Mahankali, A., Hashimoto, T. B., and Ma, T. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.

Nguyen, T., Pham, M., Nguyen, T., Nguyen, K., Osher, S., and Ho, N. Fourierformer: Transformer meets generalized fourier integral theorem. *Advances in Neural Information Processing Systems*, 35:29319–29335, 2022a.

Nguyen, T. M., Nguyen, T. M., Le, D. D., Nguyen, D. K., Tran, V.-A., Baraniuk, R., Ho, N., and Osher, S. Improving transformers with probabilistic attention keys. In *International Conference on Machine Learning*, pp. 16595–16621. PMLR, 2022b.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

Pérez, J., Barceló, P., and Marinkovic, J. Attention is turing complete. *The Journal of Machine Learning Research*, 2021.

Schlag, I., Irie, K., and Schmidhuber, J. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, pp. 9355–9366. PMLR, 2021.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *International conference on computational learning theory*, pp. 416–426. Springer, 2001.

Tsai, Y.-H. H., Bai, S., Yamada, M., Morency, L.-P., and Salakhutdinov, R. Transformer dissection: a unified understanding of transformer's attention via the lens of kernel. *arXiv preprint arXiv:1908.11775*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 2017.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023a.

von Oswald, J., Niklasson, E., Schlegel, M., Kobayashi, S., Zucchet, N., Scherrer, N., Miller, N., Sandler, M., Vladymyrov, M., Pascanu, R., et al. Uncovering mesa-optimization algorithms in transformers. *arXiv preprint arXiv:2309.05858*, 2023b.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wei, C., Chen, Y., and Ma, T. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Advances in Neural Information Processing Systems*, 35:12071–12083, 2022.

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

Wright, M. A. and Gonzalez, J. E. Transformers are deep infinite-dimensional non-mercer binary kernel machines. *arXiv preprint arXiv:2106.01506*, 2021.

Wu, J., Zou, D., Chen, Z., Braverman, V., Gu, Q., and Bartlett, P. L. How many pretraining tasks are needed for in-context learning of linear regression? *arXiv preprint arXiv:2310.08391*, 2023.

Zhang, R., Frei, S., and Bartlett, P. L. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

A. Reformulating the In-Context Loss

Lemma A.1. *Let $Z_0 \in \mathbb{R}^{(d+1)\times(n+1)}$ be the input to the Transformer (as defined in (1)):*

$$Z_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \cdots & y^{(n)} & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)}.$$

*Let $\overline{Z}_0$ be the input of the Transformer **without masking out** $y^{(n+1)}$:*

$$\overline{Z}_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \cdots & y^{(n)} & y^{(n+1)} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)},$$

*where $y^{(n+1)} = \langle w_\star, x^{(n+1)} \rangle$. Let Z_ℓ denote the output of the $(\ell-1)^{th}$ layer of the linear transformer **initialized at** Z_0 (as defined in (3)). Let $\overline{Z}_\ell$ denote the output of the $(\ell-1)^{th}$ layer of the linear transformer **initialized at** $\overline{Z}_0$ (as defined in (3)). Let $f(V, Q, K)$ denote the in-context loss defined in (5), i.e.*

$$f(V, Q, K) = \mathbb{E}_{\overline{Z}_0}\left[\left([Z_{k+1}]_{(d+1),(n+1)} + y^{(n+1)}\right)^2\right]. \tag{19}$$

Let V_ℓ satisfy Assumption 4.4. Then the in-context loss, defined in (5), has the equivalent form

$$f(A, B, C) := f(W) = \mathbb{E}_{\overline{Z}_0}\left[\mathrm{Tr}\left((I - M)\,\overline{Y}_{k+1}^\top \overline{Y}_{k+1}\,(I - M)\right)\right],$$

where $\overline{Y}_{k+1} \in \mathbb{R}^{1\times n+1}$ is the $(d+1)^{th}$ row of $\overline{Z}_{k+1}$.

Proof. Let $X_\ell \in \mathbb{R}^{d\times(n+1)}$ denote the first d rows of Z_ℓ and let $Y_\ell \in \mathbb{R}^{1\times(n+1)}$ denote the last row of Z_ℓ. Under Assumption 4.4, we can verify the following useful equivalent form of (3):

$$\begin{aligned} X_{\ell+1} &= X_\ell + A_\ell X_\ell M \tilde{h}\left(B_\ell X_\ell, C_\ell X_\ell\right) \\ Y_{\ell+1} &= Y_\ell + r_\ell Y_\ell M \tilde{h}\left(B_\ell X_\ell, C_\ell X_\ell\right) \end{aligned} \tag{20}$$

Let $c \in \mathbb{R}$ denote an arbitrary scalar. Let $\overline{X}_0 := X_0$ and let $\overline{Y}_0 = \begin{bmatrix} y^{(1)} & y^{(2)} & \cdots & y^{(n)} & y^{(n+1)} + c \end{bmatrix}$, i.e. $\overline{Y}_0$ is $\overline{Y}_0$ but with c added to its last entry. Let $\overline{X}_\ell, \overline{Y}_\ell$ evolve under identical dynamics as (20) (but with initialization at $\overline{X}_0, \overline{Y}_0$):

$$\begin{aligned} \overline{X}_{\ell+1} &= \overline{X}_\ell + A_\ell \overline{X}_\ell M \tilde{h}\left(B_\ell \overline{X}_\ell, C_\ell \overline{X}_\ell\right) \\ \overline{Y}_{\ell+1} &= \overline{Y}_\ell + r_\ell \overline{Y}_\ell M \tilde{h}\left(B_\ell \overline{X}_\ell, C_\ell \overline{X}_\ell\right). \end{aligned}$$

Then for all i, (1) $\overline{X}_\ell = X_\ell$ and (2) $\overline{Y}_\ell - Y_\ell = \begin{bmatrix} 0 & 0 & \cdots & 0 & c \end{bmatrix}$.

Statement (1) can be verified via simple induction.

Statement (2) follows from Statement (1) and induction: suppose (2) holds for some i. By definition of M, $\overline{Y}_\ell M$ has its $(n+1)^{th}$ entry zeroed out, thus by the inductive hypothesis, $\overline{Y}_\ell M = Y_\ell M$, and thus $\overline{Y}_{\ell+1} = Y_{\ell+1} + [0\,0 \cdots 0\,c]$.

The lemma statement then follows by choosing $c = y^{(n+1)}$: by (2) above, $\left[\overline{Z}_{k+1}\right]_{(d+1),(n+1)} =: \left[\overline{Y}_{k+1}\right]_{(n+1)} = [Y_{k+1}]_{(n+1)} + y^{(n+1)}$. Plugging the above into the in-context loss defined in (5) gives

$$\begin{aligned} f(V, Q, K) &= \mathbb{E}_{\overline{Z}_0}\left[\left([Z_{k+1}]_{(d+1),(n+1)} + y^{(n+1)}\right)^2\right] \\ &= \mathbb{E}_{\overline{Z}_0}\left[\left(\left[\overline{Z}_{k+1}\right]_{(d+1),(n+1)}\right)^2\right] \\ &= \mathbb{E}_{\overline{Z}_0}\left[\left(\left\|(I - M)\,\overline{Y}_{k+1}^\top\right\|\right)^2\right] \\ &= \mathbb{E}_{\overline{Z}_0}\left[\left(\mathrm{Tr}\left((I - M)\,\overline{Y}_{k+1}^\top \overline{Y}_{k+1}\,(I - M)\right)\right)^2\right] \end{aligned}$$

□

B. Proof of Proposition 3.1

We will first write down the explicit expression for (7). By Lemma G.3, for any $f \in \mathbb{H}$, $\nabla L(f) = -c\sum_{i=1}^n \left(y^{(i)} - f(x^{(i)})\right) \mathcal{K}(\cdot, x^{(i)})$, thus (7) is equivalent to

$$f_{\ell+1}(\cdot) = f_\ell(\cdot) + r'_\ell \sum_{i=1}^n \left(y^{(i)} - f_\ell(x^{(i)})\right) \mathcal{K}(\cdot, x^{(i)}). \tag{21}$$

See proof of Lemma G.3 for the explicit relation between r'_ℓ and r_ℓ.

Let X_ℓ and Y_ℓ denote the first d rows and the last row of Z_ℓ respectively, for any layer $\ell = 0 \dots k+1$ and for any $i = 0 \dots n$,

$$Y_\ell^{(i)} = y^{(i)} + \mathsf{TF}_\ell(x^{(i)}; (V, B, C), z^{(1)} \dots z^{(n)}). \tag{22}$$

In words: "$y^{(i)} - Y_\ell^{(i)}$ is equal to the predicted label for $x^{(n+1)}$, if $x^{(i)} = x^{(n+1)}$." (22) follows immediately from (3), by setting $x^{(n+1)} = x^{(i)}$, and verifying that $[Z_\ell]_{(d+1),i}$ and $[Z_\ell]_{(d+1),(n+1)}$ have identical updates across layers $\ell = 0 \dots k$.

We will now prove the lemma statement by induction. For the input, $[Z_0]_{(d+1),(n+1)} := 0 = f_0(x)$ by definition in (1), so that (8) holds. Now assume that $\mathsf{TF}_\ell(x; (V, B, C) | z^{(1)} \dots z^{(n)}) = -f_\ell(x)$ up to some layer ℓ.

By definition of the dynamics on Z_i in (3), and plugging in our choice of V, B, C, we verify that

$$\begin{aligned}
&\mathsf{TF}_{\ell+1}(x; (V, B, C) | z^{(1)} \dots z^{(n)}) \\
=&\mathsf{TF}_\ell(x; (V, B, C) | z^{(1)} \dots z^{(n)}) - r_\ell \sum_{i=1}^n Y_\ell^{(i)} \left[\tilde{h}\left(X_0, X_0\right)\right]_{i,(n+1)} && (23)\\
=&\mathsf{TF}_\ell(x; (V, B, C) | z^{(1)} \dots z^{(n)}) - r_\ell \sum_{i=1}^n Y_\ell^{(i)} \mathcal{K}(x^{(i)}, x) && (24)\\
=& - f_\ell(x) - r_\ell \sum_{i=1}^n \left(y^{(i)} - f_\ell(x)\right) \mathcal{K}(x^{(i)}, x) \\
=& - f_{\ell+1}(x).
\end{aligned}$$

In the above, the first line is by plugging in our choice of V, B, C into (3). The second line is by our assumption on $\tilde{h}$ in the lemma statement. The third line is by inductive hypothesis, along with (22). The fourth line is by (21). This concludes the proof.

C. Proof of Proposition 3.4

Let $\hat{Y} \in \mathbb{R}^n$ denote the vector of $y^{(1)} \dots y^{(n)}$. Let $\hat{\mathbb{K}}$ denote the top-left $n \times n$ block of $\mathbb{K}$. Let $\nu \in \mathbb{R}^n$ denote the vector given by $\nu_i := \mathbb{K}_{i,n+1}$. i.e. $\mathbb{K} =: \begin{bmatrix} \hat{\mathbb{K}} & \nu \\ \nu^\top & \mathbb{K}_{(n+1),(n+1)} \end{bmatrix}$. By the formula for conditional Gaussian, we know that $y^{(n+1)}$, conditioned on $y^{(1)} \dots y^{(n)}$, has Gaussian distribution with mean $\nu^\top \hat{\mathbb{K}}^{-1} \hat{Y}$. The Bayes estimator of $y^{(n+1)}$ is thus exactly this mean, which is equivalent to

$$\nu^\top \hat{\mathbb{K}}^{-1} \hat{Y} = \sum_{j,k=1}^n \mathcal{K}\left(x^{(n+1)}, x^{(j)}\right) \left[\hat{\mathbb{K}}^{-1}\right]_{jk} y^{(k)}. \tag{25}$$

Consider the construction in Proposition 3.1, i.e. $B_\ell = C_\ell = I$, $V_\ell = \begin{bmatrix} 0 & 0 \\ 0 & -r_\ell \end{bmatrix}$. For simplicity, further assume that $r_\ell = \delta$ for all ℓ, where δ is some positive constant satisfying $\delta < \|\hat{\mathbb{K}}\|_2$. For $\ell = 0 \dots k$, let $y_\ell^{(i)} := [Z_\ell]_{(d+1),i}$, and let $\hat{Y}_\ell := \left[y_\ell^{(1)} \dots y_\ell^{(n+1)}\right]$. From (3), we verify that under the above choice of Transformer weights and ,

$$y_{\ell+1}^{(i)} = y_\ell^{(i)} - \delta \sum_{j=1}^n \mathcal{K}\left(x^{(i)}, x^{(j)}\right) y_\ell^{(j)} \qquad \Leftrightarrow \qquad \hat{Y}_{\ell+1} = \left(I - \delta \hat{\mathbb{K}}\right) \hat{Y}_\ell = \left(I - \delta \hat{\mathbb{K}}\right)^\ell \hat{Y}.$$

Again by (3), we verify that $y_{\ell+1}^{(n+1)} = y_\ell^{(n+1)} - \delta \sum_{j=1}^n \mathcal{K}\left(x^{(n+1)}, x^{(j)}\right) y_\ell^{(j)}$. Rearranging terms gives $y_{\ell+1}^{(n+1)} = y_\ell^{(n+1)} - \delta\nu^\top \hat{Y}_\ell = -\nu^\top \sum_{k=0}^{\ell} \hat{Y}_k = -\nu^\top \sum_{k=0}^{\ell} \left(I - \delta\hat{\mathbb{K}}\right)^k \hat{Y}_k$. By Taylor expansion, $\hat{\mathbb{K}}^{-1} = \delta \sum_{\ell=0}^{\infty} \left(I - \delta\hat{\mathbb{K}}\right)^\ell$. Thus as $\ell \to \infty$, $y_{\ell+1}^{(n+1)} \to \nu^\top \hat{\mathbb{K}}^{-1} \hat{Y}$, which is the optimal estimator of $y^{(n+1)}$.

D. Composing multiple attention heads with different activations

Formally, we will consider Transformers with *multi-head* attention, defined by the following forward pass:

$$Z_{\ell+1} = Z_\ell + \sum_{s=1}^{H} V_\ell^s Z_\ell M \tilde{h}^s \left(B_\ell^s X_\ell, C_\ell^s X_\ell\right). \tag{26}$$

H denotes the number of heads in a layer, and $\{V_\ell^s, B_\ell^s, C_\ell^s\}_{s=1\ldots H}$ denote the {value, key, query} matrices at layer ℓ for head s. $\tilde{h}^s$ denotes the activation for head s, **which could be different for each head**. The difference between (26) and (3), is the additional summation over multiple heads $\sum_{s=1}^H$. Identical to to (4), we let TF_ℓ denote the Transformer's prediction for $-y^{(n+1)}$ at layer ℓ, given $x^{(n+1)} = x$, conditioned on $z^{(1)}\ldots z^{(n)}$ as:

$$\mathsf{TF}_\ell(x; (V, B, C) | z^{(1)} \ldots z^{(n)}) := [Z_\ell]_{(d+1),(n+1)}, \tag{27}$$

where Z_i evolves as (26), initialized at $Z_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \ldots & x^{(n)} & x \\ y^{(1)} & y^{(2)} & \ldots & y^{(n)} & 0 \end{bmatrix}$. We now present Proposition D.1 which shows that a single multi-head Transformer can perform (optimal) **functional gradient descent** with respect to a large class of RKHS metrics. Its proof is very similar to Propositions 3.1 and 3.4, we present it in Appendix D.1.

Proposition D.1. *Let $\left\{z^{(i)}\right\}_{i=1\ldots n}$ denote the in-context examples and let $L(f)$ be the empirical loss functional as defined in Proposition 3.1. For $s = 1\ldots H$, let $\mathcal{K}^s$ denote a PSD kernel function. Let $\mathcal{K}^\diamond$ be a composite kernel, defined as $\mathcal{K}^\diamond(u, v) := \sum_{s=1}^H \mathcal{K}^s(G^s u, G^s v)$, where $G^s \in \mathbb{R}^{d\times d}$ are subject to the constraint that $\mathcal{K}^\diamond$ must be PSD (but are otherwise arbitrary). Let f_ℓ denote the functional gradient descent (7) of $L(f)$, wrt the RKHS metric induced by $\mathcal{K}^\diamond$.*

*(A) **[Generalization of Proposition 3.1]** Consider the multi-head Transformer with H heads, where the s^{th} head has activation defined as $\left[\tilde{h}^s(U, V)\right]_{ij} := \mathcal{K}^s\left(U^{(i)}, W^{(j)}\right)$. Let the Transformer's parameters be $V_\ell^s = \begin{bmatrix} 0 & 0 \\ 0 & -r_\ell^s \end{bmatrix}$, $B_\ell^s = G^s$, $C_\ell^s = G^s$. Then there exist scalars $\{r_\ell^s\}_{s=1\ldots H, \ell=0\ldots k}$ such that the following holds: For any $x := x^{(n+1)}$, the Transformer's prediction for $y^{(n+1)}$ at each layer ℓ matches the prediction of the functional gradient sequence f_ℓ (27), i.e. for all $\ell = 0 \ldots k$,*

$$\mathsf{TF}_\ell(x; (V, B, C) | z^{(1)} \ldots z^{(n)}) = -f_\ell(x). \tag{28}$$

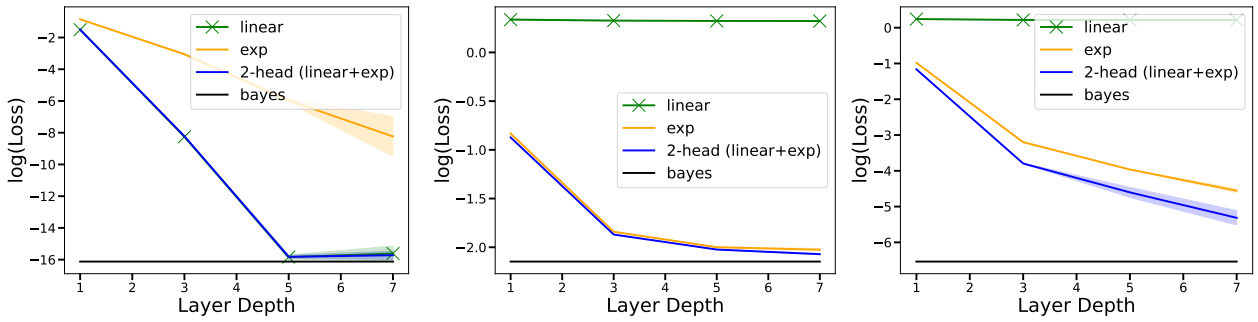
*(B) **[Generalization of Proposition 3.4]** If we additionally assume that $Y|X$ is drawn from the $\mathcal{K}^\diamond$ Gaussian Process, then as the number of layers $\ell \to \infty$, the Transformer's prediction for $y^{(n+1)}$ at layer ℓ (28) approaches the **Bayes (optimal) estimator** that minimizes the in-context loss (5).*

Remarkably, **a single** multi-head Transformer can give the near-optimal predictions over a large class of data distributions, **even without *a priori* knowledge of the data distribution**.

Figure 4 provides empirical verification of Proposition D.1: We plot the loss against number of layers for three kinds of Transformers: 1-head with $\tilde{h}^{linear}$ activation, 1-head with $\tilde{h}^{exp}$ activation, 2-head with $\tilde{h}^{linear}$ activation on the first head and $\tilde{h}^{linear}$ on the second head. Data labels are drawn from a $\mathcal{K}^\diamond$ Gaussian Process, where $\mathcal{K}^\diamond(u, v) := \alpha\mathcal{K}^{linear}(G_1 u, G_1 v) + (1-\alpha)\mathcal{K}^{exp}(G_2 u, G_2 v)$. We observe the following

1. In Figure 4(a) and 4(a), we see that the 2-head Transformer **can perform optimally on both $\mathcal{K}^{linear}$ and $\mathcal{K}^{\exp}$ data**. Specifically: In Figure 4(a), $\mathcal{K}^\diamond = \mathcal{K}^{linear}$ ($\alpha = 1$, $G_1 = G_2 = I$). The 2-head Transformer performs as well as the $\tilde{h}^{linear}$ Transformer. In Figure 4(b), $\mathcal{K}^\diamond = \mathcal{K}^{exp}$ ($\alpha = 0$, $G_1 = G_2 = I$). The 2-head Transformer performs as well as the $\tilde{h}^{exp}$ Transformer.
2. In Figure 4(c), $\mathcal{K}^\diamond(u, v) := \frac{1}{2}(u_1 v_1 + u_2 v_2) + \frac{1}{2}\exp\left(\frac{1}{2}(u_3 v_3 + u_4 v_4 + u_5 v_5)\right)$, corresponding to choosing $\alpha = 1/2$, $G_1 = \mathrm{diag}([1, 1, 0, 0, 0])$ and $G_2 = \mathrm{diag}([0, 0, 1, 1, 1])$. For this choice of $\mathcal{K}^\diamond$, the 2-head Transformer **outperforms both single-head Transformers**.

Note: the Transformer parameters **re-trained for each dataset**, so attention weights for 4(a), 4(b) and 4(c) are **different**.



4(a) $\mathcal{K}^{linear}$ 4(b) $\mathcal{K}^{exp}$ 4(c) $\mathcal{K}^{\diamond}$

Figure 4. Plot of log(test ICL loss) against number of layers. Each sub-figure samples data from a different distribution ($\mathcal{K}^{\diamond}(u,v) := \alpha\mathcal{K}^{linear}(G_1u, G_1v) + (1-\alpha)\mathcal{K}^{exp}(G_2u, G_2v)$). We compare the performance of three kinds of Transformers. The labels are generated using a $\mathcal{K}^{\diamond}$ Gaussian Process. Context length $n = 14$.

D.1. Proof of Proposition D.1

The proof of (A) is identical to the proof of Proposition 3.1 in Appendix B, up to step (23). We provide the remainder of the proof below:

$$
\begin{aligned}
&\mathsf{TF}_{\ell+1}(x;(V,B,C)|z^{(1)}\dots z^{(n)})\\
=&\mathsf{TF}_{\ell}(x;(V,B,C)|z^{(1)}\dots z^{(n)}) - r_\ell \sum_{i=1}^{n}\sum_{s=1}^{H} Y_\ell^{(i)} \left[\tilde{h}^s\left(G^sX_0, G^sX_0\right)\right]_{i,(n+1)}\\
=&\mathsf{TF}_{\ell}(x;(V,B,C)|z^{(1)}\dots z^{(n)}) - r_\ell \sum_{i=1}^{n}\sum_{s=1}^{H} Y_\ell^{(i)} \mathcal{K}^s(G^sx^{(i)}, G^sx)\\
=&\mathsf{TF}_{\ell}(x;(V,B,C)|z^{(1)}\dots z^{(n)}) - r_\ell \sum_{i=1}^{n} Y_\ell^{(i)} \mathcal{K}^{\diamond}(x^{(i)}, x)\\
=&-f_\ell(x) - r_\ell \sum_{i=1}^{n}\left(y^{(i)} - f_\ell(x)\right)\mathcal{K}^{\diamond}(x^{(i)}, x)\\
=&-f_{\ell+1}(x).
\end{aligned}
$$

We highlight in red the differences from the proof of Proposition 3.1. We use the definition of $\tilde{h}^s$ and the definition $\mathcal{K}^{\diamond}(u,v) := \sum_{s=1}^{H}\mathcal{K}^s(G^su, G^sv)$.

The proof of (B) is entirely identical to proof of Proposition 3.4 in Appendix C, only replacing $\mathcal{K}$ by $\mathcal{K}^{\diamond}$, and using (26) instead of (3).

E. Theorem E.1: Functional Gradient Descent is locally optimal under $A_\ell = 0$ constraint.

The following is the formal statement of Theorem 4.5:

Theorem E.1. *Let $\tilde{h}$ satisfy Assumption 4.3, let $x^{(i)}$'s satisfy Assumption 4.1 with matrix Σ, and $y^{(i)}$'s satisfy Assumption 4.2. With slight abuse of notation, let $f(r,B,C) := f\left(V = \left\{\begin{bmatrix}0 & 0\\ 0 & r_\ell\end{bmatrix}\right\}_{\ell=0\dots k}, B, C\right)$, where $f(V,B,C)$ is as defined in (5). Let $\mathcal{S} \subset \mathbb{R}^{(k+1)\times d\times d\times 2}$ denote a set of (Query, Key) matrices defined as follows: $(B,C) \in \mathcal{S}$ if and only if for all $\ell \in \{0\dots k\}$, there exist scalars $b_\ell, c_\ell \in \mathbb{R}$ such that $B_\ell = b_\ell\Sigma^{-1/2}$ and $C_\ell = c_\ell\Sigma^{-1/2}$. Then*

$$
\inf_{(r,B,C)\in\mathbb{R}^{k+1}\times\mathcal{S}} \sum_{\ell=0}^{k} \left(\partial_{r_\ell} f(r,B,C)\right)^2 + \left\|\nabla_{B_\ell} f(r,B,C)\right\|_F^2 + \left\|\nabla_{C_\ell} f(r,B,C)\right\|_F^2 = 0, \tag{29}
$$

where $\nabla_{B_\ell} f$ denotes derivative wrt the Frobenius norm $\|B_\ell\|_F$ (same for ∇_{C_ℓ}).

Remark E.2. By Assumption 4.3, for any invertible $\Lambda \in \mathbb{R}^{d\times d}$, $f(r, B, C) = f(r, \Lambda^\top B, \Lambda^{-1}C)$. Thus the same result holds for $S_\Lambda = \left\{B_\ell = b_\ell \Lambda^\top \Sigma^{-1/2}, C_\ell = c_\ell \Lambda^{-1}\Sigma^{-1/2}\right\}_{\ell=0\dots k}$.

E.1. Proof of Theorem E.1

Let $r(0) \in \mathbb{R}, (B(0), C(0)) \in \mathcal{S}$. Let us define the $\mathcal{S}$-gradient-flow as

$$\begin{aligned}
\frac{d}{dt} r_\ell(t) &= -\partial_{r_\ell} f(r(t), B(t), C(t)) \\
\frac{d}{dt} B_\ell(t) &= \tilde{U}_\ell(t) \\
\frac{d}{dt} C_\ell(t) &= \tilde{W}_\ell(t),
\end{aligned} \tag{30}$$

where for $\ell = 0 \dots k$, $\tilde{U}$ and $\tilde{W}$ are defined as

$$\begin{aligned}
\tilde{u}_\ell(t) &:= -\frac{1}{d} \operatorname{Tr}\left(\nabla_{B_\ell} f(r(t), B(t), C(t))\Sigma^{1/2}\right) \qquad \tilde{U}_\ell(t) := \tilde{u}_\ell(t)\Sigma^{-1/2} \\
\tilde{w}_\ell(t) &:= -\frac{1}{d} \operatorname{Tr}\left(\nabla_{C_\ell} f(r(t), B(t), C(t))\Sigma^{1/2}\right) \qquad \tilde{W}_\ell(t) := \tilde{w}_\ell(t)\Sigma^{-1/2}.
\end{aligned}$$

It follows by definition of $\tilde{U}$ and $\tilde{W}$ that $(B(t), C(t)) \in \mathcal{S}$ for all t. We will show that at any time t,

$$\begin{aligned}
\frac{d}{dt} f(r(t), B(t), C(t)) \le &- \sum_{\ell=0}^{k} \left(\partial_{r_\ell} f(r(t), B(t), C(t))\right)^2 \\
&- \sum_{\ell=0}^{k} \|\nabla_{B_\ell} f(r(t), B(t), C(t))\|_F^2 - \sum_{\ell=0}^{k} \|\nabla_{C_\ell} f(r(t), B(t), C(t))\|_F^2 .
\end{aligned} \tag{31}$$

Let $\langle A, B\rangle_{\mathrm{Tr}} := \operatorname{Tr}\left(A^\top B\right)$. By definition of the dynamics in (30),

$$\frac{d}{dt} f(r(t), B(t), C(t)) \tag{32}$$

$$= \sum_{\ell=0}^{k} \partial_{r_\ell} f(r(t), B(t), C(t)) \cdot \left(-\partial_{r_\ell} f(r(t), B(t), C(t))\right) \tag{33}$$

$$+ \sum_{\ell=0}^{k} \left\langle \nabla_{B_\ell} f(r(t), B(t), C(t)), \tilde{U}_\ell(t) \right\rangle_{\mathrm{Tr}} \tag{34}$$

$$+ \sum_{\ell=0}^{k} \left\langle \nabla_{C_\ell} f(r(t), B(t), C(t)), \tilde{W}_\ell(t) \right\rangle_{\mathrm{Tr}} . \tag{35}$$

We immediately verify that (33) $= -\sum_{\ell=0}^{k} \left(\partial r_\ell f(r(t), B(t), C(t))\right)^2$. By (38) from Proposition E.3, applied separately to each layer $\ell = 0 \dots k$,

$$\begin{aligned}
(34) &\le \sum_{\ell=0}^{k} \langle \nabla_{B_\ell} f(r(t), B(t), C(t)), -\nabla_{B_\ell} f(r(t), B(t), C(t))\rangle_{\mathrm{Tr}} \\
&= -\sum_{\ell=0}^{k} \|\nabla_{B_\ell} f(r(t), B(t), C(t))\|_F^2 .
\end{aligned}$$

Similarly, by (39) from Proposition E.3, applied separately to each layer $\ell = 0 \dots k$,

$$\begin{aligned}
(35) &\le \sum_{\ell=0}^{k} \langle \nabla_{C_\ell} f(r(t), B(t), C(t)), -\nabla_{C_\ell} f(r(t), B(t), C(t))\rangle_{\mathrm{Tr}} \\
&= -\sum_{\ell=0}^{k} \|\nabla_{C_\ell} f(r(t), B(t), C(t))\|_F^2 .
\end{aligned}$$

Combining the above bounds gives (31). Suppose (29) does not hold. Then there exists a positive constant $c > 0$ such that for all t,

$$\sum_{\ell=0}^{k} \left(\partial_{r_\ell} f(r(t), B(t), C(t))\right)^2 + \|\nabla_{B_\ell} f(r(t), B(t), C(t))\|_F^2 + \|\nabla_{C_\ell} f(r(t), B(t), C(t))\|_F^2 \geq c.$$

Then by (31), $\frac{d}{dt} f(r(t), B(t), C(t)) \leq -c$ for all t. This contradicts the fact that $f(\cdot)$ is bounded below by 0 (see (5)). Thus we prove (29).

E.2. Key Lemmas

Proposition E.3. *Let $\tilde{h}$ satisfy Assumption 4.3, let $x^{(i)}$'s satisfy Assumption 4.1 with matrix Σ, and $y^{(i)}$'s satisfy Assumption 4.2. Let $V \in \mathbb{R}^{(k+1)\times(d+1)\times(d+1)}$ satisfy, for all $\ell = 0 \dots k$, $V_\ell = \begin{bmatrix} 0 & 0 \\ 0 & r_\ell \end{bmatrix}$, where r_ℓ are arbitrary scalars. Let $(B, C) \in \mathbb{R}^{(k+1)\times d\times d\times 2}$ satisfy, for all $\ell = 0 \dots k$,*

$$B_\ell = b_\ell \Sigma^{-1/2} \qquad C_\ell = c_\ell \Sigma^{-1/2}, \tag{36}$$

where $b_\ell, c_\ell \in \mathbb{R}$ are scalars. Let $j \in \{0 \dots k\}$ be an arbitrary but fixed layer index. For $S \in \mathbb{R}^{d\times d}$, let $B_j(S) := B_j + S$, and let $B_\ell(S) := B_\ell$ for $\ell \neq j$. Let $B(S) := \{B_\ell(S)\}_{\ell=0\dots k}$. Recall $f(V, B, C)$ as defined in (5). Let $R \in \mathbb{R}^{d\times d}$ be an arbitrary matrix. Let

$$\tilde{r} := \frac{1}{d} \operatorname{Tr}\left(R\Sigma^{1/2}\right) \qquad \tilde{R} := \tilde{r}\Sigma^{-1/2} \tag{37}$$

Then

$$\frac{d}{dt} f(V, B(tR), C)\Big|_{t=0} \leq \frac{d}{dt} f(V, B(t\tilde{R}), C)\Big|_{t=0}. \tag{38}$$

Similarly, let $C_j(S) := C_j + S$, and $C_\ell(S) := C_\ell$ for $\ell \neq j$, and let $C(S) := \{C_\ell(S)\}_{\ell=0\dots k}$, then

$$\frac{d}{dt} f(V, B, C(tR))\Big|_{t=0} \leq \frac{d}{dt} f(V, B, C(t\tilde{R}))\Big|_{t=0}. \tag{39}$$

Proof of Proposition E.3. The proof of (39) is identical to that of (38), so we only present the proof of (38).

Loss Reformulation: Let us consider the reformulation of the in-context loss f presented in Lemma 5. Specifically, let $\overline{Z}_0$ be defined as

$$\overline{Z}_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \dots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \dots & y^{(n)} & y^{(n+1)} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)},$$

Let $\overline{Z}_i$ denote the output of the $(i-1)^{th}$ layer of the linear transformer (as defined in (3), initialized at $\overline{Z}_0$). For the rest of this proof, we will drop the bar, and simply denote $\overline{Z}_i$ by Z_i. Let $X_i \in \mathbb{R}^{d\times(n+1)}$ denote the first d rows of Z_i and let $Y_i \in \mathbb{R}^{1\times(n+1)}$ denote the $(d+1)^{th}$ row of Z_k. Under the assumption that $V_\ell = \begin{bmatrix} 0 & 0 \\ 0 & r_\ell \end{bmatrix}$ in the lemma statement, we verify that, for any $\ell \in \{0 \dots k\}$,

$$\begin{aligned} X_{\ell+1} &= X_0 \\ Y_{\ell+1} &= Y_\ell + r_\ell Y_\ell M \tilde{h}\left(B_\ell X_\ell, B_\ell X_\ell\right) = Y_0 \prod_{\ell=0}^{i} \left(I + r_\ell M \tilde{h}\left(B_\ell X_0, C_\ell X_0\right)\right). \end{aligned} \tag{40}$$

By Lemma A.1, the in-context loss defined in (5) is equivalent to

$$f(V, B, C) = \mathbb{E}_{Z_0}\left[\operatorname{Tr}\left((I - M) Y_{k+1}^\top Y_{k+1} (I - M)\right)\right],$$

where Y_{k+1} is as defined in (40). We will now verify (38)

We will introduce one more piece of notation: for any $S \in \mathbb{R}^{d\times d}$, let

$$G(X,S) := \prod_{\ell=0}^{k} \left(I + r_\ell M \tilde{h}\left(B_\ell(S)X, C_\ell X\right)\right),$$

so that

$$\begin{aligned} f(V, B(S), C) =& \mathbb{E}_{Z_0}\left[\mathrm{Tr}\left((I-M)\, G(X,S)^\top Y_0^\top Y_0 G(X,S)\,(I-M)\right)\right] \\ =& \mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\, G(X,S)^\top \mathbb{K}(X_0) G(X,S)\,(I-M)\right)\right], \end{aligned}$$

where recall that $\mathbb{K}(X_0) \in \mathbb{R}^{(n+1)\times(n+1)}$ is as defined in Assumption 4.2. The second equality uses the assumption on distribution of Y_0 conditioned on X_0, as specified in Assumption 4.2. Let U denote a uniformly randomly sampled orthogonal matrix. Let $U_\Sigma := \Sigma^{1/2} U \Sigma^{-1/2}$, so that $U_\Sigma^{-1} = \Sigma^{1/2} U^\top \Sigma^{-1/2}$. Using the fact that $X_0 \overset{d}{=} U_\Sigma X_0$, we can verify

$$\begin{aligned} \frac{d}{dt} f(V, B(tR), C)\Big|_{t=0} =& \frac{d}{dt}\, \mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\, G(X_0,tR)^\top \mathbb{K}(X_0) G(X_0,tR)\,(I-M)\right)\right]\Big|_{t=0} \\ =& 2\, \mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\, G(X_0,0)^\top \mathbb{K}(X_0)\, \frac{d}{dt} G(X_0,tR)\Big|_{t=0} (I-M)\right)\right] \\ =& 2\, \mathbb{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\, G(U_\Sigma X_0,0)^\top \mathbb{K}(X_0)\, \frac{d}{dt} G(U_\Sigma X_0,tR)\Big|_{t=0} (I-M)\right)\right]. \end{aligned} \tag{41}$$

The last equality uses the assumption that $\mathcal{K}\left(U_\Sigma X\right) = \mathcal{K}\left(X\right)$ from Assumption 4.2.

We will now show the following useful identities:

$$G(U_\Sigma X_0, 0) = G(X_0, 0) \tag{42}$$

$$\frac{d}{dt} G(U_\Sigma X_0, tR)\Big|_{t=0} = \frac{d}{dt} G(X_0, tU_\Sigma^\top R U_\Sigma)\Big|_{t=0} \tag{43}$$

A useful intermediate identity is

$$\begin{aligned} B_\ell U_\Sigma &= b_\ell \Sigma^{-1/2}\Sigma^{1/2} U \Sigma^{-1/2} = U B_\ell \\ C_\ell U_\Sigma &= c_\ell \Sigma^{-1/2}\Sigma^{1/2} U \Sigma^{-1/2} = U C_\ell. \end{aligned} \tag{44}$$

In the above, we crucially use the assumed form of B_ℓ, C_ℓ from the (36). We can now verify (42), which follows almost immediately from (44):

$$\begin{aligned} G(U_\Sigma X_0, 0) =& \prod_{\ell=0}^{k} \left(I + r_\ell M \tilde{h}\left(B_\ell U_\Sigma X_0, C_\ell U_\Sigma X_0\right)\right) \\ =& \prod_{\ell=0}^{k} \left(I + r_\ell M \tilde{h}\left(U B_\ell X_0, U C_\ell X_0\right)\right) \\ =& \prod_{\ell=0}^{k} \left(I + r_\ell M \tilde{h}\left(B_\ell X_0, C_\ell X_0\right)\right) = G(X_0, 0), \end{aligned}$$

where the second equality uses (44), and the third equality uses the invariance of $\tilde{h}$ from Assumption 4.3.

We now begin the verification of (43). To do so, let $\mathcal{J}_{\tilde{h}} : \mathbb{R}^{d\times(n+1)} \to \mathbb{R}^{(n+1)\times(n+1)}$ denote the Jacobian of $\tilde{h}$ wrt its first argument, evaluated at $(B_\ell X_0, C_\ell X_0)$. In more precise notation, for any $U, V, T \in \mathbb{R}^{d\times(n+1)}$, $\mathcal{J}_{\tilde{h}}(U,V)\,[T] :=$

$\frac{d}{dt}\tilde{h}\left(U+T,V\right)\Big|_{t=0}$. We verify the following useful identity: for any $S\in R^{d\times d}$,

$$\begin{aligned}
&\mathcal{J}_{\tilde{h}}(B_\ell U_\Sigma X_0, C_\ell U_\Sigma X_0)\left[SU_\Sigma X_0\right]\\
=&\frac{d}{dt}\tilde{h}\left(UB_\ell X_0+tSU_\Sigma X_0, UC_\ell X_0\right)\Big|_{t=0}\\
=&\frac{d}{dt}\tilde{h}\left(B_\ell X_0+tU^\top SU_\Sigma X_0, C_\ell X_0\right)\Big|_{t=0}\\
=&\mathcal{J}_{\tilde{h}}(B_\ell X_0, C_\ell X_0)\left[U^\top SU_\Sigma X_0\right], \qquad (45)
\end{aligned}$$

where the first equality is by (44), the second equality is by Assumption 4.3, the third equality is by definition of $\mathcal{J}_{\tilde{h}}$. The identity (43) then follows easily from chain rule and (45):

$$\begin{aligned}
&\frac{d}{dt}G(U_\Sigma X_0, tR)\Big|_{t=0}\\
=&\left(\prod_{\ell=0}^{j-1}\left(I+M\tilde{h}\left(B_\ell U_\Sigma X_0, C_\ell U_\Sigma X_0\right)\right)\right)M\mathcal{J}_{\tilde{h}}\left(B_j U_\Sigma X_0, C_j U_\Sigma X_0\right)[tR]\left(\prod_{\ell=j+1}^{k}\left(I+M\tilde{h}\left(B_\ell U_\Sigma X_0, C_\ell U_\Sigma X_0\right)\right)\right)\\
=&\left(\prod_{\ell=0}^{j-1}\left(I+M\tilde{h}\left(B_\ell X_0, C_\ell X_0\right)\right)\right)M\mathcal{J}_{\tilde{h}}\left(B_j U_\Sigma X_0, C_j U_\Sigma X_0\right)[tR]\left(\prod_{\ell=j+1}^{k}\left(I+M\tilde{h}\left(B_\ell X_0, C_\ell X_0\right)\right)\right)\\
=&\left(\prod_{\ell=0}^{j-1}\left(I+M\tilde{h}\left(B_\ell X_0, C_\ell X_0\right)\right)\right)M\mathcal{J}_{\tilde{h}}\left(B_j X_0, C_j X_0\right)\left[tU^\top RU_\Sigma\right]\left(\prod_{\ell=j+1}^{k}\left(I+M\tilde{h}\left(B_\ell X_0, C_\ell X_0\right)\right)\right)\\
=&\frac{d}{dt}G(U_\Sigma X_0, tU^\top RU_\Sigma)\Big|_{t=0}
\end{aligned}$$

In the above, the second equality uses (44) and Assumption 4.3. The third equality uses (45). The fourth equality again uses chain rule. This concludes the proof of (43).

We will now continue from (41):

$$\begin{aligned}
\frac{d}{dt}f(V,B(tR),C)\Big|_{t=0} = (41) =&2\,\mathbb{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\,G(U_\Sigma X_0,0)^\top\mathbb{K}\,\frac{d}{dt}G(U_\Sigma X_0,tR)\Big|_{t=0}(I-M)\right)\right]\\
=&2\,\mathbb{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\,G(X_0,0)^\top\mathbb{K}\,\frac{d}{dt}G(X_0,tU^\top RU_\Sigma)\Big|_{t=0}(I-M)\right)\right]\\
=&2\,\mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G(X_0,0)^\top\mathbb{K}\,\frac{d}{dt}G(X_0,t\,\mathbb{E}_U\left[U^\top RU_\Sigma\right])\Big|_{t=0}(I-M)\right)\right]\\
=&2\,\mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G(X_0,0)^\top\mathbb{K}\,\frac{d}{dt}G(X_0,t\tilde{R})\Big|_{t=0}(I-M)\right)\right]\\
=&\frac{d}{dt}f(V,B(t\tilde{R}),C)\Big|_{t=0}
\end{aligned}$$

In the above, the second equality is by plugging in (42) and (43). The third equality uses the fact that for any S, $\frac{d}{dt}G(X_0,tS)\big|_{t=0}$ is linear in S (and jointly continuously differentiable in both S and t). The fourth equality uses the definition of $\tilde{R}$ from (37). This concludes the proof of (38). □

F. Theorem F.1: characterizing local optimum when A_ℓ are unconstrained.

The following is the formal statement of Theorem 4.6:

Theorem F.1. *Let $\tilde{h}$ satisfy Assumption 4.3, let $x^{(i)}$'s satisfy Assumption 4.1 with matrix Σ, and $y^{(i)}$'s satisfy Assumption 4.2. With abuse of notation, let $f(r,A,B,C):=f\left(V=\left\{\begin{bmatrix}A_\ell & 0\\ 0 & r_\ell\end{bmatrix}\right\}_{\ell=0\ldots k},B,C\right)$, where $f(V,B,C)$ is as defined in (5).*

Let $\mathcal{S} \subset \mathbb{R}^{(k+1)\times d\times d\times 3}$ denote a set of matrices defined as follows: $(A, B, C) \in \mathcal{S}$ if and only if for all $\ell \in \{0 \dots k\}$, there exist scalars $a_\ell, b_\ell, c_\ell \in \mathbb{R}$ such that $A_\ell = a_\ell I, B_\ell = b_\ell \Sigma^{-1/2}$ and $C_\ell = c_\ell \Sigma^{-1/2}$. Then

$$\inf_{(r,A,B,C)\in\mathbb{R}^{k+1}\times\mathcal{S}} \sum_{\ell=0}^{k} \left(\partial_{r_\ell} f(r, A, B, C)\right)^2 + \|\nabla_{A_\ell} f(r, A, B, C)\|_F^2 + \|\nabla_{B_\ell} f(r, A, B, C)\|_F^2 + \|\nabla_{C_\ell} f(r, A, B, C)\|_F^2 = 0, \tag{46}$$

where $\nabla_{A_\ell} f$ denotes derivative wrt the Frobenius norm $\|A_\ell\|_F$ (same for ∇_{B_ℓ} and ∇_{C_ℓ}).

Remark F.2. By Assumption 4.3, for any invertible $\Lambda \in \mathbb{R}^{d\times d}$, $f(r, A, B, C) = f(r, A, \Lambda^\top B, \Lambda^{-1} C)$. Thus the same result holds for $S_\Lambda = \left\{A_\ell = a_\ell I, B_\ell = b_\ell \Lambda^\top \Sigma^{-1/2}, C_\ell = c_\ell \Lambda^{-1} \Sigma^{-1/2}\right\}_{\ell=0\dots k}$.

F.1. Proof of Theorem F.1

Let $r(0) \in \mathbb{R}, (A(0), B(0), C(0)) \in \mathcal{S}$. Let us define the $\mathcal{S}$-gradient-flow as

$$\begin{aligned}
\frac{d}{dt} r_\ell(t) &= -\partial_{r_\ell} f(r(t), A(t), B(t), C(t)) \\
\frac{d}{dt} A_\ell(t) &= \tilde{P}_\ell(t) \\
\frac{d}{dt} B_\ell(t) &= \tilde{U}_\ell(t) \\
\frac{d}{dt} C_\ell(t) &= \tilde{W}_\ell(t),
\end{aligned} \tag{47}$$

where for $\ell = 0 \dots k$, $\tilde{P}$, $\tilde{U}$, and $\tilde{W}$ are defined as

$$\begin{aligned}
\tilde{p}_\ell(t) &:= -\frac{1}{d} \operatorname{Tr}\left(\Sigma^{-1/2} \nabla_{P_\ell} f(r(t), A(t), B(t), C(t)) \Sigma^{1/2}\right) & \tilde{P}_\ell(t) &:= \tilde{p}_\ell(t) I \\
\tilde{u}_\ell(t) &:= -\frac{1}{d} \operatorname{Tr}\left(\nabla_{B_\ell} f(r(t), A(t), B(t), C(t)) \Sigma^{1/2}\right) & \tilde{U}_\ell(t) &:= \tilde{u}_\ell(t) \Sigma^{-1/2} \\
\tilde{w}_\ell(t) &:= -\frac{1}{d} \operatorname{Tr}\left(\nabla_{C_\ell} f(r(t), A(t), B(t), C(t)) \Sigma^{1/2}\right) & \tilde{W}_\ell(t) &:= \tilde{w}_\ell(t) \Sigma^{-1/2}.
\end{aligned}$$

It follows by definition of $\tilde{P}$, $\tilde{U}$, and $\tilde{W}$ that $(A(t), B(t), C(t)) \in \mathcal{S}$ for all t. We will show that at any time t,

$$\begin{aligned}
&\frac{d}{dt} f(r(t), A(t), B(t), C(t)) \\
\leq &- \sum_{\ell=0}^{k} \left(\partial_{r_\ell} f(r(t), A(t), B(t), C(t))\right)^2 \\
&\quad - \sum_{\ell=0}^{k} \|\nabla_{A_\ell} f(r(t), A(t), B(t), C(t))\|_F^2 \\
&\quad - \sum_{\ell=0}^{k} \|\nabla_{B_\ell} f(r(t), A(t), B(t), C(t))\|_F^2 - \sum_{\ell=0}^{k} \|\nabla_{C_\ell} f(r(t), A(t), B(t), C(t))\|_F^2 .
\end{aligned} \tag{48}$$

Let $\langle A, B\rangle_{\mathrm{Tr}} := \mathrm{Tr}\left(A^\top B\right)$. By definition of the dynamics in (47),

$$\frac{d}{dt} f(r(t), A(t), B(t), C(t)) \tag{49}$$

$$= \sum_{\ell=0}^{k} \partial_{r_\ell} f(r(t), A(t), B(t), C(t)) \cdot (-\partial_{r_\ell} f(r(t), A(t), B(t), C(t))) \tag{50}$$

$$+ \sum_{\ell=0}^{k} \left\langle \nabla_{A_\ell} f(r(t), A(t), B(t), C(t)), \tilde{P}_\ell(t) \right\rangle_{\mathrm{Tr}} \tag{51}$$

$$+ \sum_{\ell=0}^{k} \left\langle \nabla_{B_\ell} f(r(t), A(t), B(t), C(t)), \tilde{U}_\ell(t) \right\rangle_{\mathrm{Tr}} \tag{52}$$

$$+ \sum_{\ell=0}^{k} \left\langle \nabla_{C_\ell} f(r(t), A(t), B(t), C(t)), \tilde{W}_\ell(t) \right\rangle_{\mathrm{Tr}}. \tag{53}$$

We immediately verify that (50) $= -\sum_{\ell=0}^{k} \left(\partial r_\ell f(r(t), A(t), B(t), C(t))\right)^2$. By (71) from Proposition F.4, applied separately to each layer $\ell = 0 \dots k$,

$$\begin{aligned}(51) &\leq \sum_{\ell=0}^{k} \langle \nabla_{A_\ell} f(r(t), A(t), B(t), C(t)), -\nabla_{A_\ell} f(r(t), A(t), B(t), C(t))\rangle_{\mathrm{Tr}} \\ &= -\sum_{\ell=0}^{k} \|\nabla_{B_\ell} f(r(t), A(t), B(t), C(t))\|_F^2 .\end{aligned}$$

By (56) from Proposition F.3, applied separately to each layer $\ell = 0 \dots k$,

$$\begin{aligned}(52) &\leq \sum_{\ell=0}^{k} \langle \nabla_{B_\ell} f(r(t), A(t), B(t), C(t)), -\nabla_{B_\ell} f(r(t), A(t), B(t), C(t))\rangle_{\mathrm{Tr}} \\ &= -\sum_{\ell=0}^{k} \|\nabla_{B_\ell} f(r(t), A(t), B(t), C(t))\|_F^2 .\end{aligned}$$

Similarly, by (57) from Proposition F.3, applied separately to each layer $\ell = 0 \dots k$,

$$\begin{aligned}(53) &\leq \sum_{\ell=0}^{k} \langle \nabla_{C_\ell} f(r(t), A(t), B(t), C(t)), -\nabla_{C_\ell} f(r(t), A(t), B(t), C(t))\rangle_{\mathrm{Tr}} \\ &= -\sum_{\ell=0}^{k} \|\nabla_{C_\ell} f(r(t), A(t), B(t), C(t))\|_F^2 .\end{aligned}$$

Combining the above bounds gives (48). Suppose (46) does not hold. Then there exists a positive constant $c > 0$ such that for all t,

$$\begin{aligned}&\sum_{\ell=0}^{k} \left(\partial_{r_\ell} f(r(t), A(t), B(t), C(t))\right)^2 + \|\nabla_{A_\ell} f(r(t), A(t), B(t), C(t))\|_F^2 \\ &+ \|\nabla_{B_\ell} f(r(t), A(t), B(t), C(t))\|_F^2 + \|\nabla_{C_\ell} f(r(t), A(t), B(t), C(t))\|_F^2 \geq c.\end{aligned}$$

Then by (48), $\frac{d}{dt} f(r(t), A(t), B(t), C(t)) \leq -c$ for all t. This contradicts the fact that $f(\cdot)$ is bounded below by 0 (see (5)). Thus we prove (46).

F.2. Key Lemmas

Proposition F.3. *Let $\tilde{h}$ satisfy Assumption 4.3, let $x^{(i)}$'s satisfy Assumption 4.1 with matrix Σ, and $y^{(i)}$'s satisfy Assumption 4.2. Let $(A, B, C) \in \mathbb{R}^{(k+1)\times d\times d\times 3}$ satisfy, for all $\ell = 0 \dots k$,*

$$A_\ell = a_\ell I \qquad B_\ell = b_\ell \Sigma^{-1/2} \qquad C_\ell = c_\ell \Sigma^{-1/2}, \tag{54}$$

where $a_\ell, b_\ell, c_\ell \in \mathbb{R}$ are scalars. Let $V \in \mathbb{R}^{(k+1)\times(d+1)\times(d+1)}$ satisfy, for all $\ell = 0 \dots k$, $V_\ell = \begin{bmatrix} A_\ell & 0 \\ 0 & r_\ell \end{bmatrix}$, where r_ℓ are arbitrary scalars. Let $j \in \{0 \dots k\}$ be an arbitrary but fixed layer index. For $S \in \mathbb{R}^{d\times d}$, let $B_j(S) := B_j + S$, and let $B_\ell(S) := B_\ell$ for $\ell \neq j$. Let $B(S) := \{B_\ell(S)\}_{\ell=0\dots k}$. Recall $f(V, B, C)$ as defined in (5). Let $R \in \mathbb{R}^{d\times d}$ be an arbitrary matrix. Let

$$\tilde{r} := \frac{1}{d} \operatorname{Tr}\left(R\Sigma^{1/2}\right) \qquad \tilde{R} := \tilde{r}\Sigma^{-1/2} \tag{55}$$

Then

$$\frac{d}{dt} f(V, B(tR), C)\Big|_{t=0} \leq \frac{d}{dt} f(V, B(t\tilde{R}), C)\Big|_{t=0}. \tag{56}$$

Similarly, let $C_j(S) := C_j + S$, and $C_\ell(S) := C_\ell$ for $\ell \neq j$, and let $C(S) := \{C_\ell(S)\}_{\ell=0\dots k}$, then

$$\frac{d}{dt} f(V, B, C(tR))\Big|_{t=0} \leq \frac{d}{dt} f(V, B, C(t\tilde{R}))\Big|_{t=0}. \tag{57}$$

Proof of Proposition F.3. The proof of (57) is identical to that of (56), so we only present the proof of (56).

Loss Reformulation: Let us consider the reformulation of the in-context loss f presented in Lemma 5. Specifically, let $\overline{Z}_0$ be defined as

$$\overline{Z}_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \dots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \dots & y^{(n)} & y^{(n+1)} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)},$$

Let $\overline{Z}_\ell$ denote the output of the $(i-1)^{th}$ layer of the linear transformer (as defined in (3), initialized at $\overline{Z}_0$). For the rest of this proof, we will drop the bar, and simply denote $\overline{Z}_\ell$ by Z_ℓ. Let $X_\ell \in \mathbb{R}^{d\times(n+1)}$ denote the first d rows of Z_ℓ and let $Y_\ell \in \mathbb{R}^{1\times(n+1)}$ denote the $(d+1)^{th}$ row of Z_k. Under the theorem's assumption that $V_\ell = \begin{bmatrix} A_\ell & 0 \\ 0 & r_\ell \end{bmatrix}$, we verify that, for any $\ell \in \{0 \dots k\}$,

$$\begin{aligned} X_{\ell+1} &= X_\ell + A_\ell X_\ell M \tilde{h}\left(B_\ell X_\ell, C_\ell X_\ell\right) \\ Y_{\ell+1} &= Y_\ell + r_\ell Y_\ell M \tilde{h}\left(B_\ell X_\ell, B_\ell X_\ell\right) = Y_0 \prod_{\ell=0}^{i} \left(I + r_\ell M \tilde{h}\left(B_\ell X_0, C_\ell X_0\right)\right). \end{aligned} \tag{58}$$

By Lemma 5, the in-context loss defined in (5) is equivalent to

$$f(V, B, C) = \mathbb{E}_{Z_0}\left[\operatorname{Tr}\left((I - M)\, Y_{k+1}^\top Y_{k+1}\, (I - M)\right)\right]$$

We will introduce one more piece of notation: Following (58), notice that for any layer i, X_i is a function of A, B, C, X_0. Since for this part of the proof, only B is variable (function of S), and A, C are fixed, we define $X_i(X, S)$ to be "the result of evolving as (58), initialized at $X_0 = X$, where B_i is replaced by $B_i(S)$", i.e.

$$X_{i+1}(X, S) = X_i(X, S) + A_i X_i(X, S) M \tilde{h}\left(B_i(S) X_i(X, S), C_i X_i(X, S)\right) \tag{59}$$

Let us also define

$$G_i(X, S) := \prod_{\ell=0}^{i} \left(I + r_\ell M \tilde{h}\left(B_\ell(S) X_\ell(X, S), C_\ell X_\ell(X, S)\right)\right),$$

so that

$$\begin{aligned} f(V, B(S), C) =& \mathbb{E}_{Z_0}\left[\mathrm{Tr}\left((I-M)\, G_k(X,S)^\top Y_0^\top Y_0 G_k(X,S)\,(I-M)\right)\right] \\ =& \mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\, G_k(X,S)^\top \mathbb{K} G_k(X,S)\,(I-M)\right)\right], \end{aligned}$$

where recall that $\mathbb{K} \in \mathbb{R}^{(n+1)\times(n+1)}$ and $\mathbb{K}_{ij} = \mathcal{K}(\Sigma^{-1/2}x^{(i)}, \Sigma^{-1/2}x^{(j)})$ as defined in Assumption 4.2. The second equality uses the assumption on distribution of Y_0 conditioned on X_0, as specified in Assumption 4.2. Let U denote a uniformly randomly sampled orthogonal matrix. Let $U_\Sigma := \Sigma^{1/2} U \Sigma^{-1/2}$, so that $U_\Sigma^{-1} = \Sigma^{1/2} U^\top \Sigma^{-1/2}$. We will repeatedly use the following identities:

$$\begin{aligned} B_i U_\Sigma &= b_i \Sigma^{-1/2}\Sigma^{1/2} U \Sigma^{-1/2} = U B_i \\ C_i U_\Sigma &= c_i \Sigma^{-1/2}\Sigma^{1/2} U \Sigma^{-1/2} = U C_i \end{aligned} \tag{60}$$

Using the fact that $X_0 \overset{d}{=} U_\Sigma X_0$, we can verify

$$\begin{aligned} \frac{d}{dt} f(V, B(tR), C)\Big|_{t=0} =& \frac{d}{dt}\, \mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\, G_k(X_0, tR)^\top \mathbb{K}(X_0) G_k(X_0, tR)\,(I-M)\right)\right]\Big|_{t=0} \\ =& 2\, \mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\, G_k(X_0,0)^\top \mathbb{K}(X_0)\, \frac{d}{dt} G_k(X_0, tR)\Big|_{t=0} (I-M)\right)\right] \\ =& 2\, \mathbb{E}_{X_0, U}\left[\mathrm{Tr}\left((I-M)\, G_k(U_\Sigma X_0,0)^\top \mathbb{K}(X_0)\, \frac{d}{dt} G_k(U_\Sigma X_0, tR)\Big|_{t=0} (I-M)\right)\right]. \end{aligned} \tag{61}$$

The last equality uses the fact that $\mathbb{K}(U_\Sigma X_0) = \mathbb{K}(X_0)$ by Assumption 4.2.

Henceforth, assume all $\frac{d}{dt}$ occurs at $t = 0$, and we somtimes drop the explicit $|_{t=0}$ notation to save space.

X_i and $\frac{d}{dt} X_i$ under random transformation of X_0

In this part of the proof, we establish two important identities about the evolution of X_i under random rotation of its arguments:

$$X_i\left(U_\Sigma X_0, 0\right) = U_\Sigma X_i\left(X_0, 0\right), \tag{62}$$

$$\frac{d}{dt} X_i\left(U_\Sigma X_0, tR\right)\Big|_{t=0} = U_\Sigma\, \frac{d}{dt} X_i\left(X_0, tU^\top R U_\Sigma\right)\Big|_{t=0}. \tag{63}$$

We first verify (62) by induction. For $i = 0$, this identity holds by definition. Assume the identity holds for some i. Then following (59),

$$\begin{aligned} X_{i+1}(U_\Sigma X_0, 0) =& X_i(U_\Sigma X_0, 0) + A_i X_i(U_\Sigma X_0, 0) M \tilde{h}\left(B_i X_i(U_\Sigma X_0, 0), C_i X_i(U_\Sigma X_0, 0)\right) \\ =& U_\Sigma X_i(X_0, 0) + U_\Sigma A_i X_i(X_0, 0) M \tilde{h}\left(B_i U_\Sigma X_i(X_0, 0), C_i U_\Sigma X_i(X_0, 0)\right) \\ =& U_\Sigma X_i(X_0, 0) + U_\Sigma A_i X_i(X_0, 0) M \tilde{h}\left(B_i X_i(X_0, 0), C_i X_i(X_0, 0)\right) \\ =& U_\Sigma X_{i+1}(X_0, 0). \end{aligned}$$

The second equality is by the inductive hypothesis, and the fact that $A_i = a_i I$. The third equality uses (60) and Assumption 4.3.

Next, we verify (63). By definition of $X_i(X_0, S)$, the case for $i \le j$ is simple:

$$\frac{d}{dt} X_i(U_\Sigma X_0, tR)\Big|_{t=0} = 0 = U_\Sigma\, \frac{d}{dt} X_i(X_0, tU^\top R U_\Sigma)\Big|_{t=0}. \tag{64}$$

For $i = j+1$, it follows from (59) and chain rule that

$$\begin{aligned}
&\frac{d}{dt}X_{j+1}(U_\Sigma X_0, tR) \\
=&\frac{d}{dt}X_j(U_\Sigma X_0, tR) + A_j\left(\frac{d}{dt}X_j(U_\Sigma X_0, tR)\right) M\tilde{h}\left(B_j X_j(U_\Sigma X_0, 0), C_j X_j(U_\Sigma X_0, 0)\right) \\
&+ A_j X_j(U_\Sigma X_0, 0) M \frac{d}{dt}\tilde{h}\left(\underbrace{(B_j + tR)\, X_j(U_\Sigma X_0, tR)}_{S(t)}, \underbrace{C_j X_j(U_\Sigma X_0, tR)}_{T(t)}\right). \qquad (65)
\end{aligned}$$

We will now apply Lemma F.5. Let $S(t) := (B_j + tR)\, U_\Sigma X_j(U_\Sigma X_0, tR)$ and $T(t) := C_j U_\Sigma X_j(U_\Sigma X_0, tR)$. By (64), we know that $\frac{d}{dt}X_j(U_\Sigma X_0, tR)\big|_{t=0} = 0$. Thus, we can define $\tilde{S}(t) := (B_j + tR)\, X_j(U_\Sigma X_0, 0)$ and $\tilde{T}(t) := C_j X_j(U_\Sigma X_0, 0)$. Using (62) and (60), we verify that

$$\begin{aligned}
\tilde{S}(t) &= (B_j + tR)\, U_\Sigma X_j(X_0, 0) = U\left(B_j + tU^\top R U_\Sigma\right) X_j(X_0, 0) \\
\tilde{T}(t) &= (C_j + tR)\, U_\Sigma X_j(X_0, 0) = U\left(C_j + tU^\top R U_\Sigma\right) X_j(X_0, 0).
\end{aligned}$$

Let us therefore pick $\Gamma := U$. Applying Lemma F.5 and plugging into (65) gives

$$\begin{aligned}
&\frac{d}{dt}X_{j+1}(U_\Sigma X_0, tR) \\
=&U_\Sigma \frac{d}{dt}X_j\left(X_0, tU^\top R U_\Sigma\right) + A_j U_\Sigma\left(\frac{d}{dt}X_j\left(X_0, tU^\top R U_\Sigma\right)\right) M\tilde{h}\left(B_j X_j(X_0, 0), C_j X_j(X_0, 0)\right) \\
&+ A_j U_\Sigma X_j(X_0, 0) M \frac{d}{dt}\tilde{h}\left(\underbrace{\left(B_j + tU^\top R U_\Sigma\right) X_j(X_0, 0)}_{\Gamma^\top \tilde{S}(t)}, \underbrace{C_j X_j(X_0, 0)}_{\Gamma^{-1}\tilde{T}(t)}\right) \\
=&U_\Sigma \frac{d}{dt}X_{j+1}(X_0, tU^\top R U_\Sigma),
\end{aligned}$$

where the first equality also uses (60) and (62) and (64).

Finally, we need to prove (63) for the $i > j+1$ case. We will prove this by induction over i. The proof is very similar to the $i = j+1$ case:

$$\begin{aligned}
&\frac{d}{dt}X_{i+1}(U_\Sigma X_0, tR) \\
=&\frac{d}{dt}X_i(U_\Sigma X_0, tR) + A_i\left(\frac{d}{dt}X_i(U_\Sigma X_0, tR)\right) M\tilde{h}\left(B_i X_i(U_\Sigma X_0, 0), C_i X_i(U_\Sigma X_0, 0)\right) \\
&+ A_i X_i(U_\Sigma X_0, 0) M \frac{d}{dt}\tilde{h}\left(\underbrace{B_i X_i(U_\Sigma X_0, tR)}_{S(t)}, \underbrace{C_i X_i(U_\Sigma X_0, tR)}_{T(t)}\right) \\
=&\frac{d}{dt}X_i(U_\Sigma X_0, tR) + A_i\left(\frac{d}{dt}X_i(U_\Sigma X_0, tR)\right) M\tilde{h}\left(B_i X_i(U_\Sigma X_0, 0), C_i X_i(U_\Sigma X_0, 0)\right) \\
&+ A_i X_i(U_\Sigma X_0, 0) M \frac{d}{dt}\tilde{h}\left(U\underbrace{B_i X_i(X_0, tU^\top R U_\Sigma)}_{\tilde{S}(t)}, U\underbrace{C_i X_i(X_0, tU^\top R U_\Sigma)}_{\tilde{T}(t)}\right) \\
=&U_\Sigma \frac{d}{dt}X_{i+1}(X_0, tU^\top R U_\Sigma).
\end{aligned}$$

In the second equality, we apply Lemma F.5 with $\Gamma = U$. We use the inductive hypothesis to verify that $\tilde{S}'(0) = S'(0)$ and $\tilde{T}'(0) = T'(0)$. This concludes the proof of (63).

G and $\frac{d}{dt}G$ under random transformation of X_0

In this part of the proof, we establish two important identities about the evolution of G under random rotation of its arguments:

$$G_i(U_\Sigma X_0, 0) = G_i(X_0, 0), \tag{66}$$

$$\frac{d}{dt} G_i\left(U_\Sigma X_0, tR\right)\Big|_{t=0} = \frac{d}{dt} G_i\left(X_0, tU^\top R U_\Sigma\right)\Big|_{t=0}. \tag{67}$$

(66) is an immediate consequence of (62):

$$\begin{aligned} G_i(U_\Sigma X_0, 0) :=& \prod_{\ell=0}^{i} \left(I + r_\ell M \tilde{h}\left(B_\ell X_\ell(U_\Sigma X_0, 0), C_\ell X_\ell(U_\Sigma X_0, 0)\right)\right) \\ =& \prod_{\ell=0}^{i} \left(I + r_\ell M \tilde{h}\left(B_\ell X_\ell(X_0, 0), C_\ell X_\ell(X_0, 0)\right)\right) \\ =& G_i(X_0, 0), \end{aligned}$$

where the second equality uses (62), (60) and Assumption 4.3.

To verify (67), we first verify the following recursive relationship:

$$\begin{aligned} & G_i\left(U_\Sigma X_0, S\right) \\ =& \left(I + r_i M \tilde{h}\left(B_i(S) X_i(U_\Sigma X_0, S), C_i X_i(U_\Sigma X_0, S)\right)\right) G_{i-1}\left(U_\Sigma X_0, S\right) \\ \Rightarrow \quad & \frac{d}{dt} G_i\left(U_\Sigma X_0, tR\right)\Big|_{t=0} \\ =& \left(\frac{d}{dt}\left(I + r_i M \tilde{h}\left(B_i(tR) X_i(U_\Sigma X_0, tR), C_i X_i(U_\Sigma X_0, tR)\right)\right)\right) G_{i-1}\left(U_\Sigma X_0, tR\right) \\ & + \left(I + r_i M \tilde{h}\left(B_i X_i(U_\Sigma X_0, 0), C_i X_i(U_\Sigma X_0, 0)\right)\right) \frac{d}{dt} G_{i-1}\left(U_\Sigma X_0, tR\right). \end{aligned} \tag{68}$$

We will analyze the two terms in (68) separately:

$$\begin{aligned} & \frac{d}{dt}\left(I + r_i M \tilde{h}\left(B_i(tR) X_i(U_\Sigma X_0, tR), C_i X_i(U_\Sigma X_0, tR)\right)\right) \\ =& I + r_i M \frac{d}{dt} \tilde{h}\left(B_i(tR) X_i(U_\Sigma X_0, tR), C_i X_i(U_\Sigma X_0, tR)\right). \end{aligned}$$

Let $S(t) := B_i(tR) X_i(U_\Sigma X_0, tR)$ and $T(t) := C_i X_i(U_\Sigma X_0, tR)$. Let $\tilde{S}(t) := U B_i(tU^\top R U_\Sigma) X_i(X_0, tU^\top R U_\Sigma)$ and $\tilde{T}(t) := U C_i X_i(X_0, tU^\top R U_\Sigma)$. We verify that

$$\begin{aligned} S(0) =& B_i X_i(U_\Sigma X_0, 0) = U B_i\left(X_0, 0\right) = \tilde{S}(0) \\ T(0) =& C_i X_i(U_\Sigma X_0, 0) = U C_i\left(X_0, 0\right) = \tilde{T}(0). \end{aligned}$$

By chain rule,

$$\begin{aligned} S'(0) =& \left(\frac{d}{dt} B_i(tR)\right) X_i(U_\Sigma X_0, 0) + B_i \frac{d}{dt} X_i(U_\Sigma X_0, tR) \\ =& \left(\frac{d}{dt} B_i(tR)\right) U_\Sigma X_i(X_0, 0) + B_i U_\Sigma \frac{d}{dt} X_i(X_0, tU^\top R U_\Sigma) \\ =& U\left(\frac{d}{dt} B_i(tU^\top R U_\Sigma)\right) X_i(X_0, 0) + U B_i \frac{d}{dt} X_i(X_0, tU^\top R U_\Sigma) \\ =& \tilde{S}'(0). \end{aligned}$$

The second equality follows from (62) and (63). The third equality uses (60), as well as the fact that $U^\top \frac{d}{dt} B_i(tR) U_\Sigma = \frac{d}{dt} B_i(tU^\top R U_\Sigma)$; this is because for $i \neq j$, both sides are 0, and for $i = j$, $\frac{d}{dt} B_j(tR) = R$.

Similarly, we verify that

$$\begin{aligned} T'(0) =& C_i \frac{d}{dt} X_i(U_\Sigma X_0, tR) \\ =& U C_i \frac{d}{dt} X_i(X_0, tU^\top R U_\Sigma) \\ =& \tilde{T}'(0). \end{aligned}$$

Applying Lemma F.5 with $\Gamma = U$ gives

$$\begin{aligned} & \frac{d}{dt}\left(I + r_i M \tilde{h}\left(B_i(tR) X_i(U_\Sigma X_0, tR), C_i X_i(U_\Sigma X_0, tR)\right)\right) \\ =& \frac{d}{dt}\left(I + r_i M \tilde{h}\left(B_i(tU^\top R U_\Sigma) X_i(X_0, tU^\top R U_\Sigma), C_i X_i(X_0, tU^\top R U_\Sigma)\right)\right). \end{aligned}$$

Using (60) and Assumption 4.3 and the inductive hypothesis, the second term of (68) satisfies

$$\begin{aligned} & \left(I + r_i M \tilde{h}\left(B_i X_i(U_\Sigma X_0, 0), C_i X_i(U_\Sigma X_0, 0)\right)\right) \frac{d}{dt} G_{i-1}\left(U_\Sigma X_0, tR\right) \\ =& \left(I + r_i M \tilde{h}\left(B_i X_i(X_0, 0), C_i X_i(X_0, 0)\right)\right) \frac{d}{dt} G_{i-1}\left(X_0, tU^\top R U_\Sigma\right). \end{aligned}$$

Combining the above identities for each term of (68), we conclude that

$$\frac{d}{dt} G_i\left(U_\Sigma X_0, tR\right)\Big|_{t=0} = \frac{d}{dt} G_i\left(X_0, tU^\top R U_\Sigma\right)\Big|_{t=0}.$$

This concludes the proof of (67).

Putting everything together:
We will now conclude the proof of (56). Plugging in (66) and (67) into (61) gives

$$\begin{aligned} \frac{d}{dt} f(V, B(tR), C)\Big|_{t=0} =& 2\,\mathbb{E}_{X_0, U}\left[\mathrm{Tr}\left((I - M)\, G_k(U_\Sigma X_0, 0)^\top \mathbb{K}\, \frac{d}{dt} G_k(U_\Sigma X_0, tR)\Big|_{t=0} (I - M)\right)\right] \\ =& 2\,\mathbb{E}_{X_0, U}\left[\mathrm{Tr}\left((I - M)\, G_k(X_0, 0)^\top \mathbb{K}\, \frac{d}{dt} G_k(X_0, tU^\top R U_\Sigma)\Big|_{t=0} (I - M)\right)\right] \\ =& 2\,\mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I - M)\, G_k(X_0, 0)^\top \mathbb{K}\, \frac{d}{dt} G_k(X_0, t\,\mathbb{E}_U\left[U^\top R U_\Sigma\right])\Big|_{t=0} (I - M)\right)\right] \\ =& 2\,\mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I - M)\, G_k(X_0, 0)^\top \mathbb{K}\, \frac{d}{dt} G_k(X_0, t\tilde{R})\Big|_{t=0} (I - M)\right)\right] \\ =& \frac{d}{dt} f(V, B(t\tilde{R}), C)\Big|_{t=0} \end{aligned}$$

The third equality uses the fact that $\frac{d}{dt} G_k(X_0, tS)\big|_{t=0}$ is linear in S for any S. The fourth equality is by (55). This concludes the proof of (56).

□

Proposition F.4. *Let $\tilde{h}$ satisfy Assumption 4.3, let $x^{(i)}$'s satisfy Assumption 4.1 with matrix Σ, and $y^{(i)}$'s satisfy Assumption 4.2. Let $(A, B, C) \in \mathbb{R}^{(k+1)\times d\times d\times 3}$ satisfy, for all $\ell = 0 \ldots k$,*

$$A_\ell = a_\ell I \qquad B_\ell = b_\ell \Sigma^{-1/2} \qquad C_\ell = c_\ell \Sigma^{-1/2}, \tag{69}$$

where $a_\ell, b_\ell, c_\ell \in \mathbb{R}$ are scalars. Let $V \in \mathbb{R}^{(k+1)\times(d+1)\times(d+1)}$ satisfy, for all $\ell = 0 \dots k$, $V_\ell = \begin{bmatrix} A_\ell & 0 \\ 0 & r_\ell \end{bmatrix}$, where r_ℓ are arbitrary scalars. Let $j \in \{0 \dots k\}$ be an arbitrary but fixed layer index. For $S \in \mathbb{R}^{d\times d}$, let $A_j(S) := A_j + S$, and let $A_\ell(S) := A_\ell$ for $\ell \neq j$. Let $A(S) := \{A_\ell(S)\}_{\ell=0\dots k}$. Let $V_\ell(S) := \left\{ \begin{bmatrix} A_\ell(S) & 0 \\ 0 & r_\ell \end{bmatrix} \right\}$ and $V(S) = \{V_\ell(S)\}_{\ell=0\dots k}$. Let $f(V, B, C)$ e as defined in (5). Let $R \in \mathbb{R}^{d\times d}$ be an arbitrary matrix. Let

$$\tilde{r} := \frac{1}{d} \operatorname{Tr}\left(\Sigma^{-1/2} R \Sigma^{1/2}\right) \qquad \tilde{R} := \tilde{r} I \tag{70}$$

Then

$$\frac{d}{dt} f(V(tR), B, C)\Big|_{t=0} \leq \frac{d}{dt} f(V(t\tilde{R}), B, C)\Big|_{t=0}. \tag{71}$$

Proof. Proof of Proposition F.4.

Loss Reformulation: Let us consider the reformulation of the in-context loss f presented in Lemma 5. Specifically, let $\overline{Z}_0$ be defined as

$$\overline{Z}_0 = \begin{bmatrix} x^{(1)} & x^{(2)} & \dots & x^{(n)} & x^{(n+1)} \\ y^{(1)} & y^{(2)} & \dots & y^{(n)} & y^{(n+1)} \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)},$$

Let $\overline{Z}_i$ denote the output of the $(i-1)^{th}$ layer of the linear transformer (as defined in (3), initialized at $\overline{Z}_0$). For the rest of this proof, we will drop the bar, and simply denote $\overline{Z}_i$ by Z_i. Let $X_i \in \mathbb{R}^{d\times(n+1)}$ denote the first d rows of Z_i and let $Y_i \in \mathbb{R}^{1\times(n+1)}$ denote the $(d+1)^{th}$ row of Z_k. Under the assumption that $V_\ell = \begin{bmatrix} A_\ell & 0 \\ 0 & r_\ell \end{bmatrix}$, we verify that for all $i \in \{0 \dots k\}$:

$$\begin{aligned} X_{i+1} &= X_i + A_i X_i M \tilde{h}\left(B_i X_i, C_i X_i\right) \\ Y_{i+1} &= Y_i + r_i Y_i M \tilde{h}\left(B_i X_i, B_i X_i\right) = Y_0 \prod_{\ell=0}^{i} \left(I + r_\ell M \tilde{h}\left(B_\ell X_0, C_\ell X_0\right)\right). \end{aligned} \tag{72}$$

By Lemma 5, the in-context loss defined in (5) is equivalent to

$$f(V, B, C) = \mathbb{E}_{Z_0}\left[\operatorname{Tr}\left((I - M) Y_{k+1}^\top Y_{k+1} (I - M)\right)\right]$$

We will introduce one more piece of notation: Following (72), notice that for any layer i, X_i is a function of A, B, C, X_0. Since for this part of the proof, only A is variable (function of S), and B, C are fixed, we define $X_i(X, S)$ to be "the result of evolving as (72), initialized at $X_0 = X$, where A_i is replaced by $A_i(S)$", i.e.

$$X_{i+1}(X, S) = X_i(X, S) + A_i(S) X_i(X, S) M \tilde{h}\left(B_i X_i(X, S), C_i X_i(X, S)\right) \tag{73}$$

Let us also define

$$G_i(X, S) := \prod_{\ell=0}^{i} \left(I + r_\ell M \tilde{h}\left(B_\ell X_\ell(X, S), C_\ell X_\ell(X, S)\right)\right),$$

so that

$$\begin{aligned} f(V(S), B, C) &= \mathbb{E}_{Z_0}\left[\operatorname{Tr}\left((I - M) G_k(X, S)^\top Y_0^\top Y_0 G_k(X, S) (I - M)\right)\right] \\ &= \mathbb{E}_{X_0}\left[\operatorname{Tr}\left((I - M) G_k(X, S)^\top \mathbb{K} G_k(X, S) (I - M)\right)\right], \end{aligned}$$

where recall that $\mathbb{K} \in \mathbb{R}^{(n+1)\times(n+1)}$ and $\mathbb{K}_{ij} = \mathcal{K}(\Sigma^{-1/2} x^{(i)}, \Sigma^{-1/2} x^{(j)})$ as defined in Assumption 4.2. The second equality uses the assumption on distribution of Y_0 conditioned on X_0, as specified in Assumption 4.2. Let U denote

a uniformly randomly sampled orthogonal matrix. Let $U_\Sigma := \Sigma^{1/2} U \Sigma^{-1/2}$, so that $U_\Sigma^{-1} = \Sigma^{1/2} U^\top \Sigma^{-1/2}$. We will repeatedly use the following identities:

$$\begin{aligned} B_i U_\Sigma &= b_i \Sigma^{-1/2} \Sigma^{1/2} U \Sigma^{-1/2} = U B_i \\ C_i U_\Sigma &= c_i \Sigma^{-1/2} \Sigma^{1/2} U \Sigma^{-1/2} = U B_i \end{aligned} \tag{74}$$

Using the fact that $X_0 \overset{d}{=} U_\Sigma X_0$, we can verify

$$\begin{aligned} \frac{d}{dt} f(V(tR), B, C)\Big|_{t=0} =& \frac{d}{dt} \mathbb{E}_{X_0} \left[\mathrm{Tr}\left((I-M)\, G_k(X_0, tR)^\top \mathbb{K}(X_0) G_k(X_0, tR)\, (I-M) \right) \right] \Big|_{t=0} \\ =& 2\, \mathbb{E}_{X_0} \left[\mathrm{Tr}\left((I-M)\, G_k(X_0, 0)^\top \mathbb{K}(X_0)\, \frac{d}{dt} G_k(X_0, tR)\Big|_{t=0} (I-M) \right) \right] \\ =& 2\, \mathbb{E}_{X_0, U} \left[\mathrm{Tr}\left((I-M)\, G_k(U_\Sigma X_0, 0)^\top \mathbb{K}(X_0)\, \frac{d}{dt} G_k(U_\Sigma X_0, tR)\Big|_{t=0} (I-M) \right) \right]. \end{aligned} \tag{75}$$

The last equality uses the fact that $\mathbb{K}(U_\Sigma X_0) = \mathbb{K}(X_0)$ by Assumption 4.2.

Henceforth, assume all $\frac{d}{dt}$ occurs at $t = 0$, and we somtimes drop the explicit $|_{t=0}$ notation to save space.

X_i and $\frac{d}{dt} X_i$ under random transformation of X_0

In this part of the proof, we establish two important identities about the evolution of X_i under random rotation of its arguments:

$$X_i\left(U_\Sigma X_0, 0\right) = U_\Sigma X_i\left(X_0, 0\right), \tag{76}$$

$$\frac{d}{dt} X_i\left(U_\Sigma X_0, tR\right)\Big|_{t=0} = U_\Sigma\, \frac{d}{dt} X_i\left(X_0, t U_\Sigma^{-1} R U_\Sigma\right)\Big|_{t=0}. \tag{77}$$

We first verify (76) by induction. For $i = 0$, this identity holds by definition. Assume the identity holds for some i. Then following (59),

$$\begin{aligned} X_{i+1}(U_\Sigma X_0, 0) =& X_i(U_\Sigma X_0, 0) + A_i(0) X_i(U_\Sigma X_0, S) M \tilde{h}\left(B_i X_i(U_\Sigma X_0, 0), C_i X_i(U_\Sigma X_0, 0)\right) \\ =& U_\Sigma X_i(X_0, 0) + U_\Sigma A_i(0) X_i(X_0, S) M \tilde{h}\left(B_i U_\Sigma X_i(X_0, 0), C_i U_\Sigma X_i(X_0, 0)\right) \\ =& U_\Sigma X_i(X_0, 0) + U_\Sigma A_i(0) X_i(X_0, S) M \tilde{h}\left(B_i X_i(X_0, 0), C_i X_i(X_0, 0)\right) \\ =& U_\Sigma X_{i+1}(X_0, 0). \end{aligned}$$

The second equality is by the inductive hypothesis, and the fact that $A_i = a_i I$. The third equality uses (60) and Assumption 4.3.

Next, we verify (63). By definition of $X_i(X_0, S)$, the case for $i \le j$ is simple:

$$\frac{d}{dt} X_i(U_\Sigma X_0, tR)\Big|_{t=0} = 0 = U_\Sigma\, \frac{d}{dt} X_i(X_0, t U_\Sigma^{-1} R U_\Sigma)\Big|_{t=0}. \tag{78}$$

For $i = j + 1$, it follows from (59) and chain rule that

$$\begin{aligned} & \frac{d}{dt} X_{j+1}(U_\Sigma X_0, tR) \\ =& \frac{d}{dt} X_j(U_\Sigma X_0, tR) + \left(\frac{d}{dt} A_j(tR) \right) X_j(U_\Sigma X_0, 0) M \tilde{h}\left(B_j X_j(U_\Sigma X_0, 0), C_j X_j(U_\Sigma X_0, 0)\right)) \\ & + A_j(0) \left(\frac{d}{dt} X_j(U_\Sigma X_0, tR) \right) M \tilde{h}\left(B_j X_j(U_\Sigma X_0, 0), C_j X_j(U_\Sigma X_0, 0)\right)) \\ & + A_j(0) X_j(U_\Sigma X_0, 0) M \frac{d}{dt} \tilde{h}\left(B_j X_j(U_\Sigma X_0, 0), C_j X_j(U_\Sigma X_0, 0)\right). \end{aligned} \tag{79}$$

We can simplify each term on the RHS above separately:

By (74), the first, third and fourth terms are 0. By definition of A_j, $\frac{d}{dt}A_j(tR) = R$. Furthermore, using (78), (74) and Assumption 4.3, the second term simplifies to $U_\Sigma U_\Sigma^{-1} R U_\Sigma X_j(X_0, 0) M\tilde{h}\left(B_j X_j(X_0, 0), C_j X_j(X_0, 0)\right))$. Therefore,

$$\begin{aligned}\frac{d}{dt}X_{j+1}(U_\Sigma X_0, tR) =& U_\Sigma U_\Sigma^{-1} R U_\Sigma X_j(X_0, 0) M\tilde{h}\left(B_j X_j(X_0, 0), C_j X_j(X_0, 0)\right))\\ =& U_\Sigma \frac{d}{dt}X_{j+1}(X_0, tU_\Sigma^{-1} R U_\Sigma).\end{aligned}$$

We have thus verified (77) for $i \le j+1$. For $i > j+1$, we will use proof by induction. Assume (77) holds for all $\ell \le i$ for some $i \ge j+1$. Then for $i+1$,

$$\begin{aligned}&\frac{d}{dt}X_{i+1}(U_\Sigma X_0, tR)\\ =&\frac{d}{dt}X_i(U_\Sigma X_0, tR) + \left(\frac{d}{dt}A_i(tR)\right) X_i(U_\Sigma X_0, 0) M\tilde{h}\left(B_i X_i(U_\Sigma X_0, 0), C_i X_i(U_\Sigma X_0, 0)\right))\\ &+ A_i(0)\left(\frac{d}{dt}X_i(U_\Sigma X_0, tR)\right) M\tilde{h}\left(B_i X_i(U_\Sigma X_0, 0), C_i X_i(U_\Sigma X_0, 0)\right))\\ &+ A_i(0) X_i(U_\Sigma X_0, 0) M \frac{d}{dt}\tilde{h}\left(\underbrace{B_i X_i(U_\Sigma X_0, tR)}_{S(t)}, \underbrace{C_i X_i(U_\Sigma X_0, tR)}_{T(t)}\right).\end{aligned}$$

Since $i \ge j+1$, we know that $\frac{d}{dt}A_i(tR) = 0$, so the second term on RHS is 0. By the inductive hypothesis and (77), and $A_i(0) = a_i I$, and Assumption 4.3, the third RHS term can be simplified to be $U_\Sigma A_i(0)\left(\frac{d}{dt}X_i\left(X_0, tU_\Sigma^{-1} R U_\Sigma\right)\right) M\tilde{h}\left(B_i X_i(X_0, 0), C_i X_i(X_0, 0)\right))$. Finally, to simplify the last RHS term, we apply Lemma F.5. Let $S(t) := B_i X_i(U_\Sigma X_0, tR)$ and $T(t) := C_i X_i(U_\Sigma X_0, tR)$. Let $\tilde{S}(t) := U B_i X_i\left(X_0, tU_\Sigma^{-1} R U_\Sigma\right)$ and $\tilde{T}(t) := U C_i X_i\left(X_0, tU_\Sigma^{-1} R U_\Sigma\right)$. Let $\Gamma := U$. Then $\frac{d}{dt}\tilde{h}\left(B_i X_i(U_\Sigma X_0, tR), C_i X_i(U_\Sigma X_0, tR)\right) = \frac{d}{dt}\tilde{h}\left(B_i X_i(X_0, tU_\Sigma^{-1} R U_\Sigma), C_i X_i(X_0, tU_\Sigma^{-1} R U_\Sigma)\right)$. Put together, we conclude that

$$\frac{d}{dt}X_{i+1}(U_\Sigma X_0, tR) = U_\Sigma \frac{d}{dt}X_{i+1}(X_0, tU_\Sigma^{-1} R U_\Sigma).$$

We thus complete the proof of (77).

G and $\frac{d}{dt}G$ under random transformation of X_0

In this part of the proof, we establish two important identities about the evolution of G under random rotation of its arguments:

$$G_i(U_\Sigma X_0, 0) = G_i(X_0, 0), \tag{80}$$

$$\frac{d}{dt}G_i\left(U_\Sigma X_0, tR\right)\Big|_{t=0} = \frac{d}{dt}G_i\left(X_0, tU_\Sigma^{-1} R U_\Sigma\right)\Big|_{t=0}. \tag{81}$$

(80) is an immediate consequence of (76):

$$\begin{aligned}G_i(U_\Sigma X_0, 0) :=& \prod_{\ell=0}^{i}\left(I + r_\ell M\tilde{h}\left(B_\ell X_\ell(U_\Sigma X_0, 0), C_\ell X_\ell(U_\Sigma X_0, 0)\right)\right)\\ =& \prod_{\ell=0}^{i}\left(I + r_\ell M\tilde{h}\left(B_\ell X_\ell(X_0, 0), C_\ell X_\ell(X_0, 0)\right)\right)\\ =& G_i(X_0, 0),\end{aligned}$$

where the second equality uses (76), (74) and Assumption 4.3.

To verify (81), we first verify the following recursive relationship:

$$\begin{aligned}
& G_i\left(U_\Sigma X_0, S\right) \\
=& \left(I + r_i M \tilde{h}\left(B_i(S) X_i(U_\Sigma X_0, S), C_i X_i(U_\Sigma X_0, S)\right)\right) G_{i-1}\left(U_\Sigma X_0, S\right) \\
\Rightarrow \quad & \frac{d}{dt} G_i\left(U_\Sigma X_0, tR\right)\Big|_{t=0} \\
=& \left(\frac{d}{dt}\left(I + r_i M \tilde{h}\left(B_i X_i(U_\Sigma X_0, tR), C_i X_i(U_\Sigma X_0, tR)\right)\right)\right) G_{i-1}\left(U_\Sigma X_0, tR\right) \\
& + \left(I + r_i M \tilde{h}\left(B_i X_i(U_\Sigma X_0, 0), C_i X_i(U_\Sigma X_0, 0)\right)\right) \frac{d}{dt} G_{i-1}\left(U_\Sigma X_0, tR\right). \qquad (82)
\end{aligned}$$

We will analyze the two terms in (82) separately:

$$\begin{aligned}
& \frac{d}{dt}\left(I + r_i M \tilde{h}\left(B_i X_i(U_\Sigma X_0, tR), C_i X_i(U_\Sigma X_0, tR)\right)\right) \\
=& I + r_i M \frac{d}{dt} \tilde{h}\left(B_i X_i(U_\Sigma X_0, tR), C_i X_i(U_\Sigma X_0, tR)\right).
\end{aligned}$$

Let $S(t) := B_i X_i(U_\Sigma X_0, tR)$ and $T(t) := C_i X_i(U_\Sigma X_0, tR)$. Let $\tilde{S}(t) := U B_i X_i(X_0, t U_\Sigma^{-1} R U_\Sigma)$ and $\tilde{T}(t) := U C_i X_i(X_0, t U_\Sigma^{-1} R U_\Sigma)$. We verify that

$$\begin{aligned}
S(0) =& B_i X_i(U_\Sigma X_0, 0) = U B_i\left(X_0, 0\right) = \tilde{S}(0) \\
T(0) =& C_i X_i(U_\Sigma X_0, 0) = U C_i\left(X_0, 0\right) = \tilde{T}(0) \\
S'(0) =& B_i \frac{d}{dt} X_i(U_\Sigma X_0, tR) = U B_i \frac{d}{dt} X_i(X_0, t U_\Sigma^{-1} R U_\Sigma) = \tilde{S}'(0) \\
T'(0) =& B_i \frac{d}{dt} X_i(U_\Sigma X_0, tR) = U B_i \frac{d}{dt} X_i(X_0, t U_\Sigma^{-1} R U_\Sigma) = \tilde{T}'(0),
\end{aligned}$$

where the last two equalities use (77) and (74). Applying Lemma F.5 with $\Gamma = U$ gives

$$\begin{aligned}
& \frac{d}{dt}\left(I + r_i M \tilde{h}\left(B_i(tR) X_i(U_\Sigma X_0, tR), C_i X_i(U_\Sigma X_0, tR)\right)\right) \\
=& \frac{d}{dt}\left(I + r_i M \tilde{h}\left(B_i X_i(X_0, t U_\Sigma^{-1} R U_\Sigma), C_i X_i(X_0, t U_\Sigma^{-1} R U_\Sigma)\right)\right).
\end{aligned}$$

Using (74) and Assumption 4.3 and the inductive hypothesis, the second term of (82) satisfies

$$\begin{aligned}
& \left(I + r_i M \tilde{h}\left(B_i X_i(U_\Sigma X_0, 0), C_i X_i(U_\Sigma X_0, 0)\right)\right) \frac{d}{dt} G_{i-1}\left(U_\Sigma X_0, tR\right) \\
=& \left(I + r_i M \tilde{h}\left(B_i X_i(X_0, 0), C_i X_i(X_0, 0)\right)\right) \frac{d}{dt} G_{i-1}\left(X_0, t U_\Sigma^{-1} R U_\Sigma\right).
\end{aligned}$$

Combining the above identities for each term of (82), we conclude that

$$\frac{d}{dt} G_i\left(U_\Sigma X_0, tR\right)\Big|_{t=0} = \frac{d}{dt} G_i\left(X_0, t U_\Sigma^{-1} R U_\Sigma\right)\Big|_{t=0}.$$

This concludes the proof of (67).

Putting everything together:

We will now conclude the proof of (71). Plugging in (80) and (81) into (75) gives

$$\begin{aligned}
\frac{d}{dt} f(V(tR), B, C)\Big|_{t=0} =& 2\,\mathbb{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\,G_k(U_\Sigma X_0, 0)^\top \mathbb{K}\,\frac{d}{dt} G_k(U_\Sigma X_0, tR)\Big|_{t=0}(I-M)\right)\right] \\
=& 2\,\mathbb{E}_{X_0,U}\left[\mathrm{Tr}\left((I-M)\,G_k(X_0, 0)^\top \mathbb{K}\,\frac{d}{dt} G_k(X_0, tU^\top R U_\Sigma)\Big|_{t=0}(I-M)\right)\right] \\
=& 2\,\mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G_k(X_0, 0)^\top \mathbb{K}\,\frac{d}{dt} G_k(X_0, t\,\mathbb{E}_U\left[U^\top R U_\Sigma\right])\Big|_{t=0}(I-M)\right)\right] \\
=& 2\,\mathbb{E}_{X_0}\left[\mathrm{Tr}\left((I-M)\,G_k(X_0, 0)^\top \mathbb{K}\,\frac{d}{dt} G_k(X_0, t\tilde{R})\Big|_{t=0}(I-M)\right)\right] \\
=& \frac{d}{dt} f(V(t\tilde{R}), B, C)\Big|_{t=0}
\end{aligned}$$

The third equality uses the fact that $\frac{d}{dt} G_k(X_0, tS)\big|_{t=0}$ is linear in S for any S. The fourth equality is by (70). This concludes the proof of (71). □

□

Lemma F.5. *Let $S(t), T(t), \tilde{S}, \tilde{T} : \mathbb{R} \to \Gamma \in \mathbb{R}^{d\times d}$, denote arbitrary continuously differentiable, matrix-valued, functions of time. Assume that $S(0) = \tilde{S}(0)$, $T(0) = \tilde{T}(0)$, $S'(0) = \tilde{S}'(0)$ and $T'(0) = \tilde{T}'(0)$ (i.e. have the same time derivative at $t = 0$). Let $\Gamma \in \mathbb{R}^{d\times d}$ be an arbitrary invertible matrix. Then for any $\tilde{h}$ satisfying Assumption 4.3,*

$$\frac{d}{dt}\tilde{h}\left(S(t), T(t)\right)\Big|_{t=0} = \frac{d}{dt}\tilde{h}\left(\Gamma^\top \tilde{S}(t), \Gamma^{-1}\tilde{T}(t)\right)\Big|_{t=0}$$

Proof. Let ${\mathcal{J}_{\tilde{h}}}^1$ and ${\mathcal{J}_{\tilde{h}}}^2$ denote the Jacobians of $\tilde{h}(A, B)[\cdot]$ with respect to A and B respectively. Then

$$\begin{aligned}
&\frac{d}{dt}\tilde{h}\left(S(t), T(t)\right)\Big|_{t=0} \\
=& {\mathcal{J}_{\tilde{h}}}^1\left(S(0), T(0)\right)\left[S'(0)\right] + {\mathcal{J}_{\tilde{h}}}^2\left(S(0), T(0)\right)\left[T'(0)\right] \\
=& {\mathcal{J}_{\tilde{h}}}^1\left(\tilde{S}(0), \tilde{T}(0)\right)\left[\tilde{S}'(0)\right] + {\mathcal{J}_{\tilde{h}}}^2\left(\tilde{S}(0), \tilde{T}(0)\right)\left[\tilde{T}'(0)\right] \\
=& {\mathcal{J}_{\tilde{h}}}^1\left(\tilde{S}(0), \tilde{T}(0)\right)\left[\Gamma^\top \tilde{S}'(0)\right] + {\mathcal{J}_{\tilde{h}}}^2\left(\tilde{S}(0), \tilde{T}(0)\right)\left[\Gamma^{-1}\tilde{T}'(0)\right] \\
=& \frac{d}{dt}\tilde{h}\left(\Gamma^\top \tilde{S}(t), \Gamma^{-1}\tilde{T}(t)\right)\Big|_{t=0}.
\end{aligned}$$

The third equality follows from Assumption 4.3. □

G. Background on RKHS

In this section we introduce a number of results from RKHS literature which we use in several proofs.

Theorem G.1 ((Wainwright, 2019) Theorem 12.11, Kernel Reproducing Property)**.** *Given any positive semidefinite kernel function $\mathcal{K}$, defined on the cartesian product space $\mathcal{X} \times \mathcal{X}$, there is a unique Hilbert space $\mathbb{H}$ in which the kernel satisfies the reproducing property: for any $x \in \mathcal{X}$, the function $\mathcal{K}(\cdot, x)$ belongs to $\mathbb{H}$, and satisfies the relation*

$$\langle f, \mathcal{K}(\cdot, x)\rangle_{\mathbb{H}} = f(x)$$

Theorem G.2 ((Schölkopf et al., 2001) Theorem 1, Nonparametric Representer Theorem)**.** *Given $\mathcal{X}$, positive semi-definite kernel $\mathcal{K}$ over $\mathcal{X} \times \mathcal{X}$, a set of m training samples $(x_1, y_1) \ldots (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$, a strictly monotonically increasing real-valued function g on $[0, \infty]$, an arbitrary cost function $c : (\mathcal{X} \times \mathbb{R}^2)^m \to \mathbb{R}$. Then any $f \in \mathbb{H}$ minimizing the regularized risk functional*

$$c\left((x_1, y_1, f(x_1)) \ldots (x_m, y_m, f(x_m))\right) + g(\|f\|_{\mathbb{H}})$$

admits a representation of the form

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i \mathcal{K}(\cdot, x_i)$$

Finally, we establish the explicit form of steepest descent in Hilbert space. The proof is quite standard (see e.g. Martin's 241B lecture 6), and we include it for completeness.

Lemma G.3 (Steepest Descent in Hilbert Space)**.** *Given any $f \in \mathbb{H}$, let g^* denote the steepest descent direction of the weighted empirical least-squares loss wrt $\|\cdot\|_{\mathbb{H}}$, i.e.*

$$g^* := \underset{g \in \mathbb{H}, \|g\|_{\mathbb{H}}=1}{\arg\min} \frac{d}{dt} \sum_{i=1}^{n} \left(y^{(i)} - (f + tg)(x^{(i)}) \right)^2 \Bigg|_{t=0}.$$

Then $g^*(\cdot) = c \sum_{i=1}^{n} \left(y^{(i)} - f(x^{(i)}) \right) \mathcal{K}(\cdot, x^{(i)})$ for some scalar $c \in \mathbb{R}^+$ (we give explicit expression for c in the proof).

Proof. Using the method of Lagrangian multipliers, there exists some λ for which the above is equivalent to

$$\begin{aligned} g^* &= \arg\min_{g \in \mathbb{H}} \frac{d}{dt} \sum_{i=1}^{n} \left(y^{(i)} - (f + tg)(x^{(i)}) \right)^2 + \lambda \|g\|_{\mathbb{H}}^2 \Bigg|_{t=0} \\ &= \arg\min_{g \in \mathbb{H}} \sum_{i=1}^{n} -2y^{(i)} g(x^{(i)}) + \lambda \|g\|_{\mathbb{H}}^2 . \end{aligned}$$

The second line is by simple algebra.

Applying Theorem G.2 with $f := g$ and $g(r) := \frac{\lambda}{2} r^2$, we know that $g^*(\cdot) = \sum_{i=1}^{n} \alpha_i \mathcal{K}(\cdot, x^{(i)})$, for some $\alpha \in \mathbb{R}^n$. Using this together with Theorem G.1, we can write

$$\|g\|_{\mathbb{H}}^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j \left\langle \mathcal{K}(\cdot, x^{(i)}), \mathcal{K}(\cdot, x^{(j)}) \right\rangle_{\mathbb{H}} = \sum_{i,j=1}^{n} \alpha_i \alpha_j \mathcal{K}(x^{(i)}, x^{(j)}).$$

Thus $g^*(\cdot) = \sum_{i=1}^{n} \alpha^* \mathcal{K}(\cdot, x^{(i)})$. Let $Y, F \in \mathbb{R}^n$ be defined such that $Y_i = y^{(i)}$ and $F_i = f(x^{(i)})$. Then

$$\begin{aligned} \alpha^* &= \arg\min_{\alpha \in \mathbb{R}^n} \sum_{i,j=1}^{n} -2(y^{(i)} - f(x^{(i)}))\alpha_j \mathcal{K}(x^{(i)}, x^{(j)}) + \lambda \alpha_i \alpha_j \mathcal{K}(x^{(i)}, x^{(j)}) \\ &= \arg\min_{\alpha \in \mathbb{R}^n} -2 \left((Y - F) \right)^\top \mathbb{K}\alpha + \lambda \alpha^\top \mathbb{K}\alpha. \end{aligned}$$

Taking $\nabla_\alpha = 0$, we get $\alpha^* \propto (Y - F)$. Recall that our original constraint is $\alpha^\top \mathbb{K}\alpha = 1$, it follows that

$$\alpha^* = \frac{1}{(Y - F)^\top \mathbb{K} (Y - F)} (Y - F).$$

This implies that $g^*(\cdot) = \frac{1}{(Y-F)^\top \mathbb{K}(Y-F)} \sum_{i=1}^{n} \left(y^{(i)} - f(x^{(i)}) \right) \mathcal{K}\left(\cdot, x^{(i)}\right)$, which concludes our proof.

Note that the choice of α^* is in fact not unique when $rank(\mathbb{K}) < n$. To see this, let $b \in \mathbb{R}^n$ be such that $b \neq 0, \mathbb{K}b = 0$. Then by the same argument above,

$$\left\| \sum_{i=1}^{n} b_i \mathcal{K}(\cdot, x^{(i)}) \right\|_{\mathbb{H}}^2 = b^\top \mathbb{K} b = 0,$$

so that $\sum_{i=1}^{n} \left[\alpha^* + b\right]_i \mathcal{K}(\cdot, x^{(i)}) = \sum_{i=1}^{n} \left[\alpha^*\right]_i \mathcal{K}(\cdot, x^{(i)})$, where equality is in the sense of $\|\cdot\|_{\mathbb{H}}$. □

For intuition, let us consider the reduction of Lemma G.3 to the linear regression setting. Let $\theta(t) : \mathbb{R} \to \mathbb{R}^d$ denote the gradient flow of the parameter θ with respect to the empirical least-squares loss $\sum_{i=1}^n \left(y^{(i)} - \left\langle \theta(t), x^{(i)} \right\rangle\right)^2$ in Euclidean norm. It follows that

$$\begin{aligned}\frac{d}{dt}\theta(t) =& 2\sum_{i=1}^n (y^{(i)} - \left\langle \theta(t), x^{(i)} \right\rangle) x^{(i)} \\ =& 2\sum_{i=1}^n X(Y - X^\top \theta(t)).\end{aligned}$$

Now let $f_t(x) := \langle x, \theta(t)\rangle$, and let F denote the vector with $F_i = f_t(x^{(i)}) = \left[X^\top \theta(t)\right]_i$. Then $\frac{d}{dt} f_t(x) = -2\sum_{i=1}^n [Y - F]_i \mathcal{K}(\cdot, x^{(i)})$, which is equal to the direction in Lemma G.3.

H. Experiments

H.1. Experiment Details

The following are common to all experiments in this paper:

We train the Transformer to minimize the in-context loss given in (5).

Covariate Distribution
The covariates $x^{(i)} = \Sigma^{1/2}\xi^{(i)}$, where $\xi^{(i)}$ are sampled iid from the unit sphere. The dimension is $d = 5$. The covariance matrix $\Sigma = U^T D U$, where U is a uniformly random orthogonal matrix that changes across seeds, and D is a fixed diagonal matrix with entries $(1, 1, 0.25, 2.25, 1)$.

Label Distribution
Conditioned on $x^{(i)}$'s, the labels $y^{(i)}$ are jointly sampled from the $\mathcal{K}$ Gaussian Process (see Definition 3.3). We consider three choices of kernels: $\mathcal{K}^{linear}(u, v) = \langle u, v\rangle$, $\mathcal{K}^{relu}(u, v) = \mathsf{relu}(\langle u, v\rangle)$, and $\mathcal{K}^{exp}(u, v) = \exp(\langle u, v\rangle)$ (as defined (11)).

Transformer Architecture
Unless otherwise stated, we train a three-layer linear Transformer (see (3)), where the matrices are initialized by i.i.d. Gaussian matrices. We consider three different choices of nonlinearity $\tilde{h}$: linear, ReLU and softmax, defined in (12) (see also Examples 1, 2 and 3). The Transformer is parameterized by $(r_\ell, A_\ell, B_\ell, C_\ell)_{\ell=0,1,2}$. (the value matrix V_ℓ is parameterized by the A_ℓ, r_ℓ, see Assumption 4.4).

Training Algorithm
We train the Transformer using ADAM with gradient clipping. Each gradient step is computed from a minibatch of size 30000, and we resample the minibatch every 10 steps. All plots are averaged over 3 runs with different U (i.e. Σ) sampled each time, and different seeds for sampling training data.

H.2. Experiment for Theorem 4.5

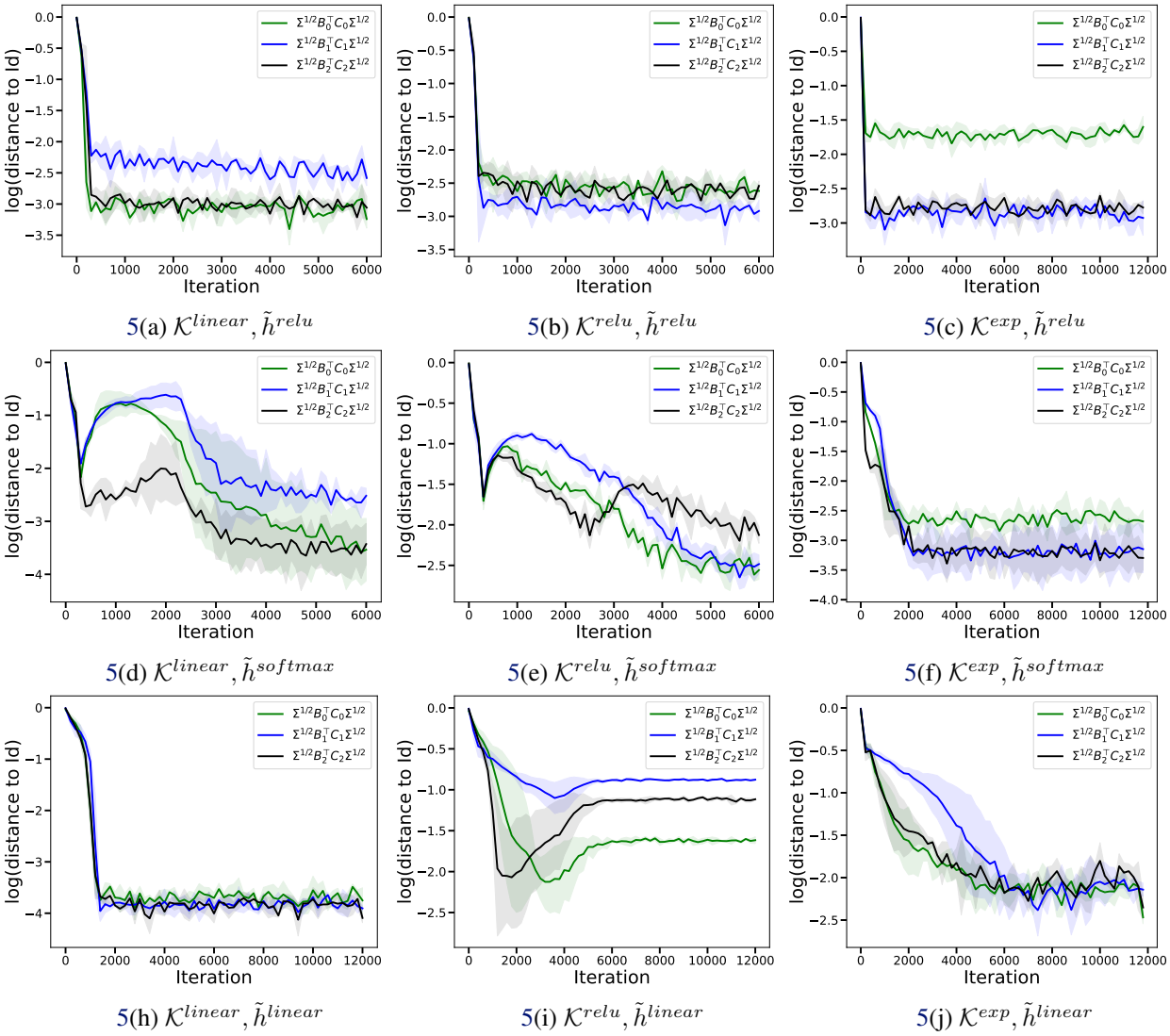
In Figure 5 below, we present empirical verification of Theorem 4.5. In addition to the setup in Appendix H.1, we additionally constrain $A_\ell = 0$ for each layer ℓ. The number of demonstrations $n = 30$.

To verify that the parameters are indeed converging to the predicted stationary point in Theorem 4.5, we plot $\mathsf{dist}\left(\Sigma^{1/2} B_i^\top C_i \Sigma^{1/2}, I\right)$, for $i = 0, 1, 2$. The *normalized Frobenius norm distance*: $\mathsf{dist}(M, I) := \min_\alpha \frac{\|M - \alpha \cdot I\|}{\|M\|_F}$, (equivalent to choosing $\alpha := \frac{1}{d}\sum_{i=1}^d M[i, i]$). This is essentially the projection distance of $M/\|M\|_F$ onto the space of scaled identity matrices.

We only verify $B^\top C$ because the network is overparameterized, and for any $\Lambda \in \mathbb{R}^{d\times d}$, (B_ℓ, C_ℓ) gives identical prediction as $(\Lambda^\top B_\ell, \Lambda^{-1} C_\ell)$. (See also Remark E.2 after Theorem E.1). It appears that in most cases, the matrices are converging to identity, which is the stationary point in Theorem 4.6. This demonstrates that Theorem E.1 holds across a broad combination of $\mathcal{K}$ and $\tilde{h}$.

We note that in the case of Figure 5(c) and 5(i), a few of the parameter matrices appear to asymptote at around 0.2 distance to identity. It is unclear if this is due to optimization difficulties, or due to convergence to stationary points different from

that proposed in Theorem 4.5.



5(a) $\mathcal{K}^{linear}, \tilde{h}^{relu}$ 5(b) $\mathcal{K}^{relu}, \tilde{h}^{relu}$ 5(c) $\mathcal{K}^{exp}, \tilde{h}^{relu}$

5(d) $\mathcal{K}^{linear}, \tilde{h}^{softmax}$ 5(e) $\mathcal{K}^{relu}, \tilde{h}^{softmax}$ 5(f) $\mathcal{K}^{exp}, \tilde{h}^{softmax}$

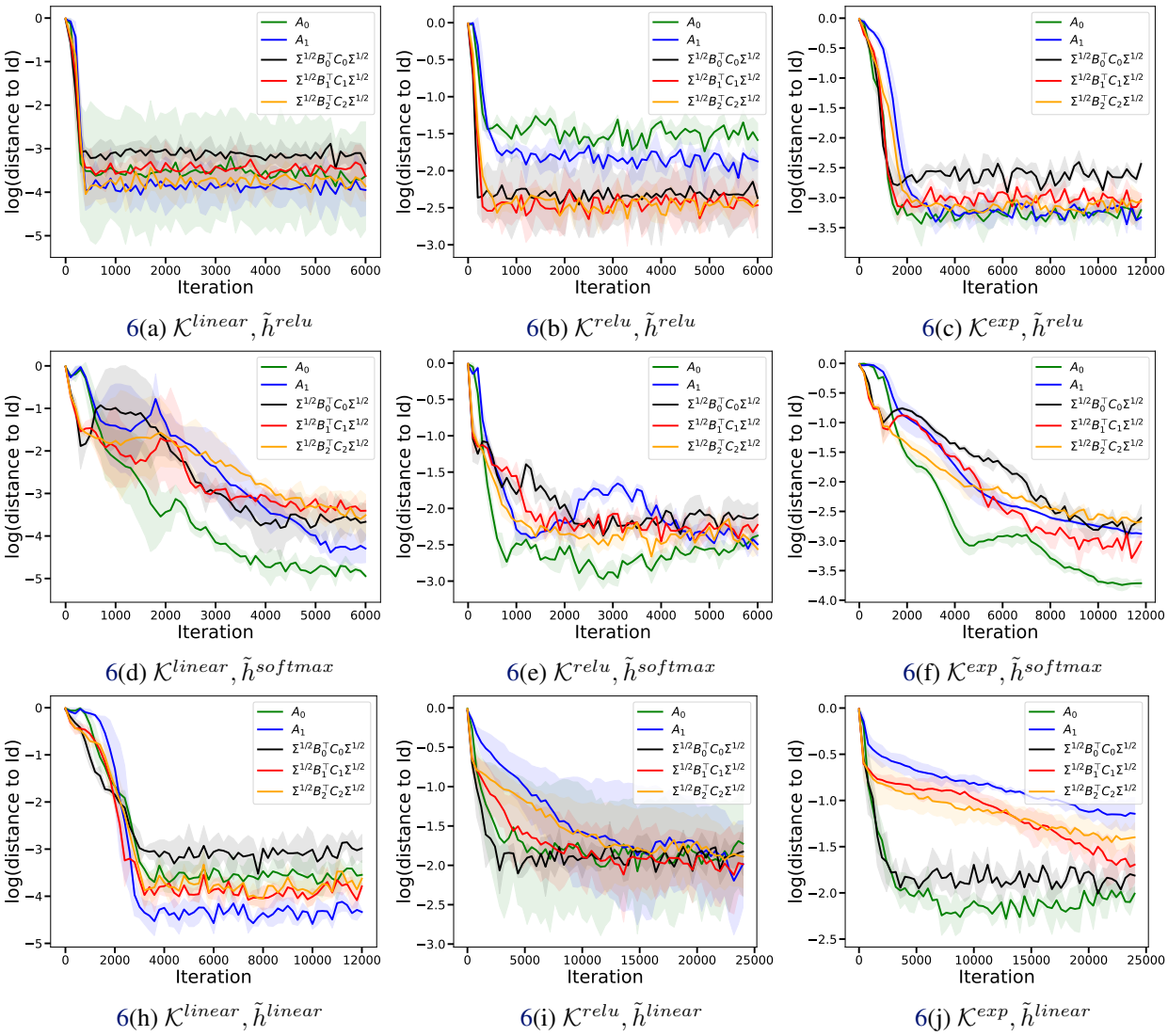
5(h) $\mathcal{K}^{linear}, \tilde{h}^{linear}$ 5(i) $\mathcal{K}^{relu}, \tilde{h}^{linear}$ 5(j) $\mathcal{K}^{exp}, \tilde{h}^{linear}$

Figure 5. Plots of $\log(\mathsf{dist}(M, I))$ for $M = \Sigma^{1/2}\left\{B_0^\top C_0, B_1^\top C_1, B_2^\top C_2\right\}\Sigma^{1/2}$ against number of training iterations. Each plot coincides with a different experiment setup, where we vary the generating distribution and the architecture. The subplot title is $(\mathcal{K}, \tilde{h})$, where $\mathcal{K}$ defines a Gaussian Process for labels, as described in Definition 3.3, and $\tilde{h}$ is the non-linear map in the Transformer's attention module. In all cases, the corresponding matrix appears to be converging to identity, which is the stationary point from Theorem 4.5.

H.3. Experiments for Theorem 4.6

In this section, we present empirical verification of Theorem 4.6. The experiment setup is as described in Appendix H.1. The number of demonstrations $n = 30$. The metric for measuring distance to identity is same as described in Section H.2, i.e. $\mathsf{dist}(M, I) := \min_\alpha \frac{\|M - \alpha \cdot I\|}{\|M\|_F}$.

Similar to Appendix H.2, it appears that in most cases the matrices are converging to identity, which is the stationary point in Theorem 4.6. We note that in the case of Figure 6(b), a few of the parameter matrices appear to asymptote at around 0.2 distance to identity. It is unclear if this is due to optimization difficulties, or due to convergence to stationary points different from that proposed in Theorem 4.6.



6(a) $\mathcal{K}^{linear}, \tilde{h}^{relu}$ 6(b) $\mathcal{K}^{relu}, \tilde{h}^{relu}$ 6(c) $\mathcal{K}^{exp}, \tilde{h}^{relu}$

6(d) $\mathcal{K}^{linear}, \tilde{h}^{softmax}$ 6(e) $\mathcal{K}^{relu}, \tilde{h}^{softmax}$ 6(f) $\mathcal{K}^{exp}, \tilde{h}^{softmax}$

6(h) $\mathcal{K}^{linear}, \tilde{h}^{linear}$ 6(i) $\mathcal{K}^{relu}, \tilde{h}^{linear}$ 6(j) $\mathcal{K}^{exp}, \tilde{h}^{linear}$

Figure 6. Plots of $\log(\mathsf{dist}(M, I))$ for $M \in \{A_0, A_1\} \cup \left\{\Sigma^{1/2}B_i^\top C_i\Sigma^{1/2}\right\}_{i=0,1,2}$ against number of training iterations. Each plot coincides with a different experiment setup, where we vary the generating distribution and the architecture. The subplot title is $(\mathcal{K}, \tilde{h})$, where $\mathcal{K}$ defines a Gaussian Process for labels, as described in Definition 3.3, and $\tilde{h}$ is the non-linear map in the Transformer's attention module. The definitions of each $\mathcal{K}$ and $\tilde{h}$ can be found in (11) and (12) respectively. In all cases, the corresponding matrix appears to be converging to identity, which is the stationary point from Theorem 4.6.

I. Miscellaneous Proofs

I.1. Verification of Example 8

Following the distributional assumption on θ_1, θ_2, we verify that

$$
\begin{aligned}
&[\mathbb{K}(X)]_{ij} := \mathbb{E}\left[y^{(i)}y^{(j)}\right] \\
&= \mathbb{E}_{\theta_1,\theta_2}\left[\mathsf{relu}\left(\theta_1 x^{(i)}\right)^\top \theta_2\theta_2^\top \mathsf{relu}\left(\theta_1 x^{(j)}\right)\right] \\
&= \mathbb{E}_{\theta_1}\left[\mathsf{relu}\left(\theta_1 x^{(i)}\right)^\top \mathsf{relu}\left(\theta_1 x^{(j)}\right)\right].
\end{aligned}
$$

Similarly, we verify that $[\mathbb{K}(UX)]_{ij} - \mathbb{E}_{\theta_1}\left[\mathsf{relu}\left(\theta_1 U x^{(i)}\right)^\top \mathsf{relu}\left(\theta_1 U x^{(j)}\right)\right] = \mathbb{E}_{\theta_1}\left[\mathsf{relu}\left(\theta_1 x^{(i)}\right)^\top \mathsf{relu}\left(\theta_1 x^{(j)}\right)\right]$, because $\theta_1 U \overset{d}{=} \theta_1$.

I.2. Functional Gradient Descent for Euclidean Inner Product Kernel

Let $\theta \in \mathbb{R}^d$ be the linear regression parameter to learn. Let $R(\theta) := \frac{1}{2}\sum_{i=1}^{n}\left(\left\langle x^{(i)}, \theta\right\rangle - y^{(i)}\right)^2$ denote the empirical least squares loss. Let θ_k denote the k^{th} iterate of gradient descent with stepsize r'_k, thus

$$\theta_{k+1} = \theta_k - r'_k \nabla R(\theta_k).$$

Let $f_k(x) := \langle \theta_k, x\rangle$. Notice that $\nabla R(\theta_k) = -\sum_{i=1}^{n}\left(y^{(i)} - f_k(x^{(i)})\right)x^{(i)}$, so that

$$f_{k+1}(x) = f_k(x) + r'_k \sum_{i=1}^{n}\left(y^{(i)} - f_k(x^{(i)})\right)\left\langle x^{(i)}, x\right\rangle$$

which is exactly the same as (7) (or more specifically, (9)), by noting that $\left\langle x^{(i)}, x\right\rangle = \mathcal{K}(x^{(i)}, x)$.